Supplementary Notes

Next-generation sequencing and sequencing data preprocessing

The Agilent (Santa Clara, CA, USA) ClearSeq Inherited Disease panel kit for enrichment was adapted on every enrolled proband. Genomic DNA was fragmented by sonication, ligated to multiplexing paired-end adapters, and amplified by PCR with indexed primers for sequencing, which were then hybridized to biotin-labelled probes, followed by paired-end sequencing (125 bp/150 bp) on the Illumina (San Diego, CA, USA) HiSeq 2000/2500 platforms. The average coverage for the yielded data was 124.05X, and the average fraction of target covered with at least 10X and 20X were 99.11% and 97.17%, respectively. Reads were cleaned to pass quality controls and were aligned to the hg19 reference genome by BWA-MEM (V.05.9-r16), sorted by SAMtools (v.1.8), and removed duplicates by Picard (v.2.20.1).

Pipeline for NGS data-based CNV detection

Data quality assessment

Data quality assessment was performed before further CNV detection. For each BAM file, the number of mapped reads on the annotated exons was calculated by BEDTools with default parameters and standardized into read depth. Mean and coefficient of variance (C. V) values for read depth across all samples were calculated and assessed on each exon. Autosomes and X chromosomes were separately calculated, since the read depth of chromosome X was adjusted according to gender information. Gene and exon annotations were based on GENCODE (v19) and downloaded from the UCSC Table Browser (wgEncodeGencodeBasicV19).

CNV calling

Both CANOES[1] and HMZDelFinder[2] were applied for the original detection of CNVs from exome sequencing data. The CANOES approach is optimized to call deletions involving three or more exons and HMZDelFinder is superior in calling single exon-level CNVs.

CANOES uses a negative binomial distribution to model read counts and detect CNVs from exome sequencing data. In the published version of CANOES, researchers stated that all CNV callings were restricted to the autosomes due to the complications resulting from the samples being of different sexes[1]. As quite a few DD-associated CNVs are located on chromosome X, we extended the CNV calling to chromosome X by adjusting the target region coverage according to the samples' sex. The detailed extension process included adjusting the X chromosome coverage according to gender and evaluating the baseline for CNV detection on the X chromosome using autosome as well as X chromosome coverage data. Both read coverage and the GC content of the exons were used to call CNVs by CANOES.

HMZDelFinder is an algorithm for the detection of rare and intragenic homozygous and hemizygous (HMZ) deletions. BAM files and VCF files of individuals were extracted as input for variant detection. The application was conducted under default-set thresholds.

CNV annotation and filtering

CNVs called from CANOES and HMZDelFinder were combined and annotated at both the gene and region levels. The annotation and filtration pipeline was called PICNIC (pipeline for clinical NGS-involved CNV detection). The PICNIC is a published automatic process that combines CNV detection results for annotation and filtration[3]. In the published work, 58 patients underwent CES-based CNV detection. Compared with the CMA result, PICNIC showed 100% sensitivity (95%CI: 92.13%-100.00%) to pathogenic/likely pathogenic CNVs, demonstrating a reliable analysis result of the CES-based CNV detection[3]. The flowchart of PICNIC is shown in **Supplementary Figure S1.**

At the gene level, PICNIC used RefSeq for mRNA annotation and OMIM, HGMD, and UniProt for functional annotation. Then, PICNIC filtered out gene deletions/duplications, which occurred in >10% of the internal samples, as their high frequency made the variation unlikely to cause rare disorders. Next, CNV-influenced genes were compared with clinical phenotypes by using the annotated HPO terms and the standardized clinical records. In this study, to reduce the amount of manual work, we generated a preliminary automatic HPO annotation system to extract HPO terms from the Chinese electronic clinical record. Clinical record sentences in Chinese were translated into English, annotated by MetaMap into UMLS standard phrases and converted to HPO terms. We also established a semantic dictionary supervised by clinicians to correct the annotation process. The final HPO assignment needed to be manually reviewed by clinicians before further analysis. A CNV would be categorized as a candidate variant if any of its affected genes could match the patient’s clinical HPO (or parental HPO) terms; otherwise, the CNV would be considered unknown or likely benign. Specifically, CNV-affected genes were compared with patient's HPO terms according to the annotations of known disease-causing genes provided by the HPO database (http://www.human-phenotype-ontology.org). Once any CNV-containing gene's HPO annotation matched with the patient's HPO terms, the CNV would be classified as a candidate variant. Considering the hierarchical and specialization differences in HPO terminology, two HPO terms that shared the same parent node in the upper three layers were considered to be consistent.

For region-level annotation, we annotated CNVs with DGV, DECIPHER and a prepublished pathogenic database[4]. The detected CNV was classified as a candidate variant if it had any overlap with an established pathogenic region or the known causative genes of an established pathogenic CNV. For other CNVs, variation regions larger than 1 Mb were also be marked as candidate variants, considering that abundant CNV-affected genes would be more likely to be pathogenic. The data of the prepublished pathogenic database are given in the **Supplementary Table S8**.

Criteria for variant classification

The variant classification, which is based on the ACMG guidelines but with some adjustments, was first described in our previous work[5] and implemented in this study. Hence, the variant classification would not always be consistent with those in the ACMG guidelines.

Specifically, the criteria of pathogenic (P) variants were as follows: 1) the variant would likely explain the indication for testing and may be responsible for the patient's clinical presentation; and 2) for SNV and small indel, the variant has the same nucleotide and amino acid change as a previously established pathogenic variant from published studies or the internal database; for CNV, the variant has the same copy-number status as a previously established pathogenic CNV and fully covers the region.

Criteria of likely pathogenic (LP) variants: 1) the variant would likely explain the indication for testing and may be responsible for the patient's clinical presentation; and 2) for SNV and small indel, the variant has the same amino acid change as a previously established pathogenic variant regardless of nucleotide change; or Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon) in a gene where loss of function (LOF) is a known mechanism of disease; for CNV, the variant overlaps with a previously established pathogenic CNV (overlapped region is >70% of the reported pathogenic CNV) and the overlapped region included a gene where LOF is a known mechanism of disease or established/predicted to be haploinsufficiency (HI); or both breakpoints of the CNV are within the same established haploinsufficiency gene; and 3) the variant is *de novo* (both maternity and paternity confirmed) in the proband with a negative family history; or is inherited from the affected parents. Specifically, if the parents are not available for the confirmation of *de novo* or compound heterozygous status of pathogenic variants identified in the proband, the variant would be downgraded and classified as 'pathogenic to likely-pathogenic' (P->LP).

Process of CNV and SNV integrated (SCI) strategy for diagnosing

The CNV and SNV integrated (SCI) strategy is a diagnosing condition that combines both

SNV and CNV. Based on the above variant classification criteria, diagnostic results for SNVs and small indels included: 1) one heterozygous P/LP variant in an autosomal dominant, or X-linked dominant gene; 2) one homozygous or two heterozygous P/LP variants (compound heterozygous) on an autosomal recessive gene; and 3) one P/LP variant in an X-linked recessive gene in males. Diagnostic results for CNV included: the P/LP CNV that explains the indication for testing and the inheritance pattern.

Moreover, for the SNVs and CNVs that failed to meet the aforementioned conditions but resulted in a rarer event: apparently homozygous (AH) variant caused by overlapping P/LP SNV on one allele and P/LP CNV on the other. These "SNV+CNV AH" cases were also included as one of the diagnosing conditions.

In addition, the initial diagnosis was issued without experimental confirmation if the variants were of high confidence[6] (SNV: coverage ≥20×, minor allele fraction ≥35%, and Phred score of variant calling ≥30; CNV: detected variation size >1 Mb).

Variant experimental validation

For SNVs, all of the identified P/LP variants were confirmed by Sanger sequencing.

For CNVs, referring to a previous study that performed an evaluation of CNV detection from panel-based next-generation sequencing data, researchers found that CNVs covering adequate exons on autosomes can be accurately detected using targeted panel sequencing data same as using CMA, while CNVs detected from sex chromosomes need further evaluation and validation. Additionally, the accuracy of CNV size improves as the size of variants increased[7]. Thus, in this study, we considered the CNVs with sizes larger than 1 Mb as variants of high confidence. The quantitative polymerase chain reaction (qPCR), multiplex ligation-dependent probe amplification (MLPA) or CMA validations were performed on CNVs detected from sex chromosomes and CNVs whose detected sizes were less than 1 Mb. Detailed information about the kits used and designed primers for CNV validations are given in **Supplementary Table S1**.

Reference

1. Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res* 2014;42:e97.

2. Gambin T, Akdemir ZC, Yuan B, Gu S, Chiang T, Carvalho CMB, Shaw C, Jhangiani S, Boone PM, Eldomery MK, Karaca E, Bayram Y, Stray-Pedersen A, Muzny D, Charng WL, Bahrambeigi V, Belmont JW, Boerwinkle E, Beaudet AL, Gibbs RA, Lupski JR. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Research* 2017;45:1633-48.

3. Qin Q, Liu B, Yang L, Wu BB, Wang HJ, Dong XR, Lu YL, Zhou WH. Application of copy number variation screening analysis process based on high-throughput sequencing technology. *Chinese Journal of Evidence -Based Pediatric* 2018;13:275-79.

4. Wang HJ, Bi WM, Wu BB, Zhou WH, Cheung SW. Application and explanation of chromosome microarray (CMA) analysis in the diagnosis of clinical genetic disease. *Chinese Journal of Evidence -Based Pediatric* 2014;9:227-35.

5. Yang L, Kong YT, Dong XR, Hu LY, Lin YF, Chen X, Ni Q, Lu YL, Wu BB, Wang HJ, Lu QR, Zhou WH. Clinical and genetic spectrum of a large cohort of children with epilepsy in China. *Genet Med* 2019;21:564-71.

6. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, Chawla A, Coffey AJ, Malhotra A, Scocchia A, Thorpe E, Dzidic N, Hovanes K, Sahoo T, Dolzhenko E, Lajoie B, Khouzam A, Chowdhury S, Belmont J, Roller E, Ivakhno S, Tanner S, McEachern J, Hambuch T, Eberle M, Hagelstrom RT, Bentley DR, Perry DL, Taft RJ. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med* 2018;21:1121–30.

7. Yao RE, Yu TT, Qing YR, Wang J, Shen YP. Evaluation of copy number variant detection from panel-based next-generation sequencing data. *Mol Genet Genom Med* 2019;7:e00513.