

Open

# Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing

Yanming Feng, PhD<sup>1</sup>, David Chen, MS<sup>1</sup>, Guo-Li Wang, PhD<sup>1</sup>, Victor Wei Zhang, MD, PhD<sup>1</sup>  
and Lee-Jun C. Wong, PhD<sup>1</sup>

**Purpose:** We aimed to demonstrate the detection of exonic deletions using target capture and deep sequencing data.

**Methods:** Sequence data from target gene capture followed by massively parallel sequencing were analyzed for the detection of exonic deletions using the normalized mean coverage of individual exons. We compared the results with those obtained from high-density exon-targeted array comparative genomic hybridization and applied similar analysis to examine samples from patients with pathogenic exonic deletions.

**Results:** Thirty-eight samples, each containing 2,134, 2,833, or 4,688 coding exons from different panels, with a total of 103,863 exons, were analyzed by capture-massively parallel sequencing and array comparative genomic hybridization. Ten deletions detected by array

comparative genomic hybridization were all detected by massively parallel sequencing, whereas only two of three duplications were detected. We were able to detect all pathogenic exonic deletions in 11 positive cases. Thirty-one exonic copy number changes from nine perspective clinical samples were also identified.

**Conclusion:** Our results demonstrated the feasibility of using the same set of sequence data to detect both point mutations and exonic deletions, thus improving the diagnostic power of massively parallel sequencing-based assays.

*Genet Med* advance online publication 17 July 2014

**Key Words:** array comparative genomic hybridization; exonic deletion; massively parallel sequencing; molecular diagnosis

## INTRODUCTION

Intragenic copy number variations (CNVs) in the human genome contribute to a significant portion of molecular defects associated with inherited genetic disorders.<sup>1,2</sup> Although traditional Sanger sequence analysis can reliably detect small genetic lesions, including point mutations and small insertions/deletions (indels), it does not detect heterozygous exonic deletions, duplications, or other rearrangements. For autosomal recessive disorders, when only a heterozygous point mutation is identified in a candidate gene by sequence analysis of the entire coding regions, the focus of the search for the second mutant allele shifts to intragenic deletions or duplications. To detect intragenic CNVs, techniques such as multiplex ligation-dependent probe amplification or exon-targeted array comparative genomic hybridization (aCGH), a completely different approach from a sequencing platform, often are applied.<sup>3,4</sup> Array-based detection of CNVs currently occurs at resolutions ranging from kilobases (kb) to megabases (Mb). Stepwise analysis using different methodologies is time-consuming and costly. Furthermore, various technologies may not be readily available to some clinical laboratories, or the assay may not include complete sets of genes or exons of interest for technical or commercial reasons.

Recent advancement in massively parallel sequencing (MPS) technologies has redefined the practice of molecular diagnosis

of human genetic diseases in clinical settings.<sup>5–8</sup> Because 85% of all known mutations are located in the coding regions and the intron/exon junctions,<sup>9</sup> capture-based target gene enrichment followed by MPS analysis is a cost-effective way to identify point mutations and small indels <20 bp. MPS with consistent exonic coverage can potentially provide an opportunity for concurrent detection of deletions and point mutations in patients with inherited disorders.

Several CNV-detection algorithms for the analysis of data produced by whole-exome sequencing have been developed.<sup>10–15</sup> Some of these algorithms are based on the coverage depth of a base, an exon, a capture, or a segment,<sup>12,14,15</sup> and many use paired-end sequencing data to detect intragenic deletions and to map breakpoints if those junction sequences are captured and sequenced.<sup>16</sup> Recent investigation of the performance of these algorithms on clinically relevant CNVs demonstrated that these coverage depth-based algorithms can detect most CNVs >200 kb. The ability to detect smaller CNVs, however, becomes much less reliable; the smallest deletion detected is at least three consecutive, relatively large exons spanning ~20–30 kb. Results from coverage depth-based analysis are consistent with those obtained by the hidden Markov model (XHMM) algorithm in a cohort of patients with autism spectrum disorders.<sup>17</sup> These algorithms used sophisticated, robust statistical analyses to interrogate CNVs from the sequence data, with high variations in coverage depth throughout the genome.

<sup>1</sup>Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, Texas, USA. Correspondence: Lee-Jun C. Wong ([ljwong@bcm.edu](mailto:ljwong@bcm.edu)); Victor Wei Zhang ([wzhang2@bcm.edu](mailto:wzhang2@bcm.edu))

Submitted 24 January 2014; accepted 29 May 2014; advance online publication 17 July 2014. doi:[10.1038/gim.2014.80](https://doi.org/10.1038/gim.2014.80)

These analyses could have been statistically or methodologically optimized for specific research purposes.<sup>18</sup> Application of whole-exome sequencing–based CNV analysis in routine clinical diagnoses and reliable detection of CNVs at the single exon level has not been reported.

In this report, we focus our investigation on the utilization of the same sequence data sets for the detection of point mutations and exonic deletions to improve diagnosis in one comprehensive approach. We compared the CNVs detected based on the analysis of MPS data to those obtained from a high-density exon-targeted aCGH to assess the quantitative performance characteristics for clinical application.<sup>3</sup> We further examined 11 samples with pathogenic deletions, and all were correctly identified.

## MATERIALS AND METHODS

### DNA samples

The analyses were performed using DNA samples collected for clinical diagnostic purposes according to protocols approved by the institutional review board at the Baylor College of Medicine. DNA was extracted from whole blood using commercially available DNA isolation kits (Gentra Systems, Minneapolis, MN) following the manufacturer's instructions.

### High-resolution oligonucleotide array (aCGH)

A custom-designed oligonucleotide CGH array (MitoMet v3) was used to detect CNVs. MitoMet v3 is a clinically validated 180K exon-targeted oligonucleotide array with complete coverage of the mitochondrial genome and 1,500 selected nuclear genes related to mitochondrial structure/function and metabolic diseases.<sup>19,20</sup> The average probe density is greater than four probes per exon, with 1-kb spacing in the intronic regions. The targeted region of the aCGH contains probes for the coding exons and 50bp of the flanking intronic regions, identical to the probes used for the target capture design for MPS studies. The criteria for a potential CNV call in the aCGH are at least two contiguous probes with a log<sub>2</sub> ratio >0.3 for duplication and <-0.3 for deletion in regions of interest.

### Target gene enrichment, library preparation of the DNA template, and sequencing

Custom-designed Roche NimbleGen SeqCap probe libraries were used to capture all coding regions and 50bp of flanking intronic regions of target genes. Genes in these libraries (<http://www.bcmgeneticlabs.org>) were selected on the basis of their structurally or functionally interactive relationship, involvement in the same or related pathways, or phenotypic similarities (**Supplementary Table S1** online). The coding regions were enriched according to the manufacturer's instructions (Roche NimbleGen Madison, WI). Details of the experimental procedures have been described previously.<sup>7</sup> Eight indexed DNA template samples were pooled at equimolar ratio and sequenced in one lane of the flow cells of HiSeq2000 (Illumina, San Diego, CA) with single-end, 75-base read lengths. Cluster generation and sequencing were carried out according to the manufacturer's instructions.

### Analysis of target capture/MPS data

Conversion of raw image data and demultiplexing were performed following Illumina's primary data analysis pipeline using CASAVA v1.7 (Illumina). Low-quality reads (Phred score < Q25) were removed during demultiplexing. The demultiplexed sequence data were further processed by NextGENe software (SoftGenetics, State College, PA) for alignment. The mean coverage of each exon was extracted from the aligned data using bed files specific for the captured genes. The exons with possible copy number changes detected by our analysis are depicted automatically for further examination. The script for the detection of CNVs is deposited at <https://sourceforge.net/projects/cnvanalysis>.

## RESULTS

### Generation of reference coverage profile, normalization, and CNV detection

We have developed more than 36 MPS-based panels for clinical usage ([www.bcmgeneticlabs.org](http://www.bcmgeneticlabs.org)). Coverage profiles vary among the different panels depending on the number and types of genes, the number of exons, and the size of the target sequences. To simplify, we use a 189-gene capture library as an example. This library contains probes for several gene panels, including 19 genes responsible for mitochondrial DNA depletion syndrome, 26 genes for metabolic myopathy and rhabdomyolysis, 25 genes for glycogen storage disorders, 17 genes for cobalamin-related disorders, and 22 genes for fatty acid oxidation. All of these genes and NM numbers are listed in **Supplementary Table S1** online. Twenty samples without CNVs, as confirmed by aCGH, were selected as reference samples for target capture/MPS. The mean sequence depth was 874–1,457× per base. The length of target exons ranged from the smallest, exon 1 (9bp) of the *MMADHC* gene (OMIM# 611935, NM\_015702.2), to the largest, exon 4 (1,739bp) of *MOGS* (OMIM# 601336, NM\_006302.2), with a mean size of 139 ± 106bp (**Supplementary Figure S1** online). The mean coverage depth of each individual exon of a sample was first normalized for the total amount of the DNA template loaded onto sequencing flow cells based on the total reads of this sample. The mean coverage of each individual exon from the reference samples thus obtained was used as the reference for a specific exon. Reference coverage profiles for a normal male and female were generated separately.

To detect CNVs, the normalized coverage of each exon of a test sample was compared to the mean coverage of the same exon in the reference file generated above. The analysis workflow for the detection of CNVs based on MPS data is illustrated in **Supplementary Figure S2** online. The theoretical ratios of 1, 0.5, and 0 represent normal copy number, loss of one copy, and loss of two copies, corresponding to normal, heterozygous deletion, and homozygous/hemizygous deletion, respectively. The actual cutoff value for heterozygous deletion was determined experimentally by the statistical analysis of a group of positive samples, which shows a range of 0.39 to 0.55, with an average of 0.45 ± 0.07. Using different cutoff values for the detection

of heterozygous deletion, the receiver operating characteristic (ROC) curve is depicted in **Supplementary Figure S3** online.

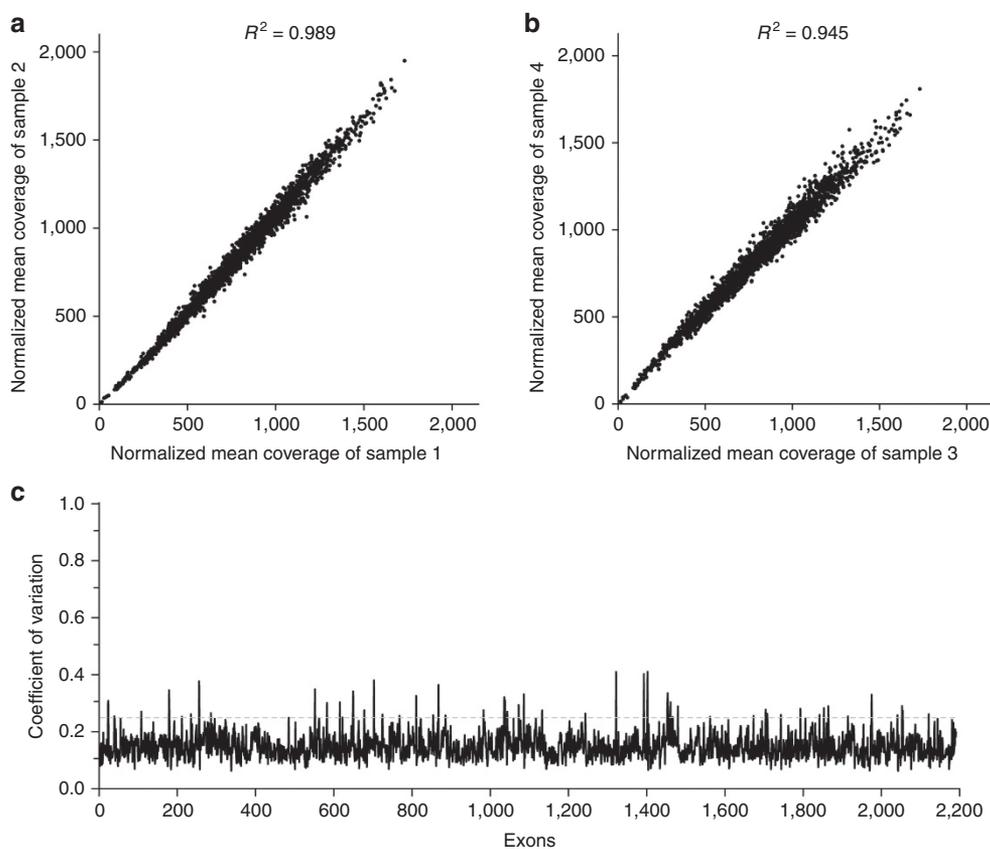
### Performance evaluation of CNV detection

We evaluated the reliability and performance characteristics of CNV detection using the simple analysis based on the deep sequencing data, as described above, by comparing the results with those obtained from high-density exon-targeted aCGH. **Figure 1** shows the consistency of the normalized mean coverage of the 2,134 coding exons in the capture library of two arbitrary samples from the same batch (**Figure 1a**) and from two different batches (**Figure 1b**).<sup>7</sup> The correlation coefficients ( $r$  values) were 0.989 and 0.945, respectively. These results suggest that the normalized mean coverage of arbitrary sample pairs is highly consistent even when the samples were prepared and sequenced at two different times. This result is highly reproducible. As shown in **Supplementary Figure S4** online, 96% of a total of 153 correlation coefficients ( $r$  values) between any 2 of the 18 arbitrary samples analyzed at different times are  $>0.95$ . Examination of the coefficient of variation of the reference files revealed that only 52 of 2,134 target exons (2.3%) have coefficient of variation values greater than 25% (**Figure 1c**). Similar variations have been observed in other test samples. Further analysis showed that exons with large coefficients of

variation often have low coverage because of high GC content (**Supplementary Figure S5** online). All exons with sufficient coverage, regardless of their coefficients of variation, were included in CNV detection analysis. Exons with coverage  $<20\times$  (consistently 1.26%, 27 of a total of 2,134 exons) were considered insufficiently covered; no-calls were made, and these exons were not included in the evaluation.

**Table 1** lists CNVs from 38 samples that have been analyzed by both aCGH and target exon capture/MPS methods. A total of 103,863 exons were included in the comparison; among those, 103,850 did not show CNVs by aCGH, whereas 102,798 exons were considered to have normal copy numbers based on MPS analysis; 962 exons were excluded because of insufficient coverage. The sensitivity at a 95% confidence interval for the detection of a deletion is 100% but is only 51.72–81.68% for duplications. The specificity for the detection of deletions and duplications is 99.42–100% and 99.17–100%, respectively (**Supplementary Table S2** online).

Our analysis was able to detect all deletions confirmed by aCGH at the single exon level without false negatives. The false-positive rate of the MPS-based method is much higher for duplications (96.77%) than deletions (75.61%). Despite the high specificity of detection of duplication, the positive predictive value is only 0–9.04% at 95% confidence intervals, which



**Figure 1** Consistency of mean coverage among different samples and quality of the reference file. Correlation of normalized mean coverage of 2,134 exons in the 189-gene capture library of two arbitrary samples (**a**) from the same batch and (**b**) from different batches. (**c**) Distribution of the coefficients of variation (CVs) among 2,134 exons calculated from reference samples. The thin gray line indicates 25% CV.

**Table 1** Comparison of CNV detection by array CGH and targeted MPS analysis for 38 validation samples

Panel	Sample ID	Exons ( <i>n</i> )	CNV detected											
			aCGH method			Targeted MPS method								
			Normal	Del	Dup	No call <sup>a</sup>	Normal	Del	Del-FN	Del-FP	Dup	Dup-FN	Dup-FP	
189-Gene	1	2,134	2,134	0	0	27	2,107	0	0	0	0	0	0	0
	2	2,134	2,134	0	0	27	2,107	0	0	0	0	0	0	0
	3	2,134	2,133	1	0	27	2,100	7	0	6	0	0	0	0
	4	2,134	2,134	0	0	27	2,106	1	0	1	0	0	0	0
	5	2,134	2,130	2	2	27	2,088	2	0	0	17	1	16	
	6	2,134	2,129	4	1	27	2,089	12	0	8	6	0	5	
	7	2,134	2,132	2	0	27	2,100	7	0	5	0	0	0	
	8	2,134	2,134	0	0	27	2,107	0	0	0	0	0	0	
	9	2,134	2,134	0	0	27	2,107	0	0	0	0	0	0	
	10	2,134	2,134	0	0	27	2,099	0	0	0	8	0	8	
	11	2,134	2,134	0	0	27	2,107	0	0	0	0	0	0	
	12	2,134	2,134	0	0	27	2,100	0	0	0	7	0	7	
	13	2,134	2,134	0	0	3	2,100	8	0	8	23	0	23	
	14	2,134	2,134	0	0	3	2,131	0	0	0	0	0	0	
	15	2,134	2,134	0	0	3	2,131	0	0	0	0	0	0	
	16	2,134	2,134	0	0	19	2,115	0	0	0	0	0	0	
	17	2,134	2,134	0	0	4	2,129	1	0	1	0	0	0	
	18	2,134	2,134	0	0	15	2,119	0	0	0	0	0	0	
	19	2,134	2,134	0	0	4	2,130	0	0	0	0	0	0	
	20	2,134	2,134	0	0	10	2,124	0	0	0	0	0	0	
	21	2,134	2,133	1	0	14	2,118	2	0	1	0	0	0	
	22	2,134	2,134	0	0	14	2,120	0	0	0	0	0	0	
	23	2,134	2,134	0	0	14	2,120	0	0	0	0	0	0	
	24	2,134	2,134	0	0	4	2,130	0	0	0	0	0	0	
256-Gene	25	2,833	2,833	0	0	32	2,800	0	0	0	1	0	1	
	26	2,833	2,833	0	0	17	2,816	0	0	0	0	0	0	
	27	2,833	2,833	0	0	17	2,816	0	0	0	0	0	0	
	28	2,833	2,833	0	0	21	2,811	1	0	1	0	0	0	
	29	2,833	2,833	0	0	38	2,795	0	0	0	0	0	0	
	30	2,833	2,833	0	0	53	2,780	0	0	0	0	0	0	
	31	2,833	2,833	0	0	45	2,788	0	0	0	0	0	0	
500-Gene	32	4,688	4,688	0	0	58	4,630	0	0	0	0	0	0	
	33	4,688	4,688	0	0	68	4,620	0	0	0	0	0	0	
	34	4,688	4,688	0	0	32	4,656	0	0	0	0	0	0	
	35	4,688	4,688	0	0	21	4,667	0	0	0	0	0	0	
	36	4,688	4,688	0	0	51	4,637	0	0	0	0	0	0	
	37	4,688	4,688	0	0	51	4,637	0	0	0	0	0	0	
	38	4,688	4,688	0	0	27	4,661	0	0	0	0	0	0	
Total		103,863	103,850	10	3	962	102,798	41	0	31	62	1	60	

aCGH, array comparative genomic hybridization; CNV, copy number variation; Del, copy number loss; Dup, copy number gain; FN, false negative; FP, false positive; MPS, massively parallel sequencing.

<sup>a</sup>Exons with coverage less than 20x.

renders the MPS-based detection of duplications less applicable for clinical use (**Supplementary Table S2** online). We found that 51% of the 31 false-positive deletions have GC content >65%, whereas 88% (53 of 60) of duplication false-positive samples have GC content >65%. Our result is also consistent with a previous report that the detection of deletions outperforms

that of duplications because of multiple factors that perturb the depth of coverage for a given exon.<sup>18</sup>

Although all copy number losses detected by aCGH were detected by MPS-based analysis, the positive predictive rate is only 17.39–47.12%. This implies that all deletions detected by the MPS-based method require further confirmation with

**Table 2** Intragenic deletions detected by targeted gene capture and sequencing

Category	Patient	Indications for study	Gene	GenBank no.	Allele 1	Allele 2	Exon(s) by NGS (n)	Size of deletion by aCGH	Breakpoint confirmation
Homozygous	P1	Mitochondrial DNA depletion syndrome	<i>MPV17</i>	NM_002437.4	E3–7 del	E3–7 del	5	1,773 bp	Yes
	P2	Myopathy, rhabdomyolysis	<i>LPIN1</i>	NM_145693.1	E18 del	E18 del	1	1,763 bp	Yes
Hemizygous	P3	Glycogen metabolism disorder	<i>PHKA2</i>	NM_000292.2	E27–30 del	NA	4	4.5 kb <sup>a</sup>	ND
Heterozygous (carrier or AD)	P4	Development delay	<i>POLG2</i>	NM_007215.3	E4–8 del	—	5	14,248 bp	Yes
	P5	Mother of P3	<i>PHKA2</i>	NM_000292.2	E27–30 del	—	4	4.5 kb <sup>a</sup>	ND
	P6	Carrier <i>ARG1</i> deletion	<i>ARG1</i>	NM_000045.2	E5 del	—	1	240 bp	Yes
	P7	Citrin deficiency	<i>SLC25A13</i>	NM_001160210.1	E3 del	—	1	4,176 bp	Yes
	P8	Mother of affected son	<i>OTC</i>	NM_000531.5	Partial E2 del	—	1	666 bp	Yes
Compound heterozygous	P9	Hyperammonemia	<i>CPS1</i>	NM_001122633.1	E9–11 del	c.2945G>T (p.Gly982Val)	3	2,395 bp	Yes
	P10	Myopathic form of mtDNA depletion syndrome	<i>TK2</i>	NM_004614.4	E1–2 del	c.635T>A (p.Ile212Asn)	2	5,828 bp	Yes
	P11	Carnitine acylcarnitine transporter deficiency	<i>SLC25A20</i>	NM_000387.4	E5–9 del	c.823C>T (p.Arg275*)	5	25,953 bp	Yes

aCGH, array comparative genomic hybridization; AD, autosomal dominant; del, deletion; mtDNA, mitochondrial DNA; ND, not determined; NGS, next-generation sequencing.

<sup>a</sup>The size of *PHKA2* deletions in P3 and P5 were determined by aCGH, but the exact breakpoint was not mapped.

a second method if the approach is to be used for clinical diagnostic purposes. In contrast, the negative predictive values for both deletions and duplications are 100%.

### Detection of pathogenic deletions in patients with inherited disorders

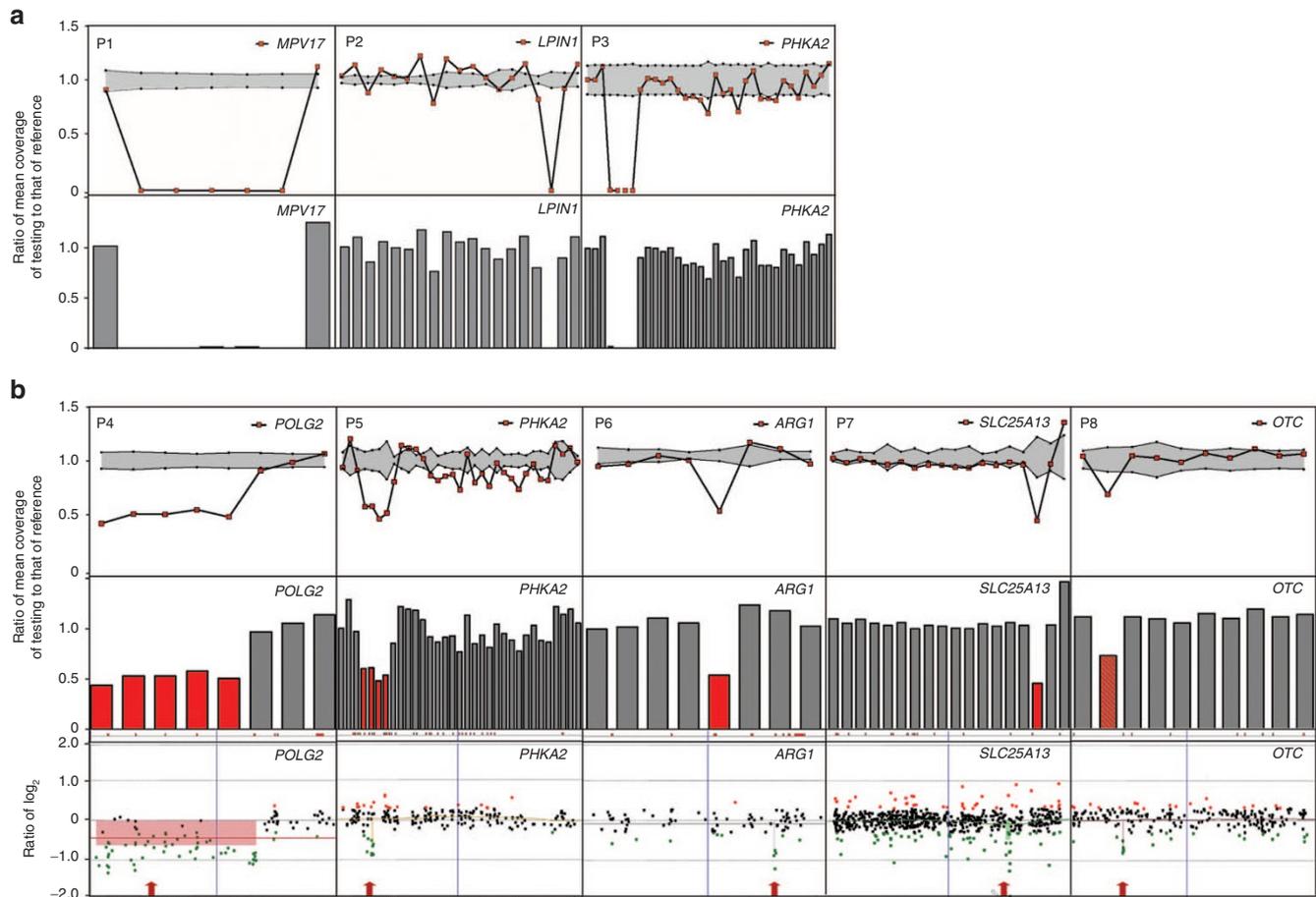
We applied a capture/MPS approach to 11 samples with pathogenic deletions identified by aCGH (Table 2). These samples represent a broad spectrum of exonic deletions, including single- or multiple-exon deletions, in heterozygous or homozygous/hemizygous states. We also identified 31 exon copy number changes in 9 prospective clinical samples (Supplementary Table S3 online).

Detection of homozygous and hemizygous intragenic deletions is relatively straightforward because the coverage of sequence reads in the deleted regions is zero. Figure 2a shows that patient (P) 1 has nearly zero sequence coverage for exons 3–7 of the *MPV17* gene (OMIM# 137960, NM\_002437.4), suggesting a homozygous deletion. Our analysis was also able to detect the homozygous deletion of a single exon of *LPIN1* (OMIM# 605518, NM\_145693.1) in P2, and a hemizygous deletion of exons 27–30 of the *PHKA2* gene (OMIM# 300798, NM\_000292.2) in P3.

Heterozygous deletions can be detected when the coverage ratio is close to 0.5. Figure 2b shows samples with heterozygous deletions detected, ranging from a single exon to multiple

exons. The samples with deletions spanning multiple exons included exons 4–8 of *POLG2* (OMIM# 604983, NM\_007215.3) in P4, exons 27–30 of *PHKA2* in P5, exons 9–11 of *CPS1* (OMIM# 608307, NM\_001122633.1) in P9, exons 1 and 2 of *TK2* (OMIM# 188250, NM\_004614.4) in P10, and exons 5–9 of *SLC25A20* (OMIM# 613698, NM\_000387.4) in P11. It is worth noting that P5, exhibiting a heterozygous deletion in *PHKA2* (Figure 2b), is the asymptomatic carrier mother of the affected male P3 in Figure 2a. Two samples showed heterozygous deletions of a single exon: P6 with exon 5 of *ARG1* (OMIM# 608313, NM\_000045.2) and P7 with exon 3 of *SLC25A13* (OMIM# 603859, NM\_001160210.1) (Figure 2b). The smallest exon deletion detected by this method is 95 bp for exon 5 of *ARG1* in P6 (Figure 2b).

Partial exonic deletion was also included to examine the potential utility of our algorithm. P8 in Figure 2b is a female affected with ornithine transcarbamylase deficiency who had a heterozygous deletion of 666 bp of the *OTC* gene (OMIM# 300461, NM\_000531.5). The deleted segment includes the last 93 bp of exon 2, which is 139 bp in length, and 573 bp of intron 2. Our deletion detection approach was able to identify an atypical copy number loss with a ratio of 0.7 instead of 0.5 for this case (Figure 2b, P8). This atypical result alerted us to further examine the coverage profile of this exon. We were able to retrieve the breakpoint information from the sequence data to confirm the partial exon loss. The normalized coverage ratio



**Figure 2** Detection of homozygous/hemizygous intragenic deletions and heterozygous deletions. (a) Top panel: the ratio of normalized mean coverage of individual coding exon of a gene to that of the reference was plotted against the exon number. The gray zone indicates the range of variation from the reference file. The individual's value is colored in red. Bottom panel: the histogram of the ratio of normalized mean coverage of the same test sample. The x-axis for both the top and bottom panels are the corresponding exons in a gene. Homozygous deletions of exons 3–7 of *MPV17* in patient (P) 1, homozygous deletion of a single exon 18 of *LPIN1* in P2, and hemizygous deletion of exons 27–30 of *PHKA2* on the X chromosome of an affected male patient (P3). (b) The corresponding array comparative genomic hybridization results are shown at the bottom panel of this figure. These known deletions are exons 4–8 of *POLG2* in P4, exons 27–30 of *PHKA2* in P5, exon 5 of *ARG1* in P6, and exon 3 of *SLC25A13* in P7. Massively parallel sequencing copy number variation analysis for P8, with heterozygous partial exonic deletion in *OTC*, is also shown. An atypical normalized coverage ratio of 0.7 suggests a heterozygous deletion of a partial exon.

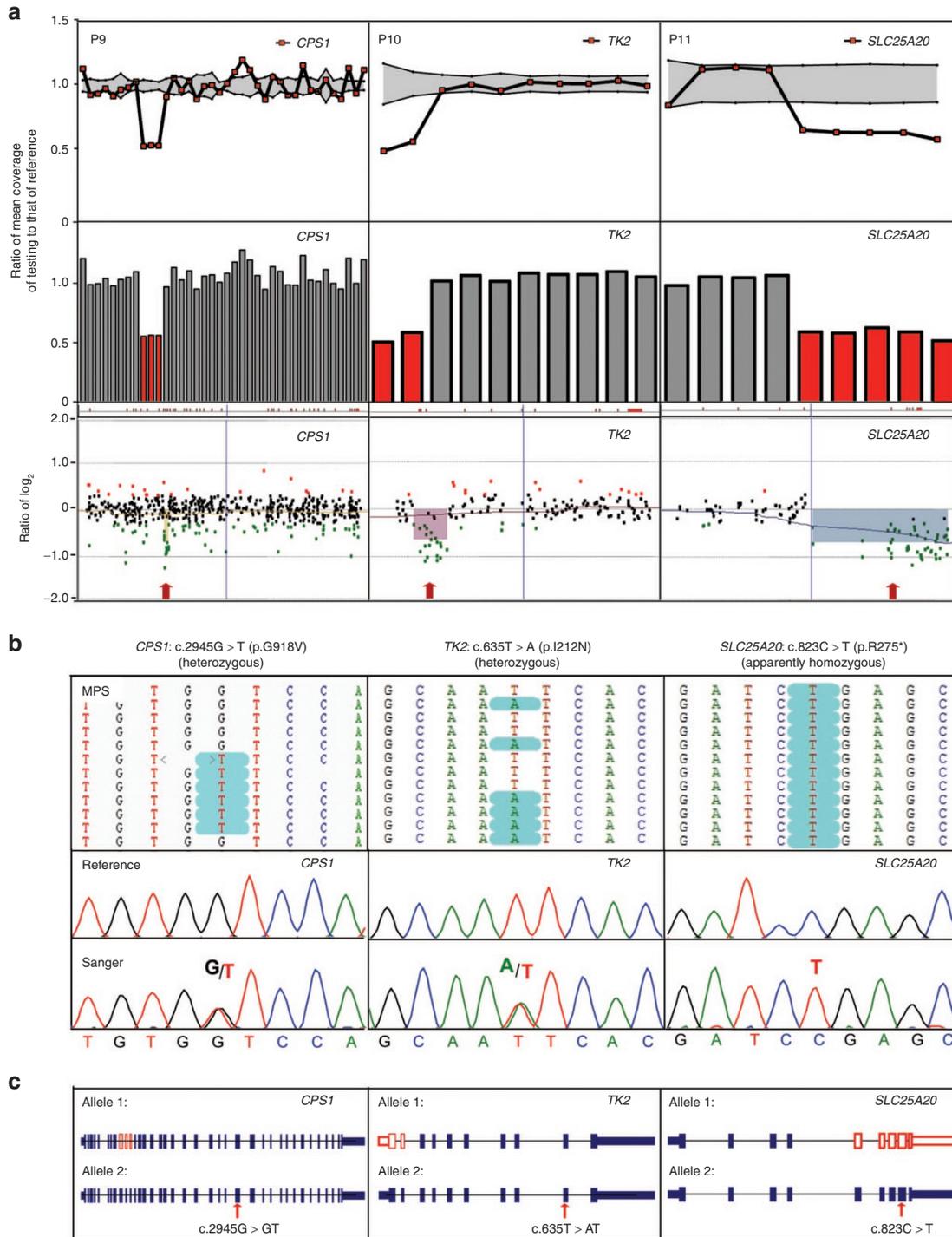
of this sample (0.7) is not within the values for our automatic CNV calls; therefore, without prior knowledge of the deletion, it is possible that this partial exonic deletion would not have been detected. Thus, manual examination is necessary to interrogate the atypical findings.

#### Compound heterozygosity and apparent homozygosity of an exonic deletion for autosomal recessive disorder

When only one heterozygous deleterious mutation is detected in a candidate gene for an autosomal recessive disorder, there is a possibility that the second mutant allele is an exonic deletion not detectable by Sanger sequencing. In this case, an exon-targeted aCGH or another method (e.g., multiplex ligation-dependent probe amplification or quantitative polymerase chain reaction) is often used to detect the presence of exonic deletions.<sup>3,4,19–21</sup> Each of the three heterozygous deletion cases described in Table 1 (P9, P10, and P11) are illustrated in Figure 3. Both P9 and P10 have a heterozygous point mutation

in a *trans* configuration with a large deletion that does not involve the exon containing the point mutation in the other mutant allele (Figure 3). MPS-based analysis detected a heterozygous c.635T>A (p.Ile212Asn) point mutation in exon 9 of the *TK2* gene (Figure 3b,c) in P10. Analysis of the same set of sequence data revealed a heterozygous intragenic deletion of exons 1 and 2 of the *TK2* gene (Figure 3a). Parental analysis confirmed that these two mutant alleles were in *trans* configuration, which is consistent with the proband's diagnosis of the myopathic form of mitochondrial DNA depletion syndrome.<sup>22</sup>

In general, homozygosity is tentative until it is proven that each chromosome carries the same mutation.<sup>23</sup> As Figure 3 illustrates, P11 is an example of an apparent homozygosity. This is a previously described patient who presented with hyperammonemia and hypoketotic hypoglycemia.<sup>24</sup> MPS-based analysis revealed an apparently homozygous point mutation, c.823C>T (p.Arg275\*) in exon 8 of the *SLC25A20* gene, encoding a carnitine transporter. Parental



**Figure 3 Compound heterozygosity of a point mutation and a large exonic deletion detected by our coverage depth analysis.** (a) The panel arrangement is the same as in Figure 2. (b) Point mutations detected by massively parallel sequencing (top) and confirmed by Sanger sequence analysis (bottom). (c) Diagram illustrating the *trans* configuration of the compound heterozygosity. The coding exons containing the point mutations in P9 and P10 are not deleted in the other mutant allele; therefore, sequencing shows heterozygous. In P11, however, the heterozygous point mutation is located in an exon that is deleted in the other allele; therefore, sequencing shows apparent homozygosity.

analysis revealed that the father carried the point mutation, but the mother did not.<sup>23,24</sup> Further analysis of the MPS data using our deletion detection algorithm revealed a heterozygous deletion of exons 5–9 in the same gene. Since the

heterozygous deletion included exon 8, the exon containing the point mutation in the other chromosome, the c.823C>T (p.Arg275\*) mutation seemed to be homozygous by both MPS and Sanger sequencing (Figure 3b,c).

## DISCUSSION

### Detection of copy number loss at the single exon level by coverage depth-based analysis

Consolidation of mutation-detection methods for the diagnosis of inherited diseases is being driven by technological advancements. It has been demonstrated that capture-based methodology can accurately detect point mutations and small indels <20 bp. CNVs, in a single exon, in multiple exons, or at the gene level, contribute to a significant portion of the causes of human inherited disorders.<sup>1,2</sup> Clinical laboratories currently apply different methodologies using multiple technological platforms and diagnostic expertise in a stepwise manner to detect different types of mutations. Capture-based target gene enrichment followed by MPS allows concurrent analysis of thousands of exons in a cost-effective manner. Detection of point mutations, indels, and CNVs in a single comprehensive assay for diagnostic purposes is greatly needed.

Using target capture/deep MPS data and normalized mean coverage per individual exon, we have shown that concurrent detection of point mutations and exonic deletions is feasible; this approach significantly improves diagnostic procedures and yields in a single assay with high sensitivity and specificity. The high false-positive rate for the detection of CNVs, however, suggests the necessity of subsequent confirmation.

Because our capture-based enrichment focuses on a small fraction (~1%) of the whole exome (>20,000 genes), the inter-probe interference is minimized and the coverage depth is increased. Thus relatively consistent and uniform coverage of all target exons is achieved, which allows for CNV analysis using the simple approach described above. This is in contrast to large coverage variations observed in whole-exome or polymerase chain reaction-based enrichment of target sequences, which requires more sophisticated analytical algorithms for CNV detection. The use of mean coverage of an exon in combination with high coverage depth increases the signal-to-noise ratio and enhances the sensitivity to detect heterozygous deletions. It has been reported that the single-nucleotide coverage depth-based analyses of whole-exome sequencing data can detect a heterozygous deletion as small as 510 bp.<sup>25</sup> Our method allowed us to detect a known heterozygous single-exon deletion as small as 95 bp (P6) and a 93-bp partial exon deletion (P8). As shown in **Supplementary Figure S1** online, the smallest exon investigated was 9 bp in exon 1 of the *MMADHC* gene, which encodes a protein responsible for methylmalonic aciduria type cblD. Thus, theoretically, deletion of this exon can be detected through our analysis.

### Simultaneous detection of point mutations and exonic deletions for autosomal recessive disorders

Exonic deletions have been recognized as an important contributor to various diseases.<sup>26–29</sup> There is general consensus that Sanger sequence analysis is not suitable for dosage analysis. For heterozygous deletions, the presence of a nondeleted allele precludes the detection of deletion by polymerase chain reaction-based methodologies. Detection of a heterozygous deletion at

the single-exon resolution has been difficult for most coverage depth-based analyses. Our experience shows that with target gene capture followed by deep sequencing (~1,000×), the normalized coverage depth of a particular exon can be highly consistent among different samples. Thus our MPS method allows simultaneous detection of a compound heterozygous point mutation and an intragenic exonic deletion, leading to the establishment of a definitive molecular diagnosis in a single assay. Our coverage depth analysis can unambiguously distinguish an apparent homozygosity from a real homozygous point mutation (P11). Our target capture MPS analysis spares the inconvenience of using two or more different diagnostic methods, greatly improving diagnosis in terms of time, cost, and yield.

### Potential limitations of capture-based MPS methods for clinical application

A common issue that limits the clinical application of CNV analysis by capture-based MPS is its inability to detect copy number changes in regions with insufficient coverage (<20×).<sup>2</sup> It is evident from our data that these problematic regions include GC-rich exons and exons containing highly repetitive sequences or sequences with high homology to regions elsewhere in the genome. These regions are difficult to capture and sequence efficiently and/or consistently. Therefore, the sequence depth generated from these regions is not reliable and has limited the ability to detect CNVs.

The detection of exonic copy number changes largely relies on the consistent coverage depth of an individual exon. The resolution of detection depends on experimental variations including genomic fragmentation, number of amplification cycles, as well as interference among different probes, that is, the so-called batch effect.<sup>30</sup> Focusing on a small group of genes is the key to reducing the unpredictable probe-probe interference and minimizing the batch effect. Using the coverage depth of each exon can significantly reduce the noise compared to using the coverage depth of each nucleotide. By doing so, the resolution is enhanced to a single exon. Our algorithm rarely provides sequence information of an intronic deletion breakpoint because only coding exons are captured and sequenced. If the breakpoints are located in the intronic regions that are not captured, there are no breakpoint sequence results from the analysis. If the breakpoints reside within exons, reads containing breakpoint sequence information can be retrieved for further analysis. We have developed an algorithm to detect breakpoint sequences by retrieving the unmatched sequence reads for realignment using less stringent parameters.<sup>8</sup>

In summary, our capture-based deep MPS approach allows concurrent detection of point mutations and exonic deletions. This is a much needed diagnostic improvement for a comprehensive one-step genetic analysis.

## SUPPLEMENTARY MATERIAL

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

## ACKNOWLEDGMENTS

We thank Eric Schmitt and Janice Smith for valuable discussions and critical editing of the manuscript. This study was partially supported by a Muscular Dystrophy Association grant to L.-J.C.W.

## DISCLOSURE

All authors are faculty members or employees in the Department of Molecular and Human Genetics at Baylor College of Medicine. The Medical Genetics Laboratories of the department offer extensive fee-based genetic tests including the use of massively parallel sequencing for molecular analyses of retinal diseases.

## REFERENCES

1. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010;61:437–455.
2. Mills RE, Walter K, Stewart C, et al.; 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.
3. Wang J, Zhan H, Li FY, Pursley AN, Schmitt ES, Wong LJ. Targeted array CGH as a valuable molecular diagnostic approach: experience in the diagnosis of mitochondrial and metabolic disorders. *Mol Genet Metab* 2012;106:221–230.
4. Wong LJ, Dimmock D, Geraghty MT, et al. Utility of oligonucleotide array-based comparative genomic hybridization for detection of target gene deletions. *Clin Chem* 2008;54:1141–1148.
5. Zhang W, Cui H, Wong LJ. Application of next generation sequencing to molecular diagnosis of inherited diseases. *Top Curr Chem* 2014;336:19–45.
6. Tang S, Wang J, Zhang VW, et al. Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects. *Hum Mutat* 2013;34:882–893.
7. Wang J, Cui H, Lee NC, et al. Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. *Genet Med* 2013;15:106–114.
8. Cui H, Li F, Chen D, et al. Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. *Genet Med* 2013;15:388–394.
9. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 2009;106:19096–19101.
10. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;28:1307–1313.
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. 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.
13. 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.
14. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;22:568–576.
15. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;27:2648–2654.
16. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;6(11 Suppl):S13–S20.
17. Poultney CS, Goldberg AP, Drapeau E, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am J Hum Genet* 2013;93:607–619.
18. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 2012;28:2711–2718.
19. Landsverk ML, Wang J, Schmitt ES, Pursley AN, Wong LJ. Utilization of targeted array comparative genomic hybridization, MitoMet, in prenatal diagnosis of metabolic disorders. *Mol Genet Metab* 2011;103:148–152.
20. Wang J, Rakhade M. Utility of array CGH in molecular diagnosis of mitochondrial disorders. *Methods Mol Biol* 2012;837:301–312.
21. Chinault AC, Shaw CA, Brundage EK, Tang LY, Wong LJ. Application of dual-genome oligonucleotide array-based comparative genomic hybridization to the molecular diagnosis of mitochondrial DNA deletion and depletion syndromes. *Genet Med* 2009;11:518–526.
22. Zhang S, Li FY, Bass HN, et al. Application of oligonucleotide array CGH to the simultaneous detection of a deletion in the nuclear TK2 gene and mtDNA depletion. *Mol Genet Metab* 2010;99:53–57.
23. Landsverk ML, Douglas GV, Tang S, et al. Diagnostic approaches to apparent homozygosity. *Genet Med* 2012;14:877–882.
24. Wang GL, Wang J, Douglas G, et al. Expanded molecular features of carnitine acyl-carnitine translocase (CACT) deficiency by comprehensive molecular analysis. *Mol Genet Metab* 2011;103:349–357.
25. Nord AS, Lee M, King MC, Walsh T. Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics* 2011;12:184.
26. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;322:881–888.
27. Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. *N Engl J Med* 2009;360:1699–1701.
28. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 2009;10:451–481.
29. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* 2006;52:103–121.
30. Karakoc E, Alkan C, O’Roak BJ, et al. Detection of structural variants and indels within exome data. *Nat Methods* 2012;9:176–178.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 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-nc-sa/3.0/>