ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Measuring coverage and accuracy of whole-exome sequencing in clinical context

Sek Won Kong, MD[1,2], In-Hee Lee, PhD[1,2], Xuanshi Liu, PhD[1,2], Joel N. Hirschhorn, MD, PhD[2,3] and Kenneth D. Mandl, MD, MPH[1,2,4]

Purpose: To evaluate the coverage and accuracy of whole-exome sequencing (WES) across vendors.

Methods: Blood samples from three trios underwent WES at three vendors. Relative performance of the three WES services was measured for breadth and depth of coverage. The false-negative rates (FNRs) were estimated using the segregation pattern within each trio.

Results: Mean depth of coverage for all genes was 189.0, 124.9, and 38.3 for the three vendor services. Fifty-five of the American College of Medical Genetics and Genomics 56 genes, but only 56 of 63 pharmacogenes, were 100% covered at 10 × in at least one of the nine individuals for all vendors; however, there was substantial interindividual variability. For the two vendors with mean depth of coverage > 120 ×, analytic positive predictive values (aPPVs) exceeded 99.1% for single-nucleotide variants and homozygous indels, and sensitivities were 98.9–99.9%; however, heterozygous indels showed lower accuracy and sensitivity. Among the trios, FNRs in the offspring were 0.07–0.62% at well-covered variants concordantly called in both parents.

Conclusion: The current standard of 120 × coverage for clinical WES may be insufficient for consistent breadth of coverage across the exome. Ordering clinicians and researchers would benefit from vendors' reports that estimate sensitivity and aPPV, including depth of coverage across the exome.

Genet Med advance online publication 12 April 2018

Key Words: breadth of coverage; depth of coverage; false negatives; pharmacogenomics; whole-exome sequencing

INTRODUCTION

Next-generation sequencing, widely used by researchers, is entering clinical care as a diagnostic test, but methods and quality vary across vendors. Customers should be fully informed of a test's accuracy in detecting coding variants, both to drive selection of high-quality vendors and to be aware of the likelihood of false positives and negatives.[1,2] The accuracy of variant detection in coding regions is lower for whole-exome sequencing (WES) than whole-genome sequencing, even at equivalent coverage.[3,4] Nonetheless, WES is widely used due to lower cost and because most disease-associated genomic variants discovered thus are in coding regions and splice sites of protein coding genes.

WES is performed as a series of biochemical and computational analytic procedures, varying from vendor to vendor, that influence exome coverage and genotype accuracy. Factors contributing to variation include (i) quality of genomic DNA,[5,6] (ii) DNA extraction methods,[7,8] (iii) sequence library preparation including exome capture[9] and polymerase chain reaction amplification,[10] (iv) the sequencing platform,[11,12] (v) short read length and depth of coverage,[12,13] (vi) computational analytical pipeline,[14] (vii) sequence contexts such as GC contents and simple repetitive DNA sequences,[11,15] and (viii) the type of variant (single-nucleotide variant (SNV), insertion/deletion (indels), and more complex variants).[16] Reanalysis of raw sequence reads using a standardized software pipeline can improve comparability between WES results[17] but cannot overcome the differences in targeted intervals, capture efficiency, and sequencing chemistry.[18]

We systematically compared the WES results from three vendors selected on the basis of varying exome capture methods with differing read-lengths and mean depths of coverage. We reprocessed raw reads using a single standard analytical pipeline to minimize variability due to bioinformatics pipelines among the vendors. For each gene, we focused on breadth of coverage at a minimum depth of 10 high-quality aligned reads. For each individual, concordant and discordant calls among the vendors were analyzed for genotype quality (GQ), depth of coverage (DP), and presence of reported minor allele frequency (MAF) to prioritize likely true-positive variants using all variant calls from the three vendors. Finally, we calculated the analytical positive predictive value (aPPV) and sensitivity for each vendor using a likely true-positive set, and estimated a lower bound for the false-negative rate (FNR) in each offspring among the three trios.

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA; [2]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA; [3]Broad Institute, Cambridge, Massachusetts, USA; [4]Department of Biomedical Informatics, Harvard Medical School, Boson, Massachusetts, USA. Correspondence: Sek Won Kong (Sekwon.Kong@childrens.harvard.edu)

Submitted 16 May 2017; accepted 16 February 2018; advance online publication 12 April 2018. doi:10.1038/gim.2018.51

MATERIALS AND METHODS

Samples and WES

Blood-derived DNA samples from three trios—hereafter referred to as trios A, B, and C—were collected and aliquoted per each vendor's specification. Exome sequencing and subsequent use for research was approved by the Boston Children's Hospital Committee on Clinical Investigation. We selected three regional vendors providing the CLIA-certified clinical sequencing service—denoted as V1, V2, and V3—all using a four-channel sequencing by synthesis technology platform (Illumina HiSeq 2500 for V1 and V2, and HiSeq 4000 for V3; Illumina, San Diego, CA, USA). Three different whole-exome capture methods were used: the Illumina Nextera Rapid capture (Nextera, V1), Agilent SureSelectXT (SureSelectXT, V2; Agilent Technologies, Santa Clara, CA, USA), and NimbleGen SeqCapEZ-MedExome (MedExome, V3; Roche NimbleGen Inc., Madison, WI, USA). Read-lengths were 76, 150, and 101 bps for V1, V2, and V3, respectively.

Comparison of capture targets, variant calling, and annotation

Comparison of exome capture target regions, calculation of coverage in target regions, and variant calling pipeline are described in the **Supplementary Methods** online. All VCF files were annotated using ENSEMBL Variant Effect Predictor release 86.[19] Rare and high-impact variants (RHIs) were defined in two ways: first, as MAF <0.005 in any population from the 1000 Genomes Project[20] or Exome Aggregation Consortium (ExAC)[21] and second, categorized as high-impact by the Variant Effect Predictor and resulting in frameshift, transcript ablation, transcript amplification, splice acceptor, splice donor, start lost, stop lost, and stop gain.

We compiled a list of 6,367 putative disease-associated genes from the following databases: the Human Gene Mutation Database (Professional 2016.02), OMIM,[22] ClinVar,[23] the Genetic Testing Registry,[24] the Cancer Gene Census,[25] and Orphanet.[26] All online databases were accessed on 27 October 2016 (**Supplementary Table S1** online). The nuclear genes implicated in clinical drug response and metabolism were collected from the Pharmacogenomics Knowledge Base's Very Important Pharmacogenes, excluding one mitochondrial gene *MT-RNR1* (PGx-VIPs, $N=63$).[27] Discordant variants in the American College of Medical Genetics and Genomics (ACMG) 56 genes[28] and PGx-VIPs were further inspected using the Integrated Genome Viewer (version 2.3.79, http://software.broadinstitute.org/software/igv/).

Analytical positive predictive value and sensitivity

To evaluate the performance of variant calls in detecting the true genotype for each vendor, we defined a "likely true-positive" set of variants among all unique variants identified by V1 and V2. In an individual, we compared variant calls between V1 and V2 in the genomic regions covered with ≥ 10 × in all three vendors (3COV) and used V3 as a tiebreaker. First, a concordant variant between V1 and V2 was considered likely true-positive. Second, a discordant variant between V1 and V2 was labeled as likely true-positive if V3 called the same variant. Finally, a discordant variant between V1 and V2 for which V3 failed to call was considered likely true-positive if $\text{GQ} \geq 20$ and $\text{MAF} > 0$ was reported in ExAC. Then we calculated aPPV (the proportion of likely true-positive variants out of total variants identified by the vendor) and sensitivity (the proportion of likely true-positive variants identified by the vendor out of total likely true-positive variants). FNR was calculated as *1–sensitivity* per exome.

False positives and false negatives among trios

We analyzed the segregation pattern in autosomes in offspring for the loci where both parents had DP ≥ 10 × in each trio. We analyzed all variants in overlapping 3COV regions of both parents where one parent was consistently called as heterozygous and the other as reference concordant homozygous in all three vendors. We then used the variant call in the offspring to estimate FNRs in these loci (see **Supplementary Methods** online).

For all summary statistics, mean values across nine individuals are shown in the Results. Standard deviations and the other descriptive statistical scores are detailed in Tables 1 and 2 and the **Supplementary Tables** online.

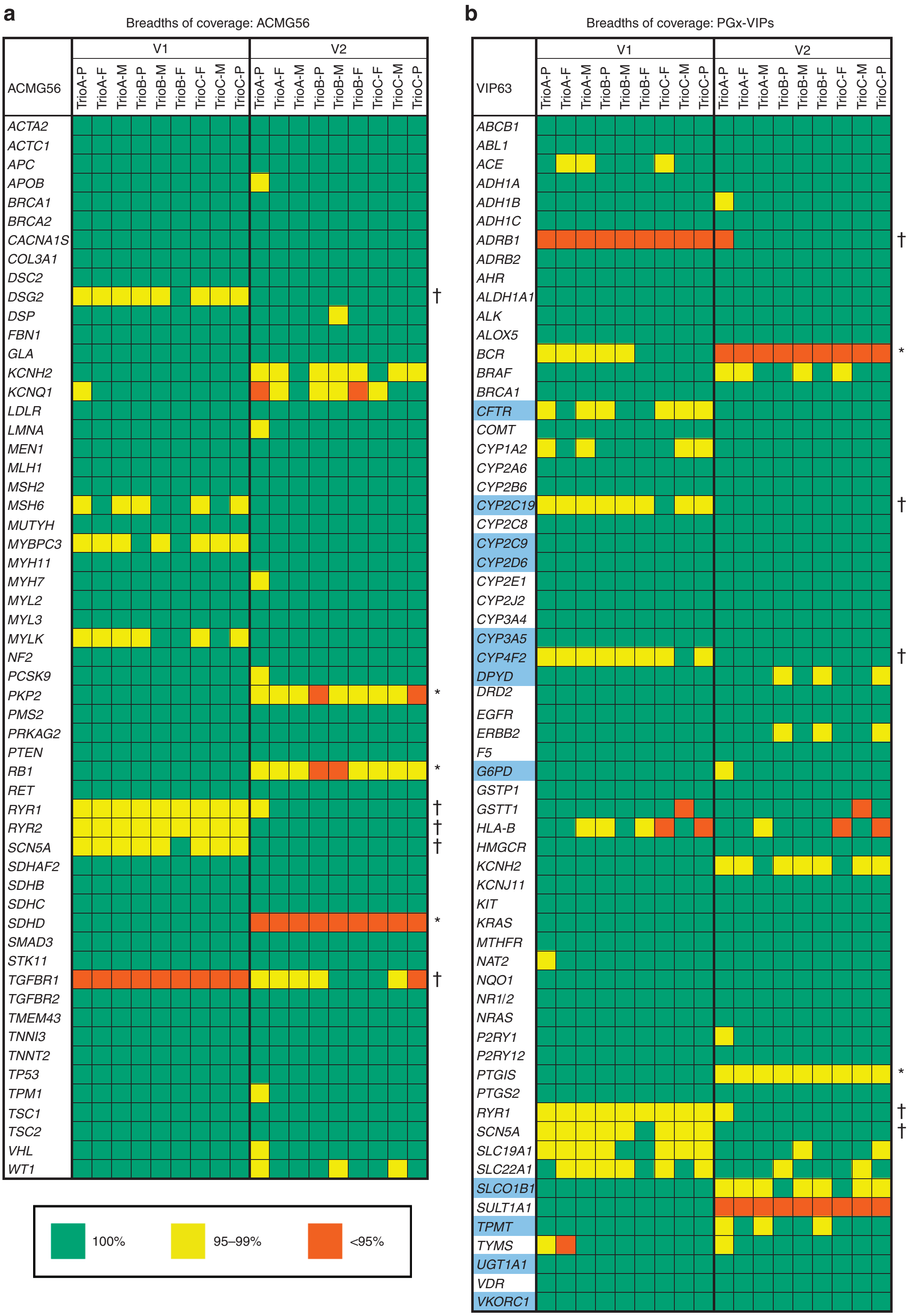
RESULTS

Depth and breadth of coverage for target regions

Compared with previous-generation technologies, current hybridization capture–based methods have superior coverage and probe design.[29] However, none of the three capture methods was designed to cover 100% of coding exons in the current version of Consensus Coding Sequence (CCDS).[30] Nonetheless, 99.97, 99.85, and 99.67% of CCDS overlapped the target regions of Nextera (V1), SureSelectXT (V2), and MedExome (V3), respectively. All three methods targeted ≥ 99.8% of coding exons of putative disease-associated genes ($N = 6{,}367$) (**Supplementary Table S2** online).

Mean depth of coverage across all CCDS genes was 189.0 (V1), 124.9 (V2), and 38.3 (V3), although there was a wide range of variation across genes (**Supplementary Figure S1** online). We analyzed coverage in more detail only for V1 and V2, because V3 did not meet the current standards typically used for WES. For each CCDS gene, we calculated the percent of exonic bases covered at ≥ 10 × that provides 95% sensitivity for heterozygous SNVs (**Supplementary Table S3** online).[4] The proportion of genes covered 100% at ≥ 10 × was 80.1% (V1) and 79.1% (V2) on average (**Supplementary Table S4** online).

For the ACMG 56 genes and PGx-VIPs, we compared the list of finished genes—i.e., 100% covered at ≥ 10 ×—in at least one of nine individuals for each vendor. Except for *RYR1*, *RYR2*, and *TGFBR1*, 53 genes were finished in at least one of the nine individuals by V1. Likewise, V2 had incomplete coverage for *PKP2*, *RB1*, and *SDHD* (**Figure 1a**). From the



PGx-VIPs ($N = 63$), 61 and 60 genes were finished by V1 and V2 respectively (**Figure 1b**). The breadth of coverage for both the ACMG 56 genes and PGx-VIPs was higher than exome-wide averages at all thresholds (Wilcoxon signed-rank tests p values <0.01, **Table 1**). For CCDS genes, V1 had consistently higher breadth of coverage compared with V2. However, no differences were found for the ACMG 56 genes and PGx-VIPs, except for PGx-VIPs at ≥10× (Wilcoxon

Figure 1 Variability in breadth of coverage for the American College of Medical Genetics and Genomics (ACMG) 56 genes and 63 pharmacogenes among the nine individuals. The percentages of coding sequence bases covered with per-site read depth ≥10× are shown for each of (**a**) ACMG 56 genes and (**b**) 63 genes from the Pharmacogenomics Knowledge Base Very Important Pharmacogenes (PGx-VIPs). Of 63 pharmacogenes, the 12 clinically actionable genes per the Clinical Pharmacogenetics Implementation Consortium guidelines are highlighted with blue background in their symbols. Each row represents a gene, and columns are grouped by the nine individuals across the three vendors. Green squares represent finished genes (i.e., 100% covered at ≥10×), and yellow (95–99%) and red (<95%) squares represent lower breadths of coverage at ≥10×. Genes with significantly different breadths of coverage between V1 and V2 are marked with * and † (Wilcoxon signed-rank tests, Bonferroni corrected *p* values <0.05). * indicates better breadth of coverage in V1 and † shows better ones in V2.

Table 1 Comparison of covered regions between two vendors

Regions of interest	Per-site effective coverage	V1 (189.0 ± 28.4)		V2 (124.9 ± 9.87)	
		Mean	SD	Mean	SD
CCDS coding sequences (total 32,279,934 bases)	≥10×	97.9	0.18	97.3	0.29
	≥20	96.9	0.40	95.6	0.62
	≥30	95.6	0.76	93.0	1.05
Putative disease-associated genes (total 14,258,789 bases)	≥10	98.9	0.13	98.7	0.19
	≥20	98.2	0.29	97.5	0.44
	≥30	98.7	0.58	95.7	1.08
ACMG 56 genes (total 196,428 bases)	≥10	99.8	0.04	99.8	0.13
	≥20	99.6	0.09	99.5	0.21
	≥30	99.2	0.28	99.0	0.32
PGx-VIPs (total 142,223 bases)	≥10	98.4	0.10	98.1	0.13
	≥20	97.7	0.23	97.6	0.23
	≥30	97.0	0.44	96.8	0.38

For each vendor, mean and standard deviation (SD) for percentage of coding sequence bases covered at different thresholds are calculated from the set of nine individuals. Mean depth of coverage is shown next to vendor identifier in top row. Consensus Coding Sequence (CCDS) genes (*N* = 18,616), putative disease-associated genes (*N* = 6,367) from disease gene databases, the American College of Medical Genetics and Genomics (ACMG) 56 genes, and the Pharmacogenomics Knowledge Base Very Important Pharmacogenes (PGx-VIPs, *N* = 63) are listed in **Supplementary Table S1** online. Sum of coding sequence bases for each region of interest is shown as total bases.

signed-rank test *p* value <0.01). Of note, *MYBPC3* and *TNNI3* were not completely targeted by SureSelectXT, but showed 100% coverage in one or more individuals. *SDHD* was not well covered by V2 (80.3% on average at ≥10×).

We observed a wide range in the breadth of coverage across the nine individuals. As such, only 47 (V1) and 41 (V2) of the ACMG 56 genes were finished in all nine individuals (**Figure 1a**). Twelve of the PGx-VIPs with the Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines (CPIC genes)[31] were finished by both V1 and V2 in at least one of the nine individuals (**Figure 1b**). Percentage of coverage of the PGx-VIPs also varied across individuals; 9 (V1) and 8 (V2) of the 12 CPIC genes were finished in all nine individuals. For the ACMG 56 genes, V1 had significantly better breadths of coverage for *PKP2*, *RB1*, and *SDHD*, and V2 had advantages for *DSG2*, *RYR1*, *RYR2*, *SCN5A*, and *TGFBR1* (Mann–Whitney–Wilcoxon tests, Bonferroni corrected *p* values <0.05). Seven of 63 PGx-VIPs showed significantly different breadth of coverage between V1 and V2 (**Figure 1b**).

We examined finished genes at ≥20× coverage that could provide 99% sensitivity for heterozygous SNVs.[4] At this threshold, 40 (V1) and 26 (V2) of the ACMG 56, and 9 (V1) and 5 (V2) of the 12 CPIC genes were finished in all nine individuals, suggesting incomplete breadth of coverage even with higher mean depth of coverage of V1 and V2 (**Supplementary Figure S2** online). At ≥20×, variability in breadth of coverage across nine individuals was significantly higher than at ≥10× (**Supplementary Figure S3** online). Breadth of coverage for four genes (*ATP7B*, *BMPR1A*, *OTC*, and *SMAD4*) recently added to the ACMG gene list for reporting secondary findings[32] is shown in **Supplementary Figure S4** online.

Concordant and discordant variant calls among the vendors in clinically implicated genes

For each variant type, we checked agreement of variant calls among the three vendors, restricting the analysis to genomic regions with ≥10× in all three vendors to minimize the effects of the lower coverage in V3 (**Table 2**). We also compared concordance rates between V1 and V2, to further minimize the effect of low coverage in V3. As expected, restricting the analysis to genomic regions with ≥10× coverage increased concordance rates for all types of variants (**Table 2** and **Supplementary Table S5** online). Of note, heterozygous indels showed the lowest concordance rates at any threshold compared with the other types of variants.

Table 2 Concordant variant calls among the vendors

		Heterozygous variants		Homozygous variants	
		SNVs	Indels	SNVs	Indels
V1-V2-V3 (≥10×)	Union of variants	18,702.4 (17,085–21,939)	1,961.4 (1,740–2,139)	9,945.3 (9,647–10,270)	607.7 (527–695)
	Concordant variants	17,478.1 (15,923–20,561)	1,051.7 (941–1,187)	9,839.6 (9,559–10,121)	501.0 (448–556)
	Concordance (%)	93.4 ± 0.85	53.7 ± 3.34	98.9 ± 0.34	82.6 ± 2.19
V1-V2 (≥10×)	Union of variants	19,369.8 (17,598–22,848)	2,298.4 (2,188–2,442)	10,397.9 (10,086–10,793)	697.9 (646–749)
	Concordant variants	18,356.2 (16,673–21,722)	1,426.9 (1,261–1,627)	10,304.0 (10,008–10,628)	580.2 (536–618)
	Concordance (%)	94.7 ± 0.68	62.0 ± 2.99	99.1 ± 0.26	83.2 ± 1.10
V1-V2 (≥20×)	Union of variants	18,488.9 (16,755–21,840)	1,888.1 (1,786–2,013)	9,923.4 (9,637–10,229)	593.9 (546–637)
	Concordant variants	17,649.0 (15,994–20,929)	1,234.8 (1,108–1,417)	9,856.4 (9,566–10,129)	515.7 (481–552)
	Concordance (%)	95.4 ± 0.71	67.7 ± 3.12	99.5 ± 0.09	89.0 ± 1.19

Overall concordance rates of called variants are significantly different across variant types. For each variant type, we calculate the proportion of concordant variants among the union of variants called by the three vendors. A total of 58,870 unique variants were called including 682 multiallelic variants. Mean and range across the nine individuals are shown for the number of union and concordant variants, and mean and standard deviation are shown for concordance rates. "V1-V2-V3" refers to results from variants covered with ≥ 10× in all three vendors, and "V1-V2" refers to results from variants covered with ≥ 10× or 20× by both V1 and V2. Concordant variants in "V1-V2-V3" include the ones with complete concordance among the three vendors.
SNV, single-nucleotide variant.

Concordance rates for the variants in clinically implicated genes such as putative disease-associated genes, the ACMG 56, and PGx-VIPs were consistently higher compared with those of exome-wide averages. When calculating the concordance rate, we excluded all variants in the major histocompatibility complex region (HMC, chr6:28,477,798–33,448,354). In 6,367 putative disease-associated genes, concordance rates between V1 and V2 with ≥ 10× were consistently higher: 96.8 (heterozygous SNVs), 99.5 (homozygous SNVs), 63.3 (heterozygous indels), and 85.2% (homozygous indels). Compared with the concordance rates calculated with VCF files generated using a single software pipeline in our study, concordance rates with vendor-provided VCF files were consistently lower (**Supplementary Table S5** online).

For the ACMG 56 genes, 98.3% of heterozygous and 100% of homozygous SNVs were concordant between V1 and V2. Discordant variants in the ACMG 56 were found in *APOB*, *KCNH2*, *PKP2*, *PMS2*, *SCN5A*, and *TSC2* (**Supplementary Table S6** online). No homozygous indel was found in the ACMG 56 genes across the nine individuals, and only three of eight heterozygous indels were concordant between V1 and V2. For the PGx-VIPs, all eight homozygous indels and 8 of 21 heterozygous indels were concordant between V1 and V2. Concordance rates were 94.6 and 99.6% for heterozygous and homozygous SNVs, respectively. Overall, concordance rate among the ACMG 56 and PGx-VIPs was higher than the exome-wide average except for indels.

Rare genetic variants with deleterious impacts on protein function are often prioritized in WES for further evaluation and validation if such variants are found in putative disease-associated genes. To minimize false negatives for this class of variants, we used all RHI variants discovered by V1 and V2 without restricting to ≥ 10× covered regions, and selected RHI variants in putative disease-associated genes. Concordance rate for heterozygous RHI SNVs was 90.4%, and all five homozygous RHI SNVs found were concordant. All RHI SNVs were concordant for genomic regions covered with ≥ 10× in both V1 and V2. There were no homozygous RHI indels, but only 11 of 46 heterozygous RHI indels from the nine individuals were concordant. We visually inspected read alignments in the regions surrounding discordant RHI variants in Integrated Genome Viewer. The discordant calls were associated with (i) low depth of coverage, (ii) allelic imbalance, (iii) strand bias, (iv) read alignments suggesting structural variation and indels, and (v) homopolymers (**Supplementary Table S6** online). Next, we examined potential false positives among concordant variant calls between V1 and V2. We restricted variant calls to heterozygous RHI variants discovered in putative disease genes (excluding MHC), which resulted in 30 SNVs and three indels in three probands. None of these were likely false positives according to segregation pattern—i.e., they were all inherited from only one of the two parents. Therefore, the proportion of likely false positives among concordant calls between V1 and V2 would be very low.

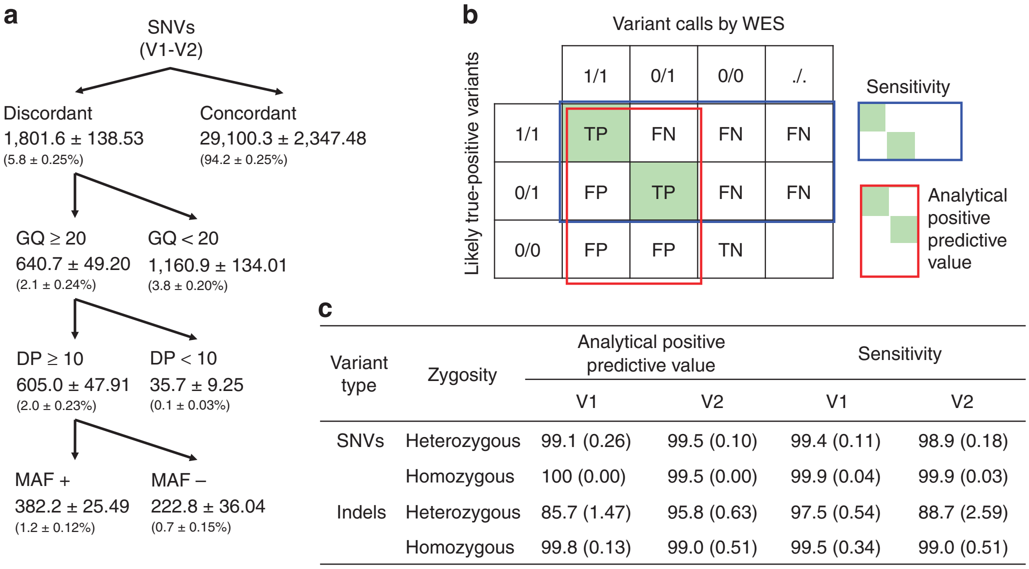
We found two discordant variants that were reported as pathogenic in ClinVar. Pathogenic heterozygous SNVs—one in *FCGR1A* and the other in *VPS13B*—were called by V1, but V2 failed to call the variants because of low coverage for both loci (**Supplementary Table S6 online**).

Analytical positive predictive value and sensitivity of variant calls

Discordant calls are due to either false-positive calls in one or more vendors, or a failure to detect a variant in one or more vendors. To further characterize discordant variant calls, we examined GQ, DP, and reported MAFs > 0 in ExAC for these. Among the discordant SNVs between V1 and V2 (5.8% of total SNVs), 3.8% had low GQ (<20) in one of two vendors. For the discordant SNVs with good GQ (≥20), 0.1% had low DP (<10) in one of two vendors. Overall, 2.0% of SNVs that were discordant between V1 and V2 had good genotype quality with ≥10× depth of coverage, of which 1.2% had reported MAFs >0 in the ExAC server (**Figure 2a** and **Supplementary Table S7 online**).

We considered a minimal set of likely true-positive calls to be with a MAF reported in ExAC, GQ ≥ 20, and DP ≥ 10. GQ and DP are often used to filter out variants with erroneous variant calls;[33] however, applying these filtering criteria could result in false negatives in WES. The proportion of retained variants among completely concordant variant calls across all three vendors-hereafter referred to as 3CON-varied by variant type and among the vendors (**Supplementary Figure S5 online**). In V1, 99.6% of heterozygous SNVs passed both DP ≥ 10 and GQ ≥ 20 thresholds. Using the same criteria, 99.3% of heterozygous SNVs retained for V2 (**Supplementary Table S8 online**).

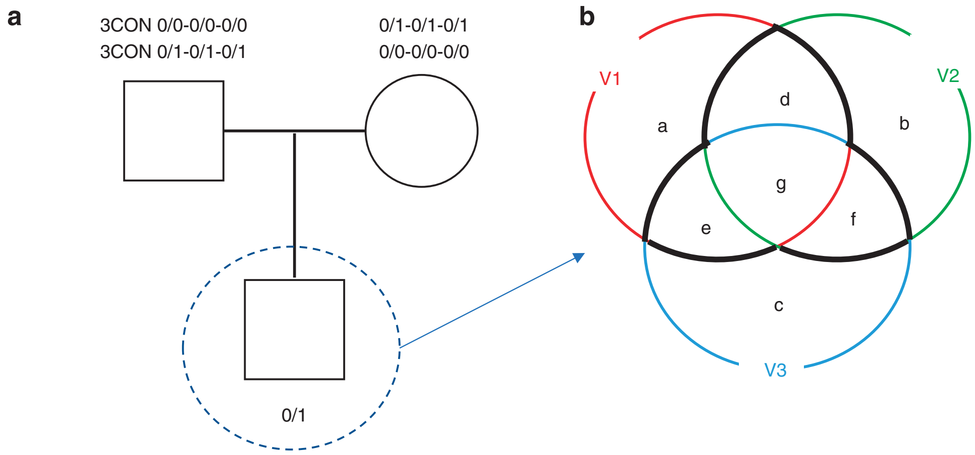
To calculate the sensitivity and aPPV of each vendor's result, we selected likely true-positive calls from aggregated variant calls from V1 and V2. The blue rectangle in **Figure 2b** represents the likely true-positive set from V1 and V2, and all variant calls from a given vendor are represented by the red rectangle. For both vendors, aPPVs and sensitivities were 99% or higher except for heterozygous indels. V1 showed higher sensitivities for heterozygous indels but lower aPPVs compared with V2 (**Figure 2c**). Of note, V2 had lower sensitivity for heterozygous indels compared with the other types of variant. V1 showed the lowest aPPV for heterozygous indels compared with the other types of variant and also compared with V2. Thus, accurate detection of heterozygous indels would be challenging for both V1 and V2. Our estimation of sensitivity represents an upper bound because true-positive variants might not have been captured and/or called by any of two vendors, and variant calls that did not pass DP and GQ thresholds could also include true-positive variants.



	1/1	0/1	0/0	./.
1/1	TP	FN	FN	FN
0/1	FP	TP	FN	FN
0/0	FP	FP	TN	

Variant type	Zygosity	Analytical positive predictive value		Sensitivity	
		V1	V2	V1	V2
SNVs	Heterozygous	99.1 (0.26)	99.5 (0.10)	99.4 (0.11)	98.9 (0.18)
	Homozygous	100 (0.00)	99.5 (0.00)	99.9 (0.04)	99.9 (0.03)
Indels	Heterozygous	85.7 (1.47)	95.8 (0.63)	97.5 (0.54)	88.7 (2.59)
	Homozygous	99.8 (0.13)	99.0 (0.51)	99.5 (0.34)	99.0 (0.51)

Figure 2 Analytical positive predictive value and sensitivity of variant calls from each vendor. (**a**) Biallelic single-nucleotide variants (SNVs) concordant and discordant between V1 and V2 were sequentially checked for genotype quality (GQ) score, per-site depth of coverage (DP), and reported minor allele frequency (MAF) in the Exome Aggregation Consortium (ExAC) server. The number of concordant and discordant SNVs is shown, as is the average number and standard deviation of variants meeting each criterion across nine individuals. In parentheses, the same statistics are expressed as a percentage of the total number of variants seen in either vendor. (**b**) The scheme for calculating analytical positive predictive value (aPPV) and sensitivity. In each individual, a likely true-positive set of variants is compiled by aggregating all unique variants seen in both V1 and V2 (blue rectangle) that had GQ ≥20, DP ≥10, and (for vendor-specific variants only) MAF >0 in ExAC. All variants discovered by a vendor are used to calculate aPPV (the green shaded squares divided by the red rectangle) and sensitivity (the proportion of the green shaded squares divided by the blue rectangle). (**c**) The mean and standard deviation across nine individuals for aPPV and sensitivity for each variant type and vendor. FN, false negative; FP, false positive; TP, true positive; WES, whole-exome sequencing.



c

Proband variant call (V1-V2-V3)		Group	Trio A	Trio B	Trio C
a	0/1-0/0-0/0	V1 specific	9	8	3
b	0/0-0/1-0/0	V2 specific	8	6	7
c	0/0-0/0-0/1	V3 specific	4	2	0
d	0/1-0/1-0/0	2CON, likey false negative of V3	23	31	38
e	0/1-0/0-0/1	2CON, likey false negative of V2	68	46	16
f	0/0-0/1-0/1	2CON, likey false negative of V1	8	13	10
g	0/1-0/1-0/1	3CON	10,907	8,092	7,873
f/(d+e+f+g)		False negative rate V1	0.07%	0.16%	0.13%
e/(d+e+f+g)		False negative rate V2	0.62%	0.56%	0.20%
d/(d+e+f+g)		False negative rate V3	0.21%	0.38%	0.48%

Figure 3 Vendor-specific false-negative rates in trio. (**a**) For each trio, we calculate number of vendor-specific false negatives in offspring by focusing on the completely concordant variant calls across all three vendors (3CON) autosomal loci where one parent is heterozygous (i.e., 0/1–0/1–0/1 for V1, V2, and V3, respectively) and the other is reference concordant homozygous. We restrict the analysis to genomic regions covered with ≥ 10 × by all three vendors. (**b**) A Venn diagram shows each subgroup of concordant and discordant variant calls for proband. The black solid line represents the total number of variants discovered by two or more vendors, which we use as a denominator to calculate vendor-specific false-negative rates (FNRs) in each trio. (**c**) FNRs and the numbers of 3CON, concordant variant calls between two vendors (2CONs), and vendor-specific heterozygous variants found in offspring. For instance, in trio A, the FNR for V1 (0.07%) is calculated as the number of loci where V1 did not detect the heterozygous variant (eight variants, corresponding to "f") divided by the total number of likely true set of heterozygous variants (i.e., 2CONs and 3CON) in the proband (23 + 68 + 8 + 10,907 = 11,006, corresponding to "d + e + f + g").

False positives and false negatives among trios

We took advantage of the family-based trio design to calculate FNRs in regions of high coverage (**Figure 3**), using the variant calls from all three vendors, but restricting the analysis to genomic regions covered with ≥ 10 × by all three vendors. In trio A, a total of 58,458 variants were found by the three vendors, of which 80.9% (47,281/58,458) were in 3COV regions for both parents (**Supplementary Table S9 online**). From these, we focused on the 22,397 loci for which both parents had 3CON and where one parent was heterozygous and the other was homozygous for the reference allele. For these variants, the offspring should be heterozygous about 50% of the time. In the offspring of trio A, 11,042 (49.3% of 22,397) were called as heterozygous in at least one vendor, of which 10,907 were 3CON. For these loci, concordance rates in the offspring of trio A were higher for both SNVs and indels—98.9 and 95.7%, respectively—compared with genome-wide concordance rates; these loci also showed higher concordance rates in the six unrelated individuals from trios B and C (98.7 and 92.4% for heterozygous SNVs and indels, respectively). We identified variants as likely false negatives if there was a no-call or a homozygous reference call in the offspring for trio A for one vendor and a heterozygous call in the other two vendors. There were 8 (V1), 68 (V2), and 23 (V3) likely false negatives, of which 0% (V1:0/8), 14.7% (V2:10/68), and 0% (V3:0/23) were due to low coverages (< 10 ×). Of eight likely false negatives from V1, seven were called as homozygous for the reference allele and one was a no-call. Of 68 loci that V2 did not call heterozygous, 6 were no-calls and 62 were called as homozygous for the reference allele. For V3, all 23 were called as homozygous for the reference allele. We did not consider the variants that were called but had discordant numbers of variant alleles among the vendors—e.g., heterozygous in two vendors and homozygous for the variant allele in the other—because these are more likely "genotype errors" rather than false negatives.

We then used these false negative loci to estimate a lower bound for FNRs for each trio. For the denominator of likely true-positive variants, we used variants that were called as heterozygous in the offspring in at least two vendors; in trio A there were 11,006 such variants. The lower bound of vendor-specific FNRs in these well-covered regions were 0.07% for V1 (8/11,006), 0.62% for V2 (68/11,006), and 0.22% for V3 (23/11,006). FNRs estimated using the same analysis strategy were 0.16% (V1), 0.56% (V2), and 0.38% (V3) for trio B, and 0.13% (V1), 0.20% (V2), and 0.48% (V3) for trio C (**Figure 3c**) (we included analysis of V3, even with a lower sequencing depth, because we focused on genomic loci covered at a minimum of 10× in all three vendors). Our estimation of FNRs is comparable with the one reported by Li et al.[34] in family samples. Variants in low-coverage regions (<10×)—comprising, for instance, 10.0% of CCDS in trio A—were not used to calculate FNRs; however, the FNR in offspring was likely very low for high-coverage regions in both parents, supporting the strength of trio sequencing for some clinical conditions,[35] at least in regions of consistently high coverage.

DISCUSSION

We performed comparative analysis of WES results from three vendors to measure the empirical coverage of medically implicated genes and concordance rate of variant calls among the vendors, using uniform variant calling methods to remove variability from software analytical pipelines. The mean percentage of CCDS coding regions covered with ≥10× was above 95% for the nine individuals and all vendors. For the ACMG 56 and PGx-VIPs, we found a wide range of difference in breadth of coverage with ≥10× across the nine individuals. The mean depth of coverage provided by V2 was typical for clinical WES;[36] however, *KCNQ1*, *PKP2*, *RB1*, and *TGFBR1* had variable coverage across the nine individuals with <95% coverage in some individuals. The depth of coverage provided by V3 would be suboptimal for clinical use.

RHI variants in putative disease-associated genes were less concordant than exome-wide averages except for homozygous SNVs, suggesting that these variants were enriched for substantial numbers of false positives, false negatives, or variant calling errors. Among these, low coverage in a vendor usually was the source of discordant calls, and the variant was likely a true positive in the other vendor(s). Moreover, structural variation, homopolymer, or simple repeats were frequently found in the flanking regions of discordant RHI variants. Therefore, further evaluation and validation including visual inspection of aligned reads and validation using an orthogonal method will be particularly important for RHI variants. Of note, some rare variants have low coverage in population-scale databases such as ExAC. For instance, only 2 of 14 discordant RHI variants had good coverage in ExAC. Conversely, estimating accuracy mostly with common variants can bias the results for well-covered genomic regions in population databases where it would be easier to call variants using next-generation sequencing. Analyzing aPPV and sensitivity of rare variants with or without good coverage in population-scale databases for research and clinical applications would be an interesting research topic but is beyond the scope of our current study.

For the sites where both parents were highly covered by the three vendors, vendor-specific FNRs in offspring were low for all vendors and three trios (0.07–0.62%). This estimation of FNR cannot be extrapolated to the rest of the genome, because false negatives are more prevalent outside of the genomic regions covered with ≥10× in 3COV regions (which comprise 10.0, 5.5, and 7.3% of CCDS in trios A, B, and C, respectively). Similarly, variants that are harder to call may have higher FNRs, as illustrated by loci in 3COV regions with discordant calls in parents (i.e., 19.1, 12.1, and 15.1% of all variants found in trio A, B, and C, respectively). Even so, our results support the strength of trio WES sequencing for molecular genetic diagnosis for most SNVs in regions of high coverage.[35]

Our study has some limitations. First of all, although the vendors covered a range of read-lengths—76 (V1), 150 (V2), and 101 bps (V3)—and mean coverage from 38.3 (V3) to 189.0 (V1), we only sampled a small number of vendors, so our results may not generalize to other vendors or sequencing platforms. To minimize additional sources of variability, we used the same DNA stocks and analytical pipeline, but it is possible that different pipelines or different quality DNA samples could have yielded different results.[37] Notably, we did not validate discordant genetic variants among the vendors (therefore the performance measures were relative between vendors), nor did we use reference genetic material for which several gold standard variant call sets are available, making it difficult to definitively classify discordant genotypes into false positives, false negatives, or genotype errors. However, our study used trio samples and multiple vendors to overcome some of these limitations. Furthermore, the utility of a gold standard set of variants from NA12878 may itself be somewhat limited because it is widely used to ensure quality (and hence may be used to optimize platforms for calling this particular set of variants) but does not cover all known disease-associated genes. Estimation of sensitivity and aPPV from an individual genome may also be difficult to generalize to a range of samples;[38] we observed substantial interindividual variation of the breadth of coverage for clinically implicated genes across the nine individuals in the current study.

Mean depth of coverage was a general indicator of overall sensitivity, but did not capture variability in coverage across potentially clinically important genes.[37] For example, V2 provided a standard mean depth of coverage (125×) for clinical research WES, but showed large variability across individuals in the breadth of coverage for the ACMG 56 and CPIC genes, and had consistently higher FNRs in three trios compared with V1. It appears that a depth of coverage for clinical WES may need to be more even, or to be closer to what V1 provided (190×) to achieve consistent coverage across actionable genes across many individuals. Higher depths of coverage (e.g., ≥20× and 30×) were required to improve concordance rates between V1 and V2 for indels. Nonetheless, exonic regions with high sequence homology are challenging to analyze using WES.[30]

Establishing the clinical utility of WES therefore requires ongoing measurement of the breadth and depth of coverage and accuracy—both across vendors and within individual vendors over time.[39,40] We observed substantial interindividual variation in coverage of medically implicated genes. Because aPPV and sensitivity are imperfectly captured by mean coverage, we suggest that clinical WES service providers should inform users as to the range of sensitivity and aPPV for different classes of variants across the sets of genes that are relevant to the particular clinical scenario, estimated across a large cohort of clinical samples. This sort of information will help clinicians both select services and also interpret clinical reports and distinguish truly negative findings from false negatives due to low coverage.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

This work was supported by the Boston Children's Hospital Precision Link initiative. S.W.K. was supported, in part, by a grant from the National Institutes of Health (R01MH107205).

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Kuderer NM, Burton KA, Blau S, et al. Comparison of 2 commercially available next-generation sequencing platforms in oncology. *JAMA Oncol* 2017;3:996–998.
2. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 2014;311:1035–1045.
3. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 2015;112:5473–5478.
4. Meynert AM, Ansari M, FitzPatrick DR & Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 2014;15:247.
5. Zhu Q, Hu Q, Shepherd L, et al. The impact of DNA input amount and DNA source on the performance of whole-exome sequencing in cancer epidemiology. *Cancer Epidemiol Biomarkers Prev* 2015;24:1207–1213.
6. Londin ER, Keller MA, D'Andrea MR, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* 2011;12:464.
7. Koshy L, Anju AL, Harikrishnan S, et al. Evaluating genomic DNA extraction methods from human whole blood using endpoint and real-time PCR assays. *Mol Biol Rep* 2017;44:97–108.
8. Ghaheri M, Kahrizi D, Yari K, Babaie A, Suthar RS & Kazemi E. A comparative evaluation of four DNA extraction protocols from whole blood sample. *Cell Mol Biol (Noisy-le-grand)* 2016;62:120–124.
9. Fisher S, Barry A, Abreu J, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 2011;12:R1.
10. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011;12: R18.
11. Meienberg J, Zerjavic K, Keller I, et al. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res* 2015;43: e76.
12. Clark MJ, Chen R, Lam HY, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;29:908–914.
13. Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 2011;12:R94.
14. Hwang S, Kim E, Lee I & Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep* 2015;5:17875.
15. Treangen TJ & Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;13:36–46.
16. Fang H, Wu Y, Narzisi G, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 2014;6:89.
17. Linderman MD, Brandt T, Edelmann L, et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics* 2014;7:20.
18. Wang Q, Shashikant CS, Jensen M, Altman NS & Girirajan S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep* 2017;7:885.
19. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17:122.
20. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.
21. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–291.
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA & McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;33(database issue): D514–517.
23. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;44 (D1):D862–868.
24. Rubinstein WS, Maglott DR, Lee JM, et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res* 2013;41(database issue):D925–935.
25. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;39(database issue):D945–950.
26. Pavan S, Rommel K, Mateo Marquina ME, Hohn S, Lanneau V & Rath A. Clinical practice guidelines for rare diseases: the Orphanet database. *PLoS One* 2017;12:e0170365.
27. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, et al. The Pharmacogenetics and Pharmacogenomics Knowledge Base: accentuating the knowledge. *Nucleic Acids Res* 2008;36(database issue):D913–918.
28. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
29. Chilamakuri CS, Lorenz S, Madoui MA, et al. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;15:449.
30. Mandelker D, Schmidt RJ, Ankala A, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med* 2016;18:1282–1289.
31. Caudle KE, Klein TE, Hoffman JM, et al. Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Curr Drug Metab* 2014;15:209–217.
32. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 2017;19:249–255.
33. Carson AR, Smith EN, Matsui H, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 2014;15:125.
34. Li B, Chen W, Zhan X, et al. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet* 2012;8: e1002944.

35. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 2014;312: 1880–1887.
36. Jamal SM, Yu JH, Chong JX, et al. Practices and policies of clinical exome sequencing providers: analysis and implications. *Am J Med Genet A* 2013;161A:935–950.
37. Roy S, Coldren C, Karunamurthy A, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* 2018;20: 4–27.
38. Weisenfeld NI, Yin S, Sharpe T, et al. Comprehensive variation discovery in single human genomes. *Nat Genet* 2014;46:1350–1355.
39. Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* 2012;30:1033–1036.
40. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013;15:733–747.