A translational bioinformatics approach to improve genetic diagnostics of hereditary cancer using next-generation sequencing data

José Marcos Moreno Cabrera

A translational bioinformatics approach to improve genetic diagnostics of hereditary cancer using next-generation sequencing data

A thesis presented by

José Marcos Moreno Cabrera

Submitted for the degree of

Doctor of Philosophy in Biomedicine

Doctor por la Universidad de Barcelona, Programa de Biomedicina

PhD thesis performed at the Bellvitge Biomedical Research Institute
and the Germans Trias i Pujol Research Institute

Dr. Bernat Gel Moreno Thesis director	Dra. Conxi Lázaro García Thesis director	Dr. Mariano Monzó Planella Thesis tutor

José Marcos Moreno Cabrera

Barcelona 2021

Acknowledgements

La ciencia es un esfuerzo coral. Explicaba John Donne en aquel poema que ningún ser humano es una isla. Estamos aquí gracias a otros; otros que estuvieron antes investigando, otros que han estado estos años compartiendo el camino conmigo. Ahora dedico unas palabras para agradecerles, no serán suficientes.

Quiero, por tanto, dar gracias a Bernat y a Conxi por darme la oportunidad de hacer investigación traslacional y útil para la sanidad pública. He aprendido mucho de vosotros. Vuestra visión y ayuda han hecho posible este trabajo. Ante la magnitud de una enfermedad como el cáncer, es un privilegio haber podido contribuir, muy modestamente, a hacer un poco más pequeña su sombra. Gracias.

En ese esfuerzo coral del que hablaba estuvieron muchos compañeros, de una u otra manera. Desarrollé esta tesis entre dos mundos, uno más allá del río Besós, llamémoslo IMPPC, otro más acá del río Llobregat, llamémoslo ICO.

De aquel lado del río Besós, gracias a Edu, Meri, Helena, Miriam, Eli, Alex, Josep, Ernest, Lucie, Izaskun. Fue una suerte trabajar con Iñaki y aquella capacidad suya de hacer fácil lo difícil. Tiene que haber en estas líneas un sitio especial para Inma por, entre otras, aquella *historia de las sillas*.

De este lado del río Llobregat, debo comenzar por Jesús, tercer *advisor* de tesis, y agradecerle su sabiduría en el mundo de los genes y su sentido del humor. Compañeros del habitáculo WHR que tantos momentos-y hasta obras navideñas- pasamos, Mariona, Xavi *beer-advisor*, Pili, Fati, Gau, Nuria, Estela, Sara. Y Eva Tornado, siempre vital. Y huracán tengo-una-idea-que-comentar Nonia. Y Gardenia power. Y Juani, Mireia, Dani, Eli, Marta, Sara. Poner vuestros nombres aquí no explica el tiempo compartido. Y gracias a Mireia Morell y Maribel por su ayuda, y a las técnicas, que han hecho un trabajo imprescindible, Raquel, Olga, Eva, Carol. *I estimats* predocs, Paula, gran Edgar, Isa, Julia; el lab está en buenas manos.

Gracias a los que hacen posible la ciencia con vocación de utilidad pública.

Gracias a los que formáis parte de mi vida más allá de la ciencia.

Gracias a Leti, por tu manera de mirar al mundo.

Y gracias a mis padres.

Table of contents

Abbreviations

aCGH - Array comparative genomic hybridization

ACMG - American College of Medical Genetics and Genomics

AMP- Association for Molecular Pathology

AD - Autosomal dominant

AR - Autosomal recessive

ASCII - American Standard Code for Information Interchange

BF - Bayesian factor

bp - Base pair

CNV - Copy number variant

CPU- Central Processing Unit

DNA - Deoxyribonucleic acid

FC - Fold change

FDR - False discovery rate

FP - False positive

FTP - File transfer protocol

GATK - Genome Analysis Toolkit

HGMD - Human Gene Mutation Database

INDEL - Insertions or deletion of up to 49 bp

LOVD - Leiden Open Variation Database

MIP - Molecular inversion probes

MLPA - Multiplex ligation-dependent probe amplification

NGS - Next-generation sequencing

PCR - Polymerase chain reaction

PPV - Positive predictive value

RNA - Ribonucleic acid

ROI - Region of interest

SAM - Sequence Alignment/Map format

SNP - Single-nucleotide polymorphism

SNV - Single-nucleotide variant

UI - User interface

VCF - Variant call format

VUS - Variant of unknown significance

WES - Whole-exome sequencing

WGS - Whole-genome sequencing

Introduction

1 Human genome

The human genome is a vast sequence of over three billion base pairs which encodes the basis of what the human being is. This sequence is contained in a DNA double-strand helix (Watson and Crick, 1953), packed into 23 chromosome pairs: one set from each progenitor. Every human cell contains a DNA copy. Although small when packed, the DNA macromolecule is approximately two meters long when stretched (Alberts *et al.*, 2002). The DNA sequence encodes around 20,000 protein-coding genes (Ezkurdia *et al.*, 2014; Salzberg, 2018) which account for a small portion of the entire genome (approximately 1%). Because of alternative splicing, these protein-coding genes encode a larger number of transcripts and, consequently, a larger number of proteins. The rest of the genome consists of regions with different functions such as introns, non-coding genes, regulatory DNA sequences, repetitive regions such as short or long interspersed nuclear elements, or intergenic regions whose function is frequently unknown.

In 1990, an international public consortium called The Human Genome Project started to obtain the complete sequence of the human genome. Eight years later, a private initiative called Celera Genomics started a parallel project to accomplish the same task. The effort of the two projects resulted in the publication of two drafts of the human genome in early 2001 (Venter *et al.*, 2001; Lander *et al.*, 2001), which covered 90% of the whole sequence. Three years later, the Human Genome Project increased the percentage of the human genome reference sequenced by resolving complex regions and other genomic areas. That genome version release, Build 35, contained 2.85 billion nucleotides interrupted by only 341 gaps (Abdellah *et al.*, 2004).

Since the human genome publication, scientists have worked towards a better description and understanding of the human genome sequence. In this regard, multiple initiatives began to identify the functional elements in the genome sequence, a task known as genome annotation. This task bridges the gap from the genome sequence to the biology of the organism (Stein, 2001). Ensembl, HAVANA, RefSeq, and GENCODE are some examples of genome annotation databases. Ensembl (Hubbard *et al.*, 2002) and HAVANA projects, developed by the European Bioinformatics Institute, use an automated and manual annotation approach, respectively. GENCODE (Harrow *et al.*, 2006), a scientific consortium currently led by the Wellcome Trust Sanger Institute, follows a combined approach by joining both Ensembl and HAVANA annotations. For its part, RefSeq (National Center for Biotechnology Information) (Pruitt *et al.*, 2007) uses a combined approach by complementing the computational annotation with the manual curation and the propagation from other already annotated genomes.

The better understanding of the human genome has allowed for a wide variety of applications and research projects. One of them is the identification of genomic variants and their relationship with human health and disease.

2 Genetic variation

A genetic variant is a specific difference between two genomes. Genetic variants cover a wide range of sizes, from single-nucleotide variants (SNVs) to those that affect entire chromosomes. From a broader perspective, the term genetic variation describes the difference in DNA sequences that occurs among individuals or populations.

2.1 Single-nucleotide variants (SNVs)

SNVs are substitutions of one nucleic acid in the genome (Figure 1). If an SNV exists in a population above a certain frequency (usually over 1%), the variant is called single-nucleotide polymorphism (SNP). SNVs are the most well-characterized and frequent variants: a human genome contains between four and five million SNVs (reviewed in Eichler, 2019).

An SNV can appear within a protein-coding region or outside it. When it appears in a protein-coding region, it can be classified into three categories depending on the protein effect. SNVs are called synonymous mutations when they do not produce an amino acid change (the new set of nucleotides encodes the same amino acid). On the contrary, if an amino acid change occurs, SNVs are called non-synonymous or missense mutations. Most known disease-causing mutations are non-synonymous SNVs (Katsonis *et al.*, 2014). On the other hand, if the nucleotide change produces a stop codon, SNVs are called non-sense or stop-gain mutations. Non-sense SNVs cause the translation process to prematurely finish the protein, which easily results in a non-functional protein.

2.2 Small insertions and deletions (INDELs)

Another variant type consists of small insertions or deletions (INDELs) (Figure 1), usually from one to 49 base pairs (bp) (Eichler, 2019). Their frequency in a human genome ranges from 700,000 to 800,000 per genome (reviewed in Eichler, 2019). When an INDEL appears in a protein-coding region, it can be classified into non-frameshift or frameshift categories depending on the effect of the variant on the reading frame. Non-frameshift INDELs are multiples of three base pairs, so they modify the protein sequence by introducing/deleting one or more amino acids, but the rest of the sequence remains unaffected. On the contrary, frameshift INDELs produce a change in the reading frame, which changes the protein sequence from the mutation point onwards and usually truncates it, affecting its function in most cases.

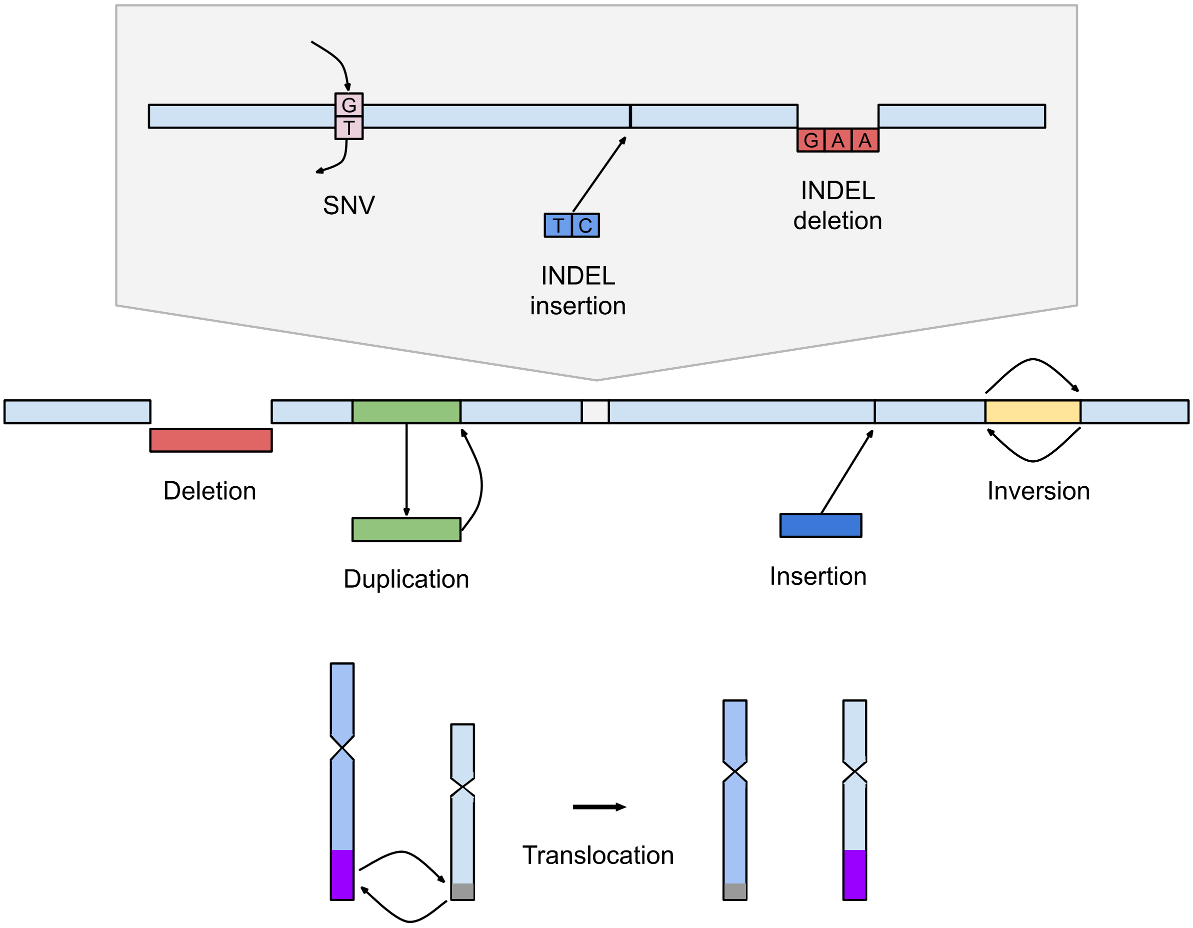


Figure 1. Classes of genetic variants. Adapted from (Smith *et al.*, 2017).

2.3 Structural variants

Structural variants are a larger type of genetic variation. This variant category includes large deletions, duplications and insertions of over 50 bp, inversions, and translocations (Figure 1). Although structural variants are less frequent, from 23,000 to 28,000 events per human genome on average, they are the variant type that most contributes to base-pair differences between two human haplotypes (reviewed in Eichler, 2019).

Within the structural variation category, copy-number variants (CNVs) can be defined as deletions (losses) or duplications (gains) of genomic regions larger than 50 bp. CNVs represent between 4.8% and 9.5% of the human genome (Zarrei *et al.*, 2015). Moreover, they play an important role in contributing to the variation within a population, as well as underlying disease phenotypes (McCarroll and Altshuler, 2007).

2.4 Somatic and germline variants

When the genetic variants appear during cell replication in somatic tissues, they are called somatic variants. This kind of variant only affects the cells descending from the original cell in which the genetic variant appeared. Their effect is limited to the tissue of the individual where the somatic variant arose, that is, they are not inherited between individuals. On the other hand, germline variants appear in germ cells, so they can be inherited between individuals. They contribute to the variability of a population and play a key role in human survival and adaptation. In fact, the accumulation of germline variants is the origin of most of the speciation processes.

3 Hereditary diseases

Genetic diseases can be associated with either somatic or germline variants. There are more than 6,800 genetic diseases for which approximately 4,400 genes have been identified (see OMIM reference). Hereditary diseases are genetic diseases caused by germline variants, so these diseases can be inherited through generations and can be classified into monogenic, polygenic, or chromosomal diseases. Monogenic diseases, also known as Mendelian or single-gene diseases, segregate according to Mendel's inheritance patterns and are caused by germline variants in a single gene. Polygenic diseases, also referred to as complex or multifactorial diseases, can be explained by germline variants in multiple genes and environmental factors. On the other hand, chromosomal diseases are due to structural differences in one or more chromosomes.

3.1 Hereditary cancer

Cancer is a genetic disease characterized by the uncontrolled proliferation of cells. It is caused by variants in genes that regulate three main processes: cell fate determination, cell survival, and genome maintenance (Vogelstein *et al.*, 2013). Genes associated with cancer can be classified into two classes: oncogenes and tumor suppressor genes. Oncogenes accelerate carcinogenesis by promoting growth and cell proliferation, while tumor suppressor genes are responsible for controlling genome integrity, and cell division and replication.

Usually, cancer-causing variants are acquired during one's lifetime due to environmental factors that damage the DNA, such as sun exposure or tobacco smoke, or due to errors during cell division. However, about 5-10% of all cancers are the consequence of germline variants; these are called hereditary cancers.

3.2 Hereditary cancer syndromes

About 200 hereditary cancer syndromes and 100 cancer predisposition genes have been described in the literature (reviewed in Nagy *et al.*, 2004; Rahman, 2014) (Table 1). Individuals carrying a pathogenic variant in a cancer-predisposition gene manifest the disease differently depending on the penetrance of the allele (Taeubner *et al.*, 2018). When the variant is in a highly penetrant gene more severe phenotypes are expected, and its frequency in the population is expected to be very low. On the contrary, low-penetrance variants have a very limited effect, and a cumulative addition with more low-risk alleles is required to significantly impact the phenotype.

Most hereditary cancer syndromes are inherited in an autosomal dominant (AD) manner with incomplete penetrance (Nagy *et al.*, 2004). Some very well studied AD hereditary cancer syndromes examples are Li-Fraumeni syndrome, caused by germline variants in *TP53* and *CHEK2* genes, familial adenomatous polyposis (the *APC* gene), Lynch syndrome (the mismatch repair genes *MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM*), and hereditary breast and ovarian cancer (*BRCA1* and *BRCA2* genes) (Sánchez, 2019). Some other hereditary cancer syndromes have an autosomal recessive inheritance, as is the case of Fanconi anemia (produced by germline variants in the *FANCA*, *FANCB*, *FANCC*, *FANCD*, *FANCE*, *FANCF*, *FANCG*, and *FANCL* genes), polyposis associated with *MUTYH*, and Werner syndrome (caused by germline variants in *WRN* gene).

Usually, hereditary cancer patients are characterized by early age at diagnosis, multiple tumors in a single patient, and other cases in the family history (Sánchez, 2019). When a patient is diagnosed with hereditary cancer, clinicians can provide specific cancer risk assessment and the establishment of appropriate surveillance measures for the patient and family members. Moreover, pathogenic variant carriers can receive targeted surgical and chemotherapeutic treatments (Pennington *et al.*, 2014; Musella *et al.*, 2015).

Table 1. Most common hereditary cancer predisposition syndromes. Adapted from (Sánchez, 2019) and guidelines from the Catalan Consensus on Hereditary cancer.

Hereditary cancer predisposition syndromes	Inheritance	Usually screened Gene(s)
Hereditary breast and ovarian cancer	AD	*ATM, BRCA1, BRCA2, BRIP1, CHEK2, MLH1, MSH2, MSH6, PALB2, RAD51C, RAD51D*
Hereditary breast cancer	AD	*ATM, BARD1, BRCA1, BRCA2, BRIP1, CHEK2, MLH1, MSH2, MSH6, NBN, PALB2, RAD51C, RAD51*
Hereditary ovarian cancer	AD	*BRCA1, BRCA2, BRIP1, MLH1, MSH2, MSH6, RAD51C, RAD51D*
Lynch syndrome	AD	*BRCA1, BRCA2, EPCAM, MLH1, MSH2, MSH6, MUTYH, PMS2, POLD1, POLE*
Familial adenomatous polyposis	AD	*APC, BMPR1A, BRCA1, BRCA2, MLH1, MSH2, MSH6, MUTYH, NTHL1, POLD1, POLE, RNF43, SMAD4*
Multiple endocrine neoplasia type 1 / 2 / 4	AD	*MEN1 / RET / CDKN1B*
Cowden syndrome	AD	*PTEN*
Von Hippel–Lindau disease	AD	*VHL*
Hereditary retinoblastoma	AD	*RB1*
Peutz-Jeghers syndrome	AD	*STK11*
Li-Fraumeni syndrome	AD	*TP53*
Gorlin syndrome	AD	*PTCH1, SUFU, PTCH2*
Tuberous sclerosis	AD	*TSC1, TSC2*
Familiar melanoma	AD	***BAP1, BRCA1, BRCA2, CDK4, CDKN2A,*** *MITF,* ***MLH1, MSH2, MSH6, POT1,*** *TERT*
Neurofibromatosis 1 / 2	AD	*NF1 / NF2*
Schwannomatosis	AD	*SMARCB1*
Familiar paraganglioma / pheochromocytoma	AD	*SDHB, SDHC, SDHB, SDHAF2*
Hereditary gastric cancer	AD	*BRCA1, BRCA2, CDH1, CTNNA1, MLH1, MSH2, MSH6*
Juvenile polyposis syndrome	AD	*SMAD4, BMPR1A*
Birt–Hogg–Dubé syndrome	AD	*FLCN*
Fanconi Anemia	AR	*FANCA-FANCM*
Bloom syndrome	AR	*RECQL3*
Carney complex	AD	*PRKRA1A*
Congenital dyskeratosis	AD, X-linked	*DKC1*
Hereditary prostate cancer	AD	*ATM, BRCA1, BRCA2, HOXB13, MLH1, MSH2, MSH6*
Werner syndrome	AR	*WRN*
Xeroderma pigmentosum	AR	*XPA-XPG, DDB2*
Ataxia–telangiectasia	AR	*ATM*
Carney-Stratakis syndrome	AD	*SDHB, SDHC, SDHD*

Hereditary renal cancer syndromes	AD	*BRCA1, BRCA2, FH, FLCN, MET, MLH1, MSH2, MSH6, SDHB, SDHC, SDHD, VHL*
Hereditary uterine and cutaneous leiomyoma	AD	*FH*
Duncan's disease	X-linked	*SH2D1A*
Sotos syndrome	AD	*NSD1*
Currarino syndrome	AD	*HLXB9*
Chediak-Higashi syndrome	AR	*LYST*
BAP1 tumor predisposition syndrome	AD	*BAP1*
Rothmund-Thomson syndrome	AR	*RECQL4*
DICER1 syndrome	AD	*DICER1*
Familiar Wilms' tumors	AD	*WT1*
Beckwith–Wiedemann syndrome	AD	*KIP2 (CDKN1C)*
Costello syndrome	AD	*HRAS*
Familiar gastrointestinal stromal tumors	AD	*KIT, PDGFRA*
Nijmegen syndrome	AR	*NBS1*
Hereditary pancreatitis	AD	*PRSS1*
Simpson-Golabi-Behmel syndrome	X-linked	*GPC3*
Familiar nonmedullary thyroid cancer	AD	*HABP2*

AD: autosomal dominant; AR: autosomal recessive

4 Genetic diagnostics of hereditary diseases

Genetic diagnostics consists of the detection of variants that predispose to or cause diseases. For hereditary diseases, genetic diagnostics comprises sample collection from the patient and the analytical procedures to detect any variants associated with the hereditary disease. The most common sample used is peripheral blood, although other body fluids or cells are also used, like saliva or oral epithelial cells. Sometimes, tumor samples are also used. Once the biological sample is obtained, nucleic acids are extracted depending on the genetic test to be performed: usually DNA, although sometimes RNA is also obtained.

4.1 SNVs, INDELs and larger events detection

Multiple methods for SNVs and INDELs detection have been developed during the last decades. Initially, methods where limited to one or a few DNA fragments, like allele-specific oligonucleotide ligation assays or Sanger sequencing (Sanger *et al.*, 1977). Sanger sequencing is still being used in genetic diagnostics, mainly for Mendelian diseases in which a mutation has been previously found in the patient’s family. Nevertheless, the arrival of Next-generation sequencing (NGS) has revolutionized genetic testing since millions of fragments can be analyzed in parallel in a single experiment (see paragraphs 5 and 6).

Identification of large rearrangements and copy number alterations require the use of other detection methods. In the early stages of DNA testing, karyotyping was used to identify chromosomal abnormalities. The introduction of fluorescence in situ hybridization (FISH) improved karyotyping performance by detecting smaller known chromosomal events (Langer-Safer *et al.*, 1982). Later, other methods contributed to the detection of smaller DNA events, like array comparative genomic hybridization (aCGH) (Pinkel *et al.*, 1998), multiplex ligation-dependent probe amplification (MLPA) (Schouten *et al.*, 2002) and single nucleotide polymorphism arrays (Chen and Sullivan, 2003). In the last few years, aCGH and MLPA have been considered the gold standards for CNV detection in genetic diagnostics (Talevich *et al.*, 2016; Kerkhof *et al.*, 2017). All mentioned methods are time-consuming and costly, so frequently diagnostic laboratories test CNV events only in a small number of genes.

In any case, when performing genetic diagnostics, some challenging aspects have to be taken into account to accomplish an effective and successful diagnosis (McPherson, 2006). First, sensitivity, since an ideal diagnostic test should not produce any false negative. Second, specificity, because a false positive (FP) involves reporting a wrong positive diagnosis to a patient. Third, clinical interpretation of the identified variants since sometimes there might not be enough evidence to assess their clinical interpretation. Fourth, cost, as genetic testing involves consumables and personnel resources that should be optimized to be cost-effective.

4.2 Variant classification

On average, each individual has up to five million germline variants. While most of them do not affect a person’s health, others might lower or increase the risk of disease. Based on its effect on the degree of likelihood of pathogenicity, a variant can be classified into five groups: benign, likely benign, variant of unknown significance (VUS), likely pathogenic, and pathogenic (Plon *et al.*, 2008; Richards *et al.*, 2015) (Table 2). This classification has been adopted internationally and is used as a standardized reporting system in genetic diagnostics, including hereditary cancer.

Table 2 Variant classification based on the degree of likelihood of pathogenicity. The pathogenicity probability was obtained from (Plon *et al.*, 2008).

Variant class	Description	Pathogenicity probability
Benign	Variants that do not increase disease risk, and are highly frequent within the population or have been demonstrated neutral in family or functional studies.	<0.001
Likely benign	Variants that are not expected to have a major effect on disease, although it cannot be proved conclusively with current evidence (under 10% certainty of being disease-causing).	0.001–0.049
Unknown Significance	Variants for which there is a lack of knowledge that prevents their classification to the other groups.	0.05–0.949
Likely Pathogenic	Variants that have over 90% certainty of being disease-causing, although additional evidence is expected to confirm their pathogenicity.	0.95–0.99
Pathogenic	Disease-causing variants.	>0.99

Variant classification is a challenging task that requires gathering information from multiple sources: functional studies, *in-silico* predictor analyses, population frequencies, and familiar cosegregation analysis. *In-silico* predictors, like SIFT or Polyphen (Ramensky *et al.*, 2002; Ng and Henikoff, 2003), are bioinformatic tools that estimate the functional impact of variants. Many other informatics and bioinformatic resources have been developed to provide useful information when classifying variants. Population databases, like the dbSNP database (SNP Consortium) or The Genome Aggregation Database (gnomAD) (Sherry *et al.*, 2001; Karczewski *et al.*, 2020) (Table 3), store variant data from numerous individuals to provide population frequencies of variants. Moreover, to unravel the relationship between variants and human diseases, there have been initiatives to aggregate clinical assertions and evidence in public databases such as ClinVar, HGMD, or LOVD (Stenson *et al.*, 2003; Fokkema *et al.*, 2005; Landrum *et al.*, 2014) (Table 4).

Multiple procedures have been proposed for variant classification in a clinical context. The International Society for Gastrointestinal Hereditary Tumors (InSIGHT) Variant Interpretation Committee in 2014 developed multiple criteria for the classification of variants in mismatch repair (MMR) genes. Similarly, the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) has been developing criteria for the classification of variants in *BRCA1* and *BRCA2* genes, the last version (2.5.1) being published in June 2017. In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) standardized the clinical interpretation of variants linked to Mendelian diseases (Richards *et al.*, 2015) (Figure 2).

Table 3. Population databases. Adapted from (Richards *et al.*, 2015).

Database	Description
Exome Variant Server	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	Database of variants found during low-coverage and high-coverage genomic and targeted sequencing from 26 populations. Provides more diversity compared to EVS but also contains lower quality data and some cohorts contain related individuals.
dbSNP	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	Database of structural variation (typically greater than 50 bp) submitted from many sources.
gnomAD	Database of variants from 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals (v2). Originally known as Exome Aggregation Consortium (ExAC).

Table 4. Disease databases. Source: (Richards *et al.*, 2015).

Database	Description
ClinVar	Database of assertions about the clinical significance and phenotype relationship of human variation.
OMIM	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database	Database of variant annotations published in the literature. Requires fee-based subscription for much of the content.
Locus/Disease/ Ethnic/Other-Specific Databases	The HGVS site developed a list of thousands of different databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the LOVD system.
DECIPHER	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very strong**
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene/gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non-predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation data with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				*De novo* (without paternity & maternity confirmed) PM6	*De novo* (paternity and maternity confirmed) PS2	
Allelic data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = patho-genic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 2. ACMG-AMP criteria for variant interpretation (Adapted from Richards *et al.*, 2015). BS: benign strong; BP: benign supporting; FH: family history; LOF: loss of function; MAF: minor allele frequency; path: pathogenic; PM: pathogenic moderate; PP: pathogenic supporting; PS: pathogenic strong; PVS: pathogenic very strong

5 Next-generation sequencing

The human genome reference obtained by the Human Genome Project allowed for the development of a second-generation genome sequencing between 2004 and 2005. Next-generation sequencing (NGS) methods, also referred to as second-generation sequencing or massive parallel sequencing, is the term used to include all the high-throughput methods that sequence several DNA fragments in parallel. Very briefly, NGS methods require the isolation and fragmentation of DNA, followed by massively parallel sequencing which produces millions of short reads, usually from 50 bp to 300 bp. These short reads are then aligned to a human genome reference, which is available because of the Human Genome Project releases.

Multiple NGS platforms have appeared, such as Illumina, the Applied Biosystems SOLiD System, 454 Life Sciences (Roche), or Life Technologies Ion Torrent, among others. Although there are multiple differences across NGS methods, most of them include these steps: library preparation, optional enrichment, sequencing, and bioinformatic analysis.

5.1 Library preparation

Library preparation starts by fragmenting the DNA sample (genomic DNA or reverse-transcribed RNA) into small pieces usually of ~500 bp or less. Then, adapter sequences are ligated to the ends of the DNA fragments. This way, each fragment becomes a template that has to be clonally amplified to allow its detection during sequencing. Template preparation strategies can be classified into emulsion PCR (454- Roche, SOLiD- Thermo Fisher, GeneReader - Qiagen, Ion Torrent- Thermo Fisher), solid-phase (Illumina), and DNA nanoball generation (Complete Genomics- BGI).

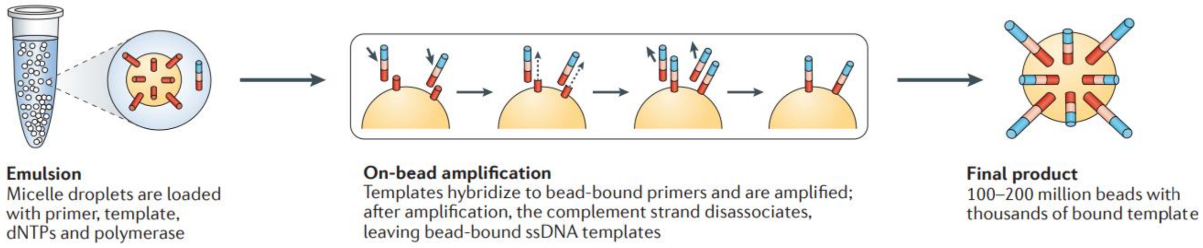

Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase
On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates
Final product
100–200 million beads with thousands of bound template

Figure 3. Emulsion PCR. Source: (Goodwin *et al.*, 2016).

With the emulsion PCR (emPCR) approach (Figure 3), each template is attached to the surface of a bead that contains oligonucleotide probes complementary to the template adaptors. The beads are then distributed into water-oil emulsion droplets where PCR amplification is produced to obtain thousands of copies of the original template.

In Solid-phase amplification (Figure 4) the process occurs on a slide. First, templates bind to complementary primers available on the slide surface and, second, templates bend over to form bridges with the adjacent primers, a process that creates clusters after several rounds. Recent NGS platforms form high-density template clusters on flow cells to achieve higher sequencing throughput.

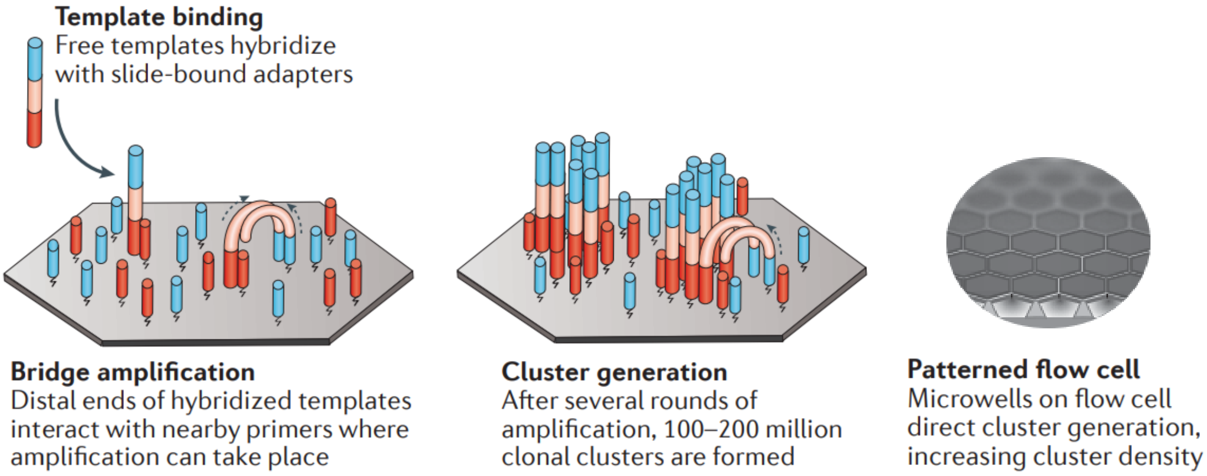

Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place
Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed
Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

Figure 4. Solid-phase amplification. Adapted from (Goodwin *et al.*, 2016).

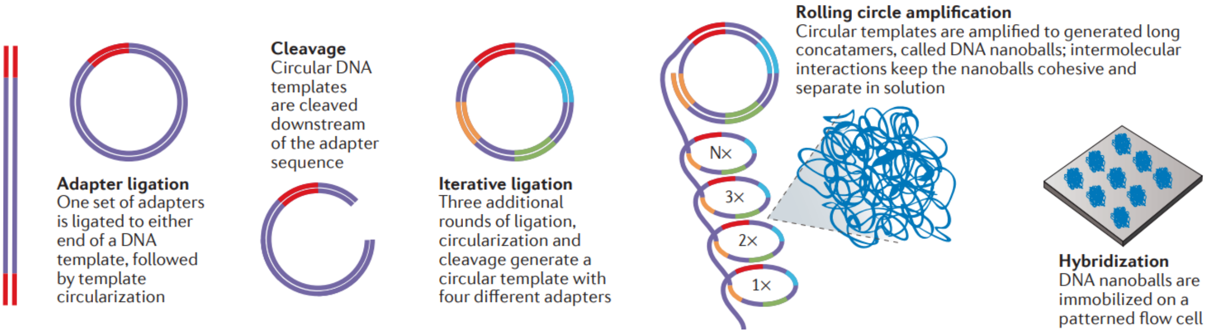

Adapter ligation
One set of adapters is ligated to either end of a DNA template, followed by template circularization
Cleavage
Circular DNA templates are cleaved downstream of the adapter sequence
Iterative ligation
Three additional rounds of ligation, circularization and cleavage generate a circular template with four different adapters
Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution
N×
3×
2×
1×
Hybridization
DNA nanoballs are immobilized on a patterned flow cell

Figure 5. DNA nanoball generation. Source: (Goodwin *et al.*, 2016).

For DNA nanoball generation (Figure 5), circular templates with four different adapters are created through four rounds consisting of adapter ligation, circularization, and cleavage. The amplification occurs due to a rolling circle amplification process that allows for the generation of large concatamers called nanoballs, which are later placed on a flow cell.

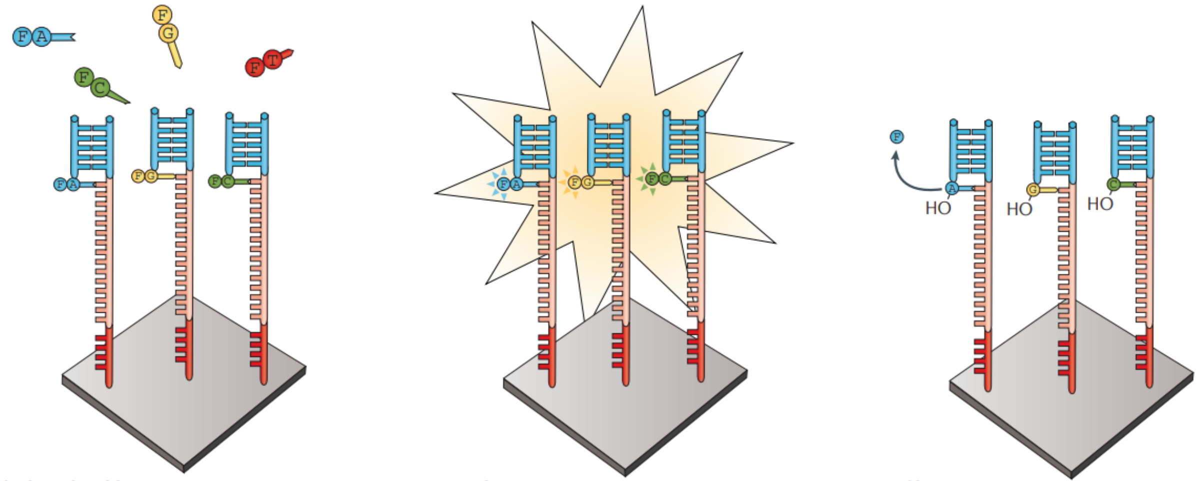
5.2 Optional target enrichment: WGS, WES or targeted gene panels

NGS based DNA-sequencing approaches can also be classified into whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted gene panels. WGS, which does not require previous targeted enrichment, entails the sequencing of the complete individual DNA, including intragenic and intergenic regions. In contrast, WES and targeted gene panels require target enrichment before sequencing. WES involves the sequencing of all the protein-coding genomic regions, that is, the exons that make up the exome (approximately) and represent a small portion of the human genome. On the other hand, in targeted gene panels, only the exons of a subset of genes of interest are sequenced. Here, the number of sequenced genes ranges from a few to hundreds of them.

Multiple target enrichment approaches have been developed based on the polymerase chain reaction (PCR), molecular inversion probes (MIP), and hybrid capture (Mamanova *et al.*, 2010). The PCR-based approach accounts for different methods: the uniplex PCR approach, for which only one amplicon is produced in each reaction, the multiplexed PCR, which allows multiple amplicons for each reaction, and the RainStorm platform approach, which allows for the generation of up to 4,000 amplicons in each reaction. In the MIP-based approach, a common linker, flanked by the ligation and extension arms, hybridizes to either side of the target genomic sequence, so the gap is filled and the target is amplified by PCR. In the hybridization-based capture approach, modified DNA libraries hybridize to target-specific probes in a solution or on a microarray surface, and the background DNA is removed by washing. Each of the approaches has advantages and disadvantages in terms of cost, sensitivity, specificity, ease of use, mass of DNA required, uniformity, and reproducibility, which have to be considered in order to better adapt to the project's needs (Mamanova *et al.*, 2010).

5.3 Sequencing

Multiple methods for parallel sequencing have been developed to date. Sequencing strategies can be summed up as sequencing by ligation, sequencing by synthesis, and single-molecule real-time methods. Sequencing by ligation (SOLiD, Complete Genomics) methods use universal sequences that flank an unknown genomic tag as anchor primer sites. The sequence of the target DNA molecule is identified by taking advantage of the mismatch sensitivity of a DNA ligase. In contrast, sequencing by synthesis methods, which can be classified into sequencing by synthesis with cyclic reversible termination [Illumina (Figure 6), Qiagen] and sequencing by synthesis with single-nucleotide addition (Roche 454, Ion Torrent) (Goodwin *et al.*, 2016), employ a DNA polymerase to perform the sequencing. Usually, the DNA polymerase is used to include a fluorescently labeled nucleotide which contains a reversible terminator. Finally, single-molecule real-time methods (Pacific Biosciences, Oxford Nanopore) do not require template amplification and perform the sequencing in real-time: there is no pause after the detection of a nucleotide or series of nucleotides.



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Figure 6. Illumina sequencing by synthesis with cyclic reversible termination. Source: (Goodwin *et al.*, 2016).

Sequencing methods do not only differ in technological and chemistry aspects, their performance also varies widely for multiple parameters such as runtime, cost, read length or throughput. For example, read length ranges from 36 bp in some Illumina platforms to over 1Mb in Oxford Nanopore MinION (Logsdon *et al.*, 2020); throughput ranges between dozens of Mb in 454 GS Junior platforms and up to 4 Tb in Oxford Nanopore Promethium (Goodwin *et al.*, 2016).

5.4 Bioinformatic analysis

The sequencing process results in the production of large amounts of data that require bioinformatic processing and analysis. The whole bioinformatic analysis can be divided into three levels : primary, secondary, and tertiary (Moorthie *et al.*, 2013) (Figure 7).

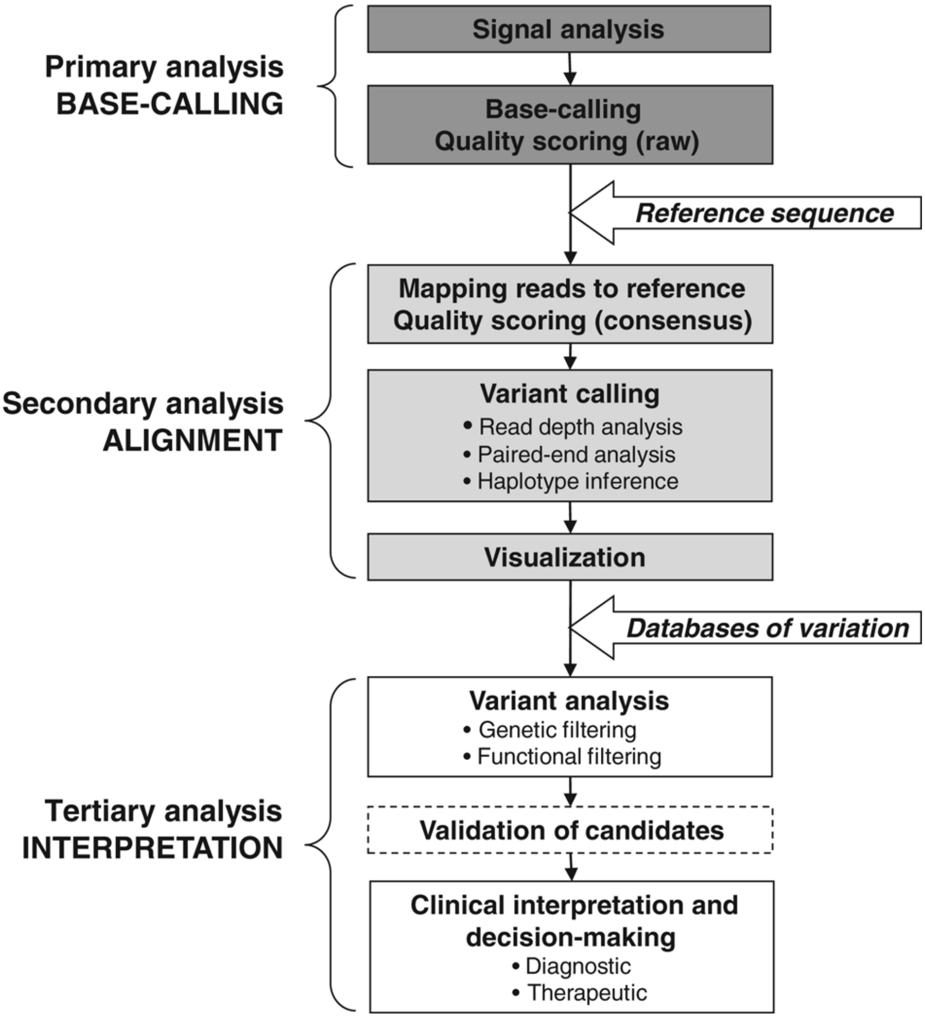


Figure 7. Primary, secondary and tertiary analysis for NGS data. Source: (Moorthie *et al.*, 2013).

The primary analysis consists of the conversion of the raw signals, like electrical current or light intensity, into sequences of nucleotides, also called reads. A single experiment can produce millions of reads that are usually stored in a specific format file: FASTQ. This text-based format stores the nucleotide sequence along with a quality score for each base call. Quality scores are encoded using ASCII character encoding to minimize the file size. After FATSQ creation, it is common to perform a quality check of the sequenced reads. Among other options, and depending on the downstream analysis, some low-quality reads can be filtered out or the adapter sequences can be trimmed. Tools like fastqc or fastp (Chen *et al.*, 2018) are used for this purpose.

In the secondary analysis, reads are mapped to a genome reference to produce the SAM files or BAM files (binary version of the former). Dozens of aligners, such as bwa or bowtie (Langmead and Salzberg, 2012; Li, 2013), have been developed to date, each of which has its own strengths and drawbacks. After read mapping, variants are called using the information provided by the reads. GATK (McKenna *et al.*, 2010), VarScan (Koboldt *et al.*, 2012), or FreeBayes (Gibbs *et al.*, 2015) are some of several tools available for variant calling. Although comma-separated and tab-separated formats are sometimes used, variant calls are usually stored in VCF files. This text format is very flexible and can store several variant fields for multiple samples.

Tertiary analysis involves the interpretation of the detected variants, that is, to explain the role that a variant plays, for example, in the development of a disease. For this purpose, each variant has to be annotated using external genomic variant databases and *in-silico* predictors that evaluate its protein impact. In clinical contexts, variant curators and clinicians will perform the clinical interpretation taking into account all the available information.

5.5 From Sanger to NGS

Until the 2000s, Sanger was the most used method for DNA sequencing (Sanger *et al.*, 1977; Kulkarni and Roy, 2015). However, this very accurate technology suffers from poor scalability, high cost, and being a time-consuming procedure. Moreover, Sanger sequencing only allows for the detection of SNVs and small INDELs. The arrival of NGS technologies revolutionized the sequencing paradigm (Figure 8). In 2005, the first NGS platforms became available and the technology evolved rapidly in terms of cost and performance. From 2005 to 2019, advances in NGS produced an 18,000-fold decrease in the cost of human genome sequencing: from more than 17 million dollars to less than one thousand dollars (Wetterstrand, 2019). The cost reduction, along with higher throughput and longer read lengths, enabled the use of NGS sequencing for multiple purposes, such as genetic diagnostics.

6 NGS in genetic diagnostics

A new healthcare model called precision medicine has been emerging during the last few years. Precision medicine aims to customize medical decisions and treatments, tailoring them to the patient (Xue and Wilcox, 2016). To achieve this customization in genetically-based diseases, such as hereditary cancer, it is key to detect genomic alterations cost-effectively. In this respect, NGS has changed the way genetic testing is performed in the laboratory routine. Due to its low cost and high throughput, NGS has expanded the number of analyzed genes to several dozens. High and moderate-risk genes have been included into the routine of genetic diagnostics laboratories, a fact that has improved the final diagnostic yield, providing additional relevant clinical information for the families (Kurian *et al.*, 2014; Feliubadaló *et al.*, 2017).

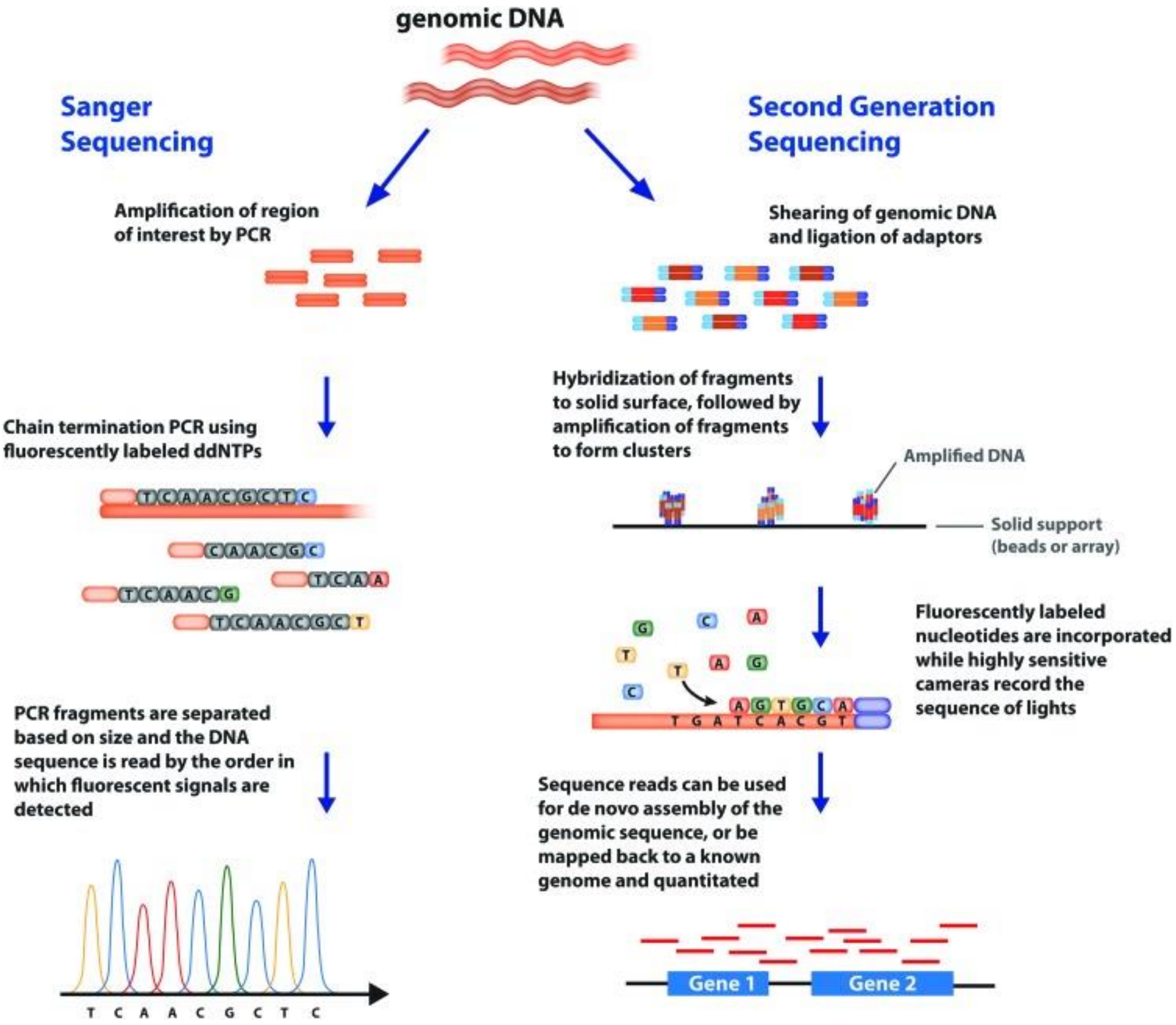

genomic DNA
Sanger Sequencing
Second Generation Sequencing
Amplification of region of interest by PCR
Shearing of genomic DNA and ligation of adaptors
Chain termination PCR using fluorescently labeled ddNTPs
Hybridization of fragments to solid surface, followed by amplification of fragments to form clusters
Amplified DNA
Solid support (beads or array)
TCAACGCTC
CAACGC
TCAA
TCAACG
TCAACGCT
Fluorescently labeled nucleotides are incorporated while highly sensitive cameras record the sequence of lights
AGTGCA
TGATCACGT
PCR fragments are separated based on size and the DNA sequence is read by the order in which fluorescent signals are detected
Sequence reads can be used for de novo assembly of the genomic sequence, or be mapped back to a known genome and quantitated
T C A A C G C T C
Gene 1
Gene 2

Figure 8. Comparison between Sanger sequencing and next-generation sequencing (NGS) technologies. Source: (Bunnik and Le Roch, 2013).

However, it is worth mentioning that NGS approaches have some limitations. Compared with Sanger sequencing, NGS error rates (~0.1-15%) are higher and read lengths, which most frequently range from 50 bp to 400 bp, are shorter (Liu *et al.*, 2012; Goodwin *et al.*, 2016). These problems make mapping and variant calling more difficult and may result in more false positives and false negatives. This is especially important in genetic diagnostics, where a clinical decision that affects a patient has to be made depending on the variants found.

The problem of false positives can be addressed by using an orthogonal validation method to confirm or discard the variants found. For example, SNVs and INDELs can be validated by Sanger sequencing, and CNVs can be validated using orthogonal methods like aCGH or MLPA. However, false negatives represent a more difficult problem. If a variant is not detected by the bioinformatic analysis pipeline, it will be missed unless a suspicion of a specific variant exists from a patient's relative. Therefore, sensitivity has to be considered as a key factor given that false negatives have to be avoided in genetic diagnostics.

Consequently, it is necessary to correctly assess the performance of any diagnostic method before using it in a clinical routine, with a special emphasis on sensitivity. Of course, specificity should ideally also be high, although the consequences of a false positive are easier to manage as we explained before. The assessment of a new method, like setting up a targeted gene panel or a bioinformatic pipeline, should include samples with known variants and a wide range of variant types, sizes, and genes. The more conditions considered, the better the genetic diagnostics method is assessed. Specific guidelines for validating NGS bioinformatic pipelines have been published (Roy *et al.*, 2018).

6.1 WGS, WES and targeted gene panels in genetic diagnostics

WGS, WES, and targeted gene panels present different advantages and limitations for genetic diagnostics. WGS is the least cost-effective approach for diagnostic purposes. For many years, its use was limited to a research context, although the drop in NGS costs facilitated its use in diagnostics, so it has been used in some cases (Van El *et al.*, 2013; Turro *et al.*, 2020). Of course, sequencing the whole genome allows for variant detection in any part of the genome, including variants in non-coding regions. Besides the higher cost of WGS, this approach suffers from other drawbacks. WGS produces an enormous amount of data, which is time-consuming and hard to manage in a demanding laboratory routine. Moreover, detecting variants in the whole genome opens the door to a much larger number of incidental findings that have to be considered in a clinical scenario; they should be discussed and agreed with the patient before initiating the genetic test (Green *et al.*, 2013; Knoppers *et al.*, 2015).

WES produces information for all the protein-coding genes of an individual, that is, a very small portion of the genome. These coding regions are the most explored and understood part of the genome, so their relationship with disease is better known. WES was first introduced as a cost-effective approach when WGS prices remained very high and when it was estimated that the exome contained about 85% of important disease-related variants (Choi *et al.*, 2009; Ng *et al.*, 2009). This approach is especially useful in the genetic diagnostics of very heterogeneous diseases. It is currently used in clinical settings, and its utility does not only apply to the diagnosis of a certain disease. In the context of a universal health system, it can be a cost-effective approach for the genetic diagnostics of multiple hereditary diseases, which may be a matter of interest in the long-term. Similar to WGS, WES results may also include some incidental findings that have to be discussed with the patient.

However, current targeted gene panels are the most cost-effective NGS approach for testing genetically heterogeneous disorders (Kurian *et al.*, 2014; Laduca *et al.*, 2014). Panels focus only on the genes of clinical interest associated with a disease, which make them the best option for genetic diagnostics in terms of cost and coverage depth (Feliubadaló *et al.*, 2017). Sequencing at a much higher depth enables the identification of some rare variants. Also, sequencing only the genes of interest has two other benefits. First, the incidental findings are limited to the genes tested. Second, more samples can be sequenced in a single run, which is an advantage in those demanding diagnostic scenarios where several samples have to be diagnosed routinely. Of course, targeted gene panels suffer from limitations that have to be considered. On one hand, non-coding regions are not included (same as WES), which limits the ability to detect structural and intronic variants. On the other hand, testing can only be successful if the gene causing the disease is included in the panel, which limits the diagnostic yield in the mid-term (Sun *et al.*, 2015).

7 CNV detection strategies from NGS data

NGS CNV detection strategies can be classified into four categories: paired-end mapping, split-read, depth of coverage, and assembly (Figure 9) (Zhao *et al.*, 2013; Pirooznia *et al.*, 2015; Mason-Suares *et al.*, 2016). Each strategy has its strengths and drawbacks, and some tools implement a combination of them. At least 81 NGS CNV detection tools have been developed to date (Tables 5-9).

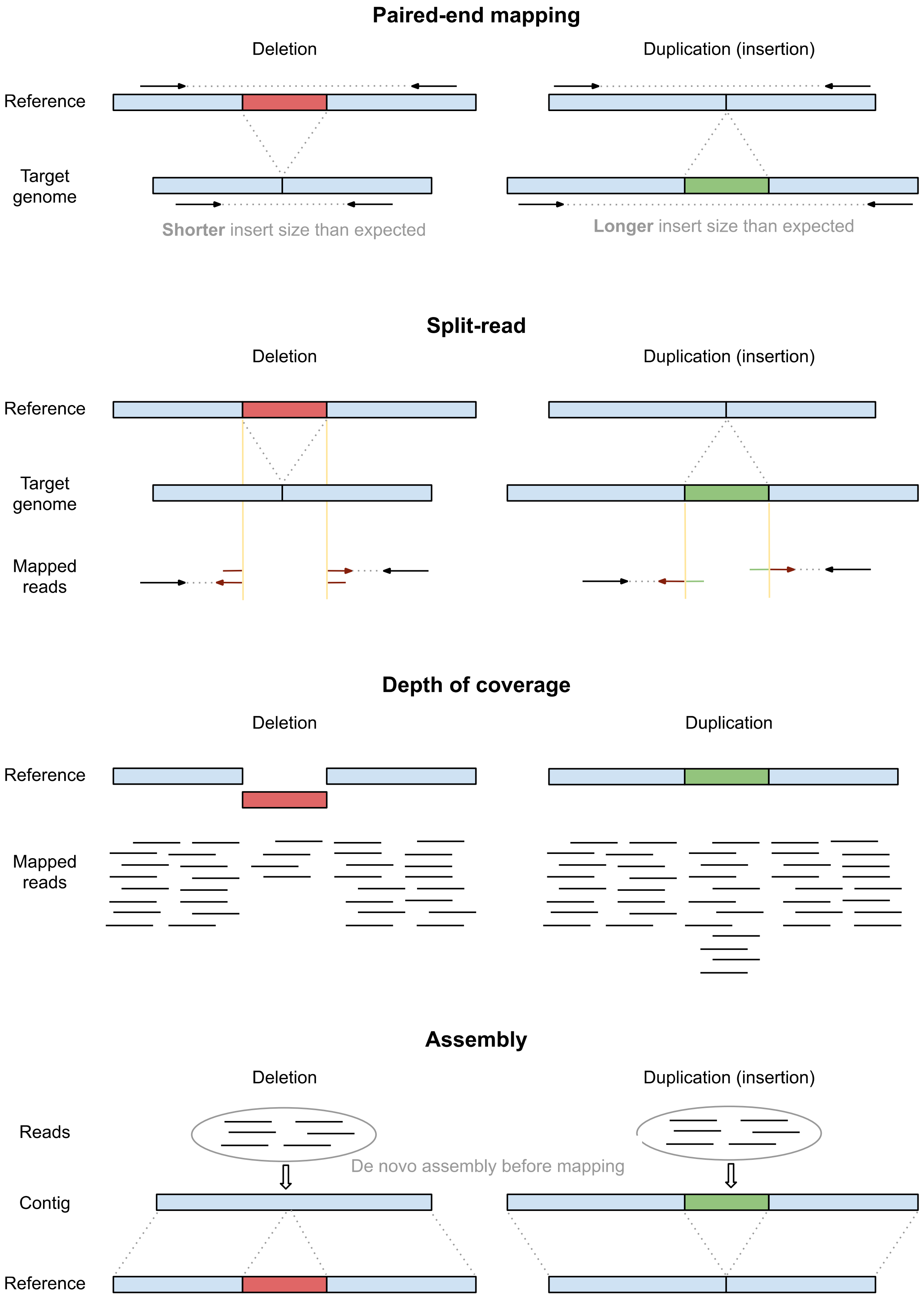


Figure 9. CNV detection methods from NGS data. Adapted from (Rausch *et al.*, 2012; Zhao *et al.*, 2013; Pirooznia *et al.*, 2015).

7.1 Paired-end mapping

Paired-end mapping methods (Table 5), also referred to as read-pair methods, work only on paired-end reads, so DNA fragments are expected to have a specific distribution around the insert size (Korbel *et al.*, 2007). Paired-end methods compare the average insert size between the sequenced paired reads with the expected size from the reference genome, so CNVs can be inferred when mapped paired reads show an unexpected insert size. By using these methods, duplications can be detected only if they are smaller than the average insert size. Moreover, small deletions and duplications cannot be detected because small insert size differences cannot be distinguished from normal variability. Also, the exact number of gains cannot be reported when using paired-end methods.

Table 5. Tools for CNV detection from NGS data using paired-end mapping strategy.

Tool	Language	Availability	Purpose
BreakDancer	Perl, C++	breakdancer.sourceforge.net/	WGS
GASV	Java	code.google.com/p/gasv/	WGS
PEMer	Perl, Python	sv.gersteinlab.org/pemer/	WGS
TARDIS	C	github.com/BilkentCompGen/tardis	WGS
VariationHunter	C	NA	WGS

NA: Not available

7.2 Split-read

Split-read methods (Table 6) also work on paired-end reads. Here, one read is perfectly mapped to the genome while the other partially or totally fails to map. This second read should contain the breakpoint produced by the CNV. Accordingly, split-read methods divide this second read into several fragments and map them to detect the exact breakpoint. The main limitation of this method is that mapping errors can produce false positives and false negatives. As a consequence, calls cannot be reliable in genomic regions where aligners struggle, such as repetitive regions.

Table 6. Tools for CNV detection from NGS data using split-read strategy.

Tool	Language	Availability	Purpose
AGE	C++	sv.gersteinlab.org/age	WGS
Pindel	C++	gmt.genome.wustl.edu/packages/pindel/	WGS
SLOPE	C++	NA	WGS
SRiC	NA	NA	WGS

7.3 Depth of coverage

Depth of coverage or read-depth methods (Table 7) can work on single-end and paired-end reads. Under the hypothesis that depth of coverage differs in the presence of CNVs, these methods call CNVs by comparing depth of coverage within a certain region with the expected calculated depth. First, reads are aligned and coverage is computed for each predefined window. Second, coverage is normalized to account for coverage biases. Third, depth of coverage methods usually use a statistical approach to predict the CNVs based on the expected coverage for a certain genomic region.

These methods can detect the exact number of CNV gains, but they are unable to report the exact breakpoints. Also, their performance is affected by coverage biases like DNA quality, batch effects, repetitive regions, and nucleotide composition. This problem is larger for WES and targeted gene panel data because, for these NGS approaches, DNA capture and amplification produce an additional bias (Tewhey *et al.*, 2009; Aird *et al.*, 2011). Anyway, the depth of coverage strategy is commonly used for CNV detection in WES and targeted gene panels because sparse data makes it difficult to detect breakpoints or consider insert sizes, which causes paired-end mapping and split-read approaches to perform poorly.

Table 7. Tools for CNV detection from NGS data using depth of coverage strategy.

Tool	Language	Availability	Purpose
Atlas-CNV	Perl, R	github.com/theodorc/Atlas-CNV	panel
BIC-seq	Perl, R	compbio.med.harvard.edu/Supplements/PNAS11.html	WGS
BIC-seq2	Perl, R, C	math.pku.edu.cn/teachers/xirb/downloads/software/BICseq2/BICseq2.html	WGS
CANOES	R	columbia.edu/~ys2411//projects/canoes/	WES
Canvas	C#, Python	github.com/Illumina/canvas	WGS / WES
CLAMMS	C, Python, R	github.com/rgcgithub/clamms	WES
CMDS	C, R	github.com/ding-lab/cmds	WGS
cn.MOPS	R	bioinf.jku.at/software/cnmops/	WGS
CNAseg	R	NA	WGS
CNVeM	C	NA	WGS
cnvHMM	C	NA	WGS
CNVkit	Python	github.com/etal/cnvkit	WES / panel
CNVnator	C++, Python	github.com/abyzovlab/CNVnator	WGS
CNVnorm	R	precancer.leeds.ac.uk/cnanorm	WGS
cnvOffSeq	Java	sourceforge.net/projects/cnvoffseq/files/cnvOffSeq/	WES
CNVPanelizer	R	bioconductor.org/packages/release/bioc/html/CNVPanelizer.html	panel
CNV-seq	Perl, R	sourceforge.net/projects/cnv-seq/	WGS

CODEX	R	bioconductor.org/packages/release/bioc/html/CODEX.html	WES
CODEX2	R	github.com/yuchaojiang/CODEX2	WES / panel
CoNIFER	Python	conifer.sf.net/	WES
CONTRA	Python	contra-cnv.sourceforge.net/	WES
Control-FREEC	C++	bioinfo-out.curie.fr/projects/freec/	WGS / WES
CoNVaDING	Perl	github.com/molgenis/CoNVaDING	WES / panel
CONVector	Java, Python, R	github.com/parseq/convector	panel
CopyWriteR	R	bioconductor.org/packages/release/bioc/html/CopywriteR.html	WES / panel
DECoN	R	github.com/RahmanTeam/DECoN/	panel
EXCAVATOR	Perl	sourceforge.net/projects/excavatortool/	WES
ExoCNVTest	Java, R	NA	WES
ExomeCNV	R	cran.r-project.org/src/contrib/Archive/ExomeCNV/	WES
ExomeCopy	R	bioconductor.org/packages/release/bioc/html/exomeCopy.html	WES
ExomeDepth	R	cran.r-project.org/web/packages/ExomeDepth/index.html	WES
FishingCNV	Java, R	sourceforge.net/projects/fishingcnv/	WES
GROM-RD	NA	NA	WGS
iCopyDAV	C++, R	github.com/vogetihrsh/icopydav	WGS
JointSLM	R	academic.oup.com/nar/article/39/10/e65/1309398#82689075	WGS
mrCaNaVar	C	mrcanavar.sourceforge.net/	WGS
panelcn.MOPS	R	www.bioinf.jku.at/software/panelcnmops/	panel
PatternCNV	Perl, R	bioinformaticstools.mayo.edu/research/patterncnv/	WES
PropSeq	R, C	bioinformatics.nki.nl/ocs/	WES
RDXplorer	Python, Shell	rdxplorer.sourceforge.net/	WGS
ReadDepth	R	code.google.com/p/readdepth/	WGS
RSICNV	C, C++	github.com/yhwu/rsicnv	WGS
rSW-seq	C	compbio.med.harvard.edu/Supplements/BMCBioinfo10-2.html	WGS
SegSeq	Matlab	broad.mit.edu/cancer/pub/solexa_copy_numbers/	WGS
SeqCNV	Java, Python	github.com/parseq/convector	WES / panel
VarScan2	Java	sourceforge.net/projects/varscan	WES
VisCap	R	github.com/pughlab/VisCap	WES / panel
XCAVATOR	Fortran, Perl, R	sourceforge.net/projects/excavatortool/	WGS
XHMM	C++	atgu.mgh.harvard.edu/xhmm/	WES

NA: Not available

7.4 Assembly

In assembly methods (Table 8), reads are first assembled to build contigs without a reference genome. Then, contigs are compared with the reference genome to discover structural variants, in particular CNVs. Since short reads are usually used in NGS, assembly methods perform poorly in complex regions like genomic repeats. Moreover, these methods are computationally demanding, so they are not frequently used.

Table 8. Tools for CNV detection from NGS data using assembly strategy.

Tool	Language	Availability	Purpose
Cortex assembler	C	cortexassembler.sourceforge.net/	WGS
Magnolya	Python	sourceforge.net/projects/magnolya/	WGS
TIGRA-SV	C	bioinformatics.mdanderson.org/public-software/archive/tigra/	WGS

NA: Not available

7.5 Combined approaches

As we have seen, each of the four strategies has advantages and disadvantages. Many tools using a combination of approaches have been developed to date (Table 9). Combined approaches aim to achieve more accurate CNV detection by covering the weaknesses of one strategy with the strengths of another.

We have also included an additional strategy in Table 9 that can provide evidence to support or discard the existence of a CNV: the use of SNVs. This source of information has not been frequently included in germline CNV calling algorithms. If a true heterozygous SNV is detected within a germline CNV deletion call, the CNV deletion call may be a false positive. Similarly, the allele frequency of the heterozygous SNVs detected within germline CNV duplications should be either close to 33% or 66%, so values close to 50% provide evidence to discard the CNV duplication call.

Table 9. Tools for CNV detection from NGS data using combined approaches.

Tool	Language	Strategy	Availability	Purpose
Clever-sv	C++, Python	PEM + SR	bitbucket.org/tobiasmarschall/clever-toolkit	WGS
CNVer	Perl, C++	PEM + DOC	compbio.cs.toronto.edu/CNVer/	WGS
cnvHiTSeq	Java	PEM + DOC + SR	sourceforge.net/projects/cnvhitseq/	WGS
CONDEX	Java	DOC + SNVs	code.google.com/p/condr/	WES
DELLY	C++/R	PEM + SR	github.com/tobiasrausch/delly	WGS
ERDS	C, Perl	DOC + SNVs	github.com/JieYang031/erds1.1	WGS
GASVPro	C++	PEM + DOC	code.google.com/p/gasv/	WGS
Genome STRiP	Java, R	PEM + DOC	broadinstitute.org/software/genomestrip	WGS
Gindel	C++	PEM + DOC + SR	sourceforge.net/projects/gindel	WGS
HadoopCNV	Java	DOC + SNVs	github.com/WGLab/HadoopCNV	WGS

HYDRA	Python	PEM + A	code.google.com/p/hydra-sv/	WGS
Hydra-Multi	C++	PEM + A	github.com/arq5x/Hydra	WGS
inGAP-sv	Java	PEM + DOC	ingap.sourceforge.net/	WGS
LUMPY	C++	PEM + DOC + SR	github.com/arq5x/lumpy-sv	WGS
Manta	C++, Python	PEM + SR	github.com/Illumina/manta	WGS
NovelSeq	C	PEM + A	novelseq.sourceforge.net/Home	WGS
PSCC	Perl	PEM + DOC	NA	WGS
SoftSearch	Perl	PEM + SR	code.google.com/p/softsearch	WGS
SVDetect	Perl	PEM + DOC	svdetect.sourceforge.net/	WGS
SVseq	C	PEM + SR	NA	WGS

NA: Not available; PEM: Paired-end mapping; SR: Split-read; DOC: Depth of coverage; A: Assembly; SNVs: Single-nucleotide variants

8 NGS CNV detection from targeted gene panel data

Many bioinformatic tools have been developed with the aim of identifying CNVs from NGS data (Zhao *et al.*, 2013; Abel and Duncavage, 2013; Mason-Suares *et al.*, 2016). However, most tools were developed to work with WGS or WES data. Also, most tools usually have a high performance when detecting large CNVs but have problems detecting small CNVs, those affecting one or a few exons. Nevertheless, these small CNVs are involved in multiple hereditary diseases (Truty *et al.*, 2019). In a diagnostic setting, MLPA and aCGH are the gold standards for CNV testing (Talevich *et al.*, 2016; Kerkhof *et al.*, 2017). Hence, it is a matter of interest to identify an NGS detection tool able to detect single-exon and multi-exon CNVs in NGS panel data with sufficient sensitivity for using as a screening step before MLPA or aCGH.

Multiple benchmarks of CNV calling tools for targeted gene panel data have been published, although they suffer from some deficiencies. Published benchmarks were performed by the own authors of the tools and executed against a single dataset (Johansson *et al.*, 2016; Fowler *et al.*, 2016; Povysil *et al.*, 2017; Kim *et al.*, 2017; Chiang *et al.*, 2019), or used mainly simulated data with a small number of validated CNVs (Roca *et al.*, 2019). Consequently, there is a gap in performing an independent benchmark of multiple CNV calling tools against different datasets generated in diagnostic settings.

9 ICO-IGTP Joint Program on Hereditary Cancer

The ICO-IGTP Joint Program on Hereditary Cancer is an initiative focused on the detection and interpretation of germline variants that predispose to hereditary cancer. Although many hereditary cancer syndromes are considered, the program specializes in the genetic diagnostics of hereditary colorectal cancer (hereditary non-polyposis colorectal cancer and familial adenomatous polyposis), hereditary breast and ovarian cancer, neurofibromatosis type 1 and type 2 (NF1, NF2) and other related disorders such as RASopathies and Phakomatoses. The clinical and genetic heterogeneity of all these syndromes requires multiple gene testing (Laduca *et al.*, 2014), for which targeted gene panel is the most cost-effective approach.

To perform the diagnostic activity, a custom gene panel was developed: I2HCP (Castellanos *et al.*, 2017; Feliubadaló *et al.*, 2017, 2019) (Figure 10), which ranged from 122 to 135 genes (v2.0-v2.2) related to hereditary cancer. The inclusion of all genes of interest in the I2HCP panel allowed for the simplification of laboratory workflows and data management when testing for different clinical conditions. A custom SureSelect bait library was designed using the Agilent eArray to cover a set of regions of interest (ROIs) obtained from translated isoforms of the Ensemble release 67 (GRCh37).

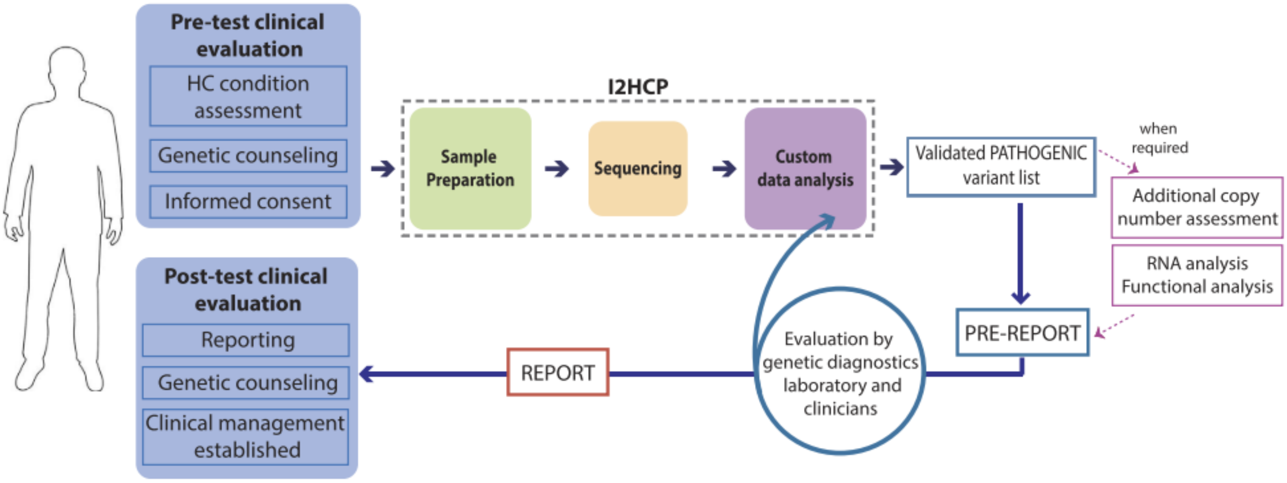

Pre-test clinical evaluation
HC condition assessment
Genetic counseling
Informed consent
I2HCP
Sample Preparation
Sequencing
Custom data analysis
Validated PATHOGENIC variant list
when required
Additional copy number assessment
RNA analysis Functional analysis
PRE-REPORT
Evaluation by genetic diagnostics laboratory and clinicians
REPORT
Post-test clinical evaluation
Reporting
Genetic counseling
Clinical management established

Figure 10. I2HCP diagnostics strategy, including pre- and post-test clinical evaluation. Source: (Castellanos *et al.*, 2017).

Hereditary cancer patients are referred through genetic counselling units based on clinical suspicion. Usually, DNA Isolation is performed from peripheral blood lymphocytes using FlexiGene DNA Kit (Qiagen GmbH, Hilden, Germany) and samples are sequenced in either a

MiSeq with 2×301 bp reads or a HiSeq with 2×251 bp reads. A custom analysis pipeline performs quality check reports, aligns FASTQ reads to the GRCh37 human genome assembly (Ensembl release 67) using BWA-mem, creates sorted bam files using samtools (Li *et al.*, 2009), and calls SNVs and INDELs calling with VarScan2 (Koboldt *et al.*, 2012). Variants obtained for each patient are then evaluated by variant curators, pathogenic and likely pathogenic variants are always confirmed using Sanger.

Aims

This PhD thesis has been carried out with the aim of improving, from a bioinformatic-based approach, the genetic diagnostics of hereditary cancer. More specifically, the aims were:

1. To perform a comprehensive evaluation of tools suitable for detecting CNVs from NGS panel data at single-exon resolution.
2. To select the best candidate tool to implement in the genetic diagnostics pipeline of the ICO-IGTP program on hereditary cancer.
3. After implementing it, to evaluate the impact of including the selected NGS CNV detection tool as a first-tier screening step prior to MLPA validation.
4. To develop a tool to identify false positives produced by germline NGS CNV detection tools.
5. To develop a web-based tool to support the entire diagnostic process during the laboratory routine.

Articles

Article 1 - Evaluation of CNV detection tools for NGS panel data in genetic diagnostics

José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro* & Bernat Gel*

European Journal of Human Genetics. Volume 28, pages 1645–1655(2020).

doi: 10.1038/s41431-020-0675-z

IF 2019 = 3.657. Rank 55/177 (GENETICS & HEREDITY), First tertile, second quartile (Q2).

Journal link: https://www.nature.com/articles/s41431-020-0675-z

Supplementary File 9 is available in Appendix B. All supplementary files (17) are fully available online:

- **Supp File 1** (pdf) Explanation of IBK141 sample exclusion from EGAD00001003335 dataset (panelcnDataset).
- **Supp File 2** (xlsx) MLPA-detected CNVs for ICR96 dataset.
- **Supp File 3** (xlsx) MLPA-detected CNVs for panelcnDataset dataset. (Indication: set of genes tested by MLPA; Truth: MLPA result [CN0: homozygous deletion; CN1: heterozygous deletion; CN2: normal; CN3: duplication (1 copy); CN4: duplication (2 copies)].
- **Supp File 4** (xlsx) MLPA-detected CNVs for in-house MiSeq dataset. (Indication: set of genes tested by MLPA; Truth: MLPA result [CN0: homozygous deletion; CN1: heterozygous deletion; CN2: normal; CN3: duplication (1 copy); CN4: duplication (2 copies)].
- **Supp File 5** (xlsx) MLPA-detected CNVs for in-house HiSeq dataset. (Indication: set of genes tested by MLPA; Truth: MLPA result [CN0: homozygous deletion; CN1: heterozygous deletion; CN2: normal; CN3: duplication (1 copy); CN4: duplication (2 copies)].
- **Supp File 6** (xlsx) Target bed file used for ICR96 and panelcnDataset datasets.
- **Supp File 7** (xlsx) Target bed file used for in-house datasets.
- **Supp File 8** (xlsx) Training and validation sets for all datasets.
- **Supp File 9** (docx) Optimization algorithm description, pseudocode and parameters.

- Supp File 10 (xlsx) Default and sensitivity-optimized parameters for all tools and datasets.
- Supp File 11 (eps) Benchmark results with default parameters: per gene metrics. Shows results when executing tools with the default parameters and computing the per gene metrics. (PPV: positive predictive value; F1: F1 score)
- Supp File 12 (xlsx) Benchmark results when evaluating tools with the default parameters.
- Supp File 13 (xlsx) Benchmark results when evaluating tools with the default parameters dividing the datasets into single-exon and multi-exon.
- Supp File 14 (xlsx) Benchmark results when evaluating tools with the default and sensitivity-optimized parameters against the validation subsets of each dataset.
- Supp File 15 (xlsx) Benchmark results when evaluating tools with the default and sensitivity-optimized parameters against the validation subsets diving them into single-exon and multi-exon.
- Supp File 16 (xlsx) Benchmark results for the diagnostics scenario: metrics on the augmented in-house datasets when executing tools with the optimized parameters in comparison to the default parameters on the validation subsets.
- Supp File 17 (xlsx) Benchmark results for the diagnostics scenario: metrics on the augmented in-house datasets when executing tools with the optimized parameters in comparison to the default parameters on the validation subsets dividing the datasets into single-exon and multi-exon.

European Journal of Human Genetics (2020) 28:1645–1655
https://doi.org/10.1038/s41431-020-0675-z

ARTICLE

Evaluation of CNV detection tools for NGS panel data in genetic diagnostics

José Marcos Moreno-Cabrera[1,2,3] · Jesús del Valle[2,3] · Elisabeth Castellanos[1] · Lidia Feliubadaló[2,3] · Marta Pineda[2,3] · Joan Brunet[2,3,4] · Eduard Serra[1,3] · Gabriel Capellà[2,3] · Conxi Lázaro[2,3] · Bernat Gel[1]

Received: 23 December 2019 / Revised: 21 April 2020 / Accepted: 28 April 2020 / Published online: 19 June 2020

Abstract

Although germline copy-number variants (CNVs) are the genetic cause of multiple hereditary diseases, detecting them from targeted next-generation sequencing data (NGS) remains a challenge. Existing tools perform well for large CNVs but struggle with single and multi-exon alterations. The aim of this work is to evaluate CNV calling tools working on gene panel NGS data and their suitability as a screening step before orthogonal confirmation in genetic diagnostics strategies. Five tools (DECoN, CoNVaDING, panelcn.MOPS, ExomeDepth, and CODEX2) were tested against four genetic diagnostics datasets (two in-house and two external) for a total of 495 samples with 231 single and multi-exon validated CNVs. The evaluation was performed using the default and sensitivity-optimized parameters. Results showed that most tools were highly sensitive and specific, but the performance was dataset dependant. When evaluating them in our diagnostics scenario, DECoN and panelcn.MOPS detected all CNVs with the exception of one mosaic CNV missed by DECoN. However, DECoN outperformed panelcn.MOPS specificity achieving values greater than 0.90 when using the optimized parameters. In our in-house datasets, DECoN and panelcn.MOPS showed the highest performance for CNV screening before orthogonal confirmation. Benchmarking and optimization code is freely available at https://github.com/TranslationalBioinformaticsIGTP/CNVbenchmarkeR.

Introduction

Next-generation sequencing (NGS) is an outstanding technology to detect single-nucleotide variants and small deletion and insertion variants in genetic testing for Mendelian conditions. However, detection of large rearrangements such as copy-number variants (CNV) from NGS data is still challenging due to issues intrinsic to the technology including short read lengths and GC-content bias [1]. Nevertheless, it is well recognized that germline CNVs are the genetic cause of several hereditary diseases [2], so their analysis is a necessary step in a comprehensive genetic diagnostics strategy.

The gold standards for CNV detection in genetic diagnostics are multiplex ligation-dependent probe amplification (MLPA) and array comparative genomic hybridization (aCGH) [3, 4]. Both methods are time consuming and costly, so frequently only a subset of genes is tested, excluding others from the analysis, especially when using single-gene approaches. Therefore, the possibility of using NGS data as a first CNV screening step would decrease the number of MLPA/aCGH tests required and would free up resources.

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0675-z) contains supplementary material, which is available to authorized users.

✉ Conxi Lázaro
clazaro@iconcologia.net

✉ Bernat Gel
bgel@igtp.cat

[1] Hereditary Cancer Group, Program for Predictive and Personalized Medicine of Cancer, Germans Trias i Pujol Research Institute (PMPPC-IGTP), Campus Can Ruti, Badalona, Spain

[2] Hereditary Cancer Program, Joint Program on Hereditary Cancer, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge—IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

[3] Centro de Investigación Biomédica en Red Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain

[4] Hereditary Cancer Program, Catalan Institute of Oncology, IDIBGi, Girona, Spain

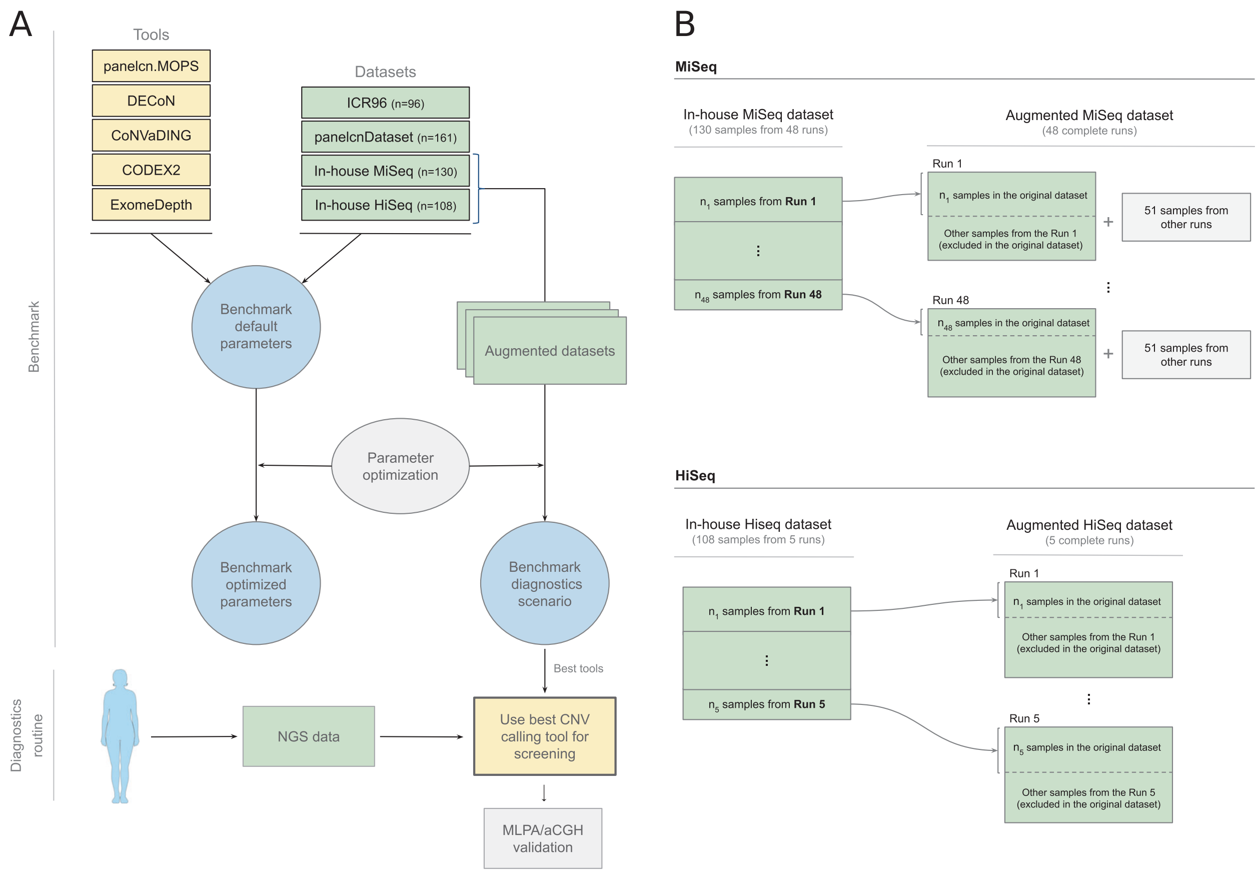

A
Tools
panelcn.MOPS
DECoN
CoNVaDING
CODEX2
ExomeDepth
Datasets
ICR96 (n=96)
panelcnDataset (n=161)
In-house MiSeq (n=130)
In-house HiSeq (n=108)
Benchmark
Benchmark default parameters
Augmented datasets
Parameter optimization
Benchmark optimized parameters
Benchmark diagnostics scenario
Best tools
Diagnostics routine
NGS data
Use best CNV calling tool for screening
MLPA/aCGH validation
B
MiSeq
In-house MiSeq dataset
(130 samples from 48 runs)
Augmented MiSeq dataset
(48 complete runs)
n1 samples from Run 1
n48 samples from Run 48
Run 1
n1 samples in the original dataset
Other samples from the Run 1 (excluded in the original dataset)
51 samples from other runs
Run 48
n48 samples in the original dataset
Other samples from the Run 48 (excluded in the original dataset)
51 samples from other runs
HiSeq
In-house Hiseq dataset
(108 samples from 5 runs)
Augmented HiSeq dataset
(5 complete runs)
n1 samples from Run 1
n5 samples from Run 5
Run 1
n1 samples in the original dataset
Other samples from the Run 1 (excluded in the original dataset)
Run 5
n5 samples in the original dataset
Other samples from the Run 5 (excluded in the original dataset)

Fig. 1 Benchmark design and augmented datasets. a The panel shows the benchmark design and the objective of applying the results in the diagnostics routine. **b** To evaluate the diagnostics scenario, a new dataset was built for each run belonging to the original dataset. The augmented datasets contained all the samples originally sequenced in the run and, in the case of the MiSeq datasets (upper), a set of 51 samples with no known CNV from different runs (MLPA multiplex ligation-dependent probe amplification; aCGH array comparative genomic hybridization; NGS next-generation sequencing; CNV copy-number variant).

Many tools for CNVs detection from NGS data have been developed [5–7]. Most of them can reliably call large CNVs (in the order of megabases) but show poor performance when dealing with small CNVs affecting only one or a few small exons, which are CNVs frequently involved in several genetic diseases [8]. In addition, most of these tools were designed to work with whole-genome or whole-exome data and struggle with the sparser data from NGS gene panels used in routine genetic testing. Therefore, the challenge is to identify a tool able to detect CNVs from NGS panel data at a single-exon resolution with sufficient sensitivity to be used as a screening step in a diagnostic setting.

Other benchmarks of CNV calling tools on targeted NGS panel data have been published. However, they were performed by the authors of the tools and executed against a single dataset [9–13], or used mainly simulated data with a small number of validated CNVs [14]. The aim of this work is to perform an independent benchmark of multiple CNV calling tools, optimizing, and evaluating them against multiple datasets generated in diagnostics settings, to identify the most suitable tools to be used for genetic diagnostics (Fig. 1).

Materials and methods

Datasets and tools

Four datasets were included in this benchmark (ICR96 exon CNV validation series [15], panelcnDataset [11], In-house MiSeq and In-House HiSeq) (Table 1) with data from two hybridization-based target capture NGS panels designed for hereditary cancer diagnostics: TruSight Cancer Panel (Illumina, San Diego, CA, USA) and I2HCP [16]. All datasets were generated in real diagnostics settings and contained single and multi-exon CNVs, all of them validated by MLPA. Negative MLPA data, meaning no detection of any CNV, were also available for a subset of genes. Detailed information on MLPA-detected CNVs for each dataset can be found in Supplementary files 2–5.

Table 1 Datasets used in the benchmark.

	Samples	Validated genes with CNV	Single-exon CNVs	Multi-exon CNVs	Deletion CNVs	Duplication CNVs	Validated genes with no CNV	Sequencing	Availability	Additional information
ICR96	96	68	25	43	51	17	1752 (96.3% of total)	TruSight Cancer Panel v2 (100 genes), HiSeq, 2 × 101 bp reads	European genome-phenome Archive EGAD00001003335	Samples obtained from one run
panelcnDataset	161	41	13	28	36	5	416 (91% of total)	TruSight Cancer Panel (94 genes), MiSeq, 2 × 151 bp reads	European Genome-phenome Archive EGAS00001002481	Only 161 of 170 samples were used. See Supplementary file 1
In-house MiSeq	130	64	19	45	56	8	167 (72.3% of total)	I2HCP Panel v2.0–v2.2 (122–135 genes), MiSeq, 2 × 300 bp reads	European Genome-phenome Archive EGAS00001004316	Samples obtained from 48 runs. Three samples had a CNV in mosaicism
In-house HiSeq	108	58	18	40	49	9	176 (75.2% of total)	I2HCP panel v2.0–v2.2 (122–135 genes), HiSeq, 2 × 251 bp reads	European Genome-phenome Archive EGAS00001004316	Samples obtained from 5 runs. Two samples had CNV in mosaicism

Samples from the In-house MiSeq and in-house HiSeq datasets were generated at the ICO-IGTP Joint Program for Hereditary Cancer and are available at the EGA under the accession number EGAS00001004316. In addition to these samples, a total of 1103 additional samples (505 MiSeq and 598 HiSeq), with no CNVs detected in the subset of genes tested by MLPA, were used to build the augmented datasets used in the diagnostics scenario analysis. Informed consent was obtained for all samples in the in-house datasets.

Five tools were tested in the benchmark (Table 2): CoNVaDING v1.2.0 [9], DECoN v1.0.1 [10], panelcn.MOPS v1.0.0 [11], ExomeDepth v1.1.10 [17], and CODEX2 v1.2.0 [18].

Data preprocessing

All samples were aligned to the GRCh37 human genome assembly using BWA mem v0.7.12 [19, 20]. SAMtools v0.1.19 [21] was used to sort and index BAM files. No additional processing or filtering was applied to the BAM files.

Regions of interest

The regions of interest (ROIs) were dependent on the dataset. For TruSight based datasets, ICR96 and panelcn-Dataset, we used the targets bed file published elsewhere [10] with some modifications: the fourth column was removed, the gene was added and it was sorted by chromosome and start position (Supplementary file 6). For in-house datasets, we generated a target bed file containing all coding exons from all protein-coding transcripts of genes in the I2HCP panel v2.1 (Supplementary file 7). These data were retrieved from Ensembl BioMart version 67 [22] (http://may2012.archive.ensembl.org). All genes tested by MLPA and used in the benchmark were common to all I2HCP versions (v2.0-2.2).

Benchmark evaluation metrics

The performance of each tool for CNVs detection was evaluated at two levels: per ROI and per gene.

Per ROI metrics treated all ROI as independent entities, assigning each of them a correctness value: true positive (TP) or true negative (TN) if the tool matched the results of MLPA, false negative (FN) if the tool missed a CNV detected by MLPA and false positive (FP) if the tool called a CNV not detected by MLPA. This is the most fine-grained metric.

Per gene metrics consider the fact that most MLPA kits cover a whole gene and so the true CNVs would be detected by MLPA when confirming any CNV call in any ROI of the affected gene. Therefore, per gene metrics assigned a correctness value to each gene taking into account all its exons: TP if one of its ROIs was a TP; FN if MLPA detected a CNV in at least one of its ROIs and none of them were detected by the tool; FP if the tool called a CNV in at least one ROI and none of them were detected by MLPA; TN if neither MLPA nor the tool detected a CNV in any of its ROIs.

For each tool against each dataset and evaluation level various performance metrics were computed: sensitivity defined as TP/(TP + FN), specificity defined as TN/(TN + FP), positive predictive value (PPV) defined as TP/(TP + FP), negative predictive value (NPV) defined as TN/(TN + FN), false negative rate (FNR) defined as FN/(FN + TP), false positive rate (FPR) defined as FP/(FP + TN), and F1 score (F1) defined as 2TP/(2TP + FP + FN).

Parameter optimization

Parameters of each tool were optimized against each dataset to maximize sensitivity while limiting specificity loss: each dataset was split into two halves, a training set used to optimize tool parameters and a validation set to evaluate them (Supplementary file 8). The optimization algorithm followed a greedy approach: a local optimization was performed at each step with the aim of obtaining a solution close enough to the global optimum. Further details of the optimization algorithm can be found in Supplementary file 9.

Benchmarking framework execution

An R framework, CNVbenchmarkeR, was built to perform the benchmark in an automatically and configurable way. Code and documentation are available at https://github.com/TranslationalBioinformaticsIGTP/CNVbenchmarkeR. Each selected tool was first executed against each dataset using default parameters as defined in tool documentation and then using the optimized parameters. Default and optimized parameter values can be found in Supplementary file 10. Tool outputs were processed with R v3.4.2, Bioconductor v3.5 [23], plyr [24], GenomicRanges [25], and biomaRt [26]. Plots were created with ggplot2 [27]. Confidence intervals (CIs) were calculated with epiR v1.0-14 at a CI of 95%. In addition, for each dataset, all executions were repeated to compare performance on two subsets: one excluding single-exon CNVs samples and one excluding multi-exon CNVs samples.

Diagnostics scenario evaluation

The In-house MiSeq and In-house HiSeq datasets were composed of a selection of samples from different sequencing runs. In a real diagnostics scenario, the objective is to analyze a new run with all its sequenced samples. To

Table 2 Tools tested in the benchmark.

	Language	Version	Number of parameters used in the benchmark	Reports no calls	Availability	Methods
CODEX2	R package	1.2.0[a]	10	No	https://github.com/yuchaojiang/CODEX2	Based on CODEX package, it models the GC content bias and normalizes the read depth data for CNV detection via a Poisson latent factor model.
CoNVaDING	Perl program	1.2.0	7	Yes	https://github.com/molgenis/CoNVaDING	Combination of ratio scores and Z-scores of the sample of interest compared to the selected normalized control samples.
DECoN	R program	1.0.1	3	Yes	https://github.com/RahmanTeam/DECoN	Modifies ExomeDepth package by altering the hidden Markov model probabilities to depend upon the distance between exons.
ExomeDepth	R package	1.1.10	4	No	https://github.com/vplagnol/ExomeDepth	Beta-binomial model with GC correction and hidden Markov model to combine likelihood across exons.
panelcn.MOPS	R package	1.0.0	13	Yes	https://github.com/bioinf-jku/panelcn.mops	Adaptation of cn.MOPS package, which decomposes variations in coverage across samples into integer copy numbers and noise by means of its mixture components and Poisson distributions.

[a] CODEX2 script for panel setting (Codex2_targeted.R) was obtained from version dated at on Sep 12, 2017.

simulate and evaluate the diagnostics scenario, we built the augmented datasets (Fig. 1), which contained all the samples from the sequencing runs instead of a selection of them. For the augmented datasets, the tools were executed against each run and metrics were computed by combining the results of all runs. Since some tools recommend more than 16 samples for optimal performance, we added 51 samples from other runs with no known CNVs when executing the tools on the runs of the augmented MiSeq dataset.

We also defined a new metric, whole diagnostics strategy, to take into account that in a diagnostics setting all regions where the screening tool was not able to produce a result (no call) should be identified and tested by other methods. Thus, any gene containing at least one positive call or no call in a ROI was considered as a positive call of the whole gene: TP if the gene contained at least one ROI affected by a CNV; FP if the gene did not contain any ROI affected by a CNV. In addition, if a tool identified a ROI both as a deletion and a duplication, it was considered a no call when computing metrics.

Results

To identify the CNV calling tools that could be used as a screening step in a genetic diagnostics setting, we needed first to select the candidate tools, and then to evaluate their performance with a special emphasis on the sensitivity, both with their default parameters and with dataset-dependent optimized parameters.

CNV calling tool selection

The first in the benchmark was to identify candidate tools that have shown promising results. After a literature search process, we selected five CNV calling tools to be evaluated (Table 2), all of them based on depth-of-coverage analysis. Three tools have been reported to perform well on NGS panel data at single-exon resolution: CoNVaDING v1.2.0 [9], DECoN v1.0.1 [10], and panelcn.MOPS v1.0.0 [11]. ExomeDepth v1.1.10 [17] was included due to its high performance in benchmarks on WES data [28, 29] and because the developers reported good performance with panel data (https://github.com/vplagnol/ExomeDepth). CODEX2 v1.2.0 was included due to the high sensitivity shown on WES data [18] and the availability of specific scripts for panel data (https://github.com/yuchaojiang/CODEX2).

Benchmark with default parameters

We executed each tool on each dataset with the default parameters and computed evaluation statistics at two levels: per ROI and per gene (see "Methods").

Regarding the per ROI metric, most tools showed sensitivity and specificity values over 0.75, with sensitivity in general over 0.9 (Fig. 2 and Table 3). However, tool performance varied across datasets. For the ICR96 and panelcnDataset datasets, specificity was always higher than 0.98, while sensitivity remained higher than 0.94 (with the exception of CODEX2). This performance was not achieved when using the in-house datasets, where lower sensitivity and specificity can be observed, and only CoNVaDING obtained sensitivity close to 1 at the expense of a lower specificity.

As expected in unbalanced datasets with a much larger number of negative elements than positive ones, NPV was higher than the PPV in all tool-dataset combinations. All NPVs were above 0.96 while PPV varied across datasets, ranging from 0.36 (CoNVaDING in ICR96) to 0.96 (ExomeDepth in In-house MiSeq). ExomeDepth had the highest PPV in all datasets.

Regarding the per gene metric, sensitivity was slightly improved compared to per ROI, and for each dataset, at least one tool showed a sensitivity of 1 and was able to identify all CNVs (Supplementary files 11 and 12).

When excluding single-exon CNVs or multi-exon CNVs, the exclusion of single-exon CNVs generally provided a better PPV, while sensitivity varied depending on the dataset (Supplementary file 13).

Benchmark with optimized parameters

In addition to evaluating the performance of the different tools tested with default parameters, we performed an optimization process to identify, for each tool and dataset, the combination of parameters that maximized the sensitivity as required for a screening tool in a diagnostics context (see "Methods" and Supplementary files 8 and 9).

Parameter optimization was performed on a subset (training) of each dataset and the optimized parameters (Supplementary file 10) were compared to the default ones on the samples not used for training (validation subset). Figure 3 shows the optimization results at the ROI level. In general, the optimization process improved sensitivity by slightly decreasing specificity. For panelcnDataset, sensitivity was increased by a higher margin driven by CODEX2, which increased its sensitivity by 58.6%. On the other hand, tools were not able to improve or showed small differences in the In-house MiSeq dataset (Supplementary files 14 and 15).

Benchmark in a diagnostics scenario

In a real diagnostic setting, all CNVs detected in genes of interest and all regions where the screening tool was not able to produce a result (no call) should be confirmed by an

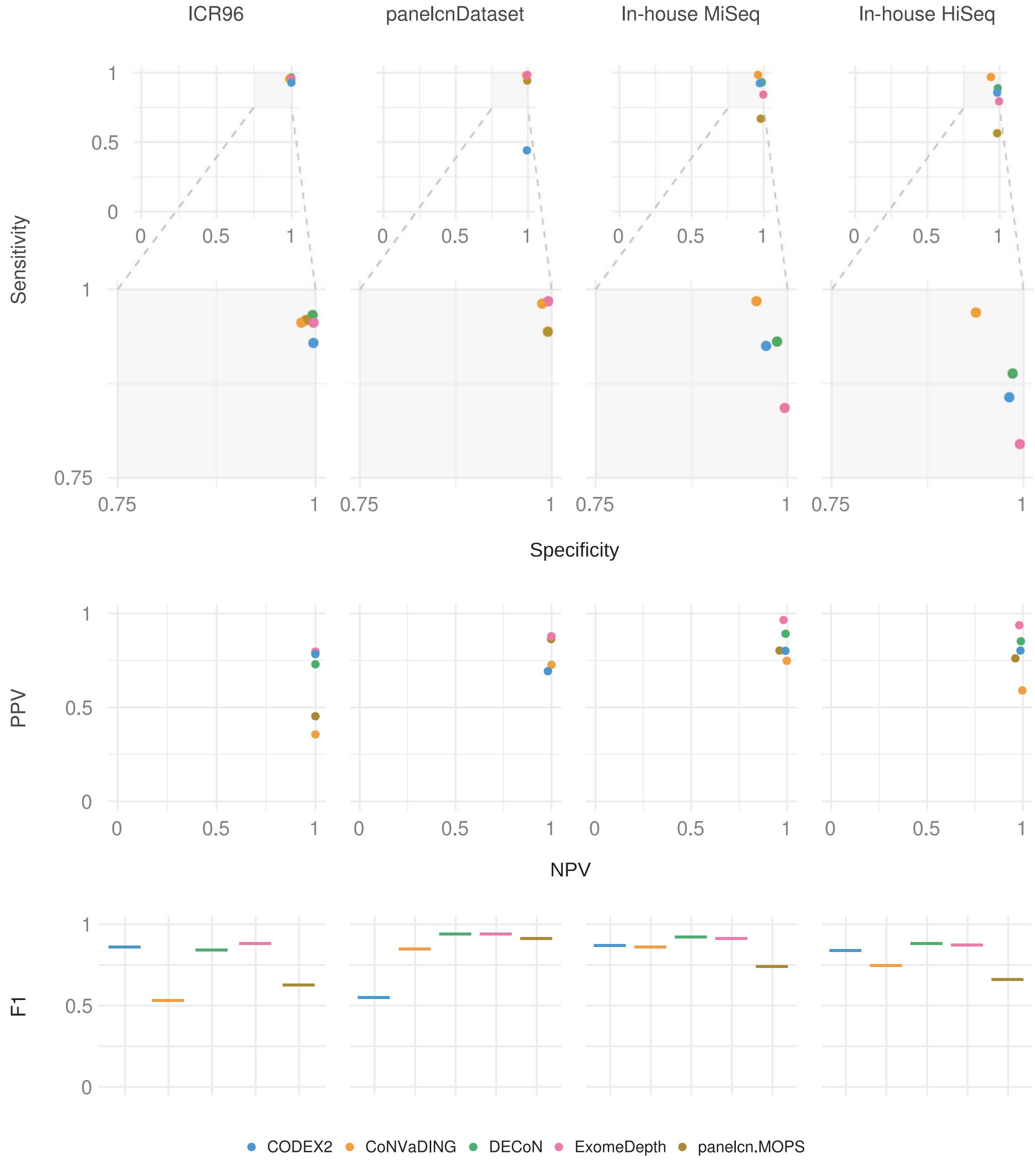


Fig. 2 Benchmark results with default parameters: per ROI metrics. Shows results when executing tools with the default parameters and computing the per ROI metrics. ExomeDepth and DECoN tools obtained same sensitivity and specificity in panelcnDataset (ROI region of interest; PPV positive predictive value; F1 F1 score).

orthogonal technique. To account for this, we evaluated the performance of all tools using the whole diagnostics strategy metric which takes the no calls into account. This evaluation was performed in a modified version of the in-house datasets, the augmented in-house datasets (Fig. 1), which contained all the samples from the original sequencing runs instead of a selection of them (see “Methods”).

Figure 4 shows sensitivity and specificity on the augmented in-house datasets when executing tools with the optimized parameters compared to the default parameters. For the In-house MiSeq dataset, two tools detected all CNVs: panelcn.MOPS achieved it with both optimized and default parameters (CI: 94.4–100%), with a specificity of 67.8% (CI: 60.3–74.8%) and 80.7% (CI: 74.0–86.3%), respectively. DECoN detected all CNVs only with the optimized parameters (CI: 94.4–100%) reaching 91.3% (CI: 86.0–95.0%) specificity. CoNVaDING also detected all CNVs, but its high no-call rate led to very low specificity, 4.1% (CI: 1.6–8.2%). For the In-house HiSeq dataset, only panelcn.MOPS detected all CNVs (CI: 93.8–100%) with an acceptable specificity (81.5% (CI: 75.0–86.9%) and 83.2% (CI: 76.8–88.3%) with the default and optimized parameters respectively). DECoN missed one CNV, being a mosaic sample, and its specificity remained high, 96.6% (CI: 92.8–98.8%) with the optimized parameters. On the other hand, CODEX2 and ExomeDepth obtained high sensitivity and specificity values for both datasets, but they did not report no calls (Table 4 and Supplementary files 16 and 17).

Table 3 Bechmark results with default parameters and per ROI metrics.

Dataset	Tool	TP	TN	FP	FN	Total	Sensitivity	Specificity	PPV	NPV	F1	FNR	FPR
ICR96	DECoN	286	28473	106	10	28875	0.9662	0.9963	0.7296	0.9996	0.8314	0.0338	0.0037
	panelcn.MOPS	284	28236	343	12	28875	0.9595	0.988	0.453	0.9996	0.6154	0.0405	0.0120
	CoNVaDING	283	28068	511	13	28875	0.9561	0.9821	0.3564	0.9995	0.5193	0.0439	0.0179
	exomedepth	283	28507	72	13	28875	0.9561	0.9975	0.7972	0.9995	0.8694	0.0439	0.0025
	CODEX2	275	28503	76	21	28875	0.9291	0.9973	0.7835	0.9993	0.8501	0.0709	0.0027
panelcnDataset	DECoN	317	9442	44	5	9808	0.9845	0.9954	0.8781	0.9995	0.9283	0.0155	0.0046
	panelcn.MOPS	304	9438	48	18	9808	0.9441	0.9949	0.8636	0.9981	0.9021	0.0559	0.0051
	CoNVaDING	316	9367	119	6	9808	0.9814	0.9875	0.7264	0.9994	0.8349	0.0186	0.0125
	exomedepth	317	9442	44	5	9808	0.9845	0.9954	0.8781	0.9995	0.9283	0.0155	0.0046
	CODEX2	142	9423	63	180	9808	0.441	0.9934	0.6927	0.9813	0.5389	0.5590	0.0066
In-house MiSeq	DECoN	486	4189	59	36	4770	0.931	0.9861	0.8917	0.9915	0.911	0.0690	0.0139
	panelcn.MOPS	349	4162	86	173	4770	0.6686	0.9798	0.8023	0.9601	0.7294	0.3314	0.0202
	CoNVaDING	513	4076	173	8	4770	0.9846	0.9593	0.7478	0.998	0.85	0.0154	0.0407
	exomedepth	440	4232	16	82	4770	0.8429	0.9962	0.9649	0.981	0.8998	0.1571	0.0038
	CODEX2	483	4128	120	39	4770	0.9253	0.9718	0.801	0.9906	0.8587	0.0747	0.0282
In-house HiSeq	DECoN	351	4197	61	44	4653	0.8886	0.9857	0.8519	0.9896	0.8699	0.1114	0.0143
	panelcn.MOPS	223	4188	70	172	4653	0.5646	0.9836	0.7611	0.9606	0.6483	0.4354	0.0164
	CoNVaDING	382	3994	265	12	4653	0.9695	0.9378	0.5904	0.997	0.7339	0.0305	0.0622
	exomedepth	314	4237	21	81	4653	0.7949	0.9951	0.9373	0.9812	0.8603	0.2051	0.0049
	CODEX2	324	4195	80	54	4653	0.8571	0.9813	0.802	0.9873	0.8286	0.1429	0.0187

TP true positive, *TN* true negative, *FP* false positive, *FN* false negative, *PPV* positive predictive value, *NPV* negative predictive value, *F1* F1 score, *FNR* false negative rate, *FPR* false positive rate.

Discussion

CNVs are the genetic cause of multiple hereditary diseases [2]. To detect them, specific tools and techniques are required. In genetic diagnostics, this is mainly done using either MLPA and aCGH or using software tools to infer copy-number alterations from NGS data generated in the diagnostics process. MLPA and aCGH are the gold standard methods [3], but both are time-consuming and expensive approaches that frequently lead laboratories to only use them in a subset of genes of interest. On the other hand, multiple tools for CNV calling from NGS data have been published [5–7], but their performance on NGS gene panel data has not been properly evaluated in a genetic diagnostics context. This evaluation is especially critical when these tools are used as a screening step in a diagnostics strategy, since a nonoptimal sensitivity would lead to a higher number of misdiagnosis.

Most CNV calling tools have not been developed to be used as a screening step in genetic diagnostics but as part of a research-oriented data analysis pipeline. Therefore, they were originally tuned and optimized for a certain sensitivity-specificity equilibrium. To be used as screening tools, we need to alter their default parameters to shift that equilibrium toward maximizing the sensitivity even at the expense of lowering their specificity. This parameter optimization must be performed in a dataset-specific way, since tools show performance differences between dataset due to dataset specificities coming from target regions composition, technical differences, or sequencing characteristics.

In this work, we selected 5 tools that have shown promising results on panel data, and we measured their performance, with the default and sensitivity-optimized parameters, over 4 validated datasets from different sources: a total of 495 samples with 231 single and multi-exon CNVs. CNVbenchmarkeR, a framework for evaluating CNV calling tools performance, was developed to undertake this task. We also evaluated their performance in a genetic diagnostics-like scenario and showed that some of the tools are suitable to be used as screening methods before MLPA or aCGH confirmation.

Benchmark with default parameters

The benchmark with default parameters showed that most tools are highly sensitive and specific, but the top performers depend on the specific dataset. Most tools performed best when using data from panelcnDataset. DECoN, ExomeDepth and CoNVaDING reached almost 100% sensitivity and specificity. A possible reason for this is that this dataset contains the lowest number of single-exon CNVs ($n = 13$), which are the most difficult type of CNVs

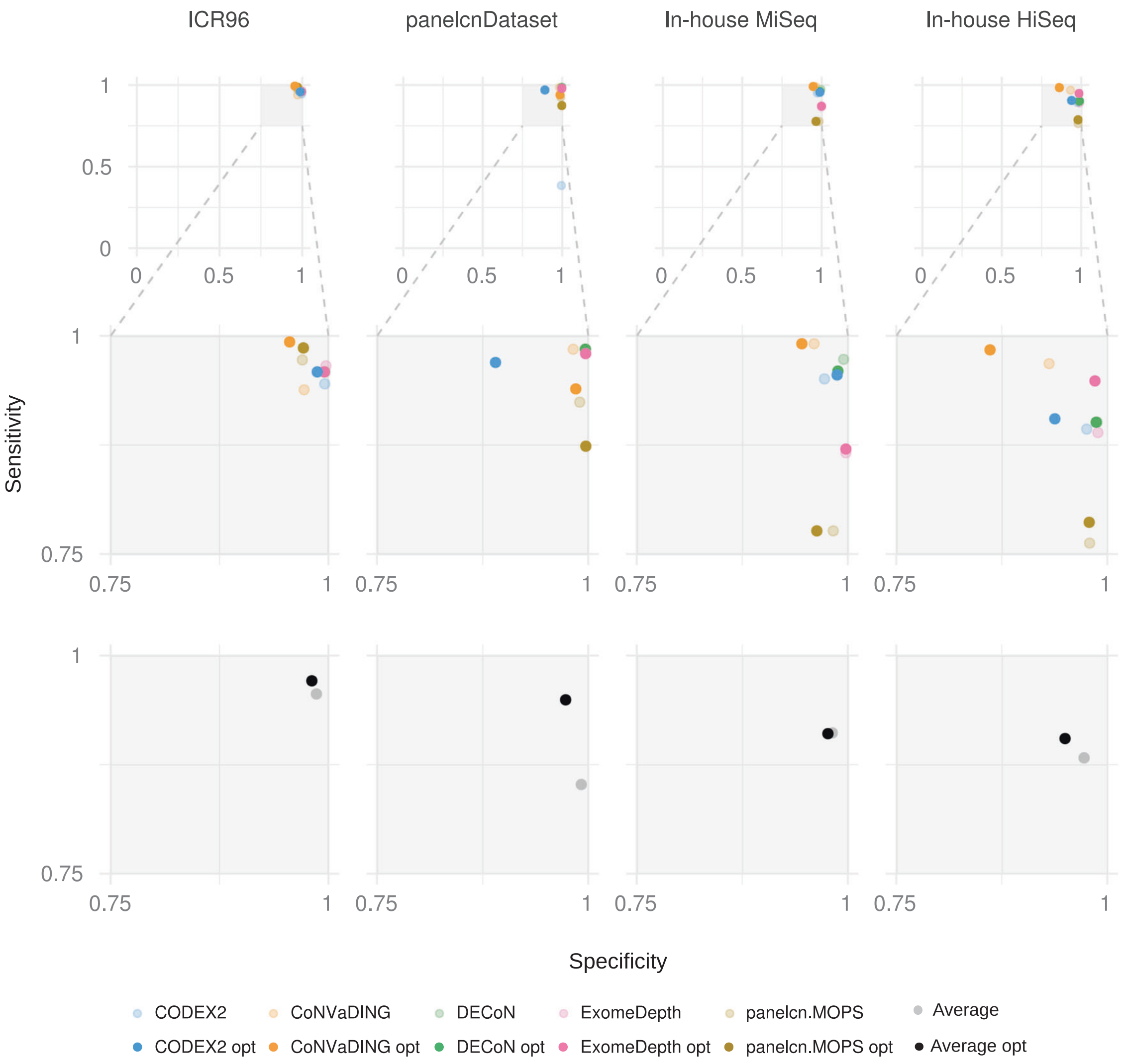


Fig. 3 Optimization results at ROI level. Shows sensitivity and specificity on validation sets when executing tools with the optimized parameters in comparison to the default parameters (ROI region of interest).

to be detected. DECoN was the best performer for ICR96, a dataset published by the same authors, but other tools obtained similar results in that dataset. CoNVaDING was the most sensitive tool when analyzing our in-house datasets but showed the lowest PPV in all datasets with the exception of panelcnDataset. ExomeDepth showed the highest PPV in all datasets, making it one of the most balanced tools regarding sensitivity and specificity. Differences in tool performance depending on the dataset were also observed in previous works [29, 30].

Optimization

The different CNV calling tools included in this work were originally designed with different aims with respect to their preferred sensitivity and specificity equilibrium or the type of CNVs they expected to detect, and this is reflected in their default parameters and their performance in the initial benchmark. Our aim with this work was to evaluate these CNV callers as potential screening tools in a genetic diagnostics setting and for this reason, we required their maximum sensitivity.

The parameter optimization process allowed us to determine the dataset-specific parameter combination maximizing their sensitivity without an excessive specificity loss. The optimization had a different impact on different tools: while CODEX2 showed a higher sensitivity in all four datasets the rest of the tools showed modest improvements. This is mainly due to the fact that sensitivity was already over 0.9 for most combinations and the number of false negatives to correctly call was small (between 4 and 8) in the per gene metric.

The final optimized parameters were dataset specific, so we do not recommend using them directly on other datasets

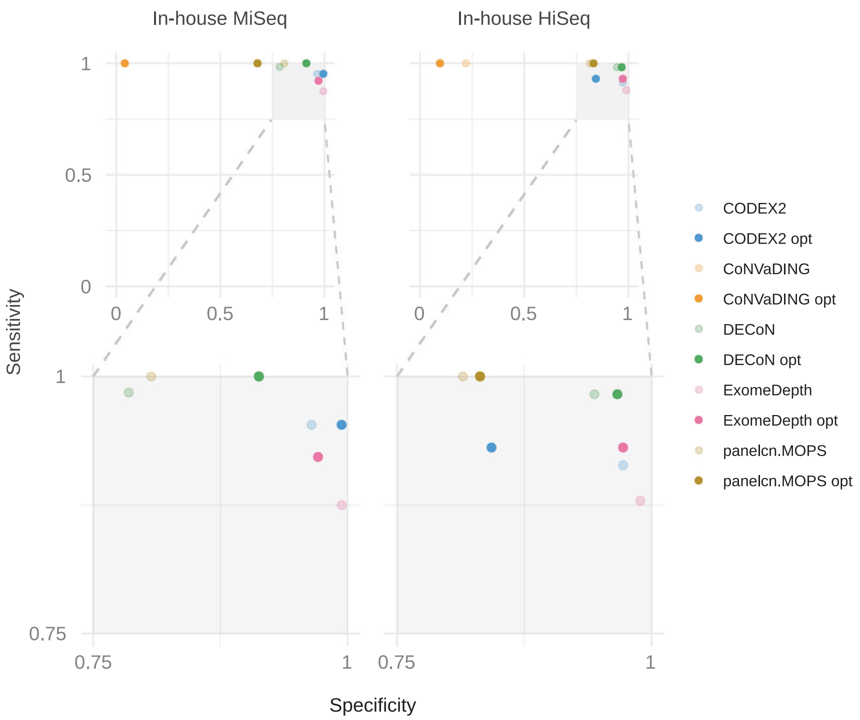


Fig. 4 Benchmark results for the diagnostics scenario: whole diagnostics strategy metrics. Shows sensitivity and specificity on the augmented in-house datasets when executing tools with the optimized parameters in comparison to the default parameters.

Table 4 Benchmark results with default and optimized parameters in the diagnostics scenario.

Dataset	Parameters	Tool	TP	TN	FP	FN	Sensitivity	Specificity	F1
In-house MiSeq	Default parameters	DECoN	63	135	37	1	0.9844	0.7849	0.7683
		panelcn.MOPS	64	138	33	0	1	0.807	0.795
		CoNVaDING	64	7	165	0	1	0.0407	0.4369
		exomedepth	56	171	1	8	0.875	0.9942	0.9256
		CODEX2	61	163	6	3	0.9531	0.9645	0.9313
	Optimized parameters	DECoN	64	157	15	0	1	0.9128	0.8951
		panelcn.MOPS	64	116	55	0	1	0.6784	0.6995
		CoNVaDING	64	7	165	0	1	0.0407	0.4369
		exomedepth	59	167	5	5	0.9219	0.9709	0.9219
		CODEX2	61	168	1	3	0.9531	0.9941	0.9683
In-house HiSeq	Default parameters	DECoN	57	168	10	1	0.9828	0.9438	0.912
		panelcn.MOPS	58	145	33	0	1	0.8146	0.7785
		CoNVaDING	58	39	139	0	1	0.2191	0.4549
		exomedepth	51	176	2	7	0.8793	0.9888	0.9189
		CODEX2	53	173	5	5	0.9138	0.9719	0.9138
	Optimized parameters	DECoN	57	172	6	1	0.9828	0.9663	0.9421
		panelcn.MOPS	58	148	30	0	1	0.8315	0.7945
		CoNVaDING	58	17	161	0	1	0.0955	0.4188
		exomedepth	54	173	5	4	0.931	0.9719	0.9231
		CODEX2	54	150	28	4	0.931	0.8427	0.7714

where the data have been obtained differently (different capture protocol or sequencing technologies, for example).

Based on our results, we would recommend optimizing the parameters for each specific dataset before adding any CNV calling tool to a genetic diagnostics pipeline to maximize its sensitivity and reduce the risk of misdiagnosis. To that end, we have developed an R framework, CNVbenchmarkeR (freely available at https://github.com/TranslationalBioinformaticsIGTP/CNVbenchmarkeR), that will help to perform the testing and optimization process in any new dataset.

Diagnostics scenario

Two tools showed performance good enough to be implemented as screening methods in the diagnostics scenario evaluated in our two in-house datasets (Fig. 4): DECoN and panelcn.MOPS. While panelcn.MOPS was able to detect all CNVs both with the default and the optimized parameters, DECoN reached almost perfect sensitivity and outperformed panelcn.MOPS specificity when using the optimized parameters, although the difference is not statistically significant. DECoN only missed a mosaic CNV affecting two exons of the NF2 gene. CoNVaDING also detected all CNVs, but the high number of no-call regions reduced its specificity to values between 4.1 and 21.9%, which rendered it non-valid as a screening tool.

The parameter optimization process improved the sensitivity of most tools. For example, for the In-house MiSeq dataset, DECoN sensitivity increased from 98.4% (CI: 91.6–100%) to 100% (CI: 94.4–100%), and the specificity increased from 78.5% (CI: 71.6–84.4%) to 91.3% (CI: 86.0–95.0%). This improvement highlights the importance of fine-tuning the tool parameters for each specific task, and shows that the optimization process performed in this work has been key for the evaluation of the different tools.

The high sensitivity reached by DECoN and panelcn. MOPS in different datasets, where they identified all known CNVs, shows that NGS data can be used as a CNV screening step in a genetic diagnostics setting. This screening step has the potential to improve the diagnostics routines. As an example, the high specificity reached by DECoN in the in-house MiSeq dataset with the optimized parameters means that around 91% of genes with no CNV would not need to be specifically tested for CNVs when using DECoN as a screening step. The resources saved by the reduction in the number of required tests could be used to expand the number of genes analyzed, potentially increasing the final diagnostics yield.

In conclusion, according to our analysis, DECoN and panelcn.MOPS provide the highest performance for CNV screening before orthogonal confirmation. Although panelcn.MOPS showed a slightly higher sensitivity in one of the datasets, DECoN showed a much higher specificity in the diagnostics scenario. Our results also showed that tools performance depends on the dataset. Therefore, it may be important to evaluate potential tools on an in-house dataset before implementing one as a screening method in the diagnostics routine.

Acknowledgements This study makes use of the ICR96 exon CNV validation series data generated by Professor Nazneen Rahman's team at The Institute of Cancer Research, London as part of the TGMI. We are grateful to the Katharina Wimmer team at Division Human Genetics, Medical University Innsbruck for providing access to the dataset deposited at EGA and hosted by the EBI, under the accession number EGAS00001002481. We thank the participating patients and all the members of the Unit of Genetic Diagnostics of the Hereditary Cancer Program of the Catalan Institute of Oncology (ICO-IDIBELL) and the Genetics Diagnostics Unit of the Hereditary Cancer Group of the Germans Trias i Pujol Research Institute (IGTP). We also thank the IGTP HPC Core Facility, Iñaki Martínez de Ilarduya and Adriana López-Doriga for their help. We thank CERCA Programme/Generalitat de Catalunya for institutional support.

Funding This work has been supported by: the Spanish Ministry of Science and Innovation, Carlos III Health Institute (ISCIII), Plan Estatal de I + D + I 2013–2016, and co-financed by the FEDER program; the Government of Catalonia, the Spanish Association Against Cancer (AECC) and Fundació La Marató de TV3. Contract grant numbers: ISCIIIRETIC RD06/0020/1051, RD12/0036/008, PI11/1609, PI13/00285, PIE13/00022, PI14/00577, PI15/00854, PI16/00563, PI19/00553, 2017SGR1282, and 2017SGR496.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28:2711–8.
2. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
3. Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, et al. Clinical validation of copy number variant detection from targeted next-generation sequencing panels. J Mol Diagn. 2017;19:905–20.
4. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol. 2016;12:1–18.
5. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinforma. 2013;14:S1.
6. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. 2013;206:432–40.

7. Mason-Suares H, Landry L, S. Lebo M. Detecting copy number variation via next generation technology. Curr Genet Med Rep. 2016;4:74–85.
8. Truty R, Paul J, Kennemer M, Lincoln SE, Olivares E, Nussbaum RL, et al. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet Med. 2019;21:114–23.
9. Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JDH, van der Hout AH, et al. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. Hum Mutat. 2016;37:457–64.
10. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016;1:1–20.
11. Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, et al. panelcn.MOPS: Copy number detection in targeted NGS panel data for clinical diagnostics. Hum Mutat. 2017;38:889–97.
12. Kim H-Y, Choi J-W, Lee J-Y, Kong G, Kim H-Y, Choi J-W, et al. Gene-based comparative analysis of tools for estimating copy number alterations using whole-exome sequencing data. Oncotarget. 2017;8:27277–85.
13. Chiang T, Liu X, Wu TJ, Hu H, Sedlazeck FJ, White S, et al. Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. Genet Med. 2019;0:1–10.
14. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutat Res/Rev Mutat Res. 2019;779:114–25.
15. Mahamdallie S, Ruark E, Yost S, Ramsay E, Uddin I, Wylie H, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome Open Res. 2017;2:35.
16. Castellanos E, Gel B, Rosas I, Tornero E, Santín S, Pluvinet R, et al. A comprehensive custom panel design for routine hereditary cancer testing: Preserving control, improving diagnostics and revealing a complex variation landscape. Sci Rep. 2017;7:39348.
17. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012;28:2747–54.
18. Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol. 2018;19:1–13.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
20. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;1303:3997v. http://arxiv.org/abs/1303.3997.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
22. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012;40:D84–90.
23. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
24. Wickham H. The split-apply-combine strategy for data analysis. J Stat Softw. 2011;40:1–29.
25. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118.
26. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. Nat Protoc. 2009;4:1184.
27. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016. https://doi.org/10.18637/jss.v077.b02.
28. de Ligt J, Boone PM, Pfundt R, Vissers LELM, Richmond T, Geoghegan J. et al. Detection of clinically relevant copy number variants with whole exome sequencing. Hum Mutat. 2013;34:1439–48.
29. Sadedin SP, Ellis JA, Masters SL, Oshlack A. Ximmer: a system for improving accuracy and consistency of CNV calling from exome data. Gigascience. 2018;7:1–11.
30. Hong CS, Singh LN, Mullikin JC, Biesecker LG. Assessing the reproducibility of exome copy number variations predictions. Genome Med. 2016;8:82.

Article 2- Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer

José Marcos Moreno-Cabrera, Jesús del Valle, Lidia Feliubadaló, Marta Pineda, Sara González, Olga Campos, Raquel Cuesta, Joan Brunet, Eduard Serra, Gabriel Capellà, Bernat Gel* & Conxi Lázaro*

Journal of Medical Genetics. Published Online First: 20 November 2020.

doi: 10.1136/jmedgenet-2020-107366

IF 2019 = 4.943. Rank 30/177 (GENETICS & HEREDITY), First tertile, first quartile (Q1).

Journal link: https://jmg.bmj.com/content/early/2020/11/20/jmedgenet-2020-107366

Supplementary File available in Appendix B.

SHORT REPORT

Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer

José Marcos Moreno-Cabrera ,[1,2,3] Jesús del Valle,[1,2] Lidia Feliubadaló,[1,2] Marta Pineda,[1,2] Sara González,[1,2] Olga Campos,[1,2] Raquel Cuesta,[1,2] Joan Brunet,[1,2,4] Eduard Serra,[2,3] Gabriel Capellà,[1,2] Bernat Gel ,[3] Conxi Lázaro [1,2]

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2020-107366).

[1]Hereditary Cancer Program, Joint Program on Hereditary Cancer, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL-ONCOBELL, L'Hospitalet de Llobregat, Spain
[2]Centro de Investigación Biomédica en Red Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain
[3]Hereditary Cancer Group, Program for Predictive and Personalized Medicine of Cancer - Germans Trias i Pujol Research Institute (PMPPC-IGTP), Campus Can Ruti, Badalona, Spain
[4]Hereditary Cancer Program, Catalan Institute of Oncology, IDIBGi, Girona, Spain

Correspondence to
Dr Conxi Lázaro, Hereditary Cancer Program, Catalan Institute of Oncology, L'Hospitalet de Llobregat, 08908 Catalunya, Spain; clazaro@iconcologia.net
Dr Bernat Gel; bgel@igtp.cat

Received 23 July 2020
Revised 23 September 2020
Accepted 24 September 2020

To cite: Moreno-Cabrera JM, del Valle J, Feliubadaló L, *et al*. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmedgenet-2020-107366

ABSTRACT

Introduction Germline CNVs are important contributors to hereditary cancer. In genetic diagnostics, multiplex ligation-dependent probe amplification (MLPA) is commonly used to identify them. However, MLPA is time-consuming and expensive if applied to many genes, hence many routine laboratories test only a subset of genes of interest.

Methods and results We evaluated a next-generation sequencing (NGS)-based CNV detection tool (DECoN) as first-tier screening to decrease costs and turnaround time and expand CNV analysis to all genes of clinical interest in our diagnostics routine. We used DECoN in a retrospective cohort of 1860 patients where a limited number of genes were previously analysed by MLPA, and in a prospective cohort of 2041 patients, without MLPA analysis. In the retrospective cohort, 6 new CNVs were identified and confirmed by MLPA. In the prospective cohort, 19 CNVs were identified and confirmed by MLPA, 8 of these would have been lost in our previous MLPA-restricted detection strategy. Also, the number of genes tested by MLPA across all samples decreased by 93.0% in the prospective cohort.

Conclusion Including an in silico germline NGS CNV detection tool improved our genetic diagnostics strategy in hereditary cancer, both increasing the number of CNVs detected and reducing turnaround time and costs.

INTRODUCTION

Germline CNVs are one of the mutation types underlying multiple hereditary diseases.[1] Currently, its detection is recommended in comprehensive genetic testing strategies. For many years, the gold standard for CNV detection for one or a few genes has been multiplex ligation-dependent probe amplification (MLPA),[2] while hybridisation arrays have also been used for comprehensive testing of dozens or hundreds of genes at once.[3] Although MLPA is relatively affordable when testing a few genes and patients, its price increases with the number of patients and genes tested, making it impractical to test a large number of genes in an extensive cohort. In addition, MLPA is time-consuming and requires a specific design for each gene of clinical interest. For these reasons, many genetic testing laboratories restrict CNV analysis to a few key candidate genes.

Nowadays, next-generation sequencing (NGS) is widely used in clinical settings due to its cost-effective yield.[4] Targeted NGS gene panels are commonly used for genetic diagnostics, containing up to hundreds of genes, depending on the test. NGS bioinformatics analyses commonly include single-nucleotide variants and small deletion and insertion variants. However, CNVs are challenging variants due to several aspects, like short read lengths and GC-content bias,[5] especially when the variant affects a single exon. Multiple CNV detection approaches have been developed over the past years, although top-performing approaches are based on comparing read depth between samples and genomic regions.[6–10] Some authors have argued against the use of an NGS CNV detection tool in a clinical setting because of performance limitations,[11 12] especially when the CNV affects a single exon. However, our recent benchmark confirmed that there are CNV calling tools with sufficient sensitivity to be used as a screening step prior to orthogonal validation, even for single exon CNVs.[13] In our diagnostics datasets, we observed that DECoN[8] detected all CNVs (except one in a mosaic sample) with a specificity of over 90%. Once this benchmark has been performed, it was of our interest to evaluate the clinical impact of its implementation, in terms of detection yield and costs.

Here, we present an evaluation of the impact of using DECoN as a screening method in a hereditary cancer genetic diagnostics setting, testing it in a retrospective and in a prospective cohort (online supplemental figure graphical abstract).

METHODS

Patients and samples

All patients were selected by our genetic counselling units based on the clinical suspicion of hereditary cancer. Genomic DNA was extracted from peripheral blood lymphocytes using the FlexiGene DNA Kit (Qiagen GmbH, Hilden, Germany). Samples were analysed using our custom hybridisation-based target capture NGS panel for hereditary cancer diagnostics, called I2HCP,[14 15] which ranged from 122 to 135 genes (V.2.0–V.2.2). Mutations were examined on a subset of genes in each patient depending on their clinical suspicion.[16] Samples were sequenced in either a MiSeq with 2×300 bp reads or a HiSeq with 2×251 bp reads. All samples were aligned to the GRCh37 human genome assembly using BWA mem V.0.7.12.[17] SAMtools V.0.1.19 was used to sort and index BAM files. No additional filtering was applied to the BAM files.

NGS CNV detection strategy

We chose DECoN for NGS CNV screening because of its performance in our previous study, where all CNVs were detected when using optimised parameters in our diagnostics datasets, except one in a mosaic sample.[13] Therefore, our NGS CNV detection strategy consisted of two steps: first, screening of all genes of clinical interest using DECoN V.1.0.1 with modified parameters (online supplemental table 1), and second, validation of putative CNVs by MLPA according to manufacturer's protocols. To discard DECoN CNV calls with low statistical support, only those with a Bayesian factor (BF) ≥2 were included for MLPA testing. We chose this cut-off after observing that all true CNVs from our previous validation study[13] had a BF value >2, except for a case of mosaicism. Additionally, we also performed MLPA when DECoN detected a failed region in a gene of clinical interest, meaning that the region coverage or the sample correlation value were below the required thresholds (see more details in online supplemental file).

Retrospective and prospective cohorts

We used our NGS CNV detection strategy to test a retrospective and a prospective cohort. The retrospective one consisted of 1860 patients, for which MLPA results of one or two genes were available in most of them, depending on their clinical characteristics (online supplemental table 1). In this cohort, we tested the remaining clinically relevant genes according to the patient's clinical phenotype. We also used DECoN to evaluate whether it could detect the 20 CNVs previously identified using MLPA. The prospective cohort consisted of 2041 patients tested for CNVs in all genes of clinical interest, according to the patient's phenotype, with our NGS CNV detection strategy.

RESULTS

CNV identification in the retrospective cohort

We used DECoN to screen all genes of clinical interest not previously analysed by MLPA in the retrospective cohort (1860 patients), and to evaluate whether the 20 CNVs detected by our previous MLPA-restricted strategy, were also detected. DECoN successfully identified the 20 CNVs previously identified. Also, six new true CNVs were identified and subsequently confirmed by MLPA (four *CHEK2*, one *RAD51C* and one *PALB2*), which represents an increase in CNV detection of 30%, from 20 to 26 CNVs (table 1). By performing this DECoN analysis, 13 687 genes were analysed across all samples. In addition to the 26 true positive signals, we obtained a total of 128 false positive calls, 68 deletions and 60 duplications that were confirmed as false positives by MLPA analysis. Furthermore, DECoN detected 87 failed regions that were tested by MLPA. In total, the number of genes tested by MLPA after DECoN in this retrospective cohort was 221, in contrast to the 2660 required with the previous MLPA-restricted strategy (table 2). Also, the average number of genes evaluated per sample with this new strategy was 7.35 compared with 1.43 using our previous detection strategy.

Table 1 New CNVs identified using DECoN

Sample	Clinical suspicion	Gene	Exons	CNV type	Classification
		Retrospective cohort			
S1	Hereditary renal cancer syndromes	*CHEK2*	3–4	Duplication	VUS
S2	Hereditary breast and ovarian cancer	*CHEK2*	3–4	Deletion	PAT
S3	Hereditary non polyposis colon cancer	*CHEK2*	3–4	Duplication	VUS
S4	Hereditary breast cancer	*CHEK2*	2	Deletion	LPAT
S5	Hereditary breast cancer	*PALB2*	8	Deletion	PAT
S6	Hereditary ovarian cancer	*RAD51C*	4	Deletion	PAT
		Prospective cohort			
S7	Hereditary breast cancer	*ATM*	63	Duplication	VUS
S8	Hereditary prostate cancer	*ATM*	27–37	Duplication	VUS
S9	Hereditary breast and ovarian cancer	*CHEK2*	3–4	Duplication	VUS
S10	Hereditary breast cancer	*PALB2*	7–11	Deletion	LPAT
S11	Hereditary breast cancer	*PALB2*	7	Deletion	VUS
S12	Hereditary breast cancer	*PALB2*	7–11	Deletion	LPAT
S13	Hereditary ovarian cancer	*RAD51C*	4–5	Deletion	PAT
S14	Polyposis	*STK11*	5–8	Deletion	PAT
S15	Polyposis	*APC*	Whole gene	Deletion	PAT
S16	Hereditary breast cancer	*BRCA1*	21	Deletion	PAT
S17	Hereditary ovarian cancer	*BRCA1*	9–13	Deletion	PAT
S18	Hereditary ovarian cancer	*BRCA1*	1–3	Deletion	PAT
S19	Hereditary ovarian cancer	*BRCA1*	24	Duplication	PAT
S20	Hereditary breast and ovarian cancer	*BRCA1*	3–5	Deletion	PAT
S21	Hereditary ovarian cancer	*BRCA1*	Whole gene	Deletion	PAT
S22	Hereditary breast cancer	*BRCA2*	2	Deletion	LPAT
S23	Hereditary non-polyposis colon cancer	*EPCAM*	8–9	Deletion	PAT
S24	Hereditary non-polyposis colon cancer	*MSH2*	4–6	Deletion	PAT
S25	Hereditary non-polyposis colon cancer	*MSH2*	7	Deletion	PAT

Samples S7–S14 contain CNVs that would not have been found with our previous MLPA-restricted strategy.
LPAT, likely pathogenic; MLPA, multiplex ligation-dependent probe amplification; PAT, pathogenic; VUS, variant of uncertain significance.

Table 2 Impact of using DECoN as CNV screening tool in the retrospective and prospective cohorts

	Retrospective		Prospective	
	MLPA-restricted strategy	With DECoN	MLPA-restricted strategy	With DECoN
Number of samples	1860	1860	2041	2041
Total number of genes tested by MLPA	2660	221	3442	240
Average number of genes tested by MLPA, per sample	1.43	0.12	1.69	0.12
Total number of genes tested by DECoN	0	13 687	0	18 836
Total number of genes covered in the testing strategy	2660	13 687	3442	18 836
Average number of genes covered per sample	1.43	7.35	1.69	9.22
Average number of kilobases covered per sample	12.94	24.26	11.22	39.27
CNVs confirmed by MLPA	20	6	11	19

The 'MLPA—Prospective' column contains an estimation of what would have happened if DECoN had not been used as a screening step prior to MLPA validation.
MLPA, multiplex ligation-dependent probe amplification.

CNV identification in the prospective cohort

We used DECoN to screen all genes of clinical interest in the prospective cohort (2041 patients). Most of the CNV calls (53.2%) were discarded because they obtained a BF <2 (online supplemental figure 1). DECoN called 158 CNVs of which 19 were confirmed by MLPA (table 1). Out of those, 8 would have not been identified using our previous MLPA-restricted approach (online supplemental table 1): 3 in *PALB2*, 2 in *ATM*, 1 in *STK11, RAD51C* and *CHEK2*. This represents an increase of 72.7% in CNV detection, from 11 to 19 CNVs. For the 18 836 genes analysed by DECoN across all samples, DECoN produced a total of 139 false positive CNV calls (71 duplications and 68 deletions) and 82 failed regions, discarded after MLPA analysis. It is worth mentioning that, compared with our previous MLPA-restricted strategy, the number of genes tested by MLPA across all samples decreased from 3442 to 240, representing a 93.0% decrease (table 2). Furthermore, the average number of genes tested per sample increased from 1.69 to 9.22 with the new strategy.

DISCUSSION

The implementation of NGS technologies in genetic diagnosis has been a breakthrough in diagnostic performance allowing the analysis of multiple genes at the same time, reducing costs and turnaround time. However, establishing accurate bioinformatics algorithms for CNV detection from NGS data has been more challenging than for other types of mutations such as point mutations or small deletions and insertions.

In this retrospective and prospective study, we evaluated the use of an NGS CNV calling tool as a screening step before MLPA validation in a hereditary cancer genetic diagnostics setting. An ideal screening tool should have 100% sensitivity to avoid missing any true positive. Therefore, we chose DECoN as a first-tier in silico screening tool based on its performance in our previous benchmarking effort.[13] Including an in silico screening step to our diagnostics strategy allowed us to analyse CNVs for all the genes of clinical interest, most of them had not tested before due to turnaround time and budget restrictions. This implementation resulted in an improved diagnostics yield in both cohorts, with up to 72.7% of additional CNVs detected. As expected, DECoN showed high sensitivity, detecting the 20 previously known CNVs in the retrospective cohort. It is extremely worth highlighting the important clinical impact of this yield improvement which allowed us to discover the genetic cause of cancer in previously uninformative families. The detection of pathogenic CNVs in clinically actionable genes is paramount for the clinical management of the patient carrying the CNV, as well as for their relatives. It allows the individualisation of cancer risk assessment for all family members as well as the establishment of specific surveillance measures and appropriate therapeutic strategies for all the carriers.

Besides the yield improvement, the use of an in silico screening tool to identify CNVs based on NGS data entailed an important decrease in the resources required for this analysis. In the prospective study, we observed a reduction of 93.0% in the number of genes requiring MLPA testing across all samples, with the associated savings in time and costs. The high specificity of DECoN, validated in our previous study, along with the introduction of a BF cut-off, made possible an important reduction in the number of genes tested by MLPA.

Although some authors have exposed limitations for the use of NGS CNV detection tools in clinical settings for different reasons,[11 12 18] others have argued in favour of its use in this context.[7–10 19 20] This work, together with our previous benchmark on CNV calling tools for genetic diagnostics,[13] shows that CNV in silico screening is viable in a genetic diagnostics setting, and results in a reduction of costs and turnaround times and, most importantly, in an increase in the diagnostic yield.

Acknowledgements The authors would like to thank the participating patients and all the members of the Unit of Genetic Diagnostics of the Hereditary Cancer Programme of the Catalan Institute of Oncology (ICO-IDIBELL) and the Genetics Diagnostics Unit of the Hereditary Cancer Group of the Germans Trias i Pujol Research Institute (IGTP). The authors would like to thank the IGTP HPC Core Facility, Lloyd Byron Goodman and Iñaki Martínez de Ilarduya for their help. The authors would also like to thank the CERCA Programme/Generalitat de Catalunya for institutional support.

Contributors JMM-C, JdV, LF, MP, SG, CLG performed the CNV analysis. OC, RC performed genetic tests. JMM-C, BG, ES, GC, JB contributed to analysis and presentation of the results. JMM-C, JdV, BG and CLG wrote the manuscript. All authors reviewed and accepted the manuscript.

Funding Contract grant sponsor: supported by the Carlos III National Health Institute funded by FEDER funds—a way to build Europe—(PI19/00553; PI16/00563, PI15/00854 and CIBERONC); the Government of Catalonia (Pla estratègic de recerca i innovació en salut (PERIS_MedPerCan and URDCat projects), 2017SGR1282 and 2017SGR496) the Spanish Association Against Cancer (AECC) and Fundació La Marató de TV3.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study protocol was approved by the Ethics Committee of IDIBELL (PR278/19). Informed written consent was obtained from all of them.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability

of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

José Marcos Moreno-Cabrera http://orcid.org/0000-0001-8570-0345
Bernat Gel http://orcid.org/0000-0001-8878-349X
Conxi Lázaro http://orcid.org/0000-0002-7198-5906

REFERENCES

1 Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 2009;10:451–81.

2 Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, Rupar CA, Adams P, Hegele RA, Lin H, Rodenhiser D, Knoll J, Ainsworth PJ, Sadikovic B. Clinical validation of copy number variant detection from targeted next-generation sequencing panels. *J Mol Diagn* 2017;19:905–20.

3 Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;12:e1004873–18.

4 Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–51.

5 Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 2012;28:2711–8.

6 Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 2013;14:S1.

7 Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JDH, van der Hout AH, Westers H, Sinke RJ, Swertz MA, Sijmons RH, Sikkema-Raddatz B. CoNVaDING: single exon variation detection in targeted NGS data. *Hum Mutat* 2016;37:457–64.

8 Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, Uddin I, Wylie H, Strydom A, Lunter G, Rahman N. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res* 2016;1:20.

9 Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, Klambauer G, Hochreiter S, Wimmer K. panelcn.MOPS: copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat* 2017;38:889–97.

10 Chiang T, Liu X, Wu T-J, Hu J, Sedlazeck FJ, White S, Schaid D, Andrade Mde, Jarvik GP, Crosslin D, Stanaway I, Carrell DS, Connolly JJ, Hakonarson H, Groopman EE, Gharavi AG, Fedotov A, Bi W, Leduc MS, Murdock DR, Jiang Y, Meng L, Eng CM, Wen S, Yang Y, Muzny DM, Boerwinkle E, Salerno W, Venner E, Gibbs RA. Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. *Genet Med* 2019;21:2135–44.

11 Mason-Suares H, Landry L, S. Lebo M. Detecting copy number variation via next generation technology. *Curr Genet Med Rep* 2016;4:74–85.

12 Yao R, Yu T, Qing Y, Wang J, Shen Y. Evaluation of copy number variant detection from panel-based next-generation sequencing data. *Mol Genet Genomic Med* 2019;7:e00513–8.

13 Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, Serra E, Capellà G, Lázaro C, Gel B. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet* 2020. doi:10.1038/s41431-020-0675-z. [Epub ahead of print: 19 Jun 2020].

14 Castellanos E, Gel B, Rosas I, Tornero E, Santín S, Pluvinet R, Velasco J, Sumoy L, Del Valle J, Perucho M, Blanco I, Navarro M, Brunet J, Pineda M, Feliubadaló L, Capellá G, Lázaro C, Serra E. A comprehensive custom panel design for routine hereditary cancer testing: preserving control, improving diagnostics and revealing a complex variation landscape. *Sci Rep* 2017;7:39348.

15 Feliubadaló L, Tonda R, Gausachs M, Trotta J-R, Castellanos E, López-Doriga A, Teulé Àlex, Tornero E, del Valle J, Gel B, Gut M, Pineda M, González S, Menéndez M, Navarro M, Capellá G, Gut I, Serra E, Brunet J, Beltran S, Lázaro C. Benchmarking of whole exome sequencing and AD hoc designed panels for genetic testing of hereditary cancer. *Sci Rep* 2017;7.

16 Feliubadaló L, López-Fernández A, Pineda M, Díez O, Del Valle J, Gutiérrez-Enríquez S, Teulé A, González S, Stjepanovic N, Salinas M, Capellá G, Brunet J, Lázaro C, Balmaña J, Campos O, Carrasco E, Cuesta R, Darder E, Gadea N, Gómez C, Grau E, Iglesias S, Izquierdo A, Llort G, Menéndez M, Moles-Fernández A, Montes E, Muñoz X, Navarro M, Catalan Hereditary Cancer Group. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int J Cancer* 2019;145:2682–91.

17 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60.

18 Ceyhan-Birsoy O, Pugh TJ, Bowser MJ, Hynes E, Frisella AL, Mahanta LM, Lebo MS, Amr SS, Funke BH. Next generation sequencing-based copy number analysis reveals low prevalence of deletions and duplications in 46 genes associated with genetic cardiomyopathies. *Mol Genet Genomic Med* 2016;4:143–51.

19 Pugh TJ, Amr SS, Bowser MJ, Gowrisankar S, Hynes E, Mahanta LM, Rehm HL, Funke B, Lebo MS. VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet Med* 2016;18:712–9.

20 Ellingford JM, Horn B, Campbell C, Arno G, Barton S, Tate C, Bhaskar S, Sergouniotis PI, Taylor RL, Carss KJ, Raymond LFL, Michaelides M, Ramsden SC, Webster AR, Black GCM. Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases. *J Med Genet* 2018;55:114–21.

Article 3- CNVfilteR: an R/Bioconductor package to identify false positives produced by germline NGS CNV detection tools

José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Eduard Serra, Gabriel Capellà, Conxi Lázaro* & Bernat Gel*

Manuscript accepted for publication in Bioinformatics (April 2021).

IF 2019 = 5.610. Rank 3/59 (MATHEMATICAL & COMPUTATIONAL BIOLOGY), First decile, first quartile (Q1).

Supplementary File 1 is available in Appendix B. All supplementary files (8) are fully available online:

- Supp File 1 (pdf) Scoring model for CNV duplications, samples evaluation details, runtime evaluation, parameters summary, and CNVfilteR use recommendations.
- Supp File 2 (xlsx) Performance evaluation metrics for the WGS samples.
- Supp File 3 (xlsx) Performance evaluation metrics for the gene-panel samples.
- Supp File 4 (bed) Target bed file used for the gene-panel samples.
- Supp File 5 (xlsx) Performance evaluation on different CNV size ranges.
- Supp File 6 (xlsx) Performance evaluation on different number of SNVs overlapping each CNV.
- Supp File 7 (xlsx) Performance evaluation metrics on different parameter values.
- Supp Filgures (pdf) Supplementary figures.

CNVfilteR: an R/Bioconductor package to identify false positives produced by germline NGS CNV detection tools

José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Eduard Serra, Gabriel Capellá, Conxi Lázaro* and Bernat Gel*

Summary: Germline copy-number variants (CNVs) are relevant mutations for multiple genetics fields, such as the study of hereditary diseases. However, available benchmarks show that all next-generation sequencing (NGS) CNV calling tools produce false positives. We developed CNVfilteR, an R package that uses the single nucleotide variant calls usually obtained in germline NGS pipelines to identify those false positives. The package can detect both false deletions and false duplications. We evaluated CNVfilteR performance on callsets generated by 13 CNV calling tools on 3 whole-genome sequencing and 541 panel samples, showing a decrease of up to 44.8% in false positives and consistent F1-score increase. Using CNVfilteR to detect false-positive calls can improve the overall performance of existing CNV calling pipelines.

Availability and Implementation: CNVfilteR is released under Artistic-2.0 License. Source code and documentation are freely available at Bioconductor (http://www.bioconductor.org/packages/CNVfilteR)

Supplementary Information: Supplementary data will be available at Bioinformatics online.

Introduction

Copy-number variants (CNVs) are a type of structural variant which has been a matter of interest in multiple genetic fields. In the research and diagnosis of hereditary diseases, where CNVs are relevant contributors (Zhang et al., 2019), the analysis of germline CNVs plays a key role. Recent improvements in next-generation sequencing (NGS) have resulted in the release of several tools for germline CNV detection on whole-genome sequencing (WGS), whole-exome sequencing, and panel data (Zhao et al., 2013; Roca et al., 2019; Mason-Suares et al., 2016).

Nevertheless, CNV detection in NGS is challenging due to aspects relative to the technology such as short read lengths or GC-content bias (Teo et al., 2012).

Available benchmarks show that all germline CNV calling tools produce false positives (Zhang et al., 2019; Kim et al., 2017; Moreno-Cabrera et al., 2020), frequently reaching high false discovery rates (FDRs). These false-positive calls impact downstream analysis. In a clinical setting, where the use of an orthogonal method is necessary to validate a CNV, false-positive calls lead laboratories to make an important effort to validate them. A tool able to identify these false-positive calls could help in this regard.

Most NGS CNV callers are based on one or more of these strategies: read-pair, split-read, read-depth, and assembly-based (Pirooznia et al., 2015). However, information from single-nucleotide variants (SNVs), usually available in NGS pipelines, is rarely used in CNV detection strategies although SNV allele frequency can provide evidence to confirm or discard CNV calls.

Here we present CNVfilteR, an R/Bioconductor package that uses SNVs to identify false positives in the output of CNV calling tools.

False-positive identification strategy

CNVfilteR uses two different strategies to identify false-positives CNV calls in diploid genomes. Heterozygous deletions are loss-of-heterozygosity regions and cannot overlap with heterozygous SNVs, since only one allele remains. If a heterozygous SNV is detected within a deleted region, either the SNV or the deletion is a false positive (Figure 1a). To account for errors in SNV calling, a CNV deletion is identified as false positive if at least a percentage of the SNVs overlapping that CNV is heterozygous, 30% by default. On the other hand, CNV duplications are evaluated using a fuzzy-logic-inspired model which scores all heterozygous SNVs overlapping the CNV. If the duplication was a true-positive, the expected allele frequency of heterozygous SNVs would be either 33% or 66%, while it would be 50% if the duplication was a false positive (Figure 1b). Therefore, each SNV is scored with a value between-1 and 1 depending on how close the allele frequency is to the nearest expected allele frequency (Figure 1c). If the sum of the scores of all the SNVs in the CNV is greater than the duplication threshold value, the CNV duplication is identified as false positive. Further details of the scoring model can be found in Supplementary File 1.

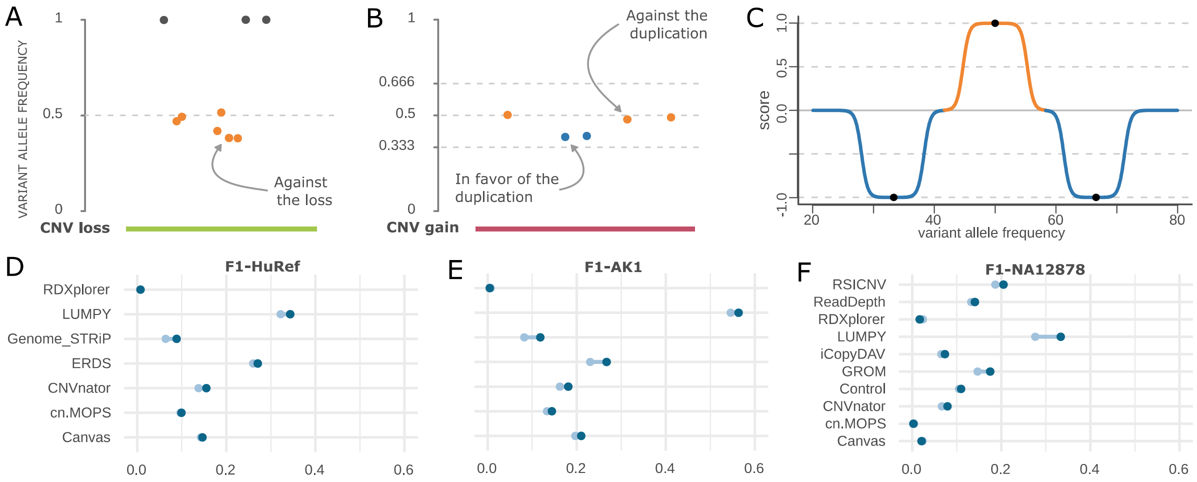


Figure 1. **(A)** CNV deletion example, adapted from CNVfilteR output. **(B)** CNV duplication example, adapted from CNVfilteR output. **(C)** Scoring model for CNV duplications, plotted by CNVfilteR. **(D-F)** F1-score differences before (light blue) and after (dark blue) removing the false-positive CNVs identified by CNVfilteR in the HuRef, AK1, and NA12878 WGS samples.

Features

Input formats

VCF format is the most common output of SNV callers and its interpretation is challenging due to the flexibility provided by the format specification. CNVfilteR provides a function to interpret automatically VCFs produced by VarScan2, Strelka/Strelka2, freeBayes, HaplotypeCaller (GATK), and UnifiedGenotyper (GATK). Output from other tools can also be loaded if adequate parameters are provided.

Visual output

Results can be plotted and customized through plotting functions based on karyoploteR (Gel and Serra, 2017) and CopyNumberPlots (https://github.com/bernatgel/CopyNumberPlots) packages (Supplementary Figure 1).

Performance evaluation

CNVfilteR was evaluated on 3 WGS samples and 541 gene-panel samples. The default parameters were chosen based on their performance in a WGS sample (HuRef sample) and a gene-panel dataset (HiSeq-panel) (Supplementary File 1).

Evaluation on WGS data

We evaluated CNVfilteR performance on three reference WGS samples: the HuRef/Venter genome (Zhou et al., 2018), the AK1 genome (Seo et al., 2016), and the NA12878 genome. The HuRef and AK1 samples were evaluated using a published reference CNV callset and the results of six CNV calling tools (Canvas, cn.MOPS, CNVnator, ERDS, Genome_STRiP, RDXplorer) (Trost et al., 2018). For these two samples, we also ran an additional CNV calling tool, LUMPY (Layer et al., 2014). On the other hand, we evaluated the NA12878 sample with a reference callset and the output of ten CNV calling tools (Canvas, cn.MOPS, CNVnator, RDXplorer, iCopyDAV, GROM-RD, Rsicnv, Control-FREEC, ReadDepth) from a previous work (Zhang et al., 2019; Parikh et al., 2016; MacDonald et al., 2014). For the three WGS samples, SNV calls were obtained using Strelka2 (Kim et al., 2018). Further details are available in Supplementary File 1.

CNVfilteR identified between 15.3% and 44.8% of the false positives and the FDR decreased for all tool-sample evaluations (up to 10.4%). Additionally, F1-score was improved in 19 out of the 24 tool-sample evaluations reaching up to 20.7% F1-score increase (Figure 1d-f). Sensitivity, however, decreased slightly: tool-sample evaluations had an absolute sensitivity decrease between 0.001 and 0.035. Metrics details are available in Supplementary File 2 and Supplementary Figures 2-7. Moreover, additional evaluations were performed to show CNVfilteR performance on different CNV size ranges, on different number of SNVs overlapping each CNV, and on different parameter values (Supplementary Figures 8-25 and Supplementary Files 5-7).

Evaluation on gene-panel data

We also evaluated CNVfilteR performance on two gene-panel targeted datasets: one containing 411 samples from different Illumina HiSeq runs (HiSeq-panel dataset) and another with 130 samples from different Illumina MiSeq runs (MiSeq-panel dataset). All samples were captured with a 135-gene panel (Castellanos et al., 2017). To evaluate CNVfilteR, previous MLPA results for a subset of genes were used as gold-standard, CNVs were called using DECoN (Fowler et al., 2016), and SNVs were called using VarScan2 (Koboldt et al., 2012) (Supplementary Files 1, 3-4).

In the HiSeq-panel and MiSeq-panel datasets, CNVfilteR identified 15% of the false-positive calls (3 out of 20 false positives) and 12.5% of the false-positive calls (2 out of 16), respectively. On both datasets, no true CNV was misidentified as false positive (Supplementary File 1), so sensitivity did not change.

Runtime

Runtime was evaluated on a subset of 79 gene-panel samples and the HuRef WGS sample. The median runtime per sample was 0.85 seconds for the gene-panel samples and 3.53 minutes for the HuRef sample (Supplementary File 1).

Conclusion

We developed CNVfilteR, an R/Bioconductor package to identify false-positive calls generated by CNV calling tools from germline NGS data using SNVs' allele frequency. CNVfilteR identified false-positive calls in all tested tools and datasets, from gene-panel to WGS, and F1-score was improved in most tool-sample combinations. CNVfilteR can be plugged in most existing CNV calling pipelines to improve calling performance at virtually no cost.

Acknowledgments

We thank all members of the Genetic Diagnostics Units at ICO and IGTP for their help.

Funding

This work has been supported by Carlos III Health Institute co-funded by FEDER– a way to build Europe; the Government of Catalonia and CERCA. Contract grant numbers: PI16/00563, PI19/00553, CIBERONC, 2017SGR1282 and 2017SGR496.

Article 4- Eleven quick tips to build a software tool for integral management of genetic diagnostics

José Marcos Moreno-Cabrera, Lidia Feliubadaló, Eli Castellanos, Jesús del Valle, Inma Rosas, Eva Tornero, Xavier Muñoz, Mireia Menéndez, Daniel Azuara, Marta Pineda, Sara González, Maribel González-Acosta, Alejandro Negro, Gabriel Capellà, Eduard Serra, Conxi Lázaro* & Bernat Gel*

Supplementary Figures available in Appendix B.

1.1 Pandora: technical tool description

We developed Pandora to support the diagnostics routine of the ICO-IGTP Joint Program on Hereditary Cancer. Pandora is a web-based tool that automates multiple tasks during the diagnostics process such as FASTQs download or sample bioinformatic analysis. It supports the process of validation or classification of a variant, and allows flexible exploitation of the patient, sample, and variant data stored in Pandora.

Pandora is a distributed system built on the Django framework, a PostgreSQL database (initially developed by Bernat Gel), and a number of Python and R cron jobs. Samples are analyzed using a custom pipeline developed by Bernat Gel (Castellanos *et al.*, 2017). The whole architecture of Pandora is described below (Figure 11).

- **Web server**. This is the server where Pandora is hosted. The web infrastructure has been implemented using the Django framework (v1.8.8), which offers a model-view-template design pattern.
    - The front-end layer is implemented in JavaScript, CSS, and HTML languages and makes use of the Django templates.
    - The back-end layer is implemented in Python (v3.4) and makes use of the Django database model to communicate with the PostgreSQL database.

  The Web server only communicates with the Database server; it reads and writes on the database. To keep the client-side updated in quasi-real time, a poll is performed every 3 seconds to get the latest database changes.

- **Database server**. Uses PostgreSQL (v9.6.2) to manage the databases described below:

  - A Django database (Users database) storing users, roles, and permissions.
  - A database (NGS database) compound of 64 tables to store, among others, variant and sample data. A trigger function was implemented to log data changes in the database to support client-side synchronization.

- Crons server. Coordinates multiple processes associated with Pandora. Six cron jobs (five implemented in Python, one in R) perform different tasks:
  - BaseSpace Downloader (Python). It searches for FASTQs to be downloaded from the Illumina BaseSpace system and the FTP server. When found, it downloads them to the Network File System (NFS) server.
  - Analysis Launcher (Python). Launches the analysis pipeline to align FASTQs, call variants and generate reports.
  - Results Files Copier (Python). Copies files requested by the user from the NFS server to the Web server.
  - Results Files Remover (Python). Removes files from the Web server that are obsolete in order to free up space.
  - Manual variant inserter (R). Inserts in the database a variant defined manually by the user.
  - Maintainer. Carries out general maintenance tasks, like updating the in-house frequencies for each variant.

- Computer cluster. The computer cluster environment where pipeline jobs are executed.

- NFS server. Stores different purpose files, from FASTQs to pipeline results. It can be accessed by any machine except the Web server.

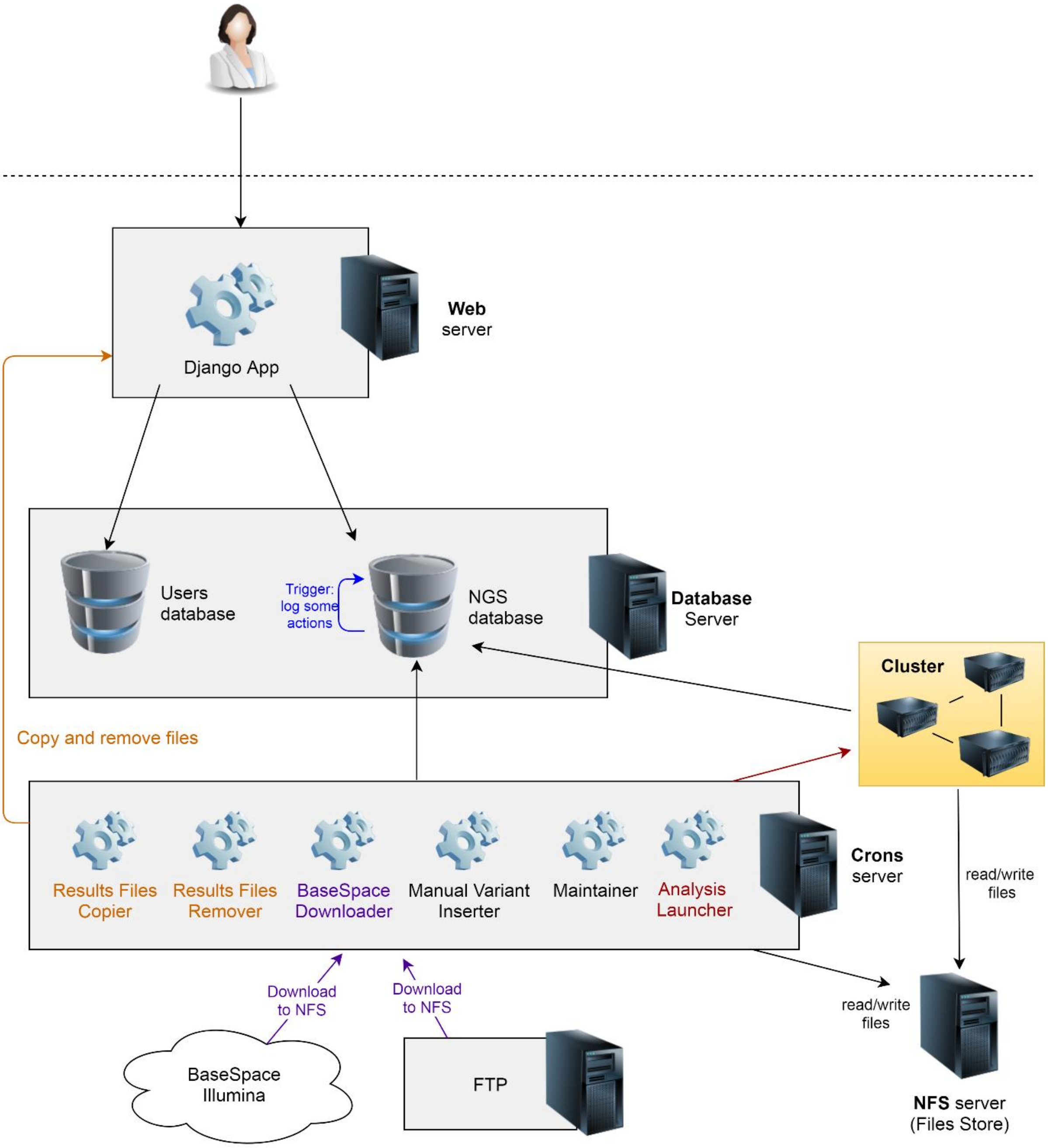


Figure 11. Pandora architecture is composed of Web server, Database server, Crons server, NFS server and computer cluster.

1.2 Manuscript

Eleven quick tips to build a software tool for integral management of genetic diagnostics

José Marcos Moreno-Cabrera, Lidia Feliubadaló, Eli Castellanos, Jesús del Valle, Inma Rosas, Eva Tornero, Xavier Muñoz, Mireia Menéndez, Daniel Azuara, Marta Pineda, Sara González, Maribel González-Acosta, Alejandro Negro, Gabriel Capellá, Eduard Serra3, Conxi Lázaro* & Bernat Gel*

Abstract. The arrival of next-generation sequencing technologies has transformed genetic diagnostics by multiplying the sequencing capacity at a limited cost. The diagnostic activity in the laboratory covers multiple steps from the patient's sample collection to the final report. A software tool to support the whole routine diagnostics has the potential to automate multiple tasks, avoid human errors and enhance data exploitation. However, building such a software tool requires careful design. We present here eleven tips to consider when designing and implementing a tool to support the genetic testing of a disease. Traceability, reproducibility, efficiency, usability and also an intensive dialog with your future users are some of the aspects that must be addressed to build a tool to enhance your genetic diagnostic workflow.

Introduction

Over the last years, next-generation sequencing (NGS) technologies have thoroughly impacted the way genetic testing of diseases is performed (Knoppers *et al.*, 2015; Xue *et al.*, 2015; Yohe and Thyagarajan, 2017). The arrival of NGS has multiplied the sequencing capacity in a very cost-effective way (Goodwin *et al.*, 2016; Pereira *et al.*, 2020). Whole-genome and whole-exome sequencing are used in diagnostic settings, although most laboratories use the smaller targeted gene-panels in their routine diagnostics. The latter allow for the cost-effective testing of up to hundreds of clinically relevant genes during the routine diagnostics, which has improved the final diagnostic yield (Teekakirikul *et al.*, 2013; Kurian *et al.*, 2014; Feliubadaló *et al.*, 2017).

The ICO-IGTP Joint Program on Hereditary Cancer focuses on the detection and interpretation of germline variants that increase the risk of developing cancer. The diagnostic activity covers all hereditary cancer syndromes, although it concentrates on hereditary colorectal cancer, hereditary breast and ovarian cancer, neurofibromatosis and other related disorders. To improve our diagnostic activity, we developed a custom gene panel: the I2HCP (Castellanos *et al.*, 2017; Feliubadaló *et al.*, 2017), a hereditary cancer gene-panel with 135 genes. To date, more than five thousand cases (1331 just in 2020) have been tested using this panel and our associated analysis pipeline.

The diagnostic workflow typically comprises DNA extraction, library preparation, NGS sequencing, primary analysis pipeline, variant analysis, variant classification and, finally, preparation of the diagnostic report. To manage and orchestrate the diagnostic routine, we developed Pandora (PlAtform for NGS Data Organization Repository and Analysis), a web-based platform built on the Django framework and a PostgreSQL database. Pandora automates multiple tasks such as sequencing data download and primary analysis, supports the process of validation and classification of a variant and offers flexible data exploitation. Pandora is currently being used in the diagnostic routine of the ICO-IGTP program and some of the data has also been opened to external institutions to help other laboratories in their variant classification efforts.

As a result of our experience designing, building and using Pandora, we present here eleven quick tips to consider when facing the development of a tool to support genetic testing of a disease. Eleven best practices or pieces of advice shared for computer scientists, biologists, bioinformaticians, clinicians and all the health care professionals involved in the process of designing and building a tool to support genetic diagnostic testing.

Tip 1: Traceability and reproducibility are essential

The diagnostic workflow finishes with a report delivered to a patient. Hence, any future modification or result re-analysis needs to retrace all the steps that led to the report. Therefore, in genetic diagnostics, it is a must to know who and when did what in the system. Traceability is especially important when validating or classifying a variant. In our laboratory routine, the validation of a variant implies the labor of two independent curators that check whether the metrics of the variants (read depth and allele frequency among others) indicate that the variant call is reliable enough, and a third curator in case the initial curators do not agree. All that information is properly stored in the Pandora database. For the classification, Pandora does not just store who and when classified a variant and why it was classified, it also stores a history of all previous classifications (Supplementary Figure 1).

Reproducibility is another matter of concern in a clinical context (Roy *et al.*, 2018). It is necessary to identify unambiguously in which computational environment a sample was analyzed: software versions, fastq files used and other input parameters. When a new- and probably better- pipeline version is available, it may be interesting re-analyzing old samples to take advantage of the new improvements. For that purpose, it is necessary to store the original fastq files, without any modification, so that new versions of the pipeline could be executed starting from the same raw data. Also, each modification in the pipeline code should entail a new pipeline version: this way all versions can be clearly identified, so any result for any sample can be reproduced with the desired pipeline version.

Tip 2: The more you automate tasks, the better

One of the core reasons to implement a piece of software to support the diagnostic routine is to enhance productivity. In this regard, you should automate every task where a human decision is not necessary. For example, the input sample sheet for Illumina sequencing platforms is automatically generated by Pandora, with the exact required format, so manual errors are avoided. Also, downloading results from either the cloud (Illumina Basespace) or an FTP server is performed by a Pandora script which scans for new results every five minutes. When sample download is finished, the analysis pipeline is automatically launched and an email is sent to the user when the results are available. The automation of all these steps speeds up the diagnostic workflow while limiting the errors that typically occur when manual tasks are performed (Figure 1).

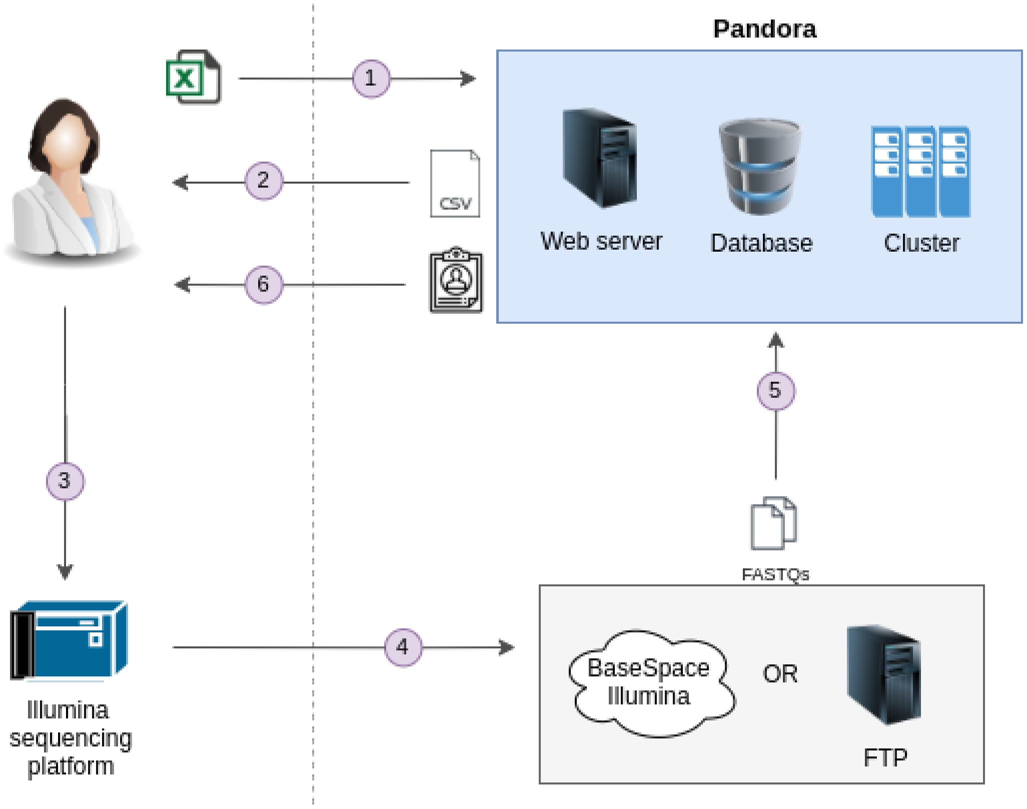


Figure 1. Pandora diagnostics workflow. **1**: The user uploads sample information to Pandora using an excel file which is parsed by Pandora. **2**: Pandora produces an Illumina sample sheet (.csv) to be downloaded. **3**: Sequencing starts using the sample sheet provided by Pandora. **4**: FASTQs are uploaded to BaseSpace or to an FTP server. **5**: FASTQs are automatically downloaded by Pandora. Pipeline analysis starts in the computing cluster. **6**: When finished, an email is sent to the curators: all results can be reviewed. Then, independent curators validate and classify sample variants and a report is sent to the clinician.

The automation of some small and less visible tasks can also improve user productivity. This is the case of the search in Google or PubMed of several nomenclatures of a certain variant, to assist in its classification. For example, for the variant selected in Supplementary Figure 2, after clicking the PubMed button the chain "ATM" "8122G>A" OR "ATM" "8122G->A" OR "ATM" "8122G-->A" OR "ATM" "8122G/A" OR "ATM" "p.Asp2708Asn" OR "ATM" "Asp2708Asn" is searched in PubMed. The same behavior applies to the Google button.

Tip 3: Listen to the user

When designing the tool, it is important to define with precision its features and appearance. Moreover, you are going to design a tool for genetic diagnostics, a new field that has rarely been traveled; in other words, you are not designing a shoe store web-site for which hundreds of examples can be found. Of course, the process of capturing future features and appearance is not easy and involves communication with future users. Apart from collecting them in a specifications document, in our experience, sketching was a useful way to explain to the user which tool we were imagining (Supplementary Figure 3) and a basis on which they could propose new ideas or modifications. The more you define and design the right tool to be developed, the fewer changes you will have to perform after the user starts using it. Also, try to develop your tool towards a functional prototype as quickly as possible, so your users will be able to give you very valuable feedback on something they can interact with. Listening to the user is not only a must when thinking and developing new software, it should also be done when the tool is finished (in production). At that point, users will start their real experience with the tool and multiple needs, previously unknown, may emerge.

In our experience, one of those emerged needs was the use of tags to label certain elements. Our initial approach was to develop a systematic and rigid system emphasizing reproducibility. However, when the users started to use the tool in a real scenario, they began to report to us the need to register in the system unforeseen conditions of certain samples or analyses. For example, users wanted to register that a defective cartridge was used when sequencing a certain group of samples or that an analysis was considered failed for not having enough reads. As a result of this feedback, we implemented tags to allow users to flexibly label and register those special conditions (Supplementary Figure 4). Thanks to this feature, adding, editing, and removing tags to samples or analysis is a powerful way to organize the data. It is possible to label the samples that are part of a certain project or those that are used as internal controls, for example. If you implement tags, consider allowing filtering for one or more of them.

Another feature requested by the users was a list of low-coverage genomic regions for each sample, also known as failed regions. For any set of samples, users can obtain the exact bases with low coverage (Supplementary Figure 5). This may be useful to decide, for example, whether some regions should be tested by Sanger sequencing or not.

Tip 4: A relational database to manage complex data

Access to and exploitation of data is determined by how data is stored. In research settings, data typically resides in the results files produced by the analysis pipelines. Exploiting these files is enough in most research contexts where the needs are very specific: usually, a single bioinformatician will process these files to generate some analyses, reports, or figures. However, using a relational database, instead of multiple files, provides multiple advantages in a diagnostic context. First, there are complex relationships underlying the data behind the diagnostic workflow that can be easily represented through a relational database. For instance, a patient can have more than one sample, the targeted enrichment process can be performed multiple times on a single sample, and the enriched DNA can be sequenced on several sequencing runs. These one-to-many relations are difficult to represent efficiently in text-based file formats but are easily implemented in a database, which allows powerful data exploitation. Second, current database engines can fluently manage tables of up to millions of rows, especially when using table indexes. This is an advantage when thousands of variants with dozens of fields have to be retrieved to review multiple patient samples. Third, a database is a prerequisite to allow multiple users to modify the same data at the same time. It is needed when, for example, multiple geneticists are validating and classifying the variants found in an urgent sample. Fourth, a relational database is the best structure to store and manage the traceability information explained in Tip 1, by creating the appropriate tables or columns. It would be unfeasible to store, access, and modify all user actions in a single or multiple conventional files.

In the Pandora database, the definition of the unique variants that have been found to the moment (what we called Variants Library) is separated from the definition of all the variants in patients with the quality metrics they were found. With this design, when a new run is analyzed, a big proportion of the variants automatically appear already classified to the user because they were previously found, classified, and stored in the Variants Library. For instance, in our last sequencing run in Pandora, 98.2% of a total of 2251 variants found across all samples appeared automatically classified to the user. This approach positively impacts the variant revision process, speeding up the reporting of variants to the patients.

Tip 5: Take care of the usability of your tool

Usability, a feature that encompasses effectiveness, efficiency, and user satisfaction, is a key factor in the success of any software tool (Yan and Guo, 2010). Some user interface (UI) design decisions can have an important impact on the user experience and the productivity of your genetic diagnostic tool. For example, chosen colors are important to facilitate human data read. To decrease the cognitive load (Wang *et al.*, 2014), try to reach a color harmony avoiding color saturation overload, and use colors to differentiate or highlight useful information on certain fields. Supplementary Figure 6 illustrates the use of colors in Pandora. As an example of highlighting, the column “Classification Date” becomes yellow if the classification is older than six months. Another feature that can improve usability is to provide a universal search box to allow quick and easy search for any item of the related table (Supplementary Figure 7).

Usability should also consider a design focused on avoiding user errors. In this regard, show a confirmation message whenever an action is important or risky, and check comprehensively for any error in the form fields before registering them in the system. Also, it is a good practice to disable those fields that shouldn't be modified. As an example, when a user downloads the sample sheet to send it to the Illumina platform, many sample fields that shouldn't be modified anymore are disabled.

For a real good user experience, an agile UI response is mandatory. If common user actions like modifying a field require seconds to be performed, the productivity and satisfaction of your users will be limited. So, emphasize the optimization of those processes that can result in time-consuming bottlenecks. Consider also to preload the most common data when the software tool is launched, even if that means an initial load of multiple seconds. In our experience, users prefer a slow initial web load of multiple seconds instead of waiting some seconds per each common action.

Tip 6: A powerful data table UI library to explore powerful data

The analysis pipeline will produce from hundreds to millions of variants that have to be properly exploited. The user may be interested in exploring, for example, all the variants identified in a sequencing run. Using a table is a good choice to manage all this information, but it must be an efficient, flexible, and fully-featured data table implementation. In our case, we chose the DataTable implemented in Webix, https://webix.com/, which supports large datasets of up to a million rows. As an example, in our Variants Library (Supplementary Figure 8) the user can quickly explore all the variants that have been found at least once in our samples: currently 65450 rows (variants) with 67 columns, that is, more than 4 million fields. Of course, it's not only about quantity. The data table UI library should allow the user to quickly filter by any column, including multiple columns, as Pandora does. Also, sorting by any column or multiple columns is a requirement to empower the user when managing the data.

Additionally, the table UI library should allow the representation of tree relationships. For instance, a patient can have multiple samples and a sample can have multiple sample captures. Note that this feature is not available in most table UI libraries. Users can also benefit from other features like fixing the position of the most useful table columns (Gene, cDNA, protein, and transcript NM in Pandora) or expanding the information shown for a certain row. The latest is a feature that we use to show the variant nomenclature for all the gene transcripts. By just clicking, the user can unfold a subview containing a list of the annotations of that variant for all the known transcripts in that gene (Supplementary Figure 9).

Tip 7: Screen is a scarce resource: optimize the information shown

Screen is the limited space in which the user will work. Accordingly, it is important to optimize the space showing only the interface elements that are useful to the user depending on the task to be done. Irrelevant elements should be hidden and you should provide mechanisms to show extra content in case it is required by the user.

Hence, consider showing customized views depending on the task. The validation and classification of a variant require some columns in common to be shown, such as the gene, the cDNA annotation, or the genomic position. However, other fields are exclusive to each task for providing the exact information useful to the user. For example, when validating a variant, the read depth, the allele frequency, and the Joint Quality Score (a custom metric to estimate the confidence of a variant call) are shown. For the classification of a variant, that information is hidden but show all the fields related to this task: classification reasoning, comments, and in-silico predictors, among others. In Pandora, the user can select the columns to be shown depending on the task (Supplementary Figure 10).

A minor but useful feature for the user is showing tooltips for any field or data, so the whole field content can be shown when the mouse is over. In fact, this feature can be extended to build complex tooltips that summarize a set of fields useful for the user (Supplementary Figure 11). Moreover, making use of context buttons helps to optimize the screen space. The idea is to offer the user only options based on the item selected and its state. For instance, if the user selects an individual, provide only the buttons applicable to this item, like removing or editing the individual entry.

Tip 8: Real-time synchronization to allow multiple users to work in parallel

The review or classification of a variant is a process that may involve multiple users. If you work with shared local excel files to check the variants found in a sample, your curators will not be able to work at the same time in the same file because the excel file will remain locked when a user opens it. As explained in Tip 4, a database allows for simultaneous access or modification of data. However, users also need to work with the last version of the data in order to make appropriate decisions. Therefore, real-time synchronization should be requisite in your tool: the user interface should be updated as soon as it is modified in the database.

Tip 9: Analysis pipeline for diagnostics

Diagnostics-oriented NGS data analysis pipelines have a number of specific characteristics and constraints that affect their design. Alterations found (or missed) by the data analysis pipeline will be reported back to patients and have a direct impact on their clinical management. Therefore, a clear and interpretable path from data to results might be required by the geneticists writing the patient's reports. In addition, the balance between specificity and sensitivity must be evaluated taking into account the different costs associated with false positives and false negatives. For example, in germline diagnostic orthogonal validation of variants is frequent, which allows for a lower specificity at the data analysis pipeline with no impact on the final reports. On the contrary, this is not always the case on somatic genetic testing, where specificity will need to be higher since orthogonal validation is more challenging.

As an example, the data analysis pipeline implemented in Pandora is essentially a standard germline small variant and copy number calling pipeline. It uses many standard bioinformatics tools as building blocks, such as bwa mem for mapping and Annovar for variant annotation. However, to improve interpretability and provide finer controls on sensitivity and specificity, Varscan2 (Koboldt *et al.*, 2012) was chosen instead of more popular choices such as GATK (McKenna *et al.*, 2010), Strelka (Kim *et al.*, 2018) or Freebayes (Garrison and Marth, 2012), since Varscan does not use complex statistical models but a simple method based on hard filters. With that, we can clearly understand why a variant is called or missed and bias the variant calling towards better sensitivity at the cost of slightly lower specificity, given that the diagnostic strategy includes the validation by Sanger sequencing of all reported variants (Castellanos *et al.*, 2017). Similarly, for CNV calling, DECoN (Fowler *et al.*, 2016) with parameters adjusted for high sensitivity (Moreno-Cabrera *et al.*, 2020) was chosen, since all reported alterations are validated by MLPA.

Finally, it is important to give the users the possibility to complement or override the results from the analysis pipeline. For example, even though your pipeline might provide perfect naming for most variants, users must be able to change them because exceptions occur. Hence, consider having in your database a user-variant nomenclature field beside the original pipeline-variant nomenclature field. Also, your tool should provide a mechanism, probably a form, to insert additional variants into the system because some patient variants might be found by other means. It is necessary to store them in the database to keep it fully updated, so data can be properly exploited. In summary, the analysis pipeline must be rigid and reproducible but the tool must provide users with sufficient flexibility to accommodate real-life complexity.

Tip 10: Your tool will be the center of everything, but it cannot be everything

Although your tool will be the orchestrator of the elements involved in the genetic diagnostics workflow, it has to be designed considering that interoperation with other tools will happen. Very likely, your users are used to work with spreadsheets like Microsoft Excel or LibreOffice Calc, so allowing data export to excel files from any of the data tables shown in your application is a must. Even if you implement many functions to operate with the data tables, like filtering or sorting by multiple columns, allowing exportation to excel-like formats allows the user to make use of more sophisticated features that you will not implement. On the other hand, excel-files will probably be the natural input for many steps in the diagnostic workflow, so prepare your code to parse them accordingly. In Pandora, the information of patients, samples, and desired analyses can be uploaded via an excel-like file, apart from adding them manually on the web tool. On the other hand, it is possible to export variants to a format compatible with Progeny, a genetic clinical software used in our laboratories. In short, prepare your tool to work harmoniously with other tools and systems that live around.

Tip 11: Adequate security protections for genomic data

The human genome has specific properties that make it sensitive information. The DNA sequence is unique for each individual, stores ethnic heritage information, allows for the identification of relatives, and provides information about the predisposition to several diseases (Naveed *et al.*, 2015; Mohammed Yakubu and Chen, 2020). Therefore, human genome data are susceptible to privacy risk and security attacks. The software you are designing must emphasize security aspects to avoid vulnerabilities (Atashzar *et al.*, 2011; Mouli and Jevitha, 2016). Security requirements should be extended to all infrastructure parts, such as the database or the web server. Here, the role of your system administrator is key to establish proper security restrictions. In Pandora, the web server is the only part that interoperates with the “outside world” (demilitarized zone). In fact, the web server cannot communicate with any other system except the database server. In addition to username/password protection, the access to the web server is restricted to a list of predefined IPs, also known as IP whitelisting. This measure might hinder access from outside of usual networks, but enhances security; anyway, external access can be solved by using a virtual private network (VPN), for example.

Conclusion

Developing a new software tool for supporting the genetic diagnostics of a disease is a challenging task that requires thinking carefully about its requirements. The tips presented in this work aim to be a useful checklist to contribute to the success of the tool you are planning. The eleven pieces of advice have the ultimate goal of providing an added value to go far beyond the excel files in which, probably, the information was previously stored. To succeed in this task, all of them should be considered to finally automate, and even enhance, the genetic diagnostic workflow of your laboratory.

Acknowledgements

We thank all the members of the Unit of Genetic Diagnostics of the Hereditary Cancer Program of the Catalan Institute of Oncology (ICO-IDIBELL) and the Genetics Diagnostics Unit of the Hereditary Cancer Group of the Germans Trias i Pujol Research Institute (IGTP). We also thank Iñaki Martínez de Ilarduya for his enormous help as system administrator. We thank CERCA Programme/Generalitat de Catalunya for institutional support.

Results summary

Director's report

As the supervisors of the doctoral thesis of José Marcos Moreno Cabrera, entitled "**A translational bioinformatics approach to improve genetic diagnostics of hereditary cancer from next-generation sequencing data**", we certify that the PhD candidate has actively participated in the design, experimental work, implementation, analysis of the results and their discussion, drawing conclusions and writing manuscripts of the work included in this thesis. Specific work contributions are listed below, along with the impact factors (IF) of the journals where the results were published (if applicable). These articles have not been presented as part of other doctoral theses.

The results obtained in this thesis are compiled in 4 manuscripts. At the time of submitting this thesis, 2 articles were published, one was accepted for publication, and another one had been submitted to scientific journals.

Article 1

Evaluation of CNV detection tools for NGS panel data in genetic diagnostics

José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro* & Bernat Gel*

European Journal of Human Genetics. Volume 28, pages 1645–1655(2020). doi: 10.1038/s41431-020-0675-z

IF 2019 = 3.657. Rank 55/177 (GENETICS & HEREDITY), Third decile, second quartile (Q2).

Contribution of the PhD candidate: systematic review of published articles, CNV calling tool candidate selection, dataset selection and preparation, benchmark with default parameters, parameter optimization using a custom greedy algorithm, diagnostics scenario evaluation, design and implementation of CNVbenchmarkeR, statistical analysis, analysis and presentation of the results, and manuscript preparation.

Article 2

Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer

José Marcos Moreno-Cabrera, Jesús del Valle, Lidia Feliubadaló, Marta Pineda, Sara González, Olga Campos, Raquel Cuesta, Joan Brunet, Eduard Serra, Gabriel Capellà, Bernat Gel* & Conxi Lázaro*

Journal of Medical Genetics. Published Online First: 20 November 2020. doi: 10.1136/jmedgenet-2020-107366

IF 2019 = 4.943. Rank 30/177 (GENETICS & HEREDITY), Second decile, first quartile (Q1).

Contribution of the PhD candidate: Germline NGS screening in the retrospective and prospective cohorts, Bayesian factor analysis, design and implementation of the NGS CNV detection strategy, analysis and presentation of the results, literature review, and manuscript preparation.

Article 3

CNVfilteR: an R/Bioconductor package to identify false positives produced by germline NGS CNV detection tools

José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Eduard Serra, Gabriel Capellà, Conxi Lázaro* & Bernat Gel*

Manuscript accepted for publication in Bioinformatics (April 2021).

IF 2019 = 5.610. Rank 3/59 (MATHEMATICAL & COMPUTATIONAL BIOLOGY), First decile, first quartile (Q1).

Contribution of the PhD candidate: literature review, design and implementation of the R/Bioconductor package, systematic evaluation of the tool, analysis and presentation of the results, and manuscript preparation.

Article 4

Eleven quick tips to build a software tool for integral management of genetic diagnostics

José Marcos Moreno-Cabrera, Jesús del Valle, Lidia Feliubadaló, Marta Pineda, Sara González, Olga Campos, Raquel Cuesta, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro* & Bernat Gel*

Submitted to The Journal of Molecular Diagnostics (April 2021).

Contribution of the PhD candidate: analysis, design and implementation of the web-based tool, improvement and maintenance of the tool, and manuscript preparation.

Conxi Lázaro García, PhD
Hereditary Cancer Program
Catalan Institute of Oncology
Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)
clazaro@iconcologia.net

Bernat Gel Moreno, PhD
Hereditary Cancer Group
Germans Trias i Pujol Research Institute (PMPPC-IGTP)
Campus Can Ruti
bgel@igtp.cat

Benchmark of NGS CNV calling tools in genetic diagnostics

This study aimed to perform a benchmark of different bioinformatic tools to identify putative CNVs from NGS data. Tools were benchmarked using up to three metrics (per gene, per region-of-interest (ROI), whole diagnostics strategy), against four datasets from different sources, and in three different contexts: using default parameters, using greedy-optimized parameters, and against the augmented datasets in the diagnostics scenario. Main results are summarized below.

- After a literature search, we selected five tools showing good performance for evaluation: CoNVaDING v1.2.0, DECoN v1.0.1, panelcn.MOPS v1.0.0, ExomeDepth v1.1.10, and CODEX2 v1.2.0.
- When performing the benchmark with default parameters:
    - Regarding the per ROI metric, most tools showed sensitivity and specificity values greater than 0.75, and sensitivity was generally over 0.9. However, tools' performance varied across datasets. Tools’ specificity remained over 0.98 and sensitivity over 0.94 when using the ICR96 and panelcnDataset datasets. In the in-house datasets, tools performed worse and only CoNVaDING obtained a sensitivity close to 1 at the expense of lower specificity.
    - Regarding the per gene metric, tools behaved slightly better compared to per ROI. At least one tool detected all CNVs in each dataset.
- When performing the benchmark with sensitivity-optimized parameters:
    - In general, the optimization process improved sensitivity by slightly decreasing specificity, but the amount of improvement was different for the different datasets. For panelcnDataset, sensitivity was improved by a higher margin because of CODEX2, which increased its sensitivity by 58.6%. On the other hand, tools didn’t improve or showed small differences in the In-house MiSeq dataset.
- When performing the benchmark in the diagnostics scenario:
    - For the In-house MiSeq dataset, two tools detected all CNVs. panelcn.MOPS reached 100% sensitivity with both optimized and default parameters, with a specificity of 67.8% and 80.7%, respectively. DECoN detected all CNVs only with

the optimized parameters achieving 91.3% specificity. CoNVaDING also detected all CNVs, but its high no-call rate led to very low specificity, 4.1%.

  - For the In-house HiSeq dataset, only panelcn.MOPS detected all CNVs with specificity over 80%: 81.5% with the default parameters and 83.2% with the optimized parameters. DECoN missed only one CNV (a mosaic sample), and its specificity remained high, 96.6% with the optimized parameters.

- We developed an R framework, CNVbenchmarkeR, to perform the benchmark in an automatic and configurable way. Code and documentation are freely available at https://github.com/TranslationalBioinformaticsIGTP/CNVbenchmarkeR.

Prospective and retrospective evaluation of DECoN as first-tier screening

Based on the results of our previous benchmark (Article 1), we chose DECoN for CNV screening in our laboratory routine. Our NGS CNV detection strategy consisted of two steps: first, screening of all genes of clinical interest using DECoN with optimized parameters and second, MLPA-validation of putative CNV calls with sufficient statistical support (BF >= 2). We applied our NGS CNV detection strategy to test a retrospective and a prospective cohort.

- CNV screening in the retrospective cohort (1860 patients):
  - For the 13687 genes analyzed by DECoN across all samples, six new true CNVs were identified and confirmed by MLPA (a 30% increase, from 20 previously known CNVs to 26 CNVs), while 128 FP calls were produced. The number of genes tested by MLPA after DECoN was 221, in contrast to the 2660 required with the previous MLPA- restricted strategy. Also, the average number of genes evaluated per sample with this new strategy was 7.35 compared with 1.43 using our previous MLPA-based strategy.
- CNV screening in the prospective cohort (2041 patients):
  - For the 18836 genes analyzed by DECoN across all samples, DECoN produced 139 FPs and 19 true positives confirmed by MLPA. Out of those true positives, 8 would have not been identified using our previous MLPA-restricted approach (a 72.7% increase, from 11 to 19 CNVs). Compared with our previous MLPA-restricted strategy, the number of genes tested by MLPA across all samples decreased from 3442 to 240, representing a 93.0% decrease. Also, the average

number of genes tested per sample increased from 1.69 to 9.22 with the new DECoN-MLPA strategy.

R/Bioconductor package to identify false positives produced by NGS CNV calling tools

- We developed an R/Bioconductor package, CNVfilteR, to identify false-positive calls generated by CNV calling tools from germline NGS data. CNVfilteR uses SNVs' allele frequency to detect both false deletions and false duplications. A CNV deletion call is identified as a false positive if there is at least a certain percentage of heterozygous SNVs in the CNV. Also, a CNV duplication call is identified as a false positive using a fuzzy-logic-inspired model which scores all heterozygous SNVs overlapping the CNV.
    - CNVfilteR provides a function to automatically interpret VCFs produced by VarScan2, Strelka/Strelka2, freeBayes, HaplotypeCaller (GATK), and UnifiedGenotyper (GATK).
    - Results can be plotted and customized via plotting functions based on karyoploteR (Gel and Serra, 2017) and CopyNumberPlots packages.
- CNVfilteR was evaluated on callsets generated by 13 CNV calling tools on multiple samples:
    - For 3 reference WGS samples, CNVfilteR identified between 17.0% and 50.4% of the FPs and the FDR decreased for all tool-sample evaluations (up to 14.0%). F1-score was improved in 19 out of the 24 tool-sample evaluations (up to 20.8% increase). On the contrary, tool-sample evaluations had an absolute sensitivity decrease of between 0.001 and 0.035.
    - For the HiSeq-panel (411 samples) and MiSeq-panel datasets (130 samples), CNVfilteR identified 15% of the FP calls (3 out of 20) and 12.5% of the FP calls (2 out of 16). Sensitivity remained the same.

Web-based software tool to manage the diagnostics process in the laboratory routine

We developed Pandora to support the diagnostics routine of the ICO-IGTP Joint Program on Hereditary Cancer. Pandora is a web-based tool built on the Django framework, a PostgreSQL database, and a number of Python and R cron jobs. It automates multiple tasks, ensures data traceability and reproducibility, supports the validation and classification of variants, and allows flexible data exploration.

- Fruit of our experience designing and building Pandora, we described eleven tips to consider when facing the development of a tool to support the genetic diagnostic testing of a hereditary disease.

Discussion

The overall goal of this PhD thesis was to improve the genetic diagnostics of hereditary cancer using multiple analytical and bioinformatic approaches. We carried out four studies to achieve this aim (Figure 12).

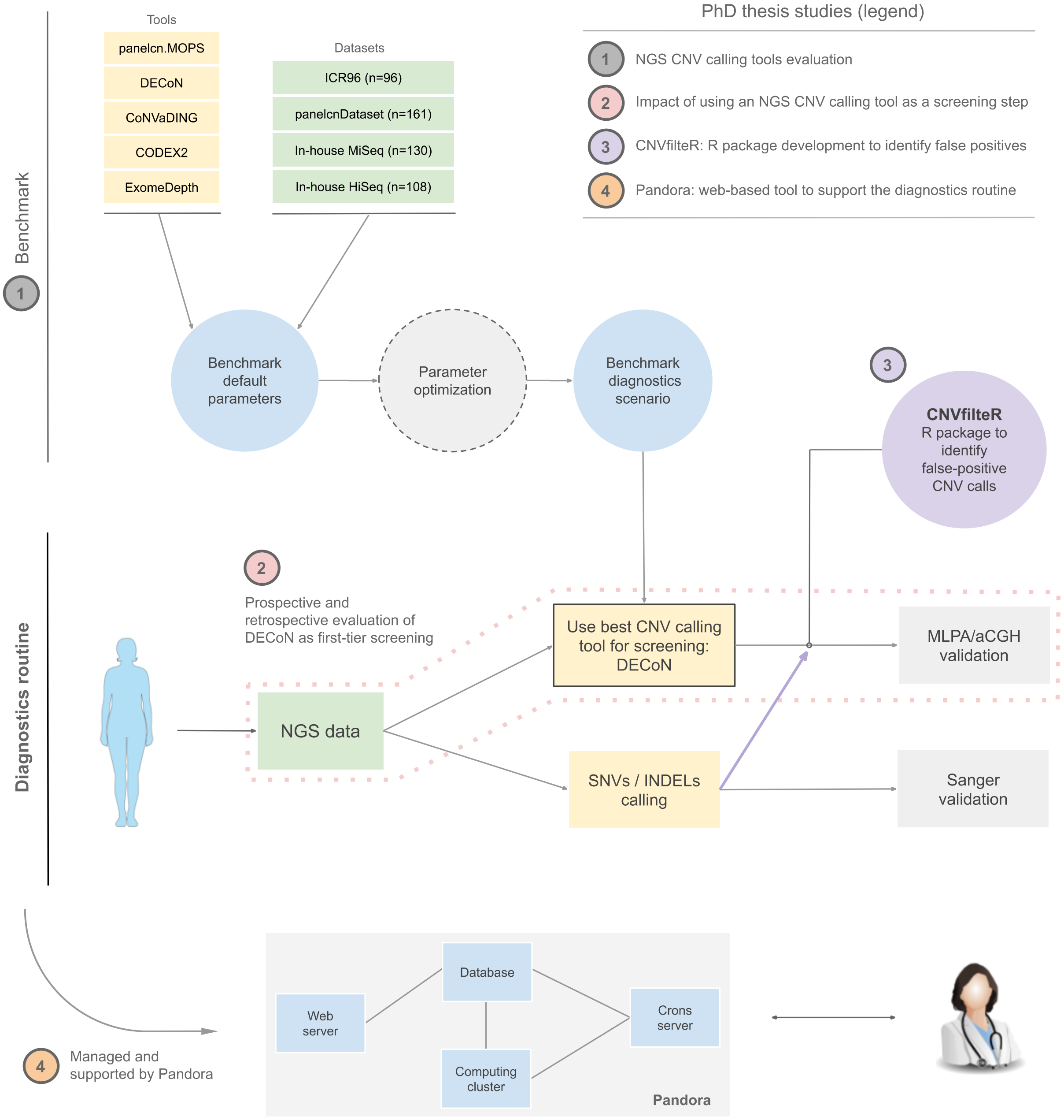

Tools
panelcn.MOPS
DECoN
CoNVaDING
CODEX2
ExomeDepth
Datasets
ICR96 (n=96)
panelcnDataset (n=161)
In-house MiSeq (n=130)
In-house HiSeq (n=108)
PhD thesis studies (legend)
1 NGS CNV calling tools evaluation
2 Impact of using an NGS CNV calling tool as a screening step
3 CNVfilteR: R package development to identify false positives
4 Pandora: web-based tool to support the diagnostics routine
Benchmark
1
Benchmark default parameters
Parameter optimization
Benchmark diagnostics scenario
3
CNVfilteR R package to identify false-positive CNV calls
Diagnostics routine
2
Prospective and retrospective evaluation of DECoN as first-tier screening
Use best CNV calling tool for screening: DECoN
MLPA/aCGH validation
NGS data
SNVs / INDELs calling
Sanger validation
4 Managed and supported by Pandora
Web server
Database
Computing cluster
Crons server
Pandora

Figure 12. Scheme showing the four studies of this thesis: the NGS CNV calling tools benchmark (1), the evaluation of an optimized NGS CNV calling tool as a screening step before MLPA validation (2), the development and implementation of a R package (CNVfilteR) to identify false positives produced by germline CNV callers (3), and the web-based tool to support the diagnostics routine, Pandora (4).

The first study aimed to fill the gap of existing NGS CNV calling-tool benchmarks on targeted gene panel data, focusing on genetic diagnostics. The second study took advantage of the benchmark results and evaluated the impact of using our selected tool as first-tier screening in our diagnostics routine. The third study focused on the identification of false positives produced by NGS CNV calling tools: an R/Bioconductor package was developed to identify them. The fourth subproject consisted of the design and implementation of a web-based tool to manage the diagnostics process during the laboratory routine, which allowed us to describe a set of recommendations to help other laboratories when building a software tool for genetic diagnostics.

The discussion is divided into four sections corresponding to each work, along with a final section addressing the translational nature of all of them.

1 Benchmark of NGS CNV calling tools in genetic diagnostics

1.1 Background and motivation

Although NGS performs well when calling germline SNVs and small INDELs, the detection of larger variants like CNVs is still challenging. The identification of such variants is a matter of interest since germline CNVs are the genetic cause of multiple hereditary diseases (Zhang *et al.*, 2009). The gold standards for CNV testing in genetic diagnostics are MLPA and aCGH (Talevich *et al.*, 2016; Kerkhof *et al.*, 2017). Both methods are time-consuming and costly, which frequently leads laboratories to test only a subset of the genes of clinical interest.

Many NGS CNV detection tools have been developed to date (Zhao *et al.*, 2013; Abel and Duncavage, 2013; Mason-Suares *et al.*, 2016). Most of them perform well on large CNVs (in the order of megabases) but are not able to reliably identify small CNVs that affect only one or a few exons. These small CNVs, however, are frequently involved in several genetic diseases (Truty *et al.*, 2019). In addition, most published NGS CNV detection tools were designed to work from WGS and WES data and perform worse on the sparser data of targeted gene panels.

In genetic diagnostics, using a testing method with low sensitivity leads to a higher number of misdiagnoses. In this context, NGS CNV calling tools have not been properly evaluated on NGS gene panel data. Although some benchmarks of CNV calling tools on targeted NGS panel data have been published, they suffer from some important limitations. They were performed by the authors of the tools and were executed against a single dataset (Johansson *et al.*, 2016;

Fowler *et al.*, 2016; Povysil *et al.*, 2017; Kim *et al.*, 2017; Chiang *et al.*, 2019) or mostly used simulated data with a small number of validated CNVs (Roca *et al.*, 2019).

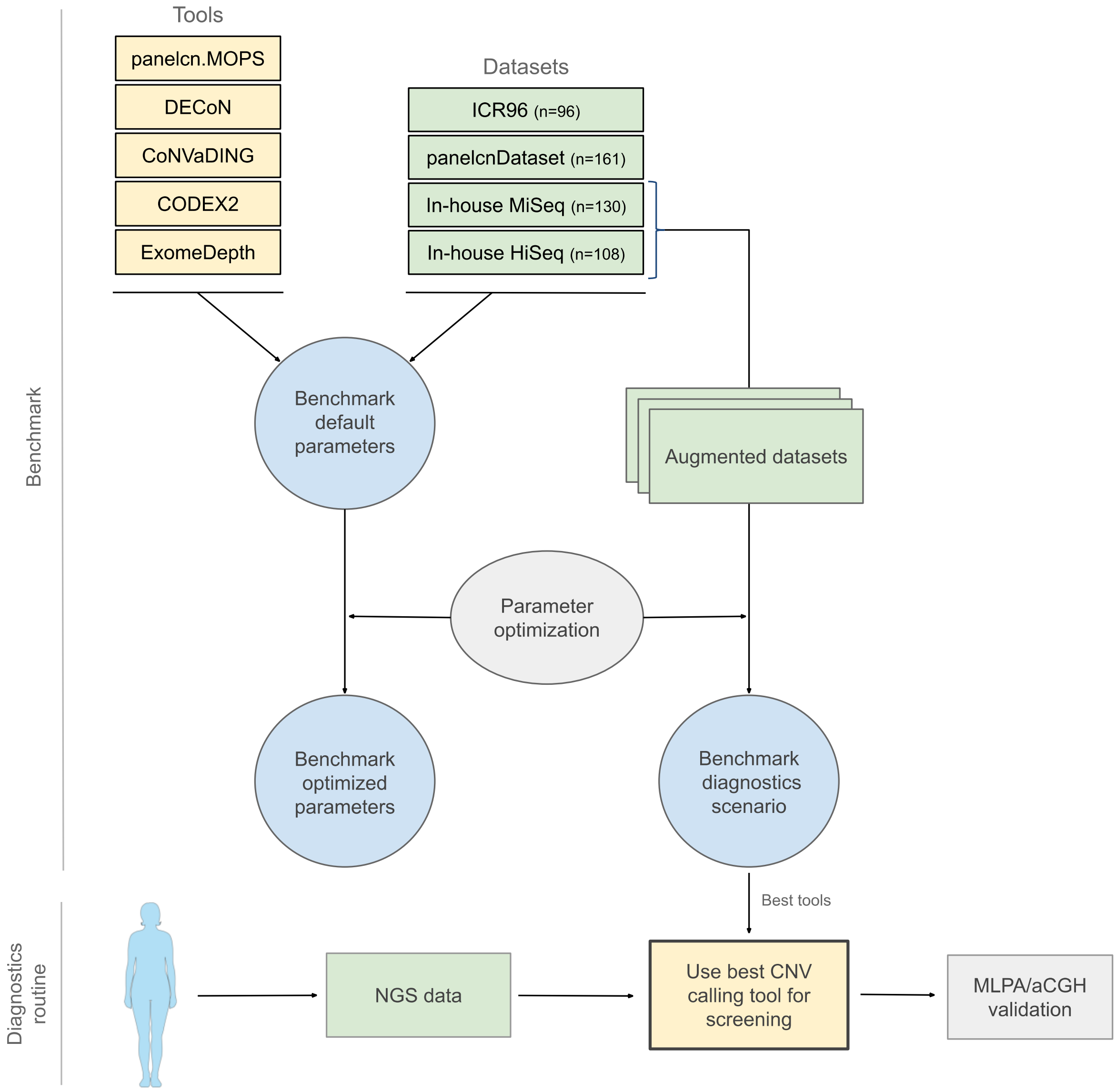


Figure 13. Benchmark design and the aim of applying the results in the diagnostics routine.

The first study of this thesis aimed to perform an independent benchmark of multiple NGS CNV calling tools, optimizing and evaluating them against multiple non-simulated and validated datasets, to identify the most suitable tools to be used for genetic diagnostics (Figure 13). In this work, we selected 5 tools (Article 1, Table 2) that had shown promising results on panel data, and we measured their performance, with the default and sensitivity-optimized parameters, on 4 validated datasets from different sources (Article 1, Table 1). We also evaluated the NGS CNV calling tool performance in a genetic diagnostics scenario and showed that some of the tools are suitable for using as screening methods before MLPA or aCGH confirmation. Apart from this, a framework for evaluating and optimizing CNV calling tool performance, CNVbenchmarkeR, was developed. Method details such as benchmark evaluation metrics or data processing are available in the original manuscript (Article 1, Methods).

1.2 Benchmark with default parameters

Most tools were highly sensitive and specific when using the default parameters in the evaluation, although the top performers depended on the specific dataset (Article 1, Figure 2). Tools performed better on the panelcnDataset dataset, where DECoN, ExomeDepth, and CoNVaDING reached almost 100% sensitivity and specificity when using the per ROI metric (Article 1, Methods). A possible explanation is that this dataset contains the lowest number of single-exon CNVs (n=13), which are the most difficult type of CNVs to detect. On the other hand, DECoN was the best performer for ICR96, a dataset published by the same authors, although other tools obtained similar results in this dataset. Regarding our in-house datasets, CoNVaDING was the most sensitive tool. This tool, however, showed the lowest positive predictive value (PPV) in all datasets with the exception of the panelcnDataset. On the contrary, ExomeDepth showed the highest PPV in all datasets, making it one of the most balanced tools regarding sensitivity and specificity. Differences in tool performance depending on the dataset were also observed in previous works (Hong *et al.*, 2016; Sadedin *et al.*, 2018).

In summary, benchmark results evidenced a very high and dataset-dependent performance of the tools. Even before any parameter modification, sensitivities were high and reached 100% for some tools and datasets in the per gene metric (Article 1, Supplementary File 11). These results showed the potential of some of these tools for using in a genetic diagnostics setting.

1.3 Benchmark with optimized parameters

Most CNV calling tools had been developed to be used in a research setting, so their parameters were optimized to meet a certain sensitivity-specificity balance. However, sensitivity is a priority in a genetic diagnostics setting. In this context, false negatives have to be avoided when using an NGS CNV calling tool as a screening step before orthogonal validation. Therefore, to optimize tools' performance for using as screening tools in genetic diagnostics, we need to modify their default parameters. The aim was to maximize the sensitivity, even at the expense of lowering their specificity. This parameter optimization must be performed in a dataset-specific way since tools perform differently depending on the dataset specificities: target region composition, technical differences, or sequencing characteristics.

We performed a parameter optimization for each tool-dataset combination. The optimization algorithm followed a greedy approach: a local optimization at each step with the aim of obtaining a solution close enough to the global optimum (algorithm details in Article 1, Supplementary File 9- Appendix B). Also, each dataset was split into two halves, a training set used to optimize tool parameters and a validation set to evaluate them.

In general, the optimization process improved sensitivity and slightly decreased specificity (Article 1, Figure 3). For one dataset (panelcnDataset), sensitivity increased notably, although most of the improvement was driven by CODEX2, which increased its sensitivity by 58.6%. The optimization had a different impact on different tools: while CODEX2 showed a higher sensitivity in all four datasets, the rest of the tools showed modest improvements. On the other hand, tools were not able to improve or showed small differences in the In-house MiSeq dataset.

One likely explanation for these improvement differences is that the tools had little room for improvement: sensitivity was over 0.9 for most dataset-tool combinations before the optimization, so the number of false negatives in each subset was very small (between 4 and 8) in the per gene metric. Moreover, there is a limitation imposed by the size of the datasets. Each tool was optimized on a training set containing only a small number of samples: between 48 and 80 samples. In an ideal scenario, having a larger dataset would allow the greedy algorithm to better train because the more samples in a dataset, the more information the algorithm has for training. Also, a larger dataset allows for better validation, which is also valuable.

The final optimized parameters were dataset specific, so we do not recommend using them directly on other datasets where the data was obtained in a different manner (different capture protocol or sequencing technologies, for example). Based on our results, we would recommend optimizing the parameters for each specific dataset before adding any CNV calling tool to a genetic diagnostics pipeline to maximize its sensitivity and reduce the risk of misdiagnosis.

1.4 Benchmark in the diagnostics scenario

As previously mentioned, one of the aims of this study was to identify the most suitable tools for using in genetic diagnostics. An NGS calling tool with a very high sensitivity could be used as a screening step prior to MLPA or aCGH validation, decreasing the number of MLPA/aCGH tests. Hence, in a real diagnostic setting, it is not only necessary to confirm, via an orthogonal method, all the CNVs detected in genes of interest, but also all regions where the screening tool was not able to produce a result (no call) should be tested via an orthogonal technique. To take this into account, we evaluated the performance of all tools using the whole diagnostics strategy metric which takes the no calls into account (Article 1, Methods). This evaluation was performed in a modified version of the in-house datasets, the augmented in-house datasets (Figure 13), which contained all the samples from the original sequencing runs instead of just a selection of them (Article 1, Methods).

Results showed that DECoN and panelcn.MOPS obtained a performance high enough to be implemented as screening methods in our two in-house datasets (Article 1, Figure 4). While panelcn.MOPS detected all CNVs both with the default and the optimized parameters, DECoN reached almost perfect sensitivity and outperformed panelcn.MOPS specificity when using the optimized parameters, although the difference was not statistically significant. In fact, DECoN only missed a mosaic CNV affecting two exons of the *NF2* gene. Since benchmarked tools use depth of coverage strategies, detecting a mosaic sample is very challenging due to the lower coverage impact of these mutations. On the other hand, CoNVaDING also detected all CNVs, but the high number of no-call regions reduced its specificity to values between 4.1 and 21.9%, which made it unsuitable as a screening tool.

The parameter optimization process improved the sensitivity of most tools. For example, for the In-house MiSeq dataset, DECoN sensitivity increased from 98.4% (CI: 91.6–100%) to 100% (CI: 94.4–100%), and the specificity increased from 78.5% (CI: 71.6–84.4%) to 91.3% (CI: 86.0–95.0%). This improvement highlights the importance of fine-tuning the tool parameters for each specific task and shows that the optimization process performed in this work has been key for the evaluation of the different tools.

The high sensitivity achieved by DECoN and panelcn.MOPS in different datasets, where they identified all known CNVs, evidences that NGS data can be used as a CNV screening step in a genetic diagnostics setting. Of course, this screening step has the potential to improve genetic diagnostics routines. For example, the high specificity achieved by DECoN in the in-house MiSeq dataset with the optimized parameters means that around 91% of genes with no CNV would not need to be specifically tested for CNVs when using DECoN as a screening step. Hence, the lower number of necessary MLPA/aCGH tests could save resources or, potentially, could be used to expand the number of genes tested with the aim of increasing the final diagnostic yield. This hypothesis was addressed in the second study presented in this PhD thesis.

1.5 CNVbenchmarkeR

Both the CNV benchmark and parameter optimization were performed using a custom R framework, CNVbenchmarkeR. The code provides a set of scripts and helpers to evaluate germline NGS CNV calling tools in different NGS datasets. The current version supports DECoN, CoNVaDING, panelcn.MOPS, ExomeDepth, and CODEX2 tools.

CNVbenchmarkeR was developed to make the automation and configuration of the benchmark easier under different combinations of tested tools, tested datasets, and tool parameters. The code has an obvious limitation or prerequisite: NGS calling tools have to be properly installed before benchmark execution. Also, it is strict in processing the input parameters: only certain formats are allowed. Open-source code, how-to-use guide, and documentation are publicly available at https://github.com/TranslationalBioinformaticsIGTP/CNVbenchmarkeR to help other laboratories perform the testing and optimization process in any new dataset.

2 Prospective and retrospective evaluation of DECoN as first-tier screening

2.1 Background and motivation

Germline CNVs are one of the genetic causes underlying hereditary diseases (Zhang *et al.*, 2009), so its testing is recommended in any comprehensive genetic testing strategy. For many years, MLPA has been the gold standard for CNV detection when testing one or a few genes (Kerkhof *et al.*, 2017), but its turnaround time and costs lead diagnostic laboratories to limit CNV testing to a few key candidate genes instead of testing all genes of clinical interest. On the other hand, the use of NGS technologies in genetic diagnosis is a very cost-effective approach that allows for the analysis of multiple genes and samples at once. However, implementing sensitive and specific bioinformatic algorithms for CNV detection from NGS data has been more challenging than for other types of mutations, such SNVs or small INDELs.

Some authors have argued against the use of an NGS CNV detection tool in a clinical setting due to performance limitations, (Mason-Suares *et al.*, 2016; Yao *et al.*, 2019) especially for single-exon CNVs. Nevertheless, our previous study (Article 1) evidenced that there are germline NGS CNV calling tools with sufficient sensitivity for using as a screening step prior to orthogonal validation, even for single-exon CNVs. In particular, DECoN (Fowler *et al.*, 2016) detected all CNVs (except one in a mosaic sample) with a specificity of over 90% in our diagnostic datasets. Once that benchmark was performed, it was of interest to evaluate the clinical impact of its implementation in the clinical routine, in terms of detection yield and costs.

Therefore, we aimed to evaluate the impact of using DECoN as a screening method in a hereditary cancer genetic diagnostics setting, testing it in a retrospective and in a prospective cohort, and comparing it against our previous MLPA-restricted strategy (Figures 14-15).

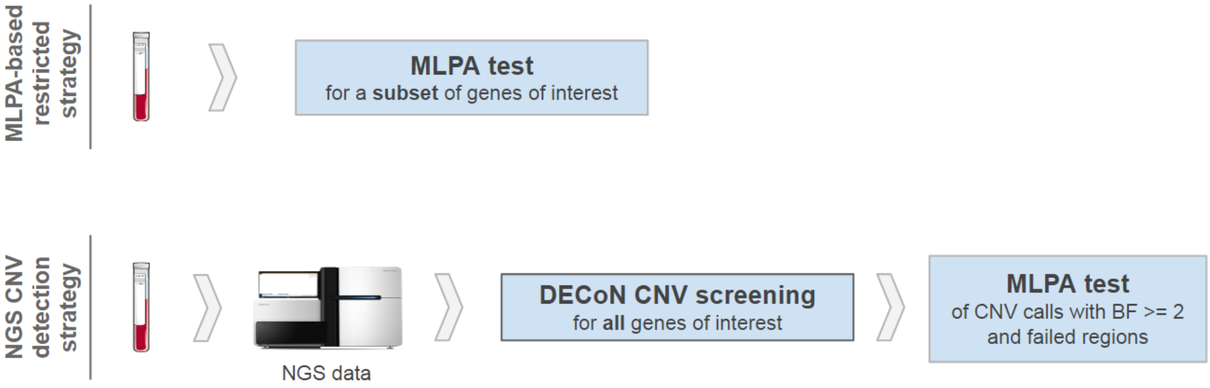
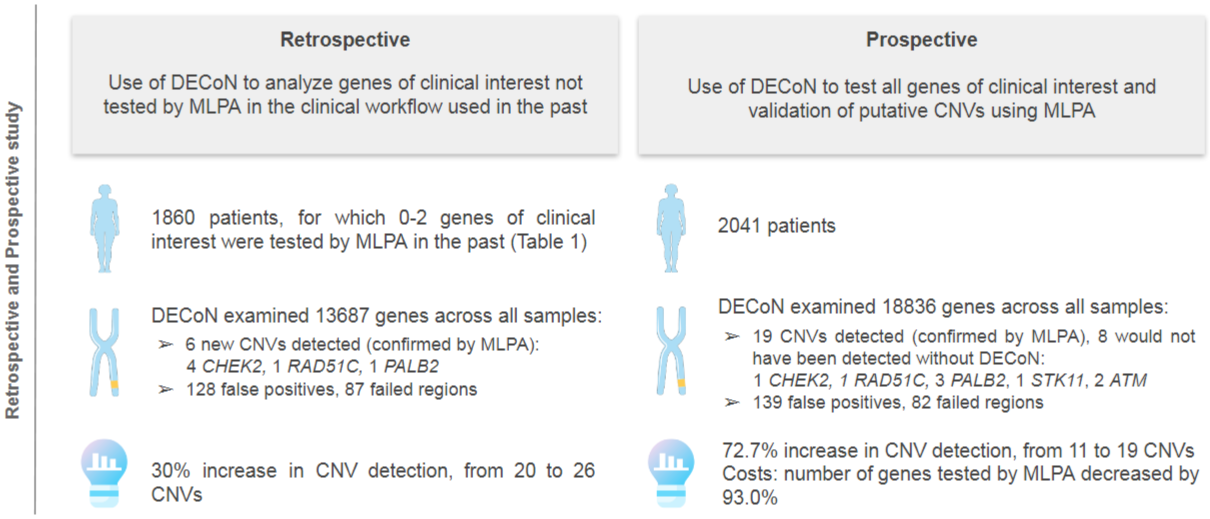

MLPA-based restricted strategy
MLPA test
for a subset of genes of interest
NGS CNV detection strategy
NGS data
DECoN CNV screening
for all genes of interest
MLPA test
of CNV calls with BF >= 2 and failed regions

Figure 14. Previous MLPA-restricted strategy and new NGS CNV detection strategy. Our NGS CNV detection strategy consisted of two steps: first, screening of all genes of clinical interest using DECoN with optimized parameters (Article 2, Supplementary File- Appendix B), and second, validation of putative CNVs (those with Bayesian factor ≥2) and failed regions by MLPA.



Retrospective

Use of DECoN to analyze genes of clinical interest not tested by MLPA in the clinical workflow used in the past

1860 patients, for which 0-2 genes of clinical interest were tested by MLPA in the past (Table 1)

DECoN examined 13687 genes across all samples:

- 6 new CNVs detected (confirmed by MLPA): 4 *CHEK2*, 1 *RAD51C*, 1 *PALB2*
- 128 false positives, 87 failed regions

30% increase in CNV detection, from 20 to 26 CNVs

Prospective

Use of DECoN to test all genes of clinical interest and validation of putative CNVs using MLPA

2041 patients

DECoN examined 18836 genes across all samples:

- 19 CNVs detected (confirmed by MLPA), 8 would not have been detected without DECoN: 1 *CHEK2*, 1 *RAD51C*, 3 *PALB2*, 1 *STK11*, 2 *ATM*
- 139 false positives, 82 failed regions

72.7% increase in CNV detection, from 11 to 19 CNVs
Costs: number of genes tested by MLPA decreased by 93.0%

Figure 15. Retrospective and prospective studies. We used our NGS CNV detection strategy in a retrospective cohort of 1860 patients where a limited number of genes were previously analyzed by MLPA, and in a prospective cohort of 2041 patients, without MLPA analysis.

2.2 Retrospective and prospective studies

We evaluated the use of DECoN with optimized parameters as a screening step before MLPA validation in our genetic diagnostics routine (Article 2, Methods). The evaluation included a retrospective cohort (1,860 patients), in which only a subset of genes of clinical interest had already been MLPA-tested, and a prospective cohort (2,041 patients) without previous MLPA analysis.

Table 10. List of genes tested for each clinical suspicion by our previous MLPA-restricted approach compared with those studied in the current NGS approach.

Clinical suspicion	Tested genes	
	MLPA-restricted strategy: without DECoN	New NGS CNV detection strategy: with DECoN
Hereditary breast cancer	*BRCA1, BRCA2*	*ATM, BRCA1, BRCA2, BRIP1, CHEK2, MLH1, MSH2, MSH6, PALB2, RAD51C, RAD51D*
Hereditary ovarian cancer	*BRCA1, BRCA2*	*BRCA1, BRCA2, BRIP1, MLH1, MSH2, MSH6, RAD51C, RAD51D*
Hereditary breast and ovarian cancer	*BRCA1, BRCA2*	*ATM, BRCA1, BRCA2, BRIP1, CHEK2, MLH1, MSH2, MSH6, PALB2, RAD51C, RAD51D*
Hereditary prostate cancer	*BRCA1, BRCA2*	*ATM, BRCA1, BRCA2, MLH1, MSH2, MSH6*
Early onset colorectal cancer	None	*BRCA1, BRCA2, EPCAM, MLH1, MSH2, MSH6, MUTYH, TP53*
Hereditary non polyposis colon cancer	One MMR gene depending on the MMR IHC pattern	*BRCA1, BRCA2, MUTYH, EPCAM, MLH1, MSH2, MSH6*
Familial malignant melanoma	None	*BAP1, BRCA1, BRCA2, CDK4, CDKN2A, MLH1, MSH2, MSH6, POT1*
Hereditary gastric cancer	None	*CDH1, BRCA1, BRCA2, CDH1, MLH1, MSH2, MSH6*
Hereditary renal cancer syndromes	*FH, FLCN* and/or *VHL*	*BRCA1, BRCA2, FH, FLCN, MET, MLH1, MSH2, MSH6, SDHB, SDHC, SDHD, VHL*
Polyposis	*APC*	*APC, BMPR1A, BRCA1, BRCA2, MLH1, MSH2, MSH6, MUTYH, NTHL1, POLD1, POLE, SMAD4*
Li-Fraumeni syndrome	*TP53*	*TP53*

MMR: Mismatch repair; IHC: immunohistochemistry

First of all, including this bioinformatic screening step in our diagnostics strategy allowed us to test for CNVs in all the genes of clinical interest instead of limiting them to a subset. Our new strategy expanded the number of tested genes from 0-2 to 6-12 genes (Table 10). Most genes had not been tested before because our previous MLPA-restricted strategy (Figure 14) did not include them: turnaround time and costs limited its use. It is worth pointing that using MLPA for testing many genes in a demanding clinical routine, which includes several samples to be tested on time, is almost an unachievable challenge in a universal public health system. In this regard, using an *in-silico* CNV calling tool offers the opportunity of expanding the number of tested genes at a low cost when NGS data is already available.

Our screening strategy resulted in an improvement in terms of diagnostic yield. In the retrospective cohort, six new true CNVs were identified and subsequently confirmed by MLPA (Article 2, Table 1). This represents an important increase in CNV detection, from 20 to 26 CNVs (30%). Additionally, the total number of genes tested across all samples in the retrospective study was 13,687, a much higher value than the 2,660 covered with our previous MLPA-restricted strategy. In the prospective study, DECoN identified 19 CNVs that were then confirmed by MLPA (Article 2, Table 1). Out of those, 8 would not have been identified using our previous MLPA-restricted approach: this represents an increase of 72.7% in CNV detection. Also, 18,836 genes were tested across all samples in the prospective study, instead of the estimated 3,442 that would have been tested with our previous MLPA-restricted strategy. Of course, this yield improvement had a clinical impact: it allowed us to discover the genetic cause of hereditary cancer in previously uninformative families. It is important to note that detecting pathogenic CNVs is key for individualized cancer risk assessment and the implementation of specific surveillance measures and therapeutic strategies for all the carriers.

In this regard, DECoN did not only detected a valuable number of CNVs, but also showed a high sensitivity by detecting the 20 previously known CNVs in the retrospective cohort. In fact, this performance was already expected since DECoN, with optimized parameters, showed very high sensitivity in our previous study (Article 1).

Besides the yield improvement, the use of an NGS CNV calling tool to screen CNVs resulted in a decrease in the resources required for CNV testing. In the prospective study, we observed a decrease of 93% in the number of genes requiring MLPA testing across all samples (Article 2, Table 2). This can be explained by two factors. First, DECoN has a high specificity in our clinical datasets, as validated in our previous study (Article 1). Second, the introduction of a BF cut-off led to an important reduction in the number of genes tested by MLPA. As an example, more than half of the CNV (53.2%) called in our prospective study were discarded because they had a Bayesian factor below 2 (see Bayesian factor discussion below).

Although some authors have argued against the use of NGS-based CNV detection tools in clinical settings for different reasons, (Mason-Suares *et al.*, 2016; Yao *et al.*, 2019) others have argued in favor of its use in this context (Zhao *et al.*, 2013; Fowler *et al.*, 2016; Pugh *et al.*, 2016; Povysil *et al.*, 2017; Ellingford *et al.*, 2018; Chiang *et al.*, 2019). This work, together with our previous study (Article 1), shows that bioinformatic NGS CNV screening is viable in a genetic diagnostics setting and contributes to decrease costs and turnaround times. In addition, NGS CNV screening allowed for an increase in the number of genes tested in our clinical setting and, consequently, of the diagnostic yield.

2.3 Bayesian factor discussion

DECoN, whose code is based on ExomeDepth, provides a Bayesian factor (BF) value to quantify the statistical support for each CNV call. Here, we briefly analyze the BF cut-off included in our NGS CNV detection strategy and how the BF value behaved in our previous study (Article 1).

To discard DECoN CNV calls with low statistical support in our NGS CNV detection strategy, only those with a BF value ≥2 were validated by MLPA. We chose 2 as a reasonable cut-off after the analysis of the distribution of the BF values from our previous benchmark (Figure 16). We observed that all the true-positive CNV calls had a BF value >2, except for a case of mosaicism. Also, the fifth percentile and first quartile for the true-positive CNV calls were placed at 5.02 and 13.67, respectively (median: 29.10). With this data, we expected that the BF values for the true-positive CNV calls would not be close to the BF cut-off. Using this threshold in our NGS CNV detection strategy allowed us to discard a large proportion of CNV calls that were not expected to be true positives. 53.2% of the CNV calls in the prospective study (Figure 16) obtained a BF <2 and, consequently, were discarded. The plot shows that most BF values of the CNV calls from the prospective study fell within the 1-3 range and there was a peak in the 1-2 range.

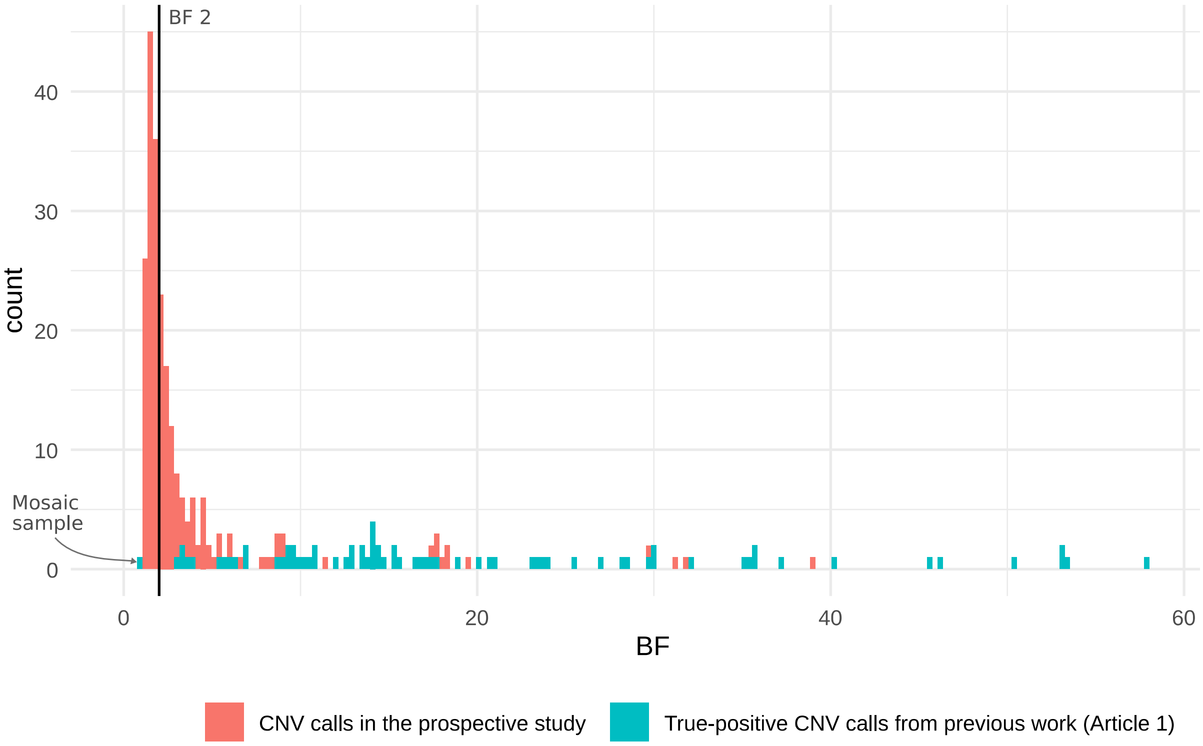


Figure 16. Histogram of BF values for the CNV calls from the prospective study and the true-positive CNV calls from the previous study (Article 1). Plot values are restricted to the 0-60 range. 53.19% of the CNV calls in the prospective study had a BF value lower than 2. In the previous study (Article 1), only one CNV call had a BF value < 2: a mosaic sample (BF 1.04).

Other factors affecting the BF value were also analyzed. Regarding the Illumina platform in which samples were sequenced, some differences were observed. Supplementary Figure 3 (Article 2) shows that the median for the HiSeq sequencing BF values, and also the lower and upper quartiles, are greater than the MiSeq ones. This suggests that, in our laboratory routine, calling a CNV in a sample sequenced in the HiSeq platform is less challenging than calling it in the MiSeq platform. A likely explanation is that all samples in each HiSeq run were analyzed together (~96 samples) whereas each MiSeq run (~16 samples) was analyzed along with 51 samples from other runs (Article 2, Supplementary File- Appendix B). Hence, each HiSeq run contained a high number of samples from a single run, so no multi-run-batch effect is expected and DECoN can find samples that correlate with each sample to analyze more easily. Additionally, the higher coverage obtained in the HiSeq samples might have positively influenced the BF values.

BF value differences can be observed in Supplementary Figure 4 (Article 2): values are lower for the true-positive duplications than the true-positive deletions. This suggests that calling a duplication CNV is more challenging than calling a deletion CNV. Differences can be observed in both MiSeq and HiSeq sequencing settings.

Finally, another interesting question is to know whether calling CNVs is more challenging in certain genes than in others, such that BF values vary depending on the gene. Supplementary Figure 2 (Article 2) shows BF distribution across genes. Although BF value differences can be easily observed, most of them have a low number of true-positive calls. In this case, the results are not conclusive.

3 R/Bioconductor package to identify false positives of NGS CNV calling tools

3.1 Background and motivation

As noted in previous sections, germline CNVs are relevant contributors to hereditary diseases. Recent NGS improvements have resulted in the release of several germline CNV calling tools for WGS, WES, and panel data (Zhao *et al.*, 2013; Mason-Suares *et al.*, 2016; Roca *et al.*, 2019). However, available benchmarks show that all germline CNV calling tools produce false positives (Kim *et al.*, 2017; Zhang *et al.*, 2019; Moreno-Cabrera *et al.*, 2020), frequently achieving high false discovery rates (FDRs). Of course, these false-positive calls negatively impact downstream analyses regardless of the context in which they were produced.

In genetic diagnostics, NGS CNV callers should be used as a first-tier screening step prior to an orthogonal validation method, such as MLPA (Article 2). Therefore, genetic diagnostics laboratories incur a significant unnecessary cost and effort with each false positive. A tool able to identify these false-positive calls could help to mitigate the problem.

NGS CNV callers perform their detection by using a paired-end, split-read, depth of coverage, assembly, or combined approach (Introduction, section 7). However, NGS CNV callers rarely use SNVs information to call CNVs (Introduction, Tables 5-9), even though SNVs are typically available in NGS pipelines. The allele frequency of the SNVs can provide valuable evidence to discard false-positive CNV calls.

The aim of this part of the thesis was to develop a tool to identify false-positive calls in the results of any germline CNV caller using SNV allele frequency information. To this end, we developed CNVfilteR, an open-source R/Bioconductor package whose code and documentation are publicly available at https://bioconductor.org/packages/release/bioc/html/CNVfilteR.html. Method details are available in the original manuscript and Supplementary File 1 (Article 3).

3.2 CNVfilteR performance and features

The performance was evaluated on callsets generated by a total of 13 CNV calling tools on 3 reference WGS and 541 gene panel samples. On the 3 WGS samples, CNVfilteR produced a consistent false-positive decrease (between 15.3% and 44.8%) at the expense of a slight decrease in sensitivity. Interestingly, CNVfilteR did not only identify a number of false-positive calls but also increased the F1-score metric in 19 out of the 24 tool-sample evaluations (up to 20.7% increase). The F1-score, defined as the harmonic mean of precision and sensitivity, is a common system to evaluate predictors. These F1-score improvements (Article 3, Figure 1d-f) evidence that CNVfilteR can improve the overall performance of CNV callers, although the effect is more notable in some tools than in others.

On the other hand, CNVfilteR also improved DECoN performance on the HiSeq-panel and MiSeq-panel datasets. In particular, CNVfilteR identified 15% of the false-positive calls (3 out of 20 false positives) and 12.5% of the false-positive calls (2 out of 16), respectively. In both datasets, no true CNV was misidentified as false positive, so sensitivity did not change. CNVfilteR performed well in these high-coverage gene panel samples, although the low number of the already known false positives prevents a more comprehensive evaluation.

Regarding CNVfilteR runtime, the tool proved a fast identification of false-positive calls on both WGS and gene panel samples. The median runtime per sample was 0.85 seconds for the gene panel samples and 3.53 minutes for the HuRef WGS sample. Hence, CNVfilteR low runtime values make it suitable for using in demanding pipelines where several samples, often of large size, have to be analyzed.

From the features that CNVfilteR offers, two should be highlighted. First, CNVfilteR provides a helper function to automatically interpret VCFs produced by some of the most used SNV callers, such as GATK or Strelka. VCF is a very flexible format and its interpretation is often challenging. The helper function provided by CNVfilteR pretends to save developers time and avoid errors when interpreting VCF fields. A second feature to highlight is the visual output produced by CNVfilteR: results can be easily plotted through customizable functions (Figure 17). Visual representation of data helps users better understand the results. CNVfilteR graphical output, which is based on karyoploteR (Gel and Serra, 2017) and CopyNumberPlots packages, help users understand which SNVs overlap a certain CNV and why that CNV call was identified as a false positive or not.

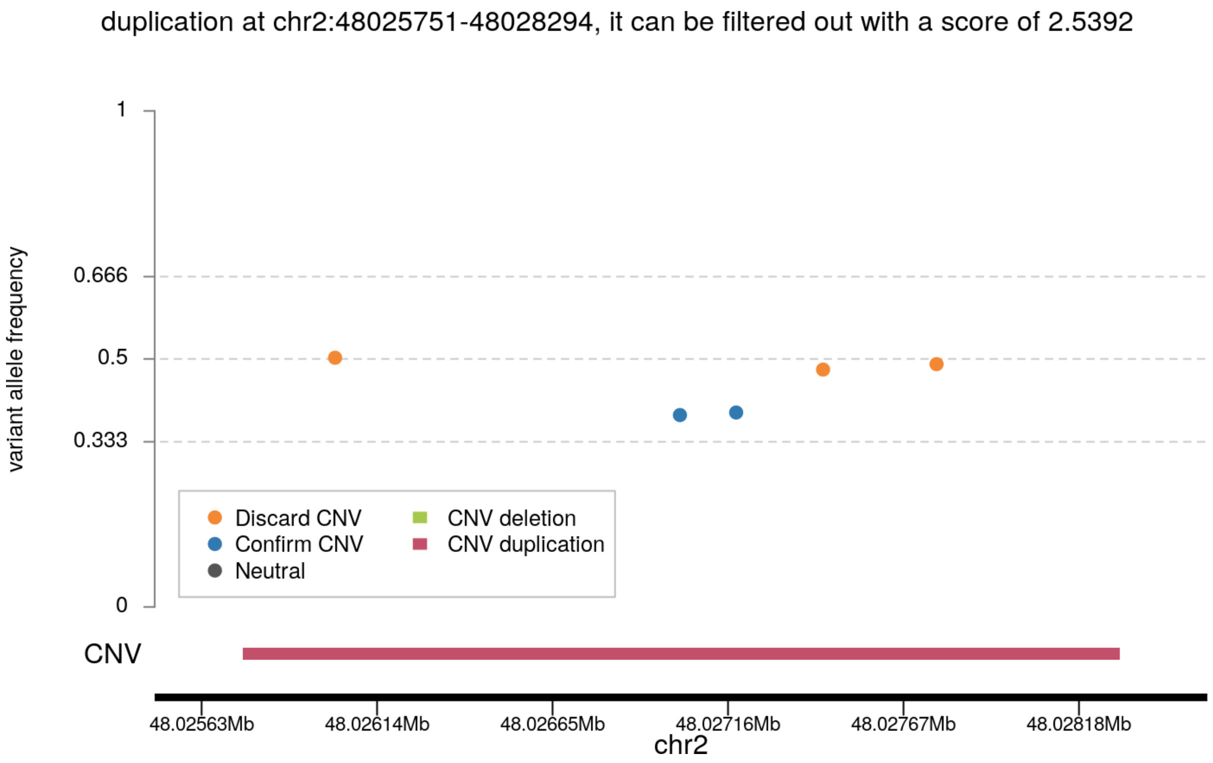


Figure 17: CNVfilteR output example for a CNV duplication. Five heterozygous SNVs overlapped the CNV duplication: three of them with an allele frequency very close to 0.5 and two near– but not close – to 0.33. In this second example, adding up the total scores provided by the CNVfilteR scoring model, the final score was higher than the duplication threshold score (0.5), so the CNV was identified as a false positive and could be discarded.

CNVfilteR performance was proven on several of the most frequently used CNV calling tools and on very different datasets, from high-coverage gene-panels samples to a few WGS samples. CNVfilteR decreases the number of false-positive CNV calls at a low computational cost. The aim was not to replace the best performing CNV calling tools, but to complement them and help users improve their CNV calling pipelines.

3.3 Limitations

The strengths of CNVfilteR have been shown in the previous section. However, the tool also suffers from some limitations that should be discussed.

First, CNVfilteR focuses on germline alterations in diploid organisms: SNV allele frequencies are expected to be close to certain theoretical values (Article 3, FP detection strategy). If a CNV duplication exists, all the overlapping SNVs are expected to have an allele frequency close to 33% or 66%. Also, if a germline CNV deletion does not exist, all the overlapping SNVs are expected to be homozygous. These assumptions are not true when working on somatic data.

In this context, both CNVs and SNVs have allele frequencies that vary in a very different manner. Therefore, CNVfilteR cannot be used on somatic data.

Another limitation to note is that CNVfilteR performance depends on SNV call reliability. In other words, any aspect affecting SNV quality should be removed or minimized in order to provide CNVfilteR reliable information to work with. CNVfilteR provides methods to manage some of these aspects. For example, the *min.total.depth* parameter allows CNVfilteR to discard SNV calls with low coverage support. The default parameter value is 10, but it should be adapted to the experiment conditions. A threshold of 10 may be appropriate for many WGS samples, but high-coverage samples may require a higher limit in order to discard the noisy false-positive SNV calls. On the other hand, the *regions.to.exclude* parameter defines the regions where the variants should be excluded. This is a key parameter since SNV callers perform worse on low complexity and repetitive genomic regions. In any case, CNVfilteR cannot ensure all aspects related to the veracity of a variant. CNVfilteR users are responsible for providing a high-quality VCF file. Many SNV callers already provide methods and parameters to account for SNVs quality and to minimize the number of false-positive SNV calls.

Lastly, CNVfilteR is not taking advantage of a potential source of information: the small INDELs. Variant callers frequently detect both SNVs and INDELs, so they are usually available in NGS pipelines. CNVfilteR provides an optional parameter to include INDELs as it does with SNVs. However, CNVfilteR does not consider INDELs by default. The reason is that calling an INDEL is more challenging than calling an SNV, so their allele frequency is usually noisier. However, we did not study how INDEL frequency actually behaved and whether very small INDELs could have allele frequencies similar to those of SNVs. Although using INDEL information might be more problematic to confirm or discard a CNV duplication, it could be still very useful to provide evidence in favor or against CNV deletions, for which the key is to determine whether a certain variant overlapping a CNV is heterozygous or homozygous. In short, although CNVfilteR allows the user to use INDEL information, we have not studied its potential and limitations in depth.

4 Web-based tool to manage the diagnostics process in the laboratory routine

4.1 Background and motivation

The arrival of NGS has multiplied the sequencing capacity in a very cost-effective way (Goodwin *et al.*, 2016; Pereira *et al.*, 2020), a fact that has impacted the way genetic testing of diseases is performed (Knoppers *et al.*, 2015; Xue and Wilcox, 2016; Yohe and Thyagarajan, 2017). Whole-genome sequencing, whole-exome sequencing, and targeted gene-panels are currently being used in diagnostics settings, but the latter, targeted gene-panels, is the technology used by most of genetic testing laboratories (Teekakirikul *et al.*, 2013; Kurian *et al.*, 2014; Feliubadaló *et al.*, 2017).

The ICO-IGTP Joint Program on Hereditary Cancer concentrates on the detection and interpretation of germline variants that increase the risk of developing cancer. Among other hereditary cancer syndromes, the diagnostic activity focuses on hereditary colorectal cancer, hereditary breast and ovarian cancer, neurofibromatosis and related disorders. The ICO-IGTP program developed a custom gene panel with 122-135 genes (depending on the version), called I2HCP (Castellanos *et al.*, 2017; Feliubadaló *et al.*, 2017), to improve its diagnostics activity.

The diagnostics workflow typically comprises DNA extraction, library preparation, NGS sequencing, analysis pipeline, variant analysis, variant classification and, finally, reporting to the health care professional who requested the test. To manage the whole process during the diagnostics routine, we developed Pandora (PlaAtform for NGS Data Organization Repository and Analysis), a web-based tool built on the Django framework and a PostgreSQL database. As a result of our experience designing and building Pandora, we presented a set of quick tips to consider when developing a tool to support genetic testing (Article 4).

4.2 Features of Pandora

Pandora works as an orchestrator of the parts involved in the laboratory diagnostics workflow. One of the main features of Pandora is the automation of multiple tasks throughout the diagnostics process, such as sequencing data download, sample sheet generation for the Illumina sequencing platforms, and execution of the analysis pipeline, among others (Article 4, Figure 1). The automation of all these steps speeds up the diagnostics workflow while avoiding the errors that typically occur when manual tasks are performed.

Another important feature of Pandora is the flexible data exploitation to ease both diagnostics and research tasks. This feature relies on two necessary technologies. First, the PostgreSQL relational database allows for easy representation of complex data relationships and efficient access to tables of up to millions of rows. Second, we use an efficient and fully-featured data table user-interface library (Webix) which supports large datasets with several built-in features. Also, Pandora supports the process of validation and classification of a variant, a key step in the diagnostics routine. This process often requires the labor of two independent curators that can work in parallel on the sample or group the samples; Pandora ensures traceability by registering who and when did what in the system.

As observed in Figure 11, Pandora interoperates with multiple parts and systems. Hence, the main limitation of a custom and architecturally complex software tool is the maintenance it requires. For example, if a new sequencing platform is used, the code of Pandora has to be properly updated. Also, changes in the diagnostics workflow, new panel designs, or external software changes may require the update of Pandora.

4.3 Eleven tips for building a tool to support the diagnostics routine

The design of Pandora has been tailored to the specific needs of the ICO-IGTP diagnostics activity. However, many design aspects can be extrapolated to the development of any software tool for genetic diagnostics. As a product of our experience building Pandora, we wrote a set of tips or recommendations to consider when facing the development of a tool for genetic diagnostics. The development of a tool in this context requires an interdisciplinary approach: the set of recommendations pretended to be useful for computer scientists, biologists, bioinformaticians, clinicians, and other health care professionals involved in that process.

Some of the tips can be considered a must in genetic diagnostics: traceability, reproducibility, data security, and custom analysis pipeline. First, it is necessary to know who and when did what in the system (Tip 1). Since a report will be delivered to a patient, any future result re-analysis needs to retrace all the steps that led to the result. Second, the computational environment in which a sample was analyzed has to be unambiguously identified to ensure reproducibility (Tip 1) (Roy *et al.*, 2018). Third, human genome data is sensitive information, so security requirements should be specially considered in this context (Tip 11). Finally, diagnostics-oriented NGS data analysis pipelines have specific characteristics (Tip 9) because any found or missed alteration will be reported back (or not) to patients, affecting their clinical management.

Other tips referred to key aspects of the operation of Pandora: automation of tasks (Tip 2), use of a relational database (Tip 4), real-time synchronization (Tip 8), and interoperability (Tip 10). Specially, Pandora could not work without considering the first two: the automation of tasks enhances productivity while avoiding human errors and a relational database is a requisite to efficiently exploit complex and large amounts of data.

The remaining tips were more specifically related to software engineering aspects; if overlooked, they can also undermine the success of the tool. Listening to the user (Tip 3) is the only way to meet the correct specifications of a tool and avoid some bad design decisions. This cross-cutting tip is also related to other aspects that limit or expand what the user can do with the tool: usability (Tip 5), the data table UI library (Tip 6), and the optimization of screen space (Tip 7).

5 Translational bioinformatics in the healthcare system

Translational bioinformatics is an emerging field in the age of precision medicine. Citing Dr Russ Altman's definition (Tenenbaum, 2016), translational bioinformatics consists of "informatics methods that link biological entities (genes, proteins, and small molecules) to clinical entities (diseases, symptoms, and drugs) or *vice versa*". Multiple actions are performed in this area. To illustrate some, translational bioinformatics deals with pathogenicity estimation of human genetic variants, develops medical tools to be used by clinicians and scientists, or performs data mining from large biomedical databases.

This PhD thesis addresses multiple bioinformatic approaches that aim to improve the genetic diagnostics of hereditary cancer. Hence, all of them have a translational dimension: the focus is on its applicability to improve the healthcare system. In fact, some research results derived from this PhD thesis already have a real impact on the genetic diagnostics activity: they have been implemented in the molecular diagnostics laboratory of the Hereditary Cancer Program at the Institut Català d'Oncologia (ICO). Among others, the results have allowed an improvement in cancer risk assessment or to guide treatment in real patients.

The first study presented here consisted of a benchmark of NGS CNV detection tools for genetic diagnostics (Article 1). It was proven that some NGS CNV calling tools can achieve very high sensitivity and specificity on gene-panel datasets. The conclusions obtained led to the implementation of one of the tools in our diagnostics routine for hereditary cancer (Article 2). However, results are applicable beyond this. Both the performance demonstrated in the benchmark and the code provided (CNVbenchmarkeR) can be used by other laboratories for the genetic diagnostics of several hereditary diseases. For example, targeted gene panels are also used for the genetic diagnostics of cardiomyopathies (Teekakirikul *et al.*, 2013), hereditary retinal dystrophies (Solebo *et al.*, 2017), or deafness (Ji *et al.*, 2014), among others.

The second study presented in this thesis consisted of the implementation and evaluation of an NGS CNV calling tool as a screening step before MLPA validation. The performance was evaluated in a real diagnostics setting: our diagnostics routine at the ICO. We observed two valuable benefits in this translational work. On the one hand, the number of genes to be tested by MLPA decreased, which drove costs and turnaround times down. Of course, available resources are an important issue in any healthcare system. They are not unlimited, a fact that has become especially evident during the current Covid-19 pandemic, which has produced enormous stress on, among others, the Spanish public healthcare system. On the other hand, including an NGS CNV calling tool for CNV screening allowed us to expand the total number of genes covered, so the final diagnostic yield was increased. New CNVs, up to 72.7% in the prospective study, were detected by DECoN in the screening step, validated via MLPA, and reported back to the clinicians and patients. This is an illustrative example of how the improvement of informatics or bioinformatic resources towards better diagnostics of hereditary cancer positively impacts the clinical management of the patient and family members. Detecting a pathogenic variant in a clinically actionable gene allows for multiple medical or healthcare improvements. All carriers can benefit from a proper cancer risk assessment, the establishment of specific surveillance measures, and the implementation of therapeutic strategies, such as prophylactic surgeries or genotype-based chemotherapeutic treatments (Pennington *et al.*, 2014; Musella *et al.*, 2015). In addition, non-carriers of variants in high-penetrance cancer predisposition variants can decrease surveillance measures according to their risk.

The third study presented here resulted in an R package, CNVfilteR, to identify false positives produced by NGS CNV calling tools. The tool can be useful for both diagnostics and research purposes. In a genetic diagnostics setting, any false-positive NGS CNV call generates an extra effort to discard it using an orthogonal method, such as MLPA. CNVfilteR can be used to identify these false-positive CNV calls and avoid additional validation tests. The tool has been already implemented as part of our diagnostics routine for hereditary cancer at the ICO laboratory to reduce the number of MLA tests. Since the package and its documentation are publicly available at the Bioconductor site, other genetic diagnostics laboratories can benefit from its use.

Finally, the fourth subproject consisted of the development of a web-based tool to manage the ICO-IGTP routine diagnostics workflow for hereditary cancer. Pandora is fully-operative in the diagnostics routine: to date, Pandora has managed the genetic testing of more than five thousand patients. Pandora has improved the diagnostics process in some aspects that were previously discussed, such as task automation, support to common tasks, and flexible data exploitation. As a result of our experience building Pandora, we also elaborated a set of recommendations to help the professionals involved in the design of a similar tool for genetic diagnostics.

Conclusions

As a result of all the work presented in this PhD thesis, we conclude that:

- The CNV tools benchmark demonstrated that DECoN and panelcn.MOPS provided the highest performance for germline CNV detection. Although panelcn.MOPS showed a slightly higher sensitivity in one of the datasets, DECoN showed a much higher specificity in a diagnostics scenario.
- The performance of the tested CNV tools depended on the dataset. Therefore, it is necessary a previous in-house validation of any CNV detection tool before using it in a clinical setting.
- The CNV bioinformatic screening improved our genetic diagnostics strategy. It contributed to decrease turnaround times and costs while allowing for an increase of tested genes and, consequently, of the diagnostic yield.
- We developed CNVfilteR, an R/Bioconductor package to identify false-positive calls generated by CNV calling tools from germline NGS data. CNVfilteR identified false-positive calls in all tested tools and datasets, from gene panel to WGS. In addition, the F1-score, a common metric to measure binary classifier performance, was improved for most tool-sample combinations. CNVfilteR can be plugged in existing CNV calling pipelines to improve calling performance at virtually no computational cost.
- We developed Pandora, a web-based tool to manage the diagnostics process in the laboratory routine. Multiple data procedures were automated while supporting common tasks and easing data exploitation. A software tool to manage genetic diagnostics should be designed considering key aspects including traceability, reproducibility, task automation, usability, interoperability, and data security. We described a set of recommendations including these, and a few more, to help genetic testing laboratories implement such systems.

References

Abdellah,Z. *et al.* (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Abel,H.J. and Duncavage,E.J. (2013) Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. *Cancer Genet.*, **206**, 432–440.

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Alberts,B. *et al.* (2002) Molecular Biology of the Cell- NCBI Bookshelf. *Garl. Sci.*

Atashzar,H. *et al.* (2011) A survey on web application vulnerabilities and countermeasures. In, *2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*. Seogwipo, Korea (South), pp. 647–652.

Bunnik,E.M. and Le Roch,K.G. (2013) An Introduction to Functional Genomics and Systems Biology. *Adv. Wound Care*, **2**, 490–498.

Castellanos,E. *et al.* (2017) A comprehensive custom panel design for routine hereditary cancer testing: Preserving control, improving diagnostics and revealing a complex variation landscape. *Sci. Rep.*, **7**, 39348.

Chen,S. *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Chen,X. and Sullivan,P.F. (2003) Single nucleotide polymorphism genotyping: Biochemistry, protocol, cost and throughput. *Pharmacogenomics J.*, **3**, 77–96.

Chiang,T. *et al.* (2019) Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. *Genet. Med.*, **0**, 1–10.

Choi,M. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 19096–19101.

Eichler,E.E. (2019) Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N. Engl. J. Med.*, **381**, 64–74.

Van El,C.G. *et al.* (2013) Whole-genome sequencing in health care. *Eur. J. Hum. Genet.*, **21**, 580–584.

Ellingford,J.M. *et al.* (2018) Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases. *J. Med. Genet.*, **55**, 114–121.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Ezkurdia,I. *et al.* (2014) Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.*, **23**, 5866–5878.

Feliubadaló,L. *et al.* (2017) Benchmarking of whole exome sequencing and Ad Hoc designed panels for genetic testing of hereditary cancer. *Sci. Rep.*, **7**, 1–11.

Feliubadaló,L. *et al.* (2019) Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int. J. Cancer*, **145**, 2682–2691.

Fokkema,I.F.A.C. *et al.* (2005) LOVD: Easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. *Hum. Mutat.*, **26**, 63–68.

Fowler,A. *et al.* (2016) Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome open Res.*, **1**, 1–20.

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

Gel,B. and Serra,E. (2017) KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, **33**, 3088–3090.

Gibbs,R.A. *et al.* (2015) xn. *Nature*, **526**, 68–74.

Gonorazky,H.D. *et al.* (2019) Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.*, **104**, 466–483.

Goodwin,S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Green,R.C. *et al.* (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.*, **15**, 565–574.

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7 Suppl 1**, S4.

Heyer,E.E. *et al.* (2019) Diagnosis of fusion genes using targeted RNA sequencing. *Nat. Commun.*, **10**, 1–12.

Hong,C.S. *et al.* (2016) Assessing the reproducibility of exome copy number variations predictions. *Genome Med.*, **8**, 82.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Ji,H. *et al.* (2014) Combined examination of sequence and copy number variations in human deafness genes improves diagnosis for cases of genetic deafness. *BMC Ear, Nose Throat Disord.*, **14**, 9.

Johansson,L.F. *et al.* (2016) CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum. Mutat.*, **37**, 457–464.

Karczewski,K.J. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

Katsonis,P. *et al.* (2014) Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Sci.*, **23**, 1650–1666.

Kerkhof,J. *et al.* (2017) Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. *J. Mol. Diagnostics*, **19**, 905–920.

Kim,D. *et al.* (2015) HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kim,H.-Y. *et al.* (2017) Gene-based comparative analysis of tools for estimating copy number alterations using whole-exome sequencing data. *Oncotarget*, **8**, 27277–27285.

Kim,S. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.

Knoppers,B.M. *et al.* (2015) Return of genetic testing results in the era of whole-genome sequencing. *Nat. Rev. Genet.*, **16**, 553–559.

Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Korbel,J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (80-.).*, **318**, 420–426.

Kress,W. *et al.* (2017) The Genetic Approach: Next-Generation Sequencing-Based Diagnosis of Congenital and Infantile Myopathies/Muscle Dystrophies. *Neuropediatrics*, **48**, 242–246.

Kulkarni,S. and Roy,S. (2015) CLINICAL GENOMICS.

Kurian,A.W. *et al.* (2014) Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *J. Clin. Oncol.*, **32**, 2001–2009.

Laduca,H. *et al.* (2014) Utilization of multigene panels in hereditary cancer predisposition testing: Analysis of more than 2,000 patients. *Genet. Med.*, **16**, 830–837.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Landrith,T. *et al.* (2020) Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. *npj Precis. Oncol.*, **4**, 1–9.

Landrum,M.J. *et al.* (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.

Langer-Safer,P.R. *et al.* (1982) Immunological methods for mapping genes on Drosophila polytene chromosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **79**, 4381–4385.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Layer,R.M. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, **1303.3997v**.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,L. *et al.* (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**.

Logsdon,G.A. *et al.* (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, **21**, 597–614.

MacDonald,J.R. *et al.* (2014) The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.

Mamanova,L. *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.

Mason-Suares,H. *et al.* (2016) Detecting Copy Number Variation via Next Generation Technology. *Curr. Genet. Med. Rep.*, **4**, 74–85.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.

McKenna,A. *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

McPherson,E. (2006) Genetic diagnosis and testing in clinical practice. *Clin. Med. Res.*, **4**, 123–129.

Mohammed Yakubu,A. and Chen,Y.P.P. (2020) Ensuring privacy and security of genomic data and functionalities. *Brief. Bioinform.*, **21**, 511–526.

Moorthie,S. *et al.* (2013) Informatics and clinical genome sequencing: Opening the black box. *Genet. Med.*, **15**, 165–171.

Moreno-Cabrera,J.M. *et al.* (2020) Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur. J. Hum. Genet.*

Mouli,V.R. and Jevitha,K.P. (2016) Web Services Attacks and Security- A Systematic Literature Review. In, *Procedia Computer Science*. Elsevier B.V., pp. 870–877.

Musella,A. *et al.* (2015) PARP inhibition: A promising therapeutic target in ovarian cancer. *Cell. Mol. Biol. Cell. Mol. Biol*, **61**, 44–61.

Nagy,R. *et al.* (2004) Highly penetrant hereditary cancer syndromes. *Oncogene*, **23**, 6445–6470.

Naveed,M. *et al.* (2015) Privacy in the genomic era. *ACM Comput. Surv.*, **48**, 1–44.

Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

OMIM reference. Entry Statistics [https://www.omim.org/statistics/geneMap] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), (Updated March 23rd, 2021)

Parikh,H. *et al.* (2016) svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, **17**, 64.

Park,E. *et al.* (2018) The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.*, **102**, 11–26.

Pennington,K.P. *et al.* (2014) Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas. *Clin. Cancer Res.*, **20**, 764–775.

Pereira,R. *et al.* (2020) Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *J. Clin. Med.*, **9**, 132.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

Pinkel,D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pirooznia,M. *et al.* (2015) Whole-genome CNV analysis: advances in computational approaches. *Front. Genet.*, **06**, 138.

Plon,S.E. *et al.* (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.*, **29**, 1282–1291.

Povysil,G. *et al.* (2017) panelcn.MOPS: Copy number detection in targeted NGS panel data for clinical diagnostics. *Hum. Mutat.*, **38**, 889–897.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Pugh,T.J. *et al.* (2016) VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet. Med.*, **18**, 712–9.

Rahman,N. (2014) Realizing the promise of cancer predisposition genes. *Nature*, **505**, 302–308.

Ramensky,V. *et al.* (2002) Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Richards,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.

Roca,I. *et al.* (2019) Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat. Res. - Rev. Mutat. Res.*, **779**, 114–125.

Roy,S. *et al.* (2018) Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagnostics*, **20**, 4–27.

Sadedin,S.P. *et al.* (2018) Ximmer: A system for improving accuracy and consistency of CNV calling from exome data. *Gigascience*, **7**, 1–11.

Sahraeian,S.M.E. *et al.* (2017) Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.*, **8**, 59.

Salzberg,S.L. (2018) Open questions: How many genes do we have? *BMC Biol.*, **16**, 94.

Sánchez,M.Á. (2019) Cáncer Hereditario (Sociedad Española de Oncología Médica).

Sanger,F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–5467.

Schouten,J.P. *et al.* (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, **30**, e57.

Seo,J.S. *et al.* (2016) De novo assembly and phasing of a Korean human genome. *Nature*, **538**, 243–247.

Sherry,S.T. *et al.* (2001) DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Smith,R.M. *et al.* (2013) Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, **14**, 1–15.

Smith,S.D. *et al.* (2017) Lightning-fast genome variant detection with GROM. *Gigascience*, **6**, 1–7.

Solebo,A.L. *et al.* (2017) Epidemiology of blindness in children. *Arch. Dis. Child.*, **102**, 853–857.

Stein,L. (2001) Genome annotation: From sequence to biology. *Nat. Rev. Genet.*, **2**, 493–503.

Stenson,P.D. *et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.

Sun,Y. *et al.* (2015) Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Hum. Mutat.*, **36**, 648–655.

Taeubner,J. *et al.* (2018) Penetrance and Expressivity in Inherited Cancer Predisposing Syndromes. *Trends in Cancer*, **4**, 718–728.

Talevich,E. *et al.* (2016) CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.*, **12**, 1–18.

Teekakirikul,P. *et al.* (2013) Inherited cardiomyopathies: Molecular genetics and clinical genetic testing in the postgenomic era. *J. Mol. Diagnostics*, **15**, 158–170.

Tenenbaum,J.D. (2016) Translational Bioinformatics: Past, Present, and Future. *Genomics, Proteomics Bioinforma.*, **14**, 31–41.

Teo,S.M. *et al.* (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, **28**, 2711–2718.

Tewhey,R. *et al.* (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, **10**, R116.

Thorvaldsdottir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.

Trost,B. *et al.* (2018) A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am. J. Hum. Genet.*, **102**, 142–155.

Truty,R. *et al.* (2019) Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet. Med.*, **21**, 114–123.

Turro,E. *et al.* (2020) Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, **583**, 96–102.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science (80-.).*, **291**, 1304–1351.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science (80-.).*, **340**, 1546–1558.

Wang,Q. *et al.* (2014) An eye-tracking study of website complexity from cognitive load perspective. *Decis. Support Syst.*, **62**, 1–10.

Watson,J.D. and Crick,F.H. (1953) The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.*,

18, 123–131.

Wetterstrand, K. A. (2019) DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). [http://www.genome.gov/sequencingcosts] National Human Genome Research Institute (updated 2019)

Xue,Y. *et al.* (2015) Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: Single-gene, gene panel, or exome/genome sequencing. *Genet. Med.*, **17**, 444–451.

Xue,Y. and Wilcox,W.R. (2016) Changing paradigm of cancer therapy: precision medicine by next-generation sequencing. *Cancer Biol. Med.*, **13**, 12–18.

Yan,P. and Guo,J. (2010) The research of web usability design. In, *2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010.*, pp. 480–483.

Yao,R. *et al.* (2019) Evaluation of copy number variant detection from panel-based next-generation sequencing data. *Mol. Genet. Genomic Med.*, **7**, 1–8.

Yohe,S. and Thyagarajan,B. (2017) Review of clinical next-generation sequencing. *Arch. Pathol. Lab. Med.*

Zarrei,M. *et al.* (2015) A copy number variation map of the human genome. *Nat. Rev. Genet.*, **16**, 172–183.

Zhang,F. *et al.* (2009) Copy Number Variation in Human Health, Disease, and Evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.

Zhang,L. *et al.* (2019) Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput. Biol.*, **15**, 1–12.

Zhao,M. *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14**, S1.

Zhou,B. *et al.* (2018) Extensive and deep sequencing of the venter/ huref genome for developing and benchmarking genome analysis tools. *Sci. Data*, **5**, 1–12.

Additional study: RNA-seq pipeline to improve genetic diagnostics of hereditary cancer

1 Introduction

The use of whole-exome sequencing (WES) and targeted gene panels have improved the cost-effectiveness of the genetic diagnostics of hereditary diseases and increased their overall diagnostic yield (Goodwin *et al.*, 2016; Kress *et al.*, 2017; Feliubadaló *et al.*, 2017). However, genetic diagnostics still suffers from a significant percentage of patients in which the genetic cause of the disease remains unknown. This fact can be explained by multiple causes. Among other reasons, there are many variants of unknown significance, whose impact on gene function is not known. The analysis of mRNA expression by RNA-seq can help determine the potential impact of these variants in the splicing of the mRNA.

RNA-sequencing (RNA-seq) provides gene and transcript expression data that can improve the diagnostic yield provided by DNA-seq (Park *et al.*, 2018; Gonorazky *et al.*, 2019; Landrith *et al.*, 2020). RNA-seq allows for the identification and quantification of known and *de novo* transcripts; aberrant splicing events can be revealed, even those caused by intronic variants missed by DNA sequencing. RNA-seq expression profiles can reveal underexpressed or overexpressed clinically relevant genes due to multiple reasons, like epigenetic mutations, variants in regulatory regions, inversions, or DNA variants that cause aberrant transcripts that may be degraded by nonsense-mediated decay. Similarly, RNA-seq data allows for the detection of imbalances in allele expression when the alterations only affect a single allele (Smith *et al.*, 2013). RNA-seq also allows for the identification of fusion genes, usually caused by translocations (Heyer *et al.*, 2019). Moreover, variant calling on RNA-seq data can be used to confirm that a certain DNA sample is the same as the RNA sample (therefore sample swap errors can be detected). Additionally, in some cases, RNA-seq could be useful to perform variant calling on regions where DNA sequencing obtained a very low coverage level.

The aim of this part of the thesis was to design, implement, and evaluate an RNA-seq analysis pipeline to complement and improve the genetic diagnostics of hereditary cancer performed from DNA-seq. Although the RNA-seq pipeline covered multiple aspects, the main focus was the detection of known and *de novo* aberrant transcripts.

2 Methods

2.1 Datasets

A total of 47 samples (41 unique), distributed into five NGS runs, were analyzed using a custom-developed NGS RNA-seq gene panel (Table 11). All 47 samples contained previously identified mutations that predispose to hereditary cancer. Of those, 31 contained a previously known mutation leading to an aberrant RNA transcript. Informed written consent for both diagnostics and research purposes was obtained from all individuals included in the study.

2.2 Sample preparation and NGS panel

RNA was extracted from cultured peripheral blood lymphocytes. Total RNA was isolated using either the TRIzol reagent or the High Pure RNA Isolation Kit (Roche Diagnostics, Manheim, Germany) after 4-6h of puromycin incubation in order to prevent the potential degradation of unstable transcripts by nonsense-mediated decay; only 6 samples from the P/NP run were not treated with puromycin (Table 11).

Table 11. List of runs analyzed using a custom NGS panel and a custom RNA-seq pipeline.

	NGF3	NGF4	NGF7	NGF8	P/NP
Total samples	4	4	11	16	12 (only 6 were unique) [1]
Samples with a mutation that leads to an aberrant isoform	2	2	9	12	12 (only 6 were unique) [1]
Sequencing platform	MiSeq, 2×300 bp	MiSeq, 2×300 bp reads	MiSeq, 2×300 bp	MiSeq, 2×300 bp	HiSeq, 2×250 bp reads
Capture version	2.0 (155 genes)	2.0 (155 genes)	2.1 (168 genes)	2.1 (168 genes)	2.1 (168 genes)
Use of puromycin	Yes	Yes	Yes	Yes	6 out of the 12 samples [1]
Comments	Standard NimbleGen SeqCap Protocol.	Fragmentation time was modified. Following runs maintained this modification.			

[1] *The same 6 samples were processed both using puromycin and without, resulting in a total of 12 samples.*

Samples were analyzed using a custom NGS panel with 155-168 hereditary cancer-associated genes (number of genes depending on the capture library version). Target enrichment was performed following the NimbleGen SeqCap RNA Target Enrichment protocol with minor modifications. Runs NGF3, NGF4, NGF7, and NGF8 were sequenced in an Illumina MiSeq platform, whereas run P/NP was sequenced in a HiSeq platform.

2.3 Bioinformatic pipeline

All samples were analyzed with a custom pipeline for RNA-seq data (Figure 18). FASTQ files were first filtered using fastp v0.20.0 (Chen *et al.*, 2018) using the following parameters: -w 10 --length_required 50 --cut_right --cut_right_mean_quality 10. FASTQ quality reports were generated with fastqc v0.11.4. FASTQ files were then mapped against the GRCh37 human genome assembly (Ensembl release 75 including SNPs and transcripts data) using HISAT2 v2.1.0 (Kim *et al.*, 2015). Sorted bam files were created with samtools v0.1.19 (Li *et al.*, 2009). Coverage metrics and alignment reports were generated using multiqc v1.7 (Ewels *et al.*, 2016). Gene expressions were obtained with stringtie v1.3.4 (Pertea *et al.*, 2015) and size-factor normalization was performed using DESeq2 package. The same stringtie version was used to identify and quantify known and *de novo* transcripts using the GRCh37 human genome assembly Ensembl release 75 as reference. DESeq2 package was used to perform size-factor normalization of the transcript expressions. RNA variant calling, including SNVs and INDELs, was performed using VarScan2 v.2.4.1 (Koboldt *et al.*, 2012) with the following parameters:--min-coverage 10--strand-filter 0--min-var-freq 5.

2.4 Transcripts filtering strategy

For each sample and gene, stringtie identified both *de novo* and already-defined Ensembl 75 transcripts. Usually, aberrant transcripts are not included in Ensembl definitions, so they were identified *de novo*. However, stringtie *de novo* transcripts discovery produced several low-support transcripts. Hence, we established a filtering transcripts strategy to discard very likely false positives. First, we removed low-support transcripts containing less than 20 reads supporting them. Second, we removed small transcripts that did not overlap any exon from any already-defined Ensembl transcript. Third, since each *de novo* transcript was quantified for all the samples, we established a 3 fold change threshold to discriminate against the rest of the noisy quantification. For each transcript, fold change (FC) was computed as the ratio between the most expressed sample and the second most expressed sample.

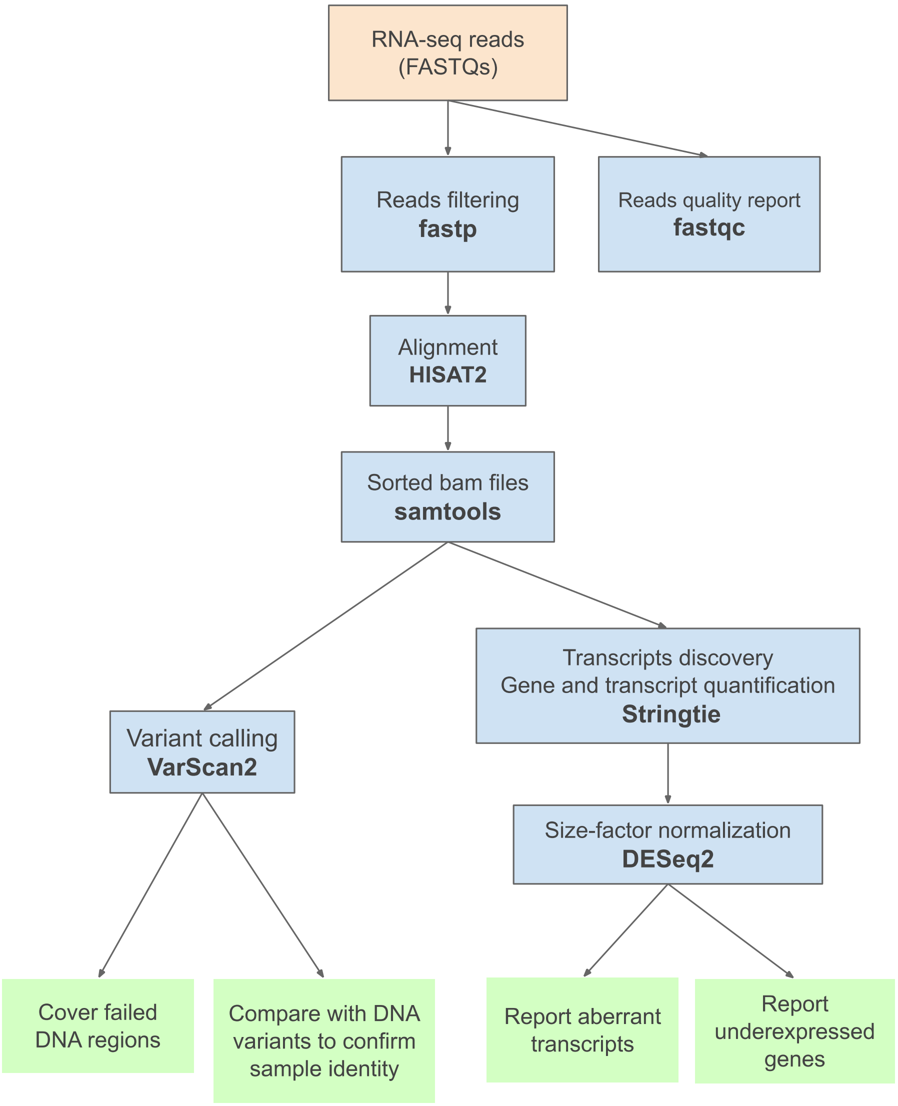


Figure 18. RNA-seq pipeline steps (blue) and applications derived from its results (green).

3 Results

3.1 Identification of aberrant transcripts

RNA-seq pipeline was used to identify aberrant transcripts, including both *de novo* and the already-defined in Ensembl 75. Pipeline performance for aberrant transcripts discovery was analyzed at two levels (Table 12). First, we analyzed the ability of stringtie to correctly identify the expected aberrant transcripts. Second, we analyzed how easily the aberrant transcripts could be discriminated from the rest of the noisy transcripts by using the FC threshold (FC = 3).

Table 12. Pipeline transcript identification for the already known mutations.

Sample	Run	cDNA change	RNA effect	Stringtie transcript identification	FC ratio > 3
S1	NGF3	*APC* c.423-3T>A	Exon skipping	Yes	Yes
S2	NGF3	*BRCA1* c.5243_5277+2788del 5277+2916_5277+2946delinsGG	Partial exon deletion and cryptic exon	Yes	Yes
S3	NGF4	*NF1* c.7908-321C>G	Cryptic exon	Yes	Yes
S4	NGF4	*BRCA1* c.212+1G>A	Intron retention	Yes	Known isoform, FC not estimated
S5	NGF7	*NF2* c.241-13T>A	Intron retention	Yes	Yes
S6	NGF7	*TSC2* c.4823_4825delACT	Partial exon deletion	Yes	No
S7	NGF7	*BRCA2* c.1763A>G	Intron retention	No	No
S8	NGF7	*BRCA1* c.5074+3A>G	Exon skipping	Yes	Yes
S9	NGF7	*BRCA1* c.(441+1_442-1)_(547+1_548-1)del	Partial exon deletion	Yes	Yes
S10	NGF7	*APC* c.1626+3A>G	Exon skipping	Yes	Yes
S11	NGF7	*MSH2* c.269A>C	Exon skipping and transposon insertion	No	No (FC 2.7)
S12	NGF7	*MSH6* c.3557-11_3557-4del	Exon skipping	No	No
S13	NGF7	*NF1* c.6117_6118	Intron retention	Yes	Yes
S14	NGF8	*NF1* c.6365-2A>G	Exon skipping	No	No
S15	NGF8	*NF1* c.733_835del	Partial exon deletion	Yes	No (FC 1.5)
S17	NGF8	*NF1* c.1466A>G	Exon skipping	No	No
S18	NGF8	*NF1* c.5205+2_5205+3dupTA	Partial exon deletion	Yes	Yes
S19	NGF8	*BRCA1* c.4484+1G>T	Exon skipping	Yes	No
S20	NGF8	*BRCA1* c.5243_5277+2788del 5277+2916_5277+2946delinsGG	Partial exon deletion and cryptic exon	No	No
S21	NGF8	*RAD51C* c.404G>A	Intron retention	Yes	Yes
S22	NGF8	*CHEK2* c.792+2T>C	Intron retention	No	No
S23	NGF8	*MSH2* c.2459-12A>G	Intron retention	Yes	No (FC 2.4)
S24	NGF8	*BRIP1* c.1628+5G>A	Exon skipping	Yes	Yes
S25	NGF8	*RAD51C* c.965+5G>A	Exon skipping	Yes	No (FC 2.2)
S26 P	P/NP	*CHEK2* c.1375G>C	Exon skipping	No	No
S26 NP				No	No
S27 P	P/NP	*BRCA1* c.213-12A>G	Intron retention	No	No
S27 NP				Yes	No
S28 P	P/NP	*BRCA2* c.-39_67del	Exon skipping	Yes	Yes
S28 NP				No	No
S29 P	P/NP	*BRCA2* c.9501+3A>T	Partial exon deletion	No	No
S29 NP				No	No
S30 P	P/NP	*MLH1* c.380G>A	Exon skipping	No	No
S30 NP				Yes	No (FC 2.1)
S31 P	P/NP	*CHEK2* c.792+2T>C	Intron retention	Yes	Yes
S31 NP				Yes	Yes

FC: Fold change ratio; P: puromycin was used; NP: puromycin was not used. Mutation in sample S4 produced an aberrant transcript already defined in Ensembl 75, so FC was not computed.

Stringtie identified 22 out of the 31 (71.0%) previously known aberrant transcripts (Table 12). All of them were identified as *de novo* transcripts, except for the transcript in sample S4 which was already defined in Ensembl 75. However, 14 out of the 31 transcripts (45.2%) achieved a transcript quantification in that sample sufficient to differentiate them from the rest of the samples through the FC ratio > 3. Also, 5 out of the 31 transcripts obtained an FC ratio between 1.5 and 3.

3.2 Variant calling

SNVs and INDEL calling was analyzed on three RNA-seq samples (run NGF3) for which the corresponding DNA-seq sample was previously available. Between 148 and 163 variants were detected for each pair of RNA and DNA samples (Table 13). 86.5-89.6% of the exonic DNA-seq variants were also detected in the RNA-seq samples. The remaining variants, those found in DNA-seq but not in RNA-seq, were placed in very lowly expressed genes, except for 2 of them.

On the other hand, all 93 variants detected only in RNA-seq data were considered likely false-positive calls for having low supporting coverage, low allele frequency, or for being likely recurrent artifacts.

Table 13. Variant calls comparison between DNA and RNA for samples S1, S2, and S32.

		S1	S2	S32
Variants called in DNA-seq and RNA-seq		148	160	163
Variants called only in DNA-seq	In a very-low expressed gene	23	18	18
	In an expressed gene	0	1	1
Variants called only in RNA-seq	Discarded for having low coverage or low allele frequency	26	32	28
	Discarded for being recurrent (likely artifacts)	2	3	2
	Probably true	0	0	0
Percentage of DNA-seq variants confirmed in RNA-seq		86.5%	89.4%	89.6%

4 Discussion

The genetic causes of hereditary diseases remains unknown for an important percentage of patients. Although this can be explained by multiple causes, analysis of RNA expression can be useful to provide insight into the genetic cause of hereditary diseases. The arrival of NGS methods reduced costs and turnaround times while increasing the number of samples sequenced at once. In this context, RNA-seq has the potential to improve the genetic diagnostics yield provided by DNA-seq (Park *et al.*, 2018; Gonorazky *et al.*, 2019; Landrith *et al.*, 2020).

4.1 Identification of aberrant transcripts

In this work, we implemented and evaluated a custom RNA-seq pipeline focusing on the identification of aberrant transcripts in genes of clinical interest. Our pipeline was able to identify 71.0% of the already-known aberrant transcripts, which highlights the potential of RNA-seq for the identification of these splicing products. However, the percentage dropped to 45.2% when considering only transcripts whose quantification was high enough to be discriminated from the rest via the FC threshold. The need to differentiate a given transcript in a sample from the rest, together with the high number of likely false positives that our pipeline discarded (Methods: transcripts filtering strategy), highlights that detecting aberrant transcripts using short-read RNA-seq is challenging. In fact, a previous benchmark (Sahraeian *et al.*, 2017) showed that the best aligner-assembler combination (HISAT2-stringtie) achieved sensitivities below 50% when reconstructing the transcriptome at a transcript level.

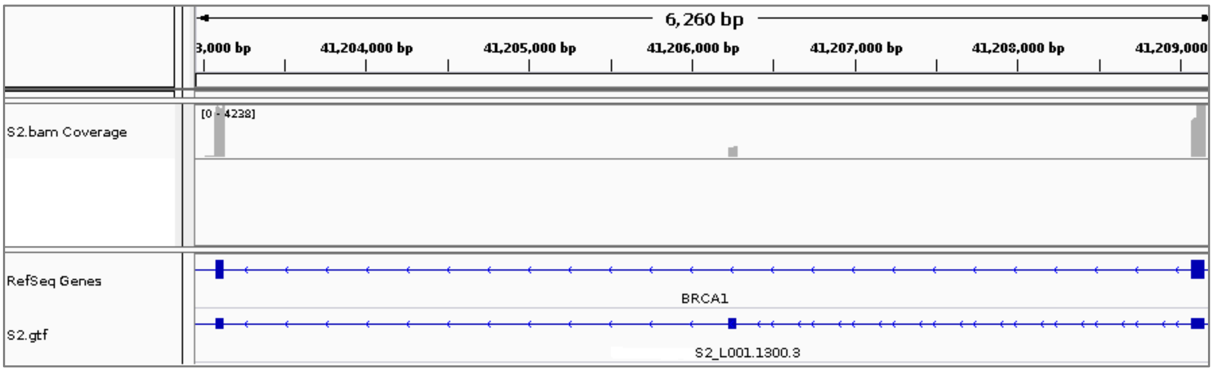


Figure 19. RNA transcript identified by stringtie in sample S2 (BRCA1). Two events (c.5243_5277+2788del and 5277+2916_5277+2946delinsGG) produced the insertion of a cryptic exon between exons 20-21 and a partial exon deletion at exon 20. The track S2_L001.1300.3 refers to the *de novo* transcript identified by stringtie. The coverage track (bam file) also supports the existence of these two events. Image obtained using IGV (Thorvaldsdottir et al., 2013).

Although transcript identification is still challenging, the percentage of transcripts successfully identified by our RNA-seq pipeline highlights the potential to improve the diagnostic yield of hereditary diseases. The identification of aberrant transcripts in clinically relevant genes provides insight into the genetic cause of inherited diseases. As an example, Figure 19 illustrates the *de novo* aberrant transcript identified in sample S2.

On the other hand, the low number of samples analyzed in this study prevents us from deriving any major conclusions regarding the factors that may affect the detection of aberrant transcripts, like the use of puromycin, the RNA effect type, or the capture version. For example, although all transcripts in runs NGF3 and NGF4 were successfully identified, no conclusions can be obtained from such a small number of samples (4).

4.2 Variant calling

Variant calling results showed that all DNA-seq variants were also called in the corresponding RNA-seq sample whenever the gene was expressed: this was the case for almost 90% of the DNA-seq variants. More DNA-RNA sample pairs should be analyzed to further confirm the results. However, results suggest that, if a gene is expressed, RNA-seq data could be used to call SNV/INDELs when DNA-seq data is not available because a region is failed for having a very low coverage level.

Moreover, since most variants were common to both RNA-seq and DNA-seq sample pairs (between 148 and 163), variant calling could be used to confirm that both samples are actually the same. Most variants should match among sample pairs; otherwise, there might have been an error during sample preparation or processing.

4.3 Other applications

RNA-seq data can be used for other purposes. Gene expression quantification allows for the detection of underexpressed or overexpressed genes when compared with other samples. In this study, gene quantification was obtained for each sample, although we did not analyze them because a larger sample size is necessary to clearly identify values placed in the extremes of the distribution. Also, other RNA-seq features were not covered in this study, such as allele-specific expression analysis and gene fusion discovery.

In summary, the RNA-seq pipeline implemented and evaluated in this study has the potential to improve our genetic diagnostics routine for hereditary cancer. The identification of aberrant transcripts, along with the RNA-seq variant calling and the discrimination of underexpressed clinically relevant genes, can provide evidence into the genetic causes of individuals with clinical suspicion of hereditary cancer.

Appendix A: additional publications

This chapter includes additional publications with the participation of the PhD candidate. In all of them, the contribution of the PhD candidate included bioinformatic support, NGS copy-number variant analysis, and manuscript review.

1 Exploring the Role of Mutations in Fanconi Anemia Genes in Hereditary Cancer Patients

Jesús del Valle, Paula Rofes, José Marcos Moreno-Cabrera, Adriana López-Dóriga, Sami Belhadj, Gardenia Vargas-Parra, Àlex Teulé, Raquel Cuesta, Xavier Muñoz, Olga Campos, Mónica Salinas, Rafael de Cid, Joan Brunet, Sara González, Gabriel Capellá, Marta Pineda, Lídia Feliubadaló and Conxi Lázaro*

Cancers 12, no. 4: 829 (2020). https://doi.org/10.3390/cancers12040829

IF 2019 = 6.126. Rank 37/244 (ONCOLOGY).

cancers

MDPI

Article

Exploring the Role of Mutations in Fanconi Anemia Genes in Hereditary Cancer Patients

Jesús del Valle [1,2,3], Paula Rofes [1,2,3], José Marcos Moreno-Cabrera [1,2,3], Adriana López-Dóriga [4,5], Sami Belhadj [1,2,3], Gardenia Vargas-Parra [1,2,3], Àlex Teulé [1,2,6], Raquel Cuesta [1,2,3], Xavier Muñoz [1,2,3], Olga Campos [1,2,3], Mónica Salinas [1,2], Rafael de Cid [7], Joan Brunet [1,2,3], Sara González [1,2,3], Gabriel Capellá [1,2,3], Marta Pineda [1,2,3], Lídia Feliubadaló [1,2,3] and Conxi Lázaro [1,2,3,*]

[1] Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL-IGTP-IDIBGI, 08908 Hospitalet de Llobregat, Spain; jdelvalle@iconcologia.net (J.d.V.); profes@iconcologia.net (P.R.); jmoreno@igtp.cat (J.M.M.-C.); samibelhadj@hotmail.com (S.B.); gvargas@idibell.cat (G.V.-P.); ateule@iconcologia.net (À.T.); rcuesta@iconcologia.net (R.C.); xmunoz@iconcologia.net (X.M.); ocampos@iconcologia.net (O.C.); msalinas@iconcologia.net (M.S.); Jbrunet@iconcologia.net (J.B.); sgonzalez@iconcologia.net (S.G.); gcapella@idibell.cat (G.C.); mpineda@iconcologia.net (M.P.); lfeliubadalo@iconcologia.net (L.F.)

[2] Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), IDIBELL, 08908 Hospitalet de Llobregat, Spain

[3] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), 28029 Madrid, Spain

[4] Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology, 08908 Hospitalet de Llobregat, Spain; alguerra@iconcologia.net

[5] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain

[6] Medical Oncology Department, Catalan Institute of Oncology, IDIBELL, 08908 Hospitalet de Llobregat, Spain

[7] Genomes for Life—GCAT lab Group, Institut Germans Trias i Pujol (IGTP), 08916 Badalona, Spain; rdecid@igtp.cat

* Correspondence: clazaro@iconcologia.net; Tel.: +34-93-2607145

Received: 7 February 2020; Accepted: 25 March 2020; Published: 30 March 2020

check for updates

Abstract: Fanconi anemia (FA) is caused by biallelic mutations in FA genes. Monoallelic mutations in five of these genes (*BRCA1, BRCA2, PALB2, BRIP1* and *RAD51C*) increase the susceptibility to breast/ovarian cancer and are used in clinical diagnostics as bona-fide hereditary cancer genes. Increasing evidence suggests that monoallelic mutations in other FA genes could predispose to tumor development, especially breast cancer. The objective of this study is to assess the mutational spectrum of 14 additional FA genes (*FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FANCP, FANCQ, FANCR* and *FANCU*) in a cohort of hereditary cancer patients, to compare with local cancer-free controls as well as GnomAD. A total of 1021 hereditary cancer patients and 194 controls were analyzed using our next generation custom sequencing panel. We identified 35 pathogenic variants in eight genes. A significant association with the risk of breast cancer/breast and ovarian cancer was found for carriers of *FANCA* mutations (odds ratio (OR) = 3.14 95% confidence interval (CI) 1.4–6.17, $p = 0.003$). Two patients with early-onset cancer showed a pathogenic FA variant in addition to another germline mutation, suggesting a modifier role for FA variants. Our results encourage a comprehensive analysis of FA genes in larger studies to better assess their role in cancer risk.

Keywords: Breast cancer risk; Breast and ovarian cancer risk; Fanconi Anemia; Hereditary Cancer; NGS panel sequencing

1. Introduction

Fanconi anemia (FA) is a rare genetic condition originated from a DNA repair deficiency that causes a broad spectrum of clinical features of variable penetrance, mainly, progressive bone marrow failure (depending on the affected gene), congenital defects and cancer predisposition [1]. FA is usually inherited as an autosomal recessive genetic disease, although X-linked inheritance and dominant inheritance have also been described.

Hitherto, 22 genes have been described as FA genes: *FANCA, FANCB, FANCC, FANCD1/BRCA2, FANCD2, FANCE, FANCF, FANCG/XRCC9, FANCI, FANCJ/BRIP1, FANCL/PHF9, FANCM, FANCN/PALB2, FANCO/RAD51C, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51, FANCS/BRCA1, FANCT/UBE2T, FANCU/XRCC2, FANCV/REV7* and *FANCW/RFWD3* [2]. The proteins encoded by these genes participate in the FA pathway involving DNA repair and genome maintenance processes when cell DNA damage occurs. These proteins are essential for inter-strand crosslink repair, and they also participate in homologous recombination and non-homologous end joining [3]. The *FANC-A, -B, -C, -E, -F, -G, -L* and *-M* genes encode the proteins that form the core complex, which monoubiquitinates the FANCI/FANCD complex formed by the dimer of FANCD2 and FANCI. The remaining proteins are downstream effectors in the FA pathway and their deficiency does not abolish the monoubiquitination of the I/D complex [4]. However, a recent publication described that biallelic FANCM mutations do not cause classical FA and therefore should not be considered a canonical FA gene [5], although these biallelic carriers showed risk for breast cancer, chemotherapy toxicity and may display chromosome fragility.

Apart from conditions caused by biallelic mutations in FA genes, it is well known that monoallelic mutations in certain FA genes (*BRCA1, BRCA2, BRIP1, PALB2* and *RAD51C*) are clearly related with hereditary breast and/or ovarian cancer predisposition [6], and these genes are bona-fide hereditary breast and ovarian cancer (HBOC) predisposition genes. Hence, cancer risks have been estimated for heterozygous mutations in these genes, and clinical management is also well established and accepted. However, the role of monoallelic mutations in the remaining FA genes regarding cancer predisposition is a matter of discussion. Over the last few years, several case-controls studies have indicated that monoallelic *FANCM* [7–15] truncating mutations are breast cancer risk factors; in addition, there are inconsistent results regarding *FANCA* [16–19], *FANCC* [20–24], *SLX4* [25–27] and *XRCC2* [28–30].

In the midst of these conflicting results, the use of comprehensive next generation sequencing (NGS) gene panels could shed some light on the role of FA genes in the context of hereditary cancer in general. For this reason, we analyzed these FA genes in our entire cohort of hereditary cancer patients, not just breast and ovarian cancer. Our I2HCP panel [31] contains, besides the five bona-fide HBOC genes, the following 14 FA genes: *FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG/XRCC9, FANCI, FANCL/PHF9, FANCM, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51* and *FANCU/XRCC2*. Here, we present the mutation profile of these 14 genes in our cohort of 1021 hereditary cancer patients and compare it with the mutational spectrum found in a control population consisting of 194 cancer-free individuals from our region as well as the GnomAD (genome aggregation database) non-cancer, European non-Finnish cohort.

2. Results

A prospective cohort of 1021 unrelated cancer cases with clinical suspicion of hereditary cancer was screened for mutations in the following 14 FA genes: *FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG/XRCC9, FANCI, FANCL/PHF9, FANCM, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51* and *FANCU/XRCC2*. The sequence of all coding regions and exon–intron boundaries (±20) was obtained by NGS and was also used to determine putative copy number variations (CNVs), which were validated by MLPA analysis. Other pathogenic variants identified in the clinical testing workflow, according to the clinical cascades presented in Feliubadaló et al. [32], are depicted in Table S1.

Our study identified 35 heterozygous carriers of 22 pathogenic/likely pathogenic variants in the patient cohort. The most frequently mutated genes were *FANCA, FANCL* and *FANCM*, whereas no mutation was identified in *FANCB, FANCD2, FANCF, FANCG, SLX4, ERCC4* and *XRCC2* (Table 1).

Six mutations were identified in our set of 194 healthy controls. Overall, a monoallelic mutation in a FA gene was identified in 3.4% of patients in our hereditary cancer cohort, a percentage very similar to that identified in the control cohort studied here (3.1%). However, distribution of mutations by clinical phenotype evidenced that pathogenic variants were mainly present in patients with a history of breast cancer, or breast and ovarian cancer. The percentage of pathogenic mutations increased to 4.6% (counting only women) in cases with breast cancer, being higher (5.5% counting only women) in those with breast and ovarian cancer (Figure 1).

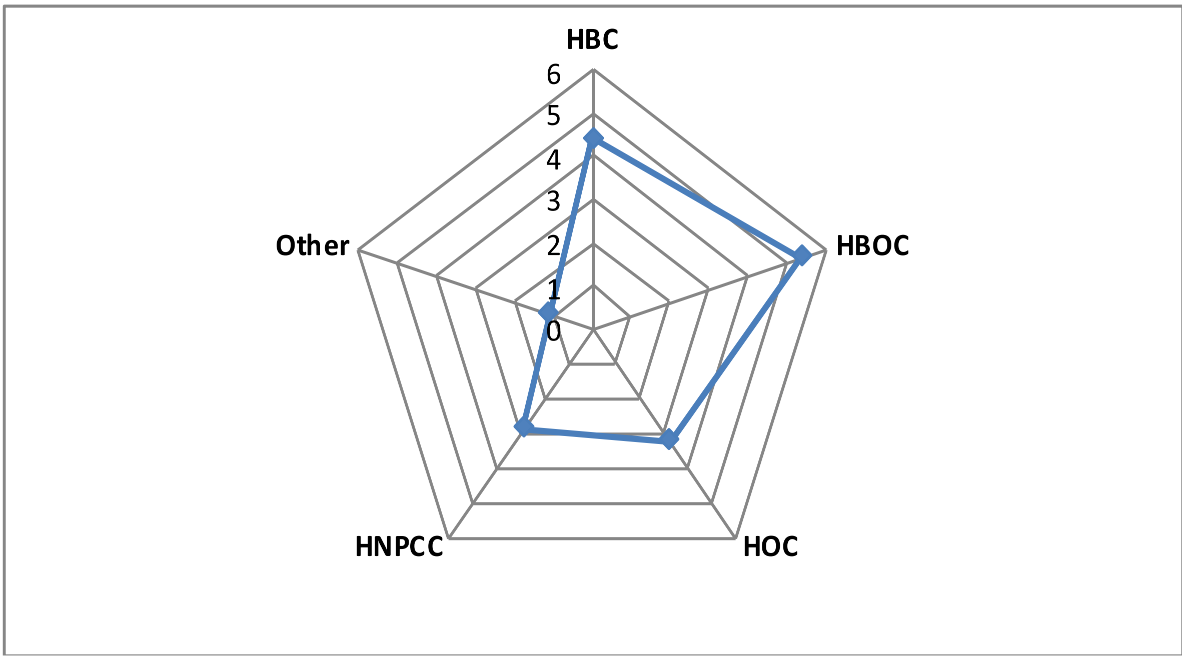


Figure 1. The diagram represents the percentage of pathogenic variants in the 14 Fanconi anemia (FA) genes analyzed per clinical suspicion group. HBC: Hereditary Breast Cancer Patients; HOC: Hereditary Ovarian Cancer Patients; HBOC: Hereditary Breast and Ovarian Cancer Patients; HNPCC: Hereditary non-polyposis colorectal cancer.

Details of all identified mutations and the clinical characteristics of the carriers are depicted in Table S2. Intriguingly, in three cases, an additional mutation in a hereditary cancer gene was also identified. Two of them were carriers of a deleterious variant in *FANCA*, one corresponds to a female with breast cancer at age 35 (patient ID 19136 in Table S2), carrier of a pathogenic variant in *ATM* and the other was diagnosed with ovarian cancer at age 49 and also harbors a mutation in *SDHB* (patient ID 6988 in Table S2). The third case, with a deleterious mutation in *FANCL*, is a Lynch syndrome patient with a mutation in *MLH1* who suffered colorectal cancer at age 29 (patient ID 19012 in Table S2). Interestingly, six of the mutations were identified in more than one individual, *p*.(Thr372Asnfs*13) in *FANCL* was identified in 10 individuals and *p*.(Arg1931*) in *FANCM* in 3 individuals, the remaining were identified in two cases each (Table S2).

DECoN (Detection of Exon Copy Number) analysis of NGS data identified 14 putative CNVs in the patient cohort that were validated by MLPA. Two turned out to be true positives consisting of a deletion of exons 6–13 in *FANCL* and a deletion of exons 11–37 in *FANCA*. Furthermore, 1605 variants of unknown significance (VUS) were identified in both cohorts, 589 unique (Table S3). Some of these VUS were predicted, by multiple in-silico tools, to alter correct splicing. Among them we were able to obtain lymphocytes for RNA analysis in five patients harboring the following mutations: *FANCA:* c.523-25_523-20delTTGTTT, c.576C > T, c.2217G > A and c.2602-9_2602-8delCT and *FANCM:* c.4222 + 5G > A. RNA analysis of these five variants did not identify any aberrant transcript (data not shown), so they remained classified as VUS.

Table 1. Summary of (Likely) Pathogenic Variants in 14 FA genes in the different clinical groups (only women are counted).

	Clinical Suspicion						GCAT Women Cohort (*n* = 100)	GnomAD European >23,000 women $^{\beta}$	Study Cohort Versus NFE $^{\gamma}$, Non-Cancer GnomAD (OR/95%CI/*p*-Value)		
Genes	**Pathogenic Variants**	**Breast (HBC)**	**Ovary (HOC)**	**Breast + Ovary (HBOC)**	**HNPCC** α	**Other**			**All Patients**	**HBC + HOC + HBOC**	**HBC + HBOC**
FANCA	10	7	0	2	1	0	3	147	1.94/0.91–3.7/0.047	2.34/1.04–4.59/0.02	3.14/1.4–6.17/**0.003***
FANCL	8	3	1	3	1	0	1	187	1.22/0.52–2.46/0.549	1.42/0.56–3/0.356	1.63/0.59–3.64/0.283
FANCM	6	2	3	0	1	0	0	159	1.07/0.38–2.39/0.828	1.19/0.38–2.85/0.618	0.63/0.08–2.34/0.774
FANCI	1	1	0	0	0	0	0	25	1.14/0.03–7/0.593	1.12/0,04–9.29/0.492	2.02/0.05–12.4/0.399
FANCE	2	1	0	0	1	0	0	17	1.14/0.03–6.97/0.593	1.52/0.04–9.3/0.492	2.03/0.05–12.4/0.399
FANCC	2	1	1	0	0	0	0	44	1.29/0.15–4.97/0.67	1.72/0.2–6.63/0.332	1.15/0.03–6.78/0.586
FANCF	1	1	0	0	0	0	0	26	1.09/0.02–6.6/0.608	1.45/0.04–8.88/0.506	1.94/0.05–11.87/0.412
RAD51	1	1	0	0	0	0	0	4	7.12/0.14–72/0.159	9.49/0.19–96/0.122	12.7/ 0.26–128/0.093
SLX4	0	0	0	0	0	0	1	36	NA	NA	NA
ERCC4	0	0	0	0	0	0	0	22	NA	NA	NA
FANCB	0	0	0	0	0	0	0	0	NA	NA	NA
FANCD2	0	0	0	0	0	0	0	21	NA	NA	NA
FANCG	0	0	0	0	0	0	0	43	NA	NA	NA
XRCC2	0	0	0	0	0	0	0	22	NA	NA	NA
TOTAL	31	17	5	5	4	0	5	753			

α Hereditary non-polyposis colorectal cancer; $^{\beta}$ The number of GnomAD non-Finnish, non-cancer women is slightly variable per gene but in all cases was greater than 23,000 γ NFE: non-Finnish European.

Lastly, we compared the mutational profile of our cohort of patients with data from the European (non-Finnish, non-cancer) GnomAD 2.1 population. After the first analysis, a possible association was only found with breast and ovarian cancer, we stratified the different populations by gender, counting only women (analysis without this stratification is shown in Table S4). By this means, only *FANCA* mutations showed a statistically significant association with an increased cancer risk (Table 1) in the combined group of hereditary breast cancer (HBC) and HBOC (odds ratio (OR) = 3.14 (95% confidence interval (CI) 1.4–6.17) p = 0.003). However, this association must be taken with caution since 3% of our in-house control cohort (from GCAT, Genomes for Life Cohort) carried deleterious *FANCA* mutations compared with 0.6% of the European non-Finnish cohort, being 0.98% in our complete cohort of hereditary cancer patients.

3. Discussion

In this study, we have evaluated the presence of deleterious mutations in 14 FA genes in a cohort of 1021 patients in the context of hereditary cancer. In total, 3.4% of the patients have a pathogenic variant in one of these genes. This percentage is higher in the group of women patients with breast cancer (4.4%) and increases in the group of women patients with a history of breast and ovarian cancer (5.4%). We analyzed these genes in two European populations, a general adult population cohort from Spain (GCAT) and in the European non-Finnish GnomAD cohort, identifying pathogenic variants in 3.1% and 3.0% of control individuals, respectively. If only women are considered, the percentages increase to 5% in GCAT and 3.2% in GnomAD. The NGS analysis performed allowed us not only to detect single nucleotide variants but also to screen for CNVs. By this means, we identified two large intragenic deletions in *FANCL* and *FANCA*, highlighting the importance of searching for this type of variant when analyzing FA genes in patients with Fanconi anemia.

In general, the genes most frequently mutated in our cohort of patients were *FANCA* (n = 10), *FANCL* (n = 10) and *FANCM* (n = 7). Few cases were identified with mutations in *FANCI, FANCE, FANCC* (n = 2, in each gene) and *FANCF* and *RAD51* (n = 1, in each gene). No mutations were identified in *FANCB, FANCD2, FANCF, FANCG, ERCC4* and *XRCC2*, and only one pathogenic variant was identified in *SLX4*, but in a sample corresponding to a healthy control. Hence, it seems that most of these 14 FA genes do not play a major role in hereditary cancer, although our data cannot discard their relation with rare cancer syndromes or their role as modifier genes. To assess these possibilities, larger cohorts of patients with different tumor types and the use of polygenic risk score methodologies should be applied.

It is worth mentioning that one of the most frequently mutated genes in our series, as well as in the European non-Finnish GnomAD cohort, is *FANCL*. This fact is due to the high number of patients carrying the c.1111_1114dup mutation. This alteration, located in the last exon of the gene, produces a frameshift that lengthens the protein by three amino acids more than wild-type. This mutation has been described in a patient with FA, a compound heterozygote with another *FANCL* mutation [33]. Functional analysis of this mutation identified a partial correction of G2/M cell cycle arrest that results in an intermediate phenotype compatible with a hypomorphic mutation. So, the contribution to cancer risk of this variant in monoallelic carriers could be very limited but deserves further study. In our series, we also detected an enrichment of the c.5791C > T variant in *FANCM*. This alteration is the most common pathogenic *FANCM* variant in Southern Europe [34] and was associated with estrogen receptorER-negative breast cancer risk (OR = 1.96; p = 0.006) in a large case-control study with more than 50,000 cases and controls [15]. However, in the present study, we could not find a significant association with breast cancer risk (odds ratio = 1.46 (95% confidence interval 0.3–4.8) p = 0.467).

4. Materials and Methods

4.1. Patients and Controls

A total of 1021 hereditary cancer-suspected index cases, referred through our genetic counselling units, that underwent NGS panel testing based on clinical suspicion [32], were included in this study (Table 2). Genetic counselors followed international guidelines to request germline genetic tests under the suspicion of a hereditary cancer syndrome. Informed written consent for both diagnostics and research purposes was obtained from all patients included in the study and the study protocol was approved by the Ethics Committee of IDIBELL (Bellvitge Biomedical Research Institute, PR278/19).

Table 2. Summary of the hereditary cancer cohort by clinical suspicion.

Clinical Suspicion	Number of Patients (Women)
Hereditary breast cancer, HBC	385 (370)
Hereditary non-polyposis colon cancer, HNPCC	210 (130)
Hereditary ovarian cancer, HOC	154 (154)
Other hereditary cancer conditions	102 (55)
Hereditary breast and ovarian cancer, HBOC	93 (90)
Familial (attenuated) adenomatous polyposis, FAP/AFAP	77 (19)
Total	1021 (818)

A set of 194 cancer-free controls (100 women) from GCAT, Genomes for Life Cohort, was also analyzed.

GCAT (Cohort Study of the Genomes of Catalonia Study) is a biomedical research project designed for the study of genetic, epigenetic and environmental factors that lead to the appearance of different complex inheritance diseases in the general population [35]. Briefly, the subjects of the present study are part of the GCAT project, a prospective study that includes a cohort of a total of 19,267 participants recruited from the general population of Catalonia, a western Mediterranean region in the Northeast of Spain. All are cancer-free general population volunteers between 40 and 65 years of age. All eligible participants signed an informed consent agreement form. The GCAT study was approved by the local ethics committee (IRB00002131) (Germans Trias University Hospital) in 2013.

4.2. DNA Isolation

Genomic DNA was extracted from peripheral blood lymphocytes using the FlexiGene DNA Kit (Qiagen GmbH, Hilden, Germany) in the patient cohort and the ReliaPrep DNA Kit (Promega, Wisconsin, USA) in the GCAT cohort.

4.3. NGS Panel Testing

All patients and controls were analyzed by our validated custom NGS panel I2HCP, which comprises 122–135 hereditary cancer-associated genes, depending on the version used [31]. This panel includes the *FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG/XRCC9, FANCI, FANCL/PHF9, FANCM, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51* and *FANCU/XRCC2* genes. Library preparation methods and bioinformatics pipeline were described previously [31]. The regions of interest analyzed include all coding regions and ±20 nucleotides intron/exon boundaries. For this study we considered as a pathogenic or likely pathogenic variant (pathogenic variant hereinafter) mutations that originate a premature stop codon, missense variants described in the literature as clearly pathogenic in FA patients and mutations affecting canonical splice site positions (+1, +2, −1,−2). All pathogenic variants were confirmed by Sanger sequencing.

Copy number analysis was performed from NGS data using the DECoN [36] tool with parameter optimization for our panel (Moreno et al., submitted manuscript). However, the *FANCB* gene was not included in this analysis as it is located on the X chromosome, which greatly complicates the identification of CNVs with our pipeline. Likewise, *FANCD2* was also excluded from this analysis due to the presence of pseudogenes, which generate false positives in both directions (deletions and duplications). For the rest of the genes, we used the Bayesian-factor value, which is a good predictor of the reliability of the DECoN's result to select the most likely true positive copy number alterations to be confirmed. All samples with a suspicion of alteration were subsequently analyzed by MLPA using custom probes according to the instructions provided by MRC-Holland in order to validate or discard the presence of CNVs (https://support.mlpa.com/downloads/files/designing-synthetic-mlpa-probes).

4.4. RNA Analysis

Lymphocytes were isolated by centrifugation of peripheral blood samples from carriers and controls. Cells were cultured in PB-Max medium for 5 to 7 days and treated with puromycin 4 to 6 h before RNA extraction in order to prevent the potential degradation of unstable transcripts by nonsense-mediated decay (NMD). Total RNA was isolated using TRIzol reagent according to the manufacturer's instructions. One microgram of total RNA was reverse transcribed using the iScript cDNA Synthesis kit (Bio-Rad Laboratories, Hercules, CA, USA). cDNA amplification was performed with specific primers that encompassed the region of interest. Transcriptional profiles from carriers were compared to those derived from control lymphocytes cultures, both by agarose gel analysis and Sanger sequencing. Primer sequences and PCR conditions are available upon request.

4.5. GnomAD Analysis

The GnomAD non-Finnish European, non-cancer subpopulation (Genome Aggregation Database, v2.1.1, http://gnomad.broadinstitute.org/) was used as a control population. Variants were exported and filtered to identify predicted loss of function variants in FA genes.

4.6. Statistical Analysis

Differences in allele frequency between cases and controls were determined by the Fisher exact test. Odds ratios (OR) and the corresponding 95% confidence intervals (CI) were determined for two by two comparisons. Statistical tests were carried out using R v.3.5.1. (R Foundation for Statistical Computing, Vienna, Austria).

5. Conclusions

Our study identified an increased number of pathogenic mutations in *FANCA* in the HBC/HBOC group (p = 0.003). In addition, we observed a higher number of mutations in the remaining genes (5.4% versus 3.2%) in the group of patients with a history of breast and ovarian cancer. Two out of the three cases with additional mutations in other moderate/high-penetrance genes, had been diagnosed with cancer at a very young age, suggesting a modifier role for FA mutations. Altogether, our results encourage further studies in larger cohorts to assess the role and risks of deleterious variants in these genes to determine their potential future use in clinical settings.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/12/4/829/s1: Table S1: List of pathogenic variants in other hereditary cancer genes, Table S2: List of (likely) pathogenic variants with clinical detail of the patients' tumors and family cancer history, Table S3: List of VUS variants identified, Table S4: Summary of (Likely) Pathogenic variants in 14 FA genes in the different groups without gender stratification.

Author Contributions: J.d.V., L.F. and C.L. conceived, designed, and planned the study; J.d.V., P.R., J.M.M.-C., A.L.-D., S.B., G.V.-P., R.C., X.M., O.C. and M.P. contributed to the acquisition, analysis, and/or interpretation of the molecular data; À.T., M.S., R.d.C., V.M., J.B., S.G. and G.C. provided samples and clinical data; J.V., P.R. and C.L. drafted the manuscript. All authors critically reviewed or revised the manuscript for intellectual content and approved the final version. All authors have read and agreed to the published version of the manuscript.

Funding: Contract grant sponsor: Supported by the Carlos III National Health Institute funded by FEDER funds—a way to build Europe—(PI19/00553; PI16/00563; PI16/01898 and CIBERONC); the Government of Catalonia (Pla estratègic de recerca i innovació en salut (PERIS_MedPerCan and URDCat projects), 2017SGR1282 and 2017SGR496) and GCAT Cession Research Project 2017_01HC_GCAT.

Acknowledgments: We thank CERCA Program/Generalitat de Catalunya for their institutional support. We wish also to thank all the members of the ICO Hereditary Cancer Program. This study makes use of data generated by the GCAT, Genomes for Life Project. Cohort study of the Genomes of Catalonia", IGTP. IGTP is part of the CERCA Program/Generalitat de Catalunya. A full list of the investigators who contributed to the generation of the data is available from www.genomesforlife.com.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mamrak, N.E.; Shimamura, A.; Howlett, N.G. Recent discoveries in the molecular pathogenesis of the inherited bone marrow failure syndrome Fanconi anemia. *Blood Rev.* **2017**, *31*, 93–99. [CrossRef] [PubMed]
2. Asur, R.S.; Kimble, D.C.; Lach, F.P.; Jung, M.; Donovan, F.X.; Kamat, A.; Noonan, R.J.; Thomas, J.W.; Park, M.; Chines, P.; et al. Somatic mosaicism of an intragenic FANCB duplication in both fibroblast and peripheral blood cells observed in a Fanconi anemia patient leads to milder phenotype. *Mol. Genet. Genom. Med.* **2018**, *6*, 77–91. [CrossRef] [PubMed]
3. Kottemann, M.C.; Smogorzewska, A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* **2013**, *493*, 356–363. [CrossRef] [PubMed]
4. Bogliolo, M.; Surrallés, J. Fanconi anemia: A model disease for studies on human genetics and advanced therapeutics. *Curr. Opin. Genet. Dev.* **2015**, *33*, 32–40. [CrossRef] [PubMed]
5. Catucci, I.; Osorio, A.; Arver, B.; Neidhardt, G.; Bogliolo, M.; Zanardi, F.; Riboni, M.; Minardi, S.; Pujol, R.; Azzollini, J.; et al. Individuals with FANCM biallelic mutations do not develop Fanconi anemia, but show risk for breast cancer, chemotherapy toxicity and may display chromosome fragility. *Genet. Med.* **2018**, *20*, 452–457. [CrossRef]
6. Daly, M.B.; Pilarski, R.; Berry, M.; Buys, S.S.; Farmer, M.; Friedman, S.; Garber, J.E.; Kauff, N.D.; Khan, S.; Klein, C.; et al. NCCN Guidelines Insights: Genetic/Familial High-Risk Assessment: Breast and Ovarian, Version 2. *J. Natl. Compr. Cancer Netw.* **2017**, *15*, 9–20. [CrossRef]
7. Gracia-Aznarez, F.J.; Fernandez, V.; Pita, G.; Peterlongo, P.; Dominguez, O.; de la Hoya, M.; Duran, M.; Osorio, A.; Moreno, L.; Gonzalez-Neira, A.; et al. Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS ONE* **2013**, *8*, e55681. [CrossRef]
8. Kiiski, J.I.; Pelttari, L.M.; Khan, S.; Freysteinsdottir, E.S.; Reynisdottir, I.; Hart, S.N.; Shimelis, H.; Vilske, S.; Kallioniemi, A.; Schleutker, J.; et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15172–15177. [CrossRef]
9. Peterlongo, P.; Catucci, I.; Colombo, M.; Caleca, L.; Mucaki, E.; Bogliolo, M.; Marin, M.; Damiola, F.; Bernard, L.; Pensotti, V.; et al. FANCM c.5791C > T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum. Mol. Genet.* **2015**, *24*, 5345–5355. [CrossRef]
10. Neidhardt, G.; Hauke, J.; Ramser, J.; Groß, E.; Gehrig, A.; Müller, C.R.; Kahlert, A.K.; Hackmann, K.; Honisch, E.; Niederacher, D.; et al. Association Between Loss-of-Function Mutations Within the FANCM Gene and Early-Onset Familial Breast Cancer. *JAMA Oncol.* **2017**, *3*, 1245–1248. [CrossRef]
11. Dicks, E.; Song, H.; Ramus, S.J.; Oudenhove, E.V.; Tyrer, J.P.; Intermaggio, M.P.; Kar, S.; Harrington, P.; Bowtell, D.D.; Group, A.S.; et al. Germline whole exome sequencing and large-scale replication identifies. *Oncotarget* **2017**, *8*, 50930–50940. [CrossRef] [PubMed]
12. Nguyen-Dumont, T.; Myszka, A.; Karpinski, P.; Sasiadek, M.M.; Akopyan, H.; Hammet, F.; Tsimiklis, H.; Park, D.J.; Pope, B.J.; Slezak, R.; et al. FANCM and RECQL genetic variants and breast cancer susceptibility: Relevance to South Poland and West Ukraine. *BMC Med. Genet.* **2018**, *19*, 12. [CrossRef] [PubMed]
13. Schubert, S.; van Luttikhuizen, J.L.; Auber, B.; Schmidt, G.; Hofmann, W.; Penkert, J.; Davenport, C.F.; Hille-Betz, U.; Wendeburg, L.; Bublitz, J.; et al. The identification of pathogenic variants in BRCA1/2 negative, high risk, hereditary breast and/or ovarian cancer patients: High frequency of FANCM pathogenic variants. *Int. J. Cancer* **2019**, *144*, 2683–2694. [CrossRef] [PubMed]
14. Nurmi, A.; Muranen, T.A.; Pelttari, L.M.; Kiiski, J.I.; Heikkinen, T.; Lehto, S.; Kallioniemi, A.; Schleutker, J.; Bützow, R.; Blomqvist, C.; et al. Recurrent moderate-risk mutations in Finnish breast and ovarian cancer patients. *Int. J. Cancer* **2019**, *145*, 2692–2700. [CrossRef] [PubMed]

15. Figlioli, G.; Bogliolo, M.; Catucci, I.; Caleca, L.; Lasheras, S.V.; Pujol, R.; Kiiski, J.I.; Muranen, T.A.; Barnes, D.R.; Dennis, J.; et al. The FANCM:p.Arg658* truncating variant is associated with risk of triple-negative breast cancer. *NPJ Breast Cancer* **2019**, *5*, 38. [CrossRef] [PubMed]
16. Seal, S.; Barfoot, R.; Jayatilake, H.; Smith, P.; Renwick, A.; Bascombe, L.; McGuffog, L.; Evans, D.G.; Eccles, D.; Easton, D.F.; et al. Evaluation of Fanconi Anemia genes in familial breast cancer predisposition. *Cancer Res.* **2003**, *63*, 8596–8599.
17. Haiman, C.A.; Hsu, C.; de Bakker, P.I.; Frasco, M.; Sheng, X.; Van Den Berg, D.; Casagrande, J.T.; Kolonel, L.N.; Le Marchand, L.; Hankinson, S.E.; et al. Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. *Hum. Mol. Genet.* **2008**, *17*, 825–834. [CrossRef]
18. Solyom, S.; Winqvist, R.; Nikkilä, J.; Rapakko, K.; Hirvikoski, P.; Kokkonen, H.; Pylkäs, K. Screening for large genomic rearrangements in the FANCA gene reveals extensive deletion in a Finnish breast cancer family. *Cancer Lett.* **2011**, *302*, 113–118. [CrossRef]
19. Abbasi, S.; Rasouli, M. A rare FANCA gene variation as a breast cancer susceptibility allele in an Iranian population. *Mol. Med. Rep.* **2017**, *15*, 3983–3988. [CrossRef]
20. van der Heijden, M.S.; Yeo, C.J.; Hruban, R.H.; Kern, S.E. Fanconi anemia gene mutations in young-onset pancreatic cancer. *Cancer Res.* **2003**, *63*, 2585–2588.
21. Couch, F.J.; Johnson, M.R.; Rabe, K.; Boardman, L.; McWilliams, R.; de Andrade, M.; Petersen, G. Germ line Fanconi anemia complementation group C mutations and pancreatic cancer. *Cancer Res.* **2005**, *65*, 383–386. [PubMed]
22. Berwick, M.; Satagopan, J.M.; Ben-Porat, L.; Carlson, A.; Mah, K.; Henry, R.; Diotti, R.; Milton, K.; Pujara, K.; Landers, T.; et al. Genetic heterogeneity among Fanconi anemia heterozygotes and risk of cancer. *Cancer Res.* **2007**, *67*, 9591–9596. [CrossRef] [PubMed]
23. Thompson, E.R.; Doyle, M.A.; Ryland, G.L.; Rowley, S.M.; Choong, D.Y.; Tothill, R.W.; Thorne, H.; Barnes, D.R.; Li, J.; Ellul, J.; et al. Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. *PLoS Genet.* **2012**, *8*, e1002894. [CrossRef] [PubMed]
24. Dörk, T.; Peterlongo, P.; Mannermaa, A.; Bolla, M.K.; Wang, Q.; Dennis, J.; Ahearn, T.; Andrulis, I.L.; Anton-Culver, H.; Arndt, V.; et al. Two truncating variants in FANCC and breast cancer risk. *Sci. Rep.* **2019**, *9*, 12524. [CrossRef] [PubMed]
25. Bakker, J.L.; van Mil, S.E.; Crossan, G.; Sabbaghian, N.; De Leeneer, K.; Poppe, B.; Adank, M.; Gille, H.; Verheul, H.; Meijers-Heijboer, H.; et al. Analysis of the novel fanconi anemia gene SLX4/FANCP in familial breast cancer cases. *Hum. Mutat.* **2013**, *34*, 70–73. [CrossRef]
26. de Garibay, G.R.; Díaz, A.; Gaviña, B.; Romero, A.; Garre, P.; Vega, A.; Blanco, A.; Tosar, A.; Díez, O.; Pérez-Segura, P.; et al. Low prevalence of SLX4 loss-of-function mutations in non-BRCA1/2 breast and/or ovarian cancer families. *Eur. J. Hum. Genet.* **2013**, *21*, 883–886. [CrossRef]
27. Shah, S.; Kim, Y.; Ostrovnaya, I.; Murali, R.; Schrader, K.A.; Lach, F.P.; Sarrel, K.; Rau-Murthy, R.; Hansen, N.; Zhang, L.; et al. Assessment of SLX4 Mutations in Hereditary Breast Cancers. *PLoS ONE* **2013**, *8*, e66961. [CrossRef]
28. Park, D.J.; Lesueur, F.; Nguyen-Dumont, T.; Pertesi, M.; Odefrey, F.; Hammet, F.; Neuhausen, S.L.; John, E.M.; Andrulis, I.L.; Terry, M.B.; et al. Rare mutations in XRCC2 increase the risk of breast cancer. *Am. J. Hum. Genet.* **2012**, *90*, 734–739. [CrossRef]
29. Hilbers, F.S.; Wijnen, J.T.; Hoogerbrugge, N.; Oosterwijk, J.C.; Collee, M.J.; Peterlongo, P.; Radice, P.; Manoukian, S.; Feroce, I.; Capra, F.; et al. Rare variants in XRCC2 as breast cancer susceptibility alleles. *J. Med. Genet.* **2012**, *49*, 618–620. [CrossRef]
30. Kluźniak, W.; Wokołorczyk, D.; Rusak, B.; Huzarski, T.; Gronwald, J.; Stempa, K.; Rudnicka, H.; Kashyap, A.; Dębniak, T.; Jakubowska, A.; et al. Inherited variants in XRCC2 and the risk of breast cancer. *Breast Cancer Res. Treat.* **2019**, *178*, 657–663. [CrossRef]
31. Castellanos, E.; Gel, B.; Rosas, I.; Tornero, E.; Santín, S.; Pluvinet, R.; Velasco, J.; Sumoy, L.; Del Valle, J.; Perucho, M.; et al. A comprehensive custom panel design for routine hereditary cancer testing: Preserving control, improving diagnostics and revealing a complex variation landscape. *Sci. Rep.* **2017**, *7*, 39348. [CrossRef] [PubMed]

32. Feliubadalo, L.; Lopez-Fernandez, A.; Pineda, M.; Diez, O.; Del Valle, J.; Gutierrez-Enriquez, S.; Teule, A.; Gonzalez, S.; Stjepanovic, N.; Salinas, M.; et al. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int. J. Cancer* **2019**, *145*, 2682–2691. [CrossRef] [PubMed]
33. Ali, A.M.; Kirby, M.; Jansen, M.; Lach, F.P.; Schulte, J.; Singh, T.R.; Batish, S.D.; Auerbach, A.D.; Williams, D.A.; Meetei, A.R. Identification and characterization of mutations in FANCL gene: A second case of Fanconi anemia belonging to FA-L complementation group. *Hum. Mutat.* **2009**, *30*, E761–E770. [CrossRef] [PubMed]
34. Figlioli, G.; Kvist, A.; Tham, E.; Soukupova, J.; Kleiblova, P.; Muranen, T.A.; Andrieu, N.; Azzollini, J.; Balmaña, J.; Barroso, A.; et al. The Spectrum of *FANCM* Protein Truncating Variants in European Breast Cancer Cases. *Cancers (Basel)* **2020**, *12*, 292. [CrossRef] [PubMed]
35. Obón-Santacana, M.; Vilardell, M.; Carreras, A.; Duran, X.; Velasco, J.; Galván-Femenía, I.; Alonso, T.; Puig, L.; Sumoy, L.; Duell, E.J.; et al. GCAT|Genomes for life: A prospective cohort study of the genomes of Catalonia. *BMJ Open* **2018**, *8*, e018324. [CrossRef]
36. Fowler, A.; Mahamdallie, S.; Ruark, E.; Seal, S.; Ramsay, E.; Clarke, M.; Uddin, I.; Wylie, H.; Strydom, A.; Lunter, G.; et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res.* **2016**, *1*, 20. [CrossRef]

2 Comprehensive analysis and ACMG-based classification of CHEK2 variants in hereditary cancer patients

Gardenia Vargas-Parra, Jesús Del Valle, Paula Rofes, Mireia Gausachs, Agostina Stradella, José Marcos Moreno-Cabrera, Angela Velasco, Eva Tornero, Mireia Menéndez, Xavier Muñoz, Silvia Iglesias, Adriana López-Doriga, Daniel Azuara, Olga Campos, Raquel Cuesta, Esther Darder, Rafael de Cid, Sara González, Alex Teulé, Matilde Navarro, Joan Brunet, Gabriel Capellá, Marta Pineda, Lídia Feliubadaló and Conxi Lázaro

Hum Mutat. 2020 Dec;41(12):2128-2142. doi: 10.1002/humu.24110.

IF 2019 = 4.124. Rank 44/177 (GENETICS & HEREDITY)

Received: 22 May 2020 | Revised: 13 August 2020 | Accepted: 6 September 2020

DOI: 10.1002/humu.24110

RESEARCH ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Comprehensive analysis and ACMG-based classification of *CHEK2* variants in hereditary cancer patients

Gardenia Vargas-Parra[1,2,3] | Jesús del Valle[1,2,3] | Paula Rofes[1,2,3] | Mireia Gausachs[1,2] | Agostina Stradella[1,2,4] | José M. Moreno-Cabrera[1,2,3] | Angela Velasco[1,2,3] | Eva Tornero[1,2,3] | Mireia Menéndez[1,2,3] | Xavier Muñoz[1,2,3] | Silvia Iglesias[1,2,3] | Adriana López-Doriga[5,6] | Daniel Azuara[1,2,3] | Olga Campos[1,2,3] | Raquel Cuesta[1,2,3] | Esther Darder[1,2,3] | Rafael de Cid[7] | Sara González[1,2,3] | Alex Teulé[1,2,3] | Matilde Navarro[1,2,3] | Joan Brunet[1,2,3,8] | Gabriel Capellá[1,2,3] | Marta Pineda[1,2,3] | Lídia Feliubadaló[1,2,3] | Conxi Lázaro[1,2,3]

[1]Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL-IGTP-IDIBGI, Badalona, Spain

[2]Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), IDIBELL, Barcelona, Spain

[3]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain

[4]Medical Oncology Department, Catalan Institute of Oncology, IDIBELL, Barcelona, Spain

[5]Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology, Barcelona, Spain

[6]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain

[7]Programa de Medicina Predictiva i Personalitzada del Càncer—Institut Germans Trias i Pujol (PMPPC-IGTP), Genomes for Life—GCAT Lab Group, Badalona, Spain

[8]Medical Sciences Department, School of Medicine, University of Girona, Girona, Spain

Correspondence

Conxi Lázaro, Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL and CIBERONC, Av. Gran Via 199-203, Hospitalet de Llobregat, Badalona, 08908 Girona, Spain.
Email: clazaro@iconcologia.net

Abstract

CHEK2 variants are associated with intermediate breast cancer risk, among other cancers. We aimed to comprehensively describe *CHEK2* variants in a Spanish hereditary cancer (HC) cohort and adjust the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) guidelines for their classification. First, three *CHEK2* frequent variants were screened in a retrospective Hereditary Breast and Ovarian Cancer cohort of 516 patients. After, the whole *CHEK2* coding region was analyzed by next-generation sequencing in 1848 prospective patients with HC suspicion. We refined ACMG-AMP criteria and applied different combined rules to classify *CHEK2* variants and define risk alleles. We identified 10 *CHEK2* null variants, 6 missense variants with discordant interpretation in ClinVar database, and 35 additional variants of unknown significance. Twelve variants were classified as (likely)-pathogenic; two can also be considered “established risk-alleles” and one as “likely risk-allele.” The prevalence of (likely)-pathogenic variants in the HC cohort was 0.8% (1.3% in breast cancer patients and 1.0% in hereditary nonpolyposis colorectal cancer patients). Here, we provide ACMG adjustment guidelines to classify *CHEK2* variants. We hope that this study would be useful for variant classification of other genes with low effect variants.

KEYWORDS

CHEK2, hereditary cancer, low penetrance, molecular diagnosis, risk allele, variant classification

Gardenia Vargas-Parra and Jesús del Valle contributed equally to the work and share first authorship.

1 | INTRODUCTION

Extensive efforts to standardize variant classification criteria in highly penetrant genes have been made by different groups, such as the joint consensus of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) (C. S. Richards et al., 2008; S. Richards et al., 2015), ENIGMA consortium for *BRCA1/2* genes (Spurdle et al., 2012), or InSiGHT variant interpretation group for MMR genes (Plon et al., 2008). However, there is still important work to be done in moderate or low-penetrance genes (Katona et al., 2018) since multigene panels for hereditary cancer (HC) include them. A recent publication proposed a framework for classification of variants in low-penetrance genes, in which a variant could be classified as an established risk allele (ERA) if it has been assessed in case–control studies of good design and data quality, demonstrated to be cancer-related and determined through robust meta-analysis (Senol-Cosar et al., 2019).

In the present work, we have focused on *CHEK2* (checkpoint kinase 2; MIM# 604373), which is a tumor-suppressor gene mainly associated with breast cancer (BC) although it has also been associated with other forms of HC, such as colorectal cancer (CRC) (Bell et al., 1999; Meijers-Heijboer et al., 2002). *CHEK2* is included in most of the in-house and commercial HC panels (Easton et al., 2015). *CHEK2* mRNA has a total length of 1844-bp distributed in 15 exons, is located at chromosome 22q12.1 and encodes for a human protein of 543-aa, an analog of the yeast checkpoint kinases Cds1 and Rad53 (Matsuoka et al., 2000). CHK2 protein is a kinase involved in several cellular processes, including the control of mitosis and meiosis progression, and plays an important role in the DNA-damage signaling network (Bartek, Falck, & Lukas, 2001; Zannini, Delia, & Buscemi, 2014). ATM activates CHK2 in response to DNA damage. Once activated, CHK2 is capable of phosphorylating many substrates involved in DNA repair, cell cycle regulation, p53 signaling, and apoptosis (Zannini et al., 2014).

A few *CHEK2* variants have been described as recurrent or founder variants in some populations. The most well-known *CHEK2* variant is c.1100delC, and it is primarily present in individuals of Northern and Eastern European descent; it results in a premature stop codon within exon 10, impairing the kinase ability of the enzyme (Wu, Webster, & Chen, 2001). A meta-analysis of 44,777 patients and 42,997 controls established a BC odds ratio (OR) of 2.26 for *CHEK2* c.1100delC carriers (Schmidt et al., 2016). Another frameshift founder mutation, the deletion of exons 9 and 10, is considered to double BC risk (Cybulski et al., 2007). The missense variant c.470T>C, p (.Ile157Thr) is described to confer lower risk compared with the two previous ones (OR of 1.58 and 1.67 for BC and CRC, respectively) (Han, Guo, & Liu, 2013). According to a study of 13,087 BC cases and 5,488 controls, the OR for 73 *CHEK2* rare missense variants was 1.36 (95% confidence interval [CI], 0.99–1.87) and 1.51 (95% CI, 1.02–2.24) if considering only variants in functional domains (Decker et al., 2017; Han et al., 2013). Furthermore, in a recent study of 1355 BC cases, the OR for *CHEK2* missense variants varied between 3.79 and 5.9 (95% CI, 1.86–7.12 and 2.38–14.78) when compared with ExAC and FLOSSIES controls, respectively (Fostira et al., 2020).

The challenge of *CHEK2* variant classification is reflected in numerous discrepancies in ClinVar classification (Decker et al., 2017), to the point of being recognized as the gene with more conflicting interpretations in HC diagnosis (Balmaña et al., 2016). Moreover, there is a current controversy about whether to use *CHEK2* missense variants at a clinical level. For instance, the National Comprehensive Cancer Network's BC management recommendations for *CHEK2* carriers only apply to carriers of truncating variants. In the same line, the UK Cancer Genetics Group decided not to take into account nontruncating variants in the clinical routine until a precise utility is stated for missense variants (Taylor et al., 2018).

Here, we present our effort to characterize the *CHEK2* mutational spectrum in Spanish HC patients, which has resulted in the need to consider refining ACMG-AMP guidelines for this gene.

2 | MATERIAL AND METHODS

2.1 | Patients and control cohort

A total of 2346 HC suspected patients were screened at two phases; first 516 cases were screened for c.1100delC, exon 9–10 deletion, and c.470T>C recurrent variants, and later 1848 HC patients and 194 healthy controls were analyzed by multigene panel testing (see Figure 1 and Supporting Information). Written informed consent was obtained from all patients, and the study protocol was approved by the Ethics Committee of IDIBELL (PR278/19).

2.2 | *CHEK2* variant annotation and collection of variant information

Variant annotation was performed using NM_007194.3 for the *CHEK2* gene (coding region and ±20bp of the intronic region). All variants identified were submitted to Alamut Software Suite v2.15.0 (Interactive Biosoftware) to retrieve population frequency and in silico prediction data. Variant classification in ClinVar, as well as literature review were collected.

2.3 | Criteria used to assess pathogenicity

2.3.1 | Very strong evidence of pathogenicity (PVS1) and PVS1_strong

Very strong evidence of pathogenicity (PVS1) and PVS1_strong were considered met according to Tayoun decision tree criteria (Abou Tayoun et al., 2018).

2.3.2 | Strong evidence of pathogenicity (PS criteria)

PS3 was weighted when a functional defect was found in at least two independent studies in the absence of discordant results. PS4

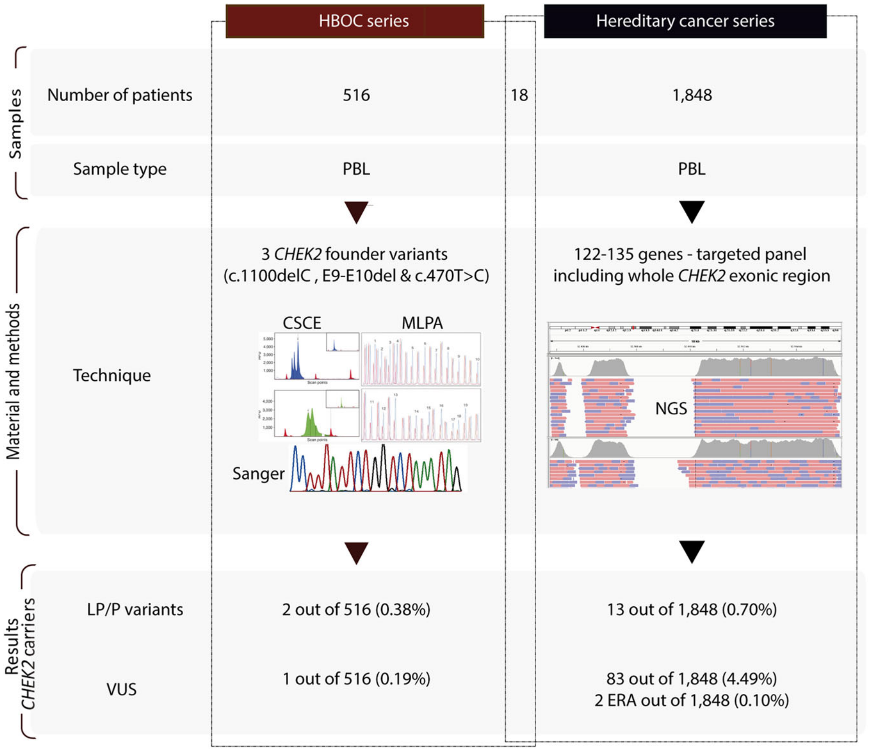


FIGURE 1 Diagram of the study. BC, breast cancer; CSCE, conformation-sensitive capillary electrophoresis; ERA, established risk allele; HBOC, hereditary breast and ovarian cancer; LP, likely pathogenic; MLPA, multiplex, ligation-dependent probe amplification; NGS, next-generation sequencing; P, pathogenic; PBL, peripheral blood lymphocytes; VUS, variant of unknown significance

was weighted for variants with an OR > 5.0 in case-control studies, PS4_moderate for low-moderate penetrant genes if the OR was between 1.5 and 5, with a *p* value < .01 as long as the phenotype was in accordance with the described for the gene.

2.3.3 | Moderate evidence of pathogenicity (PM criteria)

PM1, if the variant affected a highly conserved amino acid located in the FHA and/or kinase domain. PM2 was weighted when the variant was absent or in less than 1 out of 100,000 alleles in gnomAD v2.1.1 from “all” noncancer population data set; if present in ≥2 individuals within any subpopulation, it should be present in <1 out of 50,000 alleles in that subpopulation. Since some *CHEK2* variants in spite of being frequent in the population, the associated risk is significant; PM2_supporting was applied if the variant was present in ≤1 out of 20,000 alleles in the gnomAD v2.1.1 data set (Karczewski et al., 2019).

2.3.4 | Supporting evidence of pathogenicity (PP criteria)

PP3 was weighted if the in silico predictors suggested a splicing alteration (reduction of ≥20% in Alamut score) and/or protein function alteration according to the Varsome genome interpreter (Kopanos et al., 2018). The variant classification was performed using a different combination of rules according to classical ACMG-AMP guidelines (Richards et al., 2015), ClinGen-*TP53* suggested modifications to ACMG (https://www.clinicalgenome.org/affiliation/50013) and to ACMG-Bayesian modeling (Tavtigian et al., 2018) (Table 1).

Criteria for classification of benign and likely benign variants were applied following recommendations from ACMG-AMP guidelines (Richards et al., 2015).

Risk allele categorization was ascertained when possible, as previously described (Senol-Cosar et al., 2019) (Table S1). Accordingly, ERA classification was given to variants reported in multiple association studies or to those determined by robust meta-analysis; likely risk allele (LRA) was assigned if either the variant showed association in at least two independent studies, had been reported in a large study of high quality or in multiple studies with almost complete concordance.

3 | RESULTS

3.1 | Nature and distribution of variants and clinical classification

After *CHEK2* mutational analysis of 2346 cases with suspicion of HC and discarding benign variants, we identified 51 different variants. Sixteen of which corresponded to variants expected to produce a loss of function proteins or missense variants with conflicting interpretation in the literature (Table 1 and Figure 2, pedigrees in Figure 3). The remaining 35 variants were clearly variants of

TABLE 1 Clinico-pathological features of carriers of *CHEK2* variants with conflicting possible pathogenicity and an attempt to use ACMG-AMP guidelines

										Clinical variant classification			
Family ID	Proband gender	Tumor (age at diagnosis)	Series	cDNA variant	Expected amino acid change	Maximum GnomAD frequency	GnomAD v2.1.1 non-cancer (ALL). Gral. Freq. (*N* counts/ *N* alleles)	ACMG-AMP criteria[a]	OR	Standard ACMG-AMP	ClinGen-TP53 ACMG modifica-tions	ACMG-Bayesian	Risk allele-based[b]
Variants with PVS1 criteria: Frameshift, nonsense, canonical splice sites, initiation codon, single or multiexon deletions													
1	Male	KC (65;65); PrC (65)	HC	c.279G>A	p.(Trp93*)	0.012% AMR	0.000021117 (5/236,770)	PVS1 + PM2_supp[c]		VUS	LP	LP	NA
10	Female	BC (35)	HC	deletion exon 2	p.0?	0 out 20,000	0 (0/21,692)	PVS1_strong + PM2_supp		VUS	VUS	VUS	NA
2	Female	BC (35; 44)	HBOC	c.591delA	p.(Val198Phefs*7)	0.0098% AFR	0.000022370 (6/268,218)	PVS1 + PM2_supp		VUS	LP	LP	NA
3	Female	BC (48)	HC										
11	Female	BC (40)	HC	deletion exons 3 & 4	p.?	1 in 21476 (0.0046-%)	0.000046564 (1/21,476)	PVS1 + PM2_supp		VUS	LP	LP	NA
8	Male	CRC (44); KC (49)	HC	c.593-1G>T	p.?	0 out ~25100-0	0 (0/~251,000)	PVS1 + PM2		LP	LP	P	NA
4	Female	BC (55)	HC	c.715G>T	p.(Glu239*)	0.0029% AMR	0.000008444 (2/236,838)	PVS1 + PM2		LP	LP	P	NA
5	Female	BC (42)	HC										
9	Female	OC (22); CRC (25)	HC	c.792+2T>C	p.(Asp265Alafs*12)	0.0015% NFE	0.000007464 (2/267,922)	PVS1 + PM2		LP	LP	P	NA
6	Female	BC (48)	HC	c.1100delC	p.(Thr367Metfs*15)	0.26% NFE	0.002117640 (563/ 265,862)	PVS1 + PS3 + PS4_mod	2.89 (2.63–3.16) (Liang et al., 2018)	P	P	P	ERA
7	Female	BC (33)	HC	c.1368dupA	p.(Glu457Argfs*33)	0.0044% NFE	0.000016913 (4/236,504)	PVS1 + PS3 + PM2_supp		P	P	P	NA
12	Female	BC (42)	HBOC	whole gene deletion	p.0?	0 out 20,000	0 (0/20,000)	PVS1_SA		P	P	P	NA
Variants without PVS1 criteria: Missense variants classified as pathogenic or likely pathogenic or previously described as low-risk variants													
13	Female	BC (42)	HC	c.190G>A	p.(Glu64Lys)	0.03% NFE	0.000160283 (43/ 268,276)	PP3		VUS	VUS	VUS	NA
14	Female	OC (76); BC (81)	HC										
15	Female	PC (41)	HC										

(Continues)

TABLE 1 (Continued)

										Clinical variant classification			
Family ID	Proband gender	Tumor (age at diagnosis)	Series	cDNA variant	Expected amino acid change	Maximum GnomAD frequency	GnomAD v2.1.1 non-cancer (ALL). Gral. Freq. (*N* counts/ *N* alleles)	ACMG-AMP criteria[a]	OR	Standard ACMG-AMP	ClinGen-TP53 ACMG modifications	ACMG-Bayesian	Risk allele-based[b]
16	Female	BC (49)	HC										
17	Female	CRC (22)	HC	c.349A>G[d]	p.(Arg117Gly)	0.019% NFE	0.000111897 (30/ 268,104)	PS3 + PS4_mod + PM1 + PP3	2.26 (1.29–3.95), $p = .003$ (Southey et al., 2016)	LP	LP	LP	LRA
18	Female	BC (51)	HC										
19	Female	BC (49)	HC										
20	Female	BC (35;35)	HC	c.433C>T	p.(Arg145Trp)	0.0085% AMR	0.000041021 (11/ 268,158)	PS3 + PM1 + PP3		LP	LP	LP	NA
21	Male	TeC (25)	HC	c.470T>C	p.(Ile157Thr)	0.53% NFE	0.004878939 (1309/ 268,296)	PS4_mod + PP3	1.58 (1.42–1.75), $p < .00001$ (Han et al., 2013)	VUS	VUS	VUS	ERA
12	Female	BC (42)	HBOC	c.499G>A	p.(Gly167Arg)	0.0065% SAS	0.000016884 (4/236,906)	PS3 + PM1 + PM2_supp[c] + PP3		LP	LP	LP	NA
22	Female	BC (47)	HC	c.1427C>T	p.(Thr476Met)	0.05% NFE	0.000299455 (76/ 253,794)	PP3		VUS	VUS	VUS	NA
23	Female	BC (34)	HC										
24	Female	OC (65)	HC										

Abbreviations: ERA, established risk allele; LP, likely pathogenic; LRA, likely risk allele; NA, not enough information to be applied; P, pathogenic; VUS, variant of unknown significance.

[a] Criteria of pathogenicity very strong (PVS), strong (PS), moderate (PM), supporting (PP), taken from S. Richards et al. (2015).

[b] See Table S2.

[c] Greater frequency in a subpopulation.

[d] Variant was also found in 1 out of 194 healthy controls.

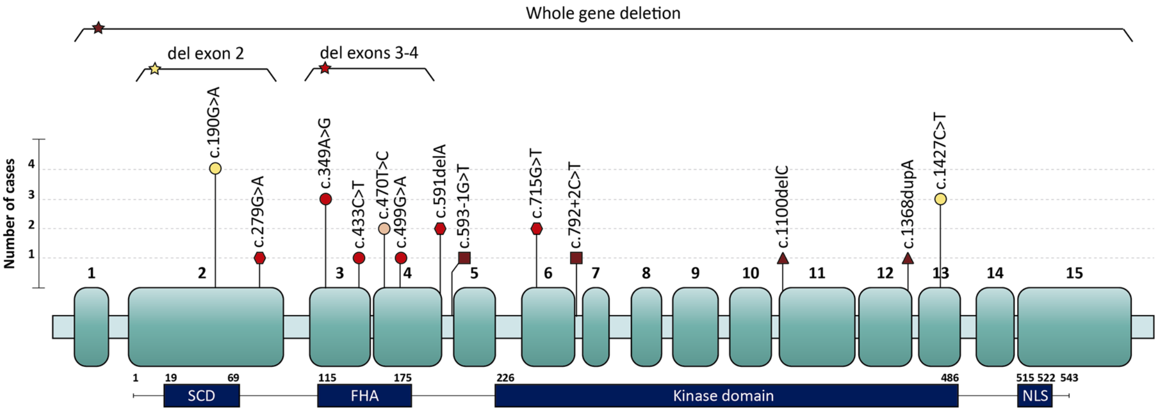


FIGURE 2 Schematic representation of *CHEK2* variants found in our cohort. Color code—dark red: pathogenic; red: likely pathogenic; pink: established risk allele; yellow: variant of uncertain significance. Shape code—diamond: nonsense variants; triangle: frameshift variants; square: splicing variants; circle: missense variants; star: copy number variants. Solid horizontal lines correspond to a copy number variant, each found in 1 index case

unknown significance (VUS) (Table S2). The control group carried one conflicting interpretation missense variant and one VUS (Tables 1 and S2).

To apply ACMG-AMP guidelines, we split them based on the presence or absence of PVS1 (criterion for a predicted loss of function variant; Table 2).

3.2 | Variants meeting PVS1 criterion

3.2.1 | Nonsense and frameshift variants

Only one patient was a carrier of the recurrent *CHEK2* c.1100delC mutation, p.(Thr367Metfs*15) (1 out of 2346, 0.04%). Given the great amount of data related to *CHEK2* c.1100delC, this variant meets PS3 (well-established functional studies) and PS4 (higher prevalence in affected individuals vs. controls), besides PVS1. However, PS4 was assigned with moderate strength (PS4_moderate), since OR > 5.0 for a moderately penetrant gene cannot be achieved. The combination of these rules classified this variant as pathogenic (P) in any combination of rules framework, and since it is well-studied and frequent in some populations, it was classified as an ERA within the Senol-Cosar framework (Tables 1 and Table S1). c.1368dupA, p.(Glu457Argfs*33) variant meets PS3 and PM2_supporting, being classified as P in all frameworks. c.715G>T, p.(Glu239*) variant meets PM2; therefore, it was classified as likely pathogenic (LP) using ACMG and ClinGen-*TP53* frameworks. According to Tavtigian's Bayes model (Tavtigian et al., 2018), it gathers enough evidence to be classified as P. Variants c.279G>A, p.(Trp93*) and c.591delA, p.(Val198-Phefs*7) were weighted PM2_supporting. The sum of PVS1 and a supporting criterion is not enough to classify a variant as LP/P using ACMG guidelines (Richards et al., 2015). However, the application of Bayesian modeling of this combination of rules gives a posterior probability of 0.988, resulting in its classification as LP, according to Tavtigian's (Richards et al., 2015; Tavtigian et al., 2018) as well as following ClinGen-*TP53* modifications (ClinGen-TP53_Expert_Panel, 2019).

3.2.2 | Canonical splice site variants

PVS1 was weighted for splicing variants predicted to produce an exon skipping with a subsequent frameshift. PM2 was weighted for c.593-1G>T and c.792+2T>C. Neither of them received PP3 to avoid redundancy with PVS1, remaining as LP according to ACMG and ClinGen-*TP53* frameworks. Notwithstanding, the combination of these rules in the Bayes model gives a posterior probability of 0.994, allowing its classification as P (Tavtigian et al., 2018). c.792+2T>C was reported in a previous study from our group (Feliubadaló et al., 2017), it produces partial retention of intron 6, decreasing the expression of wildtype. It is classified as LP by ClinVar.

3.2.3 | Copy number variants

The whole *CHEK2* deletion was weighted as PVS1 Stand-alone, as proposed for full gene deletions of known haploinsufficiency (Abou Tayoun et al., 2018), being classified P by all frameworks. The deletion of exons 3 and 4 occurs in-frame and produces the loss of the entire critical FHA domain; for this reason, PVS1 was weighted. Together with PM2_supporting, it would be a VUS with traditional ACMG combination rules but would be classified as LP following ClinGen-*TP53* as well as using Tavtigian's calculations. The deletion of exon 2 removes the first methionine and deletes 45 amino acids of the FHA domain, essential for CHK2 protein function; therefore, PVS1 was applied as “strong.” Together with PM2_supporting, it did not reach the LP/P classification in any framework.

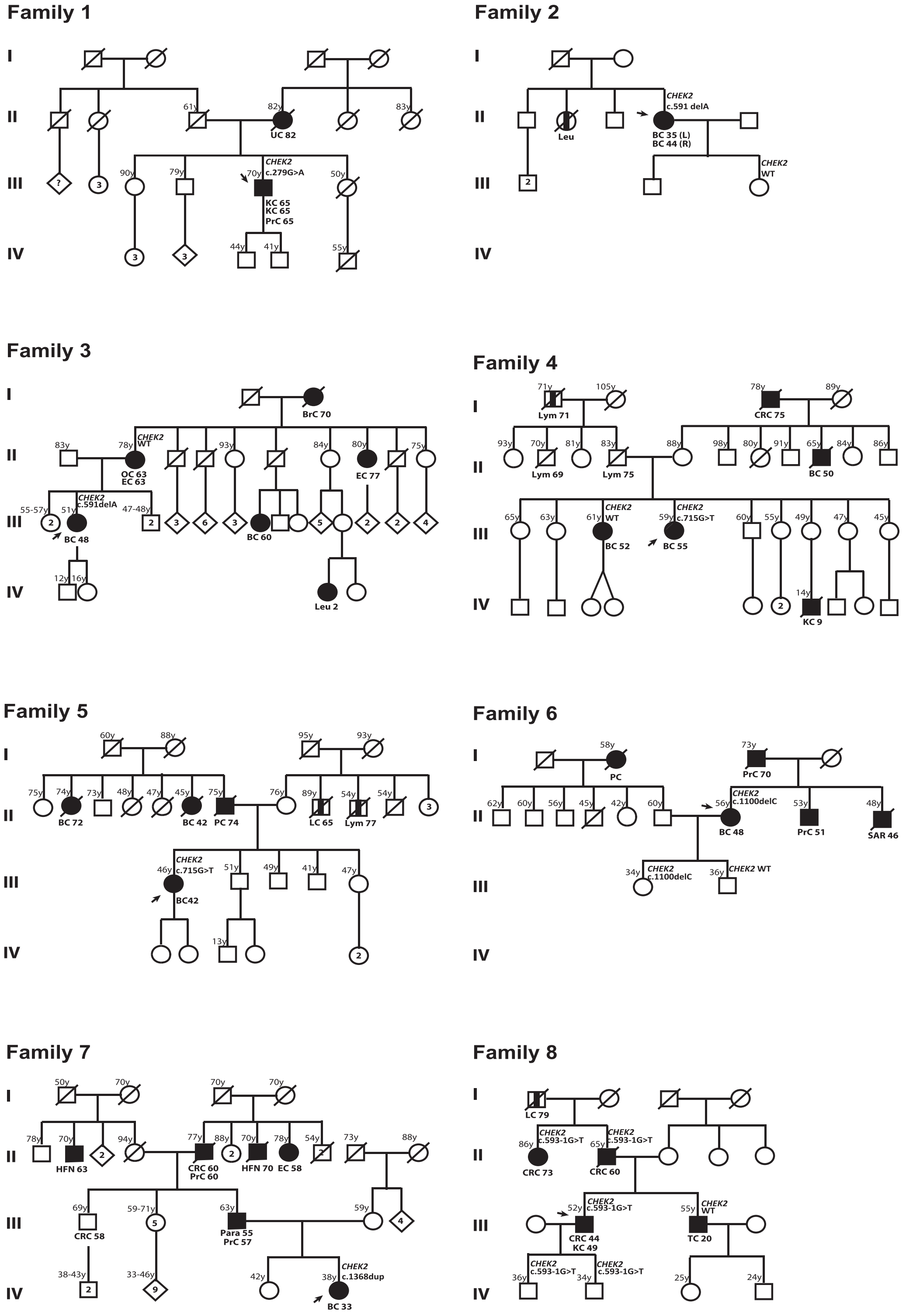


FIGURE 3 Pedigrees from families carrying 16 *CHEK2* variants discussed. Filled symbol, cancer confirmed by pathologist report; partially filled symbol, cancer referred by a relative; arrow, index case. Cosegregation results are indicated with the name of the variant if present and WT for noncarriers. Current ages and ages at death, when available, are indicated on the top-left corner of each individual's symbol. BC, breast cancer; BlC, bladder cancer; BrC, brain cancer; CRC, colorectal cancer; EC, endometrial cancer GC, gastric cancer; HFN, head/face/neck cancer; KC, kidney cancer; LC, lung cancer; Leu, leukemia; LiC, Liver cancer; Lym, Lymphoma; OC, ovarian cancer; PC, pancreas cancer; Para, parathyroid cancer; PCC, pheochromocytoma; PrC,prostate cancer; SC, skin cancer; SA, sebaceous adenoma; SAR, sarcoma; T, thyroid cancer; TeC, testicular cancer

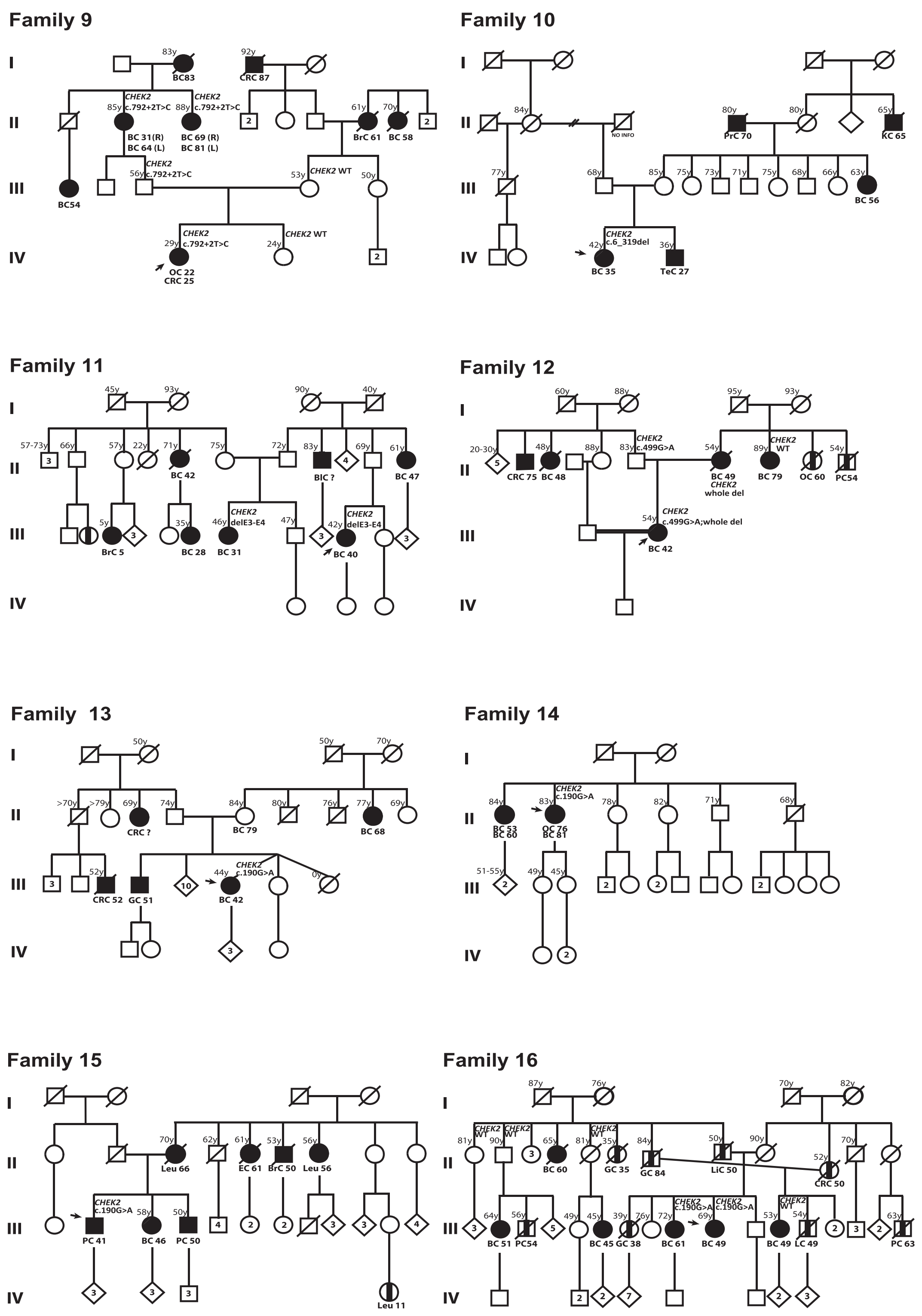


FIGURE 3 Continued

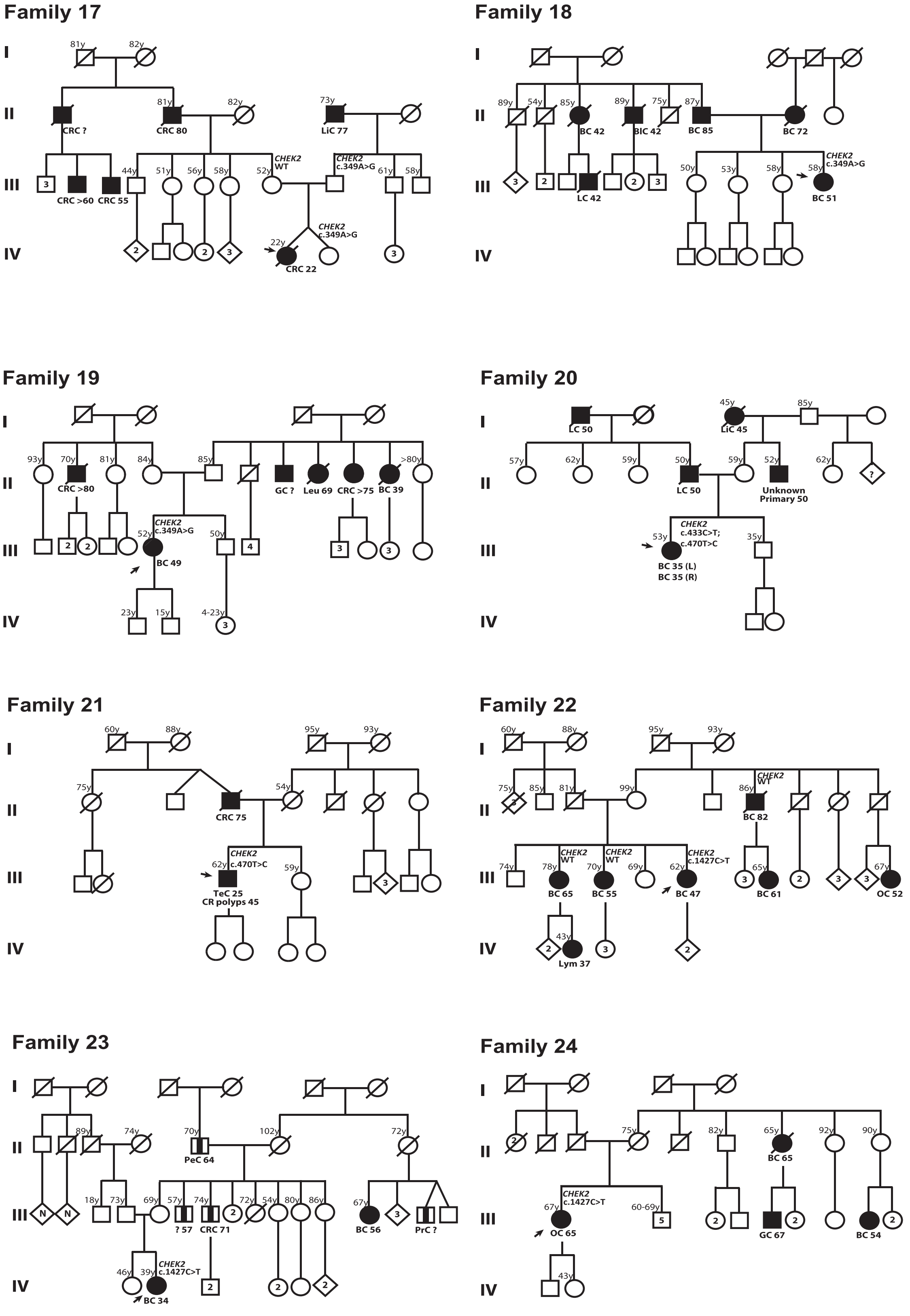


FIGURE 3 Continued

TABLE 2 Number of patients and other mutated genes by a clinical group of the hereditary cancer cohort

Clinical group	Number of patients (%)	Other mutated genes by the phenotypic cascade	Number of patients with *CHEK2* LP/P variants
Hereditary breast cancer	689 (37.3%)	68 LP/P (9.9%): 9 *ATM*, 18 *BRCA1*, 24 *BRCA2*, 3 *BRIP1*, 1 *MLH1*, 1 *MSH6*, 10 *PALB2*, 2 *RAD51C*	9 (1.3%)
Hereditary breast and ovarian cancer	217 (11.7%)	38 LP/P (17.5%): 2 *ATM*, 19 *BRCA1*, 10 *BRCA2*, 4 *BRIP1*, 1 *MLH1*, 1 *PALB2*, 1 *RAD51C*	1 (0.5%)
Hereditary ovarian cancer	248 (13.4%)	29 LP/P (11.7%): 8 *BRCA1*, 14 *BRCA2*, 3 *BRIP1*, 2 *MSH6*, 1 *RAD51C*, 1 *RAD51D*	0
Hereditary nonpolyposis colon cancer	302 (16.3%)	80 LP/P (26.5%): 18 *MLH1*, 19 *MSH2*, 41 *MSH6*, 2 *POLE*	3 (1%)
Familial (and attenuated) adenomatous polyposis	178 (9.6%)	17 LP/P (9.5%): 8 *APC*, 1 *BRCA2*, 7 *MUTYH* (biallelic), 1 *PTEN*	0
Li-Fraumeni suspected	22 (1.2%)	4 LP/P (18.2%): 1 *BRCA2*, 3 *TP53*	0
Other hereditary cancer conditions[a]	192 (10.5%)	34 LP/P (17.7%): 1 *ATM*, 1 *BRCA1*, 3 *BRCA2*, 1 *CDH1*, 4 *CDKN2A*, 3 *FH*, 8 *FLCN*, 2 *MLH1*, 2 *MSH6*, 3 *PTEN*, 1 *SDHD*, 1 *SMAD4*, 2 *STK11*, 2 *TSC2*	1 (0.5%)
Total	1848	270 (14.6%)	14 (0.8%)

Abbreviation: LP/P, (likely)pathogenic.

[a] Other hereditary conditions correspond to a group of patients with rare cancer syndromes, such as hereditary gastric cancer, hereditary melanoma, hereditary prostate cancer, Cowden syndrome, Peutz–Jeghers syndrome, tuberous sclerosis, Von Hippel–Lindau disease, multiple endocrine neoplasias, hereditary leiomyomatosis, renal cancer, juvenile polyposis, Birt–Hogg–Dubé syndrome among others.

3.3 | Variants not meeting PVS1

We found six missense variants with discordant classifications of pathogenicity in ClinVar (Table 1) in 13 unrelated patients. In addition, one of the healthy (noncancer) controls carried the *CHEK2* c.349A>G, p.(Arg117Gly) variant. To better interpret missense variants, a comprehensive review of previous functional studies was done, the main results are summarized in Tables S3 and S4. In a further effort to improve variant classification, after classical ACMG, we also followed the allele risk criteria reported recently (Senol-Cosar et al., 2019). For this, we searched for association studies of our *CHEK2* variants (Table S1).

CHEK2 c.190G>A, p.(Glu64Lys) is located in a weakly conserved amino acid in the SQ/TQ cluster domain (SCD). It is predicted deleterious by in silico analysis. It shows a partially reduced phosphorylation by ATM at the Thr68 residue, as well as partially reduced autophosphorylation and Cdc25C phosphorylation. It affects KAP1 phosphorylation and has discrepant results about DNA damage response (Table S3). Furthermore, there are no high-quality case–control studies. Therefore, this variant only meets the PP3 criterion, remaining as VUS (Table 1). Variant c.349A>G, p.(Arg117Gly) affects a highly conserved amino acid (class C65 according to GVGD) in the FHA domain. It is predicted deleterious by *in silico* analysis. It does not affect phosphorylation by ATM nor oligomerization but affects all the rest of the studied protein functions (Table S3). This variant accomplished PS3, PS4_moderate, PM1, and PP3 criteria, being classified as LP by all frameworks. It has been studied in a large high-quality case–control study, reporting a BC OR of 2.26 (95% CI, 1.29–3.95) (Table 1); therefore, it could be considered as LRA within the Senol-Cosar framework (Table S2). Variant c.433C>T, p.(Arg145Trp) is located in a moderately conserved amino acid of the FHA domain. It is predicted deleterious by in silico. It reduces CHK2 expression and stability. In functional assays, it has been consistently reported to impair kinase and DNA repair activity. Evidence for classification includes PS3, PM1, and PP3, being classified as LP by all frameworks. Variant c.470T>C, p.(Ile157Thr) lies in a weakly conserved amino acid of the FHA domain. It is predicted deleterious by in silico analysis. It has been widely studied, nevertheless, the functional assays reported to date show discordant results (Table S3). The reported OR in the biggest *CHEK2* meta-analysis was 1.58 (95% CI, 1.42–1.75); therefore, PS4_moderate was applied, but the application of PP3 was not enough to classify this variant as LP/P. However, following recommendations from Senol-Cosar et al. (2019), it would be an ERA due to the existence of multiple case-control studies. Variant c.499G>A, p.(Gly167Arg) is located in a highly conserved amino acid of the FHA domain. It is predicted deleterious by in silico analysis. Although there are only two functional studies, they both reported an impaired DNA repair activity in yeast assays (Table S3). PS3, PM1, PM2_supporting, and PP3 were assigned, being classified as LP by all frameworks. Variant c.1427C>T, p.(Thr476Met) lies in a moderately conserved amino acid of the kinase domain. It is predicted deleterious by in silico analysis. Functional assessment of KAP1 phosphorylation results in deleterious in vitro and likely benign in vivo. Furthermore, SOX phosphorylation was reported equal to that of the pathogenic c.1100delC variant. Assays on DNA repair activity have found it damaging or with intermediate activity (Table S3). Due to these discordant functional assay results, PS3 was not weighted. Classification remained as VUS since c.1427C>T only accomplished PP3.

3.4 | Variants of unknown significance

Thirty-five unique VUS (with less than 2 LP/P interpretations in ClinVar) were encountered in our cases (Table S1). We aimed to perform RNA analysis in three of these, due to in silico prediction results (c.320-5T>A [NC_000022.10:g.29121360A>T] and c.1376-8T>C [NC_000022.10:g.29090113A>G]) or to the nature of the variant (duplication of exons 3 and 4). Lymphocytes for RNA analysis were available from one carrier of the duplication of exons 3 and 4, for several samples with c.320-5T>A and were unattainable from c.1376-8T>C carriers. RNA analysis showed that the duplication of exons 3 and 4 occurs in tandem and produces ~30% of the aberrant transcript containing an in-frame insertion of 273 bp (Figure 4). This affects the region that codifies for the FHA domain; unfortunately, there were no polymorphisms in the region to perform quantitative analysis. This variant remains as VUS following all guidelines. Regarding c.320-5T>A variant, in silico programs, predicted a reduction in recognition of the splicing acceptor site of exon 3. cDNA analysis in two carriers showed the generation of an aberrant transcript, consisting in an in-frame deletion of exons 3 and 4 (Figure 5), as previously reported (Kraus et al., 2017). The amount of abnormal transcript seemed greater than 20%, although the absence of exonic polymorphisms prevented an accurate quantification. Of note, the frequency of c.320-5T>A is 0.12% in gnomAD (NFE) and of 1.35% (25 out of 1848) in our HC cohort. To understand the differences in frequency in our population with relation to international databases, we screened 1501 control samples (see Supporting Material). *CHEK2* c.320-5T>A had a frequency of 0.8% (12 out of 1501) in our controls, not a statistically significant difference, preventing it from being considered as a risk allele.

3.5 | *CHEK2* variants in the different HC groups

Applying the Bayesian combination of rules by clinical suspicion subgroups of the HC cohort, *CHEK2* LP/P variants were identified in 1.3% of HBC cases ($n = 9$), in 0.5% HBOC cases ($n = 1$), 1% of the hereditary nonpolyposis CRC patients (HNPCC, $n = 3$), and in one patient from the minority cancer group (0.5%), who had two kidney tumors, pheochromocytoma, and prostate cancer.

Among the 10 families with HBC/HBOC, 2 proband females had two variants in *CHEK2*. One female, with BC at 42, was a compound heterozygous of a whole *CHEK2* deletion and variant c.499G>A. The other patient with bilateral BC at 35 carried two *CHEK2* missense variants (c.433C>T and c.470T>C) *in trans*. Both cases were previously reported by our group (Stradella et al., 2018). In addition, a third proband diagnosed with BC at age 49 carried the *CHEK2* c.349A>G and a pathogenic variant in *ERCC3*. Interestingly, the three HNPCC patients with *CHEK2* LP/P variants developed CRC at a young age (22, 25, and 44), and their tumors were MMR proficient.

4 | DISCUSSION

We have made an effort to classify variants in the low-moderate penetrance *CHEK2* gene. For that, we analyzed the whole coding region of *CHEK2* in a large HC cohort, performed an in-depth literature review and have defined specific cutoffs for ACMG criteria to allow classification of variants with low effect. Furthermore, we applied different combinatorial rules that enabled us to compare

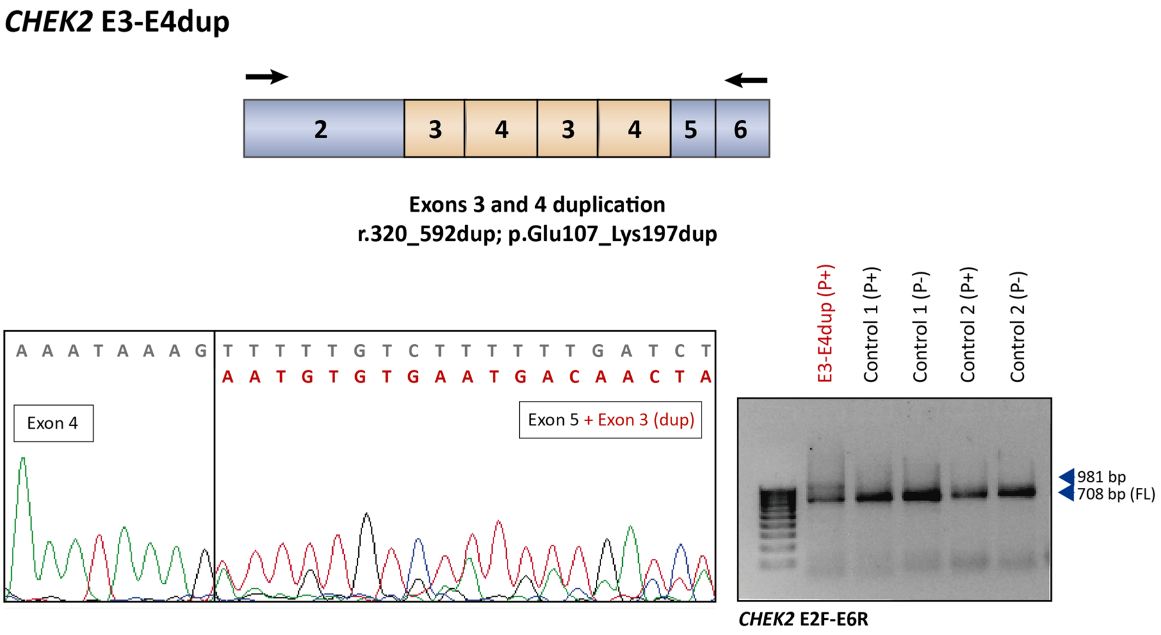

CHEK2 E3-E4dup
2
3
4
3
4
5
6
Exons 3 and 4 duplication
r.320_592dup; p.Glu107_Lys197dup
A A A T A A A G T T T T T G T C T T T T T T G A T C T
A A T G T G T G A A T G A C A A C T A
Exon 4
Exon 5 + Exon 3 (dup)
E3-E4dup (P+)
Control 1 (P+)
Control 1 (P-)
Control 2 (P+)
Control 2 (P-)
981 bp
708 bp (FL)
CHEK2 E2F-E6R

FIGURE 4 mRNA analysis of *CHEK2* E3-E4dup. Top, schematic representation of *CHEK2* E3-E4dup. cDNA amplification showed a double band, one corresponding to the full-length transcript (708 bp) and the other to the transcript carrying the duplication (981 bp), as shown in the electropherogram on the bottom left. Bottom right, agarose gel of a carrier and two controls with (P+) and without puromycin (P−)

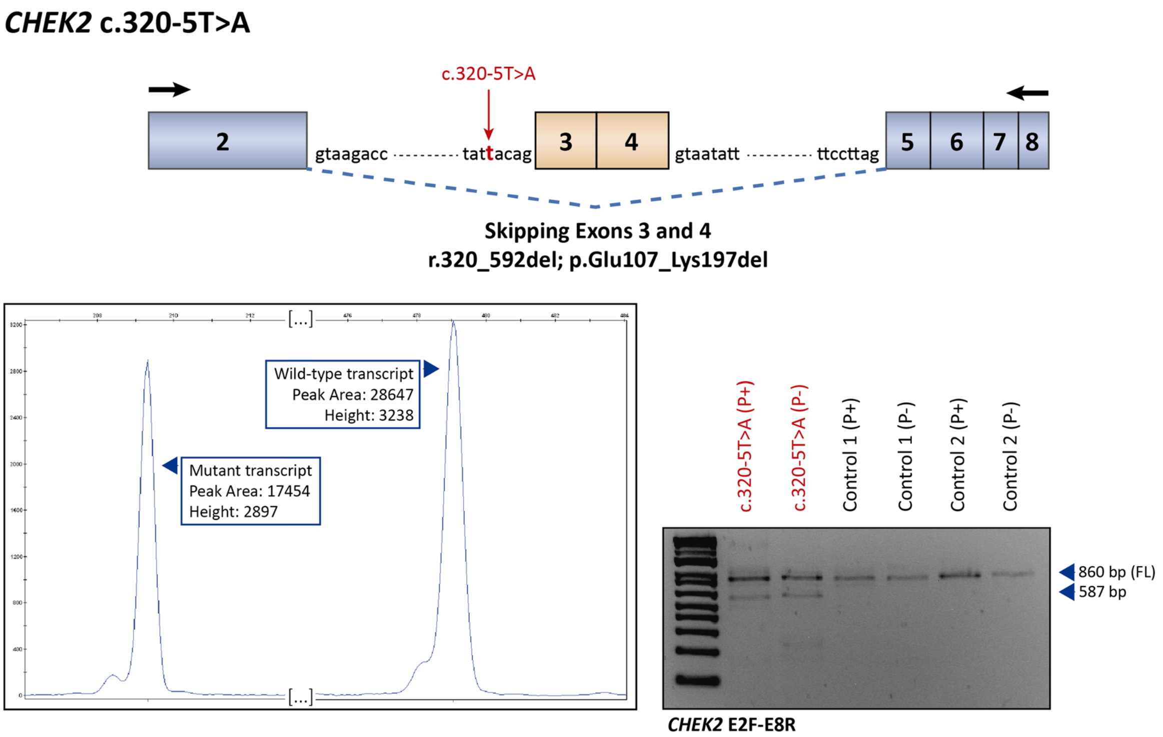


FIGURE 5 mRNA analysis of *CHEK2* c.320-5T>A. Top, schematic representation of *CHEK2* c.320-5T>A (NC_000022.10:g.29121360A>T) splicing effect. cDNA amplification showed a double band, one corresponding to the full-length transcript (860 bp) and the other to the transcript lacking exons 3 and 4 (587 bp), as shown in the electropherogram on the bottom left. Bottom right, agarose gel of a carrier and two controls with (P+) and without puromycin (P−)

classification rates—concluding that the Bayesian model is the most optimal framework to classify variants to a greater extent.

From our experience in variant classification and after a comprehensive literature review, we propose two adaptations of the ACMG criteria. Regarding PS4, we propose to score PS4_moderate for low-moderate penetrant genes if an OR is given between 1.5 and 5, with a *p* value of < .01, when the phenotype is in accordance with the previously described. In relation to PM2 evidence, in our laboratory, we use an extremely conservative approach and assign PM2 only if the variant is absent or present in less than 1 out of 100,000 alleles in gnomAD (0.001% of maximum frequency) for high penetrant genes. However, we propose to assign PM2_supporting when the variant is ≤1 out of 20,000 alleles.

Variants meeting the PVS1 criterion tend to be easier to classify as LP/P. For instance, the founder mutation c.1100delC is the most studied *CHEK2* mutation, and it has a prevalence of 0.26% in the NFE population. *CHEK2* c.1100delC has a moderate penetrance (Meijers-Heijboer et al., 2002; Oldenburg et al., 2003), conferring an increased BC risk for the overall population (OR = 2.89, 95% CI, 2.63–3.16) (Liang et al., 2018) and for carriers with familial BC (OR = 3.21, 95% CI, 2.41-4.29) (Liang et al., 2018). It has been reported absent in the Spanish population (Bellosillo et al., 2005), or with frequencies of 0.93% in the Basque population, 0.36% in the Galician population, and 0.3% in a study of *BRCA*-negative HBC Basque and Catalan families (Fachal, Santamariña, Blanco, Carracedo, & Vega, 2013; Gutiérrez-Enríquez, Balmaña, Baiget, & Díez, 2008; Martínez-Bouzas et al., 2007). In our larger cohort, only one case was identified (0.08%, 1 out of 1251 BC affected cases), confirming its low prevalence in our population. Moreover, in a recent study analyzing 15 truncating *CHEK2* variants in 213 patients and 29 control carriers, the BC risk OR was 3.11 (95% CI, 2.15–4.69) (Decker et al., 2017). Here, we identified ten proband carriers of truncating variants, eight of which developed the first tumor before the age of 50, consistent with previous findings of early cancer development in carriers of truncated variants (Decker et al., 2017; Han et al., 2013). Nonetheless, the median age at first cancer diagnosis in our study was not very different amongst carriers of truncating and missense LP/P variants, being 42 (range, 25–65) and 40 (range, 22–51) years, respectively. Bilateral BC has been mainly reported in c.1100delC carriers (M. Kriege et al., 2014), and truncating variants in this gene have been associated to other nonbreast second primary tumor diagnosis in a study using multigene panel testing (Fostira et al., 2020). In our cohort, four cases with two or multiple cancers were carriers of truncating variants, and only one was a carrier of an LP missense, confirming a higher aggressiveness of truncating variants over missense variants.

Conflicting results are common for missense hypomorphic variants and represent one of the biggest challenges we faced for *CHEK2* variant classification due to the lack of more sensitive functional assays and the use of different controls, complicating replication and therefore bypassing PS3 application. The c.470T>C founder mutation conveys a moderate susceptibility for overall cancer (OR = 1.39; $p < .00001$) and for BC only (OR = 1.58; $p < .00001$) in a large meta-analysis (Han et al., 2013). Its pathogenicity has been established for ovary cystadenomas in young Polish carriers (OR = 2.6; $p = .006$) (Szymanska-Pasternak et al., 2006) and is associated with a twofold risk of non-Hodgkin lymphoma, colon, kidney, thyroid, and prostate cancers (Cybulski et al., 2004). We found it in a male patient diagnosed with testicular cancer at 25 years. Interestingly, in a recent study of 448 Croatian testicular cancer patients, it was found in 5.1% of them, resulting in an OR of 3.93 (95% CI, 1.53–9.95) even when its population frequency is of 1–2% (AlDubayan et al., 2019). Of note, c.470T>C remains as VUS even applying PS4_moderate. To our knowledge, c.470T>C is the most studied *CHEK2* missense variant, but as shown in Table S3 it has conflictive interpretations of pathogenicity at almost all functional studies; therefore, PS3 was ruled out, remaining as VUS in the ACMG context. However, we were able to classify it as ERA according to the risk allele-based classification (Senol-Cosar et al., 2019). Of note, this variant is classified as LP by GeneDx, and as P by Ambry, Color and Invitae diagnostic laboratories (Table S3), which could convey errors in clinical interpretation. PS3 was also not possible to apply for two other missense variants: c.190G>A and c.1427C>T. *CHEK2* c.190G>A is a fairly frequent variant found in 0.03% of NFE by gnomAD, with partial reduction of Thr68 phosphorylation, autophosphorylation, and Cdc25C phosphorylation, but DNA repair assays in yeast are discordant (Table S3). Variant c.1427C>T is another relatively frequent variant present in 0.05% of NFE (gnomAD). It has been reported to affect DNA damage response in yeast at the intermediate-high level. In addition, it shows reduced SOX phosphorylation almost equally to c.1100delC. However, in vivo and in vitro studies of KAP1 phosphorylation from the same group showed discordant results of pathogenicity (Table S3). As noted in Table S2, lack of robust association studies and meta-analysis of these variants hampered the possibility of applying risk allele-based classification. Both remained as VUS in any classification framework, in spite of being classified as LP by at least two different reputable sources (Table S3).

To summarize, we describe here a comprehensive *CHEK2* mutational analysis in a large Spanish cohort of HC patients, providing full data of the actual prevalence of *CHEK2* pathogenic variants in our population. The frequency of LP/P variants in the HBC suspected cases in the whole gene analysis was 1.3% (9 out of 689), similar to the reported by Couch et al. (2017) in a study of 58,798 BC patients, in which they found 1.41% of truncating variants and 2.22% of LP/P *CHEK2* missense variants. Interestingly, three young CRC cases carried an LP/P *CHEK2* variant, and none of them had any additional pathogenic variant in our NGS panel analysis, although two of them have a nonpenetrant carrier father above the age of 50. *CHEK2* c.1100delC was reported in 6 out of 234 HNPCC families from Poland (Meijers-Heijboer et al., 2003). In their study, three of them also carried germline MMR P variants. In addition, c.470T>C has been found in familial CRC (Cybulski, Wokołorczyk, et al., 2007; Kilpivaara, Alhopuro, Vahteristo, Aaltonen, & Nevanlinna, 2006) and have been described to increase the risk of CRC among MMR-negative, HNPCC/HNPCC-related families in Poland (Suchy et al., 2010). To our knowledge, this is the largest Spanish data set presenting the sequencing of the whole *CHEK2* coding region together with the first attempt to apply ACMG-AMP guidelines for this gene. We detailed different strategies that can be helpful to classify VUS using different frameworks with the aim of being of help not only for the curation of *CHEK2* variants but also for other genes. We hope our work serves as a starting point to better tune ACMG criteria in the case of low-penetrance and low effect size variants associated with disease risk.

ACKNOWLEDGMENTS

We thank CERCA Program/Generalitat de Catalunya for their institutional support. We also wish to thank all the members of the ICO Hereditary Cancer Program. This study was supported by the Carlos III National Health Institute funded by FEDER funds—a way to build Europe—[PI19/00553; PI16/00563 and CIBERONC]; the Government of Catalonia [Pla estratègic de recerca i innovació en salut (PERIS_MedPerCan and URDCat projects), 2017SGR1282 and 2017SGR496].

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Conceptualization and design: Conxi Lázaro, Gardenia Vargas-Parra, Jesús del Valle, Mireia Gausachs. *Data curation*: Paula Rofes, Mireia Gausachs, Agostina Stradella, Angela Velasco, Eva Tornero, Mireia Menéndez, Xavier Muñoz, Silvia Iglesias, Adriana López-Doriga, Daniel Azuara, Olga Campos, Raquel Cuesta, Esther Darder, Alex Teulé, Matilde Navarro. *Formal analysis and interpretation of data*: Gardenia Vargas-Parra, Jesús del Valle, Marta Pineda, Lídia Feliubadaló, Moreno-Cabrera, Rafael de Cid, Conxi Lázaro. *Funding acquisition*: Conxi Lázaro, Joan Brunet, Gabriel Capellá, Marta Pineda, Lídia Feliubadaló. *Investigation*: Gardenia Vargas-Parra, Jesús del Valle, Mireia Gausachs, Paula Rofes, Agostina Stradella. *Methodology*: Gardenia Vargas-Parra, Jesús del Valle, Mireia Gausachs, Conxi Lázaro. *Project administration*: Conxi Lázaro. *Resources*: Rafael de Cid, Sara González. *Software*: Moreno-Cabrera, Adriana López-Doriga. *Supervision*: Conxi Lázaro. Validation: Marta Pineda, Lídia Feliubadaló. *Visualization*: Gardenia Vargas-Parra, Jesús del Valle, Paula Rofes. *Drafting of the manuscript*: Gardenia Vargas-Parra, Jesús del Valle, Paula Rofes, Conxi Lázaro. *Critical revision*: Conxi Lázaro, Joan Brunet, Gabriel Capellá, Marta Pineda, Lídia Feliubadaló.

DATA AVAILABILITY STATEMENT

The data files that support the findings of this study are available upon request.

ORCID

Gardenia Vargas-Parra http://orcid.org/0000-0003-1378-736X

REFERENCES

Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., Biesecker, L. G., … ClinGen Sequence Variant Interpretation

Working Group (ClinGen SVI). (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human Mutation*, *39*(11), 1517–1524. https://doi.org/10.1002/humu.23626

AlDubayan, S. H., Pyle, L. C., Gamulin, M., Kulis, T., Moore, N. D., Taylor-Weiner, A., …Lessel, D. (2019). Association of inherited pathogenic variants in checkpoint kinase 2 (CHEK2) with susceptibility to testicular germ cell tumors. *JAMA Oncology*, *5*, 514. https://doi.org/10.1001/jamaoncol.2018.6477

Balmaña, J., Digiovanni, L., Gaddam, P., Walsh, M. F., Joseph, V., Stadler, Z. K., …Domchek, S. M. (2016). Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing. *Journal of Clinical Oncology*, *34*(34), 4071–4078. https://doi.org/10.1200/JCO.2016.68.4316

Bartek, J., Falck, J., & Lukas, J. (2001). CHK2 kinase—a busy messenger. *Nature Reviews Molecular Cell Biology*, *2*(12), 877–886.

Bell, D. W., Varley, J. M., Szydlo, T. E., Kang, D. H., Wahrer, D. C., Shannon, K. E., …Haber, D. A. (1999). Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science*, *286*(5449), 2528–2531.

Bellosillo, B., Tusquets, I., Longarón, R., Pérez-Lezaun, A., Bellet, M., Fabregat, X., …Solé, F. (2005). Absence of CHEK2 mutations in Spanish families with hereditary breast cancer. *Cancer Genetics and Cytogenetics*, *161*(1), 93–95. https://doi.org/10.1016/j.cancergencyto.2005.01.016

ClinGen-TP53_Expert_Panel. (2019, August 6, 2020). *TP53 Rule Specifications for the ACMG/AMP Variant Curation Guidelines*. Retrieved from https://www.clinicalgenome.org/affiliation/50013

Couch, F. J., Shimelis, H., Hu, C., Hart, S. N., Polley, E. C., Na, J., … Dolinsky, J. S. (2017). Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncology*, *3*(9), 1190–1196. https://doi.org/10.1001/jamaoncol.2017.0424

Cybulski, C., Gorski, B., Huzarski, T., Masojc, B., Mierzejewski, M., Debniak, T., … Lubinski, J. (2004). CHEK2 is a multiorgan cancer susceptibility gene. *American Journal of Human Genetics*, *75*(6), 1131–1135.

Cybulski, C., Wokolorczyk, D., Huzarski, T., Byrski, T., Gronwald, J., Gorski, B., … Lubinski, J. (2007). A deletion in CHEK2 of 5,395 bp predisposes to breast cancer in Poland. *Breast Cancer Research and Treatment*, *102*(1), 119–122.

Cybulski, C., Wokołorczyk, D., Kładny, J., Kurzawski, G., Kurzwaski, G., Suchy, J., … Lubiński, J. (2007). Germline CHEK2 mutations and colorectal cancer risk: Different effects of a missense and truncating mutations? *European Journal of Human Genetics*, *15*(2), 237–241. https://doi.org/10.1038/sj.ejhg.5201734

Decker, B., Allen, J., Luccarini, C., Pooley, K. A., Shah, M., Bolla, M. K., … Easton, D. F. (2017). Rare, protein-truncating variants in. *Journal of Medical Genetics*, *54*(11), 732–741. https://doi.org/10.1136/jmedgenet-2017-104588

Easton, D. F., Pharoah, P. D. P., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., … Foulkes, W. D. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *The New England Journal of Medicine*, *372*(23), 2243–2257. https://doi.org/10.1056/NEJMsr1501341

Fachal, L., Santamariña, M., Blanco, A., Carracedo, A., & Vega, A. (2013). CHEK2 c.1100delC mutation among non-BRCA1/2 Spanish hereditary breast cancer families. *Clinical and Translational Oncology*, *15*(2), 164–165. https://doi.org/10.1007/s12094-012-0967-z

Feliubadaló, L., Tonda, R., Gausachs, M., Trotta, J. R., Castellanos, E., López-Doriga, A., … Lázaro, C. (2017). Benchmarking of whole exome sequencing and ad hoc designed panels for genetic testing of hereditary cancer. *Scientific Reports*, *7*, 37984. https://doi.org/10.1038/srep37984

Fostira, F., Kostantopoulou, I., Apostolou, P., Papamentzelopoulou, M. S., Papadimitriou, C., Faliakou, E., … Yannoukakos, D. (2020). One in three highly selected Greek patients with breast cancer carries a loss-of-function variant in a cancer susceptibility gene. *Journal of Medical Genetics*, *57*(1), 53–61. https://doi.org/10.1136/jmedgenet-2019-106189

Gutiérrez-Enríquez, S., Balmaña, J., Baiget, M., & Díez, O. (2008). Detection of the CHEK2 1100delC mutation by MLPA BRCA1/2 analysis: A worthwhile strategy for its clinical applicability in 1100delC low-frequency populations? *Breast Cancer Research and Treatment*, *107*(3), 455–457. https://doi.org/10.1007/s10549-007-9555-2

Han, F. F., Guo, C. L., & Liu, L. H. (2013). The effect of CHEK2 variant I157T on cancer susceptibility: evidence from a meta-analysis. *DNA and Cell Biology*, *32*(6), 329–335. https://doi.org/10.1089/dna.2013.1970

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210. https://doi.org/10.1101/531210

Katona, B. W., Yurgelun, M. B., Garber, J. E., Offit, K., Domchek, S. M., Robson, M. E., & Stadler, Z. K. (2018). A counseling framework for moderate-penetrance colorectal cancer susceptibility genes. *Genetics in Medicine*, *20*(11), 1324–1327. https://doi.org/10.1038/gim.2018.12

Kilpivaara, O., Alhopuro, P., Vahteristo, P., Aaltonen, L. A., & Nevanlinna, H. (2006). CHEK2 I157T associates with familial and sporadic colorectal cancer. *Journal of Medical Genetics*, *43*(7), e34. https://doi.org/10.1136/jmg.2005.038331

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2018). VarSome: The human genomic variant search engine. *Bioinformatics*, *35*(11), 1978–1980. https://doi.org/10.1093/bioinformatics/bty897

Kraus, C., Hoyer, J., Vasileiou, G., Wunderle, M., Lux, M. P., Fasching, P. A., … Reis, A. (2017). Gene panel sequencing in familial breast/ovarian cancer patients identifies multiple novel mutations also in genes others than BRCA1/2. *International Journal of Cancer*, *140*(1), 95–102. https://doi.org/10.1002/ijc.30428

Liang, M., Zhang, Y., Sun, C., Rizeq, F. K., Min, M., Shi, T., & Sun, Y. (2018). Association Between CHEK2*1100delC and Breast Cancer: A Systematic Review and Meta-Analysis. *Molecular Diagnosis & Therapy*, *22*, 397–407. https://doi.org/10.1007/s40291-018-0344-x

Kriege, M., Hollestelle, A., Jager, A., Huijts, P. E. A., Berns, E. M., Sieuwerts, A. M., … Seynaeve, C. (2014). Survival and contralateral breast cancer in CHEK2 1100delC breast cancer patients: Impact of adjuvant chemotherapy. *British Journal of Cancer*, *111*, 1004–1013. & . https://doi.org/10.1038/bjc.2014.306

Martínez-Bouzas, C., Beristain, E., Guerra, I., Gorostiaga, J., Mendizabal, J. L., De-Pablo, J. L., … Tejada, M. I. (2007). CHEK2 1100delC is present in familial breast cancer cases of the Basque Country. *Breast Cancer Research and Treatment*, *103*(1), 111–113. https://doi.org/10.1007/s10549-006-9351-4

Matsuoka, S., Rotman, G., Ogawa, A., Shiloh, Y., Tamai, K., & Elledge, S. J. (2000). Ataxia telangiectasia-mutated phosphorylates Chk2 in vivo and in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(19), 10389–10394.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., … Stratton, M. R. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature Genetics*, *31*(1), 55–59.

Meijers-Heijboer, H., Wijnen, J., Vasen, H., Wasielewski, M., Wagner, A., Hollestelle, A., … Schutte, M. (2003). The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. *American Journal of Human Genetics*, *72*(5), 1308–1314.

Oldenburg, R. A., Kroeze-Jansema, K., Kraan, J., Morreau, H., Klijn, J. G., Hoogerbrugge, N., … Devilee, P. (2003). The CHEK2*1100delC variant acts as a breast cancer risk modifier in non-BRCA1/BRCA2 multiple-case families. *Cancer Research*, *63*(23), 8153–8157.

Plon, S. E., Cooper, H. P., Parks, B., Dhar, S. U., Kelly, P. A., Weinberg, A. D., … Hilsenbeck, S. (2008). Genetic testing and cancer risk management recommendations by physicians for at-risk relatives. *Genetics in Medicine*, *13*(2), 148–154.

Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., … Ward, B. E. (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine*, *10*(4), 294–300. https://doi.org/10.1097/GIM.0b013e31816b5cae

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Schmidt, M. K., Hogervorst, F., van Hien, R., Cornelissen, S., Broeks, A., Adank, M. A., … Easton, D. F. (2016). Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. *Journal of Clinical Oncology*, *34*(23), 2750–2760. https://doi.org/10.1200/JCO.2016.66.5844

Senol-Cosar, O., Schmidt, R. J., Qian, E., Hoskinson, D., Mason-Suares, H., Funke, B., & Lebo, M. S. (2019). Considerations for clinical curation, classification, and reporting of low-penetrance and low effect size variants associated with disease risk. *Genetics in Medicine*, *21*(12), 2765–2773. https://doi.org/10.1038/s41436-019-0560-8

Southey, M. C., Goldgar, D. E., Winqvist, R., Pylkäs, K., Couch, F., Tischkowitz, M., …& Yannoukakos, D. (2016). PALB2, CHEK2 and ATM rare variants and cancer risk: Data from COGS. *Journal of Medical Genetics*, *53*(12), 800–811.

Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N. A., Nathanson, K. L., … Enigma (2012). ENIGMA—Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human Mutation*, *33*(1), 2–7. https://doi.org/10.1002/humu.21628

Stradella, A., Del Valle, J., Rofes, P., Feliubadaló, L., Grau Garces, È., Velasco, À., … Lázaro, C. (2018). Does multilocus inherited neoplasia alleles syndrome have severe clinical expression? *Journal of Medical Genetics*. https://doi.org/10.1136/jmedgenet-2018-105700

Suchy, J., Cybulski, C., Wokołorczyk, D., Oszurek, O., Górski, B., Debniak, T., … Lubiński, J. (2010). CHEK2 mutations and HNPCC-related colorectal cancer. *International Journal of Cancer*, *126*(12), 3005–3009. https://doi.org/10.1002/ijc.25003

Szymanska-Pasternak, J., Szymanska, A., Medrek, K., Imyanitov, E. N., Cybulski, C., Gorski, B., … Lubinski, J. (2006). CHEK2 variants predispose to benign, borderline and low-grade invasive ovarian tumors. *Gynecologic Oncology*, *102*(3), 429–431. https://doi.org/10.1016/j.ygyno.2006.05.040

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., … ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine*, *20*(9), 1054–1060. https://doi.org/10.1038/gim.2017.210

Taylor, A., Brady, A. F., Frayling, I. M., Hanson, H., Tischkowitz, M., & Turnbull, C., … UK Cancer Genetics Group (UK-CGG). (2018). Consensus for genes to be included on cancer panel tests offered by UK genetics services: Guidelines of the UK Cancer Genetics Group. *Journal of Medical Genetics*, *55*(6), 372–377. https://doi.org/10.1136/jmedgenet-2017-105188

Wu, X., Webster, S. R., & Chen, J. (2001). Characterization of tumor-associated Chk2 mutations. *Journal of Biological Chemistry*, *276*(4), 2971–2974.

Zannini, L., Delia, D., & Buscemi, G. (2014). CHK2 kinase in the DNA damage response and beyond. *Journal of Molecular Cell Biology*, *6*(6), 442–457. https://doi.org/10.1093/jmcb/mju045

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Vargas-Parra G, del Valle J, Rofes P, et al. Comprehensive analysis and ACMG-based classification of *CHEK2* variants in hereditary cancer patients. *Human Mutation*. 2020;41:2128–2142. https://doi.org/10.1002/humu.24110

3 BARD1 Pathogenic Variants are Associated with Triple-Negative Breast Cancer in a Spanish Hereditary Breast and Ovarian Cancer Cohort

Paula Rofes, Jesús Del Valle, Sara Torres-Esquius, Lídia Feliubadaló, Agostina Stradella, José Marcos Moreno-Cabrera, Adriana López-Doriga, Elisabet Munté, Rafael De Cid, Olga Campos, Raquel Cuesta, Álex Teulé, Èlia Gra, Judit San, Gabriel Capellá, Orland Díez, Joan Brunet, Judith Balmaña and Conxi Lázaro

Genes 2021, 12(2), 150; https://doi.org/10.3390/genes12020150

IF 2019 = 3.759. Rank 53/177 (GENETICS & HEREDITY)

genes

MDPI

Article

BARD1 Pathogenic Variants are Associated with Triple-Negative Breast Cancer in a Spanish Hereditary Breast and Ovarian Cancer Cohort

Paula Rofes [1,2,3,4,5], Jesús Del Valle [1,2,3,4,5], Sara Torres-Esquius [6], Lídia Feliubadaló [1,2,3,4,5], Agostina Stradella [1,2,3,4,7], José Marcos Moreno-Cabrera [1,2,3,4,5], Adriana López-Doriga [8,9], Elisabet Munté [1,2,3,4,5], Rafael De Cid [10], Olga Campos [1,2,3,4], Raquel Cuesta [1,2,3,4], Álex Teulé [1,2,3,4], Èlia Grau [1,2,3,4], Judit Sanz [11], Gabriel Capellá [1,2,3,4,5], Orland Díez [12,13], Joan Brunet [1,2,3,5,14], Judith Balmaña [6] and Conxi Lázaro [1,2,3,4,5,*]

[1] Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, 08908 L'Hospitalet de Llobregat, Spain; profes@iconcologia.net (P.R.); jdelvalle@iconcologia.net (J.D.V.); lfeliubadalo@iconcologia.net (L.F.); astradella@iconcologia.net (A.S.); jmoreno@igtp.cat (J.M.M.-C.); emunte@iconcologia.net (E.M.); ocampos@iconcologia.net (O.C.); rcuesta@iconcologia.net (R.C.); ateule@iconcologia.net (Á.T.); eggarces@iconcologia.net (È.G.); gcapella@iconcologia.net (G.C.); jbrunet@iconcologia.net (J.B.)
[2] Hereditary Cancer Program, Catalan Institute of Oncology, IGTP, 08916 Badalona, Spain
[3] Hereditary Cancer Program, Catalan Institute of Oncology, IDIBGI, 17007 Girona, Spain
[4] Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), IDIBELL, 08908 L'Hospitalet de Llobregat, Spain
[5] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), 28929 Madrid, Spain
[6] Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO), Medical Oncology Department, University Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, 08035 Barcelona, Spain; storres@vhio.net (S.T.-E.); jbalmana@vhio.net (J.B.)
[7] Medical Oncology Department, Catalan Institute of Oncology, IDIBELL, 08908 L'Hospitalet de Llobregat, Spain
[8] Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology, 08908 L'Hospitalet de Llobregat, Spain; alguerra@iconcologia.net
[9] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain
[10] Genomes for Life-GCAT Lab Group, IGTP, Institut Germans Trias i Pujol (IGTP), 08916 Badalona, Spain; rdecid@igtp.cat
[11] Genetic Counselling Unit, Medical Oncology Department, Althaia Xarxa Assistencial Universitària de Manresa, 08243 Manresa, Spain; jsanz@althaia.cat
[12] Catalan Health Institute, Vall d'Hebron Hospital Universitari, 08035 Barcelona, Spain; odiez@vhio.net
[13] Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO), 08035 Barcelona, Spain
[14] Medical Sciences Department, School of Medicine, University of Girona, 17007 Girona, Spain
[*] Correspondence: clazaro@iconcologia.net; Tel.: +34-93-2607145

check for updates

Citation: Rofes, P.; Del Valle, J.s; Torres-Esquius, S.; Feliubadaló, Ld.; Stradella, A.; Moreno-Cabrera, J.M.; López-Doriga, A.; Munté, E.; De Cid, R.; Campos, O.; et al. *BARD1* Pathogenic Variants are Associated with Triple-Negative Breast Cancer in a Spanish Hereditary Breast and Ovarian Cancer Cohort. *Genes* **2021**, *12*, 150. https://doi.org/10.3390/genes12020150

Academic Editors: Irmgard Irminger-Finger and Magda Ratajska

Received: 18 December 2020
Accepted: 20 January 2021
Published: 23 January 2021

Abstract: Only a small fraction of hereditary breast and/or ovarian cancer (HBOC) cases are caused by germline variants in the high-penetrance breast cancer 1 and 2 genes (*BRCA1* and *BRCA2*). BRCA1-associated ring domain 1 (*BARD1*), nuclear partner of *BRCA1*, has been suggested as a potential HBOC risk gene, although its prevalence and penetrance are variable according to populations and type of tumor. We aimed to investigate the prevalence of *BARD1* truncating variants in a cohort of patients with clinical suspicion of HBOC. A comprehensive *BARD1* screening by multigene panel analysis was performed in 4015 unrelated patients according to our regional guidelines for genetic testing in hereditary cancer. In addition, 51,202 Genome Aggregation Database (gnomAD) non-Finnish, non-cancer European individuals were used as a control population. In our patient cohort, we identified 19 patients with heterozygous *BARD1* truncating variants (0.47%), whereas the frequency observed in the gnomAD controls was 0.12%. We found a statistically significant association of truncating *BARD1* variants with overall risk (odds ratio (OR) = 3.78; CI = 2.10–6.48; $p = 1.16 \times 10^{-5}$). This association remained significant in the hereditary breast cancer (HBC) group (OR = 4.18; CI = 2.10–7.70; $p = 5.45 \times 10^{-5}$). Furthermore, deleterious *BARD1* variants were enriched among triple-negative BC patients (OR = 5.40; CI = 1.77–18.15; $p = 0.001$) compared to other BC subtypes. Our results support the role of *BARD1* as a moderate penetrance BC predisposing gene and highlight a stronger association with triple-negative tumors.

Keywords: *BARD1*; breast cancer; triple-negative breast cancer; ovarian cancer; hereditary breast and ovarian cancer; moderate cancer risk

1. Introduction

Hereditary breast and ovarian cancer (HBOC) risk has been traditionally linked to germline pathogenic variants (PVs) in breast cancer 1 and 2 genes (*BRCA1* and *BRCA2*). However, only 20–30% of high-risk families carry PVs in these genes [1]. Gradually, PVs in various other genes with different degrees of penetrance have also been associated with breast cancer (BC) and/or ovarian cancer (OC) risk [2]. Several genes that are either interacting with *BRCA1/2* or involved in DNA damage response pathways have also emerged as potential candidates that may account for some of the missing heritability of these so-called BRCAX families, although their associated risks have not been fully established [2].

BRCA1-associated ring domain 1 *(BARD1*) was first discovered in 1996 as the nuclear partner of BRCA1 and became one of the earliest candidates investigated [3]. It is localized on chromosome 2 at position 2q35 and encodes a protein of 777 amino acids that contains one N-terminal Really Interesting New Gene (RING)-finger domain, four Ankyrin (Ank) repeats and two C-terminal tandem BRCA1 C Terminus (BRCT) domains [4,5]. BARD1 shows structural homology with BRCA1 and they directly interact through their RING domains. The BARD1-BRCA1 obligate heterodimer functions as both an E3 ubiquitin ligase and as a direct mediator of homologous recombination for the recruitment of RAD51 to the sites of DNA double-strand break (DSB) [3,6,7]. Furthermore, BARD1 is also involved in other BRCA1-independent functions, including p53-mediated apoptosis [8].

To date, the role of *BARD1* in cancer predisposition remains inconclusive. Several case-control studies have reported a higher prevalence of deleterious *BARD1* variants among BC patients, supporting its role as a moderate risk predisposing gene [9–11]. An enrichment of *BARD1* PVs among triple-negative breast cancer (TNBC) cases has also been evidenced [12–14]. Contrarily, some studies have been unable to detect a significant association of *BARD1* with breast cancer risk [15,16]. Likewise, the association between *BARD1* and overall OC risk has shown controversial results [17–19]. Taken together, there is still insufficient evidence to elucidate the role of *BARD1* in breast and/or ovarian cancer predisposition. In the present study, we have investigated the prevalence of deleterious germline *BARD1* variants in a cohort of 4015 patients with clinical suspicion of hereditary breast and/or ovarian cancer, with the aim of elucidating the role of *BARD1* in cancer predisposition in the Spanish population.

2. Materials and Methods

2.1. Patients and Controls

A total of 4015 index patients with a personal or family history suggestive of hereditary BC and/or OC referring at genetic counseling units of the Catalan Institute of Oncology (ICO) and Vall d'Hebron (HVH) hospitals were included in the present study. Clinical characteristics for all enrolled patients were the following: patients with BC before 40 years; patients with TNBC before 60 years; male BC patients; patients with non-mucinous OC; patients with a family history of two cases of BC before age 50; patients with three or more cases of first-degree BC; patients with a case of bilateral BC associated with another case of BC in the family. Informed written consent for both diagnostic and research purposes was obtained from all patients, and the study protocol was approved by the ethics committee of Bellvitge Biomedical Research Institute (IDIBELL; PR278/19) and Vall d'Hebron Hospital (PRAG102-2016). A set of 194 Spanish cancer-free individuals from the Genomes For Life—Cohort Study of the Genomes of Catalonia (GCAT) cohort [20] were screened with the same cancer panel as ICO patients.

2.2. NGS Panel Testing

In the ICO cohort, genetic testing was performed on genomic DNA using the next-generation sequencing (NGS) custom panel I2HCP, which comprises 122–135 hereditary cancer (HC)-associated genes, depending on the version used [21]. Copy number analysis was performed from NGS data using DECoN [22] with parameter optimization [23]. Copy number variants (CNVs) in *BARD1* were validated using custom multiplex ligation-dependent probe amplification (MLPA) probes designed according to the instructions provided by MRC-Holland. Likewise, 26 HC-associated genes were included in the HVH NGS panel (BRCA Hereditary Cancer MASTR Plus kit, Agilent Technologies, Santa Clara, CA, USA). Copy number analysis was performed from NGS data using MASTR Reporter (Agilent Technologies, Santa Clara, CA, USA) and putative CNVs were validated by RT-PCR analysis [24]. For this study, we considered any variant that originates a premature stop codon or affects canonical splice site positions (+1,+2,−1,−2) as a pathogenic or likely pathogenic variant (pathogenic variant hereinafter); all of them were classified as (likely) pathogenic following the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines [25] and were confirmed by Sanger sequencing.

2.3. Variant Nomenclature

Human Genome Variation Society (HGVS)-approved guidelines [26] were used for *BARD1* variant nomenclature using NM_000465.2 (LRG_297). For variant numbering, nucleotide 1 is the A of the ATG translation initiation codon.

2.4. Co-Segregation Analysis and Loss of Heterozygosity (LOH)

Both analyses were performed by Sanger sequencing when samples from relatives or tumor DNA were available.

2.5. gnomAD Analysis

The Genome Aggregation Database (gnomAD) non-Finnish European population, non-cancer dataset (v2.1.1) [27] was used as a control population. Variants were downloaded and filtered to identify predicted loss-of-function variants in *BARD1*. CNV screening was performed in the gnomAD SVs v2.1 dataset.

2.6. Statistical Analysis

Differences in allele frequency between cases and controls were determined by the Fisher exact test. Odds ratios (OR) and the corresponding 95% confidence intervals (CI) were determined for two-by-two comparisons. Statistical tests were carried out using R v.3.5.1.

3. Results

In our study cohort of 4015 unrelated patients with hereditary breast and/or ovarian cancer, 476 PVs were identified as per clinical gene panel analysis (Table 1), representing 11.86% patients harboring PVs in high- to moderate-penetrance BC/OC-associated genes. In addition, with the aim of investigating the role of PVs in the *BARD1* gene, we performed an exhaustive analysis of truncating, splicing and CNVs in this gene. Nineteen patients carried heterozygous germline PVs in *BARD1*, resulting in a carrier frequency of 0.47%. Among them, one patient additionally carried a PV in the HBOC-predisposing gene *BRCA2* (patient 10; *BRCA2* c.3264dupT; p.(Gln1089Serfs*10)) (Table 2). The remaining 18 *BARD1*-mutated index patients tested negative for PVs in other BC/OC genes (for more details of the genes analyzed according to the phenotype, refer to Feliubadaló et al., 2019 [24]). Thus, after excluding carriers of other HBOC PVs, the global *BARD1* carrier frequency throughout our cohort of patients was 0.45%. The percentage of deleterious *BARD1* variants in the subset of patients with hereditary breast cancer (HBC) was 0.50%, 0.42% in hereditary ovarian cancer (HOC) cases and 0.33% in patients with HBOC (Table 1). No

BARD1 PVs were identified in our set of 194 cancer-free individuals. In order to increase the control cohort, loss-of-function *BARD1* variants were screened in the non-Finnish European gnomAD 2.1.1 (non-cancer) population, identifying a total of 61 heterozygous carriers out of 51,202 individuals (0.12%). The comparison of carrier frequencies between the patient and control cohorts revealed an overall significant association of *BARD1* PVs (OR = 3.78; CI = 2.10–6.48; $p = 1.16 \times 10^{-5}$). This association was also significant in the HBC group (OR = 4.18; CI = 2.10–7.70; $p = 5.45 \times 10^{-5}$). Moreover, deleterious *BARD1* variants demonstrated an increased risk in the HOC and HBOC groups, although the differences did not reach statistical significance (OR = 3.53, CI = 0.71–10.86, $p = 0.06$ and OR = 2.77, CI = 0.33–10.47, $p = 0.17$, respectively) (Table 1).

The clinical phenotype of *BARD1*-mutated patients is depicted in Table 2. Sixteen developed BC at a median age of 41 years (27–63), younger than the general population (median age at diagnosis 62 years old in females, according to NCI's SEER 21 2013–2017 Program). Of these, 10 were diagnosed with at least one TNBC. We compared the prevalence of deleterious *BARD1* variants between women diagnosed with TNBC and other BC subtypes and found significant differences according to the triple-negative status of carriers. deleterious *BARD1* variants were enriched in HBC families where the index case developed TNBC (OR = 5.40; CI = 1.77–18.15; $p = 0.001$) (Table 3). Regarding OC cases, three patients were diagnosed at a median age of 62 years (59–62)—two were diagnosed with high-grade ovarian serous carcinoma (HGOSC) and one with endometrioid carcinoma (EC).

Table 1. Summary of the next-generation sequencing (NGS) panel results in our hereditary breast and/or ovarian cancer (HBOC) cohort and in the control populations.

Clinical Indication	Number of Patients (%)	Genes Tested by Phenotype	Number of PVs (%)	*BARD1* (%)	*BARD1* Excluding Patients with Other PVs (%)
Only Hereditary Breast Cancer, HBC	2622 (65.31%)	*ATM, BRCA1, BRCA2, CHEK2, MLH1, MSH2, MSH6, PALB2, TP53*	270 PVs (10.30%): *ATM* (34), *BRCA1* (71), *BRCA2* (90), *CHEK2* (27), *MLH1* (3), *MSH2* (1), *MSH6* (2), *PALB2* (37), *TP53* (5)	13 (0.50%) OR = 4.18 (2.10–7.70) ** $p = 5.45 \times 10^{-5}$	13 (0.50%) OR = 4.18 (2.10–7.70) ** $p = 5.45 \times 10^{-5}$
Only Hereditary Ovarian Cancer, HOC	715 (17.81%)	*BRCA1, BRCA2, BRIP1, MLH1, MSH2, MSH6, RAD51C, RAD51D*	93 PVs (13.01%): *BRCA1* (39), *BRCA2* (35), *BRIP1* (6), *MLH1* (1), *MSH6* (4), *RAD51C* (4), *RAD51D* (4)	3 (0.42%) OR = 3.53 (0.71–10.86) $p = 0.06$	3 (0.42%) OR = 3.53 (0.71–10.86) $p = 0.06$
Hereditary Breast and Ovarian Cancer, HBOC	608 (15.14%)	*ATM, BRCA1, BRCA2, BRIP1, CHEK2, MLH1, MSH2, MSH6, PALB2, RAD51C, RAD51D, TP53*	104 PVs (17.11%): *ATM* (7), *BRCA1* (45), *BRCA2* (32), *BRIP1 (7), CHEK2* (6), *MSH2* (1), *PALB2* (3), *RAD51C* (1), *RAD51D (1), TP53* (1)	3 (0.49%) OR = 4.16 (0.83–12.79) * $p = 0.04$	2 (0.33%) OR = 2.77 (0.33–10.47) $p = 0.17$
HBC/HOC/HBOC + Other clinical indications	70 (1.74%)	Details in Ref: [24]	9 PVs (12.86%): *ATM* (2), *BRCA1* (2), *BRCA2* (1), *MSH6* (2), *PTEN* (1), *RAD51C* (1)	0 (0%)	0 (0%)
Total	4015		476 (11.86%)	19 (0.47%) OR = 3.99 (2.25–6.77) ** $p = 3.48 \times 10^{-6}$	18 (0.45%) OR = 3.78 (2.10–6.48) ** $p = 1.16 \times 10^{-5}$
			Controls studied		
			Spanish population cohort (n = 194)		0 (0%)
			gnomAD non-Finnish European, non-cancer cohort (n = 51,202)		61 (0.12%)

PV: pathogenic variant; OR: odds ratio. * α < 0.05. ** α < 0.01.

Table 2. Genotype and phenotype data of index patients carrying heterozygous germline pathogenic variants in the BRCA1-associated ring domain 1 (*BARD1)* gene.

Family	Clinical Indication	Cancer Type (Age at dx)	Tumor Phenotype	Family History (Age at dx)	*BARD1* PV (c.)	*BARD1* PV (p.)	Additional PVs
1	HBC	Breast (40,58)	ILC ER+ Her2-; TNBC	Cousin: PC (73)	c.157del	p.(Cys53Valfs*5)	
2	HBOC	Breast (30)	IDC ER+ Her2-				
3	HOC	Ovary (59)	HGOSC		c.176_177del	p.(Glu59Alafs*8)	
4 †	HBC	Breast (27,42)	ER+ BC; TNBC	Mother: Breast (44,44)	c.580_581del	p.(Arg194Glyfs*2)	

Table 2. *Cont.*

Family	Clinical Indication	Cancer Type (Age at dx)	Tumor Phenotype	Family History (Age at dx)	*BARD1* PV (c.)	*BARD1* PV (p.)	Additional PVs
5	HBC	Breast (38)	IDC ER+ Her2-	Aunt: Breast (37) ‡, Aunt: Breast (36)	c.1061C > A	p.(Ser354*)	
6	HOC	Ovary (62)	EC		c.1314+1G > A	p.?	
7	HBC	Breast (49)	TNBC		c.1349dup	p.(Asn450Lysfs*4)	
8	HBC	Breast (31)	TNBC	Aunt: Breast (64) ‡, Aunt: Breast (64) ‡	c.1652C > G	p.(Ser551*)	
9	HBC	Breast (56)	TNBC		c.1921C > T	p.(Arg641*)	
10	HBOC	Breast (54)	IDC ER+ Her2-	Mother: Ovary (63)			*BRCA2* c.3264dupT; p.(Gln1089Serfs*10)
11	HBC	Breast (63)	TNBC	Aunt: Breast (60) ‡, Cousin: Breast (54) ‡			
12	HBC	Breast (40)	IDC ER+ Her2+	Mother: EC (62), Breast (64)			
13	HBC	Breast (49)	TNBC				
14	HBC	Breast (30)	TNBC				
15	HBC	Breast (46,56,56)	IDBC; bilateral IDBC	Mother: Breast (78)			
16 ^	HBC	Breast (40,47)	TNBC; TNBC	Sister: Breast (46); Sister: Breast (48); Mother: Breast (48); Cousin: Breast (46)			
17	HBOC	Breast (42)	IDC ER+ Her2-	Uncle: Breast (71); Aunt: Ovary (62)	c.2129_2132del	p.(Asp710Valfs*3)	
18	HOC	Ovary (62)	HGOSC		c.(1568+1_1569-1)_(1810+1_1811-1)del Exons 7–8 deletion		
19	HBC	Breast (44)	TNBC	Mother: Breast (69); Aunt: Breast (60)	g.(?_215617227)_(215593730_?) Exons 7–11 deletion		

HBC: hereditary breast cancer; HBOC: hereditary breast and ovarian cancer; HOC: hereditary ovarian cancer; Dx: diagnosis; PV: pathogenic variant; HGOSC: high-grade ovarian serous carcinoma; PC: peritoneal carcinoma; EC: endometrioid carcinoma; BC: breast cancer; ILC: invasive lobular carcinoma of the breast; IDC: invasive ductal carcinoma of the breast; IDBC: intraductal breast carcinoma; TNBC: triple-negative breast cancer; ER: estrogen receptor status; Her2: human epidermal growth factor receptor 2 status. ‡ Cancer diagnosis unconfirmed. † Results previously reported in Ref [28]; ^ Results previously reported in Ref [29].

Two recurrent variants were identified in our set of samples. *BARD1* c.1921C > T; p.(Arg641*) was found in eight unrelated patients, thus representing the most frequent variant in our cohort. Besides, two unrelated patients harbored the *BARD1* c.157del; p.(Cys53Valfs*5) variant. The nine remaining variants were identified in one index case each (Figure 1). It is worth mentioning that we performed RT-PCR analysis of the splicing variant c.1314+1G > A, which causes skipping of exons 3 and 4 (r.216_1314del; p.(Ser72Argfs*37)) (data not shown). Interestingly, we identified two copy number variants (CNVs) (Table 2). One consisted in the deletion of exons 7 and 8, which was experimentally validated by RT-PCR analysis in the proband's cDNA (data not shown). This variant causes an out-of-frame deletion predicted to generate a truncated protein. The other CNV involved the loss of exons 7 to 11 and was validated using an MLPA custom probe. This deletion would presumably result in a BARD1 protein lacking both BRCT domains and the C-terminal region of the Ank domain. The screening of CNVs in the Genome Aggregation Database (gnomAD) splicing variants (SVs) dataset did not identify any CNV in the control population.

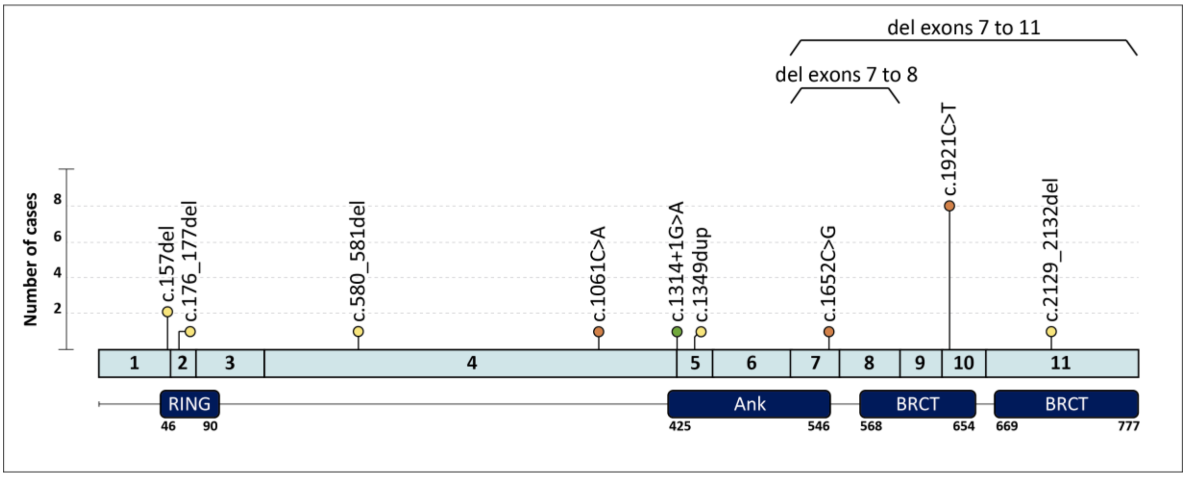


Figure 1. Spectrum of *BARD1* germline pathogenic variants found in our cohort. Locations of variants are displayed by lollipop structures with the following color code: orange for nonsense variants, yellow for frameshift variants and green for splicing variants. Horizontal lines correspond to copy number variants, each found in one index case. The different BARD1 protein domains are shown in dark blue boxes with an amino acid numbered scale.

Table 3. Summary of the triple-negative status of the hereditary breast cancer cohort.

Group	Number of Patients	***BARD1*-Mutated**
TNBC patients	680	10 (0.88%) OR = 5.40 (1.77–18.15) *p* = 0.001 **
Non-TNBC patients	2179	6 (0.28%)
Total	2859	16

TNBC: triple-negative breast cancer; OR: odds ratio. ** α < 0.01.

Regarding co-segregation and LOH studies, in a previous publication by our group, we reported the results of the co-segregation of family 16 [29]: the proband's mother, diagnosed with BC, as well as the sister and the maternal cousin, both affected by BC, had the same *BARD1* variant; the variant was also found in the proband's 39-year-old daughter, although she was asymptomatic. In the rest of the families, the co-segregation study was scarcely informative. In family 1, the proband's cousin was diagnosed with peritoneal carcinoma (PC) at age 73 and harbored the same *BARD1* PV. In families 4 and 15, the probands inherited the *BARD1* PV from their respective mothers, also affected by BC. However, in families 13 and 14, the probands inherited the *BARD1* PV from asymptomatic mothers. LOH analysis could only be performed in a tumor sample from the proband in family 14, but there was no evidence of LOH.

4. Discussion

In the present study, we performed a comprehensive analysis of the *BARD1* gene in a cohort of 4015 hereditary BC/OC patients. The screening for germline PVs evidenced that *BARD1* heterozygous carriers have an overall increased risk (OR = 3.78; CI = 2.10–6.48; $p = 1.16 \times 10^{-5}$). When stratified by clinical suspicion, the estimated risk for HBC patients resulted in a significant OR = 4.18 (CI = 2.1–7.7; $p = 5.45 \times 10^{-5}$). These results are comparable to those previously reported by several case–control studies. The largest analysis to date was performed by Couch et al. in a cohort of 28,536 BC patients, proposing *BARD1* as a moderate-risk gene with an OR = 2.16 (CI = 1.31–3.63; $p = 2.26 \times 10^{-3}$) [9]. Similarly, Slavin et al. reported an OR = 3.18 (CI = 1.34–7.36; $p = 0.012$) [10] and Weber-Lassalle et al. reported an OR = 5.35 (CI = 3.17–9.04; $p < 0.00001$) [11] in 2134 and 4469 familial BC patients, respectively. Besides, a recent meta-analysis by Suszynska and Kozlowski collected data from a total of 123 published studies and consistently reported an OR = 2.90 (CI = 2.25–3.75; $p < 0.0001$) over a cumulative cohort of ~48,700 BC patients [30]. However, there are some studies that failed to identify a significant association with BC risk, such as those published by Castéra et al. and Lu et al. [15,16].

An increase in the prevalence of PVs in *BARD1* among TNBC patients has been repeatedly suggested [12,13,31,32]. In agreement with this hypothesis, we identified ten *BARD1* PV carriers from 680 TNBC cases (carrier frequency = 0.9%), resulting in an OR = 5.40 (CI = 1.77–18.15; $p = 0.001$). Our results are comparable to the analysis of 4090 TNBC cases performed by Shimelis et al., who identified 25 individuals harboring *BARD1* PVs (0.61%) and obtained an OR = 5.92 (CI = 3.36–10.27; $p = 2.20 \times 10^{-9}$) [14], whereas a surprisingly high OR = 11.27 (CI = 3.37–25.01) was reported by Castéra et al. [15]. Despite the reduced sample size of our subset of TNBC patients, our results support that deleterious *BARD1* variants were enriched in TNBC cases. Further studies in larger cohorts will be necessary to more precisely assess the *BARD1*-associated risk with this tumor phenotype.

Our results also showed a trend, although non-significant, for HOC patients (OR = 3.53). Previous studies focusing on *BARD1* as an OC-predisposing gene have shown inconsistent results. Only Norquist et al. revealed a significant OR = 4.2 (CI = 1.4–12.5; $p = 0.02$) in 1915 OC cases [18], similar to that reported in our set of samples. Contrarily, the analysis of 3261 epithelial OC cases by Ramus et al. and 6294 OC cases by Lilyquist et al. resulted in non-significant associations of deleterious *BARD1* variants with OC risk [17,19]. The meta-analysis by Suszynska and Kozlowski could not detect an association of *BARD1* with OC risk in a cumulative set of ~20,800 OC cases either [30].

Unraveling the contribution of moderate-penetrance genes to HC predisposition is challenging, as the low incidence of PVs detected in these genes results in inaccurate estimates of their associated risks. Due to the limited number of carriers identified, increasing the study size is mandatory to improve the statistical power. Besides, case–control studies usually rely on controls from publicly available databases to reach statistical power instead of using geographically matched controls (GMCs), potentially causing an overestimation of the calculated ORs [9]. Multi-centric international studies could potentially reduce this heterogeneity by defining common inclusion criteria for patients and harmonizing the methodological features. It is also very likely that the true prevalence of *BARD1* PVs has been underrated. As a consequence of the lack of functional assays, we have not contemplated missense, synonymous and intronic variants in the risk calculations, as we cannot be certain of their pathogenicity.

It is worth emphasizing that we have performed a screening of CNVs in our cohort of HC patients, resulting in the identification of two large deletions (exons 7 to 8 and exons 7 to 11), accounting for 10.5% of the PVs. To our knowledge, only a small fraction of published studies have also performed this analysis and only seven CNVs have been identified so far: exon 1 deletion [33], exon 2 deletion [34], exon 1 to 6 deletion [35], exon 5 to 7 deletion [36], exon 8 to 11 deletion [37] and two whole-gene deletions [37,38]. While no CNVs were identified in the gnomAD SV control population dataset, analysis of *BARD1*

CNVs in HC cohorts is strongly recommended considering the significant contribution in our series of this kind of variant.

BARD1 has been included in multi-gene panels since it was regarded as a potential cancer-predisposing gene [39], despite the lack of robust risk estimates. The identification of *BARD1* PV carriers should be taken with caution, as inherited PVs in moderate- to low-penetrance genes may not necessarily be responsible for all the cancer diagnoses in a family. Nevertheless, although the clinical evidence available to date is still insufficient to impact risk management, continued testing of *BARD1* will permit access to the carrier status once recommendations for *BARD1* PV carriers become available in the future.

Taken together, our results confirm *BARD1* as a BC susceptibility gene and highlight a stronger association with triple-negative tumors. Future studies aimed at screening larger cohorts and refining the classification of *BARD1* variants will help to elucidate its role as a breast and/or ovarian cancer gene as well as define medical recommendations for *BARD1* PV carriers.

Author Contributions: *BARD1* data analysis: P.R., J.D.V., S.T.-E., L.F., J.M.M.-C., A.L.-D., R.D.C., G.C., J.B. (Joan Brunet) and J.B. (Judith Balmaña); Genetic tests: P.R., O.C., R.C. and O.D.; Collection and analysis of clinical information: P.R., J.D.V., S.T.-E., L.F., A.S., E.M., Á.T., È.G. and J.S.; Integration and comprehensive analysis of data: P.R., J.D.V. and C.L.; Writing: P.R., J.D.V. and C.L.; All authors have read and agreed to the published version of the manuscript.

Funding: Contract grant sponsor: Supported by the Carlos III National Health Institute funded by FEDER funds—a way to build Europe—(PI19/00553; PI16/00563; PI16/01898; PI16/11363; PI15/00355; PI12/02585; SAF2015-68016-R and CIBERONC); the Government of Catalonia (Pla estratègic de recerca i innovació en salut (PERIS_MedPerCan and URDCat projects), 2017SGR1282 and 2017SGR496); and CERCA Programa/Generalitat de Catalunya for institutional support. GCAT is supported by Acción de Dinamización del ISCIII-MINECO (ADE 10/00026), by the Ministry of Health of the Generalitat of Catalunya and by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (2017-SGR 529) and GCAT Cession Research Project PI-2018-09 GCAT_CM.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the ethics committee of Bellvitge Biomedical Research Institute (IDIBELL; PR278/19) and Vall d'Hebron Hospital (PRAG102-2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is contained within the article.

Acknowledgments: We thank the CERCA Program/Generalitat de Catalunya for their institutional support. We also wish to thank all members of the ICO Hereditary Cancer Program and the Hereditary Cancer Genetics Group at Hospital Vall d'Hebron. This study makes use of data generated by IGTP, specifically by its research group the GCAT—Genomes for Life Cohort study of the Genomes of Catalonia, IGTP. A full list of the investigators who contributed to the generation of the data is available from http://www.genomesforlife.com/. IGTP is part of the CERCA Program/Generalitat de Catalunya.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Brakeleer, S.; De Grève, J.; Loris, R.; Janin, N.; Lissens, W.; Sermijn, E.; Teugels, E. Cancer predisposing missense and protein truncating BARD1 mutations in non-BRCA1 or BRCA2 breast cancer families. *Hum. Mutat.* **2010**, *31*, e1175–e1185. [CrossRef] [PubMed]
2. Easton, D.F.; Pharoah, P.D.P.; Antoniou, A.C.; Tischkowitz, M.; Tavtigian, S.V.; Nathanson, K.L.; Devilee, P.; Meindl, A.; Couch, F.J.; Southey, M.; et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **2015**, *372*, 2243–2257. [CrossRef] [PubMed]
3. Wu, L.C.; Wang, Z.W.; Tsan, J.T.; Spillman, M.A.; Phung, A.; Xu, X.L.; Yang, M.C.W.; Hwang, L.Y.; Bowcock, A.M.; Baer, R. Identification of a RING protein that can interact in vivo with the BRCA1 gene product. *Nat. Genet.* **1996**, *14*, 430–440. [CrossRef] [PubMed]
4. Fox, D.; Le Trong, I.; Rajagopal, P.; Brzovic, P.S.; Stenkamp, R.E.; Klevit, R.E. Crystal structure of the BARD1 ankyrin repeat domain and its functional consequences. *J. Biol. Chem.* **2008**, *283*, 21179–21186. [CrossRef]

5. Birrane, G.; Varma, A.K.; Soni, A.; Ladias, J.A.A. Crystal structure of the BARD1 BRCT domains. *Biochemistry* **2007**, *46*, 7706–7712. [CrossRef]
6. Hashizume, R.; Fukuda, M.; Maeda, I.; Nishikawa, H.; Oyake, D.; Yabuki, Y.; Ogata, H.; Ohta, T. The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* **2001**, *276*, 14537–14540. [CrossRef]
7. Moynahan, M.E.; Chiu, J.W.; Koller, B.H.; Jasint, M. Brca1 controls homology-directed DNA repair. *Mol. Cell* **1999**, *4*, 511–518. [CrossRef]
8. Feki, A.; Jefford, C.E.; Berardi, P.; Wu, J.Y.; Cartier, L.; Krause, K.H.; Irminger-Finger, I. BARD1 induces apoptosis by catalysing phosphorylation of p53 by DNA-damage response kinase. *Oncogene* **2005**, *24*, 3726–3736. [CrossRef]
9. Couch, F.J.; Shimelis, H.; Hu, C.; Hart, S.N.; Polley, E.C.; Na, J.; Hallberg, E.; Moore, R.; Thomas, A.; Lilyquist, J.; et al. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol.* **2017**, *3*, 1190–1196. [CrossRef]
10. Slavin, T.P.; Maxwell, K.N.; Lilyquist, J.; Vijai, J.; Neuhausen, S.L.; Hart, S.N.; Ravichandran, V.; Thomas, T.; Maria, A.; Villano, D.; et al. The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *npj Breast Cancer* **2017**, *3*, 1–10. [CrossRef]
11. Weber-Lassalle, N.; Borde, J.; Weber-Lassalle, K.; Horváth, J.; Niederacher, D.; Arnold, N.; Kaulfuß, S.; Ernst, C.; Paul, V.G.; Honisch, E.; et al. Germline loss-of-function variants in the BARD1 gene are associated with early-onset familial breast cancer but not ovarian cancer. *Breast Cancer Res.* **2019**, *21*, 55. [CrossRef] [PubMed]
12. De Brakeleer, S.; De Grève, J.; Desmedt, C.; Joris, S.; Sotiriou, C.; Piccart, M.; Pauwels, I.; Teugels, E. Frequent incidence of BARD1-truncating mutations in germline DNA from triple-negative breast cancer patients. *Clin. Genet.* **2016**, *89*, 336–340. [CrossRef] [PubMed]
13. González-Rivera, M.; Lobo, M.; López-Tarruella, S.; Jerez, Y.; del Monte-Millán, M.; Massarrah, T.; Ramos-Medina, R.; Ocaña, I.; Picornell, A.; Garzón, S.S.; et al. Frequency of germline DNA genetic findings in an unselected prospective cohort of triple-negative breast cancer patients participating in a platinum-based neoadjuvant chemotherapy trial. *Breast Cancer Res. Treat.* **2016**, *156*, 507–515. [CrossRef] [PubMed]
14. Shimelis, H.; LaDuca, H.; Hu, C.; Hart, S.N.; Na, J.; Thomas, A.; Akinhanmi, M.; Moore, R.M.; Brauch, H.; Cox, A.; et al. Triple-negative breast cancer risk genes identified by multigene hereditary cancer panel testing. *J. Natl. Cancer Inst.* **2018**, *110*, 855–862. [CrossRef] [PubMed]
15. Castéra, L.; Harter, V.; Muller, E.; Krieger, S.; Goardon, N.; Ricou, A.; Rousselin, A.; Paimparay, G.; Legros, A.; Bruet, O.; et al. Landscape of pathogenic variations in a panel of 34 genes and cancer risk estimation from 5131 HBOC families. *Genet. Med.* **2018**, *20*, 1677–1686. [CrossRef] [PubMed]
16. Lu, H.M.; Li, S.; Black, M.H.; Lee, S.; Hoiness, R.; Wu, S.; Mu, W.; Huether, R.; Chen, J.; Sridhar, S.; et al. Association of Breast and Ovarian Cancers with Predisposition Genes Identified by Large-Scale Sequencing. *JAMA Oncol.* **2019**, *5*, 51–57. [CrossRef]
17. Ramus, S.J.; Song, H.; Dicks, E.; Tyrer, J.P.; Rosenthal, A.N.; Intermaggio, M.P.; Fraser, L.; Gentry-Maharaj, A.; Hayward, J.; Philpott, S.; et al. Germline mutations in the BRIP1, BARD1, PALB2, and NBN genes in women with ovarian cancer. *J. Natl. Cancer Inst.* **2015**, *107*. [CrossRef]
18. Norquist, B.M.; Harrell, M.I.; Brady, M.F.; Walsh, T.; Lee, M.K.; Gulsuner, S.; Bernards, S.S.; Casadei, S.; Yi, Q.; Burger, R.A.; et al. Inherited mutations in women with ovarian carcinoma. *JAMA Oncol.* **2016**, *2*, 482–490. [CrossRef]
19. Lilyquist, J.; LaDuca, H.; Polley, E.; Davis, B.T.; Shimelis, H.; Hu, C.; Hart, S.N.; Dolinsky, J.S.; Couch, F.J.; Goldgar, D.E. Frequency of mutations in a large series of clinically ascertained ovarian cancer cases tested on multi-gene panels compared to reference controls. *Gynecol. Oncol.* **2017**, *147*, 375–380. [CrossRef]
20. Obón-Santacana, M.; Vilardell, M.; Carreras, A.; Duran, X.; Velasco, J.; Galván-Femenía, I.; Alonso, T.; Puig, L.; Sumoy, L.; Duell, E.J.; et al. GCAT|Genomes for life: A prospective cohort study of the genomes of Catalonia. *BMJ Open* **2018**, *8*, 18324. [CrossRef]
21. Castellanos, E.; Gel, B.; Rosas, I.; Tornero, E.; Santín, S.; Pluvinet, R.; Velasco, J.; Sumoy, L.; Del Valle, J.; Perucho, M.; et al. A comprehensive custom panel design for routine hereditary cancer testing: Preserving control, improving diagnostics and revealing a complex variation landscape. *Sci. Rep.* **2017**, *7*. [CrossRef]
22. Fowler, A.; Mahamdallie, S.; Ruark, E.; Seal, S.; Ramsay, E.; Clarke, M.; Uddin, I.; Wylie, H.; Strydom, A.; Lunter, G.; et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res.* **2016**, *1*, 20. [CrossRef] [PubMed]
23. Moreno-Cabrera, J.M.; del Valle, J.; Castellanos, E.; Feliubadaló, L.; Pineda, M.; Brunet, J.; Serra, E.; Capellà, G.; Lázaro, C.; Gel, B. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur. J. Hum. Genet.* **2020**, *28*. [CrossRef]
24. Feliubadaló, L.; López-Fernández, A.; Pineda, M.; Díez, O.; del Valle, J.; Gutiérrez-Enríquez, S.; Teulé, A.; González, S.; Stjepanovic, N.; Salinas, M.; et al. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int. J. Cancer* **2019**, *145*, 2682–2691. [CrossRef]
25. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; 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.* **2015**, *17*, 405–424. [CrossRef] [PubMed]

26. den Dunnen, J.T.; Dalgleish, R.; Maglott, D.R.; Hart, R.K.; Greenblatt, M.S.; Mcgowan-Jordan, J.; Roux, A.F.; Smith, T.; Antonarakis, S.E.; Taschner, P.E.M. HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* **2016**, *37*, 564–569. [CrossRef] [PubMed]
27. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alföldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*. [CrossRef] [PubMed]
28. Bonache, S.; Esteban, I.; Moles-Fernández, A.; Tenés, A.; Duran-Lozano, L.; Montalban, G.; Bach, V.; Carrasco, E.; Gadea, N.; López-Fernández, A.; et al. Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings. *J. Cancer Res. Clin. Oncol.* **2018**, *144*, 2495–2513. [CrossRef]
29. Feliubadaló, L.; Tonda, R.; Gausachs, M.; Trotta, J.R.; Castellanos, E.; López-Doriga, A.; Teulé, À.; Tornero, E.; Del Valle, J.; Gel, B.; et al. Benchmarking of whole exome sequencing and Ad Hoc designed panels for genetic testing of hereditary cancer. *Sci. Rep.* **2017**, *7*, 1–11. [CrossRef] [PubMed]
30. Suszynska, M.; Kozlowski, P. Summary of bard1 mutations and precise estimation of breast and ovarian cancer risks associated with the mutations. *Genes (Basel)* **2020**, *11*, 798. [CrossRef]
31. Couch, F.J.; Hart, S.N.; Sharma, P.; Toland, A.E.; Wang, X.; Miron, P.; Olson, J.E.; Godwin, A.K.; Pankratz, V.S.; Olswold, C.; et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. *J. Clin. Oncol.* **2015**, *33*, 304–311. [CrossRef] [PubMed]
32. Buys, S.S.; Sandbach, J.F.; Gammon, A.; Patel, G.; Kidd, J.; Brown, K.L.; Sharma, L.; Saam, J.; Lancaster, J.; Daly, M.B. A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes. *Cancer* **2017**, *123*, 1721–1730. [CrossRef] [PubMed]
33. Tung, N.; Battelli, C.; Allen, B.; Kaldate, R.; Bhatnagar, S.; Bowles, K.; Timms, K.; Garber, J.E.; Herold, C.; Ellisen, L.; et al. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. *Cancer* **2015**, *121*, 25–33. [CrossRef] [PubMed]
34. Adedokun, B.; Zheng, Y.; Ndom, P.; Gakwaya, A.; Makumbi, T.; Zhou, A.Y.; Yoshimatsu, T.F.; Rodriguez, A.; Madduri, R.K.; Foster, I.T.; et al. Prevalence of inherited mutations in breast cancer predisposition genes among women in Uganda and Cameroon. *Cancer Epidemiol. Biomarkers Prev.* **2020**, *29*, 359–367. [CrossRef]
35. Zeng, C.; Guo, X.; Wen, W.; Shi, J.; Long, J.; Cai, Q.; Shu, X.O.; Xiang, Y.; Zheng, W. Evaluation of pathogenetic mutations in breast cancer predisposition genes in population-based studies conducted among Chinese women. *Breast Cancer Res. Treat.* **2020**, *181*, 465–473. [CrossRef]
36. Kaneyasu, T.; Mori, S.; Yamauchi, H.; Ohsumi, S.; Ohno, S.; Aoki, D.; Baba, S.; Kawano, J.; Miki, Y.; Matsumoto, N.; et al. Prevalence of disease-causing genes in Japanese patients with BRCA1/2-wildtype hereditary breast and ovarian cancer syndrome. *npj Breast Cancer* **2020**, *6*. [CrossRef]
37. Carter, N.J.; Marshall, M.L.; Susswein, L.R.; Zorn, K.K.; Hiraki, S.; Arvai, K.J.; Torene, R.I.; McGill, A.K.; Yackowski, L.; Murphy, P.D.; et al. Germline pathogenic variants identified in women with ovarian tumors. *Gynecol. Oncol.* **2018**, *151*, 481–488. [CrossRef]
38. Kwong, A.; Shin, V.Y.; Chen, J.; Cheuk, I.W.Y.; Ho, C.Y.S.; Au, C.H.; Chan, K.K.L.; Ngan, H.Y.S.; Chan, T.L.; Ford, J.M.; et al. Germline mutation in 1338 BRCA-negative Chinese hereditary breast and/or ovarian cancer patients: Clinical testing with a multigene test panel. *J. Mol. Diagn.* **2020**, *22*, 544–554. [CrossRef]
39. Alenezi, W.M.; Fierheller, C.T.; Recio, N.; Tonin, P.N. Literature review of BARD1 as a cancer predisposing gene with a focus on breast and ovarian cancers. *Genes (Basel)* **2020**, *11*, 856. [CrossRef]

Appendix B: supplementary materials

Supplementary File 9 - Article 1

Optimization algorithm

A greedy approach was used to optimize the parameters of the algorithms to maximize sensitivity while limiting specificity loss, or improving it when possible. For each dataset, we randomly select samples to define a training set with 50% of the samples to optimize algorithm parameters and a validation set with the other 50% to evaluate them.

The default parameter, D, is used as reference. For each numeric parameter, 22 values between D0.25 and D1.75 are considered (exponent values evaluated were 0.25, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.92, 0.94, 0.96, 0.98, 1, 1.02, 1.04, 1.06, 1.08, 1.1, 1.2, 1.3, 1.4, 1.5, 1.75). For CODEX2, only 9 values were considered due to its high CPU requirement (exponent values evaluated were 0.6, 0.85, 0.92, 0.97, 1, 1.03, 1.08, 1.15, 1.4). For categorical parameters, all options are considered. Initially, the best solution is the execution with the default parameters. Some numeric parameters were also restricted to a subset of options. Additionally, a min and max threshold was stablished for the numerical parameters. See “Tools parameters” section below.

The greedy algorithm behaves as follows. Optimization starts from the first parameter: the algorithm is executed evaluating the whole values range for this parameter and keeping other parameters with default values. Solution with highest whole diagnostics strategy sensitivity is chosen if specificity decreases less than 30% in comparison to the previous best solution. If the whole diagnostics strategy sensitivity cannot be improved, then the highest per gene sensitivity is chosen if specificity decreases less than 25%. Finally, if per gene sensitivity cannot be improved, the highest per ROI sensitivity is chosen if specificity decreases less than 20%. In case of a tie at any level, the solution with the best specificity is considered. The best parameter value is therefore fixed. Optimization continues from a second parameter randomly chosen. The algorithm is executed evaluating the whole values range for this parameter and keeping other parameters with default values or fixed values. The process is repeated until the last parameter is reached.

To overcome the dependence of greedy algorithms on the order of parameter selection, the whole optimization process was repeated starting from each different parameter.

Below is shown the pseudocode describing the main steps for the optimization algorithm.

Main steps of optimization algorithm

```
best_solution = execution with default parameters
for each parameter p
        D = default parameter value
        if p is numerical:
                values_range = 22 values from [$D^{0.25}$, $D^{1.75}$]
        else if p is categorical:
                values_range = all categorical values

        executions = ∅
        for each v ∈ values_range:
                v_previous = fixed values for already optimized parameters
                v_next = default values for not optimized parameters
                algorithm_parameters = v ∪ v_previous ∪ v_next
                executions = executions ∪ execute_algorithm(algorith_parameters)

        metrics = {whole strategy sensitivity, per gene, per ROI}
        for each m ∈ metrics and while !success:
                local_best = highest_sensitivity(executions, m)
                coeff = 0.7 if (m = whole strategy sensitivity)
                coeff = 0.75 if (m = per gene)
                coeff = 0.8 if (m = per ROI)

                if sensitivity(local_best, m) > sensitivity(best_solution, m)
                        and specificity(local_best, m) > coeff * specificity(best_solution, m):
                        best_solution = local_best
                        success = true
                        p is fixed with value from local_best
```

panelcn.MOPS

Parameter	default	Min	Max	Options
CN0	0.025	0	1	
CN1	0.57	0.1	2	
CN3	1.46	1	2	
CN4	2	1	3	
sizeFactor	quant			mean, median, quant, mode
norm	1			0, 1, 2
normType	quant			mean, median, quant, poisson, mode
qu	0.25	0	1	
quSizeFactor	0.75	0	1	
priorImpact	1			0, 0.5, 1, 1.5, 2, 4, 6, 8, 10, 15, 20, 30, 50, 100
minMedianRC	30	0	200	
maxControls				20, 25, 30
readLength	200	30	1000	

DECoN

Parameter	default	Min	Max	Options
mincorr	0.98			0.5, 0.75, 0.85, 0.9, 0.92, 0.94, 0.95, 0.96, 0.97, 0.975, 0.98, 0.985, 0.99, 1
mincov	100	0	1000	
transProb	0.01	0	1	

CoNVaDING

Parameter	default	Min	Max	Options
regionThreshold	20	0	100	
ratioCutOffLow	0.65	0	2	

ratioCutOffHigh	1.4	0	2	
zScoreCutOffLow	-3	-10	0	
zScoreCutOffHigh	3	0	10	
sampleRatioScore	0.09	0	1	
percentageLessReliableTargets	20	0	100	

ExomeDepth

Parameter	default	Min	Max	Options
phi.bins	1			1, 2, 3
transition.probability	0.0001	0	1	
expected.CNV.length	1000	0	1000000	
readLength	200	30	30000	

CODEX2

Parameter	default	Min	Max	Options
length_thresh_down	20	10	100	
length_thresh_up	2000	500	5000	
mapp_thresh	0.9			0.8, 0.85, 0.9, 0.95, 1
gc_thresh_down	20	5	40	
gc_thresh_up	80	60	95	
sample_reads_median_limit	50	20	100	
cn_del_factor	1.7	1	2	
cn_dup_factor	2.3	2	3	

Supplementary File - Article 2

DECoN execution

DECoN v1.0.1 was executed with customized parameters for both MiSeq and HiSeq samples, focusing on improving the sensitivity.

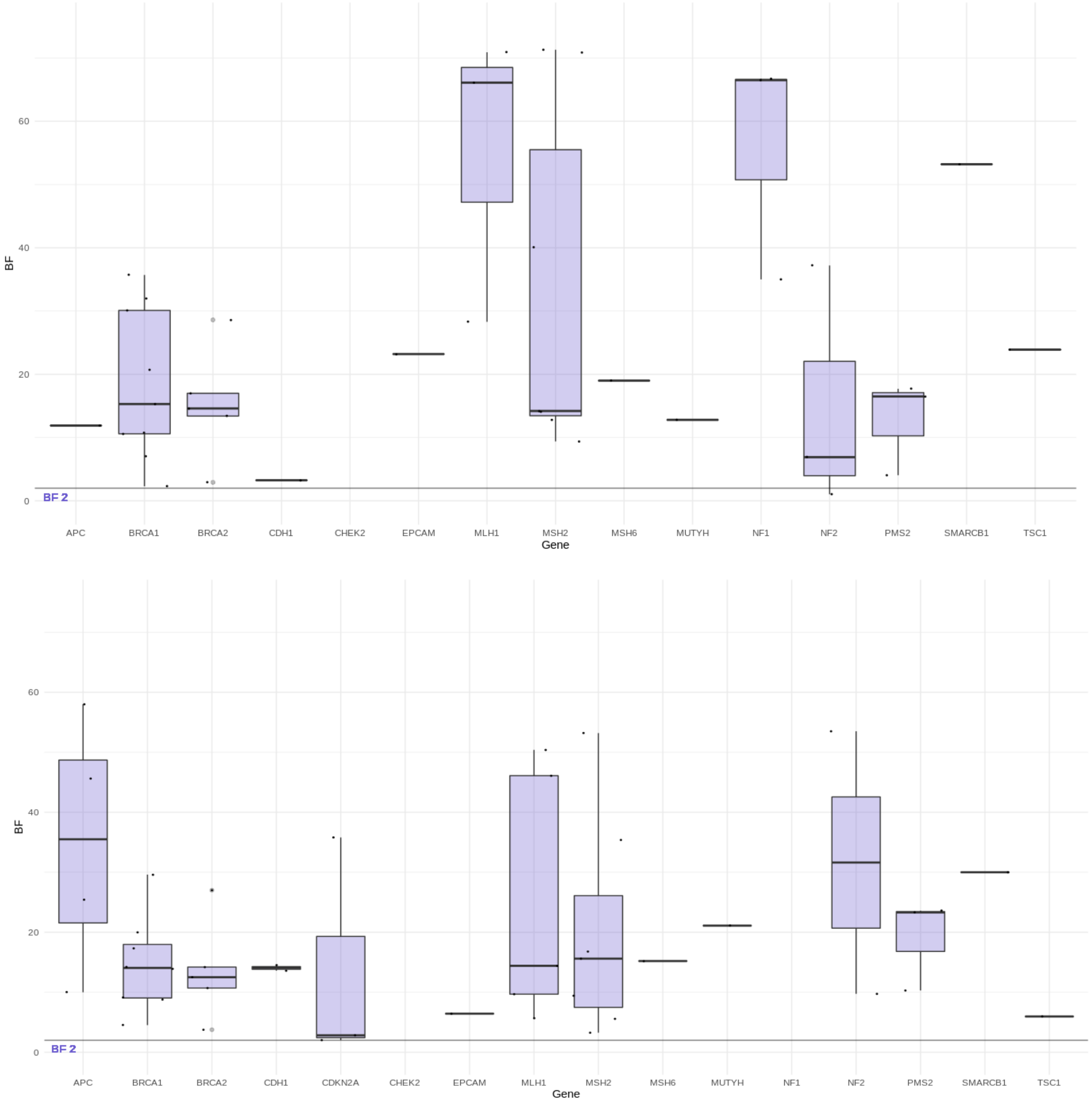
Parameter	DECoN Default	MiSeq samples	HiSeq samples
Minimum correlation (mincorr)	0.98	0.95	0.98
Minimum coverage (mincov)	100	60	80
Transition probability (transprob)	0.01	0.1585	0.1585

For the HiSeq sequencing runs, all samples (~96) of each run were analyzed together. For the MiSeq sequencing runs, all samples of each run (10 to 16 samples) were analyzed along with 51 samples from other MiSeq runs with no known CNVs. This set of 51 samples was added because read-depth based CNV calling tools work better with a higher number of samples.

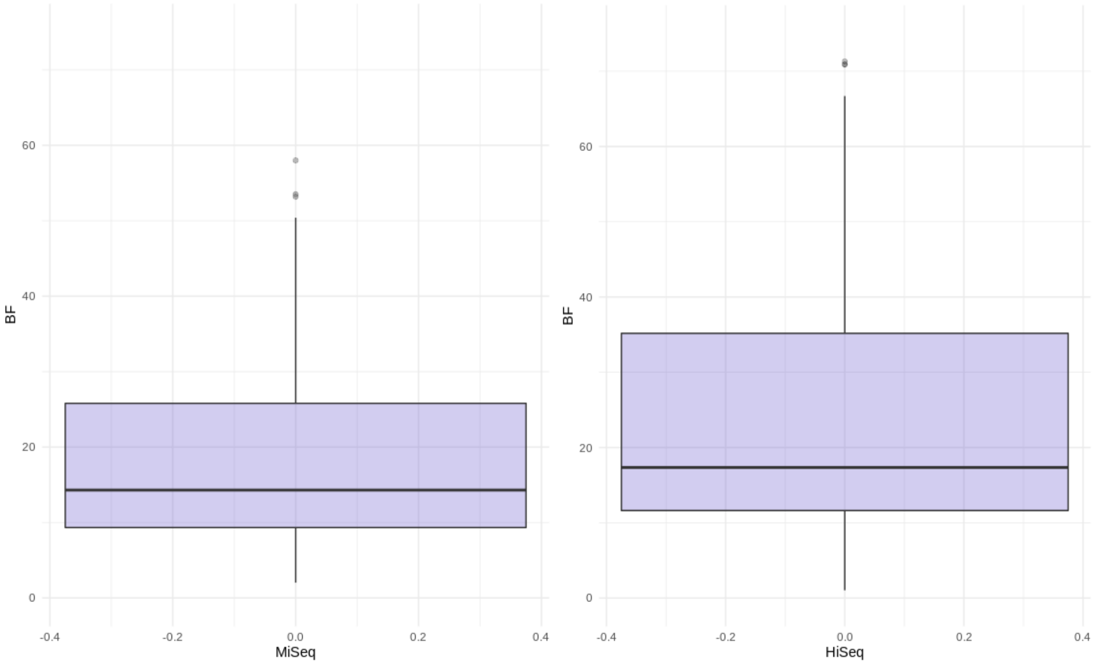
Additional N.comp discussion

For each sample, the N.comp value provided by DECoN quantifies the number of samples used as a reference set. In our study, this value ranged from 1 to 20. When plotting the N.comp value for the true-positive and false-positive calls of the prospective study (Supplementary Figure 5), we observed that most frequent values were low and the median was 5. If only the true-positive calls are plotted (Supplementary Figure 5), the N.comp value ranged from 3 to 17, although this value ranged from 1 to 20 in the true-positive CNV calls of our previous benchmark.

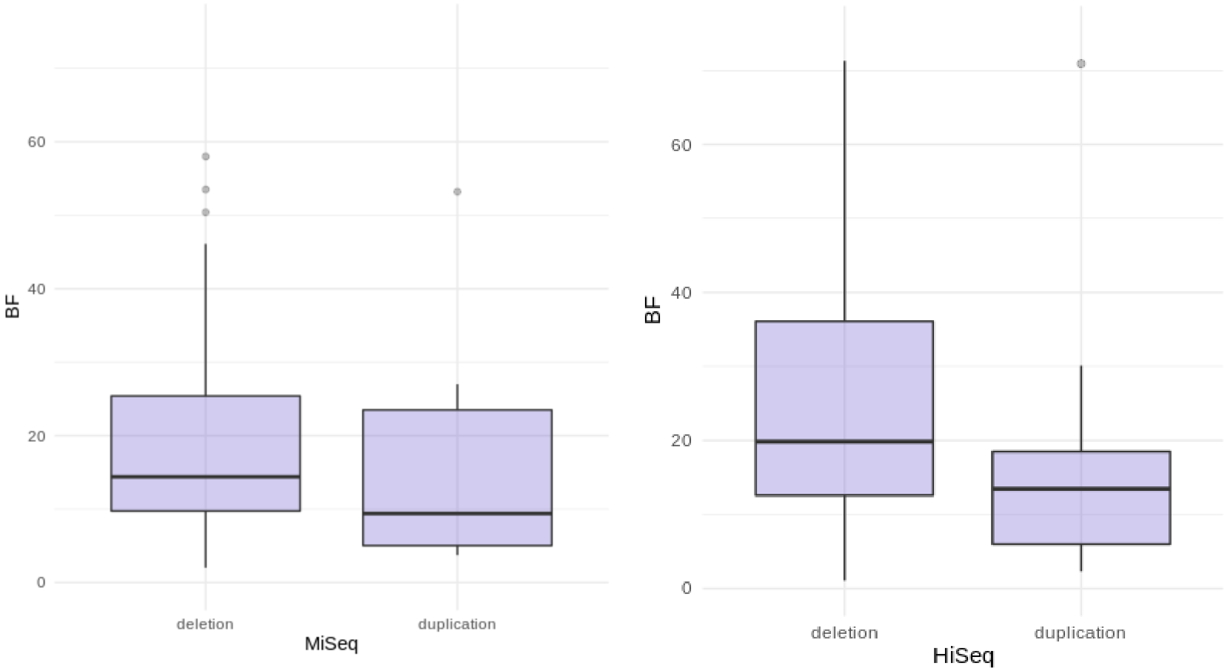
These results suggest that, in our datasets, there is not a minimum number of samples below which all CNV calls are expected to be false. In our study, DECoN was capable of detecting the true CNV calls even selecting a few samples as a reference set from the whole run where they were analyzed.



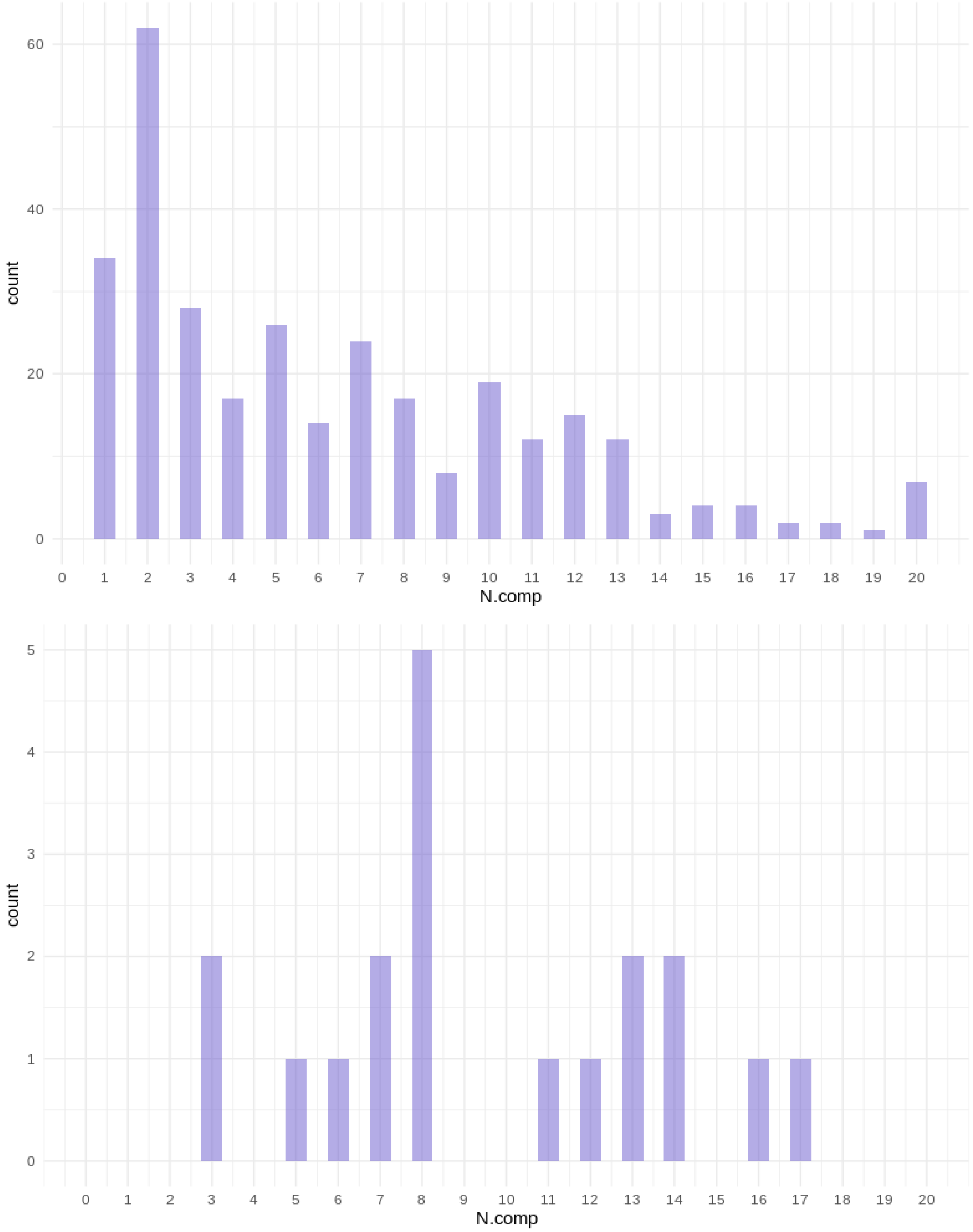
Supplementary Figure 2. BF distribution of true-positive CNV calls from our previous benchmark: gene comparison. **Upper panel**: Per gene HiSeq BF values restricted to the (0.75) range. **Lower panel**: Per gene MiSeq BF values restricted to the (0-75) range. Only one call (a mosaic sample) obtained a BF lower than 2.



Supplementary Figure 3. BF distribution of true-positive CNV calls from our previous benchmark: MiSeq vs HiSeq comparison. Values restricted to the (0-75 range). The MiSeq and HiSeq datasets contained different samples.



Supplementary Figure 4- BF distribution of true-positive CNV calls from our previous benchmark: deletions vs. duplications comparison. Values restricted to the (0-75 range).



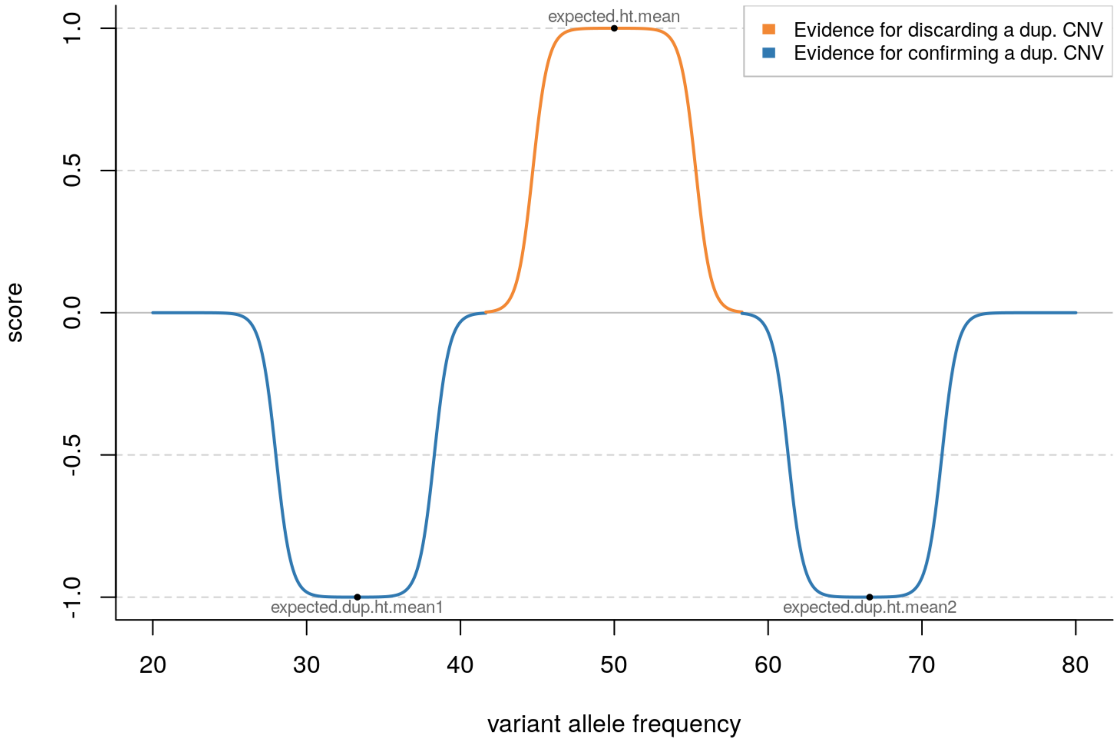
Supplementary Figure 5 – N.comp distribution of the CNV calls in the prospective study. **Upper panel**: false-positive and true-positive CNV calls are included. **Lower panel**: only true-positive CNV calls are included.

Supplementary File 1 - Article 3

Scoring model for CNV duplications

CNVfilteR identifies a certain CNV duplication as a false positive using the allele frequency of the heterozygous SNVs in that CNV. Each SNV is scored using a scoring model, and if the sum of the scores of all the SNVs in the CNV is greater than the duplication threshold score (defaults to 0.5), the CNV is identified as false positive. The scoring model is based on fuzzy logic, where elements can have any value between 1 (True) and 0 (False). A common way of applying fuzzy logic is using the sigmoid function. CNVfilteR uses the sigmoid function implemented in the pracma package, which is defined as y = 1 / (1 + e ^ (−c1(x − c2))). The scoring model is built on 6 sigmoids defined on 6 different intervals. The c1 parameter is 2 by default, and the c2 parameter is defined for the 6 sigmoids:

- First sigmoid: interval [20, 33.3], c2=28
- Second sigmoid: interval [33.3, 41.65], c2=38.3
- Third sigmoid: interval [41.65, 50], c2=44.7
- Fourth sigmoid: interval [50, 58.3], c2=55.3
- Fifth sigmoid: interval [58.3, 66.6], c2=61.3
- Sixth sigmoid: interval [66.6, 80], c2=71.3



All parameter values are customizable. Code examples of how to plot and modify the scoring model are available at Bioconductor site.

Evaluation on HuRef, AK1 and NA12878 samples

Data, tools and evaluation metrics

CNVfilteR was evaluated on the HuRef, AK1 and NA12878 genomes. Reference callsets and CNV calls from different tools were obtained from different sources (see table below).

To obtain the SNV calls for each sample, they were downloaded and aligned to the hs37d5 human genome assembly using BWA mem v0.7.13. SAMtools v0.1.8 was used to sort and index BAM files and duplicates were marked using Picard v2.18.4. Point mutations were called with Strelka v2.9.3. To enrich the CNV tools results for the HuRef and AK1 genomes, LUMPY v0.2.13 (via smoove v0.2.3) was also executed to call CNVs. Details are summarized in the following table.

	Reference call set	CNV tools results	SRA accession number
HuRef	Obtained from Trost et al. 2018 (file S1). Contains deletions and duplications ≥1 kb.	Obtained from Trost et al. 2018 (file S5). LUMPY results were obtained from our own pipeline as explained above. Only calls ≥ 500 bp were retained.	SRR7097859
AK1	Obtained from Trost et al. 2018, (file S4), which used Seo et al. 2016 as source. Contains only deletions ≥1 kb.	Same as the HuRef sample. Only deletions ≥ 500 bp were retained.	SRR3602759
NA12878	Provided by Zhang et al. 2019 authors, which used Parikh et al 2016 and MacDonald et al. 2014 as sources. Contains deletions and duplications > 1 kb. A similar version without the CNV type is also available at Zhang et al. 2019 publication.	Provided by Zhang et al. 2019. Only calls ≥ 500 bp were retained. A similar version without the CNV type is also available at Zhang et al. 2019 publication.	SRR622457

SRA: Sequence Read Archive from NCBI

Evaluation metrics

A tool call was defined as true positive (TP) if it had a 50% reciprocal overlap with any reference call, false positive (FP) otherwise. If a certain reference call had no reciprocal overlap with any tool call, it was counted as false negative (FN). Sensitivity was defined as TP / (TP + FN), false discovery rate as FP / (FP + TP) and F1-score as 2TP / (2TP + FP + FN).

Evaluation on HiSeq-panel and MiSeq-panel datasets

Datasets and tools

CNVfilteR was evaluated on 541 gene-panel samples (411 HiSeq and 130 MiSeq samples, see table below), which are a superset of the samples used in a previous work (Moreno-Cabrera et al. 2020). Both HiSeq and MiSeq datasets contained data from a hybridization-based target capture NGS panel, called I2HCP, designed for hereditary cancer diagnostics (Castellanos et al., 2017). Both datasets were generated in real diagnostics settings and contained single and multi-exon CNVs, all of them validated by MLPA. Negative MLPA data, meaning no detection of any CNV, was also available for a subset of genes. Detailed information on MLPA-detected CNVs for each dataset can be found in Supplementary File 3. Samples were generated at the ICO-IGTP Joint Program for Hereditary Cancer. All MiSeq samples and a subset of HiSeq samples are available at the EGA under the accession number EGAS00001004316. All samples were aligned to the GRCh37 human genome assembly using BWA mem v0.7.12. SAMtools v0.1.19 was used to sort and index BAM files. No additional processing or filtering was applied to the BAM files. Varscan v2.4.1 was used to call point mutations and DECoN v1.0.1 was chosen for calling CNVs.

	Samples	Validated genes with CNV	Single-exon CNVs	Multi-exon CNVs	Deletion CNVs	Duplication CNVs	Validated genes with no CNV
MiSeq dataset	130	64	19	45	56	8	167
HiSeq dataset	411	62	19	43	52	10	1076

Regions of interest

We generated a target bed file containing all coding exons from all protein-coding transcripts of genes in the I2HCP panel v2.1 (Supplementary File 4). This data was retrieved from Ensembl Biomart version 67 may2012.archive.ensembl.org). All genes tested by MLPA and used in the benchmark were common to all I2HCP versions (v2.0-2.2).

Evaluation metrics

Performance metrics were performed per gene given that most MLPA kits cover a whole gene and so the true CNVs would be detected by MLPA when confirming any CNV call in any region of interest (ROI) of the affected gene. Therefore, a CNV tool call was defined as TP if one of its ROIs was a TP; FN if MLPA detected a CNV in at least one of its ROIs and none of them were detected by the tool; FP if the tool called a CNV in at least one ROI and none of them were detected by MLPA; TN if neither MLPA or the tool detected a CNV in any of its ROIs.

Runtime

Runtime was calculated by executing CNVfilteR five times on a dataset of 79 gene-panel samples and on the HuRef WGS sample. The runtime calculations were performed on an Intel i5-2450M CPU (4 cores, 2.50 GHz) with 8 GB of RAM and an SSD disk. The median runtime per sample was 0.84 seconds for the gene-panel samples, and 3.60 minutes for the HuRef sample. See the table below for more details.

	Gene-panel samples	**HuRef WGS sample**
Number of samples evaluated	79 samples evaluated at once	1
Number of variants per sample	1554.3 (From VarScan)	4602440 (From Strelka)
Number of CNVs per sample	0.55 (From DECoN)	1362 (From LUMPY)
Total runtime (median value)	66.57 seconds	3.53 minutes
Runtime per sample (median value)	0.84 seconds	3.53 minutes

Summary of CNVfilteR parameters

Parameter	Description	Value used on WGS evaluation (default values)	Value used on gene-panel data evaluation
ht.deletions.threshold	Minimum percentage of heterozygous SNVs in a CNV deletion to filter that CNV	30	=
min.total.depth	SNV minimum total depth	10	30
dup.threshold.score	A CNV duplication is identified as false positive if the sum of the scores of all the heterozygous SNVs in the CNV is equal or greater than the *dup.threshold.score* limit.	0.5	=

margin.pct	Percentage of CNV length, from each CNV limit, where SNVs will be ignored	10	0
homozygous.range	Allele frequency interval at which SNVs are considered homozygous.	[90-100]	=
heterozygous.range	Allele frequency interval at which SNVs are considered heterozygous	[28-72]	=
expected.ht.mean	Expected heterozygous SNV allele frequency	50	=
expected.dup.ht.mean1	Expected heterozygous SNV allele frequency when the variant IS NOT in the same allele as the CNV duplication	33.3	=
expected.dup.ht.mean2	Expected heterozygous SNV allele frequency when the variant IS in the same allele as the CNV duplication	66.6	=
sigmoid.c1	Sigmoid c1 parameter	2	=
sigmoid.c2.vector	Vector containing sigmoid c2 parameters for the six sigmoid functions	(28, 38.3, 44.7, 55.3, 61.3, 71.3)	=

Two parameter values were slightly modified for the gene-panel data evaluation. We used a min.total.depth value of 10 to fit better the sample coverage and the SNV caller used (VarScan), and a margin.pct of 0 because of the small windows (regions of interest) used in gene-panel data.

CNVfilteR use recommendations

CNVfilteR uses SNVs to identify false-positive CNV calls. Therefore, its performance depends on the SNV calls quality. Some considerations can be followed in order to provide reliable SNVs to CNVfilteR:

Low complexity and repetitive regions are genome areas where SNV callers (also CNV callers) perform poorly. If possible, ignore these regions when using CNVfilteR.

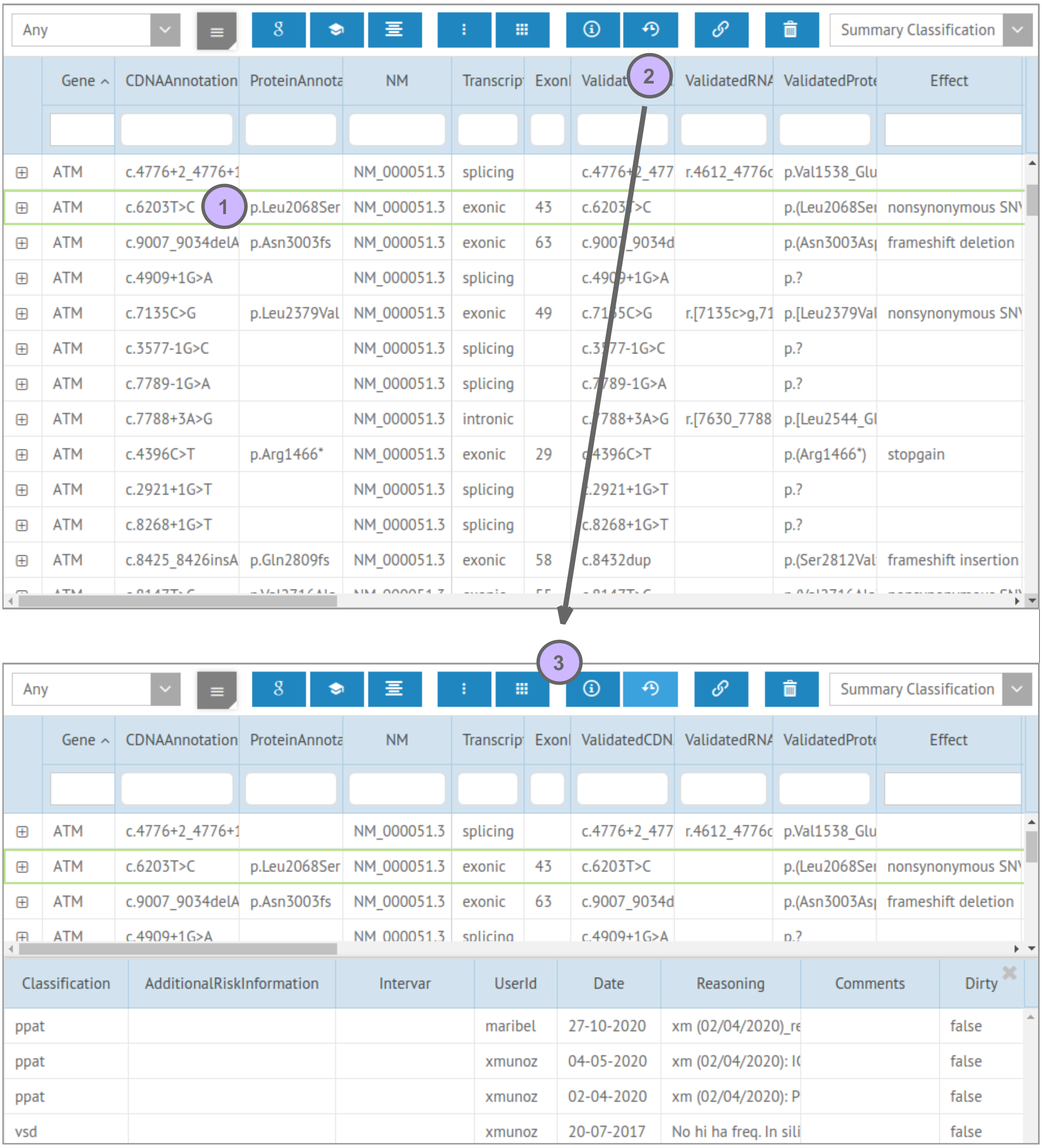
Use the min.depth parameter to discard SNVs with low depth coverage. The default value is 10, which may be appropriate in many WGS samples, but this value should be adapted to your experiment conditions.

Many CNV callers produce inaccurate CNV calls. These inaccurate CNV calls are more likely to be true (to overlap the real CNV) in the middle of the CNV than in the extremes. So, the margin.pct parameter defines the percentage of CNV (from each CNV limit) where SNVs will be ignored. By default, only 10% of SNVs from each CNV extreme will be ignored. This margin.pct parameter can be modified to better adapt it to your CNV caller. For example, we observed that DECoN produced very accurate CNV calls in our genes panel dataset, so margin.pct value was updated to 0 in this context.

A single reliable SNV can be enough to properly identify false-positive CNV calls, so there is no hard low limit on the number of SNVs required by CNVfilteR. Anyway, CNVs with a bigger number of overlapping SNVs are more likely to be correctly identified.

For other use recommendations and how-to-use guide, visit CNVfilteR vignette at https://bioconductor.org/packages/release/bioc/vignettes/CNVfilteR/inst/doc/CNVfilteR.html.

Supplementary Figures - Article 4



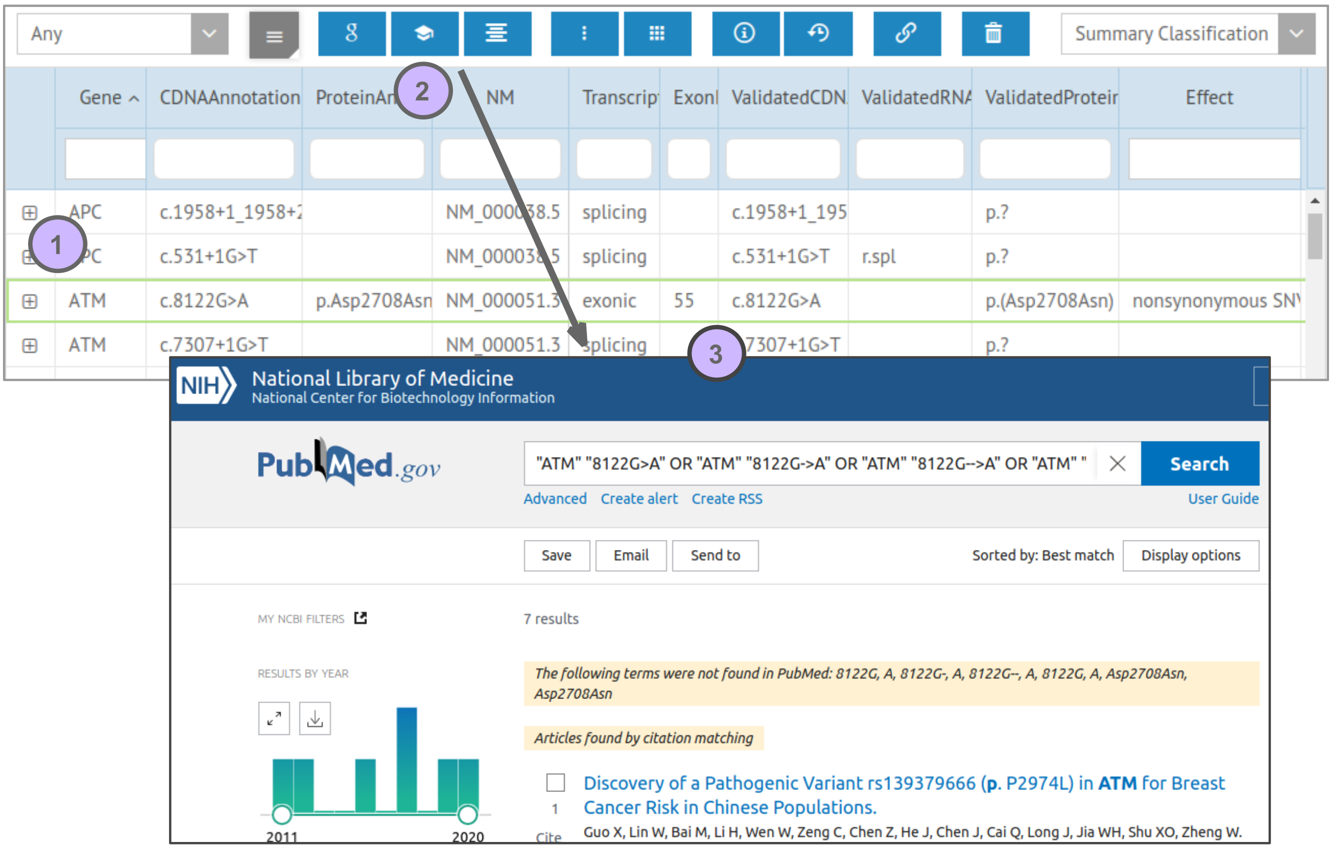
	Gene	CDNAAnnotation	ProteinAnnota	NM	Transcrip	ExonI	Validat	ValidatedRNA	ValidatedProte	Effect
⊞	ATM	c.4776+2_4776+1		NM_000051.3	splicing		c.4776+2_477	r.4612_4776c	p.Val1538_Glu	
⊞	ATM	c.6203T>C	p.Leu2068Ser	NM_000051.3	exonic	43	c.6203T>C		p.(Leu2068Ser	nonsynonymous SN
⊞	ATM	c.9007_9034delA	p.Asn3003fs	NM_000051.3	exonic	63	c.9007_9034d		p.(Asn3003As	frameshift deletion
⊞	ATM	c.4909+1G>A		NM_000051.3	splicing		c.4909+1G>A		p.?	
⊞	ATM	c.7135C>G	p.Leu2379Val	NM_000051.3	exonic	49	c.7135C>G	r.[7135c>g,71	p.[Leu2379Val	nonsynonymous SN
⊞	ATM	c.3577-1G>C		NM_000051.3	splicing		c.3577-1G>C		p.?	
⊞	ATM	c.7789-1G>A		NM_000051.3	splicing		c.7789-1G>A		p.?	
⊞	ATM	c.7788+3A>G		NM_000051.3	intronic		c.7788+3A>G	r.[7630_7788	p.[Leu2544_Gl	
⊞	ATM	c.4396C>T	p.Arg1466*	NM_000051.3	exonic	29	c.4396C>T		p.(Arg1466*)	stopgain
⊞	ATM	c.2921+1G>T		NM_000051.3	splicing		c.2921+1G>T		p.?	
⊞	ATM	c.8268+1G>T		NM_000051.3	splicing		c.8268+1G>T		p.?	
⊞	ATM	c.8425_8426insA	p.Gln2809fs	NM_000051.3	exonic	58	c.8432dup		p.(Ser2812Val	frameshift insertion

	Gene	CDNAAnnotation	ProteinAnnota	NM	Transcrip	ExonI	ValidatedCDN	ValidatedRNA	ValidatedProte	Effect
⊞	ATM	c.4776+2_4776+1		NM_000051.3	splicing		c.4776+2_477	r.4612_4776c	p.Val1538_Glu	
⊞	ATM	c.6203T>C	p.Leu2068Ser	NM_000051.3	exonic	43	c.6203T>C		p.(Leu2068Ser	nonsynonymous SN
⊞	ATM	c.9007_9034delA	p.Asn3003fs	NM_000051.3	exonic	63	c.9007_9034d		p.(Asn3003As	frameshift deletion
⊞	ATM	c.4909+1G>A		NM_000051.3	splicing		c.4909+1G>A		p.?	

Classification	AdditionalRiskInformation	Intervar	UserId	Date	Reasoning	Comments	Dirty
ppat			maribel	27-10-2020	xm (02/04/2020)_re		false
ppat			xmunoz	04-05-2020	xm (02/04/2020): I(		false
ppat			xmunoz	02-04-2020	xm (02/04/2020): P		false
vsd			xmunoz	20-07-2017	No hi ha freq. In sili		false

Supplementary Figure 1 History of variant classifications. **1**: The user selects a variant that becomes highlighted (green box). **2**: The user clicks on the History button. **3**: A subview appears at the bottom of the window containing the classification history.

]



	Gene	CDNAAnnotation	ProteinAn	NM	Transcrip	ExonI	ValidatedCDN	ValidatedRNA	ValidatedProtein	Effect
⊞	APC	c.1958+1_1958+2		NM_000038.5	splicing		c.1958+1_195		p.?	
⊞	PC	c.531+1G>T		NM_000038.5	splicing		c.531+1G>T	r.spl	p.?	
⊞	ATM	c.8122G>A	p.Asp2708Asn	NM_000051.3	exonic	55	c.8122G>A		p.(Asp2708Asn)	nonsynonymous SN
⊞	ATM	c.7307+1G>T		NM_000051.3	splicing		7307+1G>T		p.?	

National Library of Medicine
National Center for Biotechnology Information

PubMed.gov

"ATM" "8122G>A" OR "ATM" "8122G->A" OR "ATM" "8122G-->A" OR "ATM" " Search

Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Best match Display options

MY NCBI FILTERS

RESULTS BY YEAR

2011 2020

7 results

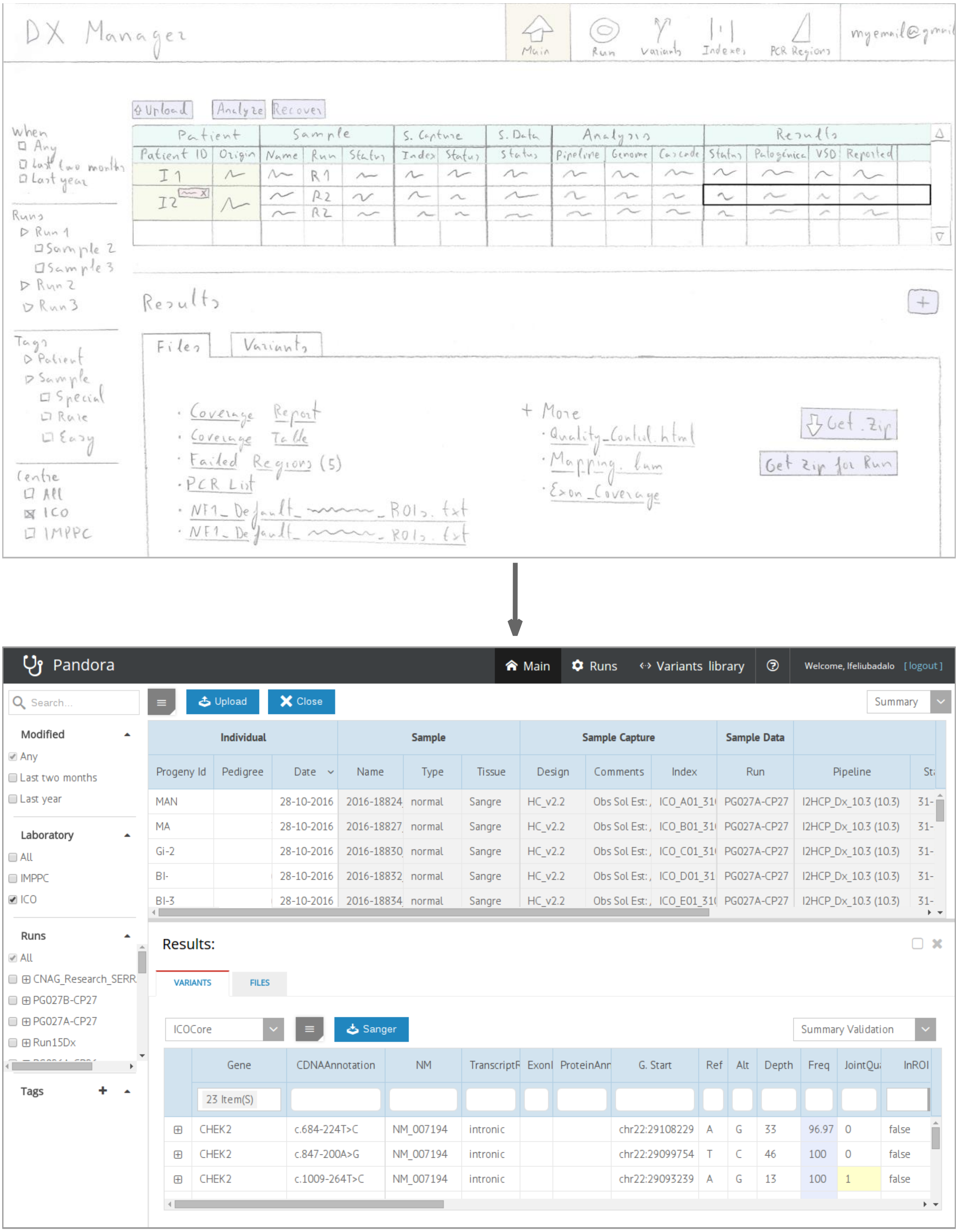
The following terms were not found in PubMed: 8122G, A, 8122G-, A, 8122G--, A, 8122G, A, Asp2708Asn, Asp2708Asn

Articles found by citation matching

□ 1 Discovery of a Pathogenic Variant rs139379666 (**p**. P2974L) in **ATM** for Breast Cancer Risk in Chinese Populations.

Cite Guo X, Lin W, Bai M, Li H, Wen W, Zeng C, Chen Z, He J, Chen J, Cai Q, Long J, Jia WH, Shu XO, Zheng W.

Supplementary Figure 2. Automatic variant search when clicking on the PubMed button. 1: The user selects a variant that becomes highlighted (green box). 2: The user clicks on the PubMed button. 3: A new tab appears containing a PubMed search including several variant nomenclatures.



Modified
☑ Any
□ Last two months
□ Last year

Laboratory
□ All
□ IMPPC
☑ ICO

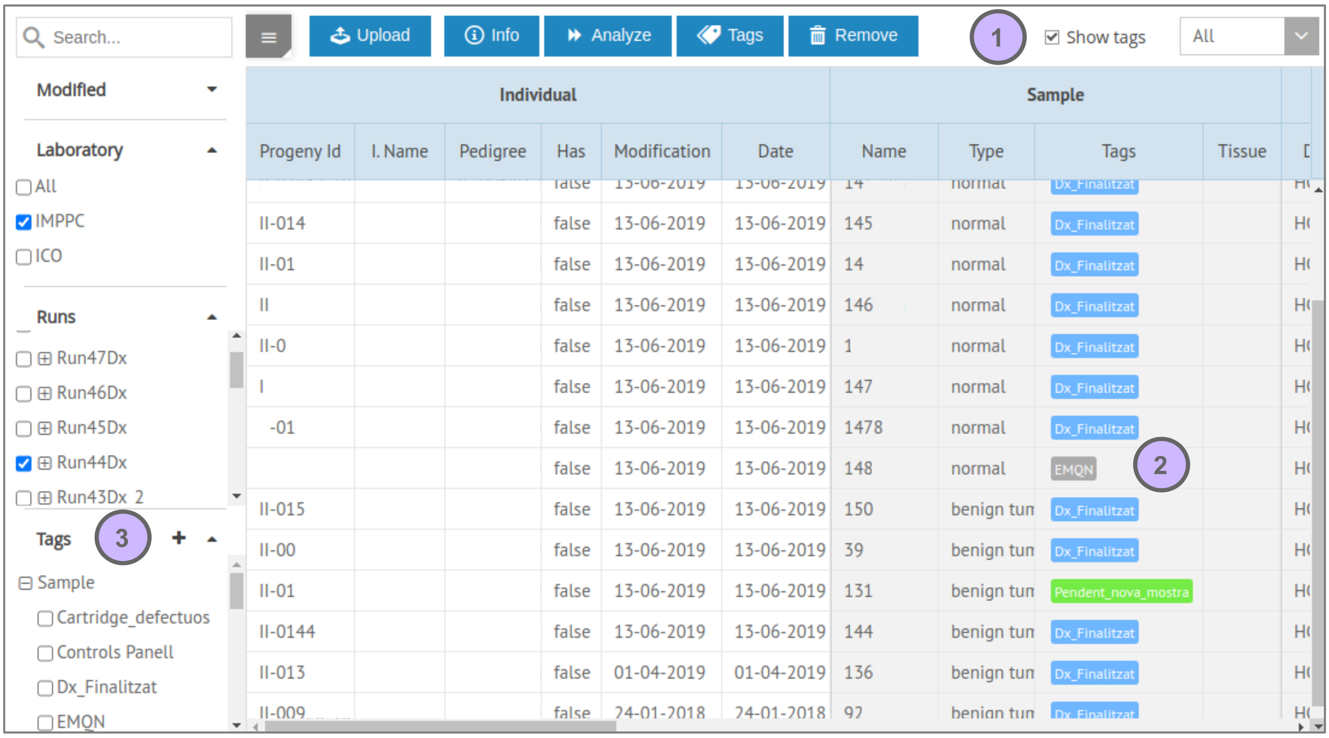
Runs
☑ All
□ ⊞ CNAG_Research_SERR
□ ⊞ PG027B-CP27
□ ⊞ PG027A-CP27
□ ⊞ Run15Dx

Tags

Individual			Sample			Sample Capture			Sample Data		
Progeny Id	Pedigree	Date	Name	Type	Tissue	Design	Comments	Index	Run	Pipeline	St
MAN		28-10-2016	2016-18824	normal	Sangre	HC_v2.2	Obs Sol Est:,	ICO_A01_31	PG027A-CP27	I2HCP_Dx_10.3 (10.3)	31-
MA		28-10-2016	2016-18827	normal	Sangre	HC_v2.2	Obs Sol Est:,	ICO_B01_31	PG027A-CP27	I2HCP_Dx_10.3 (10.3)	31-
Gi-2		28-10-2016	2016-18830	normal	Sangre	HC_v2.2	Obs Sol Est:,	ICO_C01_31	PG027A-CP27	I2HCP_Dx_10.3 (10.3)	31-
BI-		28-10-2016	2016-18832	normal	Sangre	HC_v2.2	Obs Sol Est:,	ICO_D01_31	PG027A-CP27	I2HCP_Dx_10.3 (10.3)	31-
BI-3		28-10-2016	2016-18834	normal	Sangre	HC_v2.2	Obs Sol Est:,	ICO_E01_31	PG027A-CP27	I2HCP_Dx_10.3 (10.3)	31-

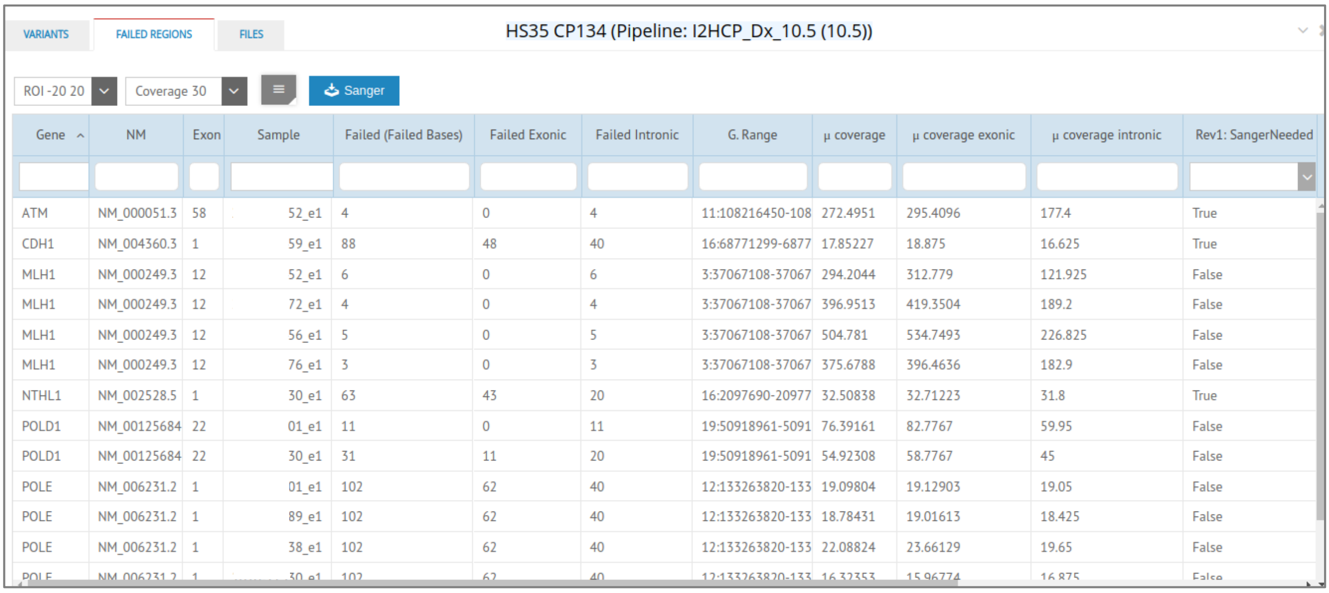
	Gene	CDNAAnnotation	NM	TranscriptF	ExonI	ProteinAnn	G. Start	Ref	Alt	Depth	Freq	JointQu	InROI
⊞	CHEK2	c.684-224T>C	NM_007194	intronic			chr22:29108229	A	G	33	96.97	0	false
⊞	CHEK2	c.847-200A>G	NM_007194	intronic			chr22:29099754	T	C	46	100	0	false
⊞	CHEK2	c.1009-264T>C	NM_007194	intronic			chr22:29093239	A	G	13	100	1	false

Supplementary Figure 3. Pandora development from the sketch (top) to the current implementation (bottom).



Individual						Sample				
Progeny Id	I. Name	Pedigree	Has	Modification	Date	Name	Type	Tags	Tissue	D
			false	13-06-2019	13-06-2019	14	normal	Dx_Finalitzat		H
II-014			false	13-06-2019	13-06-2019	145	normal	Dx_Finalitzat		H
II-01			false	13-06-2019	13-06-2019	14	normal	Dx_Finalitzat		H
II			false	13-06-2019	13-06-2019	146	normal	Dx_Finalitzat		H
II-0			false	13-06-2019	13-06-2019	1	normal	Dx_Finalitzat		H
I			false	13-06-2019	13-06-2019	147	normal	Dx_Finalitzat		H
-01			false	13-06-2019	13-06-2019	1478	normal	Dx_Finalitzat		H
			false	13-06-2019	13-06-2019	148	normal	EMQN		H
II-015			false	13-06-2019	13-06-2019	150	benign tun	Dx_Finalitzat		H
II-00			false	13-06-2019	13-06-2019	39	benign tun	Dx_Finalitzat		H
II-01			false	13-06-2019	13-06-2019	131	benign tun	Pendent_nova_mostra		H
II-0144			false	13-06-2019	13-06-2019	144	benign tun	Dx_Finalitzat		H
II-013			false	01-04-2019	01-04-2019	136	benign tun	Dx_Finalitzat		H
II-009			false	24-01-2018	24-01-2018	92	benign tun	Dx_Finalitzat		H

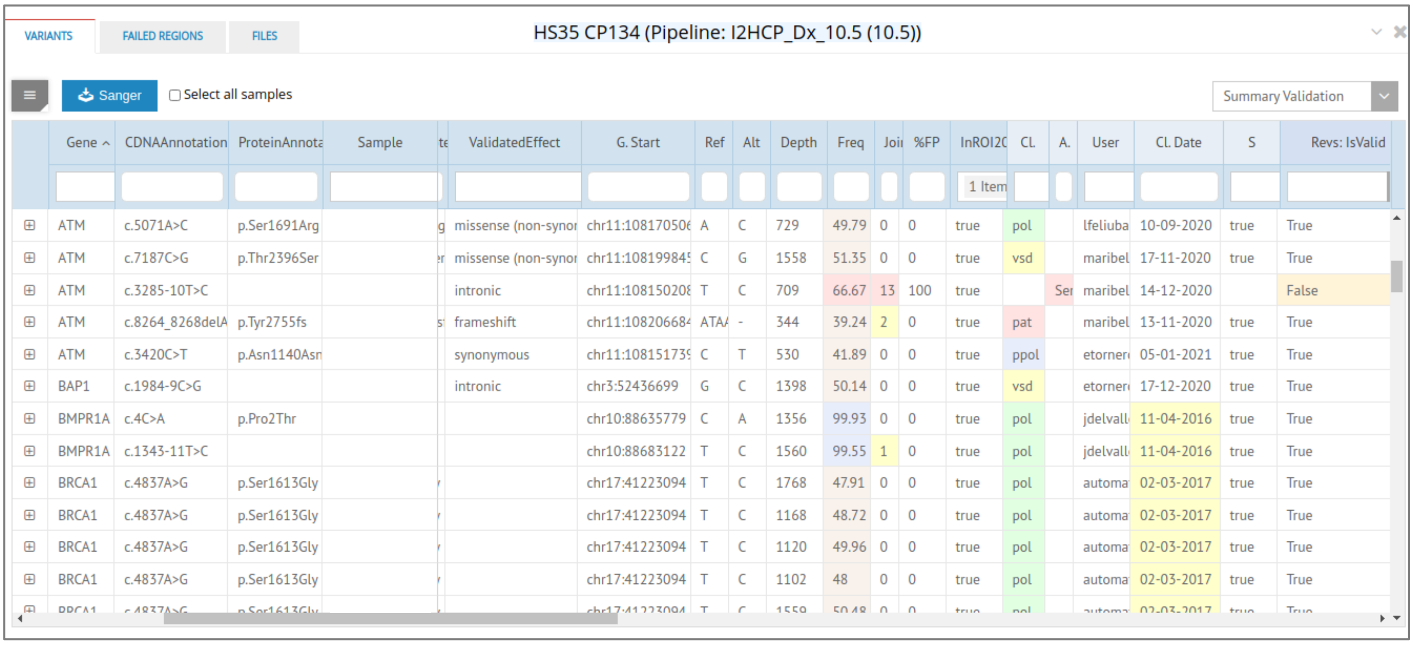
Supplementary Figure 4. Use of tags in Pandora. **1**: Tags are optional and can be enabled or disabled using a checkbox. **2**: Tags are shown in a column using its user-defined color. **3**: Rows can be filtered by tags. Also, tags can be modified, removed, or created.



HS35 CP134 (Pipeline: I2HCP_Dx_10.5 (10.5))

Gene	NM	Exon	Sample	Failed (Failed Bases)	Failed Exonic	Failed Intronic	G. Range	µ coverage	µ coverage exonic	µ coverage intronic	Rev1: SangerNeeded
ATM	NM_000051.3	58	52_e1	4	0	4	11:108216450-108	272.4951	295.4096	177.4	True
CDH1	NM_004360.3	1	59_e1	88	48	40	16:68771299-6877	17.85227	18.875	16.625	True
MLH1	NM_000249.3	12	52_e1	6	0	6	3:37067108-37067	294.2044	312.779	121.925	False
MLH1	NM_000249.3	12	72_e1	4	0	4	3:37067108-37067	396.9513	419.3504	189.2	False
MLH1	NM_000249.3	12	56_e1	5	0	5	3:37067108-37067	504.781	534.7493	226.825	False
MLH1	NM_000249.3	12	76_e1	3	0	3	3:37067108-37067	375.6788	396.4636	182.9	False
NTHL1	NM_002528.5	1	30_e1	63	43	20	16:2097690-20977	32.50838	32.71223	31.8	True
POLD1	NM_00125684	22	01_e1	11	0	11	19:50918961-5091	76.39161	82.7767	59.95	False
POLD1	NM_00125684	22	30_e1	31	11	20	19:50918961-5091	54.92308	58.7767	45	False
POLE	NM_006231.2	1	01_e1	102	62	40	12:133263820-133	19.09804	19.12903	19.05	False
POLE	NM_006231.2	1	89_e1	102	62	40	12:133263820-133	18.78431	19.01613	18.425	False
POLE	NM_006231.2	1	38_e1	102	62	40	12:133263820-133	22.08824	23.66129	19.65	False
POLE	NM_006231.2	1	30_e1	102	62	40	12:133263820-133	16.32353	15.96774	16.875	False

Supplementary Figure 5. Low-coverage genomic regions view. In the image, low-coverage bases are reported for each sample in the run HS35.

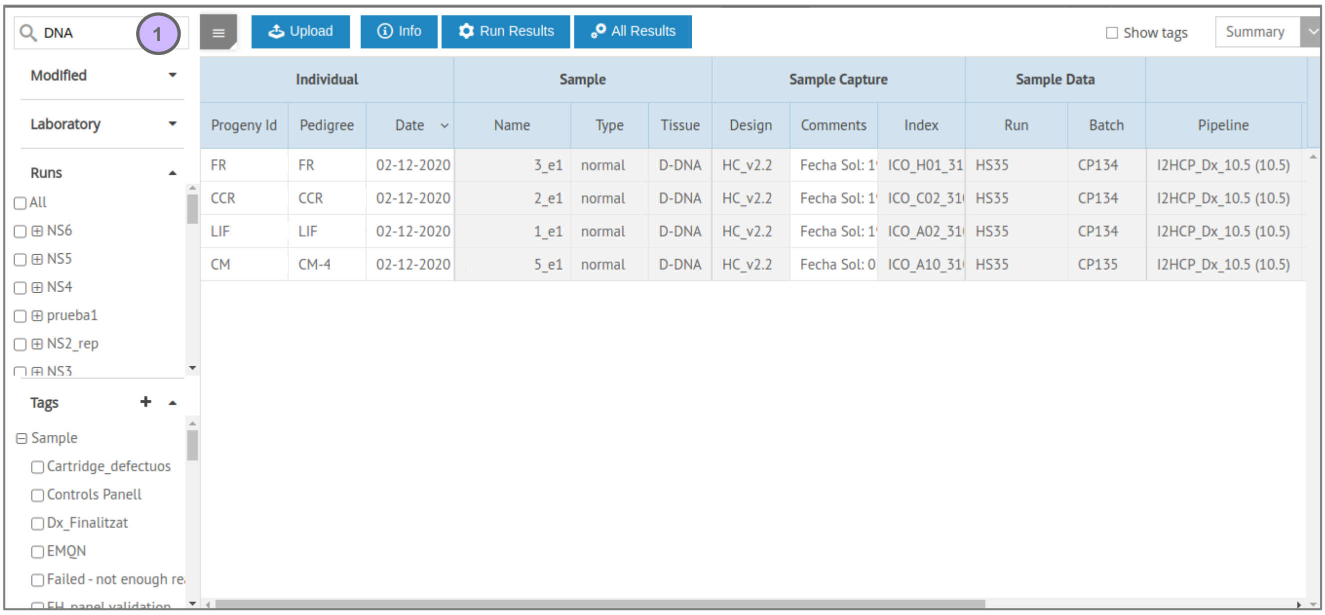


HS35 CP134 (Pipeline: I2HCP_Dx_10.5 (10.5))

□ Select all samples

	Gene	CDNAAnnotation	ProteinAnnota	Sample	te	ValidatedEffect	G. Start	Ref	Alt	Depth	Freq	Joi	%FP	InROI2C	CL.	A.	User	CL. Date	S	Revs: IsValid
														1 Item						
⊞	ATM	c.5071A>C	p.Ser1691Arg		g	missense (non-synor	chr11:108170506	A	C	729	49.79	0	0	true	pol		lfeliuba	10-09-2020	true	True
⊞	ATM	c.7187C>G	p.Thr2396Ser		er	missense (non-synor	chr11:108199845	C	G	1558	51.35	0	0	true	vsd		maribel	17-11-2020	true	True
⊞	ATM	c.3285-10T>C				intronic	chr11:108150208	T	C	709	66.67	13	100	true		Ser	maribel	14-12-2020		False
⊞	ATM	c.8264_8268delA	p.Tyr2755fs		st	frameshift	chr11:108206684	ATAA	-	344	39.24	2	0	true	pat		maribel	13-11-2020	true	True
⊞	ATM	c.3420C>T	p.Asn1140Asn			synonymous	chr11:108151739	C	T	530	41.89	0	0	true	ppol		etorner	05-01-2021	true	True
⊞	BAP1	c.1984-9C>G				intronic	chr3:52436699	G	C	1398	50.14	0	0	true	vsd		etorner	17-12-2020	true	True
⊞	BMPR1A	c.4C>A	p.Pro2Thr				chr10:88635779	C	A	1356	99.93	0	0	true	pol		jdelvall	11-04-2016	true	True
⊞	BMPR1A	c.1343-11T>C					chr10:88683122	T	C	1560	99.55	1	0	true	pol		jdelvall	11-04-2016	true	True
⊞	BRCA1	c.4837A>G	p.Ser1613Gly		'		chr17:41223094	T	C	1768	47.91	0	0	true	pol		automa	02-03-2017	true	True
⊞	BRCA1	c.4837A>G	p.Ser1613Gly		'		chr17:41223094	T	C	1168	48.72	0	0	true	pol		automa	02-03-2017	true	True
⊞	BRCA1	c.4837A>G	p.Ser1613Gly		'		chr17:41223094	T	C	1120	49.96	0	0	true	pol		automa	02-03-2017	true	True
⊞	BRCA1	c.4837A>G	p.Ser1613Gly		'		chr17:41223094	T	C	1102	48	0	0	true	pol		automa	02-03-2017	true	True
⊞	BRCA1	c.4837A>G	p.Ser1613Gly		'		chr17:41223094	T	C	1559	50.48	0	0	true	pol		automa	02-03-2017	true	True

Supplementary Figure 6. Use of colors in Pandora. **Freq column:** blue for homozygous variants (90-100%), brown for heterozygous variants (35-65%), and red for heterozygous variants with unexpected allele frequency (< 35% and 65-90%). **Cl. (Classification) column:** green for benign variants (pol), soft blue for likely benign (ppol), yellow for variants of unknown significance (vsd), orange for likely pathogenic variants (ppat), and red for pathogenic variants (pat). **A. (Additional risk information) column:** red if any risk note exists for this variant, white otherwise. **Cl. (Classification) Date column:** yellow if the last classification was more than six months ago, white otherwise.



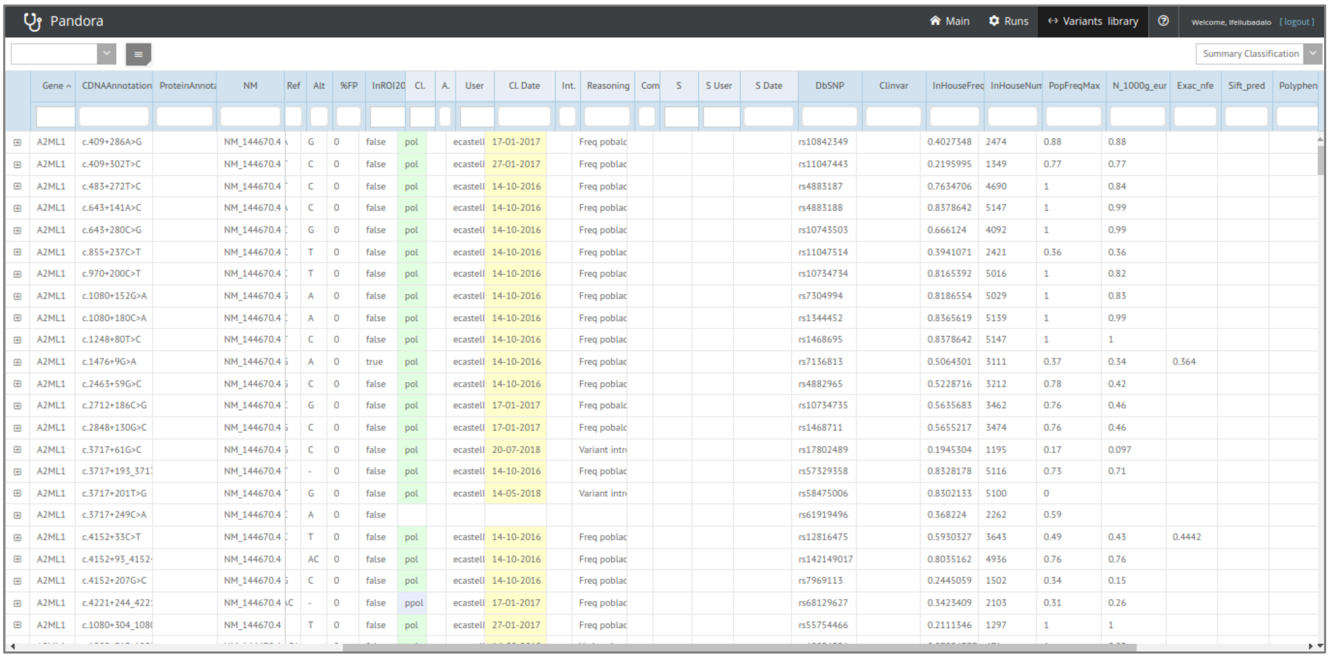
□ Show tags

□ All
□ ⊞ NS6
□ ⊞ NS5
□ ⊞ NS4
□ ⊞ prueba1
□ ⊞ NS2_rep
□ ⊞ NS3

⊟ Sample
□ Cartridge_defectuos
□ Controls Panell
□ Dx_Finalitzat
□ EMQN
□ Failed - not enough re
□ FH panel validation

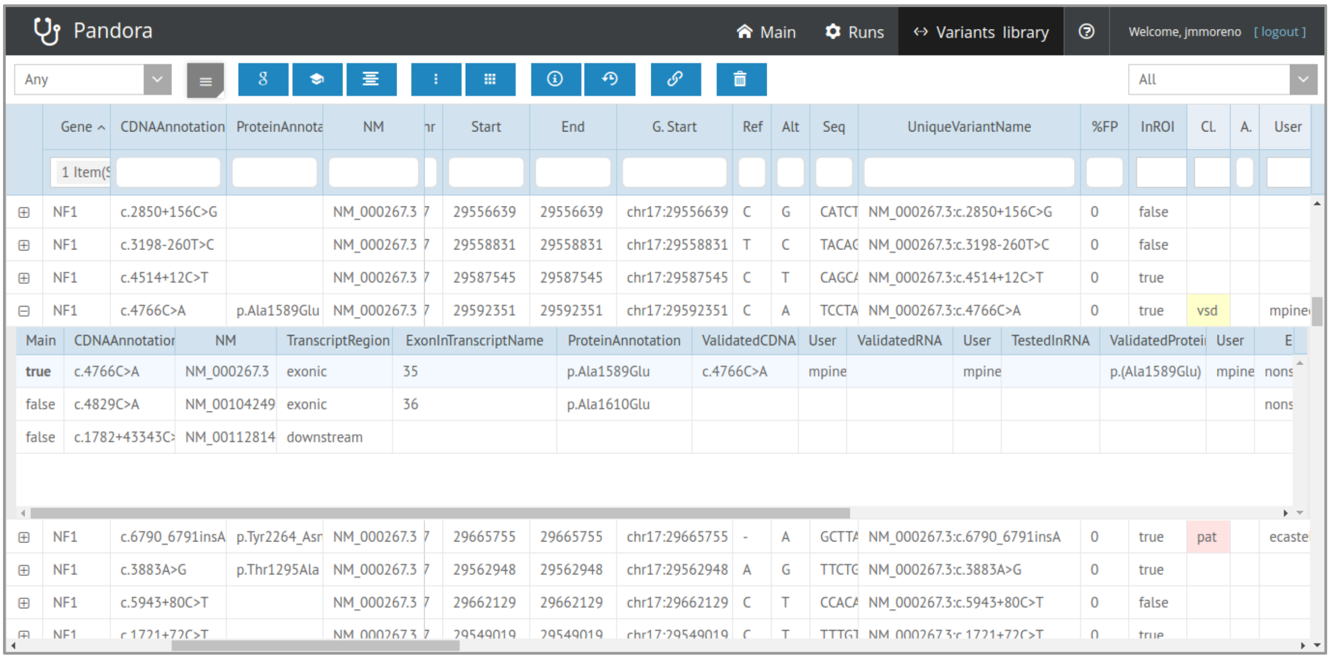
Individual			Sample			Sample Capture			Sample Data		
Progeny Id	Pedigree	Date	Name	Type	Tissue	Design	Comments	Index	Run	Batch	Pipeline
FR	FR	02-12-2020	3_e1	normal	D-DNA	HC_v2.2	Fecha Sol: 1	ICO_H01_31	HS35	CP134	I2HCP_Dx_10.5 (10.5)
CCR	CCR	02-12-2020	2_e1	normal	D-DNA	HC_v2.2	Fecha Sol: 1	ICO_C02_31	HS35	CP134	I2HCP_Dx_10.5 (10.5)
LIF	LIF	02-12-2020	1_e1	normal	D-DNA	HC_v2.2	Fecha Sol: 1	ICO_A02_31	HS35	CP134	I2HCP_Dx_10.5 (10.5)
CM	CM-4	02-12-2020	5_e1	normal	D-DNA	HC_v2.2	Fecha Sol: 0	ICO_A10_31	HS35	CP135	I2HCP_Dx_10.5 (10.5)

Supplementary Figure 7. Search box allows easy search for any field in the table.



	Gene	CDNAAnnotation	ProteinAnnota	NM	Ref	Alt	%FP	InROI20	CL.	A.	User	CL Date	Int.	Reasoning	Com	S	S User	S Date	DbSNP	Clinvar	InHouseFreq	InHouseNum	PopFreqMax	N_1000g_eur	Exac_nfe	Sift_pred	Polyphen
⊞	A2ML1	c.409+286A>G		NM_144670.4		G	0	false	pol		ecastell	17-01-2017		Freq pobalc					rs10842349		0.4027348	2474	0.88	0.88			
⊞	A2ML1	c.409+302T>C		NM_144670.4		C	0	false	pol		ecastell	27-01-2017		Freq poblac					rs11047443		0.2195995	1349	0.77	0.77			
⊞	A2ML1	c.483+272T>C		NM_144670.4		C	0	false	pol		ecastell	14-10-2016		Freq poblac					rs4883187		0.7634706	4690	1	0.84			
⊞	A2ML1	c.643+141A>C		NM_144670.4		C	0	false	pol		ecastell	14-10-2016		Freq poblac					rs4883188		0.8378642	5147	1	0.99			
⊞	A2ML1	c.643+280C>G		NM_144670.4		G	0	false	pol		ecastell	14-10-2016		Freq poblac					rs10743503		0.666124	4092	1	0.99			
⊞	A2ML1	c.855+237C>T		NM_144670.4		T	0	false	pol		ecastell	14-10-2016		Freq poblac					rs11047514		0.3941071	2421	0.36	0.36			
⊞	A2ML1	c.970+200C>T		NM_144670.4		T	0	false	pol		ecastell	14-10-2016		Freq poblac					rs10734734		0.8165392	5016	1	0.82			
⊞	A2ML1	c.1080+152G>A		NM_144670.4		A	0	false	pol		ecastell	14-10-2016		Freq poblac					rs7304994		0.8186554	5029	1	0.83			
⊞	A2ML1	c.1080+180C>A		NM_144670.4		A	0	false	pol		ecastell	14-10-2016		Freq poblac					rs1344452		0.8365619	5139	1	0.99			
⊞	A2ML1	c.1248+80T>C		NM_144670.4		C	0	false	pol		ecastell	14-10-2016		Freq poblac					rs1468695		0.8378642	5147	1	1			
⊞	A2ML1	c.1476+9G>A		NM_144670.4		A	0	true	pol		ecastell	14-10-2016		Freq poblac					rs7136813		0.5064301	3111	0.37	0.34	0.364		
⊞	A2ML1	c.2463+59G>C		NM_144670.4		C	0	false	pol		ecastell	14-10-2016		Freq poblac					rs4882965		0.5228716	3212	0.78	0.42			
⊞	A2ML1	c.2712+186C>G		NM_144670.4		G	0	false	pol		ecastell	17-01-2017		Freq pobalc					rs10734735		0.5635683	3462	0.76	0.46			
⊞	A2ML1	c.2848+130G>C		NM_144670.4		C	0	false	pol		ecastell	17-01-2017		Freq pobalc					rs1468711		0.5655217	3474	0.76	0.46			
⊞	A2ML1	c.3717+61G>C		NM_144670.4		C	0	false	pol		ecastell	20-07-2018		Variant intr					rs17802489		0.1945304	1195	0.17	0.097			
⊞	A2ML1	c.3717+193_371		NM_144670.4		-	0	false	pol		ecastell	14-10-2016		Freq poblac					rs57329358		0.8328178	5116	0.73	0.71			
⊞	A2ML1	c.3717+201T>G		NM_144670.4		G	0	false	pol		ecastell	14-05-2018		Variant intr					rs58475006		0.8302133	5100	0				
⊞	A2ML1	c.3717+249C>A		NM_144670.4		A	0	false											rs61919496		0.368224	2262	0.59				
⊞	A2ML1	c.4152+33C>T		NM_144670.4		T	0	false	pol		ecastell	14-10-2016		Freq poblac					rs12816475		0.5930327	3643	0.49	0.43	0.4442		
⊞	A2ML1	c.4152+93_4152-		NM_144670.4		AC	0	false	pol		ecastell	14-10-2016		Freq poblac					rs142149017		0.8035162	4936	0.76	0.76			
⊞	A2ML1	c.4152+207G>C		NM_144670.4		C	0	false	pol		ecastell	14-10-2016		Freq poblac					rs7969113		0.2445059	1502	0.34	0.15			
⊞	A2ML1	c.4221+244_422		NM_144670.4	C	-	0	false	ppol		ecastell	17-01-2017		Freq poblac					rs68129627		0.3423409	2103	0.31	0.26			
⊞	A2ML1	c.1080+304_108		NM_144670.4		T	0	false	pol		ecastell	27-01-2017		Freq poblac					rs55754466		0.2111346	1297	1	1			

Supplementary Figure 8. Variants Library shows all the variants that have been found in Pandora. The user can quickly explore 4.39 million fields, sorting and filtering by several columns.

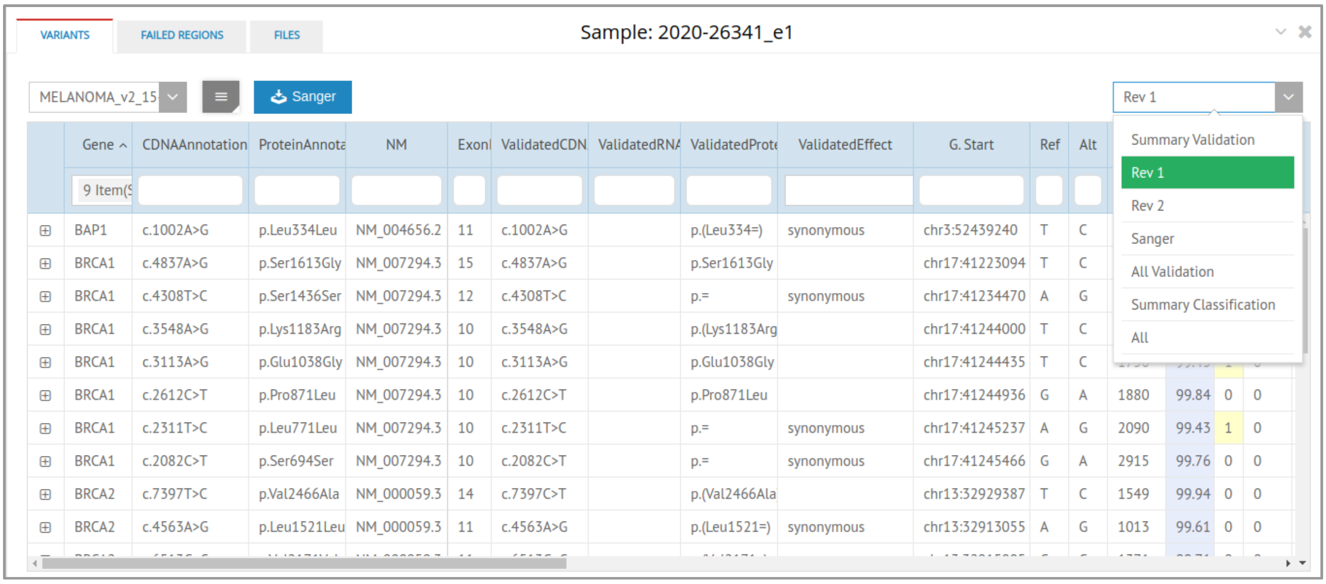


	Gene	CDNAAnnotation	ProteinAnnota	NM	hr	Start	End	G. Start	Ref	Alt	Seq	UniqueVariantName	%FP	InROI	CL.	A.	User
	1 Item(S																
⊞	NF1	c.2850+156C>G		NM_000267.3	7	29556639	29556639	chr17:29556639	C	G	CATCT	NM_000267.3:c.2850+156C>G	0	false			
⊞	NF1	c.3198-260T>C		NM_000267.3	7	29558831	29558831	chr17:29558831	T	C	TACAG	NM_000267.3:c.3198-260T>C	0	false			
⊞	NF1	c.4514+12C>T		NM_000267.3	7	29587545	29587545	chr17:29587545	C	T	CAGCA	NM_000267.3:c.4514+12C>T	0	true			
⊟	NF1	c.4766C>A	p.Ala1589Glu	NM_000267.3	7	29592351	29592351	chr17:29592351	C	A	TCCTA	NM_000267.3:c.4766C>A	0	true	vsd		mpine

Main	CDNAAnnotation	NM	TranscriptRegion	ExonInTranscriptName	ProteinAnnotation	ValidatedCDNA	User	ValidatedRNA	User	TestedInRNA	ValidatedProtei	User	E
true	c.4766C>A	NM_000267.3	exonic	35	p.Ala1589Glu	c.4766C>A	mpine		mpine		p.(Ala1589Glu)	mpine	nons
false	c.4829C>A	NM_00104249	exonic	36	p.Ala1610Glu								nons
false	c.1782+43343C>	NM_00112814	downstream										

	Gene	CDNAAnnotation	ProteinAnnota	NM	hr	Start	End	G. Start	Ref	Alt	Seq	UniqueVariantName	%FP	InROI	CL.	A.	User
⊞	NF1	c.6790_6791insA	p.Tyr2264_Asr	NM_000267.3	7	29665755	29665755	chr17:29665755	-	A	GCTTA	NM_000267.3:c.6790_6791insA	0	true	pat		ecaste
⊞	NF1	c.3883A>G	p.Thr1295Ala	NM_000267.3	7	29562948	29562948	chr17:29562948	A	G	TTCTG	NM_000267.3:c.3883A>G	0	true			
⊞	NF1	c.5943+80C>T		NM_000267.3	7	29662129	29662129	chr17:29662129	C	T	CCACA	NM_000267.3:c.5943+80C>T	0	false			
⊞	NF1	c.1721+72C>T		NM_000267.3	7	29549019	29549019	chr17:29549019	C	T	TTTG1	NM_000267.3:c.1721+72C>T	0	true			

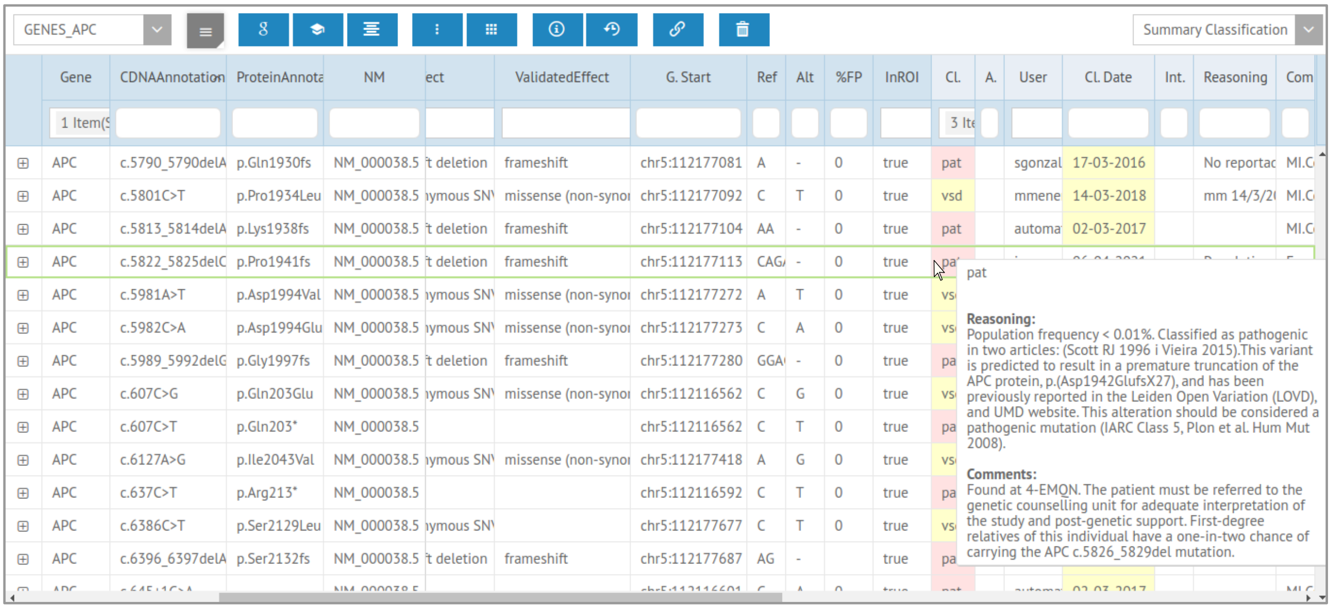
Supplementary Figure 9. For each variant, a subview can be unfolded to show the variant nomenclature for all the gene transcripts.



Sample: 2020-26341_e1

	Gene	CDNAAnnotation	ProteinAnnota	NM	ExonI	ValidatedCDN	ValidatedRNA	ValidatedProte	ValidatedEffect	G. Start	Ref	Alt				
	9 Item(S															
⊞	BAP1	c.1002A>G	p.Leu334Leu	NM_004656.2	11	c.1002A>G		p.(Leu334=)	synonymous	chr3:52439240	T	C				
⊞	BRCA1	c.4837A>G	p.Ser1613Gly	NM_007294.3	15	c.4837A>G		p.Ser1613Gly		chr17:41223094	T	C				
⊞	BRCA1	c.4308T>C	p.Ser1436Ser	NM_007294.3	12	c.4308T>C		p.=	synonymous	chr17:41234470	A	G				
⊞	BRCA1	c.3548A>G	p.Lys1183Arg	NM_007294.3	10	c.3548A>G		p.(Lys1183Arg		chr17:41244000	T	C				
⊞	BRCA1	c.3113A>G	p.Glu1038Gly	NM_007294.3	10	c.3113A>G		p.Glu1038Gly		chr17:41244435	T	C				
⊞	BRCA1	c.2612C>T	p.Pro871Leu	NM_007294.3	10	c.2612C>T		p.Pro871Leu		chr17:41244936	G	A	1880	99.84	0	0
⊞	BRCA1	c.2311T>C	p.Leu771Leu	NM_007294.3	10	c.2311T>C		p.=	synonymous	chr17:41245237	A	G	2090	99.43	1	0
⊞	BRCA1	c.2082C>T	p.Ser694Ser	NM_007294.3	10	c.2082C>T		p.=	synonymous	chr17:41245466	G	A	2915	99.76	0	0
⊞	BRCA2	c.7397T>C	p.Val2466Ala	NM_000059.3	14	c.7397C>T		p.(Val2466Ala		chr13:32929387	T	C	1549	99.94	0	0
⊞	BRCA2	c.4563A>G	p.Leu1521Leu	NM_000059.3	11	c.4563A>G		p.(Leu1521=)	synonymous	chr13:32913055	A	G	1013	99.61	0	0

Supplementary Figure 10. The set of columns shown can be selected depending on the task to be done.



	Gene	CDNAAnnotation	ProteinAnnota	NM	ect	ValidatedEffect	G. Start	Ref	Alt	%FP	InROI	Cl.	A.	User	Cl. Date	Int.	Reasoning	Com
	1 Item(S											3 Ite						
⊞	APC	c.5790_5790delA	p.Gln1930fs	NM_000038.5	t deletion	frameshift	chr5:112177081	A	-	0	true	pat		sgonzal	17-03-2016		No reportac	MI.C
⊞	APC	c.5801C>T	p.Pro1934Leu	NM_000038.5	ıymous SN\	missense (non-synoı	chr5:112177092	C	T	0	true	vsd		mmene	14-03-2018		mm 14/3/2	MI.C
⊞	APC	c.5813_5814delA	p.Lys1938fs	NM_000038.5	t deletion	frameshift	chr5:112177104	AA	-	0	true	pat		automa	02-03-2017			MI.C
⊞	APC	c.5822_5825delC	p.Pro1941fs	NM_000038.5	t deletion	frameshift	chr5:112177113	CAG/	-	0	true	pa						
⊞	APC	c.5981A>T	p.Asp1994Val	NM_000038.5	ıymous SN\	missense (non-synoı	chr5:112177272	A	T	0	true	vs						
⊞	APC	c.5982C>A	p.Asp1994Glu	NM_000038.5	ıymous SN\	missense (non-synoı	chr5:112177273	C	A	0	true	vs						
⊞	APC	c.5989_5992delG	p.Gly1997fs	NM_000038.5	t deletion	frameshift	chr5:112177280	GGA	-	0	true	pa						
⊞	APC	c.607C>G	p.Gln203Glu	NM_000038.5	ıymous SN\	missense (non-synoı	chr5:112116562	C	G	0	true	vs						
⊞	APC	c.607C>T	p.Gln203*	NM_000038.5			chr5:112116562	C	T	0	true	pa						
⊞	APC	c.6127A>G	p.Ile2043Val	NM_000038.5	ıymous SN\	missense (non-synoı	chr5:112177418	A	G	0	true	vs						
⊞	APC	c.637C>T	p.Arg213*	NM_000038.5			chr5:112116592	C	T	0	true	pa						
⊞	APC	c.6386C>T	p.Ser2129Leu	NM_000038.5	ıymous SN\		chr5:112177677	C	T	0	true	vs						
⊞	APC	c.6396_6397delA	p.Ser2132fs	NM_000038.5	t deletion	frameshift	chr5:112177687	AG	-		true	pa						

pat

Reasoning:
Population frequency < 0.01%. Classified as pathogenic in two articles: (Scott RJ 1996 i Vieira 2015).This variant is predicted to result in a premature truncation of the APC protein, p.(Asp1942GlufsX27), and has been previously reported in the Leiden Open Variation (LOVD), and UMD website. This alteration should be considered a pathogenic mutation (IARC Class 5, Plon et al. Hum Mut 2008).

Comments:
Found at 4-EMQN. The patient must be referred to the genetic counselling unit for adequate interpretation of the study and post-genetic support. First-degree relatives of this individual have a one-in-two chance of carrying the APC c.5826_5829del mutation.

Supplementary Figure 11. Complex tooltip that includes large relevant fields (Reasoning and Comments) when the mouse is over the Cl. (Classification) column.