Briefings in Bioinformatics, 2026, 27, bbag100
https://doi.org/10.1093/bib/bbag100
Published: 12 March 2026
Problem Solving Protocol

OXFORD

ML-ExonCNV: a robust XGBoost multi-expert ensemble framework for rare exon CNV detection in whole-exome sequencing data

Shuang-Hao Yang[1,2,3], Hua He[1], Shuyu Hou[1], Tuanfeng Yang[4], Zehao Yin[1], Hong-Yu Zhang[3,*], Weiyue Gu[1,2,*]

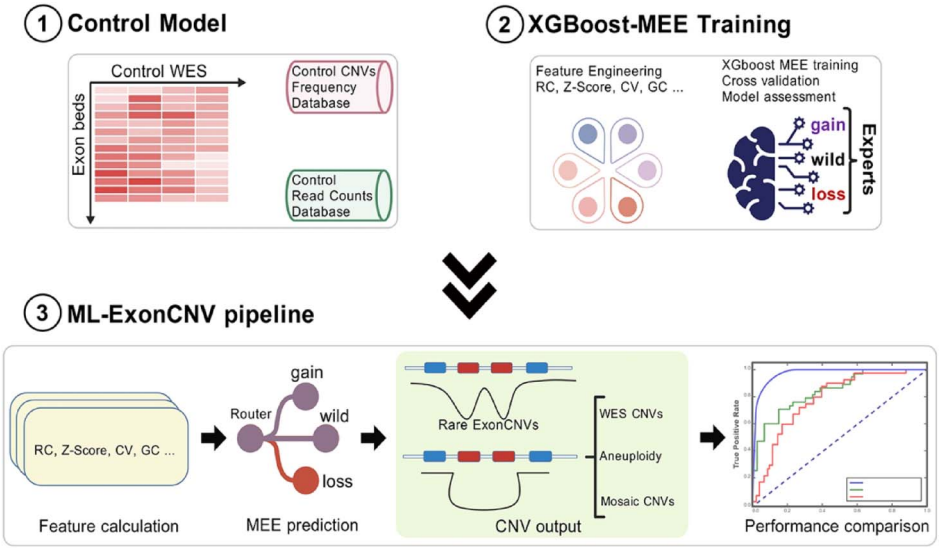
[1]Department of Bioinformatics, Chigene (Beijing) Translational Medical Research Center Co., Ltd., Beijing Yizhuang Biomedical Park, 100176, China
[2]Department of Bioinformatics, Beijing Quanpu Medical Laboratory Co., Ltd., E2, 3rd Floor, No. 88 Kechuang 6th Road, Beijing Yizhuang Biomedical Park, 100176, China
[3]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[4]Department of Neurology, Peking University International Hospital, No. 1 Life Park Road, Changping District, Beijing 102206, China

*Corresponding authors. Hong-Yu Zhang, Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. E-mail: zhy630@mail.hzau.edu.cn; Weiyue Gu, CEO, Chigene (Beijing) Translational Medical Research Center Co., Ltd., Beijing Yizhuang Biomedical Park, 100176, China. E-mail: gwyue@163.com

Abstract

Copy number variants (CNVs) have been shown to play a significant role in the pathogenesis of various human diseases. Although several tools have been developed for detecting CNVs based on whole-exome sequencing (WES) data, their performance remains suboptimal for small exon-level CNVs (exCNVs). This is primarily due to multiple technical variabilities, including probe capture efficiency, mappability, exon size, batch effects, experimental background noise bias, and control sample selection, all of which can lead to false negatives and false positives in exCNV detection. To address these challenges, we developed ML-ExonCNV, which innovatively integrates the XGBoost machine learning model with a multi-expert ensemble approach. The model was trained using 14 features derived from 22 364 real-world, quantitative polymerase chain reaction-validated rare exCNVs. Evaluation on a test set of 492 real WES and the NA12878 gold-standard dataset demonstrated that ML-ExonCNV outperformed widely used tools such as GATK-gCNV, ExomeDepth, and CNVkit. Notably, ML-ExonCNV can detect large segmental CNVs, mosaic CNVs, and breakpoint CNVs on exon region. Furthermore, our analysis revealed recurrent exCNV-associated genes and their phenotypic correlations. Neurodevelopmental and musculoskeletal abnormalities were identified as the most frequently associated phenotypes with high-recurrence exCNVs.

Graphical abstract



Keywords WES CNVs, exon-level CNVs, machine learning, XGBoost

Received: October 2, 2025. **Revised:** January 17, 2026. **Accepted:** February 7, 2026

Introduction

Copy number variations (CNVs) were initially defined as ranging from 1000 base pairs to several megabases, and later expanded to include minor variations larger than 50 base pairs [1]. Half of CNVs overlap with protein-coding regions, leading to changes in gene expression [2], and CNVs have been proven to play a significant role in human diseases. Remarkably, half of the pathogenic CNVs are associated with genetic diseases caused by exon deletions or duplications (small CNVs) [3].

The detection of CNVs has traditionally depended on low-resolution techniques such as chromosomal microarray (CMA), which provides a diagnostic yield of 15%–20% for neurodevelopmental disorders [4]. However, CMA's low resolution precludes accurate detection of small CNVs (50 bp–100K bp) and renders it incapable of detecting single nucleotide variants (SNVs) and Indels.

In contrast, whole-exome sequencing (WES) has been widely used in clinical and basic genetic research to detect SNVs/Indels. Despite comprising merely ∼1% of the genome, it is enriched with causal variants [5]. Previous studies have shown that the yield of exome sequencing (ES) was 36% overall, 31% for isolated neurodevelopmental disorder (NDD), and 53% for NDD plus associated conditions. The ES yield for NDDs is significantly higher than that of CMA (15%–20%) [6], and it has become a first-tier screen for coding short variants. In fact, WES targets 20 000 protein-coding genes and over 200 000 exon target regions, providing the potential for WES-CNV detection.

Currently, tools for detecting CNVs based on WES data have been developed, including ExomeDepth [7], GATK-gCNV [1], CNVkit [8], and ECOLE [9], among others. Most of these tools can reliably call large CNVs, but they perform poorly when dealing with smaller CNVs that affect only one or a few exons. These small exon-level CNVs (exCNVs), especially those under negative selection and rare CNVs, are often associated with various genetic diseases [10]. Traditional CNV detection methods mainly rely on read count (RC) [11], split read, or paired-end mapping to infer CNVs [12]. These methods have achieved certain results on standard datasets (such as NA12878), simulated data, and a small number of validated CNVs as test datasets [10]. However, their performance is difficult to reproduce when processing large amounts of real-world data, especially in accurately detecting rare CNV variants. It is known that variations in GC content can affect exon capture [13], polymerase chain reaction (PCR), and sequencing efficiency, ultimately strongly distorting the distribution of RCs [5]. Additionally, various technical variabilities, such as mappability, exon size, flanking exons, batch effects, bias of experimental background noise, selection of control samples, and the coefficient of variation (CV) of control samples, can all lead to false negatives and false positives in the detection of exCNVs. For example, the latest GATK-gCNV software has achieved an overall recall rate of 95%, but its precision is very low (22%) [1]. Therefore, although detecting CNVs using WES is technically feasible, accurately identifying exCNVs remains a significant challenge.

Polymorphic CNVs frequently occur in populations and are associated with low pathogenicity, thus having limited clinical significance in genetics and oncology. High-frequency CNVs, background noise, and batch effects share a common characteristic: the frequent occurrence of the same CNV segment [14]. In contrast, the identification of rare CNVs is of great significance for diagnosing genetic diseases and elucidating tumor abnormalities. Based on the aforementioned principles, this study proposes the ML-ExonCNV method to enhance the accuracy of detecting exCNVs from WES data. This method innovatively employs the XGBoost machine learning model [15] in combination with the multi-expert ensemble (MEE) approach [16]. It involves pretraining and modeling with 14 features on 22 364 real-world rare exCNVs validated by quantitative polymerase chain reaction (qPCR). The performance of the software is ensured through testing on a real-world independent WES dataset (R492) comprising 492 samples and the NA12878 gold standard dataset. The algorithm integrates GC correction [17], searching for the most similar control samples, background noise and population frequency filtering of rare CNVs, breakpoint CNV identification by Manta [18], and VCF variant support (theoretically, variants in deletion regions are hemizygous) to provide reliability grading for exCNVs (highly reliable, relatively reliable, unreliable). This significantly improves the sensitivity (95.3%) and specificity (97.6%) of exCNV detection. It also overcomes the limitations of previously mentioned software, which are unable to lock breakpoints (missed detection of exCNV breakpoints in coding regions) and lack the capability for calling mosaic CNVs. Moreover, the ML-ExonCNV software can call large segmental CNVs and chromosomal number abnormalities. Additionally, it achieves 3–22 times higher analytical efficiency compared to other software, thereby greatly enhancing diagnostic efficiency. Furthermore, through statistical analysis of real-world rare CNVs (training set), we have for the first time revealed that genes with high-recurrence exCNVs are closely related to neurodevelopmental abnormalities. In summary, our research provides novel insights into how to accurately call exCNVs from WES data.

Materials and methods

Training and test set

The training dataset consisted of 143 882 CNV-positive and 130 877 wild-type target bed labels (from 22 364 real-world exon-level CNVs) from Chigene (Beijing) Translational Medical Research Center. The independent test set (R492) comprised 492 WES samples provided by Beijing Quanpu Medical Laboratory (details in Supplementary Text).

The NA12878 dataset was generated from the Coriell Institute cell line, captured using IDT v2 probes, and sequenced with the DNBSEQ-T7 platform to produce WES data. The gold standard variant set for NA12878 included 3512 CNVs reported in previous studies [19], of which 77 CNVs intersecting with the coding regions were selected to test the performance of exCNV detection.

All research data were obtained from retrospective studies and did not involve any personal privacy data of patients. Written consent was obtained from the patients or their legal guardians, except for the NA12878 public test set. All investigations were conducted in accordance with the tenets of the Declaration of Helsinki.

Virtual target beds

Given that different capture probes may be used in WES sequencing but the target regions are invariably protein-coding areas, this study referred to the functional transcripts from the MANE Select database. The CDS regions of these transcripts were obtained from the gene annotation of GRCh37.p13 (105.20220307) and GRCh38.p14 (GCF_000001405.40-RS_2025_08). Therefore, the software supports BAM input for both hg19 and hg38 versions. The CDS intervals longer than 200 bp were segmented into 200-bp lengths. If a CDS interval was shorter than 200 bp, it was extended to 200 bp to generate synthetic

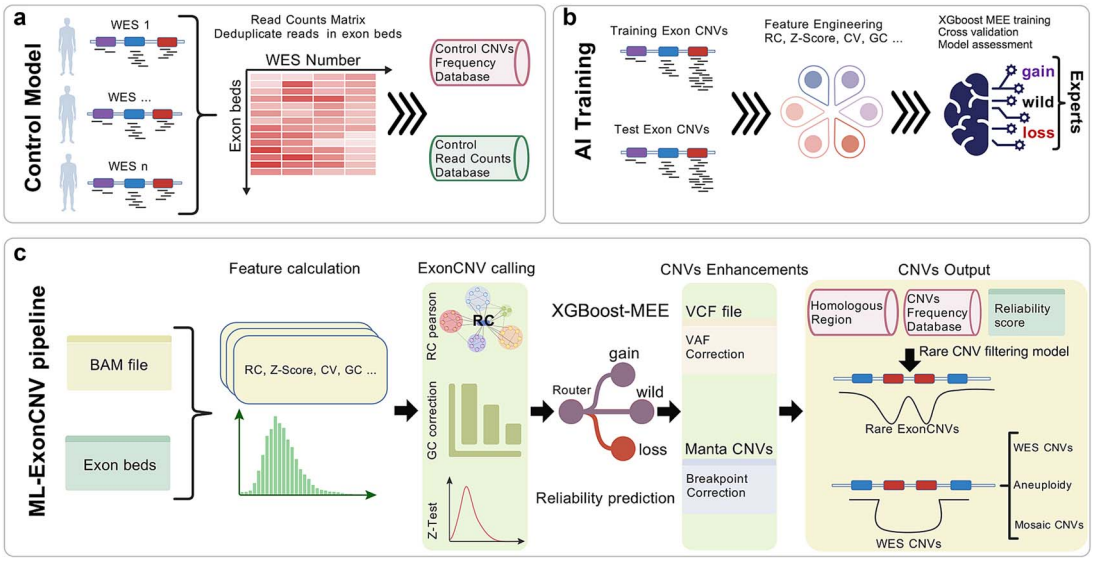


Figure 1 The ML-ExonCNV XGBoost MEE methodological framework. (a) Based on the deduplicated BAM alignment results of multiple WES sequencing samples in the control group, an RCs matrix for the virtual target beds of the control group was constructed, along with an in-house CNV frequency database. (b) Utilizing a 5-fold cross-validation method to dynamically divide the data into training and testing sets, 14 key features were calculated. The XGBoost and MEE algorithms were employed for grid search to identify the optimal hyperparameters, resulting in the training of three pretrained models for gain, wild, and loss. (c) The full ML-ExonCNV CNV calling process includes the following: inputting the BAM and BED files of the test sample, calculating the 14 feature values for each virtual target bed in both the test sample and the control library, selecting the most likely control samples from the control library based on RC values and Pearson correlation analysis, performing GC correction (loess regression algorithm) and Z-test calculations on the RC values of each virtual target bed, inputting all feature values into the pretrained XGBoost MEE model (including three expert models) to obtain the reliability of exCNVs, fine-tuning the reliability of exCNVs using VCF variants and Manta breakpoint information, and outputting rare exon CNVs after passing through the rare CNV filtering model. Adjacent exon CNVs are merged to output large segmental CNVs.

CNV target regions (virtual target beds, BED file format). This approach enhances the software's compatibility with various WES sequencing.

Construction of the read count matrix and copy number variant frequency database

The ML-ExonCNV software's control-build module can construct a multisample RC matrix based on the deduplicated BAM files (Fig. 1a). The input includes a list of control sample information (sample name, BAM path, and sex). For each sample's BAM file, the RC values for each virtual target bed are calculated using mosdepth (version 0.2.6) [20]. The in-house CNV frequency database is constructed by performing CNV calling on each control sample using ML-ExonCNV. This frequency database includes high-frequency CNVs in the population, as well as false-positive CNVs introduced by batch effects or experimental biases (Fig. 1a).

Features engineering and model training

Improving the reliability is a major challenge in detecting exCNVs through WES. To address this, we analyzed various intrasample and intersample factors that influence CNV reliability and extracted 14 key core features (Supplementary Text).

Based on the qPCR-validated CNV labels, CNVs that pass qPCR validation are flagged as positive labels, while those that do not are flagged as negative labels. We employed an MEE model combined with the XGBoost gradient-boosted decision tree algorithm for model training (Fig. 1b). According to the MEE principle and the ablation analysis of threshold selection, the training dataset was split into three expert subsets based on fold-change cutoffs of 0.75 and 1.25. These thresholds were chosen because they yield a near-optimal F1 score (Supplementary Table S10) while retaining biological interpretability in the context of CNV analysis: CNVs with a fold change (FC) less than 0.75 were assigned to the loss/del expert model dataset, those with an FC between 0.75 and 1.25 were assigned to the wild expert model dataset, and those with an FC greater than 1.25 were assigned to the gain/dup expert model dataset. Separate expert models were trained for loss, wild, and gain scenarios.

The hyperparameter settings for the XGBoost model training process were as follows: max_depth = {5, 6, 7, 8, 9, 10}, learning_rate = {0.01, 0.001}, gamma = {0.1, 0.2}, n_estimators = {300, 400, 500, 600}, subsample = {0.6, 0.7, 0.8, 0.9}, colsample_bytree = {0.6, 0.7, 0.8, 0.9}, and L2 regularization weight (reg_lambda) = {0.5, 1.0}. Hyperparameter optimization was performed using grid search and 5-fold cross-validation, with the parameter combination yielding the highest AUC selected as the optimal set. Ultimately, three expert models were generated after training. During the performance testing phase and the CNV calling process, new target intervals were routed based on FC value into the corresponding expert models for prediction. The final prediction results were then aggregated (Fig. 1c).

Technical variability between the samples

Algorithms for CNV detection from WES data based on RC are influenced by both intrinsic and extrinsic factors. Intrinsic factors encompass library preparation, capture probes, PCR processes, sequencing processes, genomic repeats, and GC content, among others. Extrinsic factors include control sample selection and CV values. Due to technical variations between samples, clustering analysis of RC for the R492 sample set reveals that identifying control samples with RC models more similar to those of the test samples determines the

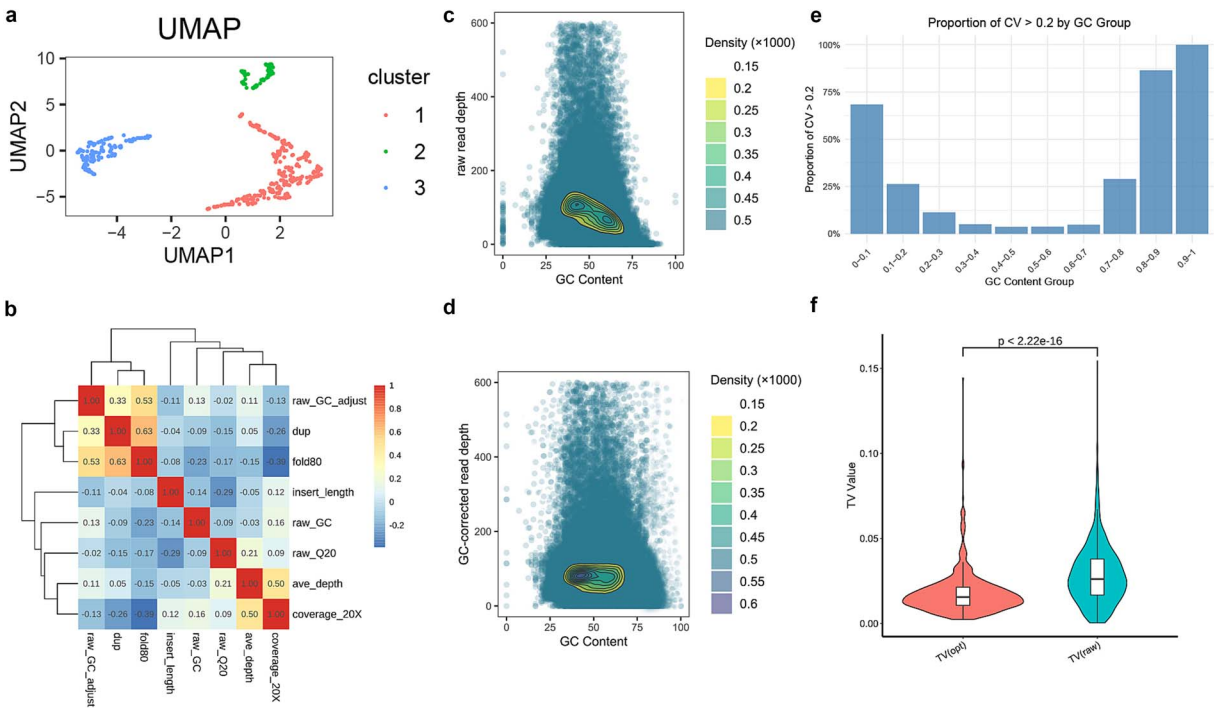


Figure 2 TV optimization of CNV calling noise. (a) Based on RCs from all target regions across 492 test samples, PCA was performed to reduce dimensionality. The top 50 principal components were selected for UMAP clustering, and K-means clustering was applied to identify three distinct clusters. Each point represents a test sample, the clusters are located on the left, upper right, and lower right of the figure. Clustering analysis revealed a clear batch effect among WES datasets. Selecting control samples with similar RC profiles can improve the accuracy of exCNV detection. (b) Pearson correlation analysis of various quality parameters in the raw data showed that PCR duplication (dup) and GC content have the highest correlation with Fold80. Duplication and GC content were identified as key experimental factors affecting uniformity. (c) The *x*-axis represents the GC content of virtual target beds, and the *y*-axis represents the read depth of each target region at that GC content. The figure shows the relationship between the original read depth and GC content for a randomly selected test sample (from R492). (d) After loess regression correction (see Supplementary Text), the relationship between read depth and GC content shows a more balanced distribution in the high GC and low GC regions compared to the original. (e) The proportion of virtual beds with a CV > 0.2 was calculated for each group of virtual target beds, categorized by GC content. The CV values were found to be higher in high and low GC regions. (f) The distribution of TV (raw) values, representing TV before optimization, is compared with the distribution of TV (opt) values after optimization for the 492 test samples. A paired *t*-test analysis revealed a significant difference brought about by the optimization.

reliability of CNVs (Fig. 2a). These dual factors severely distort the linear relationship between RC and true copy numbers.

Fold80 (defined as mean sequencing depth divided by the 20th percentile depth) is a key metric for assessing intra-sample sequencing uniformity [21]. Higher Fold80 values indicate poorer depth uniformity across exons (i.e. greater data variability). Correlation analysis of raw data quality parameters (Fig. 2b) revealed that duplication rate (dup) and GC content showed the strongest associations with Fold80. This suggests that performing read deduplication followed by GC bias correction can effectively reduce technical variability (TV). After GC correction, read distributions in both high- and low-GC regions became more balanced, approaching the mean coverage level (Fig. 2c–d).

GC content and intersample CV distribution analysis revealed significantly higher CV values in both high-GC and low-GC regions (Fig. 2e). Elevated CV values indicate greater instability in RCs across samples for the same capture region, leading to increased false negatives or false positives in CNV detection. Therefore, selecting highly sensitive genomic intervals (beds) for intersample variability assessment is critical. In this study, we defined CV-GC selected beds using the following criteria: CV > 0.3, or (GC < 0.4 & CV > 0.2), or (GC > 0.7 & CV > 0.2) (Supplementary Fig. S1A). These beds comprehensively represent regions with extreme GC content and high CV values. Subsequently, we performed Pearson correlation analysis between test samples and reference library samples using both all beds and CV-GC selected beds. Statistical analysis of Pearson distributions demonstrated that CV-GC selected beds yielded significantly higher correlation coefficients than all beds (Supplementary Fig. S1B), thereby enhancing the accuracy of CNV detection.

To quantify the combined impact of intrinsic and extrinsic factors, we proposed a novel metric: TV between samples, defined as the ratio of outlier beds in CV-GC selected beds (TV = O/A). Here, O represents the number of fold-change outlier beds (test sample FC versus control group Z-score > 1.96), and A denotes the total number of CV-GC selected beds. The TV value quantifies the TV during CNV calling for each sample. Higher TV values indicate more outlier beds in variability-sensitive regions, reflecting greater technical interference in the analytical pipeline and a higher likelihood of false-positive CNVs. Further analysis revealed that TV is a key feature in our training process, ranking as the second most important feature (Fig. 3f), our study proposes that TV optimization can reduce the noise in exCNV calling.

ML-ExonCNV calling

The strength of the ML-ExonCNVs algorithm lies not only in its ability to detect highly reliable CNVs but also in its provision of reliability grading for each CNV (highly reliable, relatively reliable,

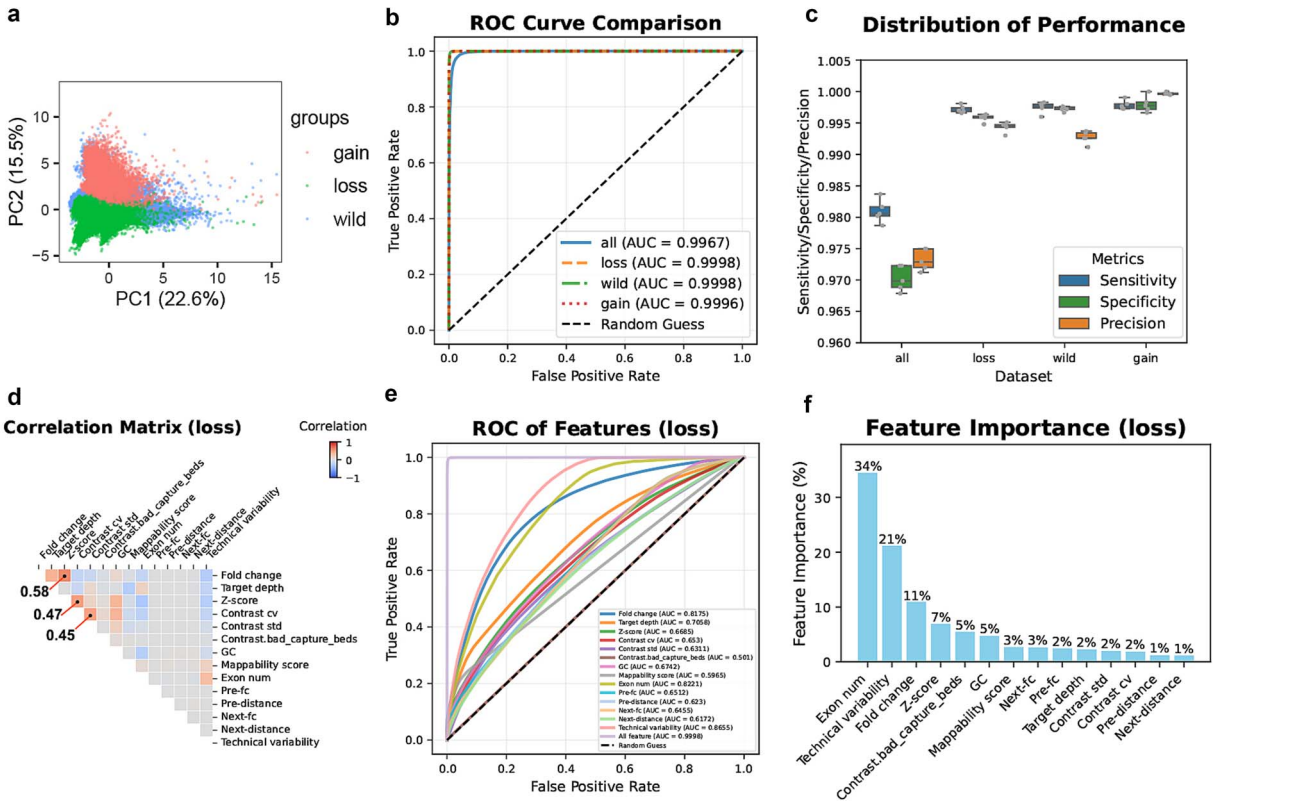


Figure 3 XGBoost MEE model training. (a) PCA was performed on the 14 features of the training dataset, with each point representing a virtual target bed. The gain beds are located in the upper part of the figure, wild beds are in the middle, and loss beds are in the lower part, showing clear stratification. (b) XGBoost training was conducted separately using all virtual target beds, gain beds, wild beds, and loss beds. The ROC curves for each model show that the all-model performs weaker than the other models. (c) Box plots of sensitivity, specificity, and precision for the all, loss, wild, and gain models trained with 5-fold cross-validation. (d) Taking the loss model as an example, Pearson correlation analysis among the 14 features revealed the highest correlation coefficient of 0.58, indicating low correlation among features, with both positive and negative correlations present. (e) In the loss model, comparison of the ROC curves for the combined performance of the 14 features versus each individual feature. (f) Ranking of the importance of the 14 features in the loss model, with exon num (exon counts) and TV being the most important model features.

unreliable). The core innovative algorithms and steps include (Fig. 1c): The pipeline begins with a BAM file and a BED file as input, calculation of 14 feature values for the test sample: Target depth, Contrast.bad_capture_beds, FC, contrast cv, contrast std, Z-score, GC, mappability score, exon num, pre-fc, pre-distance, next-fc, next-distance, TV (see Supplementary: Features for machine learning), automatic selection of at least 10 control samples with the most similar RC model to the test sample using the Pearson algorithm, GC correction algorithm to reduce the impact of GC bias on RC, calculation of the significance of each virtual target CNV using the Z-test, input of the 14 feature values into the pretrained mixture-of-experts model to predict the reliability score of the virtual target, and fine-tuning of the reliability classification of exCNVs based on the allele frequency (AF) of the exCNV regions in the test sample and breakpoint support from Manta CNV (Fig. 1c and Supplementary: ML-ExonCNV calling details). Using Rare CNV filtering model, exCNVs are filtered based on genomic homologous regions [22], CNV frequency database, and reliability classification (retaining reliable or low-frequency exCNVs) (see Supplementary: Rare CNV filtering model), and adjacent exCNVs are merged to enable our software to call large segmental CNVs, chromosomal number abnormalities, and mosaic CNVs (see Supplementary: Large segmental WES CNV calling). For more detailed methodological processes, see the Supplementary Text and the complete source code has been released on GitHub.

Results

Reduction of technical variability

Algorithms for CNV detection based on RCs are influenced by both intrinsic and extrinsic factors (Fig. 3d). This study innovatively introduces the TV value, which represents the combined impact of intrinsic and extrinsic factors on CNV calling. In the CNV calling algorithm, optimizations for intrinsic factors include: virtual target bed refinement, RCs deduplication, and GC bias correction, among others. Optimizations for extrinsic factors include: CV-GC selected beds and selection of the most similar control samples using the Pearson coefficient (Pearson correlation outperforms Spearman, details are provided in Supplementary Fig. S5), among others. Statistical analysis of the TV value distribution before and after TV algorithm optimization across 492 test samples revealed that the optimization in ML-ExonCNV significantly reduced the impact of TV (Fig. 2F).

Analysis of the Pearson coefficient and the TV for the 492 test samples revealed that the higher the Pearson coefficient between the test sample and the control sample, the more similar the selected

control sample is to the test sample. This indicates that the sample is less affected by TV (i.e. exCNV detection is more accurate) (Supplementary Fig. S1C).

XGBoost multi-expert ensemble model

Principal component analysis (PCA) was performed on the 14 features of the training dataset, and the two most critical principal components (PC1 and PC2) were extracted to plot all training data. We observed distinct stratification among loss/del, gain/dup, and wild labels (Fig. 3a). This finding inspired the introduction of an MEE model in our machine learning approach (Fig. 1). Separate XGBoost models were trained for gain, loss, and wild scenarios, achieving AUC values of 0.9996, 0.9998, and 0.9998, respectively. In contrast, a single model trained on all datasets (all sets) achieved an AUC of only 0.9967 (Fig. 3b). Comparing sensitivity, specificity, and precision during the 5-fold training process revealed that the performance of the three expert models was significantly higher than that of the single model (Fig. 3c). The SHAP plots provide a more detailed interpretation of the features and further corroborate this conclusion (Supplementary Fig. S6 A–D). Thus, ML-ExonCNV innovatively employs an XGBoost-based MEE model, which outperforms a single XGBoost model.

Taking the 14 features of the loss expert model as an example, these include eight intrinsic features characterizing CNV reliability and six extrinsic features from the control analysis process (Fig. 3d). Pearson correlation analysis revealed that the most highly correlated features were Z-score and FC, with a correlation coefficient of 0.58. This indicates low feature correlation, which contributes to model robustness. Excessive redundancy of highly similar features can increase model complexity and compromise accuracy. Additionally, the presence of both positively and negatively correlated features provides richer information, enhancing model performance. Furthermore, comparison of the AUC values for all features versus individual features demonstrated that the performance of the all-feature model far exceeded that of any single feature (Fig. 3e). Analysis of feature importance in the trained model revealed that exon number had the highest importance (34%), followed by TV, FC, Z-score, bad-capture beds, and GC content (Fig. 3f). This analysis underscores the critical importance of TV and GC bias correction in exCNV analysis. Feature figures for the other expert models (gain, wild) are provided in the supplementary figures (Figs. S2–S4).

ML-ExonCNV performance

Performance comparison on the R492 dataset

A real-world independent dataset comprising 492 WES samples (R492), containing 551 rare exCNVs validated by qPCR, was utilized as a gold standard dataset to compare the performance of various software tools, including: ML-calling step (CNVs before rare CNV filtering model), ML-ExonCNV software (after rare CNV filtering model), ExomeDepth, GATK-gCNV, and CNVkit. The comparison revealed that the ML-ExonCNV algorithm achieved the highest sensitivity (95.3%) and specificity (97.6%) among all tools on the R492 dataset (Fig. 4a–b, Table 1), demonstrating superior performance in both metrics. Other tools struggled to achieve high performance in both sensitivity and specificity simultaneously (Supplementary Table S2). GATK-gCNV exhibited good specificity (94%) but lower sensitivity (85.7%), attributed to its built-in CNV filtering steps. ExomeDepth had high sensitivity (94%) but lower specificity (80.4%), due to the lack of false-positive filtering. CNVkit showed poor sensitivity (21.9%) but acceptable specificity (88.4%). Previous studies have also found that CNVkit has poor performance [23], suggesting that CNVkit is not recommended for clinical CNV research. Analysis of sensitivity and specificity across different numbers of exons revealed that as the number of exons increased, detection performance improved for all tools, except for ExomeDepth, whose specificity decreased. This suggests that ExomeDepth may have limitations in detecting larger CNVs (Fig. 4c–d).

The key to ML-ExonCNV's excellent performance lies in its two-step approach. The initial ML-calling step achieved high sensitivity (96.3%) but lower specificity (88%). The subsequent rare CNV filtering model further filtered false positives and population-high-frequency CNVs, enhancing specificity to 97.6% and precision to 98% (Table 1). This dual-step process ensures that the ML-calling step meets the high-sensitivity requirements for scientific research, while the rare CNV filtering model satisfies the high-specificity demands for clinical diagnostics, thereby improving clinical decision-making efficiency and reducing validation time and costs. Additionally, ML-ExonCNV's high specificity is reflected in the minimal number of reported exCNVs per sample. On average, the ML-calling step reported 197.7 exCNVs per sample, compared to 761.4 and 264.7 exCNVs per sample for ExomeDepth and GATK-gCNV, respectively. After filtering, ML-ExonCNV reported an average of only 17.4 clinically relevant rare CNVs per sample (Fig. 4e, Supplementary Table S2).

ML-ExonCNV assigns a reliability score to exCNVs, categorizing them into highly reliable, relatively reliable, and unreliable groups. Using the R492 dataset, the precision ($P = TP/(TP + FP)$) of these categories was compared. The highly reliable group achieved qPCR validation rates of 99.4% for losses and 96.2% for gains (Supplementary Table S3). These high validation rates can reduce clinical validation costs and enhance decision-making efficiency.

Performance on the NA12878 test dataset

Copy number analysis from WES data is more challenging than SNV/indel analysis [14]. Therefore, the benchmark for CNVs in NA12878 is still evolving. Previous studies have shown that the CNV benchmark for NA12878 is incomplete. Comparing two different benchmark datasets revealed that 56.05% and 32.43% of CNVs were unique to each dataset, respectively [19]. This indicates that the gold standard set of true CNVs for NA12878 is not comprehensive. We used the previously established set of CNVs for the NA12878 sample, comprising 3512 CNVs as a reference dataset [19], of which 77 CNVs overlapping with coding regions were used as the evaluation set (Supplementary Table S9). We compared the sensitivity and precision of various software tools. ML-ExonCNV achieved the highest sensitivity of 45%, while ExomeDepth, GATK-gCNV, and CNVkit had sensitivities of 17%, 18%, and 3%, respectively. Due to its low sensitivity, CNVkit was excluded. In terms of precision, ML-ExonCNV, ExomeDepth, and GATK-gCNV had precisions of 12%, 4%, and 4%, respectively. This shows that our software achieved the highest sensitivity and precision in the gold standard reference dataset.

Compared with the R492 dataset, the performance of our software on the NA12878 dataset decreased significantly. The reason is that the reference CNVs in NA12878 are high-frequency CNVs in the population, and the presence of these CNV segments in the control dataset leads to decreased sensitivity. In addition, the benchmark for coding region CNVs in NA12878 is still incomplete [19].

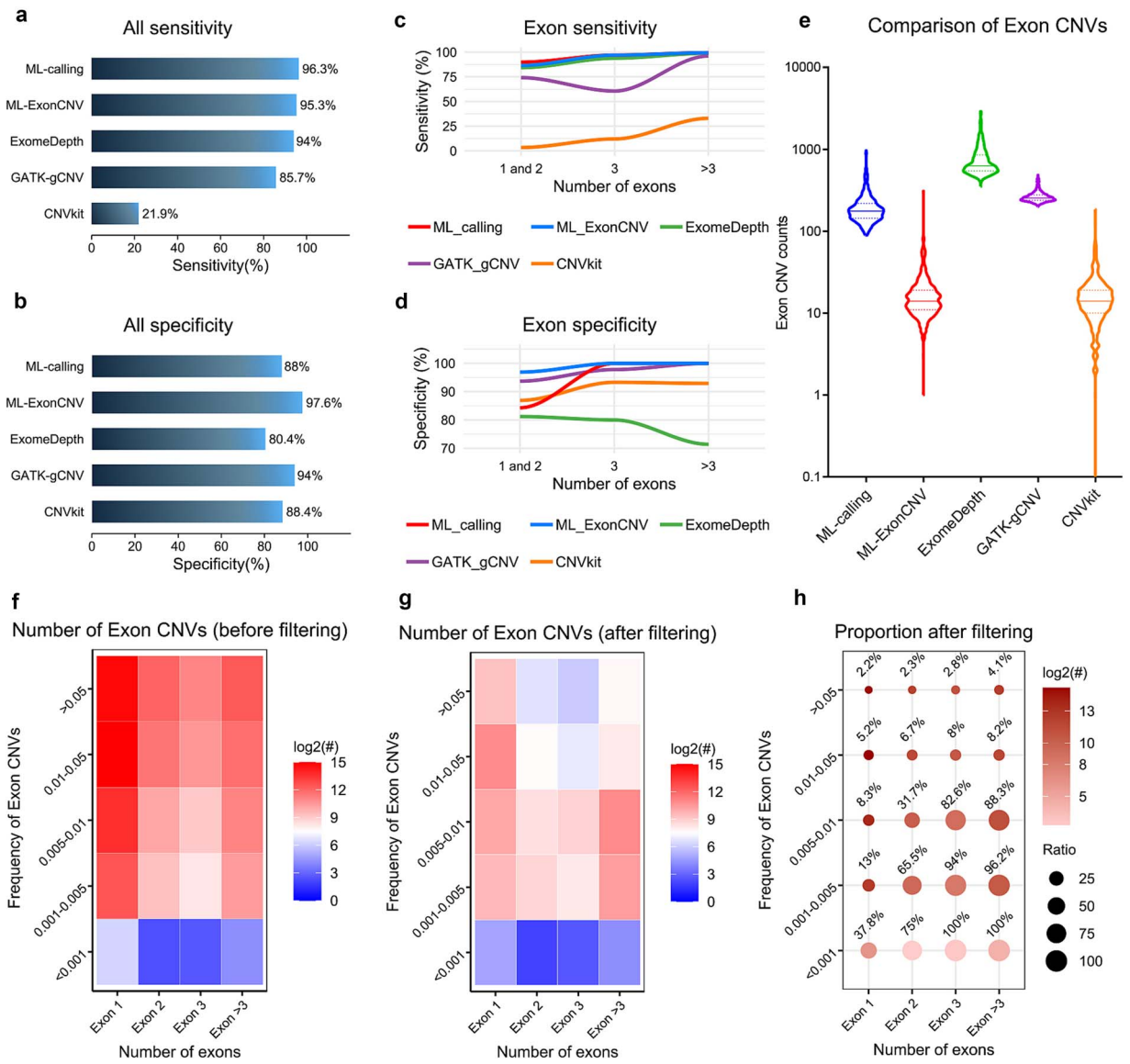


Figure 4 Performance comparison of ML-ExonCNV based on R492 test data. (a) Comparison of overall sensitivity among ML-calling(before rare CNV filtering), ML-ExonCNV (after rare CNV filtering), ExomeDepth, GATK-gCNV, and CNVkit on the R492 rare exCNVs dataset. (b) Comparison of overall specificity among the various software tools. (c) Classification by exon number (1–2 exons, 3 exons, >3 exons) and comparison of sensitivity changes for each software. (d) Classification by exon number and comparison of specificity changes for each software. (e) Distribution of reported exCNVs by each software, with fewer numbers indicating higher specificity. (f) Heatmap of exCNV distribution based on the number of exons and population frequency, before rare CNV filtering model application, using the R492 dataset following the ML-calling step. (g) Heatmap of exCNV distribution after rare CNV filtering model application. (h) Ratio of exCNV changes after and before rare CNV filtering model application, different sizes representing the retention rate after filtering (100% indicates no filtering).

In summary, the parallel comparison of CNV calling on the NA12878 test dataset across various software tools revealed that our algorithm still achieved relatively good performance on the benchmark dataset.

Rare copy number variant filtering model

Exon-level CNVs are frequently affected by experimental bias, leading to the frequent appearance of false-positive CNVs across multiple samples [14]. In contrast, pathogenic CNVs are subject to negative selection and thus have lower population frequencies. The rare CNV filtering model (Supplementary Table S7) can effectively screen for pathogenic and reliable exCNVs, thereby enhancing the specificity of exCNV results. Our study demonstrated that applying the rare CNV filtering model to the rare exCNVs in the R492 dataset significantly improved specificity, from 88.0% to 97.6% (Fig. 4b), and precision, from 90.6% to 98% (Table 1). Sensitivity only slightly decreased from 96.3% to 95.3% (Fig. 4a). Analysis revealed that the primary reason for the slight decrease in sensitivity was the filtering of the most challenging 1–2 exons del/dup events, which have higher false-positive rates (del precision 82%, dup precision 46%) (Table 1). A heatmap of exCNV distribution before filtering (Fig. 4f) showed that CNVs were predominantly located in single-exon regions with high population frequencies, suggesting that most CNVs are polymorphic or of low reliability. A heatmap of exCNV distribution after applying the rare CNV

Table 1 ML-ExonCNV performance before and after rare CNV filtering.

Step	CNV	Exon counts	Sensitivity (*N* = 301)	Specificity (*N* = 250)	Precision
ML-ExonCNV calling before rare CNV filtering	DEL	1–2 exons	89.3%	84.2%	81.7%
		3 exons	96.6%	100%	100%
		>3 exons	100%	100%	100%
	DUP	1–2 exons	92.9%	84.4%	46.4%
		3 exons	100%	100%	100%
		>3 exons	98.1%	100%	100%
	Total	All exons	96.3%	88.0%	90.6%
ML-ExonCNV after rare CNV filtering	DEL	1–2 exons	86.7%	95.8%	94.2%
		3 exons	96.6%	100%	100%
		>3 exons	100%	100%	100%
	DUP	1–2 exons	85.7%	97.9%	85.7%
		3 exons	100%	100%	100%
		>3 exons	98.1%	100%	100%
	Total	All exons	95.3%	97.6%	98.0%

filtering model (Fig. 4g) revealed a significant reduction in single-exon and high-frequency CNVs. Additionally, a comparison of CNV counts before and after filtering (Fig. 4h) (postfiltering CNV count/prefiltering CNV count) showed that CNVs involving ≥3 exons and with a population frequency < 5‰ were retained at a rate of 94%–100%, while CNVs involving 1–2 exons and with a population frequency > 1% were retained at a rate of only 2.2%–6.7%. Statistical analysis indicated that exCNVs with low population frequency (higher pathogenicity) or more exons (higher reliability) were retained at higher proportions. Ultimately, the rare CNV filtering model, through its gradient filtering algorithm, significantly enhanced result specificity while maintaining high sensitivity for rare CNVs, ensuring that clinically relevant CNVs are not missed in analysis.

Large copy number variant calling

ML-ExonCNV is capable of detecting exCNVs based on RC signals and can identify breakpoints of exCNVs when the breakpoints occur within exons. It can also detect mosaic CNVs, arm-level CNVs, and chromosomal number abnormalities. In comparative testing, one sample (case S1) was identified as female, and ML-ExonCNV detected a mosaic duplication on chr5:1-32891716, spanning 32 Mb with a copy number increase of less than one copy, indicative of a mosaic duplication (Fig. 5a). Additionally, in this female individual, a mosaic deletion was detected on the X chromosome at chrX:1-155270560 (Fig. 5a). This mosaic CNV was not detected by GATK-gCNV, ExomeDepth, or CNVkit. In the second case (S2), our software detected an arm-level CNV, specifically a duplication of the short arm of chromosome 18 (Fig. 5b). This CNV was also detected by CNVkit, but not by GATK-gCNV or ExomeDepth. In the third case (S3), ML-ExonCNV's RC algorithm detected a copy number deletion in exon 41 of the NIPBL gene. Additionally, using Manta's breakpoint algorithm, a deletion was detected at chr5:37049395-37052384 (Fig. 5c), which actually included the terminal region of exon 40 and the entirety of exon 41. Due to the short length of the exon 40 deletion region, and because the detection principles of GATK-gCNV and ExomeDepth are based on RC counting, both missed the deletion in exon 40. CNVkit, due to its lower sensitivity, did not detect any deletions. The CNV results for cases S1–S3 were all validated by qPCR. Ultimately, these three typical cases demonstrate that ML-ExonCNV outperforms other software in detecting large CNVs and breakpoint identification (Supplementary Table S4).

Summary of multidimensional comparison

On the R492 test dataset, CNV calling was performed using different software tools on a server equipped with 56 cores and 128 GB of memory. The comprehensive comparison revealed that ML-ExonCNV completed the analysis of all samples, including the construction of the control library and CNV calling, in 28 hours (Supplementary Table S5). This processing speed is 3 times faster than ExomeDepth, 11 times faster than CNVkit, and 22 times faster than GATK-gCNV (runtime parameters are provided in the Supplementary Text).

In addition, by comparing the output results of different software tools, we found that all software tools can output segmental CNVs from WES data. However, only ML-ExonCNV can output exon-level CNVs and provide reliability assessments (highly reliable, relatively reliable, unreliable), facilitating the adjustment of sensitivity and specificity according to clinical or research needs. Furthermore, none of the other software tools have the capability to detect mosaic CNVs or breakpoints (Supplementary Table S5). Regarding the detection of chromosomal number abnormalities, ExomeDepth, which is designed to detect small exon-level CNVs, lacks the capability to detect large CNVs.

Exon copy number variants high-recurrence genes and phenotypes

Statistical analysis of the 9827 qPCR-positive exCNVs in the training set of this study revealed that the gene distribution of exCNV losses and gains exhibits a long-tail effect (Fig. 6a), with some genes having a higher recurrence rate. For example, the DMD gene has the highest number of exon-level losses and gains. Other genes have different incidence rankings for losses and gains.

Based on the HPO database, phenotype enrichment analysis was conducted on the top 30 high-recurrence genes (Fig. 6b). It was found that the phenotypes most highly associated with exCNVs are as follows: intellectual disability is associated with 13 genes, seizure with 12 genes, and global developmental delay with 10 genes. The top 3 phenotypes all belong to abnormality of the nervous system (HP:

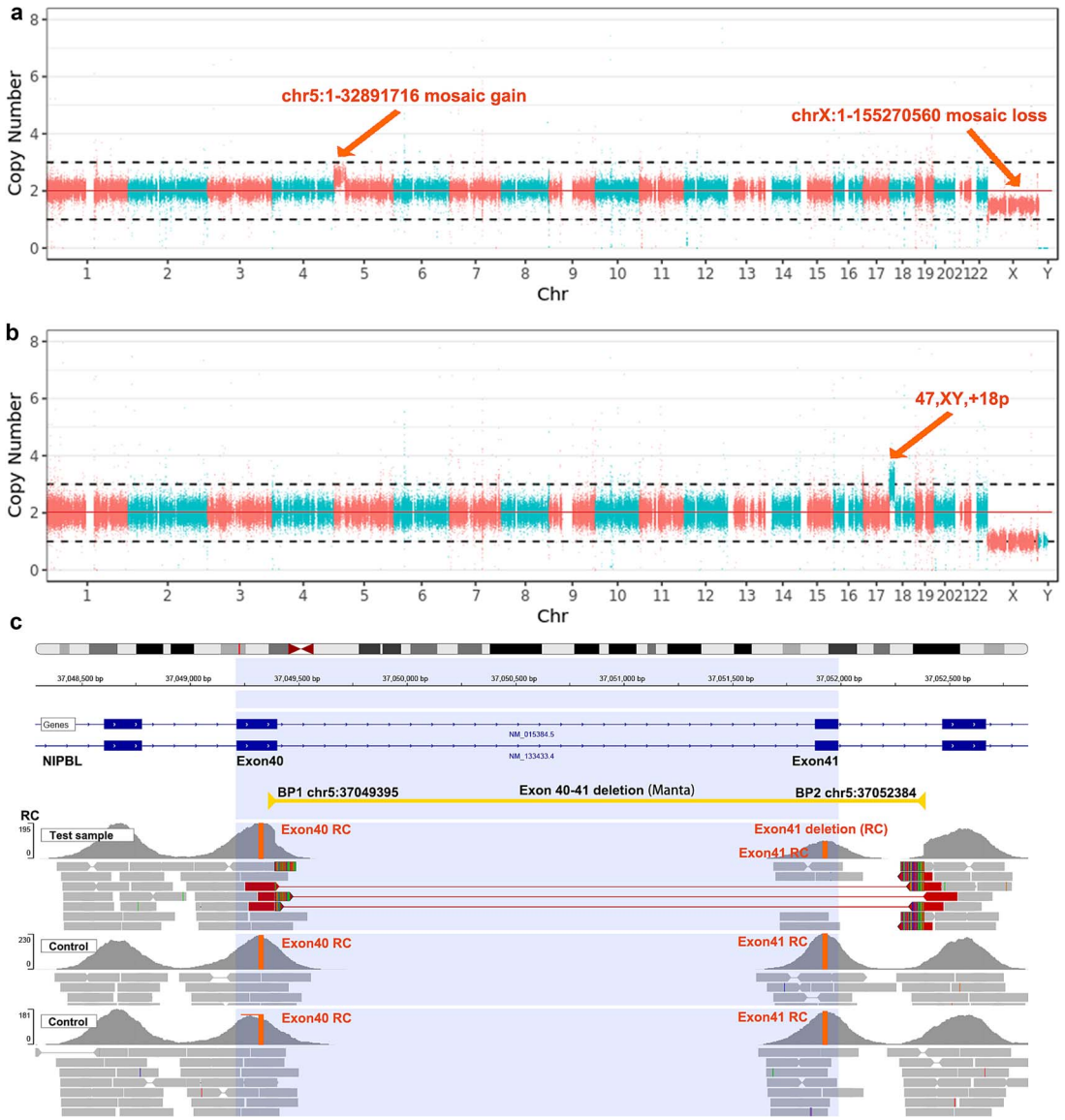


Figure 5 Large CNVs detected by ML-ExonCNV calling. (a) A case of mosaic CNV detected by the ML-ExonCNV algorithm. The *x*-axis represents different chromosomes, and the *y*-axis represents copy number. The middle line indicates two copies (i.e. normal copy number), while the black dashed lines represent an increase of one copy for gain and a decrease of one copy for loss. The arrow in the figure indicates a large segmental mosaic gain on chr5 (below the gain black dashed line) and a mosaic loss on the X chromosome (above the loss black dashed line). (b) A case of large segmental duplication detected by the ML-ExonCNV algorithm (47, XY +18p), showing a duplication of the short arm of chromosome 18 (an increase of one copy). (c) A screenshot from the Integrative Genomics Viewer (IGV) for the NIPBL gene based on the test sample and two control samples. The figure shows the read distribution for exons 40–41 of the gene (light blue background). The RC algorithm of ML-ExonCNV detected a deletion in exon 41 in the test sample, while the RC of exon 40 appeared normal (orange). Additionally, the breakpoint algorithm of ML-ExonCNV (using the Manta software) identified a deletion at chr5:37049395-37052384 (yellow line segment BP1–BP2), supported by multiple breakpoint reads (red reads in the figure), covering the entire exon 41 and part of exon 40, ultimately confirming a deletion at chr5:37049395-37052384.

0000707). The fourth and fifth ranked phenotypes are generalized hypotonia and scoliosis, both associated with 9 genes and belonging to abnormality of the musculoskeletal system (HP: 0033127). Through this enrichment analysis, our study results reveal that exCNVs are an important pathogenic mechanism in patients with genetic diseases of the nervous and musculoskeletal systems. A Sankey diagram analysis of the relationship between the top 30 high-recurrence phenotypes and genes associated with exCNVs (Fig. 6c) shows that phenotypes of nervous system abnormalities (blue) are associated with the most genes. This result clearly reveals the high-recurrence genes and phenotypes associated with rare exCNVs and their interrelationships (Supplementary Table S6), providing new findings and insights for clinical practice and research.

Discussion

In recent years, with the increase in WES data volume, it has become possible to use machine learning methods trained on real-world big data to improve the detection of exCNVs. Compared with traditional WES CNV detection methods, the ML-ExonCNV detection method introduces several innovative technologies (Fig. 1): (i) Modeling of the control library and population frequency database to filter out

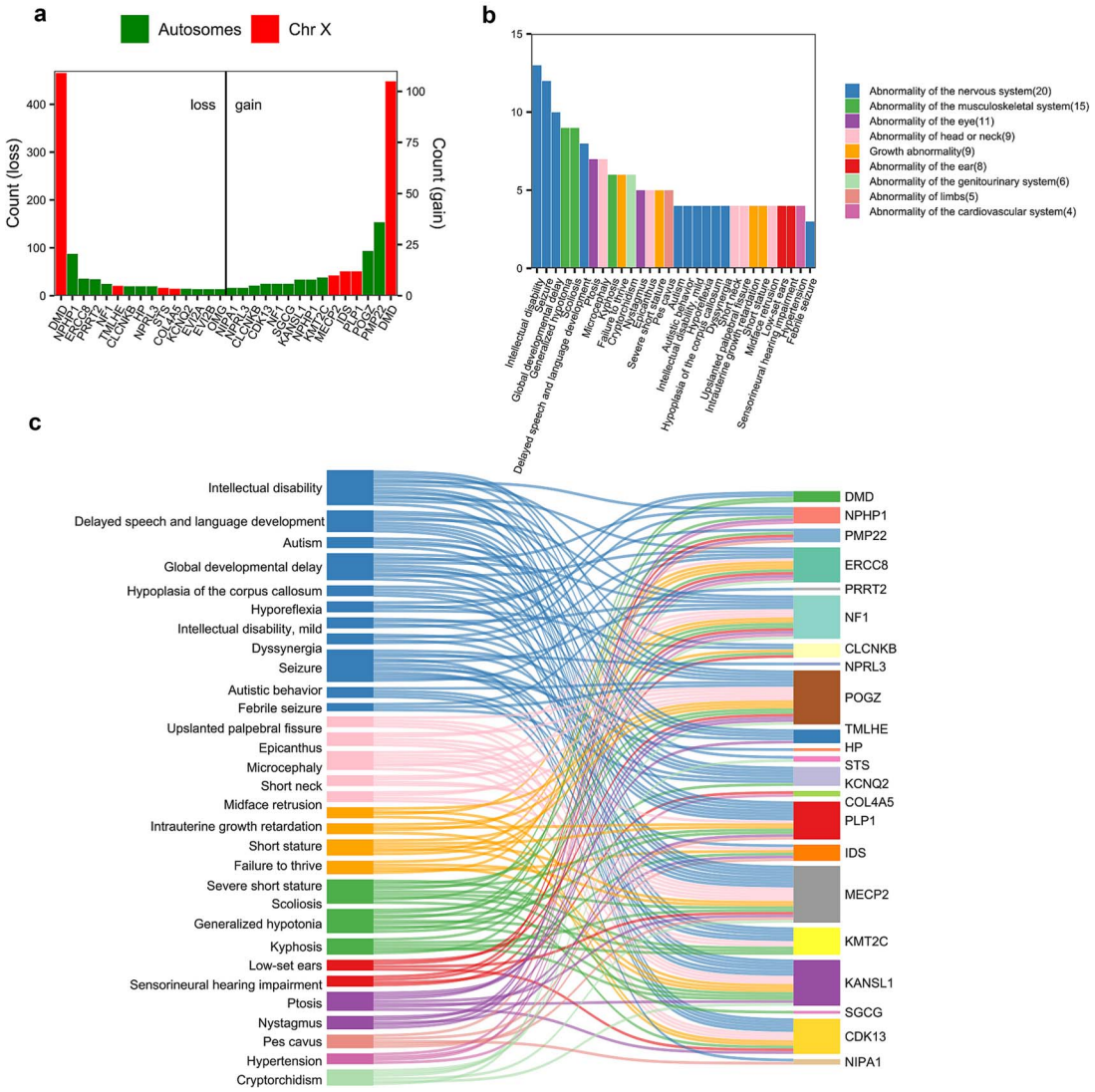


Figure 6 High-recurrence genes and phenotypes of exCNVs. (a) Based on the 9827 qPCR-positive rare exCNVs in the training dataset, the top 30 high-recurrence loss and gain genes were statistically analyzed. (b) The HPO phenotypes of the top 30 high-recurrence genes were statistically analyzed and sorted by recurrence frequency, with different colors representing different system abnormalities (e.g. blue represents nervous system abnormalities). (c) A Sankey diagram illustrates the relationship between high-recurrence genes and high-recurrence phenotypes, with different colors in the phenotypes representing different system abnormalities.

polymorphic and background noise CNVs. (ii) Pretraining modeling of 270 000 targets with known validation results based on 14 carefully selected features to establish a reliability prediction model for exCNVs. (iii) Innovatively proposing TV parameter features to measure the reliability of the entire exCNV analysis process. (iv) Pioneering the combination of the XGBoost model with the MEE model, using three expert models to predict the reliability score of exCNVs. (v) Introducing the Pearson coefficient to select likely controls in the CNV calling step, GC correction to reduce noise, and Z-score for significance filtering. (vi) Fine-tuning the reliability of exCNVs using variants from VCF files. (vii) Enhancing the accuracy of exCNV detection by integrating the breakpoint algorithm from the Manta software. (viii) Enabling ML-ExonCNV to detect mosaic CNVs and chromosomal number abnormalities through the connection of exCNVs. Ultimately, the performance of the AI-based ML-ExonCNV surpasses that of traditional software such as GATK-gCNV, ExomeDepth, and CNVkit.

In our study, the ability to detect CNVs varied significantly between the two datasets. In the R492 test dataset, ML-ExonCNV achieved the highest sensitivity (95.3%) and specificity (97.6%). ExomeDepth had relatively high sensitivity (94%) but lower specificity (80.4%), while GATK-gCNV exhibited higher specificity (94%) but slightly lower sensitivity (85.7%). We found that ExomeDepth and GATK-gCNV demonstrated good detection capabilities at the exon-level CNVs, but both tools missed large segmental CNVs (Supplementary Table S4). Although CNVkit had a low sensitivity for detecting exon-level CNVs (Fig. 4a, sensitivity 21.9%), it accurately detected the 47,XY,+18p karyotype (Supplementary Table S4). In comparison, ML-ExonCNV exhibited superior performance in detecting exon-level CNVs, large segmental CNVs, mosaic CNVs, and even CNV breakpoints (Fig. 5). Additionally, ML-ExonCNV achieved a 3- to 22-fold increase in computational speed (Supplementary Table S5).

However, this method also has limitations. Given the limited training data and the need for model interpretability, we trained three

pretrained expert models corresponding to loss, gain, and wild-type scenarios. As data volumes grow, subsequent releases of ML-ExonCNV can incorporate additional expert models to further enhance performance. During the control library modeling process in step one, it is recommended to use data from similar experimental systems for modeling, such as the same capture probes and the same experimental standard operating procedure (SOP). This can improve model accuracy.

Conclusions

ML-ExonCNV, based on real-world exon-level CNVs validated by qPCR, has achieved high sensitivity (95.3%) and specificity (97.6%) in exCNV calling using the XGBoost and MEE machine learning algorithms. Additionally, ML-ExonCNV demonstrates superior performance in detecting large segmental CNVs, mosaic CNVs, and even CNV breakpoints. It also achieves faster computational speed, saving time for users.

Our study results reveal high-recurrence genes and phenotypes in exCNV genetic diseases. We identified that in the pathogenic genes associated with exCNVs, abnormalities in neurodevelopment and the musculoskeletal system are highly recurrent phenotypes. This finding suggests that exCNV analysis in such phenotypes can yield higher diagnostic rates for clinicians and researchers.

Key Points

- ML-ExonCNV innovatively integrates an XGBoost classifier within a Multi-Expert Ensemble framework.
- We introduced technical variability as a metric to assess the reliability of CNV calls.
- ML-ExonCNV achieved both sensitivity and specificity exceeding 95%.
- ML-ExonCNV exhibited superior performance in detecting large segmental CNVs, mosaic CNVs, and pinpointing CNV breakpoints.
- High-recurrence genes and their associated phenotypes were identified from exCNVs.

Acknowledgements

We thank the Chigene (Beijing) Translational Medical Research Center and the Beijing Quanpu Medical Laboratory for providing the training and test datasets, respectively. We also thank all participants for their contributions.

Author contributions

Shuang-Hao Yang (Conceptualization, Methodology, Data Curation, Writing—original draft), Hua He (Methodology, Writing—review & editing), Shuyu Hou (Methodology, Writing—review & editing), Tuanfeng Yang (Methodology, Writing—review & editing), Zehao Yin (Methodology, Writing—review & editing), Hong-Yu Zhang (Conceptualization, Writing—review & editing), Weiyue Gu (Conceptualization, Writing—review & editing)

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Funding

None declared.

Data availability

The software ML-ExonCNV is publicly available on GitHub, complete with comprehensive usage instructions and a disclaimer (https://github.com/chigenelaby/ML-ExonCNV). To facilitate its adoption, we have additionally prepared a Docker image that can be accessed at https://hub.docker.com/r/chigenelaby/mlexoncnv/tags. The software GATK-gCNV can be obtained from https://github.com/broadinstitute/gatk. For ExomeDepth, the software is accessible at https://github.com/vplagnol/ExomeDepth, and it can be directly installed via the command install.packages(ExomeDepth). CNVkit is available for download at https://github.com/etal/cnvkit. Ninety publicly available WES samples from the 1000 Genomes Project (PMID: 25618849) are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data.

References

1. Babadi M, Fu JM, Lee SK. *et al.* GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat Genet* 2023;**55**:1589–97. https://doi.org/10.1038/s41588-023-01449-0.
2. Wei Y-C, Huang G-H. CONY: a Bayesian procedure for detecting copy number variations from sequencing read depths. *Sci Rep* 2020;**10**:10493.
3. Yang S-H, Liu J, Quan Y. *et al.* The improvement in diagnostic yield of developmental and epileptic encephalopathy by the multi-omics sequential testing method. *Biochim Biophys Acta Mol Basis Dis* 2025;**1871**:167854.
4. Miller DT, Adam MP, Aradhya S. *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**:749–64.
5. Kadalayil L, Rafiq S, Rose-Zerilli MJJ. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* 2015;**16**:380–92. https://doi.org/10.1093/bib/bbu027.
6. Srivastava S, Love-Nichols JA, Dies KA. *et al.* Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med* 2019;**21**:2413–21. https://doi.org/10.1038/s41436-019-0554-6.
7. Plagnol V, Curtis J, Epstein M. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012;**28**:2747–54. https://doi.org/10.1093/bioinformatics/bts526.
8. Talevich E, Shain AH, Botton T. *et al.* CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;**12**:e1004873. https://doi.org/10.1371/journal.pcbi.1004873.
9. Mandiracioglu B, Ozden F, Kaynar G. *et al.* ECOLE: learning to call copy number variants on whole exome sequencing data. *Nat Commun* 2024;**15**:132.
10. Moreno-Cabrera JM, Del Valle J, Castellanos E. *et al.* Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet* 2020;**28**:1645–55. https://doi.org/10.1038/s41431-020-0675-z.
11. Krumm N, Sudmant PH, Ko A. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;**22**:1525–32. https://doi.org/10.1101/gr.138115.112.

12. Karakoc E, Alkan C, O'Roak BJ. *et al.* Detection of structural variants and indels within exome data. *Nat Methods* 2011;**9**:176–8. https://doi.org/10.1038/nmeth.1810.
13. Abyzov A, Urban AE, Snyder M. *et al.* CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;**21**:974–84. https://doi.org/10.1101/gr.114876.110.
14. Gabrielaite M, Torp MH, Rasmussen MS. *et al.* A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers* 2021;**13**:6283.
15. Hijikata A, Suyama M, Kikugawa S. *et al.* Exome-wide benchmark of difficult-to-sequence regions using short-read next-generation DNA sequencing. *Nucleic Acids Res* 2024;**52**:114–24. https://doi.org/10.1093/nar/gkad1140.
16. Dai D, Deng C, Zhao C. *et al*. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models. 2024.
17. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;**40**:e72. https://doi.org/10.1093/nar/gks001.
18. Chen X, Schulz-Trieglaff O, Shaw R. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;**32**:1220–2.
19. Rao J, Peng L, Liang X. *et al.* Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms. *BMC Bioinformatics* 2020;**21**:518.
20. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 2018;**34**:867–8. https://doi.org/10.1093/bioinformatics/btx699.
21. Iadarola B, Xumerle L, Lavezzari D. *et al.* Shedding light on dark genes: enhanced targeted resequencing by optimizing the combination of enrichment technology and DNA fragment length. *Sci Rep* 2020;**10**:9424.
22. Mandelker D, Schmidt RJ, Ankala A. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med* 2016;**18**:1282–9. https://doi.org/10.1038/gim.2016.58.
23. Munté E, Roca C, Del Valle J. *et al.* Detection of germline CNVs from gene panel data: benchmarking the state of the art. *Brief Bioinform* 2024;**26**:bbae645. https://doi.org/10.1093/bib/bbae645.

Briefings in Bioinformatics, 2026, **27**, bbag100
https://doi.org/10.1093/bib/bbag100
Problem Solving Protocol