Official journal of the American College of Medical Genetics and Genomics **ORIGINAL RESEARCH ARTICLE** Genetics inMedicine

Open

Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders

Rolph Pfundt, PhD[1], Marisol del Rosario, BSc[1], Lisenka E.L.M. Vissers, PhD[1], Michael P. Kwint, BSc[1], Irene M. Janssen, BSc[1], Nicole de Leeuw, PhD[1], Helger G. Yntema, PhD[1], Marcel R. Nelen, PhD[1], Dorien Lugtenberg, PhD[1], Erik-Jan Kamsteeg, PhD[1], Nienke Wieskamp, BSc[1], Alexander P.A. Stegmann, PhD[2], Servi J.C. Stevens, PhD[2], Richard J.T. Rodenburg, PhD[3,4], Annet Simons, PhD[1], Arjen R. Mensenkamp, PhD[1], Tuula Rinne, PhD[1], Christian Gilissen, PhD[1], Hans Scheffer, PhD[1,2], Joris A. Veltman, Prof. Dr[1,2] and Jayne Y. Hehir-Kwa, PhD[1]

Purpose: Copy-number variation is a common source of genomic variation and an important genetic cause of disease. Microarray-based analysis of copy-number variants (CNVs) has become a first-tier diagnostic test for patients with neurodevelopmental disorders, with a diagnostic yield of 10–20%. However, for most other genetic disorders, the role of CNVs is less clear and most diagnostic genetic studies are generally limited to the study of single-nucleotide variants (SNVs) and other small variants. With the introduction of exome and genome sequencing, it is now possible to detect both SNVs and CNVs using an exome- or genome-wide approach with a single test.

Methods: We performed exome-based read-depth CNV screening on data from 2,603 patients affected by a range of genetic disorders for which exome sequencing was performed in a diagnostic setting.

Results: In total, 123 clinically relevant CNVs ranging in size from 727 bp to 15.3 Mb were detected, which resulted in 51 conclusive diagnoses and an overall increase in diagnostic yield of ~2% (ranging from 0 to ~5.8% per disorder).

Conclusions: This study shows that CNVs play an important role in a broad range of genetic disorders and that detection via exome-based CNV profiling results in an increase in the diagnostic yield without additional testing, bringing us closer to single-test genomics.

Genet Med advance online publication 27 October 2016

Key Words: copy-number variants; diagnostic yield; exome sequencing; read depth; structural variation

INTRODUCTION

Copy-number variation is a large source of variation in the human genome. Within an individual genome, copy-number variants (CNVs) are estimated to result in a 1.2% difference in comparison to the reference genome.[1] CNVs not only contribute to normal genomic variation but are also an important cause of disease. Clinically relevant CNVs can be identified in patients with intellectual disability (ID) and other neurodevelopmental disorders; for example, they have been identified in an estimated 10–20% of ID patients.[2–5] As a consequence, microarray-based CNV profiling has been introduced as a first-tier diagnostic test for neurodevelopmental disorders in many laboratories in the past decade. Although research studies have implied an important role for CNVs in several other diseases,[6–8] the diagnostic genetic testing strategy for nonneurodevelopmental disorders has focused primarily on the detection of small genetic variants such as (point) mutations and insertion-deletions, using, for example, Sanger sequencing. With these disorders, microarray-based CNV profiling is not part of standard clinical practice. Recently, however, next-generation sequencing (NGS) approaches such as whole-exome sequencing have been implemented for these disorders in an increasing number of diagnostic laboratories. Although exome sequencing is used mostly to study the role of small variants, it is also suitable for the detection of larger structural variation such as CNVs.

The identification of CNVs in exome data can be performed using depth-of-coverage analysis. Several algorithms have been developed using a read-depth approach to identify CNVs in exome sequencing data.[9–12] Owing to the targeted nature of exome-capture platforms, the sensitivity of identifying CNVs containing three or more exons has been estimated to be 76%, with a specificity of 94%.[11] Alternative approaches for identifying structural variants in NGS data such as split-read or discordant pairs are less suited for identifying large events in exome sequencing data due to the fragmented nature of the data.

[1]Department of Human Genetics, Donders Institute, Radboud University Medical Center, Nijmegen, The Netherlands; [2]Department of Clinical Genetics, GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands; [3]Nijmegen Center for Mitochondrial Disorders, Department of Pediatrics, Radboud University Medical Centre, Nijmegen, The Netherlands; [4]Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands. Correspondence: Jayne Y. Hehir-Kwa (Jayne.Hehir-Kwa@radboudumc.nl)

Submitted 29 March 2016; accepted 31 August 2016; advance online publication 27 October 2016. doi:10.1038/gim.2016.163

In this study, we evaluated the clinical application of exome-based CNV profiling in a cohort of 2,603 individuals who underwent exome sequencing as a diagnostic test. These individuals were affected by a broad range of diseases that fell within 1 of 13 genetic-disorder categories ranging from neurodevelopmental and sensory disorders to congenital abnormalities to hereditary cancers.

MATERIALS AND METHODS

Sample selection

We collected 2,603 samples in 13 genetic-disorder categories: neurodevelopmental disorders, blindness, deafness, immunodeficiency, movement disorders, muscle disorders, renal disorders, craniofacial anomalies, disorders of sexual developmental, mitochondrial oxidative phosphorylation disorders, and metabolic disorders (**Figure 1a**). The germ line of 73 patients with microsatellite stable colorectal cancers was also analyzed. Finally, 119 patients with complex phenotypes were included (**Supplementary Table S1** online). The exome sequencing data for all samples were previously analyzed for (potential) pathogenic small variants (mutations) that were found within a corresponding panel of disease genes as previously described.[13] In addition, the majority of the neurodevelopmental disorders samples had previously screened negative for CNVs, based mostly on Affymetrix 250k microarray data using an average resolution of 100 kb for deletions and 150 kb for duplications. CNV analysis was performed for all exome samples. First, it was performed for CNVs overlapping genes associated with the genetic-disorder category. Second, exome-wide analysis with a resolution of 200 kb was performed anonymously for all patients.

Exome sequencing

The SureSelect V4 exome kit was used for enrichment of all samples. Sequencing was performed on the Solid 5500xl for 525 samples, with sequence reads mapped using Lifescope version 1.3 (Life Technologies, MA). For 2,078 samples, Illumina HiSeq2000 sequencing was performed and reads were mapped using bwa (version 0.7.7, Illumina, CA). The exome sequencing was performed with a minimum median coverage of 80×.

CNV calling

CNV calling was performed using CoNIFER (http://conifer.sourceforge.net/)[11] for two batches based on sequencing technology. All Illumina samples were processed in one batch (n = 2,078), and all 5,500 Solid samples were processed in a second batch (n = 525). CNVs with an absolute Z-score greater than 1.7 were considered for analysis. Samples with a total number of CNVs exceeding two standard deviations from the mean number of CNV segments were presumed to have poor quality and excluded from further analysis (n = 121). This equated to samples with more than 75 CNV segments sequenced on the Solid 5500xl platform and more than 97 CNV segments in samples from the Illumina HiSeq2000 platform. All deletion events were considered disruptive, as were duplications in known fully penetrant microdeletion/duplication regions[14] and intragenic CNV duplications. CNVs that overlapped known regions of partial penetrance were considered separately.[14]

To determine the presence of a potential batch effect, principal component analysis was performed on the normalized read-depth values (rpkms) of the Illumina samples. In total, the 2,078 samples were divided into 12 batches and colored accordingly. The first two components were then plotted according to the batch in which they were processed (**Supplementary Figure S1** online).

CNV annotation

CNVs were annotated based on the number of RefSeq exons affected, frequency of CNVs within the cohort, and overlap with disease gene panels per cohort.[15] Furthermore, the overlap was calculated with OMIM (Online Mendelian Inheritance in Man) disease genes,[16] DECIPHER known syndromes (https://decipher.sanger.ac.uk/),[14] and the American College of Medical Genetics and Genomics list of 56 medically actionable genes.[17]

Validation

Clinically relevant CNVs classified as pathogenic were validated using multiple validation strategies. Affymetrix CytoScanHD (Affymetrix, CA) microarrays were used to validate events larger than 200 kb when the target region had sufficient probe coverage. For all smaller CNVs, MLPA (MRC-Holland, The Netherlands) was used based on the availability of a commercial kit; otherwise, a custom Multiplex Amplicon Quantification (MAQ) assay (Multiplicon, Belgium) was performed.[18] In addition, the inheritance was determined for potentially clinically relevant CNVs in patients from the neurodevelopmental-disorders cohort. Read counts for the chr17 H2 allele were calculated using Samtools (http://www.htslib.org/)[19] based on the number of reads mapped to chr17_ctg5. PCR validation experiments were then performed to confirm the predicted H1/H2 genotype.

RESULTS

Our patient cohort consisted of 2,603 patients who underwent exome sequencing to identify the genetic origin of their disease. In total, 13 genetic-disorder categories were included, including deafness, blindness, metabolic disorders, immunodeficiency, movement disorders, renal disorders, and hereditary cancers (**Supplementary Table S1** online; **Figure 1a**). Exome CNV analysis of the exome sequencing data was performed in two tiers, whereby all known disease genes relevant to the patient's phenotype were investigated. The disease categories were used to determine which disease gene panel should be applied when performing this first-tier targeted analysis. A subset of 42 patients was analyzed for more than one gene panel (**Supplementary Table S1** online). In a second tier, exome-wide analysis was performed depending on the informed consent given. Exome sequencing data were derived from two different NGS platforms: the 5500xl SOLiD system (Life Technologies) (n = 525) and the HiSeq2000 system (Illumina) (n = 2,078). Exome-based CNV profiling by read-depth analysis was performed with the CoNIFER algorithm[11] for all 2,603 samples.

Qualitative analysis of the data set via principal component analysis of read depth per exon values (rpkm) showed that samples processed for exome sequencing in the same time period and the same sequencing platform tend to cluster (**Supplementary Figure S1** online). Therefore, data normalization was performed using the entire cohort of samples as a reference to cover most sources of variation, as is often also done for microarray data.[20] Exome-based CNV profiling of our data set was also influenced by the genomic architecture of the 17q21.31 locus (Koolen–de Vries syndrome region (OMIM 610443)). This region in the human genome harbors a common 900-kb inversion polymorphism that results in two major haplotypes: H1 and H2 (inversion). Exome-based CNV analysis revealed numerous deletions of this region, which turned out to be false-positive events. K-means clustering of rpkm values revealed three distinct read-depth clusters that were used to distinguish carriers of the H1 and H2 microinversion alleles in this 17q21 region, all with a normal copy number (**Supplementary Figure S2** online).

Across the complete cohort (n = 2,603), exome-based CNV profiling identified an average of six CNVs per patient with a mean size of 127 kb (range, 727 bp to 15.3 Mb), affecting, on average, three genes (range, 1 to 95 genes) (**Supplementary Table S2** online).

Diagnostic yield

We first identified rare deletions and intragenic duplication events that affected one or more disease genes relevant for the patient's disorder, using the corresponding disease-category gene panels as established for diagnostic exome sequencing interpretation within our department (http://www.genomediagnosticsnijmegen.nl/services/exome-sequencing-diagnostics). Intergenic duplications were excluded in this study because of the complexities in determining the consequences at both the functional and genomic architectural levels. CNVs were considered clinically relevant if they contained one or more disease genes for which the phenotype described in the literature corresponded with the patient's phenotype. For recessive disease genes, the corresponding exome sequencing data were analyzed for the presence of a second pathogenic variant in the gene, such as a point mutation,[21] indel,[22] or mobile element insertion[23] that could contribute to a compound heterozygous event. We identified clinically relevant CNVs in 123 patients from the total cohort of 2,603 patients (**Figure 1b**). For 51 patients, a conclusive diagnosis could be made based on a validated pathogenic CNV in which the genetic diagnosis and inheritance pattern matched the clinical phenotype (**Table 1**). For another 45 patients, a CNV affected a recessive gene matching the clinical phenotype but no second mutation on the second allele was identified in the exome sequencing data that could result in a possible diagnosis.[24] In addition, for 12 patients, a CNV was identified that affected a known susceptibility locus (a known CNV region with reduced penetrance) for the disorder under investigation, which also resulted in a possible diagnosis (**Table 2**).

Across the complete cohort, the highest diagnostic yields were obtained for deaf patients (5.8%, 13 of 223 patients), patients with a complex phenotype (5.5%, 10 of 183), and patients with a renal disorder (3.6%, 2 of 56 patients). The pathogenic CNVs ranged in size from 727 bp to 15.3 Mb, with a median size of 420 kb and a standard deviation of 1.9 Mb (mean size, 1.1 Mb). No clinically relevant CNVs were identified in disorders of sexual development (31 patients), craniofacial anomalies (24 patients), metabolic disorders (38 patients), and hereditary cancers (74 patients). The large difference in the diagnostic yield between the various disease categories may be explained by different mutational mechanisms causing these diseases and by the difference in the number of known disease genes per disorder. However, the relatively small sample size per disease group is likely to have also impacted these variable diagnostic yields (**Table 1**).

Of the 2,603 patients, a subset of 1,215 had a neurodevelopmental disorder (47% of the total cohort). Within this group, exome-based CNV analysis identified 16 events that were confirmed as de novo or resulted in loss of function of an autosomal-dominant disease gene linked to the clinical phenotype. This translates in a diagnostic yield of 1.3% in this disease category (**Figure 2a**; **Table 1**). This percentage is relatively low compared with the yield of >10% reported for genomic microarrays.[2–5] This can easily be explained by the fact that the vast majority of 1,215 individuals with a neurodevelopmental disorder had previously screened negative for CNV microarray analysis, resulting in a depletion of pathogenic CNVs in this patient group. Clinically relevant CNVs were observed only in patients who had previously been screened on a (low-resolution) microarray platform or in patients who did not receive microarray-based CNV profiling. Some of these clinically relevant CNVs affected known microdeletion and/or duplication syndrome regions (https://decipher.sanger.ac.uk/disorders#syndromes/overview), such as the 22q11 VCF/DiGeorge region (1 deletion, 3 duplications) (**Figure 2a**) as well as deletions encompassing genes previously linked to intellectual disability (*ALG1*, *MSX2*, *ANKRD11*, *GATAD2B*). Notably, three events larger than 4 Mb were identified (an 18p tetrasomy, a 7.5-Mb deletion of chromosome 10 (chr10:127.5–135.0 Mb), and an 8.4-Mb deletion of chromosome 3 (chr3:182.7–191.0Mb)).

Inheritance models of pathogenic CNVs

In 51 cases, a conclusive molecular diagnosis could be made based on the detection of CNVs in the exome sequencing data. These cases encompassed 32 CNVs occurring in a heterozygous state affecting a dominant disease gene and 19 pathogenic CNVs in a recessive disease gene occurring either homozygously or in a compound heterozygous state (in combination with a second pathogenic mutation of non-CNV origin). The cases included heterozygous deletions affecting dominant disease genes, such as *SCN1A*, causing autosomal-dominant Dravet syndrome (OMIM 607208) in two neurodevelopmental patients with epilepsy and a 804-kb deletion of the *PAX2* gene causing focal segmental glomerulosclerosis

Table 1 Number of patients per indication with a clinically relevant CNV that has been identified

		CNVs overlapping		Diagnostic yield	
	Number of patients	Autosomal-dominant genes	Autosomal-recessive genes with second hit	nr CNVs	% of Cohort
Craniofacial anomalies	31	0	0	0	0.0%
Disorders of sexual development	38	0	0	0	0.0%
Immunodeficiency	24	0	0	0	0.0%
Metabolic disorders	34	0	0	0	0.0%
Hereditary cancer[a]	74	0	0	0	0.0%
Renal disorders	56	1	1	2	3.6%
Complex phenotypes[b]	183	9	1	10	5.5
Mitochondrial disorders	142	0	1	1	0.7%
Muscle disorders	171	1	0	1	0.6%
Deafness	223	0	13	13	5.8
Movement disorders	217	2	0	2	0.9%
Blindness	237	3	3	5	2.1%
Neurodevelopmental disorders	1,215	16	0	16	1.3%
Total[c]	2,645	32	19	51	2.0%

CNV, copy-number variant.

[a]Microsatellite stable colorectal cancer. [b]Patients with complex phenotypes for whom a gene panel with all known OMIM (Online Mendelian Inheritance in Man) disease genes was used for analysis. [c]Forty-two patients were analyzed for multiple disease gene panels.

Table 2 Number of patients per indication with a CNV identified in a recessive disease gene or overlapping a region with partial penetrance

		CNVs overlapping	Candidate yield		Partial penetrance yield
	Number of patients	Autosomal-recessive genes	% of Cohort	CNVs overlapping regions with partial penetrance	% of Cohort
Craniofacial anomalies	31	0	0.0%	0	0%
Disorders of sexual development	38	0	0.0%	0	0%
Immunodeficiency	24	1	4.2%	0	0%
Metabolic disorders	34	0	2.9%	0	0%
Hereditary cancer[a]	74	1	1.4%	1	1.4%
Renal disorders	56	2	3.6%	1	1.8%
Complex phenotypes[b]	183	6	3.8%	2	1.1%
Mitochondrial disorders	142	2	1.4%	1	0.7%
Muscle disorders	171	4	2.9%	5	2.9%
Deafness	223	5	3.6%	0	0%
Movement disorders	217	2	1.4%	1	0.5%
Blindness	237	2	1.3%	2	0.8%
Neurodevelopmental disorders	1,215	20	1.6%	12	1.0%
Total[c]	2,645	45	2.5%	25	1%

CNV, copy-number variant.

[a]Microsatellite stable colorectal cancer. [b]Patients with complex phenotypes for whom a gene panel with all known OMIM (Online Mendelian Inheritance in Man) disease genes was used for analysis. [c]Forty-two patients were analyzed for multiple disease gene panels.

type 7 in a patient with a renal disorder (**Figure 2b**). Likewise, homozygous deletions in recessive conditions were detected, such as a homozygous deletion of exon 45–47 in the *USH2A* gene leading to Usher syndrome in a blind patient and a homozygous deletion of the *NPHP1* gene in a patient with nephronophthisis (OMIM 256100). Exome-based CNV analysis also enabled us to identify compound heterozygous events by combining CNV and SNV events. This resulted in detection of a heterozygous known pathogenic point mutation in exon 9 of the *USH2A* gene (NM_206933.2:c.1606T>C (p.(Cys536Arg)) and a deletion removing exons 25–28 of *USH2A* on the other allele in a patient with a visual disorder. Similarly, a deletion of the first 6 exons of *PANK2* was observed in combination with a hemizygous frameshift mutation in exon 6 (c.1317del p.Arg440fs (NM_153638.2)) on the other allele in another young patient referred for visual disturbances (**Figure 2c**).

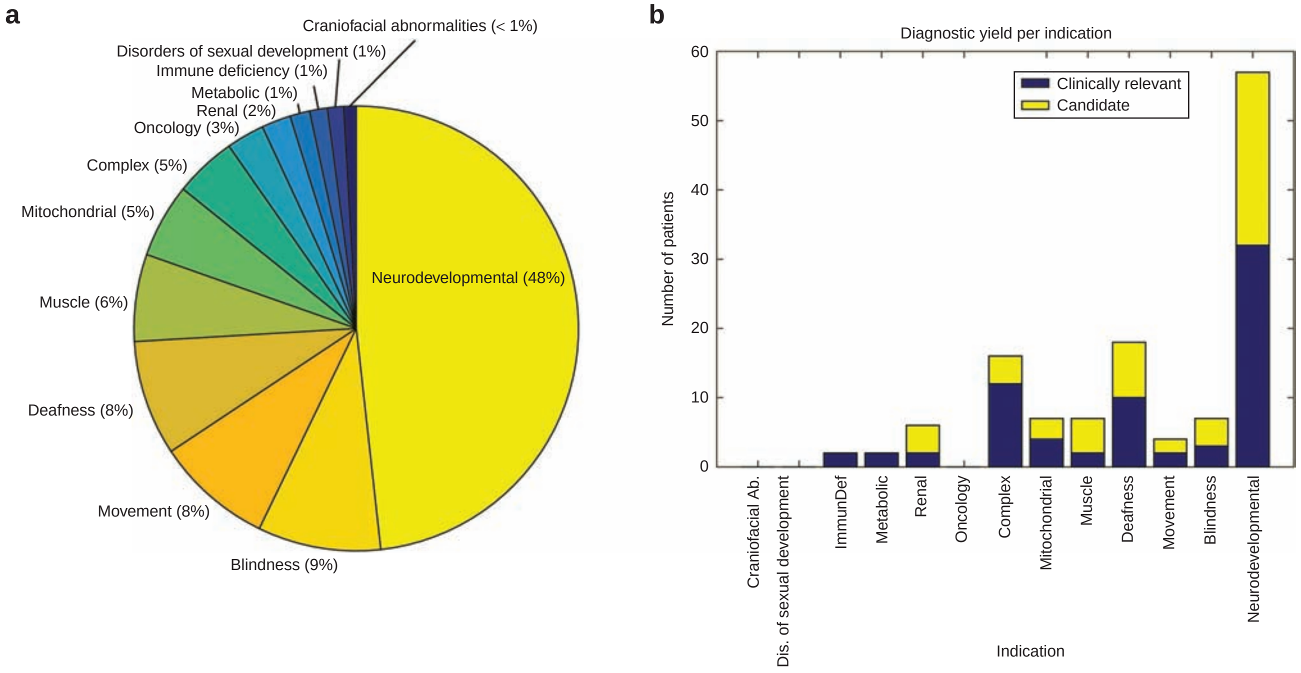


Figure 1 Detection of clinically relevant CNVs in a cohort of broad genetic disorders. (**a**) The cohort screened for large copy-number variants (CNVs) consisting of 2,603 patients spanning 13 heterogeneous disorders, including neurodevelopmental disorders, deafness, blindness, renal disorders, metabolic disease, immunodeficiency, muscle disorders, craniofacial anomalies, sex dysmorphy, movement disorders, and hereditary cancers. (**b**) The number of patients for whom a conclusive diagnosis could be made based on a pathogenic CNV and the number of candidate CNVs identified per cohort.

Recurrent CNVs

A genetic cause for deafness could be established in 13 patients. Remarkably, six of these patients carried an identical 150-kb homozygous deletion encompassing six genes, including *STRC*, a known recessive deafness gene (OMIM 603720). This deletion was also observed in four patients who carried a hemizygous missense mutation in the same gene on the other allele. The recurrent nature of this deletion, which we identified a total of 41 times in a heterozygous state (minor allele frequency 2%), is highly suggestive of a founder effect in the Dutch population (**Figure 2d**) and confirms the significance of this locus in the molecular diagnosis of this form of recessive deafness.[25] Additionally, two recurrent CNV loci affecting recessive genes *OTOA* and *NPHP1* were identified at a lower frequency.

Clinically relevant CNVs outside gene panels

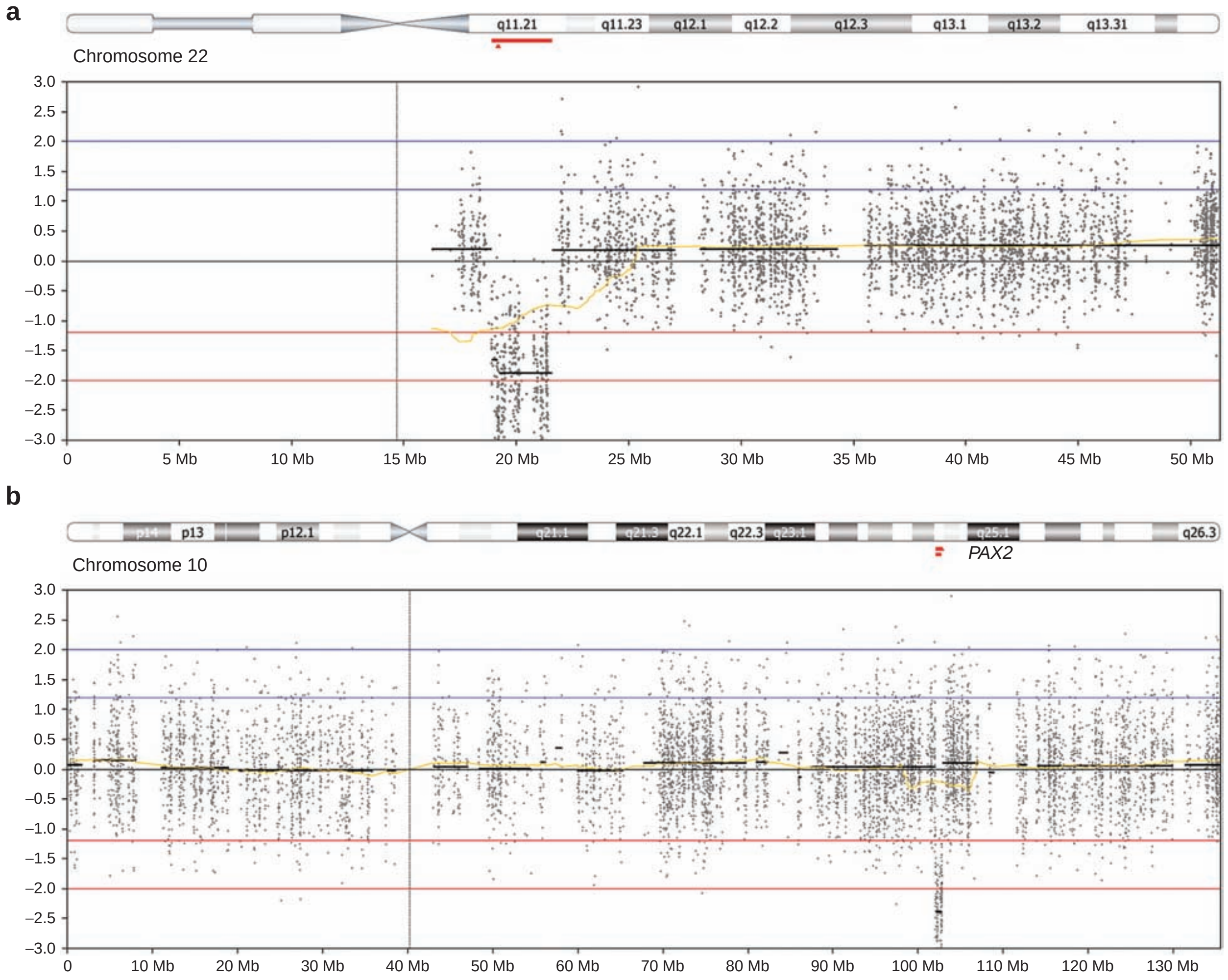
In addition to analyzing the exome sequencing data for clinically relevant CNVs that affected known disease genes, CNV analysis was expanded to an exome-wide CNV analysis for rare events larger than 200 kb. This analysis was expected to result in the identification of additional pathogenic CNVs that could be related to the clinical phenotype of the patient. In five patients with a neurodevelopmental disorder, a CNV was observed that affected a known microdeletion and/or microduplication syndrome region that did not encompass a known intellectual disability gene and therefore was not identified in the first-tier CNV analysis. In addition, the availability of exome sequencing data from patients as well as unaffected parents allowed the detection of de novo CNVs that also did not overlap a known ID gene but resulted in a likely positive diagnosis. Three patients in the remainder of the cohort received a conclusive diagnosis through exome-wide CNV analysis. This included a 1.6-Mb homozygous deletion of *HINT1* and *LYRM7* in a patient with suspected mitochondrial disease, a 377-kb deletion affecting *ATL3* in a patient with a movement disorder, and an Xp11.22 duplication (Xp11.22 microduplication syndrome OMIM 300705) in a male patient with a muscle disorder. These results indicate that, subject to proper counseling and patient's informed consent, CNV analysis preferably should not be limited to the known disease genes and should be expanded to include the entire genome, as is also done using microarray technology. Notably, as with all genome-wide tests, using a genome-wide CNV approach has the potential to uncover unsolicited findings, i.e., CNVs that are not directly related to clinical questions but are of medical importance to the patient. To estimate the incidence of such unsolicited findings during exome-based CNV analysis across the samples included in this study, we anonymized the genome-wide CNV data set to determine the number of CNV events that overlap with genes from the ACMG secondary-findings gene list.[26,27] This analysis identified four patients with an unsolicited finding (0.15% of the patient cohort) affecting the *PMS2*, *VHL*, *MYH7*, and *KCNH2* genes. Heterozygous loss of function mutations in *PMS2* have been found to contribute to hereditary nonpolyposis colorectal cancer 4 (614337).[28] Similarly, heterozygous mutations have also been described in *VHL* for Von Hippel–Lindau syndrome (193300). Heterozygous mutations in *KCNH2* have been attributed to both short (609620) and long (613688) QT syndrome, which can result in sudden cardiac death.[26,27] Furthermore, *MYH7* (160760) mutations can, for example, result in myosin storage myopathy, resulting in muscle weakness and atrophy.[29]

In addition to this ACMG list of genes, we screened for the incidence of CNV findings that corresponded to the known microdeletion and microduplication susceptibility locus (https://decipher.sanger.ac.uk/disorders#syndromes/overview), which has been reported to have partial penetrance. CNVs affecting these susceptibility regions have low penetrance and predispose to neurocognitive disorders, but they are also found at low frequency in unaffected individuals.[17] Anonymized genome-wide CNV analysis in our cohort revealed 25 patients who were found to harbor such a CNV with partial penetrance, including the 1q21, 16p13.11, and 16p11.2-p12.2 susceptibility loci[30,31] (**Supplementary Table S2** online). Notably, only 12 of these 25 patients were referred for exome sequencing because of a neurodevelopmental disorder, indicating that the other 13 patients may be nonpenetrant carriers of these susceptibility loci. Caution should be taken because other genes not linked to intellectual disability or neurodevelopmental disorders are also known to reside within these genomic loci, such as the *ABCC6* gene in the 16p13.11 recurrent microdeletion region, which causes autosomal-recessive pseudoxanthoma elasticum (264800).

DISCUSSION

In this study, we performed exome-wide CNV screening based on read-depth analysis of exome sequencing data in a large cohort of patients with a broad range of phenotypes. The aim was to assess the increase in diagnostic yield that can be achieved through exome CNV analysis. It has previously been demonstrated that exome-based CNV analysis is a robust approach to identifying genomic CNVs, although the specificity is highly dependent on the CNV detection algorithms used and estimations of the false-negative and false positive rates vary greatly.[9–12,32] Read-depth CNV analysis resulted in an overall increase in diagnostic yield of 2% in our cohort, representing a broad spectrum of genetic disorders. This is in line with the diagnostic yield obtained with high-resolution exon microarrays in similar cohorts.[32]

The majority of the 2,603 patients in the cohort analyzed in this study were not affected by a neurodevelopmental disorder. For most of these diseases, gene mutation analysis by traditional Sanger sequencing has been the main diagnostic strategy for identifying pathogenic genomic variants. CNV analysis has been largely neglected in these disorders, possibly because of limited knowledge of their clinical impact. In this study, we identified



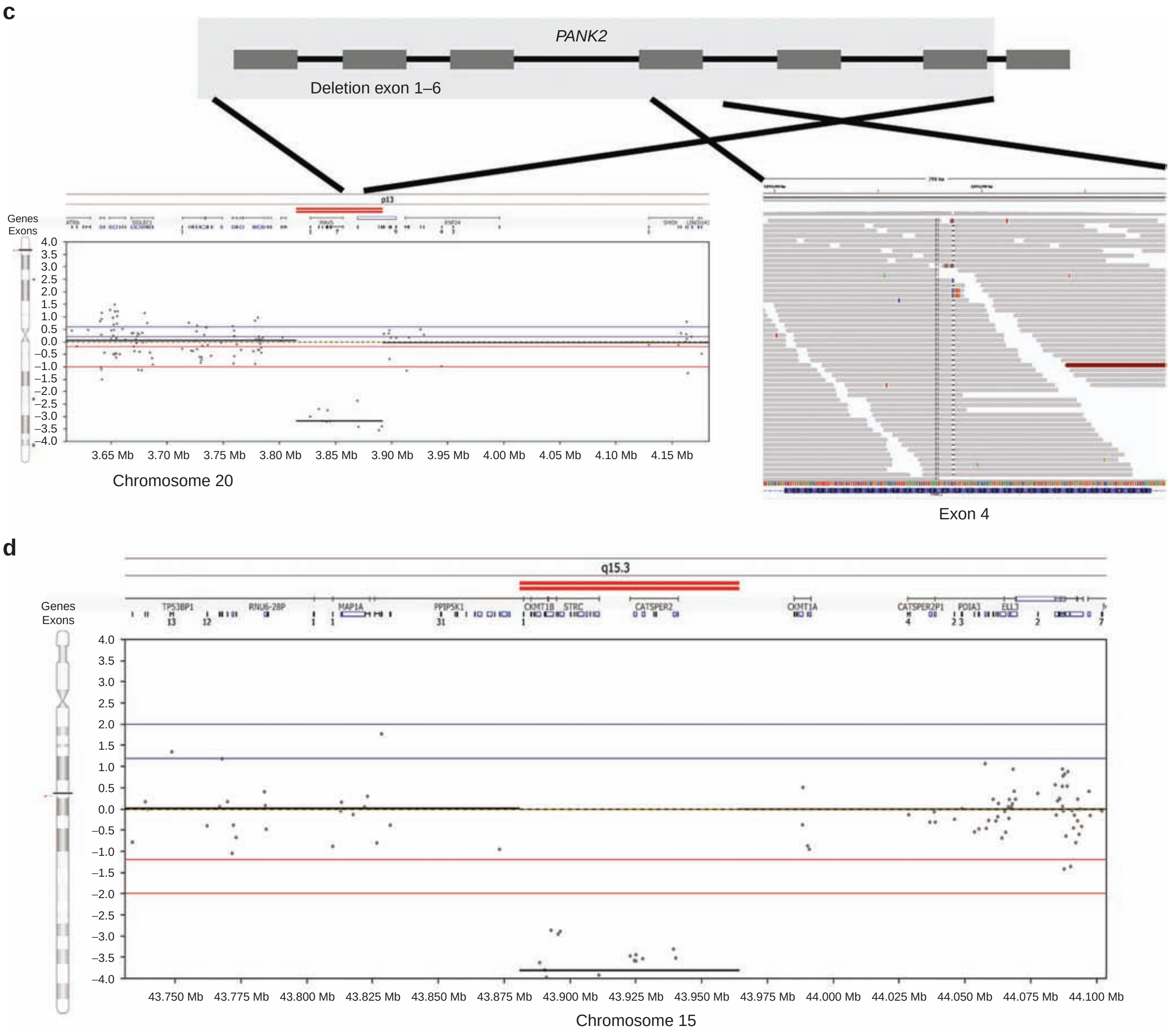


Figure 2 Examples of the inheritance patterns of pathogenic copy-number variants. (**a**) A deletion of 22q11.21 is identified in a patient with neurodevelopmental delay. (**b**) A heterozygous 800-kb deletion encompassing a dominant gene (*PAX2*) explains the phenotype of a patient with a renal disorder. (**c**) A compound heterozygous event including a deletion of the first six exons of *PANK2* (depicted in the lower left panel) in combination with a hemizygous frameshift mutation. c.1317del p.Arg440fs (NM_153638.2) reads with hemizygous mutation in exon 4 (depicted in lower right panel) in a patient with visual disturbances. (**d**) A homozygous deletion encompassing *STRC* results in a recessive disease pattern, explaining the deafness phenotype.

clinically relevant CNVs that previously would not have been discovered via targeted disease gene sequencing. These results demonstrate that large CNVs not only are an important cause of neurodevelopmental disorders but also evidently play an important role in many other genetic diseases. Of note, we identified 25 patients with known microdeletion and/or microduplication regions of partial penetrance.[14] Twelve of these patients were affected by a neurodevelopmental disorder; the remaining 13 were not. This may be due in part to the partial penetrance of these regions. However, due to the elevated rate of CNVs affecting these loci, it could suggest that the phenotypic spectrum of these syndromes is broader than initially suspected.[33]

The quality of CNV detection from exome sequencing data is affected by several technical factors, including the existence of a batch effect (**Supplementary Figure S1** online) as well as the underlying genomic architecture of regions and how this is represented in the reference genome. The latter was highlighted in our study owing to the co-occurrence of alternative H1/H2 haplotypes for the known 17q21 microdeletion syndrome.[34] The existence of alternative alleles and their inclusion in the reference genome continue to evolve[35] and fundamentally impact NGS data analysis when the sequence reads are mapped. In addition to the alternative ch17_ctg5, the human reference hg19 represents the HLA loci with six alternative contigs known to contain several disease genes relevant for severe combined immunodeficiency disorders. Hence, the accurate identification of both SNV and CNVs is important in these regions of known complexity. More recently, the GRCh38 reference genome includes 261 of these "alternative loci scaffolds" that represent 178 human chromosomal regions up to 1 Mb in size that exhibit sufficient variability to prevent adequate representation by the single primary sequence.[35] Although it is unlikely that all 178 of these loci will result in effects as strong as those demonstrated by the 17q21 region, it is important to consider the implications that these alternative loci will have for CNV detection.

In this study, we focused on the detection of large, clinically relevant CNVs from exome data via a read-depth strategy. This is one of several approaches for detecting structural variation from NGS data.[36] The detection of CNVs is also possible using discordant read pairs and a split-read strategy. However, algorithms relying solely on discordant read pairs suffer from low sensitivity due to the skewed insert size distribution as a result of shearing DNA fragments in exome sequencing procedures. Although methods using a split-read strategy have the promise of additionally detecting long indel and small deletions, particularly in the size range of 20 to 200 bp,[22,37] it is currently unknown what the impact of capture biases will be on the ability to sequence exon targets carrying structural variants in this size range. Given the high frequency of variants in this small size range, it is likely that split-read methods will be able to detect additional pathogenic variants.[38]

Although it has previously been shown that whole-genome sequencing and high-resolution microarray analysis are more sensitive than exome sequencing for the detection of subtle CNVs,[6,39,40] the sensitivity when using exome sequencing data for the detection of clinically relevant CNVs containing three exons is very high (96%).[9] With this study, we demonstrate the clinical applicability of exome-based CNV screening in genetic testing. Incorporation of CNV analysis in exome sequencing data-analysis pipelines, which until now have generally focused on SNV analysis, increases the diagnostic yield of exome sequencing by up to 6% (an average of 2%). Of importance, this increase in diagnostic yield is obtained without any additional direct laboratory costs. However, such a combined SNV and CNV analysis of exome data requires additional optimization of the exome data-analysis pipelines and subsequent interpretation, validation, and reporting stages. Nonetheless, combining SNV and CNV detection increases the suitability of exome sequencing as a first-tier diagnostic test for many, if not most, genetic disorders.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

This work was financially supported in part by grants from the Netherlands Organization for Scientific Research (NWO) (funding J.Y.H.-K. through grant 016.166.015) and the European Research Council (ERC) (funding J.A.V. through starting grant DENOVO 281964).

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Pang AW, MacDonald JR, Pinto D, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* 2010;11:R52.
2. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nat Genet* 2011;43:838–846.
3. Hochstenbach R, van Binsbergen E, Engelen J, et al. Array analysis and karyotyping: workflow consequences based on a retrospective study of 36,325 patients with idiopathic developmental delay in the Netherlands. *Eur J Med Genet* 2009;52:161–169.
4. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010;86:749–764.
5. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet Med* 2009;11:139–146.
6. Glessner J, Bick AG, Ito K, et al. Increased frequency of de novo copy number variations in congenital heart disease by integrative analysis of SNP array and exome sequence data. *Circ Res* 2014;115:884–896.
7. Orange JS, Glessner JT, Resnick E, et al. Genome-wide association identifies diverse causes of common variable immunodeficiency. *J Allergy Clin Immunol* 2011;127:1360–1367 e1366.
8. Shearer AE, Kolbe DL, Azaiez H, et al. Copy number variants are a common cause of non-syndromic hearing loss. *Genome Med* 2014;6:37.
9. de Ligt J, Boone PM, Pfundt R, et al. Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat* 2013;34:1439–1448.
10. Tan R, Wang Y, Kleinstein SE, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* 2014;35: 899–907.
11. Krumm N, Sudmant PH, Ko A, et al.; NHLBI Exome Sequencing Project. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;22:1525–1532.
12. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012;28:2747–2754.
13. de Ligt J, Willemsen MH, van Bon BW, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012;367:1921–1929.
14. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet* 2009;84:524–533.
15. Neveling K, Feenstra I, Gilissen C, et al. A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum Mutat* 2013;34:1721–1726.
16. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;33(database issue):D514–D517.
17. Green RC, Berg JS, Grody WW, et al.; American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
18. Kumps C, Van Roy N, Heyrman L, et al. Multiplex Amplicon Quantification (MAQ), a fast and efficient method for the simultaneous detection of copy number alterations in neuroblastoma. *BMC Genomics* 2010;11:298.
19. Li H, Handsaker B, Wysoker A, et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
20. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics* 2011;12:33–50.
21. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
22. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;25:2865–2871.
23. Thung DT, de Ligt J, Vissers LE, et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol* 2014;15:488.

24. Boone PM, Campbell IM, Baggett BC, et al. Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res* 2013;23:1383–1394.
25. Vona B, Hofrichter MA, Neuner C, et al. DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics. *Clin Genet* 2015;87:49–55.
26. Schimpf R, Wolpert C, Gaita F, Giustetto C, Borggrefe M. Short QT syndrome. *Cardiovasc Res* 2005;67:357–366.
27. Tester DJ, Will ML, Haglund CM, Ackerman MJ. Compendium of cardiac channel mutations in 541 consecutive unrelated patients referred for long QT syndrome genetic testing. *Heart Rhythm* 2005;2:507–517.
28. Worthley DL, Walsh MD, Barker M, et al. Familial mutations in PMS2 can cause autosomal dominant hereditary nonpolyposis colorectal cancer. *Gastroenterology* 2005;128:1431–1436.
29. Tajsharghi H, Thornell LE, Lindberg C, Lindvall B, Henriksson KG, Oldfors A. Myosin storage myopathy associated with a heterozygous missense mutation in MYH7. *Ann Neurol* 2003;54:494–500.
30. Männik K, Mägi R, Macé A, et al. Copy number variations and cognitive phenotypes in unselected populations. *JAMA* 2015;313:2044–2054.
31. Girirajan S, Rosenfeld JA, Cooper GM, et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* 2010;42:203–209.
32. Retterer K, Scuffins J, Schmidt D, et al. Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet Med* 2015;17:623–629.
33. Girirajan S, Brkanac Z, Coe BP, et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet* 2011;7:e1002334.
34. Koolen DA, Vissers LE, Pfundt R, et al. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 2006;38:999–1001.
35. Church DM, Schneider VA, Graves T, et al. Modernizing reference genome assemblies. *PLoS Biol* 2011;9:e1001091.
36. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11:31–46.
37. Rimmer A, Phan H, Mathieson I, et al.; WGS500 Consortium. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 2014;46:912–918.
38. Sudmant PH, Rausch T, Gardner EJ, et al.; 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;526:75–81.
39. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* 2014;511:344–347.
40. Turner TN, Hormozdiari F, Duyzend MH, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am J Hum Genet* 2016;98:58–74.