Human Molecular Genetics, 2017, Vol. 26, No. 10 1927–1941

doi: 10.1093/hmg/ddx102
Advance Access Publication Date: 14 March 2017
Original Article

ORIGINAL ARTICLE

Efficient CNV breakpoint analysis reveals unexpected structural complexity and correlation of dosage-sensitive genes with clinical severity in genomic disorders

Ling Zhang[1,2,3,†], Jingmin Wang[4,†], Cheng Zhang[1], Dongxiao Li[4], Claudia M.B. Carvalho[5], Haoran Ji[4], Jianqiu Xiao[1], Ye Wu[4], Weichen Zhou[1], Hongyan Wang[1,2,3], Li Jin[1,2], Yang Luo[6], Xiru Wu[4], James R. Lupski[5,7,8], Feng Zhang[1,2,3,*] and Yuwu Jiang[4,*]

[1]Obstetrics and Gynecology Hospital, State Key Laboratory of Genetic Engineering at School of Life Sciences, Institute of Reproduction and Development, Fudan University, Shanghai 200011, China, [2]Key Laboratory of Reproduction Regulation of NPFPC, Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai 200032, China, [3]Shanghai Key Laboratory of Female Reproductive Endocrine Related Diseases, Shanghai 200011, China, [4]Department of Pediatrics, Peking University First Hospital, Beijing 100034, China, [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA, [6]MOE Key Laboratory of Medical Cell Biology, The Research Center for Medical Genomics, College of Basic Medical Science, China Medical University, Shenyang 110001, China, [7]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA and [8]Texas Children's Hospital, Houston, TX 77030, USA

*To whom correspondence should be addressed at: Obstetrics and Gynecology Hospital, Fudan University, Shanghai 200011, China. Tel: +86 21 51630423; Fax: +86 21 51630607; Email: zhangfeng@fudan.edu.cn (F.Z.); Email: jiangyuwu@bjmu.edu.cn (Y.J.)

Abstract

Genomic disorders are the clinical conditions manifested by submicroscopic genomic rearrangements including copy number variants (CNVs). The CNVs can be identified by array-based comparative genomic hybridization (aCGH), the most commonly used technology for molecular diagnostics of genomic disorders. However, clinical aCGH only informs CNVs in the probe-interrogated regions. Neither orientational information nor the resulting genomic rearrangement structure is provided, which is a key to uncovering mutational and pathogenic mechanisms underlying genomic disorders. Long-range polymerase chain reaction (PCR) is a traditional approach to obtain CNV breakpoint junction, but this method is inefficient when challenged by structural complexity such as often found at the *PLP1* locus in association with Pelizaeus-Merzbacher disease (PMD). Here we introduced 'capture and single-molecule real-time sequencing' (cap-SMRT-seq) and newly developed 'asymmetry linker-mediated nested PCR walking' (ALN-walking) for CNV breakpoint sequencing in 49 subjects with PMD-associated CNVs. Remarkably, 29 (94%) of the 31 CNV breakpoint junctions unobtainable by conventional long-range PCR were resolved by cap-SMRT-seq and ALN-walking. Notably, unexpected CNV complexities, including inter-chromosomal

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Received: January 3, 2017. **Revised:** March 8, 2017. **Accepted:** March 10, 2017

rearrangements that cannot be resolved by aCGH, were revealed by efficient breakpoint sequencing. These sequence-based structures of PMD-associated CNVs further support the role of DNA replicative mechanisms in CNV mutagenesis, and facilitate genotype–phenotype correlation studies. Intriguingly, the lengths of gained segments by CNVs are strongly correlated with clinical severity in PMD, potentially reflecting the functional contribution of other dosage-sensitive genes besides *PLP1*. Our study provides new efficient experimental approaches (especially ALN-walking) for CNV breakpoint sequencing and highlights their importance in uncovering CNV mutagenesis and pathogenesis in genomic disorders.

Introduction

Genomic copy number variation (CNV) is one of the major sources of human genetic diversity (1). Large CNVs at several genomic loci result in genomic disorders, such as Charcot-Marie-Tooth disease type 1A (MIM: 118220), DiGeorge syndrome (MIM: 188400) and Pelizaeus-Merzbacher disease (PMD [MIM: 312080]) (2–4). Furthermore, CNVs can also be involved in susceptibilities to complex diseases including Alzheimer disease, autism and obesity (5–7). Notably, somatic CNV mutations can drive cancer genome evolution (8).

Array-based comparative genomic hybridization (aCGH) is the most commonly used genome technology for molecular diagnostics of genomic disorders and other CNV-related diseases (9–11). However, aCGH does not provide comprehensively the complete sequence-based variant information of CNV structures. For example, a genomic duplicated segment may be organized *in tandem* (head-to-tail orientation) or it can be inserted in a heterologous chromosome, alternatives that cannot be distinguished by array-based genome technologies (12,13). This technical limitation can sometimes hamper interpretation of experimental findings and prevent complete understanding of mutational and pathogenic mechanisms underlying CNV-related diseases. Therefore, CNV breakpoint sequencing following clinical target aCGH is needed to fully reveal CNV structures at base-pair resolution. Long-range polymerase chain reaction (PCR) amplification of CNV breakpoint junctions for Sanger sequencing is a conventional method of CNV breakpoint sequencing (14–16). However, this assay is not efficient, especially for the duplication CNVs at loci with complex genome architecture. Therefore, efficient methods of CNV breakpoint sequencing are imperative for genomic disorders, especially for PMD-associated CNVs and other challenging genomic loci. In this study, we introduced two methods for CNV breakpoint sequencing: capture and single-molecule real-time DNA sequencing (cap-SMRT-seq) and asymmetry linker-mediated nested PCR walking (ALN-walking). The majority (94%) of the challenging PMD-associated duplication breakpoints that failed to be elucidated by long-range PCR were resolved by our efficient cap-SMRT-seq and ALN-walking. Therefore, the combined approaches, i.e. high-resolution aCGH plus CNV breakpoint sequencing (especially ALN-walking), enabled sequence-based resolution of PMD-associated CNVs, uncovered unexpected CNV complexities, and revealed a correlation between the genomic contents gained by CNV with clinical severity in PMD.

Results

Complex PMD-associated CNVs identified by high-resolution aCGH

High-resolution aCGH was used to investigate 49 subjects with PMD (Table 1), for whom multiplex ligation-dependent probe amplification (MLPA) analysis indicated *PLP1* copy-number gains (Fig. 1 and Supplementary Material, Fig. S1). In 48 of 49 subjects, *PLP1* was duplicated, while *PLP1* was triplicated in subject J089 (Fig. 2 and Supplementary Material, Fig. S2). High-resolution aCGH assay also revealed potential CNV structural complexities that were not detected by low-resolution MLPA. Apparent simple duplications (i.e. continuous genomic segments with copy number gains shown by aCGH) were observed in 35/49 subjects whereas 14/49 present complex CNVs based on the aCGH analysis (Fig. 2). The PMD-associated CNV in the female subject JL21 was large and exceeded the coverage of the PMD-target aCGH utilized in this study. Therefore, the genome-wide aCGH was used for re-mapping of the CNV in JL21, resulting in the identification of a 33.5-Mb duplication of Xq21.1-Xq24.

The aCGH analyses revealed a high frequency (29%, 14/49) of complex CNV patterns suggesting complex genomic rearrangements (CGRs) (Fig. 2). These CNV complexities included interspersed duplications (DUP-NML-DUP in subject J029, J049, J068, J077, J084 and J090; DUP-NML-DUP-NML-DUP in subject J021; DUP, duplication; NML, normal) and triplications (TRPs) flanked by duplications (DUP-TRP-DUP in subject J007, J032, J061, J082, J089, J091 and J102) (Fig. 2). However, technical limitations of the aCGH assay prevented further resolution of CNV structure; especially determining the genomic orientations of these PMD-associated CNVs. Therefore, CNV breakpoint sequencing was pursued to further resolve CGR and reveal sequence-based CNV structures.

Efficient CNV breakpoint sequencing

The high-resolution aCGH results predicted 64 CNV breakpoint junctions, 11 of which were consistent with the homologous recombinations between paired low-copy repeats (LCR) (Figs 1 and 2). The remaining 53 breakpoint junctions were predicted to be produced using undetermined substrate sequences from unique genomic regions (Figs 1 and 2). Long-range PCR was conducted and allowed amplification of only 23 out of 53 putative CNV breakpoint junctions. Here, using cap-SMRT-seq and ALN-walking, we successfully resolved most (94%) of the 31 challenging junctions that failed to be obtained by conventional long-range PCR.

The cap-SMRT-seq method was modified partially based on the previously reported principle of enriching DNA fragments of multiple CNV breakpoint junctions on a DNA capture array (17). To obtain long sequencing reads for CNV breakpoint analysis in an unbiased CNV formation model, we employed the single-molecular real-time (SMRT) technology (Pacific Biosciences) in our method of cap-SMRT-seq.

Among the 31 CNV breakpoint junctions that failed to be resolved by long-range PCR in this study, cap-SMRT-seq successfully resolved 17 breakpoint junctions (Supplementary Material, Table S1). Cap-SMRT-seq is more efficient in resolving breakpoints of simple CNVs (68%, 13/19) than those of complex CNVs (33%, 4/12) (Supplementary Material, Table S1).

The remaining 14 CNV breakpoint junctions were not obtained by either long-range PCR or cap-SMRT-seq. We therefore

Table 1. Clinical features of 49 subjects with PMD

Subject	PCS	Last evaluation age	Disease onset	Psychomotor development (age at milestones acquisition)					Other symptom
				Head control	Sitting	Standing	Walking	Simple language expression	
J003	2	65 m	10 d	24 m	Not acq	Not acq	Not acq	16 m	No
J007	3	22 m	0 d	22 m	Not acq	Not acq	Not acq	22 m	No
J010	2	125 m	20 d	Not acq	Not acq	Not acq	Not acq	Not acq	No
J016	4	80 m	1 m	3.5 m	7 m	Not acq	Not acq	12 m	R
J018[a]	NA	7 m	6 m	Not acq	NA	NA	NA	NA	JC
J019	1	54 m	5 d	7.5 m	36 m	Not acq	Not acq	48 m	DK
J021[a]	NA	11 m	35 d	6 m	Not acq	Not acq	NA	NA	No
J026	5	85 m	9 m	5 m	9 m	Not acq	Not acq	24 m	No
J029	1	360 m	0 d	12 m	Not acq	Not acq	Not acq	18 m	DP, JC
J031	0	62 m	0 d	Not acq	Not acq	Not acq	Not acq	Not acq	DP
J032	2	116 m	7 d	38 m	108 m	Not acq	Not acq	12 m	No
J033	3	126 m	0 d	8 m	24 m	48 m	Not acq	29 m	JC, DK
J035	2	72 m	4 m	Not acq	Not acq	Not acq	Not acq	30 m	DP
J038	2	48 m	9 d	3 m	Not acq	Not acq	Not acq	24 m	DK
J040	0	72 m	10 d	60 m	Not acq	Not acq	Not acq	Not acq	DP
J041[a]	NA	NA	NA	NA	NA	NA	NA	NA	NA
J043	0	47 m	0 d	Not acq	Not acq	Not acq	Not acq	Not acq	DP, T
J044	1	144 m	3 m	12 m	96 m	Not acq	Not acq	36 m	DP
J046	4	48 m	0 d	7.5 m	24 m	36 m	Not acq	24 m	No
J049	1	116 m	0 d	86 m	Not acq	Not acq	Not acq	24 m	DP, T
J050	1	50 m	15 d	12 m	Not acq	Not acq	Not acq	18 m	DP, T, C
J052	1	68 m	2 m	Not acq	Not acq	Not acq	Not acq	Not acq	DP, T
J057	4	116 m	0 d	7 m	8 m	Not acq	Not acq	24 m	No
J061	3	86 m	10 d	9 m	Not acq	Not acq	Not acq	12 m	No
J064	5	26 m	2 m	3.5 m	Not acq	Not acq	Not acq	12 m	No
J065	4	48 m	1 m	7 m	12 m	Not acq	Not acq	24 m	T
J068[a]	NA	NA	NA	NA	NA	NA	NA	NA	NA
J069	3	57 m	2 m	3 m	Not acq	Not acq	Not acq	30 m	JC
J077	4	132 m	2 m	3 m	12 m	66 m	102 m	36 m	DK
J078	5	146 m	20 d	7.5 m	12 m	78 m	102 m	24 m	No
J079	4	35 m	3 m	5 m	Not acq	Not acq	Not acq	Not acq	DP, DK
J080	5	27 m	3 m	5 m	10 m	Not acq	Not acq	15 m	No
J082	5	62 m	1 m	7.5 m	12 m	Not acq	Not acq	24 m	No
J083	5	27 m	1 m	3 m	Not acq	Not acq	Not acq	Not acq	No
J084	5	106 m	21 d	4 m	8 m	Not acq	Not acq	12 m	No
J086	3	65 m	0 d	5.5 m	6 m	Not acq	Not acq	12 m	DK
J089[a]	NA	12 m	15 d	10 m	Not acq	Not acq	Not acq	NA	C
J090	4	60 m	2.7 m	8 m	Not acq	Not acq	Not acq	15 m	No
J091	1	64 m	0 d	13.5 m	Not acq	Not acq	Not acq	Not acq	No
J095	3	132 m	0 d	6 m	6 m	Not acq	Not acq	11	S
J097	4	43 m	3 m	18 m	19 m	Not acq	Not acq	8 m	No
J102	3	29 m	15 d	12 m	Not acq	Not acq	Not acq	24 m	No
J103	1	69 m	2 m	Not acq	Not acq	Not acq	Not acq	Not acq	DP, T
J104	5	20 m	2 m	4 m	20 m	Not acq	Not acq	20 m	No
J107	4	40 m	3 m	5 m	Not acq	Not acq	Not acq	24 m	No
J108[a]	NA	9 m	3 m	4 m	Not acq	NA	NA	NA	No
J109	1	47 m	4 m	15 m	Not acq	Not acq	Not acq	Not acq	DP, T
JL21[b]	NA	18 m	3 m	6 m	12 m	Not acq	Not acq	Not acq	Died at 18m
L007[a]	NA	NA	NA	NA	NA	NA	NA	NA	NA

Note: All the subjects with PCS showed nystagmus at onset except J097 whose initial symptom was developmental delay. Most cases represented hypotonia and a few cases occasionally with paroxysmal dystonia. MRI all had hypomyelination.

C, convulsion; d, day; DD, developmental delay; DK, dyskinesia; DP, dysphagia; JC, joint contracture; m, month; NA, not available; R, retrogress; S, stridor; T, tremor. 'Not acq' means 'Not acquire the milestone at the last evaluation age'.

[a]These seven subjects with incomplete clinical information were not included in the genotype–phenotype correlation study.

[b]Female case, not included in phonotype–genotype correlation analysis.

developed another experimental approach, named ALN-walking (Fig. 3 and Supplementary Material, Table S2), based on the principle of genome walking and linker-mediated nested PCR. Intriguingly, ALN-walking resolved the majority (86%, 12/14) of these most challenging breakpoint junctions (Supplementary Material, Table S1). This success rate of ALN-walking was much higher than that of long-range PCR (42%, 22/53) and cap-SMRT-seq (55%, 17/31).

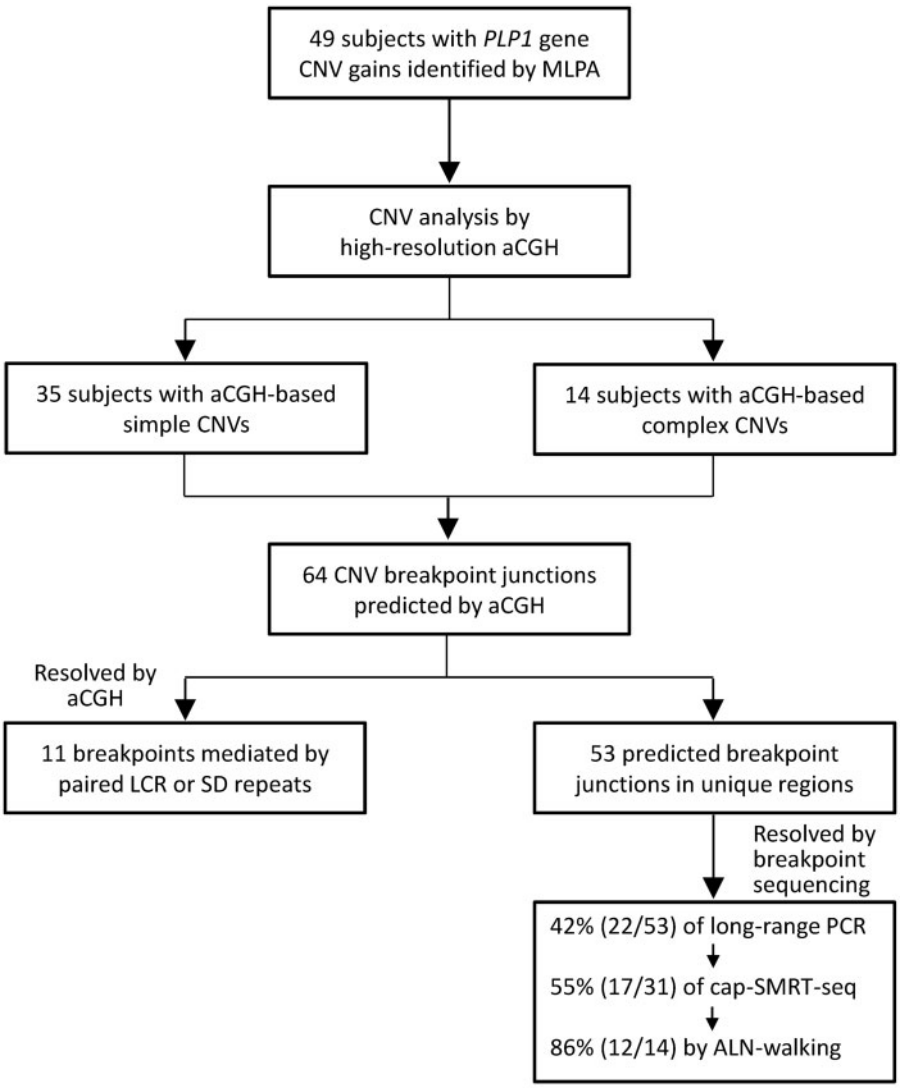

Figure 1. Workflow of CNV analyses and breakpoint sequencing for 49 subjects with PMD-associated *PLP1* gene CNVs. *PLP1* gene CNVs were identified initially by MLPA assay and were further verified by high-resolution aCGH. The sequence-based CNV structures were investigated comprehensively by CNV breakpoint sequencing. Besides the traditional long-range PCR, two efficient methods for CNV breakpoint sequencing, 'cap-SMRT-seq' and 'ALN-walking', were employed. SD, segmental duplication.

The detailed breakpoint information and proposed CNV structures comprehensively revealed by efficient breakpoint sequencing are described in Table 2, Supplementary Material, Tables S3 and S4.

Additional CNV complexities revealed by breakpoint sequencing

High-resolution aCGH revealed apparent simple CNVs in 35 subjects and complex CNVs in 14 subjects (Fig. 1). Using efficient CNV breakpoint sequencing, we identified unanticipated complex CNVs yielding CGR that could not be identified by high-resolution aCGH. These further complexities were resolved in seven apparent aCGH-based 'simple' CNVs and 2 of the 14 apparent complex CNVs, which added up to 21 (43%) sequence-based complex CNVs in the PMD subjects of this study (Table 2 and Supplementary Material, Table S4).

The newly revealed complexities could be classified into three groups. (i) Small templated insertions at the breakpoint junctions. We only considered the fragments above 20 bp at the junctions as templated insertions. The insertional events could be inverted or in a tandem orientation (Table 2). The templated insertions were found in subject J026 (30 bp and 41 bp), J068 (116 bp), J079 (40 bp and 70 bp, Fig. 4) and J090 (21 bp and 25 bp). (ii) Intra-CNV rearrangement. High-resolution aCGH showed a simple duplication at the *PLP1* locus in both J046 and J052. However, breakpoint junction sequences showed a 3.4-kb TRP buried within the duplication in J046. We also found that the duplication in J052 was actually constituted by two closely located duplications separated by a 3.3-kb normal region by breakpoint sequencing. This resulted in a DUP-TRP/INV-DUP and DUP-NML-INV/DUP structure respectively (INV, inverted). (iii) Inter-chromosomal and inter-locus rearrangement (Fig. 5). High-resolution aCGH revealed apparent single duplication in subject J003, J050 and J109 and an interspersed duplication in J077 at Xq22. The breakpoint junction sequences of these subjects suggested inter-locus and inter-chromosomal rearrangements, which were confirmed by genome-wide aCGH assay (Fig. 2 and Table 2).

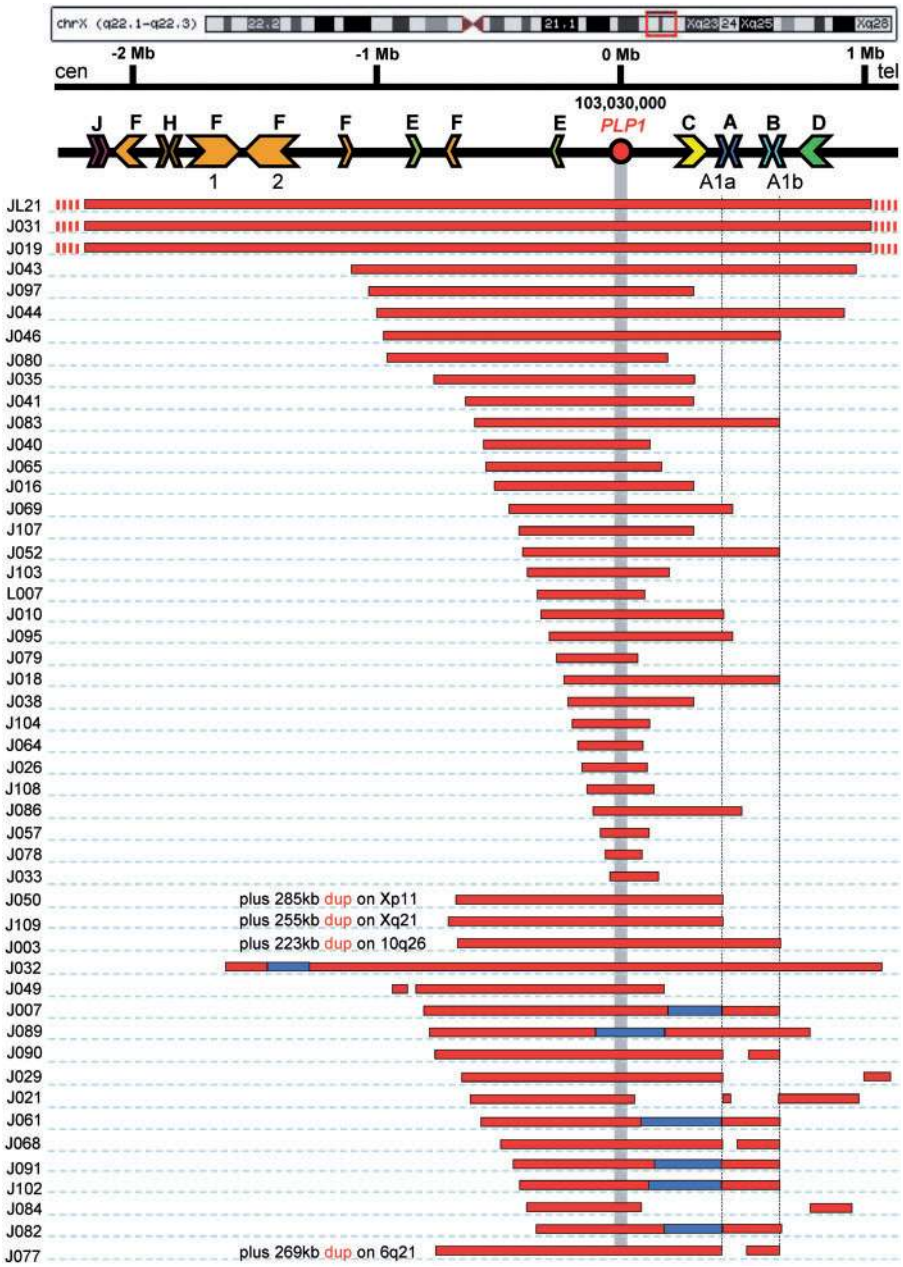

Figure 2. The CNV analysis by high-resolution aCGH identified simple and complex CNVs involving the *PLP1* gene. Human chromosome region Xq22.1–22.3 is illustrated. The red circle is the *PLP1* gene marked by a vertical gray shadow. The scale indicates the relative genomic distance to *PLP1*. The arrows in the same color indicate LCR pairs sharing high identity of DNA sequences and the arrow directions indicate the relative genomic orientations. Duplications are shown as horizontal red bars, while TRPs as blue bars. The subject indexes are given on the left. 35 subjects with aCGH-based simple *PLP1* CNVs are shown above, while 14 subjects with aCGH-based complex *PLP1* CNVs are shown below. Dotted red lines indicate the large duplications in Subjects JL21, J031 and J019. Additional CNV complexities of inter-locus or even inter-chromosomal rearrangements were revealed by breakpoint sequencing in Subjects J003, J050, J077 and J109.

Breakpoint characteristics and CNV mutational mechanisms

CNV breakpoint characteristics (Table 2, Supplementary Material, Tables S3 and S4) were informative for the investigations of CNV mutational mechanisms (1). All 28 sequence-based simple *PLP1* duplications were confirmed to be tandem duplications at the *PLP1* locus (Supplementary Material, Table S3). Microhomology (1–5 bp) or blunt ends occurred in the junction of 25/28 (89%) subjects (Supplementary Material, Table S3), indicating that microhomology-mediated break-induced replication (MMBIR) or non-homologous end-joining (NHEJ) as potential mechanisms underlying these non-recurrent junctions (1,18).

The remaining 21 sequence-based complex *PLP1* CNVs were generated by various mechanisms. Break-induced replication (BIR) paired with MMBIR and/or NHEJ (18,19) was a major contributor (52%) for complex and inverted junctions formation (in subject J007, J021, J032, J046, J052, J061, J068, J077, J082, J090 and J102) (Table 2 and Supplementary Material, Table S4). The MMBIR mechanism (18) also contributed, alone or paired with NHEJ, to form CGR with complex CNV junctions with or without

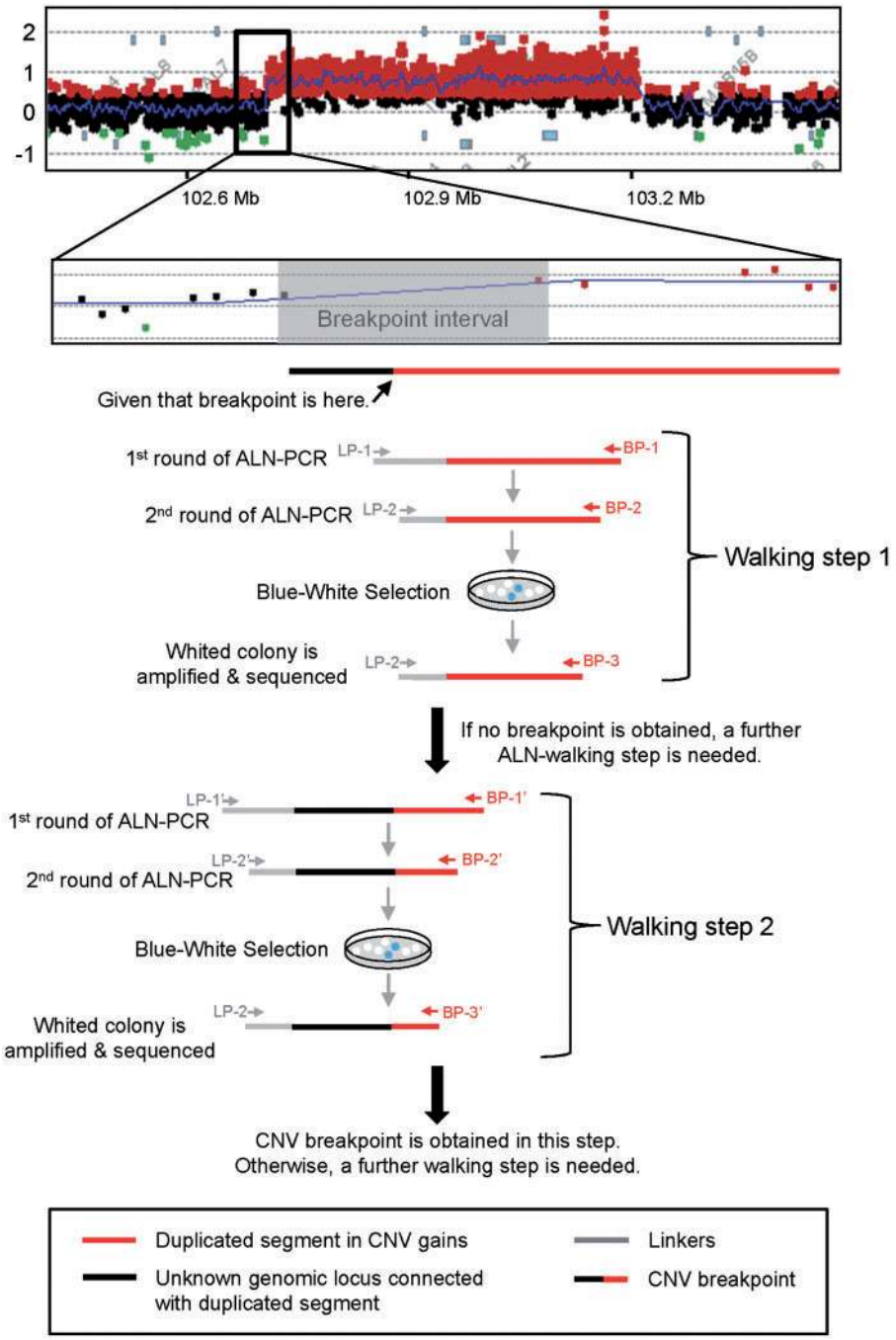

Figure 3. Schematic overview of ALN-walking for CNV breakpoint sequencing. An example of a *PLP1* gene duplication is illustrated. The patient DNA library with asymmetry linkers is constructed before ALN-walking. Each walking step includes: two rounds of asymmetry linker-mediated nested PCR, blue-white selection and clone sequencing. The first step of ALN-walking begins at the end of duplication toward breakpoint interval. The breakpoint-specific primers (BP-1, BP-2 and BP-3) were designed with melting temperatures of ~65.8, 58.9 and 55 °C, respectively according to the online tool of UCSC In-Silico PCR. The first round of nested PCR was conducted using DNA library and BP-1 (as a reverse primer) for linear amplifications of 12 cycles. After that, the linker-specific primer LP-1 was added into the reaction system as a forward primer, followed by 12 cycles for logarithmic amplifications. The PCR products were purified by the AMPure beads to remove primers and small fragments. The second round of nested PCR was conducted using the BP-2 and LP-2 primers, the purified products and the same conditions of the first round of nested PCR. The products of the second round were purified and used for blue-white selection and sequence analysis. Once no breakpoint but the human reference genome sequence is obtained, a further step of ALN-walking is needed.

inversions in subject J026, J029, J049 and J050 (Table 2 and Supplementary Material, Table S4). *Alu*- or long interspersed nuclear element (LINE)-mediated rearrangements (20) paired with MMBIR or NHEJ was also an alternative important apparent mediator of template switching resulting in complex CNV junction formation (in subject J079, J084 and J102) (Table 2 and Supplementary Material, Table S4).

Intriguingly, we found a high frequency (33%, 14/49) of single- or oligo-nucleotide changes at or near breakpoints in our PMD-associated CNVs (Supplementary Material, Tables S3

Table 2. Characteristics of sequence-based complex CNVs

Subject	Genomic Structure	Junction Position (GRCh37/hg19)	Orientation	Junction Characteristics	Mutational Mechanism
J003	NA	[Chr10:125372718]::[ChrX:103251274]	+	Microhomology	NA
		[ChrX:103251296]::[ChrX:102695417]	+	Microhomology	
		NA	+	NA	
J007	DUP-INV/TRP-DUP	[ChrX:102511787]::[ChrX:103202658]	−	Insertion	BIR paired with NHEJ
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J021	DUP-NML-INV/DUP-NML-DUP	[ChrX:103059232]::[ChrX:102788263]	+	Microhomology	BIR paired with MMBIR
		[ChrX:103494493]::[ChrX:103253841]	−	Blunt ends	and NHEJ
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J026	DUP-TRP-INV/TRP-DUP	[ChrX:103088820]::[ChrX:102848601]	+	Microhomology	MMBIR paired with NHEJ
		[ChrX:102848630]::[ChrX:102848589]	−	Microhomology	
		[ChrX:102848549]::[ChrX:102844927]	−	Insertion	
J029	DUP-NML-INV/DUP	[ChrX:104474209]::[ChrX:103242053]	−	Blunt ends	MMBIR paired with NHEJ
		[ChrX:102778931]::[ChrX:103832307]	−	Microhomology	
J032	DUP-INV/TRP-DUP	[ChrX:LCR-F1]&[ChrX:LCR-F2]	−	LCR pairs	BIR paired with *Alu*-mediated rearrangement
		[ChrX:101881087]::[ChrX:106950418]	−	*AluSg*::*AluY*	
J046	DUP-TRP/INV-DUP	[ChrX:101861310]::[ChrX:103220139]	−	Insertion	BIR paired with NHEJ
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J049	DUP-NML-INV/DUP	[ChrX:103173486]::[ChrX:102219438]	−	Microhomology	MMBIR
		[ChrX:102128770]::[ChrX:102308193]	−	Microhomology	
J050	DUP(Xp11)-NML-DUP(Xq22)	[ChrX:103235750]::[ChrX:41942956]	−	Microhomology	MMBIR
		[ChrX:41658078]::[ChrX:102615812]	−	Microhomology	
J052	DUP-NML-INV/DUP	[ChrX:LCR-BA1b]&[ChrX:LCR-AA1a]	−	LCR pairs	BIR paired with MMBIR
		[ChrX:102606726]::[ChrX:103246466]	−	Microhomology	
J061	DUP-INV/TRP-DUP	[ChrX:102847406]::[ChrX:103095764]	−	Insertion	BIR paired with NHEJ
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J068	DUP-TRP-DUP-NML-INV/DUP	[ChrX:102853876]::[ChrX:102871485]	+	Microhomology	BIR paired with MMBIR
		[ChrX:102871370]::[ChrX:103286451]	−	Microhomology	
		[ChrX:LCR-BA1b]&[ChrX:LCR-AA1a]	−	LCR pairs	
J077	DUP- NML-INV/DUP,DUP (Chr6)	[ChrX:102558373]::[Chr6:110690627]	+	Insertion	BIR paired with NHEJ
		[Chr6:110422092]::[ChrX:103264601]	−	Insertion	
		[ChrX:LCR-BA1b]&[ChrX:LCR-AA1a]	−	LCR pairs	
J079	DUP-NML-DUP-NML-INV/DUP	[ChrX:102708275]::[ChrX:102321549]	−	Microhomology	MMBIR paired with NHEJ
		[ChrX:102321588]::[ChrX:102292094]	+	Insertion	and Alu-mediated
		[ChrX:102292163]::[ChrX:103052112]	−	*AluSz6*::*AluJb*	rearrangement
J082	DUP-INV/TRP-DUP	[ChrX:102920976]::[ChrX:103190460]	−	Microhomology	BIR paired with MMBIR
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J084	DUP-NML-DUP	[ChrX:103423343]::[ChrX:102913718]	+	Insertion	LINE-mediated re-arrangement paired with NHEJ
		[ChrX:103129090]::[ChrX:103392917]	+	L1PA4::L1PA3	
J089	NA	NA	−	NA	NA
		[ChrX:LCR-C]&[ChrX:LCR-D]	−	LCR pairs	
J090	DUP-TRP-DUP-NML-INV/ DUP-NML-INV/DUP	[ChrX:102600614]::[ChrX:111314601]	−	Insertion	BIR paired with NHEJ and MMBIR
		[ChrX:111314621]::[ChrX:103234202]	−	Microhomology	
		[ChrX:103234178]::[ChrX:103284632]	−	Microhomology	
		[ChrX:LCR-BA1b]&[ChrX:LCR-AA1a]	−	LCR pairs	
J091	NA	NA	−	NA	NA
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J102	DUP-INV/TRP-DUP	[ChrX:102909074]::[ChrX:103135477]	+	Microhomology	BIR paired with MMBIR
		[ChrX:LCR-AA1a]&[ChrX:LCR-BA1b]	−	LCR pairs	
J109	DUP(Xq21)-NML-DUP(Xq22)	[ChrX:103233572]::[ChrX:87582110]	+	Insertion	*Alu*-mediated rearrangement paired with NHEJ
		[ChrX:87837563]::[ChrX:102613808]	+	*AluY*::*AluY*	

NA, not available.

::, exact junction positions determined by breakpoint-sequencing analysis.

&, predicted junctions according to reported CNV structure.

and S4). These genetic variations were absent in the databases of 1000 Genomes Project and ExAC (21,22). Because there were not enough genomic DNAs from the patients or their mothers, we did not confirm whether these nucleotide changes resulted from *de novo* mutations during CNV formation. These observations were presumably consistent with the error-prone replicative repair mechanism in CNV mutagenesis (23–25).

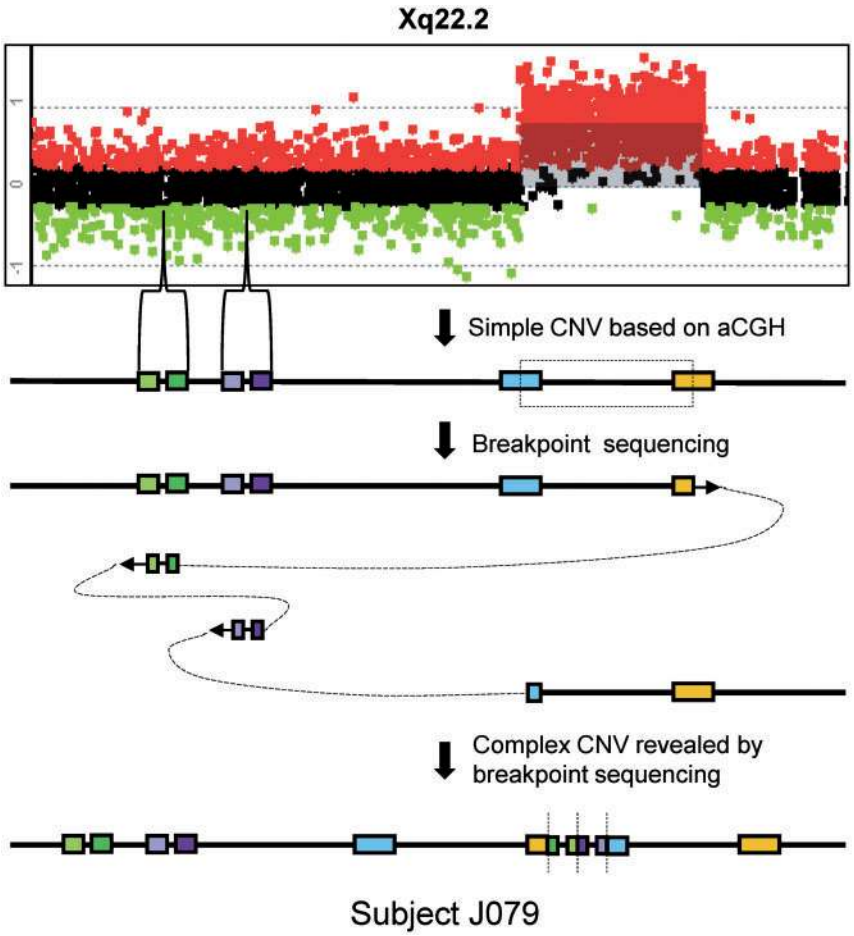

Figure 4. CNV breakpoint sequencing identified additional structural complexity in an aCGH-based apparent simple CNV. The aCGH results showed a 'simple' *PLP1* duplication in Subject J079. CNV breakpoint sequencing identified two short duplications inserted at the breakpoint junction, revealing a complex rearrangement event. These two additional duplications are less than 100 bp, which is beyond the resolution of aCGH. The corresponding genomic positions of these two duplications are shown in the aCGH plot.

CNV contents in PMD subjects

The comprehensively resolved breakpoints could clearly reveal the CNV structures and their affected genomic contents. The PMD-associated CNVs with only one breakpoint junction were tandem duplications (Supplementary Material, Table S3). In the CGR with multiple breakpoint junctions, there were multiple gained segments resulting from TRP, inverted duplication or inter-chromosomal rearrangements (Supplementary Material, Table S4). Based on the detailed breakpoint data, we could readily measure the lengths of gained segments (from 120 kb to 33.5 Mb) and their genomic contents (from 4 to 137 protein-coding genes) for all PMD-associated CNVs in this study (Table 3). These variable gain lengths and/or gained gene copy numbers might partially account for the diversity of clinical phenotype and severity between PMD subjects.

Correlation of CNV contents with clinical severity in PMD

In this study, 8 out of 49 subjects were excluded from the genotype–phenotype correlation study due to incomplete clinical information (subject J041, J068 and L007), young age (subject J018, J021 and J108), gender (a female subject JL021) and *PLP1* TRP (subject J089) (Table 1). The remaining 41 qualified subjects were categorized into two groups according to the phenotypic classification score (PCS) in this study (Table 1 and Supplementary Material, Table S5). Values from 0 to 2 of PCS represented severe phenotypes (17 subjects), while values from 3 to 5 suggested relatively milder phenotypes (24 subjects).

The potential correlation of CNV size with PMD severity was hypothesized and investigated in a previous study (26), but no statistical significance was observed based on the small sample size. Intriguingly, while *PLP1* gene duplications generally manifested PMD phenotypes in the X-linked recessive manner, a female patient caused by an extremely large *PLP1* duplication was also reported (27). This observation suggests that CNV sizes and genomic contents are likely to contribute to clinical severity in PMD. Therefore, we compared the sizes of PMD-associated CNVs between severe and mild PMD groups. Interestingly, the gained genomic segments are much longer in the severe group (PCS 0–2) than those in the mild group (PCS 3–5) ($P = 0.004$, Fig. 6A).

To further examine a potential molecular basis for the above correlation, we also counted the annotated protein-coding genes in the duplicated or triplicated genomic segments. Consistently, the subjects in the severe group gained more copies of coding genes than those in the mild group ($P = 0.009$, Fig 6B). Considering the fact that the majority of human coding

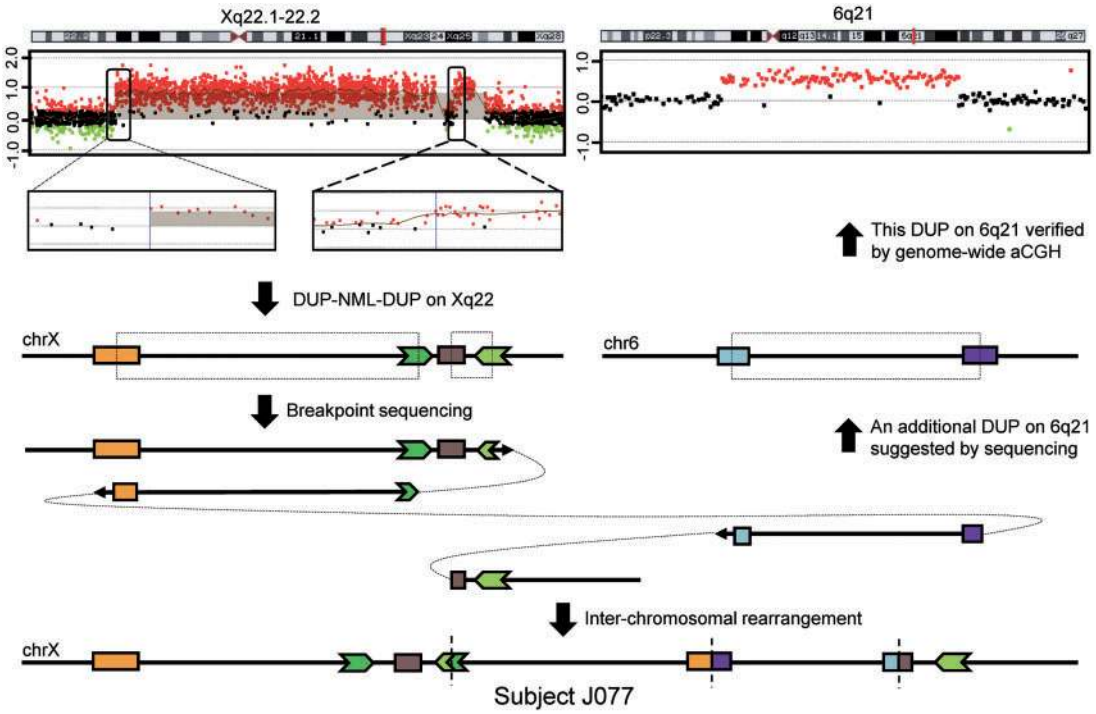

Figure 5. CNV breakpoint sequencing revealed an inter-chromosomal rearrangement beyond genomic coverage of *PLP1*–targeting aCGH. Target aCGH results of Subject J077 showed a complex structure of two interspersed duplications (DUP-NML-DUP) on Xq22. CNV breakpoint sequencing revealed that this complex CNV was not limited to the human × chromosome but an inter-chromosomal rearrangement between Xq22 and 6q21. This finding was subsequently confirmed by the Agilent human genome-wide aCGH assay. Besides the CNV complexities on Xq22, a 269-kb segment from 6q21 was also duplicated and inserted into Xq22 by a single rearrangement event. DUP, duplication; NML, normal.

genes are not dosage-sensitive and their copy-number changes may not be pathogenic (28), we investigated the dosage-sensitive genes affected by PMD-associated CNVs. The information of potential dosage-sensitive genes was obtained from the DECIPHER database (29). The genes with haploinsufficiency score of 10% or less (28,30) were taken as dosage-sensitive genes in this study. Remarkably, we found that the PMD subjects in the severe group gained more copies of dosage-sensitive genes than those in the mild group ($P = 0.015$, Fig. 6C). In addition to *PLP1*, two of these dosage-sensitive genes, *CASK* and *IL1RAPL2* are also implicated in neurodevelopmental disorders (Table 3) (31,32). Therefore, the large PMD-associated *PLP1* duplications also affecting *CASK* and/or *IL1RAPL2* may manifest more clinical severity based upon a mutational burden hypothesis in clinical variance (33–35). Intriguingly, such large *PLP1-CASK* and *PLP1-IL1RAPL2* duplications were found in 5 out of 17 subjects within the severe PMD group, while absent in the mild group (Fisher's exact test, $P = 0.008$) (Supplementary Material, Table S6). Our observations suggested that, besides the critical gene *PLP1*, other affected dosage-sensitive genes, especially those associated with neurodevelopment, might also be functional and partially contribute to variable phenotypes across PMD subjects.

Discussion

The aCGH and other array-based genome technologies are often used clinically for CNV analyses. Clinical targeted aCGH is frequently used for routine molecular diagnostics. However, these approaches cannot show the detailed CNV structure, especially the genomic orientations and base-pair resolution of breakpoint junctions. To completely resolve the CNV genotype and fully obtain CNV breakpoints at the base-pair level, efficient breakpoint sequencing methods are needed.

PMD is a classic type of genomic disorder associated with *PLP1* duplications on the human X chromosome (36). Because of complicated genome architecture such as paired LCRs, *Alu* and LINE repeats at the *PLP1* locus, PMD-associated CNVs are always non-recurrent, i.e. varying in size, structure and genomic content across PMD patients (15). In addition, the conventional approach of CNV breakpoint analysis using long-range PCR is only effective when the genomic structures can be anticipated (25). Therefore, alternative techniques are required to resolve complete structure and unanticipated complexity.

As previously reported, CNV breakpoint junctions can be covered and enriched by custom-designed DNA capture array (17,37,38). This strategy has also been employed, with modifications and improvements, in our cap-SMRT-seq method. For example, a unique index sequence was added to the DNA library for each subject, so that all the PMD samples can be pooled on the same capture array. Considering the complicated genome architecture at *PLP1* and its flanking regions, we employed the SMRT sequencing technology of Pacific Biosciences in cap-SMRT-seq to obtain long sequencing reads for breakpoint mapping. Notably, our cap-SMRT-seq is independent of prior knowledge or hypothesis of rearrangement structure; therefore, this method is applicable for CGRs and samples with complex CNV patterns. This is a technical advantage over long-range PCR, in which the hypothesized rearrangement structure is required before assay design. Due to the high costs of DNA capture array and SMRT sequencing, our cap-SMRT-seq can be applied when dozens of samples are ready for CNV breakpoint analysis. In this study, the cap-SMRT-seq method failed to

Table 3. CNV contents reveled by breakpoint analyses in 49 subjects with PMD

Subject	Length of gained segment (kb)	Gained gene		
		Coding gene	Dosage-sensitive gene including *PLP1*	Additional dosage-sensitive gene associated with neurodevelopment
J003	833	13	2	—
J007	814	13	2	—
J010	550	11	2	—
J016	608	12	2	—
J018	562	14	2	—
J019	5883	48	3	*IL1RAPL2*
J021	448	10	2	—
J026	244	7	2	—
J029	1005	12	3	*IL1RAPL2*
J031	4610	68	4	*IL1RAPL2*
J032	5491	53	3	*IL1RAPL2*
J033	183	5	2	—
J035	937	18	2	—
J038	422	8	2	—
J040	560	12	2	—
J041	841	20	2	—
J043	1823	29	2	—
J044	1387	26	2	—
J046	1448	15	2	—
J049	956	18	2	—
J050	905	13	3	*CASK*
J052	676	14	2	—
J057	165	5	2	—
J061	586	12	2	—
J064	282	7	2	—
J065	581	12	2	—
J068	389	17	2	—
J069	673	13	2	—
J077	975	18	3	—
J078	120	4	1	—
J079	344	8	1	—
J080	981	18	2	—
J082	417	10	2	—
J083	799	17	2	—
J084	246	7	2	—
J086	315	7	2	—
J089	1050	22	4	—
J090	644	13	2	—
J091	463	10	2	—
J095	552	11	2	—
J097	1176	18	2	—
J102	484	10	2	—
J103	459	9	2	—
J104	321	8	2	—
J107	524	9	2	—
J108	248	6	2	—
J109	875	12	2	—
JL21	33 519	137	12	*DACH2*, *PCDH19*, *IL1RAPL2*, *MID2*, *AMMECR1*, *PAK3*, *DCX*
L007	379	8	2	—

obtain breakpoint sequences in some of the PMD cases. It may be due to the pooled DNA libraries of different PMD cases and the diluted DNA fragments of patient-specific breakpoint junctions before DNA capture.

For the CNV studies based on a small sample size or even single case, our ALN-walking is a flexible option for breakpoint sequencing. This method is inspired by a previous study on virus integration using linker-mediated nested PCR (39). In our ALN-walking method, three tandem breakpoint-specific primers are specifically designed for each walking step with breakpoint intervals. Furthermore, the Y-shaped linkers are universal for any subject. During ALN-walking (Fig. 3), an asymmetric linear amplification using single breakpoint-specific primer was used to enrich breakpoint sequences. AMPure beads were used to purify the amplification products instead of agarose gel electrophoresis isolation. We recommend *LA Taq* polymerase in the PCR system because this polymerase will add an A deoxynucleotide to the 3′ end of DNA fragments thus

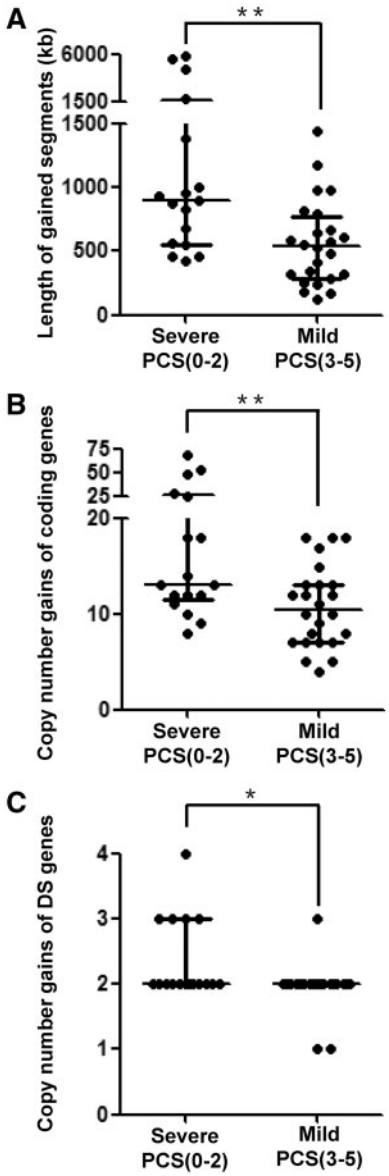

Figure 6. CNV distribution in the PMD subjects with variable clinical severity. The PCSs are available for 41 subjects with PMD in this study. These subjects can be categorized into two groups by PCS (scores 0–2 versus 3–5). Lower scores reflect more severe clinical conditions of PMD. The CNV properties, including length of gained segments **(A)**, copy number gains of coding genes **(B)**, dosage-sensitive/DS genes **(C)** and their distributions in two subject groups were investigated by the Mann-Whitney U test. The error bars show the medians with interquartile ranges (A and B) and median with range (C). The asterisks indicate the statistical significance (*$P < 0.05$; **$P < 0.01$).

producing an overhang end, which is needed in the subsequent T-A cloning for blue-white selection. The insertional fragments could be either breakpoint junctions or reference sequences within breakpoint intervals. The ratios of breakpoint junctions to reference sequences are summarized in Supplementary Material, Table S7. Here we recommend that at least 20 insertional fragments are required for each step of ALN-walking. The largest breakpoint interval that we resolved in this study is 4.5 kb, which required six ALN-walking steps to reach the breakpoint (Supplementary Material, Table S7). Therefore, ALN-walking also fits the currently used human genome-wide aCGH that can map CNV breakpoint intervals at the resolution of ~3 kb. An additional advantage of ALN-walking is the low cost for resolving CNV structures at the loci with complicated genomic architecture, in comparison to the approaches bases on high-depth whole-genome sequencing (40–43).

The complexities of PMD-associated CNVs have been comprehensively uncovered by efficient breakpoint sequencing using cap-SMRT-seq and ALN-walking. In total, we identified complex CNVs in 21 of 49 PMD subjects. As previously reported, inverted repeats at the *PLP1* locus can mediate complex CNVs with the 'DUP-TRP/INV-DUP' structure (19). Similarly, we found that the inverted repeats AA1a and BA1b were involved in six PMD-associated CNVs with TRP (subject J007, J046, J061, J082, J091 and J102) (Fig. 2 and Supplementary Material, Table S4). In addition, the inverted repeats F1 and F2 were associated with the triplicated segment in subject J032 (Fig. 2 and Supplementary Material, Table S4). All these seven CNVs with TRP were mediated by inverted repeats and showed a structure of 'DUP-TRP/INV-DUP' (19).

In the PMD cohort of this study, only subject JL21 is female. Therefore, we excluded this case from our genotype–phenotype correlation study. Generally, female individuals with *PLP1* gene duplications are only mutation carriers. However, the clinical features of JL21 met criteria for the severe connatal form of PMD (Table 1). For example, her first clinical evaluation at 14 months of age was global motor development delay and mental retardation. She showed nystagmus at 3 months of age. The ages of early milestones were 6 months of head control and 12 months of sitting without help. She also showed hypotonia. Brain MRI indicated hypomyelination. She died of respiratory failure at 18 month of age. The aCGH results showed that subject JL21 had an extremely large (33.5 Mb in length) duplication involving *PLP1* and many other genes in the human X chromosome (Fig. 2 and Table 3). Such a female PMD patient with large *PLP1* duplication has also been previously reported (27). Intriguingly, we also observed that PMD-associated CNVs affect more dosage-sensitive genes in the severe PMD group than in the mild one. In aggregate, all evidence suggested that, when *PLP1* is the critical gene for the PMD phenotype, the other dosage-sensitive genes and functional elements affected by PMD-associated CNVs are also important modifying factors potentially responsible for the variable clinical severities in PMD.

In conclusion, most (94%) of the PMD-associated CNV breakpoint junctions that failed to be elucidated by long-range PCR were successfully resolved at the base-pair level in this study. The ALN-walking method was efficient and applicable for CGRs in genomic disorders. The comprehensive sequence-based information of CNV structures can help uncover CNV mutagenesis and pathogenesis in PMD. Most importantly, the additional dosage-sensitive genes (especially those involved in neurodevelopment) affected by PMD-associated CNVs, were suggested to aggravate the clinical severity of PMD.

Materials and Methods

Subjects

Genomic DNAs were extracted from peripheral blood samples of 48 male subjects and one female subject with PMD using the DNeasy Blood and Tissue kit (QIAGEN). This study was approved by the institutional review boards of Peking University First Hospital and School of Life Sciences at Fudan University. Informed consents were obtained from each subject.

High-resolution aCGH analysis of PMD-associated CNVs

PMD-associated *PLP1* gene CNVs were initially identified by MLPA using the P071 Kit (MRC Holland). High-resolution PMD-target aCGH microarrays using probes spanning ChrX:98,028,855-113,513,744 (GRCh37/hg19, 4 × 44 K) (27) or ChrX:95,031,642-111,045,975 (GRCh37/hg19, 8 × 60 K) were designed to interrogate the *PLP1* gene and its flanking regions. Probes were designed using the Agilent eArray tool. In addition, Agilent 1 × 1M human genome-wide aCGH was used to confirm the structural complexities beyond the coverage of our PMD-target aCGH. Microarray hybridization and scanning was conducted according to the manufacturer's instructions. Agilent Genomic Workbench (version 7.0) was used for aCGH analyses.

CNV breakpoint sequencing by long-range PCR and sanger sequencing

For aCGH-based simple *PLP1* duplications, PCR primers were designed following the hypothesized tandem duplications and their head-to-tail rearrangement structures (14,15). For aCGH-based complex *PLP1* CNVs, primers were designed to amplify the unique breakpoint junctions based on various hypothesized rearrangement structures (19). Long-range PCR was performed using TaKaRa *LA Taq* polymerase and analyzed by Sanger sequencing (44). DNA sequences were analyzed by comparing to the human genome reference assembly (GRCh37/hg19) with the BLAT tool from the UCSC Genome Browser.

CNV breakpoint sequencing by cap-SMRT-seq

DNA capture array (NimbleGen, Roche) was designed to enrich breakpoint intervals and their 2-kb flanking regions (17,38). Any breakpoint mapped within genomic LCRs were excluded from capture targets. Subject DNA libraries were constructed according to the Rapid Library Preparation Method Manual (version GS FLX+ Series-XL+, May 2011). Target fragments were enriched by microarray hybridization, wash and elution according to NimbleGen Arrays User's Guide (version 3.2). The post-capture products were amplified by linker primers and then used to construct an emulsion PCR (emPCR) library according to GS FLX Titanium emPCR Method Manual (version April 2009). The emPCR products were sequenced by SMRT system (Pacific Biosciences). Breakpoint junction sequences were selected as follows. (i) Quality control on raw sequencing data: the number of bases with quality < 20 does not exceed 30% of the total number of bases in one read; and reads do not contain N bases. (ii) Cut off 300-bp segments of each end of reads longer than 1000 bp, and treat them as like-paired-end dataset; reads shorter than 1000 bp were grouped into a single-end dataset. (iii) Map like-paired-end and single-end data to the human genome reference assembly using aligner BLAT and BWA, respectively (45,46). For analysis of single-end data, we detected breakpoints from the single-end dataset via the split read method (47). Any breakpoint identified by cap-SMRT-seq was verified by Sanger sequencing.

CNV breakpoint sequencing by ALN-walking

The technical principle of ALN-walking is shown in Figure 3. For each subject, one microgram of genomic DNAs are sheared by Hydroshear (DIGILAB) into fragments of 2–3 kb in size. DNA library construction is according to the instructions of TruSeq DNA Sample Preparation Guide (Illumina, version July 2012). Breakpoint-specific primers (BP) were designed in the breakpoint-flanking regions within copy-number gained segments (48). Three adjacent tandem BPs are needed for each step of ALN-walking. Linker-specific primers (LP) are designed within one of the linkers and can be used as universal primers for all breakpoints. The melting temperatures of BPs should be approximate to those of LPs. The ALN-walking system and PCR program is shown in Supplementary Material, Figure S3. After size selection of PCR products, only the fragments of >1 kb were kept. Blue-white selection is used to separate the enrichment products. White clones are selected to amplify the insertional fragments using primers BP-3 and LP-2 with TaKaRa *LA Taq* polymerase (Fig. 3 and Supplementary Material, Table S2). The amplification products are analyzed by Sanger sequencing and comparing to the human genome reference assembly with the BLAT tool from the UCSC Genome Browser. If no breakpoint sequence but only reference sequences are obtained in 20 or more clones, a forward step of ALN-walking within the breakpoint interval using three new breakpoint-specific primers (BP-1′, BP-2′ and BP-3′) are needed until the breakpoint junction is obtained (Figs 1 and 3).

Interpretation of CNV mutational mechanisms

The mutational mechanisms of CNVs were investigated based on the sequence characteristics of breakpoint junctions. Blunt ends in the breakpoint junctions suggested the NHEJ mechanism. Microhomology of breakpoint ends suggested the mechanism of MMBIR or NHEJ. *Alu-Alu* mediated and LINE-mediated rearrangements could be identified by observing paired *Alu* and LINE repeats at breakpoint ends. The junctions located in LCR pairs could be caused by BIR, a homologous recombination mechanism to repair single-ended double-stranded breaks in a replication fork. For simple tandem duplications, only one of the above mechanisms was involved in a single mutation event. Nevertheless, for complex duplications with more than one breakpoint junction, different mechanisms could be involved in one mutation event.

Clinical severity and phenotypic classification in PMD

Detailed clinical information was available for 41 of our 49 subjects with PMD (Table 1). All 41 subjects had a clinical evaluation that included age of disease onset, head control, sitting, simple language expression and other symptoms with convulsion, dyskinesia, dysphagia, joint contracture, stridor, tremor and regression. Clinical severity of each PMD subject was investigated based on the PCS (Supplementary Material, Table S5) that was modified from the previous studies (26,49–51). PCS ranges from 0 to 5 in this study. Low PCS meant severe clinical phenotype.

Statistical Analyses

The Mann-Whitney U test was used to examine the difference of CNV properties including length of gained segments and copy of gained genes between the PMD groups with variable clinical phenotypes. The Fisher's exact test was used to examine the distribution difference of copy-number gains of neurodevelopment-related dosage-sensitive genes between the PMD groups with variable clinical phenotypes. Differences were considered statistically significant when $P < 0.05$.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We would like to thank Dr Pengfei Liu for critical comments. We also thank the Platform of Biochemistry and Plant Science, Fudan University for equipment support.

Conflict of Interest statement. None declared.

Funding

This work was supported by National Key Research and Development Program of China (2016YFC0905100), National Basic Research Program of China (2013CB945400), National Natural Science Foundation of China (31625015, 31521003, 31571297 and 81271257), National Institute of Neurological Disorders and Stroke (R01 NS058529), and National Human Genome Research Institute/National Heart Lung and Blood Institute (U54 HG006542).

Web Resources

The URLs for data presented herein are as follows:
Agilent eArray, https://earray.chem.agilent.com/earray
DECIPHER, https://decipher.sanger.ac.uk
OMIM, http://www.omim.org
UCSC Genome Browser, http://genome.ucsc.edu
UCSC In-Silico PCR, http://genome.ucsc.edu/cgi-bin/hgPcr

References

1. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
2. Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A. *et al.* (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, **66**, 219–232.
3. Driscoll, D.A., Budarf, M.L. and Emanuel, B.S. (1992) A genetic etiology for DiGeorge syndrome: consistent deletions and microdeletions of 22q11. *Am. J. Hum. Genet.*, **50**, 924–933.
4. Woodward, K., Kendall, E., Vetrie, D. and Malcolm, S. (1998) Pelizaeus-Merzbacher disease: identification of Xq22 proteolipid-protein duplications and characterization of breakpoints by interphase FISH. *Am. J. Hum. Genet.*, **63**, 207–217.
5. Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M. *et al.* (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.*, **38**, 24–26.
6. Luo, R., Sanders, S.J., Tian, Y., Voineagu, I., Huang, N., Chu, S.H., Klei, L., Cai, C., Ou, J., Lowe, J.K. *et al.* (2012) Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am. J. Hum. Genet.*, **91**, 38–55.
7. Glessner, J.T., Bradfield, J.P., Wang, K., Takahashi, N., Zhang, H., Sleiman, P.M., Mentch, F.D., Kim, C.E., Hou, C., Thomas, K.A. *et al.* (2010) A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am. J. Hum. Genet.*, **87**, 661–666.
8. Korbel, J.O. and Campbell, P.J. (2013) Criteria for inference of chromothripsis in cancer genomes. *Cell*, **152**, 1226–1236.
9. Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J. *et al.* (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.*, **86**, 749–764.
10. Wapner, R.J., Martin, C.L., Levy, B., Ballif, B.C., Eng, C.M., Zachary, J.M., Savage, M., Platt, L.D., Saltzman, D., Grobman, W.A. *et al.* (2012) Chromosomal microarray versus karyotyping for prenatal diagnosis. *N. Engl. J. Med.*, **367**, 2175–2184.
11. Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H. *et al.* (2015) TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *N. Engl. J. Med.*, **372**, 341–350.
12. Zhu, H., Shang, D., Sun, M., Choi, S., Liu, Q., Hao, J., Figuera, L.E., Zhang, F., Choy, K.W., Ao, Y. *et al.* (2011) X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. *Am. J. Hum. Genet.*, **88**, 819–826.
13. Gu, S., Szafranski, P., Akdemir, Z.C., Yuan, B., Cooper, M.L., Magrina, M.A., Bacino, C.A., Lalani, S.R., Breman, A.M., Smith, J.L. *et al.* (2016) Mechanisms for Complex Chromosomal Insertions. *PLoS Genet.*, **12**, e1006446.
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. Lee, J.A., Carvalho, C.M. and Lupski, J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, **131**, 1235–1247.
16. Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D. and Lupski, J.R. (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.*, **41**, 849–853.
17. Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J. and Hurles, M.E. (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.*, **42**, 385–391.
18. Carvalho, C.M. and Lupski, J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, **17**, 224–238.
19. 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.
20. 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.
21. 1000 Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map

of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
22. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
23. 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.
24. Carvalho, C.M., Pfundt, R., King, D.A., Lindsay, S.J., Zuccherato, L.W., Macville, M.V., Liu, P., Johnson, D., Stankiewicz, P., Brown, C.W. *et al.* (2015) Absence of heterozygosity due to template switching during replicative rearrangements. *Am. J. Hum. Genet.*, **96**, 555–564.
25. Beck, C.R., Carvalho, C.M., Banser, L., Gambin, T., Stubbolo, D., Yuan, B., Sperle, K., McCahan, S.M., Henneke, M., Seeman, P. *et al.* (2015) Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. *PLoS Genet.*, **11**, e1005050.
26. Regis, S., Biancheri, R., Bertini, E., Burlina, A., Lualdi, S., Bianco, M.G., Devescovi, R., Rossi, A., Uziel, G. and Filocamo, M. (2008) Genotype-phenotype correlation in five Pelizaeus-Merzbacher disease patients with PLP1 gene duplications. *Clin. Genet.*, **73**, 279–287.
27. Carvalho, C.M., Bartnik, M., Pehlivan, D., Fang, P., Shen, J. and Lupski, J.R. (2012) Evidence for disease penetrance relating to CNV size: Pelizaeus-Merzbacher disease and manifesting carriers with a familial 11 Mb duplication at Xq22. *Clin. Genet.*, **81**, 532–541.
28. Huang, N., Lee, I., Marcotte, E.M. and Hurles, M.E. (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.*, **6**, e1001154.
29. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M. and Carter, N.P. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.*, **84**, 524–533.
30. Ordulu, Z., Kammin, T., Brand, H., Pillalamarri, V., Redin, C.E., Collins, R.L., Blumenthal, I., Hanscom, C., Pereira, S., Bradley, I. *et al.* (2016) Structural chromosomal rearrangements require nucleotide-level resolution: lessons from next-generation sequencing in prenatal diagnosis. *Am. J. Hum. Genet.*, **99**, 1015–1033.
31. Bailey, K.A. and Aldinger, K.A. (2009) An X-linked microcephaly syndrome caused by disruptions of CASK implicates the CASK-TBR1-RELN pathway in human brain development. *Clin. Genet.*, **75**, 424–425.
32. Ferrante, M.I., Ghiani, M., Bulfone, A. and Franco, B. (2001) IL1RAPL2 maps to Xq22 and is specifically expressed in the central nervous system. *Gene*, **275**, 217–221.
33. Liu, P., Gelowani, V., Zhang, F., Drory, V.E., Ben-Shachar, S., Roney, E., Medeiros, A.C., Moore, R.J., DiVincenzo, C., Burnette, W.B. *et al.* (2014) Mechanism, prevalence, and more severe neuropathy phenotype of the Charcot-Marie-Tooth type 1A triplication. *Am. J. Hum. Genet.*, **94**, 462–469.
34. Yuan, B., Harel, T., Gu, S., Liu, P., Burglen, L., Chantot-Bastaraud, S., Gelowani, V., Beck, C.R., Carvalho, C.M., Cheung, S.W. *et al.* (2015) Nonrecurrent 17p11.2p12 Rearrangement Events that Result in Two Concomitant Genomic Disorders: The PMP22-RAI1 Contiguous Gene Duplication Syndrome. *Am. J. Hum. Genet.*, **97**, 691–707.
35. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y. *et al.* (2017) Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.*, **376**, 21–31.
36. Lupski, J.R. (2015) Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutagen.*, **56**, 419–436.
37. Sobreira, N.L., Gnanakkan, V., Walsh, M., Marosy, B., Wohler, E., Thomas, G., Hoover-Fong, J.E., Hamosh, A., Wheelan, S.J. and Valle, D. (2011) Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. *Genome Res.*, **21**, 1720–1727.
38. Wang, M., Beck, C.R., English, A.C., Meng, Q., Buhay, C., Han, Y., Doddapaneni, H.V., Yu, F., Boerwinkle, E., Lupski, J.R. *et al.* (2015) PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics*, **16**, 214.
39. Janovitz, T., Klein, I.A., Oliveira, T., Mukherjee, P., Nussenzweig, M.C., Sadelain, M. and Falck-Pedersen, E. (2013) High-throughput sequencing reveals principles of adeno-associated virus serotype 2 integration. *J. Virol.*, **87**, 8559–8568.
40. Talkowski, M.E., Ordulu, Z., Pillalamarri, V., Benson, C.B., Blumenthal, I., Connolly, S., Hanscom, C., Hussain, N., Pereira, S., Picker, J. *et al.* (2012) Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N. Engl. J. Med.*, **367**, 2226–2232.
41. Lou, H., Lu, Y., Lu, D., Fu, R., Wang, X., Feng, Q., Wu, S., Yang, Y., Li, S., Kang, L. *et al.* (2015) A 3.4-kb Copy-Number Deletion near EPAS1 Is Significantly Enriched in High-Altitude Tibetans but Absent from the Denisovan Sequence. *Am. J. Hum. Genet.*, **97**, 54–66.
42. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A. *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**, 344–347.
43. Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
44. Zhang, F., Seeman, P., Liu, P., Weterman, M.A., Gonzaga-Jauregui, C., Towne, C.F., Batish, S.D., De Vriendt, E., De Jonghe, P., Rautenstrauss, B. *et al.* (2010) Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *Am. J. Hum. Genet.*, **86**, 892–903.
45. Kent, W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*, **12**, 656–664.
46. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
47. Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S. and Zhu, M. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.*, **35**, 899–907.
48. Xiao, J., Zhang, L., Wang, J., Jiang, Y., Jin, L., Lu, J., Jin, L., Zhong, C., Xu, X. and Zhang, F. (2014) Rearrangement structure-independent strategy of CNV breakpoint analysis. *Mol. Genet. Genomics*, **289**, 755–763.
49. Cailloux, F., Gauthier-Barichard, F., Mimault, C., Isabelle, V., Courtois, V., Giraud, G., Dastugue, B. and Boespflug-Tanguy, O.

(2000) Genotype-phenotype correlation in inherited brain myelination defects due to proteolipid protein gene mutations. Clinical European Network on Brain Dysmyelinating Disease. *Eur. J. Hum. Genet.*, **8**, 837–845.

50. Hurst, S., Garbern, J., Trepanier, A. and Gow, A. (2006) Quantifying the carrier female phenotype in Pelizaeus-Merzbacher disease. *Genet. Med.*, **8**, 371–378.
51. Shimojima, K., Inoue, T., Hoshino, A., Kakiuchi, S., Watanabe, Y., Sasaki, M., Nishimura, A., Takeshita-Yanagisawa, A., Tajima, G., Ozawa, H. *et al.* (2010) Comprehensive genetic analyses of PLP1 in patients with Pelizaeus-Merzbacher disease applied by array-CGH and fiber-FISH analyses identified new mutations and variable sizes of duplications. *Brain Dev.*, **32**, 171–179.