T.R.
İSTANBUL UNIVERSITY
INSTITUTE OF SCIENCES

MASTER'S THESIS

A WHOLE EXOME SEQUENCING DATA ANALYSIS PIPELINE: AN APPLICATION ON PITUITARY ADENOMA DISEASE

Aristide IRAKOZE

Department of Informatics

Programme of Informatics

ADVISOR
Prof. Dr. Çiğdem EROL

II. ADVISOR
Assoc. Prof.Dr. Feyza NUR TUNCER

Haziran, 2024

İSTANBUL

This work is, accepted as a Master's Thesis in **Department of Informatics** Department **Programme of Informatics** Program by the following jury in 13.06.2024

Thesis Jury

Prof. Dr. Çiğdem EROL(Advisor)
Istanbul University
Faculty of Science

Prof. Dr. Sevinç GÜLSEÇENI
İstanbul University
Informatics Department

Assoc.Prof. Yalçın ÖZKAN
Istinye University
Faculty of Economics, Administrative and Social Sciences

Ethic Declaration

As required by the 9/2 and 22/2 articles of the Graduate Education Regulation which was published in the Official Gazette on 20.04.2016, this graduate thesis is reported as in accordance with criteria determined by the Institute of Sciences by using the plagiarism software to which Istanbul University is a subscriber

Research Fund

This thesis is supported by the project numbered …….of İstanbul University Scientific Research Projects Executive Secreteriat.

This thesis supported by the project numbered …….of …….

Publications Produced From Thesis

PREFACE

I would like to express my gratitude to the Almighty for allowing me to reach this stage in my life, for instilling in me a love for science, and for enabling me to realize one of my dreams. I am also grateful for connecting me with the individuals mentioned below and for preserving them.

First of all, I am deeply thankful to my supervisor, Prof. Dr. Çiğdem EROL, without whom this work would never have started or concluded. Her continuous support and wise guidance have helped me find the courage and strength necessary to confront and overcome the obstacles encountered during this endeavor. May God bless you infinitely.

I express my appreciation to my second supervisor Assoc. Prof. Dr. Feyza NUR TUNCER for her immense contribution to the completion of this work at key stages, despite her numerous commitments. Her vast experience and wisdom have guided us along the way.

I extend my gratitude to the TUBITAK ULAKBIM Institute for providing high-performance computing resources to students and academics via the TRUBA platform. Additionally, I acknowledge the support from M.Sc. Sevcan AYDIN for her valuable insights about domain knowledge and data handling, within the DTAE laboratory. I am profoundly appreciative of their contributions.

I would also like to extend my gratitude to the entire Department of Informatics at Istanbul University. I am grateful to Prof. Dr. Sevinç GÜLSEÇEN, the Department Chair for her invaluable mentorship, as well as all the teaching staff especially Dr. Murat GEZER for educating and preparing us during our academic journey.

I dedicate this work to my family, especially my dear mother, and all the individuals for their immense support throughout the entire process and for their prayers.

Haziran 2024 Aristide IRAKOZE

TABLE OF CONTENTS

Page Number

PREFACE ..... **iv**
TABLE OF CONTENTS ..... **v**
LIST OF FIGURES ..... **vii**
LIST OF TABLES ..... **ix**
ABSTRACT ..... **xi**
ÖZET ..... **xiii**
1. INTRODUCTION **1**
2. GENERAL PARTS **3**
2.1. PITUITARY ADENOMAS 3
2.1.1. Prevalence and Classification 3
2.1.2. Genetics and Heredity 4
2.2. BIG DATA IN GENOMICS 5
2.3. WES DATA MINING 8
2.3.1. Genomic variants 8
2.3.2. Variant calling from WES Databases 10
3. MATERIAL AND METHODS **15**
3.1. DATA DESCRIPTION 15
3.1.1. Whole Exome Sequencing Data 15
3.1.2. Additional resources 16
3.2. VARIANT CALLING PIPELINE 17
3.2.1. Data preprocessing 17
3.2.2. Reads mapping to a reference genome 18
3.2.3. Mapping quality and files processing 19
3.2.4. Variant calling 21
3.2.5. Pattern Evaluation and Validation 25
3.3. HARDWARE 26
4. RESULTS **28**
5. DISCUSSION AND CONCLUSION **44**
5.1. OVERLAPPING GENES AND MUTATIONS 44
5.2. CONSOLIDATION STRATEGIES 45

5.3. RECURRENCE OF GENOMIC MUTATIONS 46
5.4. LIMITATIONS 47
5.5. CONCLUSION 47
REFERENCES 49
CURRICULUM VITAE 59

LIST OF FIGURES

Figure 2.1. Statistics from the Genome Sequence Archive(GSA) and GSA-Human 6

Figure 2.2. Genomic Mutation forms ("Human Genomic Variation," n.d.) 8

Figure 2.3. The locations and consequences of a genetic mutation on the resulting protein (Bartha & Győrffy., 2019) 9

Figure 2.4. An overview of the KDD process's steps. 10

Figure 2.5. A simplified germline variant calling pipeline from WES datasets 11

Figure 2.6. Example of BWT Transform 13

Figure 3.1. A segment of the first read in the S48 sample 15

Figure 3.2. Number of reads in each sample (the first pairs) 16

Figure 3.3. Sample quality score average per base position 17

Figure 3.4. Percentage (blue color) and distribution (red color) of duplication count 17

Figure 3.5. Percentage of reads covering targets 19

Figure 3.6. Average mapping quality 20

Figure 3.7. Percentage of uniquely aligned reads 20

Figure 3.8. Number of chimeric reads per Sample 20

Figure 3.9. DeepVariant workflow overview (Poplin et al., 2018) 22

Figure 3.10. Haplotypecaller in four major steps 22

Figure 3.11. Freebayes, haplotype-based variant detection 23

Figure 3.12. Overview of the Pipeline Workflow 24

Figure 3.13. An example of a VCF file. 25

Figure 4.1. Distribution of 123 germline variants across the Samples 28

Figure 4.2. Hierarchical clustering and illustration of pairwise similarity levels 37

Figure 4.3. Mappers' contribution (before prioritization) 38

Figure 4.4. Variant callers' contributions before prioritization 38

Figure 4.5. Number of common P variants among variant callers 40

Figure 4.6. The overall number of common variants among variant callers.....40

Figure 4.7. IGV view of variant 15-42691799-G-A on CAPN3 gene.....41

Figure 4.8. IGV view of the variant 16-3293407-T-C on MEFV gene 42

Figure 4.9. IGV view of the variant 17-5071419-G-A on USP6 gene.....43

LIST OF TABLES

Table 2.1. Comparison between SW and NW algorithms (Parvez et al., 2020) 12

Table 3.1. Phred scores examples and description 15

Table 3.2. List of additional data resources 16

Table 3.3. Quality control and enhancement tools 18

Table 3.4. Command line examples for reference genome indexing and reads mapping procedures 18

Table 3.5. Software tools for the reads alignment step 19

Table 3.6. Command line examples for mapping files preprocessing using samtools 21

Table 3.7. Filtering threshold values from the GATK variant calling best practices 25

Table 3.8. Set of tools for the pattern evaluation, validation step, and results analysis 27

Table 3.9. Set of tools for results visualization 27

Table 4.4. List of Pathogenic variants 31

Table 4.5. List of LP variants 32

Table 4.6. List of LP variants (part 2) 33

Table 4.7. List of LP variants (final part) 34

Table 4.8. List of recurrent genomic variants 35

LIST OF SYMBOLS AND ABBREVIATIONS

Abbreviations	Explanations
ACMG	: American College of Medical Genetics and Genomics
ACTH	: Adrenocorticotrophic hormone
BSQR	: Base Scores Quality Recalibration
BWT	: Borrow Wheel Transform
CD	: Cushing's disease
CPU	: Central Processing Unit
DTAE	: Aziz Sancar Deneysel Tıp Araştırma Enstitüsü
FİPA	: Familial Isolated Pituitary Adenoma
FPA	: Functioning Pituitary Adenomas
FSH	: Follicle-stimulating hormone
GH	: Growth hormone
HPC	: High Performance Computing
IGV	: Integrated Genomics Viewer
KDD	: Knowledge Discovery in Databases
LH	: Luteinizing hormone
LP	: Likely Pathogenic
MAF	: Minor Allele Frequency
MEN1	: Multiple Endocrine Neoplasia Type
NMD	: Nonsense-mediated mRNA Decay
P	: Pathogenic
PA	: Pituitary Adenoma
PRL	: Prolactinomas
RAM	: Random Access Mamory
RLE	: Run-Length Encoding
SNV	: Single nucleotide variation
TRUBA	: Türk Ulusal Bilim e-Altyapısı
TSH	: Thyroid stimulating hormone
UTR	: Untranslated Region
VCF	: variant call format
VUS	: Variant of Uncertain Significance
WES	: Whole Exome Sequencing

ABSTRACT

THESIS OF MASTER

AWHOLE EXOME SEQUENCİNG DATA ANALYSİS PİPELİNE: AN APPLİCATİON ON PİTUİTARY ADENOMA DİSEASE

Aristide IRAKOZE

İstanbul University

Institute of Science

Department of Informatics

ADVISOR
Prof. Dr. Çiğdem EROL

II. ADVISOR
Assoc. Prof.Dr. Feyza NUR TUNCER

Pituitary adenomas (PAs) are benign tumors of the pituitary gland. Data science techniques are indispensable for analyzing genomic big datasets and understanding the mechanisms underlying complex diseases. We developed a modular pipeline for whole exome sequencing (WES) data mining and assessed its efficiency on a WES dataset comprising 18 pituitary adenomas (PAs) samples. The analysis identified 123 mutations across 104 genes, including notable mutations in genes such as AIP, AHNAK, and SDHB, some of which have been previously associated with PA. By employing a parallel multi-workflow pipeline with two read mappers (Bowtie, BWA-mem) and three variant callers (GATK, Freebayes, DeepVariant), we improved variant detection accuracy, confirming the utility of the approach, and drew attention to the potential risks related to a low concordance among workflows. Despite not finding mutations in some commonly reported PA genes, the study validated two significant PA-related mutations in the AIP gene: c.910C>T and c.241C>T. This work contributes valuable insights into the genetic basis of PAs, suggesting ideas for future investigation and providing recommendations for pipeline design.

Haziran 2024, 73 pages.

Keywords: Pituitary Adenoma, Whole Exome Sequencing, Data Analysis Mining

ÖZET

M.Sc. THESIS

TÜM EKZOM DİZİLEME VERİ ANALİZİ İŞ AKIŞI: HİPOFİZ ADENOMU HASTALIĞI ÜZERİNE BİR UYGULAMA

Aristide IRAKOZE

İstanbul University

Institute of Sciences

Department of Informatics

Supervisor : Prof. Dr. Çiğdem EROL

Co-Supervisor : Assoc. Prof. Dr. Feyza NUR TUNCER

Hipofiz adenomları (PA'lar) hipofiz bezinin iyi huylu tümörleridir. Veri bilimi teknikleri, genomik büyük verisetlerinin analiz edilmesi ve karmaşık hastalıkların altında yatan mekanizmaların anlaşılması için vazgeçilmezdir. Tüm ekzom dizileme (TED) veri madenciliği için modüler bir İŞ AKIŞI geliştirdik ve 18 hipofiz adenomu (PAs) örneğinden oluşan bir WES veri kümesi üzerinde, etkinliğini değerlendirdik. Analiz, bazıları daha önce PA ile ilişkilendirilmiş olan AIP, AHNAK ve SDHB gibi genlerdeki önemli mutasyonlar da dahil olmak üzere 104 gende 123 mutasyon tespit etmiştir. İki okuma hizalama aracı (Bowtie, BWA-mem) ve üç varyant arayıcı (GATK, Freebayes, DeepVariant) içeren çoklu araçlı paralel iş akışı kullanarak varyant tespit doğruluğunu artırdık, mevcut yaklaşımın faydasını teyit ettik ve iş akışları arasında düşük uyumla ilgili potansiyel risklere dikkat çektik. Yaygın olarak bildirilen bazı PA genlerinde mutasyon bulamamasına rağmen, bu tez AIP geninde PA ile ilişkili iki önemli mutasyonu doğrulamıştır: c.910C>T ve c.241C>T. Bu çalışma, PA'ların genetik temeline ilişkin değerli içgörülere katkıda bulunmakta, gelecekteki araştırmalar için fikirler önermekte ve İŞ AKIŞI tasarımı için öneriler sunmaktadır.

June 2024, 73 pages.

Anahtar kelimeler: Hipofiz Adenomu, Tüm Ekzom Dizileme, Veri Madenciliği

1. INTRODUCTION

Big data, as well as medium and small-sized data, are providing insights and significant opportunities for innovation across various domains. Advanced technologies are combined with data sciences in interdisciplinary methodologies to enhance our comprehension of the world. This thesis is located at the crossroads of data science and genomics. We investigate the data mining strategies for genomic Whole Exome Sequencing (WES) data, with a specific focus and application to Pituitary Adenomas (PA) disease.

PA originates in the pituitary gland, a vital organ situated at the base of the brain and responsible for the secretion of essential hormones(Horvath and Kovacs, 1988). These tumors lead to health complications due to hormonal imbalances and other disturbances (Aflorei & Korbonits, 2014). They represent a medical challenge because of their clinical diversity (Ho et al., 2024), complex diagnostic procedures, prevalence in the populations (Ezzat et al., 2004), etc.

Genomics and other omic sciences contribute to the understanding of the low-level mechanisms underlying disease occurrence and behavior (Khoury and Holt, 2021). However, genomic data analysis comes with several challenges arising from data generation procedures, data structure, volume, etc (Xu, 2020). Approaches such as high-performance computing (HPC), distributed computing systems, etc; coupled with optimized data analysis pipelines, are among the viable solutions to address these challenges.

Several pipelines have been described by the literature, often with specific goals and motivations. For example, *WEP* (D'Antonio et al., 2013) and *JWES* (Ahmed et al., 2021) reduce the requirement for technical expertise in conducting complex genomic analyses. *Escalibur* (Korhonen et al., 2022) provides the flexibility to run in various environments like HPC, cloud, or local environments. A standard WES data analysis pipeline combines essential operations such as quality control, read mapping, and variant calling; using efficient algorithms. Some pipelines restrict the user's control over the analysis process and the latest algorithm might not be necessarily the most efficient. It is imperative to periodically evaluate these continuously evolving algorithms (Barbitoff et al., 2022), and update the resulting data analysis pipelines.

Through an overview of the algorithmic principles underlying efficient and commonest genomic data analysis tools, we develop a parallel multi-workflow pipeline with two read mappers (Bowtie, BWA-mem) and three variant callers (GATK, Freebayes, DeepVariant). We assess its efficiency and generate useful insights, through the data mining of a WES dataset comprising 18 pituitary adenomas (PAs) samples, obtained from the Aziz Sancar DTAE Institute.

This thesis is divided into four main sections. The first, GENERAL PARTS, explores domain knowledge about Pituitary Adenomas disease, whole exome sequencing, and data mining aspect of genomic datasets. The MATERIAL AND METHODS section provides the data description and the data analysis pipeline workflows. The following chapter, RESULTS details the insights generated about the PA disease and the performance of the approaches included in the pipeline. The last section, titled DISCUSSIONS, analyzes the pipeline structure and efficiency within the context of existing literature and the results based on current scientific knowledge. The report ends with a general conclusion.

Despite not finding mutations in some commonly reported PA genes, the study validated two significant PA-related mutations in the AIP gene: c.910C>T and c.241C>T, and identified new candidate germline mutations such as c.2080A>G on the MEFV gene. Genes previously linked with PA in literature (e.g. AIP, RYR1, etc) were reaffirmed, and a comparative evaluation of algorithmic tools was provided. This interdisciplinary research contributes to the current knowledge about PA disease and showcases the power of data-driven approaches in unlocking the secrets of complex diseases. Finally, we propose ideas for future investigation and offer recommendations for pipeline design.

2. GENERAL PARTS

This chapter explores the Pituitary Adenomas disease, the big data aspect of genomic datasets, and the algorithmic approaches involved in data mining and pipeline design. WES data mining requires a deep understanding of domain knowledge and a prior comprehension of the disease aspects such as heredity, to make informed decisions.

2.1. PITUITARY ADENOMAS

The pituitary gland is a small gland located at the base of the brain, in line with the top of the nose (Daniel, 1976). It is responsible for producing a variety of essential hormones that control growth, metabolism, blood pressure, and reproductive functions such as Growth hormone (GH); Thyroid stimulating hormone (TSH); Prolactin which controls breast milk production; the Follicle-stimulating hormone (FSH) involved in the reproductive system; and ACTH or Adrenocorticotrophic hormone (Horvath and Kovacs, 1988).

The pituitary gland remotely controls the secretion of oxytocin a hormone crucially involved in childbirth and breastfeeding, and vasopressin regulating the body's balance of salt and water (Barkhoudarian and Kelly, 2017).

Pituitary Adenomas (PAs) are noncancerous tumors that develop in the Pituitary gland. By disrupting its normal function, they result in hormonal imbalances and additional symptoms such as headaches, Seizures, nausea, vomiting, and vision problems due to the pressure on the nearby nerves and organs(Aflorei and Korbonits, 2014).

A subset of adenomas may remain asymptomatic, only being incidentally discovered (Aflorei & Korbonits, 2014) while others are the main reason behind diseases such as acromegaly and CD (Cushing’s disease), hyperthyroidism, etc (Horvath & Kovacs, 1988).

2.1.1. Prevalence and Classification

Known as **Non-Functional Pituitary Adenoma** (NFPA), some tumors do not affect hormonal production and present unpleasant symptoms mostly due to the pressure on the surrounding organs (Chanson and Brochier, 2005). **Functioning Pituitary Adenomas** (FPA) are

characterized by their impact on hormonal production, and classified according to the affected hormone (Mehta and Lonser, 2016). Here are some regularly cited FPA cases:

- **Prolactinomas** or PRL-PA, related to prolactin hormone, account for up to 40% of all clinically diagnosed pituitary adenomas and they result in oligomenorrhea/amenorrhea, galactorrhea, sexual dysfunction, gynecomastia, etc (Yatavelli and Bhusal, 2024);
- **Growth hormone-secreting PAs** (GH-PA): Due to hormone overproduction, they manifest symptoms such as gigantism and acromegaly (Agarwal, 2013). A significant fraction of acromegaly cases (over 95%) are linked to GH-PA ("Acromegaly (Growth Hormone–Secreting Pituitary Adenomas) | Memorial Sloan Kettering Cancer Center," n.d.);
- **Thyroid-stimulating hormone-secreting Pas (**TSH-PA): Among the most uncommon PA tumors, studies report that they account for less than 2% of recorded cases (Byun et al., 2020).
- **LH/FSH Luteinizing hormone (LH) and follicle-stimulating hormone (FSH):** Pituitary tumors that produce these hormones are called gonadotroph adenomas (Mehta and Lonser, 2016).

PAs are also classified based on the size of the tumor, defining three groups: the smallest tumors (less than 10 mm) in diameter, known as microadenomas; giant adenomas with a diameter greater than 40 mm; and those in between, known as macroadenomas (Raverot et al., 2014).

PAs constitute the most prevalent kind of neuroendocrine tumor in the central nervous system and the third most common type of brain tumor following gliomas and meningiomas, they account for 10% to 15% of all brain tumors, (Ezzat et al., 2004). Patients with PAs are examined in the neurosurgery, ophthalmology, and pituitary endocrinology departments.

2.1.2. Genetics and Heredity

WES data analysis for variant calling defines two distinct protocols for inherited and somatic mutations. Studies about the hereditary characteristics of PAs show that these tumors can arise due to mutations that are either inherited (hereditary) or acquired during a person's lifetime (Aflorei and Korbonits, 2014).

MEN1 (Multiple Endocrine Neoplasia Type 1) is a hereditary condition characterized by the simultaneous presence of multiple endocrine neoplasia including pituitary tumors (Trump et al., 1996)**.** PAs may represent between 10 and 60% of patients with MEN1 (Syro et al., 2012).

When a minimum of 2 members of the same family are affected by PA disease without other associated syndromes, the FİPA (Familial Isolated Pituitary Adenoma) terminology can be utilized (Verloes et al., 1999). Aryl Hydrocarbon receptor-interacting protein (AIP) germline mutations have been detected in 15–30% of analyzed FIPA (Cain et al., 2010).

PAs are often associated with pheochromocytomas and paragangliomas (Pheo/PGL) in a combination designated as "the three P association" or "3PAs" (Guerrero-Pérez et al., 2019). Xekouki et al., (2015) described up to 75% of familial cases with **3PAs** harboring pathogenic variants in the **SDHx** (SDHA, SDHB, and SDHD) genes.

Peculis et al., (2022) Detected somatic variants in gene RYR1, which had been reported twice in previous studies. In the same article, additional somatic mutations were discovered in the following genes: PLA2G6, TESK, TSC, KLHL4, POLR3B, KHDRBS2, LAMA1, TUBGCP6, and AHNAK. Various epigenetic and transcriptomic mechanisms were revealed. For example, the inhibitors of the DNA-binding family (ID1, ID2, ID3, and ID4) were all down-regulated in a study by Moreno et al., (2005) on NFPA patients and Seemann et al.,(2001) identified CDKN2A gene methylation changes in a study on 72 pituitary adenomas.

Up to date, genomic variants, differential gene expressions, epigenetic changes, and further mechanisms have been identified as contributing factors to the development of PA syndromes.

2.2. BIG DATA IN GENOMICS

The advent of new-generation sequencing (NGS) technologies enabled significant advancements in various omics sciences, such as genomics. Cost reduction, high speed, and high throughput in sequencing of nucleic acids (Chen et al., 2021) are resulting in the rise of genomics Big Data. Illumina's recent launch of the NovaSeq X Series sequencers represents an illustration, of the potential to reduce the cost of whole genome sequencing to approximately $200 ("NovaSeq X Series Enables Broader, Deeper Sequencing," n.d.).

The latest genomics datasets are described using the 5V characteristics of Big Data i.e.; volume, velocity, variety, veracity, and value (Navarro et al., 2019). Current high-throughput sequencing instruments can reach huge **volumes** of genomic data such as 2 terabytes per run ("Approximate sizes of sequencing run output folders - Illumina Knowledge," n.d.). In 2015, it was believed that genomics may overtake social media, earth sciences, and astronomy fields in terms of data volume if the trend doesn't change (Stephens et al., 2015).

The following figure by (Chen et al., 2021) depicts the evolution of data generation **velocity** from 2016 to 2021 and shows how the volume of submitted data has grown over time:

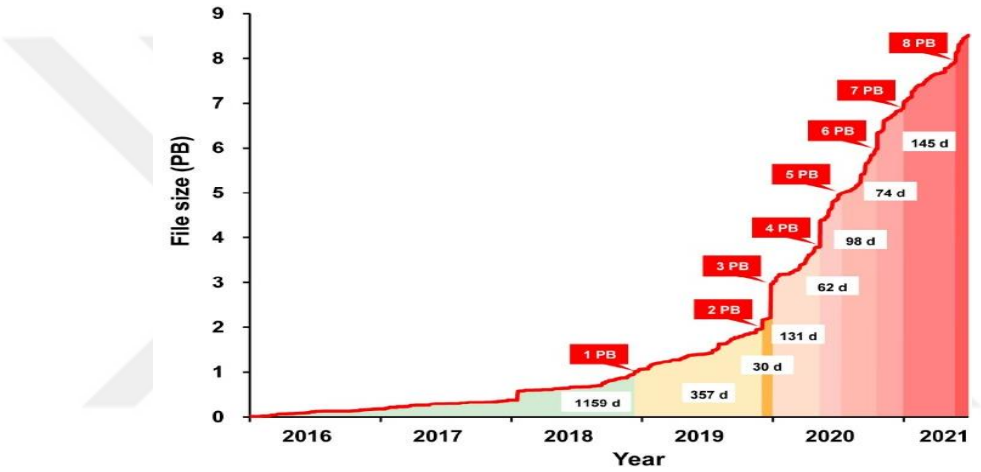


Figure 2.1. Statistics from the Genome Sequence Archive(GSA) and GSA-Human

(Chen et al., 2021)

In terms of **variety**, genomics extends beyond the common sequences of nucleotides and amino acids. An illustrating example is the graphical representation of the human pangenome (Paten et al., 2017). In deep learning-based variant calling, features derived from the alignment of sequence reads to a reference genome are encoded into pileup images, which are then used as input for Convolutional Neural Networks (Poplin et al., 2018). More types of 'omic' studies keep emerging with new data forms and analysis approaches: e.g., Transcriptomics, Epigenomics, Proteomics, Metagenomics**,** etc.

The **value** of genomics in real life relies on its several applications that span healthcare, agriculture, environmental science, and beyond. In WES scenarios, researchers can better understand the genetic basis of diseases, identify potential therapeutic targets, and improve healthcare.

Alongside the aforementioned attributes, genomics data mining comes with unique challenges, starting with data complexity.

Brusic & Zeleznikow (1999) raised four levels of **data complexity** :

- Processes that generate biological data**:** Biological data represent measurements or observations of complex biological systems. Underlying biological processes are highly interconnected and hierarchical and this complexity is usually not encoded in the data structure;
- Fuzziness**:** Biological data are obtained through diverse experimental methods, often with imperfect correlation to the underlying phenomena. Reproducing identical conditions is challenging due to various factors like differences in materials and human decisions, which can significantly impact results;
- Biases: Biological data rarely form a random sample of the entire solution space. This leads to certain regions of the solution space being extensively explored, while others remain largely unexplored;
- Noise and errors**:** Noise in biological data originates from various sources, such as errors in experimentation, measurement, reporting, annotation, and data processing; and it is challenging to estimate error levels in the datasets.

Another challenge is the **computational load**: prediction, clustering, classification and regression analysis, anomaly detection, etc are common tasks in the data mining field. Genomics involves additional tasks such as *de novo assembly*, *reads mapping,* and *variant calling*, often requiring substantial computational resources. For instance, performing a de novo assembly from long-read sequencing of a single 3 GB genome may necessitate approximately 3600 hours of CPU time and consume approximately 870 gigabytes of RAM ("GitHub - Fenderglass/Flye: De Novo Assembler for Single Molecule Sequencing Reads Using Repeat Graphs," n.d.)

An individual's genome can indicate his genetic vulnerabilities to some diseases (Peng et al., 2024), details about his ancestry (Bollas et al., 2024), and even behavioral aspects (Shnayder et al., 2022). This triggers ethical apprehensions regarding **data privacy and security**. In that context, it is essential to emphasize awareness, morality, and the prioritization of privacy in data collection and handling, particularly in the healthcare sector (Erol and Gezer, 2020).

Significant efforts, specialized algorithms, and machine learning techniques are required to extracting valuable insights from genomic datasets. In the next section, we discuss the approaches utilized in WES scenarios.

2.3. WES DATA MINING

WES is a popular genome sequencing approach, targeting the protein-coding regions of the genome to analyze the genetic abnormalities. This region represents only 1-2 percent of the human genome (Warr et al., 2015), and covers approximately 85% of the disease-causing mutations (Choi et al., 2009). Compared to other sequencing methods, WES offers advantages such as reduced sequencing costs and increased depth of coverage.

2.3.1. Genomic variants

The similarity level among two different people's genomes is estimated to be 99.9% (Jorde and Wooding, 2004). Those genetic variations account for the diversity observed in traits such as eye color, skin color, and height. They can also result from deleterious mutations leading to diseases of varying severity.



Figure 2.2. Genomic Mutation forms ("Human Genomic Variation," n.d.)

As illustrated in Figure 2.2, genetic mutations manifest in various forms:

- Single nucleotide variation (SNVs) ;
- Insertion of one or more nucleotides into the DNA sequence;
- Deletion of one or more nucleotides from the DNA sequence;
- Large-scale alterations in the DNA sequence; etc.

The impact on the resulting protein depends on the mutation form and its location (e.g., exon, intron, regulatory region, UTR, upstream and downstream) on the gene (Figure 2.3). We distinguish **Missense** mutations involving an SNV change that leads to the substitution of one amino acid in the protein sequence with another; **Nonsense** (or Stop gained): SNV mutations leading to the synthesis of an incomplete protein by introducing a premature stop codon in the mRNA sequence; **Frameshift** mutations when nucleotides are inserted or deleted from the

DNA sequence, causing a shift in the reading frame during translation; **Synonymous SNV** mutations which do not induce any change of amino acid, etc. Figure 2.3 by Bartha & Győrffy (2019) depicts the locations on the gene and impact on the protein product.

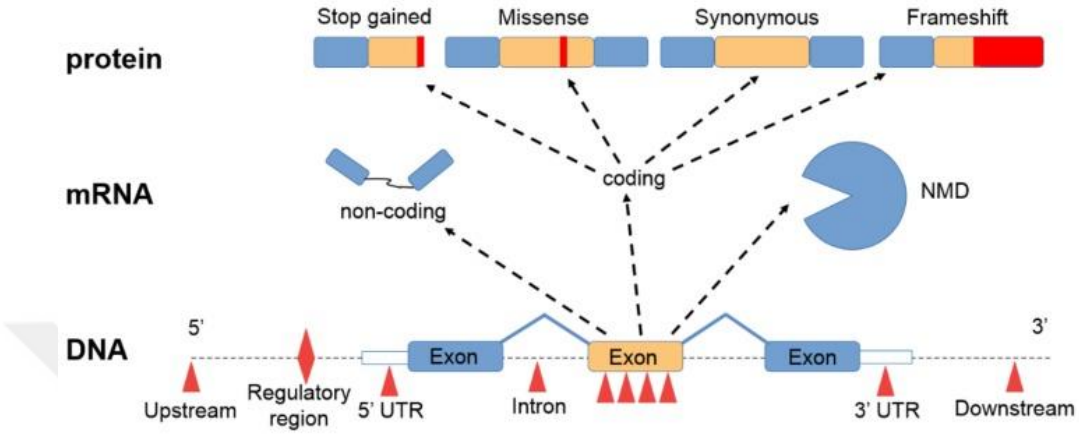


Figure 2.3. The locations and consequences of a genetic mutation on the resulting protein (Bartha & Győrffy., 2019)

Alongside the deleterious effects on the generated protein function, a gene mutation can results in the alteration of gene expression, non-coding RNA production, and Nonsense-mediated mRNA Decay (NMD). NMD is a cellular quality control mechanism that recognizes and degrades mRNA transcripts containing premature stop codons (Frischmeyer, 1999).

WES identifies clinically significant mutations in an individual's DNA sample compared to a reference genome or control tissue sample, with applications in healthcare for disease diagnosis and scientific research.

2.3.2. Variant calling from WES Databases

Knowledge Discovery in Databases (KDD) is a method for turning data into knowledge. It integrates databases, machine learning, statistics, and domain-specific knowledge to extract valuable insights (Figure 2.4) (Fayyad et al., 1996).

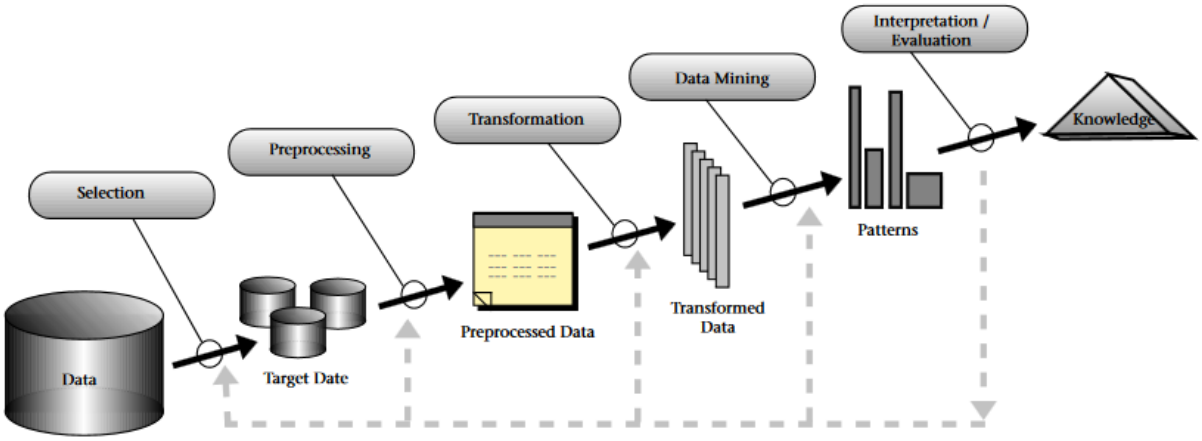


Figure 2.4. An overview of the KDD process's steps (Fayyad et al., 1996).

A standard WES data analysis pipeline (Figure 2.5) follows the KDD structure to extract meaningful biological insights, given the raw sequencing data, a reference genome, and additional resources :

- The data **preparation** encompasses the selection, cleaning, and preprocessing of raw data. WES requires filtering out low-quality reads, removing sequencing artifacts, etc;
- The data **transformation** and **integration** can correspond to the alignment of reads from the WES dataset to a reference genome;
- The data **mining** step involves the application of variant calling algorithms to identify relevant genomic mutations through adequate comparative analysis;
- Pattern **Evaluation** assesses the quality and reliability of called variants using metrics such as read depth, allele frequency, and variant validation steps;
- The insights gained from WES analysis are effectively communicated to the clinicians and researchers through knowledge **Presentation**.

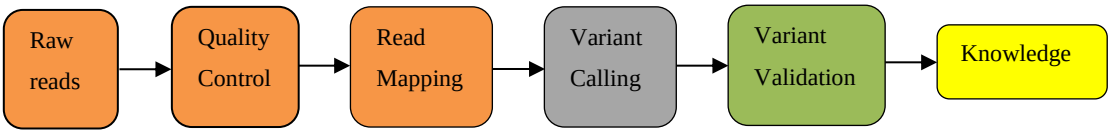


Figure 2.5. A simplified germline variant calling pipeline from WES datasets

In the following section, increased focus is dedicated to the read mapping and variant calling processes, those crucial operations require relatively more computational resources and runtime.

A. Read mapping

Genomic reads mapping is fundamentally a string pattern-matching problem, aimed at finding the optimal alignment of a sequence against another reference sequence. In the WES context, read mapping contributes to the identification of genomic variations through a comparative analysis between a sample's read sequences and a reference genome.

The Smith-Waterman or SW algorithm (Smith and Waterman, 1981) and the Needleman-Wunsch or NW algorithm (Needleman and Wunsch, 1970) are fundamental tools for sequence alignment in bioinformatics and computational biology.

Given two sequences *A* and *B* having respective lengths *n* and *m*; the SW algorithm proceeds in three steps:

1. Definition of match score *M*, mismatch *N* scores, and gap penalty *G*
2. Initialization of the similarity Matrix: $n+1 \times m+1$ matrix with all values set to zero
3. Updating of the matrix using the following formula:

$$M_{i,j} = \max\left(M_{i-1,j-1} + S(i,j), M_{i-1,j} + G, M_{i,j-1} + G, 0\right)$$

Given two values i and j such as $1 < i < n+1$ and $1 < j < m+1$

$$S(i,j) = \begin{cases} M \text{ if } A[j-1] = B[j-1] \\ N \text{ if } A[j-1] \neq B[j-1] \end{cases}$$

Each time, an arrow pointing towards the cell from where the maximum score has been determined is added to facilitate the final step: trace back the matrix.

4. The trace back begins from the position that has the highest value, pointing back with the pointers, thus finding out the possible predecessor, until we reach the score of 0

The NW algorithm performs global alignment and follows similar procedures as Smith-Waterman with the following differences (Table 2.1).

Table 2.1. Comparison between SW and NW algorithms (Parvez et al., 2020)

Table	**Smith-Waterman algorithm**	**Needleman-Wunsch algorithm**
Type	identifies regions of similarity within sequences (local alignment)	align two sequences in their entirety, to find the best overall alignment between them (global alignment)
Initialization	The first row and first column are set to 0	The first row and first column are subject to a gap penalty
Scoring	A negative score is set to 0	The score can be negative
Traceback	Begin with the highest score, and end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at the top left cell

The SW algorithm identifies regions of similarity within sequences (*local alignment*). In contrast, the NW algorithm performs global sequence alignment. It aligns entire sequences to find the optimal global alignment rather than just identifying local similarities. In the context of WES, we're more interested in the location of the entire short read on the reference genome. SW and NW algorithms are not suitable for short read mapping to a reference genome in a WES variant calling scenario for reasons including:

- Computation complexity: The NW algorithm has a time complexity of O(nm) where n and m are the lengths of the sequences to be aligned. This makes it computationally intensive for aligning short reads to a large reference genome.
- Alignment style: In WES mismatches, indels, and gaps are allowed.

Bioinformaticians designed more efficient approaches for mapping short reads to large genomes, based on heuristics, indexes, and newer data structures. Suffix trees are among the first index approaches for pattern matching, in the early ages of NGS (Musich et al., 2021). Despite offering a notable speed advantage, they represent a memory challenge. For example, the human reference genome's suffix tree occupies more than 45 GB of storage space (Kurtz, 1999).

The FM index introduced by (Ferragina and Manzini, 2004), was a great improvement in terms of speed and memory occupation. It is built on the Borrow Wheel Transform (BWT) (Burrows

and Wheeler, 1994). BWT is a sequence-to-sequence transformation technique utilized in popular read mappers (e.g., BWA (Li and Durbin, 2009)), thanks to its reversibility; data compression, and its utility as an indexing mechanism for efficient sequence search. The BWT often leads to improved compressibility of the input sequence. It tends to cluster similar characters, making the transformed sequence more suitable for entropy encoding methods such as Run-Length Encoding (RLE). This compression property is especially useful in DNA and text data, where certain patterns and repeats are prevalent;

Here are the main steps to generate a BWT of a given string:

1. Take all cyclic permutations of the word (including the original word itself).
2. Sort these cyclic permutations lexicographically.
3. Extract the last character of each sorted permutation.

The following figure illustrates the BWT transform process of the world: ‘MARMARAY’

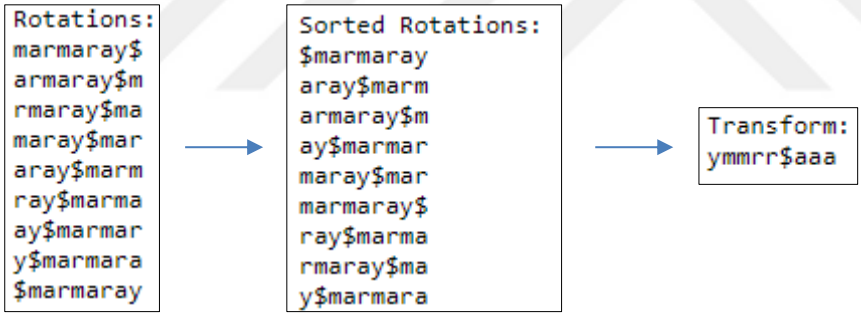


Figure 2.6. Example of BWT Transform

B. Variant Calling

Variant calling is the primary objective in WES analysis. It consists of identifying and characterizing genetic variants, within the genome protein-coding regions. The variant calling process is vulnerable to many biases:

- Allelic drop-out and amplification bias: the occasional amplification failure of one of the two alleles at a given locus during library preparation. Dropout can lead to under-detection of the rare variants and increase false negative results (Shestak et al., 2021)
- Strand bias: differences between the genotype inferred from the positive and negative strands. The impact of strand bias was evaluated by (Guo et al., 2012)
- Coverage bias: As sufficient coverage is a key parameter in variant calling, exome sequencing requires an even higher coverage to compensate for uneven coverage among selected regions (Zverinova and Guryev, 2022).

NGS reads are subject to a complex error generation process influenced by various factors, including instrument properties and preceding data processing tools (Goodwin et al., 2016; Li, 2014). This represents an additional challenge of distinguishing true genomic variants from false variations originating from sequencing errors (Goldfeder et al., 2016; Ding et al., 2012).

Genetic variant callers utilize a variety of algorithms and approaches to achieve accurate results:

- *Haplotype-based approach*: they reconstruct haplotypes from the reads and then identify variants based on differences between the reconstructed haplotypes and the reference.
- *Bayesian approach:* Bayesian variant callers use probabilistic models to estimate the likelihood of observing variants given the sequencing data and prior knowledge about the genomic variation.
- *Machine learning algorithms* to classify genomic regions as variant or non-variant based on features extracted from sequencing data (read depth, base quality, mapping quality, etc). They require large and diverse training datasets and generalize well to different genomic contexts.

3. MATERIAL AND METHODS

3.1. DATA DESCRIPTION

3.1.1. Whole Exome Sequencing Data

Whole Exome sequencing data belonging to 18 Pituitary Adenoma patients was obtained from İstanbul University Aziz Sancar Institute of Experimental Medicine (DTAE) with the contribution of Assoc. Prof. Dr. Feyza Nur TUNCER. For each sample, two WES paired-end read files were provided in FASTQ (Cock et al., 2010) format but clinical and demographic details are not available.

The FASTQ format is a common text-based file format that stores genomic sequences. Each sequence read is represented by four lines formatted as follows:

1) The header line : (starting with the **'@'** character) contains the read's unique identifier;
2) Sequence line: containing the sequence of nucleotide;
3) The '+' sign: A separator line included to separate the sequence from the quality scores;
4) The last line contains the quality scores in the Phred scale.

```
bash-4.2$ zcat   Documents/Data/p_reads/20107401648_R1.fastq.gz|head -10
@A00868:331:HKMVYDSX2:4:1101:13892:2472 1:N:0:TGCAGTGCTC+AGGTGTTCTA
CAAAGTTTTCAGTCGCAAATCATACCTTGAAAGACATAGGAGAATTCACACTGGTGAGAAACCATACAAAAGTAAGGTTT
+
FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFF
```

Figure 3.1. A segment of the first read in the S48 sample

The Phred score encodes the base call quality in an ASCII style. İt is calculated with the following formula :

$Q = -10log_{10}(P)$ where P is the probability that the base call is incorrect.

Table 3.1. Phred scores examples and description

Phred Score	Error probability	Base call accuracy	ASCII character
10	0,1	90%	+
20	0,01	99%	K
30	0,001	99,9%	U
40	0,0001	99,99%	I



Figure 3.2. Number of reads in each sample (the first pairs)

3.1.2. Additional resources

a. Reference genome: it serves as a framework for comparing the genetic material of individuals within a species. GRCh38 is the latest major assembly of the human genome released by the Genome Reference Consortium. Other genome assemblies include GRCh37 and HG19 (created by UCSC and derived from GRCh37). GRCh38 offers many enhancements over earlier versions, but it has been noted to contain an expanded range of alternative contigs, which, if not managed properly, can result in incorrect variant calls (Jia et al., 2020). In a study involving 24 academic and 4 commercial diagnostic laboratories, 54% reported having no plans to migrate to GRCh38 (Lansdon et al., 2021).

In this project, we opted for the *hs37d5* reference genome, a modified version of the hg19 reference, released by the 1000 Genome Project (Phase II). This reference genome includes extra "decoy" sequences, which can increase the accuracy of alignments and improve the precision of variant calling in analytical workflows (Li, 2014).

b. Exons capture kit coordinates: This data provides information about the positions of exons, introns, genes, and other genomic elements within a reference genome.

Table 3.2. List of additional data resources

	Source
hs37d5 **reference genome**	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
Agilent SureSelect All Exon v2.0 capture kit coordinates	https://doi.org/10.6084/m9.figshare.19115102

3.2. VARIANT CALLING PIPELINE

This section describes the main steps involved in the genetic variant calling pipeline starting from data preprocessing, quality assessment, and enhancement operation.

3.2.1. Data preprocessing

Various factors affect the quality of genomic raw datasets including, very short reads, low Phred score base qualities, high duplication levels, abnormal GC content ratios, and adapters. The quality metrics were retrieved using FastQC (Andrews, n.d.) and MultiQC (Ewels et al., 2016) tools.

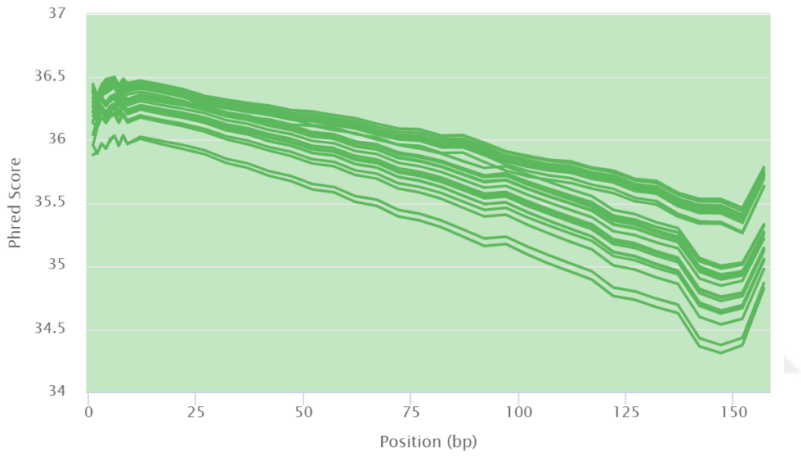


Figure 3.3. Sample quality score average per base position

The dataset exhibits a base quality range of 24 to 37, with sequence lengths varying between 10 to 156. The GC content hovers around 50%, while duplication levels exceed 20%.

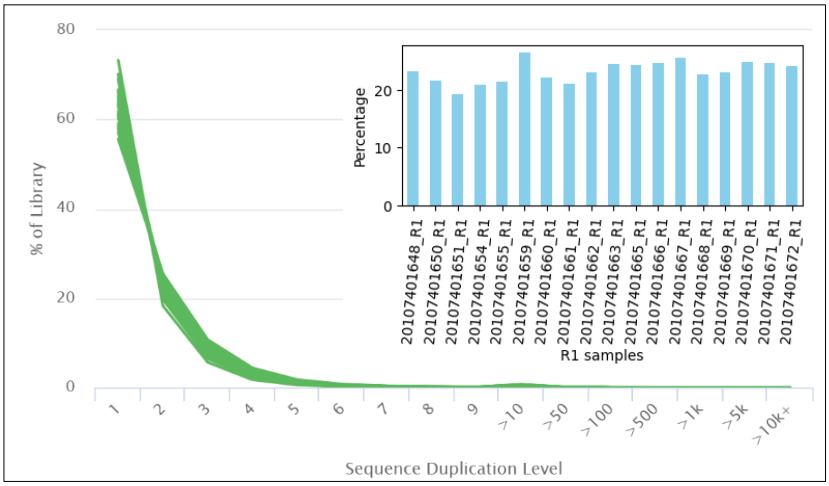


Figure 3.4. Percentage (blue color) and distribution (red color) of duplication count

The data enhancement process involves the elimination of low-quality read bases, adapters, and short reads,etc. A base quality threshold of 30 (equivalent to 99.9% accuracy) was utilized for quality improvement. This value is considered optimal in usual sequencing projects, especially in clinical research ("Sequencing Quality Scores," n.d.). Very short reads increase the risk of incorrect alignments, therefore we have established a minimum length of 150 bases. The table below provides a comprehensive list of utilized tools.

Table 3.3. Quality control and enhancement tools

	Reference	**Role**
multiQC	(Ewels et al., 2016)	Quality control
Fastqc_v0.12.1	(Andrews, n.d.)	Quality control
cutadapt	(Martin, 2011)	Detection and removal of adapters
Trimmomatic	(Bolger et al., 2014)	Read trimming
Trimgalore	(Krueger et al., 2023)	Wrapper around Cutadapt and trimmomatic

3.2.2. Reads mapping to a reference genome

Read mapping is a fundamental step in WES data analysis. It involves aligning short DNA sequences, known as reads, from sequencing to a reference genome. Pre-processed reads were mapped to the hs37d5 human reference genome using two BWT transform-based mappers: BWA-MEM (Li, 2013) and Bowtie2 (Langmead and Salzberg, 2012). The reference genome is firstly indexed using Samtools (Li et al., 2009) to enable quick and random access to the reference bases ("FASTA - Reference genome format – GATK," n.d.).

Table 3.4. Command line examples for reference genome indexing and reads mapping procedures

Samtools	BWA-mem	Bowtie2
`samtools faidx [Ref Gen]` `samtools dict[Ref Gen]`	`bwa index -b [block size][Ref Gen]`	`bowtie2-build --threads [cores] [ref gen] [ref gen]`

BWA-mem mapping	Bowtie2 mapping
`bwa mem -t [cores]-R [Read groups Info] [Ref Gen] [R1] [R2]\` `samtools [-bS -o] [output]`	`bowtie2 -p [CORES] -x [REF] -1 [R1] -2 [R2] \` `samtools [-bS -o] output_bam_file`

A reads group is a set of read sequences generated from the same run during sequencing steps. The reads group information is necessary in downstream procedures to handle potential systematic errors introduced during sequencing. Below is the list of involved software tools.

Table 3.5. Software tools for the reads alignment step

Tool	Reference	Role
Samtools-1.17	(Li et al., 2009)	Mapping files and Reference genome file manipulation (sorting, indexing, etc)
BWA-mem	(Li, 2013)	Read mapping
Bowtie2	(Langmead and Salzberg, 2012)	Read mapping

3.2.3. Mapping quality and files processing

The previous step yields two BAM (Binary Alignment/Map) (Li et al., 2009) files per sample. For each read, the corresponding DNA sequence, a CIGAR string describing how the read aligns to the reference genome with information about matches, mismatches, insertions, deletions, and skips; the alignment information (the alignment position of the read on the reference genome, such as the chromosome number and the starting position); the mate pair information (for paired-end sequencing); the mapping quality (the likelihood that the read is mapped to the correct location), and more informations are provided.

The following plots (Figure 3.5 to Figure 3.8) depict the mapping quality evaluation metrics.

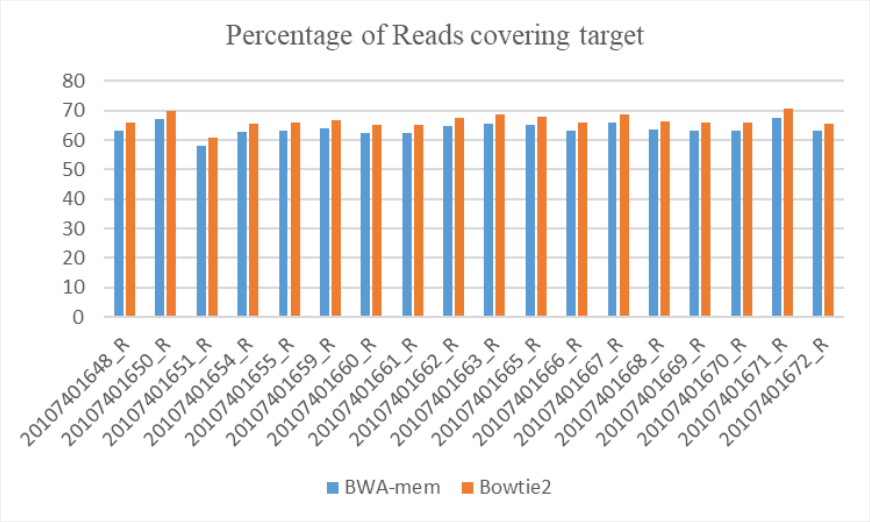


Figure 3.5. Percentage of reads covering targets

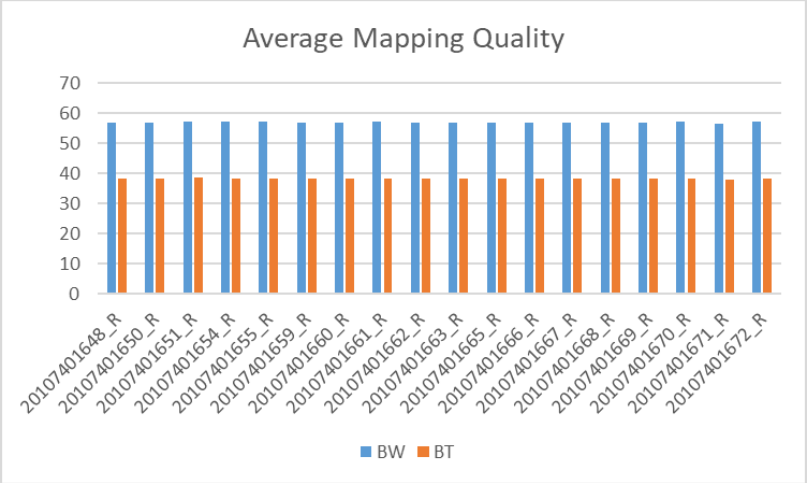


Figure 3.6. Average mapping quality

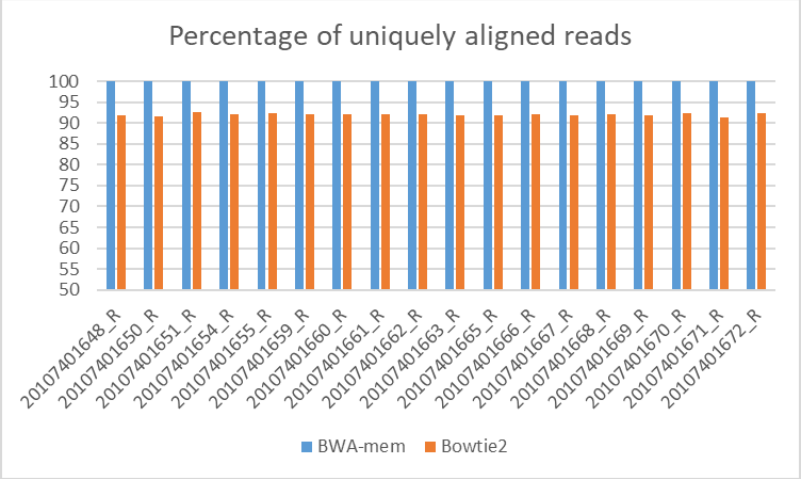


Figure 3.7. Percentage of uniquely aligned reads

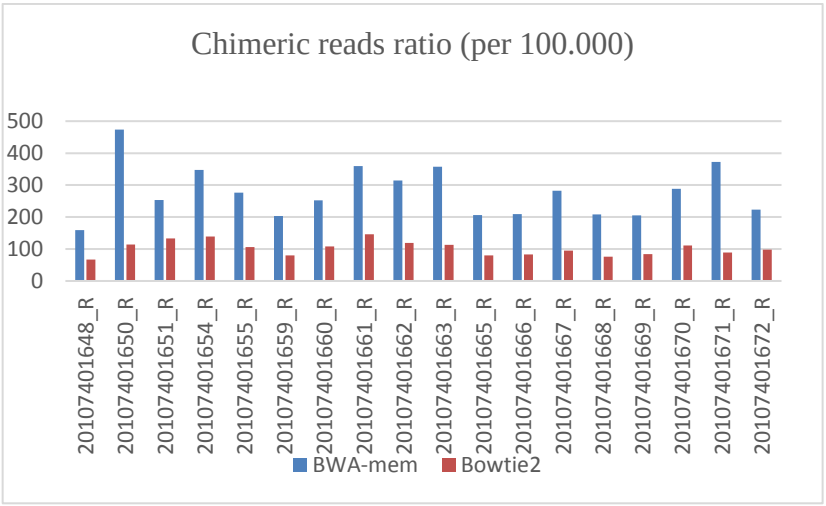


Figure 3.8. Number of chimeric reads per Sample

Two data pre-processing steps are conducted in addition to mapping quality assessment: marking duplicate reads and 'Fixmates'. Fixmates operation ensures that the attributes of each pair of reads are consistent and accurate. One additional step named BSQR (Base Scores Quality Recalibration) is recommended in GATK variant calling pipelines (DePristo et al., 2011), with two main purposes:

- Detect systematic errors in the base call accuracy estimation by the sequencing machine
- Recalibrate the scores to increase variant calling accuracy

Table 3.6. Command line examples for mapping files preprocessing using samtools

Fixmates	Mark Duplicates
`samtools sort -n -@ [cores] -m [memory] [input.bam]|\` `samtools fixmate -m -[output.bam]`	`samtools sort -@ [cores] -m [memory] input.bam"|\` `samtools markdup - [output.bam]`

3.2.4. Variant calling

Relying uniquely on one variant calling approach can lead to incomplete genomic information (Liu et al., 2020). Three different algorithmic approaches were combined to leverage their strengths while addressing their weaknesses:

- DeepVariant (https://github.com/google/deepvariant/): a deep neural network tailored for variant calling. Built on an adapted Inception-v2 architecture, the neural network is trained on sequencing datasets annotated with ground truth variant labels, to model the intricate relationship between sample reads, reference genomes, and existing genetic variants. The input (read sequence) data is transformed in a machine-learning compatible format, by first encoding the reference and read bases, qualities scores, and other read features into a red-green-blue (RGB) pileup image. As output, the CNN calculates the genotype likelihoods for the three diploid genotype states of homozygous reference (hom-ref), heterozygous (het), or homozygous alternate (hom-alt) using a 3 class-softmax layer. DeepVariant can generalize beyond its training data. A model trained on data from given samples is successfully usable for new, even non-human, samples (Poplin et al., 2018).

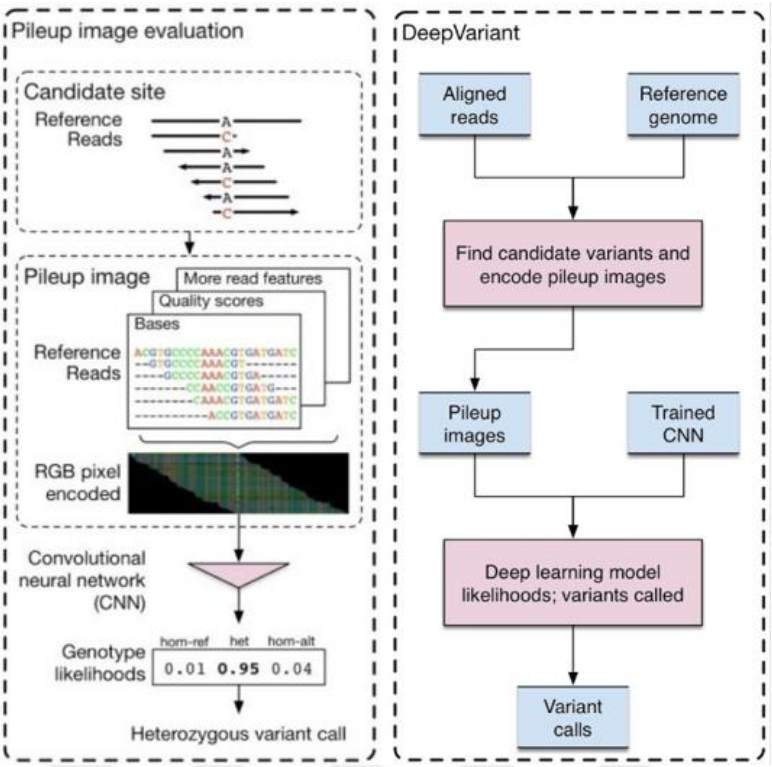


Figure 3.9. DeepVariant workflow overview (Poplin et al., 2018)

- **HaplotypeCaller**: a component of the GATK (Genome Analysis Toolkit) toolset. It incorporates many stages involving identifying potential haplotypes in the data through read assembly on the active regions then aligning reads to these haplotypes, and finally calling variants based on the alignment information and likelihood estimations ("HaplotypeCaller in a nutshell – GATK," n.d.).

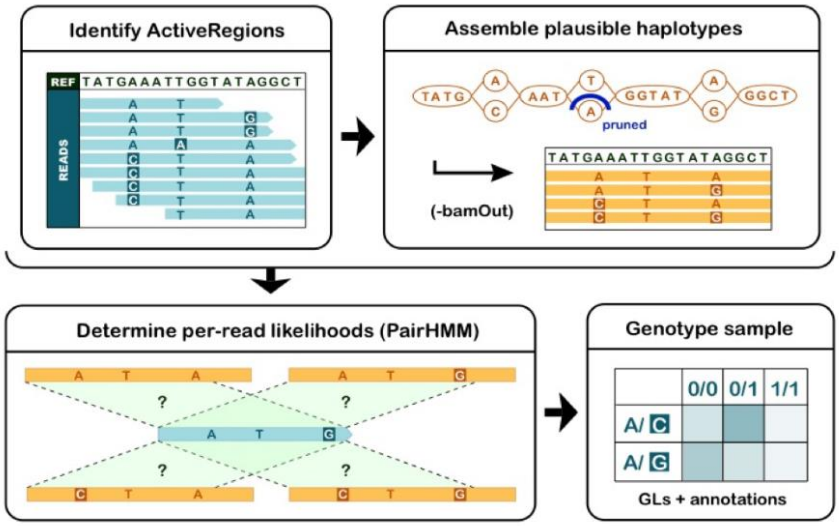


Figure 3.10. Haplotypecaller in four major steps ("HaplotypeCaller in a nutshell – GATK," n.d.)

The previous image illustrate the Haplotypecaller steps, performed based on popular algorithms such as Smith-Waterman (for aligning reads to the haplotypes), De Bruijn graph (reads de novo assembly), Pair Hidden Markov Model (PairHMM), etc (Poplin et al., 2017).

- **Freebayes**: a haplotype-based approach, like in the previous case. However, this strategy is not primarily based on aligning reads to the reference genome and then calling variants based on the alignment information. Instead, FreeBayes directly analyzes the individual sequencing reads within a specific genomic region, and uses Bayesian statistical models to infer the presence of variants (Garrison and Marth, 2012).

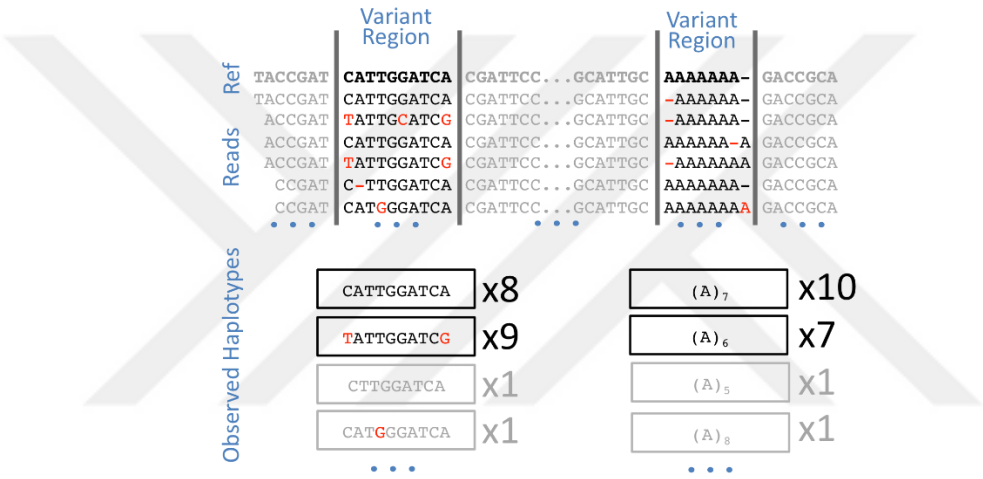


Figure 3.11. Freebayes, haplotype-based variant detection ("GitHub - freebayes/freebayes: Bayesian haplotype-based genetic polymorphism discovery and genotyping.," n.d.)

To handle the computational load resulting from the use of several approaches, a parallelization method inspired by Verrou et al., (2022) was adopted. Except for the deep learning approach, we run simultaneously multiple instances on separate regions of the genome and then concatenate the results.

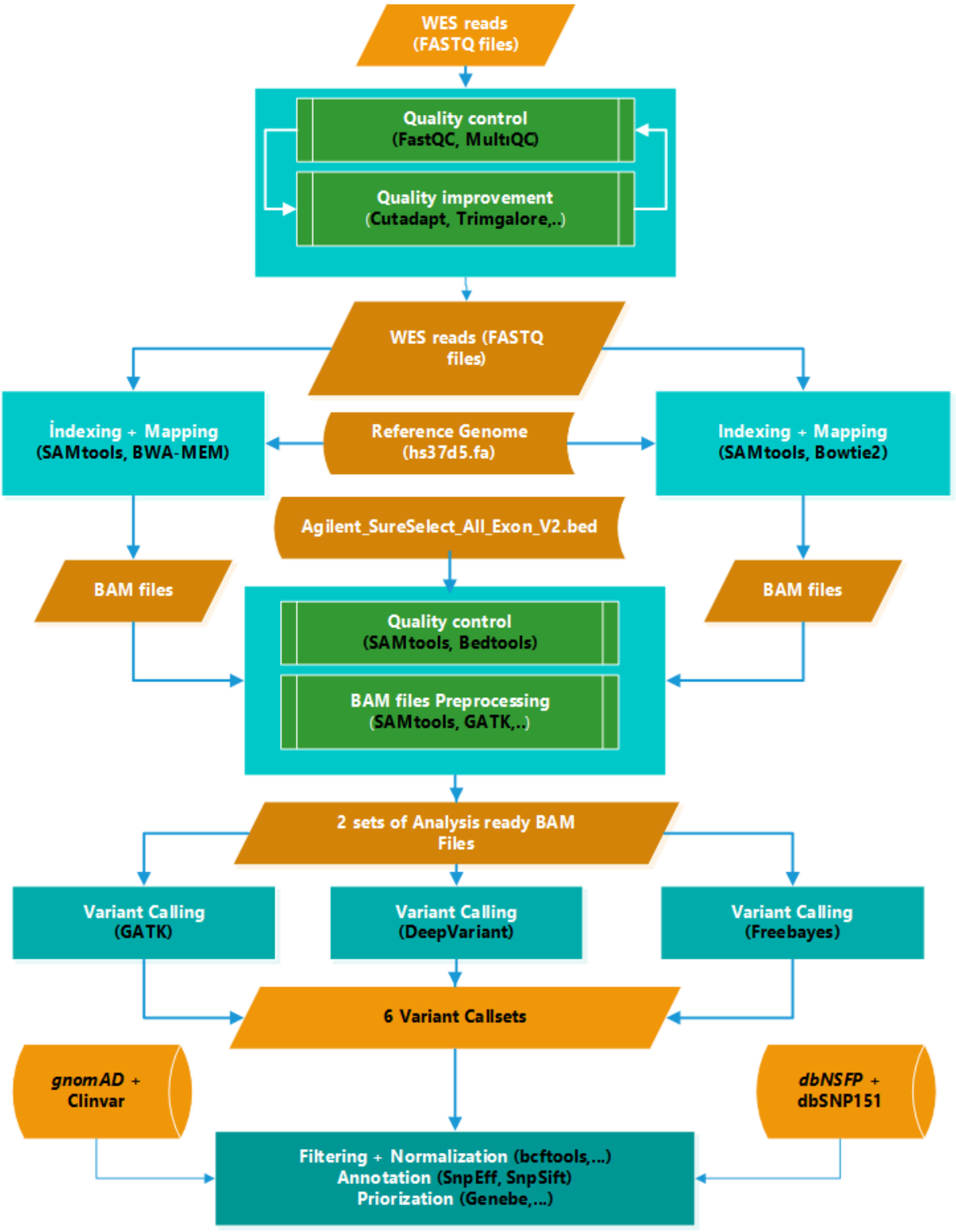


Figure 3.12. Overview of the Pipeline Workflow

3.2.5. Pattern Evaluation and Validation

The previous step yields 6 variant callsets, each containing tens of thousands of variants stored in variant call format or VCF (Danecek et al., 2011) files.

The VCF file format provides descriptive features used as filtering criteria to reduce the incidence of false positive variants. Here are examples of values located in the INFO column:

- QD: the variant confidence;
- QUAL: variant quality score, typically given in Phred-scale format;
- MQ: mapping quality over all the reads at the site;
- FS: Phred-scaled probability that there is strand bias at the site;
- StrandOddsRatio (SOR): estimate of strand bias;
- MQRankSum: the mapping qualities of the reads supporting the reference allele and the alternate allele;
- ReadPosRankSum: analyses positions of reference and alternate alleles within the reads.

The following figure by (Abdullah and Ahmet, 2020) illustrates the structure of a VCF file.

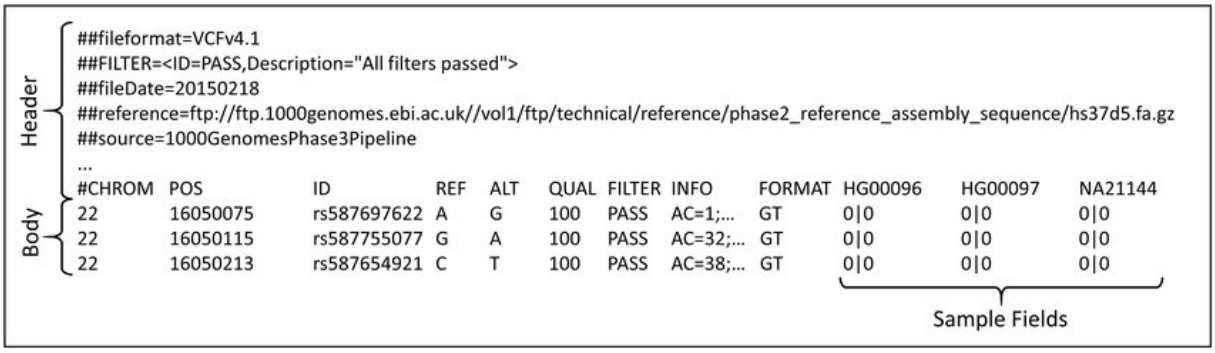


Figure 3.13. An example of a VCF file.

Table 3.7. Filtering threshold values from the GATK variant calling best practices

	QD	MQ	FS	SOR	MQRankSum	ReadPosRankSum
SNP	<2.0	<40	>60	>3	<-12.5	<-8.0
Indels	< 2.0	-	>200	-	-	< -20.0

The above table lists the threshold values recommended by GATK hard filtering best practices for genomic variant filtering ("Hard-filtering germline short variants – GATK," n.d.).

The previous filtering step selects true positive variant calls, and is followed by a variant prioritization procedure aimed to retain uniquely true variants that are also clinically significant.

It starts with variants annotation by adding necessary information such as frequencies across populations, variant impacts, computational pathogenicity and conservation scores, etc; The data was extracted from the following public datasets: *gnomAD* (Karczewski et al., 2020), *dbNSFP* (Liu et al., 2013), *dbSNP151* (Sherry, 2001) or obtained through various in silico prediction approaches.

Based on the variants annotations, they are classified into the following categories: Pathogenic (P), Likely pathogenic (LP), Uncertain significance (VUS), Likely benign (LB), and Benign (B) according to ACMG/PMG (American College of Medical Genetics and Genomics and the Association for Molecular Pathology) guidelines (Richards et al., 2015).

Variants with a Minor Allele Frequency (MAF) below 0.01 were retained, while those with a coverage less than 20 were filtered out. The candidate mutations were finally validated using the Integrated Genomics Viewer or IGV (Thorvaldsdottir et al., 2013) inspection tool.

The tools involved in pattern evaluation, validation step, and results analysis are listed in Table 3.8 and Table 3.9.

3.3. HARDWARE

The numerical computations detailed in this dissertation were conducted utilizing the resources of TUBITAK ULAKBIM's High Performance and Grid Computing Center, known as TRUBA (“Turkish National Science e-Infrastructure – TRUBA,” n.d.). TRUBA stands for Türk Ulusal Bilim e-Altyapısı which translates to Turkish National Science e-Infrastructure. All the bioinformatics intensive computations were executed on the provided storage capacity of 2TB and a total Core-Hour Usage of 210,000 hours.

After variant calling and prioritization, the remaining lightweight operations and result visualizations were performed using the free version of Google’s Compute Engine, Google Colab (“Welcome To Colab - Colab,” n.d.).

Table 3.8. Set of tools for the pattern evaluation, validation step, and results analysis

	Reference	Role
Genebe	(Stawiński and Płoski, 2024)& Płoski, 2024)	Provide access to Genebe: a web ACMG/PMG Guidelines implementation
Clinvar API	https://pypi.org/project/myvariant/	Provide access to Clinvar dataset
Liftover.1.1.17	https://pypi.org/project/liftover/	Python library for genomic coordinates conversion
LiftOver	https://liftover.broadinstitute.org/	Web platform for genomic coordinates conversion
Python 3.10	https://www.python.org/	Programming language
pandas 2.2. 2	https://pypi.org/project/pandas/	Data analysis library
NumPy	https://numpy.org/	fundamental numerical operations and data structures
scipy	https://scipy.org/	mathematical functions and algorithms for scientific computing

Table 3.9. Set of tools for results visualization

	Reference	Role
Seaborn	https://seaborn.pydata.org/	Data visualization library
matplotlib 3.8.4	https://pypi.org/project/matplotlib	Data visualisation library
upsetplot	(Lex et al., 2014)	Upsetplot generation library
Matplotlib-venn	https://pypi.org/project/matplotlib-venn/	Venn diagram generation library
Datawrapper	https://www.datawrapper.de/	Web platform for complex graphs
IGV 2.16.2	(Thorvaldsdottir et al., 2013)	Genomics data visualization and exploration

4. RESULTS

A total of 123 genetic variant mutations were obtained after variant calling and prioritization steps. They consist of 23 variants categorized as Variant of Uncertain Significance (VUS), 74 variants classified as Likely Pathogenic (LP), and 26 variants designated as Pathogenic (P).

The variants are located in 104 genes including genes previously associated with PA disease.

The variants were not uniformly distributed across the samples (Figure 4.1). The *20107401659* sample is the tumor showing the highest amount of germline variant (9 LP and 5 P) and the patient *20107401650* has only two 2 LP genomic mutations. The average is 7.1 variants per sample.

All the samples presented at least one LP variant, only 16 samples showed P mutations; and 6 samples didn't display any VUS.

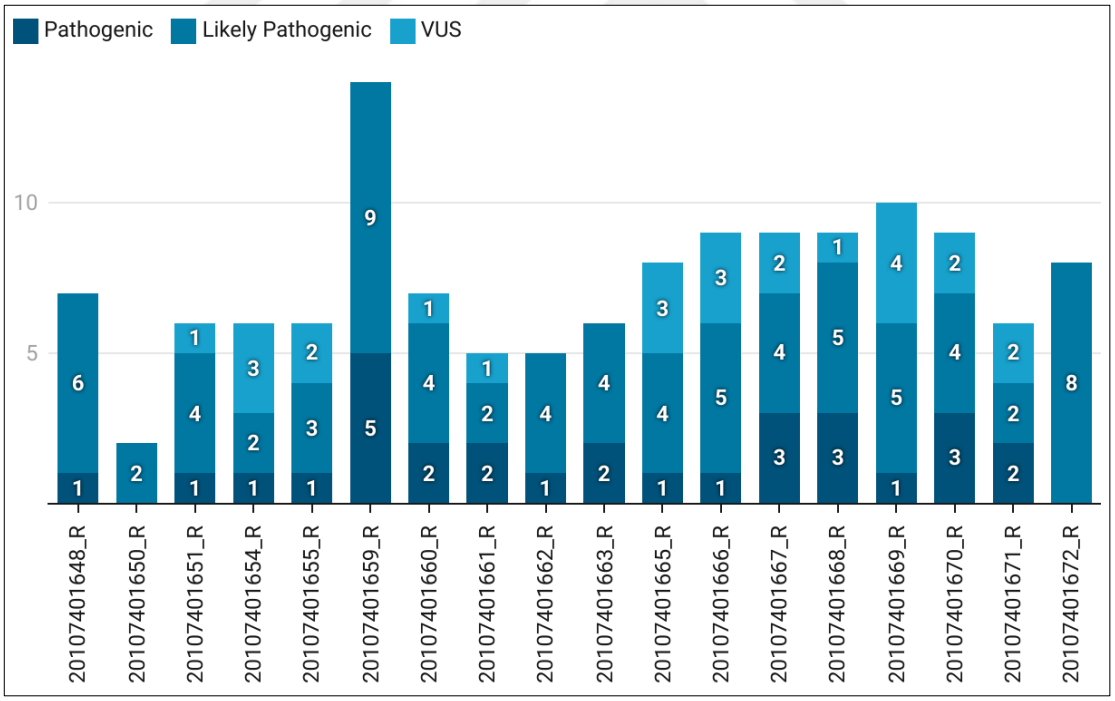


Figure 4.1. Distribution of 123 germline variants across the Samples

Single-nucleotide variants (SNVs) were the most frequently observed type of DNA mutations, accounting for over 80% of all cases. In addition, we observed 15 deletions, 3 duplications, and

2 insertions. The table below depicts the number of occurrences of each form of nucleotide change. The transitions SNV (A/G and C/T) are more represented (71.82%) than transversions.

Table 4.1. Representation of base change and SNV types

	A	C	G	T
A	0	2	20	2
C	3	0	2	18
G	32	6	0	6
T	3	9	7	0

Variant effect predictions suggest a prevalence of missense variations (N = 28), followed by frameshift and splice-donor mutations, both occurring 16 times. The other consequences are stop-gained, start-lost mutations, and splice-acceptor alteration; accounting respectively for N = 11, N = 13, and N=10. Some variants not mentioned here were predicted to have more than one effect possible.

Table 4.2. List of the recurrent genes in the variant callset

Gene	Count	VUS	LP	P	Samples
TTN*	7	7	-	-	69,71,67,55,(54,70)
CAPN3	5	0	1	4	(54, 68), 66, 71
MEFV	4	0	0	4	(59, 68,70), 63
RYR1*	4	2	2	-	67, 54
AIP*	3	-	1	2	59,62,68
PAH	2	-	-	2	61,69
ADCK2	2	-	2	-	48,50
ABCA7	2	-	2	-	(59, 66)
IPP	2	-	2	-	59,71
PLA2G6*	2	1	1	-	65,66
USP6	2	-	2	-	(59,69)
TYR	2	-	-	2	59,65

*Genes previously associated to Pas

() : Samples harboring same variant on that gene

The following tables (Table 4.3 to 4.7) list the detected variant characteristics. Some genes are notable for their established association with PA in previous studies and other genes for their repetitive occurrence across numerous samples (Table 4.2).

Table 4.3. List of germline VUS Variants

SAMPLE	Variant ID	AF (%)	DP	Nucleotide Change	Protein Change	Effect	Gene	MAF(%)	Bayes DelnoAF	phyloP 100way	ACMG
20107401651	10-21106569-T-C	47.14	70	c.2108A>G	p.Glu703Gly	missense	NEBL	0.00003184*	-0.330	5.038	0
20107401654	19-39002230-G-A	59.26	81	c.9152G>A	p.Arg3051His	missense	**RYR1**	0.00025	0.010	10.003	0
20107401654	**2-179605894-A-C**	44.85	165	c.12066T>G	p.Cys4022Trp	missense	**TTN**	0.00000342**	-0.250	0.711	3
20107401654	5-142779983-C-T	41.56	154	c.422G>A	p.Ser141Asn	missense	NR3C1	NF	-0.620	1.341	0
20107401655	1-23885763-C-T	47.55	143	c.155G>A	p.Arg52Gln	missense	ID3	0.000012	0.040	2.632	1
20107401655	2-179497668-G-T	49.35	77	c.43190C>A	p.Ser14397Tyr	missense	**TTN**	0.000004	-0.290	4.740	1
20107401660	7-141727450-G-A	56.47	85	c.1136G>A	p.Gly379Glu	missense	MGAM	0.00021	0.570	9.540	3
20107401661	1-17354257-T-C	45.21	73	c.527A>G	p.Glu176Gly	missense	SDHB	0.000003978	0.070	7.476	2
20107401665	14-65544139-G-T	43.27	171	c.276C>A	p.Ser92Arg	missense	MAX	0.000012	-0.040	4.292	3
20107401665	6-129465109-T-G	38.60	171	c.703T>G	p.Phe235Val	missense	**LAMA2**	NF	0.330	7.969	2
20107401665	9-21970957-G-C	60.00	110	c.401C>G	p.Ala134Gly	missense	CDKN2A	NF	-0.230	0.142	1
20107401666	11-62297928-G-A	40.29	139	c.3961C>T	p.Pro1321Ser	missense	AHNAK	0.000000684**	-0.310	3.119	1
20107401666	18-7023349-G-A	50.00	92	c.2515C>T	p.Pro839Ser	missense	**LAMA1**	0.0006482	0.040	9.993	1
20107401666	22-38531029-C-T	60.00	45	c.860G>A	p.Gly287Glu	missense	**PLA2G6**	0.000004	-0.220	3.447	1
20107401667	19-38996947-A-G	56.36	110	c.8546A>G	p.Tyr2849Cys	missense	**RYR1**	0.000009577**	0.100	4.631	2
20107401667	2-179446520-G-A	45.16	93	c.66475C>T	p.Pro22159Ser	missense	**TTN**	0.0000044	-0.250	9.940	2
20107401669	2-179440040-C-T	48.39	186	c.70819G>A	p.Ala23607Thr	missense	**TTN**	0.00003187*	-0.060	7.858	1
20107401669	2-47703697-G-A	49.02	102	c.2197G>A	p.Ala733Thr	missense	MSH2	0.000048	0.270	6.412	4
20107401669	5-66432450-A-C	48.18	110	c.2452A>C	p.Thr818Pro	missense	MAST4	NF	-0.020	5.171	1
20107401669	X-86890642-A-G	45.08	193	c.1792A>G	p.Ser598Gly	missense	KLHL4	0.00000182**	-0.140	8.941	1
20107401670	2-179451528-G-A	48.86	88	c.64100C>T	p.Pro21367Leu	missense	**TTN**	0.000029**	-0.130	9.540	3
20107401670	**2-179605894-A-C**	52.82	195	c.12066T>G	p.Cys4022Trp	missense	**TTN**	0.00000342**	-0.250	0.711	2
20107401671	2-141598534-G-C	48.62	181	c.5067C>G	p.Ile1689Met	missense	LRP1B	NF	0.070	2.316	1
20107401671	2-179444003-T-C	46.58	219	c.67754A>G	p.Lys22585Arg	missense	**TTN**	0.00000137**	-0.420	6.227	1

Table 4.1. List of Pathogenic variants

SAMPLE	Variant ID	AF	DP	Nucleotides Change	Protein Change	MAF	Effect	Gene	Bayes DelnoAF	phyloP 100way
20107401648	2-69581446-T-C	44.44	81	c.686-2A>G		0.000023	splice_acceptor	GFPT1	0.210	4.944
20107401651	1-17707541-G-T	52.17	46	c.436-1G>T		Not Found	splice_acceptor	PADI6	0.000	4.592
20107401654	**15-42691799-G-A**	55.56	63	c.1303G>A	p.Glu435Lys	0.000084	missense	**CAPN3**	0.370	8.056
20107401655	12-29673549-A-G	36.47	85	c.1888+2T>C		0.0000066**	splice_donor	TMTC1	0.060	8.271
20107401659	11-67254618-C-T	45.52	145	c.241C>T	p.Arg81Ter	0.00000068**	stop_gained	**AIP**	0.540	1.215
20107401659	14-31355519-G-C	47.83	92	c.1477+1G>C		0.000000684**	splice_donor	COCH	0.250	4.604
20107401659	**16-3293407-T-C**	42.51	247	c.2080A>G	p.Met694Val	0.00028	missense	**MEFV**	0.020	0.527
20107401659	9-136218980-C-T	48.10	158	c.769G>A	p.Gly257Arg	0.00000207	missense	SURF1	0.500	7.333
20107401659	11-89018108-A-G	46.83	126	c.1352A>G	p.Tyr451Cys	0.00019	missense	**TYR**	0.480	6.825
20107401660	19-7809829-G-A	55.91	93	c.898C>T	p.Gln300Ter	0.00000342**	stop_gained, splice_region	CD209	0.520	0.657
20107401660	11-64527370-T-C	66.67	87	c.1A>G	p.Met1?	0.000016	start_lost	PYGM	0.450	5.932
20107401661	12-103237568-C-T	51.85	54	c.1066-11G>A		0.00025	splice_polypyrimidine_tract, intron	**PAH**	-0.810	0.955
20107401661	9-131570091-A-G	43.66	71	c.1138-2A>G		Not Found	splice_acceptor	TBC1D13	0.310	8.088
20107401662	17-6329101-C-T	54.05	111	c.834G>A	p.Trp278Ter	0.00033	stop_gained	AIPL1	0.650	7.468
20107401663	9-133370368-G-T	51.90	79	c.1085G>T	p.Gly362Val	0.000001369**	missense	ASS1	0.470	9.294
20107401663	16-3293205-C-T	41.18	136	c.2282G>A	p.Arg761His	0.00020	missense	**MEFV**	-0.290	0.399
20107401665	11-88911936-G-A	54.05	111	c.815G>A	p.Trp272Ter	0.00000824*-	stop_gained	**TYR**	0.660	9.483
20107401666	14-61445948-C-T	52.83	106	c.667+1G>A		Not Found	splice_donor	TRMT5	0.370	7.385
20107401667	15-80465358-C-T	45.18	166	c.709C>T	p.Arg237Ter	0.000012	stop_gained,	FAH	0.450	1.975
20107401667	12-80626878-G-T	39.02	82	c.817+1G>T		0.000004304	splice_donor	OTOGL	0.340	9.332
20107401668	11-67258381-C-T	52.63	95	c.910C>T	p.Arg304Ter	0.000012	stop_gained	**AIP**	0.500	2.252
20107401668	**15-42691799-G-A**	53.09	81	c.1303G>A	p.Glu435Lys	0.000084	missense	**CAPN3**	0.370	8.056
20107401668	**16-3293407-T-C**	48.35	182	c.2080A>G	p.Met694Val	0.00028	missense	**MEFV**	0.020	0.527
20107401669	12-103237443-C-G	45.00	60	c.1180G>C	p.Asp394His	0.00000137**	missense	**PAH**	0.500	5.015
20107401670	8-143625579-A-C	58.75	80	c.4558-2A>C		Not Found	splice_acceptor	ADGRB1	-0.020	4.327
20107401670	1-154245888-T-TA	50.00	156	c.130_131insA	p.Trp44Ter	0.0000006841**	stop_gained,	HAX1		2.967
20107401670	**16-3293407-T-C**	46.98	215	c.2080A>G	p.Met694Val	0.00028	missense	**MEFV**	0.020	0.527
20107401671	15-42693840-T-TA	58.67	75	c.1357dup	p.Thr453AsnfsTer18	0.00000342**	frameshift,	**CAPN3**		9.196
20107401671	15-42693949-C-T	46.85	111	c.1465C>T	p.Arg489Trp	0.000016	missense	**CAPN3**	0.310	1.414

Table 4.2. List of LP variants

SAMPLE	Variant ID	AF	DP	Nucleotides Change	Protein Change	MAF	Effect	Gene	Bayes DelnoAF	phyloP 100way
20107401648	7-140374064-G-C	56.9	58	c.933+1G>C		0.0036	splice_donor	**ADCK2**	0.090	7.371
20107401648	17-46925761-ATTCC-A	55.6	153	c.364_367del	p.Pro122SerfsTer37	0.000001368**	frameshift	CALCOCO2		5.719
20107401648	1-15772101-G-A	60.2	98	c.649G>A	p.Gly217Ser	0.00006	missense	CTRC	0.360	10.003
20107401648	5-176829456-T-C	57.3	103	c.1685A>G	p.Asp562Gly	0.000001376**	missense	F12	0.350	4.848
20107401648	1-11863030-CT-C	50.6	172	c.143del	p.Lys48ArgfsTer2	Not Found	frameshift	MTHFR		7.460
20107401648	14-24710039-T-TC	49.3	136	c.646_647insG	p.Glu216GlyfsTer61	Not Found	frameshift	TINF2		1.373
20107401650	7-140373386-C-T	61.4	101	c.256C>T	p.Arg86Ter	0.000017	stop_gained	**ADCK2**	0.240	-0.525
20107401650	10-17659337-A-T	53.5	43	c.2T>A	p.Met1?	0.0000127**	start_lost	HACD1	-0.250	1.434
20107401651	17-79104885-T-C	43.8	32	c.1A>G	p.Met1?	0.00000214**	start_lost	AATK	-0.150	0.981
20107401651	19-35775922-T-C	51.7	89	c.232T>C	p.Cys78Arg	0.0000006845**	missense	HAMP	0.520	3.655
20107401651	13-25009613-C-T	47.0	66	c.3667-1G>A		0.00056	splice_acceptor	PARP4	-0.730	2.450
20107401651	2-61344628-AT-A	56.6	76	c.1893del	p.Asp631GlufsTer6	Not Found	frameshift	SANBR		2.127
20107401654	7-94047097-C-T	43.3	97	c.1925C>T	p.Pro642Leu	Not Found	missense	COL1A2	0.320	5.813
20107401654	17-1792138-TC-T	47.3	93	c.1546del	p.Leu516CysfsTer3	Not Found	frameshift	RPA1		1.328
20107401655	7-117251649-T-G	49.3	140	c.3154T>G	p.Phe1052Val	0.00065	missense	CFTR	0.420	7.967
20107401655	8-66616991-GCAGA-G	54.5	110	c.349_352del	p.Gln117PhefsTer13	0.0000080	frameshift	MTFR1		3.351
20107401655	15-56152667-C-A	53.3	105	c.898G>T	p.Glu300Ter	0.00000342**	stop_gained	NEDD4	0.630	5.749
20107401659	**19-1058727-C-T**	48.1	129	c.5260C>T	p.Arg1754Ter	0.00004381	stop_gained	**ABCA7**	0.620	-0.508
20107401659	11-33738938-AT-A	52.5	99	c.146del	p.Asp49ValfsTer31	0.0000006845**	frameshift	CD59		2.990
20107401659	2-71906321-T-C	43.9	139	c.6019T>C	p.Trp2007Arg	Not Found	missense	DYSF	0.430	8.017
20107401659	17-78086379-C-T	44.3	97	c.1757C>T	p.Ala586Val	0.000024	missense, splice_region	GAA	0.290	7.749
20107401659	9-6620257-A-T	53.3	165	c.397T>A	p.Tyr133Asn	Not Found	missense	GLDC	0.450	8.681
20107401659	1-46206642-CAT-C	55.6	239	c.653_654del	p.His218ArgfsTer29	0.000036	frameshift	**IPP**		7.552
20107401659	14-92472308-ACTTT-A	44.1	177	c.2008_2011del	p.Lys670LeufsTer12	0.00000137**	frameshift	TRIP11		4.952
20107401659	11-17552757-C-T	53.3	152	c.331G>A	p.Gly111Arg	0.0000006841**	missense	USH1C	0.150	7.188

Table 4.3. List of LP variants (part 2)

SAMPLE	Variant ID	AF	DP	Nucleotides Change	Protein Change	MAF	Effect	Gene	Bayes DelnoAF	phyloP 100way
20107401659	**17-5071419-G-A**	51.9	133	c.3228+1G>A		0.00051	splice_donor	**USP6**	-0.120	9.745
20107401660	11-114271053-T-C	52.8	89	c.1A>G	p.Met1?	0.00037	start_lost	C11orf71	0.320	3.190
20107401660	5-140307272-T-TG	46.7	184	c.799dup	p.Glu267GlyfsTer23	0.00039	frameshift	PCDHAC1		6.907
20107401660	19-38985262-TCG-T	53.4	73	c.6546_6547del	p.Ile2182MetfsTer70	Not Found	frameshift, splice_region	**RYR1**		9.940
20107401660	19-38985266-G-A	54.9	71	c.6548+1G>A		0.0000006841**	splice_donor	**RYR1**	0.200	9.940
20107401661	10-74235056-C-T	51.3	76	c.736-1G>A		Not Found	splice_acceptor	MICU1	0.420	7.879
20107401661	7-103214539-C-G	48.8	80	c.4511G>C	p.Arg1504Thr	0.0000006859**	missense, splice_region	RELN	0.080	7.473
20107401662	11-67257786-G-C	58.6	87	c.646-1G>C		Not Found	splice_acceptor	**AIP**	-0.390	9.540
20107401662	9-100122384-G-A	52.9	68	c.3483+1G>A		0.00012	splice_donor	CCDC180	-0.260	1.600
20107401662	7-80301310-T-G	50.5	93	c.1079T>G	p.Leu360Ter	0.00024	stop_gained	CD36	0.380	5.163
20107401662	12-31815202-G-A	54.3	70	c.314+1G>A		0.004558**	splice_donor	ETFBKMT	0.290	9.020
20107401663	2-25141615-A-C	49.7	177	c.242T>G	p.Leu81Arg	Not Found	missense	ADCY3	0.380	8.813
20107401663	6-70866121-A-T	40.9	225	c.2182A>T	p.Lys728Ter	Not Found	stop_gained, splice_region	COL19A1	0.600	3.834
20107401663	19-39440610-C-T	37.5	56	c.349+1G>A		0.000002895**	splice_donor	FBXO17	0.060	1.444
20107401663	17-7464136-T-C	50.4	125	c.737+2T>C		Not Found	splice_donor	TNFSF13	-0.260	-0.129
20107401665	5-137622217-A-G	55.8	104	c.1129T>C	p.Cys377Arg	Not Found	missense	CDC25C	0.280	7.079
20107401665	3-141497012-A-T	50.0	28	c.-113-2A>T		Not Found	splice_acceptor	GRK7	-0.860	0.304
20107401665	22-38516896-G-T	45.0	80	c.1612C>A	p.Arg538Ser	Not Found	missense	**PLA2G6**	0.25	7.250
20107401665	14-57075993-C-CCTTCA	46.1	89	c.807_811dup	p.Ser271ThrfsTer9	Not Found	frameshift	TMEM260		0.130
20107401666	**19-1058727-C-T**	100.0	73	c.5260C>T	p.Arg1754Ter	0.00004381	stop_gained	**ABCA7**	0.620	-0.508
20107401666	15-42693866-G-A	44.4	99	c.1382G>A	p.Arg461His	0.000007954	missense	**CAPN3**	0.250	7.962
20107401666	1-95699787-A-G	54.1	98	c.1A>G	p.Met1?	0.00032	start_lost	RWDD3	0.200	3.470
20107401666	1-246078929-T-C	44.7	123	c.716A>G	p.Tyr239Cys	0.00000205**	missense	SMYD3	0.550	4.948
20107401666	11-93470229-C-T	61.5	78	c.635+1G>A		0.0000044	splice_donor	TAF1D	0.120	5.062

Table 4.4. List of LP variants (final part)

SAMPLE	Variant ID	AF	DP	Nucleotides Change	Protein Change	MAF	Effect	Gene	Bayes DelnoAF	phyloP 100way
20107401667	17-4804149-A-C	55.0	149	c.856T>G	p.Phe286Val	0.00002	missense	CHRNE	0.410	6.068
20107401667	19-44160772-G-A	47.5	61	c.331C>T	p.Arg111Ter	0.00000481**	stop_gained	PLAUR	0.170	-0.104
20107401667	1-115489865-A-G	54.0	87	c.2248-2A>G		0.000036	splice_acceptor	SYCP1	0.010	6.072
20107401667	8-23002118-G-A	48.9	139	c.799C>T	p.Arg267Ter	0.000016	stop_gained	TNFRSF10D	0.000	-1.175
20107401668	6-35425745-AATGT-A	58.2	91	c.955_958del	p.Cys319ValfsTer44	Not Found	frameshift	FANCE		5.588
20107401668	2-163124596-C-T	44.3	122	c.2807+1G>A		0.0066	splice_donor	IFIH1	0.410	7.284
20107401668	3-122437806-G-A	47.5	101	c.4807+1G>A		Not Found	splice_donor	PARP14	0.180	3.331
20107401668	6-51930774-C-T	43.3	67	c.880G>A	p.Gly294Ser	0.000008	missense, splice_region	PKHD1	-0.130	5.239
20107401668	16-1555588-C-T	62.8	43	c.2020C>T	p.Gln674Ter	0.000012	stop_gained	TELO2	0.490	0.307
20107401669	22-43520031-G-T	56.6	113	c.3G>T	p.Met1?	Not Found	start_lost	BIK	0.070	3.715
20107401669	4-70823105-G-A	52.3	195	c.562C>T	p.Gln188Ter	0.000004	stop_gained	CSN2	0.290	3.559
20107401669	1-26349833-GA-G	45.5	99	c.697del	p.Arg233GlyfsTer87	Not Found	frameshift	EXTL1		-0.059
20107401669	9-35800044-A-G	47.1	119	c.1013A>G	p.Tyr338Cys	0.0000006841**	missense	NPR2	0.400	7.476
20107401669	**17-5071419-G-A**	57.5	106	c.3228+1G>A		0.00051	splice_donor	**USP6**	-0.120	9.745
20107401670	12-40716119-A-G	41.9	129	c.5318-2A>G		Not Found	splice_acceptor	LRRK2	0.290	8.616
20107401670	17-47304052-A-G	40.0	100	c.2T>C	p.Met1?	0.0000080	start_lost	PHOSPHO1	0.110	4.002
20107401670	11-117988031-TG-T	44.7	103	c.1167del	p.Leu391Ter	0.000012	frameshift	TMPRSS4		8.950
20107401671	1-46193302-C-T	48.4	95	c.1048+1G>A		Not Found	splice_donor	**IPP**	0.440	7.459
20107401671	20-35563450-C-T	51.7	118	c.491G>A	p.Arg164Gln	0.000004	missense	SAMHD1	0.600	7.905
20107401672	11-7694054-C-G	59.5	116	c.3G>C	p.Met1?	0.000007954	start_lost	CYB5R2	-0.020	1.702
20107401672	14-60622852-A-G	61.8	89	c.2T>C	p.Met1?	0.000034	start_lost	DHRS7	-0.140	3.717
20107401672	12-124413859-T-C	53.7	162	c.12344T>C	p.Met4115Thr	0.000021	missense	DNAH10	-0.120	7.951
20107401672	22-41257997-A-G	45.9	74	c.2T>C	p.Met1?	0.000014	start_lost	DNAJB7	0.030	4.617
20107401672	16-67700344-A-T	60.0	65	c.2T>A	p.Met1?	0.00052	start_lost	ENKD1	0.370	4.601
20107401672	5-140794901-TGCGGC-	48.5	270	c.2162_2166del	p.Arg721LeufsTer68	Not Found	frameshift	PCDHGA10		0.537
20107401672	X-106332067-C-A	40.8	130	c.598G>T	p.Glu200Ter	0.0000009229**	stop_gained, splice_region	RBM41	0.630	4.641
20107401672	11-113677272-AG-A	48.6	72	c.2524del	p.Leu842TrpfsTer3	Not Found	frameshift	USP28		3.565

In the above tables (Table 4.3 to 4.7), the MAF column includes details about the database name and version, by adding the following symbols on the numerical values if the variant is missing in the default dataset (gnomAD v2.1.1 Exomes):

*: gnomAD v2.1.1 Genomes,

**: gnomAD v4.1.0 Exomes,

***: gnomAD v4.1.0 Genomes,

*-: ExaC;

The variant ID represents the four values (CHROM-POS-REF-ALT) separated by the '-' character, and recurrent values are indicated in bold.

Our results report only 5 recurrent variants across the samples (Table 4.5). Detected in three different samples, rs61752717 is the most recurrent variant.

Table 4.6. List of recurrent genomic variants

Dbsnp ID	Gene	ID	Occurrences	MAF	Freq (here)
rs149914792	CAPN3	15-42691799-G-A	2	0.00008389	11.11
rs61752717	MEFV	16-3293407-T-C	3	0.0002823	16.66
rs536045893	USP6	17-5071419-G-A	2	0.0005070	11.11
rs556286113	ABCA7	19-1058727-C-T	2	0.00004381	11.11
rs931245340	TTN	2_179605894_A_C	2	0.00000342	11.11

The patient genetic profiles represent a great source of insights in clinical analysis.

Sample 60 has two LP variants on the same gene (RYR1). One mutation, identified as c.6548+1G>A, induces a splice_donor effect. It is located on chromosome 19, position 38985266. The second is a frameshift mutation denoted as c.6546_6547del (p.Ile2182MetfsTer70).

A splice-donor germline mutation was simultaneously detected in two samples: S59 and S69. The variant is registered in the gnomAD v2.1 Exomes dataset (MAF = 0.00051). Located on the 5071419 position of Chromosome 17, it is a heterozygote mutation of the USP6 gene. ACMG guidelines define the mutation as LP with the following criteria: PVS1_Moderate, PP3_Strong.

Out of 18 analyzed tumors, 3 samples harbor a germline mutation on the aryl hydrocarbon receptor-interacting protein (AIP): S68, S62, and S59.

Sample S68 harbors a stop-gained genetic mutation denoted as c.910C>T(p.Arg304Ter). Reported in literature as pathogenic for familial isolated pituitary adenomas (Thorvaldsdottir et al., 2013).

Sample 62 was identified with a different AİP gene mutation, not yet reported in the Clinvar database. Supported by 58.62 percent of the nucleotides in that position, the c.646-1G>C is a splice_acceptor homozygote variant.

The last sample affected by an AIP mutation is 59: a missense variant denoted c.241C>T(p.Arg81Ter). This variant was previously reported in Clinvar (Variation ID: 41167) as Pathogenic for somatotroph Adenoma in 4 submissions.

3 Samples 59,68 and 70 share the c.2080A>G(p.Met694Val) missense germline mutation, on the MEFV gene. That rare mutation (MAF = 0.00028) is also the most recurrent variant detected in this study.

Patients harboring similar pathogenic mutations are likely to be affected more closely by the resulting disease. The function below estimates and quantifies the similarity level between two given samples, based on their respective variants. Variants presenting deleteriousness effects (or ACMG score) on the same gene are attributed a value greater than 0. In addition to affecting the same gene, they are assigned a higher value if they are in the same position and share the same alternative alleles. The more MAF is low the more this variant is clinically meaningful.

```
Function similarity(a, b):
    res = 0
    If a['GENE'] is not equal to b['GENE']:
        res = 0
    Else if a['GENE'] is equal to b['GENE']:
        If a['variant_ID'] is equal to b['variant_ID']:
            res = a['ACMG'] + int(-10 * log10(a['MAF']))
        Else:
            res = minimum of [a['ACMG'], b['ACMG']]
    Return res
```

The following figure is the expression of pairwise similarity levels across the 18 samples. The dendrogram is generated through Hierarchical clustering.

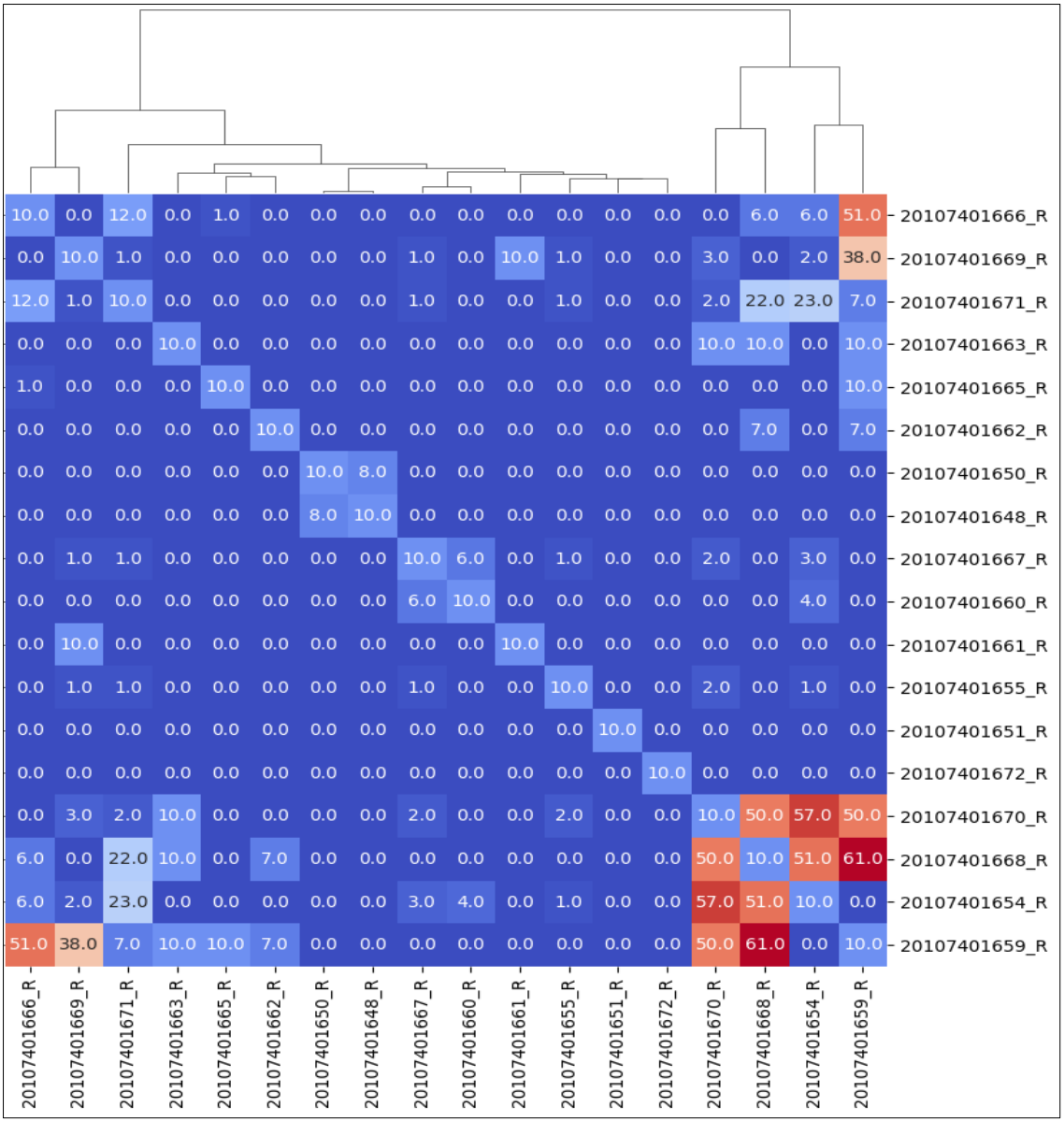


Figure 4.2. Hierarchical clustering and illustration of pairwise similarity levels.

A subgroup, constituted by samples S70, S68, S54, and S59 is identifiable in the lower right corner. Samples 54 and 68 exhibit a similar variant of the CAPN3 gene. Samples 70 and 54 share a similar pathogenic, rare variant in the gene TTN. Samples 70, 68, and 59 have a pathogenic variant in common (Table 4.2) and samples 59 and 68 have two genes (AIP and MEFV), and a rare variant (MAF = 0.0002823) in common (Table 4.2).

This thesis evaluates the performance of involved tools and compares the respective contributions to the obtained results. Following the variant calling and filtering steps, and prior to variant prioritization, a high concordance was observed between two variant mappers despite BWA-mem demonstrating a higher level of uniqueness (Figure 4.3).

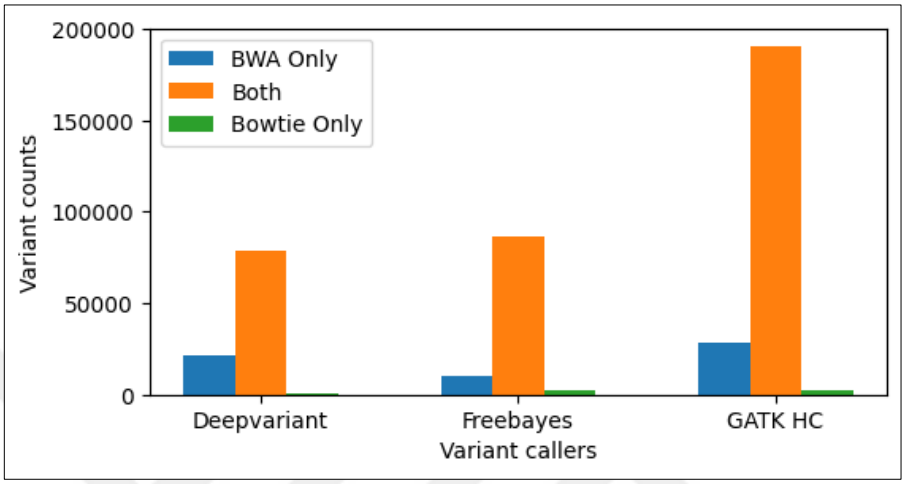


Figure 4.3. Mappers' contribution (before prioritization)

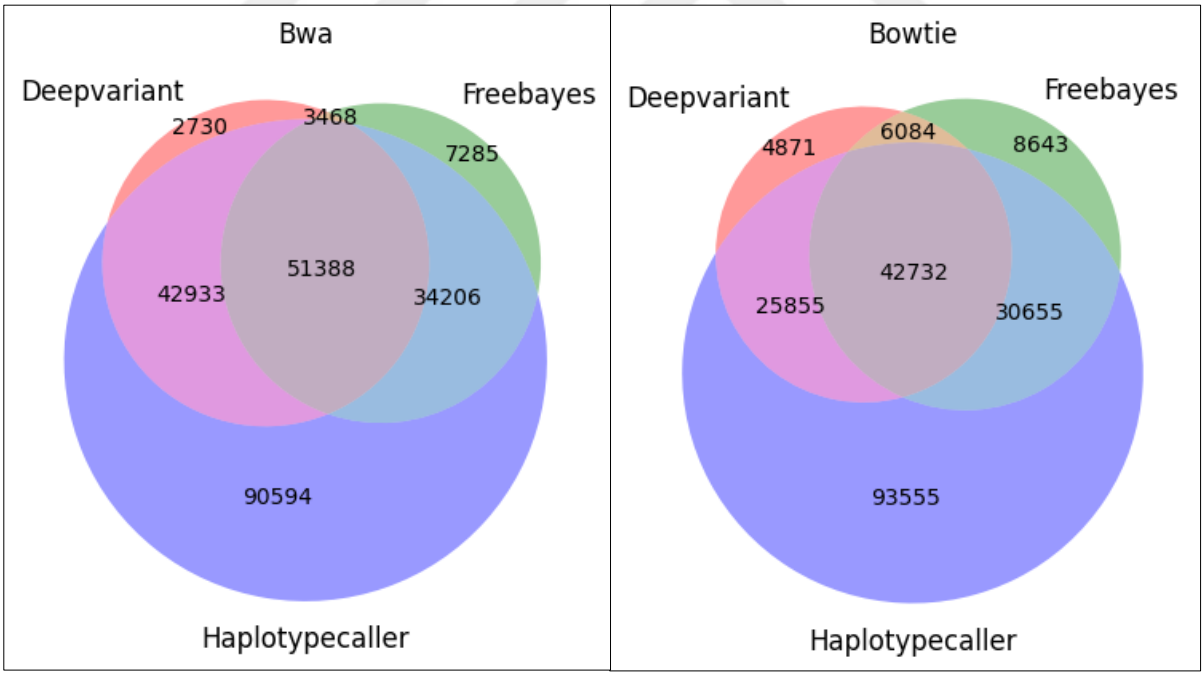


Figure 4.4. Variant callers' contributions before prioritization

The variant calling algorithms exhibit significant disparity. Common variants constitute less than one-third of the GATK Haplotypecaller's callset and approximately half of the callsets from the other algorithms (Deepvariant and Freebayes) (Figure 4.4).

I apologize for not making substantial modifications. Here is a more formally rephrased version:

Following variant prioritization, notable disparities were observed in the final results. BWA-mem-based workflows were implicated in all the VUS identifications, except for the combination with Freebayes. Workflows based on Freebayes did not detect any variants among the retained VUS (Figure 4.5).

Regarding Likely Pathogenic (LP) variants, the combination of BWA-mem and GATK Haplotypecaller identified 100% of the recorded variants. The Bowtie2-Haplotypecaller combination demonstrated nearly equivalent performance, successfully capturing 72 out of 74 variants (Figure 4.6).

About Pathogenic (P) variants, the situation resembles the previous (LP variants case), with a better performance from Bowtie2 and DeepVariant (17 or 65.38 percent) (Figure 4.7).

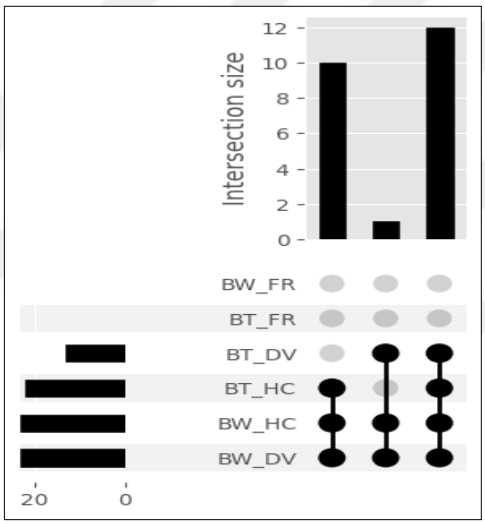


Figure 4.5. Number of common VUS variants among variant callers

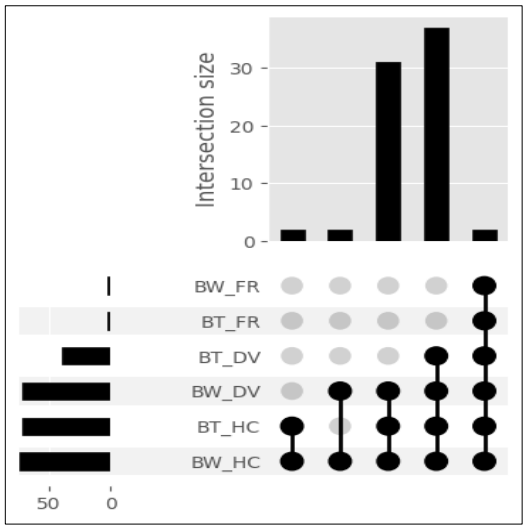


Figure 4.6. Number of common LP variants among variant callers

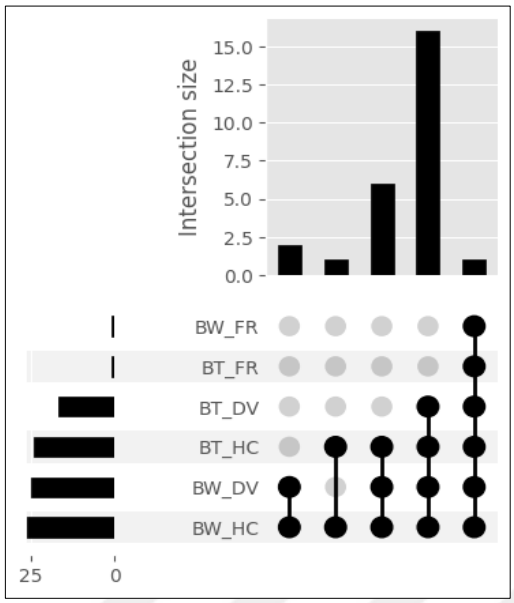


Figure 4.5. Number of common P variants among variant callers

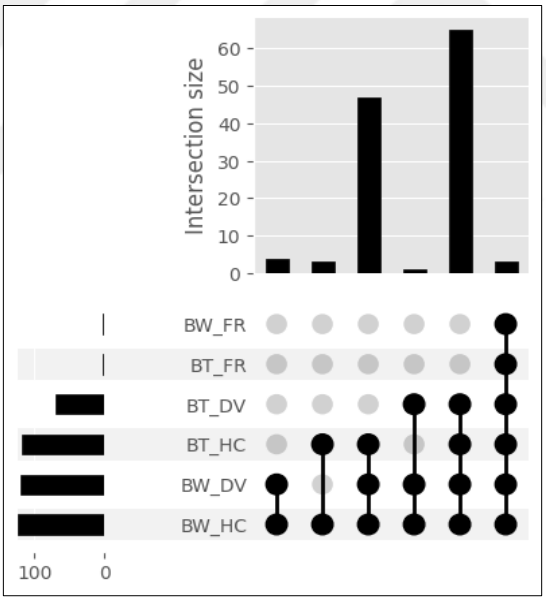


Figure 4.6. The overall number of common variants among variant callers

GATK Haplotypecaller and DeepVariant are implicated in a significant portion of the identified variants. The strategies employed in the Freebayes-based pipeline appear to be unsuitable for efficient variant calling in the current PA dataset.

IGV or Integrative Genomics Viewer (Thorvaldsdottir et al., 2013) helps in the validation of germline variants through visualization.. The next figures (Figures 4.9 to 4.13) depict two of the most recurrent variants.

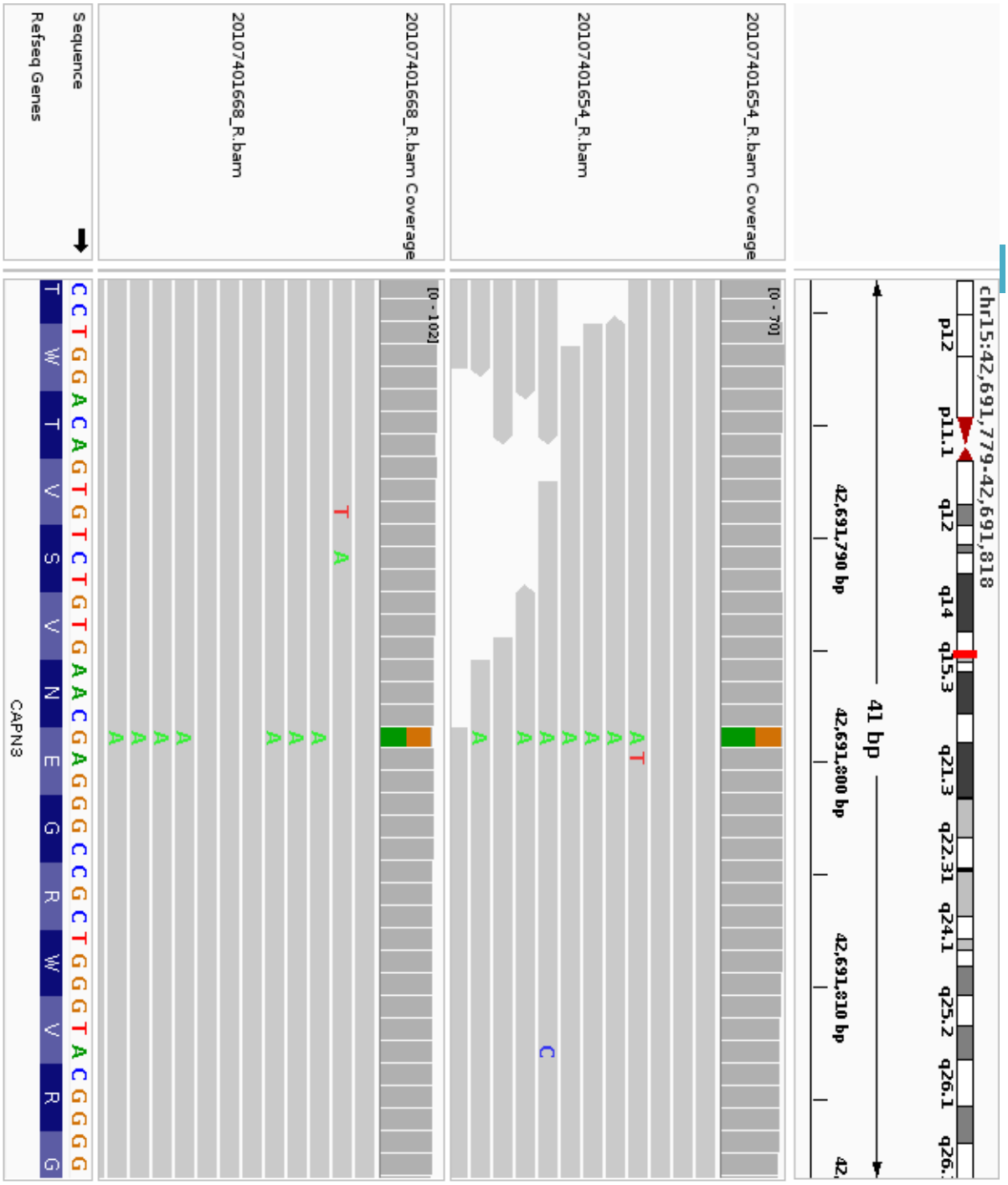


Figure 4.7. IGV view of variant 15-42691799-G-A on CAPN3 gene

Detected on samples S68 and S54, this is a heterozygote and missense variant resuting in the replacement of the Leucine (GAG) amino acid by Arginine (AAG) in the producted protein.

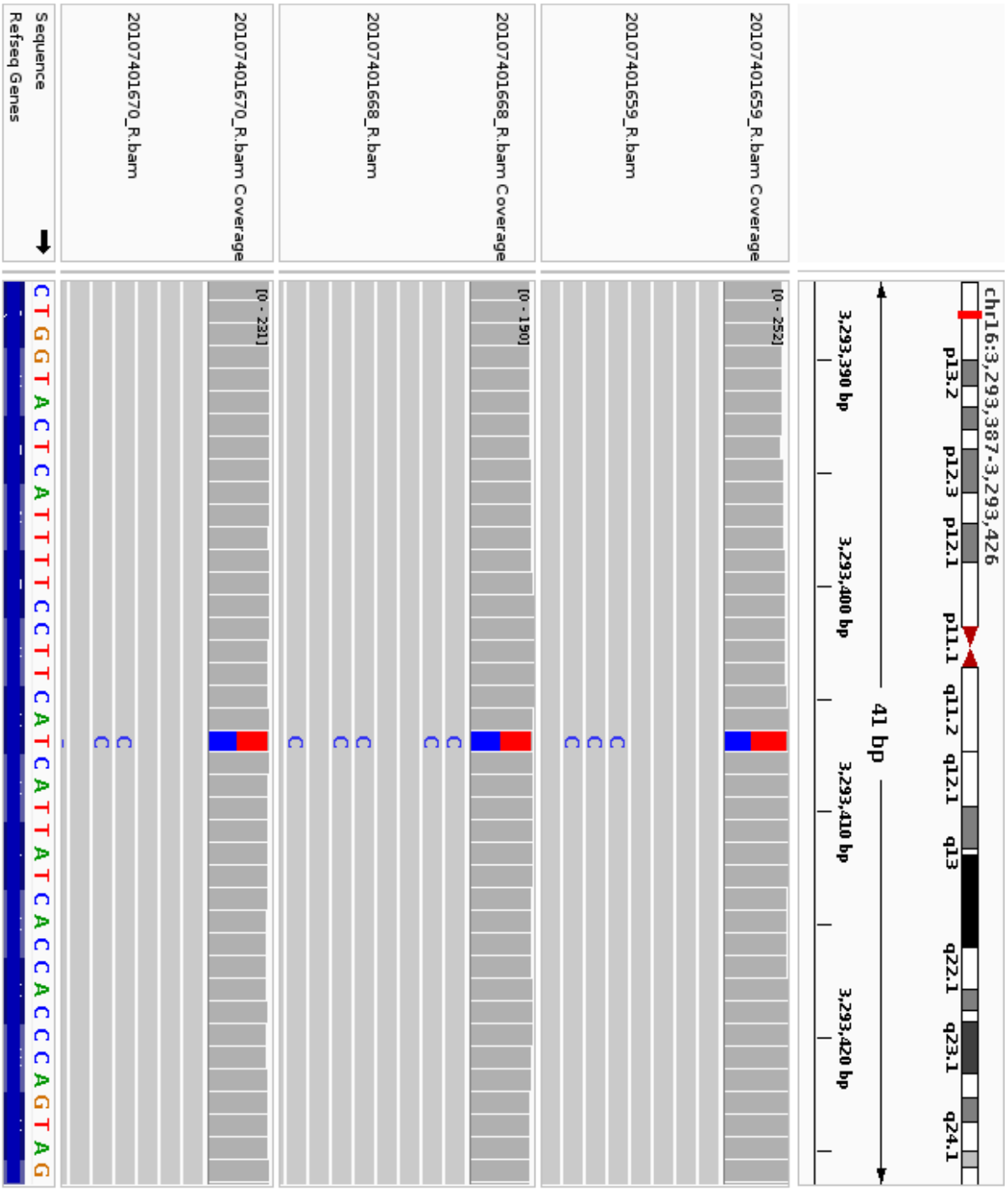


Figure 4.8. IGV view of the variant 16-3293407-T-C on MEFV gene

This variant was detected simultaneously on 3 samples: S70, S68, and S59. It is heterozygotes in all the samples, with approximately 50% of bases corresponding to the alternative allele.

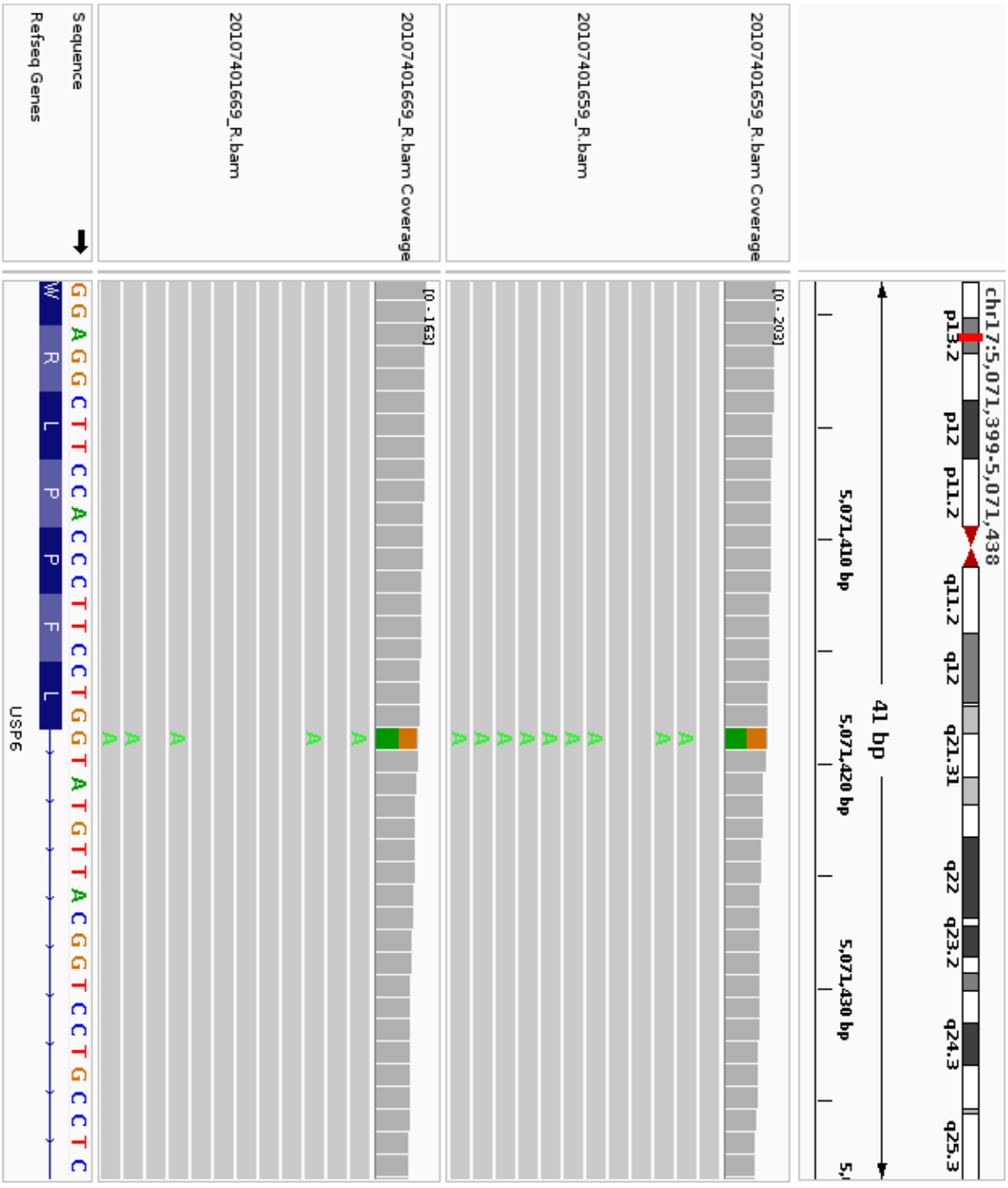


Figure 4.9. IGV view of the variant 17-5071419-G-A on USP6 gene

This mutation was detected in two samples, S59 and S69, both of which exhibit heterozygous variants. The mutation causes a splice-donor effect.

5. DISCUSSION AND CONCLUSION

Exome and genome sequencing technologies have significantly enhanced our ability to identify genes and genetic mutations contributing to complex disease mechanisms. Significant insights are generated and utilized in diagnostics, therapy, and other applications. In this thesis, whole exome sequencing data analysis was performed on 18 Pituitary adenoma samples. The results describe a total of 123 clinically significant genetic mutations located in 104 different genes.

5.1. OVERLAPPING GENES AND MUTATIONS

The literature reveals that among the detected genes, 17 genes harboring VUS genetic variants ('NEBL' 'AHNAK' 'SDHB' 'ID3' 'MAX' 'LAMA1' 'RYR1' 'LRP1B' 'TTN' 'PLA2G6' 'MSH2' 'NR3C1' 'MAST4' 'LAMA2' 'MGAM' 'CDKN2A' 'KLHL4'), and 4 genes harboring Pathogenic or Likely Pathogenic mutations (AIP, PYGM, RYR1, PLA2G6) were previously associated with PA disease. SDHB and MAX are genes commonly linked to 3P association syndrome (Tatsi and Stratakis, 2019). In their investigation of PA-associated genetic mechanisms, NEBL was among the mutated genes reported by Ye et al., (2015). Peculis et al., (2022) identified somatic variants in the following genes: LAMA1, PLA2G6, AHNAK, KLHL4, and RYR1. In a Moreno et al. (2005) study, the ID3 gene was down-regulated in many PA samples. Two missense variants in the LRP1B and MGAM genes and one frameshift mutation in the MAST4 gene were identified in invasive pituitary adenomas (Lan et al., 2016). TTN gene was the most recurrent gene (3 times) in an analysis of 42 pituitary macroadenomas (Bi et al., 2017). Bengtsson et al., (2017) identified germline mutations in the MSH2 gene while studying ACTH-secreting PA cases. The NR3C1 gene was cited in a study by Song et al., (2016), revealing its potential implication in ACTH-PA cases. Wang et al., (2019) conducted research focused on the LAMA2 gene and reported epigenetic dysfunction influencing the degree of severity of NFPA. Seemann et al., (2001) documented the inactivation of CDKN2A/p16 and its close relation with the type and size of pituitary adenomas.

SDHx are among the most prevalent genes cited in Familial Isolated Pituitary Adenomas (FIPA) and 3P syndromes (Tatsi and Stratakis, 2019). Additionally, the c.424 + 5G>C somatic variant was detected in the PYGM gene, from Growth Hormone (GH)-Secreting PA samples (Välimäki et al., 2015).

Many of the identified mutations do not match exactly those found in the previous studies, however their have clinically significant impact on genes. Additional research is necessary to elucidate their potential impact on the occurrence and behavior of PA disease.

We have validated two genetic mutations previously linked to pituitary adenomas. Both are located on the AIP gene. One is c.910C>T(p.Arg304Ter), a well-described mutation detected in sample 20107401668_R or S68 and supported by an alternative allele frequency of 52.63%. In their study, Niyazoglu et al., (2014) identified the identical heterozygous stop-gained germline mutation in 25% of the 52 family members, and seven of them had a history of pituitary adenoma. The second variant, c.241C>T(p.Arg81Ter) was detected in a GH-secreting PA sample representing a male 22 years old (Gaspar et al., 2023) and also in a case of a patient who was diagnosed with acromegaly type FIPA at the age of 34 years (Guaraldi et al., 2012). Sample 20107401659_R or S59 was the only one to exhibit this variant, which is described as an inactivating mutation that disrupts the interaction between the AIP protein and phosphodiesterase 4A (PDE4A) (Toledo et al., 2010).

Despite being a well-established gene in the context of PA, the AIP gene was detected in only three samples. The third sample (S62) harbors a non-established splice acceptor variant c.646-1G>C (58.62). Tuncer et al., (2018) also found a low frequency of the AIP gene (2.1%) in a study on Turkish patients with FPAs.

5.2. CONSOLIDATION STRATEGIES

This thesis utilizes a parallel multi-workflow pipeline incorporating two read mappers (Bowtie and BWA-mem) and three variant callers (GATH, Freebayes, DeepVariant). Previous studies, such as Peculis et al. (2022), described the utilization of multiple data analysis workflows. They differ in terms of the included tools, the number of workflows, and especially their generation of the final results strategies.

As reported by O'Rawe et al. (2013), The concordance of workflows can be quite low. The percentage of common SNVs across five pipelines was 57.4%, while the concordance for indels among three variant-calling pipelines was less than 27%. Here, using six workflows reduces significantly the number of common variants. Considering the four most effective workflows, the concordance level is close to 50%; however, if all workflows are considered, the concordance level drops to nearly zero. In the study by Peculis et al. (2022), involving two

workflows, 144 variants were collected by considering only the common variants for the final results. In the current study, all variants that passed the filtering steps were included in the final call set to avoid missing important variants not detected by lower-performing workflows.

We observed an average of 7.1 mutations per sample with a range of 2-19, which is lower than the findings of Peculis et al., (2022), who reported 9.6 mutations per sample with a range of 0–31 per PitNET. In contrast, other PA studies show a relatively lower number of mutations per sample, such as 3.5 mutations per tumor reported by Newey et al., (2013), 2.3 coding SNVs or short deletions (range : 0–4) reported by Välimäki et al. (2015) and a mean of 3.3 mutations per gene across 7PA in a study by Song et al., (2016).

5.3. RECURRENCE OF GENOMIC MUTATIONS

Out of 123 detected variants, 37 are on recurrent genes, but only 5 were detected in more than one Sample. Peculis et al., (2022) also found only 2 recurrent variants from a study on 14 PA samples. The low occurrence of recurrent gene mutations can be attributed to the limited number of samples or the complexity of the disease genomic landscape.

The most recurrent variant is c.2080A>G(p.Met694Val), a missense variant of the position 3293407 of MEFV genes. No document associating the MEFV gene with the PA case was found in the literature. Its pathogenic effect was described in conditions such as Mediterranean fever and hereditary periodic fever (Clincar ID: 2538).

In a study by Bi et al., (2017), the TTN gene was the sole gene mutated in three samples. Observed in six samples, TTN is the most frequent gene in the current research. Samples S54 and S70 share the same TTN mutation (c.12066T>G, p.Cys4022Trp). This is a rare mutation (with a Minor Allele Frequency of 0.00000342), documented in Clinvar (ID: 466797), and associated with conditions such as Autosomal Recessive Limb-Girdle Muscular Dystrophy Type 2J and Dilated Cardiomyopathy 1G.

Sample S54 shares the missense mutation c.1303G>A (p.Glu435Lys) on the CAPN3 gene with S68. While the CAPN3 gene itself is not directly associated with PA, it influences the expression levels of RYR1 (Chen et al., 2021). RYR1 is a gene strongly linked to PA (Peculis et al., 2022) and reported in samples S54 and S67. Altogether, CAPN3 carries five rare variants

in our research: four pathogenic (P) and one likely pathogenic (LP). This highlights the importance of that gene in current findings.

The c.5260C>T (p.Arg1754Ter) nonsense change on the ABCA7 gene is a likely pathogenic (LP) variant reported simultaneously in samples 59 and 66. This rare variant (MAF = 0.00004381) is cited in the literature for its implication on Alzheimer's disease (Nuytemans et al., 2016), (Sassi et al., 2016). The lack of clinical data on patients reduces our ability to assess their potential association with a specific PA type.

5.4. LIMITATIONS

Due to the lack of the sample's details, no analysis of the potential relationship between specific genetic variants and the clinical manifestations was possible. Instead, a similarity analysis was performed, revealing a small cluster of four samples (S70, S68, S54, and S59). They are linked by mutations in the following genes: CAPN3, TTN, AIP, and MEFV. The recurrence of rare variants is a low-probability event and research can focus more on developing a broader classification approach for Pas. The current classifications are based on tumor size, secreted hormone, and grade (Raverot et al., 2014). New classification approaches can integrate clinical, genetic, biochemical, radiological, pathological, and molecular information (Ho et al., 2024).

Unlike previous studies on various pituitary adenomas, we did not find mutations in previously cited genes such as MEN1(Yarman et al., 2019) and GNAS(Tatsi and Stratakis, 2019).

5.5. CONCLUSION

This thesis explores the topic of whole exome sequencing (WES) and associated data analysis methods. A dataset consisting of 18 WES samples of Pituitary Adenoma was analyzed. Two read mappers and three variant callers were included in the analysis pipeline, resulting in 6 parallel analysis workflows. The final variant callset includes all mutations that meet the prioritization and filtering criteria.

We observed a significant number of genes (n = 107), harboring clinically significant mutations. A total of 123 variants were listed with 37 of them being located in recurrent gene. Only 5 variants were recurrent. The low recurrence of genomic mutations across the samples despite many of them being located on recurrent genes suggests an intricate nature of the disease's

genomic landscape. We advocates for a disease classification approach relying on the insights generated by omic approaches, for additional and precise characterization of patients.

We evaluated the effectiveness of individual tools by assessing their standalone performance and examining how they interact when combined with other tools. For example, although the Bowtie mapping tool performs less effectively in FreeBayes-based workflows, it shows a remarkable increase in performance when combined with DeepVariant or GATK HaplotypeCaller.

Additional research should help to clarify the potential impact of detected variants on PA disease, particularly the recurrent ones. Pipeline optimization often focuses on parallelizing similar tasks, such as processing multiple samples simultaneously. Further investigation can explore the parallelization of diverse tasks, such as read quality control and genome indexing, which are independent.

REFERENCES

Abdullah, T., Ahmet, A., 2020. Extracting Insights: A Data Centre Architecture Approach in Million Genome Era, in: Hameurlain, A., Tjoa, A.M. (Eds.), Transactions on Large-Scale Data- and Knowledge-Centered Systems XLVI, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–31. https://doi.org/10.1007/978-3-662-62386-2_1

Acromegaly (Growth Hormone–Secreting Pituitary Adenomas) | Memorial Sloan Kettering Cancer Center [WWW Document], n.d. URL https://www.mskcc.org/cancer-care/types/pituitary-tumors/pituitary-adenomas/growth-hormone-producing-tumors (accessed 2.26.24).

Aflorei, E.D., Korbonits, M., 2014. Epidemiology and etiopathogenesis of pituitary adenomas. J Neurooncol 117, 379–394. https://doi.org/10.1007/s11060-013-1354-5

Agarwal, S.K., 2013. Multiple Endocrine Neoplasia Type 1, in: Stratakis, C.A. (Ed.), Frontiers of Hormone Research. S. Karger AG, pp. 1–15. https://doi.org/10.1159/000345666

Ahmed, Z., Renart, E.G., Mishra, D., Zeeshan, S., 2021. JWES: a new pipeline for whole genome/exome sequence data processing, management, and gene- variant discovery, annotation, prediction, and genotyping. FEBS Open Bio 11, 2441–2452. https://doi.org/10.1002/2211-5463.13261

Andrews, n.d. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document]. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 5.3.24).

Approximate sizes of sequencing run output folders - Illumina Knowledge [WWW Document], n.d. URL https://knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001508 (accessed 3.8.24).

Barbitoff, Y.A., Abasov, R., Tvorogova, V.E., Glotov, A.S., Predeus, A.V., 2022. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics 23, 155. https://doi.org/10.1186/s12864-022-08365-3

Barkhoudarian, G., Kelly, D.F., 2017. The Pituitary Gland: Anatomy, Physiology, and its Function as the Master Gland, in: Cushing's Disease. Elsevier, pp. 1–41. https://doi.org/10.1016/B978-0-12-804340-0.00001-2

Bengtsson, D., Joost, P., Aravidis, C., Askmalm Stenmark, M., Backman, A.-S., Melin, B., Von Salomé, J., Zagoras, T., Gebre-Medhin, S., Burman, P., 2017. Corticotroph Pituitary Carcinoma in a Patient With Lynch Syndrome (LS) and Pituitary Tumors in a Nationwide LS Cohort. The Journal of Clinical Endocrinology & Metabolism 102, 3928–3932. https://doi.org/10.1210/jc.2017-01401

Bi, W.L., Horowitz, P., Greenwald, N.F., Abedalthagafi, M., Agarwalla, P.K., Gibson, W.J., Mei, Y., Schumacher, S.E., Ben-David, U., Chevalier, A., Carter, S., Tiao, G.,

Brastianos, P.K., Ligon, A.H., Ducar, M., MacConaill, L., Laws, E.R., Santagata, S., Beroukhim, R., Dunn, I.F., 2017. Landscape of Genomic Alterations in Pituitary Adenomas. Clinical Cancer Research 23, 1841–1851. https://doi.org/10.1158/1078-0432.CCR-16-0790

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bollas, A.E., Rajkovic, A., Ceyhan, D., Gaither, J.B., Mardis, E.R., White, P., 2024. SNVstory: inferring genetic ancestry from genome sequencing data. BMC Bioinformatics 25, 76. https://doi.org/10.1186/s12859-024-05703-y

Burrows, M., Wheeler, D.J., 1994. A block-sorting lossless data compression algorithm (Technical Report No. 124). Digital Equipment Corporation, Palo Alto, CA, USA.

Byun, J., Kim, J.H., Kim, Y.-H., Cho, Y.H., Hong, S.H., Kim, C.J., 2020. Thyroid-Stimulating Hormone-Secreting Pituitary Adenomas : Single Institutional Experience of 14 Consecutive Cases. J Korean Neurosurg Soc 63, 495–503. https://doi.org/10.3340/jkns.2019.0169

Cain, J.W., Miljic, D., Popovic, V., Korbonits, M., 2010. Role of the aryl hydrocarbon receptor-interacting protein in familial isolated pituitary adenoma. Expert Review of Endocrinology & Metabolism 5, 681–695. https://doi.org/10.1586/eem.10.42

Chanson, P., Brochier, S., 2005. Non-functioning pituitary adenomas. J Endocrinol Invest 28, 93–99.

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., Dong, L., Zhang, Zhewen, Yu, C., Sun, Yanling, Chi, L., Chen, H., Zhai, S., Sun, Yubin, Lan, L., Zhang, X., Xiao, J., Bao, Y., Wang, Y., Zhang, Zhang, Zhao, W., 2021. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. Genomics, Proteomics & Bioinformatics 19, 578–583. https://doi.org/10.1016/j.gpb.2021.08.001

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., Lifton, R.P., 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U.S.A. 106, 19096–19101. https://doi.org/10.1073/pnas.0910672106

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38, 1767–1771. https://doi.org/10.1093/nar/gkp1137

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daniel, P.M., 1976. Anatomy of the hypothalamus and pituitary gland. Journal of Clinical Pathology s1-7, 1–7. https://doi.org/10.1136/jcp.s1-7.1.1

D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Elmi, B., Pallocca, M., Sanna, N., Picardi, E., Pesole, G., Castrignanò, T., 2013. WEP: a high-performance analysis pipeline for whole-exome data. BMC Bioinformatics 14, S11. https://doi.org/10.1186/1471-2105-14-S7-S11

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498. https://doi.org/10.1038/ng.806

Erol, Ç., Gezer, M., 2020. Data Collection Approaches for Artificial Intelligence Applications in Healthcare, in: Who Runs the World: Data. Istanbul University Press, pp. 227–237. https://doi.org/10.26650/B/ET06.2020.011.13

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ezzat, S., Asa, S.L., Couldwell, W.T., Barr, C.E., Dodge, W.E., Vance, M.L., McCutcheon, I.E., 2004. The prevalence of pituitary adenomas: a systematic review. Cancer 101, 613–619. https://doi.org/10.1002/cncr.20412

FASTA - Reference genome format – GATK [WWW Document], n.d. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035531652-FASTA-Reference-genome-format (accessed 4.6.24).

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. AIMag 17, 37. https://doi.org/10.1609/aimag.v17i3.1230

Ferragina, P., Manzini, G., 2004. The FM-Index: A Compressed Full-Text Index Based on the BWT.

Frischmeyer, P.A., 1999. Nonsense-mediated mRNA decayin health and disease. Human Molecular Genetics 8, 1893–1900. https://doi.org/10.1093/hmg/8.10.1893

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing.

Gaspar, L.M., Gonçalves, C.I., Saraiva, C., Cortez, L., Amaral, C., Nobre, E., Lemos, M.C., 2023. Low frequency of AIP mutations in patients with young-onset sporadic pituitary macroadenomas. J Endocrinol Invest 46, 2299–2307. https://doi.org/10.1007/s40618-023-02083-7

GitHub - freebayes/freebayes: Bayesian haplotype-based genetic polymorphism discovery and genotyping. [WWW Document], n.d. URL https://github.com/freebayes/freebayes (accessed 7.3.24).

Guaraldi, F., Corazzini, V., Gallia, G.L., Grottoli, S., Stals, K., Dalantaeva, N., Frohman, L.A., Korbonits, M., Salvatori, R., 2012. Genetic analysis in a patient presenting with meningioma and familial isolated pituitary adenoma (FIPA) reveals selective involvement of the R81X mutation of the AIP gene in the pathogenesis of the pituitary tumor. Pituitary 15 Suppl 1, S61-67. https://doi.org/10.1007/s11102-012-0391-y

Guerrero-Pérez, F., Fajardo, C., Torres Vela, E., Giménez-Palop, O., Lisbona Gil, A., Martín, T., González, N., Díez, J.J., Iglesias, P., Robledo, M., Villabona, C., 2019. 3P association (3PAs): Pituitary adenoma and pheochromocytoma/paraganglioma. A heterogeneous clinical syndrome associated with different gene mutations. European Journal of Internal Medicine 69, 14–19. https://doi.org/10.1016/j.ejim.2019.08.005

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D.C., Shyr, Y., 2012. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13, 666. https://doi.org/10.1186/1471-2164-13-666

HaplotypeCaller in a nutshell – GATK [WWW Document], n.d. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell (accessed 7.3.24).

Hard-filtering germline short variants – GATK [WWW Document], n.d. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants (accessed 4.22.24).

Ho, K.K.Y., Fleseriu, M., Wass, J., Katznelson, L., Raverot, G., Little, A.S., Castaño, J.P., Reincke, M., Lopes, M.B., Kaiser, U.B., Chanson, P., Gadelha, M., Melmed, S., 2024. A proposed clinical classification for pituitary neoplasms to guide therapy and prognosis. The Lancet Diabetes & Endocrinology 12, 209–214. https://doi.org/10.1016/S2213-8587(23)00382-0

Horvath, E., Kovacs, K., 1988. Pituitary Gland. Pathology - Research and Practice 183, 129–142. https://doi.org/10.1016/S0344-0338(88)80041-4

Jia, T., Munson, B., Lango Allen, H., Ideker, T., Majithia, A.R., 2020. Thousands of missing variants in the UK Biobank are recoverable by genome realignment. Annals of Human Genetics 84, 214–220. https://doi.org/10.1111/ahg.12383

Jorde, L.B., Wooding, S.P., 2004. Genetic variation, classification and "race." Nat Genet 36, S28–S33. https://doi.org/10.1038/ng1435

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski,

M.E., Genome Aggregation Database Consortium, Aguilar Salinas, C.A., Ahmad, T., Albert, C.M., Ardissino, D., Atzmon, G., Barnard, J., Beaugerie, L., Benjamin, E.J., Boehnke, M., Bonnycastle, L.L., Bottinger, E.P., Bowden, D.W., Bown, M.J., Chambers, J.C., Chan, J.C., Chasman, D., Cho, J., Chung, M.K., Cohen, B., Correa, A., Dabelea, D., Daly, M.J., Darbar, D., Duggirala, R., Dupuis, J., Ellinor, P.T., Elosua, R., Erdmann, J., Esko, T., Färkkilä, M., Florez, J., Franke, A., Getz, G., Glaser, B., Glatt, S.J., Goldstein, D., Gonzalez, C., Groop, L., Haiman, C., Hanis, C., Harms, M., Hiltunen, M., Holi, M.M., Hultman, C.M., Kallela, M., Kaprio, J., Kathiresan, S., Kim, B.-J., Kim, Y.J., Kirov, G., Kooner, J., Koskinen, S., Krumholz, H.M., Kugathasan, S., Kwak, S.H., Laakso, M., Lehtimäki, T., Loos, R.J.F., Lubitz, S.A., Ma, R.C.W., MacArthur, D.G., Marrugat, J., Mattila, K.M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Meigs, J.B., Melander, O., Metspalu, A., Neale, B.M., Nilsson, P.M., O'Donovan, M.C., Ongur, D., Orozco, L., Owen, M.J., Palmer, C.N.A., Palotie, A., Park, K.S., Pato, C., Pulver, A.E., Rahman, N., Remes, A.M., Rioux, J.D., Ripatti, S., Roden, D.M., Saleheen, D., Salomaa, V., Samani, N.J., Scharf, J., Schunkert, H., Shoemaker, M.B., Sklar, P., Soininen, H., Sokol, H., Spector, T., Sullivan, P.F., Suvisaari, J., Tai, E.S., Teo, Y.Y., Tiinamaija, T., Tsuang, M., Turner, D., Tusie-Luna, T., Vartiainen, E., Vawter, M.P., Ware, J.S., Watkins, H., Weersma, R.K., Wessman, M., Wilson, J.G., Xavier, R.J., Neale, B.M., Daly, M.J., MacArthur, D.G., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Khoury, M.J., Holt, K.E., 2021. The impact of genomics on precision public health: beyond the pandemic. Genome Med 13, 67, s13073-021-00886-y. https://doi.org/10.1186/s13073-021-00886-y

Korhonen, P.K., Shaban, B., Faux, N.G., Kinkar, L., Chang, B.C.H., Wang, D., Yang, B., Young, N.D., Gasser, R.B., 2022. " *Escalibur* "—A practical pipeline for the de novo analysis of nucleotide variation in nonmodel eukaryotes. Molecular Ecology Resources 22, 2120–2126. https://doi.org/10.1111/1755-0998.13600

Krueger, F., James, F., Ewels, P., Afyounian, E., Weinstein, M., Schuster-Boeckler, B., Hulselmans, G., Sclamons, 2023. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. https://doi.org/10.5281/ZENODO.5127898

Kurtz, S., 1999. Reducing the space requirement of suffix trees. Softw: Pract. Exper. 29, 1149–1171. https://doi.org/10.1002/(SICI)1097-024X(199911)29:13<1149::AID-SPE274>3.0.CO;2-O

Lan, X., Gao, H., Wang, F., Feng, J., Bai, J., Zhao, P., Cao, L., Gui, S., Gong, L., Zhang, Y., 2016. Whole-exome sequencing identifies variants in invasive pituitary adenomas. Oncology Letters 12, 2319–2328. https://doi.org/10.3892/ol.2016.5029

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Trans. Visual. Comput. Graphics 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, H., 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/ARXIV.1303.3997

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, D., Li, J., Li, N., Lu, M., Wen, S., Zhan, X., 2020. Integration of quantitative phosphoproteomics and transcriptomics revealed phosphorylation-mediated molecular events as useful tools for a potential patient stratification and personalized treatment of human nonfunctional pituitary adenomas. EPMA Journal 11, 419–467. https://doi.org/10.1007/s13167-020-00215-0

Liu, X., Jian, X., Boerwinkle, E., 2013. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human Mutation 34, E2393–E2402. https://doi.org/10.1002/humu.22376

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 17, 10. https://doi.org/10.14806/ej.17.1.200

Mehta, G.U., Lonser, R.R., 2016. Management of hormone-secreting pituitary adenomas. Neuro Oncol now130. https://doi.org/10.1093/neuonc/now130

Musich, R., Cadle-Davidson, L., Osier, M.V., 2021. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Front. Plant Sci. 12, 657240. https://doi.org/10.3389/fpls.2021.657240

Navarro, F.C.P., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., Gerstein, M., 2019. Genomics and data science: an application within an umbrella. Genome Biol 20, 109. https://doi.org/10.1186/s13059-019-1724-1

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Niyazoglu, M., Sayitoglu, M., Firtina, S., Hatipoglu, E., Gazioglu, N., Kadioglu, P., 2014. Familial acromegaly due to aryl hydrocarbon receptor-interacting protein (AIP) gene mutation in a Turkish cohort. Pituitary 17, 220–226. https://doi.org/10.1007/s11102-013-0493-1

NovaSeq X Series Enables Broader, Deeper Sequencing [WWW Document], n.d. URL https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus/applications/broad-sequencing.html (accessed 4.22.24).

Nuytemans, K., Maldonado, L., Ali, A., John-Williams, K., Beecham, G.W., Martin, E., Scott, W.K., Vance, J.M., 2016. Overlap between Parkinson disease and Alzheimer disease in ABCA7 functional variants. Neurol Genet 2, e44. https://doi.org/10.1212/NXG.0000000000000044

Paten, B., Novak, A.M., Eizenga, J.M., Garrison, E., 2017. Genome graphs and the evolution of genome inference. Genome Res. 27, 665–676. https://doi.org/10.1101/gr.214155.116

Peculis, R., Rovite, V., Megnis, K., Balcere, I., Breiksa, A., Nazarovs, J., Stukens, J., Konrade, I., Sokolovska, J., Pirags, V., Klovins, J., 2022. Whole exome sequencing reveals novel risk genes of pituitary neuroendocrine tumors. PLoS ONE 17, e0265306. https://doi.org/10.1371/journal.pone.0265306

Peng, Jiajie, Bao, Z., Li, J., Han, R., Wang, Y., Han, L., Peng, Jinghao, Wang, T., Hao, J., Wei, Z., Shang, X., 2024. DeepRisk: A deep learning approach for genome-wide assessment of common disease risk. Fundamental Research S2667325824001171. https://doi.org/10.1016/j.fmre.2024.02.015

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A., 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987. https://doi.org/10.1038/nbt.4235

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van Der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. https://doi.org/10.1101/201178

Raverot, G., Jouanneau, E., Trouillas, J., 2014. MANAGEMENT OF ENDOCRINE DISEASE: Clinicopathological classification and molecular markers of pituitary tumours for personalized therapeutic strategies. European Journal of Endocrinology 170, R121–R132. https://doi.org/10.1530/EJE-13-1031

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H.L., 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine 17, 405–424. https://doi.org/10.1038/gim.2015.30

Sassi, C., Nalls, M.A., Ridge, P.G., Gibbs, J.R., Ding, J., Lupton, M.K., Troakes, C., Lunnon, K., Al-Sarraj, S., Brown, K.S., Medway, C., Clement, N., Lord, J., Turton, J., Bras, J., Almeida, M.R., ARUK Consortium, Holstege, H., Louwersheimer, E., van der Flier, W.M., Scheltens, P., Van Swieten, J.C., Santana, I., Oliveira, C., Morgan, K., Powell, J.F., Kauwe, J.S., Cruchaga, C., Goate, A.M., Singleton, A.B., Guerreiro, R., Hardy, J.,

2016. ABCA7 p.G215S as potential protective factor for Alzheimer's disease. Neurobiol Aging 46, 235.e1–9. https://doi.org/10.1016/j.neurobiolaging.2016.04.004

Seemann, N., Kuhn, D., Wrocklage, C., Keyvani, K., Hackl, W., Buchfelder, M., Fahlbusch, R., Paulus, W., 2001. CDKN2A/p16 inactivation is related to pituitary adenoma type and size. J. Pathol. 193, 491–497. https://doi.org/10.1002/path.833

Sequencing Quality Scores [WWW Document], n.d. URL https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html (accessed 3.28.24).

Sherry, S.T., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29, 308–311. https://doi.org/10.1093/nar/29.1.308

Shestak, A.G., Bukaeva, A.A., Saber, S., Zaklyazminskaya, E.V., 2021. Allelic Dropout Is a Common Phenomenon That Reduces the Diagnostic Yield of PCR-Based Sequencing of Targeted Gene Panels. Front. Genet. 12, 620337. https://doi.org/10.3389/fgene.2021.620337

Shnayder, N.A., Novitsky, M.A., Neznanov, N.G., Limankin, O.V., Asadullin, A.R., Petrov, A.V., Dmitrenko, D.V., Narodova, E.A., Popenko, N.V., Nasyrova, R.F., 2022. Genetic Predisposition to Schizophrenia and Depressive Disorder Comorbidity. Genes 13, 457. https://doi.org/10.3390/genes13030457

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Song, Z.-J., Reitman, Z.J., Ma, Z.-Y., Chen, J.-H., Zhang, Q.-L., Shou, X.-F., Huang, C.-X., Wang, Y.-F., Li, S.-Q., Mao, Y., Zhou, L.-F., Lian, B.-F., Yan, H., Shi, Y.-Y., Zhao, Y., 2016. The genome-wide mutational landscape of pituitary adenomas. Cell Res 26, 1255–1259. https://doi.org/10.1038/cr.2016.114

Stawiński, P., Płoski, R., 2024. Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. Clinical Genetics cge.14516. https://doi.org/10.1111/cge.14516

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E., 2015. Big Data: Astronomical or Genomical? PLoS Biol 13, e1002195. https://doi.org/10.1371/journal.pbio.1002195

Syro, L.V., Scheithauer, B.W., Kovacs, K., Toledo, R.A., Londoño, F.J., Ortiz, L.D., Rotondo, F., Horvath, E., Uribe, H., 2012. Pituitary tumors in patients with MEN1 syndrome. Clinics 67, 43–48. https://doi.org/10.6061/clinics/2012(Sup01)09

Tatsi, C., Stratakis, C.A., 2019. The Genetics of Pituitary Adenomas. JCM 9, 30. https://doi.org/10.3390/jcm9010030

Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178–192. https://doi.org/10.1093/bib/bbs017

Toledo, R.A., Mendonca, B.B., Fragoso, M.C.B.V., Soares, I.C., Almeida, M.Q., Moraes, M.B., Lourenço, D.M., Alves, V.A.F., Bronstein, M.D., Toledo, S.P.A., 2010. Isolated familial somatotropinoma: 11q13-loh and gene/protein expression analysis suggests a possible involvement of aip also in non-pituitary tumorigenesis. Clinics (Sao Paulo) 65, 407–415. https://doi.org/10.1590/S1807-59322010000400010

Trump, D., Farren, B., Wooding, C., Pang, J.T., Besser, G.M., Buchanan, K.D., Edwards, C.R., Heath, D.A., Jackson, C.E., Jansen, S., Lips, K., Monson, J.P., O'Halloran, D., Sampson, J., Shalet, S.M., Wheeler, M.H., Zink, A., Thakker, R.V., 1996. Clinical studies of multiple endocrine neoplasia type 1 (MEN1). QJM 89, 653–670. https://doi.org/10.1093/qjmed/89.9.653

Tuncer, F.N., Çiftçi Doğanşen, S., Serbest, E., Tanrıkulu, S., Ekici, Y., Bilgiç, B., Yarman, S., 2018. Screening of AIP Gene Variations in a Cohort of Turkish Patients with Young-Onset Sporadic Hormone-Secreting Pituitary Adenomas. Genet Test Mol Biomarkers 22, 702–708. https://doi.org/10.1089/gtmb.2018.0133

Turkish National Science e-Infrastructure – TRUBA [WWW Document], n.d. URL https://www.truba.gov.tr/index.php/en/main-page/ (accessed 5.24.24).

Verloes, A., Stevenaert, A., Teh, B.T., Petrossians, P., Beckers, A., 1999. Familial acromegaly: case report and review of the literature. Pituitary 1, 273–277. https://doi.org/10.1023/A:1009958510378

Verrou, K.-M., Pavlopoulos, G.A., Moulos, P., 2022. Protocol for unbiased, consolidated variant calling from whole exome sequencing data. STAR Protocols 3, 101418. https://doi.org/10.1016/j.xpro.2022.101418

Wang, R.-Q., Lan, Y.-L., Lou, J.-C., Lyu, Y.-Z., Hao, Y.-C., Su, Q.-F., Ma, B.-B., Yuan, Z.-B., Yu, Z.-K., Zhang, H.-Q., Wang, D.-S., Zhu, T.-Z., Ding, Y., Zhang, N., Zhang, B., 2019. Expression and methylation status of LAMA2 are associated with the invasiveness of nonfunctioning PitNET. Therapeutic Advances in Endocrinology 10, 204201881882129. https://doi.org/10.1177/2042018818821296

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., Watson, M., 2015. Exome Sequencing: Current and Future Perspectives. G3 Genes|Genomes|Genetics 5, 1543–1550. https://doi.org/10.1534/g3.115.018564

Welcome To Colab - Colab [WWW Document], n.d. URL https://colab.research.google.com/ (accessed 5.24.24).

Xu, H., 2020. Big data challenges in genomics, in: Handbook of Statistics. Elsevier, pp. 337–348. https://doi.org/10.1016/bs.host.2019.08.002

Yarman, S., Tuncer, F.N., Serbest, E., 2019. Three Novel ***MEN1*** Variants in ***AIP***-Negative Familial Isolated Pituitary Adenoma Patients. Pathobiology 86, 128–134. https://doi.org/10.1159/000495252

Yatavelli, R.K.R., Bhusal, K., 2024. Prolactinoma, in: StatPearls. StatPearls Publishing, Treasure Island (FL).

Ye, Z., Li, Z., Wang, Yongfei, Mao, Y., Shen, M., Zhang, Q., Li, S., Zhou, L., Shou, X., Chen, J., Song, Z., Ma, Z., Zhang, Z., Li, Y., Ye, H., Huang, C., Wang, T., He, W., Zhang, Yichao, Xie, R., Qiao, N., Qiu, H., Huang, S., Wang, M., Shen, J., Wen, Z., Li, W., Liu, K., Zhou, J., Wang, L., Ji, J., Wang, Yin, Chen, H., Cheng, H., Shi, Z., Zhu, Y., Geng, D., Yao, Z., Tang, W., Lu, B., Pan, L., Zhang, Yi, Bao, W., Wu, J., Zheng, K., Shi, Y., Zhao, Y., 2015. Common variants at 10p12.31, 10q21.1 and 13q12.13 are associated with sporadic pituitary adenoma. Nat Genet 47, 793–797. https://doi.org/10.1038/ng.3322

Zverinova, S., Guryev, V., 2022. Variant calling: Considerations, practices, and developments. Human Mutation 43, 976–985. https://doi.org/10.1002/humu.24311

CURRICULUM VITAE

Personal Information	
Name Surname	Aristide IRAKOZE
Place of Birth	
Date of Birth	
Nationality	❒ T.C. ☑ Other:
E-Mail Adress	
Web Page	

Education Information	
Undergraduate	
University	University of Burundi
Faculty	Faculty of Engineering Sciences
Department	ICT Department
Year of Graduation	Tarih girmek için tıklayın veya dokunun.

Masters	
University	İstanbul Üniversitesi
Name of Institute	Institute Of Graduate Studies In Sciences
Department	Informatics
Program	Informatics

Articles and Publications