METHODS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Detection of Clinically Relevant Copy Number Variants with Whole-Exome Sequencing

Joep de Ligt,[1†] Philip M. Boone,[2†] Rolph Pfundt,[1] Lisenka E.L.M. Vissers,[1] Todd Richmond,[3] Joel Geoghegan,[3] Kathleen O'Moore,[3] Nicole de Leeuw,[1] Christine Shaw,[2,3] Han G. Brunner,[1] James R. Lupski,[2,4,5] Joris A. Veltman,[1] and Jayne Y. Hehir-Kwa[1*]

[1]Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands; [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas; [3]Roche NimbleGen, Madison, Wisconsin; [4]Department of Pediatrics, Baylor College of Medicine, Houston, Texas; [5]Texas Children's Hospital, Houston, Texas

Communicated by Johan T. den Dunnen

Received 5 April 2013; accepted revised manuscript 17 July 2013.
Published online 24 July 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22387

ABSTRACT: Copy number variation (CNV) is a common source of genetic variation that has been implicated in many genomic disorders. This has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool for CNV detection. More recently, whole-exome sequencing (WES) has been proven successful for the detection of clinically relevant point mutations and small insertion–deletions exome wide. We evaluate the utility of short-read WES (SOLiD 5500xl) to detect clinically relevant CNVs in DNA from 10 patients with intellectual disability and compare these results to data from two independent high-resolution microarrays. Eleven of the 12 clinically relevant CNVs were detected via read-depth analysis of WES data; a heterozygous single-exon deletion remained undetected by all algorithms evaluated. Although the detection power of WES for small CNVs currently does not match that of high-resolution microarray platforms, we show that the majority (88%) of rare coding CNVs containing three or more exons are successfully identified by WES. These results show that the CNV detection resolution of WES is comparable to that of medium-resolution genomic microarrays commonly used as clinical assays. The combined detection of point mutations, indels, and CNVs makes WES a very attractive first-tier diagnostic test for genetically heterogeneous disorders.
Hum Mutat 34:1439–1448, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: copy number variation; whole exome sequencing; clinical

Introduction

Whole-exome sequencing (WES) has revolutionized Mendelian disease gene identification by providing a powerful tool for exome-wide detection of single-nucleotide variants (SNVs) and small insertions and deletions (InDels) [Bainbridge et al., 2013; Bamshad et al., 2011; Gilissen et al., 2012; Hanchard et al., 2013; Ng et al., 2010; O'Roak et al., 2012]. In addition, WES is being introduced as a diagnostic procedure for genetically heterogeneous diseases in a number of laboratories [de Ligt et al., 2012; Hanchard et al., 2013; Rauch et al., 2012]. Structural variation such as copy number variants (CNVs), also contributes to these disorders [Cooper et al., 2011; Lupski, 2009; Stankiewicz and Lupski, 2010], and is currently not routinely assessed from WES data. The identification of CNVs, in addition to SNVs and InDels, would increase the versatility of WES as a genome-wide variant detection method in research and diagnostics. It would reduce the number of genomic assays required per patient to reach a diagnosis and create new possibilities to analyze the combined effects of SNVs and structural variation within an individual [Kurotaki et al., 2005].

Genomic microarray platforms based on either single-nucleotide polymorphisms (SNPs) or comparative genomic hybridization (CGH) have proven highly successful as a robust, high-throughput method for CNV detection [Boone et al., 2010; Pinkel et al., 1998; Schaaf et al., 2011; Vissers et al., 2003]. Advances in technology have resulted in an increase in the number of probes being included on a single array from hundred thousands of probes (medium resolution) to millions of probes (high resolution), resulting in both increased detection power and accuracy. The implication of CNVs in a wide range of congenital disorders including intellectual disability (ID) and developmental delay, as well as later onset common diseases such as schizophrenia and autism, has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool [Lupski, 2012; Mefford and Eichler, 2009; Miller et al., 2010; Vissers et al., 2010]. The resolution to detect CNVs using genomic microarrays is strongly governed by the spacing and number of interrogating oligonucleotide probes, and the microarray design [Boone et al., 2010; Hehir-Kwa et al., 2007; Pinto et al., 2011]. However, intragenic CNVs remain beyond the detection limit of most clinical genomic microarray analysis, with the exception of custom microarray designs with enhanced exonic coverage for selected disease genes [Boone et al., 2010].

In contrast to most available genome-wide microarrays, WES specifically targets exonic regions and is mostly blinded to the

Additional Supporting Information may be found in the online version of this article.

†These authors contributed equally to this work.

*Correspondence to: Jayne Hehir-Kwa, Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Medical Centre, Nijmegen, PO Box 9101, 6500 HB. E-mail: J.Hehir@gen.umcn.nl

Contract grant sponsors: European Union TECHGENE Project (Health-F5-2009-223143); GEUVADIS Project (Health-F7-2010-261123); European Research Council (DENOVO 281964).

remainder of the genome. The most widely applied massively parallel sequencing technologies sequence short reads (50–125 bp), either as fragments or as paired ends [Bamshad et al., 2011]. The most commonly applied methods for CNV detection in WES data are based on the analysis of the read depth, utilizing the number of fragments mapping within a genomic region as a measure of the amount of DNA present at the locus. This measure is used to determine a ratio between a test sample and reference samples [Haraksingh et al., 2011; Klambauer et al., 2012; Krumm et al., 2012; Plagnol et al., 2012], and results in an estimation of copy number for a given genomic segment, similar to what is used for array based platforms. Read count data can, however, be distorted by the capture procedure used to isolate the coding portions of the genome and by inaccurate alignment of sequencing reads to the reference genome. For example, it is well documented that the percentage of Guanine and Cytosine nucleotides in the region significantly influences the binding affinity during capture and sequencing [Metzker, 2010]. In addition, the presence of low copy repeats can negatively influence alignment of sequence reads to the reference genome and thereby distort copy number estimations of a region [Teo et al., 2012].

To date, CNV detection in next generation sequencing data has been largely limited to sporadic cases and healthy control populations in a research setting [Mills et al., 2011]. Here, we evaluate the detection of clinically relevant, rare *de novo* CNVs of varying size and copy number state via WES. We compare the performance of WES for CNV detection with that of both commercially available as well as custom designed, high-resolution array CGH enhanced for coding regions using up to 4.2 million interrogating oligonucleotides.

Materials and Methods

Sample Selection

Ten samples were selected that had previously been diagnostically reported as containing at least one clinically relevant, rare *de novo* CNV associated with ID, detected by routine microarray-based screening within the Department of Human Genetics, Radboud University Medical Centre, Nijmegen. These CNVs were chosen to represent a wide range of clinically relevant CNVs detected by microarray based analysis in our Genome Diagnostics division. The selected CNVs (1) contained at least one coding region, (2) were validated *de novo* using the same microarray platform on parental DNAs, (3) occurred across a variety of chromosomes, (4) ranged in copy number state from zero to three, and (5) ranged in genomic size from 15 kb to 24 Mb (Table 1). Eleven of these *de novo* CNVs were detected using an Affymetrix 250 k NspI (Affymetrix, Santa Clara, CA) microarray and one, in patient 1, with the Affymetrix 2.7 M microarray platform (Table 1).

WES and CNV Detection

WES was performed as described by de Ligt et al. (2012); in brief, genomic DNA from these 10 samples was isolated from blood using the QIAamp DNA Mini Kit (Qiagen, Venlo, The Netherlands). Exomes were enriched using a SOLiD-Optimized Agilent SureSelect Human Exome Kit, V2 (Agilent Technologies, Santa Clara, CA), followed by SOLiD sequencing using a 5500xl System (Life Technologies, Carlsbad, CA) to a median read depth of 67 across targeted regions. Read correction and mapping were performed with Lifescope v1.3 (Life Technologies), using default settings. The WES data were analyzed with four different published CNV detection programs; (1) cn.MOPS v1.6.4 [Klambauer et al., 2012], (2) CONTRA v2.0.3 [Li et al., 2012], (3) CoNIFER v0.2.0 [Krumm et al., 2012], and (4) ExomeDepth v0.8.4 [Plagnol et al., 2012] (see Supp. Methods), with unique hg19-based RefSeq gene exon definitions as target regions in the analysis.

Additional Genomic Microarray Studies

All samples were also analyzed on two independent, microarray platforms: (1) a high-resolution SNP microarray (Affymetrix CytoScanHD with 2.6 million probes; "CytoScanHD") (Affymetrix) and (2) a high-density CGH microarray enhanced for exonic regions (NimbleGen 4.2 million probe custom design; "ExonArray") (Roche NimbleGen, Madison, WI). Detailed experimental methods and computational approaches/software parameters are described in the Supp. Methods.

The aim of the ExonArray design was to cover each exon (Supp. Methods), and flanking sequence, with at least eight oligonucleotide probes. After testing and optimization (see Supp. Methods), the ideal coverage of eight or more probes was achieved for over 135,000 (~86%) exons; 249 (0.16%) of the exons could not be targeted at all. To test the sensitivity of the ExonArray, seven DNAs with 10 previously described CNVs (nine deletions and one duplication) with a median size of 8.5 kb (size range 1.6 kb–1.7 Mb) [Boone et al., 2013; Zhang et al., 2009] were analyzed (Supp. Fig. S1). NimbleGen performed the microarray experiments in a blinded fashion using mixed control DNAs. All the 10 CNVs were identified successfully indicating 100% sensitivity for these events, which were as small as 1.6 kb, five being smaller than 10 kb, and of which four encompassed only a single exon (Supp. Fig. S1).

CNV Annotation

Prior to annotation and interpretation, CNV calls resulting from both the WES approach and the ExonArray were subject to additional merging (Supp. Methods).

To facilitate interpretation, we annotated all CNVs for their gene content, (UCSC hg19 track GeneSymbols), the total number of genes, and the number of unique coding exons within the region. Since mapping artifacts can lead to false positive (FP) signals in sequencing data, the CNVs were annotated for features related to the uniqueness of the genomic region, the repeat content (simple and complex), and the percentage of SelfChain alignment in the region, based on the UCSC repeat tracks.

A reference set was generated to represent common CNV regions detected by the different platforms (both high-resolution microarrays and WES) and algorithms used in this study to determine which genomic regions were copy number variable (common CNVs). The reference set contained all events observed in more than one individual, by any specific platform in this study, as well as CNVs identified in our in-house set of control samples. This in-house dataset contains CNVs identified in 1,200 healthy individuals analyzed with the Affymetrix 6.0 SNP microarray platform [Franke et al., 2010] and 650 individuals analyzed with the Affymetrix CytoScanHD. The combined dataset included in total 23,125 gains and 56,066 losses.

Overall CNV Detection Power of WES

The false negative (FN) detection rate of WES was calculated by measuring the number of CNV events detected using the high-resolution microarray platforms that were missed by WES. To prevent overestimation due to platform design (exon targeted vs. whole genome), we accounted for both the exome enrichment targets and the detection power of WES. We selected CNVs that were identified

Table 1. Overview of the Detection of 12 Clinically Relevant *De Novo* CNVs

		Discovery microarray					WES read-depth algorithms			
Patient	Chromosome	Estimated start position (kb)	Estimated end position (kb)	CNV size (kb)	Copy number state	Nr. genes	CONTRA	cn.MOPS	ExomeDepth	CoNIFER
1	chr10	89,642.6	89,657.5	14.9	1	1[a]	–	–	–	–
2	chr19	33,371.1	33,394.2	23.0	0	1	–	–	V	V
3	chr8	77,745.6	77,795.2	49.6	1	1	–	–	V	V
4	chr17	1,203.6	1,516.5	312.9	3	8	–	–	V	V
5	chr16	29,673.2	29,988.3	315.1	1	16	–	–	V	V
6	chr15	43,759.8	44,862.9	1,103.2	1	24	–	–	–	V
7	chr2	233,166.3	233,886.7	720.5	3	16	–	–	V	V
8	chrX	6,495.3	7,951.7	1,456.4	0	5	–	–	V	V
9	chr2	239,952.7	241,373.1	1,420.5	3	14	–	–	V	V
	chr2	241,442.7	243,001.9	1,559.2	1	31	–	–	V	V
	chr15	60,489.7	62,906.5	24,603.6	3	210	–	–	V	V
10	chr20	77,771.0	102,374.6	2,416.8	3	91	–	V	V	V

CNVs as detected by the discovery microarray (hg19), genomic location, size, predicted copy number state and the number of genes in the region.
[a]A single exon deletion.
Detection by the different WES approaches; –, CNV is not detected with a minimum overlap of 30%, V, detected with a minimum overlap of 30%.

by at least two independent microarray platforms (minimum overlap of 30% of the CNV region, to allow for breakpoint inaccuracies due to the large differences in probe densities) and the CNV had to encompass at least three exons. For each CNV, the largest region, detected by the CytoScanHD or the ExonArray, was used for further analysis. After applying these selection criteria to the total set of 6,074 CNV identified by the different microarray experiments, the resulting consensus dataset contained 97 CNVs. Of these 97 consensus CNVs, 25 did not occur in the common CNV dataset and were considered rare CNVs. Consensus CNVs were only considered as positively detected by WES if a CNV was called in the same region and overlapped the consensus CNV region for at least 30%.

Breakpoint Analysis

To study the differences in detected CNV breakpoints across detection platforms, an overlap analysis was performed on the 11 clinically relevant CNVs. CNVs overlapping the discovery region were merged into a maximum confirmation CNV, and breakpoint differences were calculated based on the genomic coordinates of the two CNVs. The difference in genomic location was measured for each breakpoint by subtracting the genomic location as defined by the high-resolution array consensus from the location identified by the confirmation platform.

Data Availability

CNVs identified in this study by the different platforms have been submitted to dbVar under nstd84; sample identifiers correspond to those used in this paper. Detailed information on clinical presentation and the pathogenic event is available through ECARUCA for all patients under the following accession numbers (patient 1–10): 5042, 5045, 4785, 5044, 4545, 4487, 4581, 5043, 4452, and 4685, respectively. Raw data of the discovery microarray experiments are available in the Gene Expression Omnibus (GSE46060); sample identifiers correspond to those used in this paper.

Results

Our study aimed to investigate the diagnostic potential of CNV identification from short-read WES (SOLiD 5500xl) data. For this, we selected a set of 12 clinically relevant and validated, rare *de novo* CNVs detected using either an Affymetrix 250 k NspI or 2.7 M microarray, in 10 individuals with ID. This set of CNVs varied in genomic size and copy number state and incorporated both autosomal and X-linked CNVs (Table 1). WES was performed on all 10 samples and CNVs were called using four published CNV detection algorithms. In addition, high-resolution microarray experiments were performed using two independent platforms to experimentally assess the genome-wide true positive (TP), FP, and FN CNV detection rates of WES.

Detection of the Clinically Relevant CNVs

The four different WES CNV identification algorithms varied in their ability to correctly identify the 12 clinically relevant CNVs (Table 1). ExomeDepth and CoNIFER performed best, correctly identifying 10 and 11 of the 12 clinically relevant CNVs, respectively (Table 1, Fig. 1, for examples of WES-based CNV detection using CoNIFER). Of note, all WES algorithms failed to detect a clinically relevant single exon deletion (15 kb in size) in patient 1, which was originally detected using the Affymetrix 2.7 M microarray. While CONTRA and cn.MOPS often called a CNV in the relevant CNV region, the identified CNV was small and overlapped less than 30% (cut off threshold used for successful detection) with the interval identified by the discovery microarray. The copy number state reported by the WES-based CNV algorithms matched the copy number estimated by the microarrays for all CNVs.

Genome-Wide CNV Detection Using WES

The four different CNV WES detection algorithms varied widely in the total number of CNVs detected across the 10 samples; CONTRA identified 1,464 CNVs, ExomeDepth 1,482 CNVs, cn.MOPS 329 CNVs, and CoNIFER 65 CNVs in total (Supp. Table S1). All but one (99.9%) of the CNV events identified by CONTRA contained three or fewer coding exons. Similarly, many CNVs identified by cn.MOPS (56%) and ExomeDepth (58%) also contained three or fewer coding exons; in contrast, CoNIFER focuses more on detecting larger and rare CNVs, and detected only six (9%) such small CNVs (Supp. Fig. S2).

To evaluate the reliability of CNV identification using WES, we compared the results to CNVs detected by the different microarray platforms used in this study (Affymetrix 250 k NspI/2.7 M,

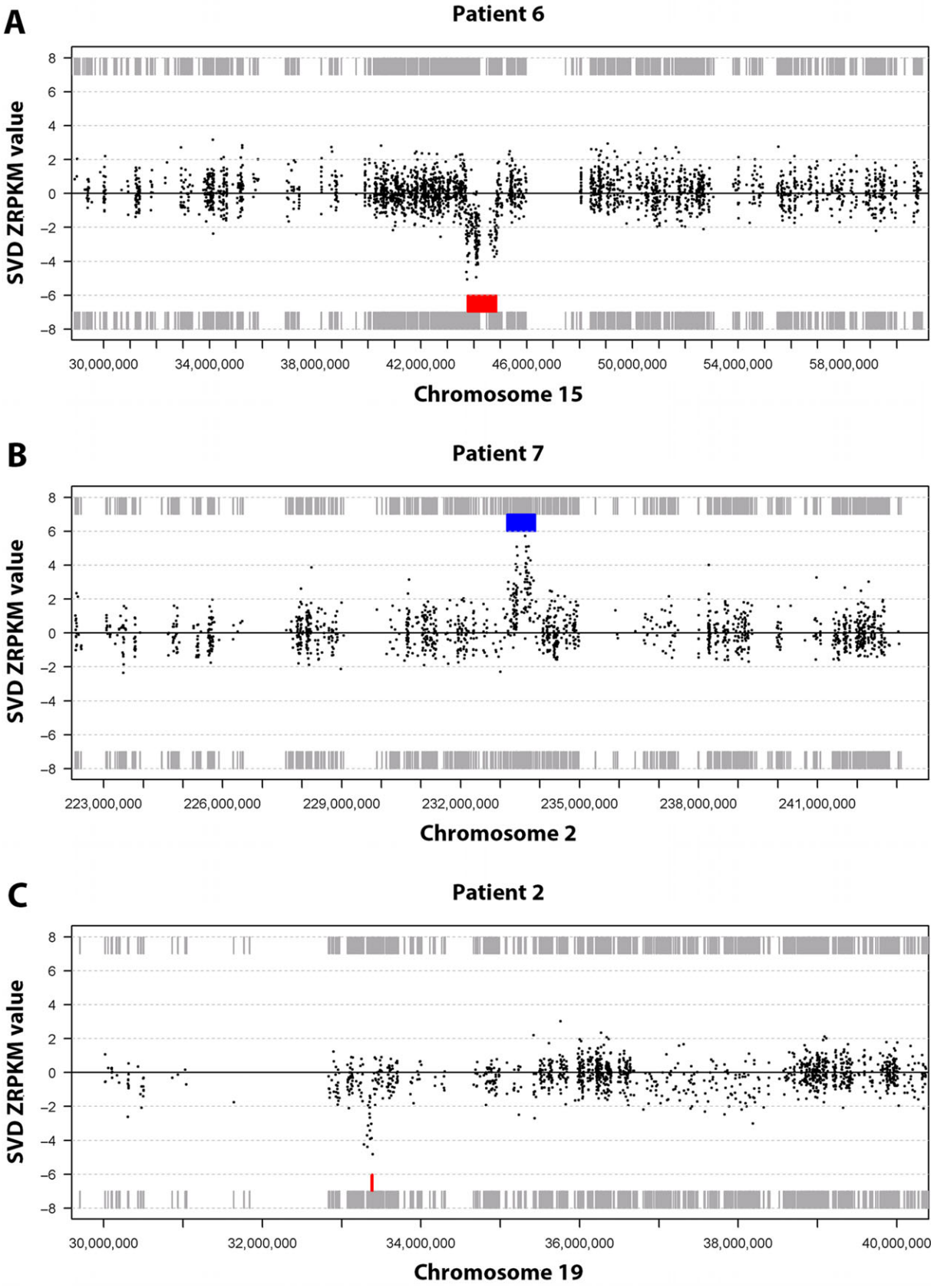

Figure 1. Detection of clinically relevant CNVs by WES. Black circles represent test over reference ratio values generated from WES data using CoNIFER; singular value decomposition (SVD) and *Z*-score adjusted read count per million (ZRPKM). Gray boxes indicate RefSeq gene exons; boxes above the ratio values represent CNV gains and below CNV deletions. **A:** A 1.1 Mb deletion including 24 genes. **B:** A 720 kb duplication including 17 genes. **C:** A 23 kb deletion containing two genes.

Table 2. Performance of WES CNV Detection Algorithms

Algorithm	Array confirmation of all WES CNVs	Array confirmation of rare WES CNVs
TP rate		
cn.MOPS (n = 329)	163 (49.5%)	33 (44.7%)
CoNIFER (n = 65)	**38 (58.5%)**	**28 (53.9%)**
CONTRA (n = 1,464)	248 (16.9%)	75 (11.0%)
ExomeDepth (n = 1,482)	234 (15.8%)	48 (6.6%)
Algorithm	Failed to confirm by WES of all array consensus CNVs n = 97	Failed to confirm by WES of rare array consensus CNVs n = 25
FN rate		
cn.MOPS	83 (86%)	23 (92%)
CoNIFER	69 (71%)	**3 (12%)**
CONTRA	97 (100%)	25 (100%)
ExomeDepth	**66 (68%)**	10 (40%)

CoNIFER has the highest TP rate and the lowest FN rate both for rare CNVs across the 10 patients. The TP rate measures the number of CNVs identified by each WES-based algorithm that were confirmed by a CNV called on a high-resolution microarray platform. The FN rate is calculated by creating a consensus of CNVs identified with the microarray platforms (*Materials and Methods*) and determining the number of CNVs not reliably detected by each WES-based algorithms. The best performing algorithm is highlighted in bold.

Affymetrix CytoScanHD and the NimbleGen 4.2 M ExonArray). In total, 38 of the 65 (59%) CNVs identified using CoNIFER were supported by one (n = 10) or more (n = 28) of the microarray platforms (Table 2). The confirmed events were larger (median 63.7 kb) and contained more exons (median 21.5) compared to the unsupported CNVs (median size 16.3 kb, median number of exons is 7). Similarly, 50% of the CNVs identified by cn.MOPS were supported by a microarray CNV, whereas a much smaller proportion of CNVs identified by CONTRA (17%) and ExomeDepth (16%) was supported by one or more microarray platforms (Table 2).

While genome-wide accuracy measures are an important indication of algorithm performance, in a clinical setting, it is important to consider the number of missed rare, genic events. To evaluate the FN rate of WES CNV detection, we investigated the detection of a CNV consensus set containing 25 rare coding CNVs detected by the two highest resolution microarray platforms used (ExonArray and CytoscanHD, see *Materials and Methods*). The overlap analysis showed the best detection rate for CoNIFER (88%), missing three of the 25 rare, genic events (Table 2). In general, the CNV events not detected by the WES approach using CoNIFER contained fewer exons (three, four, and 17 exons) than those CNVs that were detected (median 27 exons).

Overall, CoNIFER proved to be the most reliable CNV detection program for diagnostic applications based on a number of features: (1) most of the clinically relevant CNVs were detected (11 out of 12); (2) the highest percentage of CNVs were supported by an independent method (59%); and (3) the lowest number of rare consensus CNVs were missed (12%). All of the other algorithms had high FP rates and missed large numbers of rare CNVs, as well as the clinically relevant CNVs (Tables 1 and 2), making them unsuitable for diagnostic applications. Further analyses of WES CNV identification were based on the calls made by CoNIFER.

Determining the Accuracy of CNV Identification

The experimental design of WES leads to a nonuniform distribution of data points, focused only on the coding regions, whereas most genomic microarrays have a probe distribution containing a backbone covering the entire genome. To assess the effect of the unequal probe distribution of WES on the accuracy of CNV identification, we compared the breakpoints of the 11 clinically relevant CNVs detected across all four experimental platforms (discovery microarray, WES, CytoScanHD, and the ExonArray).

We generated consensus breakpoints based on the results from the highest resolution microarray platforms (CytoScanHD and ExonArray). The breakpoints of the WES CNV detection mapped within 200 kb of the consensus breakpoints for 18 of the 22 breakpoints (Table 3, Fig. 2, for example plots of CNV detection on different platforms). Three of the four breakpoints, which deviated more than 200 kb, occurred in regions with a much lower exon density as compared with the well mapped breakpoints (mean of 1.83 vs. 59.5 exons within 500 kb of the breakpoint region).

Predicting the Diagnostic Yield of CNV Identification Using WES

After determining the power to detect CNVs in WES data, we estimated the impact of using WES CNV detection on a larger set of samples within a clinical setting. For this, we compiled a list of 470 clinically relevant *de novo* CNVs detected by diagnostic microarray analysis in our center using a combination of Affymetrix 250 k NspI, 2.7 M and CytoScanHD microarray platforms. For each *de novo* CNV, the number of exons present in the sequencing capture set was calculated to determine if the CNV could be detected via WES. In total, 97% of the CNVs contained three or more exons, the minimum number required for WES-based CNV calling by CoNIFER. The majority of CNVs in this larger set of clinically relevant CNVs were larger than 200 kb in size (96%), whereas only half of CNVs from the rare consensus set (13/25) were in this size range. WES achieved a detection rate of 75% for CNVs smaller than 200 kb, and 100% for CNVs larger than 200 kb. When we apply these detection rates to the clinically relevant CNVs, it is predicted that 96% (453 CNVs) of these CNVs would have been successfully identified by WES. Based on the limited number of CNVs included in this study, this theoretical detection rate is in line with the observed experimental detection rate of 92% (i.e., 11 of the 12 clinically relevant CNVs being successfully detected).

Discussion

CNV is a common source of genetic variation that has been implicated in many genomic disorders [Cooper et al., 2011; Lupski, 2009; Stankiewicz and Lupski, 2010]. This has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool for CNV detection [Mefford and Eichler, 2009; Miller et al., 2010; Stankiewicz and Beaudet, 2007; Vissers et al., 2010]. The introduction of massive parallel sequencing approaches has provided

Table 3. Breakpoint Accuracy of WES CNV Detection

		High-resolution microarray consensus				WES detection[a]			
Patient	Chromosome	Start position (deviation) (kb)	Stop position (deviation) (kb)	Size (kb)	Nr. genes (min–max)	Start difference (kb)	Stop difference (kb)	Size difference (kb)	Nr. genes
1	chr10	x	x	x	x	x	x	x	x
2	chr19	33,352 (17)	33,396 (1)	44	2	−88.6	13.0	101.5	3
3	chr8	77,720 (0.4)	77,808 (0.3)	88	1	−100.6	1770.2	1870.8	4
4	chr17	1,142 (2.2)	1,494 (2.7)	351	9	−147.2	24.8	172.0	10
5	chr16	29,640 (11.6)	30,189 (11.3)	548	27–28	−92.9	10.9	103.7	30
6	chr15	43,714 (1.6)	44,863 (0.1)	1,150	24	−5.9	2.6	8.5	24
7	chr2	233,149 (1.1)	233,898 (2.0)	749	16–17	−34.7	1.7	36.4	17
8	chrX	6,453 (3.4)	8,001 (1.1)	1,548	5–6	−626.0	500.4	1126.4	12
9	chr2	239,945 (1.5)	241,423 (0.5)	1,478	18	71.6	−15.3	−86.9	15
	chr2	241,428 (5.0)	242,918 (134.5)	1,490	29–32	11.1	139.0	127.9	32
	chr15	77,765 (1.1)	102,415 (13.9)	24,650	200–211	5.7	80.3	74.6	213
10	chr20	60,463 (0)	62,940 (24.6)	2,478	91–92	1602.4	−40.8	−1643.2	43
Median difference (stdev)						88.5 (313)	24.7 (326)	103.7 (578)	3 (8)

The difference in breakpoints for the 11 clinically relevant CNVs detected on all platforms, compared with the average breakpoint positions detected by the two highest resolution platforms (CytoScanHD and ExonArray).
[a]WES CNV detection by CoNIFER.
Nr. genes, the number of genes based on RefSeq gene definitions (UCSC hg19); differences in breakpoint positions = WES – high-resolution microarray consensus; + indicates a higher genomic position; – indicates a lower genomic position. For size differences, – indicates undercalling of the CNV size by WES.
stdev, standard deviation.

a valuable tool for mutation identification in rare and genetically heterogeneous disorders [Bamshad et al., 2011; de Ligt et al., 2012; Gilissen et al., 2012; Gonzaga-Jauregui et al., 2012; Hanchard et al., 2013; Ng et al., 2010; O'Roak et al., 2012; Rauch et al., 2012]. For example, in a genetically heterogeneous disorder such as ID, a causal or candidate (*de novo*) mutation was identified in up to 38% of cases [de Ligt et al., 2012; Rauch et al., 2012], and it has been reported that an additional 10%–20% of ID cases can be explained by clinically relevant *de novo* CNVs [Mefford et al., 2012]. Therefore, the addition of CNV detection from WES data could achieve a diagnostic yield up to 58%, with a single test, for ID. This would represent the highest diagnostic yield of any current clinical genetic screening method for this disorder. A single genomic assay, which detects all forms of genomic variation, could decrease the time to obtain a molecular diagnosis, and reduce the diagnostic odyssey faced by patients and families.

Here, we evaluated the utility of WES to detect known clinically relevant CNVs in 10 patients. We tested four different CNV detection algorithms for WES data and compared their results to CNVs detected by three different genomic microarray platforms. These results provide insights into the possibilities and limitations of CNV detection using different experimental platforms currently available, as well as the performance of CNV identification algorithms with both WES data and high-resolution genomic microarrays.

Of the four algorithms tested in this study, CoNIFER was found to perform best with the highest TP rate and the lowest FN rate for the detection of rare coding CNVs. It is likely that CoNIFER performs especially well for rare CNVs due to the rigorous correction for systematic fluctuation, as well as enrichment, sequencing, and mapping biases, by singular value decomposition and the use of a Z-score approach, which corrects for positional fluctuation across samples. The FN rate of CoNIFER was greater for common CNVs, which is likely due to the Z-score approach applied for copy number estimation. The Z-score corrects for the fluctuation of a data point in the reference set; as a result, CNVs occurring in a region where reference samples are variable will have a lower Z-score compared with the same CNVs in a copy number stable region. Additionally, we limited our analysis to read algorithms suitable for short (50 bp) single-end reads as sequenced by SOLiD chemistry because read-depth algorithms are applicable to most WES approaches [Bamshad et al., 2011]. When longer reads or read pairs are available, more sophisticated methods can be used to increase the detection power for CNVs by combining different lines of evidence such as split read and clustering of discordant pairs [Teo et al., 2012] with a wider range of available programs [Duan et al., 2013].

Identifying CNVs in WES data is subject to a number of limitations due to the uneven spacing of exons, and thus data points, across the genome [Teo et al., 2012]. This affected the identification of the CNV segments, which in four cases were oversegmented and reported as several smaller CNVs, requiring merging during post-processing. Likewise, the unequal spacing of the genomic data points also influenced the identification of the CNV breakpoints. In general, the maximum possible size of the CNVs was reported; and in the absence of data points, segments were continued until a normal copy number signal was detected. Alternatively, CNV breakpoints can be identified based on the last occurrence of an aberrant copy number signal, the minimum CNV size. The difference between the maximum and minimum predicted CNV size as called by WES varied between 2.8 and 542.8 kb across the 11 *de novo* CNVs. Reporting both the maximum and minimum possible CNV size provides useful insights into the uncertainty of breakpoint predictions.

Most clinically relevant CNVs currently detected by routine screening are large (Supp. Fig. S2) and often contain multiple genes. Likewise, the CNVs identified in this study using WES were biased to larger CNVs containing multiple exons. However, our results using high-resolution microarrays indicate a large number of smaller single exon CNVs may exist within these samples (Supp. Table S2). Likewise, data from personal genomes [Wheeler et al., 2008], high-resolution CGH arrays [Conrad et al., 2010], and WES [Mills et al., 2011] indicate that the genomes of healthy individuals harbor 600–900 [Korbel et al., 2008; Levy et al., 2007] CNVs with a median size of 0.7 kb. Validation experiments of the 4.2 M NimbleGen microarray (ExonArray) showed that this platform has the potential to reliably detect known single exon deletions, and screening for exon level CNVs in a clinical setting has revealed multiple small, causal events [Boone et al., 2010; Whibley et al., 2010]. These small CNV events have been largely invisible to commercial genome-wide microarrays and remain challenging to detect through WES. While the detection specificity and sensitivity of the platforms used in this

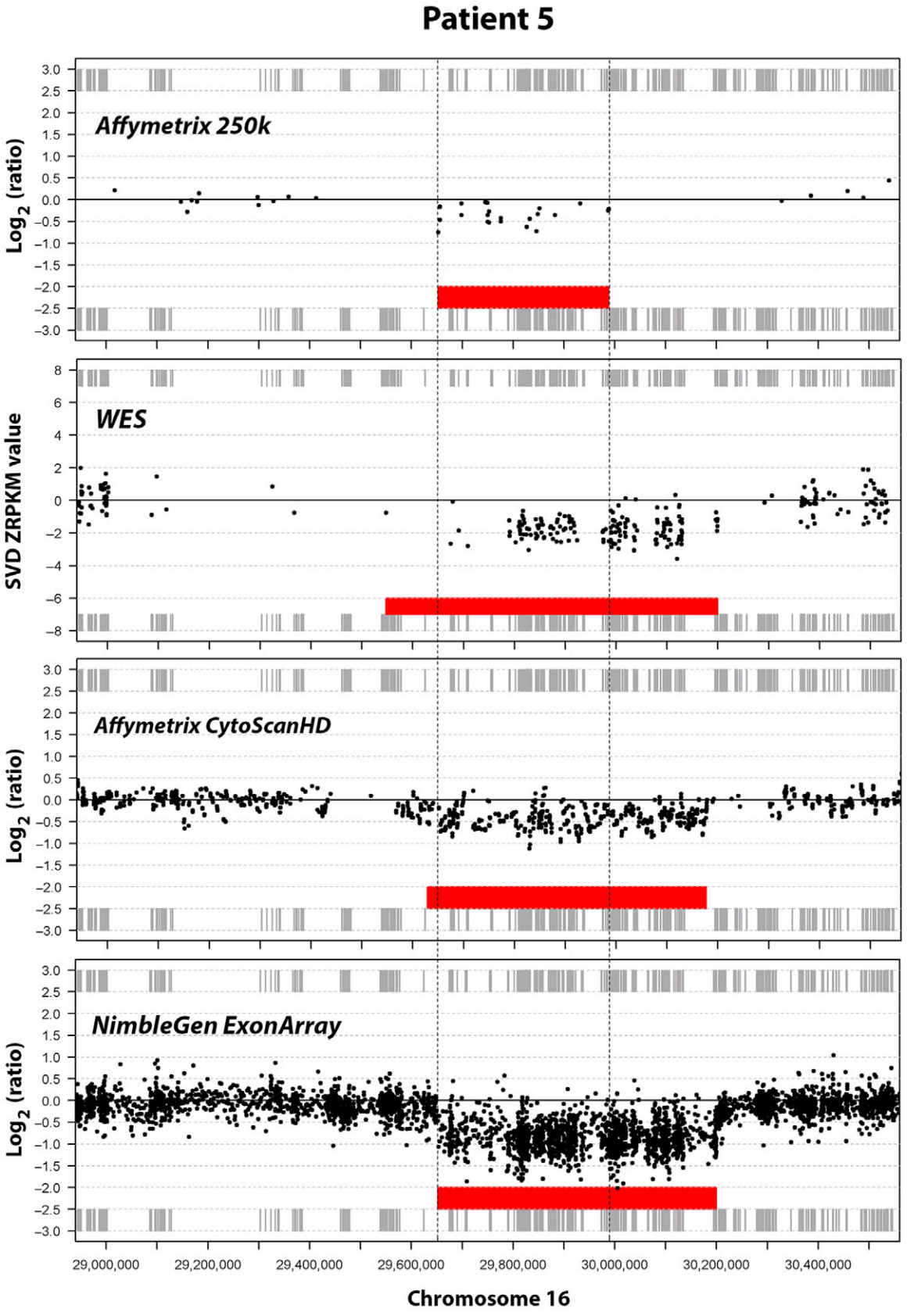

Figure 2. Detection of clinically relevant CNVs in two patients by the different platforms. Black circles are the $\log_2$ test over reference ratio values obtained through the different microarray experiments and singular value decomposition (SVD) and *Z*-score adjusted read count per million (ZRPKM) values for WES. Boxes below ratio values represent the CNV deletion as detected by the different platforms; gray boxes indicate RefSeq gene exons, and the vertical dotted lines indicate the minimally deleted region detected by the discovery microarray.

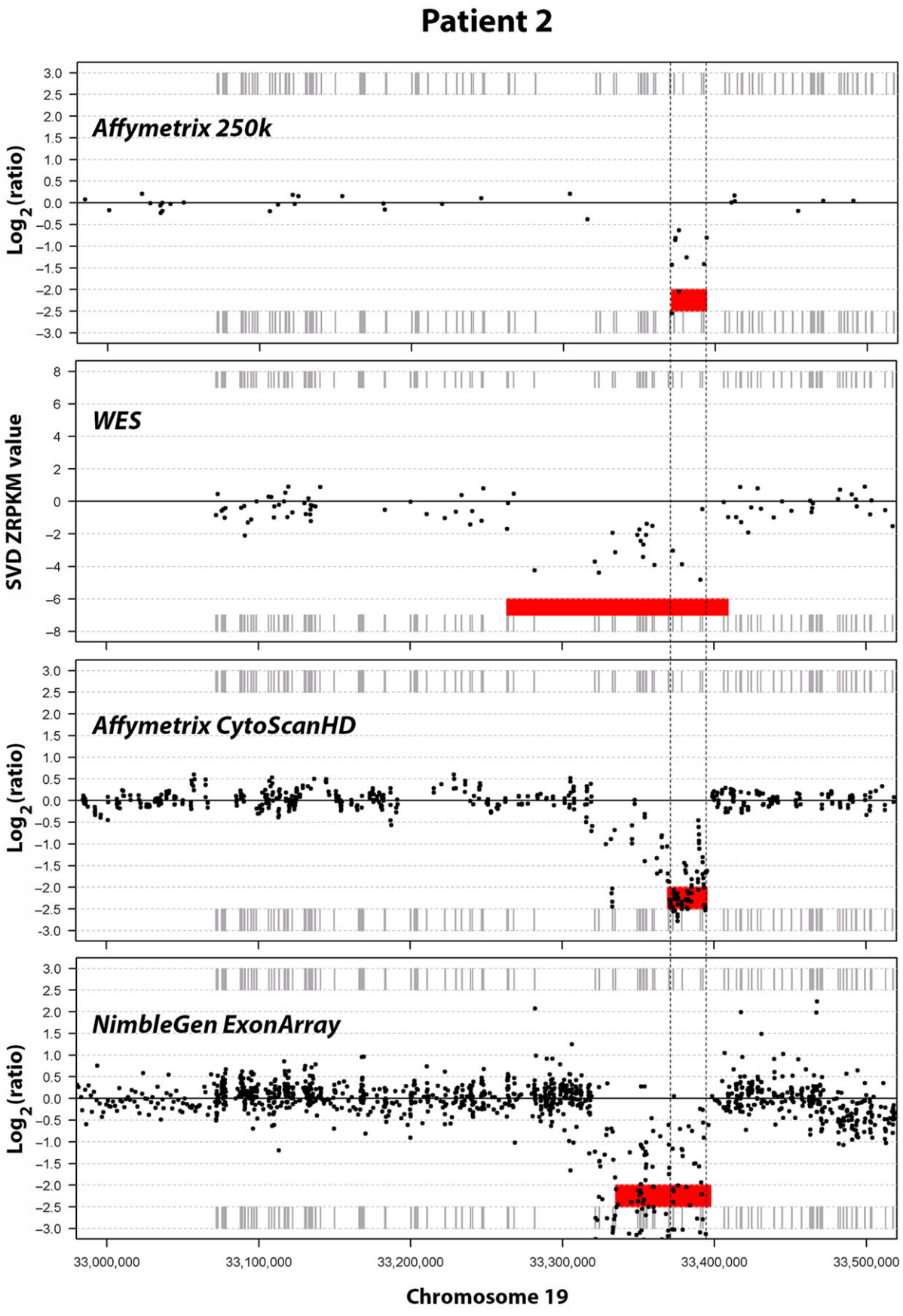

Figure 2. See figure legend on previous page.

study is unclear for these small CNVs, it is apparent that these events occur frequently and could contribute to the patient's phenotype. Thorough validation studies of these very small CNVs are required to establish their frequency and possible contribution to disease.

While the current detection power of WES, especially for single-exon CNVs, does not match that of high-resolution microarray platforms, we show that WES data are suited for the detection of large, rare, genic events that represent the majority of currently reported clinically relevant CNVs. A likely reason why the single exon 15 kb deletion included in this study was difficult to detect is that each exon represents a single data point. Detecting the difference between signal and experimental noise based on one data point requires very little fluctuation or noise. Possible solutions for larger exons include subdivision of exons into smaller regions to create multiple data points, or in the case of deletions, to include homozygosity data from SNVs into the detection algorithm. Ongoing developments in CNV identification algorithms will likely result in further performance improvements [Amarasinghe et al., 2013; Fromer et al., 2012].

The reliable detection of rare, genic CNVs is a valuable adjuvant tool within the clinical setting when WES data are available. Possibilities to enhance the detection power for CNVs of WES approaches include larger capture kits, the addition of a genomic backbone to improve genome-wide resolution, and/or the addition of intronic capture sequences to improve the accuracy in determining which exons are affected by a CNV. Improvements in data analysis could be made by applying more sophisticated normalization methods to account for biases introduced during the capture and sequencing procedures. In addition, current WES CNV detection algorithms used in this study are limited in breakpoint accuracy by the read-depth approach and could be further improved by incorporating information from genotypes, split-reads, and read-pair information to increase the detection power of WES for CNVs [Mills et al., 2011]. While these improvements are of great benefit to further increase WES-based CNV detection, the results presented in this study show that CNV detection resolution of exome sequencing is already comparable to that of medium-resolution genomic microarrays currently used as clinical assays.

Acknowledgments

The authors wish to thank Dorien Lugtenberg for her useful discussions and input during this study. We also thank the Genome Diagnostics group at UMCN for performing the Affymetrix CytoScanHD experiments and NimbleGen for their support in the validation experiments of the ExonArray.

Disclosure statement: The authors declare no conflict of interest.

References

Amarasinghe KC, Li J, Halgamuge SK. 2013. CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 14 Suppl 2:S2.

Bainbridge MN, Hu H, Muzny DM, Musante L, Lupski JR, Graham BH, Chen W, Gripp KW, Jenny K, Wienker TF, Yang Y, Sutton VR, et al. 2013. De novo truncating mutations in ASXL3 are associated with a novel clinical phenotype with similarities to Bohring-Opitz syndrome. Gen Med 5:11.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755.

Boone PM, Bacino CA, Shaw CA, Eng PA, Hixson PM, Pursley AN, Kang SH, Yang Y, Wiszniewska J, Nowakowska BA, del Gaudio D, Xia Z, et al. 2010. Detection of clinically relevant exonic copy-number changes by array CGH. Hum Mutat 31:1326–1342.

Boone PM, Soens ZT, Campbell IM, Stankiewicz P, Cheung SW, Patel A, Beaudet AL, Plon SE, Shaw CA, McGuire AL, Lupski JR. 2013. Incidental copy-number variants identified by routine genome testing in a clinical population. Genet Med 15: 45–54.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, et al. 2011. A copy number variation morbidity map of developmental delay. Nat Genet 43:838–846.

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, et al. 2012. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367:1921–1929.

Duan J, Zhang JG, Deng HW, Wang YP. 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PLoS One 8:e59128.

Franke B, Vasquez AA, Veltman JA, Brunner HG, Rijpkema M, Fernández G. 2010. Genetic variation in CACNA1C, a gene associated with bipolar disorder, influences brainstem rather than gray matter volume in healthy individuals. Biol Psychiatry 68:586–588.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, et al. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91:597–607.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. Annu Rev Med 63:35–61.

Hanchard NA, Murdock DR, Magoulas PI, Bainbridge M, Muzny D, Wu Y, Wang M, McGuire AL, Lupski JR, Gibbs RA, Brown CW. 2013. Exploring the utility of whole-exome sequencing as a diagnostic tool in a child with atypical episodic muscle weakness. Clin Genet 83:457–461.

Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. 2011. Genome-wide mapping of copy number variation in humans: comparative analysis of high-resolution array platforms. PLoS One 6:e27859.

Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA. 2007. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res 14: 1–11.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40:e69.

Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. 2008. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr Opin Struct Biol 18:366–374.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res 22:1525–1532.

Kurotaki N, Shen JJ, Touyama M, Kondoh T, Visser R, Ozaki T, Nishimoto J, Shiihara T, Uetake K, Makita Y, Harada N, Raskin S, et al. 2005. Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. Genet Med 7:479–483.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, et al. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. 2012. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28:1307–1313.

Lupski JR. 2009. Genomic disorders ten years on. Genome Med 1:42.

Lupski JR. 2012. Brain copy number variants and neuropsychiatric traits. Biol Psychiatry 72:617–619.

Mefford HC, Batshaw ML, Hoffman EP. 2012. Genomics, intellectual disability, and autism. N Engl J Med 366:733–743.

Mefford HC, Eichler EE. 2009. Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196–204.

Metzker ML. 2010. Sequencing technologies—the next generation. Nat Rev Genet 11:31–46.

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, 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.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65.

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65:6071–6079.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35.

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, et al. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338:1619–1622.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29:512–520.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, et al. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28:2747–2754.

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, et al. 2012. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 6736:1674–1682.

Schaaf CP, Wiszniewska J, Beaudet AL. 2011. Copy number and SNP arrays in clinical diagnostics. Annu Rev Genomics Hum Genet 12:25–51.

Stankiewicz P, Beaudet AL. 2007. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. Curr Opin Genet Dev 17:182–192.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455.

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 28:2711–2718.

Vissers LELM, de Vries BBA, Osoegawa K, Janssen IM, Feuth T, Choy CO, Straatman H, van der Vliet W, Huys EHLPG, van Rijk A, Smeets D, van Ravenswaaij-Arts CMA, et al. 2003. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. Am J Hum Genet 73:1261–1270.

Vissers LELM, de Vries BBA, Veltman JA. 2010. Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. J Med Genet 47:289–297.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.

Whibley AC, Plagnol V, Tarpey PS, Abidi F, Fullston T, Choma MK, Boucher CA, Shepherd L, Willatt L, Parkin G, Smith R, Futreal PA, et al. 2010. Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. Am J Hum Genet 87:173–188.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41:849–853.