Molecular Therapy
Nucleic Acids
Original Article

ifCNV: A novel isolation-forest-based package to detect copy-number variations from various targeted NGS datasets

Simon Cabello-Aguilar,[1] Julie A. Vendrell,[1] Charles Van Goethem,[2] Mehdi Brousse,[3] Catherine Gozé,[1] Laurent Frantz,[4,5] and Jérôme Solassol[1]

[1]Laboratoire de Biologie des Tumeurs Solides, Département de Pathologie et Oncobiologie, CHU Montpellier, Université de Montpellier, 34295 Montpellier, France; [2]Laboratoire de Génétique Moléculaire, CHU Montpellier, Université de Montpellier, 34295 Montpellier, France; [3]Laboratoire d'Hématologie Biologique, CHU Montpellier, Université de Montpellier, 34295 Montpellier, France; [4]Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University of Munich, Munich, Germany; [5]School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Copy-number variations (CNVs) are an essential component of genetic variation distributed across large parts of the human genome. CNV detection from next-generation sequencing data and artificial intelligence algorithms have progressed in recent years. However, only a few tools have taken advantage of machine-learning algorithms for CNV detection, and none propose using artificial intelligence to automatically detect probable CNV-positive samples. The most developed approach is to use a reference or normal dataset to compare with the samples of interest, and it is well known that selecting appropriate normal samples represents a challenging task that dramatically influences the precision of results in all CNV-detecting tools. With careful consideration of these issues, we propose here ifCNV, a new software based on isolation forests that creates its own reference, available in R and python with customizable parameters. ifCNV combines artificial intelligence using two isolation forests and a comprehensive scoring method to faithfully detect CNVs among various samples. It was validated using targeted next-generation sequencing (NGS) datasets from diverse origins (capture and amplicon, germline and somatic), and it exhibits high sensitivity, specificity, and accuracy. ifCNV is a publicly available open-source software (https://github.com/SimCab-CHU/ifCNV) that allows the detection of CNVs in many clinical situations.

INTRODUCTION

Copy-number variations (CNVs) are a class of structural variations that result from the deletion or duplication of a DNA fragment. About 1,500 CNV regions have already been discovered in humans, accounting for ~12%–16% of the entire human genome,[1] making it one of most common types of genetic variation. Although the biological impact of the majority of these CNVs remains uncertain, nearly 50% of known CNVs overlap with protein-coding regions, and many are involved in genetic diseases. Recent studies have demonstrated that CNVs can be implicated in many rare diseases, such as inherited retinal dystrophies,[2] and in diseases that involve dosage-sensitive developmental genes, such as Charcot-Marie-Tooth disease[3] and DiGeorge syndrome, among others.[4–6] CNVs, resulting from gene amplification (copy-number gain) as well as gene deletion (copy-number loss), are common in cancer cells, and multiple studies have shown that duplication or deletion of specific genes can contribute to tumor growth[7] and to resistance to anti-tumor therapies.[8,9] In cancer cells, the size of these molecular alterations can vary dramatically, from one or a few exons to an entire chromosomal arm.

Although most CNVs found in cancer cells are likely to have accumulated as a direct consequence of clonal evolution during the disease course, some have been identified as playing a role in the early development of cancer (e.g. CNVs located in *BRCA1/2* in familial breast and ovarian cancer[10]). In fact, it has been estimated that CNVs represent more than 10% of the molecular alterations linked to cancer predisposition, making their detection a priority. Detection of acquired (somatic) focal copy-number changes is also required for diagnosis, prognosis, and the therapeutic management of patients with cancer.[11] For example, loss of chromosomal arms 1p and 19q is closely associated with oligodendrogliomas, a subtype of primary brain tumors, and with a favorable prognosis in diffuse gliomas.[12] Focal copy-number increases are biomarkers predictive of responses to particular therapies; for example, patients with oncogenic *ERBB2* amplification in breast cancer respond well to trastuzumab, and acquired resistance to tyrosine kinase inhibitors is exhibited in patients with *MET*-amplified non-small cell lung carcinomas.[13,14]

Recently, the rapid implementation of high-throughput next-generation sequencing (NGS) methods, especially targeted DNA panels, in clinical laboratories has led to the emergence of a fairly large number of pipelines and algorithms able to detect CNVs from NGS data.[15–27] Most of these studies use the read-depth approach, relying on the

Received 1 March 2022; accepted 15 September 2022;
https://doi.org/10.1016/j.omtn.2022.09.009.

Corresponding author
E-mail: s-cabelloaguilar@chu-montpellier.fr

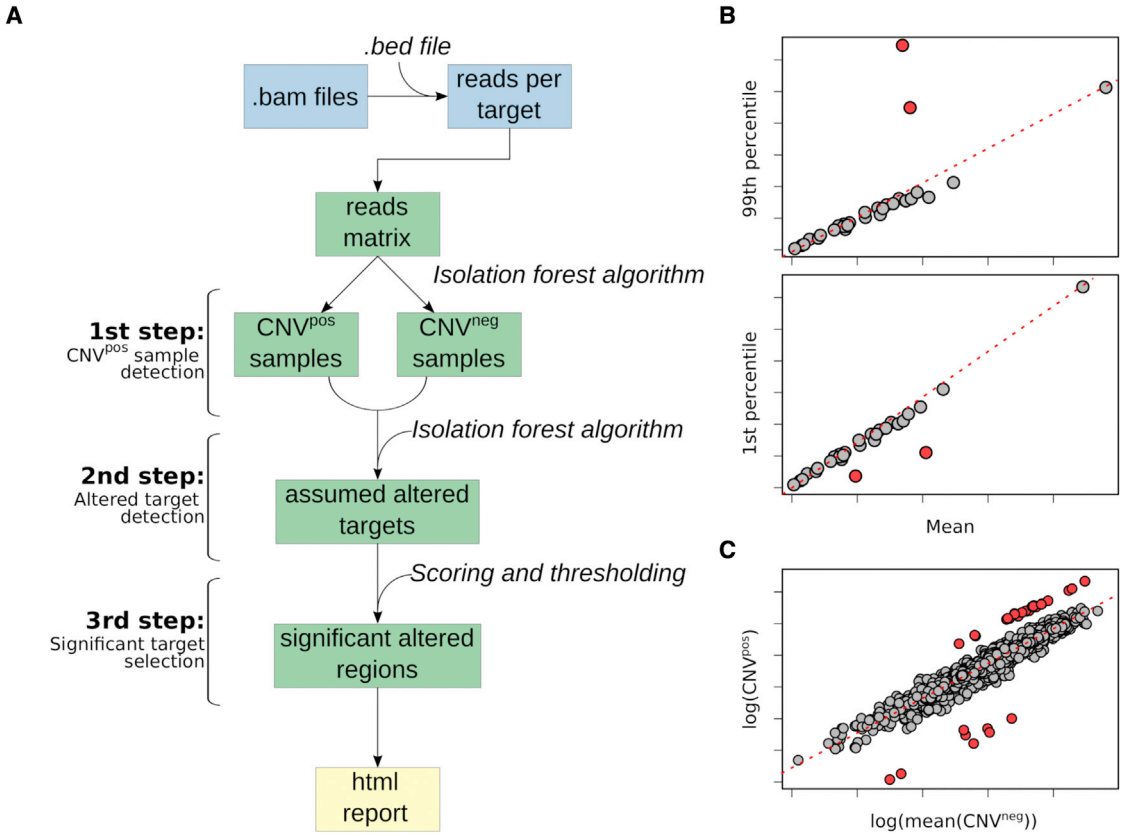

Figure 1. ifCNV workflow

(A) ifCNV is composed of three steps: the pre-processing (blue), the core algorithm (green), and the output (yellow). CNV, copy-number variation; CNV^{neg}, CNV-negative samples; CNV^{pos}, CNV-positive samples. (B) Top: 99th percentiles of the reads distribution according to the means of the reads distribution of the samples in a NGS sequencing run; bottom: 1st percentiles of the reads distribution according to the means of the reads distribution of the samples in a NGS sequencing run. The red dots correspond to the outlying samples. (C) In one CNV^{pos} sample, the logarithm of the reads per target according to the logarithm of the mean normalized normal sample. The red dots correspond to the outlying targets.

hypothesis that the number of reads aligned to a genomic region is proportional to the copy number of the region. In multiple sample methods, CNVs are detected by comparing the read counts of the sample of interest to the read counts of a reference sample. The proper building of the reference is one of the main difficulties. To that end, two main solutions exist: (1) to gather a database of normal samples or (2) to add normal samples into the NGS run. Nevertheless, both solutions come with issues, mainly the presence of a batch effect and a high cost, respectively. To avoid these problems, use of the single sample method was previously proposed, which consists of statistical modelling of the target read counts within the sample of interest to detect CNVs. Recent advances in artificial intelligence and, in particular, the availability of accessible machine-learning packages[28] have made it possible for developers to improve their algorithms in many areas. To date, only a few studies have taken advantage of these recent developments in the field of CNV detection from targeted NGS data.[15,16,26]

With careful consideration of these issues, we present ifCNV, a novel machine-learning-based software, provided as a python and R package (https://github.com/SimCab-CHU/ifCNV and https://github.com/SimCab-CHU/ifCNV-R, respectively). This approach combines several advantages, among which is that it allows detection of CNVs without the need for a reference sample and it is low resource consuming.

To validate our model and explore its limitations in clinical practice, we tested ifCNV on different synthetic datasets mimicking relevant clinical situations and on datasets obtained from amplicon- or capture-based DNA library preparation technologies.

RESULTS

ifCNV workflow

ifCNV is a CNV detection tool based on read-depth distribution obtained from NGS data (Figure 1A). It integrates a pre-processing step to create a read-depth matrix using as input the aligned binary alignment map (.bam) files and a corresponding .bed file. This reads matrix is composed of the samples as columns and the targets as rows. Next, it uses an Isolation Forest (IF) machine-learning algorithm to

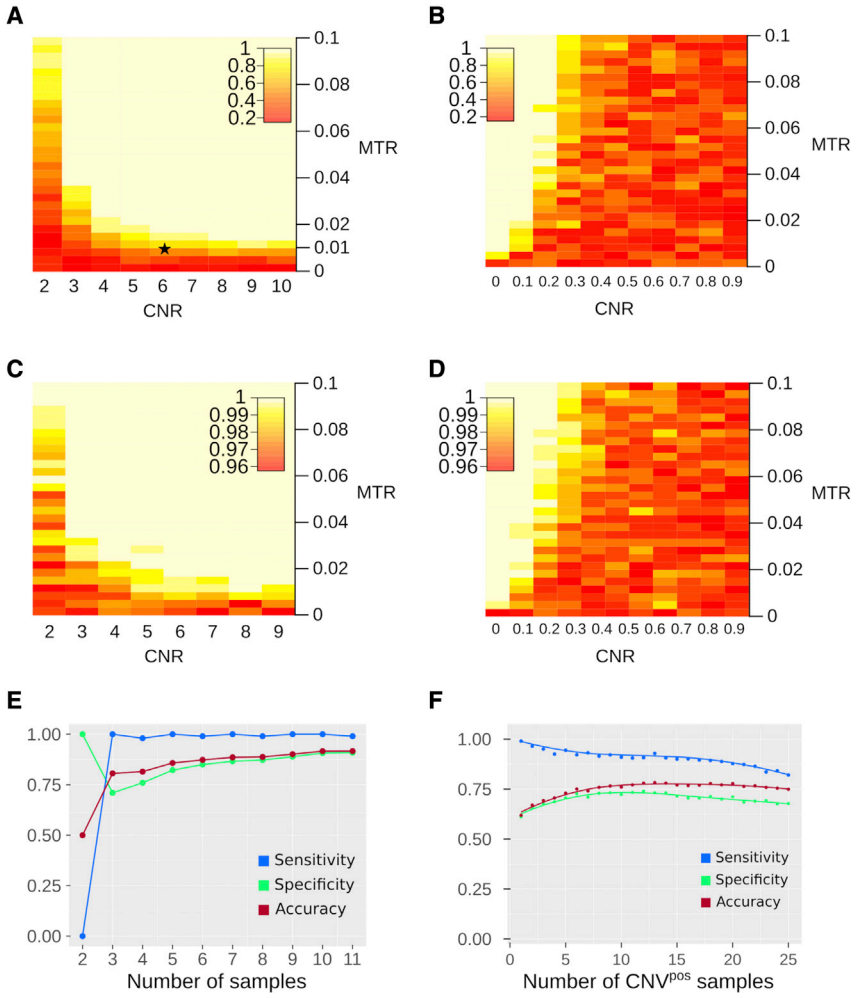

Figure 2. Performance assessment in detecting CNV^{pos} and CNV^{neg} samples

(A) Heatmap of the detection rate of CNV^{pos} samples as a function of the CNR and the MTR, for CNRs between 2 and 10. (B) Heatmap of the detection rate of CNV^{pos} samples as a function of the CNR and the MTR, for CNRs between 0 and 0.9. (C) Heatmap of the detection rate of CNV^{neg} samples as a function of the CNR and the MTR, for CNRs between 2 and 9. (D) Heatmap of the detection rate of CNV^{neg} samples as a function of the CNR and the MTR, for CNRs between 0 and 0.9. (E) Classification indicators of the detection of a single CNV^{pos} sample in a set of several CNV^{neg} samples. (F) Classification indicators of the detection of multiple CNV^{pos} samples in a set of several CNV^{neg} samples. CNR, copy-number ratio; MTR, modified target ratio.

detect the samples with a strong bias between the 99th percentile and the mean (for amplifications, Figure 1B, top plot) and the 1st percentile and the mean (for deletions, Figure 1B, bottom plot). These samples are assumed to be CNV^{pos}. The samples with no bias, which are therefore not detected by the IF as outliers, are considered CNV^{neg} samples. After filtering of the samples with a mean read depth per target less than X (X = 10 by default but can be set by the user to any value), the reads matrix is normalized by dividing each column (i.e., the reads distribution of each sample) by its median. Then, ifCNV creates a mean normalized normal sample by averaging all CNV^{neg} samples, to create the intra-run reference. The log ratio between each CNV^{pos} sample and this reference is computed, and a second IF is used to detect the outlying targets (Figure 1C). The log ratio balances the differences between ratios under 1 (deletions) and ratios over 1 (amplifications), increasing the ability of ifCNV to detect outlying targets with a ratio under 1 (data not shown).

These assumed altered targets are then used to compute the localization score per region of interest (see materials and methods). Finally, a threshold is applied on this score to select the significantly altered regions that are compiled in an html report containing a table and a graph for easy user interpretation.

Performance of ifCNV

Detection of CNV^{pos} samples

To quantify the ability of ifCNV to detect CNV^{pos} samples, we created a synthetic dataset of 1,500 targets and 30 samples in which we inserted one CNV^{pos} sample. It is noteworthy that if the copy-number ratio (CNR) or the modified target ratio (MTR; i.e., the number of altered targets, located on CNV regions, divided by the total number of targets in the panel) are low, the CNV^{pos} samples will resemble the CNV^{neg} samples and therefore will be difficult to detect. Thus, the performance of a CNV detector directly depends on the CNR and the MTR. Taking this fact into consideration, we iterated the CNR and MTR (from 0 to 10 and 0 to 0.1, respectively) and performed 1,000 simulations for each iteration (Figures 2A and 2B). The analysis of the attribution of CNV^{pos} samples for a CNR greater than 1 is shown in Figure 2A. For CNRs greater than 6, ifCNV correctly identified the abnormal sample in 99.58% of simulations when the MTR is greater than 0.01 (Figure 2A). Furthermore, if the CNR was between 4 and 6 and the MTR was over 0.01, ifCNV found the abnormal sample in 99.47% of simulations. Finally, if the CNR was over 2 and the MTR over 0.01, ifCNV detected the abnormal samples in 99.34% of simulations; this reached 99.83% when the MTR was greater than 0.035.

ifCNV was also able to detect deletion (CNRs under 0); as for CNRs greater than 1, the sensitivity was related to both the CNR and the MTR (Figure 2B). For CNRs under 0.5, ifCNV detected the abnormal sample in 92.26% of simulations. For CNRs over 0.5, ifCNV only detected the abnormal sample in 27% of simulations. Although 27% is a higher detection rate than a random choice, for which the probability

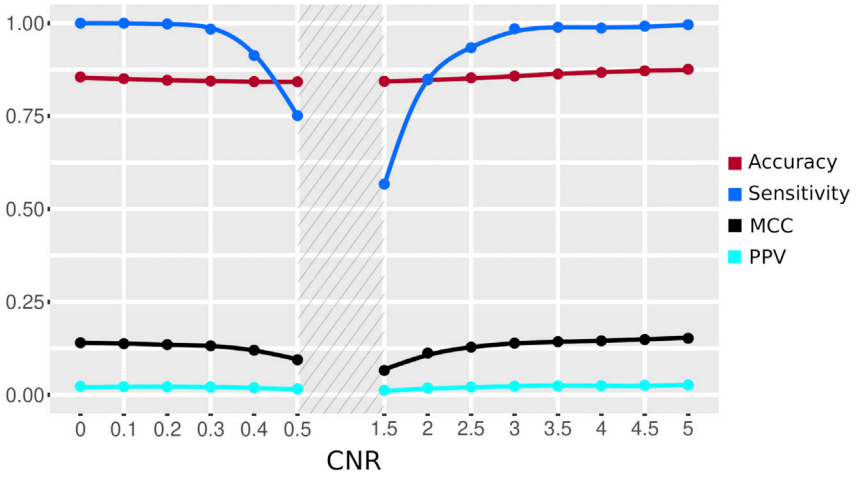

Figure 3. Classification indicators for the detection of altered targets

MCC, Matthews correlation coefficient; PPV, positive predictive value.

is 3.33% (1 sample out of 30), it is not satisfactory enough for diagnostic purposes given the importance of heterozygous deletion (hDel) in several diseases.[6] To overcome this pitfall, we assessed an alternative solution based on the ability of the IF to accurately detect the CNV^{neg} samples. Indeed, samples labeled as CNV^{neg} were considered to be normal and were adopted as an internal reference.

Identification of the CNV^{neg} samples

To quantify the ability of ifCNV to detect CNV^{neg} samples, we used the same synthetic dataset, and we also iterated the CNR and the MTR (from 0 to 10 and 0 to 0.1, respectively) and performed 1,000 simulations for each iteration (Figures 2C and 2D). ifCNV was able to correctly identify the CNV^{neg} samples in 99.87% of simulations, regardless of the CNR and MTR. For CNRs over 2, this reached 99.9%. Interestingly, for CNRs under 1, ifCNV identified the CNV^{neg} samples in 98.57% of the simulations. Finally, a value of 99.69% was obtained for CNRs under 0.5.

Detection of one CNV^{pos} sample in a set of several CNV^{neg} samples

Conventional hospital and research laboratories must determine the CNV status of numerous samples. The number of samples in a sequencing run can vary from a few to several dozen. We therefore assessed the ability of ifCNV to correctly find a unique CNV^{pos} sample in a set of several negative samples. We created synthetic datasets of 2 to 100 samples in which we inserted a CNV^{pos} sample harboring an amplification with a CNR of 5 and an MTR of 0.03. We performed 100 simulations for each and calculated the sensitivities (Se), specificities (Sp), and accuracies (Acc) of the algorithm (Figure 2E). For one CNV^{pos} sample in a set of two, ifCNV failed to label any sample as positive, leading to one true negative (TN) and one false negative (FN) (Se = 0, Sp = 1, and Acc = 0.5). Interestingly, for one CNV^{pos} sample in three, ifCNV correctly labeled the positive sample in every simulation and found one FP in less than 50% of simulations. When increasing the number of samples in the set, the Se remained close to 1 (0.992 from 3 to 100 samples), and the Sp and the Acc tended to 0.9.

Detection of multiple CNV^{pos} samples in a set of several CNV^{neg} samples

As several CNV^{pos} can be present in the same sequencing run, we tested the performance of ifCNV in such situations. We randomly chose 2 to 25 samples in a synthetic dataset of 50 samples. We then added random CNRs (from 2 to 6) to 3% of the targets of these samples and performed 100 iterations to determine the Se, Sp, and Acc of our algorithm (Figure 2F). ifCNV exhibited relatively high Se, Sp, and Acc (around 0.85, 0.7, and 0.75, respectively), regardless of the number of CNV^{pos} samples in the dataset. Notably, when half of the tested set (25/50) was CNV^{pos}, ifCNV correctly labeled a mean of 20 samples across all simulations.

Detection of altered targets

The second step of ifCNV consists of labeling the targets that are modified among the CNV^{pos} samples. To assess its performance, we created a synthetic dataset of 30 samples and 300 targets. One sample was CNV^{pos} with one modified target randomly chosen at each iteration, with a CNR from 0 to 5. We then performed 1,000 iterations and calculated the Se, Acc, positive predictive value (PPV), and Matthews correlation coefficient (MCC) (Figure 3). On the one hand, ifCNV exhibited a Se very close to 1 for CNRs from both 0 to 0.3 (deletions) and 3 to 5 (amplification), meaning that the modified target was accurately labeled in almost every simulation. On the other hand, the PPV and the MCC were low ($\sim$0.02 and $\sim$0.1, respectively), reflecting a high number of FPs. However, the Acc was $\sim$0.87 and stayed approximately unchanged, meaning that the number of TNs was high and dominated the number of FPs.

Thresholding on the localization score

To test the ability of the score to discriminate the FPs from the TPs, we used the same synthetic dataset as before and grouped targets (from 2 to 15) together to mimic regions. We then modified all the targets corresponding to a randomly chosen region. Finally, we computed the localization score and calculated the Se, Acc, PPV, and MCC before (Figure 4A, left panel) and after thresholding (Figure 4A, right panel).

As Figure 3 shows, before thresholding, the Se was close to 1, and the PPV and MCC were low (around 0.1 and 0.2, respectively). Acc was lower using the grouped targets (around 0.4) because, by construction, the total number of regions in the dataset is lower than the number of targets, and therefore the number of TN is smaller. After thresholding, we observed that the PPV, MCC, and Acc increased to reach values very close to 1 for CNRs over 2, while

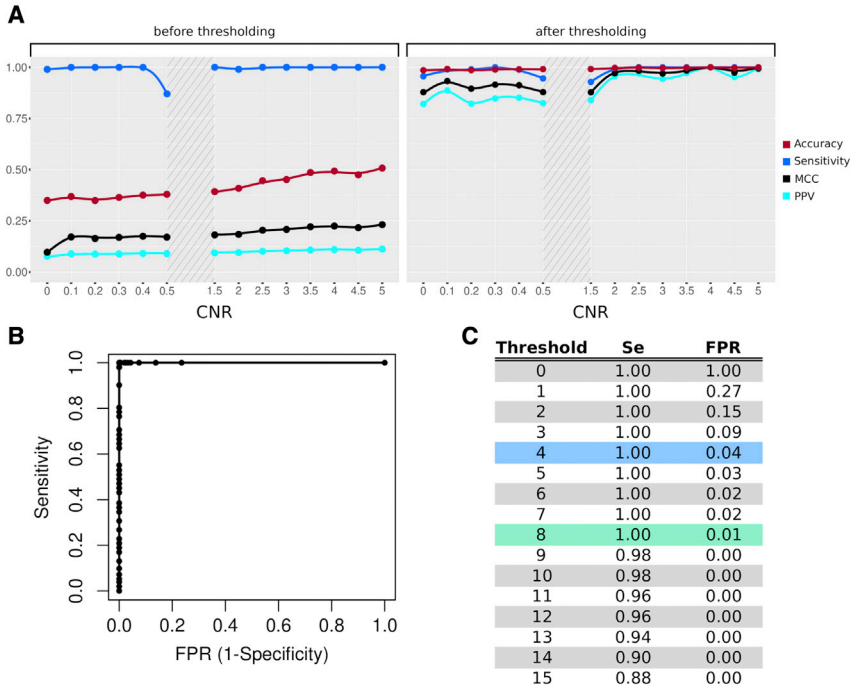

Figure 4. Performance assessment of the localization score thresholding

(A) Classification indicators for the detection of altered targets before (left panel) and after (right panel) thresholding. (B) ROC curve. (C) Associated table. FPR, false positive rate; Se, sensitivity; ROC, receiver operating characteristic.

the Se had only slightly decreased, meaning that the score thresholding enables the TPs to be kept while the FPs are discarded. The localization score thresholding approach can therefore be validated and represents an important improvement in the performance of our algorithm.

To characterize the dependence on the score, we iterated the threshold from 0 to 15, calculated the corresponding TP rate (TPR) and FP rate (FPR), and plotted a receiver operating characteristic (ROC) curve (Figure 4B). The corresponding table in Figure 4C shows that, on the simulated data, CNVs with a score over 4 were 100% TPs and that less than 5% FPs and CNVs with a score over 8 were 100% TPs and ∼1% FPs.

Evaluation of ifCNV on patient datasets

We then tested ifCNV on three real datasets. First, we used the ICR96 dataset and then used two in-house datasets for which we had NGS and array comparative genomic hybridization (aCGH) results to compare to. Tables 1 and 2 present the results.

ICR96 dataset

To compare the performance of ifCNV wtih that of other methods,[17,23–25,29] we compared our results with those obtained by Moreno-Cabrera et al.,[30] who benchmarked five widely used tools with the ICR96 dataset.[31] This dataset, however, only possess one target per exon, rendering the advantage of the ifCNV score thresholding strategy not practicable in this case. As an alternative strategy, we decided to take advantage of the ability to change the contamination parameter of the IF. This main actionable parameter is the proportion of outliers in the dataset. By default, it can be set to a value between 0 and 0.5 and to "auto," for an automatic detection of the proportion of outliers. In ifCNV, we added the ability to set this parameter to "none"; it is then calculated as 1 on the number of samples in the dataset. We thus iterated several values of the contamination parameter and calculated the correspondent binary classification indicators (Figure 5). We also compiled the results obtained with the two pre-set contamination values (auto and none) with the results from the other tools (Table 1). We observed that ifCNV exhibits performances in the same order of magnitude as the other tools, with the clear advantage of having an easily tunable parameter allowing the user to expect either no FNs (contamination = auto) or no FPs (contamination = none).

TSCA dataset

Next, we aimed to validate the performance of ifCNV on an in-house dataset (Table 2). Its particularities are (1) it is composed of tumor samples containing variable percentages of altered cells, and (2) it possesses a small number of targets per region (range: 1–40, median: 4). Using this dataset, we found that ifCNV correctly labeled 19 of the 21 amplifications present in the dataset; the 2 FNs were measured as a gain of two copies (CNR = 2) by aCGH. In addition, 14 of 17 deletions were detected with no FPs. The 3 undetected deletions were from samples that have a lower percentage of tumor cells.

Juno dataset

Finally, we also assessed our tool on a distinct library preparation approach with a larger number of targets per region composed of a larger panel than the TSCA dataset (range: 1–164, median: 17) from tumor samples (Table 2). Interestingly, we could detect all amplifications with no FPs and 17 out of the 19 deletions, leading to an overall Acc of 0.96 and MCC of 0.93. Interestingly, the 2 missed deletions are on a gene that represents only 0.4% of the panel (MTR = 0.004).

DISCUSSION

In recent years, numerous computational methods for detecting and measuring CNVs from NGS data have been developed. However, most of these are based on the use of internal or external reference samples. To date, only a few have taken advantage of easy-to-use machine-learning packages.[15,16,26]

Table 1. Description of the datasets used in the study

Datasets	Samples	Positives	Negatives	Number of alterations	Amplifications	Deletions
ICR96	96 (germline)	68	28	296	80	216
TSCA	81 (somatic)	25	56	39	21	17
Juno	43 (somatic)	26	17	28	9	19

Several artificial intelligence (AI)-based outlier detection methods exist. The main ones are the minimum covariance determinant[32] (MCD), the local outlier factor[33] (LOF), the IF,[34] and the elliptic envelope algorithm[35] (EEA). Each method has its benefits and drawbacks. Briefly, MCD and EEA were created to treat input variables with Gaussian distribution, LOF was designed for data with low dimensionality, and IF is a tree-based algorithm effective on high-dimensional data and no underlying assumption on the distribution of the data. The read-depth data obtained from targeted NGS do not follow a Gaussian distribution and can be of high dimensionality depending on the number of features of the panel and the sequencing run. Thus, the IF algorithm appears to be the most suitable to process this data. Therefore, we describe here ifCNV, a bioinformatics tool using the IF algorithm, that allows detection of CNVs without the need for a reference sample.

Moreover, in routine clinical practice, the variety of pathologies involving specific molecular alterations leads to a broad diversity in the datasets generated. Thus, in general, most CNV software that is developed for a specific data type has suboptimal reliability for use in routine practice with diversified samples. In addition, most of the genetic workflows are either developed in Python or R, and, to our knowledge, no existing CNV detection tool is available in both languages.

3ifCNV is available in both languages, making it more easily to implement in pre-existing pipelines. Also, by successfully creating its own normal reference inside each analyzed NGS run, ifCNV frees itself from any batch effect inherent to tools using external references. It also avoids the need for reference samples that are copy-number neutral to be sequenced in the same batch, which is an efficient but not cost-effective solution. ifCNV also has a simple framework: it is made up of only three main steps on which the user has full control through tunable parameters. Furthermore, its efficiency makes it possible to run on hardware with limited computing resources.

Using simulated data, we demonstrate that ifCNV is highly reliable and adapts to several relevant clinical situations including (1) when one CNV^{pos} sample is present in a set of several CNV^{neg} samples, (2) when multiple CNV^{pos} samples are present in a set of several CNV^{neg} samples, (3) when only one target is altered, and (4) when the CNRs are close to one, which can correspond to small alterations or to mixtures of normal and altered cells. ifCNV also performed well using datasets generated from amplicon- or capture-based libraries prepared from germline or somatic clinical samples.

Analyses of real data also demonstrated that ifCNV's performance was comparable to that of other widely used tools[30] but with substantial specific advantages. Our solution has a tuneable control of the FPR thanks to localization score thresholding and to the contamination parameter, which can both be optimized according to the dataset by an entry-level user. ifCNV was also able to accurately detect CNVs in difficult samples, such as those composed of a mixture of normal cells and tumoral cells, which dilutes the CNR of samples.

Even if we demonstrated that ifCNV can process various targeted dataset, as is, it cannot be applied to whole-genome sequencing (WGS) and third-generation sequencing (TGS) datasets. The concept of the method should be pertinent to treat these data types but would need further development. Indeed, pre-processing and both IF parameters will need to be adapted and benchmarked. This would be the subject of a new study.

In conclusion, ifCNV is a highly flexible tool that can detect CNVs in germline and somatic clinical samples with similar performances.

Table 2. Classification indicators for ifCNV and the tools described in Morena-Cabrera et al. on the ICR96 dataset

Tool	TP	TN	FP	FN	Total	FNR	FPR	Se	Sp	PPV	Acc	MCC
ifCNV – auto	296	27,017	1,858	0	28,875	0	0.065	1	0.9350	0.1374	0.9363	0.3586
ifCNV – none	252	28,875	0	44	28,875	0.1486	0	0.8513	1	1	0.9984	0.9219
DECoN	286	28,473	106	10	28,875	0.0338	0.0037	0.9662	0.9963	0.7296	0.9959	0.8377
panelcn.MOPS	284	28,236	343	12	28,875	0.0405	0.012	0.9595	0.988	0.453	0.9877	0.6547
CoNVaDING	283	28,068	511	13	28,875	0.0439	0.0179	0.9561	0.9821	0.3564	0.9818	0.5778
exomedepth	283	28,507	72	13	28,875	0.0439	0.0025	0.9561	0.9975	0.7972	0.9970	0.8716
CODEX2	275	28,503	76	21	28,875	0.0709	0.0027	0.9291	0.9973	0.7835	0.9966	0.8515

Acc, accuracy; FN, false negative; FP, false positive; Sp, specificity; TN, true negative; TP, true positive.

Table 3. Classification indicators on the TSCA and Juno datasets

Dataset	CNV type	TP	TN	FP	FN	FNR	FPR	Se	Sp	PPV	Acc	MCC
TSCA	amplification	19	59	1	2	0.10	0.02	0.90	0.98	0.95	0.96	0.90
	deletion	14	64	0	3	0.21	0	0.82	1	1	0.96	0.89
	total	33	123	1	5	0.15	0.01	0.87	0.99	0.97	0.96	0.90
Juno	amplification	9	19	0	0	0	0	1	1	1	1	1
	deletion	17	9	0	2	0.12	0	0.89	1	1	0.93	0.86
	total	26	28	0	2	0.08	0	0.93	1	1	0.96	0.93

ifCNV now represents an essential component of the cancer diagnosis pipeline that we routinely use to analyze samples from patients in our laboratory. We believe that the flexibility, high accuracy, easy implementation, and low hardware infrastructure afforded by our method will help other laboratories in increasing their throughput and improve disease characterization by accurate CNV detection.

MATERIALS AND METHODS

IF algorithm

The IF algorithm was developed by Liu et al.[34] It "isolates" observations using a binary tree structure called an isolation tree. In this isolation tree, anomalies are more likely to be isolated closer to the root, whereas normal points are more likely to be isolated at the deeper end. The IF algorithm builds its isolation trees for a given dataset by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length is then averaged over a forest of such isolation trees to produce a decision value. The smaller the value, the more likely it is that the sample represents an anomaly.

Synthetic datasets

To create synthetic datasets that reproduce faithfully those obtained in routine clinical practice, we selected 1,910 samples from in-house targeted NGS data with no CNVs. We extracted the total reads on each target from the aligned .bam files with the bedtools[36] (https://bedtools.readthedocs.io) multicov function and created a reference reads matrix ordered with samples as columns and targets as rows. This reference reads matrix was then normalized by dividing each column by its median. All medians were used to create a median reads distribution that was needed for the rescaling process. Next, we created a normalized target reads distribution from each row of the normalized matrix. Thus, to generate synthetic datasets, we filled each line by taking a normalized target reads distribution, in which we randomly picked a value for each column. To rescale this matrix, we multiplied each column with a value randomly picked from the median distribution. Finally, to create CNV^{pos} samples within this synthetic dataset and test the algorithm, we modified the desired number of targets by multiplying it by a factor ranging from 0 to 10.

ICR96 dataset

The dataset ICR96 exon CNV validation series[31] was downloaded from the European Genome-phenome Archive (EGA) (EGA: EGAD00001003335). This dataset consists of the read counts of 96 germline samples sequenced on a targeted panel for which the copy number, at the exon level, was validated using high-resolution multiplex ligation-dependent probe amplification (MLPA) experiments (Table 3).

In-house datasets

DNA extracted from clinical somatic samples was analyzed alongside by two molecular approaches: aCGH as a reference method, and NGS using two different library preparation protocols.

DNA extraction of formalin-fixed paraffin-embedded samples

DNA was extracted from tissue samples using the Maxwell RSC DNA FFPE Kit (Promega, Madison, WI, USA) according to the manufacturer's recommendations. DNA was quantified using the Qubit dsDNA BR Assay Kit and a Qubit Fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA).

Comparative genomic hybridization

aCGH profiling was performed with the Human Agilent Sureprint G3 8 × 60 K Microarray Kit (Agilent Technologies, Santa Clara, CA, USA). Tumor DNA was labeled with cyanine 5 (Cy5), while reference DNA from an individual of the same sex as the patient was labeled with Cy3. Sample and reference DNAs were pooled and hybridized for 24 h at 67°C on the arrays. The fluorescence was read by an Agilent SureScan Microarray scanner, and the Cy5/Cy3 fluorescence ratios were converted into log2-transformed values with Cytogenomics software (Agilent).

The threshold of the absolute value of the log2 fluorescence ratio retained to define a chromosomal anomaly was 0.25. A mean log2 ratio was calculated when, for at least three probes located on contiguous positions on the chromosome, a log2 ratio absolute value greater than 0.25 and of the same sign was measured. The minimum size of the anomalies considered in the interpretation of the results was set at 1 Mb.

The different chromosomal anomalies were defined by the Cytogenomics software according to the mean log2 ratio values, as follows: homozygous deletion for a value <-1, loss of one gene copy for a

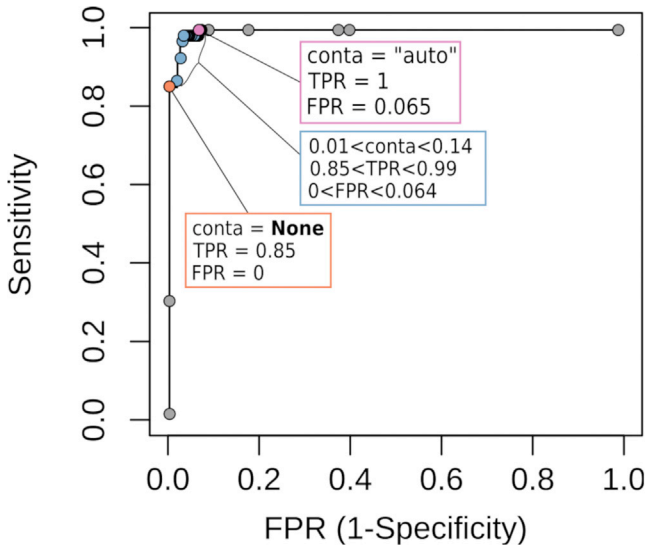

Figure 5. ROC curve for the ICR96 dataset

Some points of interest are highlighted: contamination = "auto" (pink), 0.01 < contamination < 0.14 (blue), and contamination = "none" (orange). TPR, true positive rate. ifCNV is a new software that combines artificial intelligence using two isolation forests and a comprehensive scoring method to faithfully detect CNVs among various samples. ifCNV is a publicly available open-source software (https://github.com/SimCab-CHU/ifCNV) that allows the detection of CNVs in various clinical situations.

value between −0.25 and −1, gain of one gene copy for a value between 0.25 and 1, and amplification (gain of at least five copies) for a value >2.

TruSeq Custom Amplicon library preparation assay: TSCA dataset

Library preparation was performed as previously described.[37] Briefly, extracted DNA was qualified using KAPA Sybr Fast qPCR (Kapa Biosystems, Boston, MA, USA). A home-made panel targeting specific exons of 35 clinically relevant cancer genes was used for amplification of regions of interest. For each sample, dual-strand libraries were prepared using a TruSeq Custom Amplicon protocol, as described by the manufacturer (Illumina, San Diego, CA, USA). After amplification, PCR products were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) and quantified, normalized, and pair-end sequenced on a MiSeq instrument (2 × 150 cycles, Illumina). This dataset is composed of 81 samples from 59 different sequencing runs, with 25 CNV^{pos} samples (Table 1).

Advanta Solid Tumor NGS library preparation assay: Juno dataset

Libraries were prepared using the Advanta Solid Tumor NGS Library Prep Assay with the automated Juno system on integrated fluidic circuits (LP 8.8.6 IFC) (Fluidigm, San Francisco, CA, USA) following the manufacturer's procedure. The panel is developed to allow the detection of somatic mutations in 53 oncology-relevant genes (234 kb, 1,508 assays). Briefly, the LP 8.8.6 IFCs were primed with 20 ng DNA per sample in the PCR mix. After amplification, pooled harvested samples were purified using AMPure XP beads (Beckman Coulter), and a second PCR was performed to integrate the sequencing adapters. Libraries were then quantified, normalized, and pair-end sequenced on a NextSeq instrument (2 × 150 cycles, Illumina). In this dataset, there are 43 samples with 26 CNV^{pos} from 20 different sequencing runs (Table 1).

Binary classification indicators

TPR (or sensitivity), FPR, TN rate (TNR; or specificity), FN rate (FNR), PPV, Acc, and the MCC were used to measure the performance of ifCNV. These were computed as

$$TPR = Se = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = 1 - Sp = \frac{FP}{N} = \frac{FP}{TN + FP}$$

$$TNR = Sp = \frac{TN}{N}$$

$$FNR = \frac{FN}{P}$$

$$PPV = \frac{TP}{TP + FP}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Localization scoring

Specific regions of biological significance (gene or exon) can be covered by several targets. In the event that a region is altered, all the targets in the region should be modified. By contrast, if only one target in the region is modified, it is likely to be an FP. We integrated this reasoning to develop a localization score in order to reduce the number of FPs. The localization score depends on the number of modified targets in the region, the number of targets in the region, and the total number of targets in the panel. A semi-open log scale incorporating the ratio of modified targets in the region was chosen. It is calculated as follows:

$$score = \log\left(\left(\left(\frac{n}{N}\right)^{k}.\left(1 - \frac{n}{N}\right)^{n-k}\right)^{-1}\right)\left(\frac{k}{n}\right)$$

with k the number of modified targets on the region,
n the number of targets on the region,
N the total number of targets

Pre-processing

ifCNV requires the .bam sequence files as an input but does not provide a function to create them. Therefore, the user needs to generate

the proper .bam files from the raw sequence .fastq files. ifCNV's preprocessing step uses the bedtools multicov function to generate the reads matrix. It takes as input the aligned .bam files and outputs a read-depth matrix that was used for the CNV detection analysis. In this study, the .bam files were created using the Burrows-Wheeler Alignment (BWA) tool:[38] the .fastq files were aligned against the reference genome with *bwa mem*.

Data availability statement

All the datasets used in this study are available at https://github.com/SimCab-CHU/ifCNV.

ACKNOWLEDGMENTS

This work has not benefitted from any funding, institutional or corporate.

AUTHOR CONTRIBUTIONS

Conceptualization, S.C.-A. and J.A.V.; formal analysis, S.C.-A..; investigation, S.C.-A.; methodology, S.C.-A.; project administration, J.S. and J.A.V.; resources, M.B., C.G., and J.S.; software, S.C.-A. and C.V.G.; supervision, J.S.; visualization, S.C.-A.; writing – original draft, S.C.-A.; writing – review & editing, J.A.V., L.F., and J.S.

DECLARATION OF INTERESTS

The authors do not have any conflicts of interest.

REFERENCES

1. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
2. Ellingford, J.M., Horn, B., Campbell, C., Arno, G., Barton, S., Tate, C., Bhaskar, S., Sergouniotis, P.I., Taylor, R.L., Carss, K.J., et al. (2018). Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases. J. Med. Genet. *55*, 114–121.
3. Høyer, H., Braathen, G.J., Eek, A.K., Nordang, G.B.N., Skjelbred, C.F., and Russell, M.B. (2015). Copy number variations in a population-based study of charcot-marie-tooth disease. Biomed. Res. Int. *2015*, 960404.
4. Bochukova, E.G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., et al. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. Nature *463*, 666–670.
5. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.
6. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res. *16*, 949–961.
7. Wang, H., Liang, L., Fang, J.-Y., and Xu, J. (2016). Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. Oncogene *35*, 2011–2019.
8. Jebbink, M., de Langen, A.J., Boelens, M.C., Monkhorst, K., and Smit, E.F. (2020). The force of HER2 – a druggable target in NSCLC? Cancer Treat. Rev. *86*, 101996.
9. Recondo, G., Bahcall, M., Spurr, L.F., Che, J., Ricciuti, B., Leonardi, G.C., Lo, Y.-C., Li, Y.Y., Lamberti, G., Nguyen, T., et al. (2020). Molecular mechanisms of acquired resistance to MET tyrosine kinase inhibitors in patients with MET exon 14–mutant NSCLC. Clin. Cancer Res. *26*, 2615–2625.
10. Rafii, S., Gourley, C., Kumar, R., Geuna, E., Ern Ang, J., Rye, T., Chen, L.-M., Shapira-Frommer, R., Friedlander, M., Matulonis, U., et al. (2017). Baseline clinical predictors of antitumor response to the PARP inhibitor olaparib in germline BRCA1/2 mutated patients with advanced ovarian cancer. Oncotarget *8*, 47154–47160.
11. Camidge, D.R., Otterson, G.A., Clark, J.W., Ignatius Ou, S.-H., Weiss, J., Ades, S., Shapiro, G.I., Socinski, M.A., Murphy, D.A., Conte, U., et al. (2021). Crizotinib in patients with MET-amplified NSCLC. J. Thorac. Oncol. *16*, 1017–1029.
12. Yip, S., Butterfield, Y.S., Morozova, O., Chittaranjan, S., Blough, M.D., An, J., Birol, I., Chesnelong, C., Chiu, R., Chuah, E., et al. (2012). Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. J. Pathol. *226*, 7–16.
13. Planchard, D., Loriot, Y., André, F., Gobert, A., Auger, N., Lacroix, L., and Soria, J.C. (2015). EGFR-independent mechanisms of acquired resistance to AZD9291 in EGFR T790M-positive NSCLC patients. Ann. Oncol. *26*, 2073–2078.
14. Condorelli, R., Mosele, F., Verret, B., Bachelot, T., Bedard, P.L., Cortes, J., Hyman, D.M., Juric, D., Krop, I., Bieche, I., et al. (2019). Genomic alterations in breast cancer: level of evidence for actionability according to ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). Ann. Oncol. *30*, 365–373.
15. Onsongo, G., Baughn, L.B., Bower, M., Henzler, C., Schomaker, M., Silverstein, K.A.T., and Thyagarajan, B. (2016). CNV-RF is a random forest-based copy number variation detection method using next-generation sequencing. J. Mol. Diagn. *18*, 872–881.
16. Huang, T., Li, J., Jia, B., and Sang, H. (2021). CNV-MEANN: a neural network and mind evolutionary algorithm-based detection of copy number variations from next-generation sequencing data. Front. Genet. *12*, 700874.
17. Povysil, G., Tzika, A., Vogt, J., Haunschmid, V., Messiaen, L., Zschocke, J., Klambauer, G., Hochreiter, S., and Wimmer, K. (2017). panelcn.MOPS: copy-number detection in targeted NGS panel data for clinical diagnostics. Hum. Mutat. *38*, 889–897.
18. Guo, Y., Wang, S., and Yuan, X. (2021). HBOS-CNV: a new approach to detect copy number variations from next-generation sequencing data. Front. Genet. *12*, 642473.
19. Zhao, H., Huang, T., Li, J., Liu, G., and Yuan, X. (2020). MFCNV: a new method to detect copy number variations from next-generation sequencing data. Front. Genet. *11*, 434.
20. Miller, N., Bouma, M., Sabatini, L., and Gulukota, K. (2021). SILO: a computational method for detecting copy number gain in clinical specimens analyzed on a next-generation sequencing platform. J. Mol. Diagn. S1525–S1578. 00240–3.
21. Viailly, P.-J., Sater, V., Viennot, M., Bohers, E., Vergne, N., Berard, C., Dauchel, H., Lecroq, T., Celebi, A., Ruminy, P., et al. (2021). Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers. BMC Bioinf. *22*, 120.
22. Deshpande, V., Luebeck, J., Nguyen, N.-P.D., Bakhtiari, M., Turner, K.M., Schwab, R., Carter, H., Mischel, P.S., and Bafna, V. (2019). Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat. Commun. *10*, 392.
23. Jiang, Y., Wang, R., Urrutia, E., Anastopoulos, I.N., Nathanson, K.L., and Zhang, N.R. (2018). CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol. *19*, 202.
24. Johansson, L.F., van Dijk, F., Swertz, M.A., Sijmons, R.H., Sikkema-Raddatz, B., de Boer, E.N., van Dijk-Bos, K.K., Jongbloed, J.D.H., van der Hout, A.H., Westers, H., et al. (2016). CoNVaDING: single exon variation detection in targeted NGS data. Hum. Mutat. *37*, 457–464.
25. Fowler, A., Mahamdallie, S., Ruark, E., Seal, S., Ramsay, E., Clarke, M., Uddin, I., Wylie, H., Strydom, A., Lunter, G., and Rahman, N. (2016). Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. *1*, 20.
26. Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., and Duan, J. (2021). CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. IEEE/ACM Trans. Comput. Biol. Bioinform. *18*, 539–549.
27. Quenez, O., Cassinari, K., Coutant, S., Lecoquierre, F., Le Guennec, K., Rousseau, S., Richard, A.C., Vasseur, S., Bouvignies, E., Bou, J., et al. (2021). Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. Eur. J. Hum. Genet. *29*, 99–109.

28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Python *12*, 2825–2830.

29. 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.

30. 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.

31. Mahamdallie, S., Ruark, E., Yost, S., Ramsay, E., Uddin, I., Wylie, H., Elliott, A., Strydom, A., Renwick, A., Seal, S., and Rahman, N. (2017). The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. Wellcome Open Res. *2*, 35.

32. Hubert, M., Debruyne, M., and Rousseeuw, P.J. (2018). Minimum Covariance Determinant and Extensions. Wiley Interdiscip. Rev. Comput. Stat. 1443–1471.

33. Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. (2000). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, *12*, Digital Library., ed (Association for Computing Machinery), pp. 93–104.

34. Liu, F.T., Ting, K.M., and Zhou, Z.-H. (2012). Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data *6*, 1–39.

35. Rousseeuw, P.J., and Driessen, K.V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics *41*, 212–223.

36. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

37. Vendrell, J.A., Taviaux, S., Béganton, B., Godreuil, S., Audran, P., Grand, D., Clermont, E., Serre, I., Szablewski, V., Coopman, P., et al. (2017). Detection of known and novel ALK fusion transcripts in lung cancer patients using next-generation sequencing approaches. Sci. Rep. *7*, 12510.

38. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595.