T-CNV: a robust tool for detecting and visualizing copy number variants in targeted sequencing data.

liu ye
Top Gene Tech(Guangzhou) Co., Ltd.

wu yangming
Top Gene Tech(Guangzhou) Co., Ltd.

zheng zexin
Top Gene Tech(Guangzhou) Co., Ltd.

zhou tianliangwen (✉ tian.zhou@topgene.cn)
Taylor and Francis Group https://orcid.org/0000-0003-1269-1489

Methodology article

Keywords: copy number variants, targeted sequencing panel, visualization, T-CNV

Posted Date: May 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27672/v1

T-CNV: a robust tool for detecting and visualizing copy number variants in targeted sequencing data.

Ye Liu[1,2], Yangming Wu[1,2], Zexin Zheng[1,2], Tianliangwen Zhou[1,3]

[1] R&D department, Top Gene Tech (Guangzhou) Co., Ltd., Guangzhou, Guangdong, P.R.China

[2] These authors contributed equally to this work.

[3] Corresponding author. Contact e-mail: tian.zhou@topgene.cn

Abstract

Background: Copy number variants (CNVs) are widespread among human genes, causing Mendelian or sporadic traits, or associating with complex diseases. Several tools have been developed for CNV assessment based on next generation sequencing (NGS) data using Read-depth (RD) strategy. However, maintaining high level of sensitivity and specificity is always challenging. Here, we present a novel, powerful, user-friendly and open accessed tool, T-CNV for CNV detection and visualization in targeted NGS panel.

Results: T-CNV consists of primary CNV detection and CNV candidates confirmation steps. After computing log2 values of normalized read depth ratio of tumor and normal/control sample, T-CNV confirms each possible CNV candidates by bins method, Gaussian Mixture Model (GMM) clustering approach and window-sliding method. We benchmarked its capacity with MLPA-validated dataset. Compared to three other advanced tools, T-CNV presents excellent performance with 95.42% sensitivity, 99.93% specificity and 93.63% positive predict value in MLPA-validated dataset, while achieving satisfactory performance in simulation study (sensitivity 65.95%, positive predict value 88.71% at coverage 100X).

Conclusions: T-CNV is a novel and robust tool for CNV detection and visualization in targeted NGS panel consisting of determination of possible CNV candidates and further confirmation by three different methods. It's publicly available at https://github.com/Top-Gene/T-CNV.

Keywords: copy number variants, targeted sequencing panel, visualization, T-CNV

Background

Copy number variants (CNVs) are widespread among human genes, covering 12% of human genome[1]. It can cause Mendelian or sporadic traits or be associated with complex diseases[2, 3], such as neurodevelopmental disorders[4, 5] and cancers[6]. CNV is a large category of structural variants (SVs) that first defined as a segment of DNA that is 1 kb or larger and presents at a variable copy number in comparison with a reference genome[7]. With increasing detections on human genome, CNV has widen to include unbalanced structural variants with >50bp in length[8]. It's considered that the major mechanism for phenotypes derived from CNVs relates to gene dosage effect[9]. CNVs locate either in the dosage-sensitive gene or nearby, which alters or poses an effect on the gene expression level and leads to an abnormal phenotype[10]. Several tools are routinely used for CNVs assessment, including fluorescent *in situ* hybridization (FISH), array comparative genomic hybridization (aCGH), multiplex ligation-dependent probe amplification (MLPA), and recently next generation sequencing (NGS).

Although exome- and genome-sequencing techniques are gradually being applied in clinical laboratories, disease-targeted testing still holds a firm place for its high coverage on interested region and cost-effectiveness[11]. To date, MLPA[12] is referred as the gold standard tool for CNV detection, because of its low cost, high sensitivity and specificity, and medium throughput[13]. However, the principle of MLPA mainly encompasses the weakness in lack of sensitivity of regions not directly designed in probe set[14]. In contrast, targeted panel provides a cost-effective and high-throughput way for not only determining SNP and indels, but also identifying common and novel CNVs[14, 15].

There are five strategies normally applied in present tools to detect CNVs in NGS data: (1) Paired-end mapping (PEM)[16], (2) read depth (RD)[17], (3) split-read[18], (4) *de novo* assembly[19], and (5) the combination of previously described methods[20, 21]. The concept of RD is similar to that of using density data that a lower region coverage than expected indicates deletion and a higher indicates duplication. Compared with other strategies, RD method has the ability to assess exact copy number of target regions and to detect large insertions and CNVs in complex genomic region classes[17]. Although challenges remain, RD has become the routinely used strategy for CNV assessment because of the accumulation of high-coverage NGS data. In order to obtain better prediction, mathematical models were widely used, including hidden Markov chain model (HMM)[22]and Gaussian Mixture Model (GMM)[23, 24].

In this study, we develop T-CNV as a novel and powerful tool for CNV detection and visualization in targeted NGS panel. Only autosomal chromosomes were considered in T-CNV to avoid complication of gender. With given interested regions, it utilizes read depth strategy for prior CNV detection and confirmed by three additional steps to achieve high level of sensitivity and specificity. Each possible CNV exons are further confirmed separately by non-overlapped bins method, gaussian mixture model (GMM) clustering approach and overlapped window-sliding method. Later, we conducted performance assessment with three out-of-state targeted sequencing CNV tools, DECoN[25], Atlas-CNV[26] and VisCap[27], on a MLPA-validated dataset ICR96[28] and simulation datasets. DECoN is a modified version of ExomeDepth[22], which applied beta-binomial distribution to capture the variability in read count ratio and combined the likelihood across multiple exons using HMM. The main modifications include compatibility of software and version, and alteration of HMM transition probabilities depends upon distance between exons. Atlas-CNV detects CNVs by measuring log2 ratio and C-score (Z-like-score) to reduce false discovery rate, while VisCap detects CNVs by user-defined log2 ratio threshold and boxplot method. The three tools include visualization of CNV results and are validated on targeted sequencing data. In this study, T-CNV gave satisfactory performance on CNV detection on both validated dataset and simulative datasets. In the meanwhile, T-CNV also provides comprehensive visualization result for better understanding.

Results

Overview of T-CNV

T-CNV was developed in Python 3.5.6 and R 3.5.1 or higher. It's publicly available at https://github.com/Top-Gene/T-CNV. The overview of T-CNV pipeline is illustrated in Figure 1. The input includes, (1) a bed file containing target regions, (2) DNA sequencing read alignments in BAM format for tumor samples in one pool, (3) if available, DNA sequencing read alignments in BAM format for corresponding normal/control sample. NGS data must be converted to fastq file and aligned to reference genome (hg19) before CNV assessment. Burrows-Wheeler Aligner (BWA)[29] and Genome Analysis Tool Kit (GATK)[30] are recommended for sequence alignment, realignment, recalibration.

Normalization and GC correction

T-CNV identifies CNV candidates based on read depth strategy. Thus, normalization and bias correction before identification is crucial. GC content was found to have influence on the depth of coverage of NGS data[17, 31]. Since GC content influence may vary between samples even in the same pool, LOESS for GC correction was applied at original read depth of each sample for normalization[32, 33]. We observed the dependency between tumor/normal read depth ratio and GC content (Supplementary Figure S1A). Therefore, we conducted Local Polynomial Regression (LOESS) and smoothing to reduce the influence of GC-bias on log2 values of tumor/normal ratio.

Prior researches corrected GC-content bias by performing LOESS at different level, including fragment read depth[32] and read depth ratios[23, 34]. We compared the outcome for GC correction at different level (Supplementary Figure S1A&B&C&D) and the two strategies showed highly correlated on ICR96 dataset with Pearson correlation r=0.91 in 96 samples (min. 0.76, max. 0.99) (Supplementary Figure S1E&F&G).

Quality control of ICR96 dataset

In order to reduce false discovery rate, quality control step was introduced. QC poor samples were removed from the pool for better prediction. We conducted a prior noise test on exon log2 value of tumor/normal ratio ($log_2(\frac{T}{N})_{Normalised}$) to derive an index for quality control. We first assumed a sample pool containing 50 pairs of tumor and normal samples with $log_2(\frac{T}{N})_{Normalised} = 1$ targeting 1500 exons. Secondly, random noise was computed in Python 3.5.6 by random function from 5% to 30% in 5% increments and spiked into each exon normalized tumor/normal ratio. The results (Supplementary Figure S2) showed the 50 pairs distribution overview in the left, while in the right showed the fluctuation of $log_2(\frac{T}{N})_{Normalised}$ of each exon in one pair. The worst case that samples were contaminated with 25% noise, may not influence the capacity of CNV calling under threshold of [0.32, -0.42], yielding an overall standard deviation 0.2 of $log_2(\frac{T}{N})_{Normalised}$. Thus, we considered the standard deviation of exon log2 value (Std_{Exon}) of tumor and normal ratio, after normalization and GC-bias correction, as a quality control index. A sample with Std_{Exon} > 0.2 was regarded as discordant and marked as 'QC poor'. It was removed from the pool in further assessment.

Because the samples in ICR96 dataset were single tumor samples without corresponding normal pair, T-CNV generated the control for the samples in the same pool as illustrated in section "Normalization and quality control" in Method. An overview of the log2 values distribution in pool 1 and pool 2 was illustrated in Figure 2A&B. Nine samples (4 in pool 1, 5 in pool 2) with Std_{Exon} higher than 0.2 were found in ICR96 dataset. Since further CNV assessment was based on log2 value, these nine samples would not be assessed in further performance assessment.

Optimization of bins method

Normally, RD CNV tools employ segmentation by dividing chromosome into non-overlapped segments to estimate the copy number[17]. T-CNV first identified possible CNV candidates by the exon $log_2(\frac{T}{N})_{Normalised}$. Later, non-overlap bins method was used to verify the candidates. We tuned the bin size in optimization test using ICR96 dataset. The results (Supplementary Figure S3A) indicated that the best solution was setting 30bp bin size, yielding overall sensitivity 95.42%, specificity 99.93%, positive predict value (PPV) 93.63% and negative predict value (NPV) 99.95%. Also, comparing the performance between predicting deletions (Supplementary Figure S3B) and duplications (Supplementary Figure S3C), setting bin size as 30bp presented 95.91%, 94.51% sensitivity and 92.66%, 95.56% PPV, respectively.

Confirmation by GMM clustering

Two distinct peaks were observed in the log2 value distribution histogram of genes containing CNV candidates (Supplementary Figure S4A) indicating the distribution of positive CNVs and normal exons. It therefore motivated to detect positive CNVs using GMM clustering method. T-CNV considered the log2 value from three genes (possible CNV candidate gene and other two negative genes) as clustering feature. The specific selection of genes was illustrated in Supplementary Figure S5.

Clustering results were given by GMM model and Expectation-Maximization (EM) using standardized log2 value for CNV candidate gene as single feature (Supplementary Figure S4B). Green and red dots in Supplementary Figure S4B stand for two different clusters of log2 value in NF1 in sample 17338, which indicated GMM was able to distinguish positive CNV candidates from normal exons. In accordance with our GMM sampling plan (Supplementary Figure S5), the GMM clustering result for possible CNV candidate exon 46 in gene NF1 in sample 17338 was shown in Supplementary Figure S5C&D&E.

Optimization of window-sliding method

Even though GMM clustering provides a powerful approach to identify true CNVs, the idea behind GMM is soft clustering that each point is assigned to component considering its maximum probability ($argmax_{component}(P(component|each\ point))$). In addition, T-CNV performs GMM clustering approach based on certain sampling strategy (Supplementary Figure S5). False positive would arise when the sampling strategy doesn't fit. During pre-test on ICR96dataset, GMM clustering approach gave 11.35% (32/ (250+32)) false discovery rate (FDR). Thus, other approach was necessary for better performance.

Compared to MLPA results, false positive CNVs by bins method and GMM clustering approach presented larger fluctuation (higher standard deviation) at overlapped windows (Supplementary Figure S6). We used overlapped window-sliding method to reduce FDR. Optimization of window-sliding method was conducted by using GMM clustering positive CNV candidates (310 in total), 5bp increments in windows size and 1bp increments in step length. Under 10bp window size and 6bp step length, window-sliding method reached highest area under curve (AUC) 0.82 in receiver operating characteristic curve (ROC) and cut-off value was standard deviation 0.14 (Supplementary Figure S7A&B). The precision-recall graph (Supplementary Figure S7C&D) also demonstrated the same optimal settings as 10bp window size and 6bp step with cut-off value 0.14.

Performance comparison on ICR96 dataset

To estimate the performance of our tool, we benchmarked T-CNV with a MLPA-validated ICR96 dataset, which includes 32 validated genes from prior MLPA test[28]. According to the MLPA validated results, 262 exons (171 deletions, 91 duplications) were determined as true CNV candidates in 59 samples, excluding 9 “poor-QC” samples. In addition, “Normal” exons in MLPA validated results were considered as Non-CNV. T-CNV gave an excellent performance with an overall 95.42% sensitivity (250/262), 95.91% (164/171) sensitivity for deletions and 94.51% (86/91) for duplication, respectively.

Furthermore, we compared the performance of T-CNV with other three out-of-state targeted sequencing CNV tools, including DECoN, Atlas-CNV and VisCap. We ran the other three tools under their default setting based on their protocol. Because only autosomal chromosomes are assessed in our study, we fixed the gender as female in tools setting, if necessary.

In summary, we compared four parameters, sensitivity, specificity, PPV and NPV among four tools on ICR96 dataset (Figure 3A). In accordance with Roca *et al*[35] findings, DECoN presented best performance among four tools, yielding an overall sensitivity 99.24% (260/262) and PPV 97.01 % (260/268). Compared to Altas and VisCap, T-CNV outperformed with higher sensitivity 95.42%, specificity 99.93%, PPV 93.63% and NPV 99.95% (Figure 3A). Regards to the performance on detecting deletion or duplication, T-CNV and DECoN showed no significance (Figure 3B&C), while, Atlas-CNV and VisCap, gave better prediction on duplications than deletions (Atlas-CNV: false positive rate: deletion 13.87%, duplication: 4.94%; VisCap deletion 24.31%, duplication 6.67%).

Performance assessment on simulative dataset

We generated simulative dataset containing 100 samples with random size CNVs spiked in each sample using TargetedSim (https://sourceforge.net/projects/targetedsim/files/TargetedSim/, accessed September 10, 2019). We evaluated the performance of T-CNV and DECoN on the simulative dataset at average coverage 50X, 100X and 500X in triplicate, under default settings of both tools. The result for sensitivity and PPV study (Figure 4A) demonstrated steady performance of T-CNV at different depth. While DECoN presented higher sensitivity (Figure 4B), T-CNV achieved higher PPV than DECoN, indicating that less false positive was predicted by T-CNV. T-CNV showed higher sensitive on 100X and 500X, mainly due to the minimum coverage setting 30X in default settings.

Discussion

We reported T-CNV as a powerful tool for CNV detection and visualization in targeted DNA sequencing panels. It allows a set to samples or tumor samples with their normal pair as input. Also, it accepts bam files or locus-depth files generated from SAMtools by command “depth”. When locus-depth files are not available, T-CNV converts bam files into locus-depth files automatedly and starts CNV analysis.

T-CNV implemented LOESS at log2 value level, since its result was highly correlated to the GC-correction result of LOESS at original depth level. In addition, T-CNV considered a standard deviation of log2 value as an index to separate discordant samples to achieve high sensitivity. Based on the QC criteria, we found 9 samples in ICR96 dataset were marked as poor QC. We also measured the Pearson correlation coefficient of samples and its corresponding pool control (Figure 2C). The QC poor samples also showed less correlation to their controls, indicating the noise and bias in QC poor samples were not effectively removed by previous normalization steps.

Since single exon CNV detection is challenging[36], after filtering possible CNV candidates, T-CNV preforms three critical confirmation methods at each candidate. Thus, it's compromised to give false negative prediction in a CNV covering consecutive exons (Figure 5A&B). T-CNV separates possible CNV candidates in the primary step by measuring whether the log2 value exceeds threshold, where false negative candidates might be yielded, such as NF1 exon 45 in sample 17338 (Figure 5B). Another possibility for false negative prediction results from confirmation methods. A positive candidate is identified only when all confirmation methods determine it as positive.

False positive prediction is inevitable due to cut-off setting. T-CNV has optimized thresholds and cut-off value in bins method and window-sliding method using ICR96 dataset to achieve best performance. When comparing false positive prediction with other tools, we observed that Atlas-CNV and T-CNV detected BAP1 exon 1 in sample 17340 as positive duplication (Figure 5C), while all four tools defined a CDH1 exon 1 in sample 17301 as positive deletion (Figure 5D). These false positive CNVs have small size (around 50bp), which might influence the capacity of bins method and window-sliding method due to the default size selection, resulting in false identification.

An example of the visualization results (sample 17375 from pool2 in ICR96 dataset) was shown in Supplementary Figure S8, plotting with the log2 value of all targeted exons. The predicted copy number of positive CNVs was indicated under the ploting, calculated from the log2 value.

T-CNV has been benchmarked by ICR96, achieving excellent sensitivity 95.42%, specificity 99.93%, PPV 93.63%, NPV 99.95%. Furthermore, T-CNV achieved satisfactory results in simulation study. Since GMM clustering approach requires specific sampling plan, T-CNV achieve best performance in panel with multiple genes. Thus, it's highly recommended to select appropriate target region when applying T-CNV.

Conclusions

This study reports a robust, novel and open source CNV tool, named T-CNV, consisting of determination of possible CNV candidates and further confirmation by three different methods. It accepted a pair of tumor and normal samples of a set of samples as input. It is able to automatedly convert bam file into SAMtools depth file, while depth files are acceptable. T-CNV is benchmarked by MLPA-validated dataset and also presents satisfactory performance on simulative dataset. As NGS are widely used clinically and academically, we believe T-CNV provides a parallel solution for CNV detection while calling SNPs and indels.

Methods

T-CNV pipeline

Normalization and quality control

To start CNVs calling, RD for each locus inside target region was determined by SAMtools[37]. Tumor sample and corresponding normal sample were considered as a pair in T-CNV. When normal/control sample was not available, the median RD within the interval $[mean(RD) \pm 2 \cdot standard\ deviation(RD)]$ in sample pool served as control sample.

Normalization between the RD of a pair sample (tumor and normal/control) is crucial. T-CNV computed reads per thousand bases per million reads sequenced (RPKM) at each exon by

$$RPKM = RD\ at\ exon \cdot 10^9 / (RD\ in\ target\ regions \cdot Exon\ length).$$

The normalized RD of tumor sample and normal sample was converted to log2-value $(log_2(RPKM_{tumor}/RPKM_{normal}))$. Similar to Benjamini and Speed's[33] research, log2-value was GC-bias corrected by LOESS in R 3.5.1with 75% smoothing span and degree 2.

T-CNV considered thresholds of -0.42 for deletion and 0.32 for duplication for diploid. Any exon with log2 value higher than 0.32 or lower than -0.42 was considered as possible CNV candidates. The possible CNV candidates were further confirmed by follow-up steps.

Samples with evenly distribution of exon RD were expected in a pool. T-CNV considered the standard deviation of all exon log2-value (Std_{Exon}) as an index for quality control. A noise randomly introduced test was conducted in a pool of 50 samples containing 1500 exons. In the test, 5% increment of noise was randomly introduced into all samples on log2 value using uniform function in "random" package in Python 3.5.6. The results (Supplementary Figure S2) illustrated that maximum 25% of noise, corresponding to Std_{Exon} 0.2, under threshold [-0.42, 0.32] may not influence CNV candidate identification. Therefore, to reduce false positive, samples with Std_{Exon} higher than 0.2 were considered as poor quality.

Confirmation of CNVs

Bins method

The possible CNV exons are divided into non-overlap bins with 30bp in length as default value. Average RD for each bin in tumor and normal samples were calculated and log2 value was determined. One CNV was confirmed by bins method, if 90% or more bins in the exon presented log2 values higher than 0.32 or lower than -0.42.

GMM clustering method

We first assume that the probability distribution of GC-corrected log2 value can be approximated with a set of gaussian (deletion, duplication and non-CNV). Let $\{\mu_j, \Sigma_j, \phi_j\}$ denote the mean, covariance, the weight of j^{th} component of GMM ($1 \leq j \leq K$). The likelihood of the observed GC-corrected log2 value x_i is

$$P(x_i) = \sum_{j=1}^{K} P(Z_i = j\,) \cdot P(x_i|Z_i = j) = \sum_{j=1}^{K} \phi_j \cdot \mathcal{N}(x_i|\mu_j, \Sigma_j),$$

where Z_i represents the hidden CNV state of x_i, $Z_i \in \{1, \ldots, K\}$ and $\sum_{j=1}^{K} \phi_j = 1$ $(\phi_j \geq 0, \forall j)$. Based on our assumption, observed data (GC-corrected log2 values) are 1-dimension. Therefore, the probability for each set of gaussian is

$$\mathcal{N}(x_i | \mu_j, \sigma_j^2) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp{(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2})}.$$

To confirm the possible CNV candidate, we select two other genes without CNV candidates, located in the same chromosome as observed data (illustrated in Supplement Figure S5). In this case, two-components (positive and normal) GMM ($K = 2$) is fitted with unknown parameters $\theta = \{\mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K, \phi_1, \ldots, \phi_K\}$. Given observed $X = \{x_1, x_2, \ldots, x_i, \ldots, x_N\}$, the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{N} (\ln \sum_{j=1}^{K} \phi_j \cdot \mathcal{N}(x_i | \mu_j, \Sigma_j)).$$

Our goal is to maximize the log-likelihood. Thus, we employed EM algorithm to estimate the parameter $\{\mu_j, \sigma_j, \phi_j\}$ of GMM as followed.

E-step: compute the conditional probability that x_i belongs to j^{th} component of GMM for observed X

$$P(Z_i = j | x_i) = \hat{\gamma}_{i,j} = \frac{P(Z_i = j) \cdot P(x_i | Z_i = j)}{P(x_i)} = \frac{\hat{\phi}_j \cdot \mathcal{N}(x_i | \hat{\mu}_j, \hat{\sigma}_j^2)}{\sum_{k=1}^{K} \hat{\phi}_k \cdot \mathcal{N}(x_i | \hat{\mu}_k, \hat{\sigma}_k^2)}.$$

M-step: use the updated $\hat{\gamma}_{i,j}$ calculate $\{\hat{\mu}_j, \hat{\sigma}_j^2, \hat{\phi}_j\}$ by

$$\hat{\phi}_j = \sum_{i=1}^{N} \frac{\hat{\gamma}_{i,j}}{N},$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,j} \cdot x_i}{\sum_{i=1}^{N} \hat{\gamma}_{i,j}},$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,j} \cdot (x_i - \hat{\mu}_j)^2}{\sum_{i=1}^{N} \hat{\gamma}_{i,j}}.$$

Iterate the E-step and M-step until convergence, when $\{\hat{\mu}_j, \hat{\sigma}_j^2, \hat{\phi}_j\}^{t+1} \approx \{\hat{\mu}_j, \hat{\sigma}_j^2, \hat{\phi}_j\}^{t}$. The above steps were completed by using GaussianMixture in scikit-learn[38] package in Python 3.5.6.

Window-sliding method

The possible CNV exon was divided into overlapped M bp windows with N bp steps ‘slide’ away. RD for each window and log2 values in tumor and normal samples were computed. Later, the standard deviation for log2 values of windows ($Std(log_2(windows))$) was calculated. A possible CNV candidate was confirmed as positive when the $Std(log_2(windows))$ was lower than 0.14.

A CNV candidate was confirmed as positive if above all confirmation methods identified as positive. Thus, the output of a sample includes the confirmation results from bins method, GMM clustering method and window-sliding method, quality control results and visualization file. An example of T-CNV result for CNV-positive sample was shown in Supplementary Figure S8.

MLPA validated dataset

The MLPA validated dataset ICR96[28] can be accessed through the European-Genome phenome Archive (EGA) under accession number EGAS00001002428. ICR96 consists of 96 targeted NGS samples (66 MLPA validated CNV-positive samples and 30 CNV-negative samples). Validated positive samples contains CNVs in BRCA1, BRCA2, TP53, MLH1, MSH2, PMS2, EPCAM or PTEN, which were most frequently tested in clinical practice as cancer predisposition genes.

Simulation dataset

We simulated Illumina paired-end reads datasets spiked with deletion and duplication CNVs by TargetedSim (https://sourceforge.net/projects/targetedsim/files/TargetedSim/, accessed September 10, 2019). Homozygous and heterozygous were indicated by 100% and 50% reduce/increase in depth for deletion/duplication, respectively. Simulative dataset consisted of 100 samples, randomly spiked with CNVs (3 heterozygous deletions, 3 heterozygous duplications and 2 homozygous duplications with random length from 1kb to 10kb) in each sample with targeted region of 1634 exons. To compare CNV calling performance with DECoN, we conducted analysis from dataset generation to CNV detection at coverage 50X, 100X and 500X.

In this study, true positive (TP) is defined as MLPA-validated or known CNVs, while true negative (TN) is defined as negative exons. Sensitivity (TP/ (TP + false negative (FN))), specificity (TN/ (TN + false positive (FP))), PPV (TP/ (TP+FP)) and negative predict value (NPV) (TN/ (TN+FN)) were calculated in exon basis.

List of abbreviations

CNVs: copy number variants

SVs: structural variants

FISH: fluorescent *in situ* hybridization

aCGH: array comparative genomic hybridization

MLPA: multiplex ligation-dependent probe amplification

NGS: next generation sequencing

RD: read depth

PEM: Paired-end mapping

HMM: hidden Markov chain model

GMM: Gaussian Mixture Model

BWA: Burrows-Wheeler Aligner

GATK: Genome Analysis Tool Kit

LOESS: Local Polynomial Regression

PPV: positive predict value

NPV: negative predict value

EM: Expectation-Maximization

FDR: false discovery rate

AUC: area under curve

ROC: receiver operating characteristic curve

RPKM: reads per thousand bases per million reads sequenced

TP: true positive

TN: true negative

FN: false negative

FP: false positive

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors are employed by Top Gene Tech (Guangzhou) Co., Ltd. by the time of manuscript submission. The authors declare that they have no competing interests.

Funding

Not applicable.

Availability of data and materials

The ICR96 dataset is public by Professor Nazneen Rahman's team and can be accessed through the European-Genome phenome Archive (EGA) under the accession number EGAS00001002428[28]. The simulative dataset was generated by using TargetedSim (https://sourceforge.net/projects/targetedsim/files/TargetedSim/, accessed September 10, 2019).

Authors' contributions

YL prepared data, generated all figures and wrote the manuscript. YW and ZZ designed the tool and prepared the performance assessment on ICR96 dataset. ZZ and YL prepared the performance assessment on simulative dataset. TZ supervised the project. YL, YW and ZZ contribute equally to this work. All authors have read and approved the final manuscript.

Author information

Affiliations:

R&D department, Top Gene Tech (Guangzhou) Co., Ltd., Rm H10F, GT Land Winter Plaza,
Zhujiangdong Road, Guangzhou, Guangdong, P.R.China

Ye Liu, Yangming Wu, Zexin Zheng, Tianliangwen Zhou

Corresponding author

Tianliangwen Zhou

R&D department, Top Gene Tech (Guangzhou) Co., Ltd., Rm H10F, GT Land Winter Plaza,
Zhujiangdong Road, Guangzhou, Guangdong, P.R.China

tian.zhou@topgene.cn

Acknowledgements

This study makes use of the ICR96 exon CNV validation series data generated by Professor Nazneen Rahman’s team at The Institute of Cancer Research, London as part of the TGMI.

References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.

2. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution**. *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.

3. Mikhail FM: **Copy number variations and human genetic disease**. *Curr Opin Pediatr* 2014, **26**(6):646-652.

4. Coe BP, Girirajan S, Eichler EE: **The genetic variability and commonality of neurodevelopmental disease**. *Am J Med Genet C Semin Med Genet* 2012, **160C**(2):118-129.

5. Lee JA, Lupski JR: **Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders**. *Neuron* 2006, **52**(1):103-121.

6. Mamlouk S, Childs LH, Aust D, Heim D, Melching F, Oliveira C, Wolf T, Durek P, Schumacher D, Blaker H *et al*: **DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer**. *Nat Commun* 2017, **8**:14093.

7. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nat Rev Genet* 2006, **7**(2):85-97.

8. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping**. *Nat Rev Genet* 2011, **12**(5):363-376.

9. Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, Ledbetter DH, Greenberg F, Patel PI: **Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A**. *Nat Genet* 1992, **1**(1):29-33.

10. Stankiewicz P, Lupski JR: **Structural variation in the human genome and its role in disease**. *Annu Rev Med* 2010, **61**:437-455.

11. Rehm HL: **Disease-targeted sequencing: a cornerstone in the clinic**. *Nat Rev Genet* 2013, **14**(4):295-300.

12. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification**. *Nucleic Acids Res* 2002, **30**(12):e57.

13. Homig-Holzel C, Savola S: **Multiplex ligation-dependent probe amplification (MLPA) in tumor diagnostics and prognostics**. *Diagn Mol Pathol* 2012, **21**(4):189-206.

14. Schenkel LC, Kerkhof J, Stuart A, Reilly J, Eng B, Woodside C, Levstik A, Howlett CJ, Rupar AC, Knoll JHM *et al*: **Clinical Next-Generation Sequencing Pipeline Outperforms a Combined Approach Using Sanger Sequencing and Multiplex Ligation-Dependent Probe Amplification in Targeted Gene Panel Analysis**. *J Mol Diagn* 2016, **18**(5):657-667.

15. Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, Rupar CA, Adams P, Hegele RA, Lin H *et al*: **Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels**. *J Mol Diagn* 2017, **19**(6):905-920.

16. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L *et al*: **Paired-end mapping reveals extensive structural variation in the human genome**. *Science* 2007, **318**(5849):420-426.

17. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: **Sensitive and accurate detection of copy number variants using read depth of coverage**. *Genome Res* 2009, **19**(9):1586-1592.

18. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads**. *Bioinformatics* 2009, **25**(21):2865-2871.

19. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H *et al*: **Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly**. *Nat Biotechnol* 2011, **29**(8):723-730.

20. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, Barillot E: **SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data**. *Bioinformatics* 2010, **26**(15):1895-1896.

21. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M: **Detecting copy number variation with mated short reads**. *Genome Res* 2010, **20**(11):1613-1622.

22. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ *et al*: **A robust model for read count data in exome sequencing experiments and implications for copy number variant calling**. *Bioinformatics* 2012, **28**(21):2747-2754.

23. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S: **Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data**. *Bioinformatics* 2012, **28**(1):40-47.

24. Li Y, Zhang J, Yuan X: **BagGMM: Calling copy number variation by bagging multiple Gaussian mixture models from tumor and matched normal next-generation sequencing data**. *Digital Signal Processing* 2019, **88**:90-100.

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

26. Chiang T, Liu X, Wu TJ, Hu J, Sedlazeck FJ, White S, Schaid D, Andrade M, Jarvik GP, Crosslin D *et al*: **Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel**. *Genet Med* 2019, **21**(9):2135-2144.

27. Pugh TJ, Amr SS, Bowser MJ, Gowrisankar S, Hynes E, Mahanta LM, Rehm HL, Funke B, Lebo MS: **VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data**. *Genet Med* 2016, **18**(7):712-719.

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

29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

31. Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E: **Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization**. *Bioinformatics* 2011, **27**(2):268-269.

32. Boeva V, Popova T, Lienard M, Toffoli S, Kamal M, Le Tourneau C, Gentien D, Servant N, Gestraud P, Rio Frio T *et al*: **Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data**. *Bioinformatics* 2014, **30**(24):3443-3450.

33. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing**. *Nucleic Acids Res* 2012, **40**(10):e72.

34. Talevich E, Shain AH, Botton T, Bastian BC: **CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing**. *PLoS Comput Biol* 2016, **12**(4):e1004873.

35. Roca I, Gonzalez-Castro L, Fernandez H, Couce ML, Fernandez-Marmiesse A: **Free-access copy-number variant detection tools for targeted next-generation sequencing data**. *Mutat Res* 2019, **779**:114-125.

36. de Ligt J, Boone PM, Pfundt R, Vissers LE, Richmond T, Geoghegan J, O'Moore K, de Leeuw N, Shaw C, Brunner HG *et al*: **Detection of clinically relevant copy number variants with whole-exome sequencing**. *Hum Mutat* 2013, **34**(10):1439-1448.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al*: **Scikit-learn: Machine Learning in Python**. *JMLR* 2011, **12**(85):2825–2830.

Figure Legends

Figure 1 Pipeline of T-CNV

To start CNVs assessment, read depth at each locus located with the targeted region was determined. Tumor sample and corresponding normal/control sample were considered as a pair in T-CNV, but the latter was not strictly required.Log2 value was calculated and GC- correction was computed. Any exon with log2 value higher than 0.32 or lower than -0.42 was considered as possible CNV candidates. The CNV candidates were further confirmed by the 3 methods including bins-method, GMM-clustering method and window-slicing method.

Figure 2 Box-plot for Log2 value and Pearson correlation coefficient in ICR96 dataset

(A) Four samples with in pool1 were marked as QC poor with Std_{Exon} >0.2. (B) Five samples in pool2 were marked as QC poor with Std_{Exon} >0.2. (C) Pearson correlation coefficient of each sample and its corresponding pool control in ICR96 dataset. QC poor samples are indicated with red font.

Figure 3 Performance comparison with DECoN, Atlas-CNV and VisCap on ICR96 dataset

(A) T-CNV, DECoN, Atlas-CNV and VisCap were used to detect CNVs on ICR96 dataset and their performance were presented using four parameters: sensitivity, specificity, positive predict value (PPV) and negative predict value (NPV). (B) Four tools performance on deletion CNVs. (C) Four tools performance on duplication CNVs.

Figure 4 Performance assessment of T-CNV and DECoN on simulative dataset

(A) Performance of T-CNV on simulative dataset under 50X, 100X and 500X was presented as sensitivity and PPV. "all" represented all simulative CNVs. "del" represented the simulative deletions and "dup" represented the simulative duplications. (B) Performance of DECoN on simulative dataset.

Figure 5 Example of false positive and false negative prediction in ICR96 dataset

(A) A false negative CHEK2 exon 13 in sample 17332 was identified by T-CNV. (B) Four false negative exons located in NF1 gene in 17338 sample were identified by T-CNV. (C) A false positive BAP1 exon 1 in sample 17340 was identified by T-CNV. (D) A false positive CDH1 exon 1 in sample 17301 was identified by T-CNV.

Additional file 1: Supplementary figures. Supplementary Figure S1 LOESS used in GC-content correction. Supplementary Figure S2 Random noise test. Supplementary Figure S3 Optimazation of bins method. Supplementary Figure S4 An example of GMM clustering result on sample 17338 NF1 exon 37-57 delietion. Supplementary Figure S5 Sampling plan for GMM clustering in T-CNV. Supplementary Figure S6 True positive and false positive in T-CNV window-sliding method. Supplementary Figure S7 ROC curve and PR curve for optimization of window-sliding method.

Supplementary Figure S8 The visulization result of sample 17375 from pool2 in ICR96 dataset. (DOCX 2.5MB)

Figures

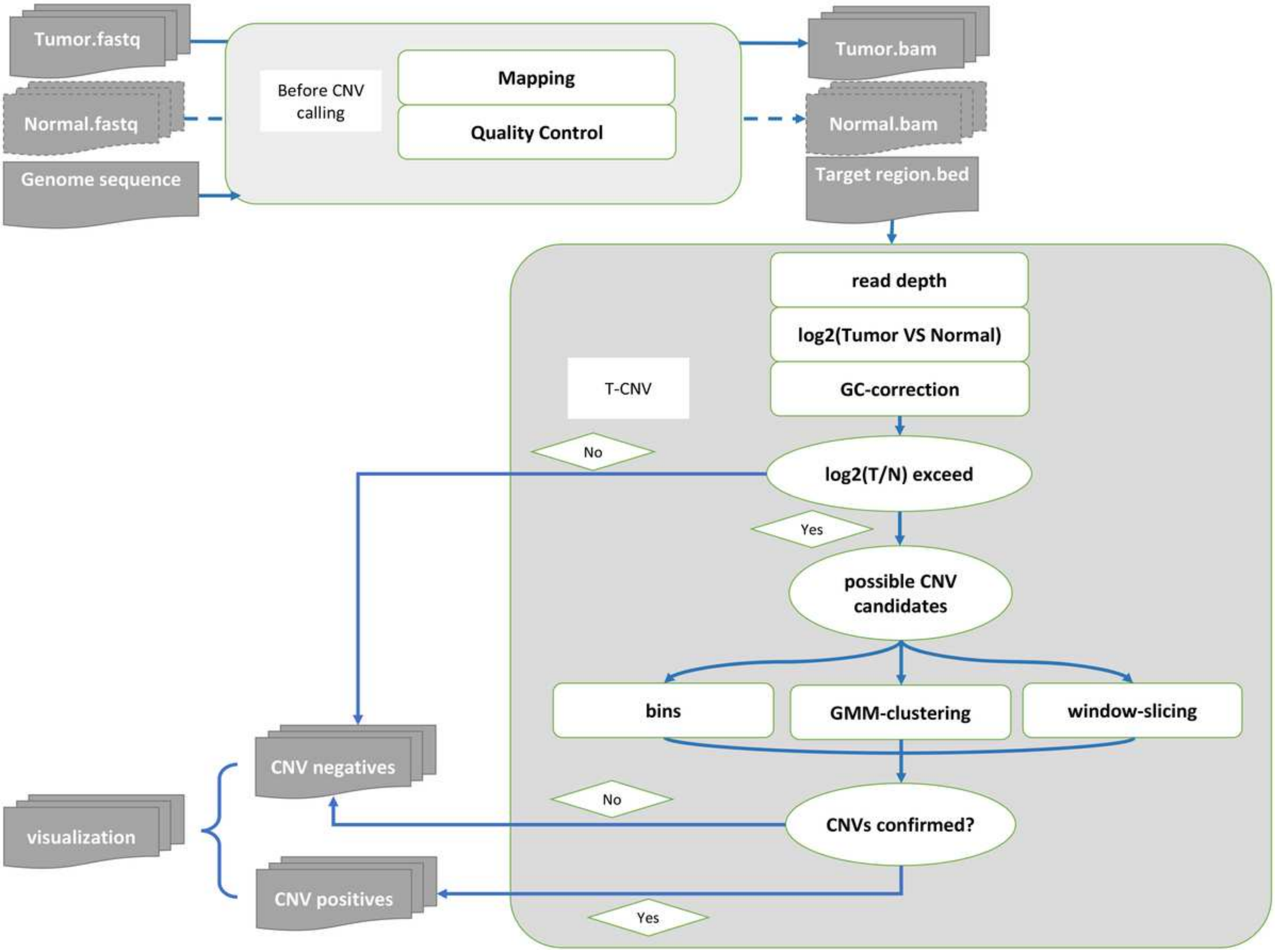

Figure 1

Pipeline of T-CNV To start CNVs assessment, read depth at each locus located with the targeted region was determined. Tumor sample and corresponding normal/control sample were considered as a pair in T-CNV, but the latter was not strictly required.Log2 value was calculated and GC- correction was computed. Any exon with log2 value higher than 0.32 or lower than -0.42 was considered as possible CNV candidates. The CNV candidates were further confirmed by the 3 methods including bins-method, GMM-clustering method and window-slicing method.

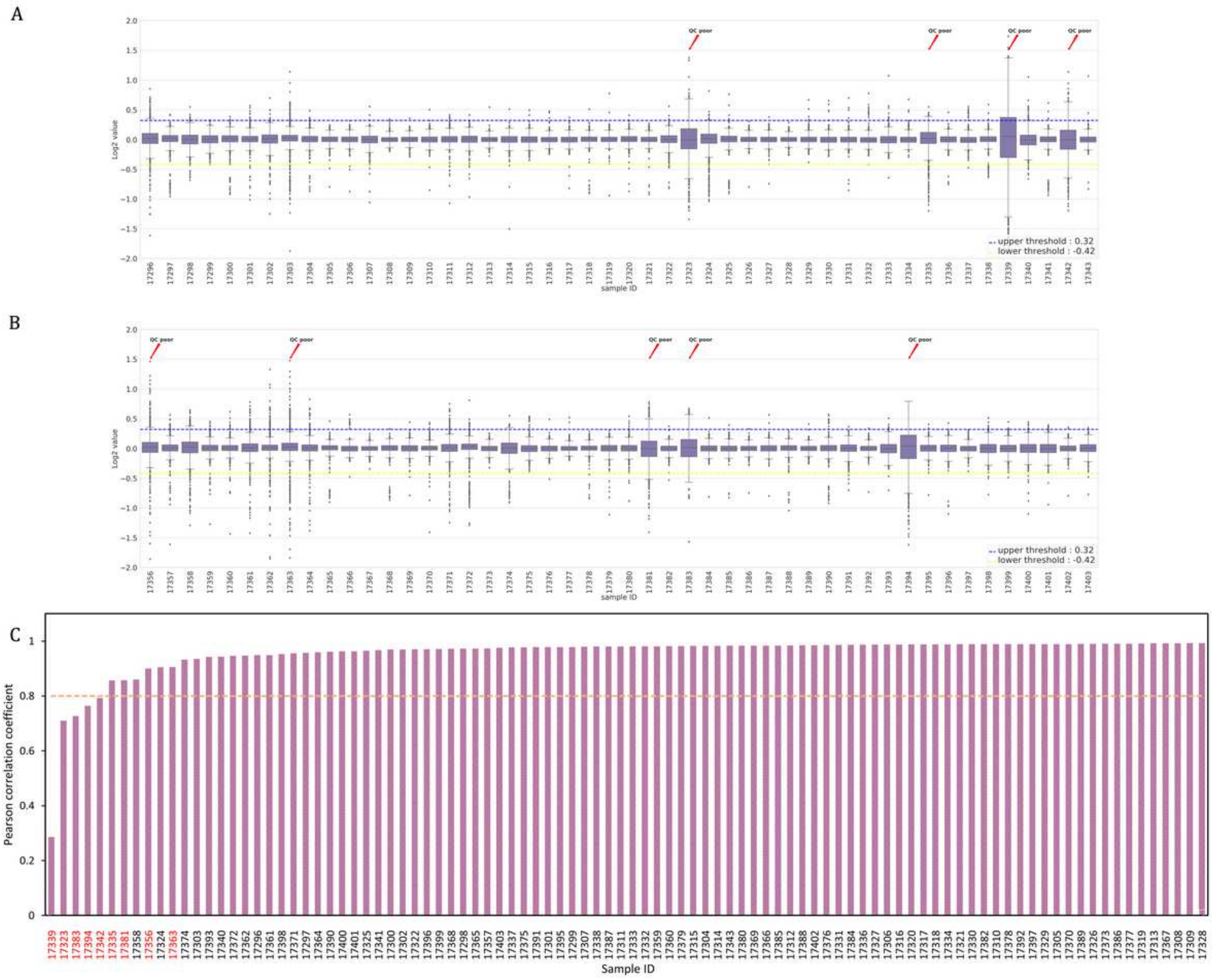

Figure 2

Box-plot for Log2 value and Pearson correlation coefficient in ICR96 dataset (A) Four samples with in pool1 were marked as QC poor with □Std□_Exon >0.2. (B) Five samples in pool2 were marked as QC poor with □Std□_Exon >0.2. (C) Pearson correlation coefficient of each sample and its corresponding pool control in ICR96 dataset. QC poor samples are indicated with red font.

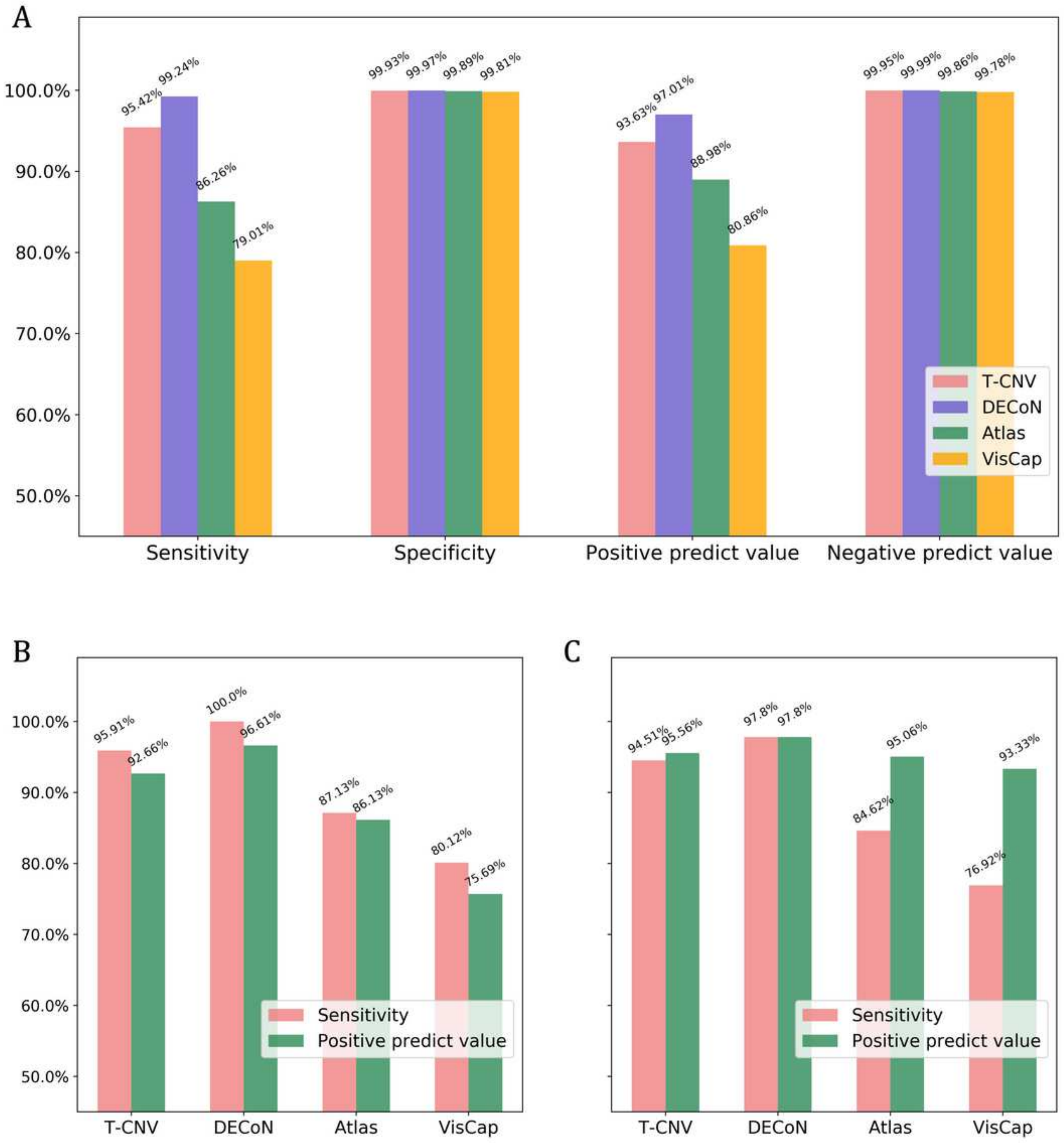

Figure 3

Performance comparison with DECoN, Atlas-CNV and VisCap on ICR96 dataset (A) T-CNV, DECoN, Atlas-CNV and VisCap were used to detect CNVs on ICR96 dataset and their performance were presented using four parameters: sensitivity, specificity, positive predict value (PPV) and negative predict value (NPV). (B) Four tools performance on deletion CNVs. (C) Four tools performance on duplication CNVs.

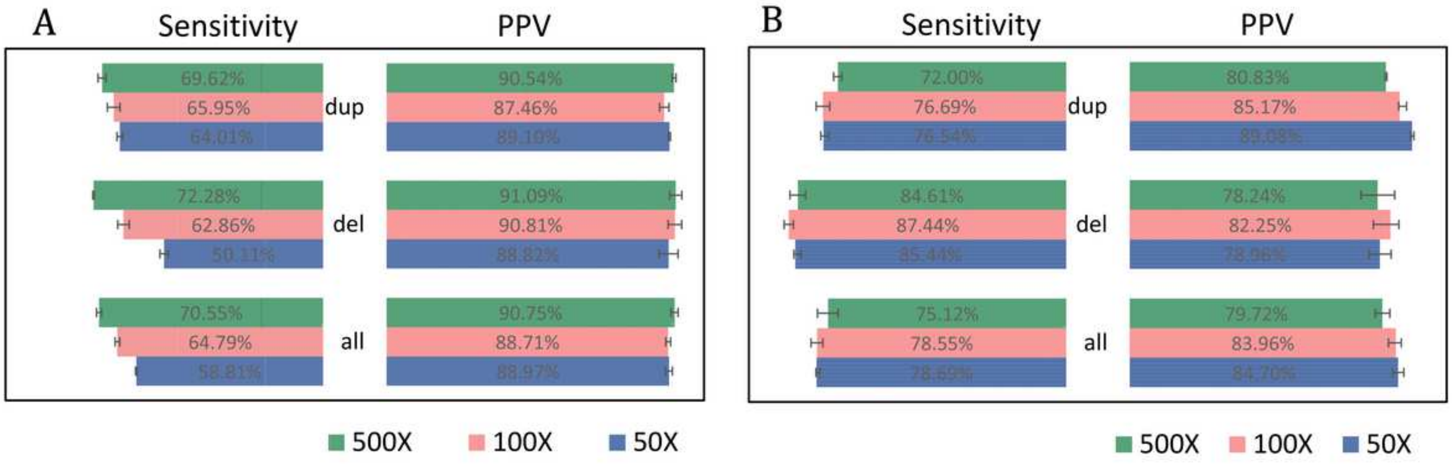

Figure 4

Performance assessment of T-CNV and DECoN on simulative dataset (A) Performance of T-CNV on simulative dataset under 50X, 100X and 500X was presented as sensitivity and PPV. “all” represented all simulative CNVs. “del” represented the simulative deletions and “dup” represented the simulative duplications. (B) Performance of DECoN on simulative dataset.

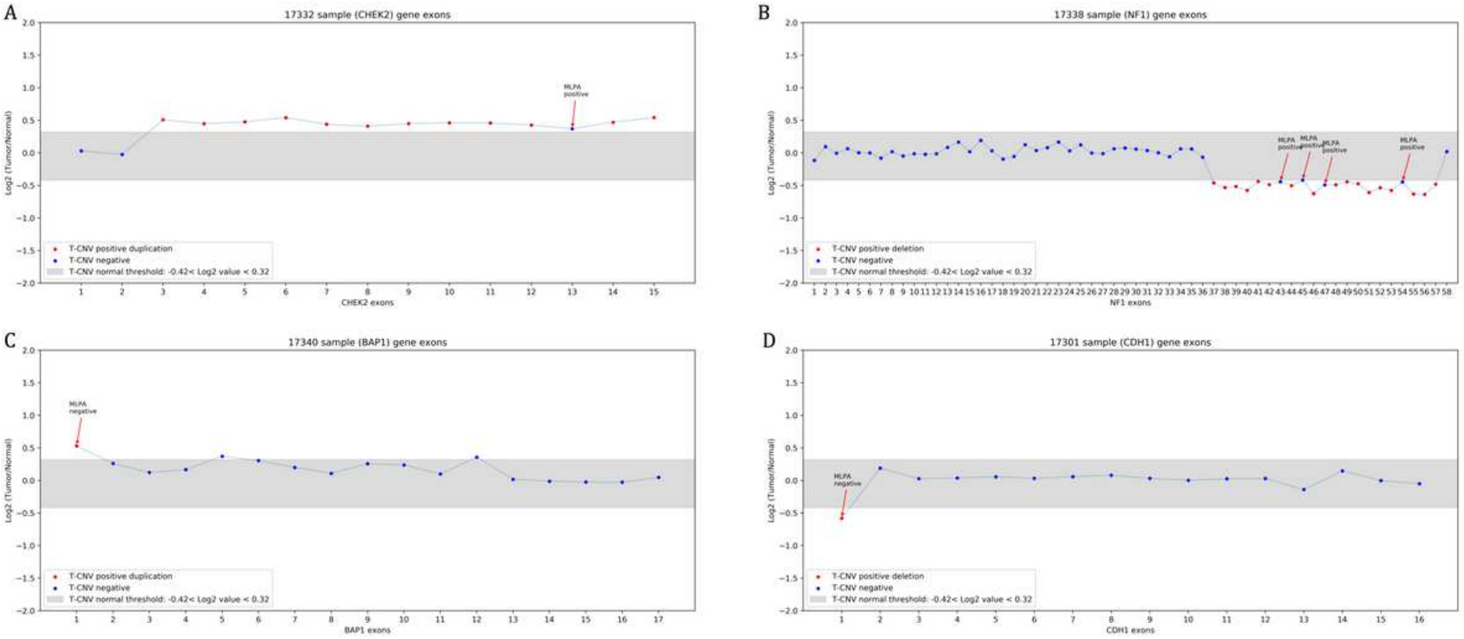

Figure 5

Example of false positive and false negative prediction in ICR96 dataset (A) A false negative CHEK2 exon 13 in sample 17332 was identified by T-CNV. (B) Four false negative exons located in NF1 gene in 17338 sample were identified by T-CNV. (C) A false positive BAP1 exon 1 in sample 17340 was identified by T-CNV. (D) A false positive CDH1 exon 1 in sample 17301 was identified by T-CNV.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryfigureS3bincomparePSupdate.png
- SupplementaryfigureS1LOESSupdate.png
- SupplementaryfigureS4GMMPSupdate.png
- SupplementaryfigureS6wintpfp.png
- Addtionalfile1.docx
- SupplementaryfigureS2noisetestPS.png
- SupplementaryfigureS817375.png
- SupplementaryfigureS7windowsliding.png
- SupplementaryfigureS5GMMsamplingPS.png