Molecular Cytogenetics

RESEARCH

Open Access

# Comparative analysis of hybrid-SNP microarray and nanopore sequencing for detection of large-sized copy number variants in the human genome

Catarina Silva[1,2], José Ferrão[1], Bárbara Marques[3], Sónia Pedro[3], Hildeberto Correia[3], Ana Valente[4], António Sebastião Rodrigues[2] and Luís Vieira[1,2*]

## Abstract

**Background** Nanopore sequencing is a technology that holds great promise for identifying all types of human genome variations, particularly structural variations. In this work, we used nanopore sequencing technology to sequence 2 human cell lines at low depth of coverage to call copy number variations (CNV), and compared the results variant by variant with chromosomal microarray (CMA) results.

**Results** We analysed sequencing data using CuteSV and Sniffles2 variant callers, compared breakpoints based on hybrid-SNP microarray, nanopore sequencing and Sanger sequencing, and analysed CNV coverage. From a total of 48 high confidence variants (truth set), variant calling detected 79% of the truth set variants, increasing to 86% for interstitial CNV. Simultaneous use of the 2 callers slightly increased variant calling. Both callers performed better when calling CNV losses than gains. Variant sizes from CMA and nanopore sequencing showed an excellent correlation, with breakpoints determined by nanopore sequencing differing by only 20 base pairs on average from Sanger sequencing. Nanopore sequencing also revealed that four variants concealed genomic inversions undetectable by CMA. In the 10 CNV not called in nanopore sequencing, 8 showed coverage evidence of genomic loss or gain, highlighting the need to improve SV calling algorithms performance.

**Conclusions** Nanopore sequencing offers advantages over CMA for structural variant detection, including the identification of multiple variant types and their breakpoints with increased precision. However, further improvements in variant calling algorithms are still needed for nanopore sequencing to become a highly robust and standardized approach for a comprehensive analysis of genomic structural variation.

**Keywords** Third-generation sequencing, Whole genome sequencing, Nanopore Sequencing, Hybrid-SNP microarray, Structural variation, Copy number variation, Variant calling

*Correspondence:
Luís Vieira
luis.vieira@insa.min-saude.pt
Full list of author information is available at the end of the article

## Background

The human genome consists of chromosomal segments whose copy number may vary from individual to individual. Copy number variations (CNV) are a subtype of genome structural variation that results from the evolution of the human species over thousands of years and contribute markedly to the diversity of its populations [1]. However, although genomic variation is the result of the evolutionary process, in certain cases structural rearrangements of a certain chromosomal segment may be associated with a genetic disease [2]. A 1.5 Mb tandem duplication at 17p12 in the peripheral neuropathy Charcot–Marie–Tooth disease type 1A and the deletion at del(17)(p11.2) in the Smith–Magenis syndrome were two of the first pathogenic CNV to be recognized [3, 4].

Although there are currently numerous cases of germline genetic disease associated with the presence of recurrent CNV, these structural variations are also present in several types of cancer. As an example, CNV are known to be associated in patients with different types of cancer with amplification of the *C-MYC* gene [5–8] or with homozygous deletions of the *CDKN2A* gene [9–12]. In the context of cancer, copy number amplification always leads to expression level upregulation whereas copy number loss results in expression level downregulation [13]. CNV have long been recognized as influencing not only the expression levels of genes present in the region of genomic variation, but also the cellular transcriptome as a whole [14].

Initial mapping studies of so-called large-scale copy polymorphisms (>100 kilobases in length) in the human genome revealed the existence of many dozens of genes with genomic imbalances in different individuals [15, 16]. A subsequent study showed that this type of variation can encompass up to 10% of the human genomic sequence and is most prevalent in the pericentromeric and subtelomeric chromosomal regions [17]. Although the definition is currently not very precise, CNV must comply with the criteria of repetitive nature, individual numerical variability and minimum length [18], which according to the most recent convention is of 50 base pairs (bp) [19]. Structural variants (SV) cover most of the variant bases of individual human genomes as opposed to single nucleotide variants plus small indels [20].

Recognition of the clinical relevance of CNV has promoted the development of sensitive technologies to identify variation of chromosomal regions in the human genome as part of the study of genetic diseases. Chromosomal microarray (CMA) and short-read sequencing (SRS) are currently the most used technologies for detecting CNV in pathologies of genetic origin. Array Comparative Genomic Hybridization (aCGH), Single Nucleotide Polymorphism (SNP) Array and a combination of both (hybrid-SNP microarray) are CMA technologies considered first-line tests for the diagnosis of a wide range of genetic pathologies [21], which allow detection of gains, losses and homozygosity of chromosomal regions. These technologies use an array of fixed oligonucleotides (probes) on a solid surface, which were designed to cover the entire human genome sequence, but with a higher-density pattern in intragenic regions. Apart from non-polymorphic marker probes that detect changes in copy number, oligonucleotides may contain within their sequence a SNP, enabling the determination of an individual sample's genotype (homozygous or heterozygous). CMA platforms can differ by the number and distribution of genome probes, which can affect the detection of small regions with gains or losses, or with loss of heterozygosity, as well as the precise location of its breakpoints.

Sequencing of the whole genome, whole exome, or gene panels on SRS platforms has also been used for the detection of SV [22–24]. Computational methods for calling SV on sequencing data are based on analysis of read pairs, read depth and split reads, on assembly of reads, or combinations of several of the underlying algorithms [22]. The first two algorithms use information from paired-end reads to verify whether they map at the expected distance and have the expected orientation, or whether the coverage depth of the reference sequence is identical in the two mapped regions, respectively. MSeq-CNV utilizes depth of coverage and insertion sizes of mate pairs to detect genomic deletions and duplications [25]. The split read strategy searches for read alignments in two distinct genomic regions, as in the case of a translocation. The assembly approach uses the generated contigs to align against the reference sequence and identify regions with structural variation. However, the length of short reads (typically 100–300 bases) leads to low mappability in repetitive sequences of the human genome, which are frequently associated with CNV [26].

Nanopore sequencing is a long read sequencing (LRS) technology that allows native DNA or RNA molecules to be sequenced when they are forced to move through the inside of a nanopore under the action of an electrical current [27, 28]. A bi-directional recurrent neural network algorithm basecalls the sequence of bases using the pattern of intensity changes in the electrical current during the migration of nucleotide chains. Nanopore sequencing can generate reads that are theoretically identical in size to the number of bases of the molecule being sequenced, typically between 1 and 100 kilobases (kb), but which can exceed 1 million bases in length [29]. This feature of the technology makes it particularly suited for discovering SV using dedicated variant calling methods and investigating the association of the different types of SV

with pathological conditions [30]. More recently, optical genome mapping (OGM) technology has emerged as an alternative imaging approach for the detection of SV whose data can be analysed in combination with sequence and/or CMA data. Although OMG can be used to detect structural variations such as inversions, translocations, and large deletions or duplications, it has a minimum resolution of about 500 bp [31].

In this study, we aimed to evaluate the performance of nanopore sequencing technology as an alternative approach to CMA for the detection of CNV in the human genome. We first used hybrid-SNP microarray to determine a high confidence "truth set" of large CNV in two different tumor cell lines, and then applied nanopore sequencing technology to the same samples to evaluate the variant calls of the truth set and characterize the differences between the two methodologies variant-by-variant.

## Methods

*Selection and preparation of biological samples.* In this work, two human cell lines acquired from the DSMZ-German Collection of Microorganisms and Cell Cultures GmbH (Leibniz Institute, Germany) were used, derived from cases of acute myeloid (eosinophilic) leukemia (cell line EOL-1) [32] and B-cell precursor acute lymphoblastic leukemia (cell line 697) [33]. EOL-1 cell line has a hyperdiploid karyotype (with 7.5% polyploidy) described as follows: 50(48–51), XY,+4,+6,+8,+19, del(4)(q12)×2, del(9)(q22). 697 cell line has a near diploid karyotype described as follows: 46(45–48), XY, t(1;19)(q23;p13), del(6)(q21). The use of two tumor cell lines was based on the following premises: (i) tumor lines that are well established in culture for many years present evidence of multiple high allele frequency genomic alterations, including CNV, that can be used to test and validate variant calling methods, (ii) each of these cell lines presents clonal alterations that result from cellular transformation processes associated with different cellular phenotypes of haematological neoplasia and which, therefore, differ in the number, type, length and genomic localization of CNV and, (iii) tumor cells in culture are a priori less contaminated with normal cells compared with primary tumor tissue, avoiding the impact of somatic mosaicism in the analysis of variants. This study focused on high allele frequency SV, i.e., variations that are detectable through hybrid-SNP microarray analysis in a manner comparable to the detection of germline structural variations in normal cells, but that may be of germline or somatic origin. Somatic alterations associated with the underlying pathologies or those that may have been acquired during clonal evolution in culture may not be technically distinguishable from germline alterations as they are also present in all, or nearly all, cells of the sample. Cell culture was performed according to the recommendations of DSMZ and to standard cell culture and harvesting procedures for cells growing in suspension. Following harvesting, ~ $5 \times 10^6$ cells were frozen in liquid nitrogen according to standard procedures. Genomic DNA extraction, quantification and quality evaluation were performed as described in 2022 by Silva et al. [34].

*Chromosomal microarray analysis and CNV validation*. Fifty nanograms of each cell line DNA were used for chromosomal microarray analysis with the Applied Biosystems CytoScan HD® array (ThermoFisher Scientific, California, USA) according to the manufacturer's recommendations (Affymetrix manual protocol Affymetrix® Cytogenetics Copy Number Assay P/N 703038 Rev. 3). The CytoScan HD® array contains 2,696,550 copy number markers including 743,304 single nucleotide markers, resulting in an average marker spacing of 1,148 bp in the complete hg19 human genome reference sequence. The raw signal intensity data (CEL files) were converted to CYCHP files using the "single sample analysis" method (which compares the values in a CEL file with the values in a reference model file for the CytoScan HD® array) and the "normal diploid analysis" method (in which over 50% of the genome is considered to be rearranged). The hg38 human genome sequence was chosen as reference. The CYCHP files were opened in Genotyping Console v4.0 and Chromosome Analysis Suite 4.0.0.385 with NetAffx na33.2 in a desktop computer with Intel Pentium CPU G3220 @ 3.00 GHz processor, 8.00 GB RAM, and 64-bit Windows 10 Enterprise version operating system. The reference model file used was CyoScanHD_Array.na33.r3.REF_MODEL. The following genome filters and corresponding values were applied: (i) minimum marker count for gains or losses = 15 and (ii) minimum segment length for gains or losses = 35,000 bp. The quality metrics analyzed were the SNP Quality Control (SNPQC), the Median of the Absolute values of all Pairwise Differences (MAPD) between $\log_2$ ratios and the waviness-Standard Deviation (waviness-SD). The thresholds used for evaluation of QC metrics were the following: MAPD ≤ 0.25, SNPQC ≥ 15 and waviness-SD ≤ 0.12. Regions called by the software as "loss of heterozygosity" and mosaic variants, were not considered in this study. An experienced cytogeneticist subsequently curated the remaining called gain and loss regions. Curation analysis included the coverage of the probes in the region covered by the variant, the contiguity of two or more nearby loss or gain regions, and the comparison of the start and end coordinates with an internal database of variants, in order to exclude uncertain calls including possible array artefacts. CNV were also compared with data available from the University of California Santa Cruz Genome Browser (https://

genome.ucsc.edu/), DECIPHER (https://www.decipher genomics.org) and ClinGen (https://www.clinicalgenome.org/). The curated and validated CNV gain and loss set ("truth set") for each cell line was exported in the VCF file format.

*Library preparation and nanopore sequencing*. Fourteen genomic libraries (8 for EOL-1 and 6 for cell line 697) were prepared for sequencing using the SQK-RAD004 Rapid Sequencing Kit (Oxford Nanopore Technologies- ONT) according to the manufacturer's instructions. Sequencing was performed on the MinION device using SpotOn Mk I R9.4 RevD flow cells (FLO-MIN106D, ONT). One of the flow cells was reloaded with a new library from the same cell line (697) following a nuclease wash with flow cell wash kit (EXP-WSH004, ONT). The MinKNOW software (MinION Release 19.06.8) was installed on a portable computer and used to program the sequencing run with default parameters. The electrical signal acquired in real-time was stored in FAST5 files.

*Structural variant data analysis workflow*. The FAST5 files were transferred to a HPE ProLiant DL385 Gen10 Plus v2 server (with 100 CPU, 766 GB RAM and Ubuntu 18.04.6 LTS operating system) to perform base calling (Guppy version v 3.0.7). The FASTQ files corresponding to each batch of 4,000 FAST5 files were concatenated into a single FASTQ file per run. Subsequently, the FASTQ files from different sequencing runs of the same cell line were concatenated into a single FASTQ file for downstream analyses. The primary assessment of sequencing data quality was performed based on metrics provided by MinKNOW and Nanoplot [35]. Read mapping was performed using lra [36]. Mapping quality metrics were obtained with qualimap2 [37]. SV were called with CuteSV [38] and Sniffles2 [39] using the BAM files produced by lra. For CuteSV, optimized calling parameters were applied, including a minimum read support of 2, a minimum variant size of 35 kb and a maximum variant size consistent with those observed in the CMA truth set for each cell line. BCFtools [40] was employed to refilter the resulting VCF files from Sniffles2 applying a minimum read support of 2, a minimum variant size of 35 kb and a maximum variant size matching the range observed in the CMA truth set for each cell line. Ribbon software version 1.0 [41] was used to visualize and interpret contradictory SV type calls between array and sequencing for the same CNV. Coverage depth calculation analysis was performed using mosdepth version 0.3.9 [42]. The Integrative Genomics Viewer (IGV) tool [43] was used to inspect and compare the CNV regions and respective flanking regions between the array and sequencing data. Command lines for running the various bioinformatics tools are described in supplementary file 1.

*Assessment of nanopore sequencing performance metrics*. Truvari v5.3.0 was used to calculate the recall performance metric depending on the reciprocal overlap (RO) threshold. The Truvari bench command was run with the following modifications: pctsize = 0, sizemax = maximum variant size, and the typeignore flag was included. The pctovl parameter ranged from 0.1 to 0.9 in steps of 0.1. For each RO threshold tested, the mean and standard deviation of the recall values for both cell lines were calculated. R software packages dplyr [44], scales [45], ggpattern [46] and ggplot2 [47] were used for computing Pearson correlations and for plotting. The VCF files generated by CuteSV and Sniffles2 for each cell line were filtered using the BCFtools v1.12 view command [40]. This included variants corresponding to breakpoint ends (BND), inversions (INV), and insertions (INS), variants that mapped to alternative hg38 contigs and the mitochondrial genome, that had a variant allele frequency < 0.15 and a size that exceeded the maximum CNV size observed in each cell line. Next, Survivor v1.0.7 [48] was used to create a merged set of CNV called by both CuteSV and Sniffles2 for each cell line. The parameters of the merge command were as follows: maximum breakpoint distance of 500 bp, SV type matching, identical strand orientation and a maximum size difference of 30% between variants called by CuteSV and Sniffles2. Variants with redundant calls were removed from the merged VCF file using Truvari v5.3.0 (collapse command) with parameters as follows: pctovl = 0.5, pctsize = 0.1, sizemax = maximum variant size and refdist = maximum distance of SV start positions (20 kb for EOL-1 and 40 kb for 697). Finally, the intersect command from BEDtools v2.30.0 [49] was used to remove the resulting non-redundant variants that overlap with genomic coordinates of "Centromeres". The bed file for these regions was retrieved from the Table Browser tool ("Mapping and Sequencing" group) of University of California Santa Cruz. Precision, recall and the F1-score performance metrics were calculated for each cell line independently. Plots were generated using ggplot2 [47]. Command lines for running the bioinformatics tools and the code for generating plots are described in supplementary file 2.

*Breakpoint mapping using polymerase chain reaction and Sanger sequencing*. Two CNV losses detected in the EOL-1 cell line and one CNV loss detected in the 697 cell line were selected for exact breakpoint localization using polymerase chain reaction (PCR) and Sanger sequencing. Sequences surrounding predicted breakpoint positions of each variant derived from SV calling results of sequencing data were inspected using IGV with respective BAM and VCF files. Forward and reverse primers were designed in the outer side of each variant breakpoint position using Primer3web version 4.1.0 at https://prime

r3.ut.ee/. PCR was performed in a SimpliAmp Thermal Cycler (Applied Biosystems) using standard reaction mixes and cycling parameters. PCR products were run on an ethidium bromide-stained agarose gel for evaluation of molecular sizes and purified using the Illustra ExoProStar 1-step, enzymatic PCR and sequence reaction clean up kit (Cytiva) according to the manufacturer´s instructions. Purified fragments were subjected to chain-termination sequencing using BigDye Terminator v1.1 Cycle Sequencing RR-100 (Applied Biosystems) and subsequently purified from unincorporated nucleotides with DyeEx 96 plate (Qiagen). Sequencing products were electrophoresed in a 3500 Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions. Basecalling was performed using Sequencing Analysis Software 6 (ThermoFisher Scientific) and sequences were compared with the human genome reference sequence hg38.

## Results

*Quality analysis of chromosomal array and sequencing data*. The quality metrics for the CMA data for each cell line are shown in Supplementary Table 1. In the two cell lines analysed, the MAPD was 0.18 (697 cell line) and 0.20 (EOL-1 cell line), indicating low noise (MAPD < 0.25) in the array data and its suitability for CNV analysis. The SNPQC was higher than 15 in cell line 697 (17.2) and slightly lower (14.1) in cell line EOL-1, indicating that the call rate of the latter may not have been optimal. The waviness-SD values were 0.20 (EOL-1) and 0.17 (697), higher than the threshold of 0.12 defined for the CytoScan HD® array. Since waviness-SD takes into account variation in $\log_2$ ratios across the genome, high waviness-SD values, in the presence of normal SNPQC and MAPD values, likely reflect the presence of multiple CNV gains and losses in the genome of both tumor cell lines. The quality data of the sequencing runs extracted by Nanoplot are described in Supplementary Table 2. A total of 7.865.129 reads (4.619.303 and 3.245.826 reads for EOL-1 and 697, respectively) were available for analysis post-mapping. The mean coverage depth in the EOL-1 and 697 cell lines was 13.1X and 11.4X, respectively. Some of the key metrics of the sequencing runs for each of the lines are displayed in Supplementary Fig. 1.

*Description of variants validated in CMA*. The number of CNV gains/losses called by the chromosomal analysis software (ChAS) in the EOL-1 and 697 cell lines were 28/23, and 24/20, respectively. Forty-eight validated variants resulted after exclusion of mosaic variants and low confidence calls, the type and size distribution of which are shown in Supplementary Fig. 2. These included 22 variants in the EOL-1 cell line and 26 variants in cell line 697, comprising 30 CNV losses and 18 CNV gains greater than 35 kb, ranging from a minimum of 35,866 bp to a maximum of 79,644,334 bp. The copy number for deletions was zero (6 variants) or 1 (24 variants) whereas the copy number for duplications was 3 (15 variants) or 4 (3 variants). This set of 48 CNV, called the "truth set", provided a range of variant types, sizes and copy numbers that served to test the performance of two SV callers on the nanopore sequencing data.

*Structural variant calling of sequencing data*. We used nanopore sequencing to generate long reads from the genomic DNA of the two cell lines with the aim of evaluating the performance of CNV calling. In this approach, we adjusted the calling parameters of CuteSV and Sniffles2 with the purpose of maximizing the calling of the CNV truth set. The SV callers called 38 of 48 of the variants (79.2%) in the truth set (Fig. 1). Five CNV losses and five CNV gains were not called by either of the 2 callers, of which two (9.1%) in cell line EOL -1 and 8 in cell line 697 (30.8%). Although both callers identified many more CNV with sizes under 35 kb, we focused only on those CNV that matched the range size determined in CMA analysis. We did not perform an analysis of variants smaller than 35 kb because data evaluation in genetics laboratories focuses on larger CNV, albeit using different analysis thresholds, and because accurate and reproducible identification of CNV depends not only on probe density but also on the minimum number of probes to ensure statistical confidence of detection.

The difference in the percentage of detection of the variants between the two cell lines is unlikely to be a result of the lower depth of coverage of the 697 cell line (only 1.7X less depth than the EOL-1 cell line), although it had better raw sequencing data quality than EOL-1. Additionally, three of the remaining 13 CNV gains of the truth set were called only by CuteSV (2 variants) or Sniffles2 (1 variant) alone, resulting in a CNV correct call percentage of 77,1% for CuteSV and 75% for Sniffles2. Considering the use of any of the SV callers and with the respective call-specific parameters, the percentage of correct calls was much higher for CNV losses (83.3%) than for CNV gains (61.1%). In particular, the five CNV losses not detected by CuteSV or Sniffles2 had a copy number of one, indicating that homozygous deletions (corresponding to a copy number of zero in the CMA data) are more easily called in the sequencing data. In contrast, either one or both of the two SV callers missed 2 out of 3 CNV gains (66.7%) with copy number 4 and 6 out of 15 gains (40.0%) with copy number 3. Although numbers are small, these results suggest a less precision of SV calling algorithm parameters for detection of different copy number gains.

*Variant size comparative analysis*. The differences in the length of the variants called by CMA and any of the SV callers were compared using different metrics.

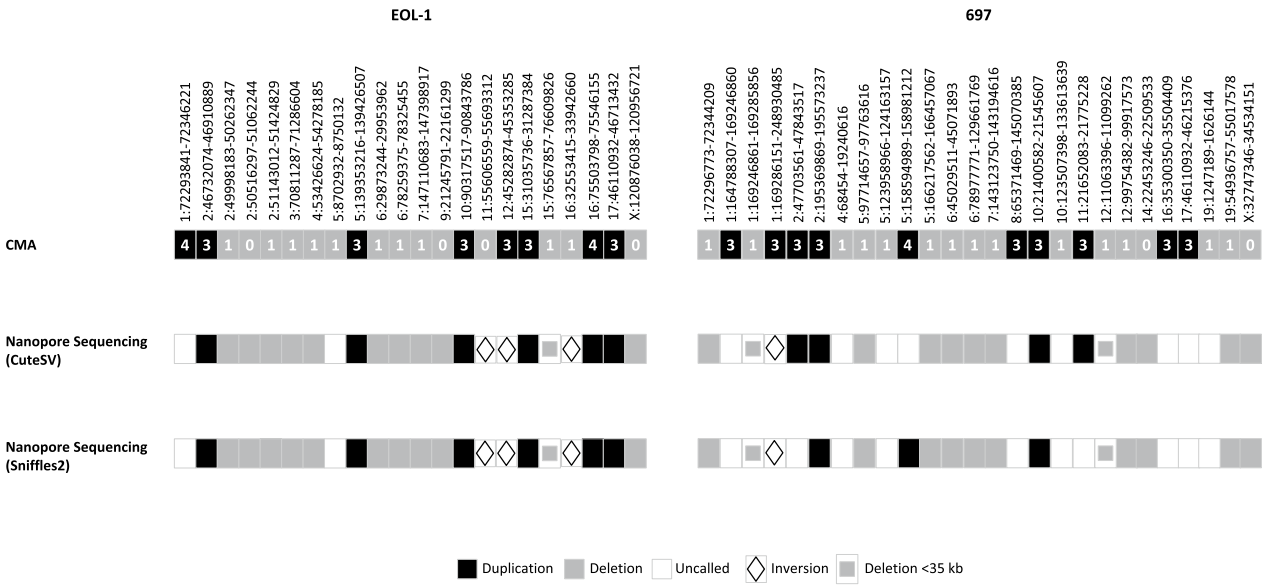


**Fig. 1** Schematic representation of the variants identified using CMA and corresponding matches/unmatches following SV calling in nanopore sequencing data for cell lines EOL-1 and 697. The chromosome and start and end genomic coordinates according to the CMA data are shown for each of the variants. The numbers inside the boxes in the CMA variants represent the copy number

Size differences were assessed for each individual cell line/SV caller combination using RO. In this analysis, the VCF files were filtered to keep only variants called by CuteSV or Sniffles2 that matched a CNV present in the CMA truth set were used, regardless of the variant type (deletion, duplication or inversion). Recall was calculated as a performance metric for each RO threshold (range: 0.1–0.9; step: 0.1). The average of the recall values obtained for the two cell lines was used to show the relationship with respect to the different thresholds. The average recall for each caller was ≥ 60% for a maximum RO = 0.9 and tends to increase similarly for both callers as the RO decreases. A ~ 85% recall was obtained for each caller taking as reference a RO = 0.5 when comparing CNV sizes obtained by CMA and nanopore sequencing (Fig. 2A). The Pearson correlation between the sizes of the CNV gains and losses called by CuteSV or Sniffles2, and the sizes of the same variants called in the CMA, using the set of variants detected in the 2 cell lines, was in both cases of 0.975 (Fig. 2B, C). The correlation was very high across the range of sizes analysed. Of notice, in three CNV losses the size of the variant called by any of the SV callers was clearly smaller than the cutoff size used for calling CNV in CMA (35 kb). In the EOL-1 cell line, the chromosome 15 CNV loss was called with a size of 41,969 bp in CMA and 12,342 bp in nanopore sequencing (Supplementary Table 3). Likewise, one of the chromosome 1 CNV losses and one of the chromosome 12 CNV losses in cell line 697 were called with sizes of 38,995 bp and 35,866 bp in CMA, compared to sizes of 22,124 bp and 5,254 bp (CuteSV) or 5,248 bp (Sniffles2), respectively, in nanopore sequencing (Supplementary Table 4). By contrast, an almost absolute correlation was observed when comparing the sizes of variants that were called by both SV callers (Fig. 2D). Of a total of 38 variants called in CMA that were also called by CuteSV, Sniffles2, or both, twenty-five (63.2%) showed a larger size in the CMA variant set. While in the 13 CNV gains the size obtained was larger in CMA in seven variants (53.8%), seventeen of 25 CNV losses (68%) presented with a larger size in CMA.

*Analysis of variant breakpoints at nucleotide level.* The breakpoints of three CNV losses called by CuteSV and Sniffles2 in cell lines 697 and EOL-1 were sequenced by the Sanger method to determine the difference in the number of nucleotides relative to the genomic breakpoint coordinates assigned by the SV callers (Fig. 3; Supplementary Fig. 3). Using the coordinates obtained with CuteSV as a reference, the average difference in the number of nucleotides for the 6 breakpoints determined by Sanger sequencing was 20 bp (range: 0–40 bp) (Supplementary Table 5). Additionally, Sanger sequencing revealed the presence of an interspersed sequence of unknown origin, 9 bp in length, between the breakpoints of the CNV loss of chromosome 5 in cell line 697.

*Discrepancies in variant types called using array and nanopore sequencing data.* Both SV callers called two CNV losses in the EOL-1 cell line, and two CNV gains, one in each cell line, as inversions. These variants showed sizes in CMA ranging from ~ 70 to ~ 79 Mb.

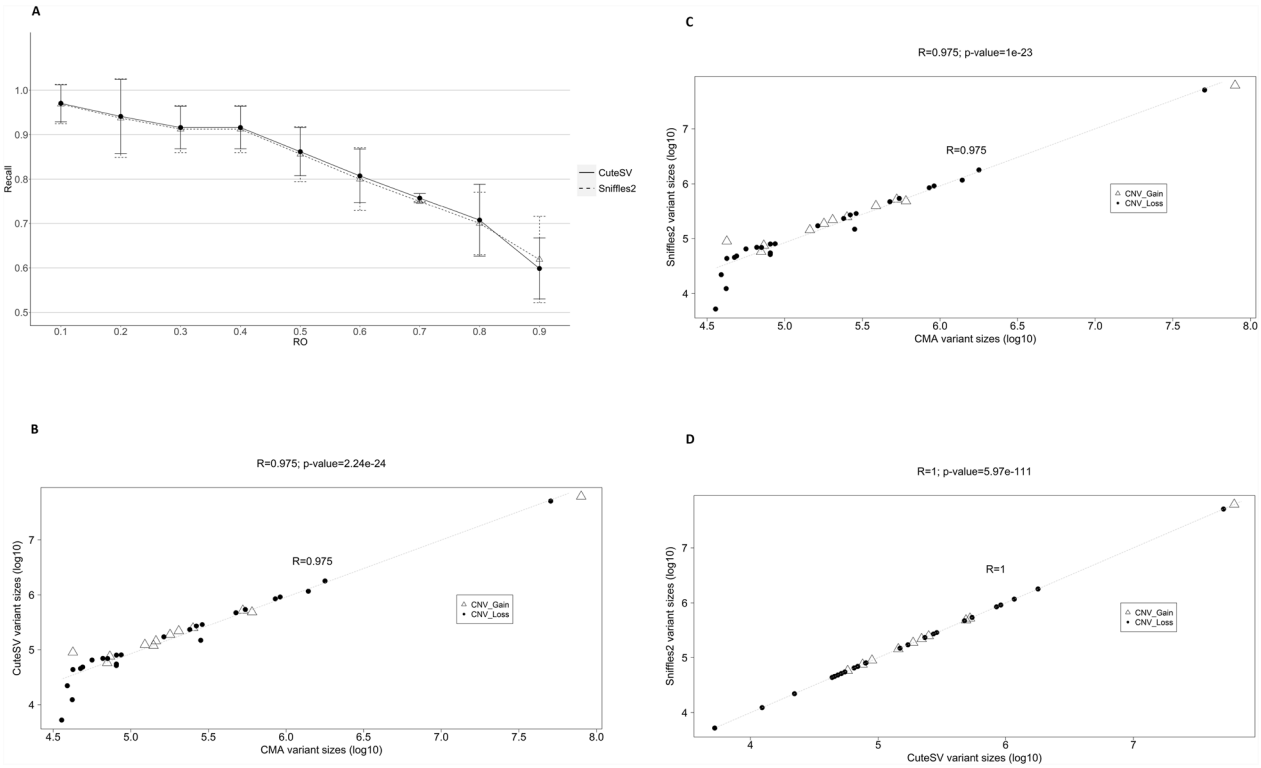


**Fig. 2** Graphical representation of variant size comparisons between variants. **A** Variation between mean recall and reciprocal overlap for each SV caller. Vertical bars indicate the standard deviation of the recall values obtained for each cell line. **B** Pearson correlation between variant sizes called in CMA and with CuteSV; **C** Pearson correlation between variant sizes called in CMA and with Sniffles2; **D** Pearson correlation between variant sizes called by CuteSV and Sniffles2. Variant sizes are shown as log(10) of basepairs

The visualization of discrepant calls in order to elucidate the mechanisms of genomic rearrangement was carried out using Ribbon, since the IGV software is not suitable for visualizing complex rearrangements involving very extensive regions. In all four CNV called by CMA, a similar pattern is evident when viewing the position and orientation of the reads in relation to the genomic coordinates of each variant (Supplementary Fig. 4). Genomic regions adjacent to the breakpoints of each variant show an inverted location in reads that overlap the call region. In all situations, the presence of a large internal deletion and, in the case of the chromosome 11 variant in the EOL-1 cell line, other small deletions close to the breakpoints, are evident. In other words, the genomic rearrangements that underlie these four variants could have involved the inversion of a genomic sequence with concomitant loss of an internal region. This evidence supports the call for CNV loss in CMA for the chromosome 11 and chromosome 16 variants in the EOL-1 cell line, since CMA does not allow the identification of inversions. In the first case, all four reads mapped in the region support the presence of the variant, which is compatible with the copy number of zero (homozygous deletion) obtained in CMA. In the second case, four reads out of a total of 14 reads mapped in the region support the existence of the variant, which is compatible with a copy number of 1 (heterozygous deletion). In the case of the chromosome 12 variant in the EOL-1 cell line, the CMA call has an additional length of ~ 14 kb at one end of the variant region, and is covered by 23 reads in nanopore sequencing, well above the mean coverage depth in the EOL -1 cell line that was 13.1X. This evidence suggests that this region is duplicated in accordance with the CMA data. In parallel, there is a total of 10 reads (43,5%) that support the existence of an inversion and absence of coverage in the region internal to the terminal coordinates of this variant, justifying the call for an inversion with an internal deletion. In the case of the CNV gain of chromosome 1 in the 697 cell line, and despite the size of the variants called in CMA and in nanopore sequencing not being very different (79 Mb versus 62 Mb, respectively), the overlapping region of the variants called in each technology is only around 37 Mb. Given that only a minority of reads (3 out of 15) support the existence of the variant, it is possible that this genomic change was not detected in CMA or it is a false call of sequencing data.

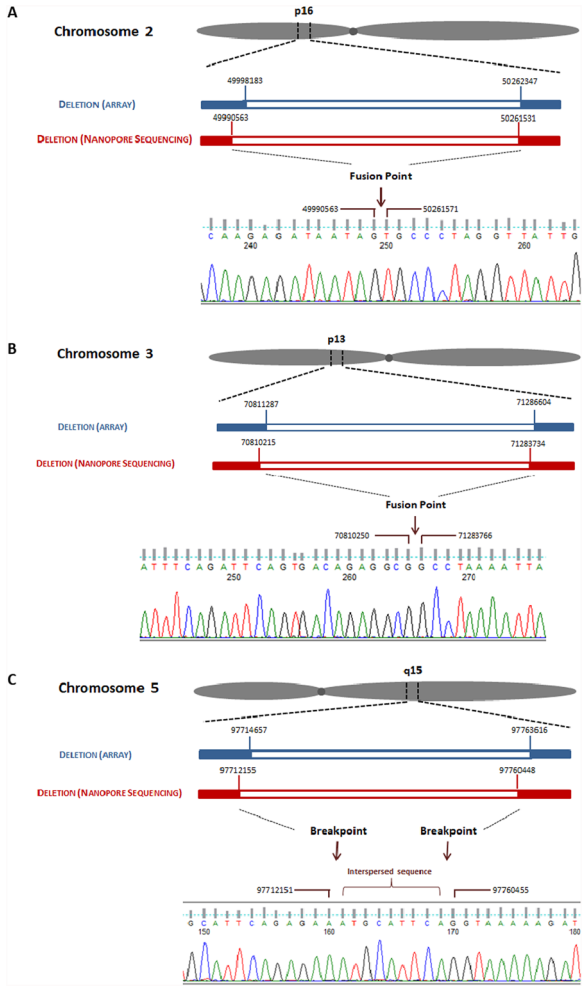


**Fig. 3** Schematic representation of breakpoint confirmation for 3 variants by Sanger sequencing. **A** Deletion on chromosome 2 in EOL-1 cell line; **B** Deletion on chromosome 3 in EOL-1 cell line; **C** Deletion on chromosome 5 in 697 cell line. Filled bars indicate the retained sequence portion, while unfilled bars correspond to the deleted sequence. Numbers correspond to genomic coordinates according to the human genome reference sequence hg38. In the case of nanopore sequencing, the coordinates correspond to those obtained using CuteSV

*Discrepancies in variant sizes between array and nanopore sequencing data*. This analysis allowed us to compare, from the reads mapped to the reference genome, the coverage of the genomic regions corresponding to the positions of the variants called in the CMA data with the coverage of the flanking genomic regions on one or both sides of the variants. This approach allows visually highlighting the existence of a CNV and its length without resorting to SV calling. In Fig. 4A, a CNV loss with copy number zero is called in nanopore sequencing data as a deletion with approximately the same size as in CMA. However, it is clear in the coverage data that there are two contiguous regions, one with coverage zero and another with approximately half the coverage of the variant flanking regions, thus revealing a homozygous deletion adjacent to a heterozygous deletion within the same variant region. Additionally, coverage analysis of the three CNV losses that were significantly smaller compared to the sizes obtained in CMA, suggests that the length of these variants is closer to that obtained using sequencing. In the CNV loss of chromosome 1 in the 697 cell line, it is visible from the coverage analysis that the region identified as being deleted in CMA has its terminal and flanking regions with identical coverage, and that the internal heterozygous deletion extends over approximately half of the length obtained in CMA (Fig. 4B). In the other two CNV losses with copy number of 1, one on chromosome 12 in the 697 cell line and another on chromosome 15 in the EOL-1 cell line, the discrepancy between the calls obtained from CMA and nanopore sequencing is more striking. In both cases, coverage analysis indicated the existence of a homozygous deletion of much shorter length than that obtained in CMA (Fig. 4C, D).

*Coverage analysis of uncalled variants in sequencing data.* Subsequently, mosdepth was used to check whether the genomic positions covered by the truth set CNV not called by CuteSV and Sniffles2 showed evidence of coverage variation. Of the 5 CNV gains uncalled by both SV callers, three of them (on chromosomes 8, 16, and 17 of cell line 697) showed an average increase in coverage that overlapped with the extent of variants called in the CMA data, indicating that both callers failed their respective calls (Supplementary Fig. 5A-C). However, in the case of variants on chromosomes 16 and 17, they appear to be more extensive on the telomeric and centromeric sides, respectively, compared to the calls made in CMA. In the remaining two CNV gains, the increase in coverage in the variant region is not as evident and may reflect potential read mapping inconsistencies in those regions. Although in the CNV gain of chromosome 1 of cell line 697 a difference in coverage is observed on the centromeric side, on the opposite side no variation in coverage is observed (Supplementary Fig. 5D). In the CNV gain of chromosome 1 in the EOL-1 cell line, it is not evident that there is a duplication in the CNV region called in CMA, which justifies the absence of calling by any of the two SV callers (Supplementary Fig. 5E). Contrary to what was observed with the CNV gains, the five uncalled CNV losses in the EOL-1 or 697 cell lines are clearly discernible in the coverage data and coincide in their extents with the called variants in CMA, showing that either SV caller failed to call these variants in the sequencing data (Supplementary Fig. 5F-J). However, it should be emphasized that three of the 10 uncalled CNV have a terminal location on the chromosome, including the CNV gain on chromosome 8

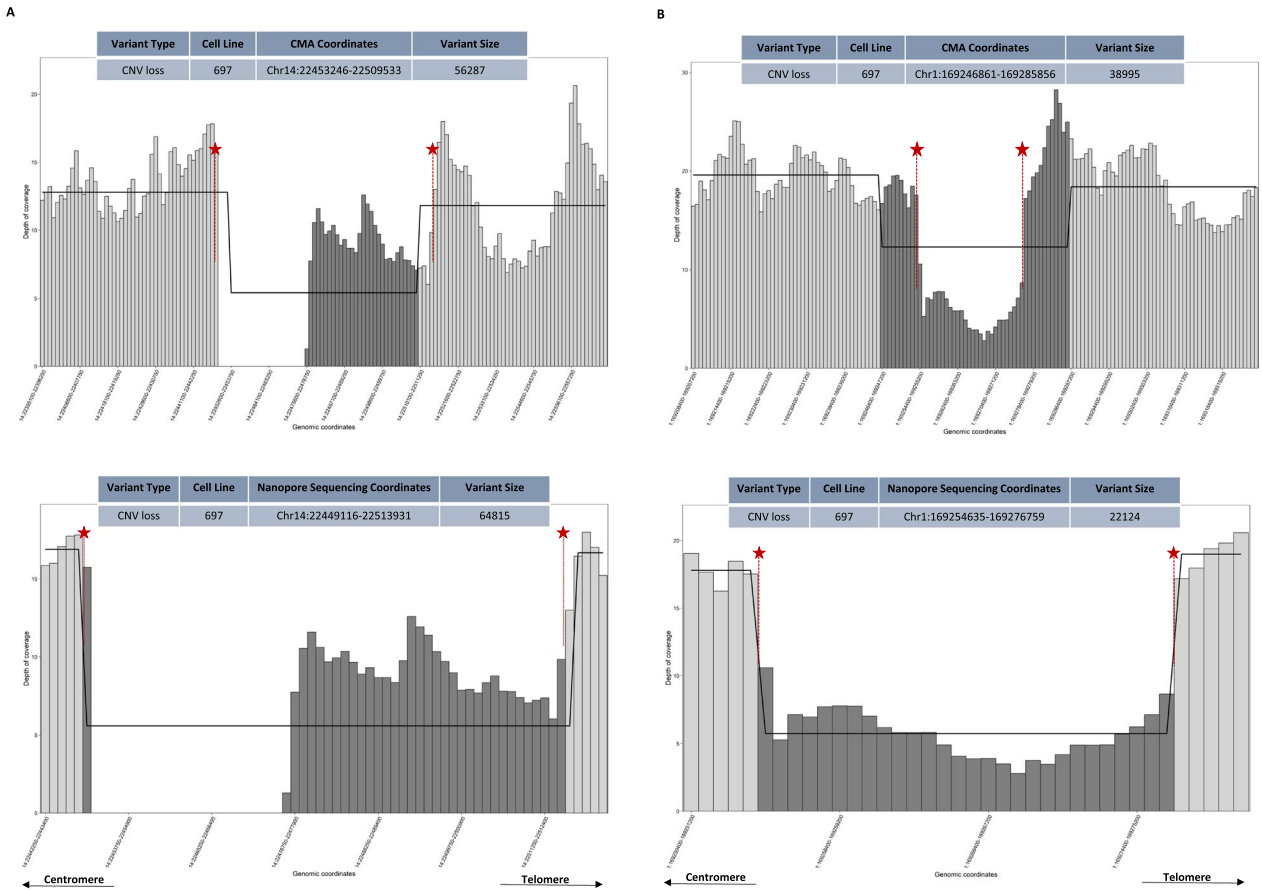


**Fig. 4** Depth of coverage bar chart of coverage analysis of CNV losses with discrepant sizes in nanopore sequencing. **A** CNV loss on chromosome 14 in cell line 697; **B** CNV loss on chromosome 1 in cell line 697; **C** CNV loss on chromosome 12 in cell line 697; **D** CNV loss on chromosome 15 in cell line EOL-1. The upper bar chart represents the average coverage of genomic bins in the variant region called in CMA (dark grey bars) and the flanking regions (light grey bars). The lower bar chart represents the magnification of the upper chart. The red stars show the terminal positions of the variant called in sequencing data. The horizontal black line represents the average depth of coverage in the variant region and in the flanking regions. The bin size was initially calculated so that each bin represents a genomic region equivalent to variant size/50, i.e., each variant is represented across ~ 50 bins. Each of the flanking regions is also represented by ~ 50 bins. Plots were generated using ggplot2

and the CNV losses on chromosomes 4 and 10 in cell line 697. In contrast, only the CNV gain of chromosome 1 in the same cell line, out of the 38 variants called by both SV callers, had a terminal location on the chromosome. However, this variant was called as an interstitial inversion on chromosome 1, which reinforces the limitation of SV callers in identifying variants that span the terminal location of chromosomes.

*Variant filtering and performance metrics evaluation.* The VCF files produced by CuteSV and Sniffles2 for each sample were filtered by variant type (BND, INV or INS), allele frequency and mapping region. The unfiltered variants from each VCF file were merged to retain only the variants called by both callers, and then redundant variants corresponding to overlapping calls were removed. Finally, variants located in centromeric regions were excluded since these regions are highly repetitive and can cause ambiguous read alignment, leading to false SV calls. The filtering process reduced the set of variants to only 65 and 70 in the EOL-1 and 697 cell lines respectively, starting with a total of 2396 (1394 for CuteSV and 1002 for Sniffles2) and 3030 (2394 for CuteSV and 636 for Sniffles2) called variants that had previously been filtered solely based on read support and minimum and maximum variant sizes. These results show that it was possible to reduce approximately 98% of the number of false positive variants, although some of the true positive variants (8 in the EOL-1 cell line and 15 in the 697 cell line) were also filtered, namely because they had a low allele frequency or were called by only one of the SV callers. The precision, recall and F1-score metrics were 16.9%, 57.9% and 26.2% for the EOL-1 cell line and 14.3%, 40.0% and 21.1% for the 697 cell line. It is worth mentioning that the values obtained for the precision metrics should

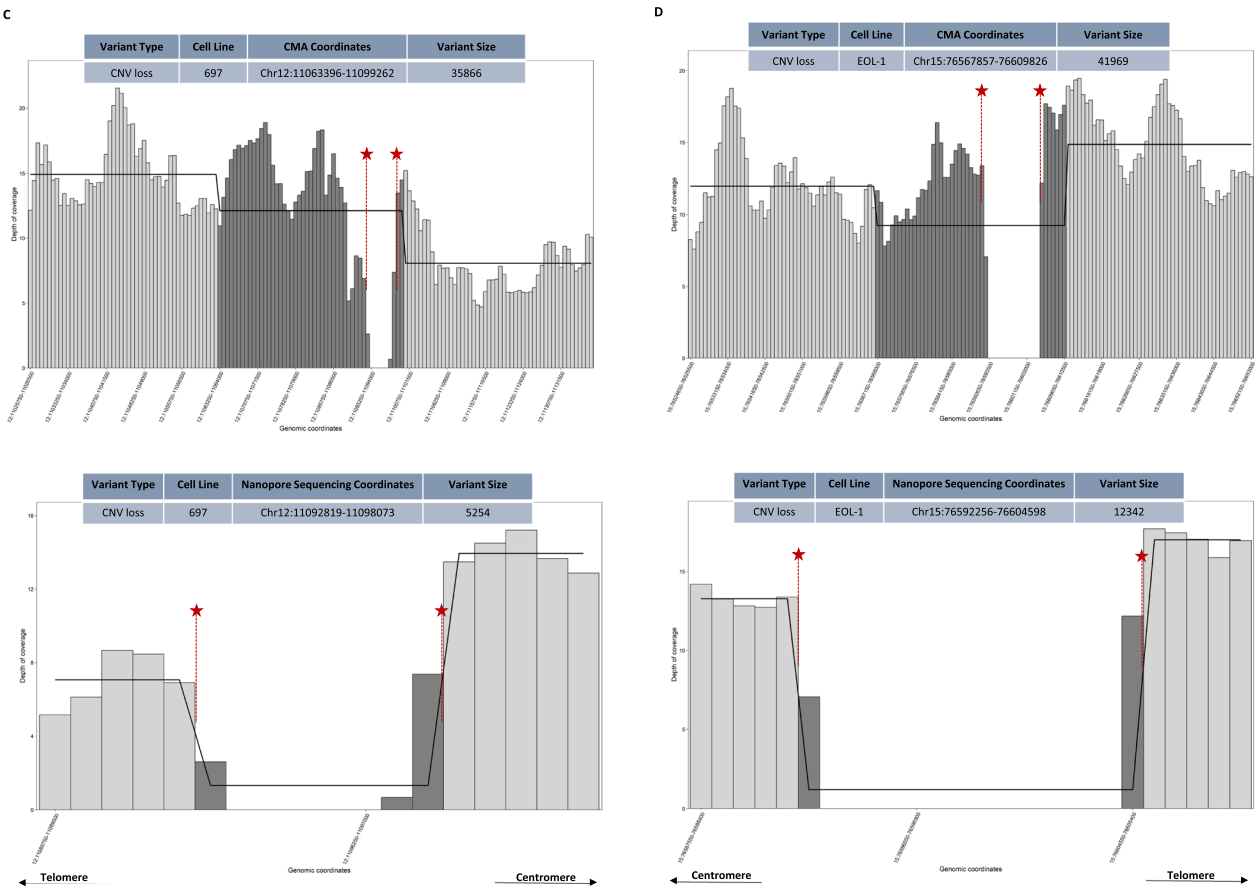


**Fig. 4** continued

not be generalized, since the number of true positive variants for each cell line is very small and, consequently, a small reduction in the number of these variants, as well as the presence of additional variants not present in the truth set, greatly affects the values obtained.

## Discussion

The study of human CNV has application in multiple areas of knowledge, such as evolution [50], susceptibility to genetic and common diseases [51, 52], diagnosis of genetic pathologies [53], impact on complex traits [54], relationship with metabolic state [55], association with the human microbiome [56] and cancer [57]. The biological and clinical relevance of CNV prompted the development of sensitive and robust technologies for their detection. After the establishment of conventional cytogenetic analysis, and the subsequent development of fluorescence in situ hybridization that allowed increased resolution at the chromosomal level, CNV analysis methods migrated to a molecular-level approach. High-resolution genomic arrays have gained an important role in the detection of CNV and are considered a gold standard approach for genetic diagnosis due to their high resolution, comprehensive analysis, clinical utility, and the possibility to detect uniparental disomy [58]. Subsequently, next-generation sequencing of short reads allowed the discovery of CNV and other types of SV in gene panel, exome and genome sequencing data, using diverse computational methods [59]. More recently, long-read sequencing on the ONT or PacBio platform has opened new possibilities for CNV identification by taking advantage of longer read lengths and new SV calling methods to characterize genomic variations [60].

In this work, we used hybrid-SNP and ONT long read sequencing of two cell lines to compare CNV calling variant-by-variant, using CMA data as the truth set. The truth set was composed of 48 validated CNV with lengths ranging from 35 kb to approximately 80 Mb. In this study, analysis of variants smaller than 35 kb was not performed for the following reasons. First, the purpose of this study was to evaluate the use of nanopore sequencing as a replacement for CMA for the genetic diagnosis of large CNV. Second, the manufacturer of the CytoScan HD® array claims that it can reliably detect

copy number changes ranging from 25 to 50 kb across the genome at high specificity. Finally, using the CytoScan HD® array and ChAS software for CNV calling, as we did in this study, it was shown that approximately 90% of non-validated variant calls (false positives) in the gold standard CNV set for the NA12878 genome were smaller than or equal to 25 kb in size [61]. Confident CNV detection depends not only on the array probe density (in this case, ~ 1 probe per kb), but also on the minimum number of probes required (between 25 and 50 probes) for statistical confidence in CNV detection. CNV smaller than 25 kb should be previously validated by orthogonal methods (e.g., quantitative PCR, multiplex ligation probe amplification, or high-coverage next-generation sequencing). Specifically, we employed two SV callers (CuteSV and Sniffles2) for the sequencing data, compared the types, extension and breakpoint location of variants called based on CMA, nanopore sequencing and/or Sanger sequencing, and analysed the coverage of the CNV and surrounding regions. The proportion of truth set variants that were called by one or both of the SV callers used in this work was 79.2%. Using only a single caller, the overall percentage of correct calls was 77.1% for CuteSV and 75% for Sniffles2, revealing a marginal increase in the detection rate using both callers. This result is in line with that obtained by other authors who demonstrated the added value of using more than one caller to improve CNV detection in whole genome sequencing data [22, 62]. Of the 13 variants not called by one or both callers, the majority (61,5%) corresponded to copy number gains, reflecting a higher rate of detection of copy number losses. Other studies have confirmed the greater difficulty in calling duplications than deletions, insertions or inversions in whole genome sequencing data produced with short reads [63]. In our study, no relationship between duplication length and CNV non-calling was evidenced for both callers, indicating that CNV size is not a factor that affects calling when working with an average coverage depth of 11-13X. Among the uncalled variants, three corresponded to terminal chromosomal deletions, while in the called variants only one other terminal deletion was present, albeit as an inversion with non-coincident breakpoints. Considering only interstitial variants within the truth set, the detection rate of CNV gains and losses using results from both SV callers increases to 81.3% and 89.3%, with an overall detection rate of 86.4%.

Helal et al. [64] recently conducted a benchmarking study where they evaluated the impact of depth of coverage on SV calling in the NA12878 sample, using different long-read aligners and SV callers. They found that using lra as the aligner, and CuteSV/Sniffles2 as SV callers, the precision, recall and F1-score decreased between 5.2%/1%, 24.5%/3.8% and 16.5%/2.5%, when the depth of coverage decreased from 30 to 10X, although the differences were also observed for other aligner/SV caller combinations. In the same study, both Sniffles and CuteSV detected a higher number of variants in the 50–250 bp range compared to larger size ranges using data from 3 different samples with 30X or 20X coverage depths. However, at 10X coverage, both SV callers detected a significantly reduced number of variants across all size ranges. Furthermore, CuteSV demonstrated consistent performance across different length groups, with F1-score values ranging from 89 to 95% using the reference sample NA24385 with high coverage depth (> 30X). In contrast, Sniffles showed unstable performance across different SV size groups, with F1-scores ranging from 42 to 73% [64]. These data show that SV calling performance metrics benefit from higher coverage depths. However, some SV aligner/caller combinations, such as those used in the present study, can exhibit highly variable performance depending on variant size.

In the present study, both lra/CuteSv and lra/Sniffles2 combinations identified, with only 10 exceptions, the majority (72,2%) of variants from both cell lines, which were > 35 kb in length and had a large size range. Although the overall performance metrics are modest, these results should be interpreted in the context of the reference dataset used for benchmarking. The ground truth set employed in this study contains only 48 validated variants, 22 in cell line EOL-1 and 26 in cell line 697. With such a small ground truth, the omission of just a few variants leads to a disproportionately large impact on recall. On the other hand, the reduced number of true positives indirectly affects precision as the method detects several additional variants (false positives). These variants may represent real variants that are not present in the ground truth set, notably because they may be located outside high-confidence regions of the genome [65] or due to benchmarking data limitations [66]. Consequently, the observed metrics likely underestimate the true performance of low-coverage nanopore sequencing in detecting large CNV (> 35 kb) compared to CMA.

CNV calling in CMA and nanopore sequencing evidenced a very high correlation in variant sizes. Still, the sizes of CNV called based on sequencing were larger in the majority (63%) of variants compared to CMA. The genomic coordinates of the breakpoints identified by SV callers are in principle more accurate than in CMA because this technology uses the coordinates located between the position of the probes immediately outside the CNV, and the position of the first probes inside the CNV, which can differ by several kilobases apart, to estimate the breakpoint locations. However, as shown in Sanger sequencing of three distinct CNV losses, there

can be a variation of up to 40 bp between the coordinate obtained by the SV caller and the true coordinate of the breakpoint. These results indicate that the accuracy of SV breakpoints in general is an aspect that could be improved in SV callers.

Three CNV losses called correctly by both SV callers revealed a size much smaller than that obtained in CMA. Analysis of sequencing coverage in the region of loss identified by CMA and its flanking regions suggests that there is no evidence in the sequencing data that supports the entire extent of the region of loss called in CMA. Coverage analysis of variant regions not identified with nanopore sequencing also revealed that 3 of the 5 CNV gains from cell line 697 had evidence of increased average coverage that is consistent with a copy number gain, even though in 2 variants the actual CNV extent may be greater than that obtained with CMA. Additionally, all five CNV losses not called by SV callers in both cell lines showed clear evidence of reduced average coverage in the loss region versus the flanking regions. These results show that further tuning of SV callers' algorithms and calling parameters is necessary to increase CNV calling accuracy. However, in the remaining 2 CNV not called by any of the SV callers, an increase in average coverage in the gain regions is not evident, which supports the hypothesis that read misalignment could have masked both regions and, consequently, providing no coverage support for CNV calling. Conversely, algorithmic limitations of SV calling rather than mappability problems likely affected the calling of CNV located in the terminal regions of chromosomes. In all three uncalled CNV with a terminal chromosomal location a reduction or an increase in coverage was clearly evidenced in the sequencing data.

One of the biggest differences that was seen between the truth set and the nanopore sequencing calls was related to four CNV that were called by the SV callers as inversions. Visualization of the genomic coordinates of the variants in the genome reference sequence, together with the location of the reads mapped to the variant region, showed that in all four cases there is a deletion internal to the inversion. This evidence supports the two CNV losses called in CMA since this methodology does not allow identification of inversions without a concurrent copy number change. In the case of chromosome 12 CNV gain in the EOL-1 cell line, a possible explanation for the discrepancy in variant type between CMA and nanopore sequencing is the simultaneous existence of both variants. The evidence of the presence of high coverage at one end of the variant called by the SV callers, and which region coincides with the additional 14 kb extension of the variant called in CMA, supports the presence of a duplicated region. On the other hand, the evidence for an inversion with internal deletion is supported by almost 50% of the reads that mapped to the region, suggesting the hypothesis of an inversion rearrangement in one allele and a duplication in the other allele, in which the regions involved are only coincident. In the CNV gain on chromosome 1 in cell line 697, the lack of coincidence in the extent of the inversion called in nanopore sequencing and the low number of reads supporting it, suggests that this variant may be present only in a minority of the cells analysed or that the call made by the SV callers is not real. In conclusion, with the exception of this last variant, nanopore sequencing provided further information on 3 variants called as CNV in CMA, which allowed elucidating the genomic events that gave rise to the divergent calls. Correct analysis of genomic alterations can have important consequences for the interpretation of the phenotypic effects caused by these variants in a clinical setting.

CMA is a well-established technology with several companies offering integrated solutions encompassing arrays with different resolutions and configurations, and automated data analysis of CNV without the need of bioinformatics expertise. Integrated laboratory and data analysis workflow is also possible in nanopore sequencing with the use of cloud-based EPI2ME for SV calling using a dedicated ONT pipeline (https://github.com/epi2me-labs/wf-human-variation). However, users often resort to locally implemented pipelines to diversify the bioinformatics methods and fine-tune analysis parameters, which may translate into a longer period for implementation of a sample-to-results workflow in nanopore sequencing compared to CMA. However, the ability of nanopore sequencing to generate real-time data and to stop sequencing runs when the desired amount of data is obtained, allows for a more adaptive and potentially faster approach than CMA in clinical diagnosis. Moreover, with the rapid library preparation protocol used in this work, the effective manual labour time is greatly reduced compared to CMA. Although whole human genome nanopore sequencing currently incurs higher costs per sample compared to CMA, especially when high genomic coverage is required, these costs have been progressively decreasing due to high competition in the global sequencing market. The lower cost of nanopore sequencing devices, particularly the GridION and the PromethION P2, is also an advantage over the cost of an array platform. In summary, nanopore sequencing is superior to CMA in accessibility and time efficiency, but is still not as competitive in terms of cost per sample for the sole purpose of detecting CNV in the human genome.

The results obtained in this work may be subject to certain limitations. Firstly, the use of tumor cell lines with significant genomic transformation may not be

considered an ideal study model. These lines can harbour subclones that alter the relative proportion of CNV in the main cell lineage or introduce new subclone-specific CNV in a small fraction of cells. A CNV present in a small allele fraction (e.g., < 20%) can escape variant calling when not parameterized accordingly, particularly when genomic coverage in these regions is low. Nevertheless, the use of tumor cell lines harbouring multiple genomic rearrangements in a single genome sequence is very useful for assessing SV callers' accuracy and for using as input for benchmarking purposes because they provide a diverse spectrum of genomic variations with large and highly variable sizes, which are not captured in a single normal individual sample. Secondly, this study was limited to two tumor cell lines (EOL-1 and 697), randomly selected from a collection of existing tumor lines in our laboratory. The main reason for using only two lines relates to the high costs of sequencing. Sequencing the complete genomes of these two cell lines required 13 MinION R9.4 flow cells at a final coverage depth of 11-13x, representing a significant cost. Despite this limitation, it was yet possible to characterize the CNV landscape that can be found in a human genome. These CNV include interstitial and terminal alterations with a variable copy number between zero and four, and with sizes spanning from 35 kb to 80 Mb in length. Although this goal was achieved, the results of the present study could also benefit from the use of the PromethION instrument instead of the MinIOn device. This solution uses higher-throughput flow cells, allowing for increased genome coverage and reducing the sequencing cost per Mb of DNA compared to the use of multiple MinIOn flow cells. Although quantitatively limited, our truth set was sufficiently comprehensive to enable a comparison between nanopore sequencing and CMA in the context of SV calling and is a proof of concept that nanopore sequencing could potentially be adopted in the future as a first-line approach for studying structural variation in the context of human disease. However, to strengthen these results, future studies should include broader validation, both by increasing the number of samples and by incorporating a wider range of sample types, particularly those commonly used for diagnosis, including formalin fixed paraffin-embedded (FFPE) and amniotic fluid samples, as well as samples produced by the Genome in a Bottle Consortium for the purpose of benchmarking human genome sequencing data.

## Conclusions

In this work we show that sequencing of the complete human genome by nanopore sequencing, even with shallow coverage (< 15X), allows identifying the majority (79%) of structural CNV than the gold standard CMA technique. Considering only CNV with interstitial chromosomal location, the percentage of detected CNV increases to over 86% using two SV callers for variant calling. The advantages of using nanopore sequencing for SV detection instead of CMA include the ability to detect all types of structural variation and their respective breakpoints with great precision. The long length of the reads obtained by nanopore sequencing also represents an advantage over short read sequencing, since it allows the identification of structural variations in individual reads, avoiding the need to sequence multiple samples to obtain a representation of the natural intra-sample coverage variation that is necessary to give greater robustness to CNV detection algorithms. Despite these advantages, mapping and variant calling algorithms still need to evolve in order that nanopore sequencing technology is used as a comprehensive methodology for studying all types of human genome variation.

**Abbreviations**

| | |
|---|---|
| CNV | Copy number variations |
| CMA | Chromosomal microarray |
| SV | Structural variants |
| Indels | Insertions and deletions |
| SRS | Short-read sequencing |
| aCGH | Array comparative genomic hybridization |
| LRS | Long read sequencing |
| SNP | Single nucleotide polymorphism |
| OGM | Optical genome mapping |
| ONT | Oxford Nanopore Technologies |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13039-025-00721-8.

Supplementary file 1.

Supplementary Figure 1: Quality analysis of nanopore sequencing reads generated using genomic DNA from cell lines EOL-1 and 697

Supplementary file 2.

Supplementary Figure 2: Representation of variant size distribution of 48 CNV gains and losses obtained using the CMA and ChAS software for both EOL-1 and 697 cell lines and validated by an experienced cytogeneticist

Supplementary file 3.

Supplementary Figure 3: Agarose gel of PCR amplification product sizes for breakpoints confirmation in 3 variants by Sanger sequencing

Supplementary file 4.

Supplementary Figure 4: Images of Ribbon plots for variants called as inversions by SV callers

Supplementary file 5.

Supplementary Figure 5: Depth of coverage bar chart of coverage analysis of CNV gains and losses not called in nanopore sequencing

Supplementary file 6.

Supplementary File 1: Bash commands

Supplementary file 7.

Supplementary File 2: R commands

Supplementary file 8.

Supplementary Table 1: Quality control metrics of the CMA analysis in cell lines EOL-1 and 697

Supplementary file 9.

Supplementary Table 2: Pre- and post-mapping sequencing statistics for cell lines EOL-1 and 697

Supplementary file 10.

Supplementary Table 3: Description of the CNV called in cell line EOL-1 using CMA and nanopore sequencing (CuteSV and Sniffles2)

Supplementary file 11.

Supplementary Table 4: Description of the CNV called in cell line 697 using CMA and nanopore sequencing (CuteSV and Sniffles2)

Supplementary file 12.

Supplementary Table 5: Breakpoint coordinates of 3 deletion variants and their respective sizes obtained by CMA, nanopore sequencing and Sanger sequencing

**Acknowledgements**
The authors are grateful to Daniel Sobral, PhD (Department of Infectious Diseases, National Institute of Health) for technical support with bioinformatics tools

**Author contributions**
CS and LV conceptualized the study; CS, BM, JF, AV and SP were responsible for data acquisition and formal analysis; LV, AR and HC were responsible for supervision; LV was responsible for manuscript preparation; CS and AR were responsible for manuscript editing and review. All authors read and approved the final manuscript.

**Funding**
This work is a result of the GenomePT project (POCI-01–0145-FEDER-022184), supported by COMPETE 2020—Operational Programme for Competitiveness and Internationalization (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by Fundação para a Ciência e a Tecnologia (FCT). This work was also supported by Fundos FEDER through the Programa Operacional Factores de Competitividade – COMPETE and by Fundos Nacionais through the Fundação para a Ciência e a Tecnologia within the scope of the project UID/BIM/00009/2019 (Centre for Toxicogenomics and Human Health -ToxOmics).

**Availability of data and materials**
Sequencing dataset presented in this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under the accession number PRJEB86107. CMA dataset presented in this study have been deposited in figshare repository at https://doi.org/10.6084/m9.figshare.28485737.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Technology and Innovation Unit, Human Genetics Department, National Institute of Health, Avenida Padre Cruz, 1649-016 Lisbon, Portugal. [2]Comprehensive Health Research Centre, NOVA Medical School, Universidade NOVA de Lisboa, Campo Dos Mártires da Pátria, 130, 1169-056 Lisbon, Portugal. [3]Cytogenetics Unit, Human Genetics Department, National Institute of Health, Avenida Padre Cruz, 1649-016 Lisbon, Portugal. [4]Research and Development Unit, Human Genetics Department, National Institute of Health, Avenida Padre Cruz, 1649-016 Lisbon, Portugal.

Received: 11 April 2025 Accepted: 10 July 2025
Published online: 23 July 2025

**References**

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444–54.
2. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61(1):437–55.
3. Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. Charcot–Marie–Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet. 1992;2(4):292–300.
4. Greenberg F, Guzzetta V, Montes de Oca-Luna R, Magenis RE, Smith AC, Richter SF, et al. Molecular analysis of the Smith-Magenis syndrome: a possible contiguous-gene syndrome associated with del(17)(p11.2). Am J Hum Genet. 1991;49(6):1207–18.
5. Alitalo K, Schwab M, Lin CC, Varmus HE, Bishop JM. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. Proc Natl Acad Sci. 1983;80(6):1707–11.
6. Favera RD, Wong-Staal F, Gallo RC. Onc gene amplification in promyelocytic leukaemia cell line HL-60 and primary leukaemic cells of the same patient. Nature. 1982;299(5878):61–3.
7. Little CD, Nau MM, Carney DN, Gazdar AF, Minna JD. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. Nature. 1983;306(5939):194–6.
8. Courjal F, Cuny M, Simony-Lafontaine J, Louason G, Speiser P, Zeillinger R, et al. Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups. Cancer Res. 1997;57(19):4360–7.
9. Worst TS, Weis CA, Stöhr R, Bertz S, Eckstein M, Otto W, et al. CDKN2A as transcriptomic marker for muscle-invasive bladder cancer risk stratification and therapy decision-making. Sci Rep. 2018;8(1):14383.
10. Bui NQ, Przybyl J, Trabucco SE, Frampton G, Hastie T, Van De Rijn M, et al. A clinico-genomic analysis of soft tissue sarcoma patients reveals *CDKN2A* deletion as a biomarker for poor prognosis. Clin Sarcoma Res. 2019;9(1): 12.
11. Guo Y, Long J, Lei S. Promoter methylation as biomarkers for diagnosis of melanoma: a systematic review and meta-analysis. J Cell Physiol. 2019;234(5):7356–67.
12. Xia L, Zhang W, Gao L. Clinical and prognostic effects of CDKN2A, CDKN2B and CDH13 promoter methylation in ovarian cancer: a study using meta-analysis and TCGA data. Biomarkers. 2019;24(7):700–11.
13. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019;20(1): 175.
14. Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, et al. Segmental copy number variation shapes tissue transcriptomes. Nat Genet. 2009;41(4):424–9.
15. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949–51.
16. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525–8.
17. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83.
18. Pös O, Radvanszky J, Buglyó G, Pös Z, Rusnakova D, Nagy B, et al. DNA copy number variation: main characteristics, evolutionary significance, and pathological aspects. Biomed J. 2021;44(5):548–59.
19. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363–76.

20. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125–38.
21. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749–64.
22. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20(1): 117.
23. Kosugi S, Kamatani Y, Harada K, Tomizuka K, Momozawa Y, Morisaki T, et al. Detection of trait-associated structural variations using short-read sequencing. Cell Genomics. 2023;3(6): 100328.
24. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. Genet Med. 2019;21(7):1603–10.
25. Malekpour SA, Pezeshk H, Sadeghi M. MSeq-CNV: accurate detection of copy number variation from sequencing of multiple samples. Sci Rep. 2018;8(1):4009.
26. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551–64.
27. Ip CLC, Loose M, Tyson JR, De Cesare M, Brown BL, Jain M, et al. MinION analysis and reference consortium: phase 1 data release and analysis. F1000Research. 2015;4:1075.
28. Lu H, Giordano F, Ning Z. Oxford nanopore minion sequencing and genome assembly. Genomics Proteomics Bioinformatics. 2016;14(5):265–79.
29. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;35(13):2193–8.
30. Romagnoli S, Bartalucci N, Vannucchi AM. Resolving complex structural variants via nanopore sequencing. Front Genet. 2023;16(14):1213917.
31. Jeffet J, Margalit S, Michaeli Y, Ebenstein Y. Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale. Essays Biochem. 2021;65(1):51–66.
32. Mayumi M. EoL-1, a human eosinophilic cell line. Leuk Lymphoma. 1992;7(3):243–50.
33. Findley HW, Cooper MD, Kim TH, Alvarado C, Ragab AH. Two new acute lymphoblastic leukemia cell lines with early B-cell phenotypes. Blood. 1982;60(6):1305–9.
34. Silva C, Machado M, Ferrão J, Sebastião Rodrigues A, Vieira L. Whole human genome 5'-mC methylation analysis using long read nanopore sequencing. Epigenetics. 2022;17(13):1961–75.
35. De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9.
36. Ren J, Chaisson MJ. lra: a long read aligner for sequences and contigs. PLOS Comput Biol. 2021;17(6):e1009078.
37. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292–4.
38. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21(1): 189.
39. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42(10):1571–80.
40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
41. Nattestad M, Aboukhalil R, Chin CS, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. Bioinformatics. 2021;37(3):413–5.
42. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34(5):867–8.
43. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
44. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.2 [Internet]. 2023. Available from: https://dplyr.tidyverse.org/
45. Wickham H, Pedersen T, Seidel D scales: Scale Functions for Visualization. R package version 1.3.0 [Internet]. 2023. Available from: https://scales.r-lib.org
46. FC M, Davis T, ggplot2 authors. ggpattern: 'ggplot2' Pattern Geoms. R package version 1.1.4–0 [Internet]. 2025. Available from: https://github.com/trevorld/ggpattern
47. Wickham H. ggplot2: elegant graphics for data analysis [Internet]. Second edition. Cham: Springer international publishing; 2016. 1 p. (Use R!). Available from: https://ggplot2.tidyverse.org
48. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8(1):14061.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
50. Hollox EJ, Zuccherato LW, Tucci S. Genome structural variation in human evolution. Trends Genet. 2022;38(1):45–58.
51. Hu L, Yao X, Huang H, Guo Z, Cheng X, Xu Y, et al. Clinical significance of germline copy number variation in susceptibility of human diseases. J Genet Genomics. 2018;45(1):3–12.
52. Auwerx C, Jõeloo M, Sadler MC, Tesio N, Ojavee S, Clark CJ, et al. Rare copy-number variants as modulators of common disease susceptibility. Genome Med. 2024;16(1):5.
53. Yuan B, Wang L, Liu P, Shaw C, Dai H, Cooper L, et al. CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. Genet Med. 2020;22(10):1633–41.
54. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of copy-number variants on complex human traits. Am J Hum Genet. 2022;109(4):647–68.
55. Lacaria M, Saha P, Potocki L, Bi W, Yan J, Girirajan S, Gu W. A duplication CNV that conveys traits reciprocal to metabolic syndrome and protects against diet-induced obesity in mice and men. PLoS Genet. 2012;8(5):e1002713.
56. Poole AC, Goodrich JK, Youngblut ND, Luque GG, Ruaud A, Sutter JL, et al. Human salivary amylase gene copy number impacts oral and gut microbiomes. Cell Host Microbe. 2019;25(4):553-564.e7.
57. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022;606(7916):984–91.
58. Xu C, Li M, Gu T, Xie F, Zhang Y, Wang D, et al. Chromosomal microarray analysis for prenatal diagnosis of uniparental disomy: a retrospective study. Mol Cytogenet. 2024;17(1):3.
59. Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, et al. A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. Cancers. 2021;13(24):6283.
60. Yuan N, Jia P. Comprehensive assessment of long-read sequencing platforms and calling algorithms for detection of copy number variation. Brief Bioinform. 2024;25(5):bbae441.
61. Haraksingh RR, Abyzov A, Urban AE. Comprehensive performance comparison of high-resolution array platforms for genome-wide copy number variation (CNV) analysis in humans. BMC Genomics. 2017;18(1):321.
62. Coutelier M, Holtgrewe M, Jäger M, Flöttman R, Mensah MA, Spielmann M, et al. Combining callers improves the detection of copy number variants from whole-genome sequencing. Eur J Hum Genet. 2022;30(2):178–86.
63. Joe S, Park JL, Kim J, Kim S, Park JH, Yeo MK, et al. Comparison of structural variant callers for massive whole-genome sequence data. BMC Genomics. 2024;25(1):318.
64. Helal AA, Saad BT, Saad MT, Mosaad GS, Aboshanab KM. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. Sci Rep. 2024;14(1):6160.
65. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32(3):246–51.
66. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines [Internet]. Bioinformatics; 2015


## Publisher's Note