CHROMOSOME STRUCTURAL VARIANTS: EPIDEMIOLOGY, IDENTIFICATION AND CONTRIBUTION TO HUMAN DISEASES

EDITED BY: Zirui Dong, Dezso David, Claudia Gonzaga-Jauregui, Cynthia Casson Morton and Cinthya Zepeda Mendoza

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-295-2
DOI 10.3389/978-2-83250-295-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

CHROMOSOME STRUCTURAL VARIANTS: EPIDEMIOLOGY, IDENTIFICATION AND CONTRIBUTION TO HUMAN DISEASES

Topic Editors:
Zirui Dong, The Chinese University of Hong Kong, China
Dezso David, Departamento de Genética Humana, Instituto Nacional de Saúde Doutor Ricardo Jorge, Portugal
Claudia Gonzaga-Jauregui, Universidad Nacional Autónoma de México, Mexico
Cynthia Casson Morton, Brigham and Women's Hospital, United States
Cinthya Zepeda Mendoza, ARUP Laboratories, United States

Citation: Dong, Z., David, D., Gonzaga-Jauregui, C., Morton, C. C., Mendoza, C. Z., eds. (2022). Chromosome Structural Variants: Epidemiology, Identification and Contribution to Human Diseases. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-295-2

Table of Contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 09 September 2022
DOI 10.3389/fgene.2022.1022918

OPEN ACCESS

EDITED AND REVIEWED BY
Maxim B. Freidin,
Queen Mary University of London,
United Kingdom

*CORRESPONDENCE
Zirui Dong,
elvisdong@cuhk.edu.hk

SPECIALTY SECTION
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

RECEIVED 19 August 2022
ACCEPTED 22 August 2022
PUBLISHED 09 September 2022

CITATION
Dong Z, David D, Gonzaga-Jauregui C, Morton CC and Zepeda-Mendoza CJ (2022), Editorial: Chromosome structural variants: Epidemiology, identification and contribution to human diseases.
Front. Genet. 13:1022918.
doi: 10.3389/fgene.2022.1022918

Editorial: Chromosome structural variants: Epidemiology, identification and contribution to human diseases

Zirui Dong[1,2,3,4]*, Dezso David[5], Claudia Gonzaga-Jauregui[6], Cynthia C. Morton[7,8,9,10,11] and Cinthya J. Zepeda-Mendoza[12]

[1]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong SAR, China, [2]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China, [3]Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China, [4]Department of Obstetrics and Gynaecology, The Fertility Preservation Research Center, The Chinese University of Hong Kong, Hong Kong SAR, China, [5]Department of Human Genetics, National Health Institute Doutor Ricardo Jorge, Lisbon, Portugal, [6]International Laboratory for Human Genome Research, Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Juriquilla, Querétaro, Mexico, [7]Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, MA, United States, [8]Department of Pathology, Brigham and Women's Hospital, Boston, MA, United States, [9]Harvard Medical School, Boston, MA, United States, [10]Broad Institute of MIT and Harvard, Cambridge, MA, United States, [11]Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, Manchester, United Kingdom, [12]Division of Laboratory Genetics and Genomics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States

KEYWORDS

structural variant (SV), methodologies & tools, SV spectrum, sequence complexity, annotation and prediction, genomic variation, genomic rearrangements, SV mechanisms

Editorial on the Research Topic
Chromosome structural variants: Epidemiology, identification and contribution to human diseases

Human chromosome structural variants (SVs) are balanced/unbalanced genomic abnormalities that include translocation, inversion, insertion, and deletion/duplication (also known as copy-number variants, CNVs) events with a size of >50 bp. Currently, the capability of genome sequencing in the research and clinical fields has increased our capacity to detect cryptic SVs and further delineate the complexity of karyotypically/microarray detectable SVs. This has increased our knowledge of pathogenicity mechanisms by considering dysregulation of gene expression through position effects and complex interactions between gene dosage and mutational burden. However, much of the contribution of SVs to human disease is left to explore, as the incidence of SVs is still underestimated owing to limitations of current sequencing technologies and analytical pipelines, and few studies have comprehensively integrated SV information with single nucleotide variants in congenital diseases. Rigorous investigation of SV pathogenicity is warranted for clinical applications.

The Research Topic in this issue is divided into three main sections: three articles demonstrate methodologies in SV identification and pathogenicity annotation; five papers discuss the spectrum of SVs in individuals with different indications; and two reports characterize sequence complexity of SVs.

Methodologies in SV identification and pathogenicity annotation

1) Chen et al. describe an optimized analytical approach in non-invasive prenatal testing (NIPT) by combining Z-score with maternal CNV analysis. In routine NIPT analysis, the calculation of Z-score approach is commonly used for determining whether the fetus has a numerical disorder. However, among those cases with outliers of Z-scores (such as $Z > 3$ or $Z < -3$), the presence of maternal CNVs should be considered. After verification with diagnostic prenatal diagnosis, the authors suggest conducting Z-score analysis together with identification of maternal CNVs to reduce significantly the false positive calling rate. 2) Guo et al. propose a new method, namely stLFRsv (single-tube Long Fragment Read), for identifying SVs with the use of co-barcoded reads. Co-barcoded reads originating from long DNA fragments provide long-range genomic information with single-base level accuracy superior to a long-read sequencing approach; however, no analytical method for SV analysis is available. The authors show a higher accuracy of SV detection utilizing co-barcoded reads through identification of abnormal large gaps between co-barcoded reads to detect potential breakpoints for reconstructing complex SVs and further filtering via haplotype phasing analysis. 3) Fino et al. present a web-based application, SVInterpreter, for annotation of both balanced and unbalanced SVs using topologically associated domains (TADs) as genome units. With the advancement of detection methods, a significantly increasing number of SVs are detected in both patients and presumably healthy individuals, and most of these SVs are interpreted as variants of uncertain significance (VUS) due to limited knowledge of their pathogenicity. By incorporating gene-associated data (as function and dosage sensitivity), phenotype similarity scores, and CNV scoring metrics, the authors demonstrate that SVInterpreter identifies the possible disease-causing candidate (such as contributed by potential position effect events) and decreases interpretations of VUS by 40%.

SV spectrum in individuals with different indications

SVs are known to contribute to genomic diversity and diseases in individuals in different developmental stages: early miscarriage, prenatal, postnatal, and adult as well as serve as markers for somatic mutagenesis after exposure to a toxic environment. 1) Gu et al. show an uneven distribution of CNVs (<3 Mb in size) in euploid products of conception (POCs) with a higher density seen in the pericentromeric and subtelomeric regions, and the genes involved are significantly enriched in biological processes and pathways important to embryonic/fetal development. 2) Chau et al. examine the landscape of rare CNVs with parental inheritance assignment in trio-based prenatal diagnosis and demonstrate among 31 pathogenic/likely pathogenic CNVs identified, over 25% are small or mosaic CNVs unlikely to be detected by routine methods. 3) Hu et al. recruit seven Chinese prenatal cases with 21q21.1–q21.2 aberrations with comprehensive pedigrees, and demonstrate a benign clinical interpretation for pathogenic assessment of 21q21.1–q21.2 duplication and deletion, which were considered VUS or likely pathogenic in previous studies. 4) Lee et al. applied an in-house bioinformatics pipeline to 1,737 cases with Alzheimer Disease (AD) and 2,063 cognitively normal controls; burden tests show that Non-Hispanic White cases on average have 16 more duplications than controls, and Hispanic cases have larger deletions than controls. 5) Meléndez-Flórez et al. show that farmers exposed to pesticides had significantly increased frequencies of chromosomal alterations/variants, instability and clonal heterogeneity when compared with controls, which might contribute to an increased risk of developing diseases.

Sequence complexity of SVs

The advancement of different methodologies helps in the delineation of sequence complexity and composition of SVs, which potentially contribute to diseases through different mechanisms such as gene disruption or dysregulation. 1) Cao et al. applied mate-pair low-pass genome sequencing in cases with developmental disorders and/or intellectual disabilities and demonstrate that a large proportion of duplications previously classified as VUS are forward tandem duplications without contributing to diseases due to gene disruption. 2) Grochowski et al. describe a 5-year-old female presenting with a constellation of clinical features consistent with a clinical diagnosis of Coffin–Siris syndrome 1 (CSS1), which is contributed by *ARID1B* gene disruption resulting from a *de novo* pericentric and multiple paracentric inversions from a chromoanagenesis-like event.

Overall, studies from this Research Topic not only provide state-of-the-art methods for identification, delineation, and pathogenicity annotation of SVs, but also elucidate the incidence, spectrum, sequence complexity and potential contribution to human diseases.

Author contributions

The authors all contributed equally to the Research Topic assembly and editing and to this editorial.

Funding

CCM acknowledges NIH P01 GM061354 and support by the NIHR Manchester Biomedical Research Centre.

Conflict of interest

CCM is a member of the Scientific Advisory Board of Luna Genetics, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 18 March 2021
doi: 10.3389/fgene.2021.636239

stLFRsv: A Germline Structural Variant Analysis Pipeline Using Co-barcoded Reads

*Junfu Guo[1], Chang Shi[1], Xi Chen[1], Ou Wang[2], Ping Liu[3], Huanming Yang[4], Xun Xu[5], Wenwei Zhang[2] and Hongmei Zhu[2]**

[1] BGI-Tianjin, BGI-Shenzhen, Tianjin, China, [2] BGI-Shenzhen, Shenzhen, China, [3] MGI, BGI-Shenzhen, Shenzhen, China, [4] Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, China, [5] Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, China*

Co-barcoded reads originating from long DNA fragments (mean length >30 kbp) maintain both single base level accuracy and long-range genomic information. We propose a pipeline, stLFRsv, to detect structural variation using co-barcoded reads. stLFRsv identifies abnormal large gaps between co-barcoded reads to detect potential breakpoints and reconstruct complex structural variants (SVs). Haplotype phasing by co-barcoded reads increases the signal to noise ratio, and barcode sharing profiles are used to filter out false positives. We integrate the short read SV caller smoove for smaller variants with stLFRsv. The integrated pipeline was evaluated on the well-characterized genome HG002/NA24385, and 74.5% precision and a 22.4% recall rate were obtained for deletions. stLFRsv revealed some large variants not included in the benchmark set that were verified by long reads or assembly. For the HG001/NA12878 genome, stLFRsv also achieved the best performance for both resource usage and the detection of large variants. Our work indicates that co-barcoded read technology has the potential to improve genome completeness.

Keywords: : human genome, co-barcoded reads, structural variation, complex variants, breakpoints

OPEN ACCESS

Edited by:
Zirui Dong,
The Chinese University of Hong Kong, China

Reviewed by:
Wan-Ping Lee,
University of Pennsylvania, United States
Cinthya Zepeda Mendoza,
ARUP Laboratories, United States

****Correspondence:***
Hongmei Zhu
zhuhongmei@genomics.cn

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *01 December 2020*
Accepted: *04 February 2021*
Published: *18 March 2021*

Citation:
Guo J, Shi C, Chen X, Wang O, Liu P, Yang H, Xu X, Zhang W and Zhu H (2021) stLFRsv: A Germline Structural Variant Analysis Pipeline Using Co-barcoded Reads. Front. Genet. 12:636239. doi: 10.3389/fgene.2021.636239

INTRODUCTION

Structural variants (SVs) represent genome variants larger than 50 bp consisting of deletions, insertions, inversions, duplications, and translocations (Feuk et al., 2006; Alkan et al., 2011). SVs contribute more genomic sequence differences than single-nucleotide polymorphisms (SNPs) or small indels between genomes (Pang et al., 2010). Some of these SVs are pathogenic variants associated with specific diseases (Singleton et al., 2003; Jongmans et al., 2006; Rovelet-Lecrux et al., 2006). Despite the importance of SVs, profiling them has been challenging.

For the last 20 years, several technologies have allowed SV annotation to improve and have helped to generate a well-characterized human genome reference sequence to facilitate the development of SV identification tools (Zook et al., 2019[Preprint]). Among these technologies, sequencing is a primary category that includes long read, short read, and co-barcoded read sequencing. Each sequencing technique has unique advantages and disadvantages that contribute to the discovery of SV profiles among populations.

Long reads or single-molecule sequencing reads usually have mean length greater than 10 kbp. These longer reads identify breakpoints more easily and may span nearby repetitive regions of

several kilobases (Jain et al., 2018). However, long reads are prone to insertion and deletion errors, and the base level accuracy is comparatively low, which leads to low accuracy for small variant (less than 200 bp) detection (Wang et al., 2019). The single-molecule circular consensus sequencing protocol, which improves base level accuracy, produces high-quality reads that average >10 kbp (Wenger et al., 2019). However, this protocol is not applicable to large-scale projects because of throughput and cost limitations.

Short reads are accurate at the base level and cost-effective. Their uniform depth and insert size can be successfully used to identify deletions and copy number variation (Layer et al., 2014; Talevich et al., 2016). Deletions are easier to detect than insertions. However, more complex variants are rarely detected with short reads because their breakpoints are usually in close proximity to regions lacking unique short read alignment.

To compensate for the lack of long-range information, co-barcoded read sequencing was developed. Co-barcoded reads are the product of novel protocols for library construction and next-generation sequencing (NGS) sequencing technology. There are two mature technologies in this category, Linked-Reads by 10× Genomics (Zheng et al., 2016; Zhang et al., 2017) and single-tube long fragment read (stLFR) by BGI (Wang et al., 2019). In both cases, all the short reads that originate from the same long DNA molecule will share a common barcode. Thus, they retain long-range genome information while maintaining base level accuracy. Only nanograms of input DNA is needed, making co-barcoding feasible for many applications. The inferred average DNA fragment length for co-barcoded reads is approximately 30 kbp, which makes it possible to sequence across even larger repetitive regions near SV breakpoints. stLFR uses a combinatorial process to generate up to 3.6 billion unique barcodes, enabling practically nonredundant co-barcoding with 50 million barcodes per sample. Compared with Linked-Reads, stLFR can achieve a much lower barcode conflict rate (how many long DNA molecules share one barcode), which is beneficial for downstream analyses.

Analysis pipelines that detect SVs with co-barcoded reads fall into three categories based on how they use barcode information. The first category identifies novel adjacency by detecting abnormal numbers of common barcodes shared between two genomic loci or bins (Spies et al., 2017; Xia et al., 2018; Marks et al., 2019). The second tests the distribution of sequenced short segments on large DNA molecules (Elyanow et al., 2018; Marks et al., 2019). The third uses barcode information to extract data for local assembly (Meleshko et al., 2019[Preprint]; Zhou et al., 2019[Preprint]).

Here, we present stLFRsv, a co-barcoded read-based SV analysis pipeline that falls into the first category and integrates the short read SV detector smoove (Brent, 2018).

METHODS

Large SVs leave apparent large gaps in long fragments based on co-barcoded read alignment (**Figure 1**). The distribution of read pairs on long fragments is approximately random, and the gap sizes between read pairs vary in a wide range. Large gaps appear in long fragments by chance. However, large SVs are likely to lead to large gap aggregation. Thus, stLFRsv detects large gaps in fragments to identify large variants. In contrast, smoove is a pipeline that uses LUMPY as its core to detect paired-end discordance and other short read signals that indicate variants (Layer et al., 2014). We use smoove to find small and mid-sized variants. There may be overlap between the two variant sets, and thus, we merge them before generating the results (**Figure 2A**). The detection process for stLFRsv is described in the following steps.

Cluster Segment Ends

We calculate an empirical gap size distribution and select a size G as cut-off such that the probability of gap sizes smaller than G is P (**Supplementary Figure 2**). Usually, P is set as 98%, which is reasonable based on statistics. When we break a long fragment at a gap larger than G, we get two sub-fragments. We define the starting and terminal positions of a sub-fragment as the left and right ends. Each end has its position on the reference. We then divide the reference sequence into consecutive bins. Each bin has a size of B *bp* based on the data profile, which holds left and right ends and serves as left and right end clusters. Additionally, B is selected in the same way as G with P set to 65%, which aims to achieve a fine cluster performance and maintain reasonable precision for end positions. All end clusters with at least one end are retained for the next step (**Figure 2B**).

Pair Up Ends

Every two end clusters are checked for common barcodes to determine whether they could form a high-quality end pair (**Figure 2C**). These end pairs with common barcodes are potential novel adjacencies and are further checked as follows. First, the sub-fragment lengths for each barcode are collected to estimate the probability $f(d)$ that one barcode is observed at both locations *locA* and *locB* with a distance of d. $f(d)$ is defined as follows:

$$f(d) = \sum_{l>d} P(l) * \frac{l-d}{l+B}$$

in which l is the sub-fragment length, $P(l)$ is the probability of length l, and B is the size for clustering mentioned above. Second, the high-quality end pairs with a distance of d are decided by the following three rules. (1) The number of shared barcodes of two end clusters is higher than the theoretical value calculated by $f(d)$. (2) The barcode counts of each end cluster are N standard deviations higher than average depth (using $N = 3$ by default in the pipeline). (3) The barcode counts of each end cluster are significantly higher than neighboring clusters with P-values less than p_th by Wilcoxon signed-rank test (using default $p\_th = 0.1$ in the pipeline).

There are four types of end pairs according to the types of the two end clusters. If the potential novel adjacency does not involve an orientation change, the end pair is a right–left or left–right. Otherwise, it is either a left–left or right–right type (**Figure 1**). If an end cluster is in pair with multiple clusters

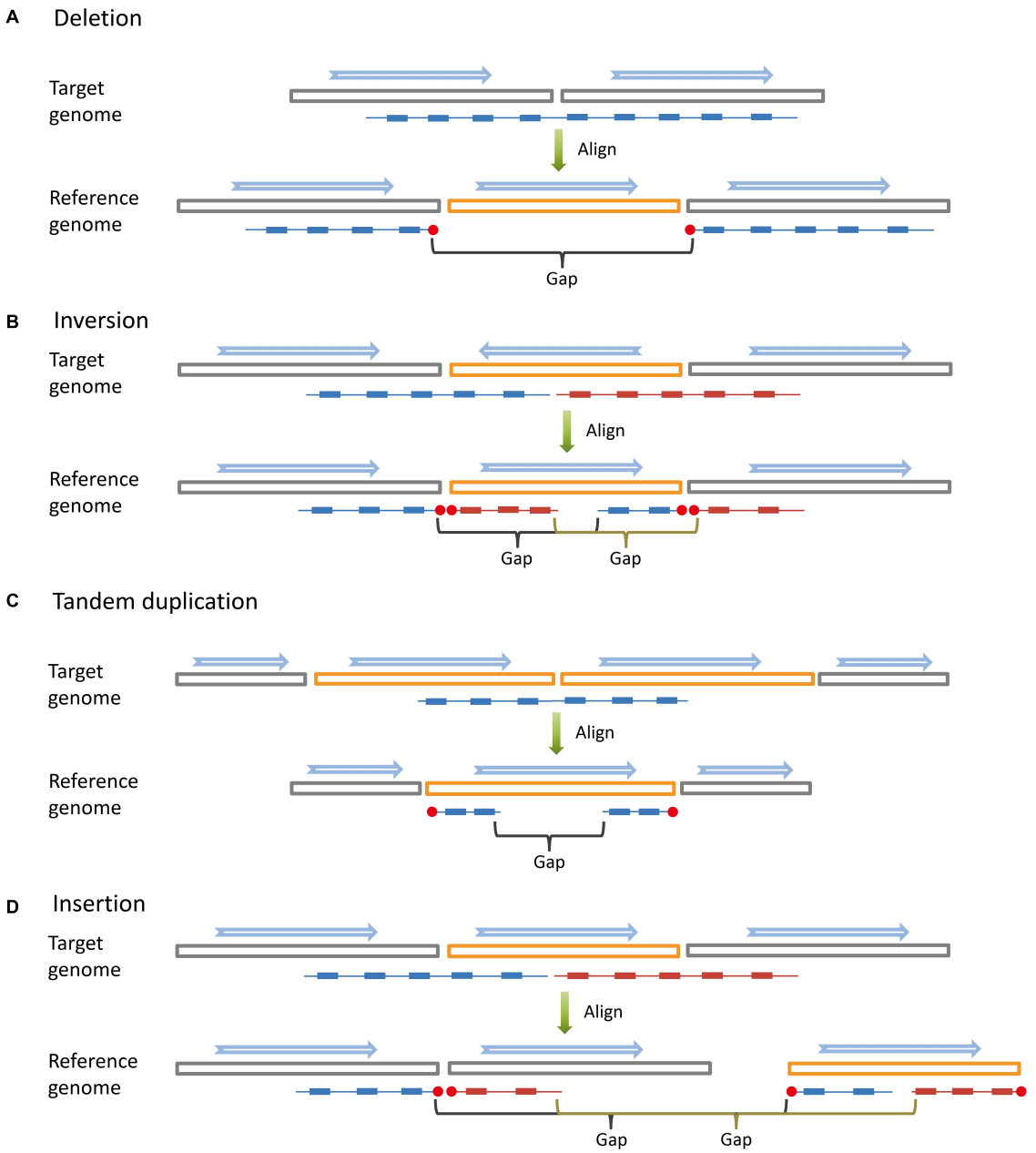

FIGURE 1 | Long DNA fragments (colored lines) are constructed by read pairs (small solid blocks) that share the same barcode. When aligned to the reference genome, long DNA fragments covering large structural variations are broken into sub-fragments by large gaps. The blue arrows indicate the directions of genome sequences (big hollow blocks). **(A)** Deletion. **(B)** Inversion. **(C)** Tandem duplication. **(D)** Insertion.

and one of the pairs is very likely to be the two ends of a sub-fragment, we unpair them.

Pair Down Candidates

Because the DNA molecules are partially sequenced, sub-fragment ends do not gather densely around a novel adjacency. They may spread in several bins and give rise to multiple end pairs. According to the gap size distribution mentioned above, a size of *Nmerge* is chosen with *P* set to 93%. To reduce redundancy, for each end pair, we recursively compare its common barcode number with that of pairs in the same type within a range of *Nmerge*, retain a representative end pair with the highest common barcode number, and refine the positions (**Figure 2D**).

Split by Haplotypes

Approximately 60% of reads can be haplotype solved, which means that those reads along with their barcodes are placed onto one of the haplotypes of each phasing block (**Figure 2E**). Thus,

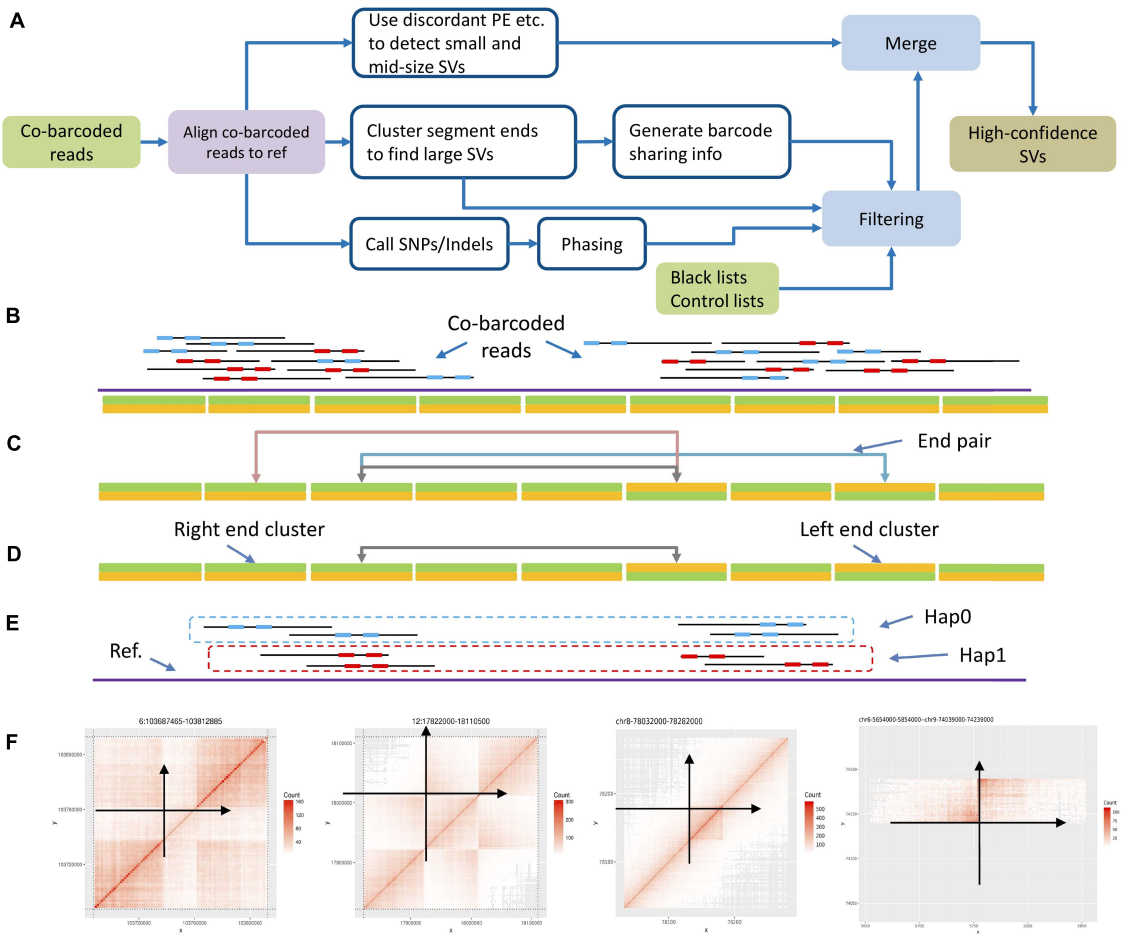

FIGURE 2 | Workflow and algorithm. **(A)** Structural variation detection workflow. **(B)** Cluster segment ends by bins: left end cluster and right end cluster. **(C)** Pair up ends by shared barcodes. **(D)** Pair down candidates by removing those nearby. **(E)** Split into haplotypes by phasing barcodes on phasing blocks. **(F)** Use barcode sharing heatmap pattern as a filter, and anchor the variation on the genome. Each point in the heatmap represents the shared barcode number at the corresponding *X*-axis and *Y*-axis positions by color depth.

the merged end pairs are checked and screened by the phasing info of their common barcodes. First, each end pair is assigned to a haplotype according to the haplotype of the common barcodes. The end pair without sufficient phased common barcodes will be assigned to one haplotype randomly. Then, the end pairs assigned to the same haplotype and sharing the same end cluster are gathered and sorted by the number of common barcodes in descending order. Finally, only the pair with the most common barcodes will be kept, because for one end cluster, a true novel adjacency only forms one end pair on the same haplotype.

Filter

Noisy signals often result in false novel adjacencies. The following noise filters can mitigate this problem.

Common Barcode Heatmap

The first filter uses the common barcode heatmap around each novel adjacency region (**Figure 2F**). A novel adjacency increases the number of common barcodes. This increase shows specific patterns in the regions in close proximity to the novel adjacency on the heatmap. Because this is not a graphic detector, we digitize the heatmap to reveal patterns. Horizontal and vertical directions intersect at the breakpoints on the heatmap, which forms four regions. For a deletion, insertion, or duplication, there is only one region showing typical adjacency barcode sharing. For an inversion, there are two regions with symmetric sharing. We collect bin-to-bin barcode sharing numbers in each region and use the Wilcoxon signed-rank test to verify the expected patterns between each two of the four regions.

Common Barcode Phase

This filter uses the phase info of the common barcodes. For each novel adjacency, if the proportion of phased common barcodes is greater than 75%, the numbers of barcodes phased to each of the two haplotypes are checked using Fisher's exact test against ideal (1|0), (0|1), and (1|1) zygosity cases. For a true novel adjacency, only one case should be significantly matched with a distinct *P*-value.

Anchor the Breakpoints

If a novel adjacency is formed by a pair of ends that are distant from each other on the reference, we would like to know whether this rearrangement results in a short interruption or a long-range SV.

Due to the limited DNA fragment lengths, the numbers of shared barcodes decrease gradually in bins further from the novel adjacency. When end pairs are placed on the target genome, they all present as a left end and a right end. If we check the common barcode numbers between the bin holding the right end and the bin holding the left end and each of the bins following the left, the common barcode numbers should show a gradual decline. We calculate the fading rates and the counts by which the observed numbers exceed the expected numbers according to the distribution $f(d)$ described above. The process is similar to that of the left end. For each end pair, we have two lists of deviations and fading rates. The end pairs are then tested by a Wilcoxon signed-rank test to detect the asymmetry of fading in both directions and a sudden loss of barcode sharing in one direction. If there is evidence of asymmetry or short-range extension, we infer that a short sequence from a distance was inserted into one direction and assign a low confidence score. Otherwise, a high confidence score is assigned. This estimation is more accurate for haplotype-solved novel adjacencies.

Map Quality

The read mapping qualities are checked within the range of *Nmerge* around the two ends of each pair, and the pairs are screened out if the low-quality ratio is above a set cut-off. A low confidence score will be given if the percentage of reads with low mapping quality is greater than 50.

Read Pairs

For regular paired-end sequencing data, the insert size of a read pair is important evidence for SV detection. The read pairs with an abnormal insert size are also checked for a novel adjacency. There is a corresponding relationship between the adjacency end orientation and the paired-end map orientation: right–left vs. forward–reverse, left–right vs. reverse–forward, left–left vs. reverse–reverse, and right–right vs. forward–forward. Four types of abnormal read pairs are counted to evaluate whether they match or conflict with the adjacency type. Additionally, if there is a match, the resolution of the adjacency will be refined from an *Nmerge* size to a normal paired-end insert size.

Black and Control Lists

Candidate pairs are filtered out in the problematic regions of the reference. These regions are defined as *black regions*, which are formed based on the reference profile and usually involve repeat sequences, mis-assembled areas, and gaps. Moreover, another set of regions defined as *control regions* is also used to filter the candidates. The *control regions* contain segmental duplications, high population frequency, and other systematic SV regions caused by the aligner, sequencer, library method, etc.

Finally, a comprehensive confidence score is generated based on the confidence scores from the filters. Then the adjacencies with high comprehensive confidence scores will be passed to downstream steps.

Merge

We extract variants below a cut-off size from smoove results and those above this cut-off from stLFRsv results and combine them by merging those with significant overlap (at least 70% overlap with respect to the longer SVs) to form the final output.

RESULTS

stLFR Co-barcoded Read Data of HG002

Data Preparation

The HG002 cell line sample was processed according to the stLFR protocol (Wang et al., 2019) and sequenced to 100× coverage. The average number of read pairs per barcode was 51. The inferred weighted fragment length was 83 kbp. The inferred mean number of fragments per barcode was 1.15. The distributions of read pair numbers, weighted fragment lengths, and fragment number per barcode are illustrated in **Supplementary Figure 1**. We down-sampled the data to 50× and 30× and called variants separately to provide guidance for applications. stLFRsv was assessed on the HG002 genome in manual parameter mode against the following four SV callers: Long Ranger, NAIBR, smoove, and GROC-SVs (Spies et al., 2017; Elyanow et al., 2018; Marks et al., 2019). The results from co-barcoded reads were also compared with SVs from 100× Nanopore long reads. The commands used to run the following pipelines are shown in **Supplementary Table 1**.

Structural Variation

The workflow of structural variation detection is illustrated in **Figure 2A**. Co-barcoded reads were aligned to hs37d5 by BWA-MEM2 (Li, 2013[Preprint]; Vasimuddin et al., 2019). Phasing was performed by HapCUT2 after SNPs were called using GATK (McKenna et al., 2010; Edge et al., 2017). The GIAB v0.6.2 structural variation set includes 7,172 insertions and 5,336 deletions. We used Truvari to align pipeline calls to the GIAB call set[1]. For Long Ranger, the alignment was performed by Lariat. For other software, the alignment results by BWA-MEM2 were used.

Seventy-nine large deletions were identified by stLFRsv. Thirty-seven of these were validated by the GIAB call set with the quality flag "PASS." Among the 42 unmatched deletions, 12 overlap with the GIAB deletion records but were failed by Truvari because of the overlap ratio. Twenty-six of the unmatched deletions overlap with the GIAB deletions with markers other than "PASS" (**Supplementary Table 2**). One is located at Chr12:11,216,856–Chr12:11,247,708 (**Figure 3C**). Several confusing signals were observed at the start of this deletion in both the co-barcoded reads and Nanopore long-read mapping results. Thus, the Nanopore assembly sequence was compared with the reference sequence. The result shows that there are two approximately 20 kbp segment duplications near

[1]https://github.com/spiralgenetics/truvari

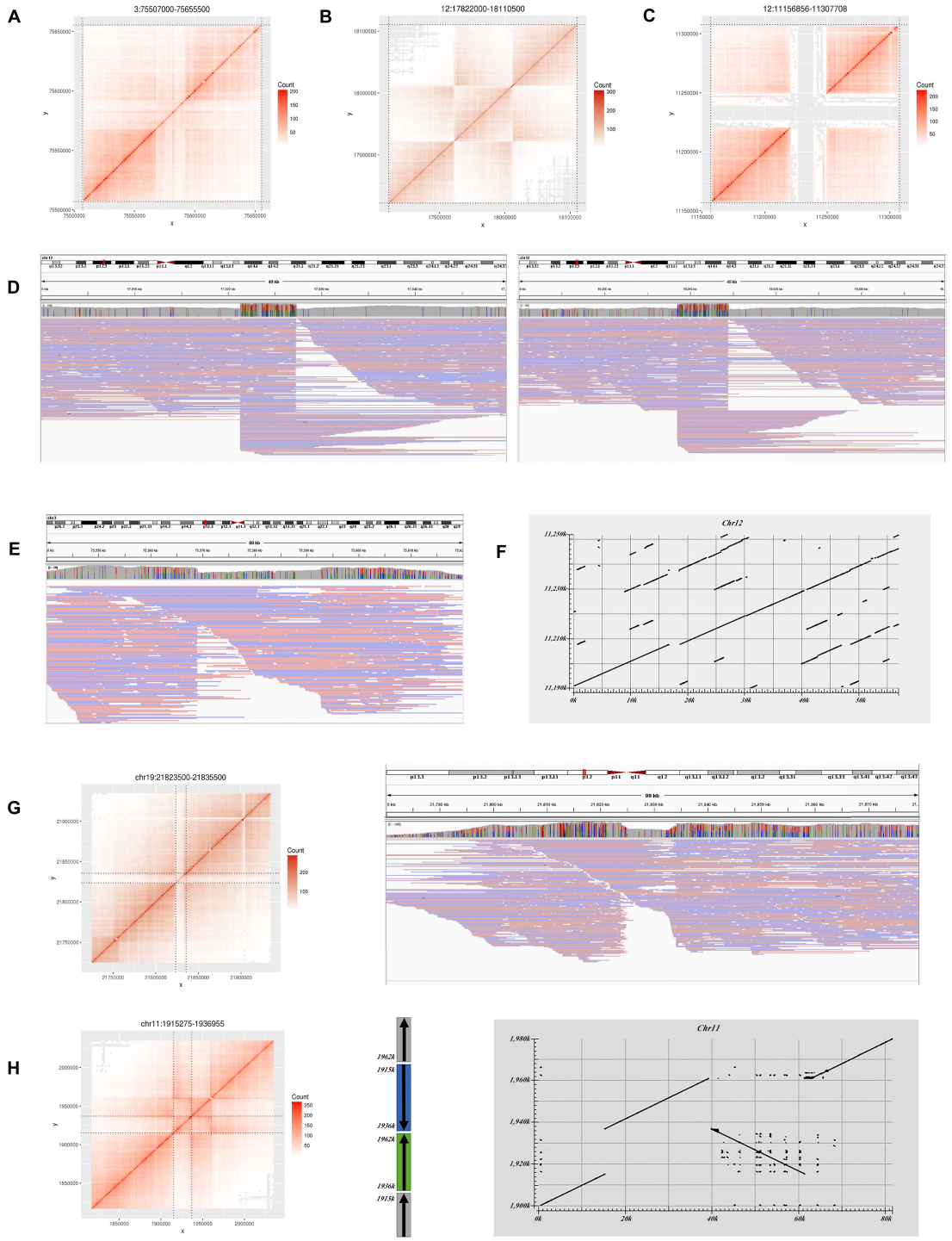

FIGURE 3 | Large variations do not match the GIAB benchmark in HG002. **(A)** Heatmap for a deletion on Chr3. **(B)** Heatmap for an inversion on Chr12. **(C)** Heatmap for a deletion on Chr12. **(D)** Long read alignment supports the inversion in **(B)**. **(E)** Long read alignment supports the deletion in **(A)**. **(F)** Assembly alignment to reference by Blast for the deletion in **(C)**. **(G)** Heatmap for a deletion on Chr19 and long read alignment. **(H)** Heatmap and structure for an inversion on Chr11 and assembly alignment.

the start and the end of this region. The downstream region is highly matched with the *hs37d5* decoy sequence, which explains the detection of this deletion (**Figure 3F**).

Four deletions do not overlap any GIAB record. Two were marked with "COMMON" by the *control list*, and the other two were marked with "PASS." Only the "PASS" two were

confirmed in the Nanopore long reads results. One is located at Chr3:75,567,000–75,595,500 and supported by Nanopore long reads (**Figures 3A,E**). The other is located at Chr19:21,822,000–21,835,500 and inferred as a heterozygous variant by long reads (**Figure 3G**).

Only three GIAB deletions larger than 10 kbp were not detected by stLFRsv, and the heatmaps for these deletions are shown in **Supplementary Figure 3A**. Two of these are in the N-regions of the reference on ChrX, and they were filtered out by the *Black list*. The third deletion is a heterozygous deletion and was undetected because the length of the DNA fragment between this deletion and the following homozygous deletion is too short for co-barcode SV detection.

In addition to deletions, stLFRsv identified 55 inversions, duplications, and translocations (**Supplementary Table 3**). Most of these are shared by multiple genomes, which indicate problematic reference regions or repeat sequences on the reference and were marked on the *Control list*. Some of them are caused by the alignment characteristics of short reads that could not be confirmed by Nanopore long reads. Others may indicate the difference between the reference and the population. For example, two inversions were also observed in HG001/NA12878 and some other samples. One has a typical inversion structure on the heatmap and was found at Chr12:17,922,000–Chr12:18,013,500 (**Figures 3B,D**). It was classified to be a homozygous variant and confirmed by Nanopore reads. The other has a more complex dual-inversion structure in which a sub-fragment of an inverted fragment reversed again and was confirmed by long read assembly (**Figure 3H**).

Furthermore, deletions (>10 kbp) that were not detected by stLFRsv but were detected by other co-barcoded read-based SV callers are listed in **Supplementary Table 4**. There are 40 deletions in total, 12 from Long Ranger, 5 from GROC-SVs, and 23 from NAIBR. Approximately 50% of these deletions were observed in stLFRsv but were filtered by the region filter (*Black list*). None were validated by the GIAB call set with a quality flag "PASS" except the three deletions mentioned above (**Supplementary Figure 3A**). Twenty-eight of these deletions are likely the result of improper short read alignments, and another eight do not overlap with GIAB call set records. One deletion at Chr8:8,032,452–Chr8:8,045,361 was chosen to evaluate the difference between the regular aligner BWA-MEM2 and the co-barcode aware aligner Lariat (aligner of Long Ranger pipeline). As shown by the heatmaps in **Supplementary Figure 3B**, although the improper alignments causing a deletion call in a complex region were corrected to a certain degree, Long Ranger still marked it as a reliable deletion. Despite its preferable performance in NGS "dead zone" genes (Mandelker et al., 2016; Marks et al., 2019), the co-barcode-aware aligner does not seem to provide significant improvements on large and complex genomic regions.

When merging deletions from stLFRsv and smoove, the size cut-off was set to 10 kbp by stLFRsv based on the data profiles. The deletion evaluation results are shown in **Table 1**. The down-sampled results are in **Supplementary Table 5**. Because few insertions were found by any of the four callers, we did not evaluate insertion results.

Unlike stLFRsv, Long Ranger, and GROC-SVs combine the co-barcode information with a local assembly strategy, which enables them to detect SVs around short sequences with high-quality alignments, such as the deletion on Chr2 shown in **Supplementary Figure 3A**, but with lower sensitivity. In contrast, NAIBR is based on a model using paired-end discordance along with co-barcode information. This model leads to higher sensitivity, especially for SVs with small size or around N-regions, such as the deletions on ChrX shown in **Supplementary Figure 3A**, but it also suffers from more false-positive SVs.

Testing Built-in Parameter Setting on Multiple HG002 Libraries

If not specified, stLFRsv offers an auto parameter mode to estimate parameters according to the following data profiles: distribution of DNA fragment length and inter-read-pair gap length. As mentioned in section "Methods," "Large-gap" size *G* to break fragment into sub-fragment, bin size *B* to cluster sub-fragment borders, and merging size *Nmerge* to merge bins into a single breakpoint are chosen based on inter-read-pair gap length distribution. These three parameters then determine the sensitivity of the pipeline and the accuracy of the breakpoint locations. In contrast, the sizes of inversion and duplication that stLFR is able to identify are dictated by the DNA fragment length distribution. Long DNA fragments only detect large inversions and duplications. The detectable deletion size should be larger than the "large-gap" size *G*.

For the HG002 cell line sample, we constructed four stLFR libraries to assess the influence of the data profile. Only high-quality reads (>4 read pairs per segment and >8 read pairs per barcode) were retained for statistical analysis. The data statistics and inferred parameters for these four HG002 libraries are illustrated in **Table 2** and **Supplementary Figure 4**. It is highly recommended, according to our tests, not only for stLFRsv but also for other co-barcoded read-based SV callers that stLFR data should have a high-quality read ratio >70%, average read pairs per segment >25, and barcode conflict <1.7 for good detection performance.

Comparison With Nanopore Long Reads

We obtained 100× Nanopore long reads of HG002 from Oxford Nanopore Technologies. The distribution of read length and percent identity are presented in **Supplementary Figure 1**. The alignment was performed with Minimap2 using default parameters (Li, 2018). SVs were detected by Sniffles with default parameters, and increasing the support read number can reduce both false positives and true positives (Sedlazeck et al., 2018). The deletion evaluation is also listed in **Table 1**. The insertion evaluation is shown in **Supplementary Table 6**. We assembled long reads by NECAT for variation validation (Chen et al., 2020).

For deletions, Nanopore long reads achieve a high sensitivity in every size level along with a number of false-positive deletions. stLFRsv attains approximately the same level of sensitivity with a lower false-positive rate for large deletions. For insertions,

TABLE 1 | Deletion evaluation on whole genome against GIAB HG002 benchmark.

		100× long reads		100× co-barcoded reads				
	Mapping	Sniffles Minimap2	Long Ranger lariat	NAIBR bwamem2	stLFRsv bwamem2	smoove bwamem2	stLFRsv + smoove bwamem2	GROC-SVs bwamem2
50–1 k	Benchmark				4,719			
	Total call	9,453	3,583	2	0	972	972	0
	TP	4,168	2,304	2	0	724	724	0
	FP	5,285	1,279	0	0	248	248	0
	FN	551	2,415	4,717	4,719	3,995	3,995	4,719
	Precision	44.09%	64.30%	100.00%	–	74.49%	74.49%	–
	Recall	88.32%	48.82%	0.04%	–	15.34%	15.34%	–
1 k–10 k	Benchmark				577			
	Total call	902	489	155	13	554	556	0
	TP	533	391	125	12	434	436	0
	FP	369	98	30	1	120	120	0
	FN	44	186	452	565	143	141	577
	Precision	59.09%	79.96%	80.65%	92.31%	78.34%	78.42%	–
	Recall	92.37%	67.76%	21.66%	2.08%	75.22%	75.56%	–
10 k–30 k	Benchmark				31			
	Total call	60	27	31	56	35	56	9
	TP	28	19	24	30	22	30	7
	FP	32	8	7	26	13	26	2
	FN	3	12	7	1	9	1	24
	Precision	46.67%	70.37%	77.42%	53.57%	62.86%	53.57%	77.78%
	Recall	90.32%	61.29%	77.42%	96.77%	70.97%	96.77%	22.58%
>30 k	Benchmark				9			
	Total call	55	14	28	23	36	23	13
	TP	9	6	8	7	7	7	4
	FP	46	8	20	16	29	16	9
	FN	0	3	1	2	2	2	5
	Precision	16.36%	42.86%	28.57%	30.43%	19.44%	30.43%	30.77%
	Recall	100.00%	66.67%	88.89%	77.78%	77.78%	77.78%	44.44%

TABLE 2 | Detection capability and estimated parameters of different HG002 libraries.

Library		HG002-1	HG002-2	HG002-3	HG002-4
Input DNA amount		1 ng	1 ng	1.5 ng	1.5 ng
Reads count		2,525,286,352	3,029,968,430	2,172,780,252	2,994,596,020
Average sequencing depth (after duplication removed)		44.34	35.77	46.73	44.38
High-quality read ratio		89.57%	78.03%	79.15%	75.55%
Read pairs per segment		32.33	18.30	18.40	17.21
Barcode conflict (segments per barcode)		1.55	1.41	2.04	1.70
Estimated parameters	*B* (bp)	1,500	1,500	2,500	1,900
	Nmerge (*B*)	4	4	4	4
	G (bp)	13,100	13,900	22,200	13,800
Detection capability	Deletion (bp)	13,500	13,500	22,500	13,300
	Inversion/duplication (bp)	48,100	28,600	46,700	32,200

Nanopore long reads show the same performance as deletion detection with small insertions but fail for large insertions just like stLFR and the other three SV callers. This result is consistent with a previous report (Fang et al., 2019) showing that the detection of large size insertion may remain a challenge for alignment-based SV callers.

Resource Usage

The resource utilization of these four callers was collected by the Linux system tool "time" (**Figure 4**), and all tests were performed on a workstation with 48 CPU cores and 256 GB memory. GROC-SVs ran very slowly because of massive assembly operations. NAIBR showed an extremely high memory

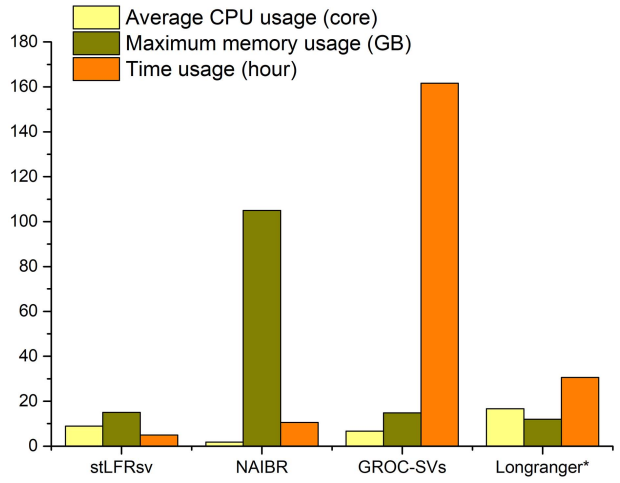

FIGURE 4 | The resource usage of four pipelines when processing 100× stLFR reads of HG002 with the given parameters in **Supplementary Table 1**. *The resource usage of Long Ranger is estimated by the log file because it is a fully integrated functional pipeline.

consumption with a low CPU load. Benefitting from the algorithm focusing only on the sub-fragment divided by large gaps, stLFRsv achieved the best performance with regard to time and memory usage while taking full advantage of the multi-core CPU.

10× Genomics Linked-Reads Data of HG001

The Linked-Reads data of HG001 downloaded from 10× Genomics official website was tested on all four co-barcoded read-based SV callers (for stLFRsv, the 10× Genomics barcode BX tag was converted to an stLFR-formatted barcode). Because there is not a well-characterized GIAB call set for HG001, only the large SVs were compared among the call sets and with the 10× Genomics SV results on the website.

There are 34 reliable large SVs in 10× Genomics Long Ranger call set, comprising 18 deletions, 12 duplications, and 4 inversions. To validate these SVs, they were checked by the heatmaps of co-barcode distribution manually and individually. The results are shown in **Figure 5A** and **Supplementary Table 7**. The performance of the four SV detectors on Linked-Reads data is consistent with that on stLFR data. stLFRsv has the highest consistency with each of the other three call sets. NAIBR presents the most SVs not detected by any other caller, and GROC-SVs has the least common SVs. Four deletions only detected by stLFRsv were all marked "COMMON" and also found in low-quality results in the Long Ranger call set. As for the duplications, only four duplications were confirmed as reliable variants, and they were all detected by read depth information without SV breakpoint details. The other three call sets provide minimal support for these duplications. All four inversions were detected by stLFRsv, three of which were marked "COMMON" and also found in HG002 results. The remaining inversion is a "DUP-INV" complex SV found only in HG001. As shown in **Figure 5B**, a DNA fragment was duplicated and inversely inserted into another genomic position of the same chromosome. Both breakpoints were detected by stLFRsv, whereas only one was reported by Long Ranger.

10× Genomics Linked-Reads Data of HX1

A Chinese individual, HX1, was studied and sequenced in several investigations (Shi et al., 2016; Fang et al., 2019), and a reliable SV call set of HX1 was established by SMRT-SV (Audano et al., 2019), a widely used long-read SV caller. Thus, stLFRsv was also tested on the Linked-Reads data of HX1. As stated in a previous report (Fang et al., 2019), duplications were barely detected by co-barcoded reads or long reads, and thus we only focused on deletions. The SMRT-SV call set has 16 large deletions (>10 kbp), 10 of which were detected by stLFRsv (**Supplementary Table 8**). The failure to detect six deletions may be the result of imprecise SV positions by stLFRsv. In other words, the size of these six deletions reported by stLFRsv may be smaller than 10 kbp, and they were accordingly found in the intermediate result file. Another deletion at Chr2:111,153,548–Chr2:111,198,923, which was missed by SMRT-SV but validated by a previous report (Fang et al., 2019), was also detected by stLFRsv (**Supplementary Figure 3C**).

DISCUSSION

We present stLFRsv, a co-barcoded read-based structural variation detector that identifies large variants with far fewer false positives than alignment-based detectors using either short reads or long reads. stLFRsv also shows the best computational performance among co-barcoded read-based SV callers. When combined with a standard short read variation caller, stLFRsv can exploit the co-barcoded reads to reveal the full spectrum of genome polymorphism. Although stLFR has decreased the average number of DNA fragments sharing the same barcode to nearly 1 and increased the coverage in "BAD" genome regions to a certain degree, co-barcoded reads have limited resolution for structural variation calling because paired reads for a long fragment only partially cover the whole sequence with unknown order and intervening distance. In contrast, the performance of single-molecule sequencing long reads has been increasing. In spite of this, discovering both base-level and very-large-scale variants simultaneously using co-barcoded sequencing technology will be promising for some clinical applications especially with lower cost and decreased turnaround time. Moreover, the greater length of DNA fragments for co-barcoded sequencing compared with single-molecule sequencing has the potential to span larger repeat regions and catch SVs missed by real long reads.

Larger variants other than deletions and insertions are needed to assess variation detection by co-barcoded reads. There are three main aspects for our future research. First, we are analyzing co-barcoded reads for clinical samples to find pathogenic balanced/unbalanced translocations, deletions, duplications, and more complex structures. This technique is

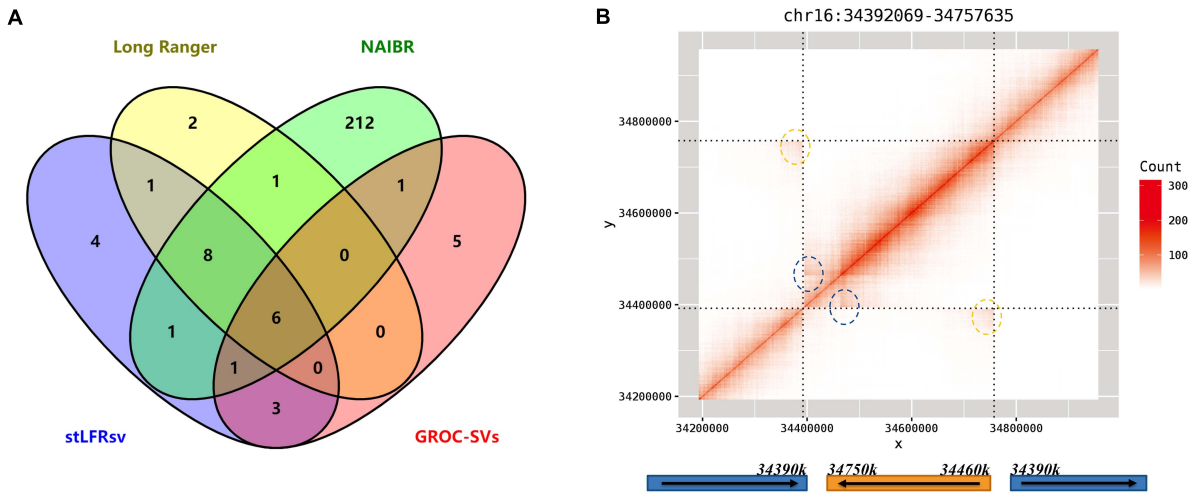

FIGURE 5 | (A) Venn diagram of detected deletions from four structural variant (SV) callers in HG001. **(B)** The heatmap and structure for a complex inversion in HG001. The signs of two breakpoints are marked by blue and yellow circles in the heatmap.

likely to provide a more precise description of such variants compared with current clinical practices by identifying more reliable breakpoints. Second, another clinical application is to associate a genetic defect with nearby alleles using co-barcoded reads, which can provide an inference as to whether an infant inherited a defect through prenatal cell-free DNA sequencing by detecting associated nearby alleles. This application benefits from the outstanding phasing ability of co-barcoded reads. The final element of future work is to add a local assembly module to enhance small variation detection (in the range of 50 bp–1 kbp).

With a cost slightly higher than standard short reads, co-barcoded reads are able to reveal much more useful information for the underlying genomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of BGI. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PL and OW conducted the co-barcoded reads library construction. JG, CS, and XC developed the pipeline and analyzed the data. XX and HY designed the research. WZ supervised and coordinated all aspects of the project. HZ wrote the manuscript. All authors revised, read, and approved the final manuscript.

FUNDING

This work was supported in part by Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014), Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011), and China National GeneBank.

ACKNOWLEDGMENTS

We would like to acknowledge the ongoing contributions and support of all BGI-Shenzhen employees.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.636239/full#supplementary-material

Supplementary Figure 1 | Basic data profile distribution for stLFR co-barcoded reads and long reads of HG002. **(A)** Read pairs per barcode. **(B)** Weighted fragment length for co-barcoded reads. **(C)** Number of fragments per barcode. **(D)** Long read length. **(E)** Long read percent identity.

Supplementary Figure 2 | An example of empirical gap size distribution of stLFR co-barcoded reads. Different selected *P*s help to decide the parameters used in the pipeline.

Supplementary Figure 3 | Heatmaps of **(A)** Three false negative large deletions in HG002. **(B)** A false positive deletion by different aligners in HG002. **(C)** A validated deletion detected by stLFRsv in HX1.

Supplementary Figure 4 | Weighted fragment length distribution for different HG002 stLFR libraries.

REFERENCES

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell* 176, 663–675.e19. doi: 10.1016/j.cell.2018.12.019

Brent, P. (2018). *Smoove*. https://brentp.github.io/post/smoove/(accessed December 24, 2020).

Chen, Y., Nie, F., Xie, S. Q., Zheng, Y. F., Bray, T., Dai, Q., et al. (2020). Fast and accurate assembly of Nanopore reads via progressive error correction and adaptive read selection. *bioRxiv* [Preprint] doi: 10.1101/2020.02.01.930107

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* 27, 801–812. doi: 10.1101/gr.213462.116

Elyanow, R., Wu, H. T., and Raphael, B. J. (2018). Identifying structural variants using linked-read sequencing data. *Bioinformatics* 34, 353–360. doi: 10.1093/bioinformatics/btx712

Fang, L., Kao, C., Gonzalez, M. V., Mafra, F. A., da Silva, R. P., Li, M., et al. (2019). LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. *Nat. Commun.* 10:5585. doi: 10.1038/s41467-019-13397-7

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97. doi: 10.1038/nrg1767

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi: 10.1038/nbt.4060

Jongmans, M. C. J., Admiraal, R. J., Van Der Donk, K. P., Vissers, L. E. L. M., Baas, A. F., Kapusta, L., et al. (2006). CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene. *J. Med. Genet.* 43, 306–314. doi: 10.1136/jmg.2005.036061

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biol.* 15:R84. doi: 10.1186/gb-2014-15-6-r84

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint] arXiv:1303.3997,

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Mandelker, D., Schmidt, R. J., Ankala, A., Gibson, K. M., Bowser, M., Sharma, H., et al. (2016). Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* 18, 1282–1289. doi: 10.1038/gim.2016.58

Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* 29, 635–645. doi: 10.1101/gr.234443.118

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meleshko, D., Marks, P., Williams, S., and Hajirasouliha, I. (2019). Detection and assembly of novel sequence insertions using Linked-Read technology. *bioRxiv* [Preprint] doi: 10.1101/551028 bioRxiv:551028,

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11:R52. doi: 10.1186/gb-2010-11-5-r52

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., et al. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* 38, 24–26. doi: 10.1038/ng1718

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15, 461–468. doi: 10.1038/s41592-018-0001-7

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* 7:12065. doi: 10.1038/ncomms12065

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., et al. (2003). [alpha]-synuclein locus triplication causes Parkinson's disease. *Science* 302:841. doi: 10.1126/science.1090278

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., et al. (2017). Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* 14, 915-920. doi: 10.1038/nmeth.4366

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* 12:e1004873. doi: 10.1371/journal.pcbi.1004873

Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). "Efficient architecture-aware acceleration of BWA-MEM for multicore systems," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (New York, NY: IEEE), 314–324. doi: 10.1109/IPDPS.2019.00041

Wang, O., Chin, R., Cheng, X., Wu, M. K. Y., Mao, Q., Tang, J., et al. (2019). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* 29, 798–808. doi: 10.1101/gr.245126.118

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. doi: 10.1038/s41587-019-0217-9

Xia, L. C., Bell, J. M., Wood-Bouwens, C., Chen, J. J., Zhang, N. R., and Ji, H. P. (2018). Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.* 46:e19. doi: 10.1093/nar/gkx1193

Zhang, F., Christiansen, L., Thomas, J., Pokholok, D., Jackson, R., Morrell, N., et al. (2017). Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol.* 35, 852–857. doi: 10.1038/nbt.3897

Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311. doi: 10.1038/nbt.3432

Zhou, X., Zhang, L., Weng, Z., Dill, D. L., and Sidow, A. (2019). Aquila: diploid personal genome assembly and comprehensive variant detection based on linked reads. *bioRxiv* [Preprint] doi: 10.1101/660605 bioRxiv:660605

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L. M., Mullikin, J. C., Xiao, C., et al. (2019). A robust benchmark for germline structural variant detection. *BioRxiv* [Preprint] doi: 10.1101/664623 BioRxiv, 664623

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

The handling editor declared a past co-authorship with the authors OW, HY, XX, and WZ.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 August 2021
doi: 10.3389/fgene.2021.708348

Chromoanagenesis Event Underlies a *de novo* Pericentric and Multiple Paracentric Inversions in a Single Chromosome Causing Coffin–Siris Syndrome

*Christopher M. Grochowski[1], Ana C. V. Krepischi[2], Jesper Eisfeldt[3,4], Haowei Du[1], Debora R. Bertola[2,5], Danyllo Oliveira[2], Silvia S. Costa[2], James R. Lupski[1,6,7,8], Anna Lindstrand[3,9] and Claudia M. B. Carvalho[1,10]**

[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, United States, [2] Department of Genetics and Evolutionary Biology, Human Genome and Stem Cell Research Center, Institute of Biosciences, University of São Paulo, São Paulo, Brazil, [3] Department of Molecular Medicine and Surgery and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, [4] Science for Life Laboratory, Karolinska Institutet Science Park, Solna, Sweden, [5] Clinical Genetics Unit, Instituto da Criança do Hospital das Clínicas, University of São Paulo, São Paulo, Brazil, [6] Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States, [7] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, United States, [8] Texas Children's Hospital, Houston, TX, United States, [9] Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, [10] Pacific Northwest Research Institute, Seattle, WA, United States

Chromoanagenesis is a descriptive term that encompasses classes of catastrophic mutagenic processes that generate localized and complex chromosome rearrangements in both somatic and germline genomes. Herein, we describe a 5-year-old female presenting with a constellation of clinical features consistent with a clinical diagnosis of Coffin–Siris syndrome 1 (CSS1). Initial G-banded karyotyping detected a 90-Mb pericentric and a 47-Mb paracentric inversion on a single chromosome. Subsequent analysis of short-read whole-genome sequencing data and genomic optical mapping revealed additional inversions, all clustered on chromosome 6, one of them disrupting *ARID1B* for which haploinsufficiency leads to the CSS1 disease trait (MIM:135900). The aggregate structural variant data show that the resolved, the resolved derivative chromosome architecture presents four *de novo* inversions, one pericentric and three paracentric, involving six breakpoint junctions in what appears to be a shuffling of genomic material on this chromosome. Each junction was resolved to nucleotide-level resolution with mutational signatures suggestive of non-homologous end joining. The disruption of the gene *ARID1B* is shown to occur between the fourth and fifth exon of the canonical transcript with subsequent qPCR studies confirming a decrease in *ARID1B* expression in the patient versus healthy controls. Deciphering the underlying genomic architecture of chromosomal rearrangements and complex structural variants may require multiple technologies and can be critical to elucidating the molecular etiology of a patient's clinical phenotype or resolving unsolved Mendelian disease cases.

Keywords: genomic inversions, structural variation, complex genomic rearrangement (CGR), chromothripsis, chromoplexy, microhomology-mediated break-induced replication (MMBIR)

OPEN ACCESS

Edited by:
Zirui Dong, The Chinese University of Hong Kong, China

Reviewed by:
Carlos Córdova-Fletes, Universidad Autónoma de Nuevo León, Mexico
Orsetta Zuffardi, University of Pavia, Italy

****Correspondence:***
Claudia M. B. Carvalho
ccarvalho@pnri.org

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *11 May 2021*
Accepted: *23 July 2021*
Published: *26 August 2021*

Citation:
Grochowski CM, Krepischi ACV, Eisfeldt J, Du H, Bertola DR, Oliveira D, Costa SS, Lupski JR, Lindstrand A and Carvalho CMB (2021) Chromoanagenesis Event Underlies a de novo Pericentric and Multiple Paracentric Inversions in a Single Chromosome Causing Coffin–Siris Syndrome. Front. Genet. 12:708348. doi: 10.3389/fgene.2021.708348

INTRODUCTION

Inversions are a unique class of structural variation (SV) that present at least two breakpoint junctions *in cis*. Although the majority of inversions are copy-number neutral (i.e., classical inversions), about 17% present with more complex structures accompanied with copy-number variants (CNVs) of a few bp to several kb in size (Pettersson et al., 2020). Inversion rearrangements can occur in a pericentric fashion when DNA is flipped 180° across the centromere or paracentric when the DNA inversion occurs on either the long (q) or short (p) chromosomal arm (Kaiser, 1984).

Historically, inversions were detected by cytogenetics with karyotyping; the resolution to detect such events is limited by the resolution of chromosomal G-banding (approximately 5–10 Mb). Routine genomic testing including array comparative genomic hybridization (aCGH) and exome sequencing (ES) will not detect most inversion events given that they are typically: (1) copy-number neutral and (2) usually do not have breakpoints within the coding regions targeted by ES (Posey, 2019; Lupski et al., 2020). The advent of short-read whole-genome sequencing (WGS) enabled detection of inversion events, though the rate of false-positives (Vicente-Salvador et al., 2017) as well as false-negatives is very high, the latter due to lack of detection of inversions with breakpoints within repetitive regions (Chaisson et al., 2019). Recently, long-read DNA sequencing, e.g., Oxford Nanopore and PacBio, and genomic optical mapping, e.g., Bionano, as well as Strand-seq have resulted in increased sensitivity of inversion detection as they allow accurate genotype and phasing of events with multiple breakpoints junctions *in cis*, including those mapping to genomic repeats (Ebert et al., 2021).

In the constitutional genome, inversions have been shown to be formed through three different molecular mechanisms sometimes acting in concert (Pettersson et al., 2020). Non-allelic homologous recombination (NAHR) is one driver of inversion formation when breakpoints are found to be part of a pair of inverted genomic segments sharing sequence homology (Flores et al., 2007; Kidd et al., 2008). Micromology-mediated end joining (MMEJ) or non-homologous end joining (NHEJ) are the most likely mechanisms generating inversions with breakpoints presenting very little or no microhomology (Pettersson et al., 2020). For copy-number associated inversions observed in complex genomic rearrangements (CGRs), replicative mechanisms, such as microhomology-mediated break-induced replication (MMBIR) play a role in the inversion formation process (Lee et al., 2007; Carvalho et al., 2011; Beck et al., 2015; Gu et al., 2015; Pettersson et al., 2020). As inversions can be formed by one or more molecular mechanisms, each individual case must be resolved to nucleotide-level resolution to infer the molecular mutational mechanism(s) that may have been involved.

Inversion formation can cause gene disruptions and amplifications and have been implicated in the evolution of novel genes and "exonization" of gene structures (Lakich et al., 1993; Carvalho et al., 2011; Zuccherato et al., 2016). Gene interrupting inversions are implicated in some genomic disorders most notably an inversion physically separating parts of the *F8* gene, the most common cause of severe hemophilia A (Lakich et al., 1993). The pathogenetic consequence of this type of structural variant may result from a breakpoint occurring within the exon of a gene or in an intragenic fashion between exons (Feuk, 2010); the end result is a gene split apart disrupting its function (Lakich et al., 1993). More cryptically, inversions may disrupt enhancer or topologically associated domains surrounding a gene, causing no change in the gene itself but leading to a pathogenic consequence through change in gene expression, a potential position effect, or other perturbations of gene regulation (Lupianez et al., 2015; Kraft et al., 2019; Sanchez-Gaya et al., 2020).

Herein, we present a patient with Coffin–Siris syndrome 1 (CSS1) and multiple inversions affecting a single chromosome. Complex structural variants have been shown to present a challenge for detection as well as molecular and genomic characterization partly due to the inability to properly phase detected variants, as well as subsequent clinical interpretation of potential contribution of variant effects to observed clinical phenotype(s) (Grochowski et al., 2018; Eisfeldt et al., 2020; Plesser Duvdevani et al., 2020). To experimentally dissect the genomic architecture of the rearranged chromosome 6 of this patient, and to explore whether genes involved in the rearrangement contributed to the observed clinical traits, we employed several technologies including karyotyping (G-banding), fluorescence *in situ* hybridization (FISH), quantitative PCR (qPCR), aCGH, WGS, and genomic optical mapping in this study. The convergence of experimental approaches allowed for DNA base-pair resolution of the genomic inversion rearrangements and revealed that an inversion caused disruption of the gene *ARID1B*, explaining the clinical phenotype in this patient. Furthermore, our studies revealed a rare chromoanagenesis event constituted by multiple copy-number neutral inversions.

MATERIALS AND METHODS

Patient Enrollment

The affected proband and unaffected sister, mother, and father were evaluated and characterized at the University of São Paulo (Protocol 2.589.398). The trio (proband, mother, and father) were subsequently enrolled under a protocol approved by the institutional review board at Baylor College of Medicine (IRB #: H-29697). Genomic DNA was extracted from peripheral blood using standard protocols.

Conventional Karyotyping and Cytogenomic Studies

GTG-banding karyotypes from cultured peripheral blood lymphocytes were obtained following standard protocols (**Supplementary Figure 1**). FISH on metaphase chromosomes was implemented using bacterial artificial chromosome (BAC) DNAs from the 1-Mb clone set[1] mapped to the long arm of

[1]http://www.ensembl.org/

chromosome 6 (RP11-506N21, RP3-336G18, and RP11-266C7). Metaphase spreads were analyzed using a Zeiss fluorescence microscope and processed using ISIS software (MetaSystem). At least 20 metaphase spreads from the patient and her parents were analyzed.

Array Comparative Genomic Hybridization (aCGH)

Initial aCGH analyses were performed using a 180K genome-wide Agilent array. A subsequent custom 180K Agilent high-resolution array was designed to interrogate both the long and short arm of chromosome 6 (AMADID#: 086000) using the Agilent e-array website[2] (Santa Clara, CA, United States) with a median probe spacing of 857 bp maximally spaced across the entire chromosome 6. Array experiments were conducted following protocols set forth by Agilent in relation to hybridization and labeling with minor modifications (Carvalho et al., 2009; **Supplementary Figure 2A**).

Short-Read WGS

Short-read WGS was performed using Illumina 30× PCR-free paired-end (PE) DNA sequencing (Hofmeister et al., 2018) at the National Genomics Infrastructure (NGI), in Stockholm, Sweden. All data obtained were processed using NGI-piper and analysis for structural variants was performed using the FindSV pipeline[3] (**Supplementary Figure 2B**). FindSV combines CNVnator V.0.3.2 (Abyzov et al., 2011) and TIDDIT V.1.1.4 (Eisfeldt et al., 2017) and produces a single variant calling format (VCF) file, subsequently annotated by variant effect predictor (VEP) and filtered based on the VCF file quality (McLaren et al., 2010). Lastly, the VCF file is sorted based on a local structural variant frequency database consisting of 351 personal genome samples of well-characterized healthy and affected individuals, and the SV of interest was identified based on the VEP annotation and variant frequency. Manual inspection and identification of split reads was performed using the Integrative Genomics Viewer (IGV)[4] (Robinson et al., 2011). Exact genomic map positions of breakpoints, at the nucleotide level, could then be determined by alignment of split reads to the Hg19/GRCh37 reference genome using the BLAST-like alignment tool (BLAT)[5] (Kent, 2002). Single-nucleotide variants (SNVs) overlapping the inversions were extracted using Tabix (Li, 2011). SNVs were called as previously described (Pettersson et al., 2020), and the resulting call sets were filtered for *de novo* SNV using BCFtools (Li et al., 2009). *De novo* and inherited SNV and indels were filtered and annotated based on the mutation identification pipeline (MIP) clinical workflow and sorted based on allele frequency, variant consequence, and CADD score.

qPCR Gene Expression Analysis

Total mRNA was extracted from peripheral blood using the RNeasy mini kit (Qiagen) following the manufacturer's instructions. After evaluating RNA integrity and concentration with a NanoDrop spectrophotometer (Thermo Fisher Scientific), 1 μg of RNA was used for cDNA synthesis with a SuperScript III First-Strand Synthesis System and oligo-dT primers (Thermo Fisher Scientific). Real-Time qPCR (RT-qPCR) experiments were performed in triplicate in a 7500 Fast Real-Time PCR System, using SYBR Green PCR Master Mix (Thermo Fisher Scientific). Primers for *ARID1B* were guided and designed using Primer3 software (forward: 5′ GGCCGTCCCGGAGTTTAATAA 3′ and reverse: 5′ CGGAGTGCATCATCCCCAT 3′), with efficiency being evaluated by serial cDNA dilutions This primer set targets a region of exon 1 in *ARID1B* of the transcript NM_001374820.1. The endogenous control *GAPDH* was used as a normalizing factor for each sample (primers: forward: 5′ GCATCCTGGGCTACACTG 3′ and reverse: 5′ CCACCACCCTGTTGCTGTA 3′). Unpaired *t*-test was applied in the statistical analyses, through SPSS V22 software.

Genomic Optical Mapping

High molecular weight (HMW) genomic DNA for use in genomic optical mapping was extracted by Histogenetics (Ossining, NY, United States) from whole blood using the Bionano Prep Blood and Cell Culture DNA Isolation Kit (Bionano Genomics). Subsequent DNA quantity and size were confirmed using a Qubit dsDNA BR Assay Kit. A total of 0.75 μg of HMW DNA was then labeled using the Bionano Prep direct label and stain (DLS) method (Bionano Genomics) and loaded onto a flow cell to run on the Saphyr optical mapping system (Bionano Genomics) (**Supplementary Figure 2C**). Approximately 230–370 Gb of data were generated per run. Raw optical mapping molecules in the form of BNX files generated from a diploid genome were parsed through a preliminary bioinformatic pipeline that filtered out molecules less than 150 kb in size and with less than nine motifs per molecule to generate a *de novo* assembly of the genome maps. Data were then aligned to an *in silico* reference genome (Hg38/GRCh38) using the Bionano Solve v3.5 RefAligner module. Structural variant calls were generated through comparison of the reference genome using a custom Bionano SV caller. Manual inspection of proposed breakpoint junctions was then visualized in the Bionano Access software program v1.5.1.

Bionano SV Analysis

Optical mapping was run on the Saphyr platform[6] at Bionano Genomics (San Diego, CA, United States). The optical maps were analyzed using the Bionano-solve pipeline[7]. Briefly, the maps were detected using AutoDetect, and assembled using the *de novo* assembly package AssembleMolecules. The resulting consensus maps were aligned to Hg19/GRCh37 using the Bionano RefAligner. Lastly, the variants of interest were visualized using Bionano Access, and the resulting smap files were converted to VCF using a custom version of the smap2vcf script[8]. *De novo* SVs were discovered by

[2]http://earray.chem.agilent.com/earray/

[3]https://github.com/J35P312/FindSV

[4]http://software.broadinstitute.org/software/igv/

[5]https://genome.ucsc.edu/cgi-bin/hgBlat

[6]https://bionanogenomics.com/support-page/saphyr-system

[7]https://bionanogenomics.com/support-page/bionano-solve

[8]https://github.com/J35P312/smap2vcf

merging these VCF files into a single trio-VCF. The SVs were merged using SVDB v2.3.0, and variants unique to the proband were discovered using the GNU grep tool (Eisfeldt et al., 2017).

De novo GATK Filtering

Individual germline SNVs and indels were called using GATK (v.4.1.3) (McKenna et al., 2010). Of note, "-GVCF" option was used for GATK haplotypecaller, which outputs a gVCF file that includes reference or variant information for all loci. The gVCF files for a family were combined and the proband's genotype was recalibrated based on parental genotype per Mendel's laws of allele transmission. Using recalibrated posterior genotype probabilities, possible *de novo* mutations were tagged. All possible *de novo* variants were filtered by an in-house developed software called DNM (*de novo* mutation)-Finder[9] that combines GATK and xAtlas (Eldomery et al., 2017).

Chromosome Rearrangement Simulation

A Monte Carlo simulation to test the likelihood of chromosomal breakpoints occurring in specific locations was designed to mirror the rearrangement observed in this patient. Briefly, the base pairs encompassing chromosome 6 (chr6:1-171,115,067) were broken into seven segments with only the first and last segment being positionally static. The remaining five segments could be randomly reshuffled with a 50% chance of inverting. The breakpoint positions of these segments were randomly and uniformly selected across chromosome 6. The simulation was run 10,000 times to statistically test for significance of clustering or enrichment of breakpoints within protein-coding genes on chromosome 6 (according to ENSEMBL release 87). The clustering of the breakpoints was assessed by computing the average distance between breakpoints; a simulated rearrangement was considered more clustered if its average breakpoint distance was smaller than the average breakpoint distance observed in the index patient. The enrichment of protein-coding genes was assessed by counting the number of breakpoint junctions carrying fusions of protein-coding genes. The scripts needed for extracting the protein coding genes and running the simulation are available on git-hub[10].

Breakpoint PCR Sequencing

The precise location of each breakpoint junction identified in the WGS data were determined and visualized with IGV. For each position, the relative strand orientation (i.e., polarity), and the genomic map position on the haploid reference human genome, of the junction was identified. Primers were designed upstream and downstream of the identified junction and PCR amplification was performed using the HotStarTaq (Qiagen) polymerase with standard conditions. Sanger-sequencing was performed at the Baylor College of Medicine Sequencing Core, and the results were visualized using the Sequencher software suite (Genecodes).

RESULTS

Pericentric and Paracentric Inversions on Chromosome 6

The 5-year-old female proband is the first child born to non-consanguineous healthy parents (29-year-old mother and 30-year-old father) at 39 weeks gestational age, i.e., full term, by cesarean section, after an uneventful pregnancy (**Figure 1A**). She has one younger sister with no history of physical or developmental abnormalities. Her birth weight was 2,345 g (<10th centile), her length was 44 cm (<10th centile), and her occipital frontal circumference (OFC) was 33.5 (50th centile). Apgar scores were 9 and 9 at 1 and 5 min, respectively. She was sent home after 3 days in the hospital. There were no major pregnancy or birth complications or any birth defects recognized on newborn examination.

The mother first noticed poor suck with hypotonia during the first week of life, evolving with poor weight gain and developmental delay: she sat unsupported at 9 months of age and crawled at 18 months. At the age of 4 years, she was not able to walk unassisted and she had not developed speech. She was evaluated by a neurologist in the first months of life and started physical therapy at 5 months of age with a treatment goal to improve her motor skills. At that time, cranial computed tomography scans and screening for inborn errors of metabolism were both normal and she never presented with any seizure disorder. An ophthalmologic evaluation disclosed strabismus, which required surgical correction at the age of 1 year and 10 months though she developed a left ptosis after the procedure.

Cardiologic evaluation disclosed an atrial septal defect (ASD), ostium secundum type, of 10 mm at 7 months of age. Further complementary exams, including audiological evaluation, abdominal ultrasound, and spine x-rays, were normal. She was evaluated by a clinical geneticist at 14 months of age and genetic tests disclosed a G-banded karyotype showing two rearrangements [46,XX, der(6)inv(6)(p23q21)inv(6)(q21q25.3)] and a normal chromosomal microarray, indicating balanced chromosomal rearrangements. Subsequent G-banded karyotyping of her mother did not indicate presence of the rearrangement. The proband also manifested premature thelarche and has been followed by an endocrinologist, with normal hormonal profile.

Physical examination at the age of 3 years showed a weight of 11.760 g (5th centile), height of 89 cm (10th centile), and OFC of 47 cm (2nd to 50th centile); there was thick hair, with sparseness in the parietal region. Facial dysmorphology was notable for bushy eyebrows, long eyelashes, and ocular asymmetry with left palpebral ptosis (**Figure 1B**). There was a long and prominent columella, widely spaced teeth, full lips with everted lower lip, and retrognathia. Palpable breast tissue was noted. Extremities were notable for hypertrichosis in upper limbs and dorsum; finger pads, single transverse palmar creases,

[9]https://github.com/BCM-Lupskilab/DNM-Finder
[10]https://github.com/J35P312/MonteSV

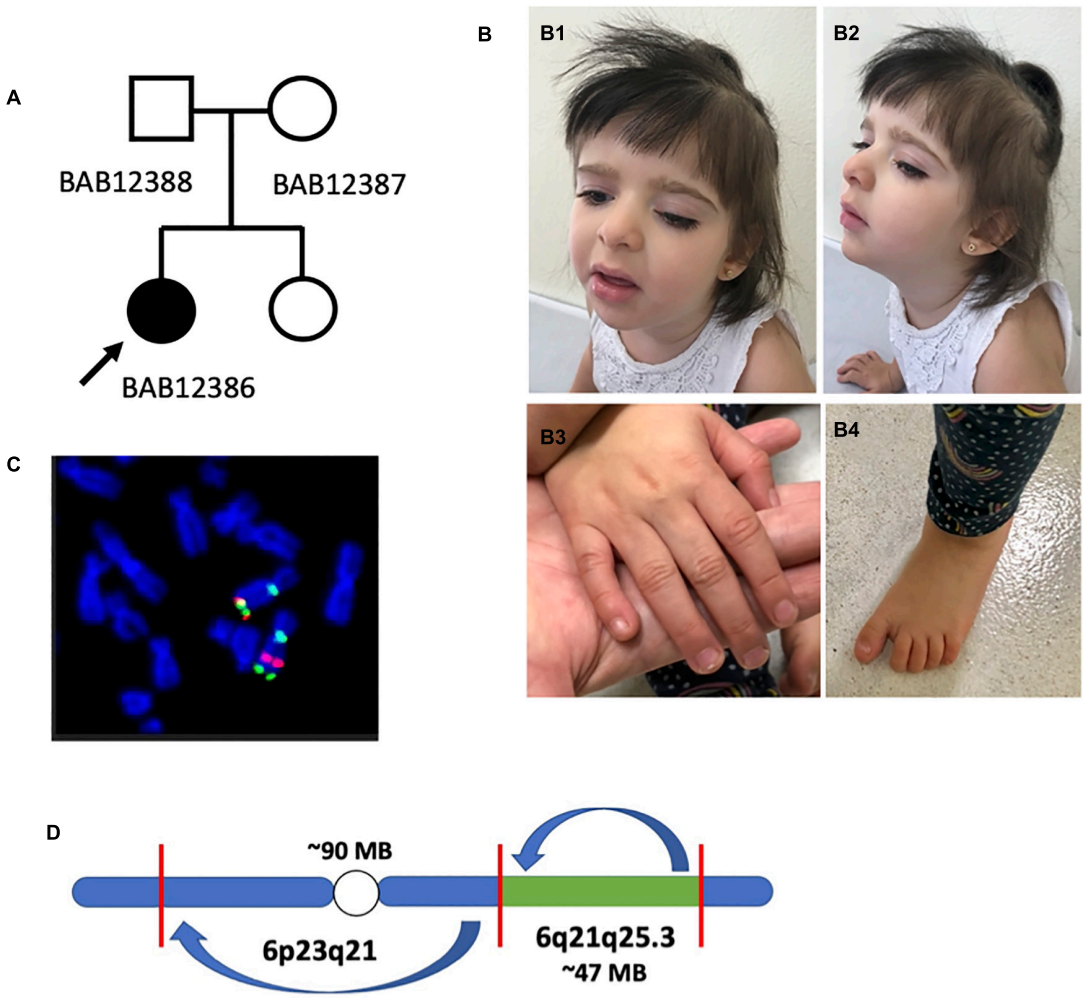

FIGURE 1 | Preliminary analysis of proband and chromosome 6 rearrangement. **(A)** Pedigree structure with the father (BAB12388), mother (BAB12387), and proband (BAB12386) as well as an unaffected sister (not enrolled). **(B)** Female proband (BAB12386) highlighting mildly dysmorphic facies and typical hand features. **(B1,B2)** Frontal and lateral view of the proband at the age of 4 years showing thick hair with sparseness in the temporal region, bushy eyebrows and long eyelashes, left palpebral ptosis, and full lips with eversion of the lower lip. **(B3,B4)** Right hand and foot depicting normal nails and increased distance between the hallux and second toe (sandal gap sign). **(C)** Fluorescence *in situ* hybridization (FISH) analysis confirming apparent pericentric and paracentric inversions present on chromosome 6 as first detected by karyotyping analysis. **(D)** Initially proposed chromosome 6 structure with a ~90-Mb and ~47-Mb inversion both present on chromosome 6.

and normal nails; and flat feet, with sandal gap deformity (**Figure 1B** and **Supplementary Figure 3**). Genitourinary exam showed hypoplastic labia minora. The diagnosis of Coffin–Siris syndrome was raised based on the clinical findings presented by the proband.

To further characterize the chromosomal abnormality, conventional clinical cytogenetics karyotyping using G-banding was repeated in the child and performed in both parents. These studies revealed a *de novo* apparently balanced rearrangement on chromosome 6 involving one pericentric and one paracentric inversion: 46,XX, der(6)inv(6)(p23q21)inv(6)(q21q25.3) (**Supplementary Figures 1**, **4**). Dual-color fluorophore FISH confirmed the two inversions and allowed mapping of one of the cytogenetic breakpoints. In the rearranged chromosome 6, the pericentromeric 6q genomic probe BAC RP11-506N21 (green) was detected on the short arm, confirming the pericentric inversion (**Figure 1C**). Regarding the two 6q25.3 probes, only the sequence RP3-336G18 (red) has moved to a location at 6q more proximal to the centromere; this result confirmed the paracentric inversion, mapping the breakpoint at 6q25.3 to a genomic segment of 1.2 Mb delimited by the clones RP3-336G18 and RP11-266C7 (**Figure 1C** and **Supplementary Figure 5**), which contains *ARID1B*, a potential candidate gene for the proband's proposed clinical diagnosis. Given this information, the original proposed architecture of chromosome 6 involved an approximately 90-Mb pericentric inversion and 47-Mb paracentric inversion based on a human haploid reference genome map (**Figure 1D**).

Evidence for Additional Chromosome 6 Inversions

We performed Illumina 30X PCR-free paired-end (PE) WGS on genomic DNA samples from the proband and parents to identify *de novo* mutational events that might be associated with the apparent sporadic disease. Subsequently, the TIDDIT structural variant caller parsed *de novo* SVs genome-wide (Eisfeldt et al., 2017). Analysis of *de novo* SVs affecting chromosome 6 confirmed the presence of the paracentric and pericentric inversions observed by cytogenetic and cytogenomic studies and revealed three additional breakpoints localized on the long arm at 6q25.3 corresponding to a potential third inversion event not observed previously (**Supplementary Table 1**). The three novel junctions are constituted of ∼1-Mb fragments mapping telomeric to the 46.21-Mb pericentric inversion on 6q. Two out of six structural variants were called as "blunt-end" by the algorithm caller and the remaining four involved in this chromosome were called as an inversion. All regions were manually inspected in IGV (**Supplementary Figure 6**) and the break disrupting the gene *ARID1B* was confirmed (Chr6:157,240,695; Hg19/GRCh37). To determine if the inversions generated were accompanied by CNVs, we performed a custom high-resolution aCGH targeting chromosome 6. No *de novo* CNVs were detected in the proband or parent genome, confirming that, indeed, these inferred SVs were copy-number neutral events affecting only chromosome 6 (**Supplementary Figures 7**, **8**). Genome-wide optical mapping and SV analysis from WGS data showed no additional potentially pathogenic variation.

GATK analysis showed approximately 61 *de novo* SNVs and indels detected genome-wide with no enrichment around the identified breakpoint junctions on chromosome 6. No other potentially pathogenic variants were detected after filtering and annotation for *de novo* or inherited variation.

Genomic Rearrangement Architecture and Recombinant Junction Sequences

Starting from the distal breakpoint position on the p arm, the pericentric inversion is highlighted as segment B (**Figure 2**). The genome map position then connects to segment C on the q arm, in an inverted orientation, which then connects to segment D also in an inverted orientation. Segments E and F are in opposite positions relative to each other with segment F connecting to segment D in the reference orientation and segment E connecting to segment F in an inverted orientation.

Sequence alignments showed that junctions 2, 4, and 6 have a blunt breakpoint junction, whereas junction 5 shows a one base pair of microhomology (G) and junction 1 had a one nucleotide insertion of a "G" (**Figure 2** and **Supplementary Figure 9**). Finally, junction 3 showed an apparent seven-nucleotide templated insertion of "TTTGAAG" likely originating from 9 bp upstream of the proximal strand. The relatively simple features (blunt fusion, microhomology, and small insertions) of the breakpoint junctions and copy-number neutral state of the rearrangement allows inference of a possible DNA NHEJ mechanism as a likely mechanism for generation of formation for this chromosomal aberration. Together, the proposed architecture using the orientation and directionality for each genomic fragment from the nucleotide-level junction alignments and the *de novo* mutation event in sporadic disease implicates this complex rearrangement as clinically relevant for this proband (**Figure 3**).

Genomic Optical Mapping Supports Genomic Orientation and Architecture

To orthogonally investigate this CGR and proposed genomic architecture of the SV haplotype involving chromosome 6, we performed DLS genomic optical mapping. After the identification and sequence alignment of the breakpoint junctions were obtained, we interrogated the genomic optical mapping data at those nucleotide positions. Although the inversion events were too large (>1 Mb) to capture on a single DNA molecule, *de novo* assembly of the patient's personal genome allowed consensus contigs to span the region upstream and downstream of each breakpoint position. Each junction orientation and connection identified in the WGS data were validated in optical genome mapping by visualizing directionality or polarity of sequence motifs in an inverted or direct recombinant join-point connection (**Figure 2** and **Supplementary Figures 10–16**). The molecules spanning the breakpoint junctions were visually inspected, and scrutinized, to parsimoniously map and positionally assign each genomic fragment visualized with optical sequence motifs consistent with the genomic fragment connection.

Inversion Results in Measurable Reduction in Gene Dosage Expression

Importantly, *ARID1B* is disrupted in one location, between the fourth and fifth exons of the transcript NM_001374820.1, and generated breakpoint junction 3 (chr6:157,240,695; Hg19/GRCh37) and junction 5 (chr6:157,240,708), *in cis* (**Figure 3** and **Supplementary Figure 17**). Disruption of the gene *ARID1B* through loss-of-function (LoF) variants has been shown to cause CSS1 (Hoyer et al., 2012; Santen et al., 2012, 2013). The expression levels of *ARID1B* were assayed, with its relative expression compared to three normal controls, to determine if the inversion splitting the gene disrupted its expression in peripheral blood. The levels were significantly ($p = 0.023$, $n = 3$) reduced 30% when compared to normal control samples against the *GAPDH* housekeeping gene.

DISCUSSION

Herein, we present a CGR involving chromosome 6 that disrupts the gene *ARID1B* causing CSS1. The initial karyotyping and FISH analysis, i.e., single cell genomics, indicated one pericentric and one paracentric inversion of chromosome 6. Higher-resolution genomic approaches including WGS and genomic optical mapping uncovered a more complex chromosomal aberration with one (∼95 Mb) pericentric inversion and three additional paracentric inversions (∼46, ∼1, and ∼1 Mb), all of which are localized to a single chromosome 6 in a *de novo* copy-number neutral mutational event. A combination

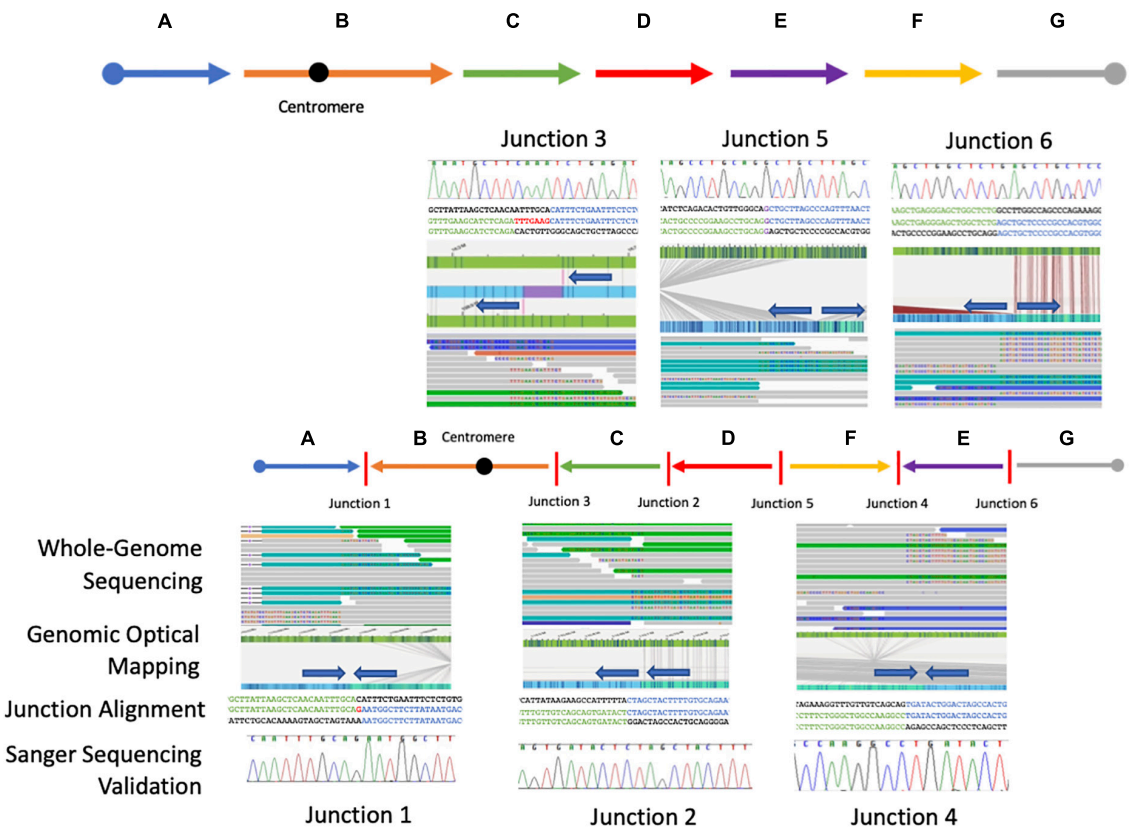

FIGURE 2 | Resolved breakpoint junctions as visualized by multiple genomic technologies. The final resolved chromosome 6 structure showing each genomic fragment orientation and the six breakpoint junctions as visualized through each technology applied including whole-genome sequencing, genomic optical mapping, breakpoint-junction alignment, and final Sanger validation. The colored arrows at the top represent the reference orientation of each genomic fragment. The arrow orientation in the middle represents the orientation of each genomic fragment in this patient with respect to reference.

of experimental methods and genomic approaches resolved the genomic structure of the derivative chromosome 6.

Coffin–Siris syndrome 1 is a clinically and genetically heterogeneous disorder with the most frequent clinically observed findings being developmental delay, coarse facial features, feeding difficulties, frequent infections, and hypoplastic or absent fingernail on the fifth digit (Fleck et al., 2001; Santen et al., 2013). In 2012, both heterozygous deletions and point mutations in the switch/sucrose non-fermentable SWI/SNF-like chromatin remodeling complex gene *ARID1B* were reported to cause CSS1 in a monoallelic, autosomal dominant trait inheritance, Mendelian model (Hoyer et al., 2012; Santen et al., 2012). Although several other genes encoding proteins in the SWI/SNF-like BAF complex including *ARID1A*, *SMARCA2*, *SMARCA4*, *SMARCB1*, and *SMARCE1* have also been shown to cause the Coffin–Siris syndrome phenotype (Santen et al., 2013), and/or a CSS-like phenotype, *ARID1B* is recognized as one of the most frequently mutated genes causing intellectual disability (Hoyer et al., 2012; Santen et al., 2014; Yang et al., 2014; Liu et al., 2019).

The proband described herein (BAB12386) presented with many of the well-characterized phenotypic features of the disease trait including developmental delay, typical craniofacial dysmorphisms, hypotonia with feeding difficulties, hypertrichosis and sparse scalp hair, and premature thelarche, the latter a rare finding reported in CSS1 (Vergano and Deardorff, 2014; **Figure 1** and **Supplementary Figure 3**). Notably absent is the hypoplastic fifth finger or toenail, which appears normal in the present patient (**Figure 1B** and **Supplementary Figure 3**), but can be observed in 81–95% of patients with clinically diagnosed CSS1 (Fleck et al., 2001; Santen et al., 2014). We cannot rule out that hypoplastic phalanges are not present in our patient, since no hand x-ray studies were performed.

There were other genes involved in the rearrangement including *ATXN1*, *CDK19*, and *SYNJ2* (**Supplementary Figure 18**). In mice, deletions of *ATXN1* have been shown to cause mild learning defects without neurodegeneration (Lu et al., 2017). Recently, missense variants in *CDK19* have been shown to cause developmental and epileptic encephalopathy (MIM:618916), though partial gene deletions have been found in healthy individuals suggesting that haploinsufficiency of *CDK19* may not be clinically relevant (Wong et al., 2007; Chung et al., 2020). *SYNJ2* has been shown to be involved in the formation of cell membrane structures though the gene has not been

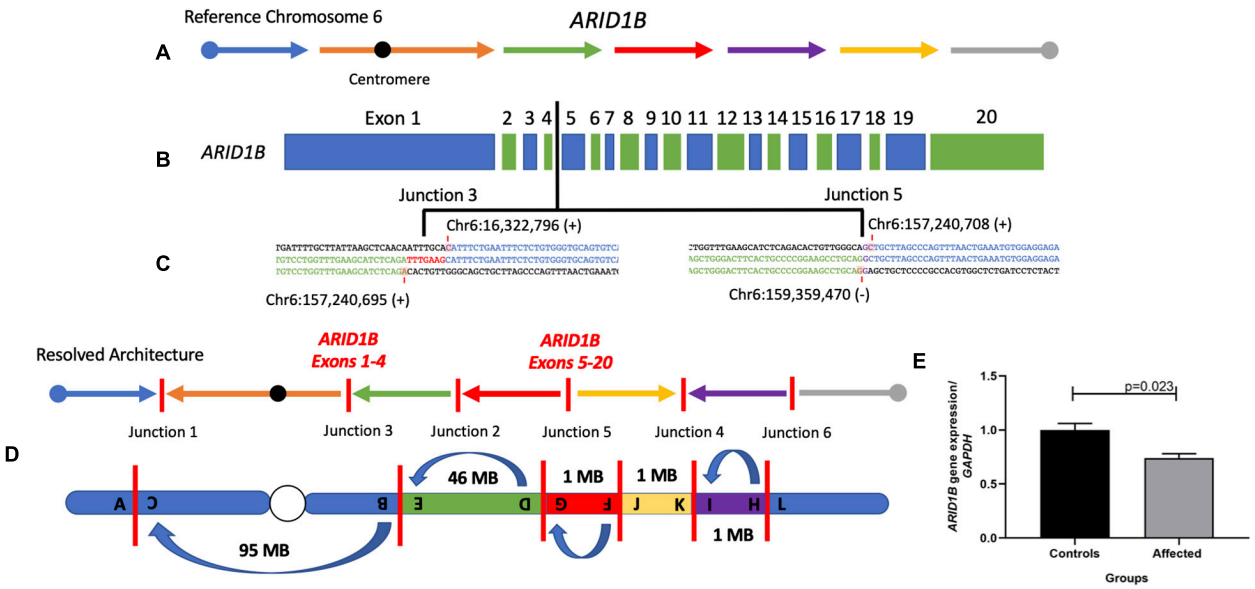

FIGURE 3 | Final chromosome 6 resolved architecture revealed disruption of the gene *ARID1B*. **(A)** Structure of chromosome 6 displaying the reference orientation of each genomic fragment as represented by arrows moving from left to right with the centromere positioned as a black circle. **(B)** Coding structure of the gene *ARID1B* (NM_001374820.1). Vertical black line indicates the inversion break that disrupted the gene between the fourth and fifth exons. **(C)** Breakpoint sequence alignments of junctions 3 and 5 indicate the nucleotide positions disrupted within *ARID1B*. **(D)** Final resolved structure of chromosome 6 showing six breakpoint junctions with one pericentric inversion and three paracentric inversions on the q arm. **(E)** qPCR analysis of *ARID1B* mRNA in patient compared to three controls reveal significant expression reduction (∼30%) in peripheral blood.

directly linked to a human disease state (Chuang et al., 2004). Therefore, disruption of *ARID1B* is a plausible explanation from the genomic and clinical points of view. Nevertheless, we cannot completely rule out a blended phenotype (Posey et al., 2017) that may occur due to the disruptions of *ATXN1* as well as *CDK19* or the contributory role of other gene loci and genetic variation potentially conferring position effects due to the complex reordered genome and chromosome structure present on chromosome 6.

Structural variation, including deletions, intragenic duplications, and translocations leading to disruptions of *ARID1B*, has been previously reported (Halgren et al., 2012; Seabra et al., 2017). The disruption of *ARID1B* that drives this patient's phenotype appears to have occurred as the result of a balanced inversion event translocating the proximal and distal *ARID1B* transcripts to two different genomic locations. This genomic rearrangement resulted in an observed 30% reduction of *ARID1B* specific mRNA dosage or expression as observed by RT-PCR in diploid cells (**Figure 3E**). It is intriguing that the levels of *ARID1B* expression in blood is reduced by 30% rather than the expected 50%. We speculate that there is higher expression of the wild-type (WT) allele in blood, perhaps due to compensation or that the qPCR experiment performed is measuring both the WT and truncated transcripts, the latter not fully degraded by nonsense-mediated decay as would be expected. Interestingly, similar ∼30% decreased mRNA expression has been detected in another patient with SV affecting *ARID1B* also clinically diagnosed with CSS1 (Halgren et al., 2012; Seabra et al., 2017). The qPCR primer sets used to assay *ARD1B* in our study as well as Seabra et al. (2017) target three out of four transcripts of the gene including the canonical transcript.

The complex genomic structure and mutational junction signatures appear to have been formed by an NHEJ mechanism generating this highly reordered chromosome. Chromoanagenesis, i.e., chromosome rebirth, encompasses the phenomena of extensive rearrangement occurring in a single burst (including chromothripsis, chromoanasynthesis, and chromoplexy), generating localized complex chromosome rearrangements identified in both somatic and germline genomes (Holland and Cleveland, 2012; Ly and Cleveland, 2017). Although this type of aberration complies with some aspects of chromothripsis, including the involvement of one chromosome and six breakpoints with genomic fragment shuffling in a balanced manner (Kloosterman et al., 2011, 2012; Maher and Wilson, 2012), the fact that the breakpoints are not clustered and appear to occur within transcriptionally active areas (four out of six breakpoints occur within genes) is also in line with a chromoplexy-type event (Shen, 2013; Redin et al., 2017). Although chromothripsis and chromoplexy were first characterized in cancer genomes, the same "mutagenic phenomenon" has been shown to underlie Mendelian diseases and genomic disorders by disruption of genes through truncating breakpoints (haploinsufficiency), by the generation of fusion genes (ectopic expression), or other position effects (Maher and Wilson, 2012; Baca et al., 2013; Redin et al., 2017; Plesser Duvdevani et al., 2020). This process may occur in a random order of DNA fusion but interestingly in this present case, almost all the inversion events happen sequentially from one another in

a potential "chained" fashion rather than a single "pulverizing" event which is more suggestive of chromoplexy (chained rearrangements) over chromothripsis (a single catastrophic event occurring).

To test the likelihood that this rearrangement is formed through a chromoplexy- versus chromothripsis-type mechanism, we performed a simulation to test for either an enrichment of breakpoints occurring within protein coding genes (which would support chromoplexy) or a clustering of breakpoints on the chromosome (which would support chromothripsis). After 10,000 simulations, we observed neither a significant enrichment of breakpoints within protein coding genes (p-value of 0.112) nor a denser clustering of breakpoints than would be expected by chance (p-value of 0.758), suggesting an expanded understanding of mutation events that appear to fall under the chromoanagenesis definition.

In summary, resolving the CGR affecting chromosome 6 required the use of multiple technologies to elucidate the structure of a derivative chromosome constituted by multiple copy-number neutral events. Resolving this genomic puzzle was key to identify the underlying molecular cause of the clinical traits in this patient. Moreover, the identification of several *de novo* inversions on a single chromosome, generated through a chromothriptic-like mutational event, suggests that such mutational process may lead to hidden complexities in seemingly "simple" structural variants. As we continue to refine and improve our ability to resolve inversions and other complex structural variants, "unsolved" Mendelian diseases should be investigated by applying new and developing genomic methodologies that allow phasing multiple breakpoint junctions *in cis* (Liu et al., 2019; Plesser Duvdevani et al., 2020).

DATA AVAILABILITY STATEMENT

Microarray data generated in this study are available through GEO under the accession number GSE180423. BAM files for the proband indicating the specified structural variants are deposited in the Sequence Read Archive (SRA), accession number PRJNA748013.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Baylor College of Medicine (IRB #: H-29697). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

CMG performed the laboratory work, analyzed and interpreted the data, and wrote the manuscript. JE and HD performed the bioinformatic analysis. ACVK, DRB, DO, and SSC provided patient samples, clinical information of patients, and/or analysis and interpretation of data. JRL and AL performed data interpretation and critical review of the manuscript. CMBC conceptualized the study, analyzed and interpreted the data, and is a major contributor in writing the manuscript. All authors have read, edited, and approved the final manuscript.

FUNDING

This study was supported in part by the United States National Institute of General Medical Sciences NIGMS R01 GM132589 (CMBC), the Swedish Brain Foundation [FO2020-0351 (AL)], and the National Institute for Neurological Disorders and Stroke [NINDS R35 NS105078 (JRL)].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.708348/full#supplementary-material

Supplementary Figure 1 | Proband (BAB12386) full karyotype highlighting chromosome 6 within the red outlined box [46,XX, der(6)inv(6)(p23q21)inv(6)(q21q25.3)].

Supplementary Figure 2 | After initial karyotyping and fluorescent *in situ* hybridization a combination of methods. **(A)** Trio custom high-resolution array comparative genomic hybridization with average probe spacing of 1 probe per 857 bp spanning chromosome 6. **(B)** Illumina 30X PCR-free paired-end (PE) trio whole-genome sequencing and **(C)** Direct Label and Stain (DLS) trio genomic optical mapping was performed on the Bionano Saphyr system.

Supplementary Figure 3 | (A) Proband (BAB12386) clinical photos denoting thick hair, with sparseness in parietal region, bushy eyebrows, long eyelashes, ocular asymmetry with left palpebral ptosis, long and prominent columella, full lips with everted lower lip, retrognathia; **(B)** hypertrichosis in upper limbs and dorsum; **(C)** finger pads, single transverse palmar creases, normal nails; flat feet, with sandal gap deformity.

Supplementary Figure 4 | Original karyotyping analysis using G-banding showing a *de novo* apparently balanced rearrangement on chromosome 6 involving one pericentric and one paracentric inversions: 46,XX, der(6)inv(6)(p23q21)inv(6)(q21q25.3).

Supplementary Figure 5 | Dual-color fluorophore FISH confirmed the two inversion events. The pericentromeric 6q genomic probe, bacterial artificial chromosome (BAC) RP11-506N21 (green) was detected on the short arm, confirming the pericentric inversion. The RP3-336G18 probe (red) has moved to a location at 6q more proximal to the centromere; this result confirmed the paracentric inversion, mapping the breakpoint at 6q25.3 to a genomic segment of 1.2 Mb delimited by the clones RP3-336G18 and RP11-266C7.

Supplementary Figure 6 | All breakpoint regions were manually inspected in the integrative genomics viewer (IGV) showing soft-clipped reads flaking each region.

Supplementary Figure 7 | (A) A high-resolution aCGH targeting the long and short arm of chromosome 6 with a median probe spacing of 857 bp across the chromosome was performed in the proband (BAB12386) as well as the mother (BAB12387) and father (BAB12388). No CNVs were detected across the chromosome or **(B)** surrounding the gene *ARID1B*.

Supplementary Figure 8 | The regions surrounding each breakpoint were scrutinized in the aCGH for the proband, mother and father. No small CNVs were detected for any of their 6 breakpoint regions.

Supplementary Figure 9 | Nucleotide-level resolution of all 6 breakpoint junctions are shown with the nucleotide position (Hg19/GRCh37) from each side as well as the directionality of the sequence (±) forming the junction.

Supplementary Figure 10 | (A) Representation of junction one showing break one in a positive orientation connecting to break 3 in a negative orientation as expected for an inversion event. **(B)** Genomic optical mapping data shows the positive and negative orientation as well as the genomic coordinates of the breakpoint connecting in an inverted manner.

Supplementary Figure 11 | Junction 1 optical mapping data showing single molecule support of the breakpoint junction architecture. A single molecule spanning the junction with a length of approximately 429 kb is highlighted by the red boxes.

Supplementary Figure 12 | Genomic optical mapping data representing junction 2 with break 4 (Chr6:111,024,035) and break 7 (Chr6:158,471,524) connecting in tandem.

Supplementary Figure 13 | (A) Genomic optical mapping data for junction 3 showing the point of connection for break 2 (Chr6:157,240,695) in tandem. **(B)** The connection between *ARID1B* (purple) to *ATXN1* (green).

Supplementary Figure 14 | Genomic optical mapping data for junction 4 showing the connection of Chr6:158,471,518 and Chr6:160,535,951 in an inverted orientation.

Supplementary Figure 15 | Genomic optical mapping data with single molecule visualization for junction 5. The connection at Chr6:158,471,518 is which is represented by the purple arrow in a tail-to-tail orientation with Chr6:157,240,708 which is represented by a red arrow.

Supplementary Figure 16 | Genomic optical mapping data showing junction 6 with the connection of Chr6:159,359,468 which is represented by a yellow arrow fused to Chr6:160,535,951 represented by a gray arrow connected in a tail-to-tail orientation.

Supplementary Figure 17 | Graphical representation of the exons forming *ARID1B* as well as the position of the inversion breakpoint disrupting the gene as denoted by the red vertical line between the 4th and 5th exon.

Supplementary Figure 18 | Out of the 6 breakpoints that occurred on this chromosome, four occurred within the genes including *ATXN1*, *CDK19*, *ARID1B*, and *SYNJ2*.

Supplementary Table 1 | TIDDIT genome-wide structural variant calls for the proband. Yellow highlights denote the variants involved in this complex rearrangement.

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., et al. (2013). Punctuated evolution of prostate cancer genomes. *Cell* 153, 666–677. doi: 10.1016/j.cell.2013.03.021

Beck, C. R., Carvalho, C. M. B., Banser, L., Gambin, T., Stubbolo, D., Yuan, B., et al. (2015). Complex genomic rearrangements at the *PLP1* locus include triplication and quadruplication. *PLoS Genet.* 11:3. doi: 10.1371/journal.pgen.1005050

Carvalho, Ramocki, M. B., Pehlivan, D., Franco, L. M., Gonzaga-Jauregui, C., Fang, P., et al. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* 43, 1074–1081. doi: 10.1038/ng.944

Carvalho, C. M., Zhang, F., Liu, P., Patel, A., Sahoo, T., Bacino, C. A., et al. (2009). Complex rearrangements in patients with duplications of *MECP2* can occur by fork stalling and template switching. *Hum. Mol. Genet.* 18, 2188–2203. doi: 10.1093/hmg/ddp151

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10:1784. doi: 10.1038/s41467-018-08148-z

Chuang, Y. Y., Tran, N. L., Rusk, N., Nakada, M., Berens, M. E., and Symons, M. (2004). Role of synaptojanin 2 in glioma cell migration and invasion. *Cancer Res.* 64, 8271–8275. doi: 10.1158/0008-5472.CAN-04-2097

Chung, H. L., Mao, X., Wang, H., Park, Y. J., Marcogliese, P. C., Rosenfeld, J. A., et al. (2020). De Novo Variants in *CDK19* Are Associated with a Syndrome Involving Intellectual Disability and Epileptic Encephalopathy. *Am. J. Hum. Genet.* 106, 717–725. doi: 10.1016/j.ajhg.2020.04.001

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372:6537. doi: 10.1126/science.abf7117

Eisfeldt, J., Pettersson, M., Petri, A., Nilsson, D., Feuk, L., and Lindstrand, A. (2020). Hybrid sequencing resolves two germline ultra-complex chromosomal rearrangements consisting of 137 breakpoint junctions in a single carrier. *Hum. Genet.* 2020:3. doi: 10.1007/s00439-020-02242-3

Eisfeldt, J., Vezzi, F., Olason, P., Nilsson, D., and Lindstrand, A. (2017). TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res* 6:664. doi: 10.12688/f1000research.11168.2

Eldomery, M. K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J. A., Gambin, T., Stray-Pedersen, A., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med.* 9:26. doi: 10.1186/s13073-017-0412-6

Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome architecture. *Genome Med.* 2:11. doi: 10.1186/gm132

Fleck, B. J., Pandya, A., Vanner, L., Kerkering, K., and Bodurtha, J. (2001). Coffin-Siris syndrome: review and presentation of new cases from a questionnaire study. *Am. J. Med. Genet.* 99, 1–7. doi: 10.1002/1096-8628(20010215)99:1<1::aid-ajmg1127<3.0.co;2-a

Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., et al. (2007). Recurrent DNA inversion rearrangements in the human genome. *Proc. Natl. Acad. Sci. U S A* 104, 6099–6106. doi: 10.1073/pnas.0701631104

Grochowski, C. M., Gu, S., Yuan, B., Tcw, J., Brennand, K. J., Sebat, J., et al. (2018). Marker chromosome genomic structure and temporal origin implicate a chromoanasynthesis event in a family with pleiotropic psychiatric phenotypes. *Hum. Mutat.* 39, 939–946. doi: 10.1002/humu.23537

Gu, S., Yuan, B., Campbell, I. M., Beck, C. R., Carvalho, C. M., Nagamani, S. C., et al. (2015). Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. *Hum. Mol. Genet.* 24, 4061–4077. doi: 10.1093/hmg/ddv146

Halgren, C., Kjaergaard, S., Bak, M., Hansen, C., El-Schich, Z., Anderson, C. M., et al. (2012). Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of *ARID1B*. *Clin. Genet.* 82, 248–255. doi: 10.1111/j.1399-0004.2011.01755.x

Hofmeister, W., Pettersson, M., Kurtoglu, D., Armenio, M., Eisfeldt, J., Papadogiannakis, N., et al. (2018). Targeted copy number screening highlights an intragenic deletion of *WDR63* as the likely cause of human occipital encephalocele and abnormal CNS development in zebrafish. *Hum. Mutat.* 39, 495–505. doi: 10.1002/humu.23388

Holland, A. J., and Cleveland, D. W. (2012). Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat. Med.* 18, 1630–1638. doi: 10.1038/nm.2988

Hoyer, J., Ekici, A. B., Endele, S., Popp, B., Zweier, C., Wiesener, A., et al. (2012). Haploinsufficiency of *ARID1B*, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. *Am. J. Hum. Genet.* 90, 565–572. doi: 10.1016/j.ajhg.2012.02.007

Kaiser, P. (1984). Pericentric inversions. Problems and significance for clinical genetics. *Hum. Genet.* 68, 1–47. doi: 10.1007/bf00293869

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56–64. doi: 10.1038/nature06862

Kloosterman, W. P., Guryev, V., van Roosmalen, M., Duran, K. J., de Bruijn, E., Bakker, S. C., et al. (2011). Chromothripsis as a mechanism driving complex *de novo* structural rearrangements in the germline. *Hum. Mole. Genet.* 20, 1916–1924. doi: 10.1093/hmg/ddr073

Kloosterman, W. P., Tavakoli-Yaraki, M., van Roosmalen, M. J., van Binsbergen, E., Renkens, I., Duran, K., et al. (2012). Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* 1, 648–655. doi: 10.1016/j.celrep.2012.05.009

Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schopflin, R., Markowski, J., et al. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* 21, 305–310. doi: 10.1038/s41556-019-0273-x

Lakich, D., Kazazian, H. H. Jr., Antonarakis, S. E., and Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* 5, 236–241. doi: 10.1038/ng1193-236

Lee, J. A., Carvalho, C. M., and Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131, 1235–1247. doi: 10.1016/j.cell.2007.11.037

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27, 718–719. doi: 10.1093/bioinformatics/btq671

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., et al. (2019). Reanalysis of Clinical Exome Sequencing Data. *N Engl J Med* 380, 2478–2480. doi: 10.1056/NEJMc1812033

Lu, H. C., Tan, Q., Rousseaux, M. W., Wang, W., Kim, J. Y., Richman, R., et al. (2017). Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. *Nat. Genet.* 49, 527–536. doi: 10.1038/ng.3808

Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025. doi: 10.1016/j.cell.2015.04.004

Lupski, J. R., Liu, P., Stankiewicz, P., Carvalho, C. M. B., and Posey, J. E. (2020). Clinical genomics and contextualizing genome variation in the diagnostic laboratory. *Expert Rev. Mol. Diagn* 20, 995–1002. doi: 10.1080/14737159.2020.1826312

Ly, P., and Cleveland, D. W. (2017). Rebuilding Chromosomes After Catastrophe: Emerging Mechanisms of Chromothripsis. *Trends Cell Biol.* 27, 917–930. doi: 10.1016/j.tcb.2017.08.005

Maher, C. A., and Wilson, R. K. (2012). Chromothripsis and human disease: piecing together the shattering process. *Cell* 148, 29–32. doi: 10.1016/j.cell.2012.01.006

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297–1303. doi: 10.1101/gr.107524.110

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069–2070. doi: 10.1093/bioinformatics/btq330

Pettersson, M., Grochowski, C. M., Wincent, J., Eisfeldt, J., Breman, A. M., Cheung, S. W., et al. (2020). Cytogenetically visible inversions are formed by multiple molecular mechanisms. *Hum Mutat*. 2020, 24106. doi: 10.1002/humu.24106

Plesser Duvdevani, M., Pettersson, M., Eisfeldt, J., Avraham, O., Dagan, J., Frumkin, A., et al. (2020). Whole-genome sequencing reveals complex chromosome rearrangement disrupting *NIPBL* in infant with Cornelia de Lange syndrome. *Am. J. Med. Genet. A* 182, 1143–1151. doi: 10.1002/ajmg.a.61539

Posey, J. E. (2019). Genome sequencing and implications for rare disorders. *Orphanet. J. Rare Dis.* 14, 153. doi: 10.1186/s13023-019-1127-0

Posey, J. E., Harel, T., Liu, P., Rosenfeld, J. A., James, R. A., Coban Akdemir, Z. H., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* 376, 21–31. doi: 10.1056/NEJMoa1516767

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* 49, 36–45. doi: 10.1038/ng.3720

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Sanchez-Gaya, V., Mariner-Fauli, M., and Rada-Iglesias, A. (2020). Rare or Overlooked? Structural Disruption of Regulatory Domains in Human Neurocristopathies. *Front. Genet.* 11:688. doi: 10.3389/fgene.2020.00688

Santen, G. W., Aten, E., Sun, Y., Almomani, R., Gilissen, C., Nielsen, M., et al. (2012). Mutations in SWI/SNF chromatin remodeling complex gene *ARID1B* cause Coffin-Siris syndrome. *Nat. Genet.* 44, 379–380. doi: 10.1038/ng.2217

Santen, G. W., Aten, E., Vulto-van Silfhout, A. T., Pottinger, C., van Bon, B. W., van Minderhout, I. J., et al. (2013). Coffin-Siris syndrome and the BAF complex: genotype-phenotype study in 63 patients. *Hum. Mutat.* 34, 1519–1528. doi: 10.1002/humu.22394

Santen, G. W., Clayton-Smith, J., and Consortium, A. B. C. (2014). The *ARID1B* phenotype: what we have learned so far. *Am. J. Med. Genet. C Semin. Med. Genet.* 166C, 276–289. doi: 10.1002/ajmg.c.31414

Seabra, C. M., Szoko, N., Erdin, S., Ragavendran, A., Stortchevoi, A., Maciel, P., et al. (2017). A novel microduplication of *ARID1B*: Clinical, genetic, and proteomic findings. *Am. J. Med. Genet. A* 173, 2478–2484. doi: 10.1002/ajmg.a.38327

Shen, M. M. (2013). Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell* 23, 567–569. doi: 10.1016/j.ccr.2013.04.025

Vergano, S. S., and Deardorff, M. A. (2014). Clinical features, diagnostic criteria, and management of Coffin-Siris syndrome. *Am. J. Med. Genet. C Semin. Med. Genet.* 166C, 252–256. doi: 10.1002/ajmg.c.31411

Vicente-Salvador, D., Puig, M., Gaya-Vidal, M., Pacheco, S., Giner-Delgado, C., Noguera, I., et al. (2017). Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum Mol Genet* 26, 567–581. doi: 10.1093/hmg/ddw415

Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., et al. (2007). A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* 80, 91–104. doi: 10.1086/510560

Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 312, 1870–1879. doi: 10.1001/jama.2014.14601

Zuccherato, L. W., Alleva, B., Whiters, M. A., Carvalho, C. M., and Lupski, J. R. (2016). Chimeric transcripts resulting from complex duplications in chromosome Xq28. *Hum. Genet.* 135, 253–256. doi: 10.1007/s00439-015-1614-x

Conflict of Interest: Baylor College of Medicine (BCM) and Miraca Holdings have formed a joint venture with shared ownership and governance of the Baylor Genetics (BG), which performs clinical microarray analysis and other genomic studies (ES and WGS) for patient/family care. JRL serves on the Scientific Advisory Board of the BG. JRL has stock ownership in 23andMe, is a paid consultant for Regeneron Pharmaceuticals, and is a co-inventor on multiple United States and European patents related to molecular diagnostics for inherited neuropathies, eye diseases, and bacterial genomic fingerprinting.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/fgene.2021.742325

Trio-Based Low-Pass Genome Sequencing Reveals Characteristics and Significance of Rare Copy Number Variants in Prenatal Diagnosis

Matthew Hoi Kin Chau[1,2,3†], Jicheng Qian[1,2†], Zihan Chen[2], Ying Li[1,2,3], Yu Zheng[1,2], Wing Ting Tse[1], Yvonne K. Kwok[1,2], Tak Yeung Leung[1,2,4], Zirui Dong[1,2,3] and Kwong Wai Choy[1,2,3,4]**

[1] Department of Obstetrics and Gynecology, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR China, [2] Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, Hong Kong, SAR China, [3] Hong Kong Hub of Pediatric Excellence, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR China, [4] The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Shatin, Hong Kong, SAR China

Background: Low-pass genome sequencing (GS) detects clinically significant copy number variants (CNVs) in prenatal diagnosis. However, detection at improved resolutions leads to an increase in the number of CNVs identified, increasing the difficulty of clinical interpretation and management.

Methods: Trio-based low-pass GS was performed in 315 pregnancies undergoing invasive testing. Rare CNVs detected in the fetuses were investigated. The characteristics of rare CNVs were described and compared to curated CNVs in other studies.

Results: A total of 603 rare CNVs, namely, 597 constitutional and 6 mosaic CNVs, were detected in 272 fetuses (272/315, 86.3%), providing 1.9 rare CNVs per fetus (603/315). Most CNVs were smaller than 1 Mb (562/603, 93.2%), while 1% (6/603) were mosaic. Forty-six *de novo* (7.6%, 46/603) CNVs were detected in 11.4% (36/315) of the cases. Eighty-four CNVs (74 fetuses, 23.5%) involved disease-causing genes of which the mode of inheritance was crucial for interpretation and assessment of recurrence risk. Overall, 31 pathogenic/likely pathogenic CNVs were detected, among which 25.8% (8/31) were small (<100 kb; $n = 3$) or mosaic CNVs ($n = 5$).

Conclusion: We examined the landscape of rare CNVs with parental inheritance assignment and demonstrated that they occur frequently in prenatal diagnosis. This information has clinical implications regarding genetic counseling and consideration for trio-based CNV analysis.

Keywords: low-pass genome sequencing, *de novo*, inherited, copy number variants, prenatal diagnosis

OPEN ACCESS

Edited by:
Gavin R. Oliver,
Mayo Clinic, United States

Reviewed by:
Keren Carss,
AstraZeneca, United Kingdom
Qi Qingwei,
Peking Union Medical College Hospital (CAMS), China
Yanmin Luo,
The First Affiliated Hospital of Sun Yat-sen University, China
Zhang Rui,
Baoan Maternal and Child Health Hospital, China

***Correspondence:**
Zirui Dong
elvisdong@cuhk.edu.hk
Kwong Wai Choy
richardchoy@cuhk.edu.hk

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *16 July 2021*
Accepted: *25 August 2021*
Published: *20 September 2021*

Citation:
Chau MHK, Qian J, Chen Z, Li Y, Zheng Y, Tse WT, Kwok YK, Leung TY, Dong Z and Choy KW (2021) Trio-Based Low-Pass Genome Sequencing Reveals Characteristics and Significance of Rare Copy Number Variants in Prenatal Diagnosis. Front. Genet. 12:742325. doi: 10.3389/fgene.2021.742325

INTRODUCTION

Prenatal genetic diagnosis is routinely performed in high-risk pregnancies to identify fetal genetic abnormalities, including chromosome aneuploidies (such as Trisomy 21) and pathogenic copy number variants (CNVs; such as deletion and duplications). Chromosomal microarray analysis (CMA) is recommended as the first-tier genetic test in the diagnostic evaluation of fetal structural abnormalities by the American College of Obstetricians and Gynecologists (Levy and Wapner, 2018). CMA provides enhanced resolution for the detection of submicroscopic deletions/duplications compared with G-banded chromosome analysis (Leung et al., 2011; Huang et al., 2014; Yang et al., 2017; Chau and Choy, 2021). The spectrum, incidence, and mode of inheritance (*de novo* or inherited) of clinically significant CNVs in prenatal diagnosis by various CMA platforms have been investigated (Chau et al., 2019). In addition, assignment of parental inheritance of CNVs is not only important for clinical interpretation, as rare *de novo* CNVs are more likely to be pathogenic (Asadollahi et al., 2014), but also essential to provide prognostic information and recurrence risk (Huijsdens-van Amsterdam et al., 2018). For instance, the incidence of *de novo* CNVs was 2.9% (14/488) in fetuses with early preterm birth (Wong et al., 2020). However, due to triplication of the experimental cost for trio-based testing (simultaneous), parental inheritance assignment is often performed sequentially, when a candidate variant of interest has been identified in the proband. In a study curating CMA results in 23,865 prenatal cases (Chau et al., 2019), more than 25% of pathogenic CNVs lacked parental inheritance assignment. Thus, comprehensive understanding of rare CNVs with the mode of inheritance is still not well studied in prenatal diagnosis.

Recent studies have demonstrated the feasibility of applying genome sequencing (GS) for CNV detection in prenatal diagnosis (Choy et al., 2019; Zhou et al., 2021), particularly using a low-pass (low-coverage and high-through) approach (Liang et al., 2014; Dong et al., 2016; Wang et al., 2018; Chau et al., 2020; Wang H. et al., 2020). It offers higher resolution of CNV detection (e.g., CNVs < 100 kb in size) and improved sensitivity in detecting low-level mosaic variants. Thus, low-pass GS provides a higher genetic diagnostic yield compared with CMA (Liang et al., 2014; Dong et al., 2016; Wang et al., 2018; Chau et al., 2020; Chaubey et al., 2020; Wang H. et al., 2020). In particular, both reagent costs and experimental repeat rates were lower compared to CMA platforms (Wang H. et al., 2020), enabling its widespread usage in clinical laboratories (Wang et al., 2018; Wang H. et al., 2020). Parental inheritance assignment of CNVs is commonly performed sequentially, after a variant of interest has been identified in the proband. However, recent studies suggested that GS-based CNV detection methods reveals a high number of small CNVs (<100 kb) (Sudmant et al., 2015; Chau et al., 2020; Collins et al., 2020; Wang H. et al., 2020), and it is difficult to determine their clinical significance with a proband-only approach. A sequential approach increases turnaround time; thus, a trio-based approach may be better suited for prenatal testing, especially when pregnancy management and decision-making are often dependent on timely results. As such, the incidence, spectrum, and mode of inheritance of rare CNVs and the proportion of cases requiring parental analysis are important considerations to guide diagnostic approaches (proband-only, sequential approach, or trio-based) by low-pass GS.

Herein, we performed a prospective trio-based study of 315 consecutive prenatal diagnosis cases to study the incidence, landscape, and characteristics of rare CNVs with mode of inheritance assignment by low-pass GS.

MATERIALS AND METHODS

Subject Enrollment, Sample Recruitment, and Preparation

This study was approved by the Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee (CREC Ref. No.: 2016.713). From January 2019 to February 2021, pregnant women referred for trio-based prenatal diagnostic testing by low-pass GS at our Prenatal Genetic Diagnosis Center, Department of Obstetrics and Gynecology, The Chinese University of Hong Kong were enrolled. Each participant provided written informed consent. Their primary referral indications included: (1) abnormal ultrasound findings, (2) positive noninvasive prenatal testing, (3) positive Down syndrome screening, (4) advanced maternal age, (5) family/adverse pregnancy history, and (6) others which included ultrasound soft markers only, maternal anxiety, and rare indications. Prenatal samples including chorionic villi samples (CVS), amniotic fluid (AF), or cord blood were collected simultaneously with parental peripheral blood samples.

Genomic DNA from prenatal and parental samples were extracted with DNeasy Blood and Tissue Kit (Cat No./ID: 69506, Qiagen, Hilden, Germany) and were treated with RNase (Qiagen). DNA was subsequently quantified using a Qubit dsDNA HS Assay kit (Invitrogen, Carlsbad, CA, United States). The DNA integrity was assessed by agarose gel electrophoresis. Quantitative fluorescent polymerase chain reaction (qfPCR) was performed using two panels of short-tandem repeat (STR) markers (P1 and XY) located on chromosomes 18, X, and Y for the detection of maternal cell admixture, polyploidy, and confirmation of biological relationships (Choy et al., 2014). G-banded chromosome analysis (karyotyping) was also performed in 205 cases (65.1%).

Low-Pass Genome Sequencing

Low-pass GS was performed on each sample essentially as previously described (Wang H. et al., 2020). In brief, 50 ng of genomic DNA was digested (200–300 bp) and repaired by fragmentation-end-repair restriction enzyme (MGI tech Co., Ltd., Shenzhen, China). Next, an A-overhang was added for adapter ligation. The DNA fragments underwent seven cycles of PCR. PCR products from each library were subsequently purified with an Agencourt AMPure XP PCR Purification Kit (Beckman Coulter, Brea, CA, United States). The concentration

of each library was measured with a Qubit dsDNA HS Assay Kit (Invitrogen). The libraries were mixed with equal molality into each pool (20–24 samples per two lanes) and were sequenced to a minimal of ∼15 million reads per sample (single-end 50 bp) on an MGISeq-2000 platform (MGI, Wuhan, China). The minimal read depth is estimated to be 0.25-fold, which is determined by multiplying the reads (15 million) and the read length (50 bp), divided by the size of human reference genome (3 Gb).

QC and Data Processing

For each sample, low-quality reads were filtered and single-end reads were aligned to the human reference genome (GRCh37/hg19) using the Burrows–Wheeler Aligner (BWA) (Li and Durbin, 2009) "Aln" and "Samse" alignment modules with the default settings. Uniquely aligned reads were deposited into adjustable sliding windows (50 kb in length with 5-kb increments) and adjustable non-overlapping windows (5 kb). The coverage of each window was calculated by the sum of read amounts after GC correction and population-scale normalization. The genome-wide standard deviation of the copy ratios from all windows except for windows on aneuploid chromosomes was used as a QC measure as previously described, and 0.1 was set as the cutoff (Chau et al., 2020; Wang H. et al., 2020).

Copy Number Variant Detection and Determination of Parental Inheritance

The detection of homozygous/heterozygous deletions and duplications/triplications was performed by our reported method (Chau et al., 2020; Wang H. et al., 2020). In brief, (1) aneuploidies were detected based on the difference between the average copy ratio for each chromosome compared to the normal copy ratio (expected as 1), where the degree of deviation from normal copy ratios was used to calculate the mosaic level; (2) regions with putative CNVs (at a resolution of 50 kb) were reported, and their precise breakpoint boundaries were determined using our in-house algorithm "Increment-Rate-of-Coverage" (Dong et al., 2016) based on the copy ratios of the adjustable non-overlapping windows; and (3) homozygous or hemizygous deletions were (at a resolution of 10 kb) called if two or more consecutive non-overlapping windows contained extremely low numbers or absence of aligned reads (copy ratio: 0.0–0.1). For mosaic CNV detection, mosaic levels were calculated as previously reported and the minimal mosaic levels of CNV detection were 30% for small CNVs (<2.5 Mb) and 20% for large CNVs (>2.5 Mb) (Chau et al., 2020).

For each CNV, population-based *U*-test, whole-sample-based *t*-test and whole-chromosome-based *t*-test were performed to eliminate false positives and common population-specific polymorphisms. In addition, CNVs with an allele frequency of <1% in our reported datasets (Dong et al., 2016; Chau et al., 2020; Wang H. et al., 2020) of ethnic Chinese fetuses ($n > 2{,}000$) were defined as rare CNVs for subsequent analyses.

Lastly, the coordinates and the variant type (homozygous/hemizygous/heterozygous deletion or duplication/triplication) of each rare CNV identified in the proband were compared to that of biological parents to determine the mode of inheritance (*de novo* or inherited).

Clinical Interpretation of Copy Number Variants

Parental inheritance assignment was required for rare CNVs that involved OMIM disease-causing genes, or disease-causing genes due to haploinsufficient/triplosensitivity in peer-reviewed publications, or by ClinGen Dosage Sensitivity Map,[1] DECIPHER,[2] or gnomAD.[3] Rare CNVs with the mode of inheritance were then classified as pathogenic, likely pathogenic (P/LP), variants of uncertain significance (VUS), likely benign, or benign based on the guidelines of the American College of Medical Genetics and Genomics (ACMG) (Riggs et al., 2019) with criteria, methods, and in-house datasets described in our previous study (Dong et al., 2016; Wang H. et al., 2020).

Verification of Copy Number Variants and the Mode of Inheritance

Rare *de novo* and P/LP inherited CNVs (**Supplementary Tables 1**, **2**) identified in this study were verified by an orthogonal approach, using either a CMA platform or quantitative PCR (qPCR). For each CNV in query, the 44K Fetal DNA Chip v1.0 (Agilent Technologies, Santa Clara, CA, United States) was assessed for sufficient probe coverage in the region of interest (at least five probes). If this criterion was satisfied, CMA was performed for both the proband and parents simultaneously according to the manufacturer's protocols. The CNVs were analyzed *via* CytoGenomics software (Leung et al., 2011; Huang et al., 2014; **Figure 1A**). For CNVs with insufficient probe coverage on our CMA platform, qPCR was performed as previously described (Wang H. et al., 2020). Primers specific to the candidate regions were designed with Primer 3 Web, Primer-Blast (NCBI), and *In Silico* PCR (UCSC) based on the reference genome (GRCh37/hg19). Melting curve analysis was carried out for each pair of primers to ensure specificity of the PCR amplification, and the standard curve method was used to determine PCR efficiency (within a range from 95 to 105%). Each reaction was performed in triplicate in 10 μl of reaction mixtures simultaneously in cases, parents, and control (in-house normal male and female controls) using the SYBR Select Master Mix (Applied Biosystems). The reactions were run on a 7900HT Real-Time PCR System (Applied Biosystems) using the default reaction conditions. The copy numbers in each sample were determined by the ΔΔCt (cycle threshold) method, which compared the difference in Ct of the targeted region with a reference primer pair targeting a universally conserved element in a case against a control. qPCR using two independent pairs of

[1]https://dosage.clinicalgenome.org/
[2]https://www.deciphergenomics.org/
[3]https://gnomad.broadinstitute.org/

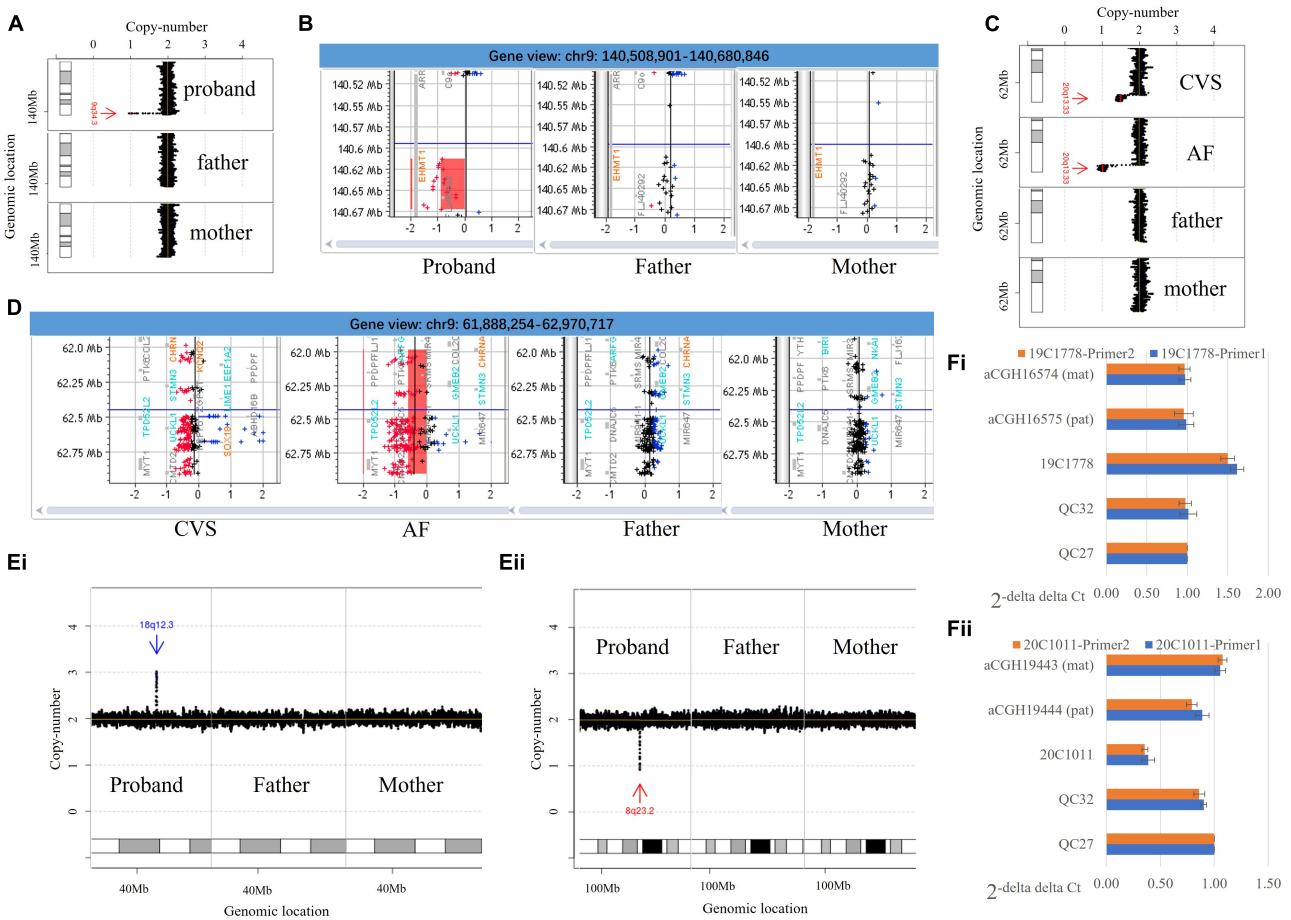

FIGURE 1 | Copy number variants identified by low-pass genome sequencing and the verification. **(A)** Low-pass genome sequencing (GS) identified a 64.7-kb *de novo* heterozygous deletion seq[GRCh37/hg19] del(9)(q34.3)dn chr9:g.140608441_140673160del in case 20C0475. The heterozygous deletion is indicated by a red arrow and was only detected in the proband. **(B)** Chromosomal microarray analysis (CMA) results confirmed a *de novo* heterozygous deletion involving *EHMT1*. **(C)** Low-pass GS detected a 1.0-Mb *de novo* deletion at approximately 50% mosaic level at 20q terminal seq[GRCh37/hg19] del(20)(q13.33)dn chr20:g.61942378_62945038del[0.5] in case 19C3563 on chorionic villi samples (CVS) samples and further identified a constitutional deletion with the same coordinates in an amniotic fluid (AF) sample submitted from the same case. **(D)** CMA analysis confirmed the findings. In panels **(A,C)**, results from low-pass GS: the *X*-axis indicates the copy number of each window (shown as black dot), while the *Y*-axis represents the genomic coordinates. In panels **(B,D)**, results from CMA: the probes in red, black, and blue represent copy number loss, neutral, and gain, respectively. The *X*-axis indicates the log(2) copy ratio of each probe, while the *Y*-axis represents the genomic coordinates. **(Ei)** Low-pass GS identified a *de novo* duplication in 18q12.3 seq[GRCh37] dup(18)(q12.3)dn chr18:g.42995657_43104692dup, which is indicated by a blue arrow. **(Fi)** Quantitative PCR (qPCR) with two independent pairs of primers targeting the region with duplication using QC27 as control. The results indicate a copy number gain in case 19C1778 and copy number neutral in QC32 as well as in each of the parents (mother: aCGH16574; father: aCGH16575). **(Eii)** Low-pass GS identified a *de novo* heterozygous deletion in 8q23.2 seq[GRCh37] del(8)(q23.2)dn chr8:g.111229945_111294607del, which is indicated by a red arrow. **(Fii)** qPCR with two independent pairs of primers targeting the deleted region using QC27 as control. The results indicate a copy number loss in case 20C1011 and copy number neutral in QC32 as well as each of the parents (mother: aCGH19443; father: aCGH19444). In panels **(A,C)**, the *X*-axis indicates the copy number of each window (shown as black dot), while the *Y*-axis represents the genomic coordinates.

primers (**Supplementary Table 3** and **Figure 1B**) was performed in triplicate to verify each rare *de novo* CNV in each trio.

Copy Number Variants Curated From ClinVar and in Other Publications

Copy Number Variants curated from ClinVar (Landrum et al., 2014) were downloaded on December 15, 2020, from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz. There were 209,120 variants in total (77,201 copy number gains/duplications and 131,919 copy number losses/deletions). CNVs with conflicting CNV classification were filtered out. There were 4,416 CNVs with sizes no less than 50-kb available for further comparison (GRCh37/hg19).

We also curated CNVs from several published studies on the spectrum of CNVs in prenatal diagnosis for comparison: (1) 428 P/LP CNVs detected in 23,865 prenatal diagnosis cases by CMA, of which clinically relevant CNVs smaller than 10 Mb were included; 25% of the CNVs did not have parental inheritance information (Chau et al., 2019), (2) 51 P/LP CNVs detected in 3,429 cases by low-coverage GS; fetuses with ultrasound anomalies were excluded and

only 20% of the CNVs had parental inheritance information (Wang et al., 2018); and (3) 217 CNVs (seven P/LP) in 111 cases by trio-based high read-depth GS (Zhou et al., 2021), in which parental inheritance was not provided for CNVs smaller than 100 kb.

Statistical Analysis

The incidence of CNVs stratified by clinical classification, mode of heritance, and referral indication for invasive testing is shown as proportions with 95% confidence intervals calculated with the Wilson score method without continuity correction. In addition, Kruskal–Wallis rank-sum test was performed to compare the CNV parameters, including the type of aberration, the size, constitutional/mosaicism, and the mode of inheritance in different studies. Lastly, chi-square test or Fisher exact test was performed to compare the incidence of small CNVs between studies. All statistical analyses were performed with the statistical software package SPSS 25.0 (IBM SPSS Statistics for Windows, version 25.0; IBM Corp., Armonk, NY, United States).

RESULTS

From January 2019 to February 2021, 315 women referred for trio-based prenatal genetic diagnosis at our clinical laboratory were enrolled. There were 54 CVS, 257 AF samples, and 4 fetal cord blood samples. Parental peripheral blood samples were available for all cases. Demographic information including maternal and paternal age, and the gestational week are shown in **Supplementary Table 1**. After exclusion of maternal cell admixture by qfPCR, all cases were subjected to low-pass GS for CNV analysis. An average of 18 million reads were generated per case, which was equivalent to 0.3-fold. Overall, trio-based low-pass GS provided a 12.4% (39/315) diagnostic yield among the 315 cases (**Table 1**).

Rare Copy Number Variants Identified in Trio-Based Analysis

Low-pass GS identified 14 constitutional or mosaic aneuploidies in 13 cases (4.1%, 13/315, **Supplementary Materials**). CNV analysis revealed 603 rare CNVs (>50 kb, homozygous/hemizygous deletion > 10 kb) including 597 constitutional and six mosaic CNVs in 272 fetuses (272/315, 86.3%, **Figure 2A**), providing roughly 1.9 rare CNVs per case (603/315). On average, 8.84 RefSeq genes were involved in each rare CNV. The majority of rare CNVs were smaller than 1 Mb (562/603, 93.2%), while the six mosaic CNVs were all larger than 1 Mb (**Figure 2A**). We further compared the size distribution of CNVs to those reported by Zhou et al. (2021) in a trio-based high read-depth GS study utilizing an independent algorithm (n = 111). The results indicated the size distributions were significantly different (Kruskal–Wallis rank sum test: $p = 0.0054$, **Figure 2B**). The large mosaic CNVs reported and the large proportion of small CNVs (<100 kb) in our study may explain the differences in size distribution (Zhou et al., 2021).

The size distribution of the 603 rare CNVs also showed significant difference compared with CNVs curated in ClinVar (n = 4,416, Kruskal–Wallis rank sum test: $p < 0.0001$, **Figure 3A**). Although our study shared a similar proportion of CNVs ranging from 100 kb to 1 Mb in size (52.74 vs. 55.75%, Chi-square test, $p = 0.1631$), the percentage of small CNVs (from 50 to 100 kb) in our study was significantly higher that of ClinVar (40.47 vs. 4.82%), with over eightfold increase (Chi-square test, $p < 0.0001$).

Mode of Inheritance

Of the 603 rare CNVs, 46 were *de novo* (in 36 cases, 11.4%, 36/315, **Supplementary Table 2**) and 557 were inherited (in 248 cases, 78.7%, **Figure 3A**). The size distribution was significantly different between *de novo* and inherited CNVs (Kruskal–Wallis rank sum test: $p < 0.0001$, **Figure 3B**). The majority of small CNVs (50–100 kb) in our study were inherited (239/244, 97.95%). In comparison, *de novo* CNVs were larger in size compared with inherited CNVs. They also involved significantly more RefSeq genes (**Supplementary Figure 1**).

Among the *de novo* CNVs, 40 were constitutional and six were mosaic (**Figure 3C**), providing a frequency of 0.15 *de novo* CNVs per case (46/315). On average, there were 92.3 RefSeq genes involved in each *de novo* CNV (median: 25 genes). All *de novo* CNVs were validated by CMA or qPCR (see "Materials and Methods" and **Figure 1**).

There was no significant difference between the size distributions of *de novo* CNVs between our study and ClinVar (Kruskal–Wallis rank sum test: $p = 0.785$, **Figure 3A**). However, the proportion of small *de novo* CNVs (50–100 kb) was significantly higher than that curated in ClinVar (10.9 vs. 2.7%, Chi-square test: $p = 0.0013$). In addition, there were no significant differences between parental age and the incidence of *de novo* CNVs (**Supplementary Figure 2**).

Rare Copy Number Variant Classification and Trio-Based Analysis

In this study, we also aimed to determine the percentage of cases with rare CNVs requiring information of parental assignment after proband-only interpretation, which is critical for genetic counseling and consideration for trio-based CNV analysis. We then classified the clinical significance of 603 rare CNVs identified in fetuses following the ACMG guidelines (Riggs et al., 2020). There were 84 rare CNVs in 74 cases that involved disease-causing genes, of which the mode of inheritance was important for the clinical interpretation and estimation of recurrence risk (23.5%, 74/315, see "Materials and Methods" and **Supplementary Table 3**). The 84 CNVs had a different size distribution compared with the overall 603 rare CNVs (median size: 725 vs. 126 kb, Kruskal–Wallis rank sum test: $p < 0.0001$, **Figure 4A**). In light of parental inheritance assignment, 31 CNVs (in 26 cases) were classified as P/LP CNVs (**Supplementary Table 3**), 18 as VUS (in 18 cases), and 35 as benign CNVs. Among the 31 P/LP CNVs (in 26 cases), 25 were *de novo* CNVs, and 6 were inherited. Low-pass GS provided a diagnostic yield of 8.2% (26/315, **Table 1**). In addition, among the 18 VUS, 5 were *de novo* and 13 were inherited. The incidence of VUS (18/315, 5.7%) was not significantly different from our previous prospective

TABLE 1 | Diagnostic yield in cases with different referral indications.

Clinical indication	Cases enrolled	Cases with diagnosis	Diagnostic yield (%)	[95% CI]	Cases with pathogenic findings (inherited or *de novo*)*	Number of Dup/del	Cases #
Abnormal ultrasound	165	15	9.09%	[5.35, 14.81]	Cases with inherited P/LP CNVs	Del: 3 Dup: 1	4(0.02)
					Cases with *de novo* P/LP CNVs	Del: 3 Dup: 4	7(0.04)
					Cases with aneuploidies	Del: 1 Dup: 3	4(0.02)
Non-invasive prenatal screening - high risk	70	19	27.14%	[17.52, 39.30]	Cases with inherited P/LP CNVs	Del: 1 Dup: 0	1(0.01)
					Cases with *de novo* P/LP CNVs	Del: 8 Dup: 1	9(0.13)
					Cases with aneuploidies	Del: 2 Dup: 7	9(0.13)
1st/2nd Trimester aneuploidy screening high risk (DSS)	16	3	18.75%	[4.97, 46.31]	Cases with inherited P/LP CNVs	Del: 1 Dup: 0	1(0.06)
					Cases with *de novo* P/LP CNVs	Del: 2 Dup: 0	2(0.13)
					Cases with aneuploidies		0(0)
Advanced maternal age	11	0	0.00%	[0, 32.14]	-		–
Family history	31	2	6.45%	[1.12, 22.84]	Cases with inherited P/LP CNVs		0(0)
					Cases with *de novo* P/LP CNVs	Del: 1# Dup: 2#	2(0.06)
					Cases with aneuploidies		0(0)
Others	22	0	0.00%	[0, 18.5]	Cases with inherited P/LP CNVs		0(0)
					Cases with *de novo* P/LP CNVs		0(0)
					Cases with aneuploidies		0(0)
Total	315	39	12.38%	[9.05, 16.65]	-		39(0.12)

**P/LP refers to pathogenic or likely pathogenic.*

#Each digit in the bracket refers to the incidence over the sample enrolled in each subgroup, #20C2527 with duplication and deletion at the same time: seq[GRCh37] dup(Y)(p11.31q11.221)dn chrY:g.2649473_19567688dup, seq[GRCh37] del(Y)(q11.221q12)dn chrY:g.19567689_59033394del.

study that performed parental inheritance assignment in a sequential approach (53/1,023, 5.2%, Chi-square test, $p = 0.7119$). Overall, the 31 P/LP CNVs also had significant differences in size distributions compared with 84 CNVs requiring parental analysis (Kruskal–Wallis rank sum test: $p = 0.0067$, **Figure 4A**). Among different classifications, *de novo* CNVs tended to be larger in size compared with inherited CNVs, particularly P/LP CNVs (**Figure 4B**).

There were five cases with 22q11.2 deletion syndrome: four cases occurred *de novo*, while one case was maternally inherited. Although all five deletions were classified as P/LP CNVs, their recurrence risks would be different (McDonald-McGinn and Zackai, 2008).

Clinical Interpretation of Copy Number Variants Based on the Mode of Inheritance

De novo CNVs in our cohort are more likely to be classified as P/LP than inherited CNVs (54.35% [25/46] vs. 1.58% [6/567], Chi-square test, $p < 0.0001$). All mosaic CNVs were classified as P/LP CNVs except the 20q13.33 deletion (VUS, **Supplementary Table 3**). Case 19C3563 was referred for cardiomegaly with abnormal tricuspid valve and abnormal ± pulmonary valve at 15 + 4 gestational weeks. A *de novo* 1.0-Mb deletion of approximately 50% mosaic level was detected, seq[GRCh37/hg19] del(20)(q13.33)dn chr20:g.61942378_62945038del[0.5], and further confirmed by CMA (**Figures 1C,D**). The gene *KCNQ2* was involved, the haploinsufficiency of which causes neonatal seizures (Heron et al., 2007) and encephalopathy (Spagnoli et al., 2018). The deletion was classified as VUS. To exclude the possibility of confined placental mosaicism, low-pass GS was further performed on the AF sample collected at a later gestational week and revealed a constitutional 20q deletion, further confirmed by CMA (**Figures 1C,D**). After genetic counseling, the couples opted for termination of pregnancy.

Among all P/LP CNVs, three were smaller than 100 kb. Two cases had Southeast Asian (SEA) type homozygous deletions resulting in α-thalassemia major (19.3-kb deletions

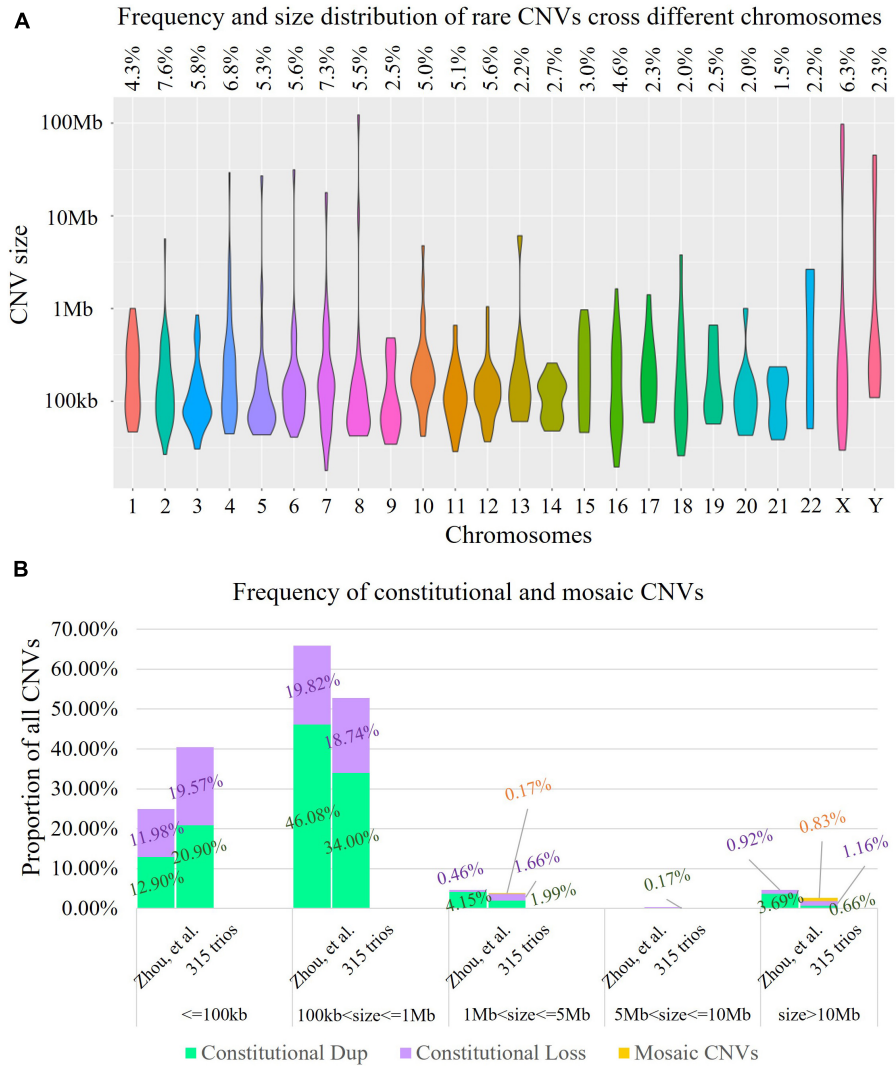

FIGURE 2 | Landscape of rare copy number variants (CNVs). **(A)** Distribution of 603 rare CNVs identified cross different chromosomes (violin plot). The *X*-axis presents different chromosomes, while the *Y*-axis indicates the number of rare CNVs identified (in log10 format). The frequency of rare CNVs in each chromosome is provided in the top panel. **(B)** Comparison of the rare CNVs identified in our study and in a trio-based high read-depth genome sequencing study (Zhou et al., 2021; $n = 111$). In each bar, each segment in green, purple, and yellow indicates the percentage of constitutional duplications (digits in dark green), constitutional deletions (digits shown in purple), and mosaic CNVs (digits in orange) identified. The results indicated that the size distributions were significantly different (Kruskal–Wallis rank sum test: $p = 0.0054$).

due to biparental inheritance), while another case had a *de novo* pathogenic deletion. Case 20C0475 was referred for invasive testing at 16 + 2 gestational weeks due to high-risk Down syndrome screening results (risk at 1:2) and advanced maternal age. Low-pass GS detected a 64.7-kb *de novo* heterozygous deletion seq[GRCh37/hg19] del(9)(q34.3)dn chr9:g.140608441_140673160del involving the exons 3–12 of *EHMT1*, which was confirmed by CMA (**Figures 1A,B**). Haploinsufficiency of *EHMT1* is known to cause Kleefstra syndrome 1 in an autosomal dominant manner (OMIM #610253). The deletion was classified as pathogenic, and the pregnancy was terminated after genetic counseling. Overall, the most common P/LP CNV identified was recurrent 22q11.2 microdeletion associated with DiGeorge syndrome ($n = 5$), while the other cases had isolated CNVs.

To further investigate whether the size distribution of P/LP CNVs in our cohort was different from previously reported studies, we further curated the CNVs reported in three prenatal studies: 23,865 cases by CMA (Chau et al., 2019), 3,429 cases by low-coverage GS (Wang et al., 2018), and 111 cases by high read-depth GS (Zhou et al., 2021) (see "Materials and Methods"). Parental confirmation was performed in a sequential manner in the first two studies. The size distributions among P/LP CNVs in all studies were significantly different (Kruskal–Wallis rank sum test: $p < 0.0001$). In addition, the size distributions of *de novo* P/LP CNVs in all studies were also significantly

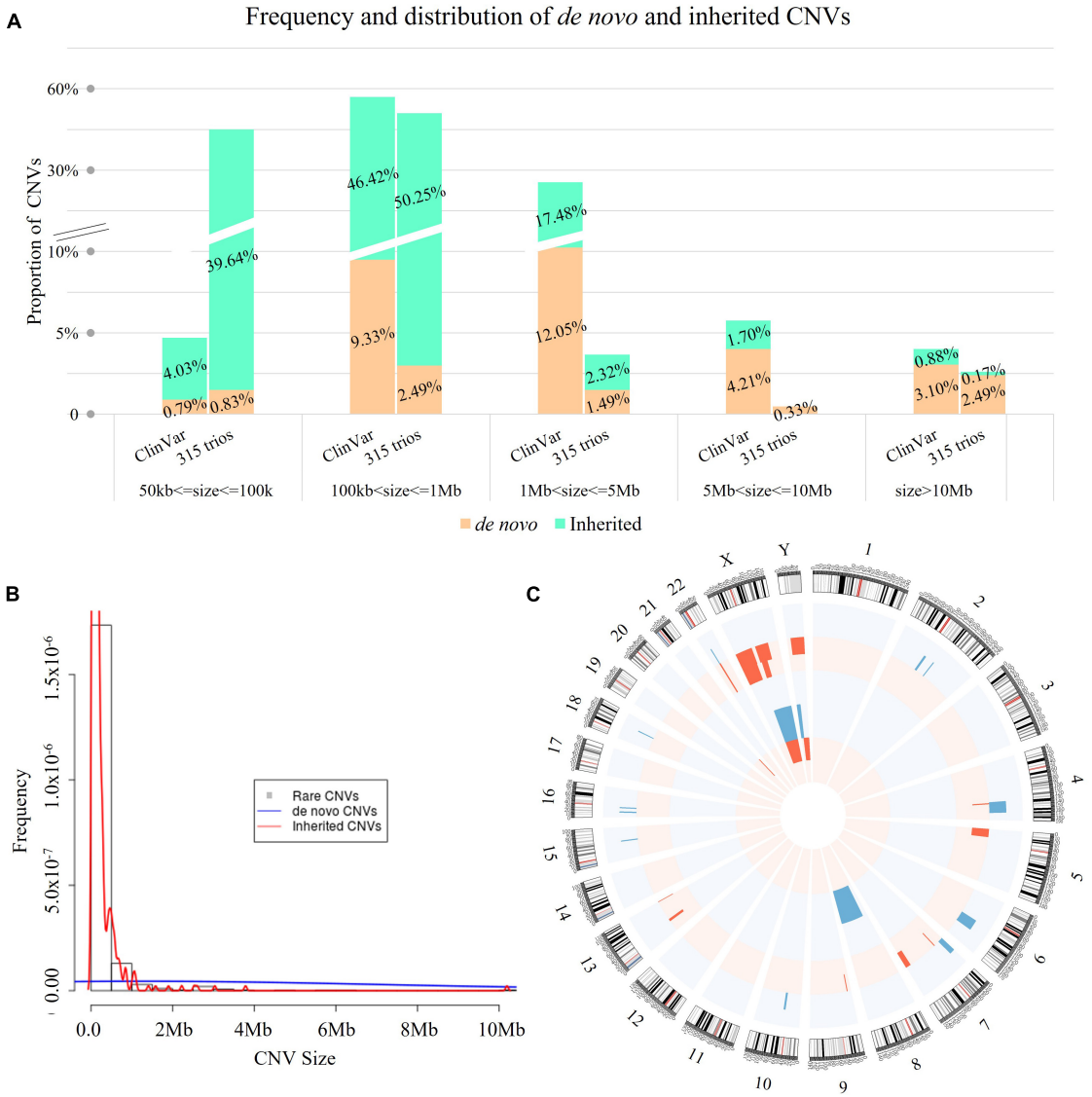

FIGURE 3 | Spectrum of rare CNVs with mode of inheritance. **(A)** Comparison of the rare CNVs identified in our study and in ClinVar (n = 4,416) with mode of inheritance. In each bar, each digit in red indicates the percentage of inherited CNVs (cyan bar), while each digit in black represents the percentage of *de novo* CNVs (tan bar). The size distribution of the 603 rare CNVs showed significant difference compared with CNVs curated in ClinVar (n = 4,416, Kruskal–Wallis rank sum test: $p < 0.0001$), but not for *de novo* CNVs Kruskal–Wallis rank sum test: $p = 0.785$). **(B)** Histogram of rare CNVs. The density lines in red and blue reflect the size distribution of inherited and *de novo* CNVs, respectively. The size distribution was significantly different between *de novo* and inherited CNVs (Kruskal–Wallis rank sum test: $p < 0.0001$). **(C)** Distribution of *de novo* CNVs identified in our study. Blue bars represent copy number gains and red bars represent copy number losses encompassing the chromosomal bands. The height represents the frequency of the pathogenic copy number variants. The outer circle indicates the distribution of mosaic *de novo* CNVs, while the inner circle presents the distribution of constitutional *de novo* CNVs.

different (Kruskal–Wallis rank sum test: $p < 0.0001$, **Figure 4C**). Particularly, the sizes of all P/LP CNVs and all *de novo* P/LP CNVs in our study were both significantly different from the ones curated in ClinVar (Kruskal–Wallis rank sum test: $p = 0.0002$ and $p < 0.0001$, respectively). In addition, both of them were also significantly different from the ones reported by the study with 3,429 cases by low-coverage GS (Wang et al., 2018) (Kruskal–Wallis rank sum test: $p = 0.0007$ and $p = 0.0013$, respectively). It could be explained by the presence of *de novo* or inherited small P/LP CNVs (<100 kb) and mosaic P/LP CNVs in our study, which accounted for 25.8% of the P/LP CNVs (8/31).

Incidence of Rare Copy Number Variants in Subgroups With Different Primary Referral Indications

In addition, we further calculated the frequency of rare CNVs and *de novo* CNVs based on the primary referral indications (**Table 2**). Subgroups of cases with abnormal ultrasound findings

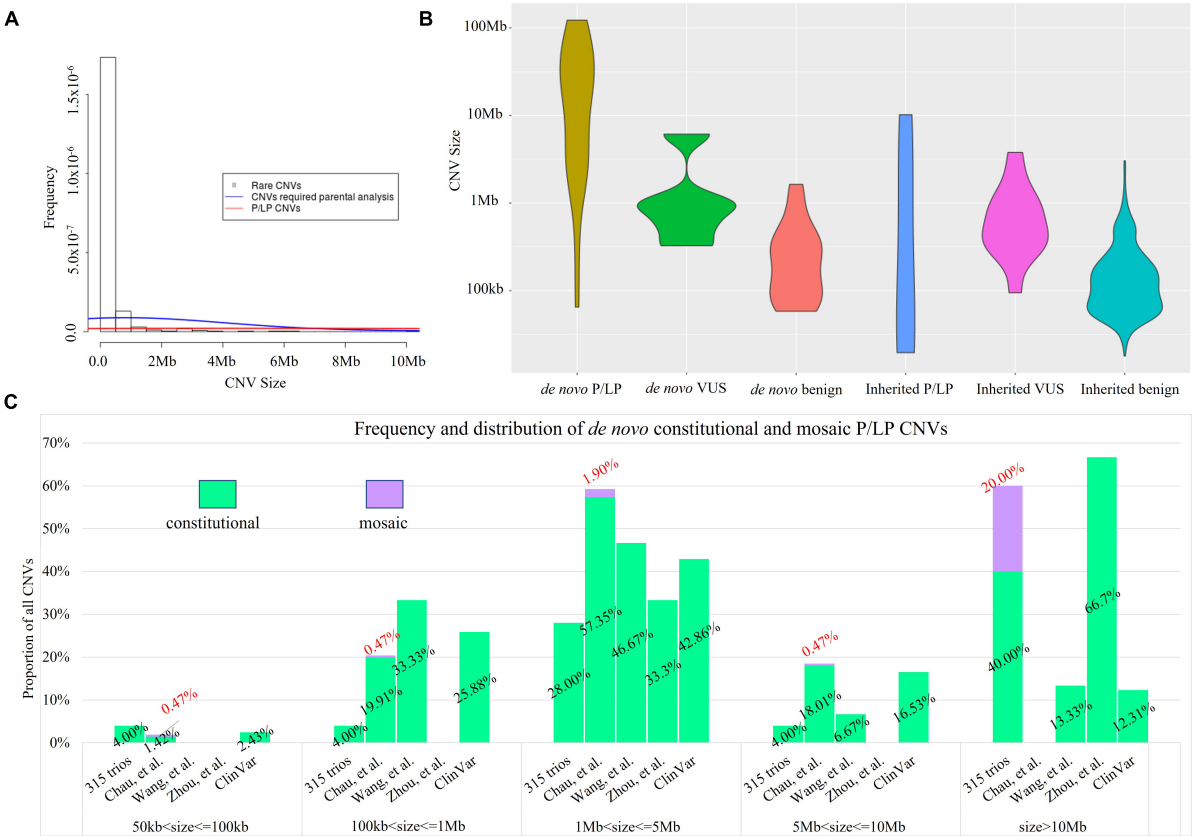

FIGURE 4 | Size distribution of rare CNVs with clinical classification. **(A)** Histogram of rare CNVs. The density lines in red and blue reflect the size distribution of 84 rare CNVs requiring parental analysis and 31 P/LP CNVs, respectively. The median size of the 84 rare CNVs was 725 kb vs. 126 kb of all 603 rare CNVs (Kruskal–Wallis rank sum test: $p < 0.0001$). **(B)** Size distribution of *de novo* and inherited CNVs and the classification. **(C)** Comparison of the size distribution of pathogenic or likely pathogenic CNVs identified in our study with ones reported in a trio-based high read-depth genome sequencing study (Zhou et al., 2021; $n = 111$), the CMA study with largest number of prenatal cases ($n = 23,865$; Chau et al., 2019), and a study with 3,429 prenatal cases by low-coverage GS (Wang et al., 2018). Bars in light green and purple indicate the percentage of CNVs identified in constitutional and mosaic form, respectively.

and cases with high risk of non-invasive prenatal testing were the two groups with the highest number of cases enrolled (165 vs. 70), and they shared similar incidences of rare CNVs (**Table 2**). However, the incidence of cases with P/LP *de novo* CNVs in high-risk cases from non-invasive testing (9/70, 12.9%) was higher than cases referred with ultrasound anomalies (4.2%, 7/165, **Table 2**). For the incidences of cases with rare CNVs with small size (<100 kb) or mosaic CNVs, cases with *de novo* small CNVs or mosaic CNVs, and cases with rare CNVs requiring parental analysis, all were similar between these two subgroups (**Table 2**).

DISCUSSION

This is a prospective study of trio-based low-pass GS in prenatal diagnosis, providing the landscape of rare CNVs and the mode of inheritance. Among the 315 fetuses, CNV analysis revealed 603 rare CNVs, namely, 597 constitutional and 6 mosaic CNVs in 272 fetuses (272/315, 86.3%). On average, 1.9 rare CNVs were detected per fetus (603/315). In a previous study on rare CNVs, the array-based method reported a frequency of 0.59 rare CNVs per case (Ruderfer et al., 2016). The average 1.9 rare CNVs identified per fetus in prenatal diagnosis is in line with expectations as GS provides improved genome coverage compared to CMA, albeit at a low-pass/low-coverage setting, shown by our previous studies (Chau et al., 2020; Wang H. et al., 2020).

The majority of CNVs detected in our study were smaller than 1 Mb (562/603, 93.2%), while 1% (6/603) were mosaic. Among all 603 rare CNVs, 46 were *de novo* (7.6%, 46/603), which were detected in 11.4% (36/315) of cases. Overall, 12.4% (39/315 vs. 13.5%, 138/1,023) of cases had pathogenic findings (aneuploidies and/or P/LP CNVs) and 5.7% (18/315 vs. 5.2%, 53/1,023) of cases had VUS, both of which were consistent with our previous study where parental inheritance assignment was performed in a sequential manner (Wang H. et al., 2020). Performing trio-based low-pass GS simultaneously or sequentially do not affect the overall diagnostic yield. However, a sequential approach would increase the turnaround time of testing. In addition, the percentage of cases with rare CNVs requiring information of parental assignment after proband-only interpretation based on ACMG guidelines was 23.5 (74/215, 84 CNVs in 74 fetuses). It

TABLE 2 | Incidence and classification of rare copy number variants (CNVs) in cases with different referral indications.

Clinical indication	Cases enrolled	Rare CNVs		*De novo* CNVs		Rare CNVs less than 100 kb or in mosaicism		*De novo* CNVs less than 100 kb or in mosaicism		Rare CNVs required parental analysis	
		Cases	Number	Cases	Number	Cases	Number	Cases	Number	Cases	Number
Abnormal ultrasound	165	141 (0.85)	304 (1.84)	17 (0.1)	21 (0.12)	89 (0.53)	127 (0.76)	5 (0.03)	5 (0.03)	41 (0.24)	47 (0.28)
Non-invasive prenatal screening – high risk	70	64 (0.91)	145 (2.07)	13 (0.18)	17 (0.24)	40 (0.57)	59 (0.84)	2 (0.02)	2 (0.02)	18 (0.25)	21 (0.3)
1st/2nd Trimester aneuploidy screening high risk (DSS)	16	15 (0.93)	31 (1.93)	3 (0.18)	3 (0.18)	9 (0.56)	12 (0.75)	1 (0.06)	1 (0.06)	6 (0.37)	6 (0.37)
Advanced maternal age	11	9 (0.81)	13 (1.18)	0 (0)	0 (0)	3 (0.27)	4 (0.36)	0 (0)	0 (0)	0 (0)	0 (0)
Family history	31	26 (0.83)	61 (1.96)	2 (0.06)	3 (0.09)	18 (0.58)	28 (0.9)	1 (0.03)	2 (0.06)	6 (0.19)	7 (0.22)
Others	22	17 (0.77)	49 (2.22)	1 (0.04)	2 (0.09)	11 (0.5)	20 (0.9)	1 (0.04)	1 (0.04)	3 (0.13)	3 (0.13)
Total	315	272 (0.86)	603 (1.91)	36 (0.11)	46 (0.14)	170 (0.53)	250 (0.79)	10 (0.03)	11 (0.03)	74 (0.23)	84 (0.26)

Each digit in the bracket refers to the incidence over the sample enrolled in each subgroup.

would provide potential clinical implications regarding genetic counseling and consideration for trio-based CNV analysis. Nonetheless, for pregnancy management and decision-making that are highly dependent on timely test results, trio-based approach may be recommended.

Among the 315 cases, 603 rare CNVs (allele frequency < 1% in our curated reference dataset of Chinese fetuses (Dong et al., 2016; Chau et al., 2020; Wang H. et al., 2020): $n > 2{,}000$) were detected, providing an incidence of 1.9 rare CNVs per case (603/315). Of these variants, 40.5% (244/603) were smaller than 100 kb. ClinVar is a database that archives reports of relationships among human genomic variants and phenotypes, with supporting evidence. However, a significant proportion of CNVs submitted to ClinVar was identified by the CMA platform. The differences in size distribution of CNVs between our study and ClinVar may be caused by platform differences. In particular, the percentage of small CNVs (from 50 to 100 kb) in our study was significantly higher than the one curated in ClinVar (40.5 vs. 4.8%) with an over eightfold increase (Chi-square test, $p < 0.0001$). In addition, the percentage of small (50–100 kb) *de novo* CNVs (5/46, 10.9%) was still significantly higher than curated in ClinVar (2.67%, 35/1302) (Chi-square test, $p = 0.0012$). This illustrates a deficiency of rare and small CNVs curated in ClinVar, which would be helpful for laboratory reference in CNV interpretation. Gradual deposition of rare and small CNVs identified by GS would benefit and facilitate prenatal diagnosis of clinically relevant CNVs. Our study not only found P/LP CNVs smaller than 100 kb (*de novo* or inherited), accounting for 9.67% of all detected P/LP CNVs (3/31), but also provided evidence that *de novo* mosaic P/LP CNVs contributed to a significant proportion of pathogenic findings (16.1%, 5/31). Both types of CNVs were not reported in a study with 3,429 prenatal cases by low-coverage GS (Wang et al., 2018). The possible reasons might include the exclusion of cases with ultrasound anomalies and limited resolution of their analysis pipeline (100 kb) (Wang et al., 2018). Significant differences in CNV size distributions between our study and previous studies with different methods (Wang et al., 2018; Zhou et al., 2021) was observed, which might be caused by the different analysis pipelines used. Our findings emphasize the important clinical implication of small CNVs and mosaic CNVs in prenatal diagnosis and warrants a CNV detection method sensitive to small and mosaic variants.

We provided the size distributions of rare CNVs, CNVs requiring parental analysis, and P/LP CNVs. The high abundance of small CNVs was largely contributed by inherited CNVs; clinical interpretation and estimation of recurrence risk largely relied on the mode of inheritance. Parental mode of inheritance assignment was important in nearly a quarter of cases. Recently, there are publications showing the performance of their in-house CNV detection methods using low-pass GS data (Wang et al., 2018; Wang et al., 2020); however, the software/pipelines are not publicly available. As our study aimed to investigate the spectrum and characteristics of rare CNVs, a fair comparison of different methods using low-coverage/low-pass GS for CNV study including pros and cons is warranted in a future study.

A major strength of this study includes the prospective study of 315 prenatal cases with a variety of different clinical indications undergoing invasive prenatal diagnosis. The rare CNV findings represent the spectrum and incidence of *de novo* and inherited CNVs identified among prenatal testing by low-pass GS. Furthermore, our analysis provided a view of rare CNVs by low-pass GS in prenatal diagnosis.

Limitations include (1) limited sample size ($n = 315$) and (2) limited CNV detection resolution (50 kb, homozygous/hemizygous deletion: 10 kb): although our study provided an enhanced resolution compared with the reported studies by GS (Wang et al., 2018) and CMA (Chau et al., 2019), the spectrum and incidence of smaller CNVs (<50 kb) are still not well studied. There are large amounts of small CNVs in human genomes (Collins et al., 2020); trio-based GS analyses using higher read-depths (increased resolution) and larger sample sizes are warranted in future studies. In addition, read-depth-based CNV analysis is unable to assemble derivative chromosomes or identify the genomic locations and orientations of copy number gains. Paired-end sequencing approaches (Talkowski et al., 2012; Dong et al., 2018, 2019, 2021) to further delineate the locations and the breakpoint junctions of CNVs may provide a more comprehensive understanding of prenatally

detected CNVs. Particularly, apparently *de novo* deletions or duplications might be caused by balanced rearrangements (such as insertions) in the parents (Nowakowska et al., 2012). Low-pass GS does not detect single-nucleotide variants (SNVs) and small insertions/deletions (indels) that can also be pathogenic in the prenatal context. Early studies have revealed promising diagnostic utility of prenatal ES for the detection of pathogenic SNVs and indels in fetuses with structural abnormalities. Further studies are warranted to examine the clinical utility of prenatal ES to guide its clinical implementation. Lastly, future studies on *de novo* variants in prenatal diagnosis may be extended to the investigation of SNVs/indels (Lord et al., 2019; Petrovski et al., 2019).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MC, YK, TL, ZD, and KC designed the study. WT and TL collected the samples and followed up. JQ, ZC, MC, and ZD performed the analysis and data interpretation. JQ, YL, and YZ conducted the validation. MC, JQ, ZD, and KC wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by the National Natural Science Foundation of China (31801042), Health and Medical Research Fund (04152666, 07180576), General Research Fund (14115418), and Direct Grant (2019.051). KC thank the Hong Kong Obstetrical & Gynaecological Trust Fund for the support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.742325/full#supplementary-material

REFERENCES

Asadollahi, R., Oneda, B., Joset, P., Azzarello-Burri, S., Bartholdi, D., Steindl, K., et al. (2014). The clinical significance of small copy number variants in neurodevelopmental disorders. *J. Med. Genet.* 51, 677–688.

Chau, M. H. K., Cao, Y., Kwok, Y. K. Y., Chan, S., Chan, Y. M., Wang, H., et al. (2019). Characteristics and mode of inheritance of pathogenic copy number variants in prenatal diagnosis. *Am. J. Obstet. Gynecol.* 221, e1–e493.

Chau, M. H. K., and Choy, K. W. (2021). The role of chromosomal microarray and exome sequencing in prenatal diagnosis. *Curr. Opin. Obstet. Gynecol.* 33, 148–155. doi: 10.1097/gco.0000000000000692

Chau, M. H. K., Wang, H., Lai, Y., Zhang, Y., Xu, F., Tang, Y., et al. (2020). Low-pass genome sequencing: a validated method in clinical cytogenetics. *Hum. Genet.* 139, 1403–1415. doi: 10.1007/s00439-020-02185-9

Chaubey, A., Shenoy, S., Mathur, A., Ma, Z., Valencia, C. A., Reddy Nallamilli, B. R., et al. (2020). Low-pass genome sequencing: validation and diagnostic utility from 409 clinical cases of low-pass genome sequencing for the detection of copy number variants to replace constitutional microarray. *J. Mol. Diagn.* 22, 823–840.

Choy, K. W., Kwok, Y. K., Cheng, Y. K., Wong, K. M., Wong, H. K., Leung, K. O., et al. (2014). Diagnostic accuracy of the BACs-on-beads assay versus karyotyping for prenatal detection of chromosomal abnormalities: a retrospective consecutive case series. *BJOG* 121, 1245–1252. doi: 10.1111/1471-0528.12873

Choy, K. W., Wang, H., Shi, M., Chen, J., Yang, Z., Zhang, R., et al. (2019). Prenatal diagnosis of fetuses with increased nuchal translucency by genome sequencing analysis. *Front. Genet.* 10:761. doi: 10.3389/fgene.2019.00761

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alfoldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nature* 581, 444–451.

Dong, Z., Chau, M. H. K., Zhang, Y., Dai, P., Zhu, X., Leung, T. Y., et al. (2021). Deciphering the complexity of simple chromosomal insertions by genome sequencing. *Hum. Genet.* 140, 361–380. doi: 10.1007/s00439-020-02210-x

Dong, Z., Wang, H., Chen, H., Jiang, H., Yuan, J., Yang, Z., et al. (2018). Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the future of clinical cytogenetics. *Genet. Med.* 20, 697–707. doi: 10.1038/gim.2017.170

Dong, Z., Yan, J., Xu, F., Yuan, J., Jiang, H., Wang, H., et al. (2019). Genome sequencing explores complexity of chromosomal abnormalities in recurrent miscarriage. *Am. J. Hum. Genet.* 105, 1102–1111.

Dong, Z., Zhang, J., Hu, P., Chen, H., Xu, J., Tian, Q., et al. (2016). Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet. Med.* 18, 940–948. doi: 10.1038/gim.2015.199

Heron, S. E., Cox, K., Grinton, B. E., Zuberi, S. M., Kivity, S., Afawi, Z., et al. (2007). Deletions or duplications in KCNQ2 can cause benign familial neonatal seizures. *J. Med. Genet.* 44, 791–796. doi: 10.1136/jmg.2007.051938

Huang, J., Poon, L. C., Akolekar, R., Choy, K. W., Leung, T. Y., and Nicolaides, K. H. (2014). Is high fetal nuchal translucency associated with submicroscopic chromosomal abnormalities on array CGH? *Ultrasound Obstet. Gynecol.* 43, 620–624. doi: 10.1002/uog.13384

Huijsdens-van Amsterdam, K., Straver, R., van Maarle, M. C., Knegt, A. C., Van Opstal, D., Sleutels, F., et al. (2018). Mosaic maternal 10qter deletions are associated with FRA10B expansions and may cause false-positive noninvasive prenatal screening results. *Genet. Med.* 20, 1472–1476. doi: 10.1038/gim.2018.32

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985.

Leung, T. Y., Vogel, I., Lau, T. K., Chong, W., Hyett, J. A., Petersen, O. B., et al. (2011). Identification of submicroscopic chromosomal aberrations in fetuses with increased nuchal translucency and apparently normal karyotype. *Ultrasound Obstet. Gynecol.* 38, 314–319. doi: 10.1002/uog.8988

Levy, B., and Wapner, R. (2018). Prenatal diagnosis by chromosomal microarray analysis. *Fertil. Steril.* 109, 201–212. doi: 10.1016/j.fertnstert.2018.01.005

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liang, D., Peng, Y., Lv, W., Deng, L., Zhang, Y., Li, H., et al. (2014). Copy number variation sequencing for comprehensive diagnosis of chromosome disease syndromes. *J. Mol. Diagn.* 16, 519–526. doi: 10.1016/j.jmoldx.2014.05.002

Lord, J., McMullan, D. J., Eberhardt, R. Y., Rinck, G., Hamilton, S. J., Quinlan-Jones, E., et al. (2019). Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* 393, 747–757.

McDonald-McGinn, D. M., and Zackai, E. H. (2008). Genetic counseling for the 22q11.2 deletion. *Dev. Disabil. Res. Rev.* 14, 69–74.

Nowakowska, B. A., de Leeuw, N., Ruivenkamp, C. A., Sikkema-Raddatz, B., Crolla, J. A., Thoelen, R., et al. (2012). Parental insertional balanced translocations are an important cause of apparently de novo CNVs in patients with developmental anomalies. *Eur. J. Hum. Genet.* 20, 166–170. doi: 10.1038/ejhg.2011.157

Petrovski, S., Aggarwal, V., Giordano, J. L., Stosic, M., Wou, K., Bier, L., et al. (2019). Whole-exome sequencing in the evaluation of fetal structural anomalies: a prospective cohort study. *Lancet* 393, 758–767. doi: 10.1016/s0140-6736(18)32042-7

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2019). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genet. Med.* 22, 245–257. doi: 10.1038/s41436-019-0686-8

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genet. Med.* 22, 245–257.

Ruderfer, D. M., Hamamsy, T., Lek, M., Karczewski, K. J., Kavanagh, D., Samocha, K. E., et al. (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* 48, 1107–1111. doi: 10.1038/ng.3638

Spagnoli, C., Salerno, G. G., Iodice, A., Frattini, D., Pisani, F., and Fusco, C. (2018). KCNQ2 encephalopathy: a case due to a de novo deletion. *Brain Dev.* 40, 65–68. doi: 10.1016/j.braindev.2017.06.008

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81.

Talkowski, M. E., Ordulu, Z., Pillalamarri, V., Benson, C. B., Blumenthal, I., Connolly, S., et al. (2012). Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N. Engl. J. Med.* 367, 2226–2232.

Wang, H., Dong, Z., Zhang, R., Chau, M. H. K., Yang, Z., Tsang, K. Y. C., et al. (2020). Low-pass genome sequencing versus chromosomal microarray analysis: implementation in prenatal diagnosis. *Genet. Med.* 22, 500–510. doi: 10.1038/s41436-019-0634-7

Wang, J., Chen, L., Zhou, C., Wang, L., Xie, H., Xiao, Y., et al. (2018). Prospective chromosome analysis of 3429 amniocentesis samples in China using copy number variation sequencing. *Am. J. Obstet. Gynecol.* 219, e1–e287.

Wang, Y., Li, Y., Chen, Y., Zhou, R., Sang, Z., Meng, L., et al. (2020). Systematic analysis of copy-number variations associated with early pregnancy loss. *Ultrasound Obstet. Gynecol.* 55, 96–104. doi: 10.1002/uog.20412

Wong, H. S., Wadon, M., Evans, A., Kirov, G., Modi, N., O'Donovan, M. C., et al. (2020). Contribution of de novo and inherited rare CNVs to very preterm birth. *J. Med. Genet.* 57, 552–557. doi: 10.1136/jmedgenet-2019-106619

Yang, X., Li, R., Fu, F., Zhang, Y., Li, D., and Liao, C. (2017). Submicroscopic chromosomal abnormalities in fetuses with increased nuchal translucency and normal karyotype. *J. Matern. Fetal Neonatal Med.* 30, 194–198. doi: 10.3109/14767058.2016.1168394

Zhou, J., Yang, Z., Sun, J., Liu, L., Zhou, X., Liu, F., et al. (2021). Whole genome sequencing in the evaluation of fetal structural anomalies: a parallel test with chromosomal microarray plus whole exome sequencing. *Genes* 12:376.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 November 2021
doi: 10.3389/fgene.2021.752390

Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole Genome Sequencing Data from the Alzheimer's Disease Sequencing Project

Wan-Ping Lee[1,2,3†], Albert A. Tucci[4†], Mitchell Conery[5,6], Yuk Yee Leung[1,2,3], Amanda B. Kuzma[1], Otto Valladares[1], Yi-Fan Chou[1], Wenbin Lu[7], Li-San Wang[1,2,3], Gerard D. Schellenberg[1,3] and Jung-Ying Tzeng[4,7*]*

[1]Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [2]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [4]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, United States, [5]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, United States, [6]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [7]Department of Statistics, North Carolina State University, Raleigh, NC, United States

Alzheimer's Disease (AD) is a progressive neurologic disease and the most common form of dementia. While the causes of AD are not completely understood, genetics plays a key role in the etiology of AD, and thus finding genetic factors holds the potential to uncover novel AD mechanisms. For this study, we focus on copy number variation (CNV) detection and burden analysis. Leveraging whole-genome sequence (WGS) data released by Alzheimer's Disease Sequencing Project (ADSP), we developed a scalable bioinformatics pipeline to identify CNVs. This pipeline was applied to 1,737 AD cases and 2,063 cognitively normal controls. As a result, we observed 237,306 and 42,767 deletions and duplications, respectively, with an average of 2,255 deletions and 1,820 duplications per subject. The burden tests show that Non-Hispanic-White cases on average have 16 more duplications than controls do (p-value 2e-6), and Hispanic cases have larger deletions than controls do (p-value 6.8e-5).

Keywords: copy number variation—CNV, Alzheimer's disease, whole-genome sequence (WGS), CNV association test, NGS—next generation sequencing

OPEN ACCESS

Edited by:
Claudia Gonzaga-Jauregui, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Audrey Qiuyan Fu, University of Idaho, United States
Nancy Monroy-Jaramillo, National Institute of Neurology and Neurosurgery, Mexico

***Correspondence:**
Wan-Ping Lee
Wan-Ping.Lee@PennMedicine.upenn.edu
Jung-Ying Tzeng
jytzeng@ncsu.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *03 August 2021*
Accepted: *11 October 2021*
Published: *04 November 2021*

Citation:
Lee W-P, Tucci AA, Conery M, Leung YY, Kuzma AB, Valladares O, Chou Y-F, Lu W, Wang L-S, Schellenberg GD and Tzeng J-Y (2021) Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole Genome Sequencing Data from the Alzheimer's Disease Sequencing Project. Front. Genet. 12:752390. doi: 10.3389/fgene.2021.752390

INTRODUCTION

Alzheimer's disorder (AD) is a devastating neurodegenerative disease and is the most common cause of dementia. Approximately 6.2 million Americans are living with AD in 2021, and it is projected to reach 12.7 million in 2050, which makes AD one of the most pressing public health issues (Alzheimer's Association, 2020). Presently, there is no known effective prevention or disease modifying therapies, and the landscape of AD drug trials is gloomy. One possible reason is that AD is a heterogeneous disorder, but trials are designed treating it as a monolithic disease. Although

lifestyle and environmental risk factors clearly affect AD, the primacy of genetic influences suggests that categorization by genetic basis should be prioritized in developing effective interventions.

AD heritability estimates range from 49–79%; however, <50% of this heritability can be explained by genome-wide association studies (GWAS) on single nucleotide variants (SNVs) (Ridge et al., 2013; Sims et al., 20202020). Taking copy number variation (CNV) into consideration may partially mitigate the problem of missing heritability and play an important role in human disease susceptibility (Cooper et al., 2011; Chung et al., 2014; McCarroll and Altshuler, 2007; Kakinuma and Sato, 2008; Cooper et al., 2011; Chung et al., 2014; McCarroll and Altshuler, 2007; Kakinuma and Sato, 2008). For neuropsychiatric disorders, such as intellectual ability, Autism Spectrum disorders, Schizophrenia, and Bipolar disorder, CNVs have given rise to a new understanding of disease etiology (Kakinuma and Sato, 2008; Malhotra and Sebat, 2012; Sullivan et al., 2012). Recently, multiple studies have highlighted the roles of CNVs in AD as well (Szigeti et al., 2013; Szigeti et al., 2014; Saykin et al., 2011; Heinzen et al., 2010; Lew et al., 2018; Zheng et al., 2015; Zhang, 2020; Heinzen et al., 2010; Saykin et al., 2011; Szigeti et al., 2013; Szigeti et al., 2014; Zheng et al., 2015; Lew et al., 2018; Zhang, 2020). For example, an intragenic CNV is found in the *CR1* gene (Brouwers et al., 2012), and people with Down syndrome have a higher chance to develop neuropathology, consistent with the observation that AD may be caused by duplications in the *APP* gene in chromosome 21 (Goate, 2006; Lanoiselée et al., 2017). However, there is no comprehensive genome-wide CNV study using whole-genome sequence (WGS) to enhance the knowledge of AD etiology and risk.

Most of the previous CNV GWAS of AD were performed using genotyping array data. Although these arrays can quickly and cost efficiently genotype large numbers of samples, there are serious technological limitations in that only large CNVs spanning multiple pre-determined probes can be reliably detected. However, WGS data allows an unbiased investigation of CNVs of all types (i.e., small and large; common and rare; within coding and non-coding regions) and provides a unique opportunity to comprehensively study CNVs in diseases. To accelerate AD genetic discovery, the Alzheimer's Disease Sequencing Project (ADSP) (Beecham et al., 2017), a strategic program funded by the National Institute on Aging (NIA), is committed to sequence AD cases, and cognitively normal elder controls from multi-ethnic populations, providing a valuable resource for genome-wide identification of CNVs.

This study utilizes the ADSP Umbrella R1 dataset (ng00067) released through the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) Data Sharing Service (Kuzma et al., 2016). After quality and relatedness checks, we had 1,737 AD cases and 2,063 cognitively normal elder controls for this study. We employed three CNV calling algorithms, CNVnator (Abyzov et al., 2011), JAX-CNV (Lee et al., 2021), and Smoove (GitHub—brentp/smoove, 2021; Layer et al., 2014) that on average detected 2,378, 25, and 4,584 CNVs, respectively, for each sample. GraphTyper2 (Eggertsson et al., 2019) was then applied for joint genotyping to generate a single VCF for all 3,800 samples in the study, which increased the number of CNVs to 280,073 average/sample; however, most of those CNVs either overlap or are adjacent to each other. After merging CNVs of the same type (deletions or duplications) and removing conflict regions with different types of CNVs, there are on average 4,075 CNVs per sample. The CNVs we identified tended to be more abundant and longer in AD cases compared to cognitively normal, elder controls, though in most cases this trend was not statistically significant.

MATERIALS AND METHODS

The analysis flow consists of two major steps; identification of CNVs from WGS from 3,800 subjects (*CNV Identification on WGS Data*), and CNV burden analysis (*CNV Burden Analysis Using PLINK*).

Figure 1 shows an overview of the flow of CNV identification on WGS data. The flow starts with alignment CRAM files and ends at the single-sample CNV list generation. The process began with a quality check (WGS Across-Chromosome Coverage Check) followed by sample-level CNV calling and project-level CNV joint genotyping (Sample-Level CNV Calling and Project-Level CNV Joint Genotyping). Finally, to meet the data format requirements of CNV burden analysis, the genotyped VCF was further split as a list in BED format per sample for region consolidation (for same-type CNVs overlapping) and removal (for different-type CNVs overlapping). Then, all BED files were merged and converted in PLINK format as the input of burden analysis (CNV List Assembling for PLINK Burden Analysis). The detailed scripts are given in supplementary material.

CNV Identification on WGS Data

WGS Across-Chromosome Coverage Check

The quality of CNV calling on WGS data is sensitive to alignment coverages across all chromosomes of a sample. Uneven coverages of chromosomes may cause false positive CNVs. Thus, before calling CNVs, it is necessary to perform a quality check of alignment coverages. Samples with uneven coverage were removed from analysis.

We developed a method (implemented as part of JAX-CNV) to first estimate the coverage of each chromosome of a sample. The method seeks 20 repetitive-free regions in each chromosome, and then calculates an average coverage of these regions to present the coverage of the chromosome. A repetitive-free region is defined as a 20k bp long region with each 25-mer (k-mer) inside the region having a unique position in the entire reference genome.

Once coverage of each chromosome was obtained, we were able to identify outlier chromosomes with unexpected high or low coverages. For example, outliers could indicate trisomy, monosomy, and other gross chromosome number anomalies. An overall average coverage of a sample was then computed by using the coverages of all chromosomes excluding outliers. A standard deviation of chromosomes coverages was employed as the metric to identify problematic samples that were removed

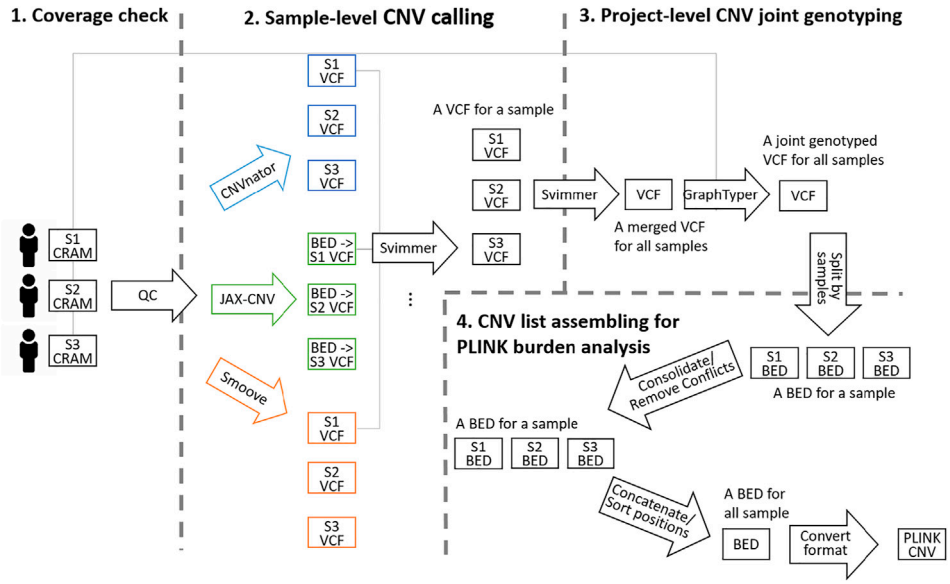

FIGURE 1 | Overview of the CNV identification workflow from WGS data consisting of the four steps. **1)** Alignment coverage check. **2)** Sample-level CNV calling including calling by CNVnator, JAX-CNV and Smoove, and merging the three callsets by Svimmer. Since Svimmer takes the VCF format as input, results of JAX-CNV in the BED format were converted to the VCF format. **3)** Project-level CNV re-genotyping. **4)** CNV list assembling for PLINK burden analysis. The illustrated three samples in the figure are notated by S1, S2 and S3 while 3,800 samples were processed in the study.

from downstream analyses. This method is fast and takes approximately 5 minutes for a 30X sequence sample.

Sample-Level CNV Calling

We employed CNVnator, JAX-CNV, and Smoove for CNV detection. CNVnator and JAX-CNV are Read-Depth-based (RD-based) algorithms while Smoove utilizes multiple signals of RD, Paired-End (PE), and Split-Read (SR). CNVnator is sensitive for CNVs sizes ranging from 1 to 50 kb; however, it may break larger CNVs into smaller pieces that introduce difficulties for downstream analyses. We included JAX-CNV in the analysis flow because it was developed to detect large (>50 kb) CNVs and resolves the issue of fine pieces from CNVnator. Smoove was recruited to strengthen small CNV (<1 kbp) identification. These three CNV calling algorithms are not only fast but also generating high-quality CNVs. Moreover, the combination of them allows us to cover the complete size spectrum of CNVs.

For each sample, we applied these three algorithms separately. Each algorithm could generate a BED (JAX-CNV) or VCF (CNVnator and Smoove) file to store a set of deletions/duplications with genomics coordinates and genotypes (homozygous or heterozygous, and copy numbers) of a sample. If a BED file was generated, we converted it to VCF format to facilitate the step of utilizing svimmer (GitHub—DecodeGenetics/svimmer, 2021) for callset merging. For variant types (deletions, duplications, inversions, and breakends) detected by Smoove, we only kept deletions and duplications. For each sample, we then applied svimmer to merge the three VCFs obtained from the three algorithms.

Project-Level CNV Joint Genotyping

Joint analysis is recommended for a dataset with multiple samples. Once variants of a sample were detected, a joint analysis step provides the ability to leverage population-wide information from multiple samples that allows us to refine low-quality genotypes and detect additional variants of a sample. For example, a joint genotyping step is suggested in the GATK best practice for SNV and INDEL detection.

Compared to SNV/INDEL joint genotyping, CNV joint genotyping is challenging since breakpoints of CNVs from short-read sequence data may be imprecise. By incorporating detected variants within the linear reference genome, the emerging methodology, Graph Genome, provides a good model for joint genotyping CNVs across multiple samples in a single step. We evaluated GraphTyper2 (Eggertsson et al., 2019), Paragraph (Chen et al., 2019), and VG (Hickey et al., 2020), and selected GraphTyper2 in the analysis flow due to its balance of required computational resource and quality of results.

As GraphTyper2 recommended, we employed svimmer (GitHub—DecodeGenetics/svimmer, 2021) to merge all sample-level VCFs and generate a single VCF that does not contain genotypes. GraphTyper2 was then applied on this merged VCF with all CRAM files for each 500kb region excluding the centromeres. GraphTyper2 generated a VCF of CNVs with genotypes of all samples. There are three models used for joint genotyping in GraphTyper2, Aggregated, Coverage, and Breakpoint, and we kept results from Aggregated model as GraphTyper2 suggests. We also applied PASS flag filter in the GraphTyper2 VCFs. Each 500kb chunk VCFs were consented using BCFtools (Danecek et al., 2021).

CNV Burden Analysis Using PLINK

CNV List Assembling for PLINK Burden Analysis

There remains a challenge in using GraphTyper2 VCF files for downstream burden analysis. Since multiple calling algorithms were applied for CNV identification, CNV lengths and breakpoints may vary. Although GraphTyper2 was applied to mitigate this situation, we still can find CNV segments overlapping each other that is not acceptable by downstream association analysis tools such as PLINK (Chang et al., 2015). To resolve overlapping segments, we first split CNVs (with PASS genotype tags) of a sample in BED format for each sample. The BED is in the format of chromosome, begin position, end position, and copy number status for each CNV. The copy number status recorded as 0, 1, 3 or 4 copies. Of note, the copy status 4 includes copy numbers equal or larger than 4. Then, we used BEDTools (Quinlan and Hall, 2010) to merge overlapping or adjacent segments. Segments were merged only if they are the same CNV type, deletions or duplications. For those regions having different CNV types, we filtered them out since the downstream association analysis would not take those regions into consideration. Once the CNV consolidation and removal were done for all samples, we then concatenated all BED files and sorted the merged BED file by CNV positions.

PLINK format, that is commonly accepted by other downstream association tools, is a tabular file format with CNV coordinates, family IDs, and sample IDs. Since there are no related samples in the dataset, we replicated sample IDs as family IDs. We then converted the BED file into a six-column with family ID, sample ID, chromosome, start position, end position, and copy number status, e.g. 0, 1, 3, or 4 copies.

Rare CNV Identification

Rare CNVs were obtained using PLINK to impose a 0.01 frequency threshold (i.e., --cnv-freq-exclude-above 38 and--cnv-overlap 0.5), which removed CNVs with >50% of its length spanning a region with >1% × 3,800 CNVs in the dataset. The same approach was applied on African American (AA) (--cnv-freq-exclude-above 9), Hispanic (--cnv-freq-exclude-above 12), and Non-Hispanic White (NHW) (--cnv-freq-exclude-above 15) samples. Then, we applied the pilot mask released by the 1,000 Genomes Project (The 1000 Genomes Project, 2010) on rare CNV lists. The pilot mask was done by looking at the amount of sequence data that aligned to any given location in the reference genome. Regions are defined inaccessible if their depth of coverages (summed across all samples in the 1,000 Genomes Project) were higher or lower than the average depth. The mask results in 5.3% of bases marked "N" (the base is an "N"), 1.4% marked "L" (coverage is low), 0.6% marked "H" (coverage is high) and 3.7% marked "Z" (many reads mapped here have zero quality). The remaining 89.0% of are marked "P" (regions are good and passed). All rare CNVs need to reside in "P" regions.

CNV Burden Analysis

We examined the burdens of all and rare CNVs in AD cases and controls using PLINK. PLINK burden analysis uses permutation tests to compute *p*-values. For our analysis, we applied 500,000 permutations. For each sample, we considered four CNV burden features: 1) number of CNV events; 2) proportion of samples with ≥1 CNV events; 3) total event length in kb; and 4) average event length in kb. The CNV events included deletions and duplications together (DelDup), deletions specific (Del), and duplications specific (Dup). We reported the CNV burdens for AA, Hispanic, and NHW separately as well as for all-combined samples (ALL), The Bonferroni threshold for multiple testing is p-value < 0.05/96 analyses = 0.000521, where the 96 analyses included the combinations from 2 sets of CNV analyses (all CNVs vs. rare CNVs), 4 burden features, 3 CNV events (DelDup, Del, and Dup) and 4 sample groups (ALL, AA, Hispanic, and NHW).

RESULTS

Dataset—3,800 WGS Samples from NIAGADS R1 Release of ADSP 5k

We used the ADSP WGS data released by NIAGADS in 2018. NIAGADS not only collected and released genetics data, but also harmonized minimal phenotypes (sex, race/ethnicity, diagnosis, *APOE* genotype) from each collocating cohort. For data harmonization, NIAGADS followed the ADSP coding scheme based on the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) (Beekly et al., 2007) definitions. We used NIAGADS and did not redefine diagnosis or ethnicities in this study.

There are 4,749 subjects and 4,788 sequenced samples (three subjects sequenced nine times and another three sequenced six times) by Illumina HiSeq 2000/2,500 or X Ten at an average of 37X coverage (the range from 10.68X to 74.16X). For the six subjects with multiple sequence sets, we picked one sequence set per subject, and removed the other 39 sequences. For the 4,749 subjects, these were 2,192 AD cases, 2,073 controls, and others 484 with diagnosis unknowns. For this study, we focused on AD cases and controls, and excluded samples with inconclusive clinical statuses.

For the remaining 4,265 samples, we performed the across-chromosome alignment coverage check (*WGS Across-Chromosome Coverage Check*) since uneven coverage may affect the quality of CNV detection. Fifteen samples were removed since their standard deviation of chromosomes coverages are greater than 15% of the average coverages, as shown in **Figure 2** where each line presents a sample, and each dot presents the alignment coverage of the sample in the chromosome on the *x*-axis.

Next, we removed 450 samples due to relatedness according to pedigree information provided by NIAGADS. Finally, we had 1,737 AD cases and 2,063 controls. The ethnicities/races are AA (n = 978), Hispanic (n = 1,247), NHW (n = 1,566), and others (n = 9), as shown in **Table 1**.

CNV Callset

We first applied CNVnator, JAX-CNV and Smoove on each CRAM file of a sample for sample-level CNV calling. CNVnator, JAX-CNV and Smoove detected an average of 2,378 (1,967 deletions and 411 duplications), 25 (12 deletions and 13

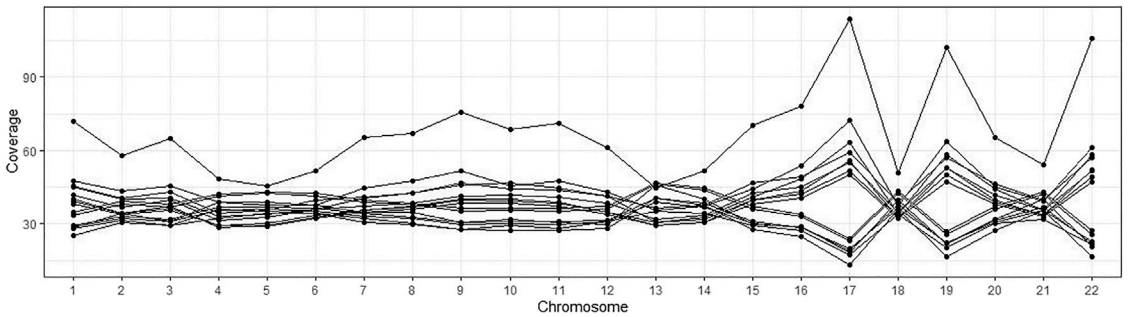

FIGURE 2 | Alignment coverages of 15 samples with uneven sequence data. Each line is a sample, and each dot presents the alignment coverage for a chromosome.

TABLE 1 | Total column denotes the number of samples remaining after each quality filtering step.

	AA			Hispanic			NHW			Others			Total
Step	**Case**	**Control**	**Unknown**	**Case**	**Control**	**Unknown**	**Case**	**Control**	**Unknown**	**Case**	**Control**	**Unknown**	
ADSP 5K	472	521	44	826	746	40	910	820	393	5	4	7	4,788
Replicate Removal	467	521	44	810	733	40	910	815	393	5	4	7	4,749
Unknown Status Removal	467	521	0	810	733	0	910	815	0	5	4	0	4,265
Uneven Coverage Removal	466	521	0	808	731	0	902	813	0	5	4	0	4,250
Relatedness Removal	457	521	0	520	727	0	755	811	0	5	4	0	**3,800**

3,800 samples remained after all filtering steps.

duplications), and 4,584 (3,876 deletions and 708 duplications) CNVs, respectively. Compared to NHW, AA and Hispanic have 141 and 122 deletions more, but 180 and 9 fewer duplications. Only Smoove yielded fewer duplications for AA and Hispanic, as shown in **Figure 3A**.

For each sample, we employed svimmer to merge the callsets from the three callers as a single VCF. Next, svimmer was applied to VCFs for all 3,800 samples to generate a combined VCF which along with all CRAM files are inputs of GraphTyper2. As described in *Project-Level CNV Joint Genotyping*, we kept Aggregated notated variants and also applied the PASS flag filter in this aggregated callset. The result was a total of 237,306 deletions and 42,767 duplications as a project-level VCF. The length distribution and allele frequency of the project-level VCF are given in **Figures 3B,C**. Lengths of deletions were presented by using negative values that were shown on the left panel of **Figure 3B**, while lengths of duplication were shown on the right panel of **Figure 3B**.

CNV Concordant Check with Other Projects

We compared our project-level callset with the 1,000 Genomes Project Phase 3 (1KG_P3) (Sudmant et al., 2015), gnomAD (Collins et al., 2020), and Decipher (Firth et al., 2009) that were obtained from dbGaP (https://www.ncbi.nlm.nih.gov/dbvar/content/human_hub/). The 1KG_P3 and gnomAD have other types of variants (insertions, inversions, mobile element deletion, and mobile element insertions) in the lists that were not used in the comparison; only autosomal copy number variations were used for the comparison. All lists were converted into the BED format for performing cross-project concordant CNV checks by using BEDTools.

We examined the overlap between our data and other call sets using either a 1bp or 50% overlap. We performed each pair of comparisons twice treating both callsets as the primary in one of the comparisons. As demonstrated in **Table 2**, each pair of comparisons is asymmetric with different concordance percentages depending upon which callset was the primary (primary callset is the one in the column). 79.9 and 76.3% of our called CNVs were found in gnomAD and Decipher when using at least 1bp overlapping criterion. However, only 39.8% were recalled in the 1KG_P3 callset. GnomAD likewise has a low concordance rate, with only 41%, of CNVs overlapping with the 1KG_P3 callset. Our callset and gnomAD callset have higher similarity and more novel CNVs compared to the 1KG_P3 and Decipher callsets.

CNV List for PLINK Burden Analysis

Since PLINK does not allow overlapping CNVs within a sample, we 1) split the project-level VCF and generated a list of CNVs for a sample in BED format, and 2) consolidated CNVs or removed conflict CNVs by the method described in Section 2.1.4. After splitting the project-level VCF for each sample, we found increased numbers of CNVs per sample (32,402 deletions and 9,131duplications) since GraphTyper2 uses a combination of the three CNV calling algorithms and leverages variant knowledge from other samples. However, most of those CNVs overlap or are adjacent to each other. Next, we consolidated

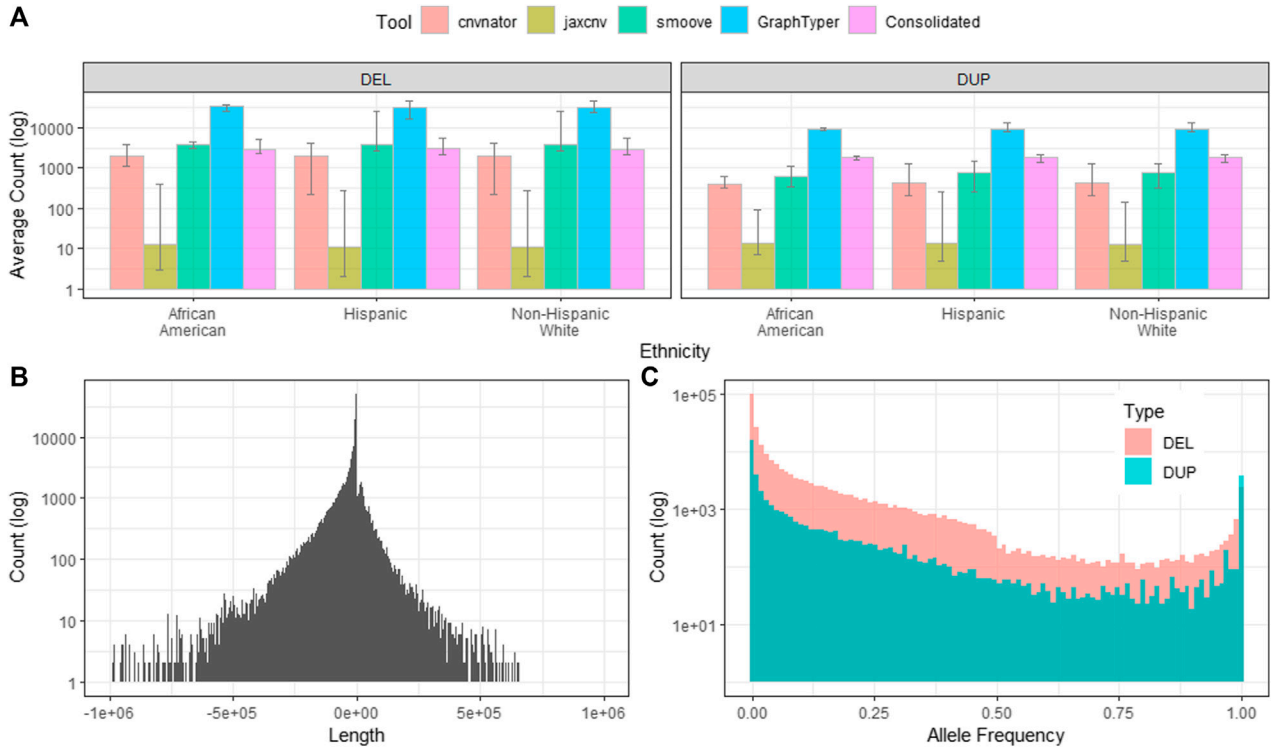

FIGURE 3 | Characteristics of the project level CNV callset. Counts shown on the *y*-axes of the sub figures are in the log10 scale. **(A).** The average deletions and duplications detected by CNVnator, JAX-CNV, Smoove and GraphTyper2. Consolidated shows CNV counts after CNV merging and conflicts removing. **(B).** Length distribution of CNVs after applying GraphTyper2 and the PASS flag filter. Lengths of deletions were presented as negative values while lengths of duplications are positive values. **(C).** Allele frequency of CNVs of GraphTyper2.

TABLE 2 | CNV concordant checks with the 1,000 Genomes Project Phase 3 (1KG_P3), gnomAD, and Decipher callsets. Each column resents the percentages of CNVs in the callset overlapping with others listed in rows.

	At least 1bp overlap					At least 50% overlap			
	Ours (280,073)	1KG_P3 (48,131)	gnomAD (188,842)	Decipher (54,422)		Ours (280,073)	1KG_P3 (48,131)	gnomAD (188,842)	Decipher (54,422)
Ours	1	0.828	0.762	0.878	Ours	1	0.772	0.726	0.816
1KG_P3	0.398	1	0.410	0.679	1KG_P3	0.293	1	0.337	0.544
gnomAD	0.799	0.861	1	0.832	gnomAD	0.668	0.767	1	0.712
Decipher	0.763	0.662	0.500	1	DECIPHER	0.724	0.600	0.458	1

overlapping/adjacent CNVs if they are the same type or removed overlapping CNVs if they are different types. This CNV consolidation step significantly reduces CNVs/sample (2,966 deletions and 1,863 duplications), as shown in **Figure 3A**.

For rare CNV analysis, we first applied the pilot mask from the 1,000 Genomes Project that further filtered about 8.4% of CNVs and became 2,255 deletions and 1,820 duplications for each sample averagely. CNVs with an allele frequency <1% were retained for analysis. The number of rare CNVs/sample ranged from 0 to 1,546 with an average of 57/sample (46 deletions and 11 duplications; median value is 44 and standard deviation is 76.58843). Among 3,800 samples, three have zero rare CNVs while four have >1,000 rare CNVs. Those four samples are all Non-Hispanic Whites (two cases and two controls), and three of the four samples. According to the final review comment have higher detected numbers of CNVs (According to the final review comment 5,809, 5,945, and 5,992) compared to average (4075.06). The three were sequenced in the earlier stage of the project by Illumina HiSeq 2000/2,500 with PCR Amplified libraries.

Burdens of All and Rare CNVs

Table 3 are the PLINK burden tests. The four burden features were considered; 1) total event numbers, 2) Proportion of samples with ≥1 events, 3) total event

TABLE 3 | The four burden features were considered; 1) total event numbers, 2) Proportion of samples with ≥1 events, 3) total event length in kb, and 4) average event length in kb.

Mean_Case Mean_Control *p*-value		DelDup		Del		Dup	
		All	**Rare**	**All**	**Rare**	**All**	**Rare**
Total event numbers	All	4,073	59.29	2,249	47.67	1823	11.62
		4,079	55.61	2,261	44.41	1818	11.2
		0.736247	0.0709259	0.876096	0.0723559	0.021826	0.132316
	AA	4,072	60.24	2,268	46.81	1803	13.43
		4,106	63.23	2,295	49.7	1811	13.53
		0.990162	0.743957	0.989694	0.753128	0.882104	0.578805
	Hispanic	4,193	42.63	2,408	33.98	1785	8.654
		4,177	59.35	2,384	48.04	1793	11.31
		0.108108	1	0.016028	1	0.974318	1
	NHW	3,991	45.33	2,129	34.1	1861	11.23
		3,972	38.81	2,127	29.01	1845	9.8
		0.158684	0.0287979	0.461645	0.0303239	**2e-06***	0.0354999
Proportion of samples with ≥1 events	All	0.9988	0.9988	0.9988	0.9988	0.9988	0.9988
		0.9995	0.9995	0.9995	0.9995	0.9995	0.9985
		0.904246	0.905054	0.905188	0.905048	0.904122	0.581927
	AA	0.9956	0.9956	0.9956	0.9956	0.9956	0.9956
		1	1	1	1	1	1
		1	1	1	1	1	1
	Hispanic	1	1	1	1	1	0.9981
		0.9986	0.9986	0.9986	0.9986	0.9986	0.9986
		0.583197	0.583439	0.582637	0.582109	0.583935	0.826018
	NHW	1	1	1	1	1	1
		1	1	1	1	1	0.9975
		1	1	1	1	1	0.269673
Total event length in kb	All	1.856e+05	1,053	2.983e+04	546.2	1.558e+05	507.2
		1.852e+05	941.4	2.974e+04	457.1	1.555e+05	484.8
		0.017098	0.01129	0.254809	0.00759198	0.0602339	0.148482
	AA	1.859e+05	1,013	3.185e+04	502.7	1.54e+05	510.8
		1.857e+05	1,055	3.175e+04	477.6	1.54e+05	577.4
		0.318897	0.704127	0.330257	0.291605	0.409045	0.942028
	Hispanic	1.83e+05	750.3	3.183e+04	408.3	1.511e+05	342.7
		1.837e+05	911.8	3.097e+04	392.5	1.527e+05	519.4
		0.982962	0.989972	**6.79999e-05***	0.356709	1	0.999998
	NHW	1.873e+05	943.1	2.725e+04	455.1	1.601e+05	487.9
		1.863e+05	713.7	2.734e+04	301.8	1.59e+05	412.9
		0.000591999	0.00145	0.670983	0.00347599	0.000116	0.013062
Average event length in kb	All	45.74	19.02	13.34	12.54	85.34	40.48
		45.57	17.96	13.24	11.43	85.47	40.56
		0.0489579	0.0469619	0.0544019	0.0573059	0.995478	0.523385
	AA	45.5	16.3	14.03	10.58	85.04	36.11
		45.29	16.09	13.89	9.459	85.01	38.94
		0.0487079	0.384237	0.108808	0.0501099	0.362197	0.935694
	Hispanic	43.67	16.67	13.26	11.59	84.68	35.79
		44	15.33	13.02	9.041	85.03	39.45
		0.9966	0.0739319	0.00848998	0.00603599	0.999998	0.966586
	NHW	47.31	20.6	12.98	12.82	85.98	41.66
		47.16	18.64	13.02	10.99	86.16	39.69
		0.22001	0.0258339	0.645417	0.0592879	0.98735	0.161868

Tests were done for all and rare CNVs as well as considering deletions and duplications (DelDup), deletions specific (Del) and duplications specific (Dup). Each cell has three values as mean of cases, mean of controls, and p-*value. Two* p-*values marked in bold indicate statistically significant.*

length in kb, and 4) average event length in kb. Tests were done for all and rare CNVs as well as considering deletions and duplications (DelDup), deletions specific (Del) and duplications specific (Dup). The results suggested two significant all-CNV burden differences between cases and controls: 1) in NHW, on average cases have 16 more duplication events compared to controls do (*p*-value 2e-6); and 2) in Hispanic, the total deletion lengths in cases is larger than in controls on average (*p*-value 6.8e-5). There are no significant differences for rare CNV burden in all aspects examined. Of note, the *p*-values from PLINK burden analysis did not account for covariates and were merely examining if the observed burden measures of cases and controls were significantly different in a marginal fashion. **Figure 4** shows the total event numbers per sample and the total event length in kb per sample.

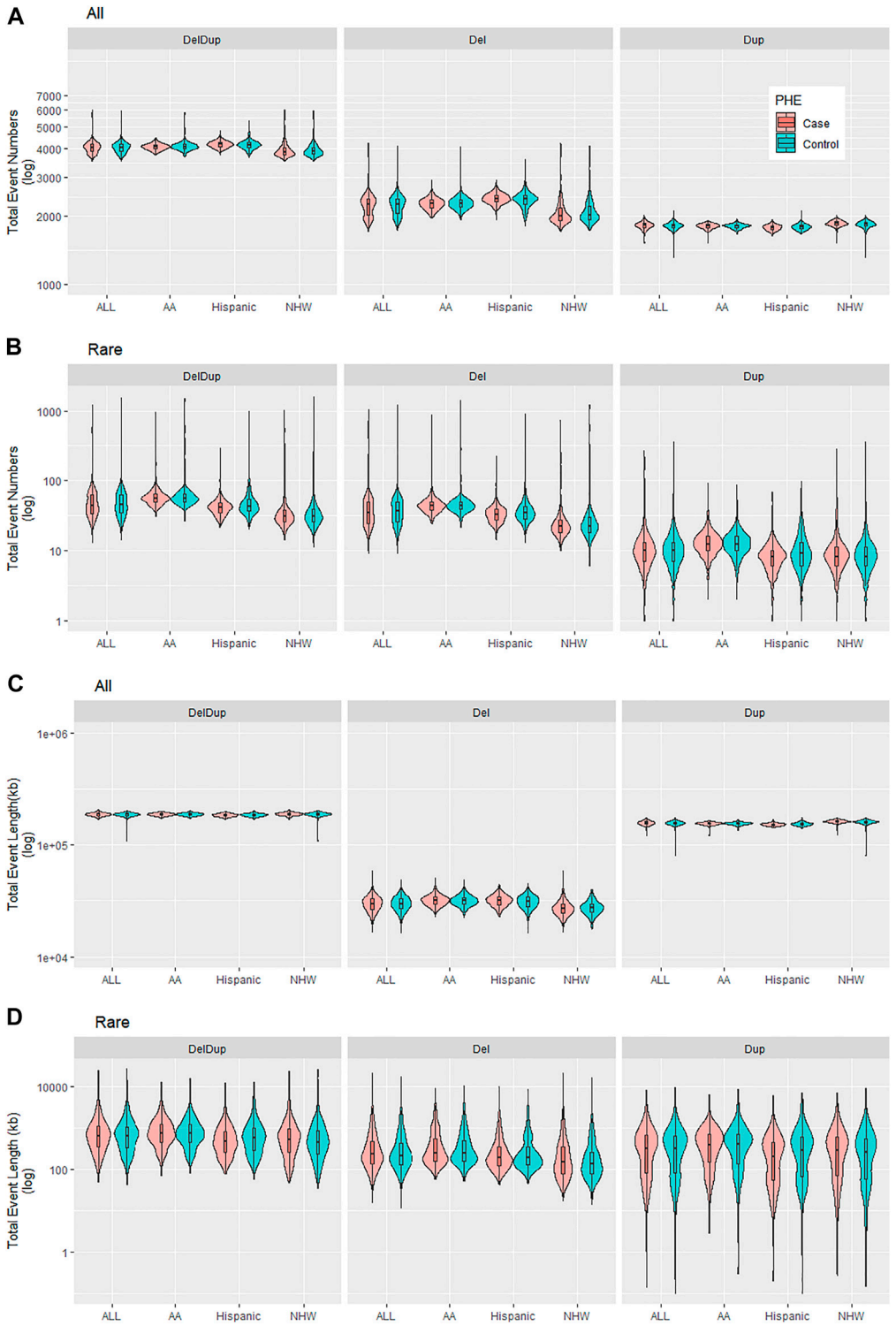

FIGURE 4 | Summary of CNV burden results for all and rare CNVs by CNV events (DelDup, Del, or Dup) and by ethnicities (ALL, AA, Hispanic, NHW). **(A).** Total event numbers per sample. **(B).** Total rare event numbers per sample. **(C).** Total event length in kb per sample. **(D).** Total rare event length in kb per sample.

DISCUSSION

We have composed a scalable bioinformatics pipeline to identify CNVs using WGS data and applied it to 1,737 AD cases and 2,063 cognitively normal controls from the ADSP. We observed 237,306 and 42,767 deletions and duplications, respectively with an average of 2,255 deletions and 1,820 duplications per subject. Although there were more and longer CNVs in AD case samples than controls, burden tests performed using all CNVs or rare CNVs (i.e., <1% in frequency) do not indicate a significant association with AD status.

The false discovery rate of detected CNVs remains uncertain despite the fact that CNVs were generated circumspectly and have been cross checked with other projects including the 1KG, gnomAD and Decipher. The callset of 1KG is smaller than ours and gnomAD's, and it is therefore expected that 1KG recalls only ~40% of ours and gnomAD's callsets, while ours and gnomAD's callsets capture 82.8 and 86.1% of 1KG's CNVs respectively. We would also like to note that 1KG processed their data several years earlier than we and gnomAD did. Since the publishing of the 1KG Phase3 callset, CNV-calling tools have moved towards integration of multiple alignment signals (such as read-depth, pair-end, and split-read signals) for calling. This concept was well-accepted before the publishing of the gnomAD callset, and could make 1KG's callset less similar to ours and gnomAD's. While extensive experimental validation of each CNV is not currently feasible, validation of significant deletions and duplications may be necessary. Alternatively, our findings could be replicated with other datasets of Alzheimer's Disease whole genome sequence data.

Joint genotyping provides the ability to leverage information from multiple samples so we could refine low-quality genotypes and detect additional variants for a sample. However, it also brings challenges when breakpoints of CNVs from different samples do not align well. The situation is even worse when using multiple calling algorithms. For this study, we employed GraphTyper2 for joint genotyping, which is a graph-genome based method and has shown an advantage for genotyping larger variants such as CNVs. However, GraphTyper2 does not provide a total solution; overlapping CNVs can still be found after joint genotyping. To address the issue, we split aggregated results to generate a CNV list for each sample and resolved overlapping CNV regions. A graph reference genome presents a variant, a CNV in our application, as a branch in the graph. For the overlapping CNV situation, the graph genome creates several similar branches in a region. The issues could be resolved in a more fundamental way by pruning unnecessary brunches of the graph genome. A slim graph genome will also improve running time and memory usage.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data is accessible from NIAGADS DSS *via* qualified access. Formal requests to access these datasets can be submitted to NIAGADS DSS: https://dss.niagads.org/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by North Carolina State University Institutional Review Board for the use of human subjects in research. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

W-PL: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing—Original Draft, Visualization, Supervision, Project administration; AT: Software, Formal analysis, Writing—Review and Editing; MC: Methodology, Software, Formal analysis, Writing—Review and Editing; YYL: Data Curation, Writing—Review and Editing, Resources; AK: Data Curation, Writing—Review and Editing, Resources; OV: Software, Data Curation; Y-FC: Software, Data Curation; WL: Conceptualization; L-SW: Conceptualization, Supervision, Writing—Review & Editing, Funding acquisition; GS: Conceptualization, Writing—Review and Editing, Funding acquisition; J-YT: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing—Original Draft, Supervision, Project administration.

FUNDING

This work has been partially supported by National Institutes of Health Grants RF1 AG074328, U54 AG052427, U24 AG041689, and P01 CA142538, and the University of Pennsylvania (Penn) Alzheimer's Disease Research Center (ADRC) Development Projects.

ACKNOWLEDGMENTS

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institutes of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Sequencing for the Follow Up Study (FUS) is supported through U01AG057659 (to Drs. PericakVance, Mayeux, and Vardarajan) and U01AG062943 (to Drs. Pericak-Vance and Mayeux). Data generation and harmonization in the Follow-up

Phase is supported by U54AG052427 (to Drs. Schellenberg and Wang). The FUS Phase analysis of sequence data is supported through U01AG058589 (to Drs. Destefano, Boerwinkle, De Jager, Fornage, Seshadri, and Wijsman), U01AG058654 (to Drs. Haines, Bush, Farrer, Martin, and Pericak-Vance), U01AG058635 (to Dr. Goate), RF1AG058066 (to Drs. Haines, Pericak-Vance, and Scott), RF1AG057519 (to Drs. Farrer and Jun), R01AG048927 (to Dr. Farrer), and RF1AG054074 (to Drs. Pericak-Vance and Beecham). The ADGC cohorts include: Adult Changes in Thought (ACT) (UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657), the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Cell Repository for Alzheimer's Disease (NCRAD) (U24 AG21886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA- LOAD) (R01 AG041797), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer's Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University HispanicEstudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904), the EU Joint Programme—Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/ grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/ Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the

responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The FUS cohorts include: the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), Alzheimer's Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer's Disease Initiative (CuAADI) (3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473, RF1AG015473, R56AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwangju Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hussman Institute for Human Genomics Brain Bank (HIHGBB) (R01AG027944, Alzheimer's Association "Identification of Rare Variants in Alzheimer Disease"), Ibadan Study of Aging (IBADAN) (5R01AG009956), Mexican Health and Aging Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, 2R01AG048927), Northern Manhattan Study (NOMAS) (R01NS29993), Peru Alzheimer's Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PR1066) (Wellcome Trust (GR066133/GR080002), European Research Council (340755)), Puerto Rican Alzheimer Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI) (U01AG052410), Rush Alzheimer's Disease Center (ROSMAP) (P30AG10161, R01AG15819, R01AG17919), University of Miami Brain Endowment Bank (MBB), and University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, R01AG028786). The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.752390/full#supplementary-material

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing. *Genome Res.* 21, 974–984. [Internet]Jun [cited 2021 Feb 15]Available from: https://pubmed.ncbi.nlm.nih.gov/21324876/. doi:10.1101/gr.114876.110

Alzheimer's Association (2020). Alzheimer's Disease Facts and Figures. *Alzheimer's Dement* 16, 391–460. [Internet]Mar 1 [cited 2021 Jul 12] Available from: https://pubmed.ncbi.nlm.nih.gov/32157811/.

Beecham, G. W., Bis, J. C., Martin, E. R., Choi, S-H., DeStefano, A. L., Duijn, C. M. van., et al. (2017). The Alzheimer's Disease Sequencing Project: Study Design and Sample Selection. *Neurol. Genet.* 3. , 2017 [Internet]Oct 1 [cited 2021 Jul 29]Available from: /pmc/articles/PMC5646177/. doi:10.1212/NXG.0000000000000194

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., et al. (2007). The National Alzheimer's Coordinating Center (NACC) Database: The Uniform Data Set. *Alzheimer Dis. Assoc. Disord.* 21, 249–258. [cited 2021 Apr 6] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5646177/. doi:10.1097/wad.0b013e318142774e

Brouwers, N., Van Cauwenberghe, C., Engelborghs, S., Lambert, J. C., Bettens, K., Le Bastard, N., et al. (2012). Alzheimer Risk Associated with a Copy Number Variation in the Complement Receptor 1 Increasing C3b/C4b Binding Sites. *Mol. Psychiatry* 17, 223–233. [Internet]Feb 15 [cited 2021 Jul 12]Available from: www.nature.com/mp. doi:10.1038/mp.2011.24

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *Gigascience* 4, 7, 2015 . [Internet]Dec 1 [cited 2021 Jul 27]Available from: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533. doi:10.1186/s13742-015-0047-8

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., et al. (2019). Paragraph: A Graph-Based Structural Variant Genotyper for Short-Read Sequence Data. *Genome Biol.* 20 (1), 1–13. [Internet]Dec 19 [cited 2021 Jun 23]. doi:10.1186/s13059-019-1909-7

Chung, B. H-Y., Tao, V. Q., and Tso, W. W-Y. (2014). Copy Number Variation and Autism: New Insights and Clinical Implications. *J. Formos. Med. Assoc.* 113, 400–408. [Internet]Jul [cited 2017 Jun 23]Available from: http://linkinghub.elsevier.com/retrieve/pii/S0929664613000570. doi:10.1016/j.jfma.2013.01.005

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., et al. (2020). A Structural Variation Reference for Medical and Population Genetics. *Nature* 581, 444–451. [Internet]May 28 [cited 2021 Apr 8]. doi:10.1038/s41586-020-2287-8

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., et al. (2011). A Copy Number Variation Morbidity Map of Developmental Delay. *Nat. Genet.* 43, 838–846. [Internet]Aug 14 [cited 2017 Jun 30]Available from: http://www.ncbi.nlm.nih.gov/pubmed/21841781. doi:10.1038/ng.909

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *Gigascience* 10, 1–4. [Internet]Jan 29 [cited 2021 Jul 20]Available from: https://academic.oup.com/gigascience/article/10/2/giab008/6137722. doi:10.1093/gigascience/giab008

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., et al. (2019). GraphTyper2 Enables Population-Scale Genotyping of Structural Variation Using Pangenome Graphs. *Nat. Commun.* 10. , 2019 [Internet]Dec 1 [cited 2021 Jun 23]Available from: /pmc/articles/PMC6881350/. doi:10.1038/s41467-019-13341-9

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., et al. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* 84 (4), 524–533. [Internet]Apr 10 [cited 2021 Jul 7]. Available from: /pmc/articles/PMC2667985/. doi:10.1016/j.ajhg.2009.03.010

GitHub - brentp/smoove (2021). Structural Variant Calling and Genotyping with Existing Tools, but, Smoothly. [Internet][cited 2021 Jun 24]. Available from: https://github.com/brentp/smoove.

GitHub - DecodeGenetics/svimmer (2021). Structural Variant Merging Tool. [Internet] [cited 2021 Jun 24]. Available from: https://github.com/DecodeGenetics/svimmer.

Goate, A. (2006). Segregation of a Missense Mutation in the Amyloid β-protein Precursor Gene with Familial Alzheimer's Disease [Internet]. *J. Alzheimer's Dis. IOS Press* Vol. 9, 341–347. [cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/16914872/. doi:10.3233/jad-2006-9s338

Heinzen, E. L., Need, A. C., Hayden, K. M., Chiba-Falek, O., Roses, A. D., Strittmatter, W. J., et al. (2010). Genome-wide Scan of Copy Number Variation in Late-Onset Alzheimer's Disease. *J. Alzheimer's Dis.* 19, 69–77. [Internet][cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/20061627/. doi:10.3233/jad-2010-1212

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., et al. (2020). Genotyping Structural Variants in Pangenome Graphs Using the Vg Toolkit. *Genome Biol.* 21, 1–17. [Internet]Feb 12 [cited 2021 Jun 23]. doi:10.1186/s13059-020-1941-7

Kakinuma, H., and Sato, H. (2008). Copy-number Variations Associated with Autism Spectrum Disorder. *Pharmacogenomics* 9, 1143–1154. [Internet]Aug [cited 2017 Jun 23]Available from: http://www.ncbi.nlm.nih.gov/pubmed/18681787. doi:10.2217/14622416.9.8.1143

Kuzma, A., Valladares, O., Cweibel, R., Greenfest-Allen, E., Childress, D. M., Malamon, J., et al. (2016). NIAGADS: The NIA Genetics of Alzheimer's Disease Data Storage Site. *Alzheimer's Dement.* 12 (11), 1200–1203. doi:10.1016/j.jalz.2016.08.018

Lanoiselée, H. M., Nicolas, G., Wallon, D., Rovelet-Lecrux, A., Lacour, M., Rousseau, S., et al. (2017). APP, PSEN1, and PSEN2 Mutations in Early-Onset Alzheimer Disease: A Genetic Screening Study of Familial and Sporadic Cases. *Plos Med.* 14. , 2017 [Internet]Mar 1 [cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/28350801/. doi:10.1371/journal.pmed.1002270

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: A Probabilistic Framework for Structural Variant Discovery. *Genome Biol.* 15, R84, 2014 . [Internet]Jun 26 [cited 2021 Feb 15]Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-6-r84. doi:10.1186/gb-2014-15-6-r84

Lee, W-P., Zhu, Q., Yang, X., Liu, S., Cerveira, E., Ryan, M., et al. (2021). JAX-CNV: A Whole Genome Sequencing-Based Algorithm for Copy Number Detection at Clinical Grade Level. medRxiv [Internet]. Mar 17 [cited 2021 Jun 23]. doi:10.1101/2021.03.16.21252173

Lew, A. R., Kellermayer, T. R., Sule, B. P., and Szigeti, K. (2018). Copy Number Variations in Adult-Onset Neuropsychiatric Diseases. *Curr. Genomics* 19, 420–430. [Internet]Mar 30 [cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/30258274/. doi:10.2174/1389202919666180330153842

Malhotra, D., and Sebat, J. (2012). CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell. Elsevier B.V.* 148, 1223–1241. doi:10.1016/j.cell.2012.02.039

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number Variation and Association Studies of Human Disease. *Nat. Genet.* 39, S37–S42. [Internet] Jul [cited 2018 Aug 23]Available from: http://www.ncbi.nlm.nih.gov/pubmed/17597780. doi:10.1038/ng2080

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* 26, 841–842. [Internet]Jan 28 [cited 2021 Jul 1]Available from: https://pubmed.ncbi.nlm.nih.gov/20110278/. doi:10.1093/bioinformatics/btq033

Ridge, P. G., Mukherjee, S., Crane, P. K., Kauwe, J. S. K., and Consortium, A. D. G. (2013). Alzheimer's Disease: Analyzing the Missing Heritability. *PLoS One* 8, e79771, 2013 . [Internet]Nov 7 [cited 2021 Sep 19]Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079771. doi:10.1371/journal.pone.0079771

Saykin, A. J., Swaminathan, S., Kim, S., Shen, L., Risacher, S. L., Foroud, T., et al. (2011). Genomic Copy Number Analysis in Alzheimer's Disease and Mild Cognitive Impairment: An ADNI Study. *Int. J. Alzheimers Dis.*, 2011, 2011 . [Internet][cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/21660214/. doi:10.4061/2011/729478

Sims, R., Hill, M., and Williams, J. (20202020). The Multiplex Model of the Genetics of Alzheimer's Disease. *Nat. Neurosci.* 23323, 311–322. [Internet]Feb 28 [cited 2021 Sep 19]. Available from: https://www.nature.com/articles/s41593-020-0599-5. doi:10.1038/s41593-020-0599-5

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* 526, 75–81. [Internet]Sep 30 [cited 2017 Jun 22]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26432246. doi:10.1038/nature15394

Sullivan, P. F., Daly, M. J., and O'Donovan, M. (2012). Genetic Architectures of Psychiatric Disorders: The Emerging Picture and its Implications. *Nat. Rev. Genet.* 13, 537–551. Nature Publishing Group [cited 2021 Jul 12]Available from: www.nature.com/reviews/genetics. doi:10.1038/nrg3240

Szigeti, K., Kellermayer, B., Lentini, J. M., Trummer, B., Lal, D., Doody, R. S., et al. (2014). Ordered Subset Analysis of Copy Number Variation Association with Age at Onset of Alzheimer's Disease. *J. Alzheimer's Dis.* 41, 1063–1071. [Internet][cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/24787912/. doi:10.3233/jad-132693

Szigeti, K., Lal, D., Li, Y., Doody, R. S., Wilhelmsen, K., Yan, L., et al. (2013). Genome-wide Scan for Copy Number Variation Association with Age at Onset of Alzheimer's Disease. *J. Alzheimer's Dis.* 33, 517–523. [Internet][cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/23202439/. doi:10.3233/JAD-2012-121285

The 1000 Genomes Project Consortium (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* 467 (7319), 1061–1073. [Internet]Oct 28 [cited 2013 May 22]. doi:10.1038/nature09534

Zhang, B. (2020). Integrative Analysis Identifies Copy Number Variations and Their Controlled Causal Molecular Networks in Alzheimer's Disease. *Alzheimer's Dement* 16, e043341, 2020 . [Internet]Dec [cited 2021 Sep 28]Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/alz.043341. doi:10.1002/alz.043341

Zheng, X., Demirci, F. Y., Barmada, M. M., Richardson, G. A., Lopez, O. L., Sweet, R. A., et al. (2015). Genome-wide Copy-Number Variation Study of Psychosis in Alzheimer's Disease. *Transl Psychiatry* 5. , 2015 [Internet]Jun 2 [cited 2021 Jul 12]Available from: https://pubmed.ncbi.nlm.nih.gov/26035058/. doi:10.1038/tp.2015.64

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 01 December 2021
doi: 10.3389/fgene.2021.757170

SVInterpreter: A Comprehensive Topologically Associated Domain-Based Clinical Outcome Prediction Tool for Balanced and Unbalanced Structural Variants

Joana Fino[1]*, *Bárbara Marques*[1], *Zirui Dong*[2,3,4] and *Dezső David*[1]*

[1]Department of Human Genetics, National Health Institute Doutor Ricardo Jorge, Lisbon, Portugal, [2]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China, [3]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China, [4]Hong Kong Hub of Pediatric Excellence, The Chinese University of Hong Kong, Hong Kong, China

OPEN ACCESS

Edited by:
Thomas Liehr,
Friedrich Schiller University Jena, Germany

Reviewed by:
Christopher Grochowski,
Baylor College of Medicine, United States
Edgar Ricardo Vázquez-Martínez,
Universidad Nacional Autónoma de México, Mexico

***Correspondence:**
Joana Fino
joana.fino@insa.min-saude.pt
Dezső David
dezso.david@insa.min-saude.pt

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 11 August 2021
Accepted: 12 October 2021
Published: 01 December 2021

Citation:
Fino J, Marques B, Dong Z and David D (2021) SVInterpreter: A Comprehensive Topologically Associated Domain-Based Clinical Outcome Prediction Tool for Balanced and Unbalanced Structural Variants. Front. Genet. 12:757170. doi: 10.3389/fgene.2021.757170

With the advent of genomic sequencing, a number of balanced and unbalanced structural variants (SVs) can be detected per individual. Mainly due to incompleteness and the scattered nature of the available annotation data of the human genome, manual interpretation of the SV's clinical significance is laborious and cumbersome. Since bioinformatic tools developed for this task are limited, a comprehensive tool to assist clinical outcome prediction of SVs is warranted. Herein, we present *SVInterpreter,* a free Web application, which analyzes both balanced and unbalanced SVs using topologically associated domains (TADs) as genome units. Among others, gene-associated data (as function and dosage sensitivity), phenotype similarity scores, and copy number variants (CNVs) scoring metrics are retrieved for an informed SV interpretation. For evaluation, we retrospectively applied SVInterpreter to 97 balanced (translocations and inversions) and 125 unbalanced (deletions, duplications, and insertions) previously published SVs, and 145 SVs identified from 20 clinical samples. Our results showed the ability of SVInterpreter to support the evaluation of SVs by (1) confirming more than half of the predictions of the original studies, (2) decreasing 40% of the variants of uncertain significance, and (3) indicating several potential position effect events. To our knowledge, SVInterpreter is the most comprehensive TAD-based tool to identify the possible disease-causing candidate genes and to assist prediction of the clinical outcome of SVs. SVInterpreter is available at http://dgrctools-insa.min-saude.pt/cgi-bin/SVInterpreter.py.

Keywords: SVInterpreter, bioinformatic web-tool, clinical outcome prediction, balanced structural variants, copy number variants, topologically associated domains, phenotypic comparison

INTRODUCTION

Structural variants (SVs) are a class of genomic alterations that include balanced (translocations and inversions) and unbalanced (deletions, duplications, and insertions), as well as complex (cx) SVs (Collins et al., 2017). Currently, genome sequencing technologies allow a broader view of genomic variation. Nevertheless, technical issues, as breakpoints located in low complexity sequence regions

that defy the bioinformatic mapping and detection tools capability, still hinder the identification of SVs (Guan and Sung, 2016).

Determining the phenotypic consequences of SVs is challenging. The diversity of its size, genomic content, location, and the intricacy of cxSV make these difficult to interpret, especially considering that they can impinge functional elements located not only within but also outside the affected genomic region (Weischenfeldt et al., 2013). Indeed, SVs alter the genome architecture of the affected regions and have a high probability of changing the position of regulatory elements, known as position effect, which may result in altered gene regulation (Spielmann et al., 2018). Previous studies showed the importance of 3D genome architecture on gene regulation, and how topologically associated domain (TAD) disruption and modification can lead to phenotypic consequences, including the alteration of chromatin loops that are recurrently associated with enhancer–promoter interaction (Lupiáñez et al., 2015; Spielmann et al., 2018)

Therefore, considering the complexity of mechanisms that can link a SV to human disease, the large number of variants identified per individual, and the substantial revision of dispersed data that this entails, ascertainment of SV pathogenicity is a daunting task (Smedley and Robinson, 2015; Zitnik et al., 2019). Furthermore, scarce integration of the available human genome annotation resources and databases also hampers clinical impact prediction of the identified variants (Lindblom and Robinson, 2011).

To date, a number of tools have been shown to tackle the role of unbalanced SVs or copy number variants (CNVs) in human diseases. Tools such as *StrVCTVRE* and *SVscore* focus on a single genomic feature to classify CNVs, as overlap with exons of important genes and precomputed pathogenicity scores of affected single nucleotide polymorphisms, respectively (Ganel et al., 2017; Sharo et al., 2020 [preprint]). *ClinTAD* provides annotation based on TAD context of each CNV, and a possible phenotypic overlap (Spector and Wiita, 2019), whereas *SVFX* uses artificial intelligence approach, based on genomic, epigenomic, and conservation features (Kumar et al., 2020).

For SVs, *AnnotSV* collects clinically relevant information on the genomic elements directly affected by breakpoints (Geoffroy et al., 2018) and *position_effect* predicts genes affected by position effects due to balanced chromosomal abnormality (BCA) breakpoints (Zepeda-Mendoza et al., 2017).

To assist the evaluation of balanced and unbalanced SVs, we previously published two useful bioinformatic tools: *TAD-GConTool* and *CNV-ConTool*. TAD-GConTool automatically defines the regions for following analysis, based on TADs affected by the breakpoints, and retrieves relevant information, whereas CNV-ConTool performs an overlap search against curated CNV databases (David et al., 2020). However, they are still limited in their scope.

Here, we present a more comprehensive tool, *SVInterpreter*, which combines the strengths of our previously published tools, with new features, to retrieve a ready-to-use data table. SVInterpreter gathers the information using breakpoints or genomic positions of balanced or unbalanced SVs, highlighting the relevant data for variant evaluation. Additionally, it performs similarity calculation between the proband's Human Phenotype Ontology (HPO)-based clinical features and those from disorders reportedly associated to genes located within the defined regions (Köhler et al., 2019). Specifically, for CNVs, it performs an overlap search with reported CNVs in public databases and establishes classification scores according to the guidelines of American College of Medical Genetics and Genomics (ACMG) (Riggs et al., 2020).

To demonstrate the robustness of SVInterpreter, we retrospectively applied it to a set of 97 balanced (including 80 translocations and 17 inversions) and 125 unbalanced (5 insertions, 60 deletions, and 60 duplications) previously published SVs as well as 145 SVs identified in 20 clinical samples, by chromosomal microarray (CMA) or genome sequencing. Overall, we demonstrated the efficacy of this tool in retrieving exhaustive genome annotation data of genomic elements affected by SVs, allowing the prediction of their clinical significance.

METHODS

Code and Data Sources

SVInterpreter is a Python-CGI developed Web application, freely available on https://dgrctools-insa.min-saude.pt/cgi-bin/SVInterpreter.py. The code is accessible at https://github.com/DGRC-PT/SVInterpreter, and can be run locally with an Apache configuration.

TAD data from 10 tissue or cell types, available at YUE Lab website[1], were accessed for SVInterpreter. The regions bordering TADs–TAD boundaries—known to potentially restrict interactions of regulatory elements, were predicted using the Dixon pipeline (Dixon et al., 2012), whereas loops were established by *Peakachu* (Salameh et al., 2020).

For the chromosome Y, the TAD average size was calculated for each tissue or cell line, varying from 815 kb for lymphoblastoid cell line GM12878 to 1.8 Mb for bladder tissue (human genome assembly GRCh38/Hg38), and used as reference (**Supplementary Table S1**).

Full description of data sources used by SVInterpreter is available in **Supplementary Table S2**.

Features and Functionality

SVInterpreter analyzes any type of balanced and unbalanced SVs larger than 1 kb (translocations, inversions, insertions, deletions, and duplications) and retrieves a table of compiled information to assist their interpretation. Complex SVs must be subdivided in distinct SVs and analyzed separately (**Supplementary Figure S1**). Optionally, the user can apply SVInterpreter to any genomic region, without specifying the SV type.

SVs can be mapped within cell- or tissue-specific TADs, using the breakpoints as signpost. In this case, by default, TADs affected by breakpoints (brTADs) are retrieved, with the possibility of including up to five additional breakpoint flanking TADs

[1]http://3dgenome.fsm.northwestern.edu/publications.html

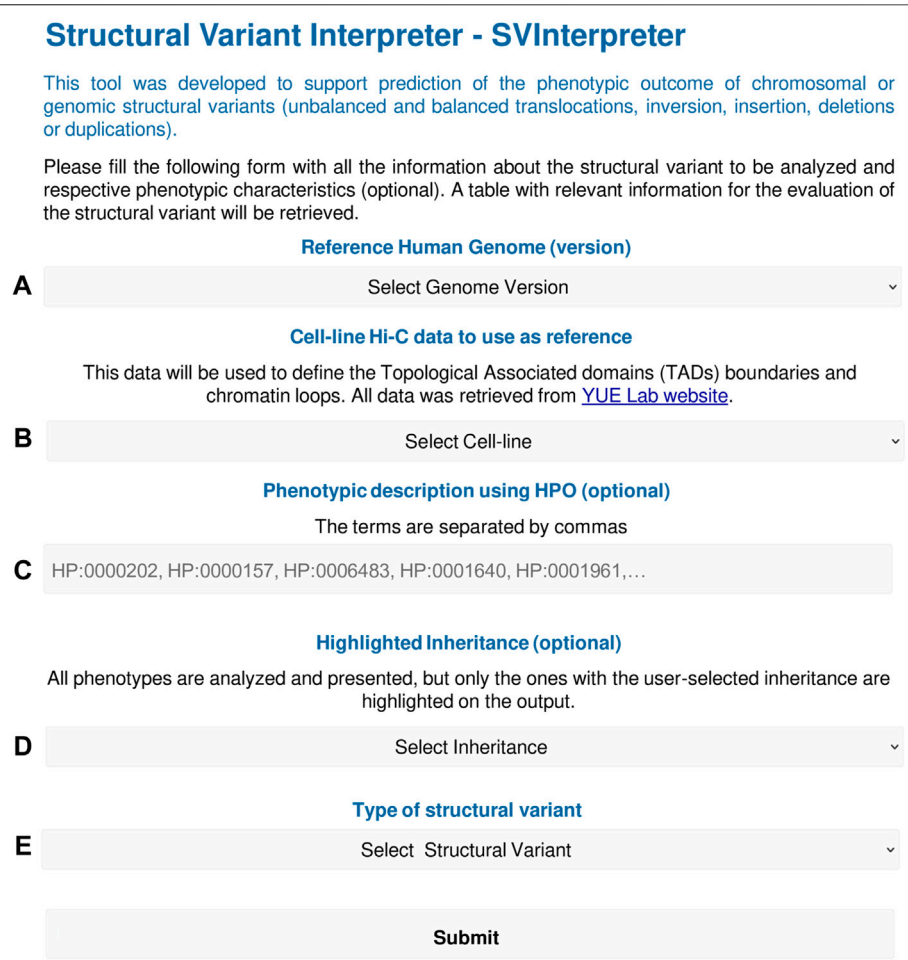

FIGURE 1 | SVInterpreter input form overview. The form starts with **(A)** the selection of the human genome version (Hg19/Hg38), and then **(B)** the tissue or cell line to use as reference for TAD and loop definition. Optionally, the user can **(C)** insert the SV-associated phenotype using HPO terms or **(D)** define an inheritance of interest that will be highlighted on the output. In **(E)**, the type of SV is chosen, which will open a submenu to input the SV-specific parameters as chromosome, breakpoints, and TAD/genomic region to analyze, among others. All SVInterpreter options are shown in detail in **Supplementary Figure S2**.

(TAD−5 to TAD+5). Alternatively, instead of a TAD based analysis, SVs can be analyzed within a genomic region defined by its genomic position (**Supplementary Figure S2**).

To run SVInterpreter, a series of general parameters, such as genome version, tissue, or cell line to be used as reference for TAD and loop definition, and SVs or genomic region-specific parameters (**Figure 1** and **Supplementary Figure S2**), are required.

From the selected specific genomic regions, SVInterpreter data were downloaded from public databases (last updated by March 31, 2021) or are automatically retrieved through an Application Programming Interface (API). From the breakpoints, all functional and non-functional genomic elements are retrieved, whereas from the remaining region, only protein-coding genes, lincRNAs, lncRNAs, functional, and non-functional genomic elements with a GTEx expression pattern are selected (Ardlie et al., 2015). Then, associated data are collected, including human disorders, cancer-specific rearrangements, phenotypes reported in animal models, genome-wide association studies (GWAS) data, and bibliography (**Supplementary Table S2**). The data are organized into a table, with indication of the breakpoint positions following the International System for Human Cytogenomic Nomenclature 2020 (McGowan-Jordan et al., 2020). In addition, to help visualization and interpretation of the SVs within the analyzed genomic regions, links to UCSC genome browser are made available on the output table. In this UCSC genome browser session, the selected genomic region is depicted, highlighting the SV (breakpoint or deleted/duplicated region). Native UCSC genome browser tracks compatible with the output table are shown, together with custom tracks, including the cell line/tissue-specific TADs and chromatin loops. Further description is available in **Supplementary Methods**.

For CNVs, SVInterpreter offers an option of performing an overlap search between the query CNV and those curated in several public CNV databases and published datasets (**Supplementary Figure S2**). The overlap specifications are similar to our previously published CNV-Content Tool (David

et al., 2020), which retrieves the best hit by database, with the respective overlap percentage and variant frequency. In addition to the SVInterpreter standard output table, a detailed overlap table is available for download on the output web page.

Furthermore, to facilitate the evaluation of CNVs according to the ACMG guidelines, together with the standard output, the scoring parameters, as presented on the CNV pathogenicity calculator[2], are retrieved. SVInterpreter outputs the scores for the parameters that are possible to be established automatically and then performs an automatic calculation of the final score and their respective class (Riggs et al., 2020).

The output table(s) are written in XLSX format and made available for download. Further description of the output, a step-by-step tutorial, and an application example is available in **Supplementary Methods** and **Supplementary Table S3**.

Phenotypic Similarity Search

Optionally, the proband's HPO-based phenotypic features can be inputted for phenotype comparison (Köhler et al., 2019).

For this, the HPO ontology provided by the HPO.db package[3] and the links between genes, diseases, and terms provided by R data file (RDA) are a prerequisite. Since these were deprecated, we developed in-house scripts and used the June 2021 HPO release data[4] (Köhler et al., 2019) to create state-of-the-art HPO.db and RDA files. The scripts and guidelines are available at https://github.com/DGRC-PT/HPOSim_Helper_Scripts.

The phenotype similarity is evaluated based on phenotype similarity score (PhenSSc), maximum similarity score (MaxSSc), and *p*-value (*p*), which are calculated for each combination of inputted phenotype and Online Mendelian Inheritance in Man (OMIM)[5] phenotype associated with functional genomic elements within the analyzed region. This is performed by *HPOSim*—getTermListSim function that calculates pairwise similarities between HPO terms, using the information content (IC) of the most informative ancestor shared by both terms (Deng et al., 2015). The IC is a numeric value associated to each term, which inversely reflects the number of diseases annotated by the term, or any of its descendent terms. That is, terms with higher ICs annotate fewer diseases, being more specific, whereas lower ICs are associated to most common terms. When comparing groups of HPO terms, the getTermListSim result is the mean of the ICs of the pairwise comparisons, and reflects the similarity between the said groups, where higher scores represent higher similarities.

For PhenSSc, the inputted clinical features and the ones associated to a disorder are compared. This score reflects the similarity between the inputted phenotypic traits and the ones used to describe the disorder.

For MaxSSc, the inputted clinical features are compared with themselves, which means that MaxSSc consists of the mean of the ICs of the inputted terms. This metric was developed by us to reflect the maximum similarity score that can be obtained from the inputted terms, and to be used in comparison with PhenSSc.

The *p*-value, which reflects the probability of obtaining the PhenSSc by random chance, was adapted from Redin et al. (2017). In sum, for each disorder that PhenSSc and MaxSSc was previously calculated, a random set of HPO terms is selected. Most importantly, this set must have the same number of terms as the input, to limit the bias. The similarity score is then calculated between this set of terms and a disorder-associated phenotype (*simulated score*). Then, this is repeated 100 (*n*) times, where each time a different set of HPO terms is selected, and a new *simulated score* is obtained. Finally, the disorder specific *p*-value is calculated as:

$$P = \frac{\sum_{i=1}^{n} [simulatedscore_i \geq PhenSSc]}{n} \quad (1)$$

Phenotypes mainly composed of terms common in a wide range of disorders, as global developmental delay, or intellectual disability (with 1,386 and 1,619 associated OMIM disorders[4], respectively), can present high PhenSSc, close to MaxSSc, and a high *p*-value as well. In these cases, the high *p*-value reflects the high probability of the phenotype to overlap by chance, warning for the limited significance of PhenSSc. Hence, ideally, the PhenSSc should be close to MaxSSc and present a *p*-value as close to zero as possible.

SVs and Clinical Cases

For retrospective analysis, 97 and 125 previously published and unpublished balanced (translocations, inversions) and unbalanced (insertions, deletions and duplications) SVs were selected, respectively (**Table 1**; **Supplementary Table S5**) (David et al., 2003, 2009, 2015, 2018, 2020; Redin et al., 2017; Riggs et al., 2020). Of note, about half of those published by Redin et al. (2017) were previously analyzed by Zepeda-Mendoza et al. (2017), with the position_effect[6] tool, for identification of additional candidate genes.

For effectiveness evaluation in clinical setting, nine prenatal cases (three without associated ultrasound abnormalities, four with isolated increased nuchal translucency, one with limb abnormalities, and another with multisystemic traits) and 11 postnatal cases (with isolated organ-specific or complex multisystem disorders) were used (**Table 1**; **Supplementary Table S5**). These were randomly selected among those referral for clinical diagnosis, of which genomic variants were identified by CytoScan 750K (nine cases), CytoScan HD microarrays (six cases), and long-insert genomic sequencing (liGS) (five cases). Microarray and liGS analysis was carried out as previously described (David et al., 2018, 2020).

Criteria for SV Interpretation and Clinical Prediction

The microarray data were processed using Chromosome Analysis Suite 4.2.0.80 with NetAffx 20200828 (GRCh37/Hg19) and with the detection criteria of, at least, 15 probes within 35 kb for gains

[2]https://cnvcalc.clinicalgenome.org/cnvcalc/

[3]http://www2.uaem.mx/r-mirror/web/packages/HPO.db/index.html

[4]https://hpo.jax.org/app/

[5]https://omim.org/

[6]https://github.com/ibn-salem/position_effect

TABLE 1 | SVs analyzed with SVInterpreter.

	Retrospectively reevaluated SVs			Clinical cases (PND; PN)[d]		Total by SV (retr. SVs/Clin. Cases)
	David et al.[a]	Redin et al.[b]	Riggs et al.[c]	Microarray (9; 6)[e]	liGS (0; 5)	
Translocation	9	71	0	0; 0	0; 2	80 / 2
Inversion	2	15	0	0; 0	0; 9	17/9
Deletion	2	0	58	26; 24	0; 24	60/74
Duplication	4	0	56	19; 21	0; 5	60/45
Insertion	4	1	0	0; 0	0; 13	5/13
cxSVs	0	0	0	0; 0	0; 2	0/2
Total by Publication	21	87	114	45; 45	0; 6	222/145

[a]*David et al., 2003, 2009, 2015, 2018, 2020*
[b]*Redin et al., 2017*
[c]*Riggs et al., 2020*
[d]*PND, Prenatal diagnosis; PN, Postnatal diagnosis*
[e]*PN diagnosis performed by Cytoscan HD with microarray; PND performed by Cytoscan 750K.*

TABLE 2 | Parameters used for the classification of SVs.

Classification	Parameters (translocation, inversion, insertion)
Pathogenic	Variant affecting or encompassing genes associated with dominant developmental disorders
Likely Pathogenic	Variant affecting or encompassing genes with a pli ≥ 0.9 not associated with disease Or Breakpoint located near a candidate gene associated with AD developmental disorders in a subject showing significant phenotype overlap with the referred disorder and predicted to impact long-range regulatory interactions
Variant of unknown significance (VUS)	All other variants not fitting Pathogenic, Likely Pathogenic, Likely Benign, and Benign parameters
Likely Benign	Variant affecting or encompassing genes only associated with AR disorder And No other data that support at least a partial overlap between the proband's phenotype and the affected genomic region
Benign	Variant not affecting or encompassing any genes And No human pathology reported to be associated with genomic elements localized within the disrupted TAD or no other data that support at least the partial overlap between the proband's phenotype and the affected genomic region

and losses. Selected SVs were manually interpreted based on the following criteria: absence/presence of OMIM genes, their association with autosomal dominant (AD) or recessive (AR) disorders, disruption of genes by the breakpoints, haploinsufficiency/triplosensitivity, and genotype/phenotype correlation (Silva et al., 2019). For this, data available at UCSC genome browser[7], Decipher[8], ClinGen[9], ClinVar[10], OMIM, DGV[11], Unique[12], and Orphanet[13] databases were used.

For liGS, SVs larger than 1 kb, and CNVs identified by discordant pair clustering and coverage analysis, were selected. Then, among these, novel variants and SVs overlapping a reported variant (Collins et al., 2017; Chaisson et al., 2019, Gnomad[14]) with a database frequency <1%, and affecting loss-of-function (LoF) sensitive genes [with an expected vs. observed ratio (oe) of LoF variants of <0.35] and/or associated to AD disorders, were indicated for clinical evaluation. On average, per individual, 11 SVs were selected for analysis.

Three evaluators classified the Riggs dataset (Riggs et al., 2020) of unbalanced SVs. Therefore, based on the following criteria: (1) a classification equal by at least two of the evaluators, or (2) a median classification that reflects dissimilar evaluations, we merged them into a consensus classification (**Supplementary Table S5**).

To allow the comparison between published predicted outcomes of SVs and SVInterpreter-based prediction, criteria for translocations, inversions, and insertions were adapted from the previously described ones (**Table 2**) (Redin et al., 2017; David et al., 2020). For CNVs, ACMG guidelines were applied (Riggs et al., 2020). In addition, the same genome version and reference cell line as in the original publications was used. If available, the proband's phenotype was inputted. For variants without pre-set of reference cell line, the human embryonic stem cell was used. By default, for all types of variants, the brTAD was used as reference, with rare exceptions. For CNVs, the overlap search against all available databases with a minimum mutual overlap of 70% was applied. The full set of variants and parameters used is available at **Supplementary Table S5**.

[7]https://genome.ucsc.edu/
[8]https://decipher.sanger.ac.uk/
[9]https://clinicalgenome.org/
[10]https://www.ncbi.nlm.nih.gov/clinvar/
[11]http://dgv.tcag.ca/dgv/app/home
[12]https://rarechromo.org/
[13]https://www.orpha.net/consor/cgi-bin/index.php
[14]https://gnomad.broadinstitute.org/

RESULTS

Retrospective Reevaluation of Published SVs

For retrospective analysis, 97 balanced and 125 unbalanced previously published SVs were reevaluated (**Table 1**;

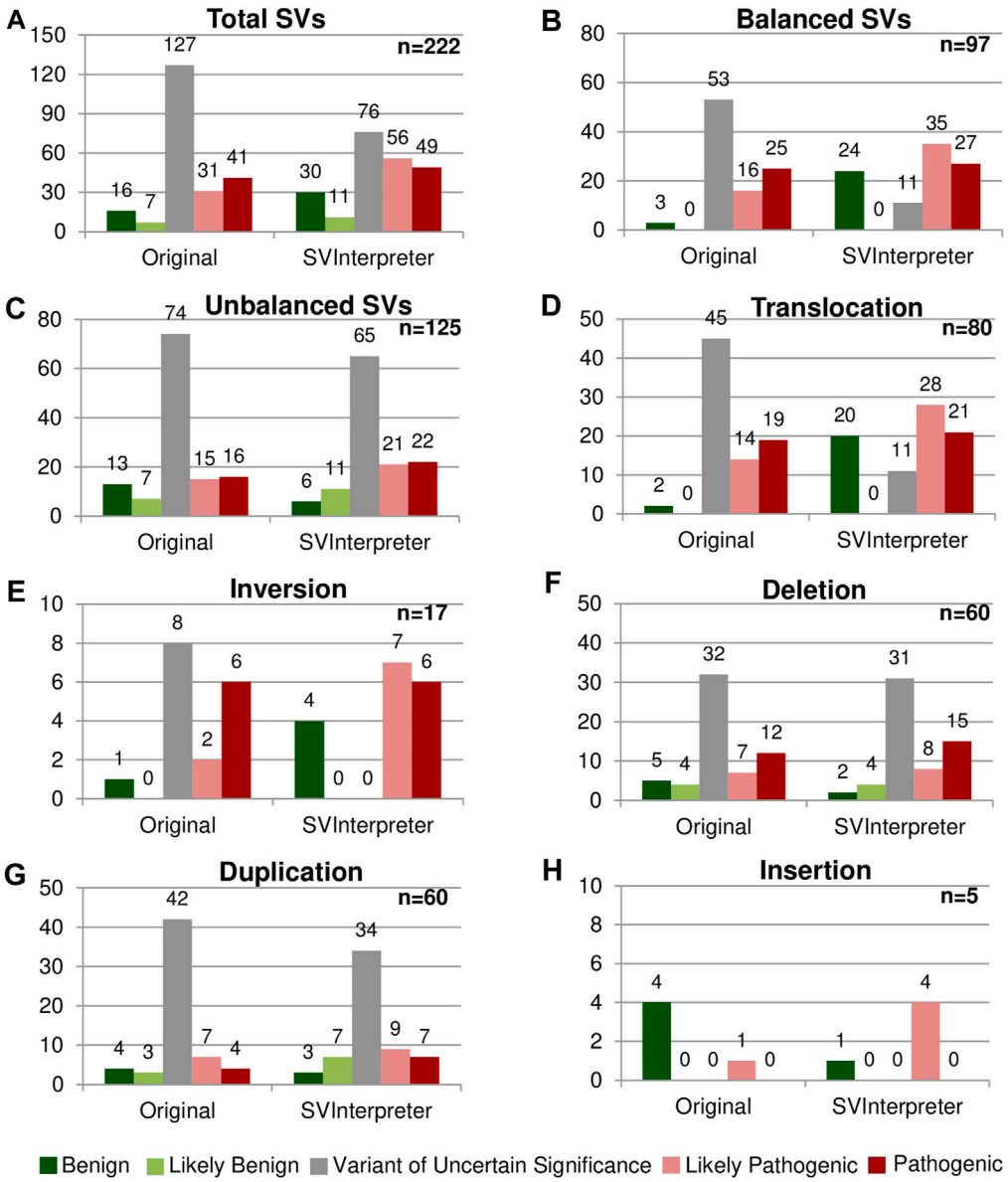

FIGURE 2 | Comparison between the original and the SVInterpreter-based clinical outcome prediction. Each graphic presents the comparison between the original classification, and tool-based clinical outcome prediction for **(A)** total of SVs, **(B)** balanced SVs, **(C)** unbalanced SVs, **(D)** translocations, **(E)** inversions, **(F)** deletions, **(G)** duplications, and **(H)** insertions. Bars are color-coded, according to the clinical outcome prediction, as benign (dark green), likely benign (light green), VUS (gray), likely pathogenic (light red), and pathogenic (dark red). Number of variants is shown above the bars.

Supplementary Table S5). With the exception of chromosome 21, SVs are distributed regularly along the genome, with an average of 12 rearrangements per chromosome. Nevertheless, the larger number of translocations (n = 15), inversions (n = 5), and insertions (n = 3) involved chromosome 1, chromosomes 2 and X, and chromosome 3, respectively (**Supplementary Table S4**).

These variants were reevaluated by SVInterpreter, and based on its retrieved data, their clinical outcome was predicted according to the established parameters (**Figure 2**; **Supplementary Table S5**).

The first level of analysis involves functional and non-functional genomic elements localized within the brTADs and their annotation data, which is usually sufficient to evaluate a SV. For clinical outcome prediction of a gene disruption, SVInterpreter retrieves gene-specific annotation data such as the LoF sensitivity, Genomics England PanelApp[15] data, its association with disorders and respective phenotypic overlap, animal model data, gene expression patters, and GWAS data.

[15]https://panelapp.genomicsengland.co.uk/

Concomitantly, the disruption of major genes by *de novo* BCA breakpoints leading to major AD developmental disorders, as retrieved by SVInterpreter, indicated the pathogenicity of *ANKRD11* (OMIM *611192; proband DGRC0016) and *WDR26* (OMIM *617424; proband DGRC0025) (David et al., 2020). In the abovementioned cases, the calculated similarity between the inputted phenotypes and of gene-associated disorders localized within the analyzed regions played a major part on the interpretation, where *ANKRD11* PhenSSc was 2.64 ($p = 0.02$; MaxSSc = 4.01) and *WDR26* PhenSSc was 2.31 ($p = 0.02$; MaxSSc = 2.91).

If the full extent of the clinical features cannot be explained by disruption or misregulation of a candidate gene, or the breakpoint is within an intergenic region, in search for potential position effect, annotation data of all genomic elements within a brTAD must be evaluated. Several data retrieved by SVInterpreter can suggest position effect events. In addition to the phenotypic overlap and expression pattern, disruption of chromatin loops and GeneHancer clusters of interactions are important signs for possible position effect (Gloss and Dinger, 2018).

DGAP131 t(1;5)(p31;q33)dn, was originally classified by Redin et al. (2017) as variant of uncertain significance (VUS). SVInterpreter showed *MEF2C*'s (OMIM *600662) GeneHancer cluster of interactions and 10 of its 14 chromatin loops disrupted by the chromosome 5 breakpoint. The PhenSSc of 1.82 ($p = 0.02$; MaxSSc = 3.1) corroborated the proposed position effect. Likewise, *MEF2C* was indicated as a potential candidate gene by Zepeda-Mendoza et al. (2017).

Then, if the protein coding genes or functional genomic elements localized within the brTADs are insufficient to explain the observed phenotype, additional upstream (−1 to −5) and downstream (+1 to +5) flanking TADs are analyzed.

Accordingly, the t(2;11)(q14.2;q14.2) breakpoints reported in proband DGRC0001 were located in intergenic regions, and no gene at the brTADs, capable of explaining the verified phenotype, was found. At TAD+1, SVInterpreter shows that the GeneHancer cluster of interactions of the proposed candidate gene *GLI2* (OMIM * 165230) was disrupted by the 2q14.2 breakpoint, confirming the previously proposed position effect (David et al., 2009). Furthermore, the involvement of *GLI2* was reinforced by its PhenSSc of 1.33 ($p = 0.3$; MaxSSc 4.2), with disorders OMIM#615849 and OMIM#610829. Ergo, the translocation was predicted to be likely pathogenic and confirmed the published assertion of the involvement of GLI2 (David et al., 2009).

Furthermore, DGAP107 t(Y;3)(p11.2;p12.3)dn, reported by Redin et al. (2017), presents, among others, neurological defects, urinary tract, and genital abnormalities. They originally classified the SV as potentially pathogenic, due to the disruption of *ROBO2* (OMIM *602431). By assessing the associated disorder (OMIM #610878), we realized that *ROB O 2* only explained the urinary tract defects (PhenSSc = 1.12; $p = 0.08$; MaxSSc 1.48). However, SVInterpreter brTAD analysis suggested a position effect on *PCDH11Y* (OMIM * 400022), which had its GeneHancer cluster of interactions disrupted. The gene function, expression pattern, and animal model data suggest its role in the development of the nervous system, and therefore may explain the neurological defects observed in the proband. Besides, Zepeda-Mendoza et al. (2017) also indicate *SRY* (OMIM *480000), located at TAD-3, as a candidate gene due to the overlap with the genital abnormalities.

The overlap search of query CNVs in public database data and the automatic ACMG scoring showed to be of utmost utility, since it can clarify immediately the potential significance of deletions and duplications, even in cases where the genomic data are scarce. As such, a 374-kb deletion, arr[GRCh37] 10q22.3(81,603,169_81,976,925)x1, in case CK without associated phenotype, was classified by Riggs et al. (2020) as VUS. According to SVInterpreter, the CNV deleted five genes that were not associated to phenotype or reported to be haploinsufficient. The CNV had 100% overlap with a likely benign ClinGen deletion (nsv3896137), and according to its ACMG CNV score of −0.9, the deletion was classified as likely benign.

Overall, more than half (57.2%) of the reevaluated SVs (45 translocations, 8 inversions, 32 deletions, and 42 duplications) were originally classified as VUS, whereas only 10.4% (23) were classified of benign and likely benign (**Figures 2A–C**). SVInterpreter-based reevaluation of published SVs provided a consistent finding with the original studies on 62.6% of all SVs (39 translocations, 9 inversions, 44 deletions, 45 duplications, and 2 insertions) (**Supplementary Table S5**). Comparatively with the original classification, the number of variants predicted as VUS decreased by 40% (from 127 to 76) (**Figure 2A**). For balanced SVs, SVInterpreter-based interpretation led to the reevaluation of 81.1% of the original VUS as pathogenic (9.4%), likely pathogenic (32.1%), and benign (39.6%) (**Figures 2B,D,E**). In addition, position effect events identified by SVInterpreter sustained the categorization of 30.2% of the potentially pathogenic balanced SVs (**Supplementary Table S5**). For deletions, the differences between published and tool-based prediction were minor, with similar results obtained by both (**Figure 2F**). Differently, 19% of the duplications categorized the VUS transited to another category, whereas only three insertions were reclassified from benign to likely pathogenic (**Figures 2G,H**).

To assess the position effect on distal genes and their contribution on the observed phenotypes, from the 87 balanced SVs published by Redin et al. (2017) and reevaluated by us, Zepeda-Mendoza et al. (2017) also analyzed 44 (**Figure 3A**). Similar candidate genes were identified in 11 of the SVs (**Figure 3B**), whereas in 5, neither of them proposed a candidate gene (**Figure 3A**). SVInterpreter and position_effect identified the same candidate genes for two originally classified VUS and two pathogenic SVs (**Figure 3B**; **Supplementary Table S5**). The position_effect tool uniquely identified 24 candidate genes in 19 SVs, where, in 14 of them, the genes were located outside the brTAD (**Figure 3C**; **Supplementary Table S5**). Based on expression, phenotypic overlap, and animal model data, SVInterpreter predicted six candidate genes not foreseen by the other two approaches, in five SVs (**Figure 3C**; **Supplementary Table S5**).

Variant Interpretation in Clinical Setting

The effectiveness of this bioinformatic tool in a clinical setting was evaluated by comparative (manual vs. SVInterpreter-based),

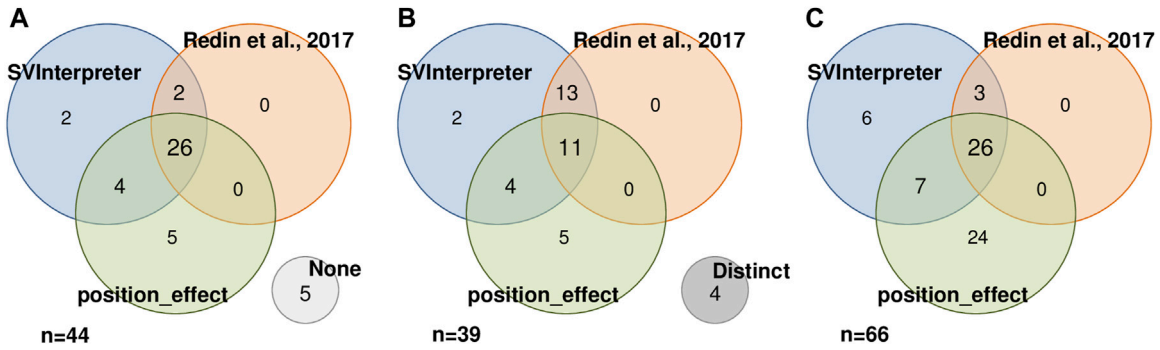

FIGURE 3 | Result comparison of 44 SVs analyzed by SVInterpreter (blue), Redin et al. (2017) (orange), and position_effect (green), in three different perspectives. **(A)** SVs with associated candidate genes, including five SVs with no candidate gene identified by any of the approaches ("None", light gray). **(B)** Similarity between the sets of candidate genes identified by SV, including four SVs where the retrieved candidate genes were different in the three approaches ("Distinct", dark gray). **(C)** Intersection between the identified candidate genes. For **(A)** and **(B)**, the numbers inside the circles correspond to the number of SVs, while in **(C)**, the numbers represent the number of candidate genes.

clinical outcome prediction of SVs identified by Cytoscan 750K microarray in nine prenatal cases, by Cytoscan HD microarray in six postnatal cases, and by liGS in five postnatal cases. Altogether, 145 variants (SVs, CNVs, and cxSVs) were analyzed (**Table 1**). The average number of SVs per individual, identified by genomic array, was 6, whereas for liGS, it was 289, with 44 balanced and 244 unbalanced variants. From the latter, on average, only 11 SVs (3 balanced and 8 unbalanced) were recognized to be potentially disease causing or pathogenic, and consequently selected for clinical outcome prediction (**Supplementary Table S5**).

Proband DGRC0004 presented a severe phenotype characterized by global developmental delay, facial dysmorphisms, and heart defects. Among other, the liGS data analysis identified a 67.3-Mb inversion inv(2)(p16.1q14.3) and a 589-kb duplication dup(2)(q21.1). Since, based on the SVInterpreter data, none of the inversion breakpoints disrupted a gene, nor any gene localized within the brTAD supported the verified phenotype, the inversion was classified as benign. Concerning the dup(2)(q21.1), although SVInterpreter identified an identical CNV in a cohort of patients with developmental delay (nsv999864) (Coe et al., 2014), the duplication has the same gene content as reported in benign SVs and did not affect triplosensitive genes, which led to its likely benign classification (ACMG CNV score = −0.9). Furthermore, none of the remaining eight clinically evaluated SVs was predicted to be likely pathogenic or pathogenic; therefore, genomic disorder was excluded in this case. Indeed, exome sequencing identified a pathogenic single-nucleotide variant within *KAT6A* (OMIM *616268) exon 18, causing AD Arboleda-Tham syndrome (OMIM #616268) (data not shown). The clinical features of this syndrome overlap that of the proband.

As CMA is the technique of choice for identification of CNVs in a clinical setting, the automatic mutual overlap search with CNV public databases and the inclusion of the ACMG scoring system is especially valuable for faster and more informed clinical outcome prediction of these.

A female in her 40s presented a dichorionic diamniotic pregnancy with an elevated risk for aneuploidy following first trimester combined screening test and normal ultrasound examination. Microarray analysis of chorionic villus sample DNA (CS750K07) identified five deletions and two duplications. By manual analysis, due to the absence of genes within the five deleted regions, these were classified as benign, whereas one of the two duplications, encompassing only a non-morbid gene, was classified as likely benign. SVInterpreter confirmed the benign and likely benign classifications, and the absence of overlapping CNVs and triplosensitive genes. In contrast, the remaining 1.1 Mb duplication at 16p13.11, arr[GRCh37] 16p13.11(15,416,498_16,527,659)x3, was classified as VUS, since the CNV was overlapped by the 16p13.11 microduplication syndrome, which likely presents an incomplete penetrance and phenotypic variability. SVInterpreter identified four overlapping disorder-associated genes, *NDE1* (OMIM *609949), *MYH11* (OMIM *160745), *ABCC1* (OMIM *158343), and *ABCC6* (OMIM *603234). Although these genes are associated to AD or AR disorders, neither of them is triplosensitive or is disrupted by the breakpoints. SVInterpreter identified overlapping duplications that were reportedly classified as pathogenic (nssv15605791), likely pathogenic (nssv15149610), likely benign (nssv15159627), and VUS (nssv15159626). In addition, automatic bibliography search identified publications that described the 16p13.11 microduplication syndrome (PMID: 30287593, PMID: 23637818). Hence, in the absence of prenatal phenotype–genotype correlation, the contradictory classifications of similar duplications, and the overlap with the microduplication syndrome, we maintained the original classification of VUS.

We confirmed the reported manual clinical prediction of SVs identified in 20 individuals analyzed in a clinical setting, with marginal variability between these two approaches (**Supplementary Table S5**).

DISCUSSION

Here, we describe SVInterpreter, a web-based tool to assist the clinical outcome evaluation of balanced and unbalanced SVs. SVInterpreter assesses the regions affected by SVs, retrieves associated genome annotation data, and organizes the results in a user-friendly table. Furthermore, it scores CNVs according to ACMG criteria and assesses the overlapped variants from public databases. SVInterpeter can be used in a straightforward identification of gene disruption, evaluation of phenotypic similarities, and the indication of potential position effects within the breakpoint or flanking TADs.

As shown by retrospective analysis of the BCA cases DGRC0016 and DGRC0025 (David et al., 2020), assessment of genotype–phenotype correlation through comparison between the probands' clinical features and of disorders caused by the disrupted genes localized within the affected genomic regions easily and quickly pointed out the pathogenicity of the analyzed variants.

Importantly, clinical setting requires tools that retrieve sufficient and adequate information to allow exclusion of the pathogenicity of SVs, in a timely fashion.

Due to the limited clinical manifestations, phenotype similarity search cannot assist in guiding the clinical outcome prediction of SVs in prenatal cases. Certainly, the availability of a dedicated fetal genotype–phenotype correlation database would further assist prenatal evaluation of SVs. Indeed, ultrasound features were absent in our prenatal sample CS750K07, making genotype–phenotype correlation practically impossible, for the 1.1-Mb duplication at 16p13.11. However, long-term follow-up would be warranted to exclude any later-onset disorder that might be associated with the SV (Halgren et al., 2018). By SVInterpreter, we were able to corroborate the manual prediction results in clinical setting, although the main advantage was essentially a more straightforward, comprehensive, and faster evaluation process.

As demonstrated by DGAP131 and DGRC0001, combination of phenotypic overlap search and identification of disrupted GeneHancer cluster of interactions and chromatin loops within the breakpoint or flanking TADs is essential for prediction of position effect events. This is true not only for breakpoints within intergenic regions where assessment of a position effect is crucial, but also for SVs where disruption of a main candidate gene is insufficient to explain the full spectrum of clinical features.

In most cases, gene disruptions or position effects within brTADs were sufficient to explain the phenotypes. Even in comparison with candidate genes uniquely identified by position_effect (**Figure 3C**), most of the ones located outside the brTAD showed to be associated to phenotypic traits that were already explained by genes inside the brTAD. However, in DGAP107, the full extent of associated clinical features was only resolved by a potential position effect on the third flanking TAD. This, combined with the current lack of knowledge in respect to TADs, shows the difficulty of establishing, at first hand, the region to be reviewed when evaluating an SV. This includes the arduousness of choosing, among the few, the adequate cell line or tissue to use as reference, as only recently has the TAD boundaries variability between tissues been documented (Sauerwald et al., 2020). SVIntepreter allows users to develop their own strategy to tackle this; nevertheless, we suggest to progressively increase the size of the analyzed genomic region, from the brTADs up to the fifth flanking TADs.

SVInterpreter retrieves the most comprehensive information, unraveling the role of genes not yet associated with disease. This was demonstrated by the identification of the potential candidate gene *PCDH11Y* in DGAP107, which was neglected by both Redin et al. (2017) analysis and position_effect (Zepeda-Mendoza et al., 2017).

For CNV analysis, SVInterpreter takes advantage of the resources available for unbalanced SVs. As displayed on CK and CS750K07, the overlap with database-classified CNVs and the automatic ACMG scoring made the evaluation much easier. Also, the automatic bibliography search complements the analysis, by presenting to the user a selection of publications of interest, which can provide data that eventually is unavailable on databases.

According to the features and results presented above, and especially the decrease of the previously classified VUS by 40%, we conclude that SVInterpreter alone provided enough support for assessment of the SVs. Nevertheless, we recognize that differences between Redin et al (2017) and our evaluation were affected by the fact that their classification criteria were more stringent and did not comprise benign and likely benign categories, and that additional knowledge has been acquired since their publication (El Mecky et al., 2019). Supporting this is the small number of deletions that were reclassified, since the ACMG criteria were equally used for the original and SVInterpreter-based analysis.

A major improvement of SVInterpreter was the inclusion of a function for phenotype comparison, developed mainly based on Köhler et al. (2009), Deng et al. (2015), and Redin et al. (2017). Since the phenotypic similarity scores are based on the HPO terms' IC (Köhler et al., 2009), the score has no scale, varying with the specificity of the term, and the number of terms used for phenotype description, making it difficult to evaluate PhenSSc by itself. Therefore, MaxSSc, which reflects the upper limit of the scale for each specific set of inputted clinical features, together with the *p*-value, which measures the probability of the PhenSSc being obtained by chance, are used to interpret the PhenSSc.

Comparatively with other recent tools that support the evaluation of SVs, such as position_effect (commit: fced2c49, 13 June 2017), AnnotSV (Version 1.0, 21 December 2017) and ClinTAD (commit: 09b4925fb, 18 September 2019), SVInterpreter seems to be more comprehensive (Zepeda-Mendoza et al., 2017; Geoffroy et al., 2018; Spector and Wiita, 2019). First, SVInterpreter showed to be the one that allows more customization and adjustments, since, for example, AnnotSV and ClinTAD only work with one genome version, and ClinTAD only uses TAD boundaries of human embryonic stem cell data. Then, SVInterpreter shows a broader view of the affected regions, accounting for both gene disruption and position effects: AnnotSV is focused on the identification of genes directly affected by a breakpoint, and position_effect was designed to identify candidate genes essentially from position effect events. In regard to phenotypic comparison, as AnnotSV does not perform any, and ClinTAD is limited to the full HPO term overlap, position_effect is the only one with a similar functionality. Also, SVInterpreter is the one that retrieves the most

information, including the position effect important data, GeneHancer cluster of interactions and chromatin loops, phenotypic data from DDG2P and clinGen, Gene-phenotype/disease associations in animal models, and GWAS data. Therefore, the existence of overlooked information by position_effect and AnnotSV, as shown in DGAP107, may contribute to limited results, biased candidate gene prioritization, and the need of additional resources.

Nonetheless, SVInterpreter still presents some limitations. The retrieved data are limited to the content of the available databases, which are regularly outdated with respect to the state of the art. This is currently remedied by the inclusion of the bibliographic search, but it can be improved by application of automatic text-mining systems (Luque et al., 2019). For cases of multisystemic phenotypes where more than one gene may be involved, the phenotypic overlap search could eventually be improved by adding individual phenotypic scores calculated for HPO supercategories. Additionally, SVInterpreter is prepared to analyze one variant at a time, which can be a disadvantage when dealing with complex rearrangements, or clinical cases with a large number of variants. Therefore, periodical update of this bioinformatic tool seems warranted.

The interpretation of any SV is not a straightforward task, even with the help of the right tools, since it is difficult to make sure that all factors are being considered. We do not expect SVInterpreter to change the result of the current SV evaluation, since it depends on the level of genome annotation, our current knowledge on pathological mechanisms in human disease, and, ultimately, reported data. Instead, this tool allows a well-informed and faster way to interpret SVs. Regardless of the bias given by the currently available data, attempts are being made to automate the clinical SV interpretation, which will change the current paradigm (Kumar et al., 2020). We believe that SVInterpreter, a tool to support the evaluation of balanced and unbalanced SVs, represents one more step towards this goal.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the National Institute of Health Doutor Ricardo Jorge. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JF developed the code of the application. DD, BM, and ZD advised on the functionality and content of the tool. DD supervised the tool's development and application. JF and BM performed the analysis of SVs. All the authors tested the tool, wrote revised, read, and approved the final article.

FUNDING

This research was supported by national funds through FCT—Fundação para a Ciência e a Tecnologia, Research Grant HMSP-ICT/0016/2013 of the Harvard Medical School—Portugal Program in Translational Research and Information.

ACKNOWLEDGMENTS

We would like to thank Márcia Rodrigues, João Freixo, Joana Paiva, and Manuela Cardoso for their feedback on the tool functionality and data, during its development. In addition, we thank the Citogenetics Unit of the Human Genetics Department of the National Health Institute Doutor Ricardo Jorge, for their contribution. Lastly, we thank the Technologies and Information Systems Unit of the National Health Institute Doutor Ricardo Jorge for the informatics support and for hosting the tool.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.757170/full#supplementary-material

REFERENCES

Ardlie, K. G., DeLuca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., et al. (2015). Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* 348, 648–660. doi:10.1126/science.1262110

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes. *Nat. Commun.* 10, 1784. doi:10.1038/s41467-018-08148-z

Coe, B. P., Witherspoon, K., Rosenfeld, J. a., Van Bon, B. W. M., Vulto-Van Silfhout, A. T., Bosco, P., et al. (2014). Refining Analyses of Copy Number Variation Identifies Specific Genes Associated with Developmental Delay. *Nat. Genet.* 46, 1063–1071. doi:10.1038/ng.3092

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., et al. (2017). Defining the Diverse Spectrum of Inversions, Complex Structural Variation, and Chromothripsis in the Morbid Human Genome. *Genome Biol.* 18, 36. doi:10.1186/s13059-017-1158-6

David, D., Almeida, L. S., Maggi, M., Araújo, C., Imreh, S., Valentini, G., et al. (2015). Clinical Severity of PGK1 Deficiency Due to a Novel p.E120K Substitution Is Exacerbated by Co-inheritance of a Subclinical Translocation t(3;14)(q26.33;q12), Disrupting NUBPL Gene. *JIMD Rep.* 23, 55–65. doi:10.1007/8904_2015_427

David, D., Anand, D., Araújo, C., Gloss, B., Fino, J., Dinger, M., et al. (2018). Identification of OAF and PVRL1 as Candidate Genes for an Ocular Anomaly

Characterized by Peters Anomaly Type 2 and Ectopia Lentis. *Exp. Eye Res.* 168, 161–170. doi:10.1016/j.exer.2017.12.012

David, D., Cardoso, J., Marques, B. á., Marques, R., Silva, E. D., Santos, H., et al. (2003). Molecular Characterization of a Familial Translocation Implicates Disruption of HDAC9 and Possible Position Effect on TGFβ2 in the Pathogenesis of Peters' Anomaly. *Genomics* 81, 489–503. doi:10.1016/S0888-7543(03)00046-6

David, D., Freixo, J. P., Fino, J., Carvalho, I., Marques, M., Cardoso, M., et al. (2020). Comprehensive Clinically Oriented Workflow for Nucleotide Level Resolution and Interpretation in Prenatal Diagnosis of De Novo Apparently Balanced Chromosomal Translocations in Their Genomic Landscape. *Hum. Genet.* 139, 531–543. doi:10.1007/s00439-020-02121-x

David, D., Marques, B., Ferreira, C., Vieira, P., Corona-Rivera, A., Ferreira, J. C., et al. (2009). Characterization of Two Ectrodactyly-Associated Translocation Breakpoints Separated by 2.5 Mb on Chromosome 2q14.1-q14.2. *Eur. J. Hum. Genet.* 17, 1024–1033. doi:10.1038/ejhg.2009.2

Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology. *PLoS One* 10, e0115692. doi:10.1371/journal.pone.0115692

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* 485, 376–380. doi:10.1038/nature11082

El Mecky, J., Johansson, L., Plantinga, M., Fenwick, A., Lucassen, A., Dijkhuizen, T., et al. (2019). Reinterpretation, Reclassification, and its Downstream Effects: Challenges for Clinical Laboratory Geneticists. *BMC Med. Genomics* 12, 170. doi:10.1186/s12920-019-0612-6

Ganel, L., Abel, H. J., and Hall, I. M. (2017). SVScore: An Impact Prediction Tool for Structural Variation. *Bioinformatics* 33, btw789–1085. doi:10.1093/bioinformatics/btw789

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., et al. (2018). AnnotSV: an Integrated Tool for Structural Variations Annotation. *Bioinformatics* 34, 3572–3574. doi:10.1093/bioinformatics/bty304

Gloss, B. S., and Dinger, M. E. (2018). Realizing the Significance of Noncoding Functionality in Clinical Genomics. *Exp. Mol. Med.* 50, 1–8. doi:10.1038/s12276-018-0087-0

Guan, P., and Sung, W.-K. (2016). Structural Variation Detection Using Next-Generation Sequencing Data. *Methods* 102, 36–49. doi:10.1016/j.ymeth.2016.01.020

Halgren, C., Nielsen, N. M., Nazaryan-Petersen, L., Silahtaroglu, A., Collins, R. L., Lowther, C., et al. (2018). Risks and Recommendations in Prenatally Detected De Novo Balanced Chromosomal Rearrangements from Assessment of Long-Term Outcomes. *Am. J. Hum. Genet.* 102, 1090–1103. doi:10.1016/j.ajhg.2018.04.005

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) Knowledge Base and Resources. *Nucleic Acids Res.* 47, D1018–D1027. doi:10.1093/nar/gky1105

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *Am. J. Hum. Genet.* 85, 457–464. doi:10.1016/j.ajhg.2009.09.003

Kumar, S., Harmanci, A., Vytheeswaran, J., and Gerstein, M. B. (2020). SVFX: a Machine Learning Framework to Quantify the Pathogenicity of Structural Variants. *Genome Biol.* 21, 274. doi:10.1186/s13059-020-02178-x

Lindblom, A., and Robinson, P. N. (2011). Bioinformatics for Human Genetics: Promises and Challenges. *Hum. Mutat.* 32, 495–500. doi:10.1002/humu.21468

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* 161, 1012–1025. doi:10.1016/j.cell.2015.04.004

Luque, C., Luna, J. M., Luque, M., and Ventura, S. (2019). An Advanced Review on Text Mining in Medicine. *Wires Data Mining Knowl Discov.* 9, e1302. doi:10.1002/widm.1302

McGowan-Jordan, J., Hastings, R. J., and Moore, S. (2020). *An International System for Human Cytogenomic Nomenclature (Iscn 2020*. Basel, Switzerland: Karger Publishers. doi:10.1159/isbn.978-3-318-06867-2

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., et al. (2017). The Genomic Landscape of Balanced Cytogenetic Abnormalities Associated with Human Congenital Anomalies. *Nat. Genet.* 49, 36–45. doi:10.1038/ng.3720

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical Standards for the Interpretation and Reporting of Constitutional Copy Number Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257. doi:10.1038/s41436-021-01150-910.1038/s41436-019-0686-8

Salameh, T. J., Wang, X., Song, F., Zhang, B., Wright, S. M., Khunsriraksakul, C., et al. (2020). A Supervised Learning Framework for Chromatin Loop Detection in Genome-wide Contact Maps. *Nat. Commun.* 11, 1–12. doi:10.1038/s41467-020-17239-9

Sauerwald, N., Singhal, A., and Kingsford, C. (2020). Analysis of the Structural Variability of Topologically Associated Domains as Revealed by Hi-C. *NAR genombioinform* 2, lqz008. doi:10.1093/nargab/lqz008

Sharo, A. G., Hu, Z., Sunyaev, S. R., and Brenner, S. E. (2020). *StrVCTVRE: A Supervised Learning Method to Predict the Pathogenicity of Human Structural Variants*. Laurel Hollow, New York: bioRxiv. preprint. doi:10.1101/2020.05.15.097048

Silva, M., de Leeuw, N., Mann, K., Schuring-Blom, H., Morgan, S., Giardino, D., et al. (2019). European Guidelines for Constitutional Cytogenomic Analysis. *Eur. J. Hum. Genet.* 27, 1–16. doi:10.1038/s41431-018-0244-x

Smedley, D., and Robinson, P. N. (2015). Phenotype-driven Strategies for Exome Prioritization of Human Mendelian Disease Genes. *Genome Med.* 7, 81. doi:10.1186/s13073-015-0199-2

Spector, J. D., and Wiita, A. P. (2019). ClinTAD: a Tool for Copy Number Variant Interpretation in the Context of Topologically Associated Domains. *J. Hum. Genet.* 64, 437–443. doi:10.1038/s10038-019-0573-9

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural Variation in the 3D Genome. *Nat. Rev. Genet.* 19, 453–467. doi:10.1038/s41576-018-0007-0

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease. *Nat. Rev. Genet.* 14, 125–138. doi:10.1038/nrg3373

Zepeda-Mendoza, C. J., Ibn-Salem, J., Kammin, T., Harris, D. J., Rita, D., Gripp, K. W., et al. (2017). Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. *Am. J. Hum. Genet.* 101, 206–217. doi:10.1016/j.ajhg.2017.06.011

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf. Fusion* 50, 71–91. doi:10.1016/j.inffus.2018.09.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

CORRECTION
published: 25 February 2022
doi: 10.3389/fgene.2022.868306

Corrigendum: SVInterpreter: A Comprehensive Topologically Associated Domain Based Clinical Outcome Prediction Tool for Balanced and Unbalanced Structural Variants

Joana Fino[1]*, *Bárbara Marques*[1], *Zirui Dong*[2,3,4] and *Dezso David*[1]*

[1]Department of Human Genetics, National Health Institute Doutor Ricardo Jorge, Lisbon, Portugal, [2]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China, [3]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China, [4]Hong Kong Hub of Pediatric Excellence, The Chinese University of Hong Kong, Hong Kong, China

Keywords: SVInterpreter, bioinformatic web-tool, clinical outcome prediction, balanced structural variants, copy number variants, topologically associated domains, phenotypic comparison

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Joana Fino
joana.fino@insa.min-saude.pt
Dezso David
dezso.david@insa.min-saude.pt

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *02 February 2022*
Accepted: *07 February 2022*
Published: *25 February 2022*

Citation:
Fino J, Marques B, Dong Z and David D (2022) Corrigendum: SVInterpreter: A Comprehensive Topologically Associated Domain Based Clinical Outcome Prediction Tool for Balanced and Unbalanced Structural Variants. Front. Genet. 13:868306. doi: 10.3389/fgene.2022.868306

A Corrigendum on

SVInterpreter: A Comprehensive Topologically Associated Domain-Based Clinical Outcome Prediction Tool for Balanced and Unbalanced Structural Variants
by Fino, J., Marques, B., Dong, Z., and David, D. (2021). Front. Genet. 12:757170. doi: 10.3389/fgene.2021.757170

In the original article, there was an error. The comparison with AnnotSV was made with its first version, which is currently out of date. Any remark made on this tool should not be considered.

The versions, commits and release dates of the tools used for comparison were not indicated.

Therefore, a correction has been made to **Discussion**, paragraph 11:

"Comparatively with other recent tools that support the evaluation of SVs, such as position_effect (commit: fced2c49, 13 June 2017), AnnotSV (Version 1.0, 21 December 2017) and ClinTAD (commit: 09b4925fb, 18 September 2019), SVInterpreter seems to be more comprehensive (Zepeda-Mendoza et al., 2017; Geoffroy et al., 2018; Spector and Wiita, 2019). First, SVInterpreter showed to be the one that allows more customization and adjustments, since, for example, AnnotSV and ClinTAD only work with one genome version, and ClinTAD only uses TAD boundaries of human embryonic stem cell data. Then, SVInterpreter shows a broader view of the affected regions, accounting for both gene disruption and position effects: AnnotSV is focused on the identification of genes directly affected by a breakpoint, and position_effect was designed to identify candidate genes essentially from position effect events. In regard to phenotypic comparison, as AnnotSV does not perform any, and ClinTAD is limited to the full HPO term overlap, position_effect is the only one with a similar functionality. Also, SVInterpreter is the one that retrieves the most information, including the position effect important data, GeneHancer cluster of interactions and chromatin loops, phenotypic data from DDG2P and clinGen, Gene-phenotype/disease associations in animal models, and GWAS data. Therefore, the existence of overlooked information by position_effect and AnnotSV, as shown

in DGAP107, may contribute to limited results, biased candidate gene prioritization, and the need of additional resources."

The authors apologize for the errors and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fgene.2021.731815

Prenatal Diagnosis and Genetic Analysis of 21q21.1–q21.2 Aberrations in Seven Chinese Pedigrees

Huamei Hu†, Rong Zhang†, Yongyi Ma, Yanmei Luo, Yan Pan, Juchun Xu, Lupin Jiang and Dan Wang**

Department of Gynecology and Obstetrics, Southwest Hospital, Third Military Medical University (Army Medical University), Chongqing, China

Background: Chromosomal aberrations contribute to human phenotypic diversity and disease susceptibility, but it is difficult to assess their pathogenic effects in the clinic. Therefore, it is of great value to report new cases of chromosomal aberrations associated with normal phenotypes or clinical abnormalities.

Methods: This was a retrospective analysis of seven pedigrees that carried 21q21.1–q21.2 aberrations. G-banding and single-nucleotide polymorphism array techniques were used to analyze chromosomal karyotypes and copy number variations in the fetuses and their family members.

Results: All fetuses and their family members showed normal karyotypes in seven pedigrees. Here, it was revealed that six fetuses carried maternally inherited 21q21.1–q21.2 duplications, ranging from 1 to 2.7 Mb, but none of the mothers had an abnormal phenotype. In one fetus, an 8.7 Mb deletion of 21q21.1–q21.2 was found. An analysis of the pedigree showed that the deletion was also observed in the mother, brother, and maternal grandmother, but no abnormal phenotypes were found.

Conclusion: This study identified 21q21.1–q21.2 aberrations in Chinese pedigrees. The carriers of 21q21.1–q21.2 duplications had no clinical consequences based on their phenotypes, and the 21q21.1–q21.2 deletion was transmitted through three generations of normal individuals. This provides benign clinical evidence for pathogenic assessment of 21q21.1–q21.2 duplication and deletion, which was considered a variant of uncertain significance and a likely pathogenic variant in previous reports.

Keywords: 21q21.1–21.2 duplication, 21q21.1–21.2 deletion, SNP array, NCAM2, prenatal diagnosis

OPEN ACCESS

Edited by:
Claudia Gonzaga-Jauregui, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Shabeesh Balan, RIKEN Center for Brain Science (CBS), Japan
Cinthya Zepeda Mendoza, ARUP Laboratories, United States

****Correspondence:***
Lupin Jiang
lpjiangcqzd@163.com
Dan Wang
wang_swh@sina.com

†These authors share first authorship

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *28 June 2021*
Accepted: *29 November 2021*
Published: *21 December 2021*

Citation:
Hu H, Zhang R, Ma Y, Luo Y, Pan Y, Xu J, Jiang L and Wang D (2021) Prenatal Diagnosis and Genetic Analysis of 21q21.1–q21.2 Aberrations in Seven Chinese Pedigrees. Front. Genet. 12:731815. doi: 10.3389/fgene.2021.731815

INTRODUCTION

To date, the 21q21 duplication and deletion have not been included in the known pathogenicity syndromes. However, early in the 1980s, Park et al. reported that partial trisomy of chromosome 21, comprising the *NCAM2* gene, results in intellectual disability but does not cause other phenotypes of Down syndrome (DS) (Park et al., 1987). Haldeman-Englert et al. revealed that a boy who was evaluated for autistic features, significant speech delay, and poor social interactions carried a *de novo* 8.8 Mb 21q21.1–q21.3 deletion involving

the *NCAM2* gene (Haldeman-Englert et al., 2009). In addition, three cases of neurodevelopmental disorders were reported, with clinical phenotypic abnormalities including global developmental delay, behavioral disorders, and impaired social interactions. All of them carried 21q21.1–21.2 deletions involving *NCAM2* (Petit et al., 2015). Another case report revealed that a boy with autism spectrum disorder and macrocephaly carried a 1.6 Mb deletion of 21q21.1–21.2, containing the *NCAM2* gene, but no other functional gene (Scholz et al., 2016). Previously, *NCAM2* was proposed as a candidate gene for autism based on genome-wide association studies (Hussman et al., 2011). Duplications, deletions, and single-nucleotide polymorphisms of the *NCAM2* gene have been found in individuals with intellectual disabilities or autism, and these studies suggest that *NCAM2* might play a role in neurodevelopmental disorders.

In our study, the carriers of 21q21.1–21.2 duplications in six pedigrees (the region of one pedigree contained *NCAM2*) showed normal phenotypes. We further identified a rare 8.7 Mb deletion of 21q21.1–21.2 containing *NCAM2*, which had been transmitted through three generations of normal individuals. These findings provide benign evidence, which is important for accurate genetic counseling on 21q21.1–21.2 aberrations in prenatal diagnosis.

MATERIALS AND METHODS

Subjects

A retrospective study was performed from January 2016 to December 2020. In total, seven cases carrying 21q21.1–21.2 deletions or duplications were selected from 11,867 pregnant women who had indications (e.g., abnormal non-invasive prenatal testing (NIPT) or fetal imaging) and underwent invasive diagnostic testing *via* amniocentesis or cordocentesis at the Prenatal Diagnosis Center of Obstetrics and Gynecology, Southwest Hospital. Informed consent for invasive prenatal diagnosis was obtained from the parents before detection. This research was approved by the Ethics Committee of Southwest Hospital, Third Military Medical University (Army Medical University). Six fetuses from six unrelated Chinese families were identified as carrying 21q21.1–q21.2 duplications, as their pedigree verification information was collected, and they were classified as pedigrees 1, 2, 3, 4, 5, and 6. In addition, a fetus carrying a 21q21.1–q21.2 deletion and its family members were labeled as pedigree 7. The pregnant women in these seven pedigrees did not have pregnancy complications and denied any related family history.

Pedigrees 1–6: The maternal age at the time of amniocentesis was between 25 and 32 years, and a gestational age ranging from 18 + 2 to 25 weeks. The pregnant woman in pedigree 3 chose amniocentesis because of pulmonary sequestration of the fetus examined by ultrasound. The others all chose amniocentesis because NIPT screening showed an abnormality on chromosome 21.

Pedigree 7: A 22-year-old woman (gravida 4, para 1) was subjected to cordocentesis at 26 + 5 gestational weeks because the bilateral ventricle of the fetus had widened, as tested by ultrasound examination (left: 14 mm, right: 14 mm).

Chromosomal Karyotyping

Approximately 0.5 ml of each peripheral blood sample and 0.4 ml of each umbilical cord blood sample were inoculated into a T-cell culture medium (BAIDI, China) and incubated at 37°C, for 3 days. Approximately 20 ml of each amniotic fluid sample was inoculated into an amniotic fluid medium (BIO-AMF™-2, BI, China) and incubated at 37°C, with 5% CO_2, for 7–10 days. Chromosomal karyotyping was performed according to the standard protocol using G-banding at a 400-banded (amniotic fluid samples) or 550-banded (blood samples) resolution, and karyotypes were described according to the International System for Human Cytogenetic Nomenclature 2016 (ISCN 2016) criteria (Stevens-Kroef et al., 2017).

Single-Nucleotide Polymorphisms Array Analysis

Uncultured amniotic fluid samples (10 ml per fetus), umbilical cord blood samples (600 μL per fetus), and peripheral blood (600 μL per person) of the pedigree members were collected, and DNA was extracted using the TIANamp Genomic DNA Kit (TIANGEN, China). The Infinium Global Screening Array (Illumina, San Diego, CA, United States) contains approximately 700,000 genome-wide tag SNPs. Genomic DNA was hybridized to an Infinium Global Screening Array as reported previously (Srebniak et al., 2011). The array was scanned with the iScan array scanning system (Illumina, San Diego, CA, United States). Molecular karyotype analysis was performed using GenomeStudio V2011.1 software (Illumina, San Diego, CA, United States), which was used for annotation. Copy number variations (CNVs) that were larger than 100 kb or affected more than 50 markers were considered and were annotated based on the GRCh37 (hg19) genome. CNVs were evaluated according to the guidelines (Richards et al., 2015; Riggs et al., 2020), scientific literature, and publicly available databases as follows: DGV (http://dgv.tcag.ca/dgv/app/home), OMIM (http://www.ncbi.nlm.nih.gov/omim), gnomAD (http://gnomad-sg.org/), DECIPHER (http://decipher.sanger.ac.uk), dbVar (http://www.ncbi.nlm.nih.gov/dbvar), ClinVar(http://www.ncbi.nlm.nih.gov/clinvar), ClinGene (https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/), and Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/). Benign or likely benign CNVs were not reported.

Prenatal and Postnatal Follow-Up Assessment

Ultrasound results of the second and third trimesters of pregnancy were collected. Postnatal clinical follow-up

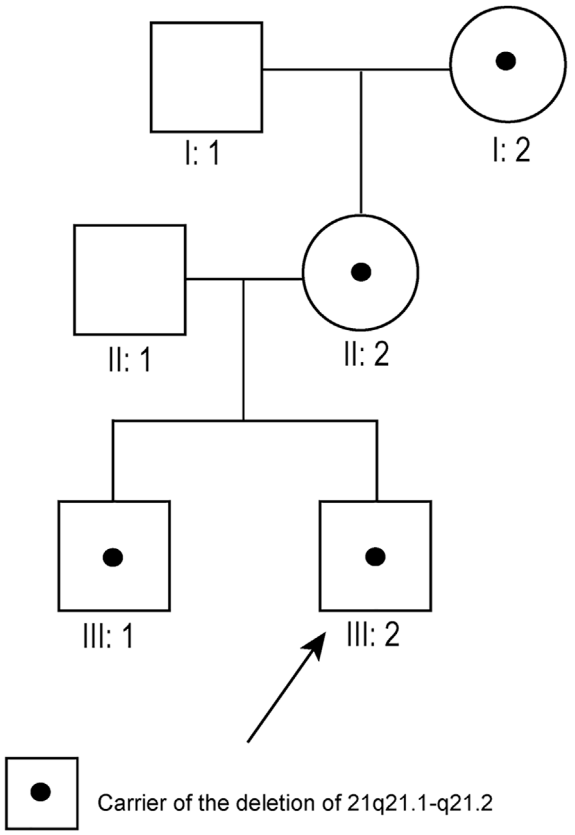

FIGURE 1 | Pedigree diagram of pedigree 7 (arrow indicates the fetus). The fetus's maternal grandmother, mother, and brother all carried the 21q21.1–21.2 deletion.

assessments *via* telephone were performed from 6 months to 3 years after birth. After obtaining their parents' informed consent, the child's healthcare data were collected to assess developmental details. General child healthcare was carried out by professional doctors according to the World Health Organization's physical and mental development table for infants aged 0–3 years. Child healthcare in tertiary hospitals was performed according to the Denver Developmental Screening Test (Wijedasa, 2012).

RESULTS

Analysis of the Chromosomal Karyotype of the Fetuses and Family Members

Amniotic fluid samples, umbilical cord blood samples, and peripheral blood samples of family members were subjected to conventional karyotyping because balanced rearrangements will escape SNP array detection (Levy and Wapner, 2018).

Pedigrees 1–6: The conventional G-banding analysis showed that the karyotypes of the fetuses and their parents were normal.

Pedigree 7: Although typical karyotypic analysis by G-banding might be able to delineate chromosomal aberrations greater than 5–10 Mb in size (Shaffer and Bejjani, 2004), the 8.7 Mb deletion of 21q21.1–21.2 was not identified in our study. Owing to the small size of chromosome 21, the deletion region could only be identified above 700-banded resolution, whereas the conventional amniotic karyotyping could only achieve 550-banded resolution at most. Therefore, the fetus (III:2, **Figure 1**), its elder brother (III:1, **Figure 1**), mother (II:2, **Figure 1**), and maternal grandmother (I:2, **Figure 1**) had normal karyotypes. Its father (II:1, **Figure 1**) and maternal grandfather (I:1, **Figure 1**) also had normal karyotypes.

Verification of SNP Array Results of the Fetuses and Family Members

Pedigrees 1–6: The fetuses of six unrelated pedigrees carried the 21q21.1–q21.2 duplications, which were inherited from their mothers, and with the same coordinates and lengths as those of their mothers. The smallest duplication length was 1 Mb (chr21:20,195,657–21,199,532, hg19 build), and the largest was 2.7 Mb (chr21:23,573,580–26,310,725, hg19 build). The duplicated regions in pedigrees 1, 2, 3, 5, and 6 did not contain any protein-coding gene, and only the duplication in pedigree 4 contained the *NCAM2* gene (**Table 1**; **Figure 2**; **Supplementary Figure S1**). All fathers were also tested with SNP arrays, and the results were negative.

Pedigree 7: SNP array results showed that the fetus (III:2, **Figure 1**) carried an 8.7 Mb deletion of chromosome 21q21.1–21.2 (chr21:16,767,983–25,441,375, hg19 build; **Figure 2**; **Supplementary Figure S1**), and the pedigree analysis found that the CNV was inherited from the mother with a normal phenotype (II:2, **Figure 1**). To obtain more genetic evidence, the elder brother and maternal grandparents of the fetus were also tested with SNP arrays. The extended analysis of the pedigree revealed that two other healthy members also carried the deletion, the elder brother, 3 years of age (III:1, **Figure 1**), and maternal grandmother, 41 years of age (I:2, **Figure 1**). In addition, the results of others (II:1 and I:1, **Figure 1**) in this pedigree were normal. Otherwise, the elder brother was found to carry another deletion, which was located on 5p15.33 (chr5:19:38,139–1,124,703, hg19 build) and was proven to be a *de novo* variation of uncertain significance (VUS) mutation. No other significant CNVs were found among the seven pedigrees.

Prenatal and Postnatal Follow-Up Assessment

Pedigrees 1–6: No abnormalities were found during the second and third trimesters of pregnancy, except for the fetuses in pedigree 3 with pulmonary isolation. Three boys (fetuses of pedigrees 1, 3, and 6) and three girls (fetuses of pedigrees 2, 4, and 5) were born at full-term delivery. Now, the youngest individual is 2.5 years of age, the oldest is 3.5 years of age, and none of them show signs of developmental delay or intellectual disability based on child's healthcare examination (**Table 2**).

TABLE 1 | Chromosomal aberrations of the fetuses in seven pedigrees.

Pedigree	Location (hg19)	Size (Mb)	Aberration type	Karyotype	Protein-coding gene content	Inheritance
Pedigree 1	chr21: 20,195,657–21,199,532	1	Duplication	46,XY	—	mat
Pedigree 2	chr21:23,573,580–24,697,989	1.1	Duplication	46,XX	—	mat
Pedigree 3	chr21: 23,288,789–25,106,099	1.8	Duplication	46,XY	—	mat
Pedigree 4	chr21: 22,734,409–25,148,429	2.4	Duplication	46,XX	NCAM2	mat
Pedigree 5	chr21:23,272,300–25,104,945	1.8	Duplication	46,XX	—	mat
Pedigree 6	chr21: 23,573,580–26,310,725	2.7	Duplication	46,XY	—	mat
Pedigree 7	chr21: 16,767,983–25,441,375	8.7	Deletion	46,XY	BTG3, C21orf91, CHODL, CXADR, NCAM2, TMPRSS15, USP25	mat

mat, Inherited from the mother.

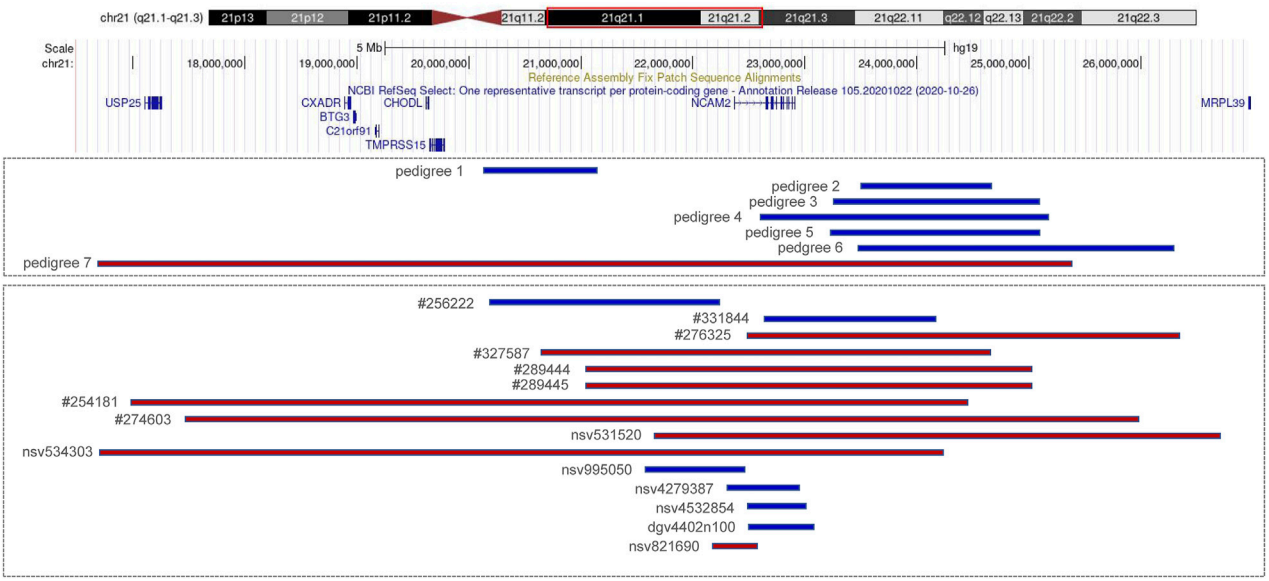

FIGURE 2 | Copy number variations (CNVs) of 21q21.1–q21.2 (blue indicates duplication and red indicates deletion). Pedigrees 1–7 are from our cases; #256222, #331844, #276325, #327587, #289444, #289445, #254181, and #274603 are recorded in Decipher; nsv995050, nsv531520, and nsv534303 are recorded in dbVar; nsv4279387and nsv4532854 are recorded in gnomAD; dgv4402n100 and nsv821690 are recorded in DGV.

Pedigree 7: Ultrasound and MRI examinations were performed at 32 weeks of gestation, and no further widening of the lateral ventricles was observed in both examinations (left: 14 mm, right: 14 mm, examined by ultrasound; left: 12.5 mm, right: 13.6 mm, examined by MRI). After genetic counseling, the pregnant woman and her husband chose to continue the pregnancy. A healthy boy was born by natural delivery at 39 gestational weeks, without any special facial features. The boy is 8 months old currently and does not have any abnormal phenotypes; moreover, the details of the child's healthcare examination were normal (**Table 2**). His elder brother is 3 years of age and also does not show developmental delay or intellectual disability.

DISCUSSION

We reported seven fetuses carrying familial 21q21.1–21.2 aberrations. The fetuses of pedigrees 1, 2, 3, 4, 5, and 6 all carried a maternally inherited 21q21.1–q21.2 duplication ranging from 1 to 2.7 Mb (**Table 1**). There are few reports about whether the duplication of this region is benign or pathogenic. In public databases such as DGV, gnomAD, DECIPHER, dbVar, and ClinVar, several significant records of 21q21.1–q21.2 duplications were found, which partially overlapped with our cases, and they were analyzed (**Figure 2**; **Table 3**). Only three records were recorded in DGV (dgv4402n100) and the gnomAD database (nsv4279387, nsv4532854), but the frequency of copy number gains in the general population had not been described. In addition, a case with intellectual disability had been reported (DECIPHER, #256222), but there is no description about its inheritance and classification of pathogenicity. Another case (DECIPHER, #331844) was described as a likely benign variant with no abnormalities other than increased nuchal translucency. A VUS variant (nsv995050) was found in the dbVar and the CinVar database, and the major phenotype was

TABLE 2 | Clinical follow-up evaluation of 7 fetuses.

Fetus	Sex	At birth		Birth with other defects	Routine child healthcare (6 m, 12 m, 24 m)	Child healthcare by DDST		At study		
		Weight (kg) (%)	Length (cm) (%)			18 m	24 m	Age (m)	Weight (kg) (%)	Height (cm) (%)
1	Male	3.3 (46)	51 (72.2)	_	Pass	NA	NA	42	15 (42.5)	102 (70.6)
2	Female	3.3 (56)	50 (67.7)	_	Pass	Pass	NA	41	15.5 (63.2)	101 (73.4)
3	Male	3.05 (26.8)	50 (52.4)	Pulmonary sequestration	Pass	NA	NA	41	14.5 (33.7)	98 (34.3)
4	Female	3.3 (56)	50 (67.7)	_	Pass	NA	NA	37	14 (50.3)	95 (44.4)
5	Female	3.5 (71.6)	51 (83.9)	_	Pass	NA	NA	33	14 (67.3)	93 (53.5)
6	Male	3 (23.3)	48 (15.9)	_	Pass	NA	Pass	32	14 (60.0)	94 (60.3)
7	Male	2.35 (1.1)	48 (15.9)	_	Pass	NA	NA	8	8.5 (49.9)	70 (48.1)

m, months; NA, not available; percentile refers to WHO, Growth Charts.

TABLE 3 | Summary of patients harboring 21q21.1–q21.2 aberrations.

Patient database	Location (hg19)	Type	Size (Mb)	Protein-coding gene	Inheritance	Pathogenicity	Phenotypes
dbVar#nsv995050	chr21: 21,601,231–22,573,421	Duplication	972 kb	NCAM2	Unknown	VUS	Developmental delay and/or other significant developmental or morphological phenotypes
Decipher#256222	chr21: 20,063,479–22,274,948	Duplication	2.2 Mb	—	Inherited from a normal parent	/	Intellectual disability
Decipher#331844	chr21: 22,782,651–24,339,651	Duplication	1.6 Mb	NCAM2	Inherited from the father	Likely benign	Increased nuchal translucency
Decipher#276325	chr21: 22,434,634–26,315,434	Deletion	3.8 Mb	NCAM2	Inherited from the affected mother	/	Behavioral abnormality, delayed speech and language development
Decipher#327587	chr21: 20,746,935–24,683,731	Deletion	3.9 Mb	NCAM2	Unknown	/	Overweight, recurrent otitis media, sandal gap, abnormal oral glucose tolerance, acanthosis nigricans, generalized non-motor (absence) seizure, generalized-onset seizure, simple febrile seizure, status epilepticus
Decipher#289444	chr21: 21,044,211–25,051,262	Deletion	4.0 Mb	NCAM2	Unknown	VUS	Abnormal facial shape, short stature, intellectual disability
Decipher#289445	chr21: 21,044,211–25,051,262	Deletion	4.0 Mb	NCAM2	Unknown	VUS	Intellectual disability
dbVar#nsv531520	chr21: 21,699,837–26,771,050	Deletion	5.1 Mb	NCAM2	Inherited from the mother	Pathogenic	Abnormality of the skeletal system, cleft palate, global developmental delay
dbVar#nsv534303	chr21: 16,714,035–24,198,636	Deletion	7.5 Mb	BTG3, C21orf91, CHODL, CXADR NCAM2, TMPRSS15, USP25	Unknown	Pathogenic	Oral cleft
Decipher#254181	chr21: 16,992,255–24,898,237	Deletion	7.9 Mb	BTG3, C21orf91, CHODL, CXADR NCAM2, TMPRSS15, USP25	Inherited from the mildly affected father	/	Epicanthic folds, long and flat philtrum, high palate, low-set ears, global developmental delay, behavioral disorder
Decipher#274603	chr21: 17,451,703–25,948,154	Deletion	8.5 Mb	BTG3, C21orf91, CHODL, CXADR NCAM2, TMPRSS15	Unknown	/	Almond-shaped eyes, hypotonia and joint laxity, global developmental delay, impaired social interactions

/, not provided by database.

developmental delay. Therefore, duplication of this region was considered a VUS in previous reports. The 21q21.1–q21.2 duplication in our study contained only one protein-coding gene, *NCAM2*, which is not predicted to be a triplosensitive gene. Jin et al. reported that a fetus and its mother both carried a 6.7 Mb duplication of 21q21.1–q21.2 including *NCAM2*, but

the phenotype was normal. The region was significantly larger than that of our cases (ranging from position 18,981,715 to 25,707,009). This study also provided benign clinical evidence for partial duplication of 21q21.1–q21.2 in prenatal diagnosis (Jin et al., 2021). The rarity of gene content might be a major factor that makes these CNV gains, shown in this study, seem benign. In addition, position effects are one of the molecular mechanisms responsible for CNVs caused by genomic rearrangements resulting in phenotypes (Zhang et al., 2009). Whether the pathogenicity can change if the chromosomal duplication is not in its original position but translocates to another chromosome requires further study.

The pathogenicity of the copy-number gain and copy-number loss might be quite different in the same region. A copy-number loss record involving *NCAM2* was found in the DGV database (nsv821690, **Figure 2**), but the frequency in the general population had not been described. Several cases of phenotypic abnormalities related to 21q21.1–21.2 deletions, and those highly overlapping with our case, were found in the public databases (**Figure 2**; **Table 3**). The sizes of these regions were approximately 4 Mb or greater. Three cases (Decipher#276325, Decipher#254181, and Decipher#274603), provided by Petit et al., their inheritance, and phenotypes had been described in detail (Petit et al., 2015). In five cases (**Table 3**), the deletion involved only *NCAM2*, and patients had abnormal phenotypes including those concerning intellectual disability, developmental delay, abnormal facial shape, and seizures. In two cases (nsv531520 and 534303), one was reported to have an abnormality of the skeletal system, cleft palate, and global developmental delay, and another had an oral cleft. They were all described as pathogenic, of which only one CNV (nsv531520) was inherited from the mother, but no information was provided about her phenotype. In summary, the 21q21.1–21.2 deletion was identified as likely pathogenic in previous reports. However, the 21q21.1–21.2 deletion in our study was not found to be associated with phenotypic consequences.

Previous and recent studies have revealed the important role of *NCAM2* in neurodevelopment (Sheng et al., 2019). In addition to *NCAM2*, there are other genes associated with clinical phenotypes in this region that deserve further analysis. This region contains seven protein-coding genes, namely, *BTG3*, *C21orf91*, *CHODL*, *CXADR*, *NCAM2*, *TMPRSS15*, and *USP25*. None of them are predicted to be haploinsufficient. Except for *C21orf91*, others are Online Mendelian Inheritance in Man (OMIM) genes. *BTG3* is a novel member of the PC3/BTG/TOB family of growth inhibitory genes (Yoshida et al., 1998) and is expressed in various human tissues. Further analysis in mice revealed that BTG3 is highly expressed in the ventricular zone of the developing central nervous system. *C21orf91* was described as having a role in defective DS neurogenesis (Li et al., 2016) and plays an important role in accurate oligodendroglial differentiation, affecting maturation capacity and axon myelination (Reiche et al., 2021). *CHODL* is a type-1A integral membrane protein and is preferentially expressed in the skeletal muscle, testis, brain, and lung (Weng et al., 2003). A recent study showed that the absence of *CHODL* leads to anatomical and functional defects of the neuromuscular synapse (Oprişoreanu et al., 2019). *CXADR* is expressed at increased levels during brain development and is considered a candidate gene in children with autism (Iourov et al., 2010). Patients carrying 21q21.1 microduplication (from 0.4 to 0.1 Mb) involving the *CXADR* gene have abnormal phenotypes such as developmental delay and intellectual disability (Li et al., 2018). *TMPRSS15* is a morbid gene, and loss-of-function variants are responsible for enterokinase deficiency (Wang et al., 2020). It is well known that *USP25* is widely expressed in the central nervous system and peripheral nervous system (Bosch-Comas et al., 2006). Recent studies have shown that USP25 plays a key role in microglial homeostasis reprogramming in Alzheimer's disease and DS (Zheng et al., 2021). Therefore, *BTG3*, *C21orf91*, *CXADR*, *NCAM2*, and *USP25* might be involved in phenotypes based on their presumed or known biological functions.

The mechanism through which aberrations do not produce clinical phenotypes is unclear. Genetic counseling in this region has become challenging, owing to limited or conflicting associations with clinical phenotypes described in the published literature and public databases. Accordingly, our study provides benign evidence for accurate genetic counseling of 21q21.1–21.2 aberrations based on prenatal diagnosis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena/browser/view/PRJEB47787.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Southwest Hospital, Third Military Medical University (Army Medical University). The number is (B)KY2021023. Written informed consent to participate in this study was provided by the participant's legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

HH designed the study and wrote the article, RZ performed chromosome analysis, YM performed sample processing and data analysis, YL performed data statistic, YP performed follow-up, JX performed cell culture, LJ proofread the paper and acquired funding, and DW performed project administration.

FUNDING

This work was supported by the grant from the National Natural Science Foundation of China (No. 81971369 to LJ.

ACKNOWLEDGMENTS

We thank Dr. Limeng Dai for his advice on the article and grammar modification.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.731815/full#supplementary-material

Supplementary Figure 1 | Single-nucleotide polymorphism (SNP) array results for the fetuses. Pedigrees 1–6: 21q21.1–21.2 duplications, with fragment sizes in the following order: 1 Mb, 1.1 Mb, 1.8 Mb, 2.4 Mb, 1.8 Mb, and 2.7 Mb. Pedigree 7: 21q21.1–21.2 deletion; the fragment size was 8.7 Mb.

REFERENCES

Bosch-Comas, A., Lindsten, K., Gonzàlez-Duarte, R., Masucci, M. G., and Marfany, G. (2006). The Ubiquitin-specific Protease USP25 Interacts with Three Sarcomeric Proteins. *Cell. Mol. Life Sci.* 63, 723–734. doi:10.1007/s00018-005-5533-1

Haldeman-Englert, C. R., Chapman, K. A., Kruger, H., Geiger, E. A., Mcdonald-mcginn, D. M., Rappaport, E., et al. (2009). A De Novo 8.8-Mb Deletion of 21q21.1-q21.3 in an Autistic Male with a Complex Rearrangement Involving Chromosomes 6, 10, and 21. *Am. J. Med. Genet.* 152A (10), 196–202. doi:10.1002/ajmg.a.33176

Hussman, J. P., Chung, R.-H., Griswold, A. J., Jaworski, J. M., Salyakina, D., Ma, D., et al. (2011). A Noise-Reduction GWAS Analysis Implicates Altered Regulation of Neurite Outgrowth and Guidance in Autism. *Mol. Autism* 2, 1. doi:10.1186/2040-2392-2-1

Iourov, I. Y., Vorsanova, S. G., Saprina, E. A., and Yurov, Y. B. (2010). Identification of Candidate Genes of Autism on the Basis of Molecular Cytogenetic and In Silico Studies of the Genome Organization of Chromosomal Regions Involved in Unbalanced Rearrangements. *Russ. J. Genet.* 46, 1190–1193. doi:10.1134/S102279541010011X

Jin, C., Gu, Z., Jiang, X., Yu, P., and Xu, T. (2021). A Prenatal Diagnosis Case of Partial Duplication 21q21.1-q21.2 with normal Phenotype Maternally Inherited. *BMC Med. Genomics* 14, 4–8. doi:10.1186/s12920-021-01013-x

Levy, B., and Wapner, R. (2018). Prenatal Diagnosis by Chromosomal Microarray Analysis. *Fertil. Sterility* 109, 201–212. doi:10.1016/j.fertnstert.2018.01.005

Li, S. S., Qu, Z., Haas, M., Ngo, L., Heo, Y. J., Kang, H. J., et al. (2016). The HSA21 Gene EURL/C21ORF91 Controls Neurogenesis within the Cerebral Cortex and Is Implicated in the Pathogenesis of Down Syndrome. *Sci. Rep.* 6, 1–14. doi:10.1038/srep29514

Li, W., Wang, X., and Li, S. (2018). Investigation of Copy Number Variations on Chromosome 21 Detected by Comparative Genomic Hybridization (CGH) Microarray in Patients with Congenital Anomalies. *Mol. Cytogenet.* 11, 1–8. doi:10.1186/s13039-018-0391-3

Oprişoreanu, A.-M., Smith, H. L., Arya, S., Webster, R., Zhong, Z., Eaton-Hart, C., et al. (2019). Interaction of Axonal Chondrolectin with Collagen XIXa1 Is Necessary for Precise Neuromuscular Junction Formation. *Cel Rep.* 29, 1082–1098. doi:10.1016/j.celrep.2019.09.033

Park, J. P., Wurster-Hill, D. H., Andrews, P. A., Cooley, W. C., and Graham, J. M. (1987). Free Proximal Trisomy 21 without the Down Syndrome. *Clin. Genet.* 32, 342–348. doi:10.1111/j.1399-0004.1987.tb03299.x

Petit, F., Plessis, G., Decamp, M., Cuisset, J.-M., Blyth, M., Pendlebury, M., et al. (2015). 21q21 Deletion Involving NCAM2: Report of 3 Cases with Neurodevelopmental Disorders. *Eur. J. Med. Genet.* 58, 44–46. doi:10.1016/j.ejmg.2014.11.004

Reiche, L., Göttle, P., Lane, L., Duek, P., Park, M., Azim, K., et al. (2021). C21orf91 Regulates Oligodendroglial Precursor Cell Fate-A Switch in the Glial Lineage? *Front. Cel. Neurosci.* 15, 1–18. doi:10.3389/fncel.2021.653075

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi:10.1038/gim.2015.30

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical Standards for the Interpretation and Reporting of Constitutional Copy-Number Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257. doi:10.1038/s41436-019-0686-8

Scholz, C., Steinemann, D., Mälzer, M., Roy, M., Arslan-Kirchner, M., Illig, T., et al. (2016). NCAM2 Deletion in a Boy with Macrocephaly and Autism: Cause, Association or Predisposition? *Eur. J. Med. Genet.* 59, 493–498. doi:10.1016/j.ejmg.2016.08.006

Shaffer, L. G., and Bejjani, B. A. (2004). A Cytogeneticist's Perspective on Genomic Microarrays. *Hum. Reprod. Update* 10, 221–226. doi:10.1093/humupd/dmh022

Sheng, L., Leshchyns'ka, I., and Sytnyk, V. (2019). Neural Cell Adhesion Molecule 2 (NCAM2)-Induced C-src-dependent Propagation of Submembrane Ca2+ Spikes along Dendrites Inhibits Synapse Maturation. *Cereb. Cortex* 29, 1439–1459. doi:10.1093/cercor/bhy041

Srebniak, M., Boter, M., Oudesluijs, G., Joosten, M., Govaerts, L., Van Opstal, D., et al. (2011). Application of SNP Array for Rapid Prenatal Diagnosis: Implementation, Genetic Counselling and Diagnostic Flow. *Eur. J. Hum. Genet.* 19, 1230–1237. doi:10.1038/ejhg.2011.119

Stevens-Kroef, M., Simons, A., Rack, K., and Hastings, R. J. (2017). Cytogenetic Nomenclature and Reporting. *Methods Mol. Biol.* 1541, 303–309. doi:10.1007/978-1-4939-6703-2_24

Wang, L., Zhang, D., Fan, C., Zhou, X., Liu, Z., Zheng, B., et al. (2020). Novel Compound Heterozygous TMPRSS15 Gene Variants Cause Enterokinase Deficiency. *Front. Genet.* 11, 1–9. doi:10.3389/fgene.2020.538778

Weng, L., Hübner, R., Claessens, A., Smits, P., Wauters, J., Tylzanowski, P., et al. (2003). Isolation and Characterization of Chondrolectin (Chodl), a Novel C-type Lectin Predominantly Expressed in Muscle Cells. *Gene* 308, 21–29. doi:10.1016/S0378-1119(03)00425-6

Wijedasa, D. (2012). Developmental Screening in Context: Adaptation and Standardization of the Denver Developmental Screening Test-II (DDST-II) for Sri Lankan Children. *Child. Care Health Dev.* 38, 889–899. doi:10.1111/j.1365-2214.2011.01332.x

Yoshida, Y., Matsuda, S., Ikematsu, N., Kawamura-Tsuzuku, J., Inazawa, J., Umemori, H., et al. (1998). ANA, a Novel Member of Tob/BTG1 Family, Is Expressed in the Ventricular Zone of the Developing central Nervous System. *Oncogene* 16, 2687–2693. doi:10.1038/sj.onc.1201805

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annu. Rev. Genom. Hum. Genet.* 10, 451–481. doi:10.1146/annurev.genom.9.081307.164217

Zheng, Q., Li, G., Wang, S., Zhou, Y., Liu, K., Gao, Y., et al. (2021). Trisomy 21-induced Dysregulation of Microglial Homeostasis in Alzheimer's

Brains Is Mediated by USP25. *Sci. Adv.* 7, 1–13. doi:10.1126/sciadv.abe1340

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 February 2022
doi: 10.3389/fgene.2022.820209

Profile of Chromosomal Alterations, Chromosomal Instability and Clonal Heterogeneity in Colombian Farmers Exposed to Pesticides

María Paula Meléndez-Flórez[1], Duvan Sebastián Valbuena[1], Sebastián Cepeda[1], Nelson Rangel[2], Maribel Forero-Castro[1], María Martínez-Agüero[3] and Milena Rondón-Lagos[1]**

[1]School of Biological Sciences, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia, [2]Departamento de Nutrición y Bioquímica, Facultad de Ciencias, Pontificia Universidad Javeriana, Bogotá, Colombia, [3]Centro de Investigaciones en Microbiología y Biotecnología-UR (CIMBIUR), Facultad de Ciencias Naturales, Universidad del Rosario, Bogotá, Colombia

OPEN ACCESS

Edited by:
Claudia Gonzaga-Jauregui, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Fabio Caradonna, University of Palermo, Italy
Grace Chappell, ToxStrategies, Inc., United States

***Correspondence:**
Milena Rondón-Lagos
sandra.rondon01@uptc.edu.co
María Martínez-Agüero
maria.martinez@urosario.edu.co

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *22 November 2021*
Accepted: *28 January 2022*
Published: *24 February 2022*

Citation:
Meléndez-Flórez MP, Valbuena DS, Cepeda S, Rangel N, Forero-Castro M, Martínez-Agüero M and Rondón-Lagos M (2022) Profile of Chromosomal Alterations, Chromosomal Instability and Clonal Heterogeneity in Colombian Farmers Exposed to Pesticides. Front. Genet. 13:820209. doi: 10.3389/fgene.2022.820209

Pesticides are a group of environmental pollutants widely used in agriculture to protect crops, and their indiscriminate use has led to a growing public awareness about the health hazards associated with exposure to these substances. In fact, exposure to pesticides has been associated with an increased risk of developing diseases, including cancer. In a study previously published by us, we observed the induction of specific chromosomal alterations and, in general, the deleterious effect of pesticides on the chromosomes of five individuals exposed to pesticides. Considering the importance of our previous findings and their implications in the identification of cytogenetic biomarkers for the monitoring of exposed populations, we decided to conduct a new study with a greater number of individuals exposed to pesticides. Considering the above, the aim of this study was to evaluate the type and frequency of chromosomal alterations, chromosomal variants, the level of chromosomal instability and the clonal heterogeneity in a group of thirty-four farmers occupationally exposed to pesticides in the town of Simijacá, Colombia, and in a control group of thirty-four unexposed individuals, by using Banding Cytogenetics and Molecular Cytogenetics (Fluorescence *in situ* hybridization). Our results showed that farmers exposed to pesticides had significantly increased frequencies of chromosomal alterations, chromosomal variants, chromosomal instability and clonal heterogeneity when compared with controls. Our results confirm the results previously reported by us, and indicate that occupational exposure to pesticides induces not only chromosomal instability but also clonal heterogeneity in the somatic cells of people exposed to pesticides. This study constitutes, to our knowledge, the first study that reports clonal heterogeneity associated with occupational exposure to pesticides. Chromosomal instability and clonal heterogeneity, in addition to reflecting the instability of the system, could predispose cells to acquire additional instability and, therefore, to an increased risk of developing diseases.

Keywords: pesticides, chromosomal alterations, chromosomal instability, clonal heterogeneity, occupational exposure

INTRODUCTION

Pesticides are a group of environmental pollutants widely used in agriculture to protect crops, so their indiscriminate use has led to a growing public awareness about the health hazards associated with exposure to these substances. Additionally, given that in Colombia one of the most important economic activities is agriculture, occupational exposure to these pesticides constitutes a risk due to their detrimental effect on human health. Currently, there are more than 1000 chemicals, which are classified as pesticides, some of them considered as potential genotoxic agents. Although the World Health Organization (WHO), groups pesticides according to their potential health risks (FAO and WHO, 2021), several of the classified as extremely toxic, are still used in our country, Colombia, including herbicide, fungicide and insecticide (mancozeb, glyphosate, malathion) (Idrovo, 2000). Pesticide exposure (absorption *via* dermal and/or respiratory routes) is now known to be associated with genotoxicity, oxidative stress, genetic damage and induction of chromosomal alterations, as well as reproductive disorders, neurodegenerative and cardiovascular diseases, and even with an increased carcinogenic risk (Cocco et al., 2013; Nicolopoulou-Stamati et al., 2016; Polito et al., 2016), especially for hematopoietic bone marrow cancers including myelodysplastic syndrome (MDS), leukemia acute myeloid (AML) and multiple myeloma (Tomiazzi et al., 2018). In fact, genetic damage constitutes an important event in the development of carcinogenesis, also correlated with the induction of genomic instability. Chromosomal damage related to pesticide exposure, has been identified in several populations, and while some researchers have reported significant differences in the frequency of chromosomal alterations (CAs) in exposed individuals compared to unexposed controls (Dulout et al., 1985; Carbonell et al., 1990; De Ferrari et al., 1991; Rupa et al., 1991; Balaji and Sasikala, 1993; Brega et al., 1998), others have not observed any association (Gomez-Arroyo et al., 2000). However, in these studies, the evaluation of chromosomal damage has been limited to the identification of chromosome gaps, breaks, sister chromatid exchange (Gomez-Arroyo et al., 2000) and micronuclei (MN), among others, so information on the type and frequency of specific CAs and chromosomal variants (CVs), as well as the level of chromosomal instability (CIN) and clonal heterogeneity (CH) induced by exposure to pesticides is scarce. In fact, one of the few studies available that indicate the type and frequency of specific chromosomal alterations induced by exposure to pesticides was reported by us, in a small group of exposed (five exposed) (Cepeda et al., 2020). Considering the importance of our previous findings (Cepeda et al., 2020) and their implications both, in the identification of cytogenetic biomarkers for the monitoring of exposed populations, and in the possibilities of their future application in early diagnostic tests, we decided to conduct a new study with a greater number of individuals exposed to pesticides. Considering the above, the aim of the present study was to evaluate the genotoxic damage (CAs, CVs, CIN and CH), in a group of thirty-four (34) farmers occupationally exposed to pesticides in the town of Simijacá, Colombia, and in a control group of thirty-four (34) unexposed individuals, by using GTG Banding and Fluorescence *in situ* hybridization (FISH). The results obtained from the analysis of a large number of metaphases, allowed to identify the type and frequency of CAs and CVs, as well the level of CIN and CH, not previously reported in farmers exposed to pesticides. Our study shows the deleterious effect of pesticides on the chromosomes of occupationally exposed individuals.

MATERIALS AND METHODS

Study Population

A total of 68 individuals were part of this study: thirty-four (34) individuals from the town of Simijacá, Colombia who were farmers routinely "exposed" to pesticides (exposed group) and thirty-four (34) individuals without indication of previous occupational exposure to pesticides (unexposed group). The exposed group consisted of men and women between 23 and 70 years old, involved in pesticide spray/handling and who had been exposed to pesticides through work for at least 3 months. The farmers' route of exposure to pesticides was mainly dermal and/or respiratory (**Table 1** and **Supplementary Table S1**). Minor routes of exposure to pesticides, including unintentional (accidental) oral exposure, ocular/ear exposure, and/or parenteral exposure (intramuscular, subcutaneous, or intravenous), were not reported by the exposed group. The unexposed group consisted of healthy men and women, without indication of previous occupational exposure to pesticides. The unexposed group had a similar age range (between 23 and 70 years old), sex distribution and life style habits as the exposed group (**Table 1** and **Supplementary Table S1**). Each subject was also required to complete a routine questionnaire to record possible confounding factors such as diseases, age, smoking and drinking habits, time of exposure to pesticides, pesticide exposure frequency, type of pesticide mixture, the dose of pesticides (expressed in kilograms/hectare) used by each exposed individual, as well as the number of hectares sprayed per day by each of them (**Table 1** and **Supplementary Table S1**). Participants suffering from cancer or had received radiotherapy, chemotherapy, or other

TABLE 1 | General characteristics of the groups studied.

	Exposed	Unexposed
Number	34	34
Age (mean ± SD)	46.64 ± 12.13	47.11 ± 11.24
Sex (n)		
Male	20	20
Female	14	14
Exposure months (mean ± SD)	133.2 ± 126.6	0
Smoking status (n)		
Smokers	4	4
Non-smokers	30	30
Drinking status (*n*)		
Drinkers	25	17
Non-drinkers	9	20

SD, standard deviation.

prolonged medical treatment, were excluded from the study. Data from the exposed individuals were compared with those of the unexposed individuals.

Blood Sampling

Five milliliters of peripheral blood, from exposed and unexposed individuals, were collected into heparinized tubes by venous puncture. The written informed consent of each subject participating in the study was obtained before the blood samples were taken.

Cytogenetic Studies and GTG Banded Karyotyping

The metaphases and interphase nuclei of the cultured peripheral blood lymphocytes were obtained using standard protocols. Briefly, lymphocyte cultures were performed by adding 1 ml of whole blood, in 5 ml of RPMI-1640 medium (Sigma, St. Louis, MO, United States), supplemented with 10% fetal bovine serum (FBS) (Sigma) and 100 µl of phytohemagglutinin-M (Gibco, Life Technologies, Nebraska, United States). The cultures were incubated at 37°C for 72 h in a 5% CO_2 atmosphere. All cultures of each individual, exposed and unexposed, were performed in duplicate. After 72 h, a solution of N-deacetyl-N-methyl colchicine (0.0001 g/ml final concentration) (Sigma) was added to the cultures for 25 min. After this time, the cells were treated with hypotonic solution (0.075 M KCl) for and fixed with carnoy fixative (3:1 methanol: acetic acid). Thus obtained, the chromosomal preparations were spread on glass slides and banded with GTG banding using trypsin (0.25%) (Gibco) and Giemsa (Sigma).

Cytogenetic Analysis

The identification of CVs and CAs (numerical and structural chromosomal alterations), by using GTG banded karyotyping was performed on a total of 2554 metaphases. Metaphase spreads were analyzed using an Olympus microscope and processed using the cytogenetic software Cytovision System 7.4 (Leica Biosystems Richmond, VA, United States). CVs [variation in length of heterochromatic segments on the long arms of chromosomes 1 (1qh+), 9 (9qh+) and 16 (16qh+)], fragilities (fra), inversion of chromosome 9 [inv(9)], chromosomal breaks (chrb) and chromatid breaks (chrb), and CAs including structural (SCAs) and numerical chromosomal alterations (NCAs) were evaluated. All CVs and CAs were described according to the International System for Human Cytogenomic Nomenclature (ISCN) 2020 (McGowan-Jordan et al., 2020).

Molecular Cytogenetics Studies (FISH)

FISH was used to evaluate CIN and CH on chromosomal spreads (metaphases and interphase nuclei) previously obtained. For the above, six (6) centromeric probes (CEP) labeled with different fluorochromes were used, for chromosomes 2 and 3 (orange fluorochrome), 8 and 17 (blue fluorochrome) and, 11 and 15 (green fluorochrome) (all from Cytocell, Cambridge). Tricolor FISH was performed on the chromosome preparations for chromosomes 2, 8, and 11, and for chromosomes 3, 15, and 17. Briefly, the chromosomal spreads were dehydrated in ethanol series, and after adding the probe mixture, they were denaturated at 75°C for 2 min and hybridized overnight at 37°C, using the Top Brite system (Resnova, Italy). After this time, the chromosome extensions were washed, dehydrated and stained with 4′, 6-diamidino-2-phenylindole (Cytocell). Finally, ten randomly selected areas of the chromosomal spreads from each exposed and unexposed individual, were acquired using an Olympus microscope and processed using the cytogenetic software Cytovision System 7.4. CIN was evaluated in a minimum of 100 intact and non-overlapping nuclei/metaphases for each chromosome. Although it has been suggested that the use of probes for only two chromosomes is sufficient to identify diploid aneuploid tumors (Fiegl et al., 2000; Takami et al., 2001), we decided to use 6 probes because the use of more than two probes allows the identification of clonal populations with greater certainty (Farabegoli et al., 2001). The CIN rate for each exposed and unexposed individual was defined first by calculating, for each of the six chromosomes separately, the percentage of nuclei with a CEP signal number different to the modal number (most frequent number of chromosomes in a cell population), and then calculating the mean CIN percentage of all six chromosomes analyzed (Lengauer et al., 1997; Munro et al., 2012). According to the level of CIN, each exposed and unexposed individual was classified as having low CIN (CIN < 25%) or high CIN (CIN ≥ 25%) (Kawauchi et al., 2010; Talamo et al., 2010). The CIN levels observed in each of exposed individuals were determined in comparison with the control group (unexposed). In order to evaluate the CH (presence of cell populations with different levels of aneuploidy in the same person), in each exposed and unexposed individual, we calculated the Shannon Diversity Index (SDI) and the true diversity index (TD) for chromosomes 2, 3, 8, 11, 15, and 17. SDI and TD integrates both the number and abundance of cell clones within each cell according to published methods (Jost, 2006; Maley et al., 2006; Roylance et al., 2011).

Data Analysis

With the aim of comparing the GTG-banding cytogenetic data with parametric and non-parametric distribution, Fisher's exact test, Student's t-test and Wilcoxon test were performed. Normality of the data was evaluated by the Shapiro Wilk test. Data from the exposed individuals were compared with those of the unexposed individuals. Student's t-test and Wilcoxon test were performed to compare CIN, SDI, and TD data with parametric and nonparametric distribution, respectively. To compare CIN, SDI, and TD between the chromosomes used in this study, the Kruskal–Wallis test was used for data with nonparametric distribution. Normality and homoscedasticity of the data were assessed by Shapiro Wilk's test and Bartlett's test, respectively. In order to establish, in each of the exposed and unexposed groups, the existence of associations between the levels of CIN and CH with variables such as age, sex, and time of exposure to pesticides (only in exposed), we perform multivariate analysis using the Pearson correlation coefficient. Data from exposed individuals were compared with those from unexposed individuals. All statistical analyses were carried out

using the R Studio version 4.0.2 and p values < 0.05 were considered as statistically significant (*$p \leq 0.05$, **$p \leq 0.01$ and ***$p \leq 0.001$). CIN, SDI and TD are expressed as means ± SD.

RESULTS

Characteristics of Study Groups

General and detailed characteristics of the groups studied (exposed and unexposed) are presented in **Table 1** and **Supplementary Table S1**, respectively. For the exposed group, the mean time of exposure to pesticides was 133.2 months, the mean age was 46.64 years (**Table 1**), and the pesticide exposure frequency was mainly once a week (**Supplementary Table S1**). The dose of pesticides (expressed in kilograms/hectare) used by each exposed individual, as well as the number of hectares sprayed per day by each of them, are also indicated in **Supplementary Table S1**. A low prevalence of alcohol consumption and cigarette smoking was reported in both groups, exposed and unexposed. The results are expressed as the mean ± standard deviation (SD) (**Table 1** and **Supplementary Table S1**). Pesticides mixtures to which farmers were mainly exposed included: fungicides (Antracol, Cymoxanil, Cymozeb, Dithane, Fitoraz, Forum, Mancozeb, Propineb), insecticides (Arrivo, Astuto, Carbosulfan, Carbofuran, Cayenne, Chlorpyrifos, Confidor, Cypermethrin, Curacron, Decis, Eltra, Engeo, Fulminator, Furadan, Imidacloprid, Karate, Lambda-cyhalothrin, Lannate, Lorsban, Match, Methyl parathion, Perban, Profenofos, Tiguvon), and herbicides (Paraquat, Cerillo) (**Supplementary Table S1**).

GTG Banding Cytogenetic Results

According to the International recommendations for the analysis of constitutional studies (CCMG-CCGM National Office, 2021; Ozkan and Marcelo, 2021), a minimum of between 10 and 20 metaphases must be analyzed for cytogenetic analysis. If in these 10 or 20 metaphases no numerical or structural alterations are observed, it is not necessary to analyze additional metaphases. If, on the contrary, numerical and/or structural alterations are observed (conditions where mosaicism is a significant possibility), examination of additional metaphases is required (minimum of 25–50 metaphases). Considering the above, we analyzed a minimum of 19 metaphases, from individuals of both groups (exposed and unexposed), in those cases in which no numerical or structural alterations were observed, and we extended the cytogenetic analysis to a maximum of 95 metaphases in the cases in which this type of alterations was observed. The difference in the number of metaphases analyzed is also due to the variation in the mitotic index in each individual included in the study. A total of 2554 metaphases were analyzed. GTG banding cytogenetic analysis for both, exposed and unexposed groups, demonstrated a modal diploid number (2n). Significantly high frequencies for CVs, fragilities, chrb, chrb, structural (SCAs) and numerical chromosomal alterations (NCAs), were found in the exposed group compared with those observed in the unexposed group (1471 and 209, respectively) ($p \leq 0.0027$**; unpaired Mann-Whitney test) (**Figure 1**).

Specifically, in the exposed group were observed: 384 numerical alterations in 32 (94.1%) individuals; 88 structural alterations in 27 (79.4%) individuals; 625 fragilities in 32 (94.1%) individuals; 107 chromatid and/or chromosomal breaks in 25 (73.5%) individuals, and 267 chromosomal heteromorphisms in 20 (58.8%) individuals (**Table 2**). While in the unexposed group, were observed 43 numerical alterations in 15 (44.1%) individuals; 13 structural alterations in 9 (26.4%) individuals; 97 fragilities in 17 (50%) individuals; 26 chromatid and/or chromosomal breaks in 14 (41.1%) individuals, and 30 chromosomal heteromorphisms in 4 (11.8%) individuals (**Figure 1** and **Table 2**). The comparison in the frequency of CVs and CAs between the exposed and unexposed groups showed statistically significant differences ($p \leq 0.01$**; Fisher's exact test) in most cases.

Within the numerical alterations, in the exposed group, monosomies (94.1%) were observed more frequently than trisomies (76.4%) (**Figure 1** and **Table 2**). The chromosomes with the highest frequency of monosomies were the chromosomes X in 11 (32.35%) exposed, and chromosome 20 in 15 (44%) exposed. Within the trisomies, marker chromosomes were observed with a higher frequency in 21 exposed (61.76%), followed by trisomy of chromosome 22 in 9 exposed (26.47%), and trisomy of X chromosome in 7 exposed (20.58%).

Numerical chromosomal alterations were also identified in the unexposed group, where monosomies (38.2%) were observed more frequently than trisomies (8.8%) (**Figure 1** and **Table 2**). Among the monosomies, the most frequent was the monosomy of the X chromosome observed in 8 individuals (23.52%), followed by monosomy of chromosome 2 (11.7%) in 6 (17.6%) unexposed individuals, monosomy of chromosome 12 (11.7%) in 4 (11.7%) unexposed individuals, and monosomy of chromosome 13 (11.7%) in 4 (11.7%) unexposed individuals.

Regarding SCAs, these were observed in the 79.41% of the exposed individuals, and in the 23.5% of unexposed individuals (**Figure 1** and **Table 2**). A total of 88 SCAs were observed in the exposed group, being the most frequent the deletions (del) (37.5%), followed by translocations (t) (14.77%) and additional material of unknown origin (add) (9.09%). Other structural alterations observed less frequently include derived chromosomes (der) (7.95%), inversions (inv) (6.81%), dicentric chromosomes (dic) (2.27%), duplications (dup) (1.13%), isochromosomes (i) (1.13%) and ring chromosomes (r) (1.13%). The chromosomes most frequently involved in SCAs were chromosomes 4, 7, and 9, followed by chromosomes 6, X, 2, and 12. While, the chromosomes least involved in SCAs were the chromosomes Y, 14, 15, and 19. No SCAs were observed affecting chromosomes 20 and 21. The following alterations: inv(9)(p21q21), del(X)(q25), del(6)(q25), del(11)(q11) and del(16)(q24) were observed in more of one exposed (**Figure 2**). Regarding specific altered chromosomal regions, we observed that chromosomal regions 6p23, 7p22, and 12p13 were commonly altered in more than one (1) exposed (E3, E21, E25, E32, E34, and E35) (**Figure 2**).

In the unexposed group were observed a total of 13 SCAs being the most frequent the deletions (del) (50%), followed by

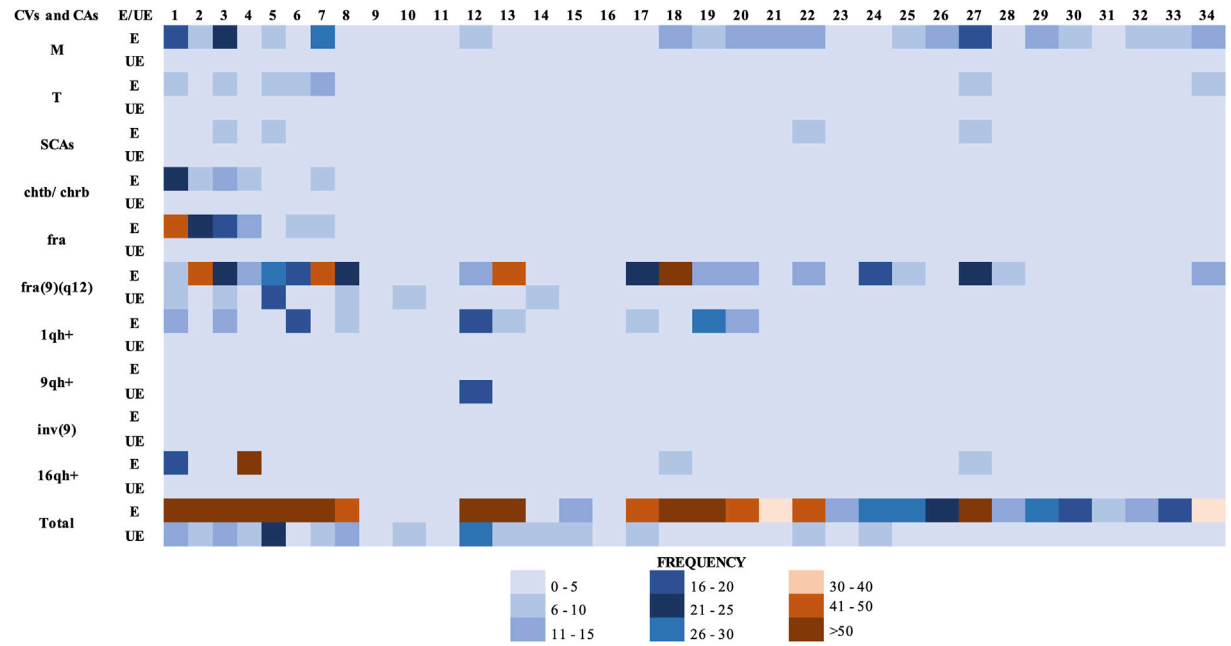

FIGURE 1 | Total chromosomal variants (CVs) and chromosomal alterations (CAs) observed in the groups studied. (E) Exposed group. (UE) Unexposed group. Each column in the figure, represents a participant in the study (34 columns in total). Abbreviations: M, monosomies; T, trisomies; SCAs, structural chromosomal alterations; chtb, chromatidic break; chrb, chromosomic break; fra, fragilities; fra(9)(q12), fragility in the long arm of chromosome 9, region 1 and band 2; 1qh+, heterochromatin increased on long arm of chromosome 1; 9qh+, heterochromatin increased on long arm of chromosome 9; inv(9), inversion of chromosome 9; 16qh+, heterochromatin increased on long arm of chromosome 16.

TABLE 2 | Frequencies and percentages of chromosomal variants (CVs) and chromosomal alterations (CAs) identified in the exposed and unexposed groups.

CVs and CAs	Number of individuals		
	Exposed n (%)	Unexposed n (%)	*p*
Monosomies	32 (94.1)	13 (38.2)	<0.0001**
Trisomies	26 (76.4)	3 (8.8)	0.0029**
SCAs	27 (79.4)	8 (23.5)	<0.0001**
chtb/chrb	25 (73.5)	14 (41.1)	0.0136**
fra	16 (47.1)	10 (29.4)	0.2118
fra(9)(q12)	32 (94.1)	17 (50)	<0.0001**
1qh+	20 (58.8)	4 (11.8)	<0.0001**
9qh+	8 (23.5)	4 (11.8)	0.3405
inv(9)	7 (20.5)	1 (2.9)	0.5118
16qh+	9 (26.4)	0 (0)	0.0021**
Total	34	34	

**Statistically significant difference relative to unexposed group at p ≤ 0.05.*
***Statistically significant difference relative to unexposed group at p ≤ 0.01 (Fisher's exact test).*
M, monosomies; T, trisomies; SCAs, structural chromosomal alterations; chtb, chromatidic break; chrb, chromosomic break; fra, fragilities; fra(9)(q12), fragility in the long arm of chromosome 9, region 1 and band 2; 1qh+, heterochromatin increased on long arm of chromosome 1; 9qh+, heterochromatin increased on long arm of chromosome 9; inv(9), inversion of chromosome 9; 16qh+, heterochromatin increased on long arm of chromosome 16; SD, standard deviation.

translocations (t) (16.7%). Other less frequently observed SCAs include inversions (inv) (8.3%), derived chromosomes (der) (8.3%) and duplications (dup) (8.3%). In addition, a higher frequency of non-clonal SCAs was identified in the both groups, being these higher in the unexposed group.

With regard fragilities (fra), a higher frequency of these were found in the exposed group (625 fragilities) compared with the unexposed group (97 fragilities) (**Figure 1**). In both groups, many of the fragilities were non-clonal. In addition, a total of 107 chromosomal (chrb) and/or chromatic (chtb) breaks were observed in the exposed group in comparison with 26 chrb and/or chtb observed in the unexposed group (**Figure 1**). In the exposed group, the chromosomal and/or chromatic breaks chtb(1)(q21), chtb(1)(q10), chrb(3)(p14), chtb(3)(p21), chtb(5)(q31), chtb(6)(p21), chrb(9)(q12), chtb(12)(q15), chtb(12)(q13), chtb(13)(q31) and chtb(19)(p10) were observed in more than one (1) exposed. Comparison of the presence of CVs, chrb/chtb, NCAs and SCAs, between exposed and unexposed groups (**Table 2**), and between paired exposed/unexposed individuals (**Table 3**) showed statistically significant differences ($p \leq 0.001$***; Fisher's exact test, and $p \leq 0.05$*, respectively). Although in all cases no statistically significant differences were observed between the exposed and unexposed individuals, the frequency of CVs, chrb/chtb, NCAs and SCAs was higher in the exposed group, evidencing chromosomal damage due to exposure to pesticides.

The evaluation of the effect of smoking and alcohol consumption as confounding factors on the frequency of CV, chrb, chtb and CCA and NCCA (numerical and structural chromosomal alterations) in all study subjects, allowed us to conclude that none of these (alcohol consumption, smoking)

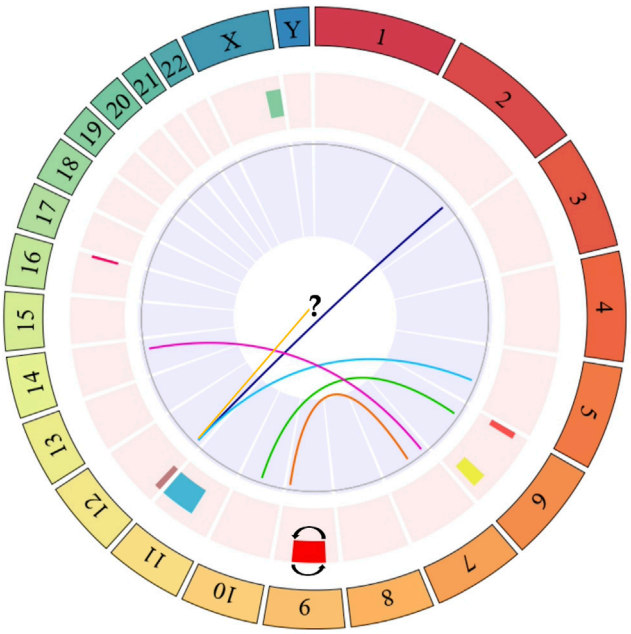

FIGURE 2 | Circos plot of specific chromosomal regions commonly altered in more than one exposed individual. The outer ring indicates the number of the chromosome. The next ring indicates chromosomal abnormalities affecting only one chromosome, or where only one chromosome was identified. These alterations include: del(X)(q25) (green bar), del(6)(p23) (red bar), del(6)(q25) (yellow bar), inv(9)(p21q22) (red bar with reverse lines), del(11)(q11) (light blue bar), add(12)(p13) (yellow line), del(12)(p13) (purple bar), and del(16)(q24) (fuchsia line). The last ring (in the center of the circos plot) indicates chromosomal alterations involving more than one chromosome. These alterations include: t(2;12)(q33;p13) (dark blue line), t(5;12)(q23;p13) (light blue line), t(6;10)(p23;q22) (green line), t(7;9)(p22;q34) (orange line), t(7;14)(p22;q12) (purple line). The question mark (?) indicates additional material of unknown origin (add) attached to the short arm of chromosome 12 [add(12)(p13)]. Dark blue, light blue, green, orange, and purple links within the circos plot show translocations. The circos plot was designed in the statistical software R using the BioCircos library, later it was edited in the power point software to add some symbols that represent some alterations, which are not found in the aforementioned library.

increases the frequency of CVs and CAs in any of the groups studied, exposed and unexposed (**Table 1** and **Supplementary Table S1**).

FISH Results

We assessed CIN in 100 interphase nuclei and some metaphases by using centromeric FISH. Exposed individuals showed a high CIN (≥22.67%) compared with a low CIN (≤13.83%) observed in unexposed individuals (**Figures 3**, **4**, and **Supplementary Table S2**). More specifically, in exposed individuals, CIN ranged between 22.67 and 47.33%, while in non-exposed individuals, CIN ranged between 0.83 and 13.83% (**Figures 3**, **4**).

The mean CIN was 34.57% ± 6.03 for exposed, and 6.48% ± 3.13 for unexposed. Student's t-test showed statistically significant differences ($p < 0.001$**) between the CIN of the exposed and unexposed individuals. These results suggest that pesticides can induce aneuploidy, which is indicative of numerical CIN.

In order to determine the most stable chromosomes in the groups studied (exposed and unexposed), we carried out the Kruskal–Wallis test. This test showed in the exposed group, a statistically significant difference ($p < 0.001$***) between chromosomes 2, 3, 11, and 15, and chromosomes 8 and 17, with chromosomes 8 and 17 being the most stable. For the unexposed group, statistically significant differences were also observed ($p < 0.001$***) between the chromosomes 3, 11, and 15; the chromosomes 2, 11, and 15 and the chromosomes 8 and 17, with chromosomes 8 and 17 being the most stable, similar to what was observed in the exposed group (**Figure 5**).

Clonal Heterogeneity

In order to determine the CH in the both groups, two different but related indices were used, the SDI and true diversity index (TD), which integrate the number and abundance of cell clones in

TABLE 3 | Frequency (n) and percentage (%) of chromosome variants (CVs) and chromosomal alterations (CAs) identified in paired exposed/unexposed individuals.

No	Exposed		Unexposed		*p*
	n	%	n	%	
1	138	9.24	14	1.31	0.01**
2	89	5.96	6	0.56	0.02*
3	109	7.30	14	1.31	0.06
4	109	7.30	7	0.65	0.01**
5	65	4.35	21	1.97	0.68
6	60	4.01	4	0.37	0.12
7	105	7.03	10	0.94	0.06
8	43	2.88	14	1.31	0.62
9	4	0.26	2	0.18	0.99
10	5	0.33	6	0.56	0.99
11	3	0.20	5	0.47	0.99
12	54	3.61	28	2.63	0.99
13	62	4.15	10	0.94	0.36
14	5	0.33	8	0.75	0.99
15	15	1.00	8	0.75	0.99
16	0	0	5	0.47	0.99
17	42	2.81	8	0.75	0.24
18	101	6.76	5	0.47	0.01**
19	52	3.48	4	0.37	0.12
10	45	3.01	0	0	0.24
21	31	2.07	1	0.09	0.49
22	43	2.88	10	0.94	0.62
23	14	0.93	0	0	0.99
24	28	1.87	7	0.65	0.49
25	30	2.00	2	0.18	0.49
26	23	1.54	1	0.09	0.49
27	72	4.82	0	0	0.05*
28	13	0.87	2	0.18	0.99
29	26	1.74	1	0.09	0.49
30	16	1.07	2	0.18	0.99
31	6	0.40	0	0	0.99
32	12	0.80	2	0.18	0.99
33	16	1.07	1	0.09	0.99
34	35	2.34	1	0.09	0.49

Statistically significant difference relative to the unexposed group at p ≤ 0.05.
***Statistically significant difference relative to the unexposed group at p ≤ 0.01 (Fisher's exact test).*
The total number of metaphases analyzed in the exposed group was 1493, while in the unexposed group (control) it was 1061.

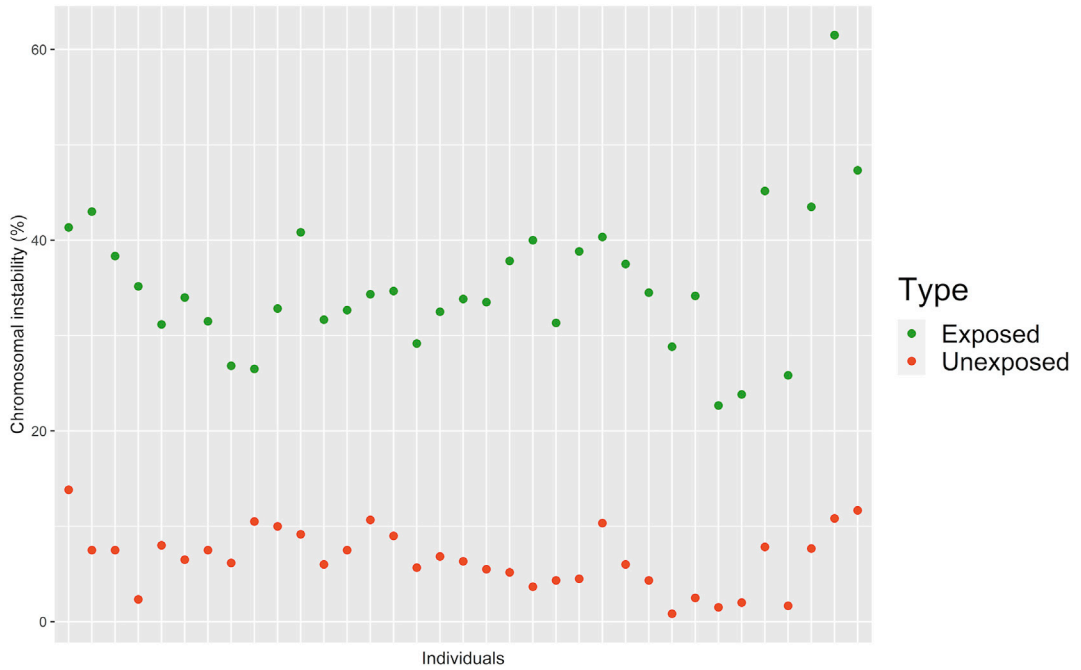

FIGURE 3 | Percentage of CIN assessed by FISH in 100 interphasic nuclei in the exposed and unexposed groups. According to the level of CIN, each exposed and unexposed individual was classified as having low CIN (CIN < 25%) or high CIN (CIN ≥ 25%).

each individual (exposed and unexposed) according to published methods (Maley et al., 2006; Jost and Gonzá lez-Oreja, 2012). CH was 1.99 higher in the exposed group than in the unexposed group. Significant statistical differences between exposed and unexposed groups for both, TD ($p < 0.001$***; Non-parametric Mann Whitney Wilcoxon) and the SDI ($p < 0.001$***; Non-parametric Mann Whitney Wilcoxon) were observed (**Figure 6**, **Supplementary Table S2**, **Supplementary Figure S1**).

Likewise, CH was also determined for each of the chromosomes studied in each group. For both groups, statistically significant differences were observed, for both TD ($p < 0.001$***; Kruskal–Wallis test) (**Supplementary Figure S2**) and for SDI ($p < 0.0016$***; Kruskal–Wallis test) (**Supplementary Figure S3**), and between the group of chromosomes 2, 3, 11, and 15 and the group of chromosomes 8 and 17, being chromosomes 8 and 17, those with the lowest CH.

Correlation of Variables

In order to establish in both groups, exposed and unexposed, the existence of associations between the levels of CIN and CH (TD), with variables such as age, sex, and time of exposure (TE) to pesticides (only in the exposed group), we perform multivariate analysis using the Pearson correlation coefficient. In both groups, a strongly positive relationship was found between the CIN and CH. However, no linear correlation was found between CIN and CH with any of the variables studied (age, sex, and TE to pesticides) (**Figure 7**). The variables smoking and drinking habits, were not evaluated due to the low prevalence reported by the two groups.

DISCUSSION

Pesticides are a heterogeneous category of chemicals specifically designed for pest control. Although its application continues to be the most effective method for protecting plants against pests, its use has been associated with harmful effects on the health of the people involved in its regular and extensive use. In fact, it has been indicated that farmers occupationally exposed to pesticides during spraying activities are more prone to genotoxicity than those not exposed. In this regard, some studies have identified chromosomal damage related to pesticide exposure in various populations, however, in these studies, information on the type and frequency of specific CAs and CVs, as well as the level of CIN and CH induced by the exposure to pesticides is scarce. In fact, one of the few available studies indicating the type and frequency of specific chromosomal alterations induced by pesticide exposure was reported by us, in a small group of exposed (five exposed) (Cepeda et al., 2020). In this study, we observed a significant increase in clonal and non-clonal chromosomal alterations in individuals exposed to pesticides compared to unexposed individuals (Cepeda et al., 2020). Considering the importance of our previous findings in the identification of cytogenetic biomarkers for the monitoring of exposed populations, we decided to conduct a new study with a greater number of individuals exposed to pesticides.

Our results indicate that occupational exposure to pesticides was associated to a significant increase in CIN, in agreement with previous reports indicating DNA damage in populations occupationally exposed to pesticides (Grover et al., 2003; Castillo-Cadena et al., 2006; Wilhelm et al., 2015). Our results

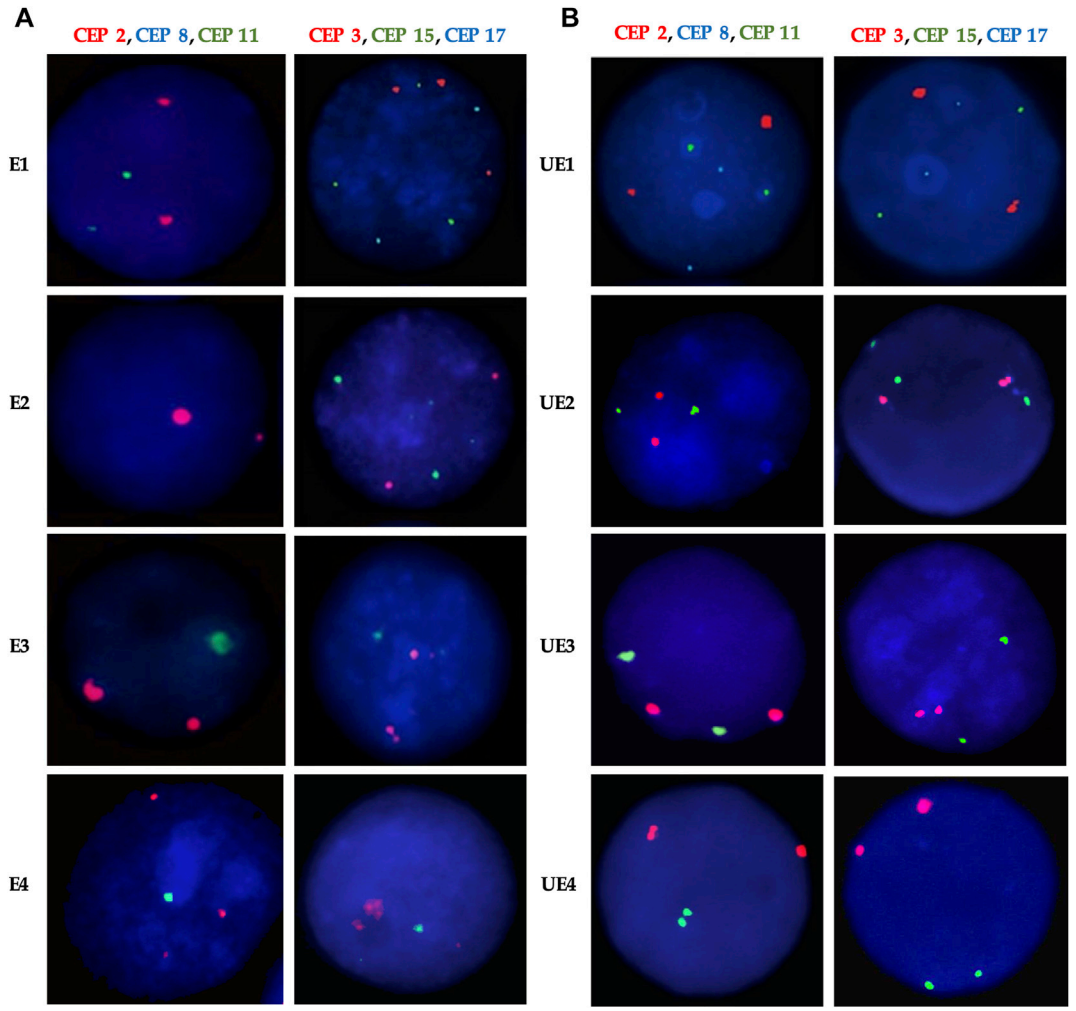

FIGURE 4 | Representative FISH images for **(A)** Exposed and **(B)** Unexposed individuals. Three-color FISH was performed on nuclei spreads for chromosomes 2, 8, and 11 and, chromosomes 3, 15, and 17 using centromeric probes (CEP) labeled with different spectrum colors: spectrum orange for CEP2 and CEP3; spectrum aqua for CEP8 and CEP17; and spectrum green for CEP11 and CEP15**.** Interphase nuclei at each treatment time point are indicated. E, Exposed; UE, Unexposed individuals.

show that individuals exposed to pesticides have a high frequency of CAs, CVs, CIN, and CH compared to low frequency observed in unexposed individuals. The mean number of CVs and CAs observed in the exposed individuals was five times higher than in the unexposed individuals. Numerical and structural chromosomal alterations were higher and with a statistically significant prevalence in the exposed group. These findings suggest a possible cytogenetic effect of pesticides on occupationally exposed individuals.

Regarding the numerical alterations identified in both study groups, a high frequency of aneuploidy, was observed in the exposed group compared to the unexposed group. Aneuploidy refers to the gain and/or loss of complete chromosome, which can be stable or unstable. Unstable aneuploidy (cell-to-cell variation in chromosome number) may favor the simultaneous growth of various cellular subpopulations leading to genomic heterogeneity (Bolt et al., 2004; Gagos and Irminger-Finger, 2005; Geigl et al., 2008; Tanaka and Hirota, 2016; Vargas-Rondon et al., 2017). Even though the mechanisms by which pesticides induce aneuploidy are not fully understood, it has been suggested that they can lead to chromosomal nondisjunction, and thus to the loss or gain of entire chromosomes, by interacting with a variety of cellular processes including, the alteration in the formation of chromosomal microtubules responsible for segregation of genetic material during cell division (Lushchak et al., 2018); the synthesis, division and functioning of centrioles, polar bodies and spindle fibers (Zijno et al., 1996); the assembly and functioning of the kinetochore proteins (Parry et al., 2002), and the centrosome activity and the modification of centromeres (Renzi et al., 1996; Mattiuzzo et al., 2006).

In addition to the numerical alterations, we also observed in the exposed group, high frequency of structural chromosomal alterations. The chromosomes most frequently involved in structural alterations were chromosomes 4, 7, and 9, followed

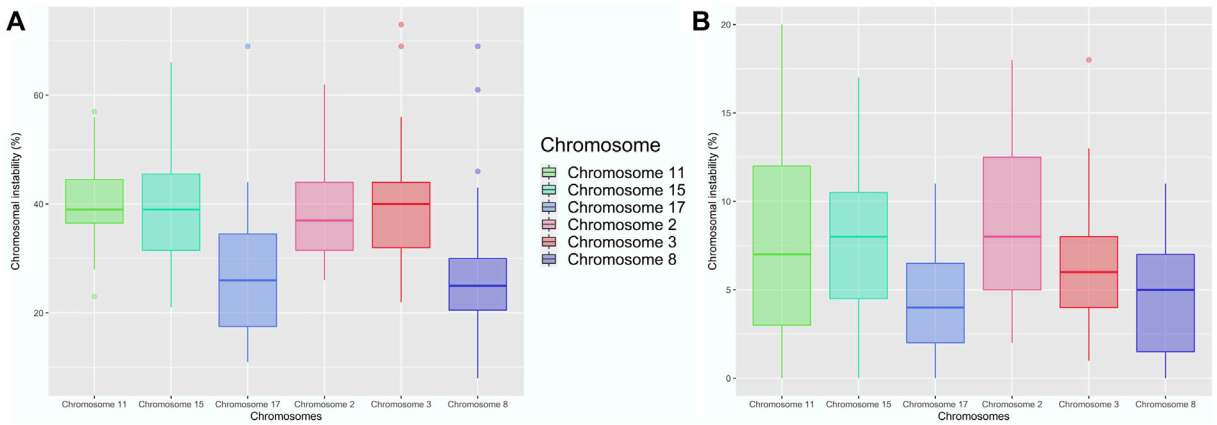

FIGURE 5 | Percentage of CIN in the Exposed **(A)** and Unexposed **(B)** groups. According to the level of CIN, each chromosome was classified as having low CIN (CIN < 25%) or high CIN (CIN ≥ 25%). The most stable chromosomes for exposed individuals were chromosome 8 and 17, and the most unstable chromosomes were chromosome 2 and chromosome 15. While for unexposed individuals, the most stable chromosome were chromosomes 8 and 17 as well, and the most unstable chromosome was chromosome 3.

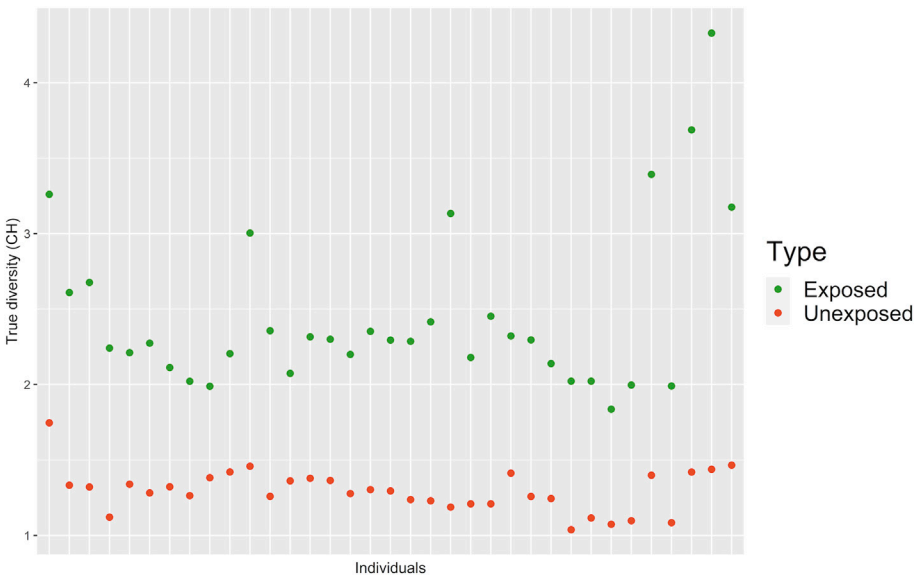

FIGURE 6 | Clonal heterogeneity (CH) determined by True Diversity (TD) for exposed and unexposed groups. Values below 1.5 were considered indicative of low CH, values between 1.6 and 2 were considered indicative of intermediate CH; and values higher than 2 were considered indicative of high CH.

by chromosomes 6, X, 2, and 12. Regarding specific chromosomal regions, we observed that chromosome regions 6p23, 7p22, and 12p13 were involved in more than one chromosomal alteration and in more than one (1) exposed. It should be noted that these affected chromosomal regions have been implicated in the development of various types of cancer (**Table 4**), evidencing the importance of their evaluation and/or identification in people exposed to genotoxics.

The implications of numerical and structural chromosomal alterations in the development of diseases could be due to the fact that chromosomal alterations can lead to altered expression of genes (proto-oncogenes and tumor suppressor genes) and variable protein concentrations, which control cell cycles and differentiation processes, and in turn may cause an unbalance at the cellular level with serious biologic consequences (Paz-y-Miño et al., 2002).

In addition to numerical and structural chromosomal alterations, a high frequency of fragilities (fra), chrb and chtb, was observed in the exposed group compared to the low frequency of the same observed in the unexposed group. Fragilities may be resulted from single-strand DNA breaks (Glover, 1998), which if not repaired, may lead to chromosome damage such as intrachromosomal gene amplification (Coquelle et al., 1997), sister chromatid

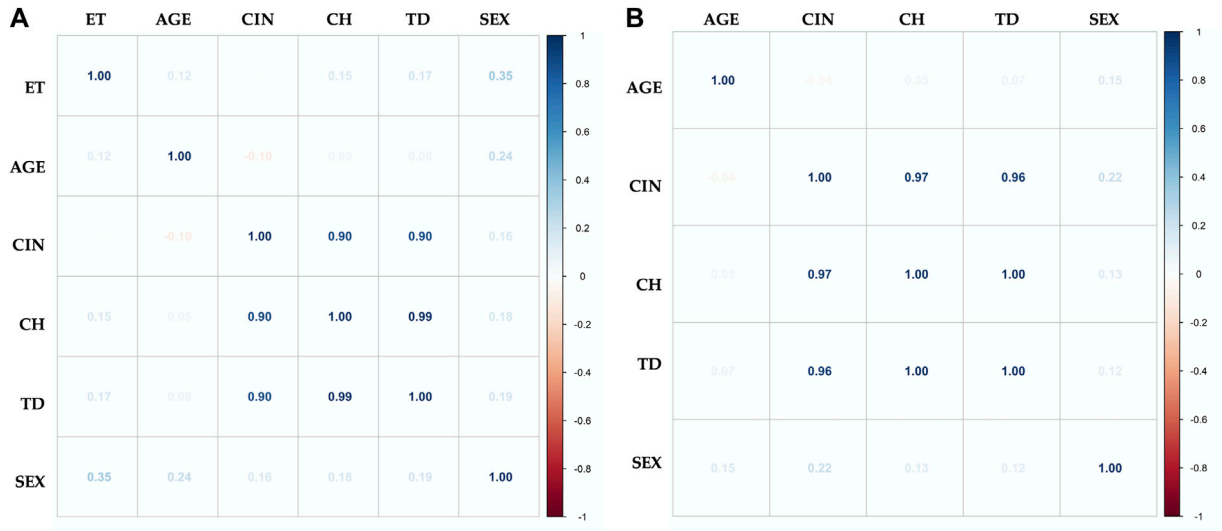

FIGURE 7 | Multivariate analysis with Pearson correlation coefficient for **(A)** Exposed and **(B)** Unexposed groups. Values greater than 0.5 are indicative of a statistically significant correlation. No linear correlation was found between chromosomal instability (CIN), clonal heterogeneity (CH) and true diversity index (TD) with any of the variables studied: time of exposure to pesticides (TE), age, and sex.

exchanges (Glover and Stein, 1987), deletions (Durkin and Glover, 2007), duplications (Hellman et al., 2002) and translocations (Re et al., 2006), among other, all of them associated with the development of cancer (Debacker and Kooy, 2007; Vincent-Salomon et al., 2013). Regarding chrb and chtb, both are chromosomal aberration that involves single and/or double stranded DNA breaks. Double-stranded DNA breaks can be induced by reactive oxygen species (ROS), which are highly reactive molecules involved in various cellular processes, causing fragmentation and oxidation of nucleic acids, proteins and lipids (Kaur and Kaur, 2018), and also associated with the exposure to pesticides (Hilgert Jacobsen-Pereira et al., 2018; Kaur and Kaur, 2018; Shah et al., 2020). Further, increased oxidative stress and ROS production due to pesticide use, has been associated with reproductive disorders in women, including cycle defects, folliculogenesis, follicular atresia, implantation defects, miscarriages and endometriosis (Bhardwaj et al., 2020). The presence of chrb and chtb in the exposed group, could predispose a greater risk to develop complex chromosomal rearrangements such as translocations, inversions, dicentric chromosomes, deletions and duplications, thus evidencing the high CIN associated with exposure to pesticides observed in our study.

Although chromosomal heteromorphisms (observed by us in higher frequency in the exposed group) are considered normal chromosomal variants, variations in size and location of the major heterochromatic regions (1qh, 9qh, 16qh) have particularly been implicated in various cancers and leukemias (Wyandt HE, 2004). For instance, Atkin (1977) first suggested susceptibility to malignancy associated with heteromorphisms in chromosome 1. In addition, rearrangements in the vicinity of the centromere of chromosome 1 have been reported as over-represented in many types of human cancers (Ji et al., 1997). Subsequent observations were reported for chromosomes 1, 9, and 16 and the Y chromosome and include observations of increased or decreased length, striking size differences between homologs (asymmetry), and pericentric inversions in heterochromatic regions. For example, an increase in heterochromatin of chromosome 16 was observed in couples with a stillborn or a malformed child (Buretic-Tomljanovic et al., 1997).

In our study, the observation of a higher frequency of chromosomal variants in the exposed group is noteworthy and could have important implications in the monitoring of populations exposed to pesticides. The above, considering not only the findings previously described, but additional studies that indicate that not all chromosomal variants involve only heterochromatin. Indeed, rearrangements in the pericentromeric region of chromosome 1 or 16, common in various types of cancers, are known to involve particular oncogenes that are close to the pericentromeric regions (Mugneret et al., 1995; Tse et al., 1995). The inversions or insertions of these genes in heterochromatin regions could possibly play a role in the activation or deactivation of these genes through positional effects (Wyandt HE, 2004).

To highlight that, while the numerical chromosomal alterations observed in the exposed group were mainly clonal (CCAs), the structural chromosomal alterations were non-clonal (NCCAs). CCAs and NCCAs can lead to clonal selection and to the expansion of chromosomal alterations, thus increasing overall heterogeneity. Both clonal selection and heterogeneity reflect the instability of the system and could lead to development of diseases by increasing the diversity of the cell population. Even though, NCCA have been considered as *in vitro* culture artifact because they are non-recurrent abnormalities, they have acquired great importance in recent years, given their correlation with both CIN

TABLE 4 | Chromosomal regions involved in chromosomal alterations in the exposed group and associated with the development of various types of cancer.

Type	Associated disease	Tumor site	Band	Abnormality	References
Unbalanced	Acute lymphoblastic leukemia/ lymphoblastic lymphoma		Xq25	del(X)(q25)	Heerema et al. (1992), Wuicik et al. (2007), Nayebbagher et al. (2020)
Unbalanced	Adenocarcinoma	Stomach, Breast	Xq25	del(X)(q25)	
Unbalanced	Acute lymphoblastic leukemia/ lymphoblastic lymphoma, Acute myeloid leukemia		6p23	del(6)(p23)	Sawyer et al. (1996), Takeshita et al. (2004), Anwar Iqbal et al. (2006)
Unbalanced	Astrocytoma, grade III-IV/ Glioblastoma	Brain	6p23	del(6)(p23)	
Unbalanced	Multiple myeloma		6p23	del(6)(p23)	
Unbalanced	Acute lymphoblastic leukemia		6q25	del(6)(q25)	Rogatto et al. (1993), Debiec-Rychter et al. (1995), Tibiletti et al. (1996), Gladstone et al. (1998), Tibiletti et al. (2000), Tibiletti et al. (2003), Amare Kadam et al. (2004), Cerretini et al. (2006), Travella et al. (2013), Sawyer et al. (2014)
Unbalanced	Adenocarcinoma	Breast, Ovary	6q25	del(6)(q25)	
Unbalanced	Astrocytoma, Glioblastoma	Brain	6q25	del(6)(q25)	
Unbalanced	Benign epithelial tumor	Breast	6q25	del(6)(q25)	
Unbalanced	Burkitt lymphoma		6q25	del(6)(q25)	
Unbalanced	Ependymoma	Cerebellum	6q25	del(6)(q25)	
Unbalanced	Multiple myeloma		6q25	del(6)(q25)	
Unbalanced	Retinoblastoma	Eye	6q25	del(6)(q25)	
Unbalanced	Teratoma	Testis	6q25	del(6)(q25)	
Balanced	Acute lymphoblastic leukemia		7p22	t(7;14)(p22;q11)	Prigogina et al. (1988), Olsson et al. (2018)
Balanced	Chronic myeloid leukemia		7p22	t(7;9;22)(p22; q34;q11)	El-Zimaity et al. (2004), Lee et al. (2012), Issa et al. (2017)
Unbalanced	Acute myeloid leukemia		9p21	46,XX,inv(9)(p21q22)	Nahi et al. (2008)
Unbalanced	Acute myeloid leukemia, Chronic lymphocytic leukemia		11q11	del(11)(q11)	Rigolin et al. (1997), Mandahl et al. (2000), Wang et al. (2001), Pantou et al. (2005), Gabrea et al. (2008), Rayeroux and Campbell, (2009), Campioni et al. (2012)
Unbalanced	Leiomyosarcoma	Soft tissue	11q11	del(11)(q11)	
Unbalanced	Multiple myeloma		11q11	del(11)(q11)	
Unbalanced	Acute lymphoblastic leukemia, Acute myeloid leukemia		12p13	add(12)(p13)	Pejovic et al. (1991), Presti et al. (1991), Bardi et al. (1993), Rodriguez et al. (1993), Hoogerwerf et al. (1994), Testa et al. (1994), Pandis et al. (1995), Bridge et al. (1997), Feder et al. (1998), Smolarek et al. (1999), Teixeira et al. (2001), Kirkhorn and Schenker, (2002), Lloveras et al. (2004), Karst et al. (2006), Kowalski et al. (2007), Al-Bahar et al. (2010), Hong et al. (2016), Ashok et al. (2017), Ampatzidou et al. (2018)
Unbalanced	Adenocarcinoma	Lung, Pancreas, Large intestine, Kidney, Breast Ovary	12p13	add(12)(p13)	
Unbalanced	Osteosarcoma	Skeleton	12p13	add(12)(p13)	
Unbalanced	Teratoma (mature and immature)	Testis	12p13	add(12)(p13)	
Unbalanced	Liposarcoma, dedifferentiated	Intraabdominal	16q24	del(16)(q24)	Pedersen et al. (1986), Macarenco et al. (2006)
Unbalanced	Malignant melanoma	Skin	16q24	del(16)(q24)	

and genomic diversity (heterogeneity) and with their involvement in the development of diseases (Rangel et al., 2017; Vargas-Rondon et al., 2017), so identifying and reporting these alterations is clinically relevant. In fact, NCCAs are the key elements that initiate the formation of CCAs (discontinuous interrupted phase) and provide the basis for the formation of diverse populations with clonal changes (gradual phase), thus leading to CIN and CH (Rangel et al., 2017; Vargas-Rondon et al., 2017). In fact, some authors have suggested that although NCCAs are not stable and cannot survive, they provide the genetic variation necessary for macrocellular evolutionary selection and for CH (Liu et al., 2014). A heterogeneity-generating event that could lead to nonclonal structural chromosomal alterations and clonal aneuploidy is the break-fusion-bridge (BFB) cycle. BFB cycles may lead to a considerable intercellular heterogeneity participating in the formation of dicentric chromosomes, ring chromosomes and/ or acentric chromosomes, among others (Gisselsson et al., 2000). At anaphase, such rearranged chromosomes frequently fail to segregate in an orderly manner, instead forming nucleoplasmic bridges (NPB) between the spindle poles (Gisselsson et al., 2001). As result of the formation of NPB, the lagging chromosome may be lost, form a micronucleus (MN), or be randomly incorporated into either of the daughter nuclei, conducing to clonal aneuploidy. Moreover, at the anaphase-telophase transition, these NPB may subsequently break, resulting in novel SCAs in the daughter cells (Gisselsson et al., 2001; Fenech et al., 2011), thus favoring the presence of non-clonal alterations. To highlight that these abnormal nuclear shapes (NPB and MN) have been considered as common features of a wide variety of unstable cells (Gisselsson et al., 2001; Caradonna, 2015). Overall, our results suggest that SCAs appear to play a major role in conferring genetic heterogeneity (NCCAs), potentially surpassing the variability observed at the numerical level (CCAs).

Additionally, the high frequency of CIN and CH observed in this study by using GTG banding was confirmed by using FISH.

FISH allows detecting the appearance of CIN, CH and clonal evolution before it is detected in metaphases. For instance, have been indicated that although the presence of a Philadelphia (Ph) chromosome was identified through the use of banding cytogenetics in peripheral blood and bone marrow samples from patients with chronic myeloid leukemia, the use of FISH assays allowed to identify a certain percentage of cells with an additional Ph + chromosome, not identified by banding cytogenetics (Bentz et al., 1994; Buno et al., 1998), which confirms the usefulness of FISH assays to identify CIN, CH and clonal evolution in peripheral blood samples.

The results obtained in our study using FISH, suggest a negative effect of occupational pesticide exposure on the stability of the chromosomes. FISH results showed that individuals exposed to pesticides have a high level of CIN (≥22.67%) compared to low CIN (≤13.83%) observed in unexposed individuals. The CIN level was 33.5 times higher in the exposed group than in the unexposed group. In addition, we also observed differences in CH levels, being it statistically higher in the exposed group than in the unexposed group. These results suggest that the high CH observed in the exposed individuals, could be the result of the high levels of CIN also presented in these individuals.

To highlight that, CH has not been evaluated in previous studies of occupational exposure to genotoxic agents, therefore, the results of our study are very important, since they show that exposure to pesticides induces CIN and CH, which in addition to reflecting the instability of the system, could predispose cells to acquire additional CIN and, therefore, to a higher risk of malignant transformation (Zhang et al., 2011; Cepeda et al., 2020). In fact, CIN has been recognized as a source of genetic variation that leads to CH, thus favoring the adaptation of cells to stressful environments and the possibility of the development of diseases, mainly cancer (Dayal et al., 2015).

In order to quantify CH, diversity measures adopted from ecology and evolution have been applied, including the SDI, which has been widely used to determine CH in cell lines (Lengauer et al., 1997; Munro et al., 2012). However, some ecologists have suggested that although the SDI is effective for measuring diversity, it does not represent diversity per se, and its misuse could lead to confusion (Jost and Gonzá lez-Oreja, 2012). Thus, we suggest the use of TD as an indicator of CH since it allows us to obtain a more realistic value of heterogeneity.

In line with previous studies (Pastor et al., 2003; Sailaja et al., 2006; Benedetti et al., 2018) we did not find associations between CIN and CH levels with variables such as sex, age, and exposure time (ET). This could suggest that the chromosomal damage induced by pesticides is independent of sex, age, and ET, and highlights the importance of identifying biomarkers that allow monitoring of exposed populations. One such biomarker is the evaluation of CIN and CH by FISH, using centromeric probes for chromosomes 2, 3, 8, 11, 15, and 17. In fact, according to our results, chromosomes 8 and 17 could be excellent biomarkers of chromosomal stability, since these chromosomes did not show great variations in the groups studied. The stability observed in chromosomes 8 and 17 could make it possible to detect damage to the genetic material by observing variations in the number of copies of these chromosomes.

Since most of the farmers who participated in our study were exposed to complex and variable mixtures of pesticides, it is not possible for us to establish whether the CAs, CVs, CIN, and CH observed in the exposed individuals are due to a single pesticide. In fact, even where associations have been seen or suspected, identifying the specific agent responsible has been difficult for a variety of reasons, including the variable exposure levels, and concurrent exposure to multiple pesticides. The above constitute a great problem and concern in public health, considering that some studies have indicated that mixtures of toxics can influence and even amplify the toxicity of the individual components through synergies, potentiation, antagonism, inhibition or effects additives (Mumtaz, 1995; Reffstrup et al., 2010). It is important to highlight that, although a limitation of our study was the impossibility of establishing associations between individual pesticides with the induction of chromosomal alterations (for the reasons indicated above), our results suggest the deleterious effect of the pesticide mixture on chromosomes. In this regard, few *in vitro* and *in vivo* studies have reported associations between some individual pesticides with the induction of chromosomal alterations. For instance, and with regard to the pesticides used most frequently by the exposed individuals included in our study, associations between mancozeb exposure with a significant increase in the frequencies of structural chromosomal alterations and genotoxic damage were reported (Jablonicka et al., 1989; Srivastava et al., 2012). In addition, *in vitro* studies in human lymphocytes demonstrated associations between exposure to paraquat and the production of isochromatic breaks (Jovtchev et al., 2010), as well as between high concentrations of chlorpyrifos with an increase in the number of numerical chromosomal alterations (Serpa et al., 2019), and between sublethal concentrations of profenofos with the induction of chromatid breaks and gaps (Prabhavathy Das et al., 2006). In the same way, *in vivo* cytogenetic analysis demonstrated the induction of chromosomal alterations and micronucleus (MN) formation in mouse bone marrow cells exposed to furadan (Chauhan et al., 2000). Unfortunately, despite the deleterious effect of pesticides on human health, only few studies have investigated the effect of individual pesticides on human chromosomes.

Some chemical classes of pesticides used by the exposed individuals, such as organophosphates and carbamates, have been reported to be genotoxic, generating free radicals that react with cell membranes and initiate the process of lipid peroxidation (Banerjee et al., 1999). In fact, it has been reported that mancozeb, one of the pesticides used by farmers in this study, is a carbamate fungicide commonly used for a wide spectrum of crops (especially soy) and contains a substance with important effects on human health: ethylene(bis) dithiocarbonate (EBCD). EBCD is easily metabolized into ethylene thiourea (ETU), which decreases the activity of tumor suppression proteins, thus facilitating tumor growth (George and Shukla, 2011; Paro et al., 2012). Paraquat, another of the pesticides used by farmers, besides being the second most widely used

prototypical agricultural herbicide (Sabarwal et al., 2018), also been associated with an increased risk of Parkinson's disease, with effects mainly in the liver and kidney (O'Leary et al., 2008), and with pulmonary fibrosis through the generation of ROS (Kirkhorn and Garry, 2000). Overall, pesticides have been associated with deleterious effects on the health of exposed people, including the interfere of the endocrine system and neurobehavioral development (LeBlanc et al., 1997), the development of respiratory symptoms and immunodeficiency (Hoppin et al., 2002), the development of diseases such as breast, lung and pancreatic cancer, lymphomas, among others, which generates a public health problem (Kawauchi et al., 2010; Arafa et al., 2013; Farkas et al., 2016).

The results of this study suggest that occupational exposure to pesticides is associated with CAs, CVs, CIN, and CH in somatic cells of Colombian farmers. Chromosomal damage is an important step in carcinogenesis and the development of many other diseases. Considering that CIN can predispose cells to additional chromosomal alterations (CH) and, therefore, to an increased risk of developing diseases, the monitoring of these markers (CAs, CVs, CIN, and CH) could be useful to estimate the genetic risk in populations exposed to pesticides. Our results highlight the need to develop educational programs aimed at controlling the use of these substances and implementing prevention and protection measures in exposed populations. Therefore, effective efforts are required to support and monitor populations exposed to pesticides, as well as implement more stringent guidelines that help reduce potential genotoxic harm. Further, early detection of chromosomic damage is crucial to implement the necessary measures to reduce or suppress the exposure to deleterious agent when the damage is still reversible, thus reduce the risk to suffer diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Universidad Pedagógica y Tecnológica de Colombia, Tunja (Colombia) (protocol code SGI 3029, April 9, 2021). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, NR and MR-L; data curation, MPM-F, DSV, SC, NR and MR-L; formal analysis, MPM-F, DSV, SC, NR and MR-L; funding acquisition, MR-L; investigation, MPM-F, DSV, SC, NR, and MR-L; methodology, MPM-F, DSV, SC, NR, and MR-L; project administration, MR-L; resources, MR-L and MM-A; supervision, NR and MR-L; validation, MPM-F, DSV, SC, NR and MR-L; visualization, MPM-F, DSV, SC, NR, MF-C, MM-A and MR-L; writing — original draft, MR-L; writing — review and editing, MPM-F, DSV, SC, NR, MF-C, MM-A and MR-L.

FUNDING

This research was funded by Universidad Pedagógica y Tecnológica de Colombia and by Universidad del Rosario.

ACKNOWLEDGMENTS

We thank all the people who were part of this study, especially the farmers of the town of Simijacá, Colombia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.820209/full#supplementary-material

REFERENCES

Al-Bahar, S., Zámečníkova, A., and Pandita, R. (2010). Frequency and Type of Chromosomal Abnormalities in Childhood Acute Lymphoblastic Leukemia Patients in Kuwait: a Six-Year Retrospective Study. *Med. Princ Pract.* 19 (3), 176–181. doi:10.1159/000285281

Amare Kadam, P. S., Ghule, P., Jose, J., Bamne, M., Kurkure, P., Banavali, S., et al. (2004). Constitutional Genomic Instability, Chromosome Aberrations in Tumor Cells and Retinoblastoma. *Cancer Genet. Cytogenet.* 150 (1), 33–43. doi:10.1016/j.cancergencyto.2003.08.015

Ampatzidou, M., Papadhimitriou, S. I., Paterakis, G., Pavlidis, D., Tsitsikas, K., Kostopoulos, I. V., et al. (2018). ETV6/RUNX1-positive Childhood Acute Lymphoblastic Leukemia (ALL): The Spectrum of Clonal Heterogeneity and its Impact on Prognosis. *Cancer Genet.* 224-225, 1–11. doi:10.1016/j.cancergen.2018.03.001

Anwar Iqbal, M., Al-Omar, H. M., Owaidah, T., Al-Humaidan, H., Bhuiyan, Z. A., and Sahovic, E. (2006). del(6)(p23) in two cases of De Novo AML - a new recurrent primary chromosome abnormality. *Eur. J. Haematol.* 77 (3), 245–250. doi:10.1111/j.1600-0609.2006.00698.x

Arafa, A., Afify, M., and Samy, N. (2013). Evaluation of Adverse Health Effects of Pesticides Exposure [Biochemical & Hormonal] Among Egyptian Farmers. *J. Appl. Sci. Res.* 9 (7), 5. Available at: http://www.aensiweb.com/old/jasr/jasr/2013/4404-4409.pdf

Ashok, V., Ranganathan, R., Chander, S., Damodar, S., Bhat, S., Nataraj, K. S., et al. (2017). Comparison of Diagnostic Yield of a FISH Panel against Conventional Cytogenetic Studies for Hematological Malignancies: A South Indian Referral Laboratory Analysis of 201 Cases. *Asian Pac. J. Cancer Prev.* 18 (12), 3457–3464. doi:10.22034/APJCP.2017.18.12.3457

Atkin, N. B. (1977). Chromosome 1 Heteromorphism in Patients with Malignant Disease: a Constitutional Marker for a High-Risk Group. *Bmj* 1 (6057), 358. doi:10.1136/bmj.1.6057.358

Balaji, M., and Sasikala, K. (1993). Cytogenetic Effect of Malathion in *In Vitro* Culture of Human Peripheral Blood. *Mutat. Res. Lett.* 301 (1), 13–17. doi:10.1016/0165-7992(93)90050-6

Banerjee, B. D., Seth, V., Bhattacharya, A., Pasha, S. T., and Chakraborty, A. K. (1999). Biochemical Effects of Some Pesticides on Lipid Peroxidation and Free-Radical Scavengers. *Toxicol. Lett.* 107 (1-3), 33–47. doi:10.1016/s0378-4274(99)00029-6

Bardi, G., Johansson, B., Pandis, N., Bak-Jensen, E., Örndal, C., Heim, S., et al. (1993). Cytogenetic Aberrations in Colorectal Adenocarcinomas and Their Correlation with Clinicopathologic Features. *Cancer* 71 (2), 306–314. doi:10.1002/1097-0142(19930115)71:2<306:aid-cncr2820710207>3.0.co;2-c

Benedetti, D., Lopes Alderete, B., de Souza, C. T., Ferraz Dias, J., Niekraszewicz, L., Cappetta, M., et al. (2018). DNA Damage and Epigenetic Alteration in Soybean Farmers Exposed to Complex Mixture of Pesticides. *Mutagenesis* 33 (1), 87–95. doi:10.1093/mutage/gex035

Bentz, M., Döhner, H., Cabot, G., and Lichter, P. (1994). Fluorescence *In Situ* Hybridization in Leukemias: 'the FISH Are Spawning!' *Leukemia* 8 (9), 1447–1452.

Bhardwaj, J. K., Mittal, M., Saraf, P., and Kumari, P. (2020). Pesticides Induced Oxidative Stress and Female Infertility: a Review. *Toxin Rev.* 39 (1), 1–13. doi:10.1080/15569543.2018.1474926

Bolt, H. M., Foth, H., Hengstler, J. G., and Degen, G. H. (2004). Carcinogenicity Categorization of Chemicals-New Aspects to Be Considered in a European Perspective. *Toxicol. Lett.* 151 (1), 29–41. doi:10.1016/j.toxlet.2004.04.004

Bréga, S. M., Vassilieff, I., Almeida, A., Mercadante, A., Bissacot, D., Cury, P. R., et al. (1998). Clinical, Cytogenetic and Toxicological Studies in Rural Workers Exposed to Pesticides in Botucatu, São Paulo, Brazil. *Cad Saude Publica* 14 (Suppl. 3), 109–115. doi:10.1590/s0102-311x1998000700011

Bridge, J. A., Nelson, M., McComb, E., McGuire, M. H., Rosenthal, H., Vergara, G., et al. (1997). Cytogenetic Findings in 73 Osteosarcoma Specimens and a Review of the Literature. *Cancer Genet. Cytogenet.* 95 (1), 74–87. doi:10.1016/s0165-4608(96)00306-8

Buño, I., Wyatt, W. A., Zinsmeister, A. R., Dietz-Band, J., Silver, R. T., and Dewald, G. W. (1998). A Special Fluorescent *In Situ* Hybridization Technique to Study Peripheral Blood and Assess the Effectiveness of Interferon Therapy in Chronic Myeloid Leukemia. *Blood* 92 (7), 2315–2321.

Buretic-Tomljanovic, A., Badovinac, A. R., Vlastelic, I., and Randic, L. J. (1997). Quantitative Analysis of Constitutive Heterochromatin in Couples with Fetal Wastage. *Am. J. Reprod. Immunol.* 38 (3), 201–204. doi:10.1111/j.1600-0897.1997.tb00299.x

Campioni, D., Bardi, M. A., Cavazzini, F., Tammiso, E., Pezzolo, E., Pregnolato, E., et al. (2012). Cytogenetic and Molecular Cytogenetic Profile of Bone Marrow-Derived Mesenchymal Stromal Cells in Chronic and Acute Lymphoproliferative Disorders. *Ann. Hematol.* 91 (10), 1563–1577. doi:10.1007/s00277-012-1500-8

Caradonna, F. (2015). Nucleoplasmic Bridges and Acrocentric Chromosome Associations as Early Markers of Exposure to Low Levels of Ionising Radiation in Occupationally Exposed Hospital Workers. *Mutagenesis* 30 (2), 269–275. doi:10.1093/mutage/geu068

Carbonell, E., Puig, M., Xamena, N., Creus, A., and Marcos, R. (1990). Sister Chromatid Exchange in Lymphocytes of Agricultural Workers Exposed to Pesticides. *Mutagenesis* 5 (4), 403–406. doi:10.1093/mutage/5.4.403

Castillo-Cadena, J., Tenorio-Vieyra, L. E., Quintana-Carabia, A. I., García-Fabila, M. M., Juan, E. R.-S., and Madrigal-Bujaidar, E. (2006). Determination of DNA Damage in Floriculturists Exposed to Mixtures of Pesticides. *J. Biomed. Biotechnol.* 2006 (2), 1–12. doi:10.1155/JBB/2006/97896

CCMG-CCGM National Office (2021). *CCMG Practice Guidelines For Cytogenetic Analysis. Recommendations For the Indications, Analysis and Reporting of Constitutional Specimens (Peripheral Blood, Solid Tissues)* [Online]. Available: https://www.ccmg-ccgm.org/images/CCMG_practice_guidelines_for_cytogenetic_analysis_B_constitutional_approved_Mar2021.pdf (Accessed January 2, 2022 2022).

Cepeda, S., Forero-Castro, M., Cárdenas-Nieto, D., Martínez-Agüero, M., and Rondón-Lagos, M. (2020). Chromosomal Instability in Farmers Exposed to Pesticides: High Prevalence of Clonal and Non-clonal Chromosomal Alterations. *Rmhp* Vol. 13, 97–110. doi:10.2147/RMHP.S230953

Cerretini, R., Noriega, M. F., Narbaitz, M., and Slavutsky, I. (2006). New Chromosome Abnormalities and Lack of BCL-6 Gene Rearrangements in Argentinean Diffuse Large B-Cell Lymphomas. *Eur. J. Haematol.* 76 (4), 284–293. doi:10.1111/j.1600-0609.2005.00616.x

Chauhan, L. K., Pant, N., Gupta, S. K., and Srivastava, S. P. (2000). Induction of Chromosome Aberrations, Micronucleus Formation and Sperm Abnormalities in Mouse Following Carbofuran Exposure. *Mutat. Res.* 465 (1-2), 123–129. doi:10.1016/s1383-5718(99)00219-3

Cocco, P., Satta, G., Dubois, S., Pili, C., Pilleri, M., Zucca, M., et al. (2013). Lymphoma Risk and Occupational Exposure to Pesticides: Results of the Epilymph Study. *Occup. Environ. Med.* 70 (2), 91–98. doi:10.1136/oemed-2012-100845

Coquelle, A., Pipiras, E., Toledo, F., Buttin, G., and Debatisse, M. (1997). Expression of Fragile Sites Triggers Intrachromosomal Mammalian Gene Amplification and Sets Boundaries to Early Amplicons. *Cell* 89 (2), 215–225. doi:10.1016/s0092-8674(00)80201-9

Dayal, J. H. S., Albergante, L., Newman, T. J., and South, A. P. (2015). Quantitation of Multiclonality in Control and Drug-Treated Tumour Populations Using High-Throughput Analysis of Karyotypic Heterogeneity. *Converg. Sci. Phys. Oncol.* 1 (2), 025001. doi:10.1088/2057-1739/1/2/025001

De Ferrari, M., Artuso, M., Bonassi, S., Bonatti, S., Cavalieri, Z., Pescatore, D., et al. (1991). Cytogenetic Biomonitoring of an Italian Population Exposed to Pesticides: Chromosome Aberration and Sister-Chromatid Exchange Analysis in Peripheral Blood Lymphocytes. *Mutat. Research/Genetic Toxicol.* 260 (1), 105–113. doi:10.1016/0165-1218(91)90086-2

Debacker, K., and Kooy, R. F. (2007). Fragile Sites and Human Disease. *Hum. Mol. Genet.* 16 (Spec No. 2), R150–R158. doi:10.1093/hmg/ddm136

Debiec-Rychter, M., Alwasiak, J., Liberski, P. P., Nedoszytko, B., Babińska, M., Mrózek, K., et al. (1995). Accumulation of Chromosomal Changes in Human Glioma Progression. A Cytogenetic Study of 50 Cases. *Cancer Genet. Cytogenet.* 85 (1), 61–67. doi:10.1016/0165-4608(95)00129-8

Dulout, F. N., Pastori, M. C., Olivero, O. A., González Cid, M., Loria, D., Matos, E., et al. (1985). Sister-chromatid Exchanges and Chromosomal Aberrations in a Population Exposed to Pesticides. *Mutat. Res. Lett.* 143 (4), 237–244. doi:10.1016/0165-7992(85)90087-9

Durkin, S. G., and Glover, T. W. (2007). Chromosome Fragile Sites. *Annu. Rev. Genet.* 41, 169–192. doi:10.1146/annurev.genet.41.042007.165900

El-Zimaity, M. M. T., Kantarjian, H., Talpaz, M., O'Brien, S., Giles, F., Garcia-Manero, G., et al. (2004). Results of Imatinib Mesylate Therapy in Chronic Myelogenous Leukaemia with Variant Philadelphia Chromosome. *Br. J. Haematol.* 125 (2), 187–195. doi:10.1111/j.1365-2141.2004.04899.x

FAO and WHO (2021). *Managing Pesticides in Agriculture and Public Health. A Compendium of FAO and WHO Guidelines and Other Resources*. Second edition. Rome, Italy: FAO and WHO. [Online]. Available: https://www.who.int/publications/i/item/9789240022478 (Accessed, 2021).

Farabegoli, F., Santini, D., Ceccarelli, C., Taffurelli, M., Marrano, D., and Baldini, N. (2001). Clone Heterogeneity in Diploid and Aneuploid Breast Carcinomas as Detected by FISH. *Cytometry* 46 (1), 50–56. doi:10.1002/1097-0320(20010215)46:1<50:aid-cyto1037>3.0.co;2-t

Farkas, G., Jurányi, Z., Székely, G., Kocsis, Z. S., and Gundy, S. (2016). Relationship between Spontaneous Frequency of Aneuploidy and Cancer Risk in 2145 Healthy Hungarian Subjects. *Mutage* 31 (5), 583–588. doi:10.1093/mutage/gew024

Feder, M., Siegfried, J. M., Balshem, A., Litwin, S., Keller, S. M., Liu, Z., et al. (1998). Clinical Relevance of Chromosome Abnormalities in Non-small Cell Lung Cancer. *Cancer Genet. Cytogenet.* 102 (1), 25–31. doi:10.1016/s0165-4608(97)00274-4

Fenech, M., Kirsch-Volders, M., Natarajan, A. T., Surralles, J., Crott, J. W., Parry, J., et al. (2011). Molecular Mechanisms of Micronucleus, Nucleoplasmic Bridge and Nuclear Bud Formation in Mammalian and Human Cells. *Mutagenesis* 26 (1), 125–132. doi:10.1093/mutage/geq052

Fiegl, M., Kaufmann, H., Zojer, N., Schuster, R., Wiener, H., Müllauer, L., et al. (2000). Malignant Cell Detection by Fluorescence *In Situ* Hybridization (FISH) in Effusions from Patients with Carcinoma. *Hum. Pathol.* 31 (4), 448–455. doi:10.1053/hp.2000.6550

Gabrea, A., Martelli, M. L., Qi, Y., Roschke, A., Barlogie, B., Shaughnessy, J. D., Jr., et al. (2008). Secondary Genomic Rearrangements Involving Immunoglobulin or MYC Loci Show Similar Prevalences in Hyperdiploid and Nonhyperdiploid Myeloma Tumors. *Genes Chromosom. Cancer* 47 (7), 573–590. doi:10.1002/gcc.20563

Gagos, S., and Irminger-Finger, I. (2005). Chromosome Instability in Neoplasia: Chaotic Roots to Continuous Growth. *Int. J. Biochem. Cel Biol.* 37 (5), 1014–1033. doi:10.1016/j.biocel.2005.01.003

Geigl, J. B., Obenauf, A. C., Schwarzbraun, T., and Speicher, M. R. (2008). Defining 'chromosomal Instability'. *Trends Genet.* 24 (2), 64–69. doi:10.1016/j.tig.2007.11.006

George, J., and Shukla, Y. (2011). Pesticides and Cancer: Insights into Toxicoproteomic-Based Findings. *J. Proteomics* 74 (12), 2713–2722. doi:10.1016/j.jprot.2011.09.024

Gisselsson, D., Björk, J., Höglund, M., Mertens, F., Dal Cin, P., Åkerman, M., et al. (2001). Abnormal Nuclear Shape in Solid Tumors Reflects Mitotic Instability. *Am. J. Pathol.* 158 (1), 199–206. doi:10.1016/S0002-9440(10)63958-2

Gisselsson, D., Pettersson, L., Hoglund, M., Heidenblad, M., Gorunova, L., Wiegant, J., et al. (2000). Chromosomal Breakage-Fusion-Bridge Events Cause Genetic Intratumor Heterogeneity. *Proc. Natl. Acad. Sci.* 97 (10), 5357–5362. doi:10.1073/pnas.090013497

Gladstone, B., Amare, P. S., Pai, S. K., Gopal, R., Joshi, S., Nair, C. N., et al. (1998). Cytogenetic Studies in Patients from India with T-Acute Lymphoblastic Leukemia. *Cancer Genet. Cytogenet.* 106 (1), 44–48. doi:10.1016/s0165-4608(98)00039-9

Glover, T. W., and Stein, C. K. (1987). Induction of Sister Chromatid Exchanges at Common Fragile Sites. *Am. J. Hum. Genet.* 41 (5), 882–890.

Glover, T. W. (1998). Instability at Chromosomal Fragile Sites. *Recent Results Cancer Res.* 154, 185–199. doi:10.1007/978-3-642-46870-4_11

Gómez-Arroyo, S., Dı´az-Sánchez, Y., Meneses-Pérez, M. A., Villalobos-Pietrini, R., and De León-Rodrı´guez, J. (2000). Cytogenetic Biomonitoring in a Mexican Floriculture Worker Group Exposed to Pesticides. *Mutat. Research/Genetic Toxicol. Environ. Mutagenesis* 466 (1), 117–124. doi:10.1016/s1383-5718(99)00231-4

Grover, P., Danadevi, K., Mahboob, M., Rozati, R., Banu, B. S., and Rahman, M. F. (2003). Evaluation of Genetic Damage in Workers Employed in Pesticide Production Utilizing the Comet Assay. *Mutagenesis* 18 (2), 201–205. doi:10.1093/mutage/18.2.201

Heerema, N. A., Palmer, C. G., Weetman, R., and Bertolone, S. (1992). Cytogenetic Analysis in Relapsed Childhood Acute Lymphoblastic Leukemia. *Leukemia* 6 (3), 185–192.

Hellman, A., Zlotorynski, E., Scherer, S. W., Cheung, J., Vincent, J. B., Smith, D. I., et al. (2002). A Role for Common Fragile Site Induction in Amplification of Human Oncogenes. *Cancer Cell* 1 (1), 89–97. doi:10.1016/s1535-6108(02)00017-x

Hilgert Jacobsen-Pereira, C., Dos Santos, C. R., Troina Maraslis, F., Pimentel, L., Feijó, A. J. L., Iomara Silva, C., et al. (2018). Markers of Genotoxicity and Oxidative Stress in Farmers Exposed to Pesticides. *Ecotoxicology Environ. Saf.* 148, 177–183. doi:10.1016/j.ecoenv.2017.10.004

Hong, M., Hao, S., Patel, K. P., Kantarjian, H. M., Garcia-Manero, G., Yin, C. C., et al. (2016). Whole-arm translocation of der(5;17)(p10;q10) with concurrent TP53 mutations in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS): A unique molecular-cytogenetic subgroup. *Cancer Genet.* 209 (5), 205–214. doi:10.1016/j.cancergen.2016.04.001

Hoogerwerf, W. A., Hawkins, A. L., Griffin, C. A., and Perlman, E. J. (1994). Chromosome Analysis of Nine Osteosarcomas. *Genes Chromosom. Cancer* 9 (2), 88–92. doi:10.1002/gcc.2870090203

Hoppin, J. A., Umbach, D. M., London, S. J., Alavanja, M. C. R., and Sandler, D. P. (2002). Chemical Predictors of Wheeze Among Farmer Pesticide Applicators in the Agricultural Health Study. *Am. J. Respir. Crit. Care Med.* 165 (5), 683–689. doi:10.1164/ajrccm.165.5.2106074

Idrovo, A. J. (2000). Surveillance of Pesticide Poisoning in Colombia. *Revista de Salud Pública* 2 (1), 10.

Issa, G. C., Kantarjian, H. M., Gonzalez, G. N., Borthakur, G., Tang, G., Wierda, W., et al. (2017). Clonal Chromosomal Abnormalities Appearing in Philadelphia Chromosome-Negative Metaphases during CML Treatment. *Blood* 130 (19), 2084–2091. doi:10.1182/blood-2017-07-792143

Jablonická, A., Poláková, H., Karelová, J., and Vargová, M. (1989). Analysis of Chromosome Aberrations and Sister-Chromatid Exchanges in Peripheral Blood Lymphocytes of Workers with Occupational Exposure to the Mancozeb-Containing Fungicide Novozir Mn80. *Mutat. Research/Genetic Toxicol.* 224 (2), 143–146. doi:10.1016/0165-1218(89)90148-1

Ji, W., Hernandez, R., Zhang, X.-Y., Qu, G.-z., Frady, A., Varela, M., et al. (1997). DNA Demethylation and Pericentromeric Rearrangements of Chromosome 1. *Mutat. Research/Fundamental Mol. Mech. Mutagenesis* 379 (1), 33–41. doi:10.1016/s0027-5107(97)00088-2

Jost, L. (2006). Entropy and Diversity. *Oikos* 113 (2), 363–375. doi:10.1111/j.2006.0030-1299.14714.x

Jost, L., and González-Oreja, J. (2012). Measuring Biological Diversity: Beyond the Shannon index. *Acta Zoológica Lilloana* 56, 11. doi:10.30550/j.azl

Jovtchev, G., Gateva, S., Stergios, M., and Kulekova, S. (2010). Cytotoxic and Genotoxic Effects of Paraquat in Hordeum Vulgare and Human Lymphocytes *In Vitro*. *Environ. Toxicol.* 25 (3), 294–303. doi:10.1002/tox.20503

Karst, C., Gross, M., Haase, D., Wedding, U., Höffken, K., Liehr, T., et al. (2006). Novel Cryptic Chromosomal Rearrangements Detected in Acute Lymphoblastic Leukemia Detected by Application of New Multicolor Fluorescent *In Situ* Hybridization Approaches. *Int. J. Oncol.* 28 (4), 891–897. doi:10.3892/ijo.28.4.891

Kaur, R., and Kaur, K. (2018). Occupational Pesticide Exposure, Impaired DNA Repair, and Diseases. *Indian J. Occup. Environ. Med.* 22 (2), 74–81. doi:10.4103/ijoem.IJOEM_45_18

Kawauchi, S., Furuya, T., Ikemoto, K., Nakao, M., Yamamoto, S., Oka, M., et al. (2010). DNA Copy Number Aberrations Associated with Aneuploidy and Chromosomal Instability in Breast Cancers. *Oncol. Rep.* 24 (4), 875–883. doi:10.3892/or.2010.875

Kirkhorn, S. R., and Garry, V. F. (2000). Agricultural Lung Diseases. *Environ. Health Perspect.* 108 (Suppl. 4), 705–712. doi:10.1289/ehp.00108s4705

Kirkhorn, S. R., and Schenker, M. B. (2002). Current Health Effects of Agricultural Work: Respiratory Disease, Cancer, Reproductive Effects, Musculoskeletal Injuries, and Pesticideâ€"Related Illnesses. *J. Agric. Saf. Health* 8 (2), 199–214. doi:10.13031/2013.8432

Kowalski, J., Morsberger, L. A., Blackford, A., Hawkins, A., Yeo, C. J., Hruban, R. H., et al. (2007). Chromosomal Abnormalities of Adenocarcinoma of the Pancreas: Identifying Early and Late Changes. *Cancer Genet. Cytogenet.* 178 (1), 26–35. doi:10.1016/j.cancergencyto.2007.06.004

LeBlanc, G. A., Bain, L. J., and Wilson, V. S. (1997). Pesticides: Multiple Mechanisms of Demasculinization. *Mol. Cell Endocrinol.* 126 (1), 1–5. doi:10.1016/s0303-7207(96)03968-8

Lee, S.-E., Choi, S. Y., Bang, J.-H., Kim, S.-H., Jang, E.-j., Byeun, J.-Y., et al. (2012). The Long-Term Clinical Implications of Clonal Chromosomal Abnormalities in Newly Diagnosed Chronic Phase Chronic Myeloid Leukemia Patients Treated with Imatinib Mesylate. *Cancer Genet.* 205 (11), 563–571. doi:10.1016/j.cancergen.2012.09.003

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1997). Genetic Instability in Colorectal Cancers. *Nature* 386 (6625), 623–627. doi:10.1038/386623a0

Liu, G., Stevens, J., Horne, S., Abdallah, B., Ye, K., Bremer, S., et al. (2014). Genome Chaos: Survival Strategy during Crisis. *Cell Cycle* 13 (4), 528–537. doi:10.4161/cc.27378

Lloveras, E., Granada, I., Zamora, L., Espinet, B., Florensa, L., Besses, C., et al. (2004). Cytogenetic and Fluorescence *In Situ* Hybridization Studies in 60 Patients with Multiple Myeloma and Plasma Cell Leukemia. *Cancer Genet. Cytogenet.* 148 (1), 71–76. doi:10.1016/s0165-4608(03)00233-4

Lushchak, V. I., Matviishyn, T. M., Husak, V. V., Storey, J. M., and Storey, K. B. (2018). Pesticide Toxicity: a Mechanistic Approach. *EXCLI J.* 17, 1101–1136. doi:10.17179/excli2018-1710

Macarenco, R. S., Erickson-Johnson, M., Wang, X., Jenkins, R. B., Nascimento, A. G., and Oliveira, A. M. (2006). Cytogenetic and Molecular Genetic Findings in Dedifferentiated Liposarcoma with Neural-like Whorling Pattern and Metaplastic Bone Formation. *Cancer Genet. Cytogenet.* 171 (2), 126–129. doi:10.1016/j.cancergencyto.2006.07.009

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., et al. (2006). Genetic Clonal Diversity Predicts Progression to Esophageal Adenocarcinoma. *Nat. Genet.* 38 (4), 468–473. doi:10.1038/ng1768

Mandahl, N., Fletcher, C. D. M., Dal Cin, P., De Wever, I., Mertens, F., Mitelman, F., et al. (2000). Comparative Cytogenetic Study of Spindle Cell and Pleomorphic Leiomyosarcomas of Soft Tissues. *Cancer Genet. Cytogenet.* 116 (1), 66–73. doi:10.1016/s0165-4608(99)00114-4

Mattiuzzo, M., Fiore, M., Ricordy, R., and Degrassi, F. (2006). Aneuploidy-inducing Capacity of Two Widely Used Pesticides. *Carcinogenesis* 27 (12), 2511–2518. doi:10.1093/carcin/bgl102

McGowan-Jordan, J., Hastings, R. J., and Moore, S. (2020). *ISCN 2020. An International System for Human Cytogenomic Nomenclature*. Basel, Switzerland: Karger.

Mugneret, F., Dastugue, N., Favre, B., Sidaner, I., Salles, B., Huguet-Rigal, F., et al. (1995). Der(16)t(l;16)(qll;qll) in Myelodysplastic Syndromes: a New Non-random Abnormality Characterized by Cytogenic and Fluorescence *In Situ*

Hybridization Studies. *Br. J. Haematol.* 90 (1), 119–124. doi:10.1111/j.1365-2141.1995.tb03389.x

Mumtaz, M. M. (1995). Risk Assessment of Chemical Mixtures from a Public Health Perspective. *Toxicol. Lett.* 82-83, 527–532. doi:10.1016/0378-4274(95)03582-6

Munro, A. F., Twelves, C., Thomas, J. S., Cameron, D. A., and Bartlett, J. M. (2012). Chromosome Instability and Benefit from Adjuvant Anthracyclines in Breast Cancer. *Br. J. Cancer* 107 (1), 71–74. doi:10.1038/bjc.2012.232

Nahi, H., Lehmann, S., Bengtzen, S., Jansson, M., Möllgård, L., Paul, C., et al. (2008). Chromosomal Aberrations in 17p Predict *In Vitro* Drug Resistance and Short Overall Survival in Acute Myeloid Leukemia. *Leuk. Lymphoma* 49 (3), 508–516. doi:10.1080/10428190701861645

Nayebbagher, T., Pashaiefar, H., Yaghmaie, M., Alimoghaddam, K., Jalili, M., Esfandbod, M., et al. (2020). Chromosomal Aberrations in Ascetic Fluid of Metastatic Gastric Cancer Patients: A Clustering Analysis. *neo* 67 (1), 185–192. doi:10.4149/neo_2019_190202N105

Nicolopoulou-Stamati, P., Maipas, S., Kotampasi, C., Stamatis, P., and Hens, L. (2016). Chemical Pesticides and Human Health: The Urgent Need for a New Concept in Agriculture. *Front. Public Health* 4, 148. doi:10.3389/fpubh.2016.00148

O'Leary, K. T., Parameswaran, N., Johnston, L. C., McIntosh, J. M., Di Monte, D. A., and Quik, M. (2008). Paraquat Exposure Reduces Nicotinic Receptor-Evoked Dopamine Release in Monkey Striatum. *J. Pharmacol. Exp. Ther.* 327 (1), 124–129. doi:10.1124/jpet.108.141861

Olsson, L., Lundin-Ström, K. B., Castor, A., Behrendtz, M., Biloglav, A., Norén-Nyström, U., et al. (2018). Improved Cytogenetic Characterization and Risk Stratification of Pediatric Acute Lymphoblastic Leukemia Using Single Nucleotide Polymorphism Array Analysis: A Single center Experience of 296 Cases. *Genes Chromosomes Cancer* 57 (11), 604–607. doi:10.1002/gcc.22664

Ozkan, E., and Marcelo, M. P. (2021). in *Genetics, Cytogenetic Testing and Conventional Karyotype* (Treasure Island, FL: StatPearls Publishing). [Internet].

Pandis, N., Jin, Y., Gorunova, L., Petersson, C., Bardi, G., Idvall, I., et al. (1995). Chromosome Analysis of 97 Primary Breast Carcinomas: Identification of Eight Karyotypic Subgroups. *Genes Chromosom. Cancer* 12 (3), 173–185. doi:10.1002/gcc.2870120304

Pantou, D., Rizou, H., Tsarouha, H., Pouli, A., Papanastasiou, K., Stamatellou, M., et al. (2005). Cytogenetic Manifestations of Multiple Myeloma Heterogeneity. *Genes Chromosom. Cancer* 42 (1), 44–57. doi:10.1002/gcc.20114

Paro, R., Tiboni, G. M., Buccione, R., Rossi, G., Cellini, V., Canipari, R., et al. (2012). The Fungicide Mancozeb Induces Toxic Effects on Mammalian Granulosa Cells. *Toxicol. Appl. Pharmacol.* 260 (2), 155–161. doi:10.1016/j.taap.2012.02.005

Parry, E. M., Parry, J. M., Corso, C., Doherty, A., Haddad, F., Hermine, T. F., et al. (2002). Detection and Characterization of Mechanisms of Action of Aneugenic Chemicals. *Mutagenesis* 17, 509–521. doi:10.1093/mutage/17.6.509

Pastor, S., Creus, A., Parron, T., Cebulska-Wasilewska, A., Siffel, C., Piperakis, S., et al. (2003). Biomonitoring of Four European Populations Occupationally Exposed to Pesticides: Use of Micronuclei as Biomarkers. *Mutagenesis* 18 (3), 249–258. doi:10.1093/mutage/18.3.249

Paz-y-Miño, C., Bustamante, G., Sánchez, M. E., and Leone, P. E. (2002). Cytogenetic Monitoring in a Population Occupationally Exposed to Pesticides in Ecuador. *Environ. Health Perspect.* 110, 3. doi:10.1289/ehp.021101077

Pedersen, M. I., Bennett, J. W., and Wang, N. (1986). Nonrandom Chromosome Structural Aberrations and Oncogene Loci in Human Malignant Melanoma. *Cancer Genet. Cytogenet.* 20 (1-2), 11–27. doi:10.1016/0165-4608(86)90103-2

Pejovic, T., Heim, S., Mandahl, N., Elmfors, B., Furgyik, S., Flodérus, U.-M., et al. (1991). Bilateral Ovarian Carcinoma: Cytogenetic Evidence of Unicentric Origin. *Int. J. Cancer* 47 (3), 358–361. doi:10.1002/ijc.2910470308

Polito, L., Greco, A., and Seripa, D. (2016). Genetic Profile, Environmental Exposure, and Their Interaction in Parkinson's Disease. *Parkinson's Dis.* 2016, 1–9. doi:10.1155/2016/6465793

Prabhavathy Das, G., Pasha Shaik, A., and Jamil, K. (2006). Cytotoxicity and Genotoxicity Induced by the Pesticide Profenofos on Cultured Human Peripheral Blood Lymphocytes. *Drug Chem. Toxicol.* 29 (3), 313–322. doi:10.1080/01480540600653093

Presti, J. C., Jr., Rao, P. H., Chen, Q., Reuter, V. E., Li, F. P., Fair, W. R., et al. (1991). Histopathological, Cytogenetic, and Molecular Characterization of Renal Cortical Tumors. *Cancer Res.* 51 (5), 1544–1552.

Prigogina, E. L., Puchkova, G. P., and Mayakova, S. A. (1988). Nonrandom Chromosomal Abnormalities in Acute Lymphoblastic Leukemia of Childhood. *Cancer Genet. Cytogenet.* 32 (2), 183–203. doi:10.1016/0165-4608(88)90281-6

Rangel, N., Forero-Castro, M., and Rondón-Lagos, M. (2017). New Insights in the Cytogenetic Practice: Karyotypic Chaos, Non-clonal Chromosomal Alterations and Chromosomal Instability in Human Cancer and Therapy Response. *Genes* 8 (6), 155. doi:10.3390/genes8060155

Rayeroux, K. C., and Campbell, L. J. (2009). Gene Amplification in Myeloid Leukemias Elucidated by Fluorescence *In Situ* Hybridization. *Cancer Genet. Cytogenet.* 193 (1), 44–53. doi:10.1016/j.cancergencyto.2009.04.006

Re, A., Cora, D., Puliti, A. M., Caselle, M., and Sbrana, I. (2006). Correlated Fragile Site Expression Allows the Identification of Candidate Fragile Genes Involved in Immunity and Associated with Carcinogenesis. *BMC Bioinformatics* 7, 413. doi:10.1186/1471-2105-7-413

Reffstrup, T. K., Larsen, J. C., and Meyer, O. (2010). Risk Assessment of Mixtures of Pesticides. Current Approaches and Future Strategies. *Regul. Toxicol. Pharmacol.* 56 (2), 174–192. doi:10.1016/j.yrtph.2009.09.013

Renzi, L., Pacchierotti, F., and Russo, A. (1996). The Centromere as a Target for the Induction of Chromosome Damage in Resting and Proliferating Mammalian Cells: Assessment of Mitomycin C-Induced Genetic Damage at Kinetochores and Centromeres by a Micronucleus Test in Mouse Splenocytes. *Mutagenesis* 11 (2), 133–138. doi:10.1093/mutage/11.2.133

Rigolin, G. M., Cuneo, A., Roberti, M. G., Bardi, A., and Castoldi, G. (1997). Myelodysplastic Syndromes with Monocytic Component: Hematologic and Cytogenetic Characterization. *Haematologica* 82 (1), 25–30.

Rodriguez, E., Houldsworth, J., Reuter, V. E., Meltzer, P., Zhang, J., Trent, J. M., et al. (1993). Molecular Cytogenetic Analysis of I(12p)-Negative Human Male Germ Cell Tumors. *Genes Chromosom. Cancer* 8 (4), 230–236. doi:10.1002/gcc.2870080405

Rogatto, S. R., Casartelli, C., Rainho, C. A., and Barbieri-Neto, J. (1993). Chromosomes in the Genesis and Progression of Ependymomas. *Cancer Genet. Cytogenet.* 69 (2), 146–152. doi:10.1016/0165-4608(93)90093-2

Roylance, R., Endesfelder, D., Gorman, P., Burrell, R. A., Sander, J., Tomlinson, I., et al. (2011). Relationship of Extreme Chromosomal Instability with Long-Term Survival in a Retrospective Analysis of Primary Breast Cancer. *Cancer Epidemiol. Biomarkers Prev.* 20 (10), 2183–2194. doi:10.1158/1055-9965.EPI-11-0343

Rupa, D. S., Reddy, P. P., Sreemannarayana, K., Reddi, O. S., and Galloway, S. M. (1991). Frequency of Sister Chromatid Exchange in Peripheral Lymphocytes of Male Pesticide Applicators. *Environ. Mol. Mutagen.* 18 (2), 136–138. doi:10.1002/em.2850180209

Sabarwal, A., Kumar, K., and Singh, R. P. (2018). Hazardous Effects of Chemical Pesticides on Human Health-Cancer and Other Associated Disorders. *Environ. Toxicol. Pharmacol.* 63, 103–114. doi:10.1016/j.etap.2018.08.018

Sailaja, N., Chandrasekhar, M., Rekhadevi, P. V., Mahboob, M., Rahman, M. F., Vuyyuri, S. B., et al. (2006). Genotoxic Evaluation of Workers Employed in Pesticide Production. *Mutat. Research/Genetic Toxicol. Environ. Mutagenesis* 609 (1), 74–80. doi:10.1016/j.mrgentox.2006.06.022

Sawyer, J. R., Roloson, G. J., Bell, J. M., Thomas, J. R., Teo, C., and Chadduck, W. M. (1996). Telomeric Associations in the Progression of Chromosome Aberrations in Pediatric Solid Tumors. *Cancer Genet. Cytogenet.* 90 (1), 1–13. doi:10.1016/0165-4608(96)00058-1

Sawyer, J. R., Tian, E., Heuck, C. J., Epstein, J., Johann, D. J., Swanson, C. M., et al. (2014). Jumping Translocations of 1q12 in Multiple Myeloma: a Novel Mechanism for Deletion of 17p in Cytogenetically Defined High-Risk Disease. *Blood* 123 (16), 2504–2512. doi:10.1182/blood-2013-12-546077

Serpa, E. A., Schmitt, E. G., Zuravski, L., Machado, M. M., and Oliveira, L. F. S. d. (2019). Chlorpyrifos Induces Genotoxic Effects in Human Leukocytes *In Vitro* at Low Concentrations. *Acta Sci. Health Sci.* 41 (1), 44291. doi:10.4025/actascihealthsci.v41i1.44291

Shah, H. K., Sharma, T., and Banerjee, B. D. (2020). Organochlorine Pesticides Induce Inflammation, ROS Production, and DNA Damage in Human Epithelial Ovary Cells: An *In Vitro* Study. *Chemosphere* 246, 125691. doi:10.1016/j.chemosphere.2019.125691

Smolarek, T. A., Blough, R. I., Foster, R. S., Ulbright, T. M., Palmer, C. G., and Heerema, N. A. (1999). Cytogenetic Analyses of 85 Testicular Germ Cell Tumors. *Cancer Genet. Cytogenet.* 108 (1), 57–69. doi:10.1016/s0165-4608(98)00113-7

Srivastava, A. K., Ali, W., Singh, R., Bhui, K., Tyagi, S., Al-Khedhairy, A. A., et al. (2012). Mancozeb-induced Genotoxicity and Apoptosis in Cultured Human Lymphocytes. *Life Sci.* 90 (21-22), 815–824. doi:10.1016/j.lfs.2011.12.013

Takami, S., Kawasome, C., Kinoshita, M., Koyama, H., and Noguchi, S. (2001). Chromosomal Instability Detected by Fluorescence *In Situ* Hybridization in Japanese Breast Cancer Patients. *Clin. Chim. Acta* 308 (1-2), 127–131. doi:10.1016/s0009-8981(01)00473-9

Takeshita, A., Naito, K., Shinjo, K., Sahara, N., Matsui, H., Ohnishi, K., et al. (2004). Deletion 6p23 and Add(11)(p15) Leading to NUP98 Translocation in a Case of Therapy-Related Atypical Chronic Myelocytic Leukemia Transforming to Acute Myelocytic Leukemia. *Cancer Genet. Cytogenet.* 152 (1), 56–60. doi:10.1016/j.cancergencyto.2003.10.002

Talamo, A., Chalandon, Y., Marazzi, A., and Jotterand, M. (2010). Clonal Heterogeneity and Chromosomal Instability at Disease Presentation in High Hyperdiploid Acute Lymphoblastic Leukemia. *Cancer Genet. Cytogenet.* 203 (2), 209–214. doi:10.1016/j.cancergencyto.2010.09.005

Tanaka, K., and Hirota, T. (2016). Chromosomal Instability: A Common Feature and a Therapeutic Target of Cancer. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1866 (1), 64–75. doi:10.1016/j.bbcan.2016.06.002

Teixeira, M. R., Tsarouha, H., Kraggerud, S. M., Pandis, N., Dimitriadis, E., Andersen, J. A., et al. (2001). Evaluation of Breast Cancer Polyclonality by Combined Chromosome Banding and Comparative Genomic Hybridization Analysis. *Neoplasia* 3 (3), 204–214. doi:10.1038/sj.neo.7900152

Testa, J. R., Siegfried, J. M., Liu, Z., Hunt, J. D., Feder, M. M., Litwin, S., et al. (1994). Cytogenetic Analysis of 63 Non-small Cell Lung Carcinomas: Recurrent Chromosome Alterations amid Frequent and Widespread Genomic Upheaval. *Genes Chromosom. Cancer* 11 (3), 178–194. doi:10.1002/gcc.2870110307

Tibiletti, M. G., Bernasconi, B., Furlan, D., Riva, C., Trubia, M., Buraggi, G., et al. (1996). Early Involvement of 6q in Surface Epithelial Ovarian Tumors. *Cancer Res.* 56 (19), 4493–4498.

Tibiletti, M. G., Sessa, F., Bernasconi, B., Cerutti, R., Broggi, B., Furlan, D., et al. (2000). A Large 6q Deletion Is a Common Cytogenetic Alteration in Fibroadenomas, Pre-malignant Lesions, and Carcinomas of the Breast. *Clin. Cancer Res.* 6 (4), 1422–1431.

Tibiletti, M. G., Bernasconi, B., Taborelli, M., Facco, C., Riva, C., Capella, C., et al. (2003). Genetic and Cytogenetic Observations Among Different Types of Ovarian Tumors Are Compatible with a Progression Model Underlying Ovarian Tumorigenesis. *Cancer Genet. Cytogenet.* 146 (2), 145–153. doi:10.1016/s0165-4608(03)00134-1

Tomiazzi, J. S., Judai, M. A., Nai, G. A., Pereira, D. R., Antunes, P. A., and Favareto, A. P. A. (2018). Evaluation of Genotoxic Effects in Brazilian Agricultural Workers Exposed to Pesticides and Cigarette Smoke Using Machine-Learning Algorithms. *Environ. Sci. Pollut. Res.* 25 (2), 1259–1269. doi:10.1007/s11356-017-0496-y

Travella, A., Ripollés, L., Aventin, A., Rodríguez, A., Bezares, R. F., Caballín, M. R., et al. (2013). Structural Alterations in Chronic Lymphocytic Leukaemia. Cytogenetic and FISH Analysis. *Hematol. Oncol.* 31 (2), 79–87. doi:10.1002/hon.2025

Tse, W., Zhu, W., Chen, H., and Cohen, A. (1995). A Novel Gene, AF1q, Fused to MLL in T(1;11) (Q21;q23), Is Specifically Expressed in Leukemic and Immature Hematopoietic Cells. *Blood* 85 (3), 650–656. doi:10.1182/blood.v85.3.650.bloodjournal853650

Vargas-Rondón, N., Villegas, V., and Rondón-Lagos, M. (2017). The Role of Chromosomal Instability in Cancer and Therapeutic Responses. *Cancers* 10 (1), 4. doi:10.3390/cancers10010004

Vincent-Salomon, A., Benhamo, V., Gravier, E., Rigaill, G., Gruel, N., Robin, S., et al. (2013). Genomic Instability: a Stronger Prognostic Marker Than Proliferation for Early Stage Luminal Breast Carcinomas. *PLoS One* 8 (10), e76496. doi:10.1371/journal.pone.0076496

Wang, R., Lu, Y.-J., Fisher, C., Bridge, J. A., and Shipley, J. (2001). Characterization of Chromosome Aberrations Associated with Soft-Tissue Leiomyosarcomas by Twenty-Four-Color Karyotyping and Comparative Genomic Hybridization Analysis. *Genes Chromosom. Cancer* 31 (1), 54–64. doi:10.1002/gcc.1118

Wilhelm, C. M., Calsing, A. K., and da Silva, L. B. (2015). Assessment of DNA Damage in Floriculturists in Southern Brazil. *Environ. Sci. Pollut. Res.* 22 (11), 8182–8189. doi:10.1007/s11356-014-3959-4

Wuicik, L., Cavalli, L. R., Cornélio, D. A., Schmid Braz, A. T., Barbosa, M. L., Lima, R. S., et al. (2007). Chromosome Alterations Associated with Positive and Negative Lymph Node Involvement in Breast Cancer. *Cancer Genet. Cytogenet.* 173 (2), 114–121. doi:10.1016/j.cancergencyto.2006.10.009

Wyandt He, P. R. (2004). "Heteromorphisms in Clinical Populations," in *Atlas of Human Chromosome Heteromorphismss*. Editor V. S.T. H E. Wyandt (Switzerland: Springer Netherlands).

Zhang, W., Jiang, F., and Ou, J. (2011). Global Pesticide Consumption and Pollution: With China as a Focus. *Proc. Int. Acad. Ecol. Environ. Sci.* 1 (2), 19. Available at: http://www.iaees.org/publications/journals/piaees/articles/2011-1(2)/Global-pesticide-consumption-pollution.pdf

Zijno, A., Marcon, F., Leopardi, P., and Crebelli, R. (1996). Analysis of Chromosome Segregation in Cytokinesis-Blocked Human Lymphocytes: Non-disjunction Is the Prevalent Damage Resulting from Low Dose Exposure to Spindle Poisons. *Mutagenesis* 11 (4), 335–340. doi:10.1093/mutage/11.4.335

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 March 2022
doi: 10.3389/fgene.2022.766492

Copy Number Variation Analysis of Euploid Pregnancy Loss

*Chongjuan Gu[1†], Huan Gao[2†], Kuanrong Li[3], Xinyu Dai[4], Zhao Yang[5], Ru Li[6], Canliang Wen[1] and Yaojuan He[1]**

[1]Department of Obstetrics and Gynecology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China, [2]Department of Toxicology, School of Public Health, Sun Yat-sen University, Guangzhou, China, [3]Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China, [4]School of Life Sciences, South China Normal University, Guangzhou, China, [5]West China Hospital, Sichuan University, Chengdu, China, [6]Prenatal Diagnostic Center, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China*

Objectives: Copy number variant (CNV) is believed to be the potential genetic cause of pregnancy loss. However, CNVs less than 3 Mb in euploid products of conceptions (POCs) remain largely unexplored. The aim of this study was to investigate the features of CNVs less than 3 Mb in POCs and their potential clinical significance in pregnancy loss/fetal death.

Methods: CNV data were extracted from a cohort in our institution and 19 peer-reviewed publications, and only those CNVs less than 3 Mb detected in euploid pregnancy loss/fetal death were included. We conducted a CNV map to analyze the distribution of CNVs in chromosomes using R packages karyoploteR_1.10.5. Gene names and annotated gene types covered by those CNVs were mined from the human Release 19 reference genome file and GENECODE database. We assessed the expression patterns and the consequences of murine knock-out of those genes using TiGER and Mouse Genome Informatics (MGI) databases. Functional enrichment and pathway analysis for genes in CNVs were performed using clusterProfiler V3.12.0.

Result: Breakpoints of 564 CNVs less than 3 Mb were obtained from 442 euploid POCs, with 349 gains and 185 losses. The CNV map showed that CNVs were distributed in all chromosomes, with the highest frequency detected in chromosome 22 and the lowest frequency in chromosome Y, and CNVs showed a higher density in the pericentromeric and sub-telomeric regions. A total of 5,414 genes mined from the CNV regions (CNVRs), Gene Ontology (GO), and pathway analysis showed that the genes were significantly enriched in multiple terms, especially in sensory perception, membrane region, and tight junction. A total of 995 protein-coding genes have been reported to present mammalian phenotypes in MGI, and 276 of them lead to embryonic lethality or abnormal embryo/placenta in knock-out mouse models. CNV located at 19p13.3 was the most common CNV of all POCs.

Conclusion: CNVs less than 3 Mb in euploid POCs distribute unevenly in all chromosomes, and a higher density was seen in the pericentromeric and sub-telomeric regions. The genes in those CNVRs are significantly enriched in biological processes and pathways that are important to embryonic/fetal development. CNV in 19p13.3 and the variations of *ARID3A* and *FSTL3* might contribute to pregnancy loss.

Keywords: copy number variant, products of conception, pregnancy loss, chromosomal array, bioinformatics, genome, fetal death

OPEN ACCESS

Edited by:
Cynthia Casson Morton, Brigham and Women's Hospital, United States

Reviewed by:
Baoheng Gui, The Second Affiliated Hospital of Guangxi Medical University, China
Darren Karl Griffin, University of Kent, United Kingdom

****Correspondence:***
Yaojuan He
miracle_he@126.com

†These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *29 August 2021*
Accepted: *24 February 2022*
Published: *23 March 2022*

Citation:
Gu C, Gao H, Li K, Dai X, Yang Z, Li R, Wen C and He Y (2022) Copy Number Variation Analysis of Euploid Pregnancy Loss. Front. Genet. 13:766492. doi: 10.3389/fgene.2022.766492

BACKGROUND

Approximately 15–20% of clinically recognized pregnancies end in pregnancy loss (Practice Committee of the American Society for Reproductive Medicine, 2012; ESHRE Guideline Group on RPL et al., 2018), and the etiology is complicated. It is evident that there are many genetic and environmental factors that are essential for a successful pregnancy, and disruption of any of them could cause pregnancy loss (Yamada et al., 2005). From the genetic perspective, abnormal number and structure of chromosomes are clearly pathogenic genetic causes, and smaller copy number variant (CNV) and mutations in genes that are important for early fetal development are also the potential genetic causes (Colley et al., 2019).

The array-based detection has been used to detect the chromosomal abnormalities of pregnancy loss owing to its higher resolution and detection rates (Hillman et al., 2011; Dhillon et al., 2014). Meanwhile, the array-based detection allows unbiased search for CNVs across the whole genome, which involves unbalanced rearrangements that increase or decrease the DNA content. CNV is associated with a wide range of human diseases, including congenital anomalies and neurodevelopmental disorders (Grayton et al., 2012; Wapner et al., 2012; Dong et al., 2016). However, owing to limited data, it is challenging for clinicians and geneticists to interpret CNVs detected in POCs. Those "pathogenic CNVs" are based on individuals with neurodevelopmental disorders and/or congenital anomalies or fetuses with ultrasound abnormalities (Riggs et al., 2020), as well as on healthy population [e.g., Database of Genomic Variants (DGV) and the 1,000 Genomes database] (Lee and Scherer, 2010; MacDonald et al., 2014), which cannot accurately interpret CNVs in demised embryos/fetuses. There is a continuous spectrum of phenotypic effects of CNV, varying from adaptive and maladaptive traits to embryonic lethality (Beckmann et al., 2007; Hurles et al., 2008). Most of the CNVs less than 3 Mb have been believed not to be associated with adverse phenotypes among healthy individuals (Zarrei et al., 2015). However, the roles of these CNVs less than 3 Mb played in pregnancy loss remain largely unexplored. We suppose that some of the small-sized CNVs detected in POCs involving embryonic lethal or placental function-specific genes have never been reported in DGV and might contribute to pregnancy loss/fetal death.

To further understand the features of CNVs less than 3 Mb detected in POCs and potential clinical roles of those CNVs in euploid pregnancy loss, we constructed a CNV map based on the data obtained from our samples and reported in the literature and analyzed the gene content and function *in silico*.

MATERIALS AND METHODS

Cohort Copy Number Variant Data

The first part of CNV data was extracted from a retrospective, hospital-based cohort of the Guangzhou Women and Children's Medical Center, a tertiary referral hospital in South China. The study protocol was approved by the Ethics Committee of the institute (2020-15001). All patients provided a written informed consent for the tests and the inclusion of results in research. All women were Han Chinese who experienced clinically confirmed pregnancy loss or fetal death according to the guideline (Doubilet et al., 2013) and underwent chromosomal microarray analysis (CMA) detection of the fresh POC sample in our hospital. The methods used for DNA extraction, maternal cell contamination test, and CMA platform have been reported in our previous publication (Gu et al., 2021). The reporting threshold of the copy number result was set at 100 kb with marker count ≥50 bp. Data were visualized and analyzed with the Chromosome Analysis Suite (ChAS) software (Affymetrix, Santa Clara, CA) based on the GRCh37/hg19 assembly. In this study, only those euploid POCs with CNV size less than 3 Mb were included.

Published Copy Number Variant Data and Quality Control

The second part of CNV data was extracted from the peer-reviewed publications. The literature search was focused on studies using microarrays and next-generation sequencing (NGS) to detect POC following pregnancy loss or fetal death. PubMed, Medline, Embase, and CNKI databases were searched electronically, with the last search updated on 30 September 2020. The complete search string is outlined in **Supplementary Table S1**. Data included in this study must meet the following criteria: 1) the subjects of the study were POCs of pregnancy loss or fetal death; 2) the methods of detection were genome-wide assessment and estimated breakpoint resolution. Studies or data would be excluded if the chromosomal karyotype was aneuploid or if the CNV length was longer than 3 Mb. Study selection was achieved independently by two investigators by screening the title, abstract, and full-text. The data of the eligible studies were documented in a table detailing the methods of the detection, chromosomal locations of CNV, sites of CNV beginning and end, and CNV gain or loss, etc. Then, quality control was performed independently by two investigators.

All CNV data were reported in the hg19 version except for two studies. In one study (Donaghue et al., 2017), CNVs were shown in OMIM, and we obtained the location information and converted it to the hg38 version according to the OMIM ID (https://omim.org/). Together with another study (Rajcan-Separovic et al., 2010), in which CNVs were also reported in the hg38 version, we converted CNV coordinates into the human assembly hg19 using the UCSC liftOver tool 18 (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

Generating the Copy Number Variant Map

To capture the maximum extent of CNVs, we combined the data from our cohort and the published data into a single map. First, we analyzed the density distribution of CNVs through locating all CNVs to the chromosomes using R packages karyoploteR_1.10.5. Second, we investigated the distribution of the CNVs in the pericentromeric and sub-telomeric regions of the genome. We used a sliding window of 5 Mb with steps of 0.5 Mb within 18 Mb from both sides of the centromeres (9 Mb from each side) and 9 Mb away from the telomeres. The percentage of un-gapped

nucleotides varying in each window was calculated per chromosome and plotted for all chromosomes. The chromosome length information and telomere and centromere position file was obtained from the UCSC database (hg19). The R packages ggplot2_3.3.0 were used to construct the histograms. Third, to explore the differences between CNVs detected in POCs and CNVs reported in human diseases, we compared the CNVs including those with data in the Database of Genomic Variants (DGV, http://dgv.tcag.ca/dgv/docs/GRCh37_hg19_variants_2020-02-25.txt), the CNVs reciprocal overlap more than 75% with the CNVs in DGV, and have the correspondent gain or loss which were considered reported CNV.

Copy Number Variant Gene Content and Gene Characteristics

Gene names and chromosomal coordinates of CNVs were mined from the human Release 19 reference genome file (https://www.gencodegenes.org/human/release_19.html), and CNV location information in the GENECODE database using the bedtools version 1.58 intersects function to investigate CNVRs coverage genes and annotate gene types. In order to explore the functional relevance of CNVs, we assessed the expression patterns of their integral genes using the TiGER database (Tissue-specific Gene Expression and Regulation: http://bioinfo.wilmer.jhu.edu/tiger/) (Liu et al., 2008) and the consequences of murine knock-out studies using the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org). Then, we focused on the genes expressed in the placenta and the genes resulted in embryonic lethality and abnormal embryonic development in knock-out murine.

Functional Gene Enrichment Analyses

Functional enrichment and pathway analysis for protein-coding genes in CNVs was performed using the clusterProfiler V3.12.0 R package 19. Gene-enrichment for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and GO terms (biological process, cellular component, and molecular function) were carried out for gain or loss CNV groups separately and together. A p value < 0.05 was considered statistically significant for GO terms and pathway analysis. Data were reported as significantly enriched GO terms and pathways.

RESULTS

Characteristics of Copy Number Variant Data

A total of 564 CNVs less than 3 Mb (mean: 690.2 Kb, ranging from 6.4 Kb to 2.98 Mb) were obtained from 442 euploid POCs, of which 176 CNVs were detected in our institution and 266 were extracted from 19 peer-reviewed publications (**Supplementary Figure S1**). All CNVs were detected using SNP array (9 research studies) (Reddy et al., 2012; Kooper et al., 2014; Wang et al., 2017; Qi et al., 2018; Zhu et al., 2018; Mao et al., 2019; Sato et al., 2019; Yang et al., 2019; Wang et al., 2020), array CGH (4 research studies) (Shimokawa et al., 2006; Deshpande et al., 2010; Rajcan-Separovic et al., 2010; Donaghue et al., 2017), or CMA (6 research studies plus our data) (Sahlin et al., 2014; Wang et al., 2014; Rosenfeld et al., 2015; Parchem et al., 2018; Chau et al., 2020; Zhang et al., 2021). The threshold of those studies called the CNVs has been reported ranging from 50 to 135 Kb. Among the 19 peer-reviewed publications, 11 were based on the report of 141 Caucasian cases in total and 8 based on the report of 125 Asian cases in total. Among the 564 CNVs, gains (microduplications) were largely more than losses (microdeletions, 349 vs. 185), and 30 CNVs were uncertain gain or loss from the articles. After removing the repeated CNVs and 30 CNVs unknown gain or loss, we compared the CNVs in this study with DGV data, and the results showed that 234 (52%) variants were reported, while 215 (48%) variants were not reported by DGV (**Table 1**).

TABLE 1 | CNV data from our institution and 19 peer-reviewed publications.

	Total	Gain	Loss	Unknown
Total	564	349	185	30
Our hospital	264	188	76	0
Published publications	300	161	109	30
Reported in DGV[1]	235	161	74	
Unreported in DGV[1]	244	129	85	

[1]*After removing the same CNVs in different cases.*
DGV, Database of Genomic Variants.

Distribution of Copy Number Variant in Chromosomes

The location and the number of all CNVs (564) for euploid POCs on chromosomes are shown in **Figure 1**. We investigated the CNVs in genomic gains and losses independently and also merged the two versions to generate a consensus map that represents all variations (**Figure 1A**). CNVs were found in all chromosomes, and the number varied from 2 CNVs in chromosome Y to 53 CNVs in chromosome 22. For gain, chromosome 22 showed the highest number (36 CNVs), followed by chromosome 19 (32 CNVs). Chromosome Y showed no CNVs gain, and chromosome 20 showed only 2 CNVs gain (**Figure 1B**). For losses, chromosome 2 showed the highest number of 20 CNVs, followed by chromosome 1, chromosome 16, and chromosome 22 that showed 16 CNVs, respectively. There was no CNV loss on chromosome 18, and only 2 CNVs loss on chromosome 20 (**Figure 1B**). **Figure 2** illustrates the distribution of CNVs in the pericentromeric and sub-telomeric regions, showing that the pericentromeric regions have a higher proportion of CNVs and the same characteristics are observed in both gain and loss in the sub-telomeric regions.

Functional Enrichment

After removing the same CNVs in different cases, 479 CNVRs remained, including 291 gains, 159 losses, and 29 CNVRs uncertain gains or losses. A total of 5,414 genes including 1,862 protein-coding genes and 1,284 noncoding genes (the categories of the 5,414 genes are shown in **Supplementary Figure S2**) were mined from the 479 CNVRs. GO and KEGG analyses of the involved protein-coding genes were performed.

FIGURE 1 | Chromosomal distribution of 564 CNVs less than 3 Mb from 422 euploid pregnancy loss/fetal death. **(A)** Overall map for the CNVs; **(B)** CNV number in each chromosome.

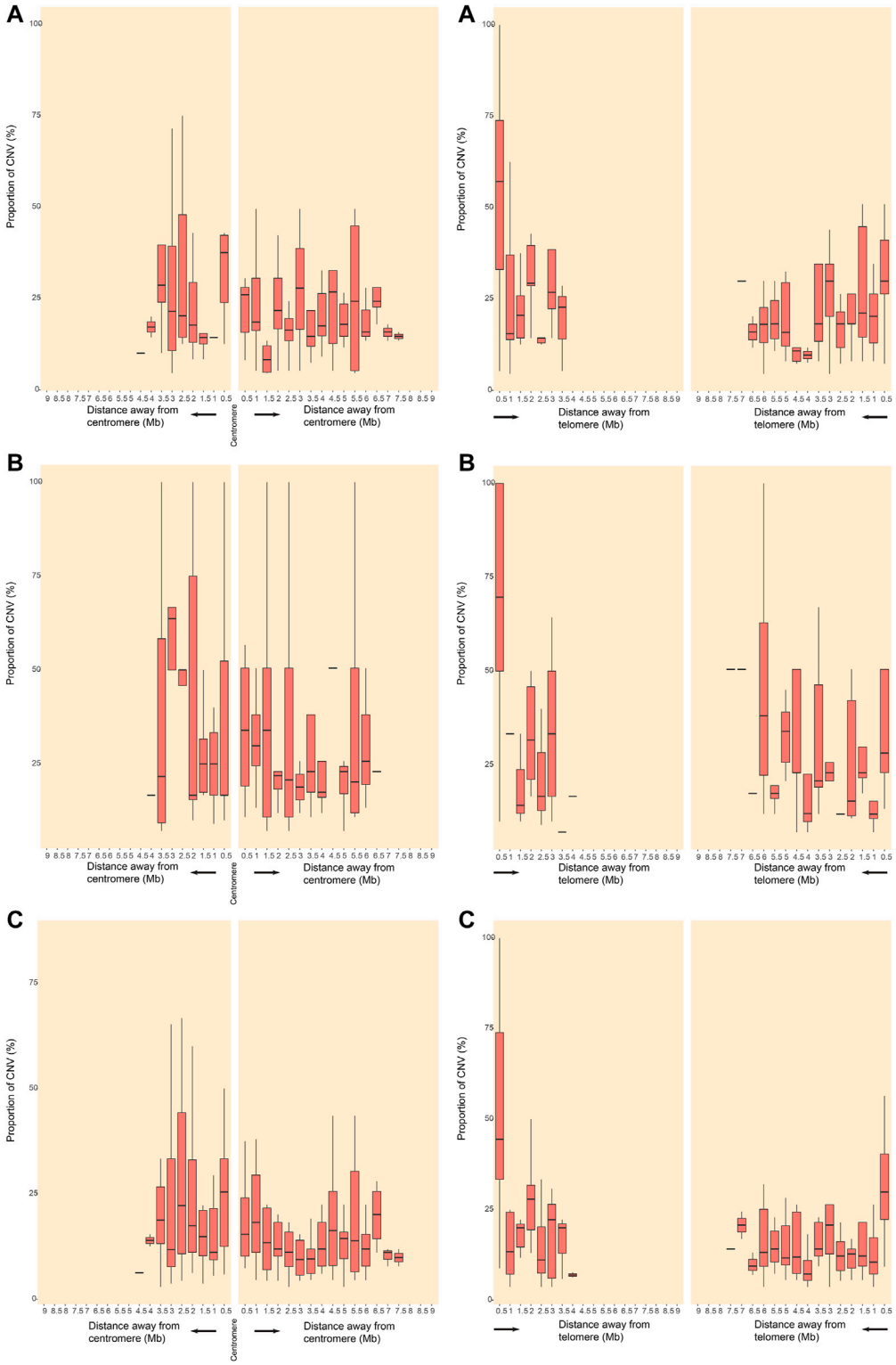

FIGURE 2 | Distribution of CNVRs in pericentromeric and sub-telomeric regions of human chromosomes. CNVRs gains **(A)**, CNVRs losses **(B)**, and CNVRs gains and losses in the inclusive map **(C)** are shown for pericentromeric regions (left panels) and sub-telomeric regions (right panels). The y axes indicate the percentage of nucleotides in each window that may involve CNVs.

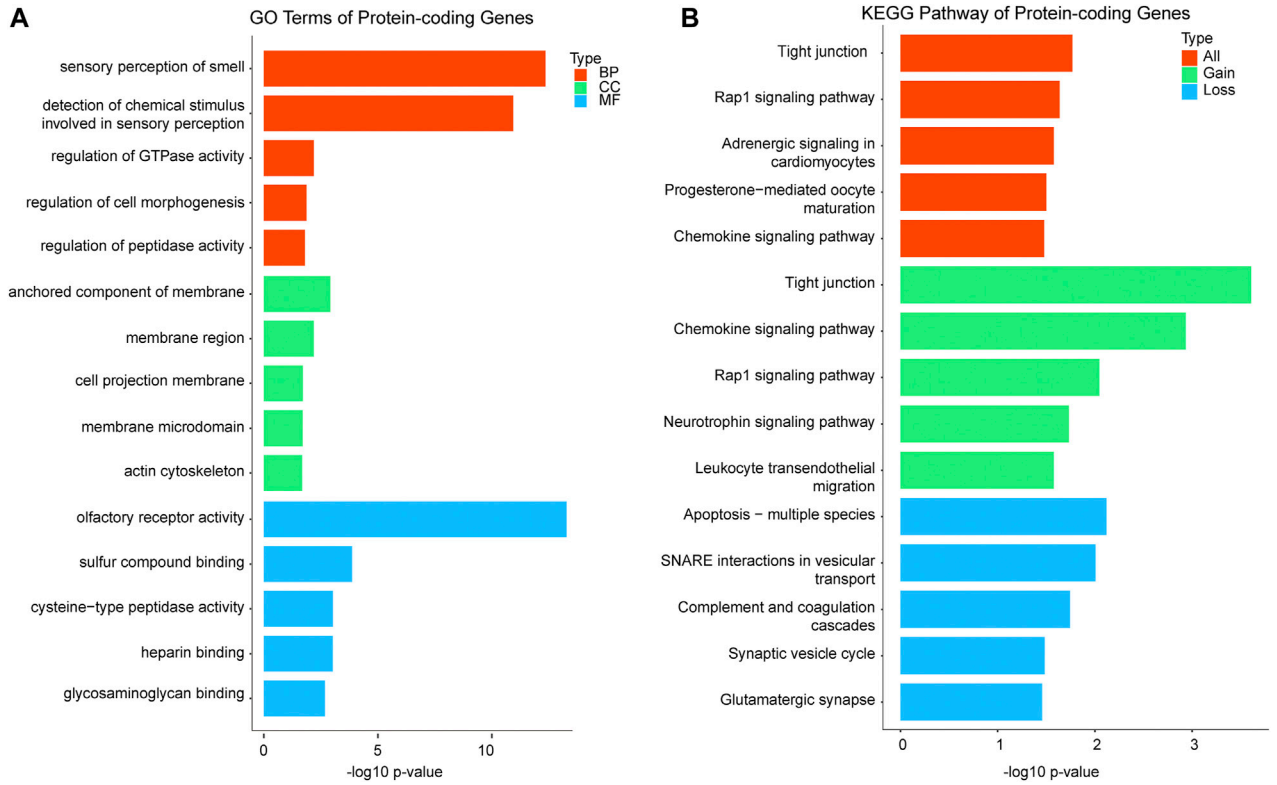

FIGURE 3 | Functional enrichment analysis of significant terms. A total of 1,862 protein-coding genes were submitted to clusterProfiler. **(A)** Top GO terms enriched for BP, CC, and MF in protein-coding genes; GO terms are assigned to y-axis, and negative log10 *p* values are assigned to x-axis; **(B)** top KEGG pathways for protein-coding genes of gains, losses, and both, and KEGG terms are assigned to y-axis, and negative log10 *p* values are assigned to x-axis; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; CC, cellular component; MF, molecular function.

For each GO analysis domain, the five top most significantly (p value < 0.05) enriched GO terms are presented in **Figure 3**. The genes in GO biological process were primarily associated with "sensory perception of smell," "detection of the chemical stimulus involved in sensory perception," and "regulation of gtpase activity." The genes in the GO cellular component were mostly enriched in "anchored component of the membrane," "membrane region," and "cell projection membrane." The genes in GO molecular function were mainly associated with "olfactory receptor activity," "sulfur compound binding," and "cysteine-type peptidase activity" (**Figure 3A**). The protein-coding genes KEGG pathway analysis indicated that the CNVRs were intensively associated with "tight junction," "Rap1 signaling pathway," "adrenergic signaling in cardiomyocytes," "progesterone-mediated oocyte maturation," and "chemokine signaling pathway" (**Figure 3B**). The details of GO terms and KEGG pathways are shown in **Supplementary Table S2**.

Gene Characteristics

Among the 1,862 protein-coding genes, 53% (995/1862) of them have been reported to present mammalian phenotypes in MGI. The number of genes that results in embryonic lethality or abnormal embryonic size/development and abnormal placental size/morphology in knock-out models were 233 and 44, respectively (**Figure 4**). The results of tissue-specific expression analysis of protein-coding genes showed that 19 genes were placental-specific or placental expression. The details of the involved genes and CNVs are shown in **Supplementary Table S3**. The most frequent CNV was located on 19p13.3, which was detected in 11 POCs with 9 gains and 2 losses, with a size ranging from 523.9 Kb to 1.5 Mb (**Supplementary Table S3**). Among CNVRs in 19p13.3, 13 genes with mammalian phenotypes in MGI caused murine embryonic lethality or abnormal embryonic/placental size/morphology in knock-out models, and 2 genes showed placental expression (**Figure 4** and **Supplementary Table S3**).

DISCUSSION

This study presents a unique analysis of CNVs less than 3 Mb detected in euploid POCs and their integral gene content in a large cohort and 19 published studies in order to evaluate their overall chromosomal distribution, genomic features, and functions based on bioinformatics. Collectively, all the chromosomes are susceptible to CNV in POCs, and CNVs distribute unevenly along the chromosomes and among

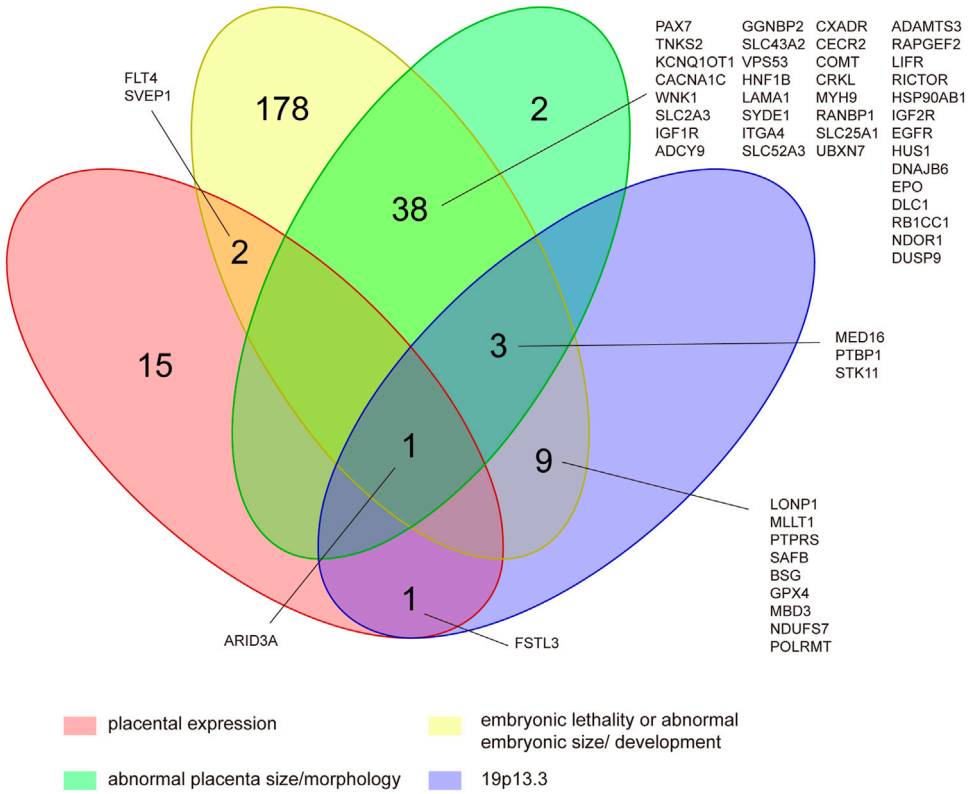

FIGURE 4 | Identification of CNV genes in POCs associated with embryonic lethality or abnormal embryonic size/development, abnormal placental size/morphology, and placental expression located in 19p13.3. This was determined by assessing 995 protein-coding genes of the CNVRs that had reported to present mammalian phenotypes in mouse knock-out studies and cataloged on MGI as well as assessing 19 human placental-expressed genes listed on TiGER.

chromosomal regions. Some CNVs might have a pathogenic role in pregnancy loss because of containing embryonic lethality genes.

Gene and segmental duplications are thought to have a significant role in gene and genome evolution and are often under positive selection, whereas deletions are biased away from certain categories and more likely to cause disease or alter the fitness (Hurles, 2004; Redon et al., 2006; Uddin et al., 2014). In our data, there are more gains than losses detected in euploid POCs (349 vs. 185), which is in contrary to CNVs in healthy individuals from various populations according to the research of Mehdi Zarrei, who analyzed 23 studies and reported that the losses were almost 10 times the gains (Redon et al., 2006). The mechanism of germ line CNVs is complicated, and it is unclear that the proportion of duplication and deletion is in the early stage of embryogenesis. There are approximately 22% of spontaneous conceptions ending in biochemical pregnancy losses (BPLs), which are poorly understood since embryonic arrest is prior to the development of a clinical pregnancy (Wilcox et al., 1988; Ellish et al., 1996; Zinaman et al., 1996). If the variational chances to duplication and deletion in embryogenesis are equal, it is possible that the embryo with CNV duplication might be more likely to cause pregnant failure than the embryo with CNV deletion.

In our data, CNVs distribute unevenly in all chromosomes, and chromosome 22 is found to have the highest variability, which is consistent with CNVs in healthy individuals (Makino et al., 2013; Zarrei et al., 2015). However, Y chromosome carries the lowest number of CNVs in our study, which is contrary to the highest proportion of CNVs in Y chromosome reported in healthy individuals (Zarrei et al., 2015). Those results indicate that the pregnancy with Y chromosome microduplications or microdeletions might not result in embryonic/fetal death. After all, the biological function of Y chromosome is believed to mainly impact male fitness such as fertility (Quintana-Murci et al., 2001; Schlegel, 2002). Our study also demonstrates that CNVs unevenly distribute within the chromosome. The pericentromeric and subtelomeric regions have a higher density of CNVs, in both gain and loss, which are same as the results of healthy individuals (Zarrei et al., 2015).

For gene functional enrichment, to our surprise, the two most significant GO germs in biological process of the protein-coding genes are involved in "sensory perception of smell" and "detection of the chemical stimulus involved in sensory perception". Sensory development is complex, with both

morphological and neural components (Clark-Gambelunghe and Clark, 2015). The tissues of the oral cavity, eye, and auditory system form the face and palate between 6 and 12 gestational weeks (Witt, 2019). The development of the nervous system and sensory perception is established throughout the fetal and postnatal period, which is important for fetal survival. We speculate that genes involved in sensory perception might be dose-sensitive and have potential to cause embryonic arrest when CNV occurs. In GO cellular component, four of top five significant terms are enriched in membrane-related components, such as "anchored component of the membrane," "membrane region," and "membrane microdomain." The genes for the cell membrane component are vital to embryonic development, and our results imply that functions of those genes might be easily affected by gene dosage. The genes in GO molecular function are significantly associated with olfactory receptor activity, sulfur compound binding, and heparin binding, and those functions are related to transmembrane transport. Our results also show that the KEGG pathway is significantly related to "tight junction." It is well known that the tight junction (TJ) is an essential component of the differentiated epithelial cell required for polarization and intercellular integrity during early development (Eckert and Fleming, 2008; Green et al., 2019). These results indicate that pregnancy with CNVs involving genes in membrane component, transmembrane transport, and TJ might relate to developmental arrest.

Among the protein-coding genes, 276 genes showed embryonic lethality or abnormal embryonic/placental size/morphology in knock-out mouse models. Theoretically, pregnancy with CNV carrying those genes could increase the risk for embryonic demise, which, however, needs to be confirmed by further studies. In addition, CNV located at 19p13.3 is found to be the most frequent one, in 11 POCs. It is interesting in our results that all the 13 genes contained in 19p13.3 that have mammalian phenotypes in MGI are shown to cause murine embryonic lethality or abnormal embryonic/placental size/morphology in the knock-out model. Chromosome 19 has the highest gene density of all human chromosomes (Grimwood et al., 2004), and CNVs in 19p13.3 have been reported in several patients with intellectual disability and congenital malformations (Orellana et al., 2015; Palumbo et al., 2016). Our study suggests that CNV in 19p13.3 might be pathogenic in pregnancy loss/fetal death.

Among those genes included in 19p13.3, two placental-expressed genes are worth of attention, namely AT-rich interaction domain 3A (*ARID3A*) and follistatin-like 3 (*FSTL3*). *ARID3A* has been reported essential to the execution of the first cell fate decision and is of importance to regulate mesoderm differentiation and nephric tubule regeneration in animal models and has a vital role in placental development (Rhee et al., 2015; Popowski et al., 2017; Suzuki et al., 2019). *FSTL3* has been demonstrated to be expressed on the maternal–fetoplacental interface in the first trimester and regulates the invasion and migration of trophoblast, which is important for establishing and maintaining normal pregnancy (Xie et al., 2018; Founds and Stolz, 2020; Xu et al., 2020). In addition, arid3a and fstl3 show abnormal phenotypes of multiple organs in knockout mouse models. Therefore, it is possible that duplication or deletion in *ARID3A* and *FSTL3* results in embryonic/fetal development arresting.

Our study has several limitations. First, CNV data were extracted from our laboratory and published studies, which were detected by different platforms, so that the potential methodological bias cannot be eliminated. Second, our study did not compare CNVs in POCs from pregnancy loss/fetal death with healthy controls, and therefore cautions should be taken in the interpretation of the pathogenic CNVs in pregnancy loss/fetal death. Third, we were unable to identify those CNVs in parents or to achieve those data about parental origin, which affects the determination of pathogenic CNV to a certain extent, especially for pathogenicity of 19p13.3.

In conclusion, this study shows that CNVs less than 3 Mb in euploid POCs distribute unevenly in all chromosomes and have a higher density in the pericentromeric and sub-telomeric regions. The CNVRs are significantly enriched in genes involving sensory perception, membrane-related components, and tight junction, and those biological processes and pathways are important for embryonic/fetal development. CNV in 19p13.3 might have a pathogenic role in pregnancy loss, and the variations of *ARID3A* and *FSTL3* might be a predisposing risk for pregnancy loss. A further study is needed to compare those CNVs with the control group and identify those CNVs in the parents for getting inheritance information.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the study protocol approved by the Ethics Committee of the institute (2020-15001). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CG and YH contributed to the conception of the study and drafted the manuscript. CG and HG contributed to the design of the research and the control of the data quality. HG performed the bioinformatics analysis. RL and CW contributed to data management and prepared the retrospective data for analysis. XD, KL, and ZY contributed to literature search and extracted data. All authors reviewed and read and approved the final manuscript.

FUNDING

The study was supported by a grant from the Guangzhou Women and Children's Medical Center, Guangzhou, China (1600067-04), and a grant from the Guangzhou Municipal Science and Technology Bureau, Guangzhou, China (202102010311).

ACKNOWLEDGMENTS

We thank all patients who paid their antenatal care in our hospital and all physicians who recorded the data on pregnancies and pregnancy outcomes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.766492/full#supplementary-material

Supplementary Figure S1 | The flew chart of the search and selection of the peer-reviewed publications.

Supplementary Figure S2 | The categories of the 5,414 genes.

Supplementary Table S1 | Complete search used in the systematic literature search.

Supplementary Table S2 | The details of GO terms and KEGG pathways.

Supplementary Table S3 | The details of placental expression genes, genes with mammalian phenotypes in MGI caused murine embryonic lethality or abnormal embryonic/placental size/morphology, as well as genes in 19p13.3.

REFERENCES

Beckmann, J. S., Estivill, X., and Antonarakis, S. E. (2007). Copy Number Variants and Genetic Traits: Closer to the Resolution of Phenotypic to Genotypic Variability. *Nat. Rev. Genet.* 8, 639–646. doi:10.1038/nrg2149

ESHRE Guideline Group on RPLBender Atik, R., Christiansen, O. B., Elson, J., Kolte, A. M., Lewis, S., Middel, dorp, S., et al. (2018). ESHRE Guideline: Recurrent Pregnancy Loss. *Hum. Reprod. Open* 2018, hoy004. doi:10.1093/hropen/hoy004

Chau, M. H. K., Wang, H., Lai, Y., Zhang, Y., Xu, F., Tang, Y., et al. (2020). Low-pass Genome Sequencing: a Validated Method in Clinical Cytogenetics. *Hum. Genet.* 139, 1403–1415. doi:10.1007/s00439-020-02185-9

Clark-Gambelunghe, M. B., and Clark, D. A. (2015). Sensory Development. *Pediatr. Clin. North America* 62, 367–384. doi:10.1016/j.pcl.2014.11.003

Colley, E., Hamilton, S., Smith, P., Morgan, N. V., Coomarasamy, A., and Allen, S. (2019). Potential Genetic Causes of Miscarriage in Euploid Pregnancies: a Systematic Review. *Hum. Reprod. Update* 25, 452–472. doi:10.1093/humupd/dmz015

Deshpande, M., Harper, J., Holloway, M., Palmer, R., and Wang, R. (2010). Evaluation of Array Comparative Genomic Hybridization for Genetic Analysis of Chorionic Villus Sampling from Pregnancy Loss in Comparison to Karyotyping and Multiplex Ligation-dependent Probe Amplification. *Genet. Test. Mol. Biomarkers* 14, 421–424. doi:10.1089/gtmb.2010.0014

Dhillon, R., Hillman, S., Morris, R., McMullan, D., Williams, D., Coomarasamy, A., et al. (2014). Additional Information from Chromosomal Microarray Analysis (CMA) over Conventional Karyotyping when Diagnosing Chromosomal Abnormalities in Miscarriage: a Systematic Review and Meta-Analysis. *Bjog: Int. J. Obstet. Gy* 121, 11–21. doi:10.1111/1471-0528.12382

Donaghue, C., Davies, N., Ahn, J. W., Thomas, H., Ogilvie, C. M., and Mann, K. (2017). Efficient and Cost-Effective Genetic Analysis of Products of conception and Fetal Tissues Using a QF-PCR/array CGH Strategy; Five Years of Data. *Mol. Cytogenet.* 10, 12. doi:10.1186/s13039-017-0313-9

Dong, Z., Zhang, J., Hu, P., Chen, H., Xu, J., Tian, Q., et al. (2016). Low-pass Whole-Genome Sequencing in Clinical Cytogenetics: a Validated Approach. *Genet. Med.* 18, 940–948. doi:10.1038/gim.2015.199

Doubilet, P. M., Benson, C. B., Bourne, T., and Blaivas, M. (2013). Society of Radiologists in Ultrasound Multispecialty Panel on Early First Trimester Diagnosis of Diagnostic Criteria for Nonviable Pregnancy Early in the First Trimester. *N. Engl. J. Med.* 369, 1443–1451. doi:10.1056/nejmra1302417

Eckert, J. J., and Fleming, T. P. (2008). Tight junction Biogenesis during Early Development. *Biochim. Biophys. Acta (Bba) - Biomembranes* 1778, 717–728. doi:10.1016/j.bbamem.2007.09.031

Ellish, N. J., Saboda, K., O'Connor, J., Nasca, P. C., Stanek, E. J., and Boyle, C. (1996). A Prospective Study of Early Pregnancy Loss. *Hum. Reprod.* 11, 406–412. doi:10.1093/humrep/11.2.406

Founds, S. A., and Stolz, D. B. (2020). Gene Expression of Four Targets *In Situ* of the First Trimester Maternal-Fetoplacental Interface. *Tissue and Cell* 64, 101313. doi:10.1016/j.tice.2019.101313

Grayton, H. M., Fernandes, C., Rujescu, D., and Collier, D. A. (2012). Copy Number Variations in Neurodevelopmental Disorders. *Prog. Neurobiol.* 99, 81–91. doi:10.1016/j.pneurobio.2012.07.005

Green, K. J., Jaiganesh, A., and Broussard, J. A. (2019). Desmosomes: Essential Contributors to an Integrated Intercellular junction Network. *F1000Res* 8, F1000. doi:10.12688/f1000research.20942.1

Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., et al. (2004). The DNA Sequence and Biology of Human Chromosome 19. *Nature* 428, 529–535. doi:10.1038/nature02399

Gu, C., Li, K., Li, R., Li, L., Li, X., Dai, X., et al. (2021). Chromosomal Aneuploidy Associated with Clinical Characteristics of Pregnancy Loss. *Front. Genet.* 12, 667697. doi:10.3389/fgene.2021.667697

Hillman, S. C., Pretlove, S., Coomarasamy, A., McMullan, D. J., Davison, E. V., Maher, E. R., et al. (2011). Additional Information from Array Comparative Genomic Hybridization Technology over Conventional Karyotyping in Prenatal Diagnosis: a Systematic Review and Meta-Analysis. *Ultrasound Obstet. Gynecol.* 37, 6–14. doi:10.1002/uog.7754

Hurles, M. (2004). Gene Duplication: the Genomic Trade in Spare Parts. *Plos Biol.* 2, E206. doi:10.1371/journal.pbio.0020206

Hurles, M. E., Dermitzakis, E. T., and Tyler-Smith, C. (2008). The Functional Impact of Structural Variation in Humans. *Trends Genet.* 24, 238–245. doi:10.1016/j.tig.2008.03.001

Kooper, A. J., Faas, B. H., Feenstra, I., de Leeuw, N., and Smeets, D. F. (2014). Best Diagnostic Approach for the Genetic Evaluation of Fetuses after Intrauterine Death in First, Second or Third Trimester: QF-PCR, Karyotyping And/or Genome Wide SNP Array Analysis. *Mol. Cytogenet.* 7, 6. doi:10.1186/1755-8166-7-6

Lee, C., and Scherer, S. W. (2010). The Clinical Context of Copy Number Variation in the Human Genome. *Expert Rev. Mol. Med.* 12, e8. doi:10.1017/s1462399410001390

Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J. (2008). TiGER: a Database for Tissue-specific Gene Expression and Regulation. *BMC Bioinformatics* 9, 271. doi:10.1186/1471-2105-9-271

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. (2014). The Database of Genomic Variants: a Curated Collection of Structural Variation in the Human Genome. *Nucl. Acids Res.* 42, D986–D992. doi:10.1093/nar/gkt958

Makino, T., McLysaght, A., and Kawata, M. (2013). Genome-wide Deserts for Copy Number Variation in Vertebrates. *Nat. Commun.* 4, 2283. doi:10.1038/ncomms3283

Mao, J., Wang, H., Li, H., Song, X., Wang, T., Xiang, J., et al. (2019). Genetic Analysis of Products of conception Using a HLPA/SNP-array Strategy. *Mol. Cytogenet.* 12, 40. doi:10.1186/s13039-019-0452-2

Orellana, C., Roselló, M., Monfort, S., Mayo, S., Oltra, S., and Martínez, F. (2015). Pure Duplication of 19p13.3 in Three Members of a Family with Intellectual Disability and Literature Review. Definition of a New Microduplication Syndrome. *Am. J. Med. Genet.* 167, 1614–1620. doi:10.1002/ajmg.a.37046

Palumbo, P., Palumbo, O., Leone, M. P., Stallone, R., Palladino, T., Zelante, L., et al. (2016). Clinical and Molecular Characterization of a De Novo 19p13.3 Microdeletion. *Mol. Cytogenet.* 9, 40. doi:10.1186/s13039-016-0252-x

Parchem, J. G., Sparks, T. N., Gosnell, K., and Norton, M. E. (2018). Utility of Chromosomal Microarray in Anomalous Fetuses. *Prenatal Diagn.* 38, 140–147. doi:10.1002/pd.5202

Popowski, M., Lee, B. K., Rhee, C., Iyer, V. R., and Tucker, H. O. (2017). Arid3a Regulates Mesoderm Differentiation in Mouse Embryonic Stem Cells. *J. Stem Cel Ther Transpl.* 1, 52–62. doi:10.29328/journal.jsctt.1001005

Practice Committee of the American Society for Reproductive Medicine (2012). Evaluation and Treatment of Recurrent Pregnancy Loss: a Committee Opinion. *Fertil. Steril* 98, 1103–1111. doi:10.1016/j.fertnstert.2012.06.048

Qi, H., Xuan, Z.-L., Du, Y., Cai, L.-R., Zhang, H., Wen, X.-H., et al. (2018). High Resolution Global Chromosomal Aberrations from Spontaneous Miscarriages Revealed by Low Coverage Whole Genome Sequencing. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 224, 21–28. doi:10.1016/j.ejogrb.2018.03.008

Quintana-Murci, L., Krausz, C., and McElreavey, K. (2001). The Human Y Chromosome: Function, Evolution and Disease. *Forensic Sci. Int.* 118, 169–181. doi:10.1016/s0379-0738(01)00387-5

Rajcan-Separovic, E., Diego-Alvarez, D., Robinson, W. P., Tyson, C., Qiao, Y., Harvard, C., et al. (2010). Identification of Copy Number Variants in Miscarriages from Couples with Idiopathic Recurrent Pregnancy Loss. *Hum. Reprod.* 25, 2913–2922. doi:10.1093/humrep/deq202

Reddy, U. M., Page, G. P., Saade, G. R., Silver, R. M., Thorsten, V. R., Parker, C. B., et al. (2012). Karyotype versus Microarray Testing for Genetic Abnormalities after Stillbirth. *N. Engl. J. Med.* 367, 2185–2193. doi:10.1056/nejmoa1201569

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global Variation in Copy Number in the Human Genome. *Nature* 444, 444–454. doi:10.1038/nature05329

Rhee, C., Lee, B. K., Beck, S., Anjum, A., Cook, K. R., Popowski, M., et al. (2015). Corrigendum: Arid3a Is Essential to Execution of the First Cell Fate Decision via Direct Embryonic and Extraembryonic Transcriptional Regulation. *Genes Dev.* 29, 1890. doi:10.1101/gad.247163.114

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical Standards for the Interpretation and Reporting of Constitutional Copy-Number Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257. doi:10.1038/s41436-019-0686-8

Rosenfeld, J. A., Tucker, M. E., Escobar, L. F., Neill, N. J., Torchia, B. S., McDaniel, L. D., et al. (2015). Diagnostic Utility of Microarray Testing in Pregnancy Loss. *Ultrasound Obstet. Gynecol.* 46, 478–486. doi:10.1002/uog.14866

Sahlin, E., Gustavsson, P., Liedén, A., Papadogiannakis, N., Bjäreborn, L., Pettersson, K., et al. (2014). Molecular and Cytogenetic Analysis in Stillbirth: Results from 481 Consecutive Cases. *Fetal Diagn. Ther.* 36, 326–332. doi:10.1159/000361017

Sato, T., Migita, O., Hata, H., Okamoto, A., and Hata, K. (2019). Analysis of Chromosome Microstructures in Products of conception Associated with Recurrent Miscarriage. *Reprod. BioMedicine Online* 38, 787–795. doi:10.1016/j.rbmo.2018.12.010

Schlegel, P. N. (2002). The Y Chromosome. *Reprod. BioMedicine Online* 5, 22–25. doi:10.1016/s1472-6483(10)61592-1

Shimokawa, O., Harada, N., Miyake, N., Satoh, K., Mizuguchi, T., Niikawa, N., et al. (2006). Array Comparative Genomic Hybridization Analysis in First- Trimester Spontaneous Abortions with 'normal' Karyotypes. *Am. J. Med. Genet. A.* 140, 1931–1935. doi:10.1002/ajmg.a.31421

Suzuki, N., Hirano, K., Ogino, H., and Ochi, H. (2019). Arid3a Regulates Nephric Tubule Regeneration via Evolutionarily Conserved Regeneration Signal-Response Enhancers. *Elife* 8, e43186. doi:10.7554/eLife.43186

Uddin, M., Tammimies, K., Pellecchia, G., Alipanahi, B., Hu, P., Wang, Z., et al. (2014). Brain-expressed Exons under Purifying Selection Are Enriched for De Novo Mutations in Autism Spectrum Disorder. *Nat. Genet.* 46, 742–747. doi:10.1038/ng.2980

Wang, B. T., Chong, T. P., Boyar, F. Z., Kopita, K. A., Ross, L. P., El-Naggar, M. M., et al. (2014). Abnormalities in Spontaneous Abortions Detected by G-Banding and Chromosomal Microarray Analysis (CMA) at a National Reference Laboratory. *Mol. Cytogenet.* 7, 33. doi:10.1186/1755-8166-7-33

Wang, Y., Cheng, Q., Meng, L., Luo, C., Hu, H., Zhang, J., et al. (2017). Clinical Application of SNP Array Analysis in First-Trimester Pregnancy Loss: a Prospective Study. *Clin. Genet.* 91, 849–858. doi:10.1111/cge.12926

Wang, Y., Li, Y., Chen, Y., Zhou, R., Sang, Z., Meng, L., et al. (2020). Systematic Analysis of Copy-number Variations Associated with Early Pregnancy Loss. *Ultrasound Obstet. Gynecol.* 55, 96–104. doi:10.1002/uog.20412

Wapner, R. J., Martin, C. L., Levy, B., Ballif, B. C., Eng, C. M., Zachary, J. M., et al. (2012). Chromosomal Microarray versus Karyotyping for Prenatal Diagnosis. *N. Engl. J. Med.* 367, 2175–2184. doi:10.1056/nejmoa1203382

Wilcox, A. J., Weinberg, C. R., O'Connor, J. F., Baird, D. D., Schlatterer, J. P., Canfield, R. E., et al. (1988). Incidence of Early Loss of Pregnancy. *N. Engl. J. Med.* 319, 189–194. doi:10.1056/nejm198807283190401

Witt, M. (2019). Anatomy and Development of the Human Taste System. *Handb Clin. Neurol.* 164, 147–171. doi:10.1016/b978-0-444-63855-7.00010-1

Xie, J., Xu, Y., Wan, L., Wang, P., Wang, M., and Dong, M. (2018). Involvement of Follistatin-like 3 in Preeclampsia. *Biochem. Biophysical Res. Commun.* 506, 692–697. doi:10.1016/j.bbrc.2018.10.139

Xu, Y., Xie, J., Wan, L., Wang, M., Xu, Y., Wang, H., et al. (2020). Follistatin-like 3, an Activin A Binding Protein, Is Involved in Early Pregnancy Loss. *Biomed. Pharmacother.* 121, 109577. doi:10.1016/j.biopha.2019.109577

Yamada, H., Sata, F., Saijo, Y., Kishi, R., and Minakami, H. (2005). Genetic Factors in Fetal Growth Restriction and Miscarriage. *Semin. Thromb. Hemost.* 31, 334–345. doi:10.1055/s-2005-872441

Yang, Y., Qu, S., Wang, L., Guo, Y., Xue, S., Cai, A., et al. (2019). Genetic Testing of Chorionic Villi from Abortuses during Early Pregnancy. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 36, 547–551. doi:10.3760/cma.j.issn.1003-9406.2019.06.004

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A Copy Number Variation Map of the Human Genome. *Nat. Rev. Genet.* 16, 172–183. doi:10.1038/nrg3871

Zhang, W., Lei, T., Fu, F., Deng, Q., Li, R., Wang, D., et al. (2021). Microarray Analysis in Fetuses with Duodenal Obstruction: It Is Not Just Trisomy 21. *Prenatal Diagn.* 41, 316–322. doi:10.1002/pd.5834

Zhu, X., Li, J., Zhu, Y., Wang, W., Wu, X., Yang, Y., et al. (2018). Application of Chromosomal Microarray Analysis in Products of Miscarriage. *Mol. Cytogenet.* 11, 44. doi:10.1186/s13039-018-0396-y

Zinaman, M. J., Clegg, E. D., Brown, C. C., O'Connor, J., and Selevan, S. G. (1996). Estimates of Human Fertility and Pregnancy loss*†*Supported by grant CR-820787 from the United States Environmental Protection Agency, Washington, D.C.†The Views Expressed in This Paper Are Those of the Authors and Do Not Necessarily Reflect the Views or Policies of the U.S. Environmental Protection Agency. The U.S. Government Has the Right to Retain a Nonexclusive Royalty-free License in and to Any Copyright Covering This Paper. *Fertil. Sterility* 65, 503–509. doi:10.1016/s0015-0282(16)58144-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 14 April 2022
doi: 10.3389/fgene.2022.803088

Investigation of Chromosomal Structural Abnormalities in Patients With Undiagnosed Neurodevelopmental Disorders

Ye Cao[1,2,3,4†], Ho Ming Luk[5†], Yanyan Zhang[2†], Matthew Hoi Kin Chau[2], Shuwen Xue[2], Shirley S. W. Cheng[5], Albert Martin Li[1,4], Josephine S. C. Chong[1], Tak Yeung Leung[2], Zirui Dong[2,3,4], Kwong Wai Choy[2,3] and Ivan Fai Man Lo[5]**

[1]Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China, [2]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong SAR, China, [3]Key Laboratory for Regenerative Medicine, Ministry of Education (Shenzhen Base), Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China, [4]Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China, [5]Clinical Genetic Service, Department of Health, Hong Kong SAR, China

OPEN ACCESS

Edited by:
Manuel Corpas,
Cambridge Precision Medicine,
United Kingdom

Reviewed by:
Xiaoli Chen,
Capital Institute of Pediatrics, China
Janani Iyer,
National Aeronautics and Space Administration (NASA), United States

***Correspondence:**
Kwong Wai Choy
richardchoy@cuhk.edu.hk
Ivan Fai Man Lo
dr.ivanlo@gmail.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 27 October 2021
Accepted: 04 February 2022
Published: 14 April 2022

Citation:
Cao Y, Luk HM, Zhang Y, Chau MHK, Xue S, Cheng SSW, Li AM, Chong JSC, Leung TY, Dong Z, Choy KW and Lo IFM (2022) Investigation of Chromosomal Structural Abnormalities in Patients With Undiagnosed Neurodevelopmental Disorders. Front. Genet. 13:803088. doi: 10.3389/fgene.2022.803088

Background: Structural variations (SVs) are various types of the genomic rearrangements encompassing at least 50 nucleotides. These include unbalanced gains or losses of DNA segments (copy number changes, CNVs), balanced rearrangements (such as inversion or translocations), and complex combinations of several distinct rearrangements. SVs are known to play a significant role in contributing to human genomic disorders by disrupting the protein-coding genes or the interaction(s) with cis-regulatory elements. Recently, different types of genome sequencing-based tests have been introduced in detecting various types of SVs other than CNVs and regions with absence of heterozygosity (AOH) with clinical significance.

Method: In this study, we applied the mate-pair low pass (~4X) genome sequencing with large DNA-insert (~5 kb) in a cohort of 100 patients with neurodevelopmental disorders who did not receive informative results from a routine CNV investigation. Read-depth-based CNV analysis and chimeric-read-pairs analysis were used for CNV and SV analyses. The region of AOH was indicated by a simultaneous decrease in the rate of heterozygous SNVs and increase in the rate of homozygous SNVs.

Results: First, we reexamined the 25 previously reported CNVs among 24 cases in this cohort. The boundaries of these twenty-five CNVs including 15 duplications and 10 deletions detected were consistent with the ones indicated by the chimeric-read-pairs analysis, while the location and orientation were determined in 80% of duplications (12/15). Particularly, one duplication was involved in complex rearrangements. In addition, among all the 100 cases, 10% of them were detected with rare or complex SVs (>10 Kb), and 3% were with multiple AOH (≥5 Mb) locating in imprinting chromosomes identified. In particular, one patient with an overall value of 214.5 Mb of AOH identified on 13 autosomal chromosomes suspected parental consanguinity.

Conclusion: In this study, mate-pair low-pass GS resolved a significant proportion of CNVs with inconclusive significance, and detected additional SVs and regions of AOH in patients with undiagnostic neurodevelopmental disorders. This approach complements the first-tier CNV analysis for NDDs, not only by increasing the resolution of CNV detection but also by enhancing the characterization of SVs and the discovery of potential causative regions (or genes) contributory to could be complex in composition NDDs.

Keywords: structural variations, mate-pair genome sequencing, neurodevelopmental disorders, Absence of heterozygosity (AOH), CNV (copy number variant), insertion, inversion, complex rearrangements

INTRODUCTION

Structural variations (SVs), including various types of DNA changes (>50bps) in the genome, are known to contribute to the genomic diversity of the populations. Some of them are also associated with various genetic diseases (Abel et al., 2020; Ho et al., 2020). SVs can be balanced where there are no major gains or losses of genomic content but change(s) the organization of chromosomal segments, such as translocations, inversions, insertions; in unbalanced forms, commonly known as copy number variations (CNVs), or in complex forms with combination of several categories even involving multiple chromosomes. Medical studies or even presumably healthy human population genomic profiling studies reveal that simple SVs defined by conventional methods, such as karyotyping or chromosomal microarray analysis (CMA), could be complex in composition by next-generation sequencing studies (Dong et al., 2021). Current studies demonstrate that SVs are frequently seen, and balanced forms would be more likely seen in asymptomatic individuals, whereas complex rearrangements involving CNVs are also commonly identified in the human germline genome (de Pagter et al., 2015; Bertelsen et al., 2016; Collins et al., 2017). Rare SVs disrupting the coding sequences or interaction with regulatory elements, or adversely affecting the expressions of those disease-associated genes, are the known underlying mechanisms contributory to human diseases (Collins et al., 2017; Pocza et al., 2021). Therefore, reliable approaches to comprehensively and cost effectively identify clinically significant SVs in human genome, which is an important type of genetic variants and largely still underappreciated by current methods, are warranted.

Neurodevelopmental disorders (NDDs) are a group of disorders primarily associated with neurodevelopmental dysfunctions such as autism spectrum disorder (ASD), developmental delay (DD), and intellectual disability (ID). It is estimated that gene dosage alterations caused by large CNVs are responsible for 10–15% of NDD cases (Miller et al., 2010; Kaminsky et al., 2011; Yuan et al., 2021), while single-nucleotide variants (SNVs) and/or small insertions/deletions (InDels) contribute to over 30% of overall NDD cases (Srivastava et al., 2019). Despite extensive research and advancements in genetic diagnosis of neurological disorders, there are still at least half of NDD patients who remain idiopathic. In the last few decades, CMA has been recommended as the first-tier test for genetic investigation of NDDs (Miller et al., 2010), while currently, exome or genome sequencing (GS) is set as the second-tier testing (Manickam et al., 2021). However, these technologies mainly detect CNVs and SNVs/InDels, but are limited in identifying the direction/orientation of CNVs, let alone those balanced SVs. For example, CMA cannot determine whether a copy number gain is a forward tandem or reverse duplication, or an insertion resulting in inconclusive classification and interpretation. In addition, structural rearrangements cryptic to conventional G-banded chromosome analysis are largely known in NDDs. Apart from affecting the protein-coding portion of the genome, SVs can cause diseases by altering the copy number or position of regulatory elements, or by reshuffling higher-order chromatin structures as demonstrated in NDDs (D'haene and Vergult, 2021). For instance, the importance of translocations, inversions, and inversion-mediated complex structural rearrangements in autism spectrum disorder (ASD) and congenital anomalies have been demonstrated to be disease related by showing gene disruption or dysregulation due to a disruption of topologically associated domains (TADs) (Talkowski et al., 2012; Collins et al., 2017; Werling et al., 2018; Pocza et al., 2021). Last, some NDDs are caused by uniparental disomy (UPD) due to the involvement of imprinting genes, while some of them are caused by the homozygous defects in autosomal recessive genes due to parental consanguinity, both of which have one or more DNA stretches with the absence of heterozygosity (AOHs) identified in the genome (Fan et al., 2013; Palumbo et al., 2015; Yu et al., 2016).

Currently, increasing studies on the development of sequencing approaches and detection algorithms show the improvement of SV detection accuracy. Particularly, our previous studies have demonstrated our in-house mate-pair library construction and low-pass genome sequencing (>4-fold) enable comprehensive detection of structural rearrangements, cryptic to conventional karyotyping, as well as long contiguous regions of AOH contributed by UPD or parental consanguinity. Herein, we aim to (1) investigate the genomic composition of deletions and duplications with inconclusive significance identified by previous CNV analysis, and (2) characterize structural rearrangements and AOHs (likely resulted from UPD or parental consanguinity) by utilizing mate-pair genome sequencing in 100 NDD cases.

MATERIALS AND METHODS

Subjects

The study was approved by the Joint Chinese University of Hong Kong—New Territories East Cluster Clinical Research Ethics

Committee (CREC Ref. No. 2019.600). DNA samples of 100 consecutive patients were retrieved for this study. These patients (1) were referred to Clinical Genetic Service, Department of Health, Hong Kong SAR during 2019–2020; (2) have major indications including developmental delay, intellectual disability, congenital abnormalities, and autism spectrum disorders; (3) with a negative or inconclusive finding from previous CNV analysis (at a resolution of 50 kb for all types of CNVs; for homozygous or hemizygous deletions, the resolution was set as 10 kb due to the absence of aligned reads) by low-pass GS (a minimal of 15 million reads) as we described previously (Wang et al., 2020). Inconclusive findings included CNVs classified as a variant of uncertain significance, such as intragenic duplications or deletions involving an autosomal recessive gene.

Mate-Pair Genome Sequencing

2 μg of genomic DNA from each case was sheared to fragment sizes ranging from 3 to 8 kb with a red mini-tube on a Covaris device (Covaris, Inc., MA, United States). The fragmented DNA was then prepared for mate-pair library construction following our reported protocols (Dong et al., 2019b). The libraries were sequenced on an MGISEQ-2000 platform (MGI Tech Co., Ltd., Shenzhen, China) for a minimum of 60 million read pairs (paired-end 100 bp) per sample, equivalent to ~ 4X sequencing read-depth.

Genomic Variant Detection

After data QC, the read-pairs were aligned to the human reference genome (GRCh37/hg19) using the Burrows–Wheeler aligner (BWA)(Li and Durbin, 2009). CNV, structural rearrangement (or structural variant, SV), and absence of heterozygosity (AOH) detection were performed according to our previously reported methods (Dong et al., 2019a; Wang et al., 2020; Dong et al., 2021).

CNV detection: Uniquely aligned reads were classified into both adjustable sliding windows (50 kb with 5 kb increments) and non-overlapping windows (5 kb), independently. Subsequently, the copy ratios of all windows were normalized by GC% and our in-house population-based dataset (Chau et al., 2020; Wang et al., 2020). Region(s) with CNV were detected, and the precise boundaries of each CNV were identified by an increment-rate-of-coverage module (Dong et al., 2016) at a resolution of 50 kb. For homozygous or hemizygous deletions, it was reported if there were more than one non-overlapping window with an extremely low number of aligned reads (0.1 as copy ratio) or even absence of aligned read (copy ratio equaled to 0). The minimal size of a reported homozygous or hemizygous deletion was approximately 10 kb.

SV identification: Chimeric read-pairs defined as read-pairs aligned to different chromosomes or to the same chromosome with a genomic distance>=10 kb were selected for event clustering. Each potential event was then filtered against a dataset of systematic errors as well as with optimized parameters (such as minimal of read-pairs supported and the orientation of aligned read-pairs) as described in our previous studies (Dong et al., 2014; Dong et al., 2019b).

AOH analysis: Reads due to PCR duplication were removed, and the coverage of each genomic location was summarized by using the Mpileup module from SAMtools. A genomic locus with a read-depth of 5- to 20-fold with read(s) covered and with at least one read supporting a mutant base type was selected for the determination of heterozygous or homozygous SNV. The number of heterozygous SNVs and homozygous SNVs were calculated per window (with fixed size: 100-kb), respectively, and normalized by the average rate in that sample. Regions with AOH were indicated by a simultaneous decrease in the rate of heterozygous SNVs and increase in the rate of homozygous SNVs (Dong et al., 2021).

Candidate CNVs, SVs, and AOHs were filtered against our in-house datasets, the 1,000 Genomes Project, and gnomAD SVs to filter the known common variants in the populations.

Variant Verification

For verification of structural rearrangements, rearrangement junction-specific PCR and Sanger sequencing were performed (Dong et al., 2014). Primers were designed by using the online software Primer3, Primer-Blast (NCBI), and *in silico* PCR (UCSC). PCR was performed in case and negative control simultaneously, and the products were sequenced on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, CA, United States). The Sanger sequencing results were aligned to the reference genome by BLAT (UCSC) for breakpoint verification and delineation.

For CNV verification, qPCR with primers targeting the candidate region was performed as previously described (Wang et al., 2020). Primers were designed with Primer 3 Web, Primer-Blast (NCBI), or *in silico* PCR (UCSC) based on the reference genome (GRCh37/hg19). The melting curve analysis was carried out for each pair of primers to ensure specificity of the PCR amplification, and the standard curve method was used to determine PCR efficiency (within a range of 95 – 105%). Each reaction was performed in duplicate in 10-μL of reaction mixtures simultaneously in case and control (in-house normal male and female controls) using the SYBR Select Master Mix (Applied Biosystems). The reactions were run on a 7900HT Real-Time PCR System (Applied Biosystems) using the default reaction conditions. The copy numbers in each sample were determined by the ΔΔ Ct (cycle threshold) method, which compared the difference in Ct of the targeted region with a reference primer pair targeting a universally conserved element in a case against control.

For verification of AOH, a well-established, customized CMA 8X60k Fetal DNA Chip v2.0 (Agilent Technologies, Santa Clara, CA, United States), containing both SNP and comparative genomic hybridization (CGH) probes, was used as previously described (Chau et al., 2019). CNV and AOH analyses were evaluated with CytoGenomics (Agilent).

Annotation and Pathogenicity Prediction

For CNVs and SVs, the breakpoints/boundaries identified by mate-pair GS were used for annotation: (1) direct disruption or involvement of gene(s), or (2) disruption of topologically associated domains (https://www.clintad.com/single/) in which

TABLE 1 | Detection results of two methods of the 24 inconclusive cases.

Case ID	Clinical details	Fetalseq CNV results	Reported results	Mate-pair genome sequencing results
		Deletion		
1	Delay	seq[GRCh37] del(5)(q14.3) chr5: g.90028949_ 90237360del	Pathogenic variant on autosomal recessive gene	seq[GRCh37] del(5)(q14.3) chr5: g.90027969_90240857del
8	Delay	seq[GRCh37] del(2)(p24.1) chr2: g.20082407_20142043del	Pathogenic variant on autosomal recessive gene	seq[GRCh37] del(2)(p24.1) chr2:g.20080939-20139774del
10	Delay	seq[GRCh37] del(6)(q12) chr6: g.65418244_65760319del	Pathogenic variant on autosomal recessive gene	seq[GRCh37] del(6)(q12) chr6:g.65415408-65763210del
21	Delay	seq[GRCh37] del(7)(q32.3q33) chr7: g.132543248_132639078del	VUS	seq[GRCh37] del(7)(q32.3q33) chr7: g.132542905-132639717del
26	Developmental delay and microcephaly	seq[GRCh37] del(12)(p11.23) chr12: g.26992893_27345229del	VUS	seq[GRCh37] del(12)(p11.23) chr12: g.26991317-27342205del
37	Bilateral severe hypoplastic vestibular nerve and global delay, ADHD	seq[GRCh37] del(22)(q11.22) chr22: g.22313025_22572225del	VUS	seq[GRCh37] del(22)(q11.22) chr22: g.22313363_22579931del
38	Delay	seq[GRCh37] del(4)(q25) chr4: g.112915276_113354558del	VUS	seq[GRCh37] del(4)(q25) chr4:g.112915198-113354258del
67	ASD, global delay	seq[GRCh37] del(8)(p21.3) chr8: g.19352596_19553354del	Pathogenic variant on autosomal recessive gene	seq[GRCh37] del(8)(p21.3) chr8:g.19352895-19553738del
71	Developmental delay	seq[GRCh37] del(9)(p24.3) chr9: g.99746_402497del	VUS	seq[GRCh37] del(9)(p24.3) chr9: g.110928_398513del
80	Autism, delay	seq[GRCh37] del(11)(p15.4) chr11: g.6907077_7058427del	VUS	seq[GRCh37] del(11)(p15.4) chr11: g.6910893_7062143del
		Duplication		
4	Delay	seq[GRCh37] dup(8)(p23.2) chr8: g.3700597_5946301dup	VUS	dup(8)(8p23.2)(pter->8p23.2(+)(5951139):: q21.3(+)(3686605)- > qter)
9	Epilepsy with mild delay	seq[GRCh37] dup(13)(q13.3) chr13: g.37265048_37433772dup	VUS	dup(13)(q13.3)(pter- > q13.3(+)(37430811):: q13.3(+)(37267951)- > qter)
13	Delay	seq[GRCh37] dup(13)(q12.3q13.2) chr13: g.30805367_34307738dup	VUS	dup(13)(q12.3q13.2)(pter- > q13.2(+)(34291095)::q12.3(+)(30797601)- > qter)
17	Delay	seq[GRCh37] dup(11)(p15.4) chr11: g.9533650_10145145dup	VUS	dup(11)(p15.4)(pter- > p15.4(+)(10148395):: p15.4(+)(9533106)- > qter)
19	Autism, developmental delay	seq[GRCh37] dup(17)(p13.1) chr17: g.6989477_7347779dup	VUS	Complex rearrangement
27	Delay	seq[GRCh37] dup(3)(q25.32) chr3: g.158051611_158591897dup	VUS	dup(3)(q25.32)(pter- > q25.32(+)(158590381):: q25.32(+)(158051006)- > qter)
29	Bilateral congenital hearing loss, history of delay	seq[GRCh37] dup(10)(q22.2) chr10: g.76002141_76107403dup	Pathogenic variant on autosomal recessive gene	dup(10)(q22.2)(pter- > q22.2(+)(76114070):: q22.2(+)(76001841)- > qter)
36	Delay	seq[GRCh37] dup(15)(q21.3) chr15: g.54467876_55401968dup	VUS	dup(15)(q21.3)(pter- > q21.3(+)(55445120):: q21.3(+)(54466811)- > qter)
40	Delay	seq[GRCh37] dup(22)(q11.23) chr22: g.23674079_25063169dup	VUS	LCR
48	Delay FTT, left corneal opacity, dysmorphism	seq[GRCh37] dup(6)(p12.3) chr6: g.46876528_47353335dup	VUS	dup(6)(p12.3)(pter- > p12.3(+)(47364590):: p12.3(+)(46875330)- > qter)
49	Delay	seq[GRCh37] dup(8)(p23.1) chr8: g.8093423_9166490dup	VUS	LCR
55	Delay, subtle dysmorphism	seq[GRCh37] dup(7)(q11.22) chr7: g.69820533_70172074dup	VUS	dup(7)(q11.22)(pter- > q11.22(+)(70166997):: q11.22(+)(69827447)- > qter)
56	Delay	seq[GRCh37] dup(4)(q32.3) chr4: g.165050961_165626257dup	VUS	dup(4)(q32.3)(pter- > q32.3(+)(165626043):: q32.3(+)(165052397)- > qter)
65	Autism, developmental delay	seq[GRCh37] dup(7)(q21.11) chr7: g.82027618_82168623dup	VUS	dup(7)(q21.11)(pter- > q21.11(+)(82155471):: q21.11(+)(82025319)- > qter)
80	Autism, delay	seq[GRCh37] dup(3)(p12.3) chr3: g.79128426_79237810dup	VUS	dup(3)(p12.3)(pter- > p12.3(+)(79237826):: p12.3(+)(79128870)- > qter)

with gene(s) involved. For CNV/SV potentially involving gene(s) that was an OMIM disease-causing gene, or a disease-causing gene due to haploinsufficient/triplosensitivity in peer-reviewed publications, or by ClinGen Dosage Sensitivity Map (https://dosage.clinicalgenome.org/), DECIPHER (https://www.deciphergenomics.org/), or gnomAD (https://gnomad.broadinstitute.org/), it was subjected for further analysis.

For AOHs, if there were multiple regions with AOH (>5 Mb) reported in a case, the overall size was calculated as the sum of all regions with AOHs excluding the ones in sex chromosomes. In contrast, if there were more than one region of AOH identified in one chromosome, uniparental disomy was suspected when the size of interstitial AOH exceeded 15 Mb or the size of terminal AOH exceeded 5 Mb based on the ACMG guideline (Del Gaudio et al., 2020).

RESULTS

Cohort Summary and Mate-Pair Genome Sequencing

In this study, 100 patients (71 male and 29 female) were recruited from 2020 to 2021. All participants were examined by clinical geneticists and received a negative result (n = 76) or an inconclusive finding (n = 24) by previous sequencing-based CNV analysis (**Table 1**). This cohort presented a spectrum of clinical features, mainly involving neurodevelopmental conditions such as intellectual disability and ASD, with or without comorbidities such as dysmorphology, seizure, and hypotonia. Among them, 13 cases (13%) had other congenital anomalies or organ-specific dysfunction (**Supplementary Table S1**).

Investigation of Inconclusive CNVs Reported in Previous Analysis

Among them, 24 cases were referred due to the inconclusive results of CNV analysis which cannot fully explain patients' phenotype, including 15 duplications and 10 deletions. In one case, patient 80 had two CNVs, including a deletion and duplication. We aimed to validate the consistency of CNV detection, and to investigate the directions/orientations of the duplications. We employed both read-depth-based and chimeric-read-pair-based algorithms for CNV detection.

Twenty-five CNVs reported in 24 cases were all detected by mate-pair GS. We also compared the locations of boundaries for the CNVs reported by each method as mate-pair GS enabled the identification of chimeric read-pairs to narrow down the candidate regions of CNV/SVs's breakpoint junctions. These two approaches yielded similar sizes of these 25 CNVs. For ten deletions, the minor discrepancies in the breakpoint coordinates did not affect the clinical interpretation of the CNVs (**Table 1**).

Among the 15 duplications with inconclusive findings, we aimed to determine the directions/orientations of these duplication segments (i.e., tandem forward or reverse duplications, insertions, or complex rearrangements) by chimeric read-pairs. Among them, 12 were identified as forward tandem duplications, and one was found to be involved in complex rearrangements (patient 19). However, the genomic compositions of the other two duplications were unable to be identified by mate-pair GS due to the presence of segmental duplications flanking the CNVs of these two regions: 22q11.23 (patient 40) and 8p23.1 (patient 49) (**Table 1**).

Additional CNV and SV Findings Among all 100 Cases

By using chimeric read-pair analysis among all 100 cases, mate-pair GS revealed five cryptic deletions from four cases, with a size ranging from 8.5 to 46 kb, and ten rare SVs detected from 10 cases including five balanced inversions, and one simple and four complex insertions (**Table 2**). Patient 15 was detected with an 8.5 kb heterozygous deletion involving exon 1 of the *ASAH1* gene which is known to be associated with autosomal recessive spinal muscular atrophy with progressive myoclonic epilepsy [MIM159950]. The deletion was classified as pathogenic CNV in an autosomal recessive gene and confirmed by qPCR (**Supplementary Figure S2**). Patient 23 was detected with 9.5 kb heterozygous deletion involving three exons of the *ANKRD26* gene, which is associated with autosomal dominant thrombocytopenia [MIM 18800] (**Table2**). It indicated a further hematological test was warranted in this patient; however, there was not enough gDNA left for validation. This deletion was further classified as VUS.

AOH Findings

The absence of heterozygosity analysis was applied to each case (n = 100) to detect constitutional and mosaic AOH with a size at 5 Mb. Three cases (3%) were detected with multiple regions with AOH (≥ 5 Mb) identified including case 41 involving imprinting chromosomes (**Supplementary Table S2**). In case 76, a three-year-old boy with autism and delay received a negative result from previous CNV analysis. However, mate-pair GS identified multiple regions with AOH, the overall size of which summed to be 214.5 Mb involving 13 autosomes (**Figure 1**). Multiple regions with AOH in this case were verified by our CMA arrays (aCGH + SNP probes). Therefore, a familial relationship between the parents was suggested. However, this patient was lost to follow-up.

DISCUSSION

Our study showed the feasibility and advantages of applying mate-pair low pass GS in a cohort of 100 patients with neurodevelopmental disorders, and congenital abnormalities with inconclusive or negative findings from previous CNV analysis. Our work also demonstrated that mate-pair GS with a large DNA insert size (~5 kb) and a minimal read-depth of 4-fold enables identification of DNA changes (CNVs or SVs) cryptic to previous CNV analysis, and delineation of the breakpoint junctions. Meanwhile, it also showed the

TABLE 2 | List of additional SVs detected in this cohort.

Case ID	FetalSeq (CNV analysis)	Additional findings	Size (bps)	Gene(s) on breakpoints
		Deletion		
42	Negative	seq[GRCh37] del(9)(q21.32) chr9:g.85918802_85929907del	11,105	*FRMD3*
15	Negative	seq[GRCh37] del(8)(p22) chr8:g.17937910_17946394del;	8,484;	*ASAH1*
		seq[GRCh37] del(1)(q21.3) chr1:g.152250046_152295889del	45,843;	*FLG*
21	VUS, seq[GRCh37] del(7)(q32.3q33) chr7: g.132543248_132639078del	seq[GRCh37] del(17)(q25.1) chr17:g.70909687_70947878del	38,191	*SLC39A11*
23	Negative	seq[GRCh37] del(10)(p12.1) chr10:g.27294954_27304416del	9,462	*ANKRD26*
		Inversion		
4	VUS, seq[GRCh37] dup(8)(p23.2) chr8: g.3700597_5946301dup	seq[GRCh37] inv(14)(q21.2)(pter- > q21.2(+)(44888815)::q21.2(-)(44950538)<-q21.2(-)(44890455)::q21.2(+)(44958120)- > qter)	69,305	—
25	Negative	seq[GRCh37] inv(1)(p22.3)(pter- > p22.3(+)(85672144)::p22.3(-)(85684901)<-p22.3(-)(85672336)::p22.3(+)(85685338)- > qter)	13,194	—
47	Negative	seq[GRCh37] inv(3)(p24.1)(pter- > p24.1(+)(94294177)::p24.1(-)(94319877)<-p24.1(-)(94296491)- > p24.1(+)(94320566)- > qter)	26,389	—
65	VUS, seq[GRCh37] dup(7)(q21.11) chr7: g.82027618_82168623dup	seq[GRCh37] inv(6)(q12)(pter- > q12(+)(66827535)::q12(-)(68075879)<-q12(-)(66828312)::q12(+)(68076174)- > qter)	1,248,639	—
66	Negative	seq[GRCh37] inv(8)(p11.1q11.1)(pter- > p11.1(+)(43669974)::q11.1(-)(48070098)<-p11.1(-)(43671748)::q11.1(+)(48071062)- > qter) inv(15)(q26.3)(pter- > q26.3(+)(100271705)::q26.3(-)(100487648)<-q26.3(-)(100272211)::q26.3(+)(100489231)- > qter)	4,401,088; 217,526	—
		Insertion		
11	Negative	seq[GRCh37] ins(5;5)(q35.3;q35.3)(pter- > q35.3(+)(180499168)::q35.3(-)(180478893)<-q35.3(+)(180416486)::q35.3(+)(180501005)- > qter) dup(5)(q35.3) chr5:g.180416486_180478893dup	20,275	*BTNL3-BTNL9*
30	Negative	Dup ins and flanking dup		
		seq[GRCh37] ins(8;8)(p23.1;p23.3)(pter- > p23.1(+)(6513172)::p23.3(-)(1543512)<-p23.3(-)(1114809)::p23.1(+)(6439080)- > pter)	428,703;	*DLGAP2;*
		dup(8)(p23.3) chr8:g.1114809_1543512dup dup(8)(p23.1) chr8:g.6439080_6513172dup	74, 039	*MCPH11*
36	VUS, seq[GRCh37] dup(15)(q21.3) chr15: g.54467876_55401968dup	Dup ins and flanking dup		
		seq[GRCh37] ins(8;8)(q23.1;q22.3)(pter- > q23.1(+)(1,10119574)::q22.3(-)(104589153)<-q22.3(-)(104465936)::q23.1(+)(109821483)- > qter) dup(8)(q22.3) chr8:g.104465936_104589153dup dup(8)(q23.1) chr8:g.109821483_1,10119574dup	123,217; 298,091	*RIMS2; TRHR*
45	Negative	Dup ins and flanking dup	42,546	*FMLN2, PRPF40A;*
		seq[GRCh37] ins(2;2)(q23.3;q23.3)(pter- > q23.3(+)(153563012)::q23.3(-)(153536242)<-q23.3(-)(153493696)::q23.3(+)(153542212)- > qter) dup(2)(q23.3) chr2:g.153493696_153536242dup		
		dup(2)(q23.3) chr2:g.153542212_153563012dup		*PRPF40A*
69	Negative	Unresolved complex rearrangement		—

robustness of AOH detection by utilizing such limited sequencing read-depth (4-fold).

Through chimeric-read-pair-based algorithm, we further confirmed the robustness of identifying the CNV boundaries by using a read-depth-based algorithm in previous CNV analysis with 0.25-fold genome sequencing data. It is consistent with our previous finding showing no significant differences of the CNV boundaries detected between two methods (Dong et al., 2016). However, the mate-pair-based algorithm shows its advantages in the following aspects.

First, it provides the genomic compositions of duplications with an inconclusive clinical significance: we re-evaluated the 15 duplications classified as VUS in this cohort detected by the previous CNV test (**Table 1**). Among these, the majority of duplications (80%, 12/15) were forward tandem duplication, the incidence of which is comparable to that previously reported (Newman et al., 2015). This suggests that genes located on the breakpoints in ~80% of duplications would be intact. In contrast, if genes located on the breakpoints of duplications cause diseases that explain the phenotypes of the patients, it would highly warrant further evaluation of the orientation of the copy number gains. Still, we have two (13%) that cannot be identified due to flanking low copy repeats, which imply the long LCR would impact the SV detection by this mate-pair genome sequencing.

Second, it identifies small CNVs (<50 kb) that go beyond the resolution of testing through 0.25-fold GS. Five cryptic exonic deletions involving single genes were identified in this study. Although two small clinically significant deletions did not fully explain the patients' neurological issue (case 15 and case 23), the accuracy of detecting such small CNVs have been confirmed by

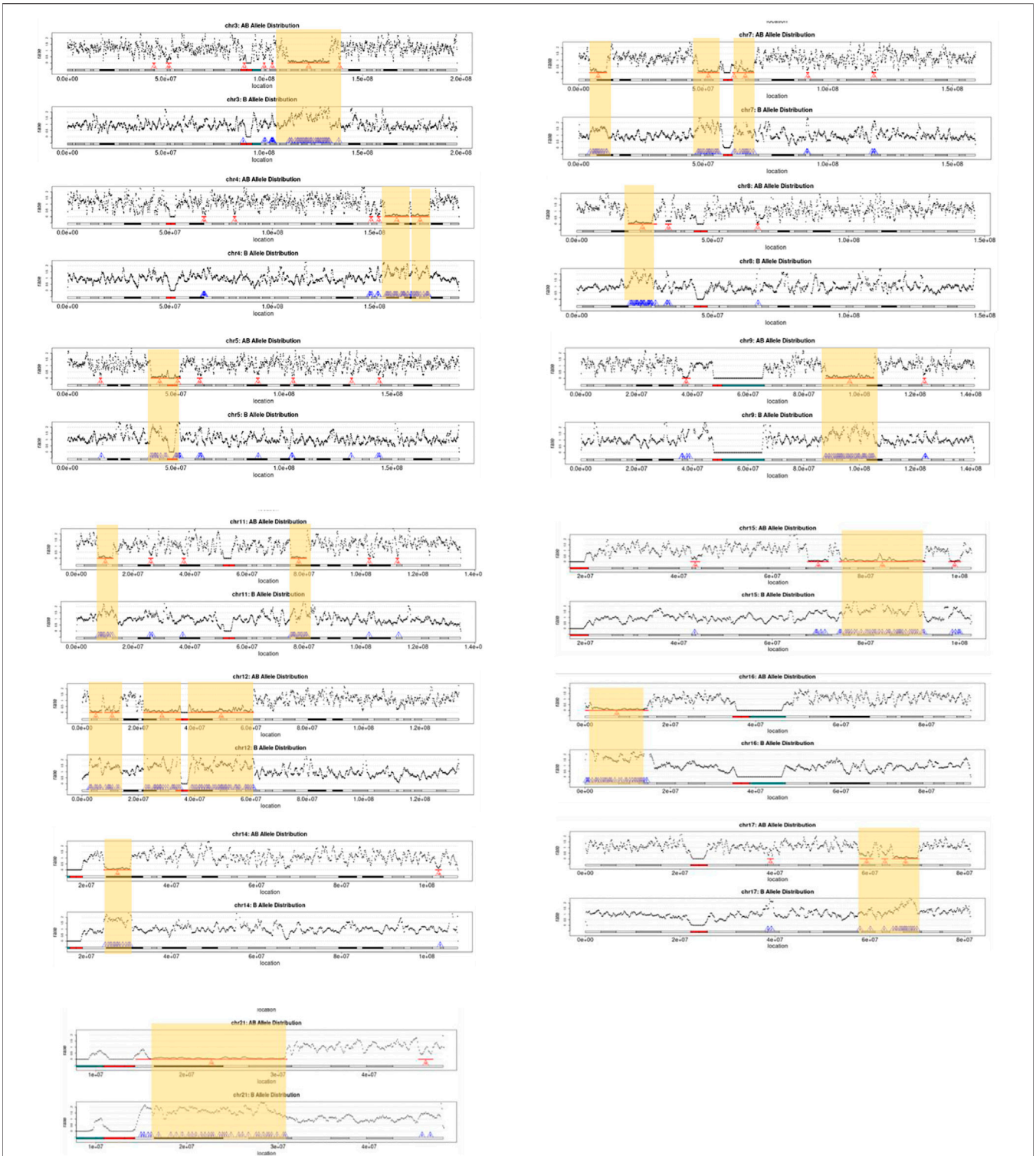

FIGURE 1 | Regions of AOH detected in thirteen chromosomes of case 76. For each chromosome, the AOH regions detected are indicated by yellow highlighted boxes and red arrows, and the number of windows that support the AOH is shown in red (upper figure in each chromosome: AB allele distribution), while windows with an increased rate of homozygous SNVs within regions reported (lower figures in each chromosome: B allele distribution) are shown by blue arrows.

qPCR. In addition, exonic CNVs related to autosomal recessive disorders are often small in size and underappreciated due to the limitations in the routine CNV detection method such as CMA (Yuan et al., 2020). Therefore, mate-pair GS might increase diagnostic yield in cases contributed by small CNVs although this might not be a common cause in NDD patients in this study.

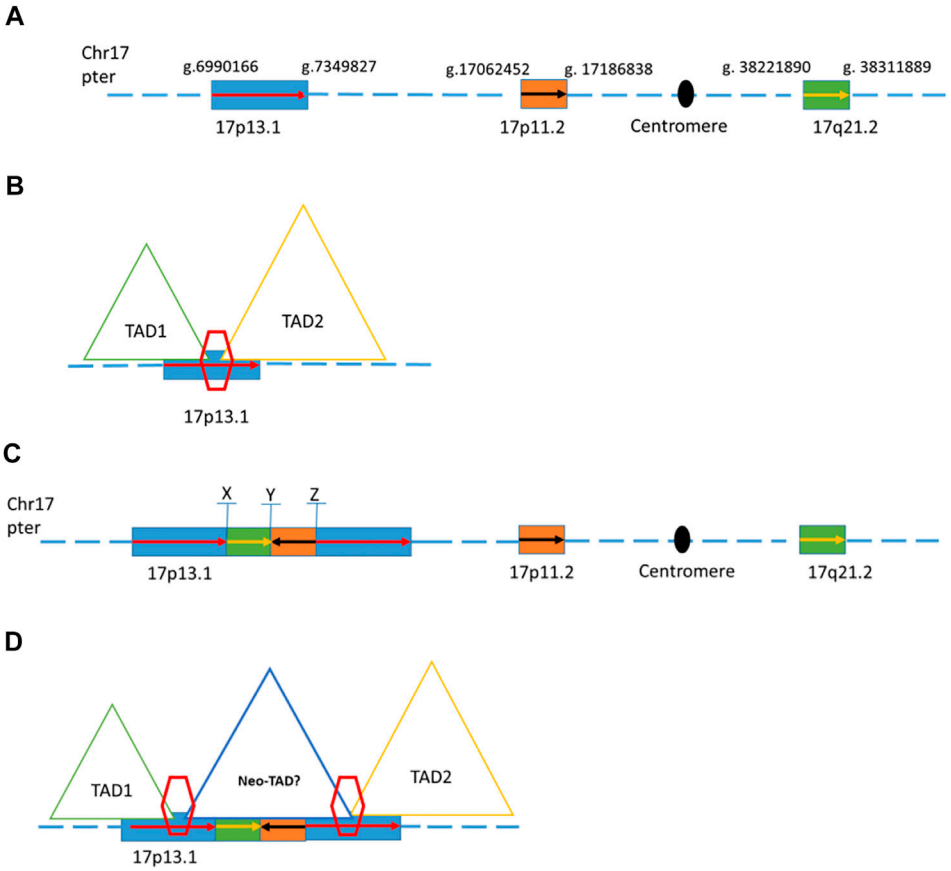

FIGURE 2 | Schematic of genome structures in case 19. **(A)** Wild type of chromosome 17 with blocks of region that involved the complex rearrangement. **(B)** Two different topologically associating domains with a boundary on the 17p13.1 **(C)**. Schematic representation of one possible complex rearrangement on 17p13.1 involving duplications and insertions from 17p13.1, 17p11.2, and 17q21.2. X, Y, and Z indicate the breakpoints within this rearrangement. **(D)** Duplications of the boundary and the flanking regions (inter-TAD duplication) were proposed to change the overall chromatin architecture of the locus, creating a new chromatin domain (neo-TAD) on this complex rearrangement region.

Third, it detects additional SVs and reveals complex rearrangements. In this cohort, five rare inversions were detected in five cases (5%), all of which were small paracentric inversions. But none of these inversions disrupted genes or their interactions with known regulatory elements. They were classified as VUS considering their rarity in the population. In addition, five insertions (5%) were detected, four out of which were involved in complex rearrangements. These five insertions were not related to those previously reported CNVs in each case. Interestingly, three of these four complex insertions were delineated as insertion (duplicated segments) with flanking duplications identified in the insertion site (**Table 2**). Although no gene disruption was observed, we still classify them as VUS. Genome-wide structure rearrangement discovery is challenging while increasingly attracting our attention with the improvement of sequencing detection methods. However, limited information about polymorphic SVs in the human genome hampers its clinical significance interpretation. Genes interrupted by the breakpoint seen in the patients would be current focus to correlate the diseases for interpretation.

One of the previously reported inconclusive CNVs was found to involve complex rearrangements based on the mate-pair GS result. Patient 19 was a five-year-old male child who showed autism, global developmental delay, and mild dysmorphic feathers. A *de novo* 358 kb duplication in chromosome 17 (seq [GRCh37] dup(17)(p13.1)dn chr17:g.6989477_7347779dup) was reported in a previous CNV analysis. Mate-pair GS detected another two genomic segments from distal location of chromosome 17 (a segment of 124 kb from 17p11.2 and a segment of 90 Kb from 17q21.2) inserted in the middle of these two copies of 358 kb segment of 17p13.1 (**Figure 2**). The composition of this complex rearrangement is shown in **Figure 2C**. The 124 kb insertion from 17p11.2 was in a reverse orientation. The three breakpoints were all validated by Sanger sequencing (**Supplementary Figure S1**). This 358 kb duplication is overlapped with a dosage-sensitive region on the 17p13.1 commonly leading to intellectual disability and microcephaly (Carvalho et al., 2014). Multiple patients with overlapping deletions or triplication changes shared microcephaly and intellectual disability, and defined the smallest region of overlapping (SRO) on the 17p13.1 as

around 156 kb in size (GRCh37/hg19, chr17:g.7055654_7212104) (Carvalho et al., 2014). In addition, defects in the *DLG4* gene of this region are known to cause intellectual developmental disorder 62 (MIM 618793) due to haploinsufficiency (Lelieveld et al., 2016; Moutton et al., 2018). Mooneyham et al. reported two patients with neurodevelopmental delays and absolute/relative macrocephaly with a shared region of 62.5 kb on the 17p13.1, suggesting that *DULLARD*, *DLG4*, and *GABARAP* genes would be the candidate genes for neurodevelopmental delays identified in this patient (GRCh37/hg19, chr17:g.7094072_7156584) (Mooneyham et al., 2014). Currently, this 358 kb duplication is known to involve a TAD boundary (**Figure 2B**). The two insertions might result in an overexpression of those genes locating in the 358 kb duplication by bringing in additional regulatory elements to possibly promote certain ectopic enhancer-promoter interactions in the neo-TAD or expression of genes in the inserted regions (**Figure 2D**).

Last, it would identify regions with AOH. Small regions with AOH (<3 Mb) in the human genome are commonly seen, while regions with AOH are also known to cause diseases by unmasking of autosomal recessive allele or imprinting region. The prevalence of UPD associated with a clinical presentation due to imprinting disorders or recessive diseases ranges from 1 in 3,500 to 1 in 5,000 (Del Gaudio et al., 2020). Studies suggest reporting terminal long continuous stretches of homozygosity (LSCH) on each chromosome at a resolution of 5 Mb and interstitial LSCH at 15–20 Mb (Hoppman et al., 2018). In this study, we applied 5 Mb as the resolution for identifying regions with AOH as demonstrated in our previous publication (Dong et al., 2021). The result showed that 3% of cases from our cohort were reported to have regions with AOH on various chromosomes more than imprinting chromosome. One of the patients was detected with multiple AOHs in 13 chromosomes, with an overall size of 214.5 Mb. These large regions of homozygosity involving multiple chromosomes indicate a consanguineous relationship between the proband's parents which was suggested to report as incidental findings based on the current laboratory's reporting policy. Such information is also important for clinicians to further evaluate the possibility of any gene locating in regions of AOH is known to be associated with a patient's presentation (Del Gaudio et al., 2020) as parental consanguinity is known to contribute to developmental delay or autism spectrum disorder due to the increased risks of autosomal recessive disorders.

In summary, this study showed the feasibility of mate-pair low-pass GS in patients with neurodevelopmental disorders who received negative or inclusive results from previous CNV analysis. This approach complements the first-tier CNV analysis for NNDs through not only increasing the resolution of CNVs detection but also better identification and delineation of chromosomal structural rearrangements as well as the discovery of potential causative regions (or genes) involved in regions with AOH.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://db.cngb.org/cnsa/; CNP0002205.

ETHICS STATEMENT

This study protocol was approved by the Ethics Committee of the Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee (CREC Ref. No. 2019.600). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YC, HL, YZ, ZD, KC, and FL designed the study. YC, HL, SC, and FL collected the samples and followed up. YC, HL, YZ, MC, SX, SC, AL, JC, TL, ZD, KC, and FL performed the analysis and data interpretation. YZ, MC, and SX conducted the validation. YC, YZ, MC, ZD, and KC wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by the Health and Medical Research Fund (07180576), Direct Grant (2020.052).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.803088/full#supplementary-material

REFERENCES

Abel, H. J., Larson, D. E., Larson, D. E., Regier, A. A., Chiang, C., Das, I., et al. (2020). Mapping and Characterization of Structural Variation in 17,795 Human Genomes. *Nature* 583, 83–89. doi:10.1038/s41586-020-2371-0

Bertelsen, B., Nazaryan-Petersen, L., Sun, W., Mehrjouy, M. M., Xie, G., Chen, W., et al. (2016). A Germline Chromothripsis Event Stably Segregating in 11 Individuals through Three Generations. *Genet. Med.* 18, 494–500. doi:10.1038/gim.2015.112

Carvalho, C. M. B., Vasanth, S., Shinawi, M., Russell, C., Ramocki, M. B., Brown, C. W., et al. (2014). Dosage Changes of a Segment at 17p13.1 lead to Intellectual Disability and Microcephaly as a Result of Complex Genetic Interaction of Multiple Genes. *Am. J. Hum. Genet.* 95, 565–578. doi:10.1016/j.ajhg.2014.10.006

Chau, M. H. K., Cao, Y., Kwok, Y. K. Y., Chan, S., Chan, Y. M., Wang, H., et al. (2019). Characteristics and Mode of Inheritance of Pathogenic Copy Number Variants in Prenatal Diagnosis. *Am. J. Obstet. Gynecol.* 221, 493.e1–493.e11. doi:10.1016/j.ajog.2019.06.007

Chau, M. H. K., Wang, H., Lai, Y., Zhang, Y., Xu, F., Tang, Y., et al. (2020). Low-Pass Genome Sequencing: a Validated Method in Clinical Cytogenetics. *Hum. Genet.* 139, 1403–1415. doi:10.1007/s00439-020-02185-9

Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., et al. (2017). Defining the Diverse Spectrum of Inversions, Complex Structural Variation, and Chromothripsis in the Morbid Human Genome. *Genome Biol.* 18, 36. doi:10.1186/s13059-017-1158-6

de Pagter, M. S., Van Roosmalen, M. J., Baas, A. F., Renkens, I., Duran, K. J., Van Binsbergen, E., et al. (2015). Chromothripsis in Healthy Individuals Affects Multiple Protein-Coding Genes and Can Result in Severe Congenital Abnormalities in Offspring. *Am. J. Hum. Genet.* 96, 651–656. doi:10.1016/j.ajhg.2015.02.005

Del Gaudio, D., Shinawi, M., Astbury, C., Tayeh, M. K., Deak, K. L., Raca, G., et al. (2020). Diagnostic Testing for Uniparental Disomy: a Points to Consider Statement from the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* 22, 1133–1141. doi:10.1038/s41436-020-0782-9

D'haene, E., and Vergult, S. (2021). Interpreting the Impact of Noncoding Structural Variation in Neurodevelopmental Disorders. *Genet. Med.* 23, 34–46. doi:10.1038/s41436-020-00974-1

Dong, Z., Jiang, L., Yang, C., Hu, H., Wang, X., Chen, H., et al. (2014). A Robust Approach for Blind Detection of Balanced Chromosomal Rearrangements with Whole-Genome Low-Coverage Sequencing. *Hum. Mutat.* 35, 625–636. doi:10.1002/humu.22541

Dong, Z., Zhang, J., Hu, P., Chen, H., Xu, J., Tian, Q., et al. (2016). Low-Pass Whole-Genome Sequencing in Clinical Cytogenetics: a Validated Approach. *Genet. Med.* 18, 940–948. doi:10.1038/gim.2015.199

Dong, Z., Yan, J., Xu, F., Yuan, J., Jiang, H., Wang, H., et al. (2019a). Genome Sequencing Explores Complexity of Chromosomal Abnormalities in Recurrent Miscarriage. *Am. J. Hum. Genet.* 105, 1102–1111. doi:10.1016/j.ajhg.2019.10.003

Dong, Z., Zhao, X., Li, Q., Yang, Z., Xi, Y., Alexeev, A., et al. (2019b). Development of Coupling Controlled Polymerizations by Adapter-Ligation in Mate-Pair Sequencing for Detection of Various Genomic Variants in One Single Assay. *DNA Res.* 26, 313–325. doi:10.1093/dnares/dsz011

Dong, Z., Chau, M. H. K., Zhang, Y., Yang, Z., Shi, M., Wah, Y. M., et al. (2021). Low-pass Genome Sequencing-Based Detection of Absence of Heterozygosity: Validation in Clinical Cytogenetics. *Genet. Med.* 23, 1225–1233. doi:10.1038/s41436-021-01128-7

Fan, Y.-S., Ouyang, X., Peng, J., Sacharow, S., Tekin, M., Barbouth, D., et al. (2013). Frequent Detection of Parental Consanguinity in Children with Developmental Disorders by a Combined CGH and SNP Microarray. *Mol. Cytogenet.* 6, 38. doi:10.1186/1755-8166-6-38

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural Variation in the Sequencing Era. *Nat. Rev. Genet.* 21, 171–189. doi:10.1038/s41576-019-0180-9

Hoppman, N., Rumilla, K., Lauer, E., Kearney, H., and Thorland, E. (2018). Patterns of Homozygosity in Patients with Uniparental Disomy: Detection Rate and Suggested Reporting Thresholds for SNP Microarrays. *Genet. Med.* 20, 1522–1527. doi:10.1038/gim.2018.24

Kaminsky, E. B., Kaul, V., Paschall, J., Church, D. M., Bunke, B., Kunig, D., et al. (2011). An Evidence-Based Approach to Establish the Functional and Clinical Significance of Copy Number Variants in Intellectual and Developmental Disabilities. *Genet. Med.* 13, 777–784. doi:10.1097/gim.0b013e31822c79f9

Lelieveld, S. H., Reijnders, M. R. F., Pfundt, R., Yntema, H. G., Kamsteeg, E.-J., De Vries, P., et al. (2016). Meta-analysis of 2,104 Trios Provides Support for 10 New Genes for Intellectual Disability. *Nat. Neurosci.* 19, 1194–1196. doi:10.1038/nn.4352

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Manickam, K., Mcclain, M. R., Demmer, L. A., Biswas, S., Kearney, H. M., Malinowski, J., et al. (2021). Exome and Genome Sequencing for Pediatric Patients With Congenital Anomalies or Intellectual Disability: An Evidence-Based Clinical Guideline of the American College of Medical Genetics and Genomics (ACMG). *Genet Med.*

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., et al. (2010). Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am. J. Hum. Genet.* 86, 749–764. doi:10.1016/j.ajhg.2010.04.006

Mooneyham, K. A., Holden, K. R., Cathey, S., Dwivedi, A., Dupont, B. R., and Lyons, M. J. (2014). Neurodevelopmental Delays and Macrocephaly in 17p13.1 Microduplication Syndrome. *Am. J. Med. Genet.* 164, 2887–2891. doi:10.1002/ajmg.a.36708

Moutton, S., Bruel, A.-L., Assoum, M., Chevarin, M., Sarrazin, E., Goizet, C., et al. (2018). Truncating Variants of the DLG4 Gene Are Responsible for Intellectual Disability with Marfanoid Features. *Clin. Genet.* 93, 1172–1178. doi:10.1111/cge.13243

Newman, S., Hermetz, K. E., Weckselblatt, B., and Rudd, M. K. (2015). Next-generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am. J. Hum. Genet.* 96, 208–220. doi:10.1016/j.ajhg.2014.12.017

Palumbo, P., Palumbo, O., Leone, M. P., Stallone, R., Palladino, T., Zelante, L., et al. (2015). Maternal Uniparental Isodisomy (iUPD) of Chromosome 4 in a Subject with Mild Intellectual Disability and Speech Delay. *Am. J. Med. Genet.* 167, 2219–2222. doi:10.1002/ajmg.a.37142

Pócza, T., Grolmusz, V. K., Papp, J., Butz, H., Patócs, A., and Bozsik, A. (2021). Germline Structural Variations in Cancer Predisposition Genes. *Front. Genet.* 12, 634217. doi:10.3389/fgene.2021.634217

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., et al. (2019). Meta-analysis and Multidisciplinary Consensus Statement: Exome Sequencing Is a First-Tier Clinical Diagnostic Test for Individuals with Neurodevelopmental Disorders. *Genet. Med.* 21, 2413–2421. doi:10.1038/s41436-019-0554-6

Talkowski, M. E., Rosenfeld, J. A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., et al. (2012). Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries. *Cell* 149, 525–537. doi:10.1016/j.cell.2012.03.028

Wang, H., Dong, Z., Zhang, R., Chau, M. H. K., Yang, Z., Tsang, K. Y. C., et al. (2020). Low-pass Genome Sequencing versus Chromosomal Microarray Analysis: Implementation in Prenatal Diagnosis. *Genet. Med.* 22, 500–510. doi:10.1038/s41436-019-0634-7

Werling, D. M., Brand, H., An, J.-Y., Stone, M. R., Zhu, L., Glessner, J. T., et al. (2018). An Analytical Framework for Whole-Genome Sequence Association Studies and its Implications for Autism Spectrum Disorder. *Nat. Genet.* 50, 727–736. doi:10.1038/s41588-018-0107-y

Yu, T., Li, J., Li, N., Liu, R., Ding, Y., Chang, G., et al. (2016). Obesity and Developmental Delay in a Patient with Uniparental Disomy of Chromosome 2. *Int. J. Obes.* 40, 1935–1941. doi:10.1038/ijo.2016.160

Yuan, B., Wang, L., Liu, P., Shaw, C., Dai, H., Cooper, L., et al. (2020). CNVs Cause Autosomal Recessive Genetic Diseases with or without Involvement of SNV/indels. *Genet. Med.* 22, 1633–1641. doi:10.1038/s41436-020-0864-8

Yuan, H., Shangguan, S., Li, Z., Luo, J., Su, J., Yao, R., et al. (2021). CNV Profiles of Chinese Pediatric Patients with Developmental Disorders. *Genet. Med.* 23, 669–678. doi:10.1038/s41436-020-01048-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fgene.2022.887176

Combining Z-Score and Maternal Copy Number Variation Analysis Increases the Positive Rate and Accuracy in Non-Invasive Prenatal Testing

Liheng Chen[1,2], Lihong Wang[3], Zhipeng Hu[1], Yilun Tao[1], Wenxia Song[4], Yu An[2,5] and Xiaoze Li[1]**

[1]Department of Medical Genetics, Changzhi Maternal and Child Health Care Hospital, Changzhi, China, [2]School of Life Sciences, Fudan University, Shanghai, China, [3]Department of Pediatrics, Changzhi Maternal and Child Health Care Hospital, Changzhi, China, [4]Obstetrics Department, Changzhi Maternal and Child Health Care Hospital, Changzhi, China, [5]Human Phenome Institute, Zhangjiang Fudan International Innovation Center, MOE Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Cynthia Casson Morton,
Brigham and Women's Hospital,
United States

Reviewed by:
Ye Cao,
The Chinese University of Hong Kong,
China
Zirui Dong,
The Chinese University of Hong Kong,
China

***Correspondence:**
Yu An
Anyu@fudan.edu.cn
Xiaoze Li
lixiaoze520@126.com

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 01 March 2022
Accepted: 02 May 2022
Published: 02 June 2022

Citation:
Chen L, Wang L, Hu Z, Tao Y, Song W, An Y and Li X (2022) Combining Z-Score and Maternal Copy Number Variation Analysis Increases the Positive Rate and Accuracy in Non-Invasive Prenatal Testing. Front. Genet. 13:887176. doi: 10.3389/fgene.2022.887176

Objective: To evaluate positive rate and accuracy of non-invasive prenatal testing (NIPT) combining Z-score and maternal copy number variation (CNV) analysis. To assess the relationship between Z-score and positive predictive value (PPV).

Methods: This prospective study included 61525 pregnancies to determine the correlation between Z-scores and PPV in NIPT, and 3184 pregnancies to perform maternal CNVs analysis. Positive results of NIPT were verified by prenatal diagnosis and/or following-up after birth. Z-score grouping, logistic regression analysis, receiver operating characteristic (ROC) curves, and S-curve trends were applied to correlation analysis of Z-scores and PPV. The maternal CNVs were classified according to the technical standard for the interpretation of ACMG. Through genetic counseling, fetal and maternal phenotypes and family histories were collected.

Results: Of the 3184 pregnant women, 22 pregnancies were positive for outlier Z-scores, suggesting fetal aneuploidy. 12 out of 22 pregnancies were true positive (PPV = 54.5%). 17 pregnancies were found maternal pathogenic or likely pathogenic CNVs (> 0.5 Mb) through maternal CNV analysis. Prenatal diagnosis revealed that 7 out of 11 fetuses carried the same CNVs as the mother. Considering the abnormal biochemical indicators during pregnancy and CNV-related clinical phenotypes after birth, two male fetuses without prenatal diagnosis were suspected to carry the maternally-derived CNVs. Further, we identified three CNV-related family histories with variable phenotypes. Statistical analysis of the 61525 pregnancies revealed that Z-scores of chromosomes 21 and 18 were significantly associated with PPV at $3 \leq Z \leq 40$. Notably, three pregnancies with $Z > 40$ were both maternal full aneuploidy. At $Z < -3$, fetuses carried microdeletions instead of monosomies. Sex chromosome trisomy was significantly higher PPV than monosomy.

Conclusion: The positive rate of the NIPT screening model combining Z-score and maternal CNV analysis increased from 6.91‰ (22/3184) to 12.25‰ (39/3184) and true positives increased from 12 to 21 pregnancies. We found that this method could improve the positive rate and accuracy of NIPT for aneuploidies and CNVs without increasing testing costs. It provides an early warning for the inheritance of pathogenic CNVs to the next generation.

Keywords: non-invasive prenatal testing (NIPT), copy number variations (CNVs), aneuploidies, prenatal diagnosis, birth defects, positive predictive value (PPV), z-scores

INTRODUCTION

Non-invasive prenatal testing (NIPT) based on high-throughput sequencing can detect fetal common chromosomal aneuploidies. The existence of placental cell-free fetal DNA (cff-DNA) fragments in the peripheral blood of pregnant women provide a basis for NIPT technology (Lo et al., 1997). Mostly, NIPT result was calculated by Z-score in which the individual sample is compared with a control group of normal (diploid) samples (Chiu et al., 2008). Numerous studies have shown that the model's accuracy is higher than that of serological screening technology, regardless of single and twin pregnancies (Zhang et al., 2015; Gil et al., 2017; Iwarsson et al., 2017; Gil et al., 2019). However, there are false results that cannot be avoided owing to the limitations of the materials and methods. Several studies indicated the accuracy and positive predictive value (PPV) of NIPT were related to Z-score; and the higher the Z-score, the greater the likelihood of true positive (Tian et al., 2018; Wan et al., 2021). Another study showed that the optimal cut-off values for trisomy (T) 21 and T18 Z-scores were 5.79 and 6.05, respectively (Zhou et al., 2021), which PPV in the group of Z-score > optimal cutoff value was higher than that in the group of 3 ≤ Z-score < optimal cutoff value. However, increasing the cut-off value will produce more false negatives. Therefore, more research is necessary on the relationship between Z-score and PPV before adjusting the cut-off value.

Maternal copy number variations (CNVs) were ignored in either NIPT or NIPT-plus except identifying of fetal *de novo* CNVs. Pathogenic CNVs cause over 300 types of chromosomal microdeletion/microduplication syndromes (MMS), with a total incidence of 3% (Weise et al., 2012; Nevado et al., 2014; Levy et al., 2018). It is very common for heterogeneity of clinical feature due to the location and size of CNVs. A few of fetal structural abnormalities resulted from microdeletion/microduplication could be found by ultrasound screening, however, most of the MMS could not be identified during pregnancy (Grati et al., 2015). Recently, there were studies to shown that the PPV of fetal CNVs detected by NIPT-plus was 20–40% (Chen et al., 2019; Hu et al., 2019). However, no research pay attention to pathogenic CNVs inherited from parents. We found maternal CNV analysis through NIPT without additional cost was meaningful for prediction of birth defects and future treatment.

This prospective study explored a new NIPT model combining Z-score and maternal CNV analysis to identify high-risk fetuses, including aneuploidies and CNVs of each chromosome. In addition, we investigated whether the Z-score was correlated with PPV to assess the accuracy of NIPT-positive results from a single center within the past 5 years. This study aimed to provide a more accurate basis for clinical genetic counselling.

MATERIALS AND METHODS

Subjects

Pregnant women selected for NIPT in Changzhi Maternal and Child Health Care Hospital were continuously included in this study from October 2016 to November 2021. Testing was successful in 61,525 pregnant women, of which 32,361 pregnancies from October 2016 to June 2019 were derived from (Li et al., in press). The study included pregnant women with aneuploidy to investigate the effects of maternal aneuploidy on the Z-score, and we also recommended that pregnant women participated simultaneously in invasive prenatal diagnosis. A total of 3,184 pregnant women were selected from August to November 2021 for NIPT combining Z-score and maternal CNV analysis (CNV > 0.5 Mb). All pregnant women voluntarily signed informed consent forms prior to the procedure. Unique identifiers were deleted before they were included in the study. All procedures were approved by the Medical Ethics Committee of Changzhi Maternal and Child Health Care Hospital (No. CZSFYLL2021017).

Noninvasive Prenatal Testing

Plasma was separated via a two-step centrifugation process within 72 h after collecting 5–8 ml maternal peripheral blood using a dedicated cell-free DNA collection tube. After cf-DNA extraction, library construction, and pooling, samples were sequenced on the Illumina NextSeq CN500 or NextSeq 550Dx platforms in collaboration with Findgene (Shanghai, China) or Biosan (Hangzhou, China). Sequences were aligned to the human genome-wide standard sequence (GRCh37) using BWA software, and Z-scores for each chromosome were obtained from bioinformatics analysis. Z-score were corrected by a series of bioinformatics methods such as normalization, GC correction and filtering out maternal CNVs. But it cannot correct for maternal aneuploidy interference. The control used a non-fixed reference set (96 experimental samples per batch) for internal comparison to eliminate batch differences. Qualified samples required the row sequencing reads greater than 3.5Mb and the fetal frequency greater than 4%. The thresholds of aneuploidy were ±3. Below the lower limit indicated a high

TABLE 1 | Results from the Cohort of 3184 Pregnancies with NIPT Combining Z-score and Maternal CNV.

Groups	Number of Outlier Z-scores						Number of maternal CNVs			Total (PR)
	T21	T18	T13	SCAs	OAAs	Total (PR)	Del	Dup	Total (PR)	
NIPT+	7	4	1	6	4	22 (6.91‰)	12	5	17 (5.34‰)	39 (12.25‰)
TP	7	1	1	2	1	12	5*	4	9*	19

**Including 2 fetuses with phenotypes related-CNV without genomic diagnosis.*

NIPT, Non-invasive Prenatal Testing; NIPT+, NIPT positive result; CNVs, copy number variations; T, trisomy; SCAs, Sex chromosome aneuploidies; OAAs, Other autosome aneuploidies (excepting Chr21, Chr18 and Chr13); PR, positive rate; Del, Microdeletion; Dup, Microduplication; TP, true positive.

risk of monosomy, above the upper limit indicated a high risk of trisomy. Between -3 and +3 represented a low risk of aneuploidy. Interpreting results from sex chromosome aneuploidies (SCAs) required combining the Z-scores of chromosome (Chr) X and ChrY. The Z-scores of ChrX and ChrY should be between −3 and 3 in normal female fetuses. The Z-scores of ChrX and ChrY in normal male fetuses should be < −3 and >3, respectively. All cases with positive results for the first time were verified in another plasma, and only the verified results were included in statistical analyses.

Analysis of Maternal Copy Number Variations

CNVs were detected using sliding window algorithm counting reads in each continuous bins (100kb size). To verify that the method for maternal CNVs is reliable, we performed genomic testing of maternal own lymphocytes by CNV-seq or SNP-array technology. CNVs were classified according to the technical standard for the interpretation and reporting of constitutional copy-number variants by the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen) version 2020 (Riggs et al., 2020). Maternal phenotype and family history were assessed through genetic counselling. Prenatal diagnosis was recommended for pregnancies with maternal CNVs.

Prenatal Diagnosis

Prenatal diagnosis was recommended in pregnant women with NIPT-positive. Amniocentesis was performed under ultrasound guidance at 18–23 gestational weeks or umbilical blood was performed when over 23 gestational weeks, with the consent of pregnant women and family members. Prenatal diagnosis techniques were chosen by one or a combination of karyotyping, SNP-array or CNV-seq. Operations and analyses were performed in accordance with relevant international and national guidelines. The detailed precedures were applied as previously reported (Ma et al., 2021).

Follow-Up

Pregnant women with NIPT-positive results who were not prenatally diagnosed at our institution were followed up to confirm the prenatal diagnosis at other institution or to perform postnatal diagnosis if they chose to continue their pregnancies.

Statistics

Logistic regression analysis was applied to associate Z-scores with the PPV of NIPT-positive results. The receiver operating characteristic (ROC) curve is a comprehensive index that reveals the relationship between sensitivity and specificity. The larger the area under the curve (AUC), the higher the accuracy. Analyses were in the R version 4.1.2. The trend of S-curves was drawn based on the Z-scores and PPV in MATLAB version R2021b with a single parameter logistic model, using the function $\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{1+\boldsymbol{a}_0 \boldsymbol{e}^{-\boldsymbol{x}}}$.

RESULTS

Efficiency of the New Model Combining Z-Score and Maternal CNV

We combined Z-score and maternal CNV to analyze data from 3,184 pregnancies with 3,090 singletons and 94 twins, and 90 pregnancies using assisted reproductive technology. Participants were 30.24 ± 4.46 years old (range, 16–46 years). As shown in **Table 1**, a total of 22 pregnancies (all singletons) were positive, with outlier Z-scores (Z < -3 or Z > 3) suggesting fetal aneuploidy. Twelve were true positives (PPV = 54.5%), including seven with T21, one with T18, one with T13, two with SCAs, and one with T9. Maternal CNV analysis showed that 17 pregnant women had pathogenic or likely pathogenic CNVs, interpreted through websites such as DECIPHER (https://www.deciphergenomics.org) and ClinGen (https://dosage.clinicalgenome.org) in **Table 2**. They did not overlap with the 22 positives analyzed using Z-scores. The NIPT-positives increased from 22 (6.91‰) analyzed using Z-scores to 39 (12.25‰) analyzed using Z-scores and maternal CNVs. The consequences of validation by maternal own lymphocytes were consistent (**Supplementary Table S1**). Among 17 maternal CNVs group, 10 pregnant women chose prenatal diagnosis, of that seven fetuses carried the same CNVs as their mother, indicating that 70% (7/10) CNVs were passed to the next generation. The other seven pregnant women refused prenatal diagnosis. Of two male fetuses undiagnosed exhibited extremely low levels of unconjugated estriol (uE3) during pregnancy, a biochemical indicator of the X-linked ichthyosis (caused by the CNVs), and present skin symptoms after birth; thus we suspected them of having the same CNVs as their mothers. Through genetic counseling, we found that some pregnant women exhibited CNV-related phenotypes in themselves or relatives what they ignored before.

We compared 17 maternal CNVs with their Z-scores to explore the effect of CNV on Z-score. Z-scores with maternal CNVs were different between before and after correction. Z-scores before correction shown that 6 out of 13 autosomes and 2 out of 4 chromosome X had outlier Z-scores. Of maternal CNVs below 2 Mb size, 9 out of 11 had Z-scores before correction in the normal

TABLE 2 | 17 Maternal CNVs Detected by NIPT and Fetal Diagnosis Results.

Sample ID	Maternal CNV	size (Mb)	Z_{bc}	Z_{ac}	CNV Interpretation	Maternal Phenotypes	Fetal Diagnosis
21J101249	seq[GRCh37]chr22: g.21706150-24644732 x1	2.94	–5.193	0.548	Contained 92 protein-coding genes, a 3-point HI gene and a 3-point HI genomic, a variable clinical phenotype such as global developmental delay, cleft lip, behavioral problems and mild dysmorphic facial features.	Slight facial asymmetry, communication impairment, her daughter suffered from cleft palate and mental/physical retardation.	seq[GRCh37]chr22: g.21702383-24620002x1
21J104405	seq[GRCh37]chr16: g.14889818-16535522 x3	1.65	1.414	0.227	Contained 16 protein-coding genes, a 2-point TS genomic region, a variable clinical presentation, lower penetrance.	No obvious abnormality.	arr[GRCh37]chr16: g.15406415-16282869x3
21J101817	seq[GRCh37]chr4: g.182695733-189079179 x1	6.38	–6.731	–0.428	Contained 37 protein-coding genes, symptomatic seizures, short stature.	Height less than 150cm, No obvious abnormality in intelligence.	arr[GRCh37]46,XN
21J104345	seq[GRCh37]chr17: g.14161233-15458439 x1	1.30	–2.137	–0.529	Contained 6 protein-coding genes, a 3-point HI gene and a 3-point HI genomic region, hereditary neuropathy with liability to pressure palsies (HNPP), few symptoms on many individuals.	No obvious abnormality; her deceased mother suffered from leg discomfort after middle age.	arr[GRCh37]chr17: g.14099565-15482833x1
21J101676	seq[GRCh37]chrX: g.2795214-17648380 x1	14.85	–7.708	0.884	Contained 66 protein-coding genes, 9 3-point HI genes and a 3-point HI genomic region, female carriers were unaffected or milder phenotypes.	Abortion history, This pregnancy is a female fetus.	seq[GRCh37]46,XN
21J105324	seq[GRCh37]chrX: g.6445119-8104085 x1	1.69	–17.549	–17.768	Contained 5 protein-coding genes, a 3-point HI genomic region, X-linked ichthyosis in males, mild or unaffected in females.	No obvious abnormality. This pregnancy was a male fetus and MoM value of uE3 was 0.08 (Standard Range, >0.7).	No prenatal diagnosis; Skin lesions on limbs 2 months after birth.
21J104580	seq[GRCh37]chr2: g.111195659-113121587 x3	1.93	–0.205	–1.512	Contained 11 protein-coding genes, a 2-point TS genomic region, a variable clinical phenotypes including developmental delay, tooth abnormalities, hypotonia, and neuropsychiatric conditions.	No obvious abnormality except tooth abnormality	No prenatal diagnosis
21J104606	seq[GRCh37]chr16: g.29410978-30305956 x3	0.89	2.325	1.792	Contained 37 protein-coding genes, a 3-point TS genomic region, a variable clinical presentation, incomplete penetrance.	No obvious abnormality	arr[GRCh37]chr16: g.29589674-30176508x3
21J107686	seq[GRCh37]chrX: g.6472218-8150233 x1	1.68	-6.918	–10.015	Contained 5 protein-coding genes, a 3-point HI gene and a 3-point HI genomic region, X-linked ichthyosis in males, mild or unaffected in females.	No obvious abnormality. This pregnancy was a male fetus and MoM value of uE3 was 0.03 (Standard Range, >0.7).	No prenatal diagnosis; Dry and rough skin 10 days after birth.
21J108961	seq[GRCh37]chr17: g.34710859-36306985 x1	1.60	–3.974	-1.550	Contained 18 protein-coding genes, contained a 3-point HI genomic region, renal cysts and diabetes syndrome, incomplete penetrance.	Renal cyst, congenital abnormal splenic structure and gallstones	No prenatal diagnosis
21J108971	seq[GRCh37]chr15: g.23990956-28419527 x3	4.43	5.978	–0.757	Contained 10 protein-coding genes, overlapped a 3-point HI genomic region, intellectual disability, psychiatric disorders, phenotypes by maternally-derived.	Mild schizophrenia, her brother had obvious mental retardation.	arr[GRCh37]chr15: g.23632678-28526905x3
21J104969	seq[GRCh37]chrX: g.2795214-16240667 x1	13.45	–9.986	–0.384	Contained 59 protein-coding genes, contained 9 3-point HI genes and a 3-point HI genomic region, Female carriers were unaffected or milder phenotypes.	Abortion history, this pregnancy is a female fetus.	No prenatal diagnosis
21J107005	seq[GRCh37]chr16: g.15112139-16561127 x1	1.45	–0.934	0.195	Contained 14 protein-coding genes, a 3-point HI genomic region, phenotypic variability, incomplete penetrance.	No obvious abnormality	No prenatal diagnosis
21J100568		1.29	–1.161	0.019	Contained 13 protein-coding genes, a 3-point HI genomic region,		

(Continued on following page)

TABLE 2 | (*Continued*) 17 Maternal CNVs Detected by NIPT and Fetal Diagnosis Results.

Sample ID	Maternal CNV	size (Mb)	Z_{bc}	Z_{ac}	CNV Interpretation	Maternal Phenotypes	Fetal Diagnosis
	seq[GRCh37]chr16: g.15142813-16428637 x1				phenotypic variability, incomplete penetrance.	No obvious abnormality, her son was intellectual and language disability.	arr[GRCh37]chr16: g.15481748-16458424x1
21J105304	seq[GRCh37]chr17: g.14126371-15556920 x3	1.43	2.386	0.170	Contained 10 protein-coding genes, a 3-point HI genomic region, Charcot-Marie-Tooth syndrome (CMT) characterized by slowly progressive.	No obvious abnormality	arr[GRCh37]chr17: g.14087918-15428901x3
21J104462	seq[GRCh37]chr2: g.111476219-113095275 x1	1.62	–3.545	–2.115	Contained 8 protein-coding genes, overlapped a 2-point HI genomic region, clinical findings are variable, non-specific dysmorphic features.	Lost to follow-up	No prenatal diagnosis
21J106665	seq[GRCh37]chr16: g.15395056-18200933 x1	2.81	–3.178	–1.059	Contained 12 protein-coding genes, a 3-point HI genomic region, phenotypic variability ,incomplete penetrance.	No obvious abnormality	arr[GRCh37]46,XN

Z_{bc}, Z-score before correction of the chromosome where the CNV is located; Z_{ac}, Z-score after correction; uE3, unconjugated estriol; MoM, multiple of median.

TABLE 3 | Characteristics and Results from the Cohort of 61525 Pregnancies

	Cohort	NIPT+	Diag	TP
Maternal and fetal characteristics				
Age (year)	29.94 ± 4.86	30.91 ± 5.67	30.67 ± 5.44	30.88 ± 5.46
GW (week)	18.97 ± 2.23	18.77 ± 2.32	18.83 ± 2.17	18.56 ± 2.18
ART (%)	1230 (2.00)	4 (1.00)	4 (1.32)	3 (1.40)
Twin (%)	1763 (2.87)	9 (2.24)	6 (1.98)	5 (2.34)
Number of Outlier Z-scores				
Total	–	402	303	214
Chr21	–	214	174	150
Chr18	–	68	48	32
Chr13	–	29	21	7
SCAs	–	57	40	20
OAAs	–	34	20	5

Diag, Fetuses with diagnostic testing; GW, Gestational week; ART, assisted reproductive technology; Chr, chromosome.

range. Of maternal CNVs above 2Mb, all of 6 were outliers with Z-scores before correction. It is noteworthy that Z-scores after correction were all negative. These CNVs would be missed if only concerned Z-scores.

General Analysis of NIPT-Positive Results Using Z-Scores

Z-score analysis without or maternal CNV analysis, was applied to the cohort of 61,525 pregnancies. Maternal and fetal characteristics, along with positive results, are shown in **Table 3**. We found outlier Z-scores in 402 pregnancies, and the positive rate was 6.53‰. Abnormalities of Chr21, Chr18, Chr13, SCAs, and other autosomes accounted for 61.03% (214/402), 16.92% (68/402), 7.21% (29/402), 14.18% (57/402), and 8.46% (34/402), respectively. Of the 402 pregnancies, 303 underwent prenatal or postnatal diagnoses, 44 pregnancies were miscarrige or induced labour due to structural abnormalities, and 55 pregnancies were lost to follow-up. PPV for twin pregnancies did not significantly differ from singletons, with 83.33% (5/6) of twin and 70.37% (209/297) of singleton. Overall PPV was 70.20%, of which Chr21, Chr18, sex chromosomes, Chr13, and other autosomes were lower successively with 86.21% (150/174), 66.67% (32/48), 50.00% (20/40), 33.33% (7/21), 25.00% (5/20), respectively.

We found outlier Z-score did not always indicate fetal aneuploidy. In this study, six pregnancies with outlier Z-scores were fetal CNVs verified by prenatal diagnosis, four were mosaic aneuploidies and three false positives were caused by maternal aneuploidies. An extreme outlier Z-scores for ChrY (Z = 362) were discovered in a pregnant woman with a history of bone marrow transplantation.

Accuracy Analysis of Z-Scores for Chr21, Chr18, and Chr13

The Chr21, Chr18, and Chr13 positive pregnancies were divided into six groups according to Z-scores: $Z \leq -3$, $3 \leq Z \leq 4$, $4 < Z \leq 5$, $5 < Z \leq 6$, $6 < Z \leq 40$, and $Z > 40$ (**Table 4**). At Z-scores of $3 \leq Z \leq 40$, PPV increased with increasing Z-scores. At the same Z-score, the PPV of

TABLE 4 | Comparison of NIPT Results and Diagnoses in Different Groups.

Chr	Group	N	Diagnosis result and Number		PPV (%)
21	Z ≤ -3	2	No abnormality	2	0
	3 ≤ Z ≤ 4	26	T21	8	30.77
			No abnormality on Chr21, but 2.2Mb deletion on 9p12	1	
			No abnormality on Chr21, but 7.7Mb duplication on Xp21.3-p21.1	1	
			No abnormality, including 1 twin	15	
			No abnormality with postpartum follow-up	1	
	4 < Z ≤ 5	7	T21	4	71.43
			Approximately 6Mb duplication on 21q21.1	1	
			No abnormality	2	
	5 < Z ≤ 6	11	T21	9	100.00
			Mosaic T21 and mosaic X0	1	
			T21 and 1.4Mb deletions on 17p12	1	
	6 < Z ≤ 40	126	T21	117	100.00
			T21 on one of twin	4	
			T21 with rob (21;21)	1	
			Mosaic T21	1	
			Approximately 10Mb duplication on 21q11.2-q21.2	1	
			T21 in postpartum follow-up	2	
	Z > 30	2	No abnormality, but T21 on mother	2	0
18	Z ≤ -3	2	6.0Mb deletion on 18p11.32-p11.31	1	50.00
			No abnormality	1	
	3 ≤ Z ≤ 4	6	T18	1	16.67
			No abnormality	5	
	4 < Z ≤ 5	10	T18	6	60.00
			No abnormality	4	
	5 < Z ≤ 6	4	T18	3	75.00
			No abnormality on Chr18, but 33.3Mb duplication on Chr11	1	
	6 < Z ≤ 40	26	T18	19	80.77
			Mosaic T18	1	
			3.1Mb duplication on 18p11.32-p11.31	1	
			No abnormality	5	
13	Z ≤ -3	2	10.6Mb deletion on 13q21.1-q21.32	1	50.00
			No abnormality on Chr13, but 0.9Mb duplication on Chr16	1	
	3 ≤ Z ≤ 4	9	1.5Mb duplication on 13q12.12	1	11.11
			No abnormality	7	
			No abnormality in postpartum follow-up	1	
	4 < Z ≤ 5	3	T13	1	33.33
			No abnormality	2	
	5 < Z ≤ 6	2	T13	1	50.00
			No abnormality	1	
	6 < Z ≤ 40	5	T13	3	60.00
			No abnormality on Chr13, but 1.4Mb duplication on 7p21.3	1	
			No abnormality	1	

Chr, chromosome; N, number; PPV, positive predictive value.

TABLE 5 | Correlation between Z-score and Positive Predictive Value (PPV) by Logistic Regression Analysis.

Group	B	Wald	OR	95% CI	*P*
T13	0.465	1.742	1.592	1.001–3.551	0.1869
T18	0.427	6.996	1.532	1.174–2.229	0.00817
T21	1.322	18.73	3.752	2.282–7.811	<0.001

B, beta coefficients; Wald, wald test; OR, odds ratio; CI, confidence interval; P, p-value; T, trisomy.

Chr21 was always the highest and that of Chr13, was lowest. Notably, two pregnancies with Z > 40 were both maternal Down syndrome. With Z-scores < -3, fetuses carried microdeletions instead of monosomy. As shown in **Table 5**, logistic regression analysis revealed that Z-scores were significantly associated with the PPV of T21 (OR = 3.752, $p < 0.001$) and T18 (OR = 1.532, $p = 0.00817$). In the ROC curves, AUCs of T21, T18, and T13 were 0.9624, 0.8043, and 0.7436, respectively. S-curves were simulated to predict the trend of PPV shown in **Figure 1**. We also revealed several special types of karyotype by diagnosis, including Robertsonian translocation, mosaic trisomy, and CNVs; these cases exhibited outlier Z-scores in NIPT.

Accuracy Analysis of Z-Scores for Sex Chromosomes and Other Autosomes

In 40 pregnancies with abnormal Z-scores for sex chromosomes, the PPV was 50.00% (20/40) shown in **Table 6**. Among them, the PPV of sex chromosome trisomy was 64.00% (16/25), and that of monosomy was 26.67% (4/15), with a significant difference ($p < 0.05$). Five out of 20 for other autosomal abnormalities were true

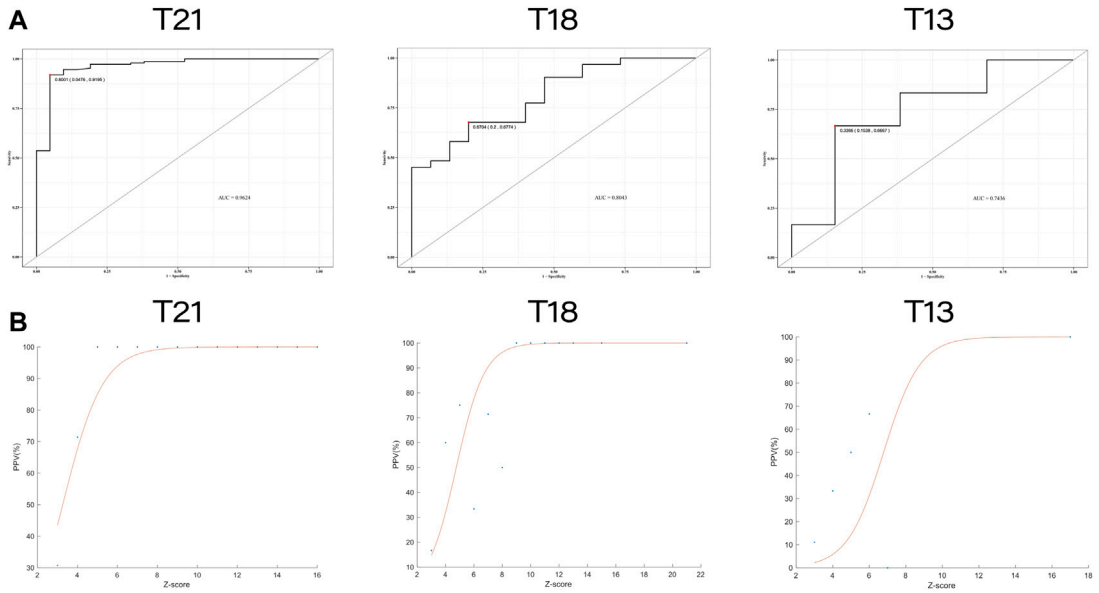

FIGURE 1 | The receiver operating characteristic (ROC) curves and trend curves between Z-score and the positive predictive value. **(A)** The ROC curves of T21, T18, and T13. The area under the curve (AUC) were 0.9624, 0.8043 and 0.7436, respectively. **(B)** The trend curves of T21, T18 and T13 with S-curve model ($\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{1+\boldsymbol{a}_0\boldsymbol{e}^{-\boldsymbol{x}}}$).

TABLE 6 | Comparison of NIPT Results and Diagnoses in Sex Chromosomes and Other Autosomes.

Chr	Type	N	Diagnostic result and Number		PPV (%)
SCAs	X0	15	X0	1	26.67%
			X0/XN	2	
			X0/X,r(X)	1	
			No abnormality	11	
	XXX	16	XXX	9	56.25
			No abnormality, but XXX is found on mother	1	
			No abnormality	6	
	XYY	1	XYY	1	100.00
	XXY	8	XXY, including 1 twin-sample	6	75.00
			No abnormality	2	
	Total	**40**		**20**	**50.00**
OAAs	T2	1	No abnormality	1	
	T3	1	No abnormality	1	
	T7	5	No abnormality	5	
	T8	3	No abnormality	3	
	M9	1	20Mb deletion on 9p24.3p21.3 and 37Mb deletion on 9q21.13q31.3	1	
	T9	1	38.6Mb duplication on 9p24.3p13.1	1	
	T10	1	0.3Mb duplication on 10q23.33	1	
	M14	1	No abnormality	1	
	T15	1	No abnormality	1	
	T16	2	T16	1	
			Mosaic T16	1	
	M16&19	1	No abnormality	1	
	T20	1	No abnormality	1	
	M22	1	No abnormality	1	
	Total	**20**		**5**	**25.00**

M, monosomy.

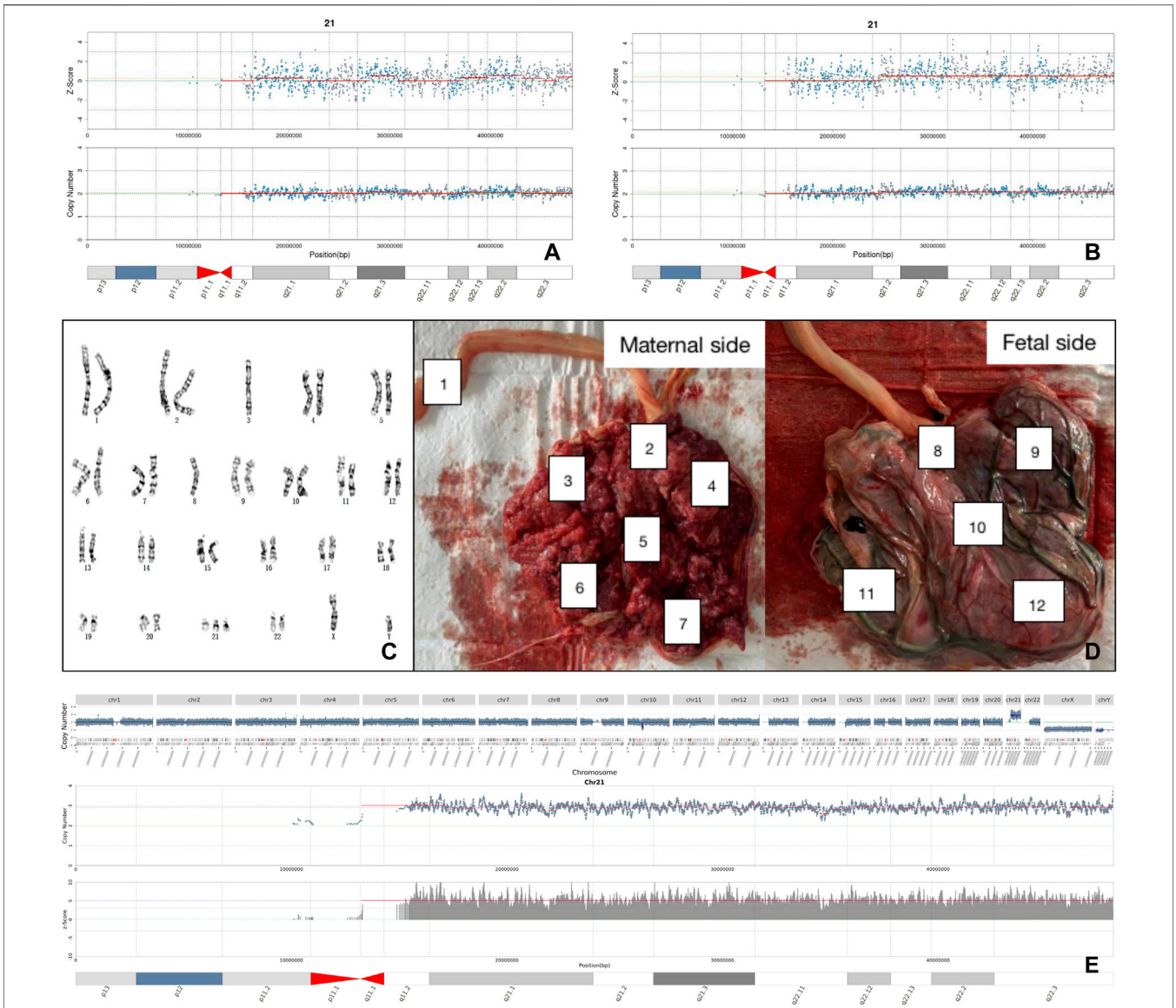

FIGURE 2 | Analysis of a false negative case. **(A)** The negative result of NIPT with first blood sampling. **(B)** The positive result of NIPT with re-sampling. **(C)** The karyotype of fetal amniotic fluid cell. **(D)** Sampling locations of placental tissue after labor induction. **(E)** Placental CNV-seq results suggested full T21.

positive (PPV = 25%) diagnosed by SNP-array or CNV-seq technology, including aneuploidies (T16 and mosaic T16) and CNVs on Chr9 and Chr10.

DISCUSSION

Because PPV indicates the possibility of true positive, it is usually used to evaluate the predictive ability of the test (Meck, et al., 2015; DiNonno, et al., 2019). PPV affected by factors such as population size and region, given that it is related to a population's basic prevalence (Monaghan, et al., 2021).

Previous studies using the ion proton semiconductor platform and the BGISEQ-500 sequencing platform suggested that Z ≥ 9/10 had a higher PPV(Tian et al., 2018; Wan et al., 2021). However, there are potential differences in low Z-scores between different sequencing platforms. In this study, we performed Z-score grouping, logistic regression analysis, ROC curves and S-curve trends to determine correlations between Z-score and PPV. There was the significant correlations between Z-scores and PPV at T21 and T18, with the exception of T13. In addition, we found that the true positives in Z < −3 were all microdeletions instead of monosomy, which was not mentioned before. Because of the diversity in sex chromosomal and other autosomal abnormalities, more data are necessary to increase the accuracy of observed trends.

Several factors affect the accuracy of NIPT. Confined placental mosaicism (CPM), with an incidence of 1–2%, is a

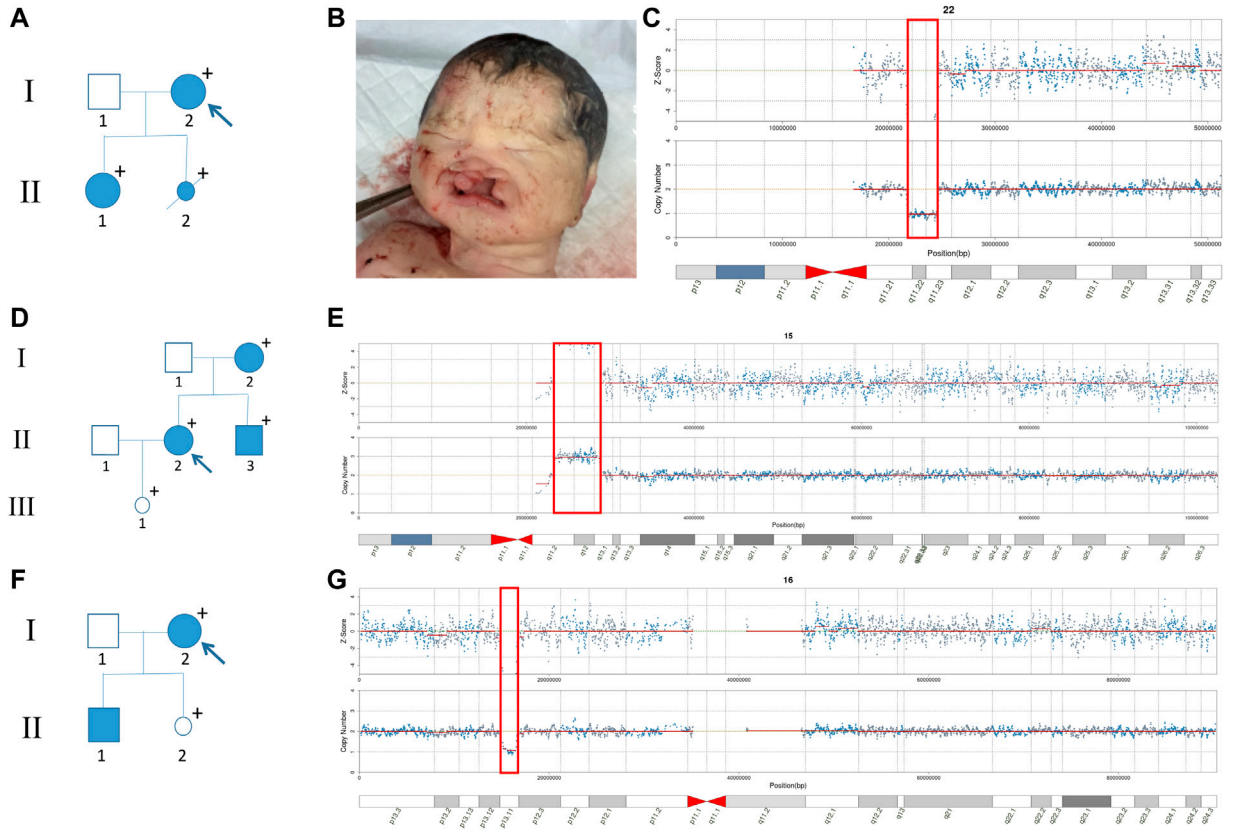

FIGURE 3 | Family pedigrees with maternal CNVs. **(A)** Family pedigrees of No. 21J101249. **(B)** The facial phenotype of post-induction fetus (No. 21J101249), including cleft lip and palate. **(C)** Maternal CNV of No. 21J101249 suggested by NIPT. **(D)** Family pedigrees of No. 21J108971. **(E)** Maternal CNV of No. 21J108971 suggested by NIPT. **(F)** Family pedigrees of No. 21J100568. **(G)** Maternal CNV of No. 21J100568 suggested by NIPT.

common reason (Malvestiti et al., 2015; Mardy et al., 2016). Another reason is maternal chromosomal abnormalities (Wang et al., 2014; Zhou et al., 2017). A study found that maternal CNVs could increase the false positive rate of NIPT by 10% (Snyder et al., 2015). Our study found three pregnancies with Z-scores > 40, were all false positive caused by maternal full aneuploidy. A common cause of false negatives is the low fetal frequency, where the cff-DNA increases as gestational weeks increase. Following previous research, we recollected blood samples in cases of inadequate fetal frequency (Wang et al., 2013). A false-negative case was found in our study due to low fetal frequency (FF=3.8%, Z < 3). Unexpectedly, the fetal frequency reached 9.6% re-collected after two weeks and Z-score of chromosome 21 was 6.66 which indicated that the fetus was T21 positive. The fetus was verified by prenatal diagnosis as 47,XN,+21. For this case, we also used CNV-seq on placental tissue to yield a results of 47,XY,+21, thus excluding CPM shown in **Figure 2**. Another factor that can interfere with NIPT results is history of transplantation (Zhu et al., 2021). In our study, a pregnant woman received a bone marrow transplant from a male donor 7 years ago. The Z-score of ChrY was 362, far exceeding Z-score from typical male fetuses (Z-score = 3–150). Prenatal diagnosis indicated a nomal fetal karyotype. We further explored cell sources of this pregnant woman's peripheral blood, oral cavity, and hair follicle cells using STR markers in sex chromosomes, and found different proportions of cell sources. The hair follicle cells were all from the pregnant woman herself, the peripheral blood lymphocytes were all from the bone marrow donor, and the oral cells were the co-existence of the two sources.

Maternal CNVs were seldom researched in NIPT or NIPT-plus technology. NIPT-plus can identify fetal de novo MMS with 0.174% positive rate (Liang et al., 2019). Considering the incidence of pathological CNVs (1.7%) in all pregnancies with normal fetal structures (Wapner et al., 2012), the inherited CNVs were underestimated. By maternal CNV analysis in this study, we found that maternally-inherited CNVs with clinically significant reached 0.28% (9/3184) comprising 7 fetuses with prenatal diagnosis and 2 with specific phenotypes after birth. It was valuable of increasing detection rate of NIPT without increasing testing costs.

Although some pregnant women did not show obvious symptoms due to incomplete penetrance of CNVs, further genetic counseling may reveal mild phenotypes or severe family histories. In our study, the pregnant woman of No. 21J101249 (**Figure 3A, I-2**) had slight facial asymmetry and

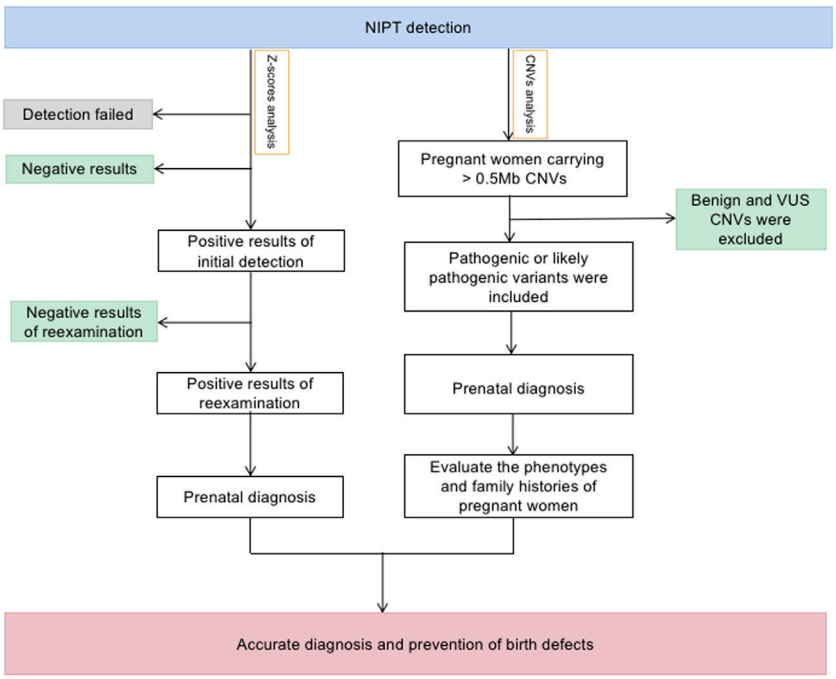

FIGURE 4 | The flow of NIPT combining Z-score and maternal CNVs analysis.

communication impairment. Her daughter (II-1) suffered from congenital incomplete cleft palate, as well as mental and physical retardation with a low developmental quotient (DQ = 62). The fetus was induced due to severe cleft lip and palate (**Figure 3B**). Among I-2, II-1, and II-2, we found an approximately 2.9 Mb heterozygous deletion in 22q11.21-q11.23 (**Figure 3C**). SMARCB1 gene and 22q11.2 recurrent region (distal type I,D-E/F) in the fragment have sufficient evidence for haploinsufficiency, which are associated with clinical phenotypes such as global developmental delay, intellectual disability, cleft lip, and behavioral problems (Mikhail et al., 2014; Holsten et al., 2018). The pregnant woman of No. 21J108971 (**Figure 3D, II-2**) had mild schizophrenia; her brother (II-3) exhibited obvious mental retardation, but her mother (I-2) had no obvious abnormality. SNP-array technology revealed that I-2, II-2, and II-3 harbored an approximately 4.4 Mb heterozygous duplication in 15q11.2-q13.1 (**Figure 3E**), and analysis of fetal amniotic fluid cells (III-1) indicated that the fetus carried the same CNV 15q11.2q13 recurrent (PWS/AS) region (Class 1, BP1-BP3) and 15q11.2q13 recurrent (PWS/AS) region (Class 2,BP2-BP3) in the fragment have sufficient evidence for triplosensitivity, associated with autism, intellectual disability, seizures, and psychiatric disorders (Christian et al., 2008; Ingason et al., 2011). Evidence suggests a parent-of-origin effect, with maternally-derived duplications being more frequently associated with abnormal phenotypes. While the pregnant woman of No. 21J100568 (**Figure 3H, I-2**) had no obvious abnormality, prenatal diagnosis indicated that the fetus (II-2) carried the same CNV on 16p13.11 (**Figure 3G**). Additionally her first son suffered from retarded intellectual and language development that had not been detected (II-1) 16p13.11 recurrent region (BP2-BP3) (includes MYH11) in the fragment has sufficient evidence for haploinsufficiency, associated with intellectual disability and/or multiple congenital anomalies (de Kovel et al., 2010; Jähn et al., 2013). It showed sex-limited effect on the penetrance with a significant enrichment among male cases (Tropeano et al., 2013). The fetus had been born for one month without abnormality. The pregnant woman of No. 21J108961 with microdeletion on 17q12 had a renal cyst, congenital abnormal splenic structure, and gallstones but continued her pregnancy without a prenatal diagnosis. The pregnant woman of No. 21J101817 with microdeletion on 4q34.3-q35.2 are less than 150 cm tall despite no obvious developmental delay. The pregnant women of No. 21J101676 and No. 21J104969 had deletions on ChrX (CNVs >10 Mb) which can cause male lethality but no effect on female carriers. The pregnant woman of No.21J105324 and No.21J107686 had microdeletions on Xp22.31 and were pregnant with male fetuses. Inheritance of these CNVs by the male fetuses could result in X-linked ichthyosis caused by haploinsufficiency of STS gene. Additionally, we found that their uE3 was far below normal. studies have shown that a characteristic of ichthyosis is significantly reduced uE3 during the embryonic period (Kashork et al., 2002).

The screening act as an early warning for parents regarding pathogenic CNVs that may be passed down to their offspring. Despite the small number of samples, our combined analysis

increased NIPT positive rates from, 6.91‰ to 12.25‰ and found 9 clinically significant fetal CNVs which inherited from mothers without increasing detection costs and placing more economic pressure on pregnant women. Therefore, we propose a new NIPT screening model that integrates Z-scores and maternal CNVs (**Figure 4**).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI SRA BioProject, accession No:PRJNA837410.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Changzhi Maternal and Child Health Care Hospital. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

XL, YA, and LC conceived the study, carried out the assays and participated in the study design. LC, LW, ZH, YT, and WS carried out the laboratory tests, statistical analysis, and followed-up.All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank all the participants for their contribution to this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.887176/full#supplementary-material

REFERENCES

Chen, Y., Yu, Q., Mao, X., Lei, W., He, M., and Lu, W. (2019). Noninvasive Prenatal Testing for Chromosome Aneuploidies and Subchromosomal Microdeletions/microduplications in a Cohort of 42,910 Single Pregnancies with Different Clinical Features. *Hum. Genomics.* 13, 60. doi:10.1186/s40246-019-0250-2

Chiu, R. W. K., Chan, K. C. A., Gao, Y., Lau, V. Y. M., Zheng, W., Leung, T. Y., et al. (2008). Noninvasive Prenatal Diagnosis of Fetal Chromosomal Aneuploidy by Massively Parallel Genomic Sequencing of DNA in Maternal Plasma. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20458–20463. doi:10.1073/pnas.0810641105

Christian, S. L., Brune, C. W., Sudi, J., Kumar, R. A., Liu, S., Karamohamed, S., et al. (2008). Novel Submicroscopic Chromosomal Abnormalities Detected in Autism Spectrum Disorder. *Biol. Psychiatry.* 63, 1111–1117. doi:10.1016/j.biopsych.2008.01.009

de Kovel, C. G., Trucks, H., Helbig, I., Mefford, H. C., Baker, C., Leu, C., et al. (2010). Recurrent Microdeletions at 15q11.2 and 16p13.11 Predispose to Idiopathic Generalized Epilepsies. *Brain* 133, 23–32. doi:10.1093/brain/awp262

DiNonno, W., Demko, Z., Martin, K., Billings, P., Egbert, M., Zneimer, S., et al. (2019). Quality Assurance of Non-invasive Prenatal Screening (NIPS) for Fetal Aneuploidy Using Positive Predictive Values as Outcome Measures. *Jcm* 8, 1311. doi:10.3390/jcm8091311

Gil, M. M., Accurti, V., Santacruz, B., Plana, M. N., and Nicolaides, K. H. (2017). Analysis of Cell-free DNA in Maternal Blood in Screening for Aneuploidies: Updated Meta-Analysis. *Ultrasound. Obstet. Gynecol.* 50, 302–314. doi:10.1002/uog.17484

Gil, M. M., Galeva, S., Jani, J., Konstantinidou, L., Akolekar, R., Plana, M. N., et al. (2019). Screening for Trisomies by cfDNA Testing of Maternal Blood in Twin Pregnancy: Update of the Fetal Medicine Foundation Results and Meta-analysis. *Ultrasound. Obstet. Gynecol.* 53, 734–742. doi:10.1002/uog.20284

Grati, F. R., Molina Gomes, D., Ferreira, J. C. P. B., Dupont, C., Alesi, V., Gouas, L., et al. (2015). Prevalence of Recurrent Pathogenic Microdeletions and Microduplications in over 9500 Pregnancies. *Prenat. Diagn.* 35, 801–809. doi:10.1002/pd.4613

Holsten, T., Bens, S., Oyen, F., Nemes, K., Hasselblatt, M., Kordes, U., et al. (2018). Germline Variants in SMARCB1 and Other Members of the BAF Chromatin-Remodeling Complex Across Human Disease Entities: A Meta-Analysis. *Eur. J. Hum. Genet.* 26, 1083–1093. doi:10.1038/s41431-018-0143-1

Hu, H., Wang, L., Wu, J., Zhou, P., Fu, J., Sun, J., et al. (2019). Noninvasive Prenatal Testing for Chromosome Aneuploidies and Subchromosomal Microdeletions/microduplications in a Cohort of 8141 Single Pregnancies. *Hum. Genomics.* 13, 14. doi:10.1186/s40246-019-0198-2

Ingason, A., Kirov, G., Giegling, I., Hansen, T., Isles, A. R., Jakobsen, K. D., et al. (2011). Maternally Derived Microduplications at 15q11-Q13: Implication Of Imprinted Genes in Psychotic Illness. *Am. J. Psychiatry* 168, 408–417. doi:10.1176/appi.ajp.2010.09111660

Iwarsson, E., Jacobsson, B., Dagerhamn, J., Davidson, T., Bernabé, E., and Heibert Arnlind, M. (2017). Analysis of Cell-free Fetal DNA in Maternal Blood for Detection of Trisomy 21, 18 and 13 in a General Pregnant Population and in a High Risk Population - a Systematic Review and Meta-Analysis. *Acta. Obstet. Gynecol. Scand.* 96, 7–18. doi:10.1111/aogs.13047

Jähn, J. A., von Spiczak, S., Muhle, H., Obermeier, T., Franke, A., Mefford, H. C., et al. (2014). Iterative Phenotyping of 15q11.2, 15q13.3 and 16p13.11 Microdeletion Carriers in Pediatric Epilepsies. *Epilepsy. Res.* 108, 109–116. doi:10.1016/j.eplepsyres.2013.10.001

Junhui, W., Ru, L., Qiuxia, Y., Dan, W., Xiuhong, S., Yongling, Z., et al. (2021). Evaluation of the Z-score Accuracy of Noninvasive Prenatal Testing for Fetal Trisomies 13, 18 and 21 at a Single Center. *Prenat. Diagn.* 41, 690–696. doi:10.1002/pd.5908

Kashork, C. D., Sutton, V. R., Fonda Allen, J. S., Schmidt, D. E., Likhite, M. L., Potocki, L., et al. (2002). Low or Absent Unconjugated Estriol in Pregnancy: an Indicator for Steroid Sulfatase Deficiency Detectable by Fluorescence *In Situ* Hybridization and Biochemical Analysis. *Prenat. Diagn.* 22, 1028–1032. doi:10.1002/pd.466

Levy, B., and Wapner, R. (2018). Prenatal Diagnosis by Chromosomal Microarray Analysis. *Fertil. Steril.* 109, 201–212. doi:10.1016/j.fertnstert.2018.01.005

Li, X., Wang, L., Yao, Z., Ruan, F., Hu, Z., and Song, W. Clinical Evaluation of Non-invasive Prenatal Screening in 32,394 Pregnancies from Changzhi Maternal and Child Health Care Hospital of Shanxi China. *J. Med. Biochem.* (in press), 41, 1-6. doi:10.5937/jomb0-33513

Liang, D., Cram, D. S., Tan, H., Linpeng, S., Liu, Y., Sun, H., et al. (2019). Clinical Utility of Noninvasive Prenatal Screening for Expanded Chromosome Disease Syndromes. *Genet. Med.* 21, 1998–2006. doi:10.1038/s41436-019-0467-4

Lo, Y. M. D., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., et al. (1997). Presence of Fetal DNA in Maternal Plasma and Serum. *Lancet* 350, 485–487. doi:10.1016/s0140-6736(97)02174-0

Ma, N., Xi, H., Chen, J., Peng, Y., Jia, Z., Yang, S., et al. (2021). Integrated CNV-Seq, Karyotyping and SNP-Array Analyses for Effective Prenatal Diagnosis of Chromosomal Mosaicism. *BMC Med. Genomics* 14, 56. doi:10.1186/s12920-021-00899-x

Malvestiti, F., Agrati, C., Grimi, B., Pompilii, E., Izzi, C., Martinoni, L., et al. (2015). Interpreting Mosaicism in Chorionic Villi: Results of a Monocentric Series of 1001 Mosaics in Chorionic Villi with Follow-Up Amniocentesis. *Prenat. Diagn.* 35, 1117–1127. doi:10.1002/pd.4656

Mardy, A., and Wapner, R. J. (2016). Confined Placental Mosaicism and its Impact on Confirmation of NIPT Results. *Am. J. Med. Genet.* 172, 118–122. doi:10.1002/ajmg.c.31505

Meck, J. M., Kramer Dugan, E., Matyakhina, L., Aviram, A., Trunca, C., Pineda-Alvarez, D., et al. (2015). Noninvasive Prenatal Screening for Aneuploidy: Positive Predictive Values Based on Cytogenetic Findings. *Am. J. Obstetrics Gynecol.* 213, e1–214. doi:10.1016/j.ajog.2015.04.001

Mikhail, F. M., Burnside, R. D., Rush, B., Ibrahim, J., Godshalk, R., Rutledge, S. L., et al. (2014). The Recurrent Distal 22q11.2 Microdeletions are Often De Novo and do not Represent a Single Clinical Entity: A Proposed Categorization System. *Genet. Med.* 16, 92–100. doi:10.1038/gim.2013.79

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., et al. (2021). Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina* 57, 503. doi:10.3390/medicina57050503

Nevado, J., Mergener, R., Palomares-Bralo, M., Souza, K. R., Vallespín, E., Mena, R., et al. (2014). New Microdeletion and Microduplication Syndromes: A Comprehensive Review. *Genet. Mol. Biol.* 37, 210–219. doi:10.1590/s1415-47572014000200007

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., et al. (2020). Technical Standards for the Interpretation and Reporting of Constitutional Copy-Number Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257. doi:10.1038/s41436-019-0686-8

Snyder, M. W., Simmons, L. E., Kitzman, J. O., Coe, B. P., Henson, J. M., Daza, R. M., et al. (2015). Copy-number Variation and False Positive Prenatal Aneuploidy Screening Results. *N. Engl. J. Med.* 372, 1639–1645. doi:10.1056/NEJMoa1408408

Tian, Y., Zhang, L., Tian, W., Gao, J., Jia, L., and Cui, S. (2018). Analysis of the Accuracy of Z-Scores of Non-invasive Prenatal Testing for Fetal Trisomies 13, 18, and 21 that Employs the Ion Proton Semiconductor Sequencing Platform. *Mol. Cytogenet.* 11, 49. doi:10.1186/s13039-018-0397-x

Tropeano, M., Ahn, J. W., Dobson, R. J., Breen, G., Rucker, J., Dixit, A., et al. (2013). Male-Biased Autosomal Effect of 16p13.11 Copy Number Variation in Neurodevelopmental Disorders. *PLoS One* 8, e61365. doi:10.1371/journal.pone.0061365

Wang, E., Batey, A., Struble, C., Musci, T., Song, K., and Oliphant, A. (2013). Gestational Age and Maternal Weight Effects on Fetal Cell-free DNA in Maternal Plasma. *Prenat. Diagn.* 33, 662–666. doi:10.1002/pd.4119

Wang, Y., Chen, Y., Tian, F., Zhang, J., Song, Z., Wu, Y., et al. (2014). Maternal Mosaicism Is a Significant Contributor to Discordant Sex Chromosomal Aneuploidies Associated with Noninvasive Prenatal Testing. *Clin. Chem.* 60, 251–259. doi:10.1373/clinchem.2013.215145

Wapner, R. J., Martin, C. L., Levy, B., Ballif, B. C., Eng, C. M., Zachary, J. M., et al. (2012). Chromosomal Microarray versus Karyotyping for Prenatal Diagnosis. *N. Engl. J. Med.* 367, 2175–2184. doi:10.1056/NEJMoa1203382

Weise, A., Mrasek, K., Klein, E., Mulatinho, M., Llerena, J. C., Hardekopf, D., et al. (2012). Microdeletion and Microduplication Syndromes. *J. Histochem Cytochem.* 60, 346–358. doi:10.1369/0022155412440001

Zhang, H., Gao, Y., Jiang, F., Fu, M., Yuan, Y., Guo, Y., et al. (2015). Non-invasive Prenatal Testing for Trisomies 21, 18 and 13: Clinical Experience from 146 958 Pregnancies. *Ultrasound. Obstet. Gynecol.* 45, 530–538. doi:10.1002/uog.14792

Zhou, L., Zhang, B., Liu, J., Shi, Y., Wang, J., and Yu, B. (2021). The Optimal Cutoff Value of Z-Scores Enhances the Judgment Accuracy of Noninvasive Prenatal Screening. *Front. Genet.* 12, 690063. doi:10.3389/fgene.2021.690063

Zhou, X., Sui, L., Xu, Y., Song, Y., Qi, Q., Zhang, J., et al. (2017). Contribution of Maternal Copy Number Variations to False-Positive Fetal Trisomies Detected by Noninvasive Prenatal Testing. *Prenat. Diagn.* 37, 318–322. doi:10.1002/pd.5014

Zhu, J., Hui, F., Mao, X., Zhang, S., Qi, H., and Du, Y. (2021). cfDNA Deconvolution via NIPT of a Pregnant Woman after Bone Marrow Transplant and Donor Egg IVF. *Hum. Genomics.* 15, 14. doi:10.1186/s40246-021-00311-w

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership