

# UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA  
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

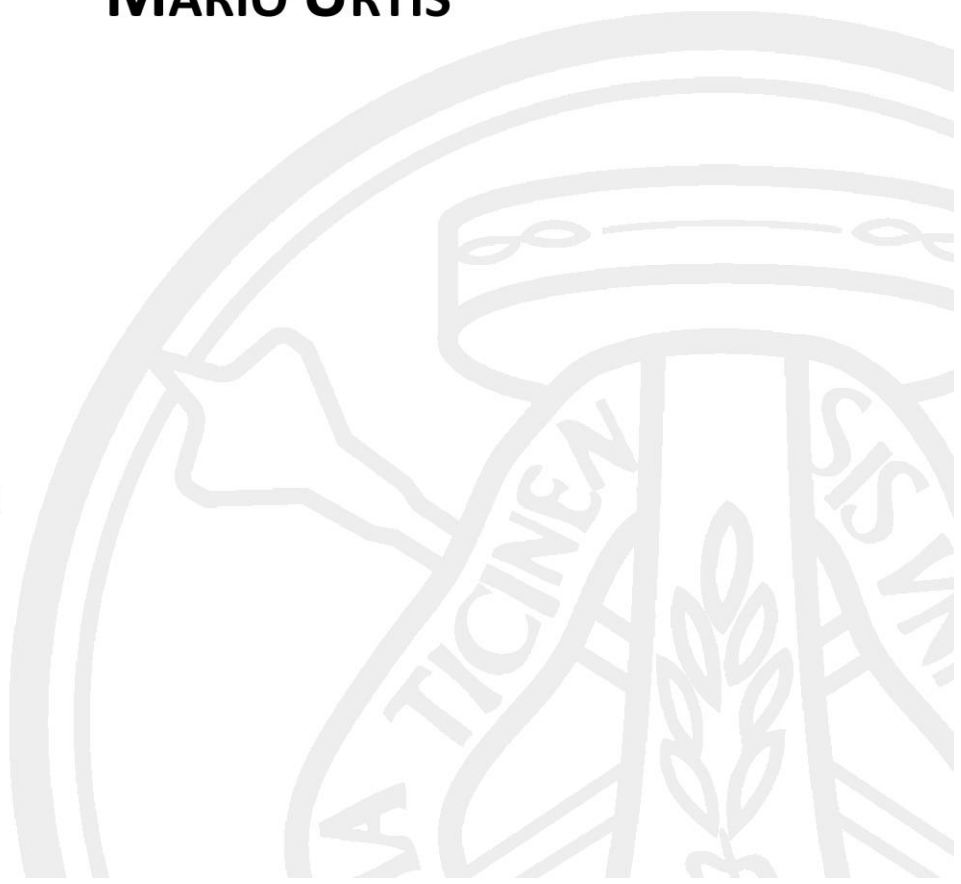
DOTTORATO DI RICERCA IN TECNOLOGIE PER LA SALUTE, BIOINGEGNERIA E BIOINFORMATICA  
XXXIV CICLO - 2021

## HELPER: A BIOINFORMATICS PLATFORM FOR CUSTOMIZATION OF NGS PIPELINES

PhD Thesis by  
**MARIO URTIS**

**Advisor:**  
Prof. Riccardo Bellazzi

**PhD Program Chair:**  
Prof. Silvana Quaglini





---

# Acknowledgments

---

This thesis is the result of the joint work of everyone who helped me with their guidance during my Ph.D., with their teachings, and with their support. For this reason, I would like to thank Professor Riccardo Bellazzi, for having confidence in my abilities since my master's degree, accepting to guide me as advisor throughout my scientific maturation process. I'd also like to thank Professor Silvana Quaglioni and Professor Paolo Magni for providing me with the opportunity to reach this important milestone.

Special thanks go to Dr. Monica Concardi, Dr. Alexandra Smirnova, Dr. Andrea Pilotto, Dr. Alessandro Di Toro, and the engineer Valentina Favalli for welcoming me among them and sharing with me all their experience and their knowledge allowing me to grow both as a man and as a scientist. My gratitude goes to every single colleague, doctor, technician, nurse, collaborator for enriching scientifically, and professionally all these years spent at the Center for inherited cardiovascular diseases.

My heartfelt thanks goes to Professor Eloisa Arbustini for teaching me what it means to be a researcher. Thanks for teaching me by example the importance of professional ethics and scientific truth, but more importantly, for being a guide in the difficult journey of becoming a scientist.

Finally, thank you very much to my family, especially to Viviana and Giovanni, who have provided me with the greatest motivation to achieve this goal, lightening my every effort and every difficulty, filling my life just with their presence, thank you.



---

## Abstract (Italiano)

---

Le tecnologie di next generation sequencing (NGS) hanno rivoluzionato il mondo della genetica e della medicina, influenzando fortemente la diagnosi delle malattie ereditarie. L'aumento della capacità di sequenziamento e l'abbattimento dei tempi d'analisi hanno permesso la diffusione delle tecnologie NGS in molti laboratori di genetica. Il grande numero di applicazioni, sia di diagnostica che di ricerca, ha inoltre generato la necessità di adattare l'analisi dei dati prodotti da queste tecnologie per ottimizzare la risposta ai problemi specifici. Il processo di analisi è implementato tramite trasformazioni consecutive dei dati genetici (pipeline) utilizzando un grande numero di tool e software bioinformatici. Spesso le performance dei diversi tool dipendono dal tipo dei dati in ingresso e l'integrazione dei software adatti ai diversi tipi di dati è diventato un passaggio critico per la qualità delle informazioni prodotte. Inoltre, l'utilizzo dei tool, la loro configurazione, la progettazione di pipeline robuste e lo sviluppo di nuove soluzioni di analisi, sono processi complessi che richiedono competenze di coding e la conoscenza dell'esteso panorama bioinformatico. In questo contesto, i bioinformatici hanno acquisito un ruolo fondamentale all'interno dei laboratori di genetica, grazie alle competenze di sviluppo di sistemi informatici unite alle capacità di comprensione dei problemi biologici e di adattamento delle analisi alle specifiche domande. I laboratori che non dispongono di queste professionalità specializzate possono incontrare difficoltà nell'ottimizzazione del workflow analitico, che spesso viene affidato a software commerciali che applicano uguali regole e sistemi a tutti i geni indistintamente. Da qui la necessità crescente di strumenti semplici e veloci che possano essere d'aiuto, anche per figure professionali con limitate competenze informatiche, alla progettazione di pipeline customizzate e al loro utilizzo nell'analisi dei dati NGS. Durante il percorso di dottorato di ricerca effettuato presso il Centro malattie genetiche cardiovascolari della Fondazione IRCCS Policlinico San Matteo di Pavia, è stata sviluppata la piattaforma Helper. Helper è nata per la progettazione e l'adattamento semplificato delle pipeline bioinformatiche dedicate all'analisi di dati NGS derivati da applicazioni di targeted sequencing. Helper è dotato di una semplice interfaccia grafica mirata a facilitare l'esperienza di sviluppo dei processi analitici bioinformatici anche per chi non possiede particolari conoscenze di sviluppo di codice. Tramite Helper è possibile scegliere quali step effettuare nel workflow di analisi e quali evitare, quali tools e software utilizzare in ogni step selezionato, e con quali argomenti settare i tool utilizzati. Helper permette inoltre di utilizzare le pipeline progettate ed effettuare l'analisi dei dati NGS, modificandole in

base all'esperimento di sequenziamento dal quale derivano i campioni e in base al tipo e all'organizzazione dei campioni. Helper può essere utilizzato sia su una workstation, sia su un comune PC, dimostrandosi compatibile con i tempi di analisi dei laboratori di genetica anche in presenza di soluzioni a bassa capacità computazionale. Nel workflow di analisi genetica, Helper è dedicato a quella che è definita come analisi secondaria, che trasforma i dati NGS grezzi in un set di varianti utili all'interpretazione del test genetico.

Il lavoro di tesi si è proposto inoltre di introdurre due ulteriori domande fondamentali per la diagnosi genetica. La prima è rappresentata dal problema della classificazione patogenica delle varianti identificate dall'analisi bioinformatica. La classificazione patogenica delle varianti è un processo delicato a causa della difficoltà esistenti nel trovare regole uniformi e robuste da applicare a tutti i difetti genici. In questa tesi viene proposto l'esempio di un sistema di classificazione per le varianti del gene DES, che prende in considerazione le caratteristiche specifiche del gene che codifica per la proteina di Desmina. Il secondo è l'identificazione dei geni responsabili di un determinato fenotipo, necessaria per l'ottimizzazione del test diagnostico e per la gestione dei pazienti. In questo contesto viene approfondito il problema dei tumori ereditari della mammella e dell'ovaio, tramite lo studio dei risultati di analisi del database genetico sviluppato presso il San Matteo per l'identificazione delle cause genetiche delle patologie oncologiche familiari, in particolare quelle clinicamente "actionable".

---

## Abstract (English)

---

Next generation sequencing (NGS) technologies have revolutionized the world of genetics and medicine, strongly influencing the diagnosis of hereditary diseases. The increase in sequencing capacity and the reduction of analysis time and costs allowed the spread of NGS technologies in many genetics laboratories. The large number of applications, both diagnosis and research, has also generated the need to adapt the analysis of the data produced by these technologies to optimize the clinical path of many human diseases. The analysis process is implemented through consecutive modifications of the genetic data (pipeline) using bioinformatics tools and software. Often, the performance of the different tools depends on the type of input data; the integration of software suitable for different types of data is a critical step for the quality of the information produced. Furthermore, the use of bioinformatics tools, their configuration, the design of robust pipelines, and the development of new analysis solutions is a complex process that requires coding skills and knowledge of the wide range of existing tools. In this context, bioinformaticians achieved a key role within genetics laboratories, thanks to the skills of developing computer systems combined with the integration of knowledge on target biology systems and related applications; these “in house” tailored activities favor the adaptation of the analyses to each specific question/objective.

Laboratories using outsourcing analysis tools or entrusting to commercial software that apply the same rules and systems to all genes, without distinction, often face difficult optimization of the analytical workflow. Hence, the growing need for simple and fast tools that can support professionals with limited computer skills in the design of customized pipelines and their use to analyze NGS data. During the PhD course carried out at the Center for Cardiovascular Genetic Diseases of the IRCCS San Matteo Hospital Foundation in Pavia, the Helper platform was developed. Helper was born for the design and simplified adaptation of bioinformatics pipelines for the analysis of NGS data derived from targeted sequencing applications. Helper is equipped with a simple graphic interface aimed at facilitating the development experience of bioinformatics analytical processes even for professionals who do not have coding knowledge.

Helper allows the selection of: the steps to carry out in the analysis workflow; the tools and software to use in each selected step; the arguments to set the tools employed in each application. Helper further allows the use of the pipelines, the design and carrying out of the analysis of NGS data; it can be modified based on the sequencing experiment from which the samples are derived, and on the basis of the organization of the samples. Helper can be used both on a workstation and on a common PC, proving to be compatible with the analysis times of the genetics laboratories even in the presence of solutions with low computational capacity. In the genetic

analysis workflow, Helper is part of the process of translating raw NGS data into a set of variants useful for the interpretation of the genetic test.

The thesis finally aimed at addressing two fundamental questions for genetic diagnosis. The first question addresses the complex issue of the variant classification as identified by bioinformatics analysis. The classification of genetic variants is a process that reflects difficulties in finding uniform and robust rules shared by all genes. In this thesis, a classification system is proposed for the variants of the DES gene, which takes into consideration the specific characteristics of the gene encoding the Desmin protein. The second question addressed the identification of the genes responsible for a specific phenotype, necessary for the optimization of the diagnostic test and for patient management. In this context, hereditary breast and ovarian tumors is investigated through the study of the results of the analysis of the genetic database developed at San Matteo for identifying the genetic basis of familial cancers, in particular clinically actionable genes and variants.



---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1.	<i>NGS-driven genetics</i>	1
1.2.	<i>NGS data analysis systems</i>	2
1.3.	<i>The aim and the structure of the thesis</i>	4
<b>2</b>	<b>Technological background</b>	<b>5</b>
2.1.	<i>NGS applications</i>	5
2.1.1.	<i>NGS re-sequencing approaches</i>	6
2.1.2.	<i>Germinal, Somatic, Cell-free DNA</i>	8
2.2.	<i>Illumina sequencing technology</i>	10
2.2.1.	<i>Library preparation</i>	11
2.2.2.	<i>Amplification and Sequencing</i>	12
2.2.3.	<i>Base-calling</i>	14
2.3.	<i>NGS data analysis</i>	15
2.3.1.	<i>Bioinformatics pipelines</i>	16
2.3.2.	<i>Alignment</i>	16
2.3.3.	<i>Pre variant calling process (pre-processing)</i>	17
2.3.4.	<i>Variant Calling</i>	19
2.3.5.	<i>Post-processing (Variant filtering)</i>	23
2.4.	<i>Variant annotation and interpretation</i>	24
2.4.1.	<i>Variant annotation</i>	24
2.4.2.	<i>Variant prioritization</i>	27
2.4.3.	<i>ACMG-AMP classification system</i>	29
<b>3</b>	<b>The Helper platform</b>	<b>33</b>
3.1.	<i>Needs and motivations</i>	33
3.2.	<i>Workflow management system</i>	34
3.2.1.	<i>Tools wrapping and parallelization</i>	35
3.2.2.	<i>The Helper structure</i>	36
3.3.	<i>Data processing workflow</i>	40
3.3.1.	<i>Pre-alignment process</i>	41
3.3.2.	<i>Alignment</i>	44
3.3.3.	<i>Pre-processing</i>	46
3.3.4.	<i>Short variant calling</i>	47
3.3.5.	<i>Post processing</i>	50
3.3.6.	<i>Variant annotation</i>	52
3.3.7.	<i>Post annotation</i>	53
3.3.8.	<i>CNV calling</i>	54
3.3.9.	<i>Sample organization and workflows</i>	56
3.4.	<i>The Helper GUI</i>	59
3.4.1.	<i>Tools setting</i>	60
3.4.2.	<i>Experiment Designer</i>	61
3.4.3.	<i>Samplesheet Designer</i>	62
3.4.4.	<i>Pipeline designer</i>	64
3.4.5.	<i>Analysis Settings</i>	65
3.5.	<i>Workflow performance study</i>	66
3.5.1.	<i>Trusight Cardio and Trusight Cancer Panels</i>	67

3.5.2. Computing performance study .....	67
3.5.3. CNV Analysis .....	71
<b>4 Clinical Applications .....</b>	<b>78</b>
4.1. <i>Variant interpretation - the case of Desmin</i> .....	78
4.1.1. Clinical and genetic background .....	79
4.1.2. The CMGCV-DES system .....	81
4.1.3. The DES-dataset .....	89
4.1.4. Benchmark study .....	89
4.1.5. The final classification .....	94
4.1.6. The importance of clinical and pathology studies.....	97
4.1.7. Clinical features of variant's groups .....	98
4.2. <i>Variants study in breast and ovarian cancer families</i> .....	100
4.2.1. Introduction to hereditary cancer.....	100
4.2.2. Genetic and clinical background .....	100
4.2.3. The reasons for genetic testing.....	101
4.2.4. The clinical and molecular genetic path at the OSM.....	101
4.2.5. NGS sequencing and analysis pipeline .....	102
4.2.6. Results of genetic testing.....	104
4.2.7. Family segregation and familiarity for BROVCA tumors.....	119
<b>5 Conclusions and future implementations.....</b>	<b>124</b>
<b>References.....</b>	<b>127</b>

---

# Chapter 1

---

## Introduction

### 1.1. NGS-driven genetics

Next-generation sequencing has definitively revolutionized genetic testing in human as well as non-human pathology. Earlier, in the 1970s, the Sanger methodology [1] had provided a new way of directly searching for the genetic basis of hereditary diseases. The process of genetic diagnosis before NGS technologies was a difficult path, due to the lack of information supporting the interpretation of findings and to the low throughput sequencing potential of the tools available. The search for gene defects was based on gene-by-gene sequencing, exon by exon, in a long and costly process if performed on a large scale. Sanger sequencing was used for more than 15 years as the unique sequencing tool at the Centro Malattie Genetiche CardioVascolari (CMGCV) of the IRCCS Policlinico San Matteo of Pavia (OSM), later integrated with Roche 454 sequencer (2011-2014), and finally by Illumina MiSeq from 2015 to date; it is still used as confirmatory second tool for the diagnostic test, as requested by Region Lombardia rules for genetic testing and guidelines for genetic testing by scientific societies.

The Sanger sequencing limitations were overcome with NGS tools that parallelize the sequencing of a large number of genes in a pool of samples at the same time, lowering the costs and times of genetic analysis and opening new horizons only glimpsed until then. The enormous amount of data that has reached the scientific community in just less than 20 years is one of the main effects of the expansion of NGS technologies. The evolution of population genetics gained by large-scale genetic studies led to the development of large population databases [2] as well as to the origin and implementation of clinical and genetic association databases (for example Decipher [3], HGMD [4] or ClinVar [5][6]). The broadcasting of shared genetic repositories around the world contributed to the exponential expansion of the methodology. An example of the impact of NGS technologies on the study of rare diseases is the growth of the Online Mendelian Inheritance in Man (OMIM) database [7] in which the number of inherited phenotypes with a known genetic basis has nearly doubled since

2007. In parallel, the number of genes associated with rare diseases has grown at an impressive proportion. The opportunities that NGS offers to the scientific community are not easily quantifiable: the possibility of discovering the causes of hereditary diseases that are still orphan, or of identifying more than one genetic disease in a single individual; the discovery of genetic markers of predisposition to quantify the risk of developing more common diseases [8][9]; the study the genetic makeup of tumors and the identification of targets to develop disease-specific medications and preventive surgery [10]. The cascade of the benefits of the new knowledge deepens molecular mechanisms of diseases thus finally translating into human care. These are some of the examples of the impact of NGS technology on patient management. NGS technology overturned the paradigm that guided genetic diagnosis. Moving from clinically-driven genetics to genetically-driven clinics, making the reverse phenotyping process possible [11]. NGS sequencing has in fact made it possible to identify the causes of genetic diseases that can be detected far before the phenotype develops, and therefore to optimize clinical management by anticipating the effects of the disease, significantly improving human lives. Now a new calling for a third step is needed, from genetically-driven clinics to molecular clinics, when new disease classifications incorporate their genetic basis (for example Desminopathy as the disease caused by DES defect shown in chapter 4).

However, easy access to these sequencing technologies introduces the risks of moving genetics away from the clinics. In recent years, huge amounts of data have been produced supporting associations between genes and diseases that have often proved inconclusive, complicating the genetic diagnosis process, and confusing the clinic. The large number of scientific papers that analyze the genetic bases of the diseases generate a “jungle” of contents that remain largely unconfirmed and non-validated. Hence, in recent years, the need to put an order within the genetic knowledge has arisen, which has favored the birth of projects such as ClinGen [12] aimed at providing precise rules for the interpretation of genetic data and standards for scientific communication. The method applies a robust process of curation of the literature, returning to paying particular attention to the clinic and refocusing on specific genes.

## 1.2. NGS data analysis systems

The exploitation in NGS sequencing led to the development of the systems necessary to analyze the large amount of data produced. The new sequencing platforms have the potential to produce terabytes of output files. Whole-genome sequencing projects can generate a huge amount of data that turns NGS analysis management into a big data problem. The challenges include the implementation of analysis processes suitable for the different

applications of NGS sequencing; the development of hardware systems, specific computing, and storage structures for the analysis of big data; the training of professional figures capable of responding effectively to the technological and biological problems typical of the NGS world. Over the years, a large number of tools and software dedicated to NGS data analysis have been developed, thanks also to special contests that engage the scientific community to collaboratively solve fundamental biomedical questions and focus the attention to bioinformatics problems [13]. The possibility of exploiting different tools made possible the adaptation of the analysis process (pipeline) to the different types of data produced by NGS sequencing. Although the application of the same pipeline is advisable to obtain repeatable and robust results, the adaptation of the workflow to each specific problem is often crucial for the final result. For this reason, in the last 5 years, various solutions have been developed aimed at customizing the NGS analysis process. Furthermore, the storage and processing of NGS data require computational structures that often are not readily available. Within labs with NGS machines, the available computing resources should match the computational needs of the instruments. In some cases, a workstation is the most cost-effective solution; in other cases, high-performance computing (HPC) resources are needed, such as cluster or server solutions. Cloud computing solutions may help to overcome the issues related to the purchase of expensive and difficult-to-manage solutions such as cluster servers. The option of paying based on the computing resources effectively utilized for the analysis helps to reduce costs for large sequencing projects and many companies, including Illumina, have adopted this solution to release accessible services to all customers. Finally, the adoption of NGS technology involves a series of difficulties that are not always within the reach of traditional figures in genetics laboratories, such as doctors, biologists, and laboratory technicians. The challenges include the development of analysis systems, the selection, and management of calculation tools, the design of new methods of extracting information from the raw data combined with the ability to fully understand the biological problem and to succeed to communicate effectively with the biomedical world. These complex challenges generate the need for new professionals whose contribution is now central to the management of NGS technologies. Their role is to effectively interface in all the steps of the genetic diagnosis process, from the evaluation of the clinical parameters of the patients to the extraction of information from the genetic data. These professionals may not be present in all laboratories, supporting the need for developing simplified systems that can help design NGS data analysis pipelines, and that can have an educational role in understanding the bioinformatics processes.

### **1.3. The aim and the structure of the thesis**

This thesis presents Helper, a platform for the simplified development of customized pipelines aimed at analyzing NGS data derived from DNA target-sequencing applications. The idea of a platform for the customization of analysis workflows was born within the highly multidisciplinary context of the CMGCV of the OSM Foundation. For more than 35 years, the center has been dealing with genetic diseases, including heritable cardiomyopathies, aneurysmal diseases, hereditary-familial tumors, and other rare and ultra-rare conditions. Within the genetics laboratory, the large number of experiments, pilot studies, and research projects that require NGS sequencing has generated the need for a fast and flexible system for adapting NGS analysis pipelines to different needs. The thesis reflects the experience of the bioinformatician within the CMGCV, invested with the technical role of developing tailored analysis solutions, incorporating all investigation tools that can contribute to the interpretation process of the genetic data. The next chapter (chapter 2) describes the technological background that describes the applications of NGS systems, the work-path of the Illumina sequencing technology, and the methods of NGS data analysis aimed at identifying the genetic basis of hereditary diseases. Chapter 3 describes the Helper platform and discusses the structure and the workflow management system, as well as the graphical interface for preparing the analyses. The chapter further discusses the results of the performance of the platform, considering times for analysis, and the accuracy of the results of the variant calling of the allelic copy number (CNV), a hot topic for the scientific community. Finally, chapter 4 describes two clinical-genetic applications: one exemplifies a rare monogenic disease with complex gene analysis and interpretation (Desmin), and one shows the germinal genetic basis of familial Breast and Ovarian Cancers.

# Chapter 2

---

## Technological background

This chapter describes the state of the art of NGS as applied to the analysis of the human genome. The aim is to show the technological path leading to the identification of disease-causing variants, for both research and diagnostic applications. The chapter briefly shows the scenarios of NGS applications for DNA sequencing, the Illumina sequencing technology, and the NGS data analysis workflow, from the structure of the bioinformatics pipelines to the interpretative path of the genetic data.

### 2.1. NGS applications

The potential of NGS technology is still evolving; despite being a relatively young technology, dozens of applications are described in the literature [14]. Many applications are now used on a large scale and have made the success of NGS. In short, NGS introduced a revolution in genome studies, greatly increased the potential for identifying gene variants, simplified the sequencing of new genomes, made it possible to carry out transcriptomics and gene expression studies, allowed the identification of the epigenetic changes of DNA and better understand DNA-protein interactions. For example, Bisulfite sequencing (methylation seq) is used to determine methylation patterns that regulate gene expression [15][16]; the ChIP (chromatin immunoprecipitation) seq is a sequencing technique used to study protein–DNA relationships. ChIP seq determines the sequence of the binding sites of DNA-associated proteins and maps these regions precisely in the genome [17][18]; the RNA sequencing is used to identify and quantify the transcripts that are expressed in tissues or single-cell sequences [19] as well as their changes over time. This technique allows studying alternative splicing effects in genes, gene fusion, transcriptional modifications, and the effect of genetic variants on the RNA product. RNA seq is used to identify medications in small-RNA, miRNA, tRNA, and rRNA or to find new RNA molecules [14][20].

Although epigenetics and transcriptomics are now commonplace in research laboratories, the two major applications of NGS sequencing remain the sequencing of new genomes (De-novo sequencing) and re-sequencing. De-novo sequencing has the primary objective of discovering the sequence of novel genomes -never been previously studied-without reference sequence, which must be generated. De-novo sequencing also contributed to improving and completing the genome sequencing of known organisms and to elucidating the structure of highly repetitive complex areas of DNA. It is usually applied to small bacterial and viral genomes and has fundamental importance in phylogenetic studies. Re-sequencing is vice versa defined as sequencing aimed at identifying variations of a genome when compared with a reference genome. The most common applications include the identification of the genetic causes of hereditary diseases, the discovery of new gene-phenotype associations, the calculation of the risk predisposition to different diseases, and pharmaco-genetics.

### **2.1.1. NGS re-sequencing approaches**

Re-sequencing applications are based on different genome interrogation strategies. The choice of the genome sequencing strategy is a fundamental step in the design of the study and takes into account factors such as the throughput capacity of the instruments, the number and type of samples to be sequenced, the costs, and the impact of the strategy on the goal of the project. The possible strategies include the sequencing of the whole genome (Whole Genome Sequencing - WGS) or the analysis of a pool of target genome regions of interest (targeted sequencing). The most extensive application of the targeted strategy is the sequencing of all coding regions of the genes (Whole Exome Sequencing - WES). However, often in research practice and diagnostics, sequencing of restricted genomic targets is used to specifically address the research or clinical aims (Gene Panels or Hot spot arrays).

The WGS represents the most comprehensive method for studying the genome: in human genetics, WGS may apply to both chromosomal and mitochondrial DNA. The WGS is the most effective application for characterizing the patient's genomic profile, due to its ability to identify defects in coding zones and in the intronic zones that contain regulatory transcription sequences. In recent years, the costs of high-throughput NGS technologies (e.g., Illumina NovaSeq and BGI platforms) have fallen below € 1,000, encouraging its use also in diagnostics [21]. The main obstacles related to WGS are the difficulties faced by many institutions in supporting the costs of consumables, either maintenance of the instruments or outsourced sequencing services, and finally managing a large amount of



data. The WGS generates a huge amount of data which, to be analyzed and stored, requires adequate infrastructure, which may not be easy to implement in all research labs. In clinical settings, a key point against the use of WGS in diagnostic contexts is the difficult interpretation of data, which currently prefers other more convenient and feasible sequencing strategies.

WES is a less thorough approach than WGS as it only provides sequences of the coding regions (exomes) of the genes. The WES, despite not including intronic regions, covers about 20,000 genes that code for proteins and whose defects cause a large number of known Mendelian hereditary diseases; in addition, WES may contribute to discovering new genes associated with the studied phenotype. Although limited by the absence of information on introns and some regulatory areas, the WGS guarantees an excellent cost-effectiveness compromise of the test. For this reason, the practice of exome sequencing is now entering genetic diagnostic paths.

The basis of Gene Panel Sequencing (GPS) is the selective study of genes or genomic regions known to be associated with diseases, or biological pathways pertinent with the given disorders, as suggested by previous studies of WGS, WES, or linkage analysis. Regions commonly studied include exons, introns, promoter sequences, or other highly conserved regions with biological significance and pertinence with the phenotypes. This method is the most widely used in the field of precision medicine for the detection of genetic variants associated with monogenic diseases or genetic risk factors, in which the variants are directly associated with specific genomic regions [22]. The advantage of GPS is the restriction of the analysis to target genes and to reduce the number of unneeded information that negatively affects the genetic diagnosis of specific diseases, syndromes, or phenotypes. In addition, an increasing number of guidelines/recommendations/position statements are generated to focus the clinical applications to those genes that are progressively proven and confirmed to play a deterministic role in the pathogenesis of the disease. This is because in the recent past, many “new disease genes” remained unconfirmed, not validated, and their defects were unsupported by functional studies. In addition, by reducing the target, sequencing costs are amortized, and the computational resources needed to manage, analyze and store genetic data are reduced. Compared to WGS and WES that usually require advanced analysis systems such as cluster servers or cloud systems and extensive bioinformatics work, the approach of sequencing gene panels reduces management costs [23] and analysis times. Despite the convenience, GPS is connected to some difficulties related to the composition of the gene panels and the limited detection capacity. In both cases, the efficacy of GPS is closely linked to the level of knowledge of the genetic basis of the diseases by the designer who must know how to identify the optimized target in order to meet the needs of the study. The use of scientific literature alone may not be sufficient for this purpose, which is

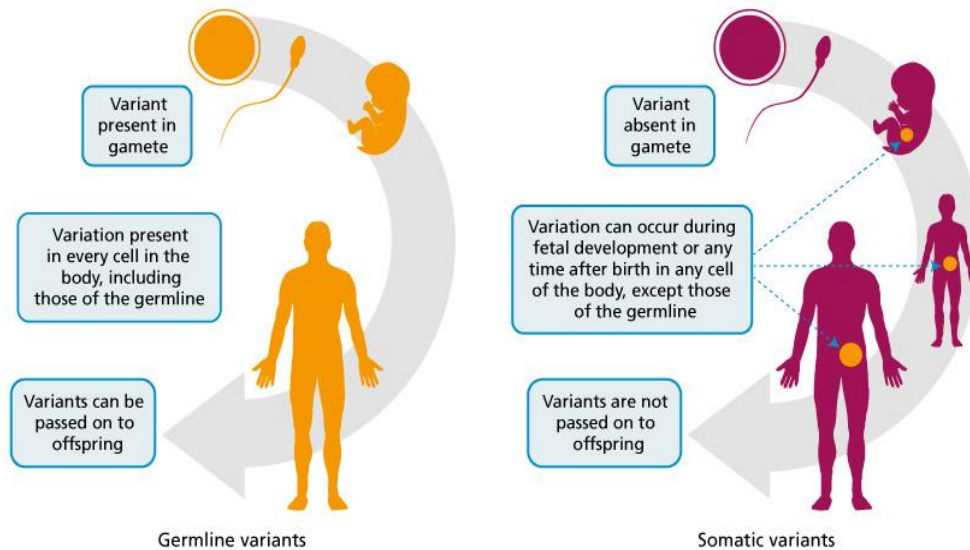
sometimes achieved only by integrating the use of multiple functional research methods and clinical studies capable of supporting and confirming the choice of clinically actionable genes (the thesis provides the example of Desmin gene). For other applications (e.g., malignancies), the choice of genes must be even more robust, because today the applications go beyond the diagnostic impact on patients and families but strictly concern preventive medical and surgical therapy (the thesis provides the example of Breast and Ovarian cancers). These issues have implications for liability, costs, and reimbursements, both diagnostic and therapeutic appropriateness, as well as impacting the health of patients and families.

### **2.1.2. Germinal, Somatic, Cell-free DNA**

The applications of NGS technologies also vary according to the type of sample to be analyzed. The germinal or constitutive DNA is inherited from the parents and represents the common genetic source for all the cells of the body. Germline DNA mutations are the cause of inherited genetic diseases and are the main target of the genetic diagnosis process for familial phenotypes. The identification of the causative variants of Mendelian diseases is the main goal of NGS sequencing in clinical practice. Defining the cause of a disease or the predisposition to develop a disease allows optimizing the clinical and therapeutic management of the patient and his family. For diploid organisms such as humans, the DNA defects can be inherited from one parent or both parents, and the allelic status of the variant can be heterozygous (one in two mutated alleles), or homozygous (inheritance of both parental mutated alleles). Inherited variants are found at the same allelic frequency in all cells of the body.

During fetal development and throughout the lifespan, genomic sequence variations occur in an individual's DNA due to random errors in the DNA replication process or damage caused by exposure to environmental factors such as harmful radiation, chemical or physical injuries/exposures, incorrect lifestyles, etc. Variants acquired post-zygotically are referred to as somatic variants. The characteristic of somatic variants is that they cannot be passed on to subsequent generations if they are absent in the progenitor cells of the gametes. The median somatic mutation rate for variants affecting a single nucleotide span in the order of  $3 \times 10^{-7}$  [24], therefore an accumulation of variants in the DNA of the cells is expected to occur during life, generating genetic heterogeneity within the same tissue or between different tissues of an individual. This process of genetic differentiation due to somatic variants is called somatic mosaicism [25]. The more a variant occurs early in the cell differentiation process, the more it is represented in the cellular populations

of the organism while the variants located in specific districts have a more recent temporal origin [26].



**Figure 2.1:** Germline vs Somatic variants

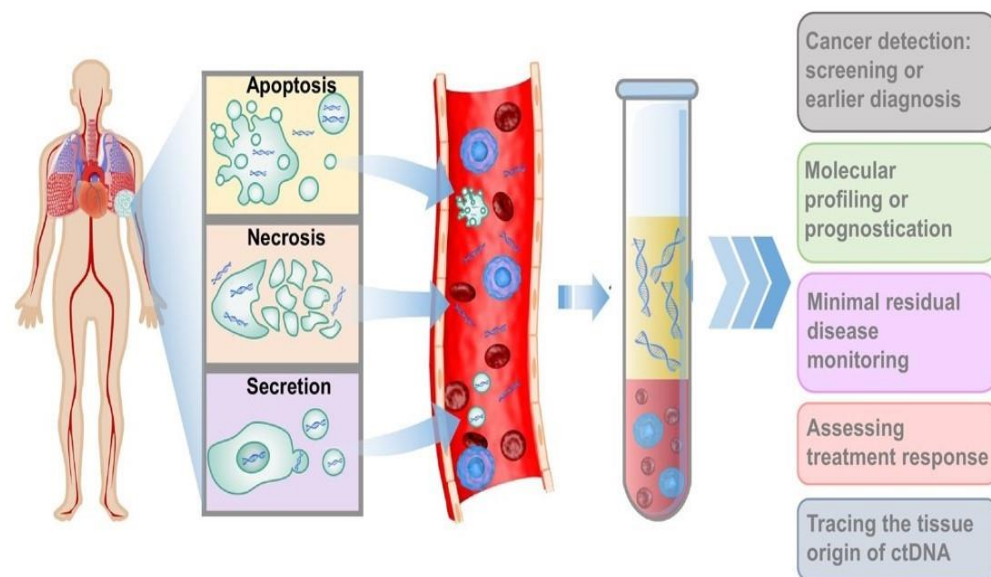
(Image from <https://www.genomicseducation.hee.nhs.uk/cancer-genomics/>)

The most common field of investigation of somatic DNA is cancer. Cancer is caused by a progressive accumulation of somatic mutations (sometimes favored/promoted by a germ gene defect) that generates high genetic heterogeneity and causes continuous clonal differentiation. Clones in which mutations that provide individual cell evolutionary advantages, reproduce, and become increasingly aggressive [27]. The somatic variants can be exploited as pharmacological targets in the treatment of some oncological diseases [28][29] and can be used as a marker of tumor evolution or adaptation from a prognostic perspective [30].

The search for somatic variants, their burden, and distribution in a tissue sample or single cells is far more complex than the search for germline variants. Somatic variants are involved in many diseases and cell aging processes, but their impact must also be assessed based on the fraction of mutated cells. Unlike the germline variants that can be identified by DNA from any nucleated cell, the somatic variants must be searched within the affected tissue and their measured allelic frequency depends on the number of mutated cells that are sampled for sequencing.

A further method for studying somatic DNA is the sequencing of cell-free DNA (cfDNA). CfDNA is composed of somatic DNA fragments released into the bloodstream following cell damage caused by trauma, sepsis, aseptic inflammation, myocardial infarction, stroke, transplantation, diabetes, sickle cell disease, and cancer [31].

Major sources of cfDNA are massive cellular apoptosis and necrosis that occur in the exponential growth of tumors which releases a very high amount of fragmented DNA into the plasma compared to the physiological baseline levels [32]. CfDNA reflects the genetic makeup of the cells that release it and can be used as a marker for the early diagnosis of cancer and relapse [33][34][35], for the identification of pharmacological markers for target treatments and for monitoring the evolution of the disease as well as the minimal residual disease [36] (Figure 2.2).



**Figure 2.2:** Cell-free DNA sources and analysis applications (Figure modified from [34])

## 2.2. Illumina sequencing technology

Illumina (San Diego, CA) is an American company that develops systems for the analysis of genetic variation and biological function since 1998. Since then and very quickly, Illumina gained the market leadership in NGS machines, and its platforms are still a technological reference despite the continuous evolution of the NGS and market competition. Illumina boasts a range of tools that cover all the possible needs of a laboratory, thus managing to achieve the largest proportion of the worldwide market [37]. Illumina sequencing technology is defined as a second-generation technology and is based on the clonal amplification of DNA fragments on a solid support and the generation of read sequences of 100-300 bp (short reads). The NGS Illumina sequencing workflow can be divided into three steps common to all the platforms produced by the company: 1. Preparation of the libraries; 2. Sequencing; 3. Base-calling.

### 2.2.1. Library preparation

Library preparation starts from the DNA molecule and transforms it into a pool of fragments (genomic library) ready to be uploaded on the instrument and then, sequenced. The entire DNA molecule is too large to be sequenced using Illumina instruments. For this reason, once the DNA has been isolated, a fragmentation step is performed that generates millions of small fragments. There are several methods of DNA fragmentation: sonication and enzymatic methods are those commonly used in most labs. The fragments need the addition of adapters that bind fragments to the flowcell (see Chapter 2.2.2) and indexes useful to identify the sample of origin of the fragment in case of multiplexing sequencing. For some applications, additional indices called Unique Molecular Identifiers (UMI) are added which represent a unique code for each fragment and are useful for increasing error correction and accuracy. They can reduce false-positive variant calls and increase variant detection sensitivity.

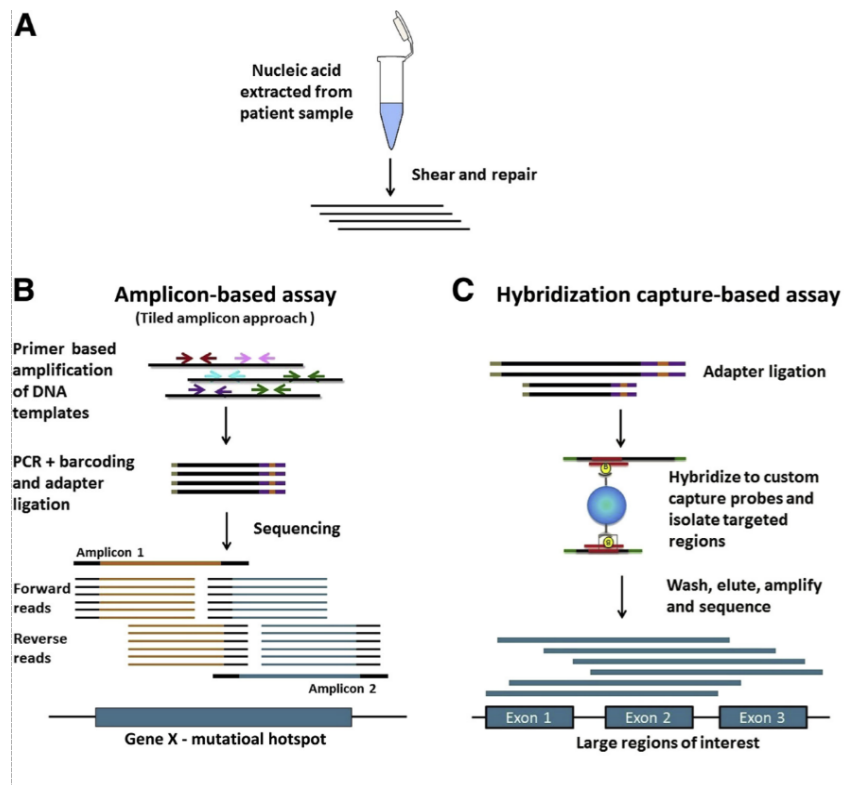
In the case of WGS sequencing, the library is amplified through PCR cycles to increase the signal readable by the instrument. Then, the length distribution of the fragments that characterize the library is analyzed, and finally, the optimal sample quantitation is defined, which is loaded onto the sequencer. In the case of a targeted approach, a selection step is performed for the fragments that cover the target of interest before the amplification and analysis of the distribution of lengths. The two most common target selection methods are that based on fragment capture (Hybridization capture) and on copying fragments (Amplicon) (Figure 2.3) [38].

The hybridization capture-based method uses long oligonucleotide probes to hybridize and capture fragments. Because the DNA is randomly sheared during library preparation, captured fragments are partially overlapping and unique. Overlapping allows coverage of the target, even in the event of problems with some nearby probes, and the uniqueness helps to identify possible sequencing errors. With the capture enrichment method, coverage of certain particular regions, such as genes with pseudogenes, highly repeated regions, and regions with high GC content - can be difficult. Furthermore, it could be affected by differences in the affinity of the different probes, thus impairing the coverage of the target. These problems can be solved during the design of the target. Areas known to be associated with these problems can be covered using second methodologies such as Sanger sequencing.

Methods based on target enrichment amplification methods (Amplicon) employ a set of PCR primers to generate PCR products -size 150–400 base pair- starting from the ends of the fragments. Each fragment of interest is cloned (a high number of times) to be read by the sequencer. Amplicon sequencing is usually a faster process than hybridization capture with the

same samples and guarantees a higher fraction of sequences within the target (in-target reads) than the capture method thanks to the specificity of the primers used. Amplicon methods, however, like all PCR-based methods, are sensitive to allele dropout which can be caused by variants present at the primer hybridization site. The dropout allele can generate the loss of coverage of entire fragments and the loss of the ability to identify variants. Another problem of Amplicon-based methods is the amplification of the error due to the PCR reaction which increases the probability of false-positive findings [39].

Amplicon sequencing is optimal for efficiently sequencing small targets such as small gene panels (1-25 genes) and mutational hotspot panels and is preferable for deep sequencing applications. For larger targets, the method based on hybridization capture enrichment is usually preferred, capable of providing a more uniform coverage distribution and mitigating problems due to the quality of the probes [38].

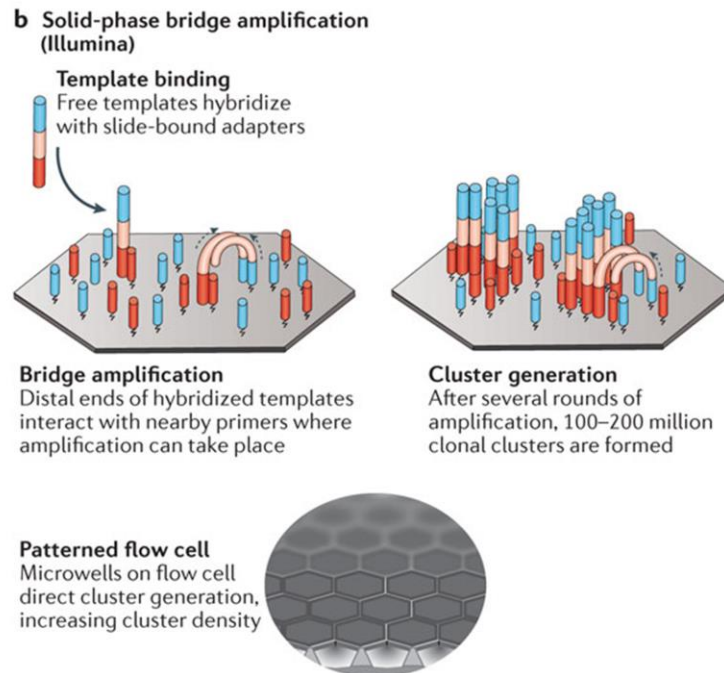


**Figure 2.3:** Target selection methods in NGS library preparation (Figure modified from [40])

### 2.2.2. Amplification and Sequencing

Once the NGS library is loaded onto the sequencer, the amplification needed to distinguish the sequencing signal from the background noise

occurs. In the Illumina solid-phase bridge amplification, the fragmented library is linked to primer immobilized on a solid support, such as a patterned flowcell. The free end of the fragment interacts with other nearby primers, forming a bridge structure. Using PCR, a second strand from the immobilized primers is created, and unbound DNA is removed (Figure 2.4) [41]. The process is repeated to generate a cluster of clones for each fragment that is bound to the flowcell.

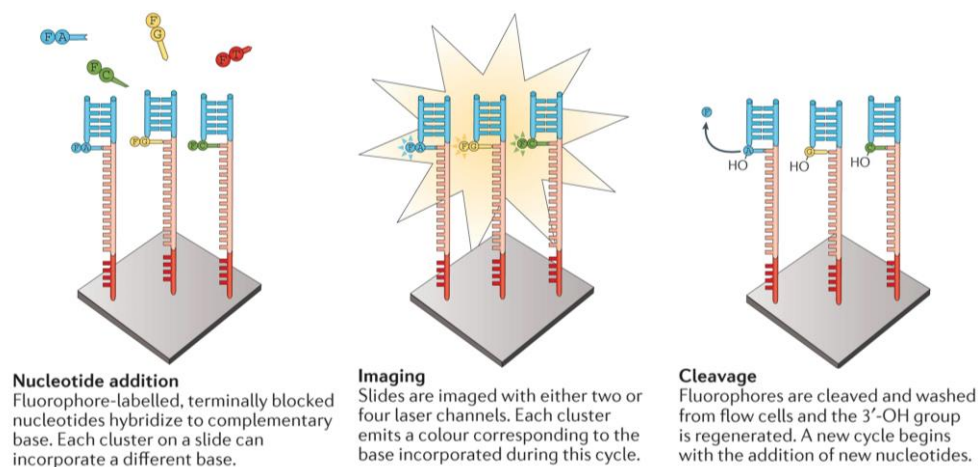


**Figure 2.4:** Illumina amplification system (Figure modified from [41])

Illumina's NGS technology is a Sequencing By Synthesis (SBS) method [41] with a fluorescent-labeled reversible terminator technology (RT). In brief, each fragment becomes a template and is copied by an enzyme, the DNA polymerases, capable to incorporate a complementary nucleotide to that of the template. The procedure is composed of cycles of three phases: 1. Addition of a single nucleotide; 2. Creation of the image of the binding signal; 3. Cleavage of the terminator and washing of the flowcell.

When a single dNTP linked to the reversible terminator is incorporated into the sequence of a cluster, a fluorescent light signal is emitted at a wavelength that differs for each nucleotide. The signals emitted on the flowcell are recorded by an optical system (Charge-Coupled Device - CCD camera) which captures an image for each emission wavelength of the nucleotides. The terminator does not allow the polymerase to incorporate other nucleotides and further elongate the sequence so that only one

nucleotide per cycle can be incorporated. During the last phase, the terminator and the fluorescent dye are split from the incorporated dNTP to allow the addition of the next labeled dNTP. The unused nucleotides are removed by washing the flowcell; then a new cycle restarts (figure 2.5). The procedure takes place in parallel on all the clusters present on the flowcell. Each cycle, therefore, corresponds to four images that appear dotted, one for each nucleotide, where the dots represent the clusters that have incorporated the specific dNTP. The result of the sequencing step is a number N of flow-cell image quadruplets, where N corresponds to the number of cycles performed and therefore to the length of the read. Illumina NGS platforms are capable of sequencing both ends of each DNA fragment (paired-end sequencing) increasing sequencing quality and target read depth.

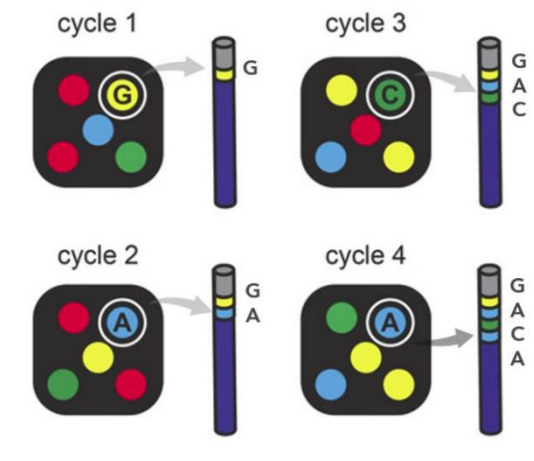


**Figure 2.5:** The Illumina sequencing cycle (Figure modified from [41])

### 2.2.3. Base-calling

Images acquired during sequencing cycles are then analyzed by the Illumina proprietary software installed on the sequencing platform. The images are first filtered to eliminate background noise, the light signals are identified and improved, the positions of the clusters in the flowcell are identified. For each image generated by a single machine cycle, the base that is most likely to be identified is assigned to each cluster (figure 2.6). The result is that each cluster is represented by a sequence of ACGTs corresponding to the nucleotides incorporated during sequencing. Each of these sequences is defined as a read. A Phred quality score is assigned to each base of the read, which represents in logarithmic scale the probability that the base has been erroneously assigned. All the reads generated by the sequencing are stored in a text file according to the Fastq format.





**Figure 2.6:** Illumina base-calling process (Figure modified from [42])

## 2.3. NGS data analysis

NGS generates massive amounts of data that require multiple computationally intensive steps for appropriate analysis to be performed. The analysis of NGS data is considered an integral part of the NGS sequencing process. The analysis workflow is specific for the type of DNA sequenced, for the library preparation method, for the sequencing technology of the instrument, and the amount of data produced. At the current state of the art, there are many tools and software for NGS data analysis, and the choice of the best solution is essential to perform a robust and cost-effective workflow. An example is that of the choice of computing resources to be used to analyze the data. High-performance computing (HPC) systems such as cloud services or server clusters, allow us to face the big problems of computing and storage resources typical of big data derived from large sequencing experiments (e.g., WGS) and to break down the costs on large numbers. The management of these systems is complex and requires dedicated expertise. Less complex solutions such as workstations, on the other hand, are better suited to the analysis of small experiments (e.g., small targeted-seq) and can be managed more easily even by less specialized figures.

The NGS data analysis process starts from the sequencing tool. The primary analysis of the data is represented by the transformation of the fluorescence signal acquired during sequencing through the base calling process, the calculation of the Base quality, and the production of Fastq files. The consecutive manipulation of the different types of files until reaching a useful result for the experiment is instead defined as secondary analysis. For genotyping applications, the secondary analysis starts from the Fastq files up to a set of variants contained in the sample under analysis. The transformation of data into knowledge useful for the interpretation of the

results is instead called tertiary analysis. The primary analysis, being implemented within the sequencer, is now a robust and reproducible process, while secondary and tertiary analyses are highly customizable.

### **2.3.1. Bioinformatics pipelines**

The secondary analysis is typically performed through progressive steps that process the sequencing data and transform them using multiple tools and software components. This process, in which the genetic data output of one tool becomes the input of another to be manipulated by sequential modules, is called "bioinformatics pipeline". The goal of bioinformatics pipelines is to perform the analysis process in an automatic and reproducible manner and ensure the greatest achievable robustness and accuracy. NGS bioinformatics pipelines are frequently platform-specific and may be customizable based on the experiment design and laboratory needs. A typical clinical implementation of a bioinformatics pipeline to search for variants in DNA samples consists of five major steps: 1. Alignment; 2. Pre-processing; 3. Variant calling; 4. Variant post-processing; 5. Variant annotation.

The internal workflow of each of the major steps is complex and may vary according to the application. To make the result of the different pipelines reproducible, best practices have been developed over the years for the different types of NGS data [43]. The workflow for genotyping applications of NGS data produced by re-sequencing experiments to identify the genetic causes of a particular phenotype is introduced below.

### **2.3.2. Alignment**

The first fundamental step for the study of NGS data is the Alignment which consists in recomposing the sequenced genome starting from the reads present in the Fastq files. For re-sequencing applications, the alignment of reads is a process facilitated by the presence of a standard genome to which it is possible to refer to find the right position of every single read. There are reference genomes for many organisms: they are updated cyclically to improve their accuracy. The latest reference genome for humans, GRCh38, was released in 2013 but many laboratories still use the previous GRCh37 or hg19.

There are several alignment algorithms, but most use the Burrows-Wheeler transform (BWT), or techniques based on hash tables [44]. BWT-based aligners are memory-efficient and work faster than hash-based aligners, but they are less accurate. In contrast, hash table algorithms tend to be slower, but more sensitive. The choice of the aligner is a key step that can influence the result of the analysis. The algorithms are evaluated based on the accuracy in finding the right position of the reads, but also in terms of efficiency (speed of execution) and scalability (storage capacity). Some benchmarking studies have compared the alignment tools found in the

literature [45][46] demonstrating that algorithm performance depends on input data and that there is no better one for all scenarios. Many tools are highly configurable to increase their adaptability to a particular application and it is up to the researcher to find the right set-up to optimize the analysis considering the possible obstacles.

One of the main challenges for alignment is the presence within the genome of repetitive or low-complexity regions. This often leads the reads to be mapped in different areas of the genome with the same reliability. The result is an ambiguous alignment that potentially leads to errors in the variant detection process. Longer reads and paired-end sequencing can help in improving alignment in these particular regions. The length of the reads and their complexity in terms of the sequence are directly proportional to the quality of the alignment. The presence of paired reads instead increases the available information (e.g., orientation and distance between read pairs) to improve mapping performance.

At the end of the alignment process, the Mapping quality score is calculated for each read, indicating the accuracy of the chromosomal position assigned by the algorithm. Reads enriched with further information on the mapping are stored in a Sequence Alignment Map (SAM) (specifications in [47] for SAM format description) format.

### **2.3.2.1. Post Alignment process**

After the alignment, it is possible to make changes to the mapped sequences which commonly include the conversion of the SAM files in the compressed version in BAM (Binary Alignment Map) form, in the sorting of the reads inside the BAM files to optimize the analysis and in the general assessment of the alignment by issuing a report. It is possible to evaluate the alignment by viewing some software that shows the reads mapped on the reference genome such as the Broad Institute's Integrative Genomics Viewer (IGV) [48][49] but it is a difficult process for quality control of large targets.

### **2.3.3. Pre variant calling process (pre-processing)**

In order to improve the variant identification process, some data optimization steps are recommended. The most important ones involve identifying duplicates from PCR, alignment artifact correction, and sequence quality score recalibration.

#### **2.3.3.1. The marking of duplicates reads**

The amplification step usually concludes the preparation of libraries (see Chapter 2.2.1) and is useful to get a greater sequencing yield. The amplification generates clones of the fragments contained in the library which are randomly immobilized on the flowcell and sequenced. When

multiple copies of the same original fragment bind at different points in the flowcell, they give rise to separate clusters and are sequenced independently. This process generates duplicated reads that can introduce a bias in the analysis that causes false high coverage of some areas and rises false-positive variant calls due to errors that occurred during library preparation, and that have been propagated to PCR duplicates. The percentage of duplicates depends on the characteristics of the NGS library and the loading phase of the instrument. If the amount of starting sample is small, the amplification step of the library must be greater thus increasing the duplication rate, furthermore, the smaller fragments are amplified more and can be over-represented. Finally, if the amount of library loaded on the instrument is lower than expected, a higher percentage of clones bind to the flowcell and are sequenced as duplicates.

Given that, PCR duplicates originate from the same DNA fragment, their mapping positions can be used to identify and either mark or entirely remove these duplicates, retaining only the highest quality read. The duplicate reads removal step is strongly recommended in workflow analysis of NGS data generated from the Hybridization capture-based method but not for Amplicon sequencing. In the Amplification based method, reads start and end at the same positions by design and duplicates removal should be disabled because otherwise, it will remove most aligned reads.

### 2.3.3.2. Indel Realignment

Because alignment algorithms map each read individually to the reference genome, reads spanning insertions or deletions (Indels) are often misaligned and it commonly results in mismatches. The tools that call variants could be fooled by mismatches and could call an insertion or deletion (Indels) in the sequence as a set of SNVs, increasing the error rate. To recognize and eliminate these artifacts, the local realignment process around the indels is performed, which is divided into two steps. In the first phase, suspicious intervals are defined in three ways: sites where there are frequent Indels in the population databases such as dbSNP [50] and 1000G [51], Indels seen in original alignments, and sites where some evidence suggests a hidden Indel. In the second step, the optimal consensus sequence is determined, and the local realignment of reads around the site is performed.

The entire process of Indel Realignment is computationally intense and for high coverage, sequencing is very time-consuming. The latest software for calling variants have implemented a local realignment step to improve the accuracy and quality of the variants identified. If these tools are used, realignment is no longer an essential step and can be avoided by saving time and resources. However, it remains recommended because it improves the Base Quality Score Recalibration process.

### 2.3.3.3. Base quality score recalibration

Base quality scores are per-base estimates of error emitted by the sequencing machines and express how confidently the called base is deemed correct. The base quality score is a fundamental factor that is used by variant callers to decide whether a variant really exists or is an error and is a main feature for filtering false positives. The BQ score emitted by the sequencing machine is often inaccurate and is subject to various systematic errors due to the sequencing reaction (e.g., machine cycle and sequence context) and to small defects in the instrumentation that cause it to be incorrectly estimated. For this reason, a score recalibration step is essential that re-evaluates the error probability of the called base using several features including starting quality score, the machine cycle, and the dinucleotide sequence context (the current and the previous bases).

### 2.3.4. Variant Calling

The key step in the analysis of NGS data is the identification of the variants present in the sample. The variants can be of three types:

- Point Variants or Single Nucleotide Variants (SNV): these are substitutions of single nucleotides in the DNA sequence.
- Short Insertions or Deletions (InDels): they are caused by an insertion or loss of some nucleotides respectively.
- Structural Variants (SV): are large genomic rearrangements affecting extended areas of the genome from hundreds of nucleotides to entire chromosomal segments. SVs include Translocations, Inversions, or variations of the copy number of a DNA stretch (CNV).

#### 2.3.4.1. SNV / InDel calling

The tools calling short variants compare the aligned sequences contained in the BAM file against the reference genome and identify the variants using different approaches. Numerous tools have been developed to identify single nucleotide variants (SNVs) and short insertions/deletions (indels) from aligned NGS data [52][53]. The tools use different methods to perform variant calling, some are based on heuristic methods, some use probabilistic models, other machine learning algorithms.

Heuristic methods call variants based on multiple information sources associated with the structure and quality of mismatches. For variant detection, a heuristic algorithm determines the genotype based on thresholds for coverage, base quality, and variant allele frequency. These tools usually use statistical tests (i.e., Fisher's exact test on the reads covering the variant) to assess the call quality.

Probabilistic methods instead provide measures of statistical uncertainty for called genotypes. Probabilistic tools use Bayes' theorem to calculate the

genotype likelihood for each possible genotype at each base (a homozygote for the reference allele, a homozygote for the alternative allele, or a heterozygote). The algorithm calculates the *a priori* probabilities of the genotypes and infers the posterior probabilities using the information from the quality scores and allele counts. The genotype with the highest posterior probability is chosen and the ratio between the highest and the second-highest probabilities may be used as a measure of confidence. Some Bayesian tools also implement a local realignment or assembly of suspicious reads to increase the accuracy of the variant call. Variants identified during variant calling are reported in the variant calling format (VCF) [54].

### 2.3.4.2. Individual versus joint variant calling

Many tools provide the ability to analyze both a single sample at a time and a cohort of samples simultaneously. Single sample analysis produces reproducible and repeatable results because it does not depend on other samples. This approach is the simplest and the least computationally expensive, but it may cause some information to be lost. In fact, in the VCF file all the sites in which a variant has been identified in the sample are reported, but in all the other sites not reported it is not clear whether the sample is homozygous and not mutated or if the coverage is not sufficient to perform a call. Conversely, multi-sample calling involves a simultaneous identification of variants in several individuals, and it is much more CPU time- and resource-consuming than individual variant calling. It produces genotypes for every sample at all variant positions by differentiating, for samples that do not carry the variant, between not mutated homozygotes and those with insufficient coverage. Furthermore, the joint analysis allows a variant caller to minimize the issue of variant representation differences that affects particularly complex variants and to use multi-sample information to improve the genotype likelihood calculation. Finally, multi-sample analysis can help in trio sequencing, enabling direct inference of the *cis* or *trans* status of two heterozygous variants.

### 2.3.4.3. Germline versus Somatic variant calling

A particular case of variant calling is the search for variants in somatic DNA samples. The call of the germline variants is relatively simple, identifying the mismatches between the sample sequence and the reference sequence that exceed a certain probability of not being sequencing errors or alignment bias. As far as somatic variant calling is concerned, the matter is more complex. Usually, the search for somatic variants is performed through a case-control analysis in which the case is represented by the somatic sample and the control is a sample taken as Germline or as Normal (in the case of tumors). Therefore, the somatic tools search for the mismatches with respect to the reference genome that are present in the somatic sample and

identify which of these variants are of somatic or germinal nature. Sometimes, it is possible to analyze only the somatic sample against the reference, but it is not recommended due to the high number of false positives that are produced. The greatest difficulty is given by the nature of the samples; using the example of tumors, it is common for the somatic tissue sample taken with a biopsy to be contaminated with normal cells and vice versa, causing changes in the allelic fraction in both tumor and healthy samples. Furthermore, the tumor could be subject to clonal heterogeneity or be affected by structural events and by changes in the number of copies of a given region. The consequences of these factors are an allelic fraction of the somatic variants that can reach very low values (even below 1%) and the presence of somatic variants of the germ samples due to tissue contamination problems. The biggest challenge for somatic variant callers is to recognize variants with lower allelic fractions and rule out sequencing, alignment, and cross-contamination artifacts.

#### 2.3.4.4. CNV calling

Multiple tools have been generated to detect CNVs in NGS data. Their approaches can be categorized into five different strategies (figure) that have advantages and limitations:

- Paired-End Mapping (PEM);
- Split Read (SR);
- Read Depth (RD);
- Assembly-based (AS).

The PEM uses distances between paired-end reads and is not applicable with single-end reads. In paired-end sequencing, the libraries prepared using the same protocol have similar fragments length, and consequently similar distances between paired reads, distribution. PEM identifies the distances of the mapped paired reads that are significantly different from the expected insert size and infers the presence of a CNV event. The main limitation is that it cannot detect CNVs in regions with segmental duplication.

The SR identifies possible CNV events using read pairs. The SR method identifies paired read in which one read is uniquely aligned to the reference genome and the paired one is unmapped or only partially maps to the genome. The assumption is that the read fails in perfect mapping because of the presence of a breaking point. SR strategy split the mis-mapped reads into multiple fragments and re-aligns the first and the last parts providing the precise start and end positions of the CNV events. The SR method is also affected by the nucleotide composition of the interrogated area as well as by the length of the reads.

The RD method is based on the assumption that the presence of a CNV is related to the variation of the read depth in the region of the event. In case of an allele deletion, a significant decrease in coverage should be observed

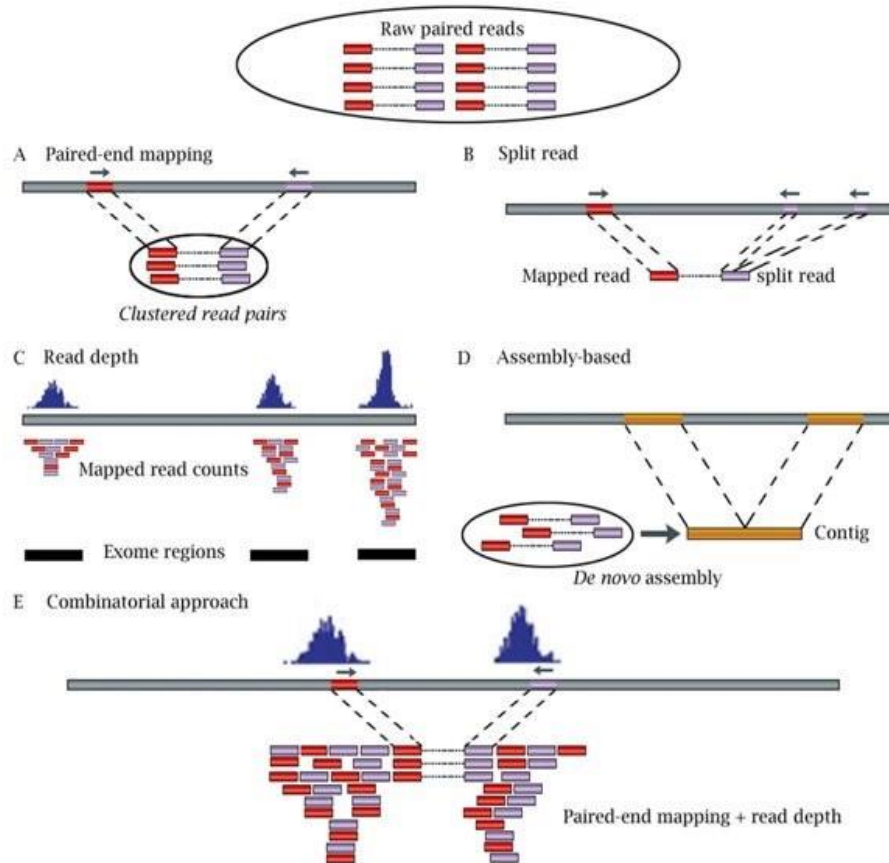
and vice versa a duplication should increase the coverage of the duplicated zone. The read depth procedure for CNV detection consists of four steps. The CNV caller tool calculates the read depth in the predefined window, normalizes and estimates the number of copies in the window, and finally merges all regions with a similar copy number detecting CNV events [55]. Generally, RD-based tools use a cohort of samples to improve the quality of the normalization step and the calling of the CNVs. The coverage normalization is performed as a function of the distribution of coverage in the single sample by correcting it according to the characteristics of the analyzed genomic region that may have introduced a coverage bias (e.g., GC content, repeated regions), and using the overall depth distribution from more samples to improve the result. The RD approach works well with high coverage samples and allows to mitigate the problems related to difficult areas of the genome and it better quantifies the extent of the structural event than the other methods.

The assembly methods first reconstruct DNA contigs performing the assembly of overlapping reads, then identify CNV events comparing the assembled regions with the reference genome. The assembly can also use the reference genome as a template to improve the quality of the contigs and the computational efficiency.

As for SNV and short Indels callers, even for CNV callers, despite the large number of tools developed, there is no gold standard. PEM-based methods can detect all types of SVs (even single exons) but not insertions that exceed the average insert size of the library. In addition, the estimate of the number of copies of the event cannot reach the quality of RD-based methods. Read depth-based methods can be useful in target sequencing applications especially for identifying larger CNVs. However, with RD methods the identification of small CNV (<1 kb), inversion or translocations, and the precise breakpoint sites. On the other hand, SR-based methods provide a very high resolution in finding breakpoints but not repeated or low-complexity genomic regions.

New tools adopt combined strategies to exploit the advantages of multiple methods while reducing their weaknesses. In any case, the choice of the CNV caller and the detection method must adapt to the needs of the specific application.





**Figure 2.7:** CNV calling strategies (Figure from [56])

### 2.3.5. Post-processing (Variant filtering)

This phase aims to increase the calling accuracy and eliminate “residual” artifacts. The filtering of VCF files is an important step in the bioinformatics pipelines because it guarantees the most accurate set of variants in the output, minimizing the number of false variants without excluding any of those actually present in the sample. Short variant filtering strategies can be classified into two groups; 1. filtering based on the quality threshold criteria (hard filtering) and 2. machine learning systems for automatic filtering of variants (soft filtering). Hard filtering (HF) is a system of rules that discriminates artifacts and variants by evaluating a set of quality indicators. Characteristics commonly assessed with HF include variant coverage and allelic frequency in the sample, variant base quality, and mapping quality scores as well as related differences with reference allele, and genotype quality score. An additional index is the Imbalance strand specificity because a true variant is expected to be equally represented on both forward and reverse strands. The thresholds for discriminating false positives should be modified according to the type of application desired. For example, for the search for somatic vs. germline variants, coverage and frequencies filters are different, or the criterion that evaluates the imbalance strand in the case of

amplicon or capture enrichment sequencing differ. Furthermore, it is necessary to carefully evaluate the result of filtering in complex areas of the genome where true variants could appear as artifacts using the same thresholds on the whole target. The major flaw of Hard filtering is that evaluates each threshold independently, and the discrimination rules, although considering one or more criteria at the same time, fail to grasp the interdependencies between indicators, producing effective but coarse discrimination. Soft filtering is a more sophisticated approach that leverages the capabilities of machine learning to identify patterns within data, combining different indicators, and performing a finer classification than hard filtering. Soft filtering methods build a supervised classification system by training the model on a set of known variants and artifacts. This model estimates the probability that a variant is really present and allows filtering at various confidence levels. Soft filtering is especially useful for low coverage samples [57] but its performance is influenced by the need for a large training dataset which is often not available, especially for targeted sequencing applications.

## 2.4. Variant annotation and interpretation

The format used by variant callers to report the variants describes the internal characteristics of the sample and are useful for discriminating artifacts and true variants but does not allow us to understand their role in the carrier phenotype. The annotation of the variants together with the process of interpreting the genetic data constitute the tertiary analysis of the NGS data and are essential for identifying the causes of hereditary diseases in the genetic diagnosis process.

### 2.4.1. Variant annotation

Once the calling and variant filtering process has been completed, the last step of an analysis pipeline is functional annotation. The annotation aims at enriching each variant with useful information to explore the impact of the genotype on the phenotype.

Different types of information can be associated with each variant and may help in better understand their role. The first level of information concerns the affected genomic area. In fact, a variant can fall into an intergenic region between two different genes (intronic variants), or it can affect a protein-encoding gene (exonic variant). Since several transcripts can be associated with a unique gene and the variant may fall into different functional zones according to the analyzed transcript, precise information is generated for each transcript. The choice of transcripts is a relevant contributor to the interpretation of the genetic test as a variant may have different roles in different transcripts of the same gene. Several transcript databases (Ensembl [58], RefSeq [59], and UCSC [60]) with which variants

can be annotated exist. The second level concerns the functional description of the variant with respect to the transcript. Essential information include: the type of consequence of the variant on the transcript (e.g., synonymous, missense, stop gain, etc.), the nucleotide changes in the coding sequence (HGVS nomenclature for cDNA sequence changes - HGVS<sub>c</sub>) and amino acid change in the protein (HGVS nomenclature for protein sequence changes - HGVS<sub>p</sub>). Other key annotations are those obtained from variant databases. Many databases (table 2.1) provide information of different nature: clinical databases such as ClinVar and Uniprot [61] contain information on the impact of variants on clinical phenotypes, population databases such as dbSNP, 1000 Genomes Project database, ExAC [62], and GnomAD [63] report the frequency with which the variant was observed in large groups of subjects, and finally, databases such as OMIM that contains information about Gene-disease associations. Some databases cover specific genes such as BRCA Exchange [64] (which reports information on BRCA1 and BRCA2), others such as COSMIC [65] contain a multitude of information only on genes and variants identified on somatic tissue. The last level of annotation is the one based on the tools that provide a damage prediction score generated with different approaches, such as protein structure, sequence homology, evolutionary conservation or statistical prediction based on known mutations. In table 2.2 are reported some of in silico prediction tool commonly used for variants annotation.

**Table 2.1:** Useful databases for variant interpretation

<b>Population Databases</b>	
Exome Aggregation Consortium <a href="http://exac.broadinstitute.org/">http://exac.broadinstitute.org/</a>	Database of variants found during exome sequencing of 61,486 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Pediatric disease subjects as well as related individuals were excluded.
Genome Aggregation Database <a href="https://gnomad.broadinstitute.org/">https://gnomad.broadinstitute.org/</a>	Database of variants found during exome sequencing of several large cohorts of individuals of European and African American ancestry. Includes coverage data to inform the absence of variation.
1000 Genomes <a href="http://browser.1000genomes.org">http://browser.1000genomes.org</a>	Database of variants found during low-coverage and high coverage genomic and targeted sequencing from 26 populations.
dbSNP <a href="http://www.ncbi.nlm.nih.gov/snp">http://www.ncbi.nlm.nih.gov/snp</a>	Database of short genetic variations (typically 50 bp or less) submitted from many sources. May lack details of originating study and may contain pathogenic variants.
dbVar <a href="http://www.ncbi.nlm.nih.gov/dbvar">http://www.ncbi.nlm.nih.gov/dbvar</a>	Database of structural variation (typically greater than 50 bp) submitted from many sources.

<b>Disease Databases</b>	
ClinVar <a href="http://www.ncbi.nlm.nih.gov/clinvar">http://www.ncbi.nlm.nih.gov/clinvar</a>	Database of assertions about the clinical significance and phenotype relationship of human variation.
OMIM <a href="http://www.omim.org">http://www.omim.org</a>	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database <a href="http://www.hgmd.org">http://www.hgmd.org</a>	Database of variant annotations published in the literature. Requires fee-based subscription for much of the content.
<b>Sequence Databases</b>	
NCBI Genome <a href="http://www.ncbi.nlm.nih.gov/genome">http://www.ncbi.nlm.nih.gov/genome</a>	Source of full human genome reference sequences.
RefSeqGene <a href="http://www.ncbi.nlm.nih.gov/refseq/rsg">http://www.ncbi.nlm.nih.gov/refseq/rsg</a>	Medically relevant gene reference sequence resource
MitoMap <a href="http://www.mitomap.org/MITOMAP/HumanMitoSeq">http://www.mitomap.org/MITOMAP/HumanMitoSeq</a>	Revised Cambridge reference sequence (rCRS) for the Human Mitochondrial DNA

**Table 2.2:** In silico prediction tools

Name	Basis
<b>Missense prediction</b>	
ConSurf <a href="https://consurf.tau.ac.il/">https://consurf.tau.ac.il/</a>	Evolutionary conservation
FATHMM <a href="http://fathmm.biocompute.org.uk/">http://fathmm.biocompute.org.uk/</a>	Evolutionary conservation
PANTHER <a href="http://www.pantherdb.org/">http://www.pantherdb.org/</a>	Evolutionary conservation
SIFT <a href="https://sift.bii.a-star.edu.sg/">https://sift.bii.a-star.edu.sg/</a>	Evolutionary conservation
SNPs&GO <a href="https://snps-and-go.biocomp.unibo.it/">https://snps-and-go.biocomp.unibo.it/</a>	Protein structure/function
Align GVGD <a href="http://agvgd.hci.utah.edu/">http://agvgd.hci.utah.edu/</a>	Protein structure/function and evolutionary conservation
MAPP <a href="http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html">http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html</a>	Protein structure/function and evolutionary conservation
MutationTaster <a href="https://www.mutationtaster.org/">https://www.mutationtaster.org/</a>	Protein structure/function and evolutionary conservation
MutPred <a href="http://mutpred.mutdb.org/">http://mutpred.mutdb.org/</a>	Protein structure/function and evolutionary conservation
PolyPhen-2 <a href="http://genetics.bwh.harvard.edu/pph2">http://genetics.bwh.harvard.edu/pph2</a>	Protein structure/function and evolutionary conservation

PROVEAN <a href="http://provean.jcvi.org/index.php">http://provean.jcvi.org/index.php</a>	Alignment and measurement of similarity between variant sequence and protein sequence homolog
Condel <a href="https://bbglab.irbbarcelona.org/fannsdb/help/condel.html">https://bbglab.irbbarcelona.org/fannsdb/help/condel.html</a>	Combines SIFT, PolyPhen-2 and MutationAssessor
CADD <a href="https://cadd.gs.washington.edu/">https://cadd.gs.washington.edu/</a>	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants
<b>Splice site prediction</b>	
GeneSplicer <a href="https://ccb.jhu.edu/software/genesplicer/">https://ccb.jhu.edu/software/genesplicer/</a>	Markov models
Human Splicing Finder <a href="http://www.umd.be/hسف">http://www.umd.be/hسف</a>	Position-dependent logic
MaxEntScan <a href="http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html">http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html</a>	Maximum entropy principle
NetGene2 <a href="https://services.healthtech.dtu.dk/service.php?NetGene2-2.42">https://services.healthtech.dtu.dk/service.php?NetGene2-2.42</a>	Neural networks
NNSplice	Neural networks
<b>Nucleotide conservation prediction</b>	
GERP <a href="http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html">http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html</a>	
PhastCons <a href="http://compugen.cshl.edu/phast">http://compugen.cshl.edu/phast</a>	
PhyloP <a href="https://ccg.epfl.ch/mga/hg19/phyloP/phyloP.htm">https://ccg.epfl.ch/mga/hg19/phyloP/phyloP.htm</a>	

## 2.4.2. Variant prioritization

Once all the information about the functional effect of the variant in the different transcripts, its frequency in the population, the damage prediction scores on the protein, and the notes about the gene-disease association have been collected, these are useful for the prioritization process. The goal of prioritization is to pass from thousands of variants to a small group of one or two variants that are candidates to be responsible for the observed phenotype. The prioritization process can be seen as a cascade of filters applied to variant annotations, guided by a reasonable method based on the specific clinical context.

For hereditary diseases, the in-depth phenotyping of the carrier of the variants and his family members, together with the study of the pedigrees, is the starting point for validating the result of the prioritization process

obtained with functional information and bioinformatics data. The definition of the inheritance model of the disease allows excluding all the variants that do not respect the principles of co-segregation. For recessive diseases homozygous variants inherited from two parents (often consanguineous), or two heterozygous variants, each passed by a different parent, are sought; for a dominant inheritance model, the candidates will be heterozygous variants inherited from a sick parent or private variants (de novo variant). More attention is needed for the evaluation of X-linked phenotypes or mitochondrial diseases associated with mutations in mitochondrial DNA.

Once the characteristics of the phenotype and its inheritance model have been evaluated, it is possible to interrogate genes that are plausible candidates in case of a causative variant. A thorough understanding of the molecular causes of the phenotypes is essential in order to narrow the spectrum of genes of interest. The gene-disease association data extracted from clinical databases such as ClinVar or OMIM, the information reported in the literature, and the geneticist's experience can guide the ranking of the genes to be investigated, excluding all those related to phenotypes far from the one under investigation.

One of the primary criteria for predicting if a variant is likely to have a functional effect on the encoded protein is a rarity. A commonly used threshold to exclude a variant from potentially harmful ones is a Minor Allele Frequency (MAF) in population databases greater than 1%. This threshold may vary according to the incidence of the disease and the level of penetrance expected for the phenotype. Often the causative variants of the disease are extremely rare and unreported in the population databases, while others are instead observed in various subjects considered healthy at the date of clinical control but who may have developed the disease later during the life. If a variant is common in the population, it almost certainly has a neutral effect on the protein, but a rare variant may still be benign. For this reason, the choice of the MAF threshold to be applied for prioritization must adapt to specific issues.

Another contributor helping the characterization of a variant is its functional impact on the transcript. In fact, different types of variants are associated with different levels of protein damage: intronic variants far from gene regulation sites, variants in UTR, and synonymous variants have a very low probability of causing disease; missense, exonic insertions, deletions, stop losses and start losses, carry a greater potential of damaging the protein function; while stop gains, frameshifts, and splice site variants are of primary interest for their protein-truncating effect. Also, in this case, special caution must be considered in filtering out the variants: it is possible that some variants with a low impact potential may instead be the cause of hidden protein damage. An important example is the case of synonymous variants that cause a cryptic splicing site within an exon [66][67].

The functional impact prediction from *in silico* tools can help refine prioritization. Many prediction tools and the diversity of algorithms with which the damage score is calculated, can cause interpretation difficulties linked to the discordant predictive results. For this reason, there is no

standard filtering strategy for this type of data, but it must be adapted to the molecular context of the disease.

The result of the prioritization process is composed of a narrow subset of variants with a reasonably high probability of causing the phenotype, and a larger group of variants with a low probability of affecting protein function. This ranking process has no standard rules and is subject to many variables that could change the accuracy of the result. Furthermore, the candidate variants are not interpreted as causative or neutral, but their role is defined in a descriptive way.

### **2.4.3. ACMG-AMP classification system**

If the prioritization aims at minimizing the number of variants that disease-causing candidates, the classification of the variants is the process aimed at interpreting their specific role on the phenotype. The functional filtering process of variants is guided by rules that often vary from laboratory to laboratory, producing heterogeneity in interpretation. The need for common rules to homogenize the results of genetic tests has prompted the scientific community to devise a robust method for pathogenic classification. In 2015, the American College of Molecular Genetics (ACMG) together with the College of American Pathologists (AMP) developed guidelines for the interpretation of the role of Mendelian and mitochondrial variants (68). The old term that defined "mutation" as a causative variant and "polymorphism" as a non-causative variant has been replaced by a system based on 5 classes: 1. Benign (B - non-causative variant of disease), 2. Likely Benign (LB - probably not causing disease); 3. A variant of Uncertain Significance (VUS); 4. Likely Pathogenic (LP - probably causative of disease); 5. Pathogenic (P - Definitely causative of disease).

The pathogenicity class is defined on a system that evaluates the combinations of 28 criteria activated by the different types of information available on the variant and their relative strength. Different sources of information are evaluated:

- Population data such as frequencies of variants in large populations and prevalence in control groups.
- Computational and prediction data that consider the functional effect of the variant and damage mechanism to which the affected gene is sensitive, the existence of variants whose role is established affecting the same nucleotide or amino acid, and the results of the in-silico protein damage prediction tools.
- Functional data produced through in-vivo, ex-vivo, and in-vitro studies, aimed at determining the consequence of the variant on the affected protein and on the cellular phenotype.
- Clinical and segregation studies aimed at verifying the specificity of the clinical picture of the carrier and family members, the co-segregation of the variant with the phenotype in case of hereditary

disease, the possibility that a mutational event occurred de novo, the presence of other variants that potentially cause the phenotype, and finally the cis or trans status of two variants identified in the same gene.

- Information derived from reputable sources such as peer-reviewed literature and from curated disease databases such as ClinVar.

The criteria are divided into ones in favor of the benign role of the variant (12 criteria) and criteria in favor of the pathogenic role (16 criteria) based on the evidence reported by the analyzed information. Each criterion is associated with a weight (strength) that reflects the level of evidence in favor of the benign or pathogenic interpretation, and which determines the strength with which the single criterion guides the final classification. Each pathogenic criterion is weighted as very strong (PVS1), strong (PS1–4), moderate (PM1–6) or supporting (PP1–5) and each benign criterion is weighted as stand-alone (BA1), strong (BS1–4), or supporting (BP1–7). The default strength levels were calculated during the validation phase of the ACMG system, but to improve the flexibility of the model, the weights can be modified based on the evidence supporting each criterion.

	Benign			Pathogenic		
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
<b>Population Data</b>	MAF is too high for disorder <i>BA1/BS1</i> OR observation in controls inconsistent with disease penetrance <i>BS2</i>			Absent in population databases <i>PM2</i>	Prevalence in affecteds statistically increased over controls <i>PS4</i>	
<b>Computational And Predictive Data</b>		Multiple lines of computational evidence suggest no impact on gene /gene product <i>BP4</i> Missense in gene where only truncating cause disease <i>BP1</i> Silent variant with non predicted splice impact <i>BP7</i>	Multiple lines of computational evidence support a deleterious effect on the gene /gene product <i>PP3</i>	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before <i>PM5</i> Protein length changing variant <i>PM4</i>	Same amino acid change as an established pathogenic variant <i>PS1</i>	Predicted null variant in a gene where LOF is a known mechanism of disease <i>PVS1</i>
<b>Functional Data</b>	Well-established functional studies show no deleterious effect <i>BS3</i>		Missense in gene with low rate of benign missense variants and path. missenses common <i>PP2</i>	Mutational hot spot or well-studied functional domain without benign variation <i>PM1</i>	Well-established functional studies show a deleterious effect <i>PS3</i>	
<b>Segregation Data</b>	Non-segregation with disease <i>BS4</i>		Co-segregation with disease in multiple affected family members <i>PP1</i>	Increased segregation data →		
<b>De novo Data</b>				<i>De novo</i> (without paternity & maternity confirmed) <i>PM6</i>	<i>De novo</i> (paternity & maternity confirmed) <i>PS2</i>	
<b>Allelic Data</b>		Observed in <i>trans</i> with a dominant variant <i>BP2</i> Observed in <i>cis</i> with a pathogenic variant <i>BP2</i>		For recessive disorders, detected in <i>trans</i> with a pathogenic variant <i>PM3</i>		
<b>Other Database</b>		Reputable source w/out shared data = benign <i>BP6</i>	Reputable source = pathogenic <i>PP5</i>			
<b>Other Data</b>		Found in case with an alternate cause <i>BP5</i>	Patient's phenotype or FH highly specific for gene <i>PP4</i>			

**Figure 2.7:** Data sources and level of strength for ACMG criteria (Figure from [68])



Each single criterion is not able of determining the class of pathogenicity alone, except for the one that evaluates the allele frequency in the population. In fact, a MAF such as to define the variant as common is a sufficient criterion for the Benign classification (Stand-Alone strength). The pathogenic class is assigned based on the combination of weighted criteria that are activated during the analysis (table 2.3). If a variant does not fulfill the criteria to gain a benign or pathogenic class, or the evidence for benign and pathogenic is conflicting, the variant must be classified as Uncertain Significance (VUS).

Table 2.3: ACMG criteria combination for determining pathogenicity class

<b>Pathogenic</b>
1 Very Strong (PVS1) <i>AND</i> $\geq 1$ Strong (PS1–PS4) <i>OR</i> $\geq 2$ Moderate (PM1–PM6) <i>OR</i> 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) <i>OR</i> $\geq 2$ Supporting (PP1–PP5) $\geq 2$ Strong (PS1–PS4) <i>OR</i> 1 Strong (PS1–PS4) <i>AND</i> $\geq 3$ Moderate (PM1–PM6) <i>OR</i> 2 Moderate (PM1–PM6) <i>AND</i> $\geq 2$ Supporting (PP1–PP5) <i>OR</i> 1 Moderate (PM1–PM6) <i>AND</i> $\geq 4$ Supporting (PP1–PP5)
<b>Likely Pathogenic</b>
1 Very Strong (PVS1) <i>AND</i> 1 Moderate (PM1–PM6) <i>OR</i> 1 Strong (PS1–PS4) <i>AND</i> 1–2 Moderate (PM1–PM6) <i>OR</i> 1 Strong (PS1–PS4) <i>AND</i> $\geq 2$ Supporting (PP1–PP5) <i>OR</i> $\geq 3$ Moderate (PM1–PM6) <i>OR</i> 2 Moderate (PM1–PM6) <i>AND</i> $\geq 2$ Supporting (PP1–PP5) <i>OR</i> 1 Moderate (PM1–PM6) <i>AND</i> $\geq 4$ Supporting (PP1–PP5)
<b>Benign</b>
1 Stand-Alone (BA1) <i>OR</i> $\geq 2$ Strong (BS1–BS4)
<b>Likely Benign</b>
1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) <i>OR</i> $\geq 2$ Supporting (BP1–BP7)

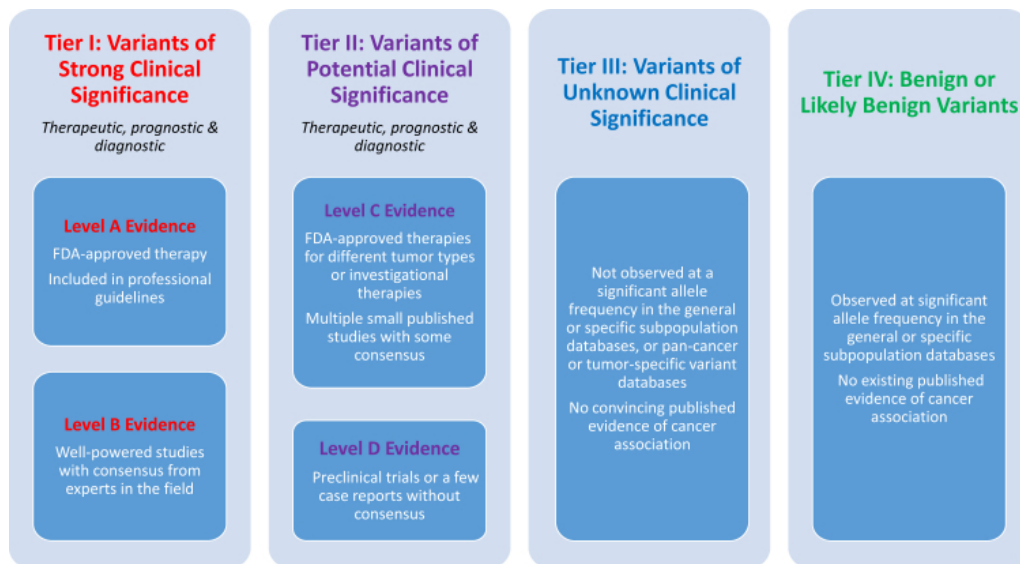
The variant classification process is a very hot topic for genetics. The exponential increase in genetic tests and the ever-increasing sequencing capacity of the new tools has caused a sharp increase in the time and costs required for the interpretation of the results. In response to this need, many tools have been developed to simplify and speed up the decision-making process. These tools, both commercial and open sources [69][70][71], group in a functional way all the information available on the variant and implement automatic algorithms for “activating” the criteria and modulating

their weights, improving the data analysis capacity and simplifying the interpretation of large sets of variants.

### 2.4.3.1. Somatic Variants

The evaluation of the somatic variants is instead carried out on two different levels, especially in cancer sequencing applications. The first type is a functional type on the nature of cancer and is carried out to distinguish which acquired variants give an evolutionary advantage to the tumor cells and are drivers for the generation of new and more aggressive subclones, and which ones have a neutral impact in the progress of the illness (passenger). Driver variants usually have a gain-of-function effect for proto-oncogenic genes (commonly missense variants) and a loss-of-function effect for tumor suppressor genes.

The second type is focused on the variant impact on clinical care. An actionable variant can be considered a predictive biomarker for sensitivity or resistance to therapies, can be targeted for new generation drugs, can take on prognostic significance, facilitate early diagnosis and guide preventive actions. Based on the available evidence, a clinical impact-driven categorization system has been proposed [72] based on 4 classes of variants: tier I, variants with strong clinical significance; tier II, variants with potential clinical significance; tier III, variants with unknown clinical significance; and tier IV, variants that are benign or likely benign (figure 2.8).



**Figure 2.8:** Evidence-based somatic variant categorization in cancer sequencing. (Figure from [72])

# Chapter 3

---

## The Helper platform

In the background chapter, the technological scenario of NGS technology was introduced in a comprehensive synthetic presentation. It was shown which are the fundamental steps in the process of identifying the genetic causes for hereditary diseases and which is the role of bioinformatics analysis in the diagnostic process. The present chapter describes the Helper platform, developed to simplify the design and execution of bioinformatics pipelines for NGS sequencing data. The chapter introduces the needs, and the solutions present in the literature for pipeline development, the implementation structure of Helper and its operating principles. The implemented algorithms, and how it works in the pipelines are described for each tool that can be used in the Helper platform. Then the Helper's graphical interface is presented, and finally the results of the performances, in terms of processing time and CNV calling, are discussed for a pipeline developed for the analysis of NGS libraries commonly used at OSM.

### 3.1. Needs and motivations

The development of new pipelines addressing specific issues involving NGS sequencing is an ever-evolving field. There is an increasing need for simple systems for customizing bioinformatics analyses, overcoming the coding difficulties. One of the pioneering projects that promoted this trend of making user-friendly both bioinformatics and pipeline development, is Galaxy [73]. Galaxy is a web-based platform developed for making analysis completely reproducible and accessible to all researchers.

Galaxy implements a large number of bioinformatics solutions for the analysis and manipulation of data from different types of experiments (Genomics, RNA-seq, Chip-seq, etc.). Galaxy is an open system that provides a large choice of tools, excellent documentation, ability to run the analysis in cloud, and the support of an extensive community.

Despite all the advantages of systems like Galaxy, tools dedicated to specific applications are often required. Dedicated systems focus on the problem, simplifying the user experience in terms of understanding the processes and using the platform. In recent years, various systems for the

customization of bioinformatics analysis have been created for different NGS applications [74][75][76].

The Helper platform fits into this landscape as a solution designed to simplify the development of new bioinformatics pipelines for the analysis of NGS data from Illumina sequencing of DNA samples for targeted sequencing applications. The need for a simple and fast tool for the development of new pipelines arose because of the various research projects active at the Center for Inherited Cardiovascular Diseases of the OSM. The heterogeneity of different projects, of the different analysed samples, and of the different technologies of the sequencing kits, translates into greater complexity of adaptation of the bioinformatics pipelines. The development of pipelines must take into account a multitude of factors that affect the design of the project and requires several phases, such as code development, testing, debugging and validation of results. Without adequate coding experience, the development of new bioinformatics pipelines could be a difficult path. Helper aims to relieve the users from writing new code by guiding them through a simple graphical interface in the implementation and use of new pipelines easily adaptable to the context and specific needs. Unlike systems such as Galaxy, which provide tools for multiple bioinformatics applications, Helper is a platform dedicated to the analysis of data derived from DNA target re-sequencing experiments. This specific setting guarantees an optimized management of the analyzed data and a user-friendly experience in using Helper. The user is able to choose which steps to include in the pipeline, which software to use in the different steps of the analysis, and which parameters to run the different tools, in the context of a workflow aimed to the identification and interpretation of variants. The Helper platform is deposited in the OSM repository (protocol number 0102850/21) and is accessible upon request.

### 3.2. Workflow management system

Helper is a platform developed in Python3 compatible with Ubuntu 16 and 18 operating systems. Helper requires the installation of a few dependencies:

- *PyQT5* for the execution of the graphic interface.
- *Json* for the decoding and encoding of the configuration files necessary for the operation of Helper.
- *Argparse* for the implementation of the Helper main script argument system.
- *Subprocess* for the execution and parallelization of tools.

The tools used in the bioinformatics pipelines are many, developed using different languages and each need specific dependencies in order to be used. To safely install the tools and their dependencies without incurring the danger of changing the work environment, it is advisable to use an

environment management system such as Conda (<https://docs.conda.io/>). Conda is an open-source package management system that allows to quickly install, run and update packages and their dependencies. The list of tools that can be used for the execution of pipelines with Helper, and the description of the workflow is described in Chapter 2.3.

### 3.2.1. Tools wrapping and parallelization

Integration of software developed in different languages is not always simple and represents an obstacle for an inexperienced bioinformatician. The wrappers have been designed to simplify the use of these tools and integrate them more easily. In the world of software, a wrapper is a code that wraps or covers other functions or tools. It can be thought of as a sandwich that contains several ingredients, making them easier to use.

The Helper platform uses a complex wrapper system that implements the functions of 25 external tools and scripts and allows their calling using just Python language. The tools must be installed locally and have to be compatible with the versions supported by the platform. Within Helper, tools are coded as an object that contains a series of methods called through the wrappers. Before being able to call a function of a tool, it is necessary to initialize it by supplying the path of the main script responsible for the execution of the tool, the amount of RAM to be dedicated, the number of threads to be used (in case the tool that implements the multithreading), and the set of parameters needed for execution. The wrappers implemented in Helper all have a similar structure (Figure 3.1); they have three ports:

- A setting port to which the tool initialization information is provided;
- An input port to which the files to be processed are provided, the accessory files required by the tool function, the log file in which to keep track of the processing result, the working directory in which to save the output files.
- An output port that is used by the wrapper to return the files produced by the execution of the function.

The module used to manage the wrapping is *subprocess* that allows to call and run external software, connect inputs, outputs, and errors in pipe, and monitor the status of the process. Furthermore, the *subprocess* module provides the ability to improve workflow efficiency thanks to the parallel execution of processes. Each function is performed on several samples at the same time, significantly reducing processing times.

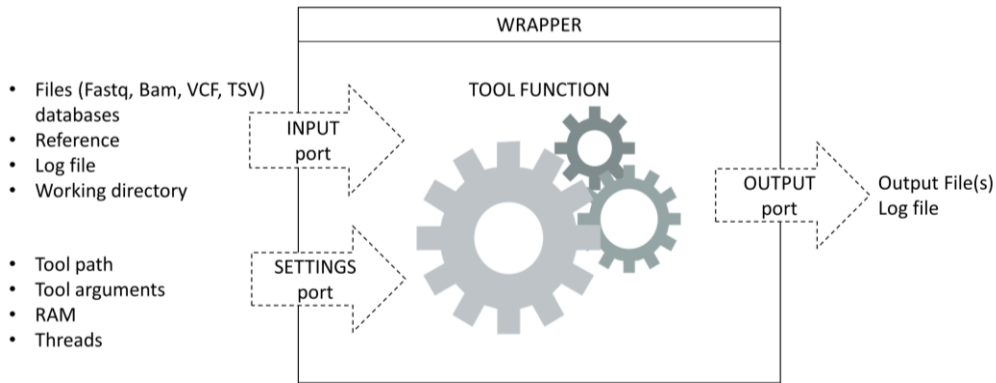


Figure 3.1: The structure of the wrappers

### 3.2.2. The Helper structure

Helper consists of a simple Graphical User interface (GUI) for the preparation of the configuration files necessary for the execution of the analysis, and of a back-end for the execution of the analysis process. The back end of Helper is composed of the main script in which the execution of the pipelines is managed (*pipeline.py*), and of four libraries of functions. The *function.py* library includes all the functions for managing directories within the analysis folder, the functions to support the reading of configuration files and those for interpreting the samplesheet. In the *tools.py* and *parallel\_tools.py* libraries are implemented the wrappers of the tools and software necessary for the serial and parallel analysis, correspondingly. Finally, the *scripts.py* library contains all the wrappers that call the in-house scripts dedicated to the execution of some Helper steps.

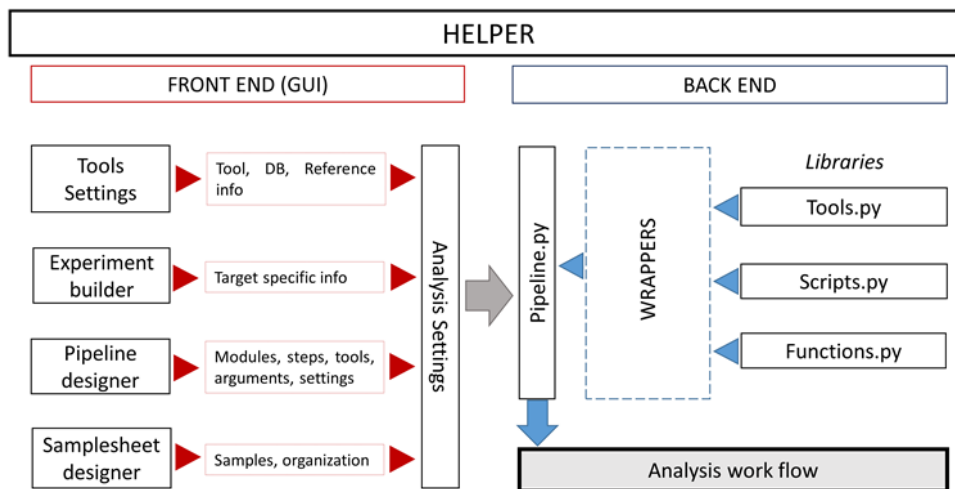


Figure 3.2: The Helper structure.

### 3.2.2.1. The configuration files

Helper takes all the information necessary to carry out the analysis from the configuration files (CF). All CFs are implemented in JSON format [<https://www.json.org>] with information nested at different levels. For the execution of Helper 4 CF are required:

1. The tools configuration file
2. The experiment configuration file
3. The pipeline configuration file
4. The samplesheet

The tools configuration file (*Tools.cfg*) contains useful information to recall the tools, databases, and reference genomes during the analysis. The CF contains the list of the tools implemented in Helper (*tools.list*), and a set of lists of tools that specify which one can be used during each step of the analysis (for example *tools.fastq\_alignment* or *tools.variant\_calling*). Similar lists for databases and reference genomes are present in the CF (*databases.list*, *genomes.list*). Furthermore, in the CF all the information necessary to use each tool, database, and reference genome, is specified. For the tools, the path of the main script or jar file (for example *GATK.path*), the tag that indicates in which step of the analysis it can participate (for example *GATK.tag*), and the version of the tool (*GATK.version*) are indicated. For databases, the path to the DB, the tags of the steps in which they can be used, and other accessory information are specified. For the reference genomes, the path of the fasta file, the “dict” file, and the version must be indicated. During the implementation phase of the pipeline via GUI, Helper checks which tools can be used in each step and proposes them to the user. During the analysis processing, at the beginning of each step of the pipeline, the main script reads the need information about selected tools from the *tools.cfg*.

Example of tool:

```
"GATK v.4.1": {
  "path": "/NGS_TOOLS/GATK/v4.1.2.0/gatk",
  "version": "4.1.2.0",
  "tags":
  "preprocessing,variantcalling,cnv_calling"}
```

Example of database:

```
"dbNSFP": {
  "path": "/dbSNFP/dbNSFP4.0/dbNSFP4.0a_hg19.gz",
  "files": "/dbSNFP/dbNSFP_replacement_logic",
  "version": "4.0",
  "tags": "database"}
```

Example of genome reference:

```
"GRch37": {
  "fasta": "/NGS_REF/hg19/GRch37.fasta",
  "dict": "/NGS_REF/hg19/GRch37.dict",
  "version": "37"}
```

The experiment configuration file contains information regarding the sequencing target. For each experiment contained in the list, the following information are reported: the identification name of the panel; the technology used to select the target during the sample preparation step (Capture Enrichment or Amplicon); the list of genes (one entry per line); the list of transcripts to be analyzed (one per gene and one per row); files describing the target and “BED” format and in “List” format. The last fields are specific for the analysis of the CNVs, and in the example shown below, the directories of models and the target needed by GATK to perform the CNV call are indicated.

```
"TrusightCardio": {
  "panel_name": "TrusightCardio",
  "panel_technology": "Capture Enrichment",
  "gene_list":
"/TARGET/gene_list_Trusightcardio.txt",
  "transcripts_list":
"/TARGET/transcriptList_Trusightcardio.txt",
  "target_list":
"/TARGET/Trusightcardio_manifest.list",
  "target_bed":
"/TARGET/Trusightcardio_manifest.bed",
  "cnv_calls_model": "/CNV/GATK/Trusightcardio-
model",
  "cnv_ploidy_model": "/CNV/GATK/Trusightcardio-
model",
  "cnv_target_list":
"/CNV/Trusightcardio_manifest.CNV.list"
}
```

The pipeline configuration contains information about the workflow and which tools have to be used. The CF contains fields that specify the ID of the pipeline, the type of analysis (Germline or Somatic), the name of the genome reference (as reported on the tools CF), and the workflow as a list of modules that the pipeline must execute. Each module contained in the workflow list is then described in an independent block with a series of nested information: the RAM and the number of threads to be used during the module processing, the workflow of the module reported as a list of steps, and the information about each step as separate blocks. For each step, the tool (or the list of tools in the case that the step can be executed by several tools such as variant calling) and the tool parameters (see example below) are reported. The modules and the steps are explained in Chapter 2.3.

```
Pipeline configuration file example:
"id": "helper_test",
"analysis": "Germline",
"reference_version": "hg19",
"workflow": ["alignment"],

"alignment": {
```



```

        "workflow": ["fastq_alignment",
"sam_to_bam", "sortSam", "bam_QC"],
        "threads": "2",
        "ram": "1g",
        "fastq_alignment": {
            "tool": "BWA",
            "BWA v.0.7.17": {"args": [], "algorithm":
"mem"}
        },
        "sam_to_bam": {
            "tool": "PICARD v.2.7.1",
            "PICARD v.2.7.1": {"args": []}
        },
        "sortSam": {
            "tool": "PICARD v.2.7.1",
            "PICARD v.2.7.1": {"args":
["SORT_ORDER=coordinate"]}
        }...
    }
}

```

The last configuration file is the samplesheet which contains information on the samples. The samplesheet contains the list of samples to be analyzed (sample\_list), the organization of the samples (sample\_organization) and the files about the samples organized in blocks, one for each module (module block) of the pipeline. Within each module block, samples are structured into blocks (sample block) based on sample organization: "only case" (example 1), "case-control" (example 2), or trio (example 3). Each sample block is identified by the ID of the case sample. Sample blocks contains information about each sample associated with the case sample, and identified by role (case, control, or parent). The information concerns the name of the sample and the files to be processed in the specific module of the pipeline.

Example 1: "only case" sample organization

```

"variantcalling": {
    "sample1": {
        "case": {
            "sample_name": " sample1",
            "bam": " sample1.bam"},
        },
    "sample2": {
        "case": {
            "sample_name": " sample2",
            "bam": " sample2.bam"},
        },
    ...
}

```

Example 2: "case-control" sample organization

```

"variantcalling": {
    " sample1": {
        "case": {
            "sample_name": " sample1",
            "bam": " sample1.bam"},
        },
    ...
}

```

```

        "control": {
            "sample_name": " sample2",
            "bam": " sample2.bam"},
        },
        ...
    }

```

Example 3: "trio" sample organization

```

"variantcalling": {
    "sample1": {
        "case": {
            "sample_name": " sample1",
            "bam": " sample1.bam"},
        "parent1": {
            "sample_name": " sample2",
            "bam": " sample2.bam"},
        "parent2": {
            "sample_name": " sample3",
            "bam": " sample3.bam"}
    },
    ...
}

```

### 3.3. Data processing workflow

The main script that performs sample analysis is the *Pipeline.py*. The script is executed by the Analysis\_designer interface or using a simple command line. The required input arguments are summarized in the following table:

**Table 3.1:** Input arguments to *Pipeline.py* script

Argument	Description	Comments
<i>--tools_cfg</i> <b>[file]</b>	Tools config file path	Default: Helper_dir/configs/tools_cfg/tools.cfg
<i>--samplesheet</i> <b>[file]</b>	Samplesheet file path	<i>Required argument</i>
<i>--panel</i> <b>[string]</b>	The experiment name contained in the Experiment.list file	<i>Required argument</i>
<i>--pipeline</i> <b>[file]</b>	Pipeline config file path	<i>Required argument</i>
<i>--run_id</i> <b>[string]</b>	The identification name of the analysis	<i>Required argument</i>
<i>--workdir</i> <b>[directory]</b>	The path of the working directory	<i>Required argument</i>
<i>--workflow</i> <b>[string]</b>	The list of major steps confirmed in the analysis launcher	Default: all the major steps
<i>--parallel</i> <b>[bool]</b>	Enable parallel analysis	Default: Not activated (False)

<code>--del_temp</code> <i>[bool]</i>	Delete temporary files	Default: Not activated (False)
--	------------------------	--------------------------------

The structure of Pipeline.py is modular and the workflow is managed in automatic way. The complete workflow provided by Helper is composed of seven major modules (pre-alignment, alignment, preprocessing, variant calling, CNV calling, post processing, annotation, post-annotation), each of which includes a variable number of other sub-steps. Helper extracts the complete workflow from the pipeline configuration file but only executes the modules confirmed by the “--workflow” argument. For example, if the Pre-alignment module is provided in the pipeline workflow, but the user wants to start analyzing the data directly from the Alignment, he should omit the pre-alignment in the “--workflow” argument. It is important that compatibility is maintained between files produced by the previous module and entering the next. For example, it is not possible to omit the variant calling module if you want to proceed later with the post-processing or annotation phase. In fact, these last two modules require VCF files that can only be produced by variant callers.

The files produced by a module are traced by updates of the samplesheet. Each module receives as input the updated samplesheet from the previous module with information on the files produced; the first module of the pipeline uses the samplesheet provided when launching the analysis. Once all the steps, provided in the module, have been performed, the last files produced are stored in the dedicated directories and their path are indicated in the updated samplesheet. In order to track the workflow and to simplify troubleshooting, Helper implements a log file system in which the STD-OUT and STD-ERROR of the tools are printed. If an error occurs in reading an input file or due to an incorrect parameter, Helper specifies in which step of the pipeline the problem occurred and reports the error message, issued by the failing tool, in the log file of the specific step.

### 3.3.1. Pre-alignment process

The first module of the workflow is the processing of the Fastq files before the alignment of the sequences, and is needed to improve the quality of the data. The pre-alignment consists of four possible sub-steps that are performed on each sample: Adapter Trimming, Fastq filtering (using Read mean quality or read length), and Fastq quality control.

#### 3.3.1.1. Trimming of adapters

Removal of adapter sequences (read trimming or clipping) is the first steps in analyzing NGS data. Adapter contamination will lead to NGS alignment errors and an increased number of unaligned reads, since the adapter sequences are synthetic and do not occur in the genomic sequence. In Illumina sequencing, adapter sequences will only occur at the 3' end of the

read and only if the DNA fragment is shorter than the number of sequencing cycles. For applications where the fragment size is well calculated, adapter contamination is expected to be small, and the adapter removal step can be skipped saving time and efforts.

The tools implemented in Helper among which it is possible to choose for the removal of adapter sequences are AGeNT [77] and Cutadapt [78]. The Agilent Genomics NextGen Toolkit (AGeNT) is a Java-based (Java 8) software module that processes specifically sequencing data obtained via Agilent libraries (SureSelect and Haloplex) and should not be used for experiments performed with other kits (e.g., Illumina, IDT, and Roche). AGeNT is a command-line tools collection that contains a module for managing molecular barcodes (LocatIt) and a module that removes the adapter sequences (Trimmer). To use AGeNT Trimmer function, the indication of adapters sequences it is not necessary, because they are automatically recognized by indicating, the specific tag (as input to the tool) of the library used for the preparation of the samples. Cutadapt is a very simple tool to use and allows trimming both single and paired-end reads obtained from any type of NGS library. Cutadapt, unlike AGeNT, requires in input the sequences of the adapters to be removed, indicating the position in the read (in 3', 5' or in the middle of the read). In the case of paired-end sequencing, the reverse strand adapters are also required. The wrapper for the AGeNT Trimmer function (*AGeNT.Trimmer*) takes in input the path of the AGeNT tool, the Fastq paired files, the specific tag for the NGS library, the address of the reference genome, and other ancillary topics. The wrapper for the Cutadapt trimming function (*Cutadapt.Trim\_Adapters*) requires the path of the Cutadapt tool, the number of threads to be used for analysis, Fastq paired files, the adapter sequence in 3' forward and the sequence of the adapter in 3' reverse, the path to the reference genome, and other optional arguments. In output from the trimming step of the adapters, you get the two trimmed Fastq paired files.

### 3.3.1.2. Fastq filtering

The filtering of Fastq files is a useful step to improve the quality of the data, excluding reads with low base quality and selecting the reads within a certain length range. The Fastq filter step potentially increases the accuracy of the NGS analysis but in the case of good quality sequencing experiments it is possible to skip it.

To perform the filtering of the Fastqs, Cutadapt was implemented using two different wrappers, one for filtering using mean quality (*Cutadapt.Fastq\_fiter\_Qual*) and one for filtering using read length (*Cutadapt.Fastq\_fiter\_Len*). *Cutadapt.Fastq\_fiter\_Qual* requires in input A novel cryptic splice site mutation in ut the path of the tool, how many threads to use for analysis, Fastq paired files, the Base Qual threshold to be used to filter reads, and other ancillary arguments. In input, *Cutadapt.Fastq\_filter\_Len* requires the path of the tool, the number of

threads to use for analysis, the Fastq paired files, the maximum and minimum length thresholds that delimit the optimal range, and the other ancillary arguments. The output from each filtering step contains two filtered paired Fastq files.

### 3.3.1.3. Quality control

The Quality control step performs simple checks to ensure that the raw data are good and there are no problems or biases potentially affecting results. Typical metrics analyzed to assess the quality of NGS data are: quality base distribution in reads, GC mean content, contamination with adapter sequences and biases in base composition, sequence duplication, and reads length distribution [79].

The FAsTQC [80] tool has been implemented in Helper to perform the Quality control on Fastq files. FastQC is a perl script that parses Fastq, SAM and BAM file and produces an HTML report file that reports the data for the sample evaluation in graphical format and a zipped folder containing the results in TXT format. FastQC uses multiple modules to calculate the statistics:

1. The “Basic Statistics” module generates a descriptive summary of the analyzed sample indicating the file name and the file type, the encoding of quality values (e.g., Illumina), the total number of sequences processed, the number of sequences flagged as poor quality, the min and max read length in the sample, and the overall% GC of all bases in all sequences.

2. The “Per Base Sequence Quality” module shows an overview of the range of quality values across all bases at each position in the FastQ file.

3. The “Per Sequence Quality Scores” module reports the mean quality score distribution over all reads. If a significant proportion of the sequences in the run have overall low quality, then this could indicate some kind of systematic problem.

4. The “Per Base Sequence Content” module calculates the proportion of each called nucleotide for each base position in reads. An unbalance can indicate an overrepresented sequence which is contaminating library.

5. The “Per Sequence GC Content” module measures the GC content across all reads and compare it with a normal distribution from a random library. An unusually shaped distribution may indicate a contaminated library or some other kinds of biased subset.

6. The “Per Base N Content” module reports the percentage of N calls at each position. N are called when the sequencer is unable to make a base call with sufficient confidence and a increased percent of N suggest a low-quality sequencing.

7. The “Sequence Length Distribution” module generates a distribution of fragment sizes. More than one peak in the distribution means different sizes in the libraries that can introduce analysis biases.

8. The “Duplicate Sequences” module counts the degree of duplication for every sequence and calculates a distribution of duplication level.

9. The “Overrepresented Sequences” module finds all of the sequences which make-up more than 0.1% of the total and that can indicate some source of contamination.

### 3.3.2. Alignment

The next module is the alignment of the sequences into the Fastq files to obtain BAM files that contain the aligned reads. The workflow of the alignment phase consists of four mandatory steps (`Fastq_alignment`, `sam_to_bam`, `sortSam`, `indexBam`) and a fifth step of Bam quality control that is possible (but not recommended) to skip.

#### 3.3.2.1. Fastq\_alignment

`Fastq_alignment` is the step responsible for aligning the sequences contained in the Fastq files against the reference genome. The aligners implemented in Helper are BWA [81] and Bowtie2 [82]. BWA and Bowtie2 are two tools that implement alignment algorithms based on the Burrows-Wheeler Transform (BWT), they work well with paired-end reads, and are widely used for their accuracy and mapping speed. BWA implements three different algorithms: a) BWA-backtrack, b) BWA-sw, and c) BWA-mem. The BWA-backtrack algorithm is designed for short Illumina reads (up to 100bp), while both BWA-mem and BWA-sw are implemented for longer sequences and are very similar. BWA-mem is the last implemented algorithm, it is faster and more accurate than the other two and it is the generally recommended algorithm.

Bowtie2 allows choosing between two alignment algorithms: a) End-to-end alignment and b) Local alignment. End-to-end alignment is the one performed by default in Bowtie2, and it searches for alignments involving all the read bases without trimming the reads (untrimmed alignment), while Local alignment maximizes the alignment score by trimming some bases.

In order to perform the alignment, both tools require that the FASTA file containing the reference genome be indexed, each aligner using its own function.

```
// Building a reference index with bwa and bowtie2
bwa index [options] reference.fasta
bowtie2-build [options] reference.fasta output_dir
```

In Helper, the wrappers for BWA and Bowtie2 are implemented with the functions `Bwa.align_fastq` and `Bowtie2.align_fastq`, correspondingly. The functions ask in input the paths to the executables of the tools, the Fastq paired, the path to the reference genome for mapping the reads (the index files of the genome for the respective tools must be present in the same folder), the optional arguments of the tool, the alignment algorithm (only for

*Bwa.align\_fastq*), the log file, and the output directory. The output of the Fastq file alignment step is a file in SAM format containing the aligned reads.

### 3.3.2.2. Sam to Bam conversion

This step converts the format from SAM to BAM via the *SamFormatConverter* function of Picard toolkit [83]. Picard is a JAVA package of command line tools for manipulating files containing NGS data that has become part of the GATK best practices [43] for the implementation of NGS analysis pipelines.

The *Picard.SamFormatConverter* wrapper requests the path to the Picard Jar file, the SAM file to be converted, and the amount of RAM to use for processing. The converted BAM file is returned in output.

### 3.3.2.3. Sort Sam files

After Fastq alignment, the read contained in the SAM/BAM files are sorted in random order according to their positions in the Fastq files. In order to be usable in the subsequent steps of pre-processing and variant calling, the reads in Bam files must be ordered according to the chromosomal coordinates of the region in which they are mapped.

In Helper, the process of sorting Bam files is implemented through Picard tool and the wrapper for its sorting function (*picard.SortSam*) requires the path to the Picard Jar file, the Bam file, the amount of dedicated RAM and the optional arguments of the tool, the log file and the directory in which the sorted file has to be saved. The wrapper default is to sort by coordinates (*SORT\_ORDER=coordinates*). In output, a Bam file containing the reads sorted first by the reference sequence name (RNAME field), then by the mapping position (POS field), is released.

### 3.3.2.4. Index Bam files

Indexing a sorted Bam file allows a quick access to reads that are mapped in particular genomic regions and to extract alignment information quickly. The index file acts like an external table of contents and allows programs to jump directly to specific parts of the Bam file without reading through all of the sequences. Many tools require Bam files to be indexed in order to read them. The indexing of Bam files is performed using a Picard function. The *Picard.BuidBamIndex* wrapper requests the Bam file to be indexed and the amount of RAM to use and returns a file with the same name as the Bam file suffixed with “bai”.

### 3.3.2.5. Bam quality control

Bam files are evaluated using parameters similar to those of Fastq QC. In addition, the following indexes of quality are assessed: the average coverage of the target; the uniformity of coverage calculated as the percentage of the target with coverage in the range between 80% and 120% of the average coverage of the sample, and how much of the target exceeds the minimum acceptable coverage threshold (which depends on the application); the quality of alignment of the reads on the target in terms of fraction of reads that are mapped to the target (in-target and off-target reads); finally, the areas with low coverage compared to the rest of the target (gaps) are identified. In Helper the Bam quality control is implemented through FastQC. The *Fastqc.bam\_diagnosis* wrapper requests the input of the Bam file to be analyzed and returns the statistics for the evaluation of the sample.

### 3.3.3. Pre-processing

#### 3.3.3.1. Add readgroups to Bam file

Adding read groups to Bam files is a step that facilitates the analysis of samples by subsequent tools. The function is implemented through the *picard.AddOrReplaceReadGroups* wrapper, which requests the Bam file to be modified, the experiment ID, the analysis ID, and the sample\_name. The function modifies the fields present in the Bam file according to the following table.

**Table 3.2** - Tags modified during Add readgroups to Bam file

BAM TAG	FIELD
RGID	Sample name
RGPL	'ILLUMINA'
RGSM	Sample name
RGLB	Experiment ID
RGPU	Analysis ID

#### 3.3.3.2. Mark pcr duplicates

The marking of duplicate reads within the Bam file is implemented through Picard's *MarkDuplicates* function. The *picard.MarkDuplicates* wrapper asks in input the Bam file to be analyzed and the optional arguments to run the tool. The function works by comparing sequences in the 5-prime positions of both reads and read-pairs. The tool output is a new Bam file, in which duplicates have been identified (not deleted) using SAM flags field for each read, and a metrics file indicating the numbers of duplicates reads. The wrapper assumes the Bam is sorted using chromosome coordinates (*ASSUME\_SORT\_ORDER = coordinates*).



### 3.3.3.3. Realignment around InDels

The tool used for realigning the reads around the InDels is GATK v3. This step is no longer implemented in GATK v4 as it is replaced by the local realignment of sequences directly in the variant calling step.

The wrapper *GATK.IndelRealigner* requests as input the Bam to be analyzed, the BED or LIST file that contains the coordinates of the sequencing target, and the Fasta file of the reference genome. To improve accuracy, the database containing known InDel sites (e.g., *mills\_and\_1000G\_gold\_standard.indels.hg19.sites*) can also be provided. The realignment process implemented in GATK consists of two steps: 1. the sites where it is probably necessary to realign through the *RealignerTargetCreator* function are identified, and 2. the candidate sites are realigned through the *IndelRealignerfunction*. The output of the wrapper is a Bam file in which potentially problematic sites due to the presence of an InDel have been realigned.

### 3.3.3.4. Base Quality Score Recalibration

The recalibration of the Base quality scores step is implemented using GATK v3 or v4. The recalibration process consists of two phases: 1. First, GATK calculates for each mismatch found in the Bam file a series of statistics and covariates and generates a recalibration table file; 2. Then, GATK uses these tables to calculate the new quality score for the bases contained in the Bam file.

The wrapper *GATK.BaseRecalibrator* requires reads data in Bam format whose base quality scores need to be assessed, one or more databases of known polymorphic sites that can be useful to improve the process quality (e.g., *mills\_and\_1000G\_gold\_standard.indels.hg19.vcf* or *dbsnp.vcf*), the target file in BED or LIST format, and the reference Fasta file. The first step of generating the recalibration table is implemented through the *BaseRecalibrator* function for both versions of GATK, while the second step is performed by the *PrintReads* function for GATK v3 or the *ApplyBQSR* function for GATK v4. The output of the *GATK.BaseRecalibrator* is a Bam file with recalibrated quality scores.

## 3.3.4. Short variant calling

The variant calling step is performed differently based on the type of samples (germline or somatic) and their organization in the samplesheet (only case, case-control, trio). In the case of germline analyses, variant calling can be performed in single-sample, in cohort and in trio modalities, while in somatic analysis the samples are organized in case-control modality (figure).

In order to perform variant calling, Helper implements 3 tools: GATK (HaplotypeCaller + GenotypeGVCF) [84], Freebayes [85], and VarScan2 [86]. For somatic variant calling, GATK (Mutect2), VarScan2, and Vardict [87] are implemented.

HaplotypeCaller and Freebayes are two variant callers based on a local de-novo assembly of the suspicious regions and implement a method of detection of the probable haplotypes, present in the target analyzed, a priori from the alignment information contained in the Bam files. Both algorithms identify regions that show sufficient evidence to hypothesize the presence of a variant and construct a window of interest (“ActiveRegion” for GATK) around the candidate region. All possible haplotypes observed within the window of interest are calculated and the one with the greatest likelihood of actually being present in the sample is considered. GATK calculates the likelihood after a local realignment step of the haplotypes, while Freebayes performs a count of the frequencies of the observed haplotypes. Using the information on the haplotypes, the probability of the genotype for each potential variant site is inferred using Bayesian methods. Finally, the most likely genotype is assigned to each genomic position within the haplotype considered (an example of haplotype-based workflow is shown in figure). In both HaplotypeCaller and Freebayes, the genotype Quality score associated with the variant is provided as the difference between the likelihood of the chosen genotype and the second most probable.

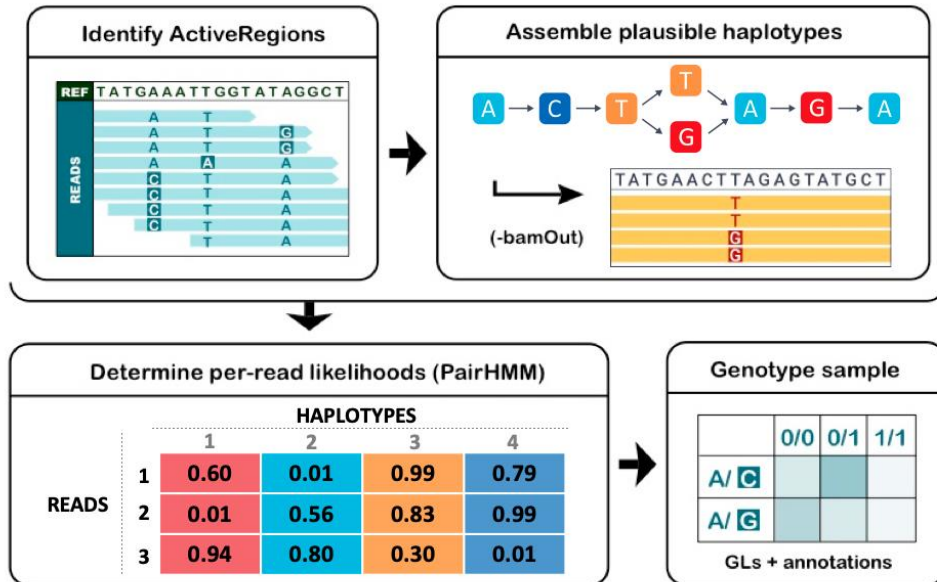


Figure 3.3: The HaplotypeCaller genotyping workflow (Figure from [84])

Mutect2 is the variant caller implemented in GATK for the analysis of somatic samples. The workflow for identifying variants is similar to those of

HaplotypeCaller. Both search for the active regions, perform the assembly de novo to reconstruct the haplotypes, calculate the probability associated with each identified haplotype, and estimate the most likely genotype. The main difference is represented by the model used to calculate the likelihood of the genotype. HaplotypeCaller relies on a fixed ploidy assumption to calculate the genotype likelihood, Mutect2 instead does not use a fixed ploidy model in order to ensure greater accuracy in calling variants with a lower allelic frequency. This allows Mutect2 to gain greater flexibility in the evaluation of samples with problems of fractional purity, sub-clonality, and copy number variations common in cancer sequencing applications.

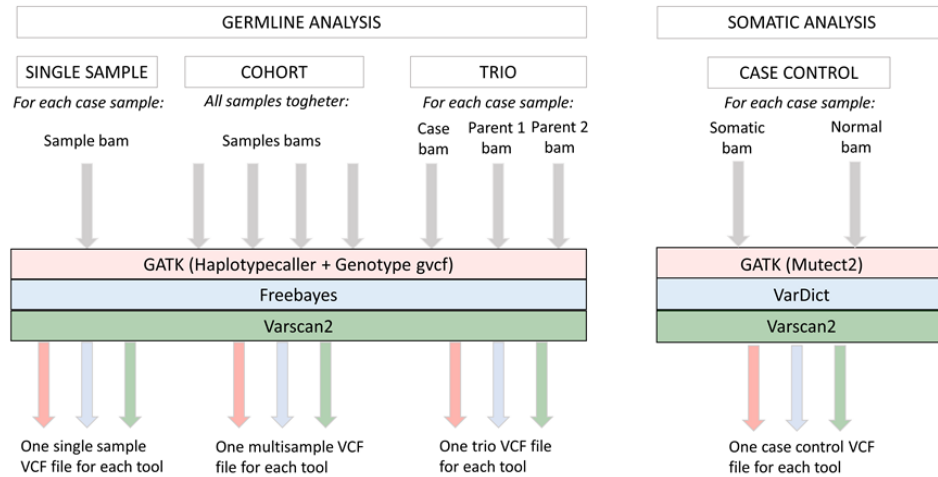
On the other end, VarScan is a tool that implement a robust heuristic approach to call variants. Unlike the Bayesian tools, it does not calculate the probability of the genotype based on the observations on the sample but evaluates the mismatches with the reference genome on a base-by-base basis through a threshold system. For each possible variant site, apply a cascade of quality filters to evaluate the parameters supporting each observed allele. The bases that mismatch and that exceed the coverage and base quality thresholds on the reads with non-null mapping score are examined, the others are recognized as non-variant positions. The alleles identified in the position under examination are tested based on the number of supporting reads, allele frequency, strand balance, and the p-value derived from a Fisher's Exact Test on the observations on the variant and on the reference. The genotype of the evaluated site is determined based on the frequency of the variant allele. If the allelic frequency exceeds a certain threshold (0.80 by default) then the genotype is Homozygous, otherwise heterozygous. Although GATK HaplotypeCaller and FreeBayes are two tools that generally work well, Varscan guarantees a different “point of view” useful for increasing the number of detected variants in the case of joint multi-tool variant calling.

VarDict is a variant caller developed for cancer sequencing applications and allows to perform analysis of paired samples (tumor and matched normal samples) to detect germline, somatic and loss of heterozygosity (LOH) variants. Similar with Varscan, VarDict uses a heuristic approach for the identification of variants and implements an algorithm specially developed for the detection of InDels hard to evaluate and to estimate their allelic frequency with greater accuracy. Taking advantage of the fact that InDels often cause misalignments and clipped reads, VarDict performs two types of local realignments based on the size of InDels:

1. For small InDels, a supervised method is used that realigns clipped reads in the around variants already identified, improving the estimate of the allele frequency.

2. To search for new larger InDels, the near clipped reads areas are monitored, a consensus sequence is generated which VarDict realigns (unsupervised realignment) within a window of variable size based on the length of the InDels to be identified. Based on the result of the alignment of the consensus, deletions, insertions or complex variants are called even if they are larger than the length of the reads.

When evaluating paired samples, VarDict performs for each identified variant, a fisher's exact test to determine if the difference of allele frequency between case and control samples is significant. Variants present only in the case sample are called somatic, variants present in both samples are called germline, and variants in heterozygous state in control sample that become homozygous in the case sample are called LOH.



**Figure 3.4:** Variant calling process based on the sample organization

### 3.3.5. Post processing

#### 3.3.5.1. VCF normalization

During variant calling, the same indel can often be reported multiple times and with different starting positions in the VCF file. The standard convention with VCF is to place an indel at the left-most position to define a unique record. The VCF normalization step is necessary to ensure that the InDels described in the VCFs are reported in the left-most standard. In Helper this step is performed by GATK v4 or Bcftools.

The two wrappers *gatk.LeftAlignAndTrimVariants* and *bcftools.norm* require input the VCF file to normalize and the reference genome. Both wrappers implement by default the splitting of multiallelic sites (sites where are reported more than one alternate alleles) in biallelic sites. The normalized VCF is returned as output.

#### 3.3.5.2. VCF filtering

VCF file filtering is implemented through the *VariantFiltration* tool of GATK v3 and GATK v4. VariantFiltration allows to perform a Hard filtering of the variants using a threshold system on the information contained in the

VCF file. It was decided not to implement filtering methods based on the machine learning approach (such as GATK's *VQSR*) in the v1.0 version of Helper as it is a framework dedicated to the analysis of target sequencing data. In fact, The ML filtering algorithms are poorly performing for this type of applications.

The wrapper for the filtering function (*gatk.VariantFiltration*) asks for the VCF file to be filtered, the reference genome, and filters that can be provided as an argument to the tool at the pipeline design time. In output the function produces a VCF file that contains the FILTER field modified with the respective filtering tags if the variants that do not exceed the thresholds, and with PASS in the others.

### 3.3.5.3. VCF split by samples

This step is needed only in the case in which the germinal variant calling is performed in Cohort mode. The joint VCF that contains all the samples of the cohort is split into several VCFs containing a single sample in order to be analyzed individually.

The *scripts.Filter\_by\_sample* wrapper calls a script developed in-house (*vcf\_split\_by\_sample.py*) which requests the joint VCF file and the sample name used for information extraction and with which to rename the filtered VCF. The script filters the VCF joint using the sample name, extrapolates all the variants identified in the sample excluding sites with wild type (0/0) or unknown (./.) genotype. In output, a VCF file is obtained, in which the fields of the chromosomal position (CHROM and POS), of the alleles (ID, REF, and ALT), of the FILTER, and of the INFO remain unchanged with respect to the joint VCF starting file and the field FORMAT that reports only the information of the sample of interest.

### 3.3.5.4. VCF merge

The VCF merge step generates a single VCF using calls from multiple variant callers. The *scripts.merge\_vcfs* wrapper calls an in-house script (*merge\_vcfs.py*) which requests as input the VCFs issued by the individual variant callers: GATK, Freebayes, and Varscan for germline analysis, and GATK, Varscan, and Vardict for somatic variant calling. For each variant it extracts and processes the information contained in the different VCFs and outputs a new merged VCF file. The new VCF contains the set of variants identified by at least one of the variant callers. The new FORMAT field is the result obtained by averaging the values of the FORMAT fields of the three software, and the most represented genotype across the three tools is chosen. The new INFO field shows the INFO fields of the other VCFs with a prefix indicating the source software (for example GATK\_AC). The INFO field also reports the FORMAT fields of each VCF in order to track original

values (for example GATK\_FORMAT). The new file respects the VCFv4.2 format and is compatible with the most software that analyze VCF files.

### 3.3.5.5. VCF to TSV conversion

This pipeline step converts the variant format from VCF to TSV. The TSV format is easier to read and can be parsed like an Excel or Calc worksheet. Also, in this case the *vcf\_to\_tsv.py* script that deals with the conversion of the format is a script developed in-house and requires the VCF file to be converted, the FORMAT and INFO fields to be reported in the TSV file, and the name of the output file. The FORMAT and INFO fields to be reported in the TSV file, must be indicated as a comma-separated list under the “--format” and “--info” parameters. Alternatively, it is possible to provide a file (“--tag\_file”) containing the list of FORMAT and INFO fields of interest. Within the tag\_file, the fields must be indicated as a list of elements (one entry per line) consisting of FORMAT or INFO and the name of the field of interest separated by TAB. The nested INFO fields, such as the formats of the individual Variant callers in the case of VCF merged (e.g., GATK\_FORMAT), must be reported indicating the name of the nested field like GATK\_FORMAT and the name of the field of interest contained in the nested one as GT, separated by ":".

Examples of entries in the tag\_file:

```
FORMAT GT
FORMAT AD
INFO AC
INFO GATK_FORMAT:GT //nested field
INFO FREEB_FORMAT:DP //nested field
```

The output TSV file contains a Header that includes the mandatory descriptive fields about the chromosomal position of the variant, the alleles, the filters ('CHROM', 'POS', 'ID', 'REF', 'ALT', 'FILTER'), and the whole list of fields extrapolated from the separate FORMAT and INFO TAB. The information corresponding to the fields of the Header, separated by TAB, are shown for each variant (one per line).

The wrapper *scripts.Vcf\_to\_tsv* asks as input the path to the script, the VCF file to convert, the tag\_file, the list of fields of the FORMAT and the list of fields of the INFO report in the TSV file.

### 3.3.6. Variant annotation

The variant annotation step is implemented in Helper using Variant Effect Predictor (VEP) [88] and Annotate Variant (ANNOVAR) [89]. Both tools annotate variants locally by extracting information from precompiled databases. Both VEP and ANNOVAR are two tools widely used for their

ease of use and completeness of annotation. ANNOVAR performs three levels of annotation using three different scripts or using a single line command: 1. Gene-based annotation to identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected; 2. Region-based annotation to identify variants in particular genomic regions, for example, conserved regions, predicted transcription factor binding sites, segmental duplication regions, or many other annotations on genomic intervals. 3. Filter-based annotation for information about the presence of the variant in the various databases (dbSNP, population database, etc.), or to extract the scores from the damage prediction tools. VEP, on the other hand, generate the entire set of annotations with a single line of code, and allows information on the gene, the region and the various clinical and population databases to be integrated with other custom databases external to the pre-compiled one, increasing the quantity of potentially obtainable information.

The wrapper for VEP (*vep.vcf\_annotation*) requests the VCF file to be annotated, the reference genome, the assembly with which you want to annotate the variants (the precompiled VEP database), the species (the default specie in Helper is “homo\_sapiens”), the fields with which you want to note the variants, the additional plugins (optional). The additional plugins must be present locally and must be indicated in the tools configuration file as a database (for example dbSNFP).

The ANNOVAR wrapper (*annovar.vcf\_annotation*) requests in input the VCF file to be annotated, the reference genome, the path to the precompiled database, and the protocol with which to annotate the variants. In Helper, the annotation using the single command line is implemented for ANNOVAR.

The output of the module is an annotated VCF file.

### **3.3.7. Post annotation**

#### **3.3.7.1. Report annotation in TSV format**

This step reports in TSV format the variants contained in the annotated VCF. The `scripts.add_Annotation` wrapper calls an in-house script (`annotation_extractor.py`) which requests the annotated VCF, the file containing the list of annotations to be extracted, the TSV format file generated in the `vcf_to_tsv` step of the post-processing module (optional), the file containing the list of main transcripts from which to extract the annotation information (optional), as well as the log file and the working directory. The script first filters the transcripts for each variant, considering only those provided in input, or alternatively the canonical ones; then look for the annotation tags provided in input with the annotation list file; finally, if the TSV file produced by post-processing is supplied to him, the annotations extracted are added directly to this file, otherwise the variants are reported in new file in TSV format.

### 3.3.8. CNV calling

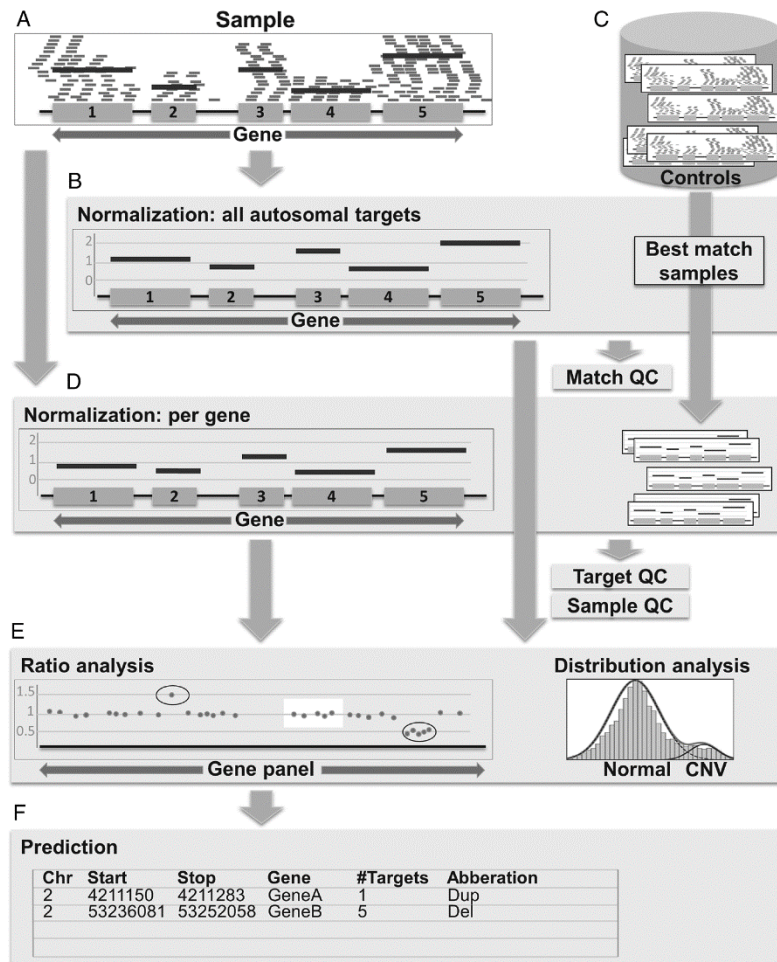
In Helper the CNV call module is implemented using GATK, Decon (90), CoNVaDING [91], and CNVkit [92].

GATK calls CNVs using a Read Depth (RD) based method for both WGS and targeted sequencing applications. The GermlineCNVCaller algorithm uses a Bayesian approach to calculate the likelihood of the ploidy of the regions of interest and call the CNVs. The algorithm generates a coverage model by calculating all the descriptive parameters of the distribution of read depth, variance and bias within the target through the comparative analysis of a training dataset that contains a series of similar samples (same sequencing platform, same library preparation protocol, and same capture kit). From the training dataset, GermlineCNVCaller also infers the ploidy status of the target contigs and uses them as the baseline copy number state for a Hidden Markov Model (HMM). The HMM algorithm uses information on the region of interest and the parameters of the coverage model to calculate the probability that a change in copy number status may have occurred in one or more adjacent regions of the target. GermlineCNVCaller can run in Cohort mode or Case mode: in cohort mode the coverage model and its parameters are calculated directly from the samples that are part of the cohort to be analyzed and based on these parameters it calls the CNVs; in Case mode the model is built on a cohort different from the samples to be analyzed but which must be compatible in terms of library preparation and sequencing platform. The cohort mode needs a large number of samples to be analyzed in parallel in order to work at its best (recommended 30 WES or WGS samples), while the Case mode allows you to analyze even a single sample at a time as long as you have a model trained with a sufficient number of compatible samples. The result of the CNV call is linked to the quality of the coverage model both in the number of samples that compose the training cohort and in the estimation of the hyperparameters that can be modified by the user and must be evaluated on a case-by-case basis. The GermlineCNVCaller tool has computational resource requirements to consider that scale linearly with the number of analyzed samples and the complexity of the trained model.

CoNVaDING is a CNV caller that implements an RD-based algorithm based on comparative analysis with a group of control samples. The CoNVaDING workflow consists of several steps, starting from the selection of the best control group, composed of samples generated with the same library preparation protocol and sequenced with the same platform. The tool performs two coverage normalizations for each region contained in the target: a normalization on the whole sample using the average coverage of the entire target, and a normalization on each gene, comparing the coverage of the single exons with the average coverage of the entire gene. The most informative samples are chosen, based on the similarity in terms of coverage with the sample under examination, to be used as a reference set for calling the CNVs. The CNV call is made for each target region by comparing the normalized coverage of the sample under examination and the average



coverage of the reference set and calculating the Z-score between the normalized coverage distribution (first on the whole sample and then on the specific gene) in the sample under examination and the distribution of the control group. The CNVs are called by combining information on the coverage ratio and distributions in a different way based on the magnitude of the event identified. CoNVaDING filters the called CNVs by dividing them into 3 sets of different sensitivity and specificity based on the quality control results on the samples.



**Figure 3.5:** The CoNVaDING workflow (Figure from [91])

Decon is an ExomeDepth based tool [93], optimized for target sequencing applications that implement an RD type approach. CNVkit calculates a coverage metric called the fragment per kilobase and million base pairs (FPKM) for each exon in the target. The FPKM normalizes the number of reads that map the analyzed region based on the length of the exon and the total number of samples reads. CNVkit works in batch mode and requires a minimum of input samples to ensure call quality (the number of samples depend on the experiment). The call of the CNVs is made through an HMM which considers the FPKM, the quality of the analyzed region, the quality of

the entire sample in terms of coverage and correlation with the other samples of the incoming cohort.

CNVkit uses both the on-target reads and the nonspecifically captured off-target reads to identify CNVs for each sample. Both the on- and off-target locations are separately used to calculate the mean read depth within each interval. In fact, for each Bam file, CNVkit computes the log<sub>2</sub> mean read depth in each on and off- target bin. CNVkit uses a copy number reference in order to correct the results of test samples. The reference profile is estimates using samples derived from same NGS protocol and analyzed using same sequencer. The number of reference samples depends on the applications; it is possible to generate a reference using just a sample, but more samples are recommended. Additional information can be associated with each bin in order to perform GC bias, and repetition bias correction. The CNV calling is performed after a read depth fixing step. The single sample's on- and off-target read data are combined, then CNVkit removes bins that fail quality check, performs the correction of systematic biases, subtracts the reference read depth from each bin, and finally median-centers the corrected copy ratios. The sample's copy ratios are segmented into discrete copy-number regions and the report containing CNV calls is emitted. The segmentation step can be performed using a set of algorithms (CBS, HaarSeg, HMM) in order to adapt the analysis to different applications.

The CNV callers implemented in Helper can work both in single sample mode and in batch mode. Whereas the single sample mode uses samples from other experiments as a reference in order to compare the analyzed sample and identify variants, the batch mode uses samples in the same cohort as reference. In Helper, the choice between the two calling modalities depends on the Experiment configuration file. If the fields that concern CNV tools reference files and directories (for example GATK ploidy and call models, or the control samples directory for CoNVaDING) are empty, then Helper performs CNV calling in batch mode, otherwise uses the single sample mode.

### **3.3.9. Sample organization and workflows**

The sample organization is important for the workflow setting (Figure 3.6). The pre-alignment, the alignment, and the pre-processing modules are performed ever in the same way, based on the workflow set in the pipeline configuration file. Files from all samples are analyzed step by step independently from the sample organization. On the other hand, variant calling strongly depends on the sample's organization: “single-sample” or “cohort” modality of germline variant calling can be performed with “only-case” sample organization, while trio germline analysis and case-control somatic variant calling are performed in case of “trio” and “case-control” organization, respectively.

In case of single sample variant calling using multiple tools, VCF files of the same sample are merged in a single VCF file; otherwise, the merging

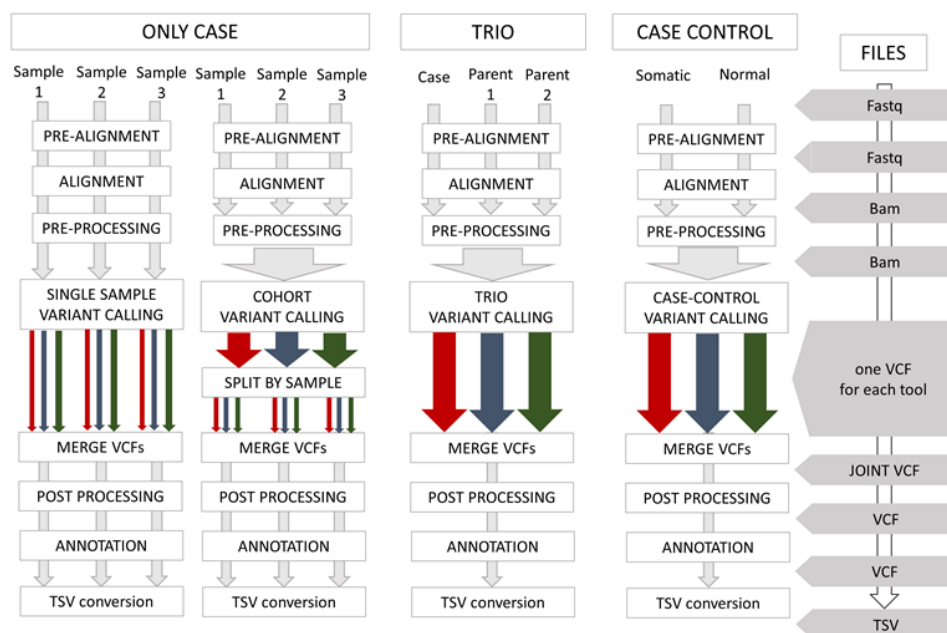
step is bypassed. This file is optionally processed, filtered, annotated, and converted in TSV format. The final TSV file contains information about variants in a single sample.

In case of cohort variant calling, a multi-sample VCF is produced for each used tool. Joint VCF files are split in single sample VCFs; thereafter, the workflow proceed per single-sample analysis and therefore, each final TSV contains the variants of a single sample, to which useful information about allele count, allele number, and allele fraction in the cohort are added.

In case of trio variant calling, the split step is bypassed. VCF files from the tools are merged in a single VCF file, and then analyzed as a single sample VCF. The final TSV file contains variants from the case sample and from both parent samples. Information about the genotype and quality scores are reported for the three samples, in order to understand which samples are carriers of the variants and to reconstruct the variant segregation.

The case of case-control variant calling, is similar to the trio one. The splitting step is bypassed, VCF files from different tools are merged in a single VCF and analyzed as a single-sample VCF, and the final TSV files contains variants from both the somatic and the control samples. In this case is important to understand which variants are present only in the somatic sample and which are also present in the control sample. Information about genotype and quality scores are reported for both samples to facilitate the identification of true somatic variants from artifacts.

In this version of Helper, the CNV calling module is performed ever as single-sample or cohort (batch) modality. In case of trios or case-control samples organization, CNV module consider all samples as only-case mode. For each tool a file report is generated, and the comparison between child and parents, or tumor vs normal sample have to be performed manually. In the next version, this comparison step will be implemented.



**Figure 3.6:** Workflows based on the sample's organization

**Table 3.3:** The table reports tools implemented in Helper, the version, the module, and the steps where can be used.

<b>Tool name</b>	<b>Version</b>	<b>Module</b>	<b>Step</b>
AGeNT	v.3.5.1.46	pre-alignment	Trimming of adapters
CUTADAPT	v.1.13	pre-alignment	Trimming of adapters
		pre-alignment	Fastq filtering
FastQC	v.0.11.8	pre-alignment	Fastq QC
		pre-processing	Bam QC
BWA	v.0.7.17	alignment	Fastq alignment
BOWTIE2	v.2.3.5.1	alignment	Fastq alignment
PICARD	v2.7.1	alignment	Sam to Bam conversion
		alignment	Bam sorting
		alignment	Bam indexing
		pre-processing	Add or replace read group
		pre-processing	Duplicates marking
GATK v.3	v.3.7	pre-processing	Indel realignment
		pre-processing	Base quality score recalibration
		variant calling	short variant calling
		post-processing	VCF filtration
GATK v.4	v.4.1	pre-processing	Base quality score recalibration
		variant calling	short variant calling
		variant calling	CNV calling
		post-processing	VCF filtration
		post-processing	VCF normalization
Freebayes	v.1.1	variant calling	short variant calling
Varscan2	v.2.3.9	variant calling	short variant calling
VarDict-Java	-	variant calling	short variant calling
Samtools	v.1.3.1	variant calling	short variant calling
Bcftools	v.1.5	post-processing	VCF normalization
VEP	-	annotation	VCF annotation
Annovar	-	annotation	VCF annotation
Decon	v.1.0.2	variant calling	CNV calling
CoNVaDING	v.2.3.2	variant calling	CNV calling
CNVkit	-	variant calling	CNV calling

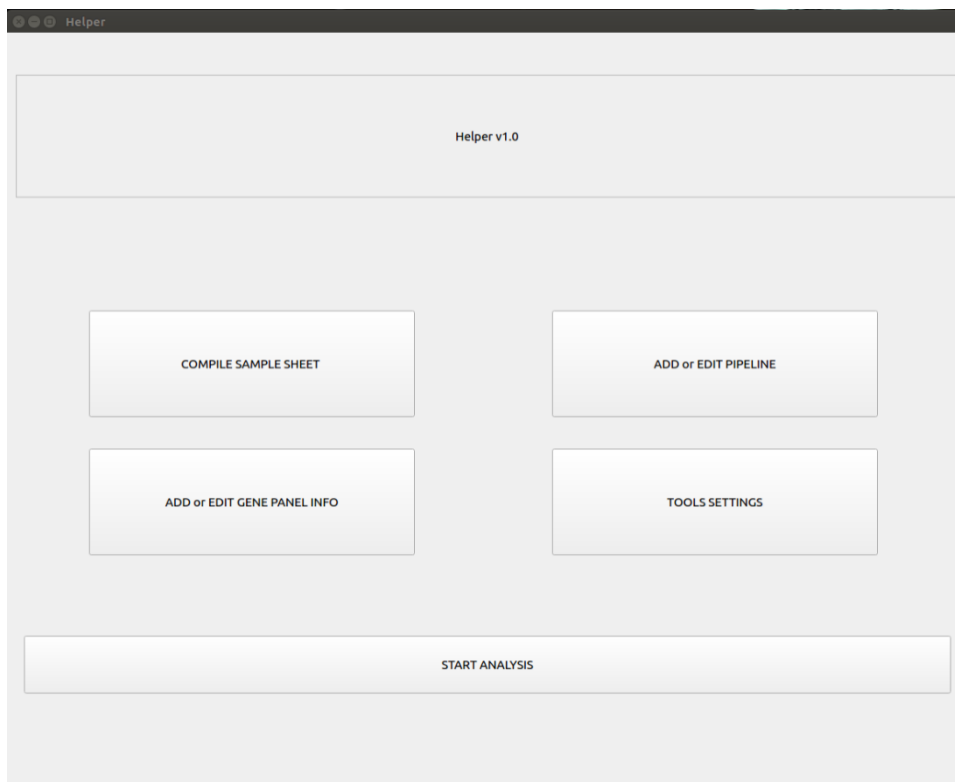
### 3.4. The Helper GUI

To facilitate the user experience in the setup of the configuration files In order to facilitate the user experience in the compiling of the configuration files, in the preparation of the samplesheet file, and in the setup of the analysis of the samples, a simple graphical user interface (GUI) is implemented in Helper. The GUI was developed in Python 3 language using PyQt5 python module. The GUI can be called using simple command line:

```
Python /paht/to/Helper.py
```

The first window that appears is the main Helper window and contains five buttons, each of which opens another window:

1. The button labeled "compile shamplesheet" opens the Samplesheet designer window;
2. The button labeled "Add or Edit gene panel info" opens the Experiment designer window;
3. The button labeled "Add or edit pipeline" opens the Pipeline designer window;
4. The button labeled "Instrument settings" opens the Samplesheet designer window;
5. The button labeled "Start analysis" opens the Analysis settings window.



**Figure 3.7:** The Helper's main window

### 3.4.1. Tools setting

The tool setting window allows the compilation of the Tools config file. This window automatically loads the Tools configuration file into the */pathtoHelper/config/tools* folder. Any changes made to the information and settings of the tools are automatically saved in that configuration file.

The main window of the Tool settings contains the list of tools implemented in Helper, the list of databases that can be used in the various steps of the pipeline by the tools, and the list of reference files. When you select a tool, a database, or a reference genome, the information present in the configuration file is displayed in the "Settings" table.

The "Add" button opens a window (Add tool info window, for the button dedicated to tools) in which you can indicate: the tool name field that identifies the tool within Helper; the tool version for tracing tools with similar names (e.g., GATK v3 and GATK v4); the path of the main script in case of tools developed in Python, R, Perl, or Bash, and of the jar file in case of Java tools; the tags that indicate in which steps the tool can be used. The tags can be entered manually, or through the "Add\_tags" window which can be accessed via the appropriate button. The Add Tags window contains the list of possible pipeline steps. By selecting the steps and confirming, the tags will be automatically added to the tool. Using the save button, information contained in the Add tools window is saved in the tool's configuration file and the new tool is added to the list in the main window.

The "Set" button opens a window that contains the same fields as the "Add tool" window. In this case the fields are pre-filled with the information of the selected tool extracted from the configuration file. Clicking the save button, the information in the configuration file is overwritten by the modified one.

Using the Delete button, the selected tool is deleted from the list of tools and from the configuration file.

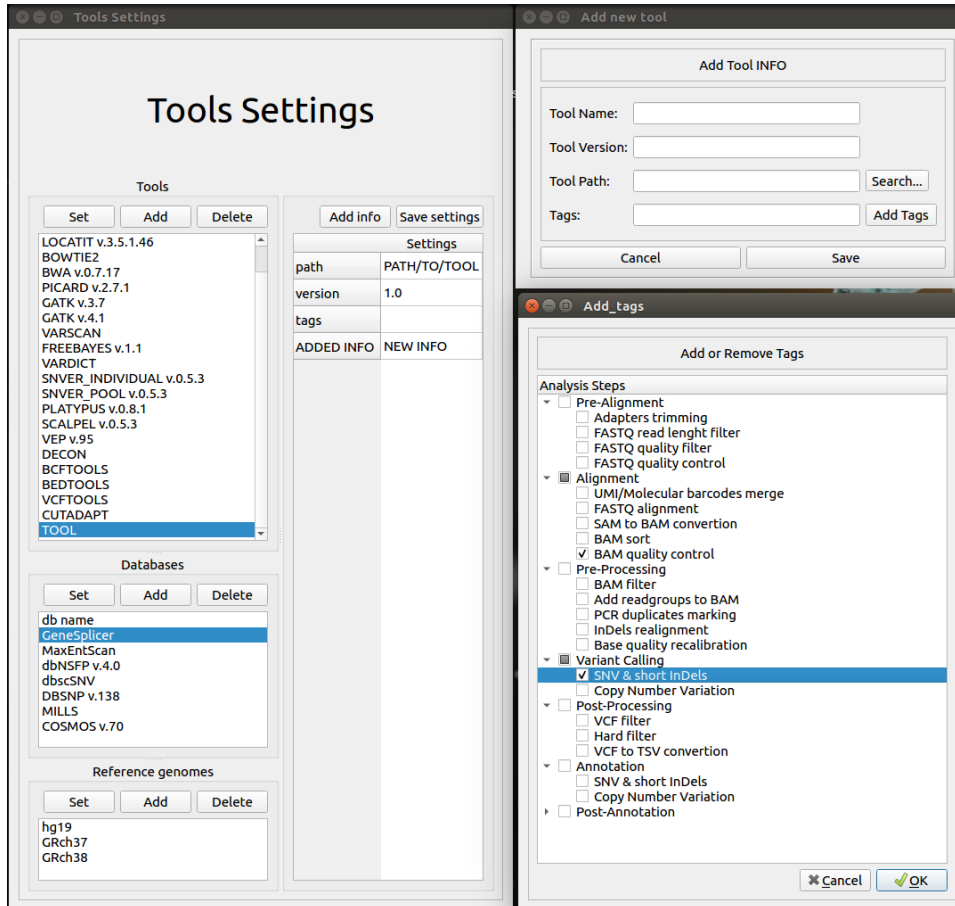
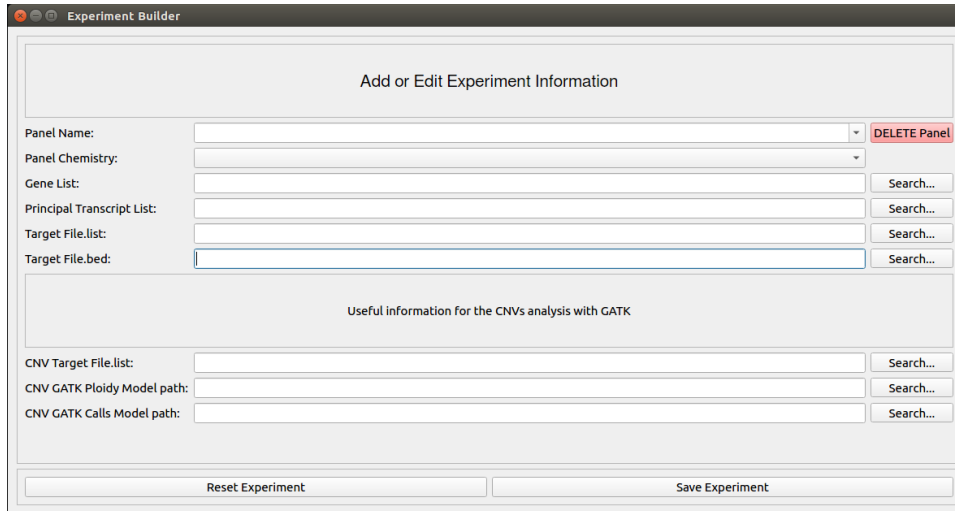


Figure 3.8: The Helper's Tool Settings window

### 3.4.2. Experiment Designer

In the experiment designer window, all the information contained in the Experiment configuration file must be specified. Using the “Panel Name” drop-down menu it is possible to choose an existing experiment to modify it, or to create a new one by simply entering a new experiment ID. From the drop-down menu “Panel chemistry” it is possible to select the type of sample preparation: “Capture Enrichment” or “Amplicon”. In the "Gene List" field it is possible (optional) to indicate the list of genes contained in the panel used for the experiment; in the “Principal Transcript List” field it is possible (optional) to indicate the list of transcripts necessary to filter the annotations of the variants in the post-annotation form; In the "Target file" fields it is necessary to indicate the file containing the target in LIST format and in BED format. Finally, it is necessary to indicate the directories and files essential to the tools to make the CNV call. Each tool needs specific files in order to perform the analysis. The example in the figure shows the fields dedicated to GATK: The target file in LIST format dedicated to the CNV call, the Ploidy model used by the algorithm for calculating the likelihood of the genotype, and the Call model needed with the 'Single sample' CNV calling

mode. The "Reset Experiment" button deletes all the fields of the selected experiment, while the "Save Experiment" button allows you to save the changes made or to save the new experiment in the Experiment configuration file.



The screenshot shows the 'Experiment Builder' window with the title 'Add or Edit Experiment Information'. It contains several input fields and search buttons:

- Panel Name: [text input] [DELETED Panel]
- Panel Chemistry: [text input]
- Gene List: [text input] [Search...]
- Principal Transcript List: [text input] [Search...]
- Target File.list: [text input] [Search...]
- Target File.bed: [text input] [Search...]

Below these fields is a section titled 'Useful information for the CNVs analysis with GATK' containing:

- CNV Target File.list: [text input] [Search...]
- CNV GATK Ploidy Model path: [text input] [Search...]
- CNV GATK Calls Model path: [text input] [Search...]

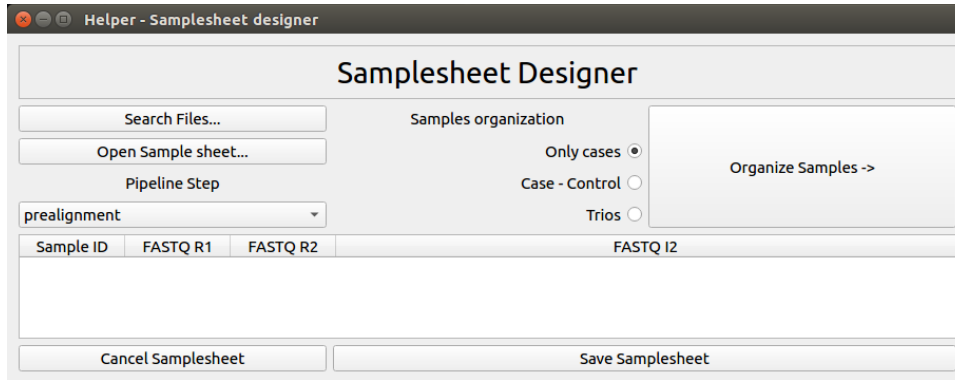
At the bottom of the window are two buttons: 'Reset Experiment' and 'Save Experiment'.

**Figure 3.8:** The Helper's Experiment Designer window

### 3.4.3. Samplesheet Designer

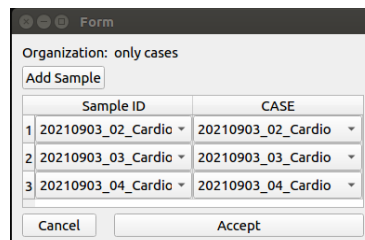
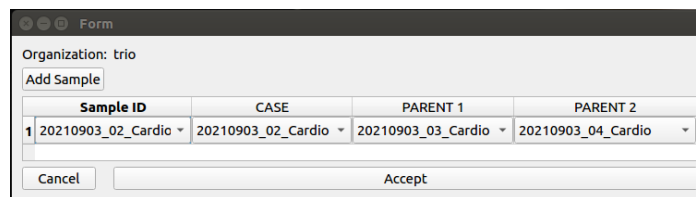
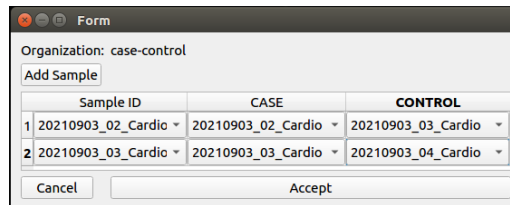
The samplesheet designer allows the organization of sample files in such a way that they can be analyzed by Helper. The main window has two search buttons. The "search files" button allows you to search for files one by one to add them to the list and create a new samplesheet; if multiple files are selected at the same time, they will be added together. The drop-down menu of the Pipeline step determines in which module the files will be saved. For example, the selection of prealignment, allows the selection of Fastq files only (which are files compatible with the prealignment module); they will be organized in such a way as to distinguish Fastq R1 from R2 (and Fastq I2 if needed). The sample ID will be inferred directly from the files but can be changed later. It is possible to add one sample at a time by right clicking on the table and selecting add sample; in that case the Fastq files will be added individually by double clicking on the specific box. From the drop-down menu it is possible to choose the four starting modules of the pipeline: prealignment, alignment, preprocessing, variant calling (short variants and CNV). For prealignment and alignment only Fastq files can be selected, while for preprocessing and variant calling only Bam files can be selected. The second button (open samplesheet) allows you to open and edit an existing samplesheet. In this case, the drop-down menu allows you to switch between the different modules within the samplesheet.





**Figure 3.9:** The Helper’s Samplesheet Designer window

The "Sample organization" radio button is used to indicate the organization of the samples in the pipeline. Depending on the organization, once the setting window is opened using the "Organize samples" button, you can indicate the role of each sample. If the organization is "only case", each sample will be considered independently; if it is case control, you must indicate which sample is the case and which is the control; in the case of "trio" it is necessary to indicate who is the case and who are the relatives. In order to save the pipeline, it is necessary to perform this sample organization step.



**Figure 3.10:** The “Organize Samples” window

### 3.4.4. Pipeline designer

The Pipeline Designer window allows the design of new pipelines or the modifications of existing ones. Through the drop-down menu "Pipeline" the pipeline can be chosen. Through the "Tools cfg" field it is possible to choose the tool configuration file from which the tools, that can be used in each step of the pipeline, are extracted. The two drop-down menus "Analysis" and "Reference version" indicate the analysis type (somatic or germline) and the version of the reference genome to use, correspondingly. The central core of the interface is the tree of the analysis steps. The user decides which steps to enable and disable in the pipeline: if a step belonging to a module is enabled, the module is also automatically enabled; if all the steps of a module are disabled, the module itself is also disabled; if you enable or disable the module directly, all the steps of the module are enabled or disabled. When a step is selected, the tools provided and the specific settings for the selected step are displayed in the next window. Through the buttons "Use / add this tool" you choose which tool to use to perform the step, in case of multi-tool step the selected tool is added to the list. Using the button "Don't use this tool" you remove the tool from those provided in the step. When a tool is chosen to perform the selected step, it is also indicated in the "Analysis step" tree, under the "Tools" field. The settings compiled in the Step settings table are used by all the tools included in the list, while the Tool settings table contains those relating to the single selected tool. In this table you can enter all the input arguments and parameters of the selected tool, to further customize the analysis. Finally, the "Delete Pipeline" button deletes the selected pipeline, the "Cancel" button deletes all unsaved changes, and the "Save" button saves the pipeline as a Json format.

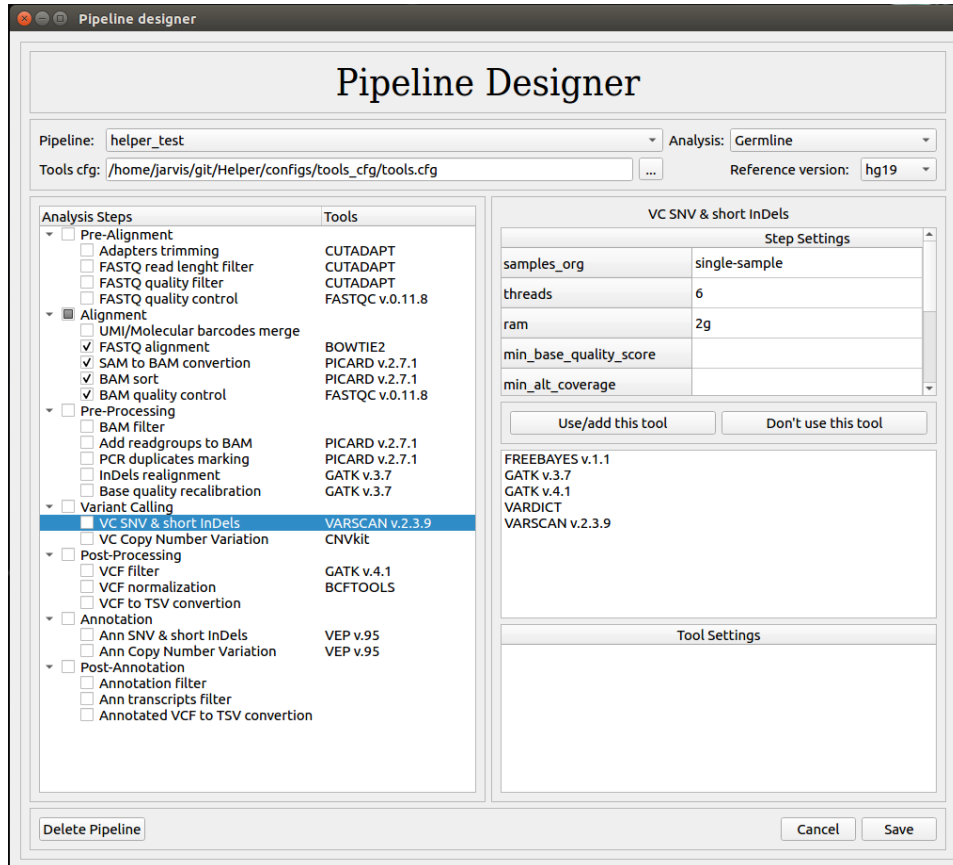
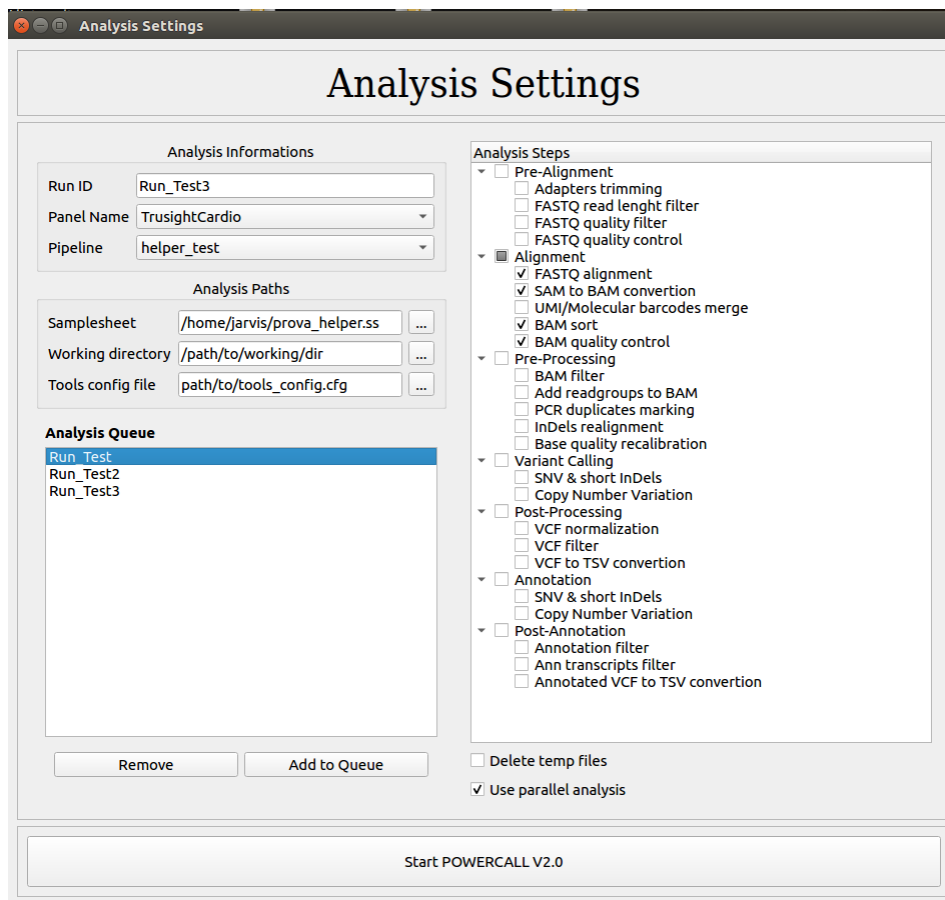


Figure 3.11: The Helper's Pipeline Designer window

### 3.4.5. Analysis Settings

Finally, the Analysis settings window allows you to execute the Helper main script (pipeline.py) and to launch the analysis. In order to start the pipeline, it is necessary to specify the ID of the single run, the experiment (Panel name) from which the samples are derived, and the pipeline configuration file. The samplesheet file, the working directory and the tools configuration file are also required. After the pipeline configuration file is selected, the steps provided by the chosen pipeline appear in the "Analysis steps" tree. By deselecting the modules within the Analysis steps tree, you choose which steps to process in the specific workflow. For example, if my pipeline includes the complete workflow but there is a need to start the analysis from the alignment, the pre-alignment module must be deselected; or, if for the specific run it is not necessary to call the CNVs, deselect Copy number variation in the Variant calling module. This strategy allows you to restart the analysis from the module in which it stopped, or to save time if you do not need to run a particular module. The two flags "Delete temp files" and "Use parallel analysis" activate the elimination of the temporary files of the modules and launch the pipeline by activating the parallel processing of the samples, respectively. The temporary files are all those files produced by

the intermediate steps within the modules and which do not need to be subtracted in the long term. Deleting these files saves a large amount of memory space. Parallel analysis allows to reduce sample processing times but requires adequate hardware resources. Once the configuration files have been selected, the workflow selected and the paralleling and deletion parameters of the temp files have been set, the analysis can be added to the queue. You can prepare multiple runs and add them to the queue before starting sample processing; in the event of multiple analyzes in the queue, these will be processed in series, in the order in which they were added. The start Analysis button calls pipeline.py and provides it with all the parameters necessary to perform the analysis.



**Figure 3.12:** The Helper’s Analysis Settings window

### 3.5. Workflow performance study

Understanding whether it is possible to use Helper within a clinical and research context is a necessary step to test the potential of the software. For this reason, an ad hoc pipeline was developed for the needs of the genetics laboratory of the CMGCV. The CMGCV mainly uses two gene panels for diagnostic and research routines: the Illumina Trusight Cardio kit for the

study of hereditary cardiomyopathies and aneurysmal connective tissue diseases, and the Illumina Trusight Cancer kit for the study of hereditary tumor pathologies.

### **3.5.1. Trusight Cardio and Trusight Cancer Panels**

The TruSight Cardio (TSCardio) is a gene panel that provides comprehensive coverage of 174 genes with known associations to 17 inherited cardio-vascular conditions, including cardiomyopathies, arrhythmias, aortopathies, and more. Genes were expertly selected with researchers at the National Heart Center Singapore and Imperial College of London. The TruSight Cancer (TSCancer) is a gene panel developed in collaboration with cancer genomics experts, that includes 94 genes and 284 single nucleotide polymorphisms (SNPs) associated with a predisposition towards cancer. The TSCardio target is 571,897 bp long and allows to sequence 12 samples per run with Illumina v2 reagents (based on 300x mean coverage of targeted content). The TSCancer target is 252,835 bp long and allows to sequence 24 samples per run with Illumina v2 reagents (based on 250x mean coverage of targeted content).

### **3.5.2. Computing performance study**

Critical problems in the management of NGS data within the laboratories include the quantification and selection of the suitable computing resources supporting the sequence analysis. There is no optimal solution for each case, but the the computing capacity to has to be optimized for the specific needs. Some laboratories produce a mass of data compatible only with high-performance computing systems, but in many other cases it is possible to adopt fewer demanding solutions such as workstations or personal computer stations. The analysis of gene panels such as Trusight cardio (TSCardio) and Trusight cancer (TSCancer), for example, does not require high computational performance and could be performed in stations with hardware features that are now common on the market and at low cost.

To verify the analysis capabilities of a common computer (PC) against a workstation (WS) designed ad hoc for targeted sequencing applications, the performance study of a pipeline implemented through Helper was performed, evaluating the analysis times of the samples sequenced using TSCardio and TSCancer gene panels. 10 cohorts of samples were selected for each of the two Trusight panels. Each of the cohorts is derived from a sequencing experiment performed on MiSeq Illumina. The cohorts prepared with the TSCardio contain of 12 samples, while those prepared with the TSCancer are composed of 24 samples. The workstation (WS) used for the tests has 64 GB (4 x 16, DDR4) of RAM memory and an Intel core i9-10940X processor with 3.30 Ghz and 28 threads. To simulate the use of a PC, analysis of the samples was started in serial mode, using 2 threads for each step and limiting the

amount of usable RAM to 8 GB. Instead, to calculate the performance of the workstation, the analysis of the samples was started in parallel, using 2 threads for each step and without limiting the use of RAM. The average times, calculated on the analysis of the 10 different cohorts for both panels, were recorded both for the single step and for each module of the pipeline by comparing the serial mode with the parallelized one.

The pipeline used is the “`trusight_germline`”. The `trusight_germline` was developed through Helper and is implemented within the platform as a precompiled pipeline. The workflow of the pipeline include:

- Alignment of Fastq files using BWA-mem.
- Sam to Bam files conversion, bam sorting, and marking of duplicates using Picard tool.
- Realignment around indels and Base Quality Score Recalibration using GATK v.3.7.
- Joint Variant calling using GATK v.4.1 and Freebayes.
- Annotation using VEP.
- VCF to TSV files conversion.
- CNV calling using GATK v4.1.

For the TSCardio, the computing capacity of the workstation is sufficient to analyze all 12 samples in parallel using 2 threads tools for each instance. Also for the TSCancer, the workstation can analyze 12 samples simultaneously, which however represent half of the samples in the cohort. The results are summarized in tables 3.4 and 3.5.

### 3.5.2.1. `Trusight_germline` runtime

The first module is the Alignment; the entire module is processed with average times of 3 min and 14 sec for a single TSCardio sample and 1 min and 28 sec for each TSCancer sample. The step that takes highest time fraction is the alignment of the Fastq files with BWA (2 min, 36 sec). The most time-consuming module is the preprocessing of the Bam files, which takes about 15 min for each TSCardio sample and almost 8 min for each TSCancer sample. By itself, the Indel realignment step represents about 30% of the processing time of the entire preprocessing module; and skipping it would allow a significant time saving (see chapter 2.3.3). The Base quality recalibration step takes a long time to perform (9 min for each TSCardio sample and 4 min for TSCancer) and represents almost 60% of the module time and between 25 and 30% of the processing time of the entire pipeline for calling and annotating short variants. The variant calling module performed using two tools takes about 5 min and 3 min and 40 sec per sample for TSCardio and TSCancer, respectively, and the processing time of the two tools is almost the same. The Annotation module of the variants performed using VEP takes about 3 min for the TSCardio and about 2 min and 30 sec for the TSCancer, while the post-processing module, in which the

information on the variants in TSV format is reported, takes just over 1 minute. In total, the complete workflow, excluding CNV calling, of the Trusight\_germline pipeline takes about 30 minutes to analyze a single TSCardio sample, while for the TSCancer it takes just over 16 minutes. The difference is due to the size of the two targets; the TSCardio has an almost double target than the TSCancer, and the processing times of the modules that analyze the entire target (alignment, preprocessing, and variant calling) reflect this proportion. The average performance recorded for the simulated PC considering the analysis of all the samples in the cohort, demonstrate a pipeline processing time of approximately 342 minutes for the TSCardio and approximately 393 minutes for the TSCancer. Both experiments take 6 to 7 hours to complete the pipeline. This is because, although the target of the TSCardio is almost double than the target to the TSCancer, the TSCancer cohort contains twice as many samples as the TS cardio. The step of CNV analysis was excluded from this calculation as it is always carried out in parallel on all samples and does not respect the design of the experiment based on the analysis in series vs. in parallel. However, taking into account the analysis of the CNVs, a PC could take about 10 hours to complete the entire workflow, which corresponds to an acceptable time for an overnight analysis. The comparison of the performance in terms of timing between the use of a PC, which can potentially analyze the samples only in series, and a workstation that can take advantage of the paralleling of the processes, demonstrates that the WS is able to perform each step over the entire cohort in the same amount of time that a PC analyzes a single sample. The PC perform the workflow of the Trusight\_germline on the entire cohort 10 to 12 times slower than the WS. Despite the significant time savings that are achieved by using a workstation, Helper can also be used in laboratories where high computing solutions are not available.

**Table 3.4:** Processing time for Trusight Cardio panel

TRUSIGHTCARDIO			
	SERIAL		PARALLEL
STEP	1 SAMPLE	12 SAMPLES	12 SAMPLES
ALIGNMENT			
BWA MEM	2 min 36 sec	31 min 12 sec	2 min 41 sec
SAM TO BAM	0 min 12 sec	2 min 24 sec	0 min 13 sec
SORT BAM	0 min 26 sec	5 min 12 sec	0 min 29 sec
TOTAL	3 min 14 sec	38 min 48 sec	3 min 23 sec
PREPROCESSING			
ADD READ GROUP	0 min 50 sec	10 min 0 sec	0 min 53 sec
MARK DUP	1 min 11 sec	14 min 12 sec	1 min 15 sec
INDEL REALIGNMENT	4 min 15 sec	51 min 0 sec	4 min 22 sec
QB RECALIBRATION	9 min 15 sec	111 min 0 sec	9 min 13 sec
TOTAL	15 min 21 sec	186 min 12 sec	15 min 43 sec
VARIANT CALLING			
GATK	2 min 32 sec	30 min 24 sec	2 min 28 sec

FREEBAYES	2 min 55 sec	32 min 56 sec	3 min 07 sec
TOTAL	5 min 27 sec	63 min 20 sec	5 min 35 sec
ANNOTATION			
VEP	3 min 12 sec	38 min 24 sec	3 min 25 sec
POST ANNOTATION			
VCF TO TSV	1 min 13 sec	14 min 36 sec	1 min 16 sec
OVERALL TOTAL			
OVERALL TOTAL	28 min 27 sec	342 min 16 sec	29 min 22 sec
CNV CALLING			
GATK	-	258 min 56 sec	259 min 32 sec

**Table 3.5:** Processing time for Trusight Cancer panel

TRUSIGHTCANCER			
STEP	SERIAL		PARALLEL
	1 SAMPLE	24 SAMPLES	2X12 SAMPLES
ALIGNMENT			
BWA MEM	1 min 0 sec	24 min 0 sec	2 min 12 sec
SAM TO BAM	0 min 10 sec	4 min 0 sec	0 min 28 sec
SORT BAM	0 min 18 sec	7 min 23 sec	0 min 41 sec
TOTAL	1 min 28 sec	35 min 23 sec	3 min 23 sec
PREPROCESSING			
ADD READ GROUP	0 min 47 sec	18 min 48 sec	1 min 41 sec
MARK DUP	0 min 54 sec	21 min 36 sec	2 min 01 sec
INDEL REALIGNMENT	2 min 12 sec	52 min 28 sec	4 min 38 sec
QB RECALIBRATION	4 min 01 sec	96 min 24 sec	8 min 25 sec
TOTAL	7 min 44 sec	185 min 50 sec	16 min 45 sec
VARIANT CALLING			
GATK	1 min 54 sec	45 min 36 sec	4 min 43 sec
FREEBAYES	1 min 46 sec	42 min 18 sec	4 min 12 sec
TOTAL	3 min 40 sec	87 min 54 sec	8 min 55 sec
ANNOTATION			
VEP	2 min 28 sec	59 min 12 sec	5 min 32 sec
POST ANNOTATION			
VCF TO TSV	1 min 1 sec	24 min 24 sec	2 min 12 sec
OVERALL TOTAL			
OVERALL TOTAL	16 min 21 sec	392 min 43 sec	36 min 47 sec
CNV CALLING			
GATK	-	190 min 8 sec	194 min 41 sec



### 3.5.3. CNV Analysis

One of the critical points of the bioinformatics pipeline is the analysis concerning the CNVs in the samples studied for targeted sequencing applications. The potential of the analysis of CNVs on NGS samples can be assessed in economic and time terms. The CNVs are in fact studied mainly through MLPA (multiplex ligation-dependent probe amplification), which still today represents the gold standard method, through real time PCR (rtPCR), or through array CGH (aCGH). All three methods have substantial flaws, including the need to prepare an additional experiment, which increases the costs of studying the sample and lengthens reporting times. These methods also have problems with the accuracy of the result and still require validation of the findings. The call of the CNVs in the same assay, in which the short genomic variants are studied, becomes essential in the diagnostic path of genetic diseases, but requires particular attention in the validation of the results to better understand the expected error range that must be calculated when issuing a report.

To understand the difficulties related to the detection of the CNVs, a performance study of the tools dedicated to the call of the CNVs, implemented in Helper, was performed. In order to compare copy-number-variation (CNV) detection methods, for targeted NGS panel data in a clinical diagnostic setting, 3 CNV callers were evaluated on 3 CNV datasets validated using MLPA, rtPCR, or aCGH methods. The tools used are GATK V4 in cohort and single sample mode, CoNVaDING, and CNVkit.

#### 3.5.3.1. Datasets and tools

Three datasets were included in this benchmark, 2 with data from TSCancer sequencing panel, and 1 from TScardio panel:

- The panelcnDataset (IBK) [94][95] contains 170 samples that were processed using the Illumina Trusight Cancer and sequenced using Illumina MiSeq instrument. The dataset contains single exon CNV (n=19), multi exons CNV (n=22), and whole gene CNV (n=6) validated using MLPA assays. The panelcnDataset is accessible on the European Genome-Phenome Archive (EGA) using the EGAD00001003400 dataset ID. (<https://ega-archive.org/datasets/EGAD00001003400>).
- The OSM-TSCancer dataset contains 70 samples from the OSM population. Samples were processed using the Illumina Trusight Cancer and sequenced using Illumina MiSeq instrument. The CMGCV-TSCancer dataset contains 19 samples with CNV, including single exon (n=6), multiple exon (n=11), and whole gene CNV (n=2), and 51 samples without CNV in analyzed genes.

- The OSM-TSCardio dataset contains 150 samples from the OSM population. Samples were processed using the Illumina TruSight Cardio and sequenced using Illumina MiSeq instrument. The CMGCV-TSCancer dataset contains 70 samples with CNV, including single exon (n=11), multiple exon (n=38), and whole gene CNV (n=21), and 80 samples without CNV in analyzed genes.

For each dataset, the samples without known CNVs were considered as the control population. For the IBK dataset, the control population is composed of 123 samples, for OSM-TSCancer 51 samples, and for OSM-TSCardio 80 samples.

For GATK tools in single sample modality (GATK-ss) the read depth values for each sample were calculated; those belonging to the control group were used for the calculation of the ploidy model and the Call model; subsequently the call was made on the test samples. This process was done for all three datasets. For GATK in cohort modality (GATK-cohort) the same read depth calculation process was performed, but both test and control samples were used to calculate the ploidy model and call the CNVs.

For CoNVaDING, the control samples separated from the test samples, were used to call the CNVs. From the control set, 30 best match samples were chosen for each test sample to improve the accuracy of the analysis. For the evaluation of the results, the set of CNVs contained within the extended list produced by CoNVaDING was used.

Also for CNVkit, the samples of the control set were used to generate a reference. CNVkit in addition to the analysis target, also requires studying the off-target coverage to improve the call of the CNVs. The reference in target and the reference off target were used to study each sample separately and to generate the variant call. All the tools were used with the default settings, in order to compare the finding without altering the result by customizing the analysis.

### 3.5.3.2. Benchmark evaluation metric

The performances of each tool for CNVs detection were evaluated considering the calling sensitivity defined as  $TP / (TP + FN)$ . Each validated CNV that is identified by the tools represents a True Positive call (TP), while each validated CNV not found is considered as True Negative call (TN). The CNV calls made by the tools that concern genes other than those containing the validated CNVs were not considered as False Positive calls because it is not possible to define which ones have actually been studied with a second method.

### 3.5.3.3. CNV calling sensitivity

In the sensitivity test, the call performance of the CNVs of the four tools, on the three datasets, are evaluated. GATK in cohort mode and GATK in single sample mode are considered as two different tools. The results refer to both the sensitivity level of the tools, and the number of CNVs identified or missed by CNV callers. This is because, in addition to performance statistics, each CNV missed has an important weight within the diagnostic workflow.

For the IBK Dataset, which contains 47 validated CNVs (40 deletions and 7 duplications) the tool that identifies the greatest number of TPs is GATK, which identifies 42/47 CNV in both cohort and single sample mode. Among these 42 CNVs 36 are deletions and 6 are duplications. GATK misses the detection of 4/40 deletions and 1 duplication. CNVkit and CoNVaDING only call 38 and 39 CNV, respectively. Both tools identify 4 out of 7 duplications in the IBK dataset, CNVkit identifies 34/40 deletions, while CoNVaDING identifies 35/40. Comparing the detection capacity levels of the tools on the entire IBK dataset, it is noted that GATK is the most performing tool, with a sensitivity of 0.894 against the 0.818 of CNVkit and 0.864 of CoNVaDING. Despite the few duplications present in the IBK dataset, it is interesting to note that the missing rate of CNVkit and CoNVaDING for this type of CNV exceeds 40% against 15% for GATK.

For the OSM-TSCancer Dataset, which contains 18 deletions and 1 duplication, the tool that performs best is CoNVaDING. CoNVaDING identifies 18 CNVs, missing only 1 deletion in the BRCA2 gene. Also in this case, GATK-cohort and GATK-ss show the same performances, identifying 17/19 CNV, missing the same 2 deletions, and identifying the only duplication present in the dataset. CNVkit identifies 16/19 variants, with a missing rate of approximately 5%. The CoVading sensitivity goes from 0.864 on the IBK dataset to 0.947 on the OSM-TSCancer, with a missing rate of 13.6% on the two datasets together. As regards GATK, both in cohort and in single sample modality, the sensitivity remains unchanged on the two datasets, considered separate or considered together, with a missing rate of 10.6% on the datasets composed of Trusight Cancer samples. CNVkit is the least performing tool on both datasets, with a sensitivity on the OSM-TSCancer of 0.842 and a missing rate on the two datasets of 18.2%.

The OSM-TSCardio dataset contains 71 CNVs of which 42 deletions and 29 duplications. The tool that performs best on this dataset is GATK-ss, which identifies 70/71 CNV, calling all deletions, and missing the 1 duplication detection. GATK-cohort and CoNVaDING identify 68/70 CNVs, but GATK calls 28/29 duplications and 40/42 deletions, while CoNVaDING calls 27 duplications and 41 deletions. CNVkit is the tool with the highest number of missed CNVs, identifies 40 deletions and 27 duplications, with a missing rate of 5.6%. The sensitivity index on the TSCardio dataset is greater than the two TSCancer datasets, for all 4 tools. GATK-ss is the best tool, with a sensitivity of 0.986, followed by GATK-cohort and CoNVaDING which identify 95.8% of the variants, and finally, by CNVkit with a sensitivity of 0.944.

Considering all three datasets together, the best performing tool is confirmed to be GATK in single sample modality, with a sensitivity of 0.942 and a missing rate of 5.8%. GATK-cohort and CoNVaDING demonstrate very similar performances on all CNVs, with a sensitivity of 0.927 and 0.912, respectively. The results change considering the type of CNV, GATK-cohort has a duplication detection rate higher (94.6%) than CoNVaDING (86.5%), but identifies 1% fewer deletions (92% GATK vs 93% CoNVaDING). CNVkit has an overall sensitivity of 0.883, which translates into a double missing rate compared to GATK-ss (11.7% CNVkit vs 5.7% GATK). Even CNVkit, like CoNVaDING, demonstrates a higher difficulty in identifying duplications than deletions with a sensitivity of 0.865 and 0.890, respectively.

**Table 3.6:** CNV TP calls

Dataset	CNV Type	Validated CNV	GATK-Cohort	GATK-ss	CNVkit	Convading
OSM -TSCancer	DEL	18	16	16	15	17
OSM -TSCancer	DUP	1	1	1	1	1
OSM -TSCancer	ALL	19	17	17	16	18
IBK	DEL	40	36	36	34	35
IBK	DUP	7	6	6	4	4
IBK	ALL	47	42	42	38	39
OSM-TSCardio	DEL	42	40	42	40	41
OSM-TSCardio	DUP	29	28	28	27	27
OSM-TSCardio	ALL	71	68	70	67	68

**Table 3.7:** Sensitivity of CNV callers

Dataset	CNV Type	GATK-Cohort	GATK-ss	CNVkit	Convading
OSM -TSCancer	DEL	0.889	0.889	0.833	0.944
OSM -TSCancer	DUP	1.000	1.000	1.000	1.000
OSM -TSCancer	ALL	0.895	0.895	0.842	0.947
IBK	DEL	0.900	0.900	0.850	0.875
IBK	DUP	0.857	0.857	0.571	0.571
IBK	ALL	0.894	0.894	0.818	0.864
OSM-TSCancer+ IBK	DEL	0.897	0.897	0.845	0.897
OSM-TSCancer+ IBK	DUP	0.875	0.875	0.625	0.625
OSM-TSCancer+ IBK	ALL	0.894	0.894	0.818	0.864
OSM-TSCardio	DEL	0.952	1.000	0.952	0.976
OSM-TSCardio	DUP	0.966	0.966	0.931	0.931
OSM-TSCardio	ALL	0.958	0.986	0.944	0.958

3 DATASETS	DEL	0.920	0.940	0.890	0.930
3 DATASETS	DUP	0.946	0.946	0.865	0.865
3 DATASETS	ALL	0.927	0.942	0.883	0.912

In addition to considering the type of CNV, it is useful to understand the detection capacity based on the size of the variant. It is well known that the CNVs of a single exon are difficult to identify, due to the lack of informativity compared to the larger CNVs [(94)]. To evaluate this aspect, given that the performances of each tool are similar on the two datasets composed of samples sequenced with the TSCancer, the CNVs of the IBK and OSM-TSCancer datasets were considered as a single dataset.

The TSCancer is made up of 25 single exons, 8 full gene, and 33 multiple exons CNVs. All the tools prove to have greater difficulty in identifying CNVs composed of a single exon: GATK-cohort and GATK-ss identify 19/25 variants, CoNVaDING 17/25 and CNVkit 14/25. All the tools are able to identify 100% of the CNVs that affect the whole gene, while GATK-cohort, CoNVaDING, and CNVkit miss 1 CNV that spans over more exons. The missed multi-exon CNV is the same for all three tools, it is a deletion of 2 exons (exons 8 and 9) in the EPCAM gene related to colorectal carcinoma.

The TSCardio is composed of 11 single exon, 21 full gene, and 39 multiple exon CNVs. The ability to detect single exon CNVs in this gene panel is greater than in TSCancer. GATK-ss can identify 10/11, GATK-cohort and CoNVaDING identify 9/10, and CNVkit calls 8/11. All the tools identify 100% of the full CNV genes, while only GATK-ss can find 100% of the CNVs composed of more than one exon. The other 3 tools identify 70 out of 72.

As expected, the single exon CNVs are the ones that put the CNV callers in greater difficulty. The tool that demonstrates the best performances is, also in this case, GATK-ss, which identifies all the CNVs involving more than one exon, with a sensitivity of 1,000. The detection rate of GATK-ss is lower for single exon CNVs, the tool calls only 80% of the variants. The performances have a similar trend also for the other tools, with a sensitivity for multi exon CNV of 0.972 and a very high single exon CNV missing rate. GATK-cohort misses 23% of the variants, CoNVaDING 27.8%, and CNVkit nearly 40%.

**Table 3.8:** Number of called CNV based on CNV length

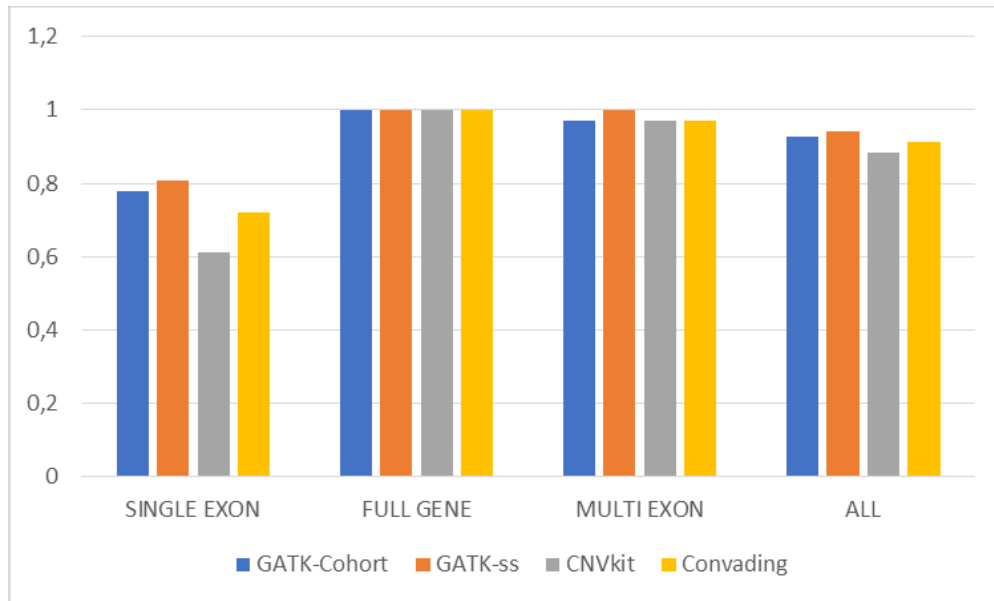
Datase	CNV Type	Validated CNV	GATK-Cohort	GATK-ss	CNVkit	Convading
TSCancer	SINGLE EXON	25	19	19	14	17
TSCancer	FULL GENE	8	8	8	8	8
TSCancer	MULTI EXON	33	32	33	32	32

## The Helper platform

TSCardio	SINGLE EXON	11	9	10	8	9
TSCardio	FULL GENE	21	21	21	21	21
TSCardio	MULTI EXON	39	38	39	38	38
3 DATASETS	SINGLE EXON	36	28	29	22	26
3 DATASETS	FULL GENE	29	29	29	29	29
3 DATASETS	MULTI EXON	72	70	72	70	70

**Table 3.9:** Sensitivity of CNV callers based on CNV length

<b>Dataset</b>	<b>CNV Type</b>	<b>GATK-Cohort</b>	<b>GATK-ss</b>	<b>CNVkit</b>	<b>Convading</b>
TSCancer	SINGLE EXON	0,760	0,760	0,560	0,680
TSCancer	FULL GENE	1,000	1,000	1,000	1,000
TSCancer	MULTI EXON	0,970	1,000	0,970	0,970
TSCardio	SINGLE EXON	0,818	0,909	0,727	0,818
TSCardio	FULL GENE	1,000	1,000	1,000	1,000
TSCardio	MULTI EXON	0,974	1,000	0,974	0,974
3 DATASETS	SINGLE EXON	0,778	0,806	0,611	0,722
3 DATASETS	FULL GENE	1,000	1,000	1,000	1,000
3 DATASETS	MULTI EXON	0,972	1,000	0,972	0,972



**Figure 3.13:** The CNV callers performances in terms of True call sensitivity for each of the CNV type

The performance test of the tools has shown that it is necessary to pay attention to the call results of the NVCs. The single tools alone cannot guarantee a detection rate of 100%, especially for single exon CNVs. The results show that the CNVs contained in the samples sequenced with the TSCancer are identified with more difficulty than those of the TSCardio dataset. This could be related to the difference in target coverage by the two panels, in fact the TSCardio sequencing panel has an in-target coverage of about 82% of the aligned reads, and the TSCancer has an in-target of about 70%. The percentage of on and off target could have an important impact on the result of the call of the CNVs. Furthermore, the evaluation of False Positives was excluded from the performance analysis. The PFs increase the uncertainty about the result by decreasing the total accuracy of the analysis. CNVs are variants that often have an important impact on the phenotype of carriers, and it is essential to be able to identify them with certainty. Even if a missed CNV has a greater weight than False positives calls, the presence of the latter, generates the need to confirm the result with a second method. For this reason, it is necessary to identify the right set-up for each tool to maximize the accuracy of the CNV call.

# Chapter 4

---

## Clinical Applications

In the previous chapter, a novel solution to adapt bioinformatics analysis to different target-sequencing applications was presented. Customizing pipelines is just one of the challenges that must be faced in the path of genetic test optimization. When the sample is sequenced, analyzed, and a narrow set of variants has been identified, the further step is the classification in order to correlate the finding with the patient's phenotype. Understanding the role of variants within the gene, and the role of different genes in the disease, are two fundamental processes both for diagnosis and for research in the molecular-genetic field. In this chapter, I will present two examples of Helper applications: the case of optimized interpretation of variants in the specific field of Desminopathies, and the process of exploring the heterogeneous genetic bases for hereditary Breast and Ovarian cancer syndrome.

### 4.1. Variant interpretation - the case of Desmin

The criteria for the classification of variants generated by the ACMG (see variant classification in chapter 2.7) has introduced a conservative and robust framework for the interpretation of the genetic data in the scientific community. The system was planned in such a way as to standardize the variant classification path regardless of the gene and the disease under examination. However, neither genes nor diseases are generalizable. For this reason, the current trend is to modify the ACMG system by adapting the strength of the criteria to increase the accuracy of classification of variants present in a specific gene (e.g., *MYH7*) or in a group of genes associated with a particular disease [96][97]. This chapter aims at describing a Desmin-specific adaptation system of the ACMG rules, and includes:

1. the description of the gene and of the clinical issues related to the defects of DES, which concerns a subgroup of highly malignant heart diseases.
2. the path that led to the development of the adapted system.
3. the dataset of variants in DES identified within the cohort of patients cared for desminopathies at the CMGCV.



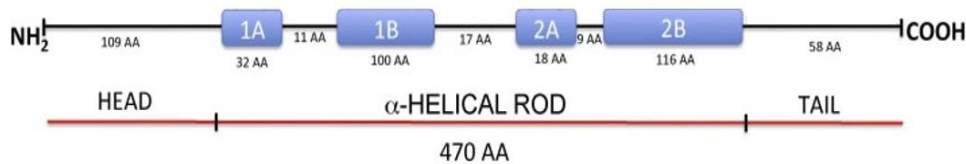
- the comparative analysis of the ACMG-based classification results using 3 commercial software, and the OSM system adapted for DES and using the clinical and pathological findings that establish the precise diagnosis.

The goal is to provide the rules for definite and irrefutable diagnosis of Cardiodesminopathy.

#### 4.1.1. Clinical and genetic background

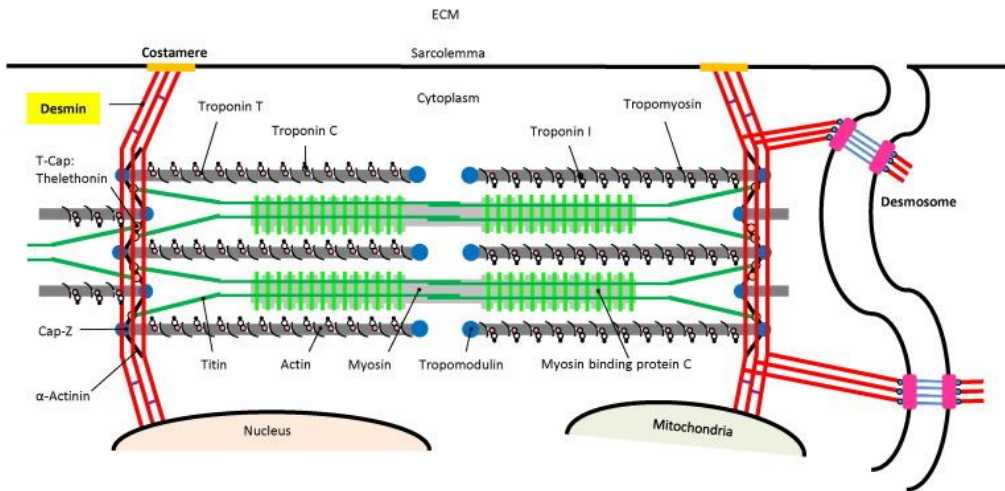
##### 4.1.1.1. The DES gene and the Desmin protein

The DES gene maps in the chromosome 2 (2q35); it consists of 9 exons, about 8.4 kilo bases [kb]. The mature protein contains 470 amino acids and is composed of a non-helical amino-terminal domain (Head), a central alpha helical rod and a non-helical domain carboxy-terminal (Tail). The central rod domain is composed of four helices (coil-1A, coil-1B, coil-2A, coil-2B) interspersed with 3 short non-helical linkers (L1, L12, L2).



**Figure 4.1:** The Desmin protein primary structure (Figure modified from [98]).

The DES gene encodes the class III intermediate filament (IF) protein Desmin that plays a central role in the cytoskeleton structure of the cells. IFs are constituted of highly flexible non-globular protein units that form an elastic scaffold connecting most of the cytoplasm structures of skeletal myocytes and cardiomyocytes. Desmin is expressed in cardiac, skeletal, and smooth muscle cells. Within these cells, its first function is myofibril stabilization, by inter-connection of Z-disks and forming a three-dimensional network that extends from the nucleus to the junctional structures such as desmosomes and adhesion structures such as costamers; the second function is the transmission of the mechanical force of cellular contraction to the extracellular matrix and to the other adherent cells; finally, Desmin regulates the distribution and modulates the function of the mitochondria within the cytoplasm.



**Figure 4.2:** The intracellular organization of myocytes and the connecting role of the Desmin. (Figure from [98])

#### 4.1.1.2. Phenotypes related to DES defect

DES defects were first reported in relation to semi-dominant Myofibrillar Myopathy (MFM). Desmin-related MFM - also called desminopathy - defines a set of inherited muscle diseases primarily characterized by abnormal aggregates of misfolded Desmin in the cellular cytoplasm. The desminopathy phenotype is characterized by progressive muscle weakness, cardiomyopathy, and abnormalities of cardiac rhythm.

In the current literature, DES gene defects are associated with phenotypically heterogeneous cardiomyopathies, which include Dilated cardiomyopathy (DCM), Hypertrophic cardiomyopathy (HCM), Restrictive cardiomyopathy (RCM), and Arrhythmogenic cardiomyopathy (ARVC) (98) (99). The ClinGen expert panel (<https://www.clinicalgenome.org/>) classified the DES defects as strongly associated with DCM, ARVC, and MFM. However, the most typical cardiac phenotype caused by Desmin-related MFM is RCM associated with atrio-ventricular conduction delay that evolves over time to Atrioventricular Block (AVB). Desminopathy evolves to progressive heart failure and, in many cases, to the need for heart transplantation (HTx).

#### 4.1.1.3. Genetic complexity of Desminopathy

Interpretation of DES variants is especially complex due to the heterogeneity of the phenotypes reported to date as associated with *DES* gene defects. Many associations remain questionable mainly due to the lack of demonstration of the misfolded Desmin within the cardiac myocytes. The complexity increases when considering that variants reported to date in amino acid residues close to each other are related to different types of cardiomyopathy, or that different diseases within the same family are

associated with the same DES variant. This latter evidence, together with the presence of unaffected carriers, due to the heterogeneity of the onset age and the penetrance of desminopathy (about 80%), makes difficult the interpretation of family studies and complicates the proper management of families. Although all above issues are reported in the literature, they do not convince on a clinical level, and are difficult to support when the information needs to be translated from the scientific papers to patients. The concept is that when a genetic variant is found in the DES gene, for example in a patient with a typical dilated phenotype, and the same variant is reported in a different phenotype, neither observed case nor reported cases are sufficient to close the diagnostic dilemma on pathogenicity of the given variant.

#### **4.1.2. The CMGCV-DES system**

To break down the interpretative uncertainty associated with DES variants, all ACMG criteria were analyzed and investigating the literature, clinical databases such as Clinvar, and population databases such as ExAC and GnomAD, a DES adapter system was developed. The result is the CMGCV-DES system, in which some ACMG criteria can be activated in a specific way using dedicated thresholds, some criteria are not recommended, while others can be used without particular precautions, according to the characteristics of the DES gene.

##### **4.1.2.1. Variant type and location (PVS1 / PM1 / PM4 / PP2)**

The PVS1 criterion is activated in the presence of a variant that induces Loss of function (LoF) in the protein and LoF is a known mechanism of disease. Variants that induce the loss of protein function are the so-called "null variants" (nonsense, frameshift, canonical  $\pm$  1 or 2 splice sites, initiation codon, single or multi-exon deletion). To quantify the tolerance of a gene to LoF variants, two indices are usually evaluated: the probability score (pLI) of Intolerance to LoF variants calculated from ExAC data and the ratio between observed and expected LoF variants (o/e constraint metric) calculated on gnomAD data. While a pLI close to 1 is usually dichotomized using 0.9 as threshold (pLI > 0.9 identifies intolerance to LoF), the index o/e, or rather the LOEUF (a more conservative estimate which is equivalent to the upper bound of the confidence interval of the Poisson distribution constructed on o/e), is a continuous value that indicates different degrees of tolerance. Although pLI or LOEUF are two easy-to-use numerical methods, caution is needed during the evaluation of their reliability for most adult-onset Mendelian disorders [100]. Evaluating the fraction of LoF variants classified as P or LP within databases such as ClinVar can help in understanding the effect of these variants on gene function. In ClinVar, for Des gene, 37 LoF variants are described; they are distributed in all exons of the gene and 30/37 (81%) are classified as conflict-free P or LP. Although

pLI = 0.047 (<0.9) and LOEUF = 0.596 indicate low or moderate intolerance, clinical correlation data between LoF variants and diseased phenotype extracted from ClinVar indicate that the *DES* gene is sensitive to the damage mechanism induced by null variants.

The CMGCV-DES system applies PVS1 by modulating the strength of the criterion according to the zone in which the null variant falls, considering the nonsense-mediated decay that could limit the protein damage as suggested by the recommendations for this specific criterion [101].

The PP2 criterion is activated in the presence of missense variants in a gene that has a low rate of benign missense variations and in which missense variants are a common mechanism of disease. As a general recommendation, based on ExAC and gnomAD data, GlinGen suggests the use of the z-score as an index of tolerability of the gene to missense variants. The z-score is calculated by comparing the observed missense variants to the expected ones and is directly proportional to the sensitivity of the gene to the presence of the missense variant. As for the pLI, a threshold can also be applied to the z-score to define tolerance/intolerance (threshold z-score = 3.09), but it may also be useful to evaluate the number of missenses classified in the databases as P or LP compared to those classified as B, LB or VUS to refine the rule for activating Score PP2.

The z-scores for missense variants in the *DES* gene calculated from ExAC and gnomAD are 2.45 and 1.7, respectively (<3.09). ClinVar reports 252 missenses of which 36/252 are classified without conflict of interpretation: 35/36 P or LP, 1/36 LB or B. Although the z-scores do not reach the recommended threshold of 3.09, they do not exclude a certain sensitivity of *DES* to missense (2.45 > the upper limit of the confidence interval on all z-scores calculated on ExAC), this data is confirmed by the fraction of missense variants classified in the literature as pathogenic compared to the benign ones. Furthermore, the large number of missenses classified as VUS or with conflict of interpretation shows that the uncertainty that accompanies the presence of a missense variant in *DES* is still high. For these reasons, the CMGCV-DES system activates the PP2 criterion using a Supporting strength.

The PM4 criterion is activated in the presence of an in-frame change in the length of the protein due to In-frame Deletions and Insertion that do not fall into a homopolymer zone (in the case of a repeated zone, BP3 is activated) or a Stop loss variant. ClinVar describes 12 in-frame InDels, of which 5 are classified as P / LP and none as LB / B, and 2 LP Stop loss. These types of variants are poorly described for *DES* and appear to have a harmful impact on the protein. For this reason, the CMGCV-DES system activates PM4 in the presence of in-frame InDels or Stop loss variant, using a Moderate strength.

The PM1 criterion is activated when a non-null and non-synonymous variant falls into a mutational hotspot or a functional domain important for protein function validated by experimental evidence. In the absence of robust regions intolerant to variations, alternative methods have been described in the literature that can be used to infer the presence of fragile areas of the

gene by analyzing the distribution of variants classified as P or LP and B or LB [102][103]. These systems identify exonic regions in which only P / LP variants are present with an increased density compared to control series. These systems generalize the PM1 activation method for all gene types. The PM1 score has a Moderate strength, and it is often the tip of the balance in the classification of variants that are not described in the literature, because it can shift the interpretation of a new variant from VUS to LP in the absence of experimental or clinical data. For this reason, caution should be applied in evaluating the presence of a fragile site or hotspot using only the positional information of the variants, without validated data about the characteristics of the analyzed gene. Given that the knowledge on *DES* fragile sites is not yet robust and the Moderate strength of PM1 is a decisive factor that could lead to overestimating the pathogenic interpretation if wrongly activated, the CMGCV-DES system does not apply the PM1 criterion.

#### **4.1.2.2. Same residue as known pathogenic (PS1 / PM5)**

The PS1 criterion is activated when another variant, that affects the same nucleotide of the analyzed variant, is known to be pathogenic, while the PM5 criterion is activated when the same affected amino-acid residue changes in another variant defined as pathogenic.

In ClinVar, reports for DES include 41 amino acid positions where the same amino acid varies in at least two different residues, and 14/41 (34.15%) involve at least one variant called P or LP with no classification conflict. None of the latter (0/14) is involved in a second P or LP variant, without conflict. Furthermore, among the P or LP variants, the distribution of the involved amino acids does not demonstrate the presence of a starting residue that, when mutated, is particularly harmful, while the introduction of a Proline in the amino acid sequence could have a detrimental effect on protein folding [99]. Due to the lack of evidence to support the PM5 criterion, the CMCV-DES system excludes it from the evaluation of the variants, while it does apply the PS1 criterion without modifications.

#### **4.1.2.3. Population frequency (PM2 / BS1 / BA1)**

The population frequency provides significant data for the interpretation of the variants. The criteria in favor of the benign interpretation vary from BA1 (Allele frequency too high in the control population), whose strength is Stand Alone, and therefore alone can be considered a filter to distinguish the benign variants that are too frequent in controls, to BS1 (allele frequency too high for the disorder) which corresponds to a Very Strong strength and which, if activated by itself in the absence of other scores, would move the pathogenicity class to LB. The score that considers the MAF in favor of the pathogenicity of a variant is PM2 (Absent or extremely rare from large population studies). PM2 has moderate strength and is activated when the

variant is absent or very rare in the control population. A determining factor for the evaluation of allele frequencies for the classification of variants is that MAFs are calculated in robust population datasets composed of at least 2000 alleles derived from unrelated subjects [104]. Since the frequency in the population is decisive for the classification of a variant and the use of the same thresholds for all genes and diseases could generate misinterpretation, the ClinGen expert groups have suggested methods to optimize the MAF thresholds for each of the three criteria (BA1, BS1, PM2) as a function of the phenotype and gene studied. An example is the case of *MYH7* for which the CMP-EP has developed a method for calculating the BA1 and BS1 thresholds that take into account the disease prevalence, gene contribution to disease, and estimated penetrance of the variant [105]. Using a conservative prevalence among all *MYH7*-associated phenotypes (1/400 chromosomes), a contribution of *MYH7* to HCM of 10.6%, and a mean penetrance of all *MYH7* variants of 30%, the following MAF thresholds were obtained: for BA1 it is a MAF > 0.1% (0.001), for BS1 the threshold is > 0.02% (0.0002) and for PM2 the threshold is < 0.004% (0.00004). The same thresholds are generally used for the evaluation of the variants present also in the other genes associated with hereditary cardiomyopathies [106].

The cardiac phenotype associated with *DES* with the highest prevalence is DCM. DCM has a prevalence of 1/2500 but may be higher according to some updated estimates. The contribution of *DES* to dilated cardiomyopathies is very low, around 1% [107][108], and the mean penetrance for DCM of variants in *DES* is unknown. Although, the value of the thresholds for BA1 and BS1, obtained using the specific data of *DES*, are an order of magnitude smaller than those of *MYH7*, it was preferred to start from the values commonly used for cardiomyopathies by inserting some adaptations: the BA1 criterion is activated if the MAF is > 0.001 or has been observed in a homozygous state on gnomAD (Number of Homozygous, NoH  $\geq 1$ ), and the BS1 criterion is activated if the MAF is > 0.0001.

Furthermore, in ClinVar there are 76 variants in Desmin (SNV and Short InDels) classified P and LP without interpretation conflicts, 34 are missense (1 of which involves 2 consecutive amino acids), 33 are null variant, 2 are non-canonical splices, 6 are inframe InDels and 1 is a synonym. Of all these, only 6 of the 33 null variants are reported in GnomAD, and c.194dup (p.Leu66fs) is the P variant with the highest MAF among all *DES* variants (MAF of 0.000032), while the one with the greatest allelic count (AC) is the variant of the canonical splice site c.1288+1G>A (AC = 6/282842 alleles). In GnomAD, 281 variants are reported for *DES* (Nonsense, missense, non-canonical splice, inframe delins) with a median MAF of 0.00000795 (CI 0.0000040 - 0.0000278); of these, 157/281 (55.8%) variants have an AC equal to 1 and 247/281 (87.9%) have MAF < 0.00004.

Taking into account that:

- missense, indels, and non-canonical splice sites defined as Pathogenic in ClinVar are not present in gnomAD,

- a nonsense variant is present with an allele count equal to 6 (all heterozygous subjects),
- the percentage of very rare variants in *DES* is high,
- some variants are recessive,

we have decided to apply PM2 in a different way between the null variant and the other variants also depending on their transmission (Table 4.1):

- Non-canonical splicing variants, missense, delins, and splices AD activate the PM2 criterion if they are absent in the control population, while those with AR transmission activate PM2 if they have a count  $\leq 1$  allele in gnomAD.
- The nonsense variants (Stop, FS, canonical splice, start loss) activate PM2 if the MAF is  $< 0.00004$  with an allele count  $\leq 1$ .

Table 4.1 - Rules for activating BA1, BS1, and PM2 criteria

Variant type	Inheritance	PM2	BA1	BS1
Missense NC splice InDels Synonymous	AD	Absent in control population	MAF $\geq 0.001$ or NoH $\geq 1$	MAF $\geq 0.0001$
	AR	AC $\leq 1$	MAF $> 0.001$	MAF $\geq 0.0001$
Stop gain Frameshift Canonical splice Start loss	AD	MAF $< 0.00004$ and AC $\leq 1$	MAF $\geq 0.001$ or NoH $\geq 1$	MAF $\geq 0.0001$

#### 4.1.2.4. Homozygous status (PM3)

PM3 is activated when a variant is found in trans with another pathogenic for a recessive disease. Desminopathy is known to be a semi-dominant disease caused by both heterozygous and homozygous variants. In the presence of a recessive variant in a homozygous state, CMGCV-DES applies the PM3 criterion downgraded to Supporting strength, but it can be upgraded to Moderate in the presence of more observations in favor of the recessive transmission of the variant [109].

#### 4.1.2.5. Specific phenotype (PP4/BS2)

The PP4 criterion is activated in the presence of a phenotype closely related to the mutated gene. Generally, for the evaluation of variants in genes of cardiomyopathies, it is recommended not to use PP4 due to the lack of

specificity of the genetic causes. Among the DES-related phenotypes, DCM and ARVC have multiple genetic causes and are attributable to Desmin defects in only a small percentage of cases. The cardiac expression of the MFM includes RCM + AVB associated with myopathy (detected by increased serum CK). These phenotypes considered individually are not specific enough to be associated to *DES* mutation but, considered as a complex phenotype, they can easily be related to Desminopathy. The main difficulty in evaluating the phenotypes associated with MFM is their clinical identification in the different stages of the disease: RCM is not always full-blown and can often be mistaken as mild concentric HCM, if the atrial chambers are not correctly evaluated; cardiac filling patterns evolve from semi-normal to restrictive over time; the AVB may have a late-onset, but it is always anticipated by a conduction delay that can be found as a long PQ wave in the ECG trace. These intermediate phenotypes could be due to an early stage desminopathy but cannot be considered specific enough to activate PP4. In conclusion, the CMGCV-DES system considers the RCM + AVB + myopathy complex phenotype as closely related to MFM-Desmin related and activates PP4; on the contrary, mild phenotypes do not activate the PP4 criterion, but can be used in co-segregation studies when parents show desminopathy and sons present the mild phenotype.

The BS2 criterion is activated in the presence of healthy adults. Due to the phenotypic heterogeneity associated with Desmin mutations, the CMGCV-DES system considers "healthy" only adult patients who do not report the clinical characteristics of onset of cardiomyopathy on specific instrumental tests and who do not have a family history related to inherited cardiomyopathy.

### **4.1.2.6. Functional studies (PS3 / BS3)**

The PS3 and BS3 criteria are activated if well-performed functional studies demonstrate a correspondingly harmful or neutral effect on the protein and phenotype. The strength with which these criteria are applied is modulated according to the type of functional studies performed and the robustness of their results, and the certainty of the pathological effect of the mutated protein [110].

Functional studies for DES variants are commonly performed on in vitro cell models, in vivo mouse models, and on tissue from affected patients. The in vitro models use different cell lines for the evaluation of the structural conformation of the cytoplasm and the integrity of the cytoskeleton by fluorescence microscopy. Mouse models allow the evaluation of both the cellular structure and the phenotype induced by the mutation. The pathological assessment of the tissue of affected patients, on the other hand, allows the effects of cell damage on humans to be investigated directly in vivo and to have a direct comparison with the clinical phenotype.

The intracellular granulo-filamentous accumulations (or myofibrillar material MFM) characteristic of Desminopathy are easily diagnosed with



electron microscopy (EM), especially if performed through ultrastructural immuno-histochemistry (U-IHC) methods that allow to specifically mark the Desmin protein inside the cells. The same accumulations observed in light microscopy, both in bright field and in fluorescence, lose some of their specificity and can be used as distinctive signs in skeletal muscle cells but not in myocardial tissue [111]. The difference in diagnostic capacity is due to the characteristics and the location of the accumulation in the different cell types. In skeletal myocytes, Desmin aggregates are localized at subsarcolemmal level with unique characteristics when labeled with anti-Desmin antibodies. On the contrary, in the myocardium, the accumulations are arranged in a diffuse manner in the cytoplasm of the cardiomyocyte cells and can be confused with the contracture bands due to the action of the biopptome in the biopsy site.

The CMGCV-DES system evaluates:

- The Immuno-ultrastructural study as diagnostic test to determine the presence of MFM and activate the PS3 with Stand Alone strength.
- The EM study without the use of highly specific Desmin markers for the characteristics of the accumulations in ultrastructure and activates the PS3 with Very Strong strength.
- The study of optical IHC, in bright field and in fluorescence, on skeletal myocytes as sufficiently robust for diagnosis and activates the PS3 with Strong strength.
- The study of optical IHC, in bright field and in fluorescence, on cardiomyocytes as not robust enough and does not activate the PS3.

As far as studies with animal and cell models are concerned, the presence of robust and validated results makes possible the activation of PS3 according to the recommendations of the scientific society [110].

The BS3 criterion is activated with Strong strength, as it is not possible to exclude causative damage of the variant for the DCM or ARVC phenotypes even in the absence of specific accumulations for MFM. There is no scientific evidence of the structural characterization of myocytes and cardiomyocytes in the presence of these phenotypes.

#### **4.1.2.7. In-silico prediction (PP3 / BP4)**

The PP3 and BP4 criteria are activated if the in-silico tools that predict the impact of the variant gives a result in favor of or against pathogenicity. It is not clear how to evaluate the different tools, some software uses a majority rule to activate the criteria, others activate PP3 and BP4 exclusively, and others activate them at the same time. This criterion is difficult to apply. The CMGCV-DES system uses 9 prediction tools for the evaluation of the missense variants and 2 tools for the evaluation of the splicing variants.

CMGCV-DES activates the PP3 and BP4 criteria exclusively taking into account the evaluation trend of the different tools, using following rules:

- criterion PP3 is activated if at least 8/9 software predict a harmful impact for the missense variants.
- criterion PP3 is activated if 2/2 software predict a damaging impact for variants in canonical and non-canonical splicing sites.
- the BP4 criterion is activated if at least 8/9 software predict a benign impact for the missense variants.
- criterion BP4 is activated if 2/2 software predicts a benign impact for variants in canonical and non-canonical splice sites.

**Table 4.2:** The adapted criteria of the CMGCV-DES system

CRITERIA	CHANGED	ADAPTION
PATHOGENIC CRITERIA		
PVS1	not changed	modulated using [101]
PS3	changed	Stand_alone strength if U-IHC results positive; Very_strong strength if EM shows aggregates; Strong strength if LM in skeletal myocytes shows aggregates
PM1	changed	not applicable
PM2	changed	see Table 4.1
PM3	changed	homozygous variants activate PM3_Supporting
PM5	changed	not applicable
PP3	changed	Missense: 8/9 tools predict damage Splicing: 2/2 tools predict damage
PP4	changed	if MFM phenotype (RCM+AVB+sPK+) is present
PS1, PS2, PS4, PM4, PM6, PP1, PP2, PP5	not changed	applicable
BENIGN CRITERIA		
BA1	changed	see Table 4.1
BS1	changed	see Table 4.1
BS2	changed	if present in multiple controls
BS3	changed	not applicable
BP4	changed	Missense: 8/9 tools predict no damage Splicing: 2/2 tools predict no damage
BS4, BP1, BP2, BP3, BP5, BP6, BP7	not changed	applicable

### 4.1.3. The DES-dataset

At the CMGCV, from 2015 to 2021, 2562 unrelated subjects with hereditary cardiomyopathy and controls with other genetic diseases were studied. All probands were tested using NGS sequencing; library preparation was performed using the Illumina Trusight Library Preparation Kit in combination with the Illumina Trusight cardio probes. Samples, after quantification and library quality control, were sequenced using MiSeq Illumina in a pool of 12 samples per run as per Illumina protocol.

The bioinformatics analysis was performed through Helper using the Trusight analysis pipeline described in chapter 4.1. All variants identified in the cohort were grouped by the affected gene in the CMG-CardioDB. The variants affecting the *DES* gene have been subjected to a review aimed at identifying the candidates to be causative of Desminopathy. The final *DES*-dataset consists of 41 variants (Table 4.3): 33 are missense variants (1 of these involves two consecutive amino acids), 3 variants affecting canonical splice sites, 2 non-canonical splicing variants, 1 frameshift, 1 stop gain, and 1 exon deletion.

### 4.1.4. Benchmark study

To better understand how the non-adapted ACMG rules classify *DES* variants, variant classification was performed using three commercial software, commonly used in genetics laboratories, which support the annotation and interpretation of variants: Varsome [71], eVai [112], and Franklin [113]. The classification process was carried out using the criteria compiled automatically by the three software:

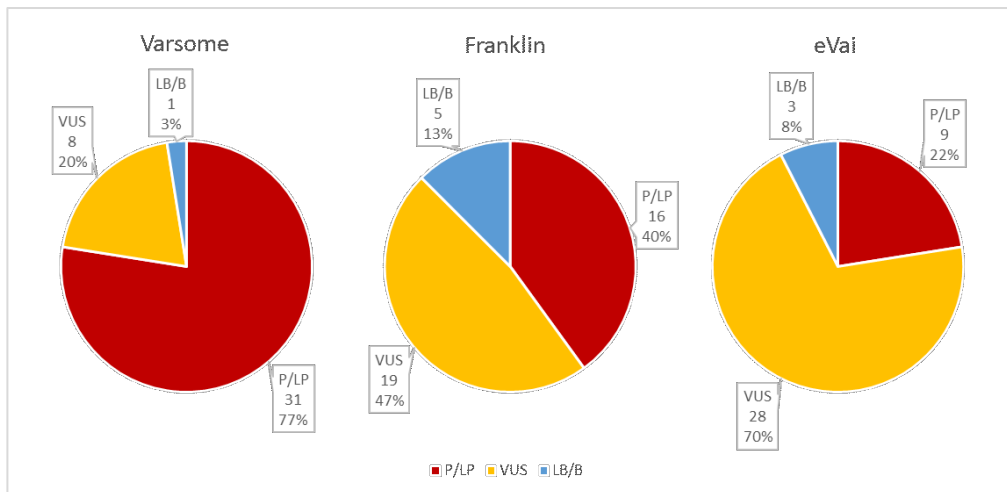
- Criteria about variant type and location (PVS1, PM1, PM4, PP2, BP1, BP7).
- Criteria about variant MAF (PS4, PM2, BA1, BS1).
- Criteria about functional studies (PS3, BS3).
- Criteria about residues (PS1, PM5).
- Criteria about in silico tools (PP3, BP4).
- Criteria about reputable sources (PP5, BP6).

The result was analyzed to evaluate if the three software agree in classifying the variants and to understand which criteria are activated in a different way and which can cause misclassification of the *DES* variants. Subsequently the variants were classified using the adapted criteria of the CMGCV-*DES*. Finally, by integrating the information on the phenotype of carriers (PP4, BS2, BP5), on the family study (PS2, PM6, PP3, BS4), and on the functional studies on myocardial tissue carried out at OSM, a robust classification of the variants contained in the *DES*-dataset was provided.

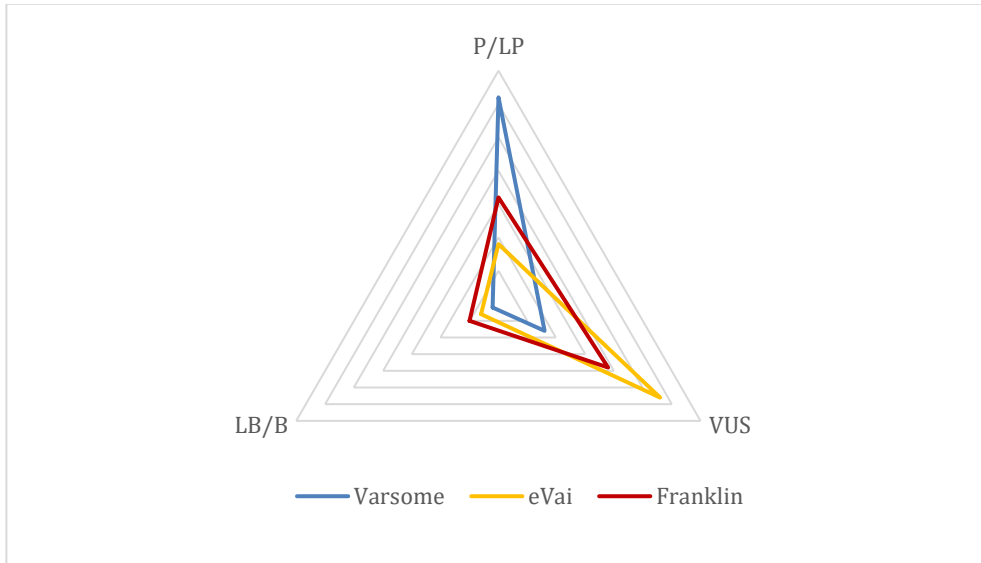
#### 4.1.4.1. Differences between software before patient and family evaluation

All the 40 short variants extracted from the DES-dataset were classified using the three software (Figure 4.3, 4.4). The CNV was excluded from the evaluation because not all software provide interpretation of this type of variant by applying the 28 criteria-based ACMG system, rather preferring to apply the rules dedicated to structural variants [114].

- Varsome classifies 31 variants as P / LP (77.5%), 8 as VUS (20%), 1 as LB / B (2.5%);
- Franklin classifies 16 variants as P / LP (40%), 19 as VUS (47.5%), 5 as LB / B (12.5%);
- eVai classifies 9 variants as P / LP (22.5%), 28 as VUS (70%), 3 as LB / B (7.5%).

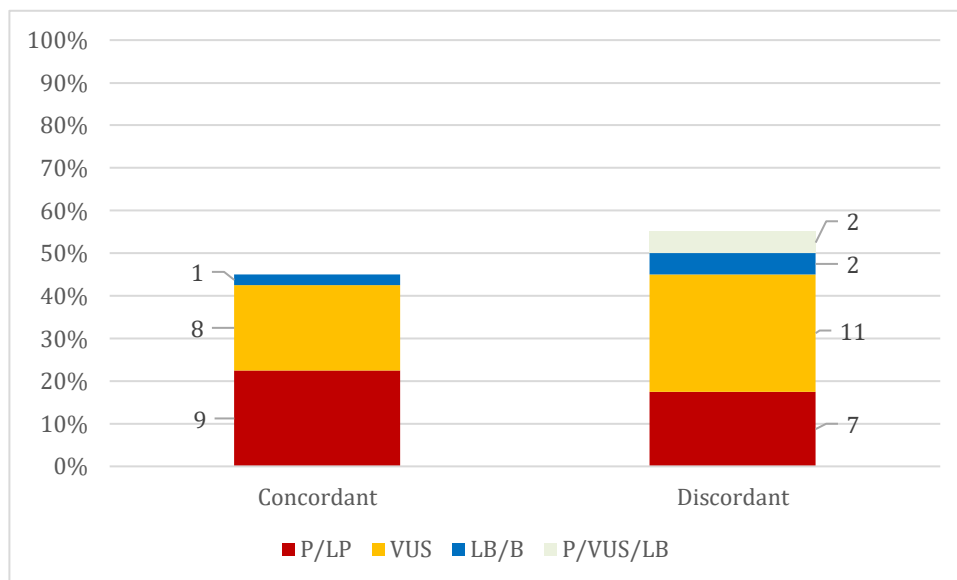


**Figure 4.3:** Distribution of variant classification of the three software



**Figure 4.4:** Trends in the classification of the three software

The 3 software agree in classifying 18/40 (45%) variants (9 as P / LP, 8 as VUS, and 1 as LB / B), while for 22 (55%) variants the classification is discordant. As the 22 variants discordantly classified are concerned, for 20/40 variants (50%) 2 out of 3 software agree in the classification and classify by majority 7/20 as P / LP, 11/20 as VUS, and 2/20 as LB / B. The interpretation from 3 software are completely different only for 2 (5%) variants: Varsome classify them as P / LP, eVai as VUS, and Franklin as LB / B (Figure 4.5).



**Figure 4.5:** Concordant and discordant classification between the three software. Variants without the full agreement between software are classified using the 2 out of 3 rule.

These classification differences are due to the different strategies with which the software activates the ACMG rules criteria.

All software agrees to use the criteria concerning the type of variant (PVS1 or PP2) with the only difference that Franklin applies the recommendations for the use of PVS1 by modulating its strength based on the prediction of the nonsense-mediated decay as per recommendations [101].

The other criteria activated in a discordant way between the software are:

1. The criterion that evaluates the prediction tools (PP3 / BP4).
2. The criteria that evaluate the MAF thresholds to determine if a variant is to be considered rare or common (PM2 / BA1 / BS1).
3. The criterion that evaluates the literature and clinical database data (PP5 / BP6).

The different use of PP3 or BP4 is due to the different sets of prediction tools that each software queries and to the rule with which they activate the score. Specifically, eVai activates PP3 and BP4 independently and, in some cases, simultaneously if even a single tool is in favor of the deleterious or benign effect. Varsome uses a majority ranking among the results of all tools to determine whether to activate PP3 or BP4 and activate them exclusively. Franklin calculates a machine learning-based meta-score of the predictions of the interrogated tools and interprets the impact through ranges of benignity or deleterious effect.

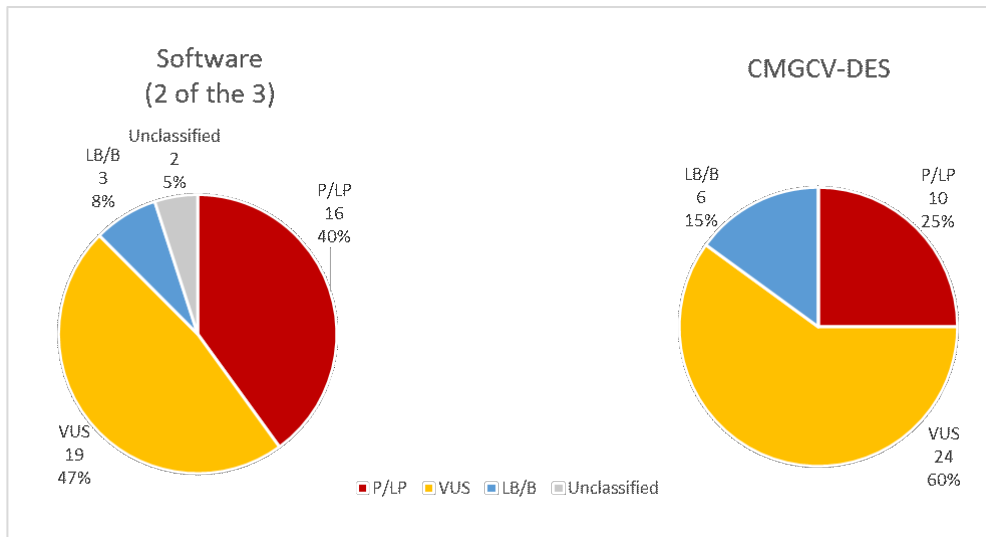
In interpreting the MAF calculated from the control population, the software agrees to classify 17/22 variants as rare and activate PM2, while calculating the threshold for BS1 differently. Franklin, in fact, activates the BS1 for 3 variants, determining the LB / B class, while Varsome and eVai do not activate the BS1 for any variant. The rule that most causes the classification differences between the 3 software is the PM1 which in 21/22 variants switches the classification from VUS to LP. The software that most applies the PM1 criterion is Varsome: it classifies the 21 LP variants using the PM1 compared to 7/21 classified as P / LP by Franklin, and 0/21 eVai. EVai does not activate the PM1 criterion for any variant, while Franklin only activates PM1 for 6/21 variants.

### **4.1.4.2. The CMGCV interpretation**

By applying the CMGCV-DES system, the 40 variants of the DES-dataset were classified as 10 LP-P (25%), 24 VUS (60%), and 6 LB-B (15%), while the three software, considering at least 2 out of 3 software, classified the variants as 16 LP-P (40%), 19 VUS (47.5%), 3 LB-B (7.5%) and 2 LP-VUS-LB (5%) (Figure 4.6). Through the CMGCV-DES system, the CNV DEL exon 3, excluded from the evaluation of the 3 software, was classified as LP.

Of the 38 variants with a concordant class calculated by 2 out of 3 software, the CMGCV-DES system classifies 31/38 (81.6%) in agreement

with software: 10 LP-P, 18 VUS, and 3 LB-B. Of the 7/38 (18.4%) variants classified differently, 6 were classified LP-P by the software and VUS by the CMGCV-DES, while 1 is VUS for software and LB for CMGCV-DES. The differences in classification are due to the difference in the interpretation of critical areas (hotspots or functional domains) within the Desmin gene (PM1), to the interpretation of the MAF of the variant (PM2 / BS1), and to the criterion that evaluates a second variant which afflicts the same amino acid (PM5).



**Figure 4.6:** Distribution of variant classification of the three software (using the 2 out of 3 rule) and of the CMGCV-DES system.

#### 4.1.4.3. The impact of the pathological and clinical study of subjects and the analysis of families on classification

The final classification was optimized using clinical information of variant carriers, familiar co-segregation information, and functional tissue pathology studies where available. The immuno-ultrastructural pathological study for the identification of Desmin aggregates in myocardial tissue was performed on the carriers of 9 variants, 7 of which classified as LP and 2 as VUS by the CMGCV-DES system. The carriers of 8 of the 9 tested variants demonstrated specific accumulations due to the Desmin defect, while in the carrier of 1 variant no pathological findings of intra-myocyte myofibrillar aggregates were found.

Clinical evaluation of patients revealed the compound phenotype RCM + AVB (and serum PK+) in probands carrying 10/41 variants, and none of these demonstrated the presence of a second variant that can be considered an alternative cause of the phenotype. Carriers with DCM, HCM, and ARVC (but also carriers without cardiomyopathy) were identified for 14/41 variants; in subjects with cardiomyopathy carriers of 5/14 variants, a second variant classified as P or LP was identified in a disease, probably the

principal cause of the phenotype. Finally, 17/41 variants were identified in control subjects with clinically proven absence of the phenotype, and clinical and family history negative for cardiomyopathy and myopathy.

The family study identified 4/41 de-novo variants; for 1/41 variant recessive transmission was demonstrated in two unrelated families with homozygous subjects affected by RCM and AVB and healthy heterozygous subjects; for 2/41 variants segregation with the phenotype with the dominant transmission was observed, while for the remaining variant carriers the families proved to be not informative or not accessible. By adding this information, the pathogenicity class changes for 16/41 variants. For 8 variants the pathogenic interpretation was strengthened passing from LP to P, 2 variants were confirmed benign passing from LB to B, while from 6 variants they changed class passing from VUS to P (n = 1) and from VUS to LB (n = 5).

### 4.1.5. The final classification

Considering the above-presented data, the 41 variants can be divided into 4 groups:

1. The first group comprises variants whose carriers are all affected by RCM with AVB, myopathy, and Desmin accumulations. Some of these have led to end-stage disease and heart transplantation. This group is composed of 9 variants, 4 null variants and 5 missenses (1 of which replaces 2 amino acids) distributed in the Head (n = 1), in the coil1B (n = 4), Coil2B (n = 1), and in the Tail (n = 3) of the Desmin protein; all 9 variants are absent from the population databases but 4/9 are described in ClinVar as P or LP and 1/9 are described as P and VUS. In subjects carrying 8/9 variants of this group, the immuno-ultrastructural investigation was performed which revealed the presence of Desmin accumulations inside the cardiomyocytes demonstrating an "aggregate-forming" effect of these variants. The final interpretation of the variants of the first group is Pathogenic for MFM Desmin related.

2. The second group includes variants identified in subjects with cardiomyopathy related to the Desmin defect (HCM, DCM, and ARVC) but non-specific for Desminopathy and who do not have a second variant that could be the main cause of the phenotype. The second group is composed of 9 variants, 6 missenses, 2 truncating variants, and 1 variant of non-canonical splice site, distributed in the Head (n = 3), in Coil1A (n = 1), in Coil1B (n = 3), and in the Tail (n = 2) of the Desmin protein. The EM study carried out in the myocardial tissue of the carrier of 1 variant (c.323A> G; p.Glu108Gly) did not show the presence of Desmin accumulations excluding an aggregate-forming effect of the variant. Of this group 6/9 are variants absent from population databases, 1/9 is described in ClinVar as P, 1/9 as LP and VUS, and 1/9 as LB and VUS. The final classification determines the LP-P class for 2/9 variants, 6/9 variants are classified as VUS and 1/9 as LB. Variants in this group do not appear to have an aggregate-forming impact but may play a causative role for other DES-associated cardiomyopathies.



3. The third group is composed of 6 missense variants identified in subjects with cardiomyopathy associated with a Desmin defect (HCM, DCM, ARVC, RCM) but who have a variant in another gene that is likely to cause the cardiac phenotype. The variants of the third group are distributed in the Head (n = 2), in the Coil 1B (n = 1), in the Coil 2B (n = 1), and in the Tail (n = 2) of the Desmin. 5/6 variants are present in the population databases with a number of alleles  $\geq 4$ , while 4/5 are described in ClinVar as B and VUS (n = 2) and VUS (n = 2). The clinical study shows both carriers affected by cardiomyopathy (6 of these are mutated in other phenotype-related genes) and healthy subjects from a cardiac point of view. In this group, 5/6 variants are classified LB or B and 1/6 is VUS. The variants of the third group are likely to have no impact on the MFM phenotype and do not appear to cause cardiomyopathy either.

4. The fourth group includes 17 variants identified in control subjects without CMP and myopathy. This group is constituted of 16 missense variants and 1 non-canonical splice site variant (c.579-4C> T) distributed in the Head (n = 2), in Coil 1A (n = 3), in Coil 1B (n = 7), in Coil 2B (n = 3), and in Tail (n = 2). Out of 17 variants, 5/17 are not represented in the control populations and 7/17 are described in ClinVar as VUS (n = 5) and VUS and LB (n = 3). All carriers of the variants contained in this group were considered healthy controls without cardiomyopathy. Of this group, 5/17 variants meet the criteria to be classified as LB or B, and 12/17 remain VUS. Although the interpretation tends towards kindness. The variants of the fourth group are most likely benign and have a neutral role towards Desminopathy.

**Table 4.3:** DES-dataset variants classification

VARIANT	EVAI	VAR SOME	FRANKLIN	CMGCV-DES	FINAL CLASS
AGGREGATE FORMING VARIANTS					
c.46C>T p.Arg16Cys	VUS	LP	LP	LP	LP → P PM2,PP2,PP3,PP5 + PS3_Verystrong, PM3_Supporting,PP4
c.536_551del p.Glu179fs	P	P	P	LP	LP → P PVS1,PM2 + PP4
c.641_735+1del p.Asp214_Glu245del	-	-	-	LP	LP → P PVS1,PM2 + PS3_StandAlone,PP4
c.735+2_735+11del -	P	P	LP	LP	LP → P PVS1,PP5,PP3 + PS3_StandAlone, PP4
c.735+1G>A -11111111\	P	P	P	LP	LP → P PVS1,PM2,PP3,PP5_S trong + PS3_StandAlone, PP4

## Clinical Applications

c.1216C>T p.Arg406Trp	LP	P	P	LP	LP → P PM2,PP2,PP3,PP5_Strong + PS3_StandAlone, PP4
c.1358C>T p.Thr453Ile	VUS	LP	LP	LP	LP → P PM2,PP2,PP3,PP5 + PS3_StandAlone, PP4
c.1360C>T p.Arg454Trp	LP	P	P	LP	LP → P PM2,PP2,PP3,PP5_Strong + PS3_StandAlone, PP4
c.1398_1399delGCinsTT p.GlnHis466HisTyr	VUS	VUS	VUS	VUS	VUS → P PM2,PP2 + PS3_StandAlone, PP4
VARIANTS IN CMPs WITHOUT OTHER MUTATIONS					
c.250G>A p.Gly84Ser	LB	VUS	LB	LB	LB → LB PP2,BS1,BP6 + PP4
c.322G>A p.Glu108Lys	VUS	LP	LP	VUS	VUS PM2,PP2,PP3
c.323A>G p.Glu108Gly	VUS	LP	LP	VUS	VUS PM2,PP2,PP3
c.380G>C p.Arg127Pro	VUS	LP	LP	LP	LP PM2,PP2,PP3,PP5
c.517C>A p.Arg173Ser	VUS	LP	VUS	VUS	VUS PM2,PP2,PP3
c.634C>T p.Arg212Ter	P	P	P	VUS	VUS PVS1,PP5
c.749T>C p.Leu250Ser	VUS	LP	VUS	VUS	VUS PM2,PP2,PP3
c.1289-3C>T -	VUS	VUS	VUS	VUS	VUS PM2
c.1371+2T>C -	P	P	LP	P	P PVS1,PP5,PP3
VARIANTS IN CMPs WITH OTHER MUTATIONS					
c.170C>T p.Ser57Leu	LB	P	LB	LB	LB PP2,BS1,BP6
c.238C>T p.Pro80Ser	VUS	VUS	VUS	VUS	VUS → LB PM2,PP2,BP4 + BP5
c.635G>A p.Arg212Gln	VUS	LP	VUS	LB	LB → B PP2,PP3,BS1,BP6 + BS2,BP5
c.1123C>T p.Arg375Trp	VUS	LP	LP	VUS	VUS → LB PP2,PP3 + BS2,BP5
c.1286G>A p.Arg429Gln	VUS	LP	VUS	VUS	VUS → LB PP2 + BS2,BP5
c.1334C>T p.Thr445Met	VUS	LP	VUS	VUS	VUS → VUS PP2
VARIANTS IN CONTROLS					

c.173G>A p.Arg58His	VUS	VUS	VUS	VUS	VUS → VUS PM2,PP2 + BS2
c.216C>A p.Ser72Arg	VUS	VUS	VUS	VUS	VUS → LB PP2,BS1 + BS2
c.404C>T p.Ala135Val	LB	LP	LB	LB	LB → B PP2,BS1,BP4,BP6 + BS2
c.407T>A p.Leu136His	VUS	LP	LP	VUS	VUS → VUS PP2,PP3 + BS2
c.415G>C p.Glu139Gln	VUS	LP	VUS	VUS	VUS → VUS PP2,PP3 + BS2
c.460C>A p.Leu154Ile	VUS	VUS	VUS	VUS	VUS → VUS PP2 + BS2
c.538C>T p.Arg180Cys	VUS	LP	VUS	VUS	VUS → VUS PM2,PP2,PP3 + BS2
c.543C>G p.Asp181Glu	VUS	LP	VUS	VUS	VUS → VUS PM2,PP2,PP3 + BS2
c.546C>G p.Asn182Lys	VUS	LP	VUS	VUS	VUS → VUS PM2,PP2,PP3 + BS2
c.577A>C p.Lys193Gln	VUS	LP	VUS	VUS	VUS → VUS PM2,PP2,PP3 + BS2
c.579-4C>T -	VUS	VUS	VUS	VUS	VUS → LB BP4 + BS2
c.643G>T p.Val215Leu	VUS	LP	VUS	VUS	VUS → VUS PP2,PP3 + BS2
c.935A>C p.Asp312Ala	VUS	LP	LB	B	B → B PP2,PP3,BA1,BP6 + BS2
c.1064G>A p.Arg355Gln	LP	LP	LP	VUS	VUS → VUS PP2 + BS2
c.1180G>A p.Val394Met	VUS	LP	B	B	B → B PP2,PP3,BA1,BP6 + BS2
c.1189G>A p.Ala397Thr	VUS	LP	VUS	VUS	VUS → VUS PP2,PP3 + BS2
c.1361G>A p.Arg454Gln	LP	LP	LP	VUS	VUS → VUS PP2 + BS2

#### 4.1.6. The importance of clinical and pathology studies

The difficulties related to understanding the impact of Desmin-gene variants, and consequently on the clinical path of carriers, make DES a key example of the need to deepen the methods of interpretation of genetic tests. The presence of commercial or free-to-use software for the prioritization of variants has facilitated the genetic diagnosis process by implementing the ACMG rules and making easier the collection and interpretation of the necessary information. Despite the concrete help that derives from the use of

these systems, the implementation of the criteria for the pathogenic classification is still not very robust, especially using partial information. The analysis of the results of the three software has shown that the classification of pathogenicity of the variants, provided only by the information available *in silico*, is strongly influenced by the algorithms with which the rules are interpreted. Each software uses different logics for the activation of each criterion and small differences in implementation can generate important classification discrepancies, giving rise to more or less conservative or unbalanced interpretations. Depending on the software used, there is a risk of over-interpreting the variants, altering the calculation of the predisposition to develop the disease, with a deleterious impact both for the patient and for the healthcare system. On the other hand, a too conservative classification that interprets as VUS variants that hardly have a pathogenic effect, leaves the genetic report pending due to lack of certainties. These problems give rise to the need to integrate as much knowledge as possible about the clinical case, its genetics, its family, and dedicated functional experiments into the interpretative process. The interpretative problem affects almost all genes but is crucial for those associated with rare diseases characterized by a lack of genetic, clinical, and segregation information. For this reason, it becomes essential to adapt the algorithms according to the genes being analyzed to maximize accuracy. The CMGCV-DES system avoids overestimating the impact of the variants and increases the number of likely benign or benign classification, preferring a more conservative interpretation in the absence of functional or segregation tests to confirm the pathogenic assessment. Although the ACMG rules help to make the interpretation of DES variants more robust, only clinical and pathology can shed light on the real role of the variant on the phenotype. It is, in fact, essential to be able to understand which variants actually have an "aggregate-forming" effect causing the intra-myocytes accumulation of Desmin, and which ones do not have this role despite being classified as pathogenic.

### **4.1.7. Clinical features of variant's groups**

As a conclusion of the evaluations on the variants present in the subjects belonging to our center, we decided to organize the series of variants according to 4 distinct groups based on the characteristics of the patients and the variants themselves.

All the variants of which we are certain of pathogenicity and of the role of the aggregate forming cause of myofibrillar myopathy belong to the first group. The clinical picture of carriers of these variants has a very similar evolution, which begins with a delay in conduction and slight concentric hypertrophy of the ventricles and evolves into RCM and AVB. All variant subjects in this group underwent PM implantation and many of them underwent cardiac transplantation. Evidence of the intra-myocyte Desmin aggregates labeled with anti Desmin antibodies during the EM study excludes any doubts about the effect of the variants of this group. While the

aggregate forming variants are relatively simple to identify by deepening the clinical aspects of the patient and through robust and well-performed functional studies, the variants that do not generate intra-myocyte accumulations of Desmin but that decrease its functionality and cause nonspecific phenotypes, are still a challenge. The definition of the cause-and-effect relationship between the genetic defect in DES and myocardial pathology in subjects who do not have a second mutation capable of explaining the phenotype is essential in the path of understanding the disease, but it is still difficult to evaluate due to the lack of markers, pathological and functional tests. The variants of the second group within our series are an example of this. Carriers in fact demonstrate a range of cardiomyopathies, mostly sporadic, which includes DCM, HCM, ARVC, conduction dysfunctions, and extensive myocardial fibrosis; moreover, they are not affected by other pathogenic variants and the families are not very informative due to the lack of phenotypic transmission. The effect of this uncertainty is reflected in the final classification of these variants, which are mostly referred to as VUS precisely because of the lack of crucial evidence of a damaging effect. The lack of functional data does not allow to assess the aggregate forming role, and the clinical study on the proband and on his family does not clarify the ideas on the pathogenic impact. The meaning changes for the variants of the last two groups in patients with cardiomyopathy caused by a mutation in another gene strongly related to the phenotype or in control subjects who do not have cardiomyopathy and who come from phenotypically healthy families for hereditary cardiomyopathy. The CMGCV-DES system with the addition of information on the patient and families, classifies these variants as (likely) Benign or as VUS tending to Benign. This group includes variants that affect an amino acid that also changes into a pathogenic variant. Patients' clinics orient the assessment towards a likely neutral impact, confirming the CMGCV-DES interpretation.

## **4.2. Variants study in breast and ovarian cancer families**

### **4.2.1. Introduction to hereditary cancer**

Since 2013 the Centre for Genetic Diseases of the OSM Foundation, in collaboration with the Breast Cancer and Gynecology and Obstetrics units of the Hospital, has developed a path of clinical and molecular genetics to support and enhance the diagnosis and care of women at high risk of developing breast and ovarian cancer. In the same year, a research project was launched aimed at identifying the genetic causes of breast and ovarian cancer and HBOC syndrome. These paths were provided to over a thousand women with Breast and Ovarian Cancer (BROVCA) who received highly specialized multidisciplinary care aimed at the most advanced management (diagnosis and treatment) of malignancies (OSM PDTA EUSOMA; EU certification with annual confirmation). Within this process, an integrated clinical and genetic database of patients suffering not only from BROVCA, but also from non-BROVCA hereditary cancers, was created in order to incorporate the data in an easy and fast management, the data analysis and interpretation.

This chapter describes the database developed at the CMGCV of San Matteo and the results obtained from the analysis of genetic data conducted with NGS analyzes from 2015 to 2020. The primary objective is to understand the genetic makeup of patients with BROVCA compared to patients with other malignant neoplasms; the secondary objective is the assessment of the feasibility of family segregation studies.

### **4.2.2. Genetic and clinical background**

Breast cancer (BR) is the most common cancer in women. The World Health Organization (WHO) has estimated that it accounts for more than 25% of all new cancer cases per year in women and 10% of all cancers when men are included in the estimates [115]. Ovarian cancer (OV) is less common than breast cancer. The latest estimate in the "Global Cancer Statistics 2020" report [116], generated by the American Cancer Society (ACS) and the International Agency for Research on Cancer (IARC), shows that ovarian cancer accounts for approximately 3.4% of all female cancers, globally. Most BR and OV appears as sporadic without an obvious genetic etiology. A smaller proportion - between 5 and 15% of BR cancers and 6-25% of OV cancers - is linked to a strong hereditary "predisposition". The genetic causes of these two types of cancer overlap, and both Breast and Ovarian cancers (BROVCA) are often observed in family members, carriers of the same genetic defect. This familial predisposition syndrome to develop BROVCA cancers is called hereditary breast and ovarian cancer (HBOC).

Although most cases of HBOC syndrome are associated with mutations in BRCA1 or BRCA2 genes, defects in these two genes explain about 15-25% of cases. The study of the HBOC BRCA negative families led to the identification of additional 25 genes associated with hereditary predisposition to BROVCA cancers; still far from a completion of the predisposing genes, and in the context of this high genetic heterogeneity, further studies are ongoing to identify all the possible disease-genes thus providing precise molecular diagnostics with complete lists of the genetic causes of HBOC syndrome [115][117].

#### **4.2.3. The reasons for genetic testing**

Genetic testing for HBOC syndrome is integrally part of the management of patients who develop BROVCA cancers and their families. The most common test is limited to BRCA1 and BRCA2 because most familial BROVCA are associated with mutations in these 2 genes. However, the new discoveries on the genetic causes of non-BRCA BROVCA cancers demonstrated the clinical relevance to other genes as well. The result of the genetic investigation is clinically actionable as it has an immediate impact on surgery, oncology treatment, clinical management of family members, as well as on the quality of life of unaffected carriers. Immediately after the diagnosis of cancer, the detection of pathogenic or likely pathogenic variants in a relevant gene can support surgical decision, from a conservative quadrantectomy to a total mastectomy and, in some cases, to a preventive bilateral mastectomy [118]. Prophylactic bilateral salpingo-oophorectomy is recommended in genetically predisposed and aged individuals at risk and can reduce the risk of ovarian cancer by up to 80% [118]. Cancers related to BRCA1 or BRCA2 defects are commonly susceptible to carboplatin which is considered the first line treatment for genetic cancers. In addition, some ovarian tumors, and recently also BR cancers, associated with defects in the homologous recombination pathway (HRD) are targets of a new line of PARP inhibitors that has shown promising effects [119][120]. Finally, probands and family carriers exposed to the risk of HBOC syndrome enter personalized prevention monitoring programs to ensure early diagnosis and increase the probability of survival and quality of life.

#### **4.2.4. The clinical and molecular genetic path at the OSM**

The path is structured as indicated by the Regione Lombardia rules and by guidelines from scientific societies. The first step is genetic counselling, with examination of individual clinical data and of clinical data from relatives (patients are asked to trace and collect clinical records of relatives before accessing the center for counselling). In cases in which the genetic test is appropriated, informed consent for testing is collected and then the patient undergoes blood sampling. The blood sample is transferred to the

laboratory where it is processed for NGS tests. After running the NGS, data are analyzed as described in the prior chapters. The file including all variants selected by the analysis is addressed to the lab for Sanger confirmation, as per request of the Regione Lombardia rules. After completion of the confirmation, a report describing the results is generated according to the rules indicated by scientific societies. In parallel, family segregation studies are performed, in particular in case of variants that are defined as VUS based on ACMG rules due to the lack of any prior description. Reports on known and proven pathogenic variants are usually released immediately after Sanger confirmation. The center acknowledges ACMG criteria for variant reclassification and ACMG indications for recalling patients when reclassification provides new evidence of variant actionability. Each patient then receives three written reports: the pre-test genetic counselling, the genetic test report and the post-test counselling report. The reports and the related information are directly transferred to the patients during the post-test counselling. Each patient signs a further form in which she/he declares the receipt of the reports and the full understanding of the information provided during post-test counselling.

### **4.2.5. NGS sequencing and analysis pipeline**

The workflow of NGS-based germline genetic analysis of samples was validated within the accreditation process of the European Society of Breast Cancer Specialists (EUSOMA) [121] for the genetic diagnosis of hereditary breast cancers.

#### **4.2.5.1. Wet process**

By protocol, DNA is isolated from whole blood by the Promega Maxwell® RSC automatic extractor and quantified by NanoDrop™.

NGS libraries are prepared using the Illumina Trusight Rapid Capture kit in combination with the Trusight Cancer (illuminate) probes. The libraries undergo a process of quantification, quality control and selection of fragments using the BioAnalyzer (Agilent), before being loaded onto the Illumina MiSeq sequencer in groups of 24 samples per run, as per the Illumina protocol.

#### **4.2.5.2. NGS data analysis**

Fastq files are analyzed via the `trusight_germline` pipeline implemented using Helper platform. The workflow of the `trusight_germline` pipeline includes:

1. Fastq QC using FastQC tool.
2. Alignment of Fastqs using BWA-mem.



3. Sam conversion, bam sorting, and marking of duplicates using Picard tool.
4. Realignment around InDels and Base Quality Score Recalibration using GATK v.3.7.
5. Joint Variant calling using GATK v.4.1 and Freebayes in cohort modality.
6. Variant filtering using GATK v.4.1.
7. Annotation using VEP.
8. Transcript selection using Canonical Transcript for all genes.
9. CNV calling using GATK v4.1.

The NGS samples are evaluated through target coverage quality parameters and the identified variants undergo a prioritization process through a cascade of sequential filters.

Variants with the following characteristics are excluded:

1. Heterozygous in more than two samples or homozygous in more than one sample within the cohort consisting of 24 samples.
2. Intronic, UTR, intergenic and non-coding variants distant from the splice sites.
3. Variants with MAF greater than 0.005 in at least one population database including 1000G, ExAC, and GnomAD.
4. Benign or Likely Benign, ascertained.
5. Synonymous or missense variants in genes with a high mutation frequency and with an established loss of function as mechanism of damage.

The remaining variants were classified according to the ACMG-AMP guidelines and only variants with a Likely Pathogenic and Pathogenic significance are considered.

#### **4.2.5.3. The CMG-Cancer DB**

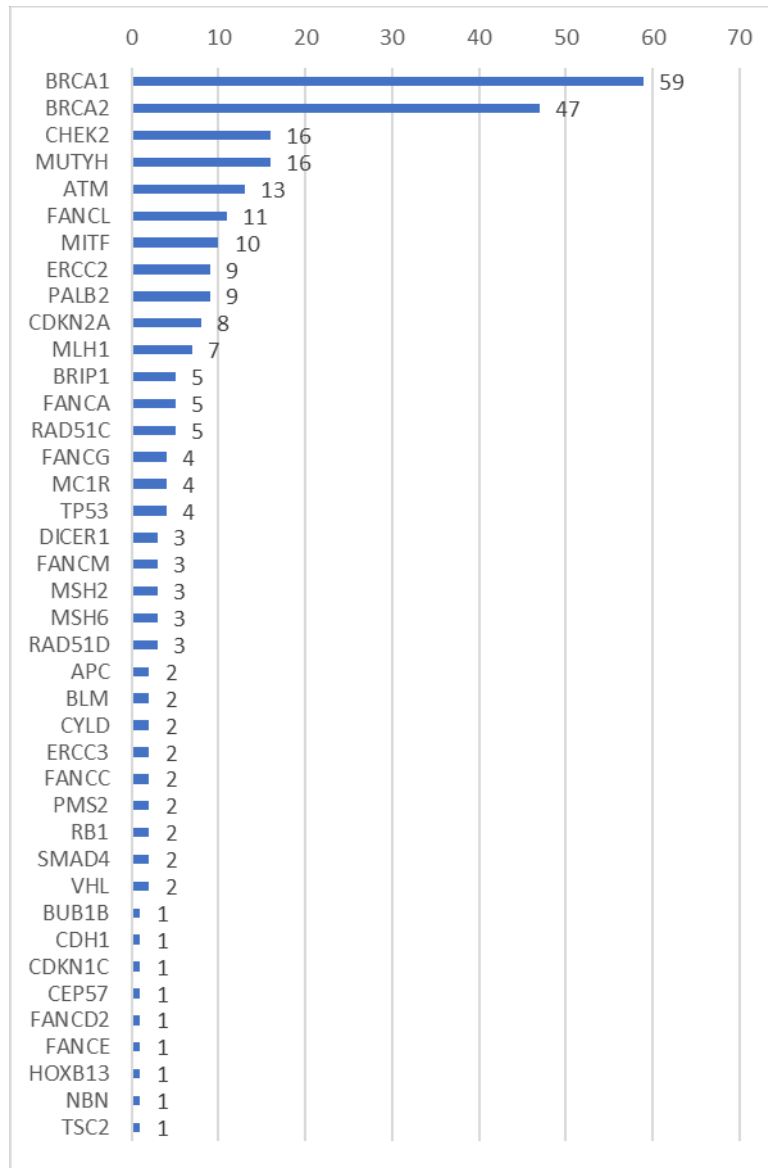
For the present study, the clinical and family information on patients, obtained at the pre-test genetic counseling, and the variants identified through the NGS analysis, were included in the CMG-CancerDB. Clinical data such as the site of the tumor, the age of onset, the characteristics of malignancy and information on family history have been entered manually and are subjected to a continuous review process. The genetic information of the probands obtained through NGS analysis was integrated with an automatic system and was updated periodically.

Cancer DB contains information from 1320 unrelated probands, addressed to the CMG for cancer (n = 1225; 92.8%) and eligible for genetic testing, as well as unaffected subjects (n = 95; 7.2%) who underwent genetic testing for positive family history suggestive for Hereditary Cancer syndrome (HCS). The series analyzed in this study includes 253 males and 1067 females, with a mean age of 55 years at the date of the consultation (C.I. 95% = 32-77

years) (Table). Based on tumor site and characteristics, probands with prior malignancy were grouped into probands with BROVCA (p-BO) (n = 825; 62.5%) and probands with other non-BROVCA tumors (p-NBO). The latter, together with the unaffected probands, were considered as a single group without prior BROVCA (p-NBO) (n = 495; 37.5%). For the p-NBO group, family history was assessed and subjects with at least one BROVCA relative in the first (parents, siblings, children) generation or in the second generation (uncles, grandparents, grandchildren) were grouped as probands no BROVCA in BROVCA family (p-NBO in f-BO) (n = 173; 34.9% of p-NBO). The remaining patients without prior BROVCA in the family were grouped together as probands no BROVCA in no BROVCA family (p-NBO in f-NBO) (n = 322; 65.1% of p-NBO). To carry out the classification, only subjects with primary breast or ovarian cancer were considered as BROVCA and metastatic lesions were excluded. Non-HBOC tumors such as Sertoli-Leydig cell tumor (non-germ cell neoplasm of the ovary and testis) were considered among the non-BROVCA tumors due to the unique characteristics and well-defined genetic causes (e.g., Diceropathies).

### **4.2.6. Results of genetic testing**

Although the genes reported as associated with increased risk of BROCA are a small group, all genes present in the Trusight Cancer panel were considered for this study. Of the 1320 tested probands, 274 (20.8%) carried a P or LP variant in a gene associated with HCS Syndrome, with a slightly higher percentage of mutants among the p-BO probands (n = 182 / 825; 22.1%) compared with p-NBOs (n = 92/495; 18.6%).



**Figure 4.7:** The distribution of genes within the entire population of CMG-CancerDB.

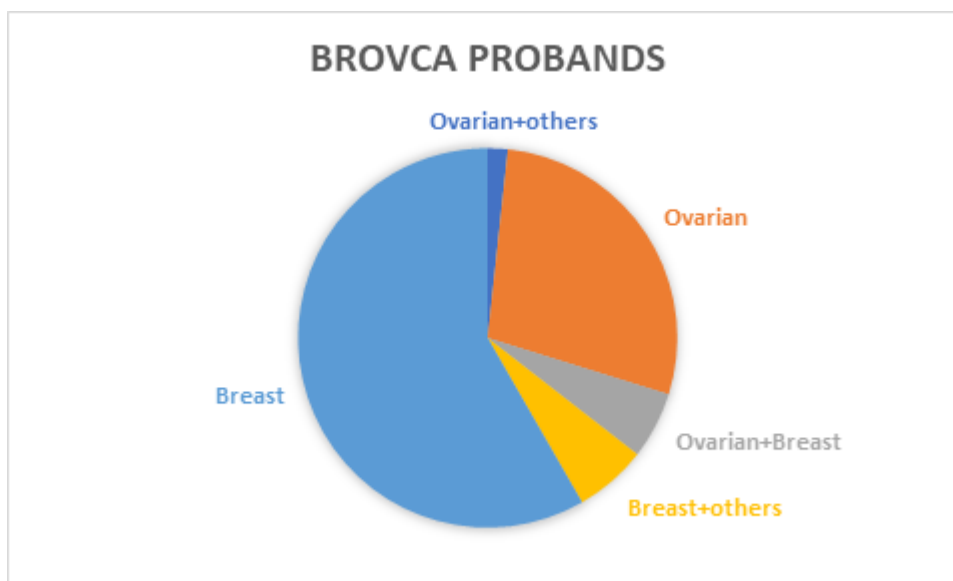
The list of genes with at least one P/LP variant includes 40 genes with at least one P or LP variant (Figure 4.7). Most common genes are BRCA1 ( $n = 59$ ) and BRCA2 ( $n = 47$ ) which, by themselves, comprise 38.7% of the causative disease variants in the series.

Among the BROVCA probands, 91/180 (50.5%) carry a P / LP variant in the BRCA1 or BRCA2 genes and 89/180 (49.5%) in other genes. The percentage decreases in the non-BROVCA group of probands ( $n = 15/93$ ; 16.3%); 11/15 (73.3%) were identified in unaffected probands from high-risk families.

#### 4.2.6.1. BRCA vs other genes in BROVCA probands

For a more in-depth analysis of BRCA1 and BRCA2 genes vs. other genes in probands with previous breast or ovarian cancer, the 825 BROVCA probands were divided into 5 subgroups based on the location of the malignancies:

- ovarian cancer alone (n = 232/825, 28.1%)
- ovarian cancer plus other types of cancer (n = 14/825, 1.7%)
- breast cancer alone (n = 481/825, 58.3%)
- breast cancer plus other types of cancer (n = 51/825, 6.2%)
- breast + ovarian cancer (n = 47/825, 5.7%).



**Figure 4.8:** The distribution of malignancy types in BROVCA probands

The highest percentage of mutated probands (35.7%) was found in the group of patients with previous ovarian cancer and at least one more non-BROVCA tumor. In this group, 40% of the mutants are carriers of a P / LP variant in BRCA1 or BRCA2, while 60% carry pathogenic variants in other genes. The representativeness of the Ovarian + other tumors group is limited by the low number of probands. The probands affected by Ovarian cancer alone are 232: 37 (50.7% of the mutants) are mutated in BRCA1 / 2 and 36 (49.3%) in other genes, with overall 73/232 mutated (31.5%). Considering all probands who had Ovarian but no Breast cancer (n = 246/825, 29.8%), 78/246 (31.7%) were mutated in BRCA1/2 (n=39/78, 50%) and in other genes (n=39/78, 50%).

In 481 (90.4%) of the 532 probands, breast cancer was the only malignancy, while the remaining 9.4% had had at least one further non-BROVCA malignancy. Among the probands who had only breast cancer, 77

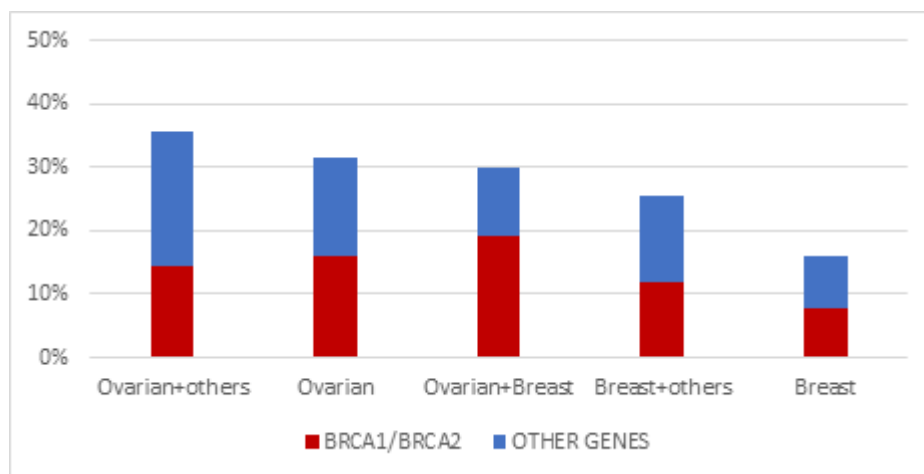
(16.0%) had a positive genetic test, and 37 (48.1%) were carriers of BRCA1/2 mutation. Of the 51 patients who had more than one malignancy, 13 carried pathogenic variants (25.5%), of which 6/13 (46.2%) in BRCA1/2, and 7/13 (53.8%) in other genes. Overall, 16.9% probands with breast cancer (with or without other tumors) carry BRCA defects. Finally, the probands who developed both breast and ovarian cancer are 47: 14 (29.8%) were carriers of mutations in at least one gene of the panel; 9/14 (64.3%) tested positive for BRCA1 or BRCA2.

Within the BROVCA probands, the groups show a similar percentage of P / LP variants in BRCA1/2 versus all other 38 genes. This trend reflects the high correlation between ovarian and breast cancer with the two BRCA genes and is best seen among patients who developed lesions in both sites. This last group, despite the low number, showed a higher frequency of variants classified as causing disease in BRCA genes (64.3%).

Overall, the diagnostic yield in BRCA for patients with ovarian cancer is about one positive out of three mutated ones, while it drops dramatically in breast cancer.

**Table 4.4:** The genetic distribution into the BROVCA group

	MUTATED			NON MUTATED
	BRCA1/BRCA2	OTHER GENES	ALL GENES	
OVARIAN+OTHERS	2 (40%)	3 (60%)	35,7%	64,3%
OVARIAN	37 (50.7%)	36 (49.3%)	31,5%	68,5%
OVARIAN+BREAST	9 (64.3%)	5 (35.7%)	29,8%	70,2%
BREAST+OTHERS	6 (46.2%)	7 (53.8%)	25,5%	74,5%
BREAST	37 (48.1%)	40 (51.9%)	16,0%	84,0%



**Figure 4.9:** The BRCA1/2 vs other genes distribution into the BROVCA group

#### **4.2.6.2. Non-BROVCA probands**

In probands without BROVCA, the mutation frequency of BRCA versus other genes was assessed by comparing probands from BROVCA families, as previously defined, and probands without family history of breast and ovarian cancer. In the cohort of non-BROVCA probands, regardless of family history, 91 subjects (18.6%) were mutated, with a higher percentage of defects in the non-BRCA gene group (n = 77; %) than in BRCA1 and BRCA2 (n = 15; %). All 15 BRCA positive probands had a positive family history of BROVCA cancer, while BRCA mutants were not identified in the group of subjects from non-BROVCA families. In the non-BROVCA group in BROVCA families, the percentage of BRCA defects is 36.8% versus the 63.2% in other genes.

This result suggests that in probands without previous ovarian or breast cancer and without any first or second-degree family member affected by BROVCA, the probability that the oncology risk in the family is associated with BRCA defect is almost nil.

#### **4.2.6.3. Mutated non affected probands**

In our series, 15 probands had no previous malignancies, but were carrier of a P or LP variant in a gene associated with HCS: 11 probands in BRCA1 (n = 6) and BRCA2 (n = 4), 1 carries both BRCA2 and MSH6 mutations, 1 in FANCG, 1 in HOXB13, and 2 in MUTYH. Among the 11 probands with a variant P or LP in BRCA1 / 2, 8 are Female and 3 are Male. The females - mean age 44.5 years (CI-0.95 = 28.2-62.5) - all have a strong family history of BROVCA; 6 of 8 have mothers with a past breast or ovary and 2 have at least 1 sister with breast or ovarian cancer. The 3 BRCA1-positive males have a strong family history of BROVCA (2 breast cancers in the mother and 1 ovarian and breast cancer of a sister). The age at counseling ranges from 22, 41, to 60 years. As for the 4 subjects mutated in non-BRCA genes, only 2 of 4 had at least one first-degree relative with a previous BROVCA.

#### **4.2.6.4. BROVCA in male probands**

Men who have developed breast cancer during their lifetime deserve a careful analysis. In our series 17 men have breast cancer, with a mean age at counselling = 65.7 years (C.I. 95% = 43.8-72.5). They constitute 1.6% of all tested subjects and 2.1% of BROVCA probands. 12 out of 17 probands had only breast cancer while 5 had malignancies in other districts as well: one breast and prostate, two had kidney and breast cancers and two had kidney, breast and colon cancers. Only 4 (23.5%) male probands were found to be

mutated in the genes contained in the panel, one in BRCA1, one in BRCA2, one in ERCC2, and one in FANCA. The two BRCA carriers developed both Breast and clear cells Renal cancer (one of them also a colon adenocarcinoma) while the other two only breast cancer.

#### 4.2.6.5. Gene pathways in breast and ovarian cancers

The evaluation of the mutation fraction of the two BRCA genes vs. all the other malignancy genes has showed a balanced relationship in BROVCA probands, with the percentage of positive BRCA decreasing in the non-BROVCA probands who were members of BROVCA families, up to zero in BRCA negative probands within the subgroup of the non-BROVCA families.

Within the group of BROVCA with mutations in non-BRCA genes, there are 38 different genes, some are associated with specific syndromic malignancies (e.g., TSC2, DICER1, RB1, MC1R), while others are related to syndromes associated with the risk of malignancy in different districts organism (e.g., TP53, CHEK2). However, it should be noted that syndromic malignancy can be suspected after genetic counseling and visit, and that when the suspect is strong, the genetic testing can be performed by sequencing the given suspected gene (e.g., Carney Complex, DICER1, or NF1, NF2, etc.). To carry out an in-depth analysis of the mutational profile of the 3 clinical groups (p-BO, p-NBO in f-BO and p-NBO in f-NBO) the genes were grouped according to the biochemical pathways.

In fact, genetic susceptibility to HBOC is caused by defects in the genes that participate in maintaining the stability of the genome, that is, DNA error identification and correct nucleotide sequence restoration. The major pathways involved in the protective mechanisms of the human genome include the homologous recombination repair (HRR), the mismatch repair (MMR), the cell checkpoint pathways (CKP), and the Fanconi anemia pathway (FA) [117]. In short, there are four major mechanisms of maintenance of the genome involving genes whose defects in one or more than one mechanism cause HBOC.

The first mechanism -the HRR- intervenes in case of double-strand damage (DSB). In DSB, the checkpoint system detects an error and promotes the removal of the ends of both strands of the damaged sequence. The HRR complex is recalled, which uses the complementary sequence of the paired chromatid as a reference and repairs the damage with a copy and paste action.

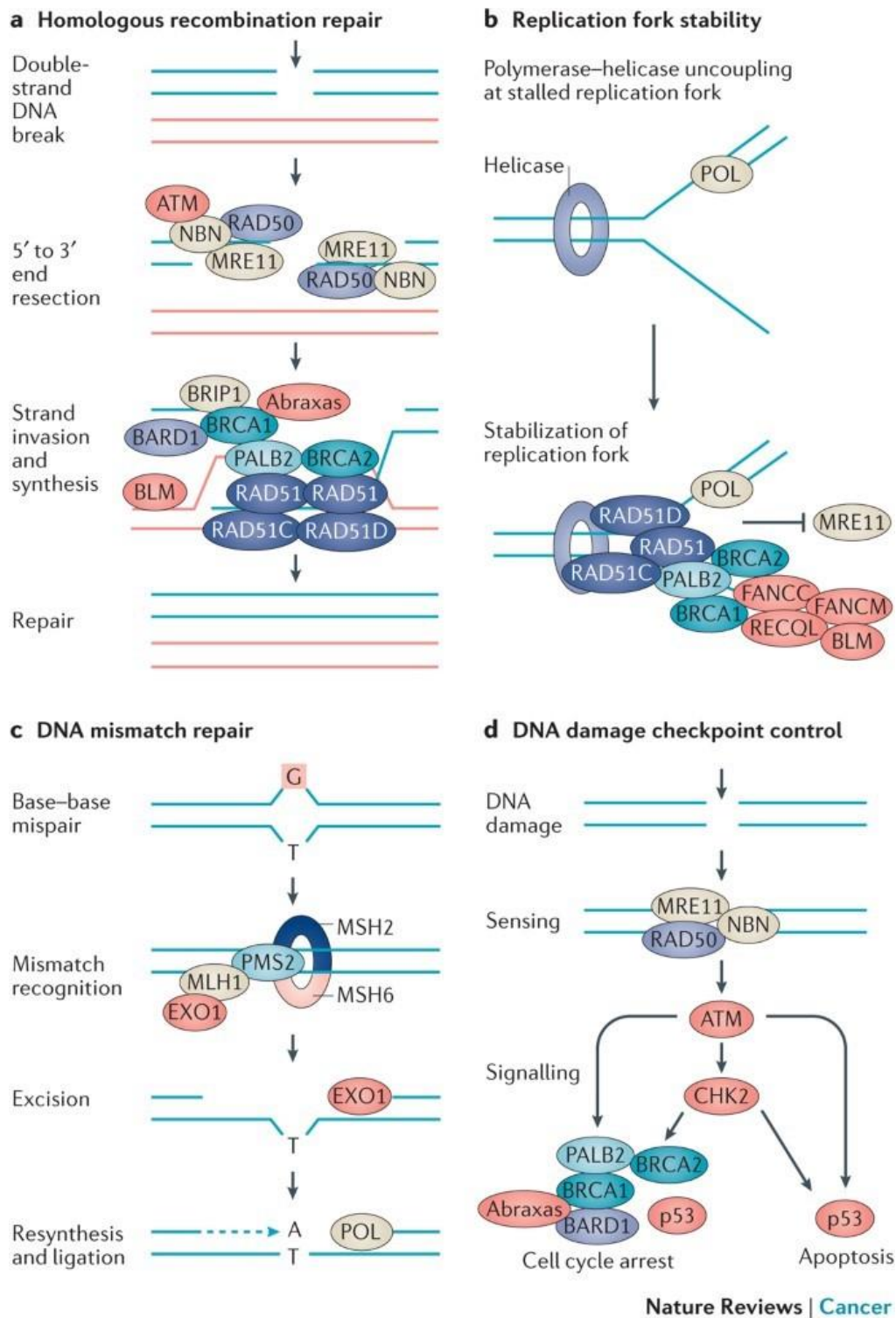
The protection system of replication fork stability limits the erroneous degradation of newly synthesized DNA sequences. The nascent DNA sequences are protected via the stabilization effect of the replication fork, which prevents them from becoming subject to the action of nucleases. Both HRR genes and Fanconi anemia complex are involved in this mechanism.

The mismatch repair mechanism corrects base-base mispairing as it recognizes and repairs erroneous insertion, deletion and misincorporation of nucleotides. In addition to monitoring the entire sequence, the MMR controls the HRR mechanism. In case of an excessive number of mismatches in the

copy process, the MMR disables the HRR, preventing the accumulation of further errors and decreasing the possibility of potentially damaging effects.

The last mechanism - the cell checkpoint pathway (CPK) - supports DNA repair pathways and includes DNA damage signaling, checkpoint control and cell death. When genetic defects affect the function of error detection in the homologous recombination process, in parallel to the recalling of the HRR complex, CPK activates the checkpoint system in cascade that pauses the progression of the cell cycle to allow DNA repair. Alternatively, the accumulation of errors in cellular DNA leads the cell to senescence and the cell death control system promotes apoptosis. When the checkpoint and senescence process are damaged, the DNA would continue to accumulate errors without the cell being induced to death and could acquire changes that promote uncontrolled proliferation.





**Figure 4.6:** Genome maintenance mechanisms (Figure from [115])

The genes involved in the HRR process present in the Trusight Cancer Illumina panel are BRCA1, BRCA2, BRIP1, PALB2, RAD51C, RAD51D, and BLM. BLM is not part of the main HRR complex, but its support action

is of great importance. These genes are major players related to HBOC and have been grouped in the HRR cluster (Table 4.5).

The genes involved in the MMR mechanism are MLH1, MSH2, MSH6, and PMS2. Typically, these genes are strongly associated with malignancies of the intestinal tract but also influence the predisposition to BROVCA cancers. The mismatch repair genes were considered as a single group and are reported in Table 4.6.

The genes involved in the error sensing, cell checkpoint and cell death system, present in our series are ATM, NBN (sensing and signaling), CHEK2 (checkpoint control) and TP53 (apoptosis promoter). All these genes demonstrate variable correlation with breast and ovarian cancer and make up the CKP group (Table 4.7). In addition, the CDH1 gene was added to CKP cluster due to its association with breast cancer.

Genes playing in the Fanconi Anemia complex deserve specific comments. FA genes are numerous and act differently in co-operation with genes belonging to other cellular repair pathways (e.g., HRR complex). The Fanconi Anemia Core Complex (FACC) consists of FANCA, FANCB, FANCC, FANCE, FANCG, FANCI, FANCL, FANCM, and FANCD2 (Table 4.8).

All genes that are not involved in genome stability mechanisms and that are primarily related to different malignancies compared to BROVCA cancers, were grouped all together in the cluster termed “Others”. Genes included in this group may predispose to rare tumors such as those seen in Diceropathies (DICER1) [122] and retinoblastoma (RB1) [123]; genes associated with increased risk of familial melanoma such as MITF, MC1R, CDKN2A; genes associated with increased risk of gastrointestinal tumors such as APC, MUTYH and SMAD4, or of kidneys, such as VHL. Other genes are related to syndromic malignancies such as TSC2, ERCC2, ERCC3, CYLD, CDKN1C, and BUB1B (Table 4.9).

**Table 4.5:** Genes in the homologous recombination repair pathway

HOMOLOGOUS RECOMBINATION REPAIR - HRR		
GENE	GENE NAME	OMIM CANCER/DISEASE ASSOCIATION
BRCA1	Breast cancer-1 gene	Breast-ovarian cancer (604370); Pancreatic cancer (614320)
BRCA2	BRCA2 gene	Breast-ovarian cancer (612555); Fanconi anemia D1 (605724); Prostate cancer (176807); Breast cancer male (114480); Wilms tumor (194070); Medulloblastoma (155255); Glioblastoma 3 (613029); Pancreatic cancer 2 (613347)
BRIP1	BRCA1-associated C-terminal helicase 1	Breast cancer, early-onset (114480); Fanconi anemia J (609054)

PALB2	Partner and localizer of BRCA2	Fanconi anemia N (610832); Breast cancer (114480); Pancreatic cancer (613348)
RAD51C	RAD51, <i>S. cerevisiae</i> , homolog of, C	Fanconi anemia O (613390); Breast-ovarian cancer (613399)
PAD51D	RAD51, <i>S. cerevisiae</i> , homolog of, D	Breast-ovarian cancer (614291)
BLM	DNA helicase, RecQ-like 3	Bloom syndrome (210900)

**Table 4.6:** Genes in the mismatch repair pathway

MISMATCH REPAIR - MMR		
GENE	GENE NAME	OMIM CANCER/DISEASE ASSOCIATION
MLH1	mutL, <i>E. coli</i> , homolog of, 1	Colorectal cancer (609310); Mismatch repair cancer syndrome (276300)
MSH2	mutS, <i>E. coli</i> , homolog of, 2	Colorectal cancer (609310); Mismatch repair cancer syndrome (276300)
MSH6	MutS, <i>E. coli</i> , homolog of, 6	Colorectal cancer (609310); Mismatch repair cancer syndrome (276300)
PMS2	Postmeiotic segregation increased, <i>S. cerevisiae</i> , 2, homolog of	Colorectal cancer (614337); Mismatch repair cancer syndrome (276300)

**Table 4.7:** Genes in the pathway of sensing, signaling, checkpoint, and cell death control

SENSING, SIGNALING, CHECKPOINT CONTROL, CELL DEATH - CKP		
GENE	GENE NAME	OMIM CANCER/DISEASE ASSOCIATION
ATM	Ataxia-telangiectasia mutated (includes complementation groups A, C, D, and E)	Ataxia-telangiectasia (208900); Breast cancer (114480)
CHEK2	Checkpoint kinase 2, <i>S. pombe</i> , homolog of (RAD53, <i>S. cerevisiae</i> , homolog of)	Li-Fraumeni syndrome (609265); Breast cancer (114480); Prostate cancer (176807)
TP53	Tumor protein p53	Colorectal cancer (114500); Li-Fraumeni syndrome (151623); Hepatocellular carcinoma (114550); Osteosarcoma (259500); Choroid plexus papilloma (260500); Nasopharyngeal carcinoma (607107); Pancreatic cancer (260350); Adrenal cortical carcinoma (202300); Breast

## Clinical Applications

		cancer (114480); Basal cell carcinoma (614740); Glioma susceptibility (137800)
CDH1	Cadherin 1	Blepharocheilodontic syndrome 1 (119580) Gastric cancer (137215) Ovarian cancer (167000) Breast cancer (114480) Prostate cancer (176807)
NBN	Nibrin	Nijmegen breakage syndrome (251260); Aplastic anemia (609135); Leukemia, acute lymphoblastic (613065)

**Table 4.8:** Genes in the Fanconi Anemia Core Complex

FANCONI ANEMIA CORE COMPLEX - FACC		
GENE	GENE NAME	OMIM CANCER/DISEASE ASSOCIATION
FANCA	Fanconi anemia, comp. group A	Fanconi anemia A (227650)
FANCC	Fanconi anemia, comp. group C	Fanconi anemia C (227645)
FANCE	Fanconi anemia, comp. group E	Fanconi anemia E (600901)
FANCG	X-ray repair, repair cross comp. 9	Fanconi anemia G (614082)
FANCL	PHD finger protein 9	Fanconi anemia L (614083)
FANCM	FANCM gene	-
FANCD2	Fanconi anemia, comp. group D2	Fanconi anemia D2 (227646)

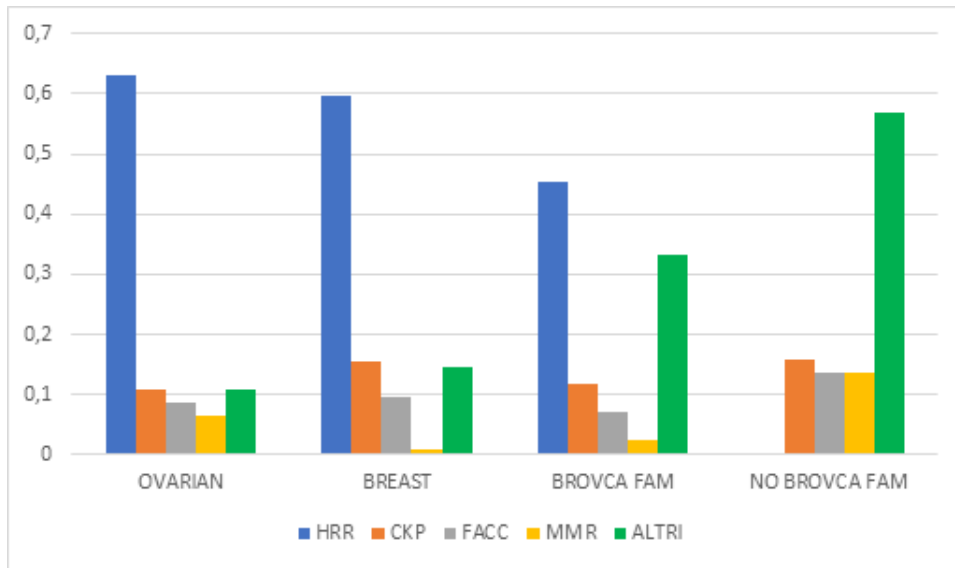
**Table 4.9:** Genes in other pathways associated with non-breast and non-ovarian cancer.

OTHER GENES ASSOCIATED WITH NO BROVCA		
GENE	GENE NAME	OMIM CANCER/DISEASE ASSOCIATION
APC	Adenomatous polyposis coli	Adenomatous polyposis coli (175100); Desmoid disease (135290); Brain tumor-polyposis syndrome 2 (175100); Gardner syndrome (175100)
BUB1B	Budding uninhibited by benzimidazoles 1, <i>S. cerevisiae</i> , homolog of, beta	Mosaic variegated aneuploidy syndrome 1 (257300)
CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	Beckwith-Wiedemann syndrome (130650); IMAGE syndrome (614732)

CDKN2A	Cyclin-dependent kinase inhibitor 2A (p16, inhibits CDK4)	Melanoma (155601); Pancreatic cancer (606719)
CEP57	Centrosomal protein, 57-KD	Mosaic variegated aneuploidy syndrome 2 (614114)
CYLD	CYLD gene	Cylindromatosis (132700); Brooke-Spiegler syndrome (605041); Trichoepithelioma (601606)
DICER1	Dicer, Drosophila, homolog of, 1	Pleuropulmonary blastoma (601200); Goiter with or without Sertoli-Leydig cell tumors (138800); Rhabdomyosarcoma (180295)
ERCC2	Excision repair cross complementing rodent repair deficiency, complementation group-2	Xeroderma pigmentosum (278730)
ERCC3	Excision-repair cross-complementing rodent repair deficiency, complementation group 3	Xeroderma pigmentosum (610651)
HOXB13	HOMEODOMAIN BOX B13	Prostate cancer (610997)
MC1R	Melanocortin-1 receptor	Melanoma (613099)
MITF	Microphthalmia-associated transcription factor	Melanoma (614456)
MUTYH	MutY, E. coli, homolog of	Colorectal cancer (132600)
RB1	Retinoblastoma-1	Retinoblastoma (180200)
SMAD4	Mothers against decapentaplegic, Drosophila, homolog of, 4	Juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome (175050); Myhre syndrome (139210)
TSC2	Tuberin (tuberous sclerosis 2 gene)	Tuberous sclerosis (613254)
VHL	VHL gene	von Hippel-Lindau syndrome (193300)

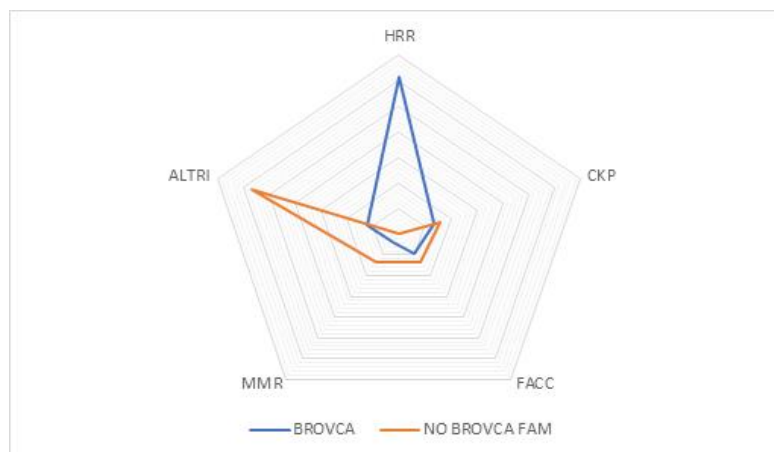
The distribution of the gene clusters in the different clinical groups shows similar profiles for probands with Ovarian and Breast cancer, but it varies in non-BROVCA subjects. The BROVCA patient group demonstrates a high percentage of carriers of P variants in HRR pathway compared with other gene clusters. A slight difference is observed between ovarian vs breast cancer concerning the pathway of mismatch repair. The fraction of probands with breast cancer carrying damage to the MMR is almost nil, while among the ovarian cancer it increases significantly. On the other hand, the fraction of probands presenting a P/LP variant in “Other genes”, is greater in the breast than in the ovarian group. The distributions change further considering probands without positive family history of BROVCA. This latter group demonstrates an opposite profile to the BROVCA probands, without carriers of HRR pathway deficiency and a very high percentage of mutation carriers in the group “Other genes”. Subjects belonging to the BROVCA family's group show an intermediate mutational profile between the p-BO and the p-

NBO in f-NBO groups, retaining a high fraction of carriers of HRR damage together with an increased percentage of other gene defects.



**Figure 4.7:** Distribution of groups of genes in clinic groups

The distribution of damage among the different pathways in the clinical groups highlighted an orthogonality of the mutational profile between the probands with previous Breast or Ovarian Cancer and the subjects unaffected or with malignancy in other locations and without family history of BROVCA (figure 4.8). The two groups share a small part of the defective genes belonging almost entirely to the mismatch repair, checkpoint control and FA core complex groups, while they have an opposite profile with respect to the HRR genes and the group of other genes.



**Figure 4.8:** The distribution of damage among the different pathways in p-BO vs p-NBO in f-NBO groups.

#### 4.2.6.6. From genetics to clinical groups

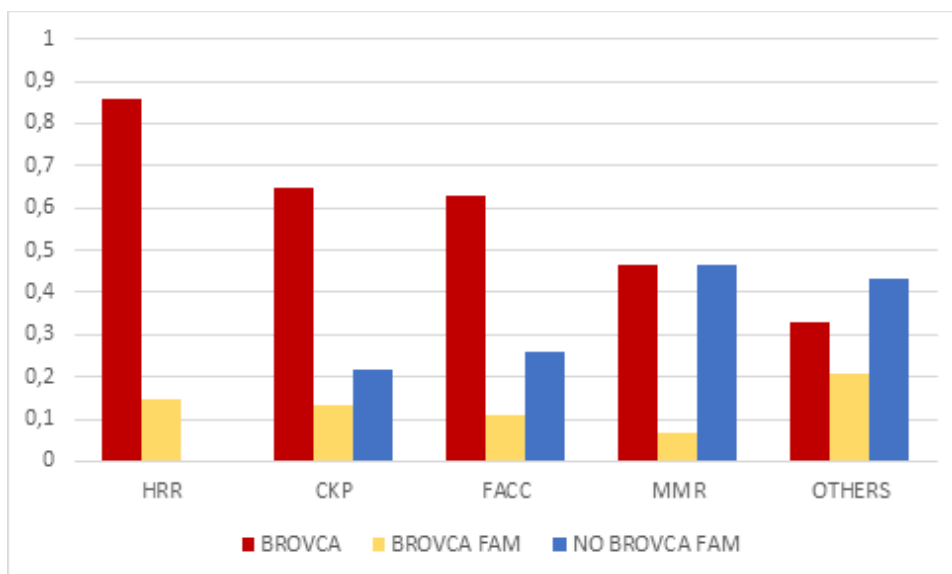
A further point of view that can help to understand better the correlations between genes and tumor type is to study the distribution of clinical groups according to the damaged gene pathway.

When the damage is charged to the HRR genes, almost all carriers have BROVCA cancer and probands unaffected or with tumors in other sites have a positive family history for BROVCA. None of the carriers of the homologous recombination pathway deficiency has a negative familiarity for HBOC, demonstrating a high specificity of mechanism of increased risk for both ovarian and breast cancer.

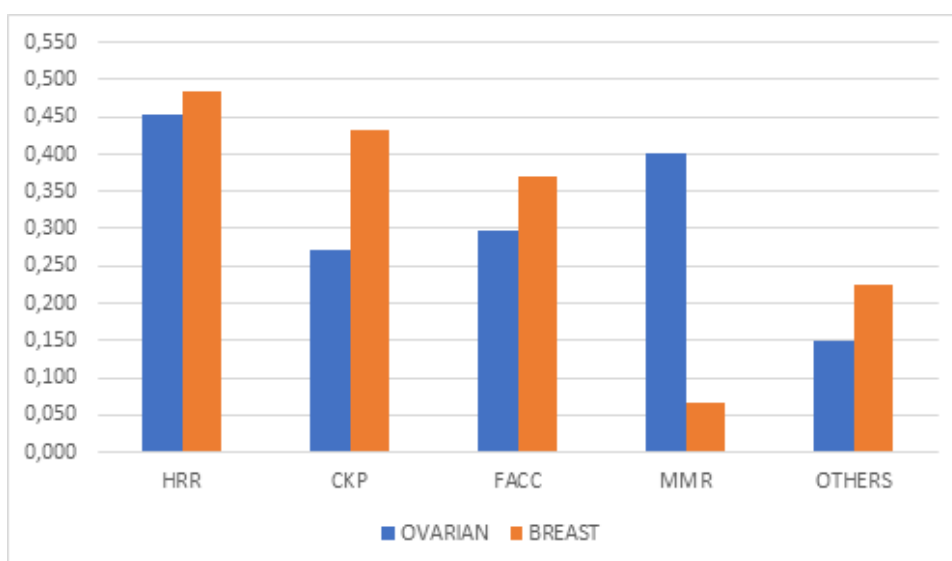
The distribution profile of clinical groups between the CKP and FACC gene groups is remarkably similar to each other and is characterized by a higher fraction of BROVCA tumors, a smaller proportion of non-BROVCA probands in BROVCA families, and a percentage between 20 and 30% of tumors in other sites with negative familiarity for HBOC. In this case, the fraction of breast cancer is greater than ovarian cancer.

The balance between p-BO e p-NBO in f-NBO is obtained in the pathways of mismatch repair, in which the fraction of BROVCA probands is equal to that of non-BROVCA probands in no BROVCA families (approximately 45% for both groups). The mismatch genes are primarily associated in the literature with colorectal cancer but have been shown to cause an increased risk also for BROVCA tumors with a strong imbalance in favor of ovarian tumors compared to breast.

Considering the group of other genes not primarily associated with ovarian and breast cancer, the number of subjects with previous cancer of another type or not affected significantly increases and the proportion of probands BROVCA decreases. In this case, the fraction of no BROVCA in BROVCA families is about 20%, which added to the share of no BROVCA families, reaches the threshold of 2 mutated out of 3. Although the genes of this group have at most a minimal correlation with HBOC, the percentage of carriers nevertheless develops breast or ovarian cancer.



**Figure 4.9:** Distribution of the fraction of BROVCA, NON BROVCA, and not affected probands within the groups of genes.



**Figure 4.10:** Percentage of Ovarian and Breast cancer among the mutated in the diverse groups of genes

**Table 4.10:** Distribution of clinic groups into genetic pathways

	BROVCA	NO BROVCA	NO BROVCA	
			BROVCA FAM	NO BROVCA FAM
HRR	110 (0.859)	19 (0.148)	19 (0.148)	0 (0.000)
CKP	24 (0.647)	10 (0.351)	3 (0.135)	7 (0.216)
FACC	17 (0.630)	8 (0.370)	1 (0.111)	7 (0.259)

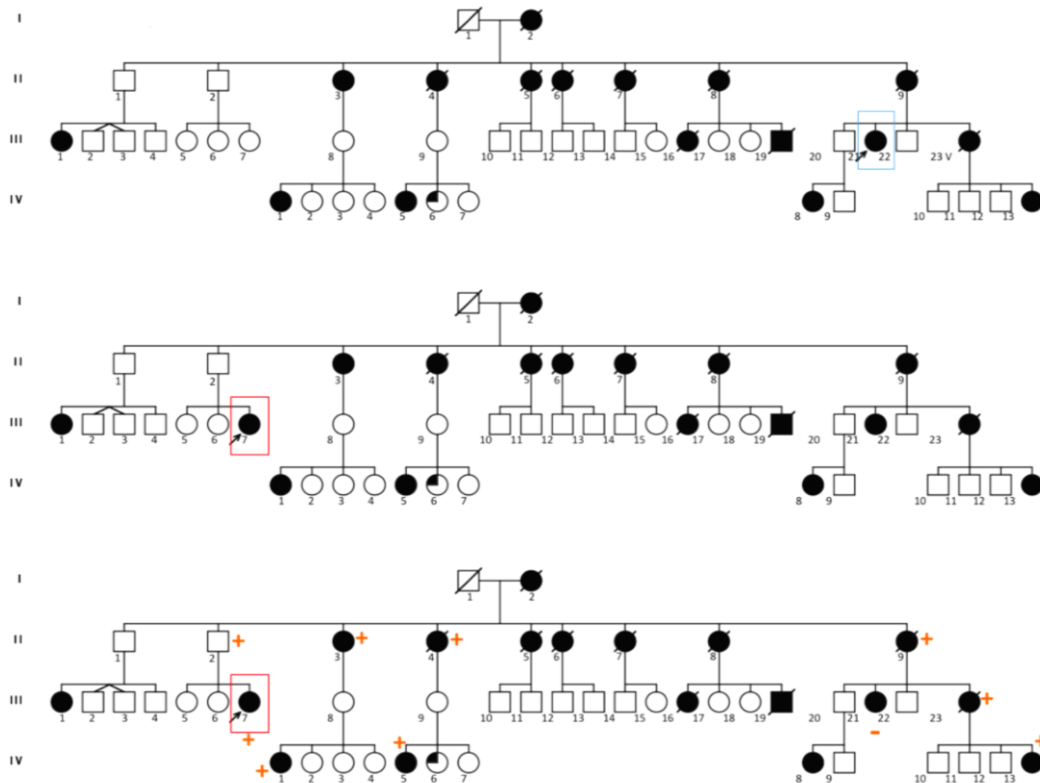


MMR	7 (0.467)	13 (0.533)	5 (0.067)	8 (0.467)
ALTRI	22 (0.328)	43 (0.642)	14 (0.209)	29 (0.433)

#### 4.2.7. Family segregation and familiarity for BROVCA tumors

HBOC families are often characterized by a high number of family members affected by either BROVCA or other types of tumors. Some families are highly informative in terms of predisposition to hereditary cancer and when an affected member tests positive for a pathogenic or likely pathogenic variant, the family management path is simplified. Greater difficulties are encountered in the presence of VUS in the proband: to better understand its role, in the absence of specific *in vitro* tests, the only option potentially contributing to the given variant interpretation is the segregation study in the family. The feasibility of the segregation studies largely depends on the number of living family members, both affected and non-affected, in particular for *de novo* variants (extremely rare in BRCA genes), on the possibility of testing clinically unaffected parents. Cascade genetic testing can be performed in relatives of BRCA-positive probands, thus contributing to the ACMG segregation criteria (PP3). We regularly performed segregation studies in families (both genetic testing, clinical evaluation, tracing clinical reports or biological samples -when possible and feasible- of deceased relatives).

In families with a high number of affected members, a negative test in the proband should not discourage the search for gene defects in other affected relatives. The figure below (figure 4.11) shows that family member III:21 tested negative. When the previously unaffected member III:7 developed triple negative breast cancer, genetic test was performed and identified a BRCA2 pathogenic variant. The segregation study in the family demonstrated the absence of the variant in III:21 but the segregation of the cancer with the BRCA2 mutation in the rest of the family.



**Figure 4.11:** Evolution of a genetic pedigree in a BRCA1 family may show HBOC syndrome despite the presence of a non-mutated proband.

However, this clinical contribution to variant interpretation, which is often feasible for most Mendelian diseases, can be especially difficult in cancer families because of several reasons including:

- High death rate in families
- Age of family members available for genetic testing
- Difficult joining of family members
- Probands with paternal inheritance of BRCA mutations

High death rate in families: Breast and ovarian cancers usually develop from the third decade onwards, with a probability that increases with age. Often, affected probands come to counseling at an adult or advanced age. Parents are often unavailable for genetic testing, especially those who developed and died for cancer. This is a major limitation for segregation studies. Offspring from the familial lineage of an affected and deceased parent can contribute when cascade genetic testing demonstrate a carrier status, because it adds the information on the obligate carrier status of an affected uncle or aunt. An example is the family shown in the figure below. The first pedigree evaluation (figure 4.12) shows:

- The BROVCA proband that carries a BRCA pathogenic variant (indicated with the arrow).
- Two mutated daughters without malignancies.
- Three not affected brothers: two not mutated and 1 not tested.
- The mutated but not affected mother. The mother was subjected to hysterectomy and bilateral salpingo-oophorectomy at the age of 45 years.
- The grandmother deceased due to BROVCA and not tested.

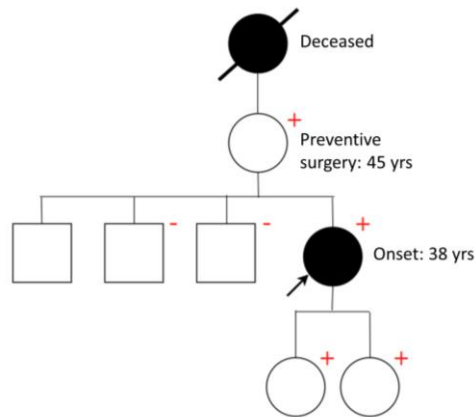
The segregation study of this family pedigree is uninformative due to the early age of the two daughters, the impossibility of testing the affected but deceased grandmother, and the preventive surgery of the mother. The segregation information is not sufficient to apply the PP3 criterion. By expanding the family study with the genetic test of a family member (III:2) suffering from BROVCA, the pedigree becomes informative (figure 3.13). The genetic test reveals that the family member carries the same variant of the proband, therefore the subject II:1 (great-aunt of the proband with BROVCA) is an obligate carrier, like the grandmother of the proband (II: 4).

The ACMG classification system recommends quantifying the co-segregation in order to shift the strength of the PP3 criterion, based on the cumulative number of meiosis occurring between mutated affected family members and the proband [68]: PP3\_supporting with 3 meiosis, PP3\_Moderate with 5 meiosis, and PP3\_Strong with 7 meiosis.

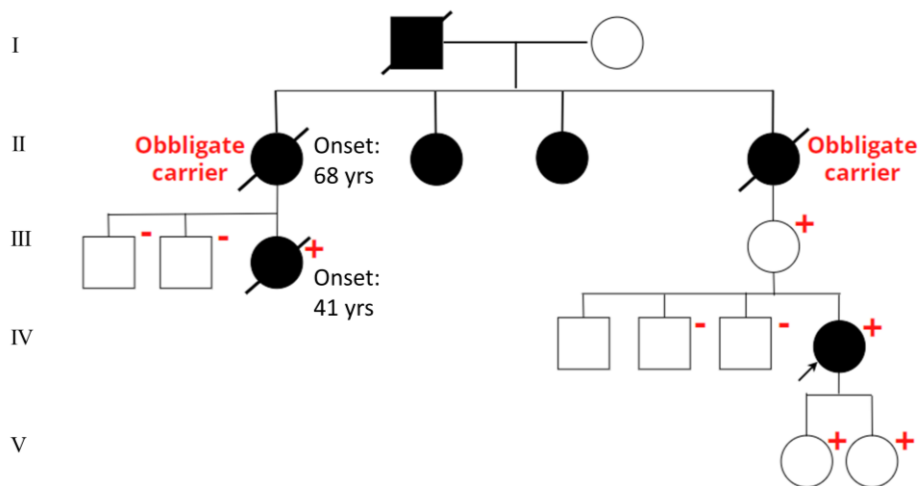
Considering that, between mother and children, and between siblings a meiosis a single meiosis occurs, the quantification of the co-segregation of the variant with breast or ovarian malignancies, takes in consideration:

- 2 meiosis from the proband and the grandmother (II:4);
- 3 meiosis from the proband and the great-aunt (II:1);
- 4 meiosis from the proband and the III:2 family member.

The cumulative number of meiosis that support the co-segregation is 9 meiosis. This result is sufficient to activate PP3 using Strong weight (PP3\_Strong).



**Figure 4.12:** The first evaluation of the family pedigree



**Figure 4.13:** Expanded family pedigree shows strong co-segregation.

Age of family members available for genetic testing: The relatives who perform the targeted test to identify if they carry the variant of the affected family member are often the children or grandchildren of the proband. The probability that they have developed BROVCA tumor at the date of the consultation is low and, regardless of the result of the genetic test, this type of subject is not very informative.

Difficult joining of family members: Genetic testing is commonly perceived by patients as a personal and private examination and the result is hardly communicated to more distant family members. Even in the presence of potentially highly informative families because they are made up of many subjects, perhaps with a significant share of BROVCA cancer, it is often impossible to extend the test to a sufficient number of family members to better understand the effect of the variant identified in the proband.

Probands with paternal inheritance of BRCA mutations: when the carrier is the father, the inheritance may escape attention. Family studies are uniquely useful to characterize the parental lineage and to activate protection plans for family members who are exposed at the risk of malignancy. Given that, recent data provide evidence of an increased risk of prostatic cancer in males carriers of pathogenic mutations in BRCA genes, healthy mutated fathers of affected daughters are now entering novel prevention surveillance plans for prostatic cancer.

#### **4.2.7.1. Multiple families affected by the same variant**

One possibility is to perform a segregation study using multiple families affected by the same variant. This strategy could increase the possibility of encountering sufficient information subjects to define the pathogenic role of the variant. However, the frequency of pathogenic variants in the population is extremely low and this contributes to increasing the difficulty in finding multiple carrier families of the same genetic defect.

In conclusion, although many families have a high potential in terms of informativeness regarding the analysis of co-segregation between BROVCA cancer and the family variant, being able to reach enough relatives of the proband to obtain a significant result is very difficult. The study of familiarity can be useful to hypothesize the segregation of HBOC in the family, but the identification in the proband of a variant of uncertain significance that is a candidate for causing the syndrome can hardly make use of sufficient data to reclassify it. This can have negative implications in terms of choosing the appropriate therapy or clinical management and monitoring of unaffected carriers, decreasing the effectiveness of dedicated care pathways.

# Chapter 5

---

## Conclusions and future implementations

The path followed during the dissertation of the thesis opened with the introduction of the fundamental concepts underlying the knowledge on NGS sequencing, its applications and the analysis of the data produced by this technology. The discussion of the background has focused on the applications of DNA targeted sequencing. In the context of targeted sequencing, Helper was developed as a solution for the simplified customization of bioinformatics pipelines.

The implementation structure of Helper was detailed in chapter 3, and the modules and steps that compose the pipeline workflow have been discussed, paying particular attention to the characteristics of the integrated tools. The graphic interface developed to simplify the experience of using the platform was shown, and two performance tests were conducted using the pipeline developed for the CMGCV of the San Matteo Hospital in Pavia. Helper is now an essential part of the genetic units, both for clinics and research. The possibility of using a tool that simplifies the analysis of NGS samples facilitate the approach to bioinformatics of professionals who have little expertise with code management and NGS data analysis. Helper therefore have a dual function: the intuitive development of bioinformatics pipelines and the teaching role to explain how a bioinformatics pipeline is developed. Helper can also be used both on a workstation and on a common PC, ensuring analysis times consistent with laboratory times for reporting results even in the case of low computation potential. The next Helper development step is to adapt the system also to HPC solutions such as cluster servers or Cloud computing in order to expand the potential of the platform to WES and WGS applications.

The last chapter presented two examples of practical application of the analysis customization based on specific needs, and of in-depth study of the genetic causes of a disease. The development of a classification system of the ACMG rules, adapted to the specific problem of Desmin variants, has shown how it is necessary to focus attention on the unique characteristics of genes related to genetic diseases, in order to better understand the genotype-phenotype correlation. The support that the aid systems for the interpretation of genetic data provide to the molecular genetics laboratories is essential to

simplify the collection of information and accelerate the decision-making process. However, the comparative evaluation of the data has shown that the result is highly dependent on the software used for the bioinformatics analysis, and that the non-correct interpretation of the variants may depend upon incomplete data used for the classification. The development of gene specific systems reduces misclassification (in our specific DES example to prevent the over-interpretation of LP variants). In addition, the inconsistencies between geno-phenotypes in both patients and relatives, as well as the detection of second pathogenic variants in non-DES gene segregating with the phenotype in the family adds further contribution to the variant interpretation.

The exclusion of some VUS - which through the adapted system become LB or B - increases the informativeness of the genetic test, decreasing the uncertainty. The example of the DES variants confirms the central role of integration of genetic, clinic, and pathology data in unravelling the real cause of the disease and strengthens the clinical actionability. Understanding when a pathological phenotype is related to a given mutated gene further contributes to disease classification according to genetic causes and to effectively schedule the clinical follow-up for patients and families. From this point of view, deep phenotyping, carried out through an in-depth study of the patient and its monitoring over time, can help to better define some features of the disease that may appear non-specific or non-informative in early stages of the disease, but which can later reveal the consistency of the genotype with the phenotype. The future goal is to improve the CMGCV-DES model and to extend the adaptation of the ACMG rules to other disease genes.

The in-depth study of the genetic causes of HBOC and BROVCA tumors through the family survey, demonstrates the need to depart from the past restrictive guidelines that limited the genetic testing to BRCA1 and BRCA2, and expand the test to other cancer genes. BRCA1 and BRCA2 pathogenic variants are actionable for both prevention and treatment (risk-reducing surgery and medication-PARP-INHIBITORS) of the proband and family, preventing, and monitoring plans; defects in other malignancy-related genes may equally become potentially actionable for treatments and surgical decisions, as well as for family care.

Understanding the role of other genes and other pathways on pathology is essential in order to calculate the risk of malignancy throughout life. For example, our BROVCA tumor study not only demonstrated that the HRR system is actually related to breast and ovarian cancer, but also strengthened this correlation, showing that in the presence of a defect in HRR genes, the familial BROVCA risk is very high. Alternatively, in families without members suffering from BROVCA, the probability of defects in genes acting in the HRR pathway is very low. Having a complete scenario of the genetic makeup of the different types of tumors contributes to better define potential diagnostic targets and provide optimally interpreted genetic data to the clinical and scientific community.

The content of the thesis demonstrates the essential role of bioinformaticians / bioengineers in the genetic paths of mendelian diseases. The ability to combine development skills of software systems and tools to simplify complex processes such as the design of bioinformatics pipelines, the possibility of carrying out technological consultancy on calculation and analysis systems, and the ability to interact effectively in highly multidisciplinary contexts, make the bioinformatician / bioengineer an important member of the team dealing with the diagnostic and research process in the field of molecular genetics.



---

## References

---

- [1] F. Sanger, S. Nicklen, A. R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74, 5463-5467 (1977).
- [2] J. Straiton, T. Free, A. Sawyer, J. Martin, From Sanger sequencing to genome databases and beyond. *Biotechniques* 66, 60-63 (2019).
- [3] H. V. Firth et al., DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 84, 524-533 (2009).
- [4] P. D. Stenson et al., The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133, 1-9 (2014).
- [5] M. J. Landrum et al., ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44, D862-868 (2016).
- [6] M. J. Landrum et al., ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42, D980-985 (2014).
- [7] V. A. McKusick, Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* 80, 588-604 (2007).
- [8] D. Klarin, P. Natarajan, Clinical utility of polygenic risk scores for coronary artery disease. *Nat Rev Cardiol*, (2021).
- [9] Polygenic Risk Score Task Force of the International Common Disease Alliance, Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* 27, 1876-1884 (2021).
- [10] K. Stanek et al., Bilateral Prophylactic Nipple-Sparing Mastectomy: Analysis of the Risk-Reducing Effect in BRCA1/2 Mutation Carriers. *Aesthetic Plast Surg*, (2021).

- [11] A. Fernandez-Marmiesse, S. Gouveia, M. L. Couce, NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Curr Med Chem* 25, 404-432 (2018).
- [12] P. Pawliczek et al., ClinGen Allele Registry links information about genetic variants. *Hum Mutat* 39, 1690-1701 (2018).
- [13] P. Meyer, J. Saez-Rodriguez, Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. *Cell Syst* 12, 636-653 (2021).
- [14] J. Shendure, E. Lieberman Aiden, The expanding scope of DNA sequencing. *Nat Biotechnol* 30, 1084-1094 (2012).
- [15] D. R. Masser, D. R. Stanford, W. M. Freeman, Targeted DNA methylation analysis by next-generation sequencing. *J Vis Exp*, (2015).
- [16] L. Feng, J. Lou, DNA Methylation Analysis. *Methods Mol Biol* 1894, 181-227 (2019).
- [17] D. Schmidt et al., ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* 48, 240-248 (2009).
- [18] T. H. Kim, J. Dekker, ChIP-seq. *Cold Spring Harb Protoc* 2018, (2018).
- [19] T. Stuart, R. Satija, Integrative single-cell analysis. *Nat Rev Genet* 20, 257-272 (2019).
- [20] R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years. *Nat Rev Genet* 20, 631-656 (2019).
- [21] K. J. van Nimwegen et al., Is the \$1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clin Chem* 62, 1458-1464 (2016).
- [22] M. Gulilat et al., Targeted next generation sequencing as a tool for precision medicine. *BMC Med Genomics* 12, 81 (2019).
- [23] P. Marino et al., Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study. *Eur J Hum Genet* 26, 314-323 (2018).
- [24] B. Milholland et al., Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* 8, 15183 (2017).

- [25] Y. Dou, H. D. Gold, L. J. Luquette, P. J. Park, Detecting Somatic Mutations in Normal Cells. *Trends Genet* 34, 545-557 (2018).
- [26] S. Oota, Somatic mutations - Evolution within the individual. *Methods* 176, 91-98 (2020).
- [27] P. C. Nowell, The clonal evolution of tumor cell populations. *Science* 194, 23-28 (1976).
- [28] P. Vikas, N. Borcharding, A. Chennamadhavuni, R. Garje, Therapeutic Potential of Combining PARP Inhibitor and Immunotherapy in Solid Tumors. *Front Oncol* 10, 570 (2020).
- [29] P. G. Pilié, A. George, T. A. Yap, Patient selection biomarker strategies for PARP inhibitor therapy. *Ann Oncol* 31, 1603-1605 (2020).
- [30] D. Xiao et al., High Tumor Mutation Burden and DNA Repair Gene Mutations are Associated with Primary Resistance to Crizotinib in. *Onco Targets Ther* 14, 4809-4817 (2021).
- [31] L. Kananen et al., Circulating cell-free DNA level predicts all-cause mortality independent of other predictors in the Health 2000 survey. *Sci Rep* 10, 13809 (2020).
- [32] H. Osumi, E. Shinozaki, K. Yamaguchi, H. Zembutsu, Clinical utility of circulating tumor DNA for colorectal cancer. *Cancer Sci* 110, 1148-1155 (2019).
- [33] S. De, Signatures Beyond Oncogenic Mutations in Cell-Free DNA Sequencing for Non-Invasive, Early Detection of Cancer. *Front Genet* 12, 759832 (2021).
- [34] Z. B. Huang, H. T. Zhang, B. Yu, D. H. Yu, Cell-free DNA as a liquid biopsy for early detection of gastric cancer. *Oncol Lett* 21, 3 (2021).
- [35] C. Bailleux, L. Lacroix, E. Barranger, S. Delaloge, Using methylation signatures on cell-free DNA for early cancer detection: a new era in liquid biopsy? *Ann Oncol* 31, 665-667 (2020).
- [36] H. Luo, W. Wei, Z. Ye, J. Zheng, R. H. Xu, Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends Mol Med* 27, 482-500 (2021).

- [37] R. K. Ravi, K. Walton, M. Khosroheidari, MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Methods Mol Biol* 1706, 223-232 (2018).
- [38] I. Kozarewa, J. Armisen, A. F. Gardner, B. E. Slatko, C. L. Hendrickson, Overview of Target Enrichment Strategies. *Curr Protoc Mol Biol* 112, 7.21.21-27.21.23 (2015).
- [39] J. M. Kebschull, A. M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 43, e143 (2015).
- [40] L. J. Jennings et al., Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* 19, 341-365 (2017).
- [41] S. Goodwin, J. D. McPherson, W. R. McCombie, Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17, 333-351 (2016).
- [42] J. A. Reuter, D. V. Spacek, M. P. Snyder, High-throughput sequencing technologies. *Mol Cell* 58, 586-597 (2015).
- [43] G. A. Van der Auwera et al., From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43, 11.10.11-11.10.33 (2013).
- [44] K. Reinert, B. Langmead, D. Weese, D. J. Evers, Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet* 16, 133-151 (2015).
- [45] J. Shang et al., Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int* 2014, 309650 (2014).
- [46] G. Highnam et al., An analytical framework for optimizing variant discovery from personal genomes. *Nat Commun* 6, 6275 (2015).
- [47] H. Li et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079 (2009).
- [48] H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14, 178-192 (2013).

- [49] J. T. Robinson, H. Thorvaldsdóttir, A. M. Wenger, A. Zehir, J. P. Mesirov, Variant Review with the Integrative Genomics Viewer. *Cancer Res* 77, e31-e34 (2017).
- [50] S. T. Sherry et al., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29, 308-311 (2001).
- [51] A. Auton et al., A global reference for human genetic variation. *Nature* 526, 68-74 (2015).
- [52] Á. Bartha, B. Györfy, Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers (Basel)* 11, (2019).
- [53] A. Supernat, O. V. Vidarsson, V. M. Steen, T. Stokowy, Comparison of three variant callers for human whole genome sequencing. *Sci Rep* 8, 17851 (2018).
- [54] P. Danecek et al., The variant call format and VCFtools. *Bioinformatics* 27, 2156-2158 (2011).
- [55] A. Magi, L. Tattini, T. Pippucci, F. Torricelli, M. Benelli, Read count approach for DNA copy number variants detection. *Bioinformatics* 28, 470-478 (2012).
- [56] M. Zhao, Q. Wang, P. Jia, Z. Zhao, Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14 Suppl 11, S1 (2013).
- [57] A. Y. Cheng, Y. Y. Teo, R. T. Ong, Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* 30, 1707-1713 (2014).
- [58] F. Cunningham et al., Ensembl 2022. *Nucleic Acids Res*, (2021).
- [59] N. A. O'Leary et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44, D733-745 (2016).
- [60] W. J. Kent et al., The human genome browser at UCSC. *Genome Res* 12, 996-1006 (2002).
- [61] U. Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49, D480-D489 (2021).
- [62] M. Lek et al., Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291 (2016).

- [63] K. J. Karczewski et al., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434-443 (2020).
- [64] M. S. Cline et al., BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet* 14, e1007752 (2018).
- [65] J. G. Tate et al., COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 47, D941-D947 (2019).
- [66] W. Zhu et al., GNE myopathy caused by a synonymous mutation leading to aberrant mRNA splicing. *Neuromuscul Disord* 28, 154-157 (2018).
- [67] A. El-Gazzar et al., A novel cryptic splice site mutation in COL1A2 as a cause of osteogenesis imperfecta. *Bone Rep* 15, 101110 (2021).
- [68] S. Richards et al., 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. *Genet Med* 17, 405-424 (2015).
- [69] Q. Li, K. Wang, InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* 100, 267-280 (2017).
- [70] G. Nicora et al., CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. *Hum Mutat* 39, 1835-1846 (2018).
- [71] C. Kopanos et al., VarSome: the human genomic variant search engine. *Bioinformatics* 35, 1978-1980 (2019).
- [72] M. M. Li et al., Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn* 19, 4-23 (2017).
- [73] V. Jalili et al., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* 48, W395-W402 (2020).

- [74] T. Joo et al., SEQprocess: a modularized and customizable pipeline framework for NGS processing in R package. *BMC Bioinformatics* 20, 90 (2019).
- [75] J. Lin et al., Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. *BMC Genomics* 18, 378 (2017).
- [76] M. D'Antonio et al., RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC Genomics* 16, S3 (2015).
- [77] Agilent, The Agilent Genomics NextGen Toolkit (AGeNT), [https://www.agilent.com/en/product/next-generation-sequencing/hybridization-based-next-generation-sequencing-ngs/ngs-software/agent-232879#features](https://www.agilent.com/en/product/next-generation-sequencing/hybridization-based-next-generation-sequencing/ngs/ngs-software/agent-232879#features).
- [78] M. Martin. (*EMBnet.journal*, 2011), vol. 17, pp. 10-12.
- [79] U. H. Trivedi et al., Quality control of next-generation sequencing data without a reference. *Front Genet* 5, 111 (2014).
- [80] S. W. Wingett, S. Andrews, FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* 7, 1338 (2018).
- [81] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).
- [82] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359 (2012).
- [83] Broad Institute, (<https://broadinstitute.github.io/picard/>), (Github repository, 2019).
- [84] R. Poplin et al., Scaling accurate genetic variant discovery to tens of thousands of samples, (bioRxiv), vol. 201178.
- [85] E. Garrison., M. Gabor., Haplotype-based variant detection from short-read sequencing, (arXiv, 2012), vol. arXiv:1207.3907v2.
- [86] D. C. Koboldt et al., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22, 568-576 (2012).

- [87] Z. Lai et al., VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 44, e108 (2016).
- [88] W. McLaren et al., The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122 (2016).
- [89] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, e164 (2010).
- [90] A. Fowler et al., Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res* 1, 20 (2016).
- [91] L. F. Johansson et al., CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum Mutat* 37, 457-464 (2016).
- [92] E. Talevich, A. H. Shain, T. Botton, B. C. Bastian, CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* 12, e1004873 (2016).
- [93] V. Plagnol et al., A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 28, 2747-2754 (2012).
- [94] G. Povysil et al., panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat* 38, 889-897 (2017).
- [95] J. M. Moreno-Cabrera et al., Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet* 28, 1645-1655 (2020).
- [96] B. D. Gelb et al., ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med* 20, 1334-1345 (2018).
- [97] M. J. Patel et al., Disease-specific ACMG/AMP guidelines improve sequence variant interpretation for hearing loss. *Genet Med* 23, 2208-2212 (2021).
- [98] V. Azzimato, N. Genneback, A. M. Tabish, B. Buyandelger, R. Knöll, Desmin, desminopathy and the complexity of genetics. *J Mol Cell Cardiol* 92, 93-95 (2016).



- [99] A. Brodehl, A. Gaertner-Rommel, H. Milting, Molecular insights into cardiomyopathies associated with desmin (DES) mutations. *Biophys Rev* 10, 983-1006 (2018).
- [100] ClinGen Variant Curation SOP Committee, ClinGen General Sequence Variant Curation Process Standard Operating Procedure Version 2.0, (The Clinical Genome Resource, <https://clinicalgenome.org/docs/variant-curation-standard-operating-procedure-version-2/>), (2021).
- [101] A. N. Abou Tayoun et al., Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* 39, 1517-1524 (2018).
- [102] A. Waring et al., Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy. *J Med Genet* 58, 556-564 (2021).
- [103] E. Persyn et al., DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. *PLoS One* 12, e0179364 (2017).
- [104] R. Ghosh et al., Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum Mutat* 39, 1525-1530 (2018).
- [105] M. A. Kelly et al., Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genet Med* 20, 351-359 (2018).
- [106] R. Walsh et al., Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: the case of hypertrophic cardiomyopathy. *Genome Med* 11, 5 (2019).
- [107] A. Pérez-Serra et al., Genetic basis of dilated cardiomyopathy. *Int J Cardiol* 224, 461-472 (2016).
- [108] M. R. Taylor et al., Prevalence of desmin mutations in dilated cardiomyopathy. *Circulation* 115, 1244-1251 (2007).
- [109] The Sequence Variant Interpretation (SVI) Working Group, SVI Recommendation for in trans Criterion (PM3) - Version 1.0 (2019).

- [110] S. E. Brnich et al., Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med* 12, 3 (2019).
- [111] E. Arbustini et al., Desmin accumulation restrictive cardiomyopathy and atrioventricular block associated with desmin gene defects. *Eur J Heart Fail* 8, 477-483 (2006).
- [112] EnGenome, eVai software platform, ([www.engenome.com](http://www.engenome.com))
- [113] Genoox, Fanklin (<https://franklin.genoox.com/>)
- [114] E. R. Riggs et al., Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med* 22, 245-257 (2020).
- [115] R. Yoshida, Hereditary breast and ovarian cancer (HBOC): review of its molecular characteristics, screening, treatment, and prognosis. *Breast Cancer* 28, 1167-1180 (2021).
- [116] H. Sung et al., Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 71, 209-249 (2021).
- [117] F. C. Nielsen, T. van Overeem Hansen, C. S. Sørensen, Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer* 16, 599-612 (2016).
- [118] L. C. Hartmann, N. M. Lindor, The Role of Risk-Reducing Surgery in Hereditary Breast and Ovarian Cancer. *N Engl J Med* 374, 454-468 (2016).
- [119] S. Banerjee et al., Maintenance olaparib for patients with newly diagnosed advanced ovarian cancer and a BRCA mutation (SOLO1/GOG 3004): 5-year follow-up of a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol*, (2021).
- [120] A. N. J. Tutt et al., Adjuvant Olaparib for Patients with. *N Engl J Med* 384, 2394-2405 (2021).
- [121] L. Biganzoli et al., The requirements of a specialist breast centre. *Breast* 51, 65-84 (2020).

- [122] J. Azzollini et al., Clinical heterogeneity and reduced penetrance in DICER1 syndrome: a report of three families. *Tumori*, 3008916211058788 (2021).
- [123] M. Tanwar, S. Balaji, A. Vanniarajan, U. Kim, G. Chowdhury, Parental age and retinoblastoma-a retrospective study of demographic data and genetic analysis. *Eye (Lond)*, (2021).