HHS Public Access

Author manuscript

*Am J Med Genet A*. Author manuscript; available in PMC 2026 June 22.

# *De Novo* Complex Genomic Rearrangement Spanning 2q31.1 in a Proband With Congenital Malformations: Genotype–Phenotype Correlation and Development of a CGR Detection Pipeline

**Katherine Helle**[1], **Jesse D. Bengtsson**[1], **Mira Gandhi**[1], **Christopher M. Grochowski**[2,3], **Ming Yin Lun**[1], **Neha Sudhir**[1], **Shalini N. Jhangiani**[3], **Fritz J. Sedlazeck**[2,3], **Seema R. Lalani**[2], **Neil A. Hanchard**[2], **Claudia M. B. Carvalho**[1,2]

[1]Pacific Northwest Research Institute, Seattle, Washington, USA

[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

[3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA

## Abstract

The 2q31 region is commonly associated with pathogenic alleles of the HOXD cluster leading to various clinical phenotypes related to skeletal development. We present a proband with tetralogy of Fallot and multiple congenital anomalies. Genomic variant screening including an in-house CGR detection pipeline pairing genome sequencing (GS) structural variant calls with read-depth data revealed a *de novo* complex genomic rearrangement (CGR) spanning 2.7 Mb across 2q31 characterized by a series of duplications and triplications including the HOXD gene cluster. The genomic structure was assembled by applying combined methodologies including short-read and long-read GS, and optical genome mapping (OGM). This in-house CGR pipeline detected five additional rearrangements throughout the genome confirmed by orthogonal methodologies as inherited and unlikely to impact this individual's phenotype. Importantly, these catastrophic genomic events, chromoanasynthesis-like, are surprisingly commonly observed in the genome, inherited and involve large regions of the genome. Moreover, such inherited CGRs often include copy-number gains that partially affect disease-causing genes which complicate clinical interpretation. Overall, we show the utility of short-read sequencing to uncover *de novo* and



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

Author Contributions

K.H. analyzed and interpreted data and wrote the manuscript. J.D.B. contributed to analysis and interpretation of data and contributed to writing the manuscript. M.G. performed chromosomal microarray analysis. M.Y.L. developed the CGR pipeline and processed the genomic data. N.S. contributed to analysis and interpretation of SNVs from short-read data. S.N.J. and F.J.S. coordinated genome sequencing experiments. C.M.G. assisted with optical mapping analysis. S.R.L. and N.A.H. provided patient samples, clinical information, and contributed to writing the manuscript. C.M.B.C. conceptualized the experiments, overviewed data analysis and interpretation, and contributed to writing the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethics Statement

This study is approved by the WIRB for the Pacific Northwest Research Institute (IRB Protocol #H-47127/20202158). The research was conducted in compliance with the principles of the Helsinki Declaration.

inherited chromoanasynthesis events and established genotype–phenotype correlation in a proband with multiple congenital malformations.

**Keywords**

chromothripsis; copy-number variants; dosage effects; genomic disorders; long-read sequencing; PacBio HiFi; rare genetic diseases; structural variants

## 1 | Introduction

In relevance to human diseases, structural variants (SVs) include copy-number variations (CNVs), such as deletions, duplications, triplications, added amplifications, or other large-scale variants, and copy number-neutral inversions, insertions, and translocations (Schuy et al. 2022). Complex genomic rearrangements (CGRs) are a class of SVs that contain more than one breakpoint junction, leading to a structure with more than one SV *in cis* (Carvalho and Lupski 2016; Schuy et al. 2022). CGRs are rare causes of genomic disorders, a group of diseases caused by rearrangements of the human genome due to genomic instability (Lupski 1998), but they can be responsible for up to 40% of pathogenic SVs in genetic disorder cohorts (Schuy et al. 2022) and act as modifiers in neurodevelopmental diseases (Sanders et al. 2020). In fact, CGRs modify phenotypic severity due to increased gene dosage (Carvalho et al. 2011; Regier et al. 2018) or can lead to pleiotropy of psychiatric phenotypes (Grochowski et al. 2018), with a possibility of rescue and treatment when the underlying biology is known (Farek et al. 2023). Moreover, CGRs may change gene splicing, lead to fusion gene formation and exonization of intronic mobile elements (Zuccherato et al. 2016), or alter the expected patterns of imprinting with pathogenic consequences for the carriers (Carvalho et al. 2019).

Due to the presence of multiple structural variants in one event, CGRs may reshape the overall genomic structure of a given locus, which increases the possibility of dysregulation of genes *in cis* or *in trans* (Grochowski et al. 2021). Therefore, detecting and resolving rare CGRs may be clinically important. Before the widespread application of next-generation sequencing (NGS), CGR breakpoints were identified by fluorescence in situ hybridization (FISH), multi-color banding, karyotyping, or chromosomal microarray (CMA). Such events were considered rare, with a 0.0026% prevalence in 269,371 prenatal samples (Giardino et al. 2009). However, recent developments in genome sequencing (GS) in both congenital disorders and somatic events in cancer genomes (Cosenza et al. 2022; Schuy et al. 2022) have increased our ability to detect and map complex genomic events, revealing a growing prevalence of CGRs that may be clinically relevant (Pagnamenta et al. 2024). These updates include application of long molecules and combination of next-generation sequencing approaches to map breakpoints in one molecule of DNA, particularly those mapping to complex regions, such as telomeres, centromeres, repeat elements, and low-copy repeats (LCRs) or segmental duplications (SDs) (Beck et al. 2019; Grochowski et al. 2024). Narrowing down junctions to the nucleotide resolution is necessary to assemble the derivative chromosomal structures that will inform hypothesis-driven experiments to investigate functional consequences of CGRs (Grochowski et al. 2021; Schuy et al. 2024).

Both germline and somatic CGRs can be grouped as having intermediate or extreme complexity, which loosely correlates with the number of breakpoints and the size of the genomic segments involved (Ly and Cleveland 2017; Schuy et al. 2022). Intermediate CGRs carry < 5 breakpoints and mostly involve a single chromosome, whereas extremely complex CGRs carry > 5 to hundreds of breakpoints and sometimes involve multiple chromosomes. Extreme CGRs are classified under the umbrella of chromoanagenesis (chromosome re-birth), which includes chromothripsis, chromoplexy and chromoanasynthesis (Nazaryan-Petersen et al. 2018). In constitutional diseases, chromoanagenesis has been sporadically reported as a driver of congenital phenotypes (Eisfeldt et al. 2019; Grochowski et al. 2021), mostly constituting a one-off family with an SV affecting a dosage sensitive gene (Carvalho et al. 2015; Liu et al. 2017), therefore unlikely to be a prevalent mechanism in genetic disorders. In contrast, intermediate CGRs are frequently found in genomic disorders and constitute drivers of genetic diseases, particularly those caused by copy-number gains reviewed in (Carvalho and Lupski 2016).

Herein we present a patient with global developmental delay, tetralogy of Fallot, and multiple other congenital anomalies carrying a *de novo* 2.7 Mb CGR involving 2q31.1. To analyze the genomic architecture of the impacted chromosome 2 and explore whether genes involved in this rearrangement are contributing to the observed clinical phenotype, we used several genomic technologies including chromosomal microarray, Illumina short-read genome sequencing, PacBio long-read genome sequencing, and Bionano optical genome mapping. This combination of tools allowed us to resolve the breakpoint junctions at the base-pair level resolution, revealing the architecture of the arrangement and confirming dosage alteration of several genes within 2q31 which may impact the patient's clinical traits. Additionally, to investigate whether this proband carries additional CGRs, either of intermediate or higher number of breakpoint junctions, we developed a computational pipeline based on short-read sequencing that leverages distinct information provided by read-depth and split-read sequencing analysis. This pipeline utilizes the molecular features of chromoanasynthesis, including formation of copy-number variants (CNVs) together with other types of SVs *in cis*, to establish the number of CGR events this proband carries. We then combined NGS technologies to assemble derivative chromosomes and defined genetic segregation. In all, we show the utility of short-read sequencing to uncover *de novo* and inherited chromoanasynthesis and established genotype–phenotype correlation in a proband with multiple congenital malformations.

## 2 | Materials and Methods

### 2.1 | Editorial Policies and Ethical Considerations

Patient SEA101 was found to carry CGRs by chromosome microarray analysis (CMA), completed as part of the proband's clinical evaluation through Baylor Genetics. Informed consent was obtained for research participation of the proband, mother (SEA102), and father (SEA103) under protocol H-47127, approved by the institutional review board of Baylor College of Medicine and WCG (IRB Protocol #20202158).

### 2.2 | Chromosomal Microarray Analysis (CMA)

Initial CMA analysis was performed using a custom-designed 400 K genome-wide Agilent CMA-HR + SNP Screen array with 60,000 single nucleotide polymorphism (SNP) probes at Baylor Genetics (Houston, TX, United States). Following this, a custom 180 K high-resolution microarray (AMADID#: 086747) was designed using the Agilent Sure Design Website v.6.9.11 (https://earray.chem.agilent.com/suredesign/) on NCBI Build 37. We selected 153,450 probes to explore chr2: 93866425–243199372, with a median probe spacing of 804 bp. The array was run according to the Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol (Agilent Technologies) with some modifications (Beck et al. 2019).

### 2.3 | Short-Read Whole-Genome Sequencing

The Human Genome sequencing center (HGSC) at Baylor College of Medicine performed whole-genome sequencing using genomic DNA isolated from the frozen blood of the trio (proband, mother, and father). The BCM-HGSC protocol was followed, using 750 ng DNA with KAPA Hyper PCR-free reagents to prepare short-read libraries (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). Dual-indexed libraries were multiplexed at 81 libraries per pool and sequenced across three S4 flow cells to target 30× minimum sequence per sample. This was achieved using a two-step process: in Step 1, 81 libraries were sequenced on a flow cell to assess pooling uniformity (the calibration pool), which was adjusted in step 2 if re-pooling was necessary. To generate 150 bp, dual-indexed and paired-end sequence reads for all samples in a format of multiplexed pools, the HGSC Illumina NovaSeq 6000 instrument fleet was utilized, producing an average of 35.5× coverage. The average sequencing yield was 115.2 GB, meaning the samples achieved 97.4% of the bases covered to a depth of 20× or greater. To process the raw sequence data, the HGSC HgV analysis pipeline (Farek et al. 2023; Regier et al. 2018; Reid et al. 2014) (https://www.hgsc.bcm.edu/software/mercury) which applies to BWA-MEM software was used. This executed base calling, mapping, merging, variant calling, post-processing and QC metric collection for all sequencing events. To conclude, a suite of annotation tools was used to add annotation data to the VCF; “Cassandra” brings together function, frequency, and other relevant information obtained from using AnnoVar with UCSC and RefSeq gene models, as well as many other internal and external data resources. The Fluidigm SNPtrace method for rapidly genotyping 96 SNP sites was then employed to ensure sample identity and integrity, verify gender, and detect contamination (Liang-Chu et al. 2015). The Error Rate In Sequencing (ERIS) software, developed at the HGSC, then verified assay sample identity.

### 2.4 | Data Processing and Analysis of Short Read Sequencing Data

The bioinformatic pipeline Parliament 2 (Zarate et al. 2020), consisting of Manta, Delly, Lumpy, Breakdancer, Crest, and Pindel, was performed to capture genomic coordinates of SV variant calls. These results were visualized using an in-house tool, VizCNV (Du et al. 2025), which is also publicly available (https://github.com/Carvalho-Lab). VizCNV incorporates read-depth information from the proband and the parents to facilitate *de novo* SV detection, as well as BAF (B-allele frequency) of inherited SNV alleles to identify CNV

origin. Mosdepth v0.3.3 (Pedersen and Quinlan 2018) was used to calculate the average read depth from the short-read sequencing in 1 kb windows. Then, the median read depth of each chromosome was used to normalize the read depth of each chromosome and the $\log_2$ ratio for each 1000 bp window was calculated. SLM Suite implemented in the DNAcopy Bioconductor package was applied to create the segmented $\log_2$ ratio profiles (Orlandini et al. 2017; Venkatraman and Olshen 2007). The individual and segmented $\log_2$ ratios could then be visualized together using ggplot2. A $\log_2$ ratio of 0 represents a diploid copy number state, ~0.58 represents a copy number gain of $n = 3$ (duplication in autosomes), and − 1 represents a heterozygous copy number loss. To call individual germline SNVs and indels, we used GATK v.4.1.9.0 (McKenna et al. 2010). The "-GVCF" option was used for GATK haplotypecaller, which yields a gVCF file including references or variant information for all loci. The gVCF files for the family were combined and the parental genotypes were used to recalibrate the proband's genotype. We used function FindCNV in VizCNV to identify *de novo* CNVs > 10 kb and filtered out calls overlapping with repetitive sequences. Breakpoint junctions of the *de novo* CGR calls identified by Parliament 2 were analyzed from short-read GS BAM files using Integrated Genomics Viewer (IGV) (https://igv.org); information from split-reads were used to make junction alignments. BAF analysis is performed in conjunction to the CNV calls to provide support (Du et al. 2025). BAF is calculated as $B/(A + B)$, where $B$ is the signal of the B allele and A is the signal of the A allele. The typical BAF is 0.50 for heterozygous autosomal regions, however the expected BAF for autosomal duplications are 0.33 and 0.67 and for some types of triplications are 0.25 and 0.75. For heterozygous deletions, the expected BAF are 0 and 1 (Carvalho et al. 2019).

### 2.5 | Single Nucleotide Variants and Indel Analysis Using Seqr

The trio's genomic data were analyzed on web-based variant prioritization tool seqr v1.0-fa92af86 (Pais et al. 2022). In built search filters "*de novo* dominant" and "recessive" (restrictive and permissive) were utilized for family based variant search. Rare, moderate to high impact variants with good sequencing quality (allelic balance > 0.4) and minor allele frequency of < 0.0001 gnomAD (v4.1) (https://gnomad.broadinstitute.org) for the *de novo* model and < 0.01 for the recessive model were carefully reviewed for phenotypic associations. Also, based on phenotype, a more permissive approach was taken into consideration for rasopathy genes. Variants with minor allele frequency ≤ 0.01 gnomAD (v4.1.0) were filtered and assessed with respect to American College of Medical Genetics guidelines (Richards et al. 2015).

### 2.6 | Short-Read-Based Complex Genomic Rearrangement (CGR) Detection Pipeline

The general expectation for CNV detection is that they will be defined by one breakpoint junction (e.g., deletions and tandem duplications) or by two breakpoint junctions (e.g., insertions, inversions, balanced translocations). For those types of "simple" SVs, read-depth calls and discordant reads/soft-clipped reads will provide redundant information with a consistent match regarding breakpoint and breakpoint junction coordinates. In contrast, per definition CGRs contain more than one SV *in cis*, often constituted by multiple breakpoint junctions. In such events, the CNV breakpoint coordinates will be displaced regarding the SV breakpoints, and they also may appear to represent unrelated events from the VCF

outputs. Based on those features, we developed an automated pipeline (Figure S1) to screen for unbalanced CGRs resulting in chromoanasynthesis events.

First, we identified all CNVs and SVs independently. CNVs were detected by VizCNV whereas SVs were detected by split-reads based callers in Parliament 2, details described in Methods section “data processing and analysis of short read sequencing data”. Then, we excluded CNVs overlapping segmental duplications (SDs) by at least 98% as they are likely to be false-positives (Du et al. 2025). For each filtered CNV, we performed a spatial search for overlapping or proximal SVs. This process involved breakpoint mapping, where for every CNV breakpoint, the “geographically” closest SV breakpoint was identified, as well as SV type mapping, where the type of SV is considered. A CNV and SV were considered a “match” only if they satisfied two primary conditions: the left and right breakpoints of the CNV and SV were identical and they both shared the same variant type (e.g., both classified as a DEL). Subsequently, we removed CNVs that have matching SV calls. From the set of CNVs that have non-matching SV calls, we applied a stringent proximity criterion, filtering CNVs with SVs located within 2.0 kilobases (kb) from either breakpoint (Figure S1).

Following the automated portion of the analysis in SEA101, we manually inspected each CGR candidate call using IGV and VizCNV to evaluate the CNV and SV calls regarding $\log_2$ ratio, split reads, BAF, and repeat annotation to confirm likelihood of the call to reflect CGR events not representing mapping artifacts. We then analyzed the CGR and assembled the genomic region by using long-read sequencing data and/or optical genome mapping.

As we focused on chromoanasynthesis events, i.e., unbalanced CGRs, this pipeline will not detect balanced CGRs. Moreover, VizCNV was optimized to detect CNVs > 10 kb (Du et al. 2025), therefore smaller CGRs will not be detected. Finally, short-read sequencing maps poorly within segmental duplications and SV callers show limited detection of breakpoint of variants overlapping segmental duplications. Therefore CNV calls mapped within those regions were not included in the analysis.

### 2.7 | Long-Read Sequencing and Data Processing

The Human Genome sequencing center (HGSC) at Baylor College of Medicine completed whole-genome sequencing using long reads from the Pacific Biosciences sequencing platform. The DNA quality was assessed using Qubit and the Femto Pulse system (Agilent). Afterwards, 15 μg of the DNA and the SMRTbell Express Template Kit 2.0 were used to construct a library, averaging a fragment length of 15 kb. Two SMRTcells were then sequenced per library, using the PacBio Sequel IIe instrument with SMRTlink v10, yielding an average of 25 Gb of HiFi yield and 1.6 million HiFi reads per sample.

### 2.8 | Optical Genome Mapping

Ultra-high molecular weight (UHMW) genomic DNA was extracted by Bionano (San Diego, CA, United States) from frozen whole blood using the Bionano Prep SP Blood and Cell Culture DNA Isolation Kit (Bionano Genomics) to be used for optical genome mapping (OGM). The DNA from this procedure was then quantified and sized using the Qubit dsDNA BR Assay Kit. The 0.75 μg of the UHMW gDNA was then labeled using the Bionano Prep Direct Label and Stain (DLS) Protocol (Bionano Genomics), which was then

loaded into a Saphyr Chip flow cell to run on the Saphyr optical mapping system (Bionano Genomics). Each run generated approximately 1800 Gb of data. The raw optical mapping molecules were then filtered via a preliminary bioinformatics pipeline to sort out molecules smaller than 150 kb in size and with less than nine labeling motifs. This generates a *de novo* assembly of the genome. The Bionano Solve v3.7 RefAligner module was used to align the data to an *in silico* reference genome (hg19/GRCh37). A custom Bionano SV caller then generated structural variant calls through comparison to the reference genome which proposed breakpoint junctions to be manually inspected via visualization on the Bionano Access software program v1.7.2.

### 2.9 | Bionano SV Analysis

Optical mapping was run on the Saphyr platform by Bionano Genomics (San Diego, CA, United States) and analyzed using the Bionano-Solve pipeline. The maps were detected using AutoDetect, then assembled with the *de novo* assembly package AssembleMolecules. The consensus maps resulting from this were aligned to a reference genome (hg19/GRCh37) with Bionano RefAligner and variants of interest were visualized with Bionano Access.

## 3 | Results

### 3.1 | Clinical Report

This case reports a 7-year-old female proband (SEA101) born to non-consanguineous healthy parents (33-year-old mother, SEA102, and 33-year-old father, SEA103). She has two healthy siblings, neither having any history of medical or developmental concerns (Figure 1A).

She was born at 38 weeks gestational age by spontaneous vaginal delivery, after a pregnancy complicated by a prenatal diagnosis of tetralogy of Fallot (TOF) and renal pelviectasis. Postnatally, she was found to have dysmorphic features, bilateral vesicoureteral reflux, and hydronephrosis, mild anorectal stenosis, bilateral nasolacrimal duct obstruction, and mild laryngomalacia. She underwent TOF variant repair around 6 months of life.

She had strabismus corrective surgery within the first year of life. She was additionally diagnosed with severe central and obstructive sleep apnea. Brain MRI showed mild restricted diffusion within the bilateral globus pallidi with subtle T2/FLAIR hyperintense signal. Her formal developmental evaluation at 3 years and 7 months (43 months) of age demonstrated her gross motor skills to be at the 24-month level, her visual motor problem-solving skills to be at the 28.5-month level, and her speech and language skills to be at the 25-month level, consistent with mild global developmental delay. On physical examination at 4 years and 6 months of age, her weight was at the 37th percentile, height was at the 10th percentile, and head circumference was at the 47th percentile. She had dysmorphic features, including mild coarse facial features, prominent forehead, hypertelorism, arched eyebrows, flat nasal bridge, upturned nose, carp-shaped mouth, high-arched palate, small umbilical hernia, and mild fifth digit clinodactyly (Table S1).

### 3.2 | Copy-Number Analysis Revealed a Complex Structural Variant at chr2q31

As part of her clinical evaluation, a clinical oligonucleotide-based whole genome CMA analysis was completed, which revealed a copy number gain of ~2.6 Mb on the long arm of chromosome 2, characterized as a triplication-duplication-triplication (TRP-DUP-TRP) (Figure 1B). The affected region included the 2q31.1 and 2q31.2 cytogenetic bands. Further testing with a higher-resolution customized CMA, with probes specifically designed to interrogate this region, revealed the structure to be more complex, including two more regions of duplication (Figure 1C). The CNV was resolved to be a DUP-TRP-DUP-TRP-DUP spanning 2.7 Mb (Table S2). Short-read data was used to calculate the read-depth ratio, which confirmed the DUP-TRP-DUP-TRP-DUP structure (Figure 1D). Trio B-allele frequency analysis confirmed the copy number of each segment and demonstrated that the affected alleles were inherited from the same maternal chromosome (Figure 1D), suggesting an intrachromosomal CGR occurred in the maternal gametogenesis or during embryonic development.

The short reads spanning the SV were analyzed in IGV. Breakpoint junctions were identified by soft-clipped reads corresponding to the copy-number changes followed by downloading the sequencing of those reads for sequencing alignment. These reads were mapping to locations on the chromosome as far as 2.7 Mb away, indicating that there were junctions connecting distant regions (Figure 2). PacBio long-read whole genome sequencing and optical genome mapping were used as an orthogonal technology to confirm the junctions at the nucleotide level resolution and the genomic structure, respectively (Figure 2). Analysis of the variant calls with the Access software identified two of the three junctions. Mapping the contigs generated by the *de novo* assembly pipeline to hg19 identified the three junctions found with GS. Manual inspection of the molecules used to build the assembly confirmed the three junctions had support from molecules spanning said junctions. With these three tools supporting the location of the breakpoints, a final structure DUP-TRP-DUP-TRP-DUP could be proposed consisting of three copies of segments B and D and two copies of segments A, C, and E (Figure 2).

### 3.3 | Published Patients Carrying chr2q31 Duplications

Following the characterization of the CGR on chromosome 2, we used the proposed structure and junctions to analyze the genetic content impacted by the rearrangement. Using UCSC genome browser, we determined which genes span the CGR, specifically focusing on those previously associated with diseases in the OMIM database (Hamosh et al. 2005). In the impacted region, there are nine OMIM genes known to be previously associated with diseases. Literature search and analysis of the DECIPHER database (https://www.deciphergenomics.org/browser) identified 14 individuals carrying pathogenic or likely pathogenic copy-number gain overlapping chr2q31 (Bragin et al. 2013). Of these reports, the phenotypic similarities were very limited; SEA101 is unique regarding the size and complexity of the CGR, gene content, as well as clinical phenotype. Published individuals with duplications spanning this chromosomal region present with variable clinical phenotypes, with two families reported with mesomelic dysplasia, Kantaputra type (MDK, MIM#156232) (Cho et al. 2010; Ghoumid et al. 2011; Kantaputra et al. 2010),

while another family presented bilateral 3–4 finger cutaneous syndactyly and nystagmus (42) (Figure S2).

### 3.4 | In-House CGR Pipeline Revealed Additional CGRs in SEA101

The in-house chromoanasynthesis, CGR detection pipeline developed here called a total of 38 events (Table S3, Figure S1). This number is consistent with the mean (39.5) number of CGR calls per individual in our database of 399 individuals with short-read sequencing from families with Mendelian diseases (Figure S1). Of the 38 calls, IGV and VizCNV analysis based on inconsistent $\log_2$ ratio concerning the called CNV, overlap with large SDs or with pericentric/centromeric regions and inconsistent B-allele frequency across the CNV eliminated 21 calls. From the 17 remaining calls, 7 were part of the same CGR on chromosome 2. A total of five additional CGRs were detected (Figure 3) by the short-read CGR pipeline, confirmed and assembled (Figure 3, Table S2) by using long-read sequencing analysis: a paternally inherited DEL-INV-DEL on 2q31.2 (Figure S3), a maternally inherited DUP-NML-DUP on 7p11.2 (Figure S4), a paternally inherited LINE-mediated DEL-INV-DEL-NML-INV-DEL on 11q11 (Figure S5), a paternally inherited INV-DUP on Xp21.3 (Figure S6), and a paternally inherited *AluY*-mediated, DUP-NML-DUP-NML-DUP on Xp21.2-p21.1 (Figure S7). One additional CGR event consistent with triplication and inversion spanning 1.2 Mb of 16p12.2 could not be assembled due to its size and the presence of large SDs at the CNV junctions.

## 4 | Discussion

CGRs characterized by clustered unbalanced gains and losses carrying multiple breakpoint junctions, often more ≥ 5, involving a single chromosome have been defined as chromoanasynthesis (Liu et al. 2011). Those events are very rare and not observed consistently in one given genomic disorder. However, chromoanasynthesis-like events with ≥ 2 and < 5 breakpoint junctions are often observed in genomic disorders including Yuan-Harel-Lupski syndrome (MIM#616652), *MECP2* duplication syndrome (MIM#300260) and Pelizaeus-Merzbacher disease (MIM#312080). In fact, they constitute 28%–40% of pathogenic CNVs in those loci (reviewed in Schuy et al. 2022) (Schuy et al. 2022) with some specific emerging structural patterns such as inverted triplications flanked by duplications (DUP-TRP/INV-DUP) and interspersed duplications (DUP-NML-DUP) (Bahrambeigi et al. 2019; Grochowski et al. 2024; Gu et al. 2015; Pehlivan et al. 2024; Yuan et al. 2015). Inverted repeat pairs in the form of LCRs or SDs (Carvalho et al. 2011; Grochowski et al. 2024) or *Alu* and LINE (Gu et al. 2015) mediate formation of those structures by break-induced replication (BIR) or microhomology-mediated break-induced replication (MMBIR), often accompanied by a second junction formed by canonical nonhomologous end joining (c-NHEJ) or microhomology-mediated end-joining (MMEJ) (reviewed in Carvalho and Lupski 2016) (Carvalho and Lupski 2016).

Although the prevalence of intermediate chromoanasynthesis CGR structures for specific genomic disorder loci is known, estimate for the number of those events' genome-wide is unclear. Here we investigate a proband with multiple phenotypic anomalies carrying a *de novo* CGR mapping to chromosome 2q31.1. Initial CMA revealed a 2.6 Mb rearrangement

characterized as a TRP-DUP-TRP. Further analyses of the impacted region using higher resolution genomic approaches revealed a more complex rearrangement than originally proposed, with two additional duplications flanking the ends, making the final proposed structure spanning 2.72 Mb. Three breakpoint junctions were characterized, one of them shows 4 bp of microhomology whereas the other two show non-templated insertions varying from 4 to 17 bp (Table S2). The occurrence of amplification (duplications and triplications) of multiple segments originating from the same maternal chromosome supports a mitotic error during gametogenesis or early embryo development potentially involving MMBIR.

Pipeline development of a chromoanasynthesis screening based on independent CNV and SV detection revealed that the proband carries a total of 7 CGRs, encompassing 10.7 kb to 8.1 Mb, affecting segments at chromosomes 2, 7, 11, 16 and X. Six out of seven could have inheritance established. Four out of six CGRs were inherited from paternal chromosomes (2q31.2, 11q11, Xp21.3 and Xp21.2-p21.1), one from the maternal chromosome (7p11.2) and one was generated *de novo* (2q31.1) (Figure 3, Table S2). Two to four breakpoint junctions per CGR were detected and confirmed by both short-read and long-read sequencing alignments in both proband and parental DNA when inherited. Three out of six resolved CGRs show inversions as part of the rearrangement; inverted segments vary in size from 16 to 21.6 kb. Microhomology varies from 1 to 31 bp, the latter generating a DUP-NML-DUP-NML-DUP event at Xp21.2-p21.1 mediated by an *AluY* pair sharing 88% of nucleotide similarity. Gene content analysis of those inherited CGRs present on chromosomes 2, 7, 11, 16 and X do not support contribution to proband's clinical phenotype. Populational-based studies indicate that CGRs account for approximately 1.6%–3.3% of all identified SVs, depending on the methodology employed (Abel et al. 2020; Collins et al. 2020, 2017). Direct comparison between those estimates and our findings is not appropriate, as our pipeline is specifically designed to enrich for unbalanced CGRs. Nevertheless, these population studies support the observation that a subset of SVs in human genomes represents inherited CGRs that do not result in clinical consequences for carriers.

The proband described herein presented with TOF and dysmorphic features. Comparison with others presenting pathogenic CNVs spanning chr2q31.1 (Table S2) did not reveal obvious triplosensitive genes that could have contributed to those traits. Considering the size of the DUP-TRP-DUP-TRP-DUP and the lack of similarities to other rare patients with reported overlapping CNVs, it is plausible that the displayed phenotypes may be arising from one or more dosage-sensitive genes within the CGR (Table S4). *LNPK* within this interval encodes a transmembrane protein with a role in the endoplasmic reticulum tubular network maintenance. Homozygous pathogenic variants affecting *LNPK* have been shown to cause a neurodevelopmental disorder with epilepsy and hypoplasia of the corpus callosum (NEDEHCC, MIM#618090) (Breuss et al. 2018). Specific symptoms include poor growth, absent speech, inability to walk, and delayed psychomotor development, none of which were part of the proband's clinical presentation. However, we cannot rule out the possibility of its contribution to the neurological phenotype in a dominant mode.

In SEA101, the duplicated segment C includes most of the HOXD gene cluster except *HOXD1*, *HOXD3*, and *HOXD4*, which are included in the triplicated segment D (Figure 2). Nonetheless, the proband shows no significant skeletal phenotype. Patients carrying

duplications involving the HOXD gene cluster can exhibit mesomelic dysplasia (MDK), as reported in one carrying a CGR consisting of two interspersed large duplications of 474 kb and 511 kb and another carrying a 1 Mb duplication (Figure S2); (Cho et al. 2010; Ghoumid et al. 2011; Kantaputra et al. 2010). Lack of skeletal phenotype in large duplications spanning the HOXD cluster was reported in a father and daughter carrying a 3.8 Mb duplication encompassing 2q31.1-q31.2, in whom the skeletal phenotype was limited to bilateral 3–4 cutaneous syndactyly of hands (Cho et al. 2010; Ghoumid et al. 2011; Kantaputra et al. 2010). Importantly, control of the HoxD cluster expression is performed by nearby regulatory elements that act in a regulated mode during limb development in both space and time, which implicates that rearrangement of the genomic structure within this chromosomal region may cause misexpression due to position effects (Kantaputra et al. 2010; Tarchini and Duboule 2006). In fact, a small inversion of 22.9 kb disrupting *HOXD13* was recently shown to segregate with mesomelic dysplasia, consistent with this position effect hypothesis (Pagnamenta et al. 2024). Overall, the mild skeletal developmental phenotype shown by the proband reported herein supports the hypothesis of position effects due to disconnection of the regulatory elements relevant to limb and hand development rather than dosage effects due to DNA copy number gain of the HOXD cluster. The larger CGR (2.7 Mb) carried by SEA101, similar to the 3.8 Mb duplication reported by (Cho et al. 2010; Ghoumid et al. 2011; Kantaputra et al. 2010), do not seem to lead to relevant gene expression dysregulation that results in limb malformations.

In conclusion, CGR screening using mismatched CNV and SV calls is an effective approach to investigating chromoanasynthesis events. The data reported here supports the hypothesis that: i. chromoanasynthesis CGRs with less than 5 breakpoint junctions are more prevalent than those involving multiple breakpoint junctions, ii. CGRs are observed genome-wide involving few kb to several Mb of DNA; iii. Analysis in conjunction with CNVs and SVs based on independent calls is an effective approach for CGR discovery using short-read sequencing data. The contribution of inherited CGRs to the clinical phenotype of probands with genomic disorders needs to be further investigated.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The authors would like to thank the family for their continued support and participation in this research effort.

### Funding

This work was supported by the National Institute of General Medical Sciences (R01 GM132589).

## Data Availability Statement

Sequencing data from the regions around detected CGRs is available in the SRA database under accession code PRJNA1257858. Microarray data is available in the Gene Expression Omnibus under accession code GSE296122. The code used for data analysis is available at https://github.com/Carvalho-Lab.

## References


Abel HJ, Larson DE, Regier AA, et al. 2020. "Mapping and Characterization of Structural Variation in 17,795 Human Genomes." Nature 583: 83–89. 10.1038/s41586-020-2371-0. [PubMed: 32460305]

Bahrambeigi V, Song X, Sperle K, et al. 2019. "Distinct Patterns of Complex Rearrangements and a Mutational Signature of Microhomeology Are Frequently Observed in *PLP1* Copy Number Gain Structural Variants." Genome Medicine 11: 80. 10.1186/s13073-019-0676-0. [PubMed: 31818324]

Beck CR, Carvalho CM, Akdemir ZC, et al. 2019. "Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2." Cell 176: 1310–1324. 10.1016/j.cell.2019.01.045. [PubMed: 30827684]

Bragin E, Chatzimichali EA, Wright CF, et al. 2013. "Decipher: Database for the Interpretation of Phenotype-Linked Plausibly Pathogenic Sequence and Copy-Number Variation." Nucleic Acids Research 42: D993–D1000. 10.1093/nar/gkt937. [PubMed: 24150940]

Breuss MW, Nguyen A, Song Q, et al. 2018. "Mutations in *LNPK*, Encoding the Endoplasmic Reticulum Junction Stabilizer Lunapark, Cause a Recessive Neurodevelopmental Syndrome." American Journal of Human Genetics 103: 296–304. 10.1016/j.ajhg.2018.06.011. [PubMed: 30032983]

Carvalho C, Coban-Akdemir Z, Hijazi H, et al. 2019. "Interchromosomal Template-Switching as a Novel Molecular Mechanism for Imprinting Perturbations Associated With Temple Syndrome." Genome Medicine 11: 1–14. 10.1186/s13073-019-0633-y. [PubMed: 30609936]

Carvalho C, Ramocki MB, Pehlivan D, et al. 2011. "Inverted Genomic Segments and Complex Triplication Rearrangements Are Mediated by Inverted Repeats in the Human Genome." Nature Genetics 43: 1074–1081. 10.1038/ng.944. [PubMed: 21964572]

Carvalho CM, and Lupski JR. 2016. "Mechanisms Underlying Structural Variant Formation in Genomic Disorders." Nature Reviews Genetics 17: 224–238. 10.1038/nrg.2015.25.

Carvalho CM, Pfundt R, King DA, et al. 2015. "Absence of Heterozygosity due to Template Switching During Replicative Rearrangements." American Journal of Human Genetics 96: 555–564. 10.1016/j.ajhg.2015.01.021. [PubMed: 25799105]

Cho TJ, Kim OH, Choi IH, et al. 2010. "A Dominant Mesomelic Dysplasia Associated With a 1.0-Mb Microduplication of *HOXD* Gene Cluster at 2q31.1." Journal of Medical Genetics 47: 638–639. 10.1136/jmg.2009.074690. [PubMed: 20577005]

Collins RL, Brand H, Karczewski KJ, et al. 2020. "A Structural Variation Reference for Medical and Population Genetics." Nature 581: 444–451. 10.1038/s41586-020-2287-8. [PubMed: 32461652]

Collins RL, Brand H, Redin CE, et al. 2017. "Defining the Diverse Spectrum of Inversions, Complex Structural Variation, and Chromothripsis in the Morbid Human Genome." Genome Biology 18: 36. 10.1186/s13059-017-1158-6. [PubMed: 28260531]

Cosenza MR, Rodriguez-Martin B, and Korbel JO. 2022. "Structural Variation in Cancer: Role, Prevalence, and Mechanisms." Annual Review of Genomics and Human Genetics 23: 123–152. 10.1146/annurev-genom-120121-101149.

Du H, Lun MY, Gagarina L, et al. 2025. "An Integrated Platform for Concurrent Structural and Single-Nucleotide Variants Improves Copy-Number Detection and Reveals Pathogenic Alleles in Undiagnosed Mendelian Families." Genome Medicine 18: 16. 10.1186/s13073-025-01593-8. [PubMed: 41470026]

Eisfeldt J, Pettersson M, Vezzi F, et al. 2019. "Comprehensive Structural Variation Genome Map of Individuals Carrying Complex Chromosomal Rearrangements." PLoS Genetics 15: e1007858. 10.1371/journal.pgen.1007858. [PubMed: 30735495]

Farek J, Hughes D, Salerno W, et al. 2023. "xAtlas: Scalable Small Variant Calling Across Heterogeneous Next-Generation Sequencing Experiments." *GigaScience* 12: giac125. 10.1093/gigascience/giac125.

Ghoumid J, Andrieux J, Sablonnière B, et al. 2011. "Duplication at Chromosome 2q31.1-q31.2 in a Family Presenting Syndactyly and Nystagmus." European Journal of Human Genetics 19: 1198–1201. 10.1038/ejhg.2011.95. [PubMed: 21654727]

Giardino D, Corti C, Ballarati L, et al. 2009. "*De Novo* Balanced Chromosome Rearrangements in Prenatal Diagnosis." Prenatal Diagnosis 29: 257–265. 10.1002/pd.2215. [PubMed: 19248039]

Grochowski CM, Bengtsson JD, du H, et al. 2024. “Inverted Triplications Formed by Iterative Template Switches Generate Structural Variant Diversity at Genomic Disorder Loci.” Cell Genomics 4: 100590. 10.1016/j.xgen.2024.100590. [PubMed: 38908378]

Grochowski CM, Gu S, Yuan B, et al. 2018. “Marker Chromosome Genomic Structure and Temporal Origin Implicate a Chromoanasynthesis Event in a Family With Pleiotropic Psychiatric Phenotypes.” Human Mutation 39: 939–946. 10.1002/humu.23537. [PubMed: 29696747]

Grochowski CM, Krepischi ACV, Eisfeldt J, et al. 2021. “Chromoanagenesis Event Underlies a *de Novo* Pericentric and Multiple Paracentric Inversions in a Single Chromosome Causing Coffin-Siris Syndrome.” Frontiers in Genetics 12: 708348. 10.3389/fgene.2021.708348. [PubMed: 34512724]

Gu S, Yuan B, Campbell IM, et al. 2015. “Alu-Mediated Diverse and Complex Pathogenic Copy-Number Variants Within Human Chromosome 17 at p13. 3.” Human Molecular Genetics 24: 4061–4077. 10.1093/hmg/ddv146. [PubMed: 25908615]

Hamosh A, Scott AF, Amberger JS, Bocchini CA, and McKusick VA. 2005. “Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders.” Nucleic Acids Research 33: D514–D517. 10.1093/nar/gki033. [PubMed: 15608251]

Kantaputra PN, Klopocki E, Hennig BP, et al. 2010. “Mesomelic Dysplasia Kantaputra Type Is Associated With Duplications of the *HOXD* Locus on Chromosome 2q.” European Journal of Human Genetics 18: 1310–1314. 10.1038/ejhg.2010.116. [PubMed: 20648051]

Liang-Chu MM, Yu M, Haverty PM, et al. 2015. “Human Biosample Authentication Using the High-Throughput, Cost-Effective Snptracetm System.” PLoS One 10: e0116218. 10.1371/journal.pone.0116218. [PubMed: 25714623]

Liu P, Erez A, Nagamani SCS, et al. 2011. “Chromosome Catastrophes Involve Replication Mechanisms Generating Complex Genomic Rearrangements.” Cell 146: 889–903. 10.1016/j.cell.2011.07.042. [PubMed: 21925314]

Liu P, Yuan B, Carvalho CM, et al. 2017. “An Organismal CNV Mutator Phenotype Restricted to Early Human Development.” Cell 168: 830–842. 10.1016/j.cell.2017.01.037. [PubMed: 28235197]

Lupski JR 1998. “Genomic Disorders: Structural Features of the Genome Can Lead to DNA Rearrangements and Human Disease Traits.” Trends in Genetics 14: 417–422. 10.1016/s0168-9525(98)01555-8. [PubMed: 9820031]

Ly P, and Cleveland DW. 2017. “Rebuilding Chromosomes After Catastrophe: Emerging Mechanisms of Chromothripsis.” Trends in Cell Biology 27: 917–930. 10.1016/j.tcb.2017.08.005. [PubMed: 28899600]

McKenna A, Hanna M, Banks E, et al. 2010. “The Genome Analysis Toolkit: A Mapreduce Framework for Analyzing Next-Generation DNA Sequencing Data.” Genome Research 20: 1297–1303. 10.1101/gr.107524.110. [PubMed: 20644199]

Nazaryan-Petersen L, Eisfeldt J, Pettersson M, et al. 2018. “Replicative and Non-Replicative Mechanisms in the Formation of Clustered CNVs Are Indicated by Whole Genome Characterization.” PLoS Genetics 14: e1007780. 10.1371/journal.pgen.1007780. [PubMed: 30419018]

Orlandini V, Provenzano A, Giglio S, and Magi A. 2017. “SLMSuite: A Suite of Algorithms for Segmenting Genomic Profiles.” BMC Bioinformatics 18: 1–5. 10.1186/s12859-017-1734-5. [PubMed: 28049414]

Pagnamenta AT, Yu J, Walker S, et al. 2024. “The Impact of Inversions Across 33,924 Families With Rare Disease From a National Genome Sequencing Project.” American Journal of Human Genetics 111: 1140–1164. 10.1016/j.ajhg.2024.04.018. [PubMed: 38776926]

Pais LS, Snow H, Weisburd B, et al. 2022. “Seqr: A Web-Based Analysis and Collaboration Tool for Rare Disease Genomics.” Human Mutation 43: 698–707. 10.1002/humu.24366. [PubMed: 35266241]

Pedersen BS, and Quinlan AR. 2018. “Mosdepth: Quick Coverage Calculation for Genomes and Exomes.” Bioinformatics 34: 867–868. 10.1093/bioinformatics/btx699. [PubMed: 29096012]

Pehlivan D, Bengtsson JD, Bajikar SS, et al. 2024. “Structural Variant Allelic Heterogeneity in *MECP2* Duplication Syndrome Provides Insight Into Clinical Severity and Variability of Disease Expression.” Genome Medicine 16: 146. 10.1186/s13073-024-01411-7. [PubMed: 39696717]

Regier AA, Farjoun Y, Larson DE, et al. 2018. "Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling Across Human Genetics Projects." Nature Communications 9: 4038. 10.1038/s41467-018-06159-4.

Reid JG, Carroll A, Veeraraghavan N, et al. 2014. "Launching Genomics Into the Cloud: Deployment of Mercury, a Next Generation Sequence Analysis Pipeline." BMC Bioinformatics 15: 1–11. 10.1186/1471-2105-15-30. [PubMed: 24383880]

Richards S, Aziz N, Bale S, 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." Genetics in Medicine 17: 405–424. 10.1038/gim.2015.30. [PubMed: 25741868]

Sanders AD, Meiers S, Ghareghani M, et al. 2020. "Single-Cell Analysis of Structural Variations and Complex Rearrangements With Tri-Channel Processing." Nature Biotechnology 38: 343–354. 10.1038/s41587-019-0366-x.

Schuy J, Grochowski CM, Carvalho CMB, and Lindstrand A. 2022. "Complex Genomic Rearrangements: An Underestimated Cause of Rare Diseases." Trends in Genetics 8: 1134–1146. 10.1016/j.tig.2022.06.003.

Schuy J, Sæther KB, Lisfeld J, et al. 2024. "A Combination of Long- and Short-Read Genomics Reveals Frequent p-Arm Breakpoints Within Chromosome 21 Complex Genomic Rearrangements." Genetics in Medicine Open 2: 101863. 10.1016/j.gimo.2024.101863. [PubMed: 39669604]

Tarchini B, and Duboule D. 2006. "Control of *HOXD* Genes' Collinearity During Early Limb Development." Developmental Cell 10: 93–103. 10.1016/j.devcel.2005.11.014. [PubMed: 16399081]

Venkatraman ES, and Olshen AB. 2007. "A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data." Bioinformatics 23: 657–663. 10.1093/bioinformatics/btl646. [PubMed: 17234643]

Yuan B, Harel T, Gu S, et al. 2015. "Nonrecurrent 17p11.2p12 Rearrangement Events That Result in Two Concomitant Genomic Disorders: The *PMP22-RAI1* Contiguous Gene Duplication Syndrome." American Journal of Human Genetics 97: 691–707. 10.1016/j.ajhg.2015.10.003. [PubMed: 26544804]

Zarate S, Carroll A, Mahmoud M, et al. 2020. "Parliament2: Accurate Structural Variant Calling at Scale." *GigaScience* 9: giaa145. 10.1093/gigascience/giaa145.

Zuccherato LW, Alleva B, Whiters MA, Carvalho CM, and Lupski JR. 2016. "Chimeric Transcripts Resulting From Complex Duplications in Chromosome Xq28." Human Genetics 135: 253–256. 10.1007/s00439-015-1614-x. [PubMed: 26667017]

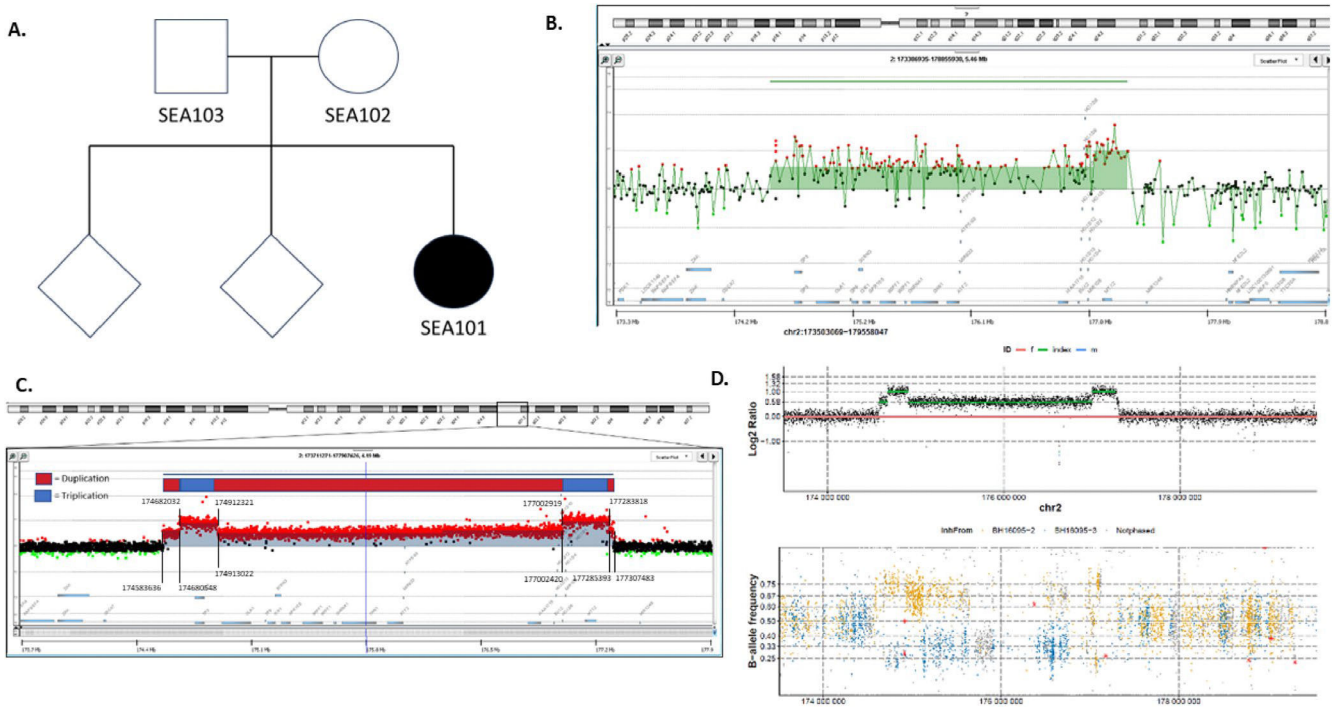


**FIGURE 1 |.**

Preliminary analysis of proband. (A) Pedigree with father (SEA103), mother (SEA102), and proband (SEA101) as well as two unaffected siblings. (B) Original 30 kb array CMA that was characterized as a TRP-DUP-TRP by the diagnostic laboratory. (C) Detailed custom CMA shown in Agilent Genomic Work Bench with an average spacing of 964 bp with 700 bp resolution between probes, detailing a DUP-TRP-DUP-TRP-DUP event. (D) Read depth graphs shown in VizCNV show increased $\log_2$ ratio consistent with a DUP-TRP-DUP-TRP-DUP structure, with trio analysis confirming the structural variant to be *de novo*. The B-allele frequency plot, generated using SNP data from GATK haplotype caller, demonstrates the affected alleles originated from the maternal chromosome.

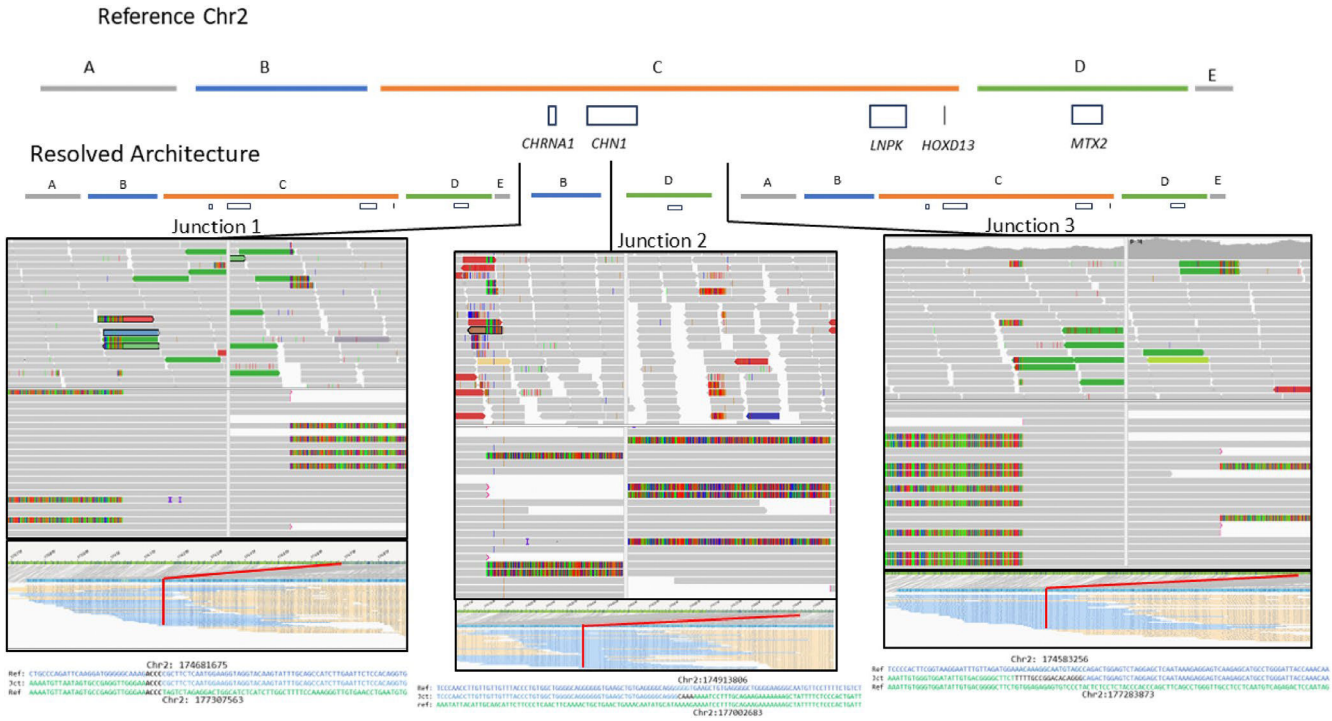


**FIGURE 2 |.**

Resolved structure for the *de novo* 2q31.1 complex genomic rearrangement (CGR). The resolved structure shows each breakpoint as indicated by short-read whole-genome sequencing, PacBio long-reads, optical genome mapping, and breakpoint junction alignment. Rectangles below the structures represent the relative position of five genes out of 23 (as defined by NCBI RefSeq Select and MANE, hg19) present overlapping this CGR.

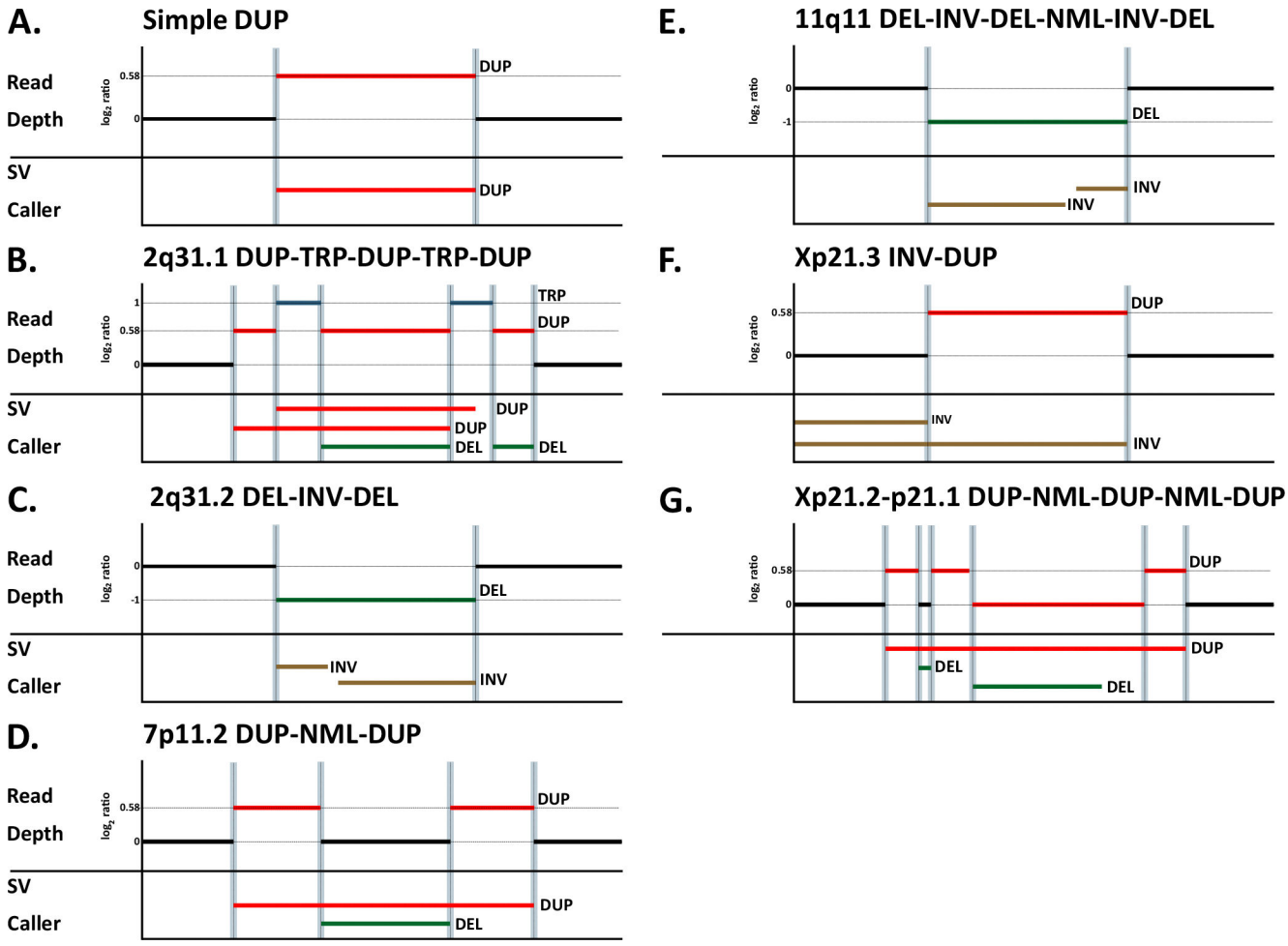


**FIGURE 3 |.**

Representative plots of one simple and six complex genomic variants computationally identified in SEA101. For each plot: Top displays the $\log_2$ ratio based on read depth (CNV calls) including duplications (0.58, DUP, horizontal red bar) or triplications (1, TRP, horizontal dark blue bar), deletion (−1, DEL, horizontal green bar); Bottom displays genomic coordinates relative to calls from split-read-based structural variant callers in Parliament 2. Vertical light gray bars indicate overlap between read depth breakpoint and SV call breakpoints. (A) Example of a Simple DUP structure shows matching CNV and SV calls; (B–G) Examples of inconsistent CNV and paired SV calls flagged as CGRs in SEA101. Note that size and coordinates match for at least one breakpoint within 2 kb window, but they either do not match for the type of CNV/SV call or the second breakpoint is inconsistent between CNV and SV. The examples shown here correspond to the CGRs reported in Table S2. Resolved CGR genomic structures based on short-read, long-read sequencing and optical mapping analysis are described on top of each plot with corresponding derivative genome assembled structures displayed in Figures 2 and S3–S7.