

Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints

Scott Newman,^{1,2} Karen E. Hermetz,^{1,2} Brooke Weckselblatt,¹ and M. Katharine Rudd^{1,*}

Interpreting the genomic and phenotypic consequences of copy-number variation (CNV) is essential to understanding the etiology of genetic disorders. Whereas deletion CNVs lead obviously to haploinsufficiency, duplications might cause disease through triplosensitivity, gene disruption, or gene fusion at breakpoints. The mutational spectrum of duplications has been studied at certain loci, and in some cases these copy-number gains are complex chromosome rearrangements involving triplications and/or inversions. However, the organization of clinically relevant duplications throughout the genome has yet to be investigated on a large scale. Here we fine-mapped 184 germline duplications (14.7 kb–25.3 Mb; median 532 kb) ascertained from individuals referred for diagnostic cytogenetics testing. We performed next-generation sequencing (NGS) and whole-genome sequencing (WGS) to sequence 130 breakpoints from 112 subjects with 119 CNVs and found that most (83%) were tandem duplications in direct orientation. The remainder were triplications embedded within duplications (8.4%), adjacent duplications (4.2%), insertional translocations (2.5%), or other complex rearrangements (1.7%). Moreover, we predicted six in-frame fusion genes at sequenced duplication breakpoints; four gene fusions were formed by tandem duplications, one by two interconnected duplications, and one by duplication inserted at another locus. These unique fusion genes could be related to clinical phenotypes and warrant further study. Although most duplications are positioned head-to-tail adjacent to the original locus, those that are inverted, triplicated, or inserted can disrupt or fuse genes in a manner that might not be predicted by conventional copy-number assays. Therefore, interpreting the genetic consequences of duplication CNVs requires breakpoint-level analysis.

Introduction

Genomic copy-number variation (CNV) is a major cause of birth defects, intellectual disability, autism spectrum disorders, psychiatric disorders, and other neurodevelopmental disabilities. Approximately 10%–15% of children referred for diagnostic CNV testing have a rare deletion or duplication responsible for their phenotype.^{1,2} Clinical interpretation of germline CNVs is based on genomic size, gene content, and segregation of the CNV with phenotype.³ Recurrent CNVs with common breakpoints can define genomic disorders characterized by particular clinical features because the same genes are deleted or duplicated.⁴ However, even recurrent deletions and duplications exhibit variable expressivity and incomplete penetrance among individuals with the same CNV.^{5–7} Of the approximately 75% of germline CNVs that are non-recurrent,^{8–10} some share a critical region that segregates with a particular phenotype, but the pathogenicity of others cannot be easily inferred from the genes deleted or duplicated. Thus, interpreting the phenotypic consequences of CNVs is challenging.

Haploinsufficiency for genes within a deletion CNV is a well-recognized cause of genetic disease. Duplication CNVs can lead to triplosensitivity for some genes, among them *CREBBP*¹¹ (MIM 600140), *LMNB1*¹² (MIM 150340), *MECP2*¹³ (MIM 300005), and *PLP1*¹⁴ (MIM 300401), but

the pathogenicity of most duplications is not explained by an extra copy of one gene. Larger CNV size and greater gene number correlate with duplication pathogenicity,^{2,8} consistent with deleterious consequences from extra copies of many genes. In addition, phenotypes could be due to disruption or misregulation of genes that span duplication breakpoints; however, it is impossible to infer the effects of duplications on gene structure without resolving breakpoints and determining the orientation and location of the duplicated segment. Though many deletion breakpoints have been sequenced, sequencing duplication CNVs has proved more of a challenge.^{15–17} Thus, many questions remain about the genomic organization and genetic consequences of duplication CNVs.

In this study, we fine-mapped 184 clinically relevant duplications and sequenced 130 breakpoint junctions. This large-scale analysis revealed that most duplications are tandem in direct orientation adjacent to the original locus. Intragenic duplications disrupt the reading frame of at least some gene isoforms. Intergenic direct duplications might disrupt or fuse genes at breakpoint junctions, but leave one intact gene copy on the duplication allele. Inverted and inserted duplications have the potential to disrupt genes at breakpoint junctions without preserving an intact copy (Figure 1). Thus, determining the orientation and location of duplication CNVs is essential to interpret their effects on genes and correlate with phenotypes.

¹Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

²These authors contributed equally to this work

*Correspondence: katie.rudd@emory.edu

<http://dx.doi.org/10.1016/j.ajhg.2014.12.017>. ©2015 by The American Society of Human Genetics. All rights reserved.

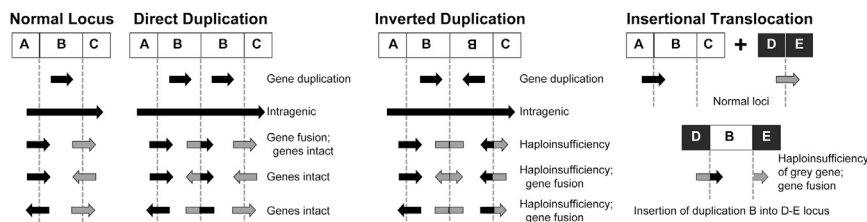


Figure 1. Genetic Outcomes for Duplication CNVs

Duplication of region B can be in direct or inverted orientation or can be inserted at another locus. Genes (arrows) and duplication breakpoints (dashed lines) are shown. Whole-gene duplication can lead to triplosensitivity, whereas intragenic duplications can disrupt the reading frame and cause loss of function. Direct intergenic

duplications can generate a nonfunctional gene at the breakpoint junction while maintaining intact genes at the edges of the duplication. Intergenic duplications with breakpoints in two different genes can create a gene fusion if the genes are in the same orientation and the reading frame is maintained. Inverted intergenic duplications can create a fusion gene at the junction and will mutate one gene (gray) without retaining an intact copy at the locus. Loss of one gene copy through inverted duplication can lead to haploinsufficiency. Insertional translocations can disrupt or fuse genes at the site of insertion (gray).

Subjects and Methods

Human Subjects

This study was approved by the Institutional Review Board (IRB) at Emory University. Individuals were referred for clinical microarray testing with indications including but not limited to intellectual disability, developmental delay, autism spectrum disorders, congenital anomalies, and dysmorphic features. Duplications were initially identified via diagnostic chromosomal microarray analysis (CMA) performed at Emory Genetics Laboratory (EGL). Clinical microarrays have genome-wide coverage with one oligonucleotide probe per ~75 kilobases and greater probe density in targeted regions.¹⁸ The genomic coordinates of duplications identified by CMA are listed in [Table S1](#).

High-Resolution Array CGH

We designed custom 60K CGH arrays with oligonucleotide probes targeted to the 250 kb surrounding proximal and distal ends of duplications identified by clinical CMA (Agilent Technologies). Oligonucleotide arrays were designed with the Agilent eArray program; array design ID (AMADID) numbers are listed in [Table S1](#). DNA extraction, microarray hybridization, scanning, and analysis were performed as described previously.¹⁹

Next-Generation Sequencing

Once we fine-mapped breakpoints by high-resolution array CGH, we targeted regions 20 kb proximal to 20 kb distal of breakpoints with SureSelect Target Enrichment probes (Agilent Technologies). We designed three SureSelect libraries that encompass 1.8, 2.3, and 2.7 Mb with 3× tiling to capture breakpoints from 190 subjects (ELID 393121, 397531, and 404011, respectively). Breakpoints mapped by array CGH and the corresponding SureSelect libraries are listed in [Table S1](#). SureSelect capture and sequencing were performed by the Genomic Services Lab at Hudson Alpha Institute for Biotechnology (Huntsville, AL). Five to seven genomic DNA samples were multiplexed per SureSelect capture using one of the three custom bait libraries. The resulting 38 capture libraries were barcoded and pooled, four libraries per sequencing lane. We performed 100-bp paired-end sequencing in 9.5 lanes of an Illumina HiSeq 2000 instrument.

Our structural variation (SV) pipeline identifies sequence reads that span duplication breakpoints. First, we aligned paired-end fastq files to the GRCh37/hg19 reference genome by using BWA-0.5.9²⁰ and identified improperly aligned pairs with the SAMtools-0.1.18 filter function.²¹ Discordant read pairs were clustered to predict structural variants.²² By using CIGAR scores, we identified split reads where only part of the read aligns to the refer-

ence genome and inspected these regions by IGV.²³ To map duplications in the pseudoautosomal region, sequence reads were aligned to a human reference genome without the Y chromosome, and reads were processed as above.

We also sequenced five genomes via WGS at Complete Genomics.²⁴ Using Complete Genomics Analysis Tools (CGA Tools), we converted mappings and reads to .bam files to analyze discordant and split reads with our SV pipeline.

Sanger Confirmation of Breakpoint Junctions

To confirm breakpoints mapped from NGS and WGS, we attempted to PCR and Sanger sequence breakpoint junctions.¹⁹ We downloaded the reference sequences surrounding predicted SV junctions from Ensembl and designed primers with Primer 3. For standard breakpoint PCR, we performed 30 cycles of 98°C for 10 s, 57°C for 15 s, and 68°C for 3 min. We also performed long-range PCR for some duplication junctions ([Table S1](#)). Starting with breakpoints identified by high-resolution array CGH, we designed multiple primer pairs spaced at approximately 3 kb intervals. We performed PCR using all possible primer combinations with touchdown PCR. An initial denaturing step at 95°C for 3 min was followed by 10 cycles of denaturation at 95°C for 30 s, annealing at 72°C (decreasing 1°C every cycle) for 45 s, and elongation at 72°C for 10 min. The remaining 25 cycles had an annealing temperature of 57°C with denaturation and elongation as above. We purified PCR products from agarose gels, and cloned and sequenced the products according to standard methods. DNA sequences were aligned to the human genome reference assembly GRCh37/hg19 with the BLAT tool at the UCSC Genome Browser.

Microhomology Simulations

To investigate the amount of homology shared between sequences brought together by duplication junctions, we generated a control dataset of simulated tandem, direct duplications. We applied the random number function within a custom Perl script to generate a list of 1,000 genomic regions from random chromosomes of random sizes between 14.7 kb and 25.3 Mb. We downloaded each genomic region from the GRCh37/hg19 reference genome using “getfasta” from BedTools.²⁵ For the 1,000 regions, we used a custom Perl script to array simulated duplications in direct orientation and count microhomology at the junctions. Code is available at SourceForge.

Mapping Insertions

We performed BLAT alignments to determine the origin of inserted sequence.²⁶ For insertions that were too short to

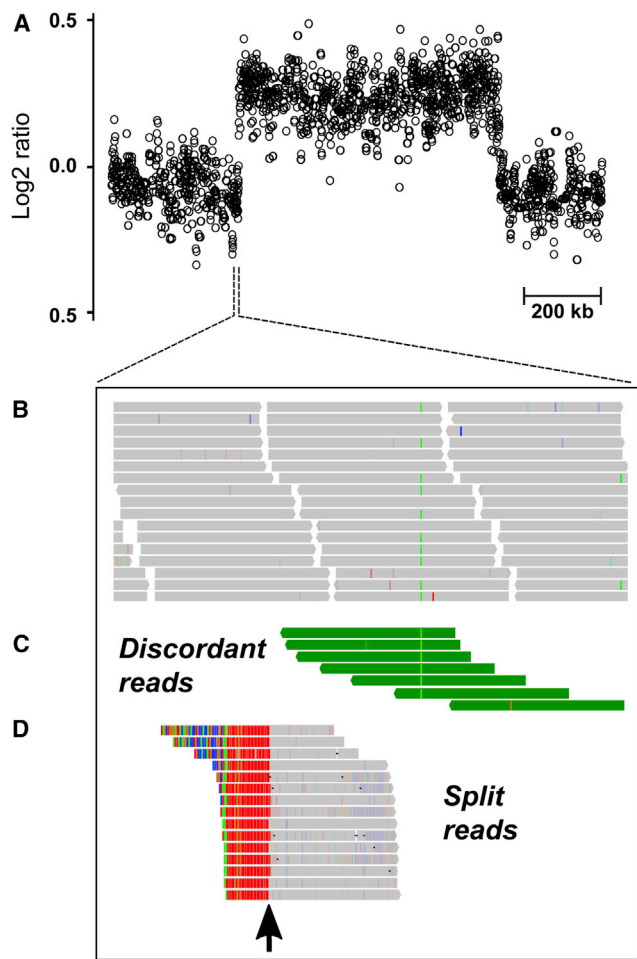


Figure 2. Duplication Breakpoint Sequencing

(A) High-resolution array CGH of genomic DNA from subject EGL464 fine maps the 568-kb duplication. Log₂ ratio of subject versus control signal intensity is shown on the y axis.

(B) SureSelect target enrichment of the 20-kb region surrounding breakpoints (dashed lines) followed by next-generation sequencing and alignment of paired-end reads (gray) reveals sequences from the normal chromosome 1 (chr1: 46,084,756–46,085,053).

(C) Discordant reads (green) that map to this region have mate pairs that map to the positive strand at chr1: 46,652,055–46,652,356, consistent with a direct, tandem duplication.

(D) Split reads that span the duplication junction misalign (colored vertical lines) to the reference genome at the site of the breakpoint (arrow; chr1: 46,084,825).

BLAT, we searched for nearby matching sequences. We downloaded from Ensembl 10 kb of genomic sequence proximal and distal of breakpoint junctions. For each 20-kb region, we searched for sequences similar to the inserted sequence using Perl regular expressions, allowing up to 30% of bases to mismatch.

Fusion Gene Prediction

For duplications with genes at both breakpoints, we analyzed the gene orientation and reading frame to predict fusions. We included all isoforms from Ensembl release 75 (GRCh37.p13) and counted in-frame fusions as those with the same exon phase.²⁷

Results

Duplication Cohort

We analyzed the genomic structure of 184 duplications from 170 unrelated individuals tested at Emory Genetics Laboratory (EGL) between 2007 and 2012 (Table S1). Fourteen individuals had two duplications detected by clinical array CGH testing that are derived from different chromosomes ($n = 7$) or the same chromosome arm ($n = 7$). We included duplications that were reported as pathogenic or of uncertain clinical significance and excluded common CNVs present in the general population.^{8,28,29} We also excluded CNVs with known etiologies: recurrent duplications mediated by non-allelic homologous recombination (NAHR) between segmental duplications, inverted duplications adjacent to terminal deletions, and copy-number gains due to supernumerary chromosomes, unbalanced translocations, and trisomy. As determined by clinical array testing, duplications ranged from 14.7 kb to 25.3 Mb, with mean and median sizes of 1.15 Mb and 532 kb, respectively (Table S1). We performed fluorescence in situ hybridization (FISH) or microarray testing on parent(s) of 78 probands to determine duplication inheritance. Five were de novo, 41 were maternally inherited, and 28 were paternally inherited. In four families, only the mother was tested, and the duplication was not maternally inherited.

Duplication Breakpoint Analysis

To fine map duplications, we designed custom high-resolution oligonucleotide microarrays that target ~250 kb around each breakpoint, with one probe per ~300 bp (Figure 2). We manually inspected array CGH data and called CNV boundaries as previously described¹⁹ (Table S1). Based on duplication breakpoints from high-resolution array CGH, we performed SureSelect Target Enrichment to capture sequence from 20 kb proximal to 20 kb distal of breakpoints. Because individuals in our cohort have a range of duplications, we were able to multiplex genomic DNA from five to seven subjects with different duplications per SureSelect capture and identify subject-specific junctions during sequence analysis. After sequence capture, we barcoded and pooled four SureSelect libraries per HiSeq lane and sequenced 100-bp paired end reads. We implemented a bioinformatics pipeline to identify discordant read pairs: those that mapped too far apart, too close together, in the wrong orientation, or to different chromosomes. Out of 181, a total of 131 (62%) targeted breakpoints were supported by unique discordant read pairs (mean = 12.6; median = 9) that mapped aberrantly compared to the reference genome. 91 of those junctions were also supported by split reads that spanned the breakpoint junction, and three junctions were supported by split reads but no discordant reads.

For 97 out of 116 junctions (84%) identified by discordant reads and/or split reads, we confirmed the breakpoints by Sanger sequencing. Some junctions failed to

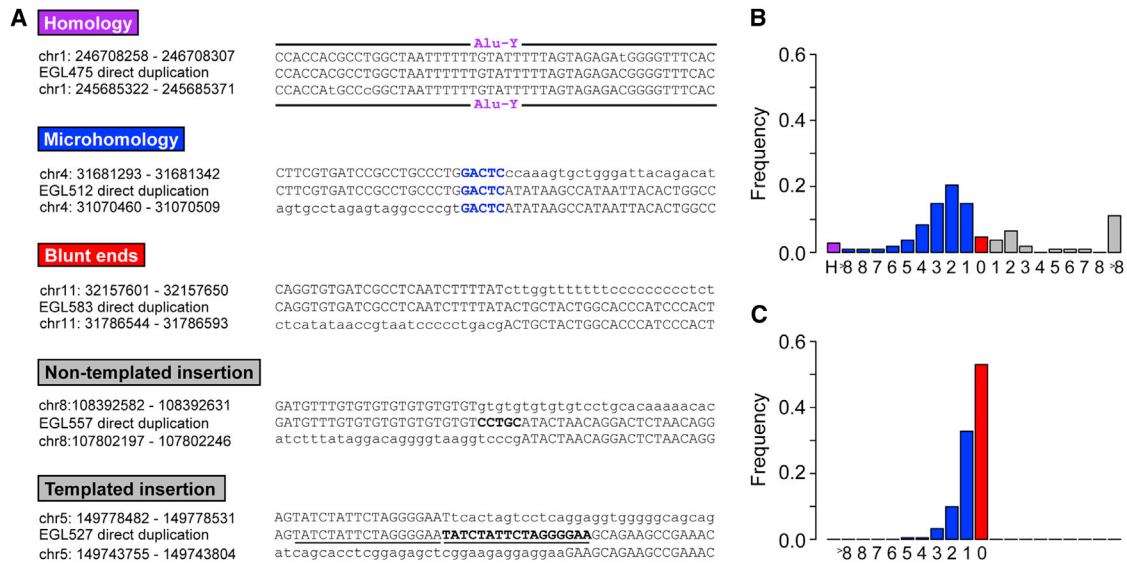


Figure 3. Breakpoint Junctions Reveal Signatures of DNA Repair

(A) Examples of junctions with *Alu-Alu* homology (purple), microhomology (blue), blunt ends, and insertions (bold) are shown. Duplication breakpoint junctions are shown as the middle sequence, aligned to the reference genome at the two sides of the direct duplications. Underlined sequence shows the origin of the templated insertion in EGL527.

(B) Frequency of *Alu-Alu* homology (H), microhomology (1 to >8 bp), blunt ends (0), and insertions (1 to >8 bp) at sequenced junctions. Colors are the same as in (A).

(C) Breakpoints from 1,000 simulated duplications have a different distribution of microhomology and blunt ends compared to observed junctions in (B) ($p = 5.117 \times 10^{-12}$).

confirm due to lack of genomic DNA, and others failed after multiple PCR attempts. In addition, we sequenced nine breakpoints by long-range PCR and Sanger sequencing. In these cases, the breakpoints delineated by high-resolution array CGH were sufficiently resolved to predict and sequence breakpoint junctions without SureSelect. We designed primers to PCR amplify duplications in direct or inverted orientation.¹⁵

We also performed WGS for five subjects with duplications.²⁴ Duplications in EGL698, EGL823, EGL824, EGL825, and LM223 were identified with CGAtools (Complete Genomics) and our SV pipeline. 11–847 pairs of discordant reads supported the duplication junctions in EGL823, EGL824, EGL825, and LM223. CGAtools called EGL698's 165-kb duplication by read-depth but not junction reads (Table S2).

We analyzed the 118 breakpoint junction contigs from Sanger sequencing or NGS split reads (Tables S3 and S4, respectively). Ten duplication junctions were shared between at least two individuals, so there were a total of 108 unique duplication junctions. 3 out of 108 (2.8%) junctions span homologous *Alu* repeats in the same orientation at the two sides of the duplication, consistent with *Alu-Alu* recombination. Duplication breakpoints from EGL475, EGL577, and EGL671 had 248 bp, 285 bp, and 267 bp of homology between recombining *Alus*. 28 junctions had short insertions (1–187 bp long) at the breakpoints. Five insertions are homologous to sequence at the breakpoint, eight are homologous to sequence 67–5,345 bp from the breakpoint, and 15 are of unknown

origin (Table S5). Five junctions had blunt ends, and 72 had microhomology 1–15 bp long (Figure 3). For the 77 junctions without insertions, we compared the length of microhomology (0–15 bp) to microhomology from 1,000 simulated tandem duplication junctions. Observed microhomology was significantly different from simulated microhomology according to the Student's t test with Welch's correction for unequal variances ($p = 5.117 \times 10^{-12}$).

Fine-mapping duplications to the base-pair level by high-resolution array CGH followed by breakpoint junction sequencing revealed greater complexity than recognized by clinical microarray testing. We analyzed 130 breakpoints with discordant, split, and/or Sanger sequence support from 112 subjects with 119 CNVs to interpret duplication organization and orientation (Table S1). 99 out of 119 (83%) were tandem duplications in direct orientation, whereas others were more complex rearrangements, including triplications (10), adjacent duplications (5), insertional translocations (3), an inverted duplication adjacent to a cryptic terminal deletion (LM223), and a duplication with unknown structure (EGL414).

Complex Duplications

Six individuals had two duplications derived from regions 300 kb to 2.63 Mb apart (Table S6). According to microarray analysis, these CNVs have a characteristic duplication-normal-duplication (DUP-NML-DUP) copy-number pattern.^{30–32} We fine mapped six DUP-NML-DUP rearrangements by high-resolution arrays and sequenced the breakpoint junctions of five. Five DUP-NML-DUPs were

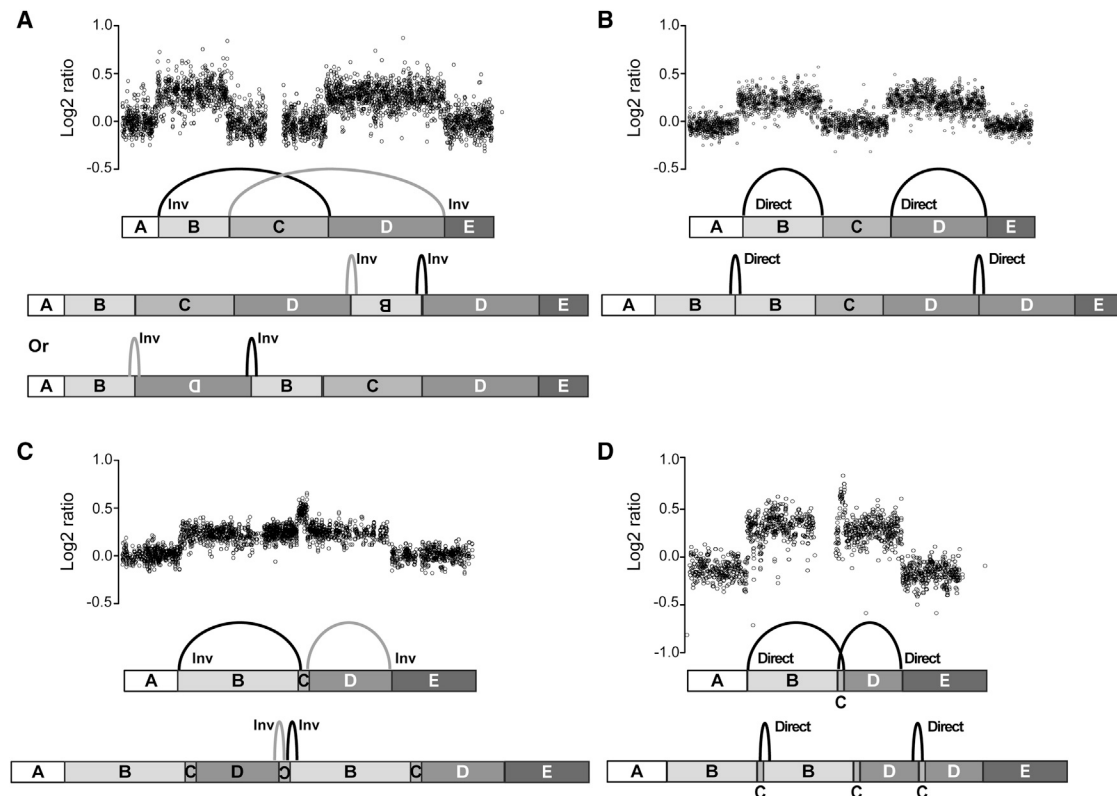


Figure 4. DUP-NML-DUP and DUP-TRP-DUP Organization

High-resolution array CGH reveals duplications and/or triplications in EGL515 (A), EGL559 (B), EGL688 (C), and EGL407 (D). Log₂ ratio of subject versus control signal intensity is shown on the y axis. Normal copy number, duplicated, and triplicated segments are labeled A–E for DUP-NML-DUP (A and B) and DUP-TRP-DUP (C and D) rearrangements. Gray arches connect sequenced junctions relative to the reference genome (above) and the rearrangement (below). Duplicated and triplicated segments can be inverted (Inv) or in direct orientation.

visible by clinical microarray, but EGL586's 82.8-kb and 65.2-kb duplications were originally detected as a single 412-kb duplication (Table S1). Sequencing breakpoint junctions from EGL515 and EGL605 revealed a common DUP-NML-DUP structure, where an inverted duplication is sandwiched in between another direct duplication. Both of the duplications in EGL559 were in direct orientation, and the two sequenced junctions did not link the duplications (Figure 4). This could indicate two independent direct duplications or a complex CNV that formed two nearby duplications as a single event. Because parents were not available for testing, we cannot distinguish these possibilities. EGL586 and EGL629 have DUP-NML-DUPs, with one direct and one inverted junction. We sequenced two junctions of EGL586's DUP-NML-DUP but could not interpret the structure because the breakpoints map outside of the duplicated segments. As for any CNV breakpoint study, there may be additional breakpoint junctions that we failed to sequence. This is probably the case for EGL414's duplication, which appears to be simple by array CGH but has a junction connecting one end of the duplication to a site within the duplication.

We identified ten triplications embedded within duplications and sequenced at least one junction from each of the DUP-TRP-DUPs. Duplications ranged from 557 kb to

5.43 Mb, and triplications were 26.0 kb to 4.97 Mb (Table S7). For six DUP-TRP-DUPs, we sequenced two breakpoint junctions and could infer the orientation of the triplication relative to the duplication. EGL600, EGL688, and EGL824 have inverted triplications, whereas EGL407, EGL543, and EGL544 have direct triplications (Figure 4). Triplications in EGL481, EGL501, EGL577, and EGL690 are supported by only one sequenced junction, so we cannot infer their complete genomic organization (Table S1).

Three duplications were inserted into different chromosomes. EGL526's 24.6-Mb duplication of chromosome 5q23.3–q33.2 inserted into chromosome 9q21.13. The insertion was confirmed by chromosome banding, and sequencing of both insertion junctions revealed that the duplicated segment was inserted in chromosome 9 in the same orientation as its original locus on chromosome 5. There is a 5.8-kb deletion of chromosome 9q21.13 at the insertion site; however, this deletion does not fall within a gene.

EGL483 has a 25.4-Mb duplication of chromosome 2p16.1–p12 and a 6.67-Mb duplication of chromosome 2q22.1–q22.2 inserted into chromosome 6. Though FISH confirmed that both duplications inserted into the long arm of chromosome 6, we did not sequence breakpoint

junctions that connected the two segments of chromosome 2 to chromosome 6. Instead, we captured and sequenced an inverted junction between the 2q22.1 duplication breakpoint and a region 1.67 Mb away. This insertional translocation probably has additional breakpoints consistent with an even more complex rearrangement. EGL483's mother carries a balanced form of the insertional translocation, where both regions of chromosome 2 are inserted in chromosome 6 and are missing from one chromosome 2.

We did not confirm EGL701's 522-kb insertion of Xq22.3 into 9q34.11 by FISH or chromosome banding. However, junction sequencing revealed an inverted insertion of Xq22.3 into 9q34.11. Breakpoints on chromosomes 9 and X lie in *USP20* (MIM 615143) and *COL4A6* (MIM 303631), respectively (Figure 5). EGL701 has one intact copy of *COL4A6* on his X chromosome, one intact copy of *USP20* on one chromosome 9, and disruption of *USP20* on the derivative chromosome 9 that carries the insertion. Based on the orientation of the genes and the inverted insertion of Xq22.3, this is predicted to result in an in-frame fusion of exons 1–2 of *COL4A6* and exons 4–26 of *USP20*.

Though we excluded obvious inverted duplications adjacent to terminal deletions from this study, breakpoint sequencing revealed that LM223's duplication was inverted and adjacent to a very small terminal deletion that was not visible by CMA testing. This type of rearrangement forms via a distinct mechanism involving a dicentric chromosome intermediate that breaks to give rise to a characteristic terminal deletion adjacent to inverted duplications separated by a short disomic spacer.³³ Another duplication that appeared to be terminal by microarray analysis turned out to be in direct orientation. WGS of EGL825's terminal duplication suggested an interchromosomal duplication between chromosomes 9 and 12 (Table S2); however, FISH confirmed that this is an intrachromosomal duplication of the short arm of chromosome 9. Because the distal breakpoint lies in subtelomeric segmental duplications, it is not surprising that the direct duplication junction mapped to a different chromosome.

Gene Fusions and Phenotypes

We analyzed genes at the 118 breakpoint junctions with contiguous sequence to infer effects of duplication breakpoints on gene structure (Table S1). 90 out of 118 breakpoints do not fuse genes in the same direction and are not predicted to generate fusion transcripts. Five sequenced duplications lie within a single intron and do not include splice sites. Intragenic duplications in EGL456 and EGL527 are predicted to result in out-of-frame transcripts of *CNTN4* (MIM 607280) and *TCOF1* (MIM 606847), respectively. EGL456 was referred for testing because of infantile cerebral palsy. *CNTN4* lies within the region deleted in 3p– syndrome (MIM 613792), and rearrangements involving *CNTN4* have been described in children with developmental delay, speech delay, or ASD.^{34–36}

EGL527's referring diagnosis of cleft palate is probably due to loss of function of *TCOF1*, which causes autosomal-dominant Treacher Collins syndrome (MIM 606847).

Intergenic duplications with genes spanning both breakpoints can generate fusion genes if the genes are in the same orientation and the reading frame is maintained. Six fusions are predicted to be in-frame and 15 are predicted to be out-of-frame (Table 1). EGL480's tandem duplication juxtaposes exons 1–6 of *SOS1* (MIM 182530) to exons 2–33 of *MAP4K3* (MIM 604921) in-frame (Figure 5). Gain-of-function missense mutations in *SOS1* cause Noonan syndrome.^{37,38} Although EGL480 does not have a formal diagnosis of Noonan syndrome, he does exhibit hypertelorism, seizures, and developmental delay that could be related to gain of function in the *SOS1-MAP4K3* fusion product. EGL605's DUP-NML-DUP fuses the *KCNH5* (MIM 605716) and *FUT8* (MIM 602589) genes and is predicted to be in-frame (Figure 5). A de novo missense variant in *KCNH5* has been reported in a child with epilepsy.³⁹ EGL605 was tested because she presented failure to thrive as an infant and we do not know whether she developed seizures later. EGL701 had a referring diagnosis of developmental delay, short stature, and multiple congenital anomalies that might not be related to the maternally inherited duplication that produces a putative *COL4A6-USP20* fusion at the chromosome 9 insertion site. The phenotypic consequences of the putative *TRPV3-TAXIBP3* (EGL413) and *LTBP1-BIRC6* (EGL415, EGL478) fusions are difficult to predict because these genes have not been implicated in neurodevelopmental disorders (MIM 607066, 150390, 605638).

Duplications with Common Breakpoints

Most of the duplication breakpoints we sequenced (108/118) were unique to a single individual in our cohort (Table S1). Identical breakpoint junctions in unrelated individuals are consistent with inherited CNVs present in the population rather than new duplication events. Such duplications might have no phenotypic consequences, or they could confer a subtle disease susceptibility risk. Despite their common origin, duplications with identical breakpoints might appear to be different CNVs due to variable array platforms in different studies. Thus, sequencing breakpoint junctions can consolidate common CNVs and clarify genotype-phenotype correlations. We describe seven duplication CNVs with common breakpoints confirmed by Sanger sequencing junctions.

EGL594, EGL595, EGL596, and EGL597 carry identical 1.3-Mb duplications of chromosome 12p11.1 as measured by high-resolution array CGH. Based on the gene content and abundant normal variation in this region, all four duplications were interpreted as likely benign; however, the large size met our criteria for clinical reporting. We sequenced four duplications with identical junctions. This pericentromeric duplication has been reported in databases of normal variation with slightly different breakpoints^{8,28,29,40} and is probably a benign CNV. The 704-kb

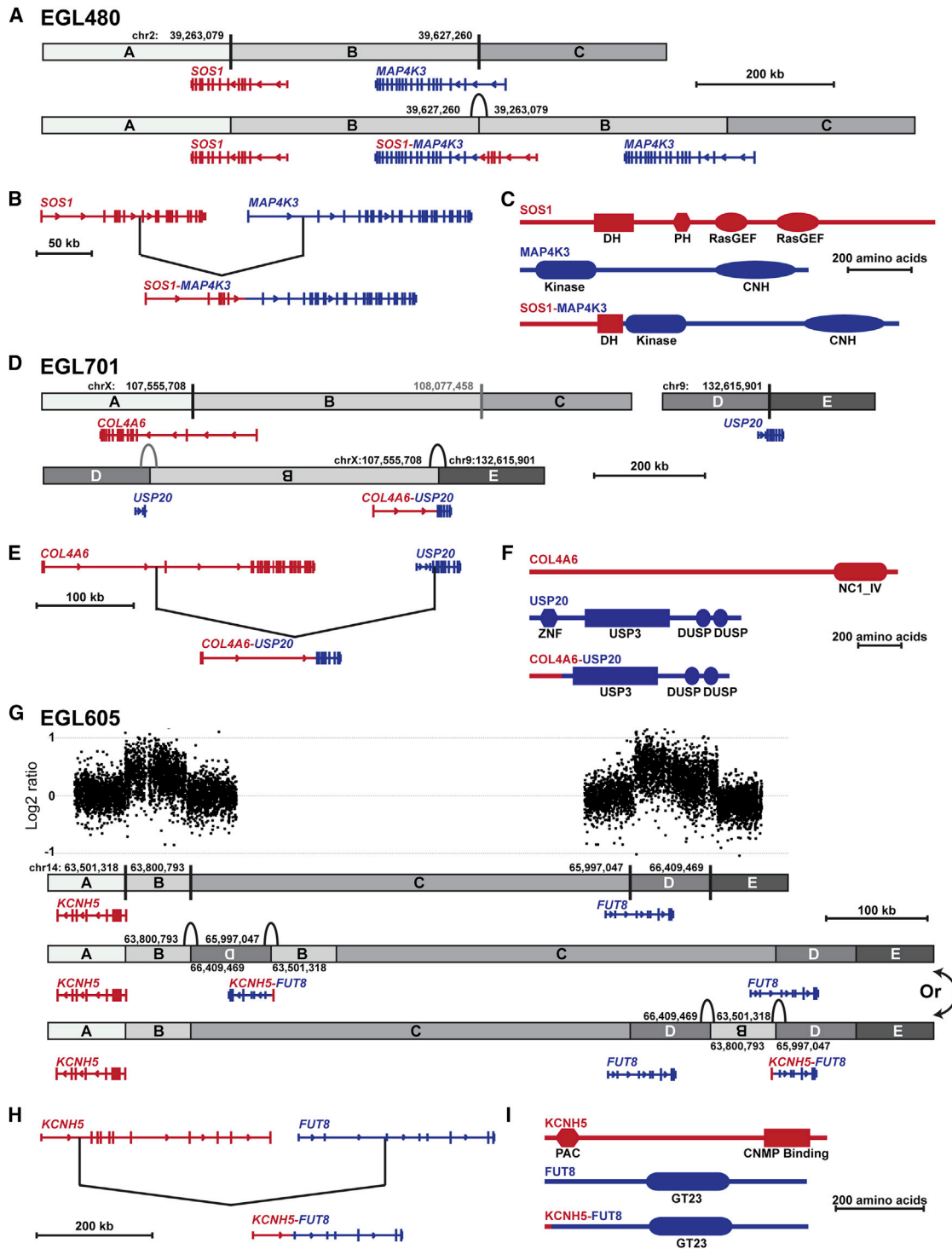


Figure 5. In-Frame Fusion Genes Predicted at Duplication Junctions

(A, D, and G) Genes that cross breakpoints are shown relative to the reference genome (above) and the duplication (below). The genomic coordinates of breakpoints have been confirmed by sequencing (black) or high-resolution array CGH (gray).
 (A) EGL480's direct duplication of chromosome 2p22.1.
 (B) The direct duplication fuses *SOS1* to *MAP4K3*.
 (C, F, and I) Domains of the fusion proteins in EGL480 (C), EGL701 (F), and EGL605 (I). We predicted fusion protein motifs by entering fusion cDNA sequence from Ensembl 75 into ScanProsite.
 (D) EGL701's duplication of the X chromosome is inverted and inserted into chromosome 9.
 (E) *COL4A6* is fused to *USP20* at the insertion site.
 (G) Array CGH (above) and breakpoint sequencing (below) of EGL605's DUP-NML-DUP. There are two possible structures for this rearrangement, and both predict a *KCNH5-FUT8* fusion.
 (H) *KCNH5* fuses to *FUT8* at the inverted junction of the two duplications.

Table 1. Predicted Fusion Genes at Duplication Breakpoints

Subject ID	Duplication Structure	Predicted Frame	Fusion Gene
EGL456	intragenic, direct	out of frame	<i>CNTN4</i>
EGL527	intragenic, direct	out of frame	<i>TCOF1</i>
EGL413	intergenic, direct	in frame	<i>TRPV3-TAX1BP3</i>
EGL415	intergenic, direct	in frame	<i>LTBP1-BIRC6</i>
EGL478	intergenic, direct	in frame	<i>LTBP1-BIRC6</i>
EGL480	intergenic, direct	in frame	<i>SOS1-MAP4K3</i>
EGL605	DUP-NML-DUP	in frame	<i>KCNH5-FUT8</i>
EGL701	insertional translocation	in frame	<i>COL4A6-USP20</i>
EGL403	intergenic, direct	out of frame	<i>ADD2-EXOC6B</i>
EGL408	intergenic, direct	out of frame	<i>H6ST2-GPC4</i>
EGL465	intergenic, direct	out of frame	<i>LPHN2-IFI44</i>
EGL473	intergenic, direct	out of frame	<i>SHDC-LY9</i>
EGL492	intergenic, direct	out of frame	<i>BARD1-FN1</i>
EGL500	intergenic, direct	out of frame	<i>RAF1-TMEM40</i>
EGL509	intergenic, direct	out of frame	<i>WHSC1-FGFR3</i>
EGL542	intergenic, direct	out of frame	<i>CACNA2D1-PCLO</i>
EGL572	intergenic, direct	out of frame	<i>LMX1B-MVB12B</i>
EGL582	intergenic, direct	out of frame	<i>TEAD1-MICAL2</i>
EGL598	intergenic, direct	out of frame	<i>PDZRN4-CNTN1</i>
EGL617	intergenic, direct	out of frame	<i>TRAP1-UBLAD1</i>
EGL668	intergenic, direct	out of frame	<i>PNPLA4-KAL1</i>
EGL683	intergenic, direct	out of frame	<i>TAB3-DMD</i>
EGL692	intergenic, direct	out of frame	<i>XIST-FTX</i>

duplications of chromosome 2p22.3 in EGL415 and EGL478 had the same breakpoint junctions and were both inherited from parents. EGL460 and EGL461 have identical 582-kb duplications of chromosome 1p36.32 that have one breakpoint in the PR domain-containing protein 16 (*PRDM16*) gene. Heterozygous deletions and mutations in *PRDM16* have been described in individuals with left ventricular noncompaction or cardiomyopathy (MIM 605557).⁴¹ Because this duplication is in direct orientation, it does not disrupt *PRDM16*. Further, neither of our subjects had a referring diagnosis of heart disease. The 668-kb duplication of chromosome 12p12.1 is identical in EGL408 and EGL592 and similar to duplication CNVs in control databases.^{8,29}

The duplications of chromosome 21q22.11–q22.12 in EGL653 and EGL655 have identical breakpoints. *KCNE1* (MIM 176261) and *KCNE2* (MIM 603796) lie within the duplicated region, and mutations in both genes have been associated with long QT syndrome. EGL653 has left pulmonary arterial atresia, and EGL655 has an atrial septal defect. Duplication of this region, including *KCNE1* and

KCNE2, has not been reported in cohorts of children with congenital heart disease^{42,43} but could be a risk factor.

EGL627 and EGL823 have duplications of part of intron 1 of *PFAFH1B1*, also known as *LIS1* (MIM 601545). Sanger sequencing confirmed that the entire 32.8-kb tandem duplication lies within intron 1. These duplications were interpreted as being of uncertain clinical significance because mutations in *PFAFH1B1* cause autosomal-dominant lissencephaly. The two unrelated individuals share the same duplication, and one was inherited from an unaffected mother, so it is unlikely that this finding is related to their clinical features. The indications for testing were seizures (EGL627) and neurological disorder, newborn apnea, and feeding difficulties (EGL823).

EGL543 and EGL544 have identical DUP-TRP-DUPS of chromosome 7q21.12. This appeared to be a simple 1.5-Mb duplication by clinical array CGH, but fine mapping and sequencing revealed a DUP-TRP-DUP structure (Table S7). Similar CNVs have been reported in databases of normal variation, suggesting that this finding is not related to the clinical presentations of EGL543 and EGL544.^{8,28,29}

Discussion

Chromosome duplications can cause phenotypes through loss of function, gain of function, triplosensitivity, and/or misregulation of genes within or near the duplicated region. Whereas deletions have a straightforward genomic structure, duplications can have very different effects on gene function depending on the duplication breakpoints, duplication location, and gene reading frame. Our large-scale study of chromosome duplications revealed that most interstitial duplications are tandem and in direct orientation relative to the original locus. The only inverted duplications are those that are part of more complex rearrangements, including insertional translocations, inverted duplications adjacent to terminal deletions, DUP-NML-DUPS, and DUP-TRP-DUPS.

Triplications embedded within duplications have been described at a number of loci.^{30,44–48} In most cases, the triplicated segment is inverted relative to the tandem duplication, and this conformation is known as DUP-TRP/INV-DUP. Breakpoint analysis of six DUP-TRP-DUPS in our study revealed that half of the triplications are in direct orientation and half are inverted relative to the duplication (Table S7). As shown by high-resolution CGH and breakpoint sequencing, all ten of the triplications in our study lie within larger duplications. This is characteristic of the type II triplication structure, whereas type I triplications are made up of head-to-tail triplicated copies separated by segmental duplications.⁴⁹ Some type II triplications are flanked by inverted repeats; however, we did not detect this feature at any of the DUP-TRP-DUP boundaries in our study. Triplications derived from regions rich in segmental duplications might be more likely to be

mediated by homology between inverted repeats.⁴⁴ Though some of our DUP-TRP-DUP breakpoints lie in genes, none are predicted to form fusion transcripts.

DUP-NML-DUPs might also exist in direct or inverted orientation. Breakpoint sequencing revealed that duplications in some DUP-NML-DUPs are connected. On the other hand, duplications of different chromosome arms are not usually part of the same CNV. Duplicated segments were connected by inverted junctions in DUP-NML-DUPs from EGL515 and EGL605, whereas direct junctions in EGL559 did not connect the nearby duplications. Similar DUP-NML-DUPs have been described at the *MECP2* locus³¹ and other sites.^{30,32} The in-frame fusion of *KCNH5* and *FUT8* was not recognized until we sequenced EGL605's DUP-NML-DUP.

Insertional translocations make up 2.8% of the sequenced duplications in our study. Other groups have also found ~2% of clinically relevant duplications to be inserted in other loci.^{50–52} In two out of three insertional translocations, we performed FISH to identify the location of the duplicated material, and in one case the insertion was detected only by breakpoint sequencing. EGL701 inherited this duplication of Xq22.22 from his mother, and based on CMA we assumed that it was tandem. Instead, the duplication is inserted into chromosome 9 and produces a putative *COL4A6-USB20* fusion at the insertion site. This fusion is predicted to be in-frame and could create a unique fusion protein.

Intragenic duplications can disrupt gene reading frames, leading to loss-of-function mutations. Breakpoint analysis of direct duplications in EGL425, EGL588, EGL627, EGL684, and EGL823 confirmed that these duplications lie within a single intron. Though these duplications are not predicted to disrupt the reading frame, in some cases intronic insertions can affect splicing of flanking exons.^{53,54} On the other hand, intragenic duplications in EGL456 and EGL527 are predicted to result in out-of-frame transcripts in *CNTN4* and *TCOF1*, respectively. Genes at sequenced intergenic duplication breakpoints are predicted to generate 6 in-frame fusions and 15 out-of-frame fusions (Table 1). Breakpoints that fuse genes with the same exon phase can create unique in-frame fusion genes (Figure 5). For example, the direct duplication in EGL480 can produce a fusion of *SOS1* and *MAP4K3*. Structural rearrangements that fuse kinase genes are an important class of oncogenes in leukemia and solid tumors.⁵⁵ It is tempting to speculate that the germline *SOS1-MAP4K3* fusion gene also plays a role in EGL480's clinical presentation. In addition, transcripts that we predict to be out-of-frame might produce proteins by alternative splicing using cryptic splice donor and/or acceptor sites. Future mRNA and protein studies are necessary to determine the functional consequences of genes fused at duplication breakpoints.

Analysis of breakpoint junctions can shed light on CNV mechanisms. Most sequenced breakpoints (67%) had short microhomology between the two sides of the duplication, and 26% had short insertions at the breakpoint junctions.

These junction signatures are consistent with nonhomologous end-joining (NHEJ) or microhomology-mediated break-induced replication (MMBIR).⁵⁶ Similar junctions have been described at tandem duplications of *MECP2*,⁵⁷ *LMNB1*,¹² *PLP1*,¹⁴ *HUWE1*⁵⁸ (MIM 300697), and other loci. Regions that are enriched in paralogous segmental duplications or interspersed repeats give rise to more duplications via NAHR.^{49,59} We detected three duplications flanked by pairs of *Alu* repeats that are 75%–88% identical and that generate a hybrid *Alu* at the breakpoint junction (Table S1). Similar homology has been described for other *Alu-Alu* recombination events that give rise to interstitial deletions and duplications.^{19,60,61}

Almost all duplication CNVs in our study were inherited from a parent. For 69 out of 74 (93%) trios tested, the CNV was inherited, but because most parents have not been assessed clinically, we cannot determine the penetrance or expressivity of the duplication CNV. In general, duplication CNVs are less penetrant than deletion CNVs.^{2,7,62} Though de novo CNVs are more likely to be disease related, de novo duplications in our study were not particularly large or complex. Two of the five de novo CNVs in our study were complex (EGL501, EGL824), and the other three were direct duplications 900 kb–1.2 Mb in size (EGL617, EGL662, EGL825). Some of the largest duplications were inherited from parents (e.g., EGL568, 8.4 Mb; EGL641, 11.0 Mb), so duplication size does not correlate with those that are de novo. In addition, duplication breakpoints did not change from parent to offspring. We sequenced 17 breakpoints from family members with the same duplication, and all the junction sequences were conserved (Table S1).

The clinical significance of duplication CNVs is difficult to interpret. Genomic gains detected by diagnostic CMA testing might represent a number of different chromosome rearrangements that vary in pathogenicity and recurrence risk. Inverted duplications adjacent to terminal deletions have a characteristic appearance via microarray analysis and in almost all cases occur de novo.^{33,63} Terminal gains are most often unbalanced translocations, but in rare cases might be inverted or direct intrachromosomal duplications. Chromosomes with a terminal duplication of one end and a terminal deletion of the other end can be generated by recombination within an inversion loop. Because parents of children with unbalanced translocations and recombinant inversion chromosomes might carry balanced forms of the rearrangements, their recurrence risk for another child with a chromosome rearrangement is significant. Interstitial duplications are often inherited from parents, so predicting outcomes for future pregnancies is complicated by incomplete penetrance and variable expressivity. Unlike recurrent duplications, those in our study are too rare to compare the phenotypes of multiple individuals.

Most interstitial duplications are tandem and lie in direct orientation. More complex DUP-NML-DUP, DUP-TRP-DUP, and insertional translocation CNVs can be detected

by clinical CMA and FISH, but without sequencing breakpoints it is impossible to determine the orientation of duplication segments that could disrupt or fuse genes. Furthermore, these complex duplications were interpreted as pathogenic ($n = 5$) or uncertain clinical significance ($n = 14$) and were either de novo ($n = 2$), maternal ($n = 6$), paternal ($n = 2$), or of unknown origin ($n = 9$). Thus, even duplications with recognized complexity can be difficult to interpret. As NGS and WGS become routine for copy-number analysis, it will be possible to capture CNV and breakpoint junction data at the same time.^{64–69} These breakpoint analyses, as well as future RNA and protein studies, are essential to determine the functional consequences of duplication CNVs.

Accession Numbers

Microarray data are deposited in the NCBI Gene Expression Omnibus under accession number GSE62657. Breakpoint junction sequences have been submitted to GenBank under BankIt1750132 with accession numbers KP007212–KP007329. NGS data were submitted to the Sequence Read Archive (SRA) under accession number PRJNA264978. WGS data are available through the database of Genotypes and Phenotypes (dbGaP) under accession number phs000845.v1.p1.

Supplemental Data

Supplemental Data include seven tables and can be found with this article online at <http://dx.doi.org/10.1016/j.ajhg.2014.12.017>.

Acknowledgments

We thank Madhuri Hegde and Arun Ankala for scientific discussions on duplication formation. Kelly Shaw, Michael Christopher, and Alev Cagla Ozdemir performed breakpoint junction experiments. We thank Cheryl Strauss for editorial assistance. This study was supported by a grant from the NIH (MH092902 to M.K.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Received: October 31, 2014

Accepted: December 15, 2014

Published: January 29, 2015

Web Resources

The URLs for data presented herein are as follows:

Agilent eArray, <https://earray.chem.agilent.com>

Breakpoint Simulator, <http://sourceforge.net/projects/breakpoint-simulator/>

Database of Genomic Variants (DGV), <http://dgv.tcag.ca/dgv/app/home>

dbGaP, <http://www.ncbi.nlm.nih.gov/gap>

Ensembl Genome Browser, <http://www.ensembl.org/index.html>

GenBank, <http://www.ncbi.nlm.nih.gov/genbank/>

Gene Expression Omnibus (GEO), <http://www.ncbi.nlm.nih.gov/geo/>

OMIM, <http://www.omim.org/>

Primer3, <http://bioinfo.ut.ee/primer3-0.4.0/primer3/>

ScanProsite, <http://prosite.expasy.org/scanprosite/>

Sequence Read Archive (SRA), <http://www.ncbi.nlm.nih.gov/sra>

UCSC Genome Browser, <http://genome.ucsc.edu>

References

1. Neill, N.J., Torchia, B.S., Bejjani, B.A., Shaffer, L.G., and Ballif, B.C. (2010). Comparative analysis of copy number detection by whole-genome BAC and oligonucleotide array CGH. *Mol. Cytogenet.* 3, 11.
2. Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. *Nat. Genet.* 43, 838–846.
3. Kearney, H.M., Thorland, E.C., Brown, K.K., Quintero-Rivera, F., and South, S.T.; Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* 13, 680–685.
4. Watson, C.T., Marques-Bonet, T., Sharp, A.J., and Mefford, H.C. (2014). The genetics of microdeletion and microduplication syndromes: an update. *Annu. Rev. Genomics Hum. Genet.* 15, 215–244.
5. Cook, E.H., Jr., and Scherer, S.W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455, 919–923.
6. Deak, K.L., Horn, S.R., and Rehder, C.W. (2011). The evolving picture of microdeletion/microduplication syndromes in the age of microarray analysis: variable expressivity and genomic complexity. *Clin. Lab. Med.* 31, 543–564, viii.
7. Rosenfeld, J.A., Coe, B.P., Eichler, E.E., Cuckle, H., and Shaffer, L.G. (2013). Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet. Med.* 15, 478–481.
8. Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., et al. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* 84, 148–161.
9. Rudd, M.K., Keene, J., Bunke, B., Kaminsky, E.B., Adam, M.P., Mulle, J.G., Ledbetter, D.H., and Martin, C.L. (2009). Segmental duplications mediate novel, clinically relevant chromosome rearrangements. *Hum. Mol. Genet.* 18, 2957–2962.
10. Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 143, 837–847.
11. Thienpont, B., Béna, F., Breckpot, J., Philip, N., Menten, B., Van Esch, H., Scalais, E., Salamone, J.M., Fong, C.T., Kussmann, J.L., et al. (2010). Duplications of the critical Rubinstein-Taybi deletion region on chromosome 16p13.3 cause a novel recognizable syndrome. *J. Med. Genet.* 47, 155–161.
12. Giorgio, E., Rolyan, H., Kropp, L., Chakka, A.B., Yatsenko, S., Di Gregorio, E., Lacerenza, D., Vaula, G., Talarico, F., Mandich, P., et al. (2013). Analysis of LMNB1 duplications in autosomal dominant leukodystrophy provides insights into duplication mechanisms and allele-specific expression. *Hum. Mutat.* 34, 1160–1171.

13. Van Esch, H., Bauters, M., Ignatius, J., Jansen, M., Raynaud, M., Hollanders, K., Lugtenberg, D., Bienvenu, T., Jensen, L.R., Geç, J., et al. (2005). Duplication of the MECP2 region is a frequent cause of severe mental retardation and progressive neurological symptoms in males. *Am. J. Hum. Genet.* *77*, 442–453.
14. Woodward, K.J., Cundall, M., Sperle, K., Sistermans, E.A., Ross, M., Howell, G., Gribble, S.M., Burford, D.C., Carter, N.P., Hobson, D.L., et al. (2005). Heterogeneous duplications in patients with Pelizaeus-Merzbacher disease suggest a mechanism of coupled homologous and nonhomologous recombination. *Am. J. Hum. Genet.* *77*, 966–987.
15. Arlt, M.F., Mülle, J.G., Schaibley, V.M., Ragland, R.L., Durkin, S.G., Warren, S.T., and Glover, T.W. (2009). Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am. J. Hum. Genet.* *84*, 339–350.
16. Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J., and Hurler, M.E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* *42*, 385–391.
17. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.; 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* *470*, 59–65.
18. Baldwin, E.L., Lee, J.Y., Blake, D.M., Bunke, B.P., Alexander, C.R., Kogan, A.L., Ledbetter, D.H., and Martin, C.L. (2008). Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray. *Genet. Med.* *10*, 415–429.
19. Luo, Y., Hermetz, K.E., Jackson, J.M., Mülle, J.G., Dodd, A., Tsuchiya, K.D., Ballif, B.C., Shaffer, L.G., Cody, J.D., Ledbetter, D.H., et al. (2011). Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. *Hum. Mol. Genet.* *20*, 3769–3778.
20. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* *25*, 1754–1760.
21. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* *25*, 2078–2079.
22. Ng, C.K., Cooke, S.L., Howe, K., Newman, S., Xian, J., Temple, J., Batty, E.M., Pole, J.C., Langdon, S.P., Edwards, P.A., and Brenton, J.D. (2012). The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J. Pathol.* *226*, 703–712.
23. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nat. Biotechnol.* *29*, 24–26.
24. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* *327*, 78–81.
25. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* *26*, 841–842.
26. Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* *12*, 656–664.
27. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* *42*, D749–D755.
28. Shaikh, T.H., Gai, X., Perin, J.C., Glessner, J.T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L.K., D'Arcy, M., et al. (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.* *19*, 1682–1690.
29. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* *42*, D986–D992.
30. Liu, P., Erez, A., Nagamani, S.C., Dhar, S.U., Kołodziejaska, K.E., Dharmadhikari, A.V., Cooper, M.L., Wiszniewska, J., Zhang, F., Withers, M.A., et al. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* *146*, 889–903.
31. Carvalho, C.M., Pehlivan, D., Ramocki, M.B., Fang, P., Alleva, B., Franco, L.M., Belmont, J.W., Hastings, P.J., and Lupski, J.R. (2013). Replicative mechanisms for CNV formation are error prone. *Nat. Genet.* *45*, 1319–1326.
32. Brand, H., Pillalamarri, V., Collins, R.L., Eggert, S., O'Dushlaine, C., Braaten, E.B., Stone, M.R., Chambert, K., Doty, N.D., Hanscom, C., et al. (2014). Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am. J. Hum. Genet.* *95*, 454–461.
33. Hermetz, K.E., Newman, S., Conneely, K.N., Martin, C.L., Ballif, B.C., Shaffer, L.G., Cody, J.D., and Rudd, M.K. (2014). Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet.* *10*, e1004139.
34. Fernandez, T., Morgan, T., Davis, N., Klin, A., Morris, A., Farhi, A., Lifton, R.P., and State, M.W. (2004). Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. *Am. J. Hum. Genet.* *74*, 1286–1293.
35. Roohi, J., Montagna, C., Tegay, D.H., Palmer, L.E., DeVincent, C., Pomeroy, J.C., Christian, S.L., Nowak, N., and Hatchwell, E. (2009). Disruption of contactin 4 in three subjects with autism spectrum disorder. *J. Med. Genet.* *46*, 176–182.
36. Cottrell, C.E., Bir, N., Varga, E., Alvarez, C.E., Bouyain, S., Zernzach, R., Thrush, D.L., Evans, J., Trimarchi, M., Butter, E.M., et al. (2011). Contactin 4 as an autism susceptibility locus. *Autism Res.* *4*, 189–199.
37. Tartaglia, M., Pennacchio, L.A., Zhao, C., Yadav, K.K., Fodale, V., Sarkozy, A., Pandit, B., Oishi, K., Martinelli, S., Schackwitz, W., et al. (2007). Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nat. Genet.* *39*, 75–79.
38. Zenker, M., Horn, D., Wiczorek, D., Allanson, J., Pauli, S., van der Burgt, I., Doerr, H.G., Gaspar, H., Hofbeck, M., Gillissen-Kaesbach, G., et al. (2007). SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *J. Med. Genet.* *44*, 651–656.
39. Veeramah, K.R., Johnstone, L., Karafet, T.M., Wolf, D., Sprissler, R., Salogiannis, J., Barth-Maron, A., Greenberg, M.E., Stuhlmann, T., Weinert, S., et al. (2013). Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* *54*, 1270–1281.
40. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shaperro, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature* *444*, 444–454.

41. Arndt, A.K., Schafer, S., Drenckhahn, J.D., Sabeh, M.K., Plovie, E.R., Caliebe, A., Klopocki, E., Musso, G., Werdich, A.A., Kalwa, H., et al. (2013). Fine mapping of the 1p36 deletion syndrome identifies mutation of PRDM16 as a cause of cardiomyopathy. *Am. J. Hum. Genet.* *93*, 67–77.
42. Soemedi, R., Wilson, I.J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., et al. (2012). Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* *91*, 489–501.
43. Warburton, D., Ronemus, M., Kline, J., Jobanputra, V., Williams, I., Anyane-Yeboah, K., Chung, W., Yu, L., Wong, N., Awad, D., et al. (2014). The contribution of de novo and rare inherited copy number changes to congenital heart disease in an unselected sample of children with conotruncal defects or hypoplastic left heart disease. *Hum. Genet.* *133*, 11–27.
44. Carvalho, C.M., Ramocki, M.B., Pehlivan, D., Franco, L.M., Gonzaga-Jauregui, C., Fang, P., McCall, A., Pivnick, E.K., Hines-Dowell, S., Seaver, L.H., et al. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* *43*, 1074–1081.
45. Giorda, R., Beri, S., Bonaglia, M.C., Spaccini, L., Scelsa, B., Manolagos, E., Della Mina, E., Ciccone, R., and Zuffardi, O. (2011). Common structural features characterize interstitial intrachromosomal Xp and 18q triplications. *Am. J. Med. Genet. A.* *155A*, 2681–2687.
46. Shimojima, K., Mano, T., Kashiwagi, M., Tanabe, T., Sugawara, M., Okamoto, N., Arai, H., and Yamamoto, T. (2012). Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. *Eur. J. Med. Genet.* *55*, 400–403.
47. Fujita, A., Suzumura, H., Nakashima, M., Tsurusaki, Y., Saitsu, H., Harada, N., Matsumoto, N., and Miyake, N. (2013). A unique case of de novo 5q33.3-q34 triplication with uniparental isodisomy of 5q34-qter. *Am. J. Med. Genet. A.* *161A*, 1904–1909.
48. Soler-Alfonso, C., Carvalho, C.M., Ge, J., Roney, E.K., Bader, P.I., Kolodziejaska, K.E., Miller, R.M., Lupski, J.R., Stankiewicz, P., Cheung, S.W., et al. (2014). CHRNA7 triplication associated with cognitive impairment and neuropsychiatric phenotypes in a three-generation pedigree. *Eur. J. Hum. Genet.* *22*, 1071–1076.
49. Liu, P., Carvalho, C.M., Hastings, P.J., and Lupski, J.R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.* *22*, 211–220.
50. Kang, S.H., Shaw, C., Ou, Z., Eng, P.A., Cooper, M.L., Pursley, A.N., Sahoo, T., Bacino, C.A., Chinault, A.C., Stankiewicz, P., et al. (2010). Insertional translocation detected using FISH confirmation of array-comparative genomic hybridization (aCGH) results. *Am. J. Med. Genet. A.* *152A*, 1111–1126.
51. Neill, N.J., Ballif, B.C., Lamb, A.N., Parikh, S., Ravnan, J.B., Schultz, R.A., Torchia, B.S., Rosenfeld, J.A., and Shaffer, L.G. (2011). Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. *Genome Res.* *21*, 535–544.
52. Nowakowska, B.A., de Leeuw, N., Ruivenkamp, C.A., Sikkema-Raddatz, B., Crolla, J.A., Thoelen, R., Koopmans, M., den Hollander, N., van Haeringen, A., van der Kevie-Kersemaekers, A.M., 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.
53. Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E.Y., and Ast, G. (2008). Intronic Alu influence alternative splicing. *PLoS Genet.* *4*, e1000204.
54. Hellsten, U., Aspden, J.L., Rio, D.C., and Rokhsar, D.S. (2011). A segmental genomic duplication generates a functional intron. *Nat. Commun.* *2*, 454.
55. Medves, S., and Demoulin, J.B. (2012). Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies. *J. Cell. Mol. Med.* *16*, 237–248.
56. Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.* *10*, 551–564.
57. Carvalho, C.M., Zhang, F., Liu, P., Patel, A., Sahoo, T., Bacino, C.A., Shaw, C., Peacock, S., Pursley, A., Tavyev, Y.J., 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.
58. Froyen, G., Belet, S., Martinez, F., Santos-Rebouças, C.B., Declercq, M., Verbeeck, J., Donckers, L., Berland, S., Mayo, S., Rosello, M., et al. (2012). Copy-number gains of HUWE1 due to replication- and recombination-based rearrangements. *Am. J. Hum. Genet.* *91*, 252–264.
59. Giorda, R., Bonaglia, M.C., Beri, S., Fichera, M., Novara, F., Magini, P., Urquhart, J., Sharkey, F.H., Zucca, C., Grasso, R., et al. (2009). Complex segmental duplications mediate a recurrent dup(X)(p11.22-p11.23) associated with mental retardation, speech delay, and EEG anomalies in males and females. *Am. J. Hum. Genet.* *85*, 394–400.
60. Bose, P., Hermetz, K.E., Conneely, K.N., and Rudd, M.K. (2014). Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. *PLoS ONE* *9*, e101607.
61. Boone, P.M., Yuan, B., Campbell, I.M., Scull, J.C., Withers, M.A., Baggett, B.C., Beck, C.R., Shaw, C.J., Stankiewicz, P., Moretti, P., et al. (2014). The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am. J. Hum. Genet.* *95*, 143–161.
62. Kaminsky, E.B., Kaul, V., Paschall, J., Church, D.M., Bunke, B., Kunig, D., Moreno-De-Luca, D., Moreno-De-Luca, A., Mülle, J.G., Warren, S.T., 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.
63. Rudd, M.K. (2011). Structural variation in subtelomeres. In *Genomic Structural Variants: Methods and Protocols*, L. Feuk, ed. (New York: Springer Science+Business Media).
64. Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.M., Lee, E., Zhang, J., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Gibbs, R.A., et al. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. USA* *108*, E1128–E1136.
65. Michaelson, J.J., and Sebat, J. (2012). forestSV: structural variant discovery through statistical learning. *Nat. Methods* *9*, 819–821.
66. Krumm, N., Sudmant, P.H., Ko, A., O’Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., and Eichler, E.E.; NHLBI Exome Sequencing Project (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* *22*, 1525–1532.
67. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O’Donovan,

- M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* *91*, 597–607.
68. Poultney, C.S., Goldberg, A.P., Drapeau, E., Kou, Y., Harony-Nicolas, H., Kajiwara, Y., De Rubeis, S., Durand, S., Stevens, C., Rehnström, K., et al. (2013). Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.* *93*, 607–619.
69. Fromer, M., and Purcell, S.M. (2014). UsingXHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet.* *81*, 1, 21.