Published in final edited form as:
Hum Mutat. 2018 December ; 39(12): 1916–1925. doi:10.1002/humu.23608.

LINE-and *Alu*-containing genomic instability hotspotat 16q24.1 associated with recurrent and nonrecurrent CNV deletions causative for ACDMPV

Przemyslaw Szafranski[1], **Ewelina Kośmider**[1,2], **Qian Liu**[1], **Justyna A. Karolak**[1,3], **Lauren Currie**[4], **Sandhya Parkash**[4], **Stephen G. Kahler**[5], **Elizabeth Roeder**[1,6], **Rebecca O. Littlejohn**[6], **Thomas S. DeNapoli**[7], **Felix R. Shardonofsky**[8], **Cody Henderson**[6,9], **George Powers**[6,9], **Virginie Poisson**[10], **Denis Bérubé**[10], **Luc Oligny**[10], **Jacques L. Michaud**[10], **Sandra Janssens**[11], **Kris De Coen**[12], **Jo Van Dorpe**[13], **Annelies Dheedene**[11], **Matthew T. Harting**[14], **Matthew D. Weaver**[14], **Amir M. Khan**[14], **Nina Tatevian**[14], **Jennifer Wambach**[15], **Kathleen A. Gibbs**[16], **Edwina Popek**[17], **Anna Gambin**[2], and **Paweł Stankiewicz**[1]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland [3]Department of Genetics and Pharmaceutical Microbiology, Poznan University of Medical Sciences, 60-781 Poznan, Poland [4]Maritime Medical Genetics Service, IWK Health Centre, Halifax, Nova Scotia, NS B3K 6R8, Canada [5]Section of Genetics and Metabolism, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, USA [6]Department of Pediatrics, Baylor College of Medicine, San Antonio, Texas 78207, USA [7]Department of Pathology, Children's Hospital of San Antonio, San Antonio, Texas, 78207, USA [8]Pediatric Pulmonary Center, Children's Hospital of San Antonio, San Antonio, Texas, 78207, USA [9]Neonatal-Perinatal Medicine, Children's Hospital of San Antonio, San Antonio, Texas, 78207, USA [10]CHU Sainte-Justine, Montreal, Québec, QC H3T 1C5, Canada [11]Center for Medical Genetics, Ghent University, Ghent, Belgium [12]Department of Neonatal Intensive Care, Ghent University, Ghent, Belgium [13]Department of Pathology, Ghent University, Ghent, Belgium [14]McGovern Medical School at UTHealth, Houston, Texas 77030, USA [15]Edward Mallinckrodt Department of Pediatrics, Washington University School of Medicine, St. Louis, Missouri 63110, USA [16]Children's Hospital of Philadelphia, and University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA [17]Department of Pathology and Immunology, Baylor College of Medicine, Houston, Texas 77030, USA

Abstract

Correspondence: Paweł Stankiewicz, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA., pawels@bcm.edu.

SUPPORTING INFORMATION
Additional Supporting Information can be found online in the supporting information tab for this article.

DISCLOSURE STATEMENT
The authors declare no conflict of interest.

DATA DEPOSITION
CNV deletions associated with ACDMPV were submitted to dbVar (http://ncbi.nlm.nih.gov/dbvar): dbvar - ticket #28045–259747.

Transposable elements modify human genome by inserting into new loci or by mediating homology-, microhomology-, or homeology-driven DNA recombination or repair, resulting in genomic structural variation. Alveolar capillary dysplasia with misalignment of pulmonary veins (ACDMPV) is a rare, lethal, neonatal developmental lung disorder caused by point mutations or copy-number variant (CNV) deletions of *FOXF1* or its distant tissue-specific enhancer. Eighty five per cent of 45 ACDMPV-causative CNV deletions, of which junctions have been sequenced, had at least one of their two breakpoints located in a retrotransposon, with more than half of them being *Alu* elements. We describe a novel ~35 kb-large genomic instability hotspot at 16q24.1, involving two evolutionarily young LINE-1(L1) elements, L1PA2 and L1PA3, flanking *Alu*Y, two *Alu*Sx, *Alu*Sx1, and *Alu*Jr elements. The occurrence of L1s at this location coincided with the branching out of the *Homo-Pan-Gorilla* clade, and was preceded by the insertion of *Alu*Sx, *Alu*Sx1, and *Alu*Jr. Our data show that, in addition to mediating recurrent CNVs, L1 and *Alu* retrotransposons can predispose the human genome to formation of variably sized CNVs, both of clinical and evolutionary relevance. Nonetheless, epigenetic or other genomic features of this locus might also contribute to its increased instability.

Keywords

DNA repair; genome instability; nonrecurrent structural variants

1 | INTRODUCTION

Approximately 45% of the human genome is comprised of transposable elements (TEs), a small fraction of which is still capable of undergoing transposition in both germline and somatic cells (Beck et al., 2011; Boissinot & Sookdeo 2016; deKoning et al., 2011; Furano 2000; Helman et al., 2014; Ivancevic et al., 2016; Jurka 2000; Lander et al., 2001; Lee et al., 2012). The presence of TEs has profound implications as they contribute to genome evolution, and disease (Beck et al., 2010; Callinan & Batzer 2006; Gogvadze & Buzdin 2009; Hancks & Kazazian 2016; Iskow et al., 2010; Kazazian & Moran 2017; Richardson et al., 2015).

In addition to insertional mutagenesis and non-pathogenic intra-and inter-individual variation, mobile elements can act as substrates for homology-driven rearrangements. Similar to low-copy repeats or segmental duplications, LINE-1 (L1) and endogenous retroviral (ERV) elements can predispose the genome to copy-number variant (CNV) deletions and reciprocal duplications via nonallelic homologous recombination (NAHR) (Belancion et al., 2009; Boone et al., 2014; Burwinkel & Kilimann 1998; Campbell et al., 2014; Gilbert et al., 2005; Hedges & Deininger 2007; Hehir-Kwa et al., 2016; Higashimoto et al., 2013; Kohmoto et al., 2017; Lupski 2010; Quadri et al., 2015; Startek et al., 2015; Szafranski et al., 2016; Temtamy et al., 2008; Vissers et al., 2009). Other rearrangements mediated by L1s and ERVs include translocations (Buysse et al., 2008; Robberecht et al., 2013), insertions (Gu et al., 2016), inversions (Kidd et al., 2010), and complex genomic rearrangements (Gu et al., 2015; Liu et al., 2011).L1s, ERVs, and *Alu* elements also predispose the genome to structural variants via DNA break repair-or replication-associated processes (Carvalho & Lupski 2016). Due to the high copy-number of retrotransposons, CNVs mediated by them remain challenging for detection using chromosomal microarray analysis or next generation sequencing that rely on sequence uniqueness to identify assay results by specific genomic coordinates (Hehir-Kwa et al., 2016; Thung et al., 2014).

Recently, we have compiled 49 CNV deletions in the *FOXF1* locus causative for alveolar capillary dysplasia with misalignment of pulmonary veins (ACDMPV, MIM# 265380) (Szafranski et al., 2016). ACDMPV is a lethal neonatal lung developmental disorder characterized by severe respiratory failure and refractory pulmonary hypertension (Bishop et al., 2011; Langston 1991). The vast majority of patients with ACDMPV had point mutations or CNV deletions in *FOXF1* or its distant upstream enhancer on 16q24.1 (Stankiewicz et al., 2009; Sen et al. 2013; Szafranski et al. 2013; Szafranski et al., 2016). Interestingly, over three-fourths of the ACDMPV causative deletions, for which breakpoints were sequenced, involved retrotransposons; in 30% of those cases, L1 was present at least at one of the CNV two breakpoints, and half of the deletions were *Alu*-mediated.

Here we describe a novel genomic instability hotspot at 16q24.1, featuring L1 and *Alu* elements located at the distal edge of the *FOXF1* enhancer region, and show that itis involved in formation of same-and variably-sized pathogenic and benign CNVs.

2 | METHODS

2.1 | Human subjects

ACDMPV patients and their parents were recruited and tissue specimens were collected after obtaining informed consents, following protocols approved by the IRB for Human Subject Research at Baylor College of Medicine (H-8712).

2.2 | Lung autopsy and biopsy

Histopathological initial evaluations and subsequent verification were done using formalin-fixed paraffin-embedded (FFPE) tissue specimens from lung biopsies or autopsies stained with hematoxylin and eosin.

2.3 | DNA isolation

DNA was extracted from peripheral blood or FFPE lung tissue using Gentra Purgene Blood Kit (Qiagen, Germantown, MD, USA) and DNeasy Blood and Tissue Kit (Qiagen), respectively.

2.4 | Array comparative genomic hybridization

CNV deletions were identified by comparative genomic hybridization (CGH) using custom-designed high-resolution,16q24.1 region-specific oligonucleotide microarrays (4 × 180K) (Agilent Technologies, Santa Clara, CA, USA). Array CGH (aCGH) was performed according to the Agilent Technologies aCGH protocol v3.5.

2.5 | Sequencing of deletion breakpoints

Deletion junctions were amplified by long-range PCR using LA Taq DNA polymerase (TaKaRa Bio, Madison, WI, USA). Cycling conditions were 94ºC for 30 s, and 68ºC for 7 min, repeated 30 times. Primers were designed with Primer3 (http://frodo.wi.mit.edu/primer3) using up to 10 kb-large breakpoint-containing regions determined by aCGH. PCR products were treated with ExoSAP-IT (USB, Cleveland, OH, USA) and directly Sanger sequenced. Sequences were assembled using Sequencher v.4.8 (GeneCodes, Ann Arbor, MI, USA) and the reference human genome version GRCh37/hg19 (http://genome.ucsc.edu).

2.6 | Parental origin of deletions

Parental origin of the deletions was determined using informative microsatellites or single nucleotide polymorphism (SNP) mapping to the deleted genomic interval.

2. 7 | Distribution of the recombination-associated motif along 16q24.1

The copy number of 7-mer 5’-CCTCCCT-3’ motif along the 16q24.1 region was compared with the expected copy number of this motif estimated by simulation assuming its uniform distribution. We checked 1000 randomly sampled regions equal in length to the 16q24.1 region and calculated the number of the recombination-associated motifs along the analyzed region. We justified the evidence of enrichment of the 7-mer recombination motif by checking the frequency of several randomly chosen 7-mers.

2.8 | *In silico* phylogenetic analyses of ACDMPV-linked L1PA2 and L1PA3 and *Alu* elements

BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was conducted for homologs of L1PA3 (chr16:86,266,902–86,272,916) and L1PA2 (chr16:86,295,780–86,301,803) on chromosome 16. Sequences with length cutoff of 5 kb and identity cutoff of 96% were aligned using Clustal Omega(http://www.ebi.ac.uk/Tools/msa/clustalo). Phylogenetic reconstruction was then performed using the maximum-likelihood (ML) method implemented in the R ‘phangorn’ package (https://cran.r-project.org/web/packages/phangorn/phangorn.pdf) with GTR + Γ + I model of evolution (the general time reversible model with corrections for invariant characters and gamma-distributed rate heterogeneity). The tree was rooted in the L1PA4 consensus sequence (Khan et al., 2006). The non-human primate evolution of *Alu* elements in the described locus was reconstructed by sequence comparison.

2.9 | PCR analyses of syntenic genomic regions in non-human primates

Experimental verification of genome integration times for L1PA2 and L1PA3 at 16q24.1 was done by determining the presence of their orthologs in syntenic chromosomal regions in chimp, gorilla, orangutan, and macaque by long-range PCR. Primers used for amplifications were designed from unique sequences flanking L1 elements of interest at locations syntenic for human 16q24.1.

3 | RESULTS

3.1 | A novel LINE and *Alu* genomic instability hotspot at 16q24.1

In addition to 12 previously reported CNV deletions with one breakpoint mapping at the distal edge or within the *FOXF1* upstream enhancer region (Dello Russo et al., 2015;

Szafranski et al., 2016), using aCGH and Sanger sequencing, we have now identified eight novel 16q24.1 deletions. The distal breakpoints of six deletions map within either L1PA2 (chr16:86,295,780–86,301,803) (pt 153.3) or L1PA3 (chr16:86,266,902–86,272,916) (pts 54.3, 155.3, 165.3, 177.3, and 179.3). These two full-length L1s are located ~ 22.9 kb apart, are directly-oriented, contain PolII promoters at their 5' end (Figure 1; Table 1; Supp. Figure S1), and both are included in the L1Base2 database of ~13,000 full-length FLn1–L1s; http://L1base.charite.de (Penzkofer et al., 2017).

In total, we have sequenced 12 ACDMPV CNV deletions with their distal breakpoints located within 16q24.1 L1PA2 (six) or L1PA3 (six), delimiting one side of the *FOXF1* upstream enhancer region (Figure 1). In nine of these 12 cases, the proximal breakpoint maps to a directly-oriented full-length or incomplete L1, exhibiting 91–97% sequence identity with L1 harboring the distal breakpoint and displaying 27–149 bp microhomology at the deletion junction sites. In the three remaining cases, the proximal breakpoint is located within non-homologous repetitive sequence (*Alu*Y, LTR/ERVL, or a simple repeat $(TTCC)_n$) with 2 bp-or no microhomology (Szafranski et al., 2014, 2016).

Interestingly, we have found that L1 and *Alu* content in the *FOXF1* locus is significantly lower than that estimated for the entire genome (Supp. Table S1). We have next inquired whether distribution of the breakpoints along L1 sequences is random or it correlates with the presence of some DNA structural features. We have found that breakpoints of four CNV deletions whose proximal breakpoint L1 element was complete (pts 60.4, 165.3, 177.3, and 179.3) map in 5' portion of the L1, whereas breakpoints of deletions with proximal breakpoint mapping to incomplete L1 (pts 54.3, 57.3, 127.3, 153.3, and 155.3) or non L1 sequence (pts 111.3, 119.3, and 139.3) clustered within 3' one-third portion of the L1PA2 or L1PA3 (Figure 2A).To shed more light on structural features within L1PA2and L1PA3 that might be causatively linked to the observed non-random distribution of DNA breakpoints along L1 sequence and L1's susceptibility to DNA breaks in general, locations of deletion breakpoints were analyzed in the context of GC content (https://www.biologicscorp.com/tools/GCContent), GC skewness (http://stothard.afns.ualberta.ca/cgview_server) (Grigoriev 1998), potentia to form palindromic structures (Grechishnikova & Poptsova 2016), and the presence of homologous recombination-associated,PRDM9-binding 7-mer 5'-CCTCCCT-3' or degenerate 13-mer 5'-CCNCCNTNNCCNC-3'motif (Billings et al., 2013; Myers et al., 2008). The average GC content around sequenced breakpoints (regions of microhomology or, in its absence, those flanking breakpoints by 20 bp on each side) is 39% (SD±2%), thus similar to overall 42% GC content of each of these two L1PAs (Figure 2A). We have also identified a negative GC composition bias in both L1s. We have not found any correlation between the location of the L1 breakpoints and the conserved stem-loops. Interestingly, the L1PA2 and L1PA3 breakpoints map within 1.6kb (SD±0.5 kb, n=10) of a 7-mer, 5'-CCTCCCT-3'of the recombination-associated motif (chr16:86,299,271–86,299,277 and chr16:86,270,389–86,270,395, respectively). This motif is also located 121 bp upstream of L1PA2 and in opposite orientation 236 bp downstream of L1PA3. In total, seven copies of 5'-CCTCCCT-3' are located between the two L1s. We have also found an enrichment of the 7-mer recombination motif in the entire 16q24.1 (P=0.004) (Supp. Figure S2).

One of the two CNV deletion breakpoints in four previously reported ACDMPV patients (28.7, 64.5, 95.3, and 117.3) (Szafranski et al., 2016) and in two newly reported patients (147.3 and 158.3) map within ~ 22.9 kb genomic interval between L1PA2 and L1PA3 harboring five different *Alu* elements (Figure 1; Table 1; Supp. Figure S3). Three of these breakpoints (pts 64.5, 95.3, and 147.3) map to the same *Alu*Sx (chr16:86,287,015–86,287,326), one (pt 158.3) maps to *Alu*Y (chr16:86,284,317–86,284,617), and one (pt 117.3) maps to *Alu*Sx1 (chr16:86,288,115–86,288,338). All those *Alu* elements are directly oriented with regard to each other and their partners at the other breakpoints. Thus, those deletions represent *Alu/Alu*-mediated genomic rearrangements (Song et al. 2018). In patient 28.7, breakpoint-containing regions were narrowed by aCGH to chr16:86,140,499 and chr16:86,285,499, but could not be sequenced (Stankiewicz et al., 2009). The GC content of the identified microhomologies around the deletion breakpoints was 48% (SD±16%), similar to 54% (SD±2%) average GC content for those three *Alu*s (Figure 2B). We found that the locations of deletion breakpoints do not correlate with the presence of a particular *Alu* stem-loop structure. *Alu*Y and *Alu*Sx each harbor PolIII promoter regions, thus similarly as L1PA2 and L1PA3, they might be transcribed. None of seven copies of the 7-mer recombination-associated motif, located between L1PA2 and L1PA3, maps to *Alu* element.

Besides ACDMPV-causing deletions, query of the Database of Genomic Variants (DGV) database of polymorphic CNVs (http://dgv.tcag.ca/dgv/app/home) revealed 48 small, presumably non-pathogenic deletions, and three reciprocal duplications, all with breakpoints mapping within this ~ 35 kb hotspot region (Figure 1). Although them of those CNVs were not sequenced, based on array CGH data, the majority if not all of them are likely located within L1PA2, L1PA3, *Alu*Y, or *Alu*Sx.

Of note, the identified 16q24.1 instability hotspot resides in the intron 3 of an ~ 61 kb-large lncRNA gene *LINC01081* oriented in the same direction as all *Alu*s and oppositely to L1s. All pathogenic CNV deletions discussed here arose *de novo*, on the maternally inherited chromosome 16. In one case (pt 179.3), the parental chromosome origin of de novo CNV deletion was not determined.

3.2 | Evolutionary origin of ACDMPV-linked L1PA2, L1PA3 and *Alu* elements

BLAST analyses of ACDMPV-linked L1PA2 and L1PA3 at 16q24.1 revealed that they share 97% sequence identity. PCR and *in silico* phylogenetic analyses of these L1s indicated that they arose in the human-chimpanzee-gorilla lineage after its split from the orangutan lineage, most likely 7–12 million years ago (Supp. FigureS4; Supp. Figure S4). We confirmed by PCR the presence of L1PA2 orthologs in the syntenic genomic regions of chimpanzee and gorilla and their absence in orangutan and macaque. However, we were able to amplify an ortholog of human L1PA3 only from chimpanzee, which suggests evolutionarily more recent arrival of this L1 at 16q24.1.

Sequence comparison of the non-human primate genomic regions syntenic with the human 16q24.1 instability hotspot (http://genome.ucsc.edu) (Supp. Figure S5) showed that the presence of *Alu*Sx1 and *Alu*Sx in this region dates around the time of the establishment of the Old World Monkey and the New World Monkey clades, respectively. The evolutionarily youngest *Alu*Y was found only in this genomic location only in humans. Interestingly, analysis of the database of polymorphic CNVs (Figure 1) showed that this *Alu*Y element may be polymorphic in different world populations.

4 | DISCUSSION

LINE/*Alu* hotspot at the *FOXF1* locus on 16q24.1

We describe a novel ~35 kb in size genomic instability hotspot on 16q24.1 that includes two L1s, L1PA2 and L1PA3, and five *Alu*s located in between. L1PA2, L1PA3, and *Alu*Sx are evolutionarily young elements that harbor recurrent breakpoints of both recurrent and nonrecurrent CNV deletions. We propose that recurrent DNA breaks in the described genomic instability hotspot might have been repaired using DNA sequence homology or homeology in other directly oriented full-length or truncated L1 partner (NAHR) or (ii) microhomology in shorter homologous or non-homologous sequences (i.e., MMBIR, MMEJ, or SSA), or by non-homologous end joining (NHEJ) (Carvalho & Lupski 2016; Song et al. 2018) (Table 1). Analyses of the *SPAST* locus at 2p22.3 also implicated *Alu*s in generation of recurrent DNA breaks leading to nonrecurrent CNVs (Boone et al., 2014).

L1 and *Alu* features that may predispose the genome to local instability

We found that the location of the deletion breakpoints along L1PA2 and L1PA3 in 16q24.1 correlates with the length of homology shared by flanking L1s. Breakpoints of CNVs with full-length L1 at their ends are located closer to the 5’ end of L1, whereas breakpoints of CNVs with L1 only at one of their two breakpoints are located closer to the 3’ end of L1.

Grechishnikova and Poptsova (2016) bioinformatically predicted potential of the evolutionarily young L1HS and L1PA1–L1PA8 elements and *Alu* repeats to adopt stem-loop structure. For instance, three conserved stem-loop clusters could form at L1’s 5’UTR, two in the middle of the ORF2, two at the end of ORF2, one at the 3’UTR, and numerous less conserved palindromes along the entire L1 length. We did not find correlation between location of L1PA2 and L1PA3 breakpoints and stem-loop structures, or G-quadruplex structures (Sahakyan et al., 2017). However, we have identified GC skewing along the length of L1PAs and *Alu*s, suggesting more frequent presence of their DNA in a single-stranded form (due to, e.g., their relatively more frequent replication or transcription) that may be easier to fold into non-B DNA structures predisposing to DNA breaks.

It has been suggested that the high frequency of LINE-or *Alu*-mediated CNVs may result from replication-transcription collisions (Hastings et al., 2009; Carvalho & Lupski 2016; Szafranski et al., 2016). We propose that secondary structures of L1s and *Alu*s might contribute to those events by slowing down or stopping progression of transcription or replication. Transcription, especially of the longer genes, results in prolonged chromatin opening and formation of R-loops, and may persist into the S phase of the cell cycle, thus increasing the chance of replication fork stalling followed by illegitimate template switching or fork collapse with broken DNA ends (Hastings et al., 2009). The genomic instability hotspot described here overlaps a long non-coding RNA gene, *LINC01081*, transcriptionally codirectional-directionally with *Alu*s and L1’s antisense promoter. Such genomic arrangement may lead not only to replication-transcription, but also transcription-

transcription collisions. Similarly, late replication increases chances of its interference with transcription, leading to stalled RNA polymerase complexes, and increasing the likelihood of template switching or the occurrence of DNA breaks within non-B DNA regions.

Of additional interest is general enrichment of 16q24.1 in recombination-associated 7-mer motif, in particular the presence of several copies of this motif within the described instability hotspot (one within each of the L1PAs and seven between them), suggesting that in some cases CNV formation might involve generation of double-strand breaks (DSBs), potentially initiated by meiosis specific SPO11 (Myers et al., 2008). Another possible scenario might involve generation of two DSBs in the vicinity of L1 elements, followed by resection of the annealing of two heterologous repeats by single strand annealing mechanism.

Conclusions

We demonstrate that the 16q24.1 genomic instability hotspot, harboring evolutionarily young L1s and *Alu*s, predisposes the genome to formation of same-and variably-sized CNV deletions via both homology-and non-homology-based mechanisms. As the detection of transposons and other repetitive elements is often challenging, we predict that a systematic genome-wide search for CNV breakpoint clusters will reveal more L1 and *Alu* genomic instability hotspots.

From the evolutionary perspective, TEs had contributed to development of hundreds of thousands of novel regulatory elements in the primate lineage and reshaped the human transcriptional landscape (Jacques et al. 2015). More recently, Trizzino et al. (2017) speculated that TEs, including L1s and *Alu*s, are the primary source of novelty in primate gene regulation. L1s and *Alu*s appeared at 16q24.1 location relatively recently during primate evolution, and substantial fraction of CNVs that they mediate are nonrecurrent. We hypothesize that formation of variably-sized CNVs catalyzed by recurrent DNA breaks within TEs in unstable genomic loci may have even facilitated evolution of environmental adaptation when compared to the same-sized CNVs occurring by NAHR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by grants awarded by the US National Heart, Lung, and Blood Institute (NIH grant R01HL137203) to P.St., the National Organization for Rare Disorders (2014 and 2016 16001 NORD grants) to P.Sz., and the Polish National Science Center (2012/06/M/ST6/00438) to A.G.

We thank Drs. Christine R. Beck, Grzegorz Ira, and James R. Lupski for helpful discussion.

Funding information

This work was supported by grants awarded by the US National Heart, Lung, and Blood Institute (NIH grant R01HL137203) to P.St., the National Organization for Rare Disorders (2014 and 2016 16001 NORD grants) to P.Sz., and the Polish National Science Center (2012/06/M/ST6/00438) to A.G.

REFERENCES

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, … Moran JV (2010). LINE-1 retrotransposition activity in human genomes.Cell,141, 1159–1170. [PubMed: 20602998]

Beck CR, Garcia-Perez JL, Badge RM, &Moran JV (2011). LINE-1 elements in structural variation and disease. Annual Reviews in Genomics and Human Genetics,12, 187–215.

Belancion VP, Deininger PL, &Roy-Engel AM (2009). LINE dancing in the human genome: transposable elements and disease. Genome Medicine, 1, 97. [PubMed: 19863772]

Billings T, Parvanov ED, Baker CL, Walker M, Paigen K, &Petkov PM (2013). DNA binding specificities of the long zinc-finger recombination protein PRDM9. Genome Biology, 14, R35. [PubMed: 23618393]

Bishop NB, Stankiewicz P, &Steinhorn RH (2011). Alveolar capillary dysplasia. American Journal of Respiratory and Critical Care Medicine,184: 172–179. [PubMed: 21471096]

Boissinot S,& Sookdeo A (2016). The evolution of LINE-1 in vertebrates. Genome Biology and Evolution,8, 3485–3507. [PubMed: 28175298]

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC… Lupski JR (2014). The *Alu*-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles.American Journal of Human Genetics,95, 143–161. [PubMed: 25065914]

Burwinkel B, &Kilimann MW (1998). Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. Journal of Molecular Biology,277, 513–517. [PubMed: 9533876]

Buysse K, Crepel A, Menten B, Pattyn F, Antonacci F, Veltman JA,… Speleman F(2008). Mapping of 5q35 chromosomal rearrangements within a genomically unstable region. Journal of Medical Genetics,45, 672–678. [PubMed: 18628311]

Callinan PA, &Batzer MA (2006). Retrotransposable elements and human disease. Genome Dynamics, 1, 104–115. [PubMed: 18724056]

Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P,… Stankiewicz P(2014). Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination.BMC Biology,12, 74. [PubMed: 25246103]

Carvalho CM, &Lupski JR (2016). Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics,17, 224–238.

Dello Russo P, Franzoni A, Baldan F, Puppin C, De Maglio G, Pittini C, … Damante G (2015). A 16q deletion involving *FOXF1* enhancer is associated to pulmonary capillary hemangiomatosis. BMC Medical Genetics, 16, 94. [PubMed: 26462560]

de Koning APJ, Gu W, Castoe TA, Batzer MA, &Pollock DD(2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genetics,7, e1002384. [PubMed: 22144907]

Furano AV (2000). The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Progress in Nucleic Acid Research and Molecular Biology,64, 255–294. [PubMed: 10697412]

Gilbert N, Lutz S, Morrish TA, &Moran JV (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Molecular and Cellular Biology,25, 7780–7795. [PubMed: 16107723]

Gogvadze E, &Buzdin A (2009). Retroelements and their impact on genome evolution and functioning. Cellular and Molecular Life Sciences,66, 3727–3742. [PubMed: 19649766]

Grechishnikova D, &Poptsova M (2016). Conserved 3’ UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition. BMC Genomics,17, 992. [PubMed: 27914481]

Grigoriev A (1998). Analyzing genomes with cumulative skew diagrams.Nucleic Acids Resarch,26, 2286–2290.

Gu S, Szafranski P, Akdemir ZC, Yuan B, Cooper ML, Magriñá MA,… Lupski JR(2016). Mechanisms for complex chromosomal insertions.PLoS Genetics,12, e1006446. [PubMed: 27880765]

Gu S, Yuan B, Campbell IM, Beck CR, Carvalho CM, Nagamani SC, … Lupski JR (2015). *Alu*-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3.Human Molecular Genetics,24, 4061–4077. [PubMed: 25908615]

Hancks DC, &Kazazian HH, Jr. (2016). Roles for retrotransposon insertions in human disease. Mobile DNA,7, 9. [PubMed: 27158268]

Hastings PJ, Lupski JR, Rosenberg SM, &Ira G (2009). Mechanisms of change in gene copy number. Nature Reviews Genetics,10, 551–564.

Hedges DJ, &Deininger PL (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutation Research,616, 46–59. [PubMed: 17157332]

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, … Guryev V (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants.Nature Communications,7, 12989.

Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, &Meyerson M (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Research,24, 1053–1063. [PubMed: 24823667]

Higashimoto K, Maeda T, Okada J, Ohtsuka Y, Sasaki K, Hirose A, … Soejima H(2013). Homozygous deletion of *DIS3L2* exon 9 due to non-allelic homologous recombination between LINE-1s in a Japanese patient with Perlman syndrome. European Journal of Human Genetics, 21, 1316–1319. [PubMed: 23486540]

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF,… Devine SE (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell,141, 1253–1261. [PubMed: 20603005]

Ivancevic AM, Kortschak RD, Bertozzi T, &Adelson DL (2016). LINEs between species: Evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. Genome Biology and Evolution,8, 3301–3322. [PubMed: 27702814]

Jacques PÉ, Jeyakani J, &Bourque G (2015). The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genetics,9, e1003504.

Jurka J (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics,16, 418–420. [PubMed: 10973072]

Kazazian HH, Jr, &Moran JV (2017).Mobile DNAin health and disease.New England Journal of Medicine, 377, 361–370. [PubMed: 28745987]

Khan H, Smit A, & Boissinot S (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Research, 16, 78–87. [PubMed: 16344559]

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M,… Eichler EE (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell,143, 837–847. [PubMed: 21111241]

Kohmoto T, Naruto T, Watanabe M, Fujita Y, Ujiro S, Okamoto N,…Imoto I (2017). A 590 kb deletion caused by non-allelic homologous recombination between two LINE-1 elements in a patient with mesomelia-synostosis syndrome. American Journal of Medical Genetics A,173A, 1082–1086.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, …Snyder M(2007). Paired-end mapping reveals extensive structural variation in the human genome. Science,318, 420–426. [PubMed: 17901297]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, …(2001). Initial sequencing and analysis of the human genome. Nature, 409, 860–921. [PubMed: 11237011]

Langston C (1991). Misalignment of pulmonary veins and alveolar capillary dysplasia. Pediatric Pathology, 11, 163–170. [PubMed: 2014189]

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, … Park PJ (2012). Landscape of somatic retrotransposition in human cancers. Science,337, 967–971. [PubMed: 22745252]

Liu P, Erez A, Nagamani SC, Dhar SU, Kołodziejska KE, Dharmadhikari AV, …Bi W (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell,146, 889–903. [PubMed: 21925314]

Lupski JR (2010). Retrotransposition and structural variation in the human genome. Cell,141, 1110–1112. [PubMed: 20602993]

Myers S, Freeman C, Auton A, Donnelly P, &McVean G (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nature Genetics,40, 1124–1129. [PubMed: 19165926]

Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, & Zemojtel T (2017). L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Research, 45, D68–D73. [PubMed: 27924012]

Quadri M, Vetro A, Gismondi V, Marabelli M, Bertario L, Sala P, … Ranzani GN (2015). APC rearrangements in familial adenomatous polyposis: heterogeneity of deletion lengths and breakpoint sequences underlies similar phenotypes. Familial Cancer,14, 41–49. [PubMed: 25159889]

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, &Moran JV (2015). The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiology Spectrum, 3, MDNA3–0061-2014.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, & Vermeesch JR (2013). Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. Genome Research, 23, 411–418. [PubMed: 23212949]

Sahakyan AB, Murat P, Mayer C, &Balasubramanian S (2017).G-quadruplex structures within the 3’ UTR of LINE-1 elements stimulate retrotransposition.Nature Structural & Molecular Biology,24, 243–247.

Sen P, Yang Y, Navarro C, Silva I, Szafranski P, Kolodziejska KE, … Stankiewicz P (2013). Novel *FOXF1* mutations in sporadic and familial cases of alveolar capillary dysplasia with misaligned pulmonary veins imply a role for its DNA binding domain. Human Mutation, 34, 801–811. [PubMed: 23505205]

Song X, Beck CR, Du R, Campbell IM, Coban-Akdemir Z, Gu S, … Lupski JR (2018). Predicting human genes susceptible to genomic instability associated with *Alu*/*Alu*-mediated rearrangements Genome Research, (in press).

Stankiewicz P, Sen P, Bhatt SS, Storer M, Xia Z, Bejjani BA,…Shaw-Smith C (2009). Genomic and genic deletions of the FOX gene cluster on 16q24.1 and inactivating mutations of *FOXF1* cause alveolar capillary dysplasia and other malformations. American Journal of Human Genetics, 84, 780–791. [PubMed: 19500772]

Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA… Gambin A (2015). Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. Nucleic Acids Research,43, 2188–2198. [PubMed: 25613453]

Szafranski P, Dharmadhikari AV, Brosens E, Gurha P, Kolodziejska KE, Zhishuo O, … Stankiewicz P (2013). Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. Genome Research, 23, 23–33. [PubMed: 23034409]

Szafranski P, Dharmadhikari AV, Wambach JA, Towe CT, White FV, Grady RM, … Stankiewicz P (2014). Two deletions overlapping a distant *FOXF1* enhancer unravel the role of lncRNA *LINC01081* in etiology of alveolar capillary dysplasia with misalignment of pulmonary veins. American Journal of Medical Genetics A, 164A, 2013–2019.

Szafranski P, Gambin T, Dharmadhikari AV, Akdemir KC, Jhangiani SN, Schuette J, …Stankiewicz P (2016). Pathogenetics of alveolar capillary dysplasia with misalignment of pulmonary veins. Human Genetics,135, 569–586. [PubMed: 27071622]

Temtamy SA, Aglan MS, Valencia M, Cocchi G, Pacheco M, Ashour AM,…Ruiz-Perez VL (2008). Long interspersed nuclear element-1 (LINE1)-mediated deletion of *EVC*, *EVC2*, *C4orf6*, and *STK32B* in Ellis-van Creveld syndrome with borderline intelligence. Human Mutation,29, 931–938. [PubMed: 18454448]

Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, … Hehir-Kwa JY (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biology,15, 488. [PubMed: 25348035]

Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, … Brown CD (2017). Transposable elements are the primary source of novelty in primate gene regulation. Genome Research,27, 1623–1633. [PubMed: 28855262]

Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, … Stankiewicz P (2009). Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. Human Molecular Genetics, 18, 3579–3593. [PubMed: 19578123]

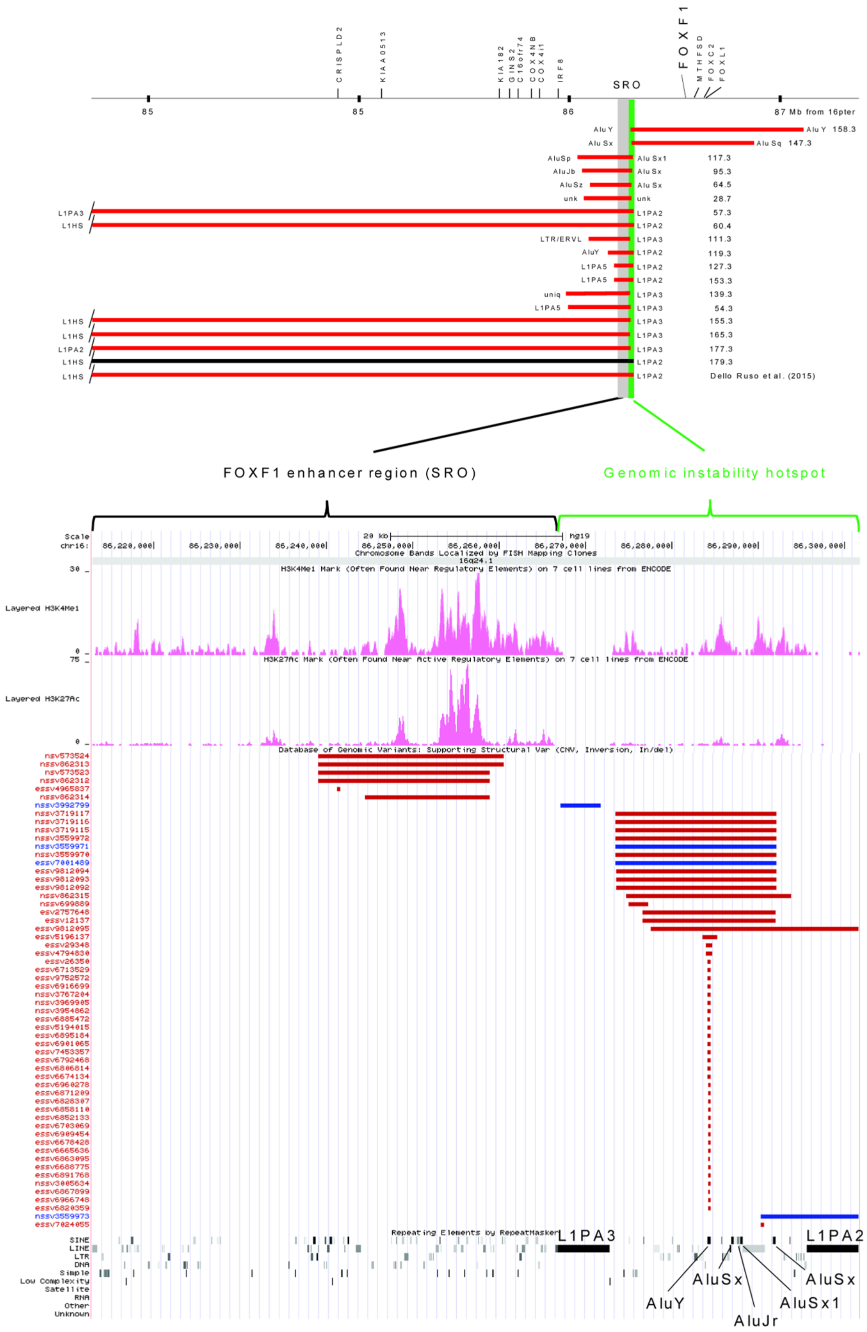

Figure 1. The LINE-and *Alu*-containing genomic instability hotspot at the *FOXF1* locus on 16q24.1.

Nineteen CNV deletions causative for ACDMPV have one of the two breakpoints mapping in L1PA2, L1PA3, or *Alu*s located in between at 16q24.1 (pts 179.3 and Dello Russo et al. (2015) had pulmonary hypertension and capillary hemangiomatosis, respectively). Genomic location of the locus is marked with a vertical green bar at the distal edge of the ~ 60 kb tissue-specific *FOXF1* enhancer region (SRO, smallest region of deletion overlap) (Szafranski et al., 2016). Notably, all but one (pt 179.3) ACDMPV-causing deletions arose *de novo* on the maternal chromosome 16. The lower panel shows the *FOXF1* enhancer region, the described ~35 kb genomic instability hotspot located at its distal end, and DGV CNV deletions (red) and duplications (blue),further indicating instability at this genomic hotspot. Epigenetic features (H3K27ac and H3K4me1) are shown in the middle.

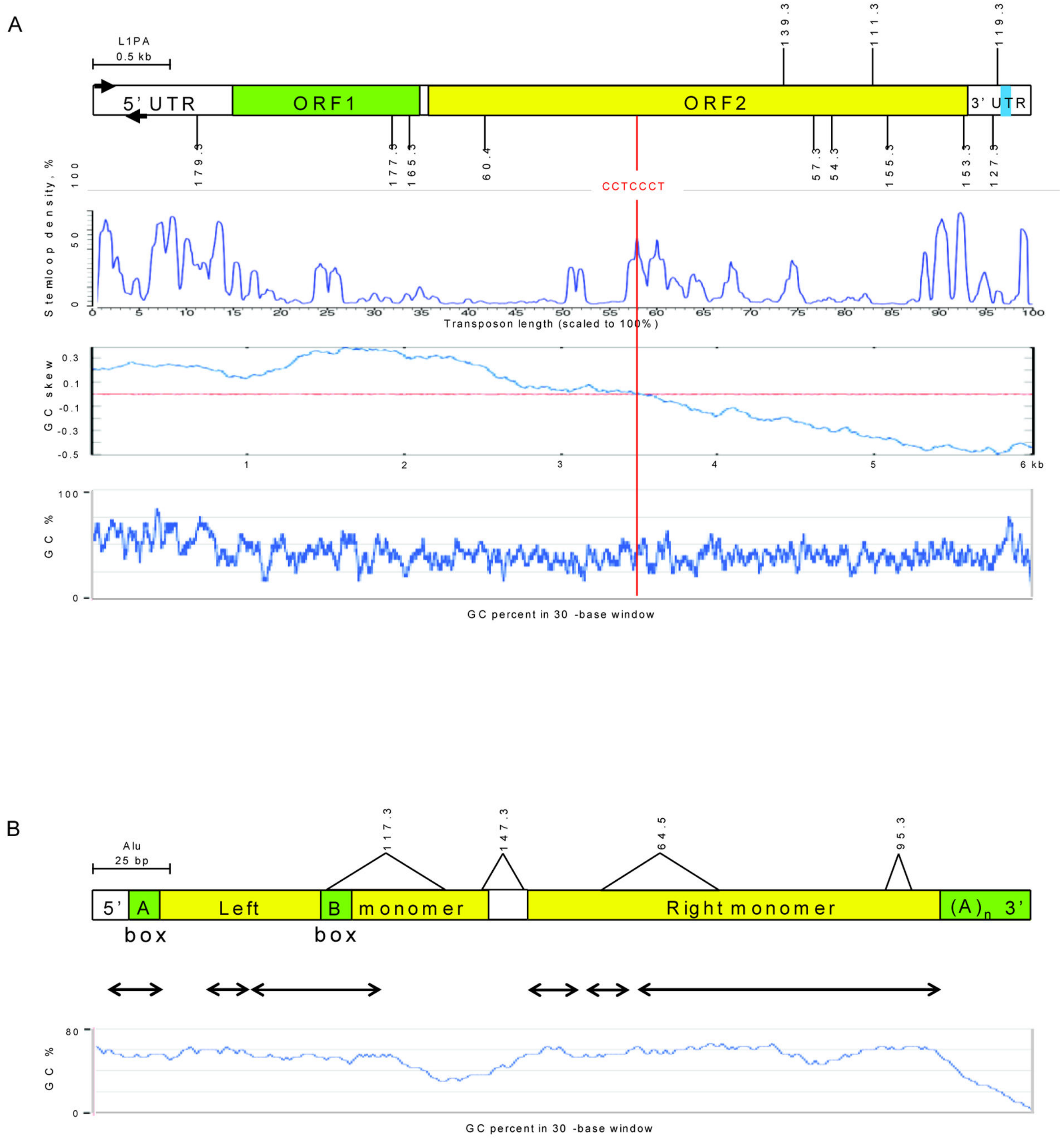

Figure 2. Distribution of the 16q24.1 CNV deletion breakpoints along retrotransposon consensus sequences.

(A) Location of the deletion breakpoints in L1s. Breakpoints of nine CNVs with the proximal breakpoint located within full length or incomplete L1, and three CNVs with proximal breakpoints located in non L1 sequence are shown below and above, respectively. Blue box in the 3’ UTR refers to G-rich sequence. The stem-loop density along L1 is shown for L1PA3and is similar to those for other L1PA1–L1PA4s (Grechishnikova & Poptsova 2016). The cumulative GC skewness calculated for L1PA3 as the sum of (G-C)/(G+C) of the adjacent 5-base windows sliding along L1 sequence, and its profile is similar to that calculated for the neighboring L1PA2. **(B)** Location of the deletion breakpoints along the consensus *Alu* element. Positions of microhomologies around the breakpoints are indicated above the *Alu* diagram. Components of the RNA PolIII promoter are labeled A Box and B Box. Conserved stem-loop structures are indicated by arrows.

Table 1.

Localization of 38 breakpoints of ACDMPV-causative CNV deletions at the LINE and *Alu* genomic instability hotspot on 16q24.1.

Sizes of the truncated, but highly similar and directly oriented LINE elements, are shown in parenthesis. LINEs and *Alu* repeats constituting the hotspot are shown in bold.

ACDMPV pt	Deletion coordinates	Repetitive element containing breakpoint				Proposed mechanism of CNV deletion formation[*]
		Proximal	Distal	Identity between LINEs or *Alus* (%)	Microhomology (bp)	
28.7	~ chr16:86,140,499–86,285,499	unk	unk	unk	unk	unk
64.5	chr16:86,147,527/566–86,287,120/159	*Alu*Sz	**_Alu_Sx**	84	38	MMBIR, MMEJ, or SSA
95.3	chr16:86,118,131/141–86,287,054/064	*Alu*Jb	**_Alu_Sx**	75	9	MMBIR, or MMEJ
117.3	chr16:86,055,159/200–86,288,226/268	*Alu*Sp	**_Alu_Sx1**	87	41	MMBIR, MMEJ, or SSA
147.3	chr16:86,287,188/199–86,848,466/477	**_Alu_Sx**	*Alu*Sq	82	10	MMBIR, or MMEJ
158.3	chr16:86,284,317/617–87,137,455/746	**_Alu_Y**	*Alu*Y	89	unk	MMBIR, MMEJ, or SSA
54.3	chr16:85,910,504/580–86,271,634/710	L1PA5 (1.7 kb)	**L1PA3**	93	75	NAHR
57.3	chr16:82,014,639/716–86,300,403/481	L1PA3 (2.6 kb)	**L1PA2**	97	77	NAHR
60.4	chr16:83,673,382/476–86,298,284/378	L1HS	**L1PA2**	97	93	NAHR
111.3	chr16:86,077,955/958–86,271,915/918	LTR/ERVL	**L1PA3**	-	2	NHEJ, MMEJ, or MMBIR
119.3	chr16:86,148,250–86,301,591	*AluY*	**L1PA2**	-	7 bp insertion at the deletion junction	NHEJ
127.3	chr16:86,209,157/194–86,301,558/595	L1PA5 (0.6 kb)	**L1PA2**	91	36	NAHR
139.3	chr16:85,877,831–86,271,338	simple repeat $(TTCC)_n$	**L1PA3**	-	0	NHEJ
153.3	chr16:86,208,967/995–86,301,369/397	L1PA5 (0.6 kb)	**L1PA2**	91	27	NAHR
155.3	chr16:84,491,194/238–86,271,998/272,042	L1HS (2.1 kb)	**L1PA3**	97	43	NAHR
165.3	chr16:83,672,829/882–86,268,857/910	L1HS	**L1PA3**	97	52	NAHR
177.3	chr16:82,174,710/852–86,268,760/909	L1PA2	**L1PA3**	96	149	NAHR
179.3	chr16:83,671,523/574–86,296,427/478	L1HS	**L1PA2**	97	51	NAHR
Dello Russo et al. 2015	~ chr16:83,676,990–86,292,585	unk	unk	unk	unk	unk

[*] MMBIR, microhomology-mediated break-induced replication; MMEJ, microhomology-mediated end joining; NAHR, nonallelic homologous recombination; NHEJ, nonhomologous end joining; SSA, single strand annealing, unk, unknown