

Open

Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort

Kyle Retterer, MS¹, Julie Scuffins, MS¹, Daniel Schmidt, MS¹, Rachel Lewis, MS¹, Daniel Pineda-Alvarez, MD¹, Amanda Stafford, MS¹, Lindsay Schmidt, MS¹, Stephanie Warren, BS¹, Federica Gibellini, PhD¹, Anastasia Kondakova, BS¹, Amanda Blair, BS¹, Sherri Bale, PhD¹, Ludmila Matyakhina, PhD¹, Jeanne Meck, PhD¹, Swaroop Aradhya, PhD¹ and Eden Haverfield, PhD¹

Purpose: Detection of copy-number variation (CNV) is important for investigating many genetic disorders. Testing a large clinical cohort by array comparative genomic hybridization provides a deep perspective on the spectrum of pathogenic CNV. In this context, we describe a bioinformatics approach to extract CNV information from whole-exome sequencing and demonstrate its utility in clinical testing.

Methods: Exon-focused arrays and whole-genome chromosomal microarray analysis were used to test 14,228 and 14,000 individuals, respectively. Based on these results, we developed an algorithm to detect deletions/duplications in whole-exome sequencing data and a novel whole-exome array.

Results: In the exon array cohort, we observed a positive detection rate of 2.4% (25 duplications, 318 deletions), of which 39%

involved one or two exons. Chromosomal microarray analysis identified 3,345 CNVs affecting single genes (18%). We demonstrate that our whole-exome sequencing algorithm resolves CNVs of three or more exons.

Conclusion: These results demonstrate the clinical utility of single-exon resolution in CNV assays. Our whole-exome sequencing algorithm approaches this resolution but is complemented by a whole-exome array to unambiguously identify intragenic CNVs and single-exon changes. These data illustrate the next advancements in CNV analysis through whole-exome sequencing and whole-exome array.

Genet Med advance online publication 6 November 2014

Key Words: array CGH; deletion; duplication; exome; next-generation sequencing

INTRODUCTION

Significant advances in copy-number detection have broadened the mutation spectrum for many clinical genetic disorders.^{1,2} Intragenic deletion mutations are of considerable frequency in many disease genes, such as *PAX6*, *CDKL5*, and *STXPB1*. Recurrent rearrangements between segmentally duplicated sequences are also associated with a number of syndromic disorders.³ For these known disorders, targeted gene testing by multiplex ligation-dependent amplification or exon-focused arrays has been useful. With the increasing uptake of exome sequencing into the clinical diagnostic approach, the need for testing previously uncharacterized genes for pathogenic copy-number variation (CNV) is a significant consideration, not only to detect aberrations in genes that may cause disease when haplo-insufficient but also in genes associated with recessive disorders for which the mutation has been identified in only one of the alleles by exome sequencing.⁴

Whereas exome sequencing is still gaining popularity as a powerful clinical tool, whole-genome chromosomal microarray analysis (CMA) has become an indispensable screening method that is now routinely used as a first-tier test for children with intellectual disability, developmental delay, or congenital anomalies.⁵ In less than 10 years, the CMA designs

have evolved from low-resolution arrays containing large bacterial artificial chromosome clones or <100,000 oligonucleotide probes to high-resolution versions with more than 1 million probes.⁶ As a result, several groups have identified single-gene pathogenic aberrations that boost the analytical sensitivity of CMA.⁷ However, although some of these more recent arrays have higher density at disease genes, they do not all cover every exon in those genes and are therefore not capable of detecting some pathogenic intragenic mutations. Separately, data from exon-focused arrays have shown that up to 40% of intragenic mutations can involve just one or two exons within a gene, and therefore it is essential to cover all exons within targeted genes.¹

Copy-number detection in clinical genetic testing eventually will occur entirely through examination of next-generation data, whereas array comparative genomic hybridization (aCGH) and other assays will serve as complementary and confirmatory methods.⁸ To complement whole-exome sequencing (WES) or whole-genome sequencing data in a meaningful way, an array with coverage of virtually all exons is essential. Until the time that WES/whole-genome sequencing can be used routinely and reliably for copy-number detection, a whole-exome array can be used as the ultimate whole-genome CMA platform.

¹GeneDx, Gaithersburg, Maryland, USA. Correspondence: Kyle Retterer (kyle@genedx.com) or Eden Haverfield (eden@genedx.com)

Submitted 19 June 2014; accepted 1 October 2014; advance online publication 6 November 2014. doi:[10.1038/gim.2014.160](https://doi.org/10.1038/gim.2014.160)

We analyzed data from 28,000 individuals tested by targeted exon-focused arrays or whole-genome cytogenetic arrays to identify CNVs that provide a deep perspective of the intragenic copy-number mutation spectrum in the human genome. Based on expectations of analytical sensitivity needed to reliably survey that spectrum and previously described methods,⁹ we have developed copy-number detection by WES and demonstrate its use in conjunction with a whole-exome array.

MATERIALS AND METHODS

Subjects

For targeted exon-focused microarray or whole-genome CMA testing, we analyzed data derived from blood or DNA samples submitted for clinical testing. Informed consent for testing was obtained on the test requisition. If peripheral blood samples were submitted, then DNA was extracted using a QIAcube robot (Qiagen, Venlo, The Netherlands).

Targeted exon-focused arrays

The design principle and processing of exon-focused arrays have been described previously by our group.⁸ Testing of the 14,228 individuals was performed on different array versions with a common design principle, as described previously. Mutations were confirmed in a second DNA sample by multiplex ligation-dependent amplification, quantitative polymerase chain reaction, or repeat aCGH.

Whole-genome CMA (cytogenetic) arrays

Whole-genome CMA was performed on samples from 14,000 unrelated individuals using a variety of designs, including arrays with 44,000, 105,000, or 180,000 oligonucleotide probes. All array designs were based on the principle described in several publications^{5,6} and covered the complete unique sequence of the genome at a resolution of 100 to 250 kb genome wide and at 250 bp to 10 kb within targeted regions. The densely probed targeted regions included individual disease-related genes, microdeletion/duplication syndrome regions, and subtelomeric and pericentromeric regions. Arrays were cohybridized with a half microgram of DNA each from the referred individual and a sex-matched pooled genome reference (Promega, Madison, WI). The hybridization, washing, and scanning were performed according to the manufacturer's protocol (Agilent Technologies, Santa Clara, CA). The resulting data were imported for analysis into DNA Analytics/Agilent Genomic Workbench software. All arrays, hybridization and scanning equipment, and software were from Agilent Technologies.

Whole-exome array CGH

The design for the exome array included four probes per exon plus 250 bp of flanking intronic sequence in virtually all genes in the GRCh37/hg19 genome build (~220,000 exons targeted). The mucin, olfactory receptor, and testis antigen gene families were excluded, as were a select group of genes in regions known to vary frequently in copy number. A total of 989 disease genes in the GeneTests.org database¹⁰ were targeted, with an average

of seven probes per exon. Larger exons were targeted with additional probes to maintain a minimum of one probe per 250 bp. A small number of exons did not have sufficient unique probes available because of homology or small exon size, and these were targeted with fewer probes. A whole-genome backbone covering intergenic regions was added with 500 kb of spacing between probes. Promoter regions were covered with five probes within the first kilobase upstream of each gene. The 5' untranslated regions were covered, but the 3' untranslated regions were not targeted. Regions with significant homology elsewhere in the genome were also excluded. The total number of probes (986,000) was placed on a one-array format on the Agilent Technologies platform. The array was hybridized according to the manufacturer's protocol as described in the "Whole-Genome CMA (Cytogenetic) Arrays" section above.

Whole-exome sequencing

Exome sequencing was performed on exon targets isolated by capture using the Agilent SureSelect Human All Exon V4 (50 Mb) kit (Agilent Technologies). One microgram of DNA from the blood specimen was sheared into 350- to 400-bp fragments, which were then repaired, ligated to adapters, and purified for subsequent polymerase chain reaction amplification. Amplified products were then captured by biotinylated RNA library baits in solution following the manufacturer's instructions. Bound DNA was isolated with streptavidin-coated beads and reamplified. The final isolated products were sequenced using the Illumina HiSeq 2000 sequencing system with 100-bp paired-end reads (Illumina, San Diego, CA). DNA sequence was mapped to the published human genome build University of California, Santa Cruz, hg19/GRCh37 reference sequence.¹¹ Targeted coding exons and splice junctions of known protein-coding RefSeq¹² genes were assessed for average depth of coverage with a minimum depth of 10× required for inclusion in downstream analysis. A custom-designed bioinformatics analysis pipeline was used to evaluate sequence changes.

Detection of copy number from next-generation sequencing data

The algorithm for CNV detection relied on the reproducibility of relative per-target coverage from sample to sample. Each exon was defined as a target with the location and number of probes per target determined by Agilent's proprietary capture design. The per-target coverage for each sample was normalized to the sample's overall exome coverage depth, and that normalized coverage was then compared with a baseline distribution of normalized per-target scores of previously sequenced samples.

Assuming a Gaussian distribution of the normalized coverage for each target, the SD can be used as an effective measure of the allowable deviation for each target. The level of deviation from the norm was calculated as a fractional increase or decrease, hereafter referred to as the "deviance." The degree of deviance from the norm predicted the presence of a CNV and the type, e.g., heterozygous deletion (−0.5 deviance), homozygous

deletion (−1.0 deviance), or duplication (+0.5 deviance or some multiple thereof).

To assess whether an apparent deviance was real, a *P* value was calculated to estimate the likelihood that the observed deviance was the result of random chance based on the previously observed normalized coverage distribution for that target. A single-sample Z-test scored a given sample against the baseline. The tuning of the optimal baseline size is discussed later. The Z-test was chosen for its speed and simplicity because the baseline sample size was large and the population variance was known. A minimum threshold of >0.25 absolute deviance with a corresponding *P* value of ≤0.01 was used to classify CNVs as significant.

A custom Perl script was used to chain adjacent significant targets into larger calls. When CNVs of three sequential targets are detected, the algorithm opens gaps to allow for calling events across low-coverage targets or those with nonsignificant *P* values but similar levels of deviance. This allows testing for larger CNVs through regions that cannot be accurately measured because of high noise or lack of coverage.

RESULTS

Intragenic copy-number mutations in 14,228 individuals

We previously described results from testing 3,018 individuals by exon array.⁸ Here we present data from an additional 11,210 individuals to provide a deeper picture of the types and frequencies of intragenic deletions and duplications. A list of all mutations not reported previously by our group is provided in **Supplementary Table S1** online. Among the 14,228 individuals tested, 343 had a pathogenic or likely pathogenic copy-number mutation, representing an analytical sensitivity of 2.4% overall. Of the mutations, 318 were deletions and 25 were duplications. One hundred thirty-four (39%) mutations involved only a single exon or two exons. Although it is possible to detect purely intronic changes in near-exonic regions, the great majority of these events encompass at least one exon.

<500 kb CNVs detected by CMA

Whole-genome CMA in 14,000 individuals identified 27,677 CNVs (data not shown); 24,276 CNVs were smaller than 500 kb, 4,350 of these affected a single OMIM gene, and 1,570 of the OMIM single-gene CNVs were deletions, accounting for 5.7% of all CNVs observed (**Supplementary Table S2** online). Examples of pathogenic CNVs included deletions affecting *NRXN1*, *CNTNAP2*, *DMD*, and *AUTS2*. We also found pathogenic copy-number variants in *STS*, *NSD1*, *OTX2*, *ANKRD11*, *EHMT1*, and a variety of other genes associated with syndromic disorders. Separately, we also identified individuals who were heterozygotes for intragenic deletions in *DOCK8*, *LARGE*, *SRD5A2*, and other genes associated with autosomal recessive disorders.

Detection of copy-number variants from exome sequencing

Nine samples previously tested by exon-focused aCGH or whole-genome CMA were blinded and tested by WES. Twelve

definitive CNVs within coding portions of genes had been detected previously in these samples. The CNVs ranged in size, extending from one to dozens of exons, and included both heterozygous deletions and duplications. We analyzed the WES data for CNVs using our algorithm and found that all 12 CNVs had deviance scores consistent with their corresponding array result. Four intragenic CNVs included three exons or more, and each of these CNVs had a *P* value of <0.01, meeting our threshold for significance. Four two-exon CNVs were also detected, but only three of them had *P* values of <0.01. Finally, all four of the single-exon CNVs were detected but had *P* values >0.01, indicating that the deviance was not significant when compared with the baseline distribution at those exons. Assessment of this small CNV set suggested that although it is possible to detect single-exon CNVs using our method, the likelihood of reproducibly and reliably identifying them is low. However, for CNVs that include three exons or more, the algorithm is significantly more reliable.

We also applied the reverse approach, verifying CNVs called from WES data using aCGH. We selected 11 random WES cases, each having a predicted significant CNV of at least three exons. These samples were tested by exome aCGH (see next section). The predicted CNVs in the WES data were divided into three groups: single-exon, two-exon, and three or more-exon CNVs. Two of the 11 samples had more than five times as many exons with significant deviance scores as the others and were thus classified as outliers and excluded from further analysis. Additional discussion of these noisy samples is included later.

A total of 15,370 CNVs were predicted at a significant *P* value among the remaining nine samples. Eighty-nine percent of deletion CNVs that included three or more exons were confirmed by aCGH, whereas only 60% of similarly sized duplications were confirmed. Only 11% of two-exon duplications and deletions, 3.3% of single-exon duplications, and 0.29% of single-exon deletions found by WES were confirmed.

These data indicate a high false-positive rate for CNV detection from WES data for events of fewer than three exons; the false-positive rate is more pronounced for duplications than deletions. The results of this pilot study are given in **Table 1**, which demonstrates the difference in specificity between single-exon events and events involving three or more exons. **Figure 1** illustrates the detection of intragenic copy-number mutations by WES and exome array.

After confirming the prospective CNV results, an optimal control baseline size of 500 samples was established using this set of known positive controls by progressively increasing the baseline set to 10, 50, 100, 200, 500, and 1,000 samples. Each progressive step reduced the total number of prospective CNVs called while keeping the number of known true positives fixed. The difference in noise reduction observed when moving from 500 to 1,000 samples was negligible compared with the increase in computational time, thus the optimal baseline was set at 500 samples.

After demonstrating the potential utility of this method for detecting CNVs from WES data, we began applying it

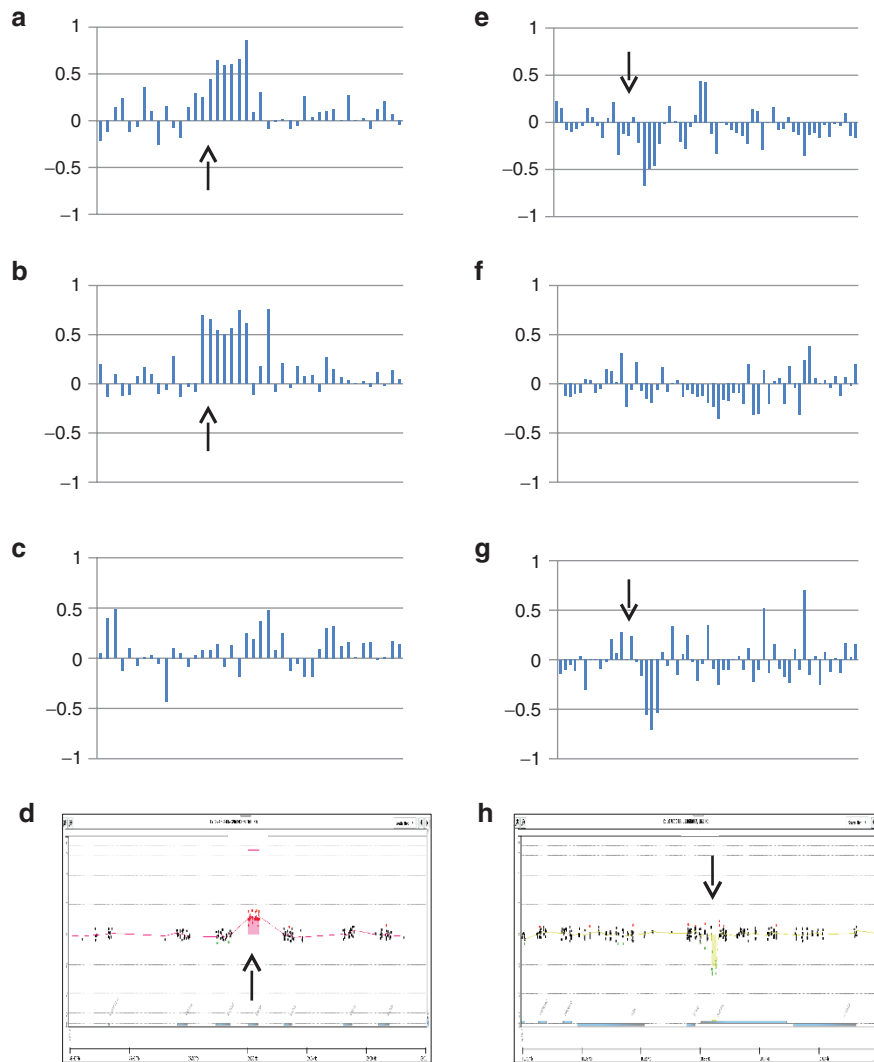


Figure 1 Examples of intragenic deletion and duplication detected by WES and confirmed by exome aCGH. Each bar in the graphs (a)–(c) and (e)–(g) represents an exon. (a–c) WES data from a family trio in which the (a) proband has inherited a whole-gene duplication of *KRT34* from the (b) father, whereas the (c) mother shows normal copy number at that gene. (e–g) WES data from a family trio in which the (e) proband has inherited a partial-gene heterozygous deletion in the *SYCP2L* gene from the (g) mother, whereas the (f) father shows normal copy number at those exons. Each dot in panels d and h represents an oligonucleotide probe in the gene of interest on the exome array, with a duplication shown by probes deviating to a positive log₂ ratio (marked in red) and a deletion shown by probes deviating to a negative log₂ ratio (marked in green). Panels d and h show confirmation of the *KRT34* duplication and the *SYCP2L* deletion, respectively, by exome aCGH. aCGH, array comparative genomic hybridization; WES, whole-exome sequencing.

prospectively to clinical WES samples. Examination of 1,500 samples identified 54 likely pathogenic CNVs that we attempted to confirm by aCGH, multiplex ligation-dependent amplification, or quantitative polymerase chain reaction; we chose the most appropriate method based on the size of the CNV and assay cost. Table 2 shows that 34 of the 54 CNVs were confirmed. Of these, 20 were deemed pathogenic. These 20 confirmed pathogenic CNVs detected in 1,500 samples represent an analytical sensitivity of 2.3%. The 34 confirmed CNVs range from single-exon to megabase-scale events.

As seen in Table 2, the confirmation rate was significantly higher for one-exon and two-exon events than predicted by our earlier cross-comparison. Multiple factors likely contributed to this enhanced specificity. First, CNVs were selected for

confirmation based on patient phenotypes and, when available, observed inheritance patterns in trios and duos. This effectively acted as a selection bias because additional information beyond the deviance scores and *P* values was used to choose these CNVs for confirmation. Second, the new prospective CNVs were detected in samples prepared with an automated library preparation method, whereas the original test set was prepared using a manual protocol. The automated protocol likely provided more consistent hybridization performance from one sample to the next, thereby reducing the amount of random variation in the prospective CNV set.

We observed that 10.3% of the samples in our cohort produced high levels of CNV noise, with two to three orders of magnitude more CNV calls than other samples. The cause of these noise

Table 1 Results from calling CNVs from WES data on the initial 11-sample pilot and the confirmation rate for three or more exon events via subsequent exome aCGH

Sample	Exons with $P \leq 0.01$	Exons with $P > 0.01$	3+ Exon dels	3+ Exon dups	3+ Exon dels confirmed	3+ Exon dups confirmed
1	329	183,839	1	4	1	3
2	962	183,206	0	3	0	2
3	1,033	183,135	0	3	0	3
4	1,132	183,036	0	5	0	2
5	1,439	182,729	2	0	2	0
6	1,682	182,486	1	1	0	0
7	1,830	182,338	0	7	0	5
8	3,233	180,935	1	2	1	0
9	3,730	180,438	2	2	2	1
10	13,211	170,957	1	45	NA	NA
11	16,083	168,085	49	85	NA	NA
Confirmation rate					86%	59%

The number of significant CNV calls is reduced significantly by requiring events to be of at least three exons in size. Note that the last two samples were excluded from confirmation because of the high noise levels.

aCGH, array comparative genomic hybridization; CNV, copy-number variation; dels, deletions; dups, duplications, NA, not applicable; WES, whole-exome sequencing.

Table 2 Confirmation rates by size for CNVs detected in 1,500 WES probands

CNV size	Attempted confirmations	Confirmed	Confirmation rate, %
1 Exon	22	11	50
2 Exons	14	11	79
3+ Exons	18	12	67
All	54	34	63

CNV, copy-number variation; WES, whole-exome sequencing.

spikes is unclear, but it effectively renders these samples blind to our method. To determine whether this noise was related to total sample coverage, we assessed the number of CNVs called against the unnormalized average coverage for each sample. Excluding the high-noise outliers gave a correlation coefficient of -0.26 , indicating no significant correlation. Because our method is subject to reproducibility of relative coverage from one sample to the next, we hypothesize that this noise is caused by poorly controlled differences in the capture hybridization conditions, such as ambient temperature fluctuations. Susceptibility to these noise spikes is a significant drawback to our method.

In addition to some samples being noisy, the normalized coverage distribution of some targets is too variable to ever score a deviation from the norm as highly significant. **Figure 2** illustrates the spectrum of variability across all exon targets. CNVs in targets with high coefficients of variation are effectively blind to the method described here. Because we are looking for deviations of 0.5 from the norm, a coefficient of variation of ≥ 0.25 restricts our ability to call single-copy changes with any confidence. Sixteen percent of all exon targets had coefficients of variation of ≥ 0.25 and are effectively blind to this method of interrogation. Low coverage of an exon target may also limit our ability to detect CNVs. The spectrum of per-target coverage (**Figure 3**) indicates that 21% of targets are covered below 40 \times , making detection of copy-number changes more difficult

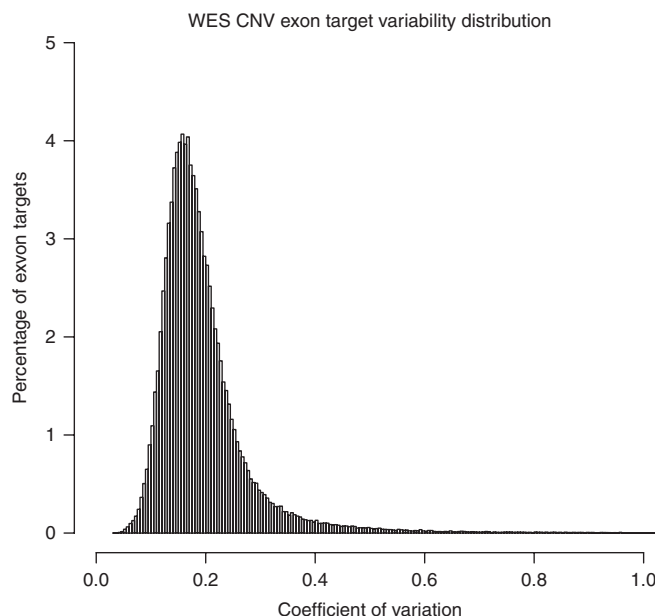


Figure 2 The relative frequency distribution of the coefficient of variation of the normalized relative coverage for all 180,000 exon targets across a 500-sample baseline. Sixteen percent of exons have coefficients of variation of 0.25 or greater, indicating that a -0.5 change in deviance will not be scored as highly significant; 3.1% of exons have coefficients of variation of 0.5 or greater, indicating extreme relative coverage variability at that target.

because true deviation and random noise are more difficult to distinguish at lower coverage levels. Although individual exons may not score as significant, deviations across several exons may still be detected through our chaining method, provided that the majority of the exons are well covered and not noisy.

Exome array CGH

We first tested 25 DNA samples with known CNVs on the custom whole-exome array design to identify deletions or

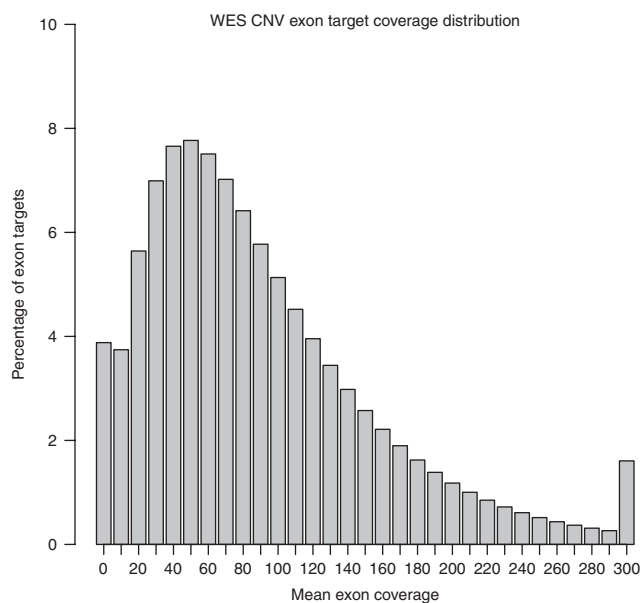


Figure 3 The absolute coverage distribution of all exon targets over a 500-sample baseline for all 180,000 exon targets. Coverages are means across each exon target and are binned into groups of 0–10, 10–20, and so forth, with all values ≥ 300 in the final bin.

duplications that ranged from 3kb to whole chromosomes. The data were consistent with previous whole-genome CMA or exon-focused array results and reliably showed the known CNVs. We observed a high incidence of reproducible artifacts (abnormal positive deviation \log_2 ratios), typically at the GC-rich 5' ends of genes as described previously,⁸ and these artifacts were excluded from analysis. We then performed a blinded study on nine DNA samples with previously characterized pathogenic deletions and duplications that ranged from 760 bp to several kilobases. The nine samples were hybridized multiple times, with one intra-array replicate and two inter-array replicates, to assess reproducibility. The whole-exome array design showed 100% concordance with previous data and across multiple hybridizations. The exome array results also refined the breakpoints of some CNVs because of higher exonic probe content. We also tested DNA samples for which WES data showed predicted CNVs. This effort is described in the previous section.

DISCUSSION

Deletions and duplications in the genome have long been investigated as causes for recognizable genetic disorders.³ The extent and variety of CNV in the human genome have become more apparent, and solutions to address them in a clinical context are improving rapidly. With ongoing integration of next-generation sequencing in clinical laboratories, the ability of this technology to provide copy-number information alongside sequence information is important because it brings us closer to a single platform that can identify most of the currently known molecular causes for genetic disorders. We have presented data here to provide a deeper picture of the types and variety of copy-number mutations at disease loci in the human genome, establishing

the expectations for extracting copy-number information from capture-based next-generation sequencing. We have also used an exome array to demonstrate its utility to complement exome sequencing and to provide a method to test virtually any gene for deletions or duplications.

Data from 14,228 individuals tested on exon-focused arrays showed a positive detection rate of 2.4% for copy-number mutations in the genes tested. This observation of significant analytical sensitivity added by exon aCGH has been elaborated on in other reports.² These data also show that a high number of intragenic copy-number mutations encompass just a single exon, thereby setting an expectation of sensitivity for any method used for deletion/duplication analysis. Other groups have also emphasized the need for exon-level CNV detection at disease genes.⁷

In our study results described here, copy-number detection directly from whole-exome sequencing successfully identified deletions that were three exons or bigger with a statistically significant $P \leq 0.01$. Smaller events may also be detected, but with uncertain reliability.

There are some limitations to obtaining copy-number information from WES data. For instance, there are some targets that have a low signal-to-noise ratio such that it cannot be easily determined whether a true deviation is the result of random chance. Second, there are some samples with abnormal relative coverage distributions. Such samples will produce one or two orders of magnitude more false-positive events than true-positive events and are effectively useless for CNV detection. Third, targets that are highly polymorphic (e.g., common duplications) will skew the distribution such that those targets effectively fall into the first scenario described above. This, however, can be a benefit in clinical WES because it effectively excludes common polymorphisms from the final variant list. Although there are known limitations to this method of detecting CNVs directly from WES data, and although it cannot yet serve as a substitute for aCGH, it provides added value to WES diagnostics.

Other methods for detecting CNVs directly from WES data have been published since we developed our algorithm. Many, such as XHMM¹³ and EXCAVATOR,¹⁴ use hidden Markov model approaches, whereas others such as CoNIFER¹⁵ use single-value decomposition. CoNIFER, interestingly, offers the ability to group multiple runs of a single sample together to help reduce batch effects, which we believe are a significant source of noise in our own data set. The method described here differs from these approaches in the simplicity, and thus speed, of its statistical test and its reliance on a large control baseline for comparison, which includes inherent removal of common polymorphisms. A detailed comparison of all published methods is beyond the scope of this paper.

An exome array is a useful adjunct tool to confirm findings by exome sequencing or to use when analysis is restricted to one or a handful of genes for which other copy-number analysis methods are not available. Moreover, it can reliably detect CNVs smaller than three exons that may be missed by exome sequencing and in exons that may be poorly covered or highly variable via WES. The whole-exome array can also be a useful

tool when one mutation in a gene for an autosomal recessive disorder is found by sequencing and a deletion/duplication on the other allele is suspected.⁸

Importantly, an exome array with a backbone of pan-genomic coverage of intergenic regions, as used in this study, is the ideal clinical CMA because it offers the highest resolution possible for cytogenetic analysis. In addition, the array design tested in this study also included single-nucleotide polymorphism probes to provide genotyping information to detect long continuous stretches of homozygosity, which may represent uniparental disomy or identity by descent.¹⁶ The added sensitivity of exon-level analysis of the entire genome is unclear at present because it has not been used widely before, but it would likely augment the 15–20% sensitivity now routinely observed with traditional whole-genome CMA designs for diagnosing intellectual disability, developmental delay, and congenital anomalies.⁵ To that end, various groups have already described intragenic deletions in a wide variety of genes.^{7,17,18}

Some limitations persist for detecting CNVs for both aCGH and WES. Difficult sequence characteristics, such as GC-rich content, pseudogenes, and segmental duplications, present a significant challenge to cover the entire coding region of the genome. Even when CNV detection is possible at some loci, the significance of specific CNVs can be difficult to interpret. For example, CNVs that include noncoding elements of a gene, such as promoters, untranslated regions, and intronic sequences, are invariably classified as variants of uncertain significance or dismissed as likely benign. One reason for that difficulty is the absence of robust CNV data at the resolution of a single gene. Despite the tremendous progress made in recent years in collecting CNV data from clinically affected individuals and from control individuals, the resolution of the arrays used in those studies was not high enough, and therefore the spectrum of CNV variation is largely restricted to deletions or duplications in the range of 250kb and longer. However, this gap will close relatively quickly with increasing use of WES in clinical and research environments and further improvements in copy-number detection algorithms.

Data from whole-genome CMA and from exon-focused aCGH indicate that intragenic CNV is not uncommon. The data presented here provide insight into the incidence of intragenic CNV in a clinical population and show that exome sequencing with a depth of coverage of 10× is sufficient to detect many pathogenic multi-exon deletions and duplications. With a higher depth of coverage, the resolution of WES will increase and may detect single-exon deletions or duplications routinely, although this would significantly increase the cost per sample at present. Until the time that WES can reliably detect intragenic CNV mutations, a whole-exome array is a sensitive tool to detect intragenic deletions or duplications in virtually any gene in the genome.

SUPPLEMENTARY MATERIAL

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

DISCLOSURE

The authors are employees of GeneDx, a subsidiary of Bioreference Laboratories.

REFERENCES

1. Yalçın O. Genes and molecular mechanisms involved in the epileptogenesis of idiopathic absence epilepsies. *Seizure* 2012;21:79–86.
2. Tayeh MK, Chin EL, Miller VR, Bean LJ, Coffee B, Hegde M. Targeted comparative genomic hybridization array for the detection of single- and multiexon gene deletions and duplications. *Genet Med* 2009;11:232–240.
3. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;61:437–455.
4. Abecasis GR, Auton A, Brooks LD, et al.; 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56–65.
5. 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.
6. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 2010;12:e8.
7. Boone PM, Bacino CA, Shaw CA, et al. Detection of clinically relevant exonic copy-number changes by array CGH. *Hum Mutat* 2010;31:1326–1342.
8. Aradhya S, Lewis R, Bonaga T, et al. Exon-level array CGH in a large clinical cohort demonstrates increased sensitivity of diagnostic testing for Mendelian disorders. *Genet Med* 2012;14:594–603.
9. Wu J, Grzeda KR, Stewart C, et al. Copy number variation detection from 1000 Genomes Project exon capture sequencing data. *BMC Bioinformatics* 2012;13:305.
10. Pagon RA. GeneTests: an online genetic information resource for health care providers. *J Med Libr Assoc* 2006;94:343–348.
11. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006.
12. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012;40(Database issue):D130–D135.
13. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012;91:597–607.
14. Magi A, Tattini L, Cifola I, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* 2013;14:R120.
15. 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.
16. Kearney HM, Kearney JB, Conlin LK. Diagnostic implications of excessive homozygosity detected by SNP-based microarrays: consanguinity, uniparental disomy, and recessive single-gene mutations. *Clin Lab Med* 2011;31:595–613, ix.
17. Celestino-Soper PB, Shaw CA, Sanders SJ, et al. Use of array CGH to detect exonic copy number variants throughout the genome in autism families detects a novel deletion in TMLHE. *Hum Mol Genet* 2011;20:4360–4370.
18. Lee JH, Ki CS, Kim HJ, et al. Analysis of copy number variation using whole genome exon-focused array CGH in Korean patients with primary congenital glaucoma. *Mol Vis* 2011;17:3583–3590.



This work is licensed under a Creative Commons Attribution 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/>