NAR Genomics and Bioinformatics, 2024, **6**, lqae033
https://doi.org/10.1093/nargab/lqae033
Advance access publication date: 17 April 2024
Standard Article

A method to comprehensively identify germline SNVs, INDELs and CNVs from whole exome sequencing data of BRCA1/2 negative breast cancer patients

Andrea Bianchi[1,†], Veronica Zelli[2,*,†], Andrea D'Angelo[1], Alessandro Di Matteo[1], Giulia Scoccia[1], Katia Cannita[3], Antigone S Dimas[4], Stavros Glentis[4,5], Francesca Zazzeroni[2], Edoardo Alesse[2], Antinisca Di Marco[1] and Alessandra Tessitore[2]

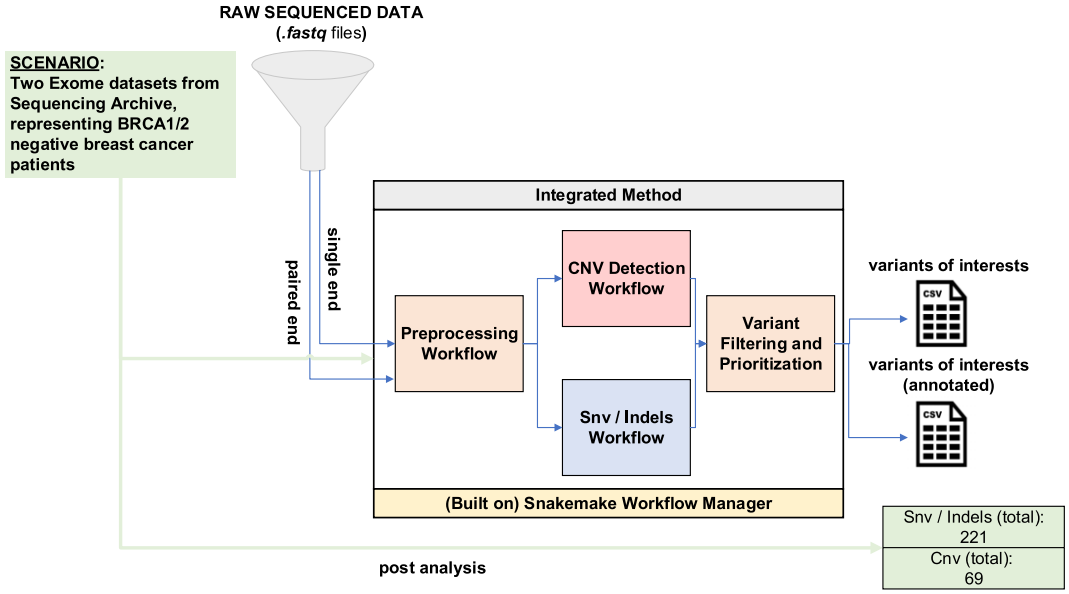
[1]Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila 67100, Italy
[2]Department of Biotechnological and Applied Clinical Sciences, University of L'Aquila, L'Aquila 67100, Italy
[3]Oncology Division, Mazzini Hospital, ASL Teramo, Teramo 64100, Italy
[4]Institute for Bioinnovation, Biomedical Sciences Research Center, Alexander Fleming, Vari 16672, Greece
[5]Pediatric Hematology/Oncology Unit (POHemU), First Department of Pediatrics, University of Athens, Aghia Sophia Children's Hospital, Athens 11527, Grece

[*]To whom correspondence should be addressed. Tel: +39 0862 433506; Fax: +39 0862 433523; Email: veronica.zelli@univaq.it
[†]The first two authors should be regarded as Joint First Authors.

Abstract

In the rapidly evolving field of genomics, understanding the genetic basis of complex diseases like breast cancer, particularly its familial/hereditary forms, is crucial. Current methods often examine genomic variants—such as Single Nucleotide Variants (SNVs), insertions/deletions (Indels), and Copy Number Variations (CNVs)—separately, lacking an integrated approach. Here, we introduced a robust, flexible methodology for a comprehensive variants' analysis using Whole Exome Sequencing (WES) data. Our approach uniquely combines meticulous validation with an effective variant filtering strategy. By reanalyzing two germline WES datasets from *BRCA1/2* negative breast cancer patients, we demonstrated our tool's efficiency and adaptability, uncovering both known and novel variants. This contributed new insights for potential diagnostic, preventive, and therapeutic strategies. Our method stands out for its comprehensive inclusion of key genomic variants in a unified analysis, and its practical resolution of technical challenges, offering a pioneering solution in genomic research. This tool presents a breakthrough in providing detailed insights into the genetic alterations in genomes, with significant implications for understanding and managing hereditary breast cancer.

Graphical abstract



Received: November 29, 2023. **Revised:** March 22, 2024. **Editorial Decision:** March 27, 2024. **Accepted:** April 3, 2024

Introduction

In the modern era of genomics, the capability to perform a comprehensive analysis of genetic alterations is crucial in deciphering the genetic architecture of complex diseases, in particular cancer (1). Among these, Single Nucleotide Variants (SNVs), insertions and deletions (Indels) and Copy Number Variations (CNVs) are critically important for understanding pathogenesis and refining diagnosis (2). Recent studies have underscored the importance of an integrated approach to analyze these genomic alterations (3–5). However, a common shortfall is their focus on either SNVs/Indels or CNVs separately, thus omitting a comprehensive pipeline that addresses all variants simultaneously (6). Moreover, the lack of available codes for reproducibility further complicates the landscape for researchers, especially those not well-versed in bioinformatics. It is noteworthy that in different laboratories, it is common to use in-house pipelines for the detection of genetic alterations (7) or https://www.alamyahealth.com/next-era-whole-genome-sequencing/. However, these pipelines can often be proprietary, with settings and algorithms that are not fully accessible to the broader scientific community (8). This lack of full accessibility poses significant challenges, as it hinders the ability to validate, reproduce, and potentially improve upon these methodologies. Existing literature reflects the struggle in harmonizing the analysis of these variants, especially when relying on custom solutions or established platforms like Galaxy (https://usegalaxy.org/), which may not always meet specific research needs. This gap in genomic research highlights the need for a user-friendly, integrated approach. Our work addresses this need by introducing a novel pipeline that encapsulates the analysis of SNVs, Indels, and CNVs within a single framework. This integrated approach is not only comprehensive but also ensures the reproducibility of results, a crucial aspect often overlooked in existing studies. Our study leverages online archives to re-analyze Next Generation Sequencing (NGS) data. We specifically focus on Whole Exome Sequencing (WES) datasets from familial/hereditary breast cancer (BC) patients who do not show *BRCA1/2* mutations (non-BRCA). This targeted approach allows us to test our methodology in a well-defined and controlled setting. It is known that a large fraction of familial BC cases lack lesions at the level of the most penetrant susceptibility genes *BRCA1/2*, and, despite the identification of other less-penetrant genes (e.g. *PALB2, CHEK2, ATM, BARD, RAD51*) (9), approximately 70–75% of familial cases remains unexplained at the genetic level (10). Therefore, there is an urgent need to identify novel predisposing factors whose genetic variants could explain familial BC susceptibility in non-BRCA patients.

Materials and methods

Methodological approach

We constructed an integrated approach aimed at the simultaneous identification of SNVs, Indels and CNVs within WES data. This holistic approach encompasses a pipeline dedicated to the identification of germline SNVs and Indels, alongside a parallel pipeline aimed at determining CNVs. Despite their distinct end goals, both pipelines share common preliminary steps which ensure uniform processing of the data. Although the tools used within our pipelines are crucial in defining the existing internal workflows, they can be quickly replaced or changed with only minor configuration modifications. This adaptability extends to the references used, including genomes, targeted sections, and known sites, all of which can be simply switched out to meet the needs of various projects. All data underlying this article are available at https://github.com/anbianchi/IntegratedSNVINDELSandCNV/.

Dataset

The datasets employed encompass two WES datasets: PRJEB3235 (36 items) (11) and PRJEB31704 (7 items) (12). The dataset corresponding to id PRJEB3235 provides sequencing data for eleven BC cases: 07S240, DAD1, family F2887 (F2887-13 and -24), family F3311 (F3311-5 and -43), I-1408, family RUL036 (RUL036-2 and -7), family RUL153 (RUL153-2 and -3) and seven HapMap controls. Regarding the dataset with id PRJEB31704, we employed seven samples: family F11 (BC patients F11S01 and F11S02 and their informative relatives F11S03 and F11S04), family F12 (BC patients F12S01, F12S02 and their informative relative F12S03). Both the datasets correspond to samples sequenced on the Illumina platform. They were chosen for their potential since they contain novel genetic variants associated with BC susceptibility, as they included data from families of patients who were negative for *BRCA1/2* mutations.

Software

The technological foundation of our strategy is Snakemake (13), a workflow management system well known for encouraging sustainable, scalable, and repeatable data analysis. Within this management framework, we integrated a suite of tools for all the various stages of our analysis. Common tools shared along the integrated approach include FastQC for quality control (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trimmomatic for read trimming (http://www.usadellab.org/cms/?page=trimmomatic), Samtools for indexing and statistics on BAM files (http://www.htslib.org/), Picard for sorting and duplicate removal (https://broadinstitute.github.io/picard/), and Genome Analysis Toolkit (GATK) (14) for recalibration of bases. These tools serve as the backbone for our approach, ensuring a consistent and robust processing pipeline from raw reads (in terms of .fastq files). Following the realignment of bases, our strategy diverges, employing specialized tools contingent on the specific objectives of each sub-pipeline. Regarding short variant calling of SNVs and Indels, we employed the well-known GATK germline short variant discovery pipeline (https://tinyurl.com/ypcx7fnx). About CNV calling, we utilized two well-known CNV tools: ExomeDepth (15) and cn.mops (16). For the annotation of SNV/Indels and CNVs, we respectively used Annovar (version 2020-06-08) (17) and AnnotSV (version 3.2.3) (18). The details regarding these specialized tools, along with their respective versions, are provided in Table 1 .

Hardware

The research was conducted on Caliban, a cluster environment provided by the DISIM Department of the University of L'Aquila and comprising multiple nodes. Specifically, the experiments were executed on a system running CentOS Linux release 7.4.1708 (Core) and powered by an Intel(R) Xeon(R) CPU E5-2698 v4, operating at 2.20 GHz, with an available RAM of 141 GB.

Table 1. List of tools, version number and parameters employed in our pipeline

Scope	Tool	Version	Parameters
Preprocessing	Trimmomatic	0.39	MAXINFO:40:0.9 MINLEN:36
	Trimmomatic PE	0.39	MAXINFO:40:0.9 MINLEN:36
	BWA-MEM	0.7.17	Default
	Samtools Flagstat	2.6.0	Default
	Mark Duplicates	3.0.0	remove_duplicates: true create_index: true validation_stringency: silent
	Picard Sortsam	1.17	sort_order:coordinate extra:create_index true
	Gatk Base Recalibrator	4.4.0.0	intervals: 100bp_exon.bed known-sites: Mills_and_1000G.ind.hg19
Snv/Indels	GATK HaplotypeCaller	4.4.0.0	-ERC GVCF
	GATK GenotypeGVCFs	4.4.0.0	Default
	GATK VQSR	4.4.0.0	-mode SNP, -mode INDEL
	GATK VariantFiltration	4.4.0.0	-
	Annovar	2020-06-07	-build hg19
CNV	Gatk Apply BQSR	4.4.0.0	extra = –intervals 100bp_exon.bed
	ExomeDepth	1.1.16	100bp_exon.bed
	cn.mops	1.44.0	100bp_exon.bed
	AnnotSV	3.2.3	Default

Resources

Reference genome

The reference genome used is hg19, taken from UCSC Genome Browser (https://genome.ucsc.edu/). In particular, the *hg19.fa* file we used is in (https://tinyurl.com/4rjnnetn).

Targeted regions (bed files)

To accurately specify the regions of interest for our tools, we utilized a bed file format. Given the possibility of identifying high-quality off-target variants from WES, as reported in (19), we obtained the input by using exon sites with a 100bp flanking region. These sites were retrieved using UCSC's Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables).

Known sites

Known sites were utilized for various steps including base recalibration and variant filtering. We sourced these from GATK bundle (https://tinyurl.com/3f5n4cy7).

Implementation

The integrated method for SNV, Indels and CNV detection we used is depicted in Figure 1. It is organized into three primary segments: *Preprocessing*, *SNV and indels identification* and *CNV detection*. They are detailed in the following sections.

Preprocessing

The origin of our integrated method encompasses a series of crucial preprocessing steps, forming a common foundation for the ensuing specialized analyses aimed at SNVs, Indels and CNVs detection. The primary objective of these steps is to refine and structure the data, paving the way for precise variant detection in the later stages.

- Raw data quality control: The approach begins with a quality control check on the raw WES data using FastQC (https://tinyurl.com/333dk7kp).

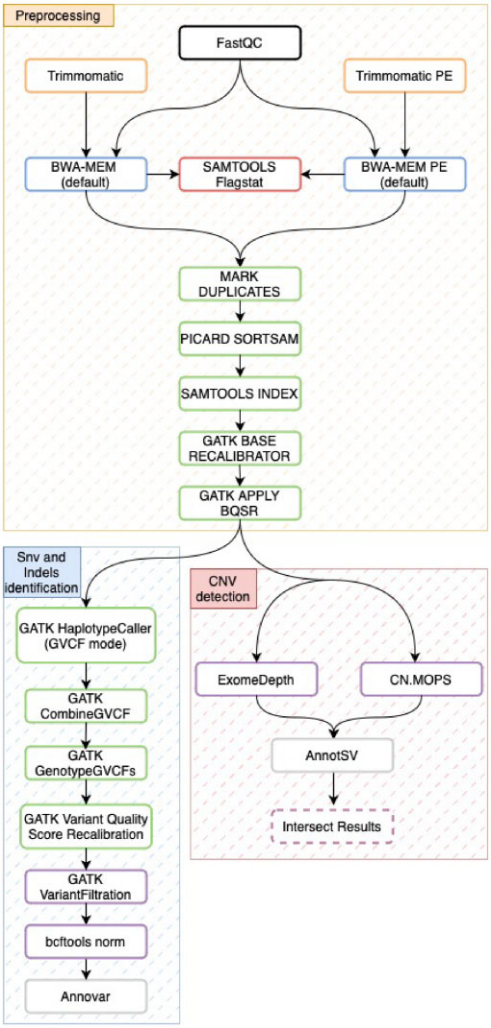


Figure 1. Proposed method for SNV/indels and CNV detection.

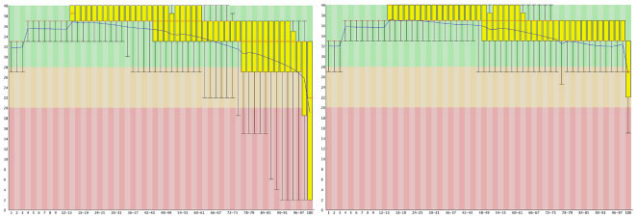


Figure 2. Pre-trimming (left) and post-trimming (right) base read quality of a representative sample. Values in red denote poor quality. After trimming, the read quality is greatly increased.

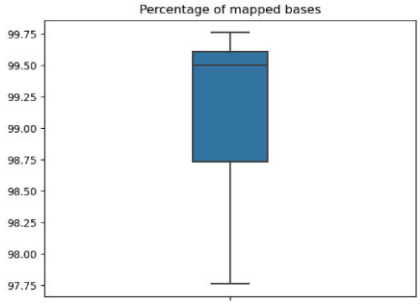


Figure 3. Boxplot of the percentage of mapped bases for all samples of the two datasets. The y-axis shows the percentage of properly mapped bases.

- Read trimming: Following the quality control, the reads undergo trimming to eliminate low-quality bases and adapter sequences, employing Trimmomatic with parameters set to MAXINFO:40:0.9 and MINLEN:36. These parameters, specifically chosen based on the nature of our data, but adaptable also to other data, enhance the overall read quality (Figure 2).
- Read alignment: The high-quality trimmed reads are then aligned to the reference genome using the BWA-mem tool with default settings (https://bio-bwa.sourceforge.net/bwa.shtml). Accurate mapping of reads to their respective genomic locations is pivotal at this juncture (Figure 3).
- Duplicate removal and sorting: Duplicate reads are identified and removed with MarkDuplicates. This process curtails redundancy and ensures accurate coverage calculations at specific genomic loci. Then, the aligned reads are sorted by genomic coordinates using Picard.
- Alignment quality assessment: The alignment quality is assessed using Samtools Flagstat to generate statistics on the mapped files. A high percentage of properly mapped bases attests the reliability of the alignment process.
- Base Quality Score Recalibration: Lastly, a base quality score recalibration is executed using the BaseRecalibrator tool from GATK to amend any biases or errors in the quality scores assigned by the sequencer. Known sites from the GATK bundle are utilized to provide additional contextual information for the recalibration process.

SNV and indels identification pipeline

In this segment of our pipeline, depicted in Figure 1 (left side, bottom), we adhere to GATK4 best practices for discerning germline single SNVs and Indels. Briefly, GATK process for short variant discovery encompasses a streamlined process for accurately identifying SNVs and Indels from high-throughput sequencing data. The workflow initiates with quality control and alignment of sequencing reads to a reference genome. Subsequent steps include preprocessing tasks such as marking duplicates and recalibrating base quality scores to enhance data quality. Variant discovery is conducted using the HaplotypeCaller in GVCF mode, enabling comprehensive variant evidence capture for each sample. This facilitates accurate joint genotyping across the cohort, combining individual GVCFs into a single, refined variant call set. Variant calls are further polished using Variant Quality Score Recalibration (VQSR) and/or hard filtering, improving call precision. The final step involves annotating variants for biological relevance, aiding in the interpretation of their impact. This compact yet comprehensive approach ensures high-quality, reliable variant discovery for genetic research. The tools incorporated are enumerated in Table 1, alongside their version numbers and Extra parameters. Here, SNVs and Indels are identified using HaplotypeCaller in GVCF mode. Post-HaplotypeCaller execution, joint genotyping is carried out with GenotypeGVCFs, generating a variant list. Herein, we apply both VQSR and hard filtering to sift through the variants. This double filtering enhances the reliability of our variant calls. Finally, variants are annotated with Annovar to provide more context and insights on the discovered variants.

In order to further discern and refine the selection to ensure more focused variants, we used the Franklin's tools by Genoox (https://franklin.genoox.com/clinical-db/home), based on the following: (i) variants located in exonic regions and splicing sites, (ii) rare heterozygous variants with minor allele frequency (MAF) <0.01 in ExAC, GnomAD and 1000 genomes databases, (iii) start-loss, stop-gain, stop-loss and frameshift variants, along with missense variants flagged as deleterious by at least half (6 out of 11) of the *in silico* prediction tools considered (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, LR, VEST3, CADD). To diminish the false call rate, a read depth above 10, and a quality per depth above 10, are also taken into account (20).

In the variant and gene prioritization phase, targeting the identification of candidate BC risk loci, we consider: (i) variant classification (pathogenic, likely/possibly pathogenic, uncertain significant) by the American College of Medical Genetics and Genomics-Association for Molecular Pathology (ACGM) guidelines (21), identified through the Franklin ACMG annotation, (ii) variants characterized by evidence of pathogenicity or conflicting interpretation of pathogenicity in the Clinvar database, (iii) variants in known/candidate cancer predisposition genes (22). A manual check on the published data regarding filtered variants is also conducted. Please note that the utilization of Genoox's tool served as an additional step to our primary methodology, aimed at prioritizing the variations identified through the annotation of VCF files. This aspect of the process is out of the scope of the this study but rather serves to validate the discovered variants.

CNV pipeline

The ensuing section elaborates on the CNV facet of our approach, depicted in Figure 1 (right side, bottom). All tools encompassed in this segment, alongside their version numbers, are itemized in Table 1. For CNV detection, we adopted two distinct tools. The first tool, ExomeDepth, is esteemed for its

capability in discerning CNVs, particularly excelling in detecting rare variants (23). By leveraging statistical models and read-depth data, ExomeDepth renders predictions of CNV occurrences throughout the genome. Conversely, cn.mops, the second tool employed, is tailored for CNV detection, employing a distinctive methodology that accommodates variations in read depth. Informed by the importance of employing multiple CNV callers to mitigate biases inherent in CNV detection, our approach strategically incorporated two of the most frequently used, cited and reliable CNV calling tools (24).

The preprocessing stage of our pipeline yields mapped files (.bam extension), which are subsequently inputted into both ExomeDepth and cn.mops for CNV calling. These tools also accept bed files as input, utilizing them to read exon coordinates that delineate the specific genomic regions of interest to be examined. Discrepancies in read counts enable the pipeline to spotlight regions potentially containing CNVs. The ensuing CNV calls are ranked and archived based on their confidence and relevance levels, ensuring that the most significant and reliable CNV calls are prioritized for further scrutiny. The output files (.bed format) from both tools are further annotated via AnnotSV to get additional information on the genes encompassing the variants. To obtain highly reliable data, we filtered out alterations with allele frequency lower than 0.01, in order to remove those considered as common from the analysis. We also assessed metrics calculated by ExomeDepth for identified deletion and duplication events, in particular a read ratio (reads expected vs observed) between 0.4 and 0.7 for deletions and greater than 1.3 for duplications. Quality parameters, frequency within the analyzed BC cases and consistency of filtered CNVs were also manually checked. Note that only the CNVs identified by both tools are retained post the ranking intersection. Mirroring the identification process of SNVs and short Indels, we incorporated a prioritization step for potential genes of interest, predominantly based on ACMG classification and the selection of known/candidate cancer susceptibility genes (22).

Validation analysis

For the comprehensive validation of our pipeline, we utilized the NA12878 sample from the Genome in a Bottle Consortium (GIAB), as suggested in (25). The NA12878 sample is derived from a well-studied human cell line, extensively sequenced across multiple platforms. All relevant materials, including the raw data and the truth set of VCFs, are available in the official BioProject (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA200694). Specifically, we obtained the raw data and the truth set VCFs from the VCF folder within the official FTP repository (https://ftp-trace.ncbi.nlm.nih.gov/giab/), ensuring our validation process leverages the most accurate and comprehensive reference data available. The initial step in our validation process involved downloading the raw sequences and the truth sets of variant calls for the NA12878 sample from the GIAB's FTP official site (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome). Following data acquisition, the validation of SNV and Indels was performed using the *hap.py* tool developed by Illumina, as suggested in (26). This tool is specifically engineered to adeptly handle complex variant types, enabling a distinct evaluation of SNVs and Indels. For the evaluation, we compared our pipeline's filtered call set against the real truth set provided in the FTP site, specifically named `project.NIST.hc.snps.indels.vcf`. This truth set is recognized as the gold standard for variant calling in the NA12878 sample, comprising high-confidence SNVs and Indels identified through rigorous consensus methodology among leading genomics research institutions. For the benchmarking of CNVs, we adopted a reference SV baseline callset for NA12878, courtesy of the Mt. Sinai School of Medicine. This callset, derived from approximately 44× coverage of PacBio data, encompasses a merged SV VCF file (high-confidence reference callset), which is crucial for a comprehensive CNV analysis. Data, including the merged SV VCF, are publicly accessible at the GIAB repository under NA12878 PacBio data (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/), providing a valuable resource for high-confidence variant calls. To compare our pipeline's CNV calls against this high-quality reference set, we employed Truvari (27), a tool specifically designed for the evaluation and comparison of genomic variants. Truvari facilitates the assessment of concordance between our detected CNVs and the reference SV callset, enabling a detailed analysis of our pipeline's performance. The performance of our pipeline was evaluated in terms of Recall, Precision and F1 Score.

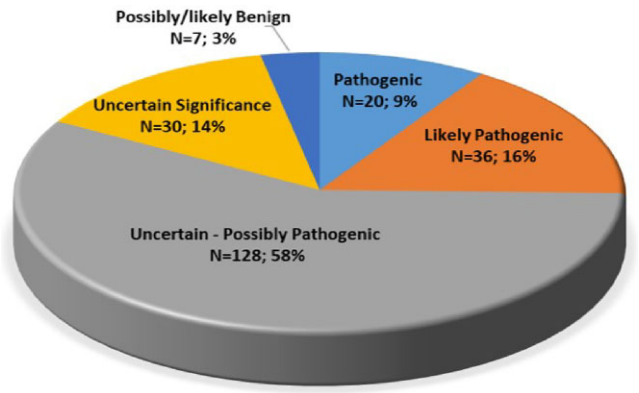


Figure 4. SNVs and indels identified in this study after variant filtering, based on ACMG classification.

Results

Indels and SNV

Post-calling and filtering, 221 variants were identified with 191 being unique (Supplementary Table S1). These variants were categorized as per the ACGM classification and the distribution is illustrated in Figure 4. To validate, we cross-verified with variants from the original publications (11,12). In particular, we were able to confirm the presence of the two most relevant variants previously described: *CHEK2* c.1100delC, p.Thr367fs in family RUL153 and *FANCM* c.5791C>T, p.Arg1931* in sample DAD1 (11) (Table 2).

Several interesting variants, unmentioned in original publications, were also detected (Table 2). Notably, we found missense variants in the *RBBP8* (c.298C>T, p.Arg100Trp, sample F3311_5) and *LEPR* genes (c.1835G>A, p.Arg612His, sample 07S240), previously described in high-risk, BRCA-negative BC cases (30,31). Frameshift mutations in tumor-related genes such as *DLEC1* (c.5068_5071dupAACA, p.Ser1691fs, sam-

Table 2. List of the most interesting pathogenic (P), likely pathogenic (LP) and uncertain significance (VUS) variants identified in this study

Gene	Transcript	Nucleotide Change	AA Change	Effect	ClinVar Classification	Franklin ACMG classification	Sample	Notes
AIM2	NM_004833.3	c.1027delA	p.Thr343fs	Frameshift	N.A.	VUS	DAD1/RUL153_3	Gene involved in cell cycle regulation, suppression of tumor proliferation (28)
ATM	NM_000051.3	c.8560C>T	p.Arg2854Cys	Missense	Conf. int. of Path.	VUS (possibly pathogenic)	F3311_5	Known BC-related gene (22)
ATM	NM_000051.3	c.572T>A	p.Ile191Asn	Missense	VUS	VUS	F2887_24	Known BC-related gene (22)
CHEK2	NM_007194.4	c.1100delC	p.Thr367fs	Frameshift	P	P	RUL153_2/ RUL153_3	Rreported in the original publications (11)
DLEC1	NM_007335.3	c.5068_5071dupAACA	p.Ser1691fs	Frameshift	N.A.	VUS (possibly pathogenic)	F2887_24	Tumor suppressor gene involved in DNA damage response (22)
EWSR1	NM_005243.3	c.1843C>T	p.Arg615*	Stop Gain	N.A.	LP	F2887_13	Gene linked with BRCA1/2 pathway and associated with non-medullary thyroid cancer susceptibility (29)
FANCM	NM_020937.4	c.5791C>T	p.Arg1931*	Stop Gain	P/LP	P	DAD1	Rreported in the original publications (11)
LEPR	NM_002303.5	c.1835G>A	p.Arg612His	Missense	Conf. int. of Path.	VUS (possibly pathogenic)	07S240	Same variant identified in high-risk, non-BRCA BC case (30)
PDGFRA	NM_006206.6	c.2411G>A	p.Arg804Gln	Missense	VUS	VUS	I1408	Gene associated with prostate cancer and sarcoma predisposition (22)
RBBP8	NM_002894.3	c.298C>T	p.Arg100Trp	Missense	Conf. int. of Path.	LP	F3311_5	Same variant identified in high-risk, early onset, non-BRCA BC case (31)
RECQL	NM_002907.4	c.401C>T	p.Thr134Ile	Missense	Conf. int. of Path.	VUS	I1408	Known BC-related gene(22)
SEC23B	NM_006363.6	c.2035G>T	p.Glu679*	Stop Gain	N.A.	LP	F3311_5	Gene associated with cowden syndrome and sporadic thyroid cancer (32)
TP53AIP1	NM_022112.2	c.63dupG	p.Gln22fs	Frameshift	N.A.	VUS (possibly pathogenic)	RUL36_7/RUL153_2	Gene associated with melanoma susceptibility (22)
TSC2	NM_000548.5	c.748A>G	p.Lys250Glu	Missense	VUS	VUS	F2887_13	Gene associated with colorectal and gland cancers predisposition (22)

N.A.: not available. Conf. int. of Path.: conflicting interpretation of pathogenicity.

ple F2887_24) and *AIM2* (c.1027delA, p.Thr343fs, samples DAD1/RUL153_3) were also found. In addition, we detected variants of potential interest in known BC-related genes: these included the variants c.572T>A (p.Ile191Asn, sample F2887_24) and c.8560C>T (p.Arg2854Cys, sample F3311_5) in the *ATM* gene and c.401C>T (p.Thr134Ile, sample I1408) in the *RECQL* gene.

Overall, we identified numerous Variant of Uncertain Significance (VUS), largely deemed possibly pathogenic by the ACMG classification (Figure 4). While not directly linked to disease, they underscore the samples' genetic complexity and offer potential research avenues.

CNV

For CNVs identification, we considered only coherent results from ExomeDepth and cn.mops analysis. The merging operation we implemented reduced the number of CNVs from 102.359 to only 983. We then considered parameters, including allele frequency in the population and within the analyzed BC cases, metrics calculated by ExomeDepth and consistency of filtered CNVs, further reducing the number of CNVs to a total of 69 CNVs affecting 103 genes (Supplementary Table S2).

The ExomeDepth algorithm had an average Bayes factor of 19.56 and read count ratios of 0.56 for deletions and 1.36 for duplications. CNVs were unevenly distributed across chromosomes (chr), with more duplications in chr 1 and 12, and deletions in chr X (Figure 5). Notably, most X-related CNVs were due to a large region (48316827_49144655) deletion in sample F11S02. We prioritized CNVs based on ACMG annotations and genes linked to cancer predisposition. This analysis revealed the 29496809–29509783 pathogenic deletion on chr 17, overlapping with *NF1* gene and the 86368073–86408032 deletion on chr 9, overlapping with *GKAP1*, both in sample F11S01. Moreover, sample F12S01 showed CNVs overlapping with known cancer-related genes, including *KRAS*, *RBBP8*, *ARNT2*, *ESR1*, *SYCP1* and *XPO1*.

Pipeline validation and performance metrics

The detailed performance of our pipeline in detecting SNVs and Indels, evaluated using the gold standard reference sample NA12878, is summarized in Table 3. These results underscore the strengths of our pipeline in SNV detection, marked by high precision (97.86%) and a robust F1 score (91.12%). However, the relative challenge in Indel detection, especially in achieving a higher recall (now 71.13%), indicates an area for further research and development.

The validation of CNVs yielded the values in Table 4. These results, while indicating a solid foundation for CNV detection,

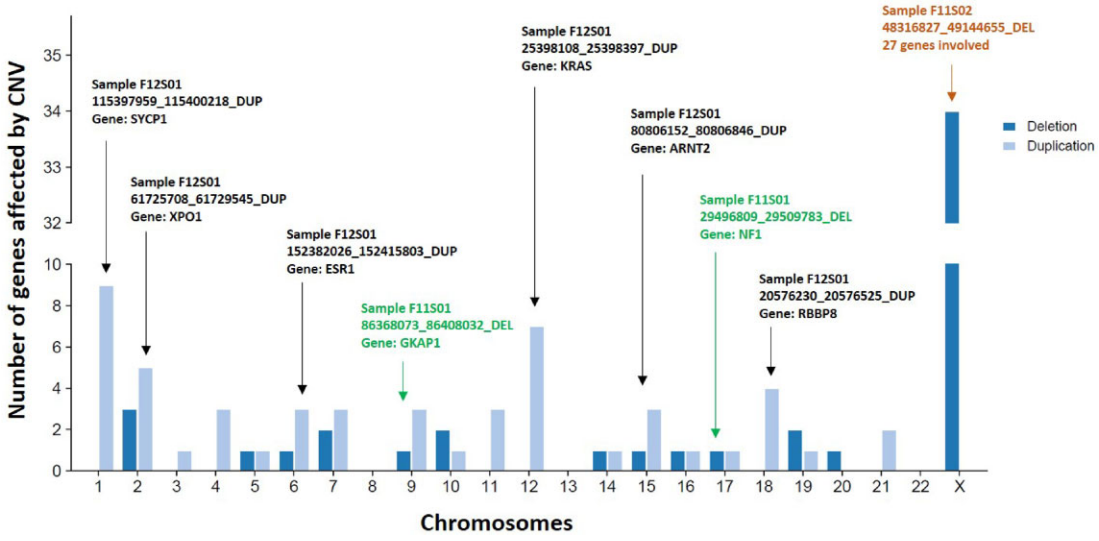


Figure 5. Chromosomal distribution of CNVs. Most relevant CNVs detected in samples F11S01 (green), F11S02 (orange) and F12S02 (black) are shown.

Table 3. Validation results for SNVs and indels

Variant type	Recall	Precision	F1 score
SNVs	85.25%	97.86%	91.12%
Indels	71.13%	81.09%	75.78%

Table 4. Validation results for CNVs

Variant type	Recall	Precision	F1 score
CNV	60.53%	72.72%	66.07%

also highlight specific areas for improvement and refinement within our methodology. In particular, a recall rate of 60.53% suggests that, although the pipeline is capable of identifying a majority of the true positive CNVs present within the dataset, a significant proportion remains undetected. This gap may be attributed to the inherent complexities associated with CNV detection. The challenges posed by these complexities are further compounded by variations in sequencing data quality and the limits of current detection algorithms' sensitivity. The precision rate of 72.72% demonstrates the pipeline's effectiveness in distinguishing true CNVs from false positives with good accuracy. The F1 score, which is a harmonic mean of precision and recall, stands at 66.07%.

Discussion

In personalized medicine, the identification of novel putative genes or variants implicated in BC susceptibility would have a significant impact on clinical practice (9). WES analysis has proven to be a suitable procedure for detecting disease-causing variants and discovering new target genes (33). Notably, several interesting and promising new putative genes/variants, such as *RCC1* and *SERPINA3*, are emerging in BC predisposition by using this method (34,35). However, few studies about WES analysis in non-BRCA patients, mainly in a limited number of families, are available (36). Our study's main contribution is the creation of a specialised method, included in a replicable and adaptable tool we made available to the scientific community, created for processing WES datasets to comprehensively detect SNVs, Indels and CNVs.

In this study, using the proposed method, we re-analyzed two WES datasets (PRJEB3235 project (11) and PRJEB31704 project (12)) to look for germline alterations in non-BRCA patients potentially predisposing to familial BC, with the aim of detecting SNVs, Indels and CNVs.

Interestingly, putatively relevant variants, undetected in the original published studies, emerged in the present work: these occurred in genes recently associated with BC predisposition and pathogenesis, including *LEPR* and *RBBP8* (30,31), or genes involved in cancer-related pathways, including *AIM2, EWSR1, DLEC1, SEC23B* and *TP53AIP1* (22,28,29,32). With this regard, the c.298C>T (p.Arg100Trp) variant in the *RBBP8* gene, which is involved in the homologous recombination DNA repair mechanism, was recently identified in a high-risk, early onset BC case negative for mutations in *BRCA1/2* genes (31). Similarly, the c.1835G>A (p.Arg612His) variant in the *LEPR* gene, which encodes for the leptin receptor involved in the regulation of lipid metabolism, was identified in a high-risk-familial, BRCA-negative BC patient (30).

CNV detection accuracy varies with the bioinformatics tools and settings utilized. For optimal results, it's advised to merge algorithms from different methods (37). In our approach, we integrated ExomeDepth and cn.mops. ExomeDepth is considered one of the most balanced tools for sensitivity and specificity (38), supporting its use in routine targeted NGS diagnostic services for Mendelian diseases (39). Similarly, cn.MOPS shows the best performance when the size of targeted CNVs is between 100 kb and 10 Mb, but it is also a suitable choice for unknown research, as its accuracy is globally satisfactory (37). Based on a prioritization scale, the most interesting CNV detected in this study was a deletion on chr 17, which includes the known BC susceptibility gene *NF1* (22). The same sample (F11S01) also showed a deletion on chr 9, overlapping with *GKAP1*, a gene recently suggested to be a candidate susceptibility factor in esophageal squamous cell carcinoma (40). The deletion of a large genomic region on the X chromosome was also observed in the sample F11S02. This chromosome carries a significant number of oncogenes and tumor suppressor genes, the genetic alteration or dysregulation

of which, both at germline and somatic level, has been associated with the development and progression of different cancer types, including BC (41). Finally, sample F12S01 revealed duplication CNVs overlapping several known cancer-related genes, not only with an oncogenic role such as *ESR1* and *KRAS*, but also identified as tumor suppressors, including the aforementioned *RBBP8* and *ARNT2*. Deletion CNVs result in haploinsufficiency, while duplications can cause triplosensitivity, gene fusion, or disruption. Though most duplications are adjacent to the original locus, inversions or intragenic variations can disrupt genes. Therefore, predicting genetic consequences of such alterations requires specific breakpoint-level analysis (42). Of note, key CNVs were found in samples without significant SNVs-indels, indicating that in these familial BC cases, susceptibility may arise from these alterations over nucleotide-based variants.

Overall, our methodology has been meticulously crafted with adaptability, allowing for straightforward adjustments to cater to various genomes or resources utilized during research. Importantly, it is capable of adeptly handling both single end-reads and paired-end reads, demonstrating its flexibility in accommodating different data configurations. Employing stringent quality control parameters at various junctures of the pipeline, our tool minimizes the risk of false positives and negatives. This process allowed us to confirm the presence of all the more prominent variants in genes known to be involved in BC susceptibility, such as *CHECK2* and *FANCM*, previously described (11), as well as highlighting the good performance of our method, as demonstrated by the analysis of the reference sample NA12878.

In particular, the accuracy metrics here obtained, make the performance of our pipeline comparable to others already described in literature for SNV-indels detection (43). Our ongoing efforts aim to enhance the algorithm's sensitivity to indels (recall 71.13%, precision 81.09%, F1 score 75.78%) without compromising the high precision observed in SNV detection (recall 85.25%, precision 97.86%, F1 score 91.12%). At the same time, enhancing the pipeline's ability to detect CNVs (recall 60.53%, precision 72.72%, F1 score 66.07%), particularly by improving recall, is a pivotal area for future development. Strategies to achieve this may include refining existing detection algorithms, integrating additional CNV-specific quality metrics, and leveraging advanced computational techniques to better interpret complex genomic regions (44). Of note, CNV callers tends to be more challenging both because: (i) these variants are more difficult to accurately detect using short read sequencing data, which makes structural variants calling more error-prone than small variants calling, (ii) the precise breakpoints for CNVs are not always well defined, which makes comparison between call-sets more complex (45). Therefore, even the best performing structural variants callers with whole-genome sequencing data achieve F1 scores of 80–90% (45). However, to date, WES is the most widely used NGS approach in clinical diagnostics and academic research (46), so there is a need to find accurate solutions and strategies for detecting CNVs also from WES data. Here, using two well-known CNV detection tools (ExomeDepth and cn.mops), taking into account quality parameters and retaining only the CNVs identified by both tools, we obtained performance metrics higher to others already described in literature based on WES data (47).

This pipeline, enclosed within the Snakemake (13) workflow management system, thus offers a turnkey solution for researchers, particularly in the realm of BC genetics. Snakemake is scalable by design, facilitating the management of workflows ranging from small-scale analyses to large, high-throughput computing environments. By consolidating the analytical process into one unified system, we significantly reduce the time and expertise required to configure and run genomic analyses. This approach allows researchers to swiftly apply our pipeline to their datasets, enabling a more focused investigation of genetic underpinnings in diseases without the burden of technical complexities often associated with genomic data analysis. The openness and accessibility of our pipeline contrast sharply with the closed nature of many in-house tools (8), providing a valuable resource for the wider research community to engage in collaborative improvements and benchmarking efforts.

A limitation of our study is the inability to directly validate new variants with Sanger sequencing and MLPA (Multiplex Ligation-dependent Probe Amplification). To address this, we experimentally benchmarked our data against the NA12878 reference material provided by the GIAB, serving as a standard for assessing the accuracy of our NGS approach. Furthermore, despite the validation constraints, a study validating 1109 NGS variants from 825 clinical exomes reported 100% concordance for SNV and Indel variants and 95.65% for CNVs with traditional methods. This suggests that, especially with high-quality NGS data, confirmatory analysis might not always be essential (48). Of note, while our initial focus was on a specific BC subgroup, we underline that the methodology we set-up has broader applications and it can certainly be considered a suitable tool for advancing our understanding of other high-impact complex polygenic diseases (e.g. cardiovascular, neurological, diabetes).

Conclusion

In this study, our primary objective was to design a specialized, consistent, and flexible integrated method, tailored for advanced WES data analysis, emphasizing germline variants in non-BRCA familial BC patients. This innovation offers a comprehensive tool for analyzing SNVs, short indels and CNVs. Through strict quality control, detailed cross-verification and validation analysis, we've significantly reduced false positives and negatives, ensuring pinpoint accuracy in variant identification. It reaffirmed known significant variants and highlighted novel variants potentially linked to BC pathogenesis. Our study complements and expands on previous WES-based BC predisposition research. The evolving NGS technology, paired with updated gene databases and cutting-edge variant classification tools, has amplified our capacity to revisit whole genome/exome NGS data from non-BRCA patients.

Accurately detecting CNVs remains a focal point in related research. Our strategy—merging ExomeDepth and cn.mops and focusing on common variants—proves a hopeful approach to address this. Distinguishingly, our tool's broad analysis range surpasses traditional methods often limited to specific genes or genetic alterations.

In future research, our aim is to enhance our tool's capabilities, refine its precision across diverse datasets, and cultivate partnerships for integration into holistic genomic studies. Such endeavors aim to progressively unveil the intricate genomic backdrop of various cancers and genetic anomalies.

Data availability

The data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.10078336. They can be accessed also from GitHub, at https://github.com/anbianchi/IntegratedSNVINDELSandCNV/.

The datasets were derived from sources in the public domain: BioProject PRJEB3235 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB3235) and BioProject PRJEB31704 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31704).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

Author contributions: A.B: Writing – original draft, Software, Methodology, Formal Analysis. V.Z.: Writing – original draft, Investigation, Visualization. A.D, Al.DM. and G.S.: Writing - review & editing, Formal Analysis. K.C., F.Z. and E.A.: Conceptualization, Writing – review & editing, AS.D and S.G.: Data curation, Writing – review & editing. An.D.M. and A.T.: Conceptualization, Project administration, Supervision, Funding acquisition, Resources, Writing – review & editing. All authors read and approved the final version of the manuscript. V.Z. was supported by Carispaq Foundation L'Aquila 2020. All the numerical simulations have been realized on the Linux HPC cluster Caliban of the High-Performance Computing Laboratory of the Department of Information Engineering, Computer Science and Mathematics (DISIM) at the University of L'Aquila.

Funding

European Union—NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR)—Project: "SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics", [IR0000013]; European Union—NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem [ECS00000041 - VITALITY].

Conflict of interest statement

None declared.

References

1. Berger,M.F. and Mardis,E.R. (2018) The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.*, **15**, 353–365.
2. Valsesia,A., Macé,A., Jacquemont,S., Beckmann,J.S. and Kutalik,Z. (2013) The growing importance of CNVs: new insights for detection and clinical interpretation. *Front. Genet.*, **4**, 92.
3. Pfundt,R., Del Rosario,M., Vissers,L.E., Kwint,M.P., Janssen,I.M., De Leeuw,N., Yntema,H.G., Nelen,M.R., Lugtenberg,D., Kamsteeg,E.-J., *et al.* (2017) Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.*, **19**, 667–675.
4. Qi,Q., Jiang,Y., Zhou,X., Meng,H., Hao,N., Chang,J., Bai,J., Wang,C., Wang,M., Guo,J., *et al.* (2020) Simultaneous detection of CNVs and SNVs improves the diagnostic yield of fetuses with ultrasound anomalies and normal karyotypes. *Genes*, **11**, 1397.
5. Yuan,B., Wang,L., Liu,P., Shaw,C., Dai,H., Cooper,L., Zhu,W., Anderson,S.A., Meng,L., Wang,X., *et al.* (2020) CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. *Genet. Med.*, **22**, 1633–1641.
6. Minoche,A.E., Lundie,B., Peters,G.B., Ohnesorg,T., Pinese,M., Thomas,D.M., Zankl,A., Roscioli,T., Schonrock,N., Kummerfeld,S., *et al.* (2021) ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. *Genome Med.*, **13**, 32.
7. Reid,J., Kachhia,S., Dougall,P., Shovelton,J., Molha,D., Taylor,C., Kasturiarachchi,J., Holdstock,J., Pullabhatla,V., Parkes,L., *et al.* (2020) A next generation sequencing solution to detect copy number variants, single nucleotide variants and loss of heterozygosity in intellectual disability and developmental delay samples. *Mosaic*, **5**, 100.
8. Bademci,G., Foster,J., Mahdieh,N., Bonyadi,M., Duman,D., Cengiz,F.B., Menendez,I., Diaz-Horta,O., Shirkavand,A., Zeinali,S., *et al.* (2016) Comprehensive analysis via exome sequencing uncovers genetic etiology in autosomal recessive nonsyndromic deafness in a large multiethnic cohort. *Genet. Med.*, **18**, 364–371.
9. Breast Cancer Association Consortium, Dorling,L., Carvalho,S., Allen,J., González-Neira,A., Luccarini,C., Wahlström,C., Pooley,K.A., Parsons,M.T., Fortuno,C., *et al.* (2021) Breast cancer risk genes - association analysis in more than 113,000 women. *N. Engl. J. Med.*, **384**, 428–439.
10. Keeney,M.G., Couch,F.J., Visscher,D.W. and Lindor,N.M. (2017) Non-BRCA familial breast cancer: review of reported pathology and molecular findings. *Pathology*, **49**, 363–370.
11. 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.* (2013) Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PloS one*, **8**, e55681.
12. Glentis,S., Dimopoulos,A.C., Rouskas,K., Ntritsos,G., Evangelou,E., Narod,S.A., Mes-Masson,A.-M., Foulkes,W.D., Rivera,B., Tonin,P.N. and et,al. (2019) Exome sequencing in BRCA1-and BRCA2-negative Greek families identifies MDM1 and NBEAL1 as candidate risk genes for hereditary breast cancer. *Front. Genet.*, **10**, 1005.
13. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
14. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
15. Plagnol,V., Curtis,J., Epstein,M., Mok,K.Y., Stebbings,E., Grigoriadou,S., Wood,N.W., Hambleton,S., Burns,S.O., Thrasher,A.J., *et al.* (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, **28**, 2747–2754.
16. Klambauer,G., Schwarzbauer,K., Mayr,A., Clevert,D.-A., Mitterecker,A., Bodenhofer,U. and Hochreiter,S. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.
17. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
18. Geoffroy,V., Herenger,Y., Kress,A., Stoetzel,C., Piton,A., Dollfus,H. and Muller,J. (2018) AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, **34**, 3572–3574.
19. Guo,Y., Long,J., He,J., Li,C.-I., Cai,Q., Shu,X.-O., Zheng,W. and Li,C. (2012) Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, **13**, 194.
20. Mehandziska,S., Stajkovska,A., Stavrevska,M., Jakovleva,K., Janevska,M., Rosalia,R., Kungulovski,I., Mitrev,Z. and

Kungulovski,G. (2020) Workflow for the implementation of precision genomics in healthcare. *Front. Genet.*, **11**, 619.
21. Richards,S., Aziz,N., Bale,S., Bick,D., Das,S., Gastier-Foster,J., Grody,W.W., Hegde,M., Lyon,E., Spector,E., *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**, 405–423.
22. Rotunno,M., Barajas,R., Clyne,M., Hoover,E., Simonds,N.I., Lam,T.K., Mechanic,L.E., Goldstein,A.M. and Gillanders,E.M. (2020) A systematic literature review of whole exome and genome sequencing population studies of genetic susceptibility to cancer. *Cancer Epidem. Biomar.*, **29**, 1519–1534.
23. Rajagopalan,R., Murrell,J.R., Luo,M. and Conlin,L.K. (2020) A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med.*, **12**, 14.
24. Gordeeva,V., Sharova,E., Babalyan,K., Sultanov,R., Govorun,V.M. and Arapidi,G. (2021) Benchmarking germline CNV calling tools from exome sequencing data. *Sci. Rep.*, **11**, 14416.
25. Zook,J.M., Catoe,D., McDaniel,J., Vang,L., Spies,N., Sidow,A., Weng,Z., Liu,Y., Mason,C.E., Alexander,N., *et al.* (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.
26. Haraksingh,R.R., Abyzov,A. and Urban,A.E. (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*, **18**, 321.
27. English,A.C., Menon,V.K., Gibbs,R.A., Metcalf,G.A. and Sedlazeck,F.J. (2022) Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.*, **23**, 271.
28. Liu,Z.-Y., Yi,J. and Liu,F.-E. (2015) The molecular mechanism of breast cancer cell apoptosis induction by absent in melanoma (AIM2). *Int. J. Clin. Exp. Med.*, **8**, 14750.
29. Srivastava,A., Giangiobbe,S., Skopelitou,D., Miao,B., Paramasivam,N., Diquigiovanni,C., Bonora,E., Hemminki,K., Försti,A. and Bandapalli,O.R. (2021) Whole genome sequencing prioritizes CHEK2, EWSR1, and TIAM1 as possible predisposition genes for familial non-medullary thyroid cancer. *Front. Endocrinol.*, **12**, 600682.
30. Aloraifi,F., McDevitt,T., Martiniano,R., McGreevy,J., McLaughlin,R., Egan,C.M., Cody,N., Meany,M., Kenny,E., Green,A.J., *et al.* (2015) Detection of novel germline mutations for breast cancer in non-BRCA 1/2 families. *FEBS J.*, **282**, 3424–3437.
31. Zarrizi,R., Higgs,M.R., Voßgröne,K., Rossing,M., Bertelsen,B., Bose,M., Kousholt,A.N., Rösner,H., Ejlertsen,B., Stewart,G.S. and et,al. (2020) Germline RBBP8 variants associated with early-onset breast cancer compromise replication fork stability. *J. Clin. Invest.*, **130**, 4069–4080.
32. Yehia,L., Niazi,F., Ni,Y., Ngeow,J., Sankunny,M., Liu,Z., Wei,W., Mester,J.L., Keri,R.A., Zhang,B., *et al.* (2015) Germline heterozygous variants in SEC23B are associated with Cowden syndrome and enriched in apparently sporadic thyroid cancer. *Am. J. Hum. Genet.*, **97**, 661–676.
33. Kiezun,A., Garimella,K., Do,R., Stitziel,N.O., Neale,B.M., McLaren,P.J., Gupta,N., Sklar,P., Sullivan,P.F., Moran,J.L., *et al.* (2012) Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, **44**, 623–630.
34. Riahi,A., Radmanesh,H., Schürmann,P., Bogdanova,N., Geffers,R., Meddeb,R., Kharrat,M. and Dörk,T. (2018) Exome sequencing and case–control analyses identify RCC1 as a candidate breast cancer susceptibility gene. *Int. J. Cancer*, **142**, 2512–2517.
35. Koivuluoma,S., Tervasmäki,A., Kauppila,S., Winqvist,R., Kumpula,T., Kuismin,O., Moilanen,J. and Pylkäs,K. (2021) Exome sequencing identifies a recurrent variant in SERPINA3 associating with hereditary susceptibility to breast cancer. *Eur. J. Cancer*, **143**, 46–51.
36. Zelli,V., Compagnoni,C., Cannita,K., Capelli,R., Capalbo,C., Di Vito Nolfi,M., Alesse,E., Zazzeroni,F. and Tessitore,A. (2020) Applications of next generation sequencing to the analysis of familial breast/ovarian cancer. *High-throughput*, **9**, 1.
37. Zhao,L., Liu,H., Yuan,X., Gao,K. and Duan,J. (2020) Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*, **21**, 97.
38. Moreno-Cabrera,J.M., Del Valle,J., Castellanos,E., Feliubadaló,L., Pineda,M., Brunet,J., Serra,E., Capellà,G., Lázaro,C. and Gel,B. (2020) Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur. J. Hum. Genet.*, **28**, 1645–1655.
39. Ellingford,J.M., Campbell,C., Barton,S., Bhaskar,S., Gupta,S., Taylor,R.L., Sergouniotis,P.I., Horn,B., Lamb,J.A., Michaelides,M., *et al.* (2017) Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur. J. Hum. Genet.*, **25**, 719–724.
40. Donner,I., Katainen,R., Tanskanen,T., Kaasinen,E., Aavikko,M., Ovaska,K., Artama,M., Pukkala,E. and Aaltonen,L.A. (2017) Candidate susceptibility variants for esophageal squamous cell carcinoma. *Genes Chromosom. Cancer*, **56**, 453–459.
41. Achilla,C., Papavramidis,T., Angelis,L. and Chatzikyriakidou,A. (2022) The implication of X-linked genetic polymorphisms in susceptibility and sexual dimorphism of cancer. *Anticancer Res.*, **42**, 2261–2276.
42. Newman,S., Hermetz,K.E., Weckselblatt,B. and Rudd,M.K. (2015) Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am. J. Hum. Genet.*, **96**, 208–220.
43. Abdelwahab,O., Belzile,F. and Torkamaneh,D. (2023) Performance analysis of conventional and AI-based variant callers using short and long reads. *BMC Bioinformatics*, **24**, 472.
44. Mandiracioglu,B., Ozden,F., Kaynar,G., Yilmaz,M.A., Alkan,C. and Cicek,A.E. (2024) ECOLE: Learning to call copy number variants on whole exome sequencing data. *Nat. Commun.*, **15**, 132.
45. Koboldt,D.C. (2020) Best practices for variant calling in clinical sequencing. *Genome Med.*, **12**, 91.
46. Zhao,L., Liu,H., Yuan,X., Gao,K. and Duan,J. (2020) Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*, **21**, 97.
47. Gordeeva,V., Sharova,E., Babalyan,K., Sultanov,R., Govorun,V.M. and Arapidi,G. (2021) Benchmarking germline CNV calling tools from exome sequencing data. *Sci. Rep.*, **11**, 14416.
48. Arteche-López,A., Ávila-Fernández,A., Romero,R., Riveiro-Álvarez,R., López-Martínez,M., Giménez-Pardo,A., Vélez-Monsalve,C., Gallego-Merlo,J., García-Vara,I., Almoguera,B., *et al.* (2021) Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. *Sci. Rep.*, **11**, 5697.

Received: November 29, 2023. **Revised:** March 22, 2024. **Editorial Decision:** March 27, 2024. **Accepted:** April 3, 2024