Assessing the capability of massively parallel sequencing for opportunistic pharmacogenetic screening

David Ng, MD[1], Celine S. Hong, PhD[1], Larry N. Singh, PhD[1], Jennifer J. Johnston, PhD[1], James C. Mullikin, PhD[2,3], Leslie G. Biesecker, MD[1,2]; on behalf of the NISC Comparative Sequencing Program

Purpose: The aim of the study was to assess exome data for pre-emptive pharmacogenetic screening for 203 clinically relevant pharmacogenetic variant positions from the Pharmacogenomics Knowledgebase and Clinical Pharmacogenetics Implementation Consortium and identify copy-number variants (CNVs) in *CYP2D6*.

Methods: We examined the coverage and genotype quality of 203 pharmacogenetic variant positions in 973 exomes compared with five genomes and with five genotyping chip data sets. Then, we determined the agreement of exome and chip genotypes by evaluating concordance in a three-way comparison of exome, genome, and chip-based genotyping at 1,929 variant positions in five individuals. Finally, we evaluated the utility of exomes for detecting *CYP2D6* CNVs.

Results: For 5 individuals examined for 203 pharmacogenetic variants ($5 \times 203 = 1,015$), 998/1,015 were identified by genome,

849/1,015 were identified by exome, and 295/1,015 by genotyping chip. Thirty-six pharmacogenetic star allele variants with moderate to strong Clinical Pharmacogenetics Implementation Consortium (CPIC) therapeutic recommendations were identified in 973 exomes. Exomes had high (98%) genotype concordance with chip-based genotyping. *CYP2D6* CNVs were identified in 57/973 exomes.

Conclusions: Exomes outperformed the current chip-based assay in detecting more important pharmacogenetic variant positions and *CYP2D6* CNVs for preemptive pharmacogenetic screening. Tools should be developed to derive pharmacogenetic variants from exomes.

Genet Med advance online publication 18 August 2016

Key Words: massively parallel sequencing; pharmacogenetic screening

INTRODUCTION

One of the goals of precision medicine is to use pharmacogenomics to optimize treatment efficacy and minimize adverse drug reactions. Barriers to the implementation of pharmacogenomics-guided therapy include the turnaround time for obtaining a pharmacogenetic (PGx) result[1] and the clinical utility of returning PGx variants.[2] One recommendation for avoiding treatment delays is to implement preemptive PGx testing.[1] Current PGx testing uses array-based genotyping platform— e.g., Affymetrix DMET Plus (Drug Metabolizing Enzymes and Transporters array)—screens for a predefined set of PGx variants.[3,4] Genomic testing platforms such as exome sequencing (ES) or genome sequencing (GS),[5] also called massively parallel sequencing (MPS), have potentially wider utility than the aforementioned genotyping platforms, and this begs the question of whether MPS sequence data could be used for preemptive PGx testing. Part of the larger challenge for the field of medical genomics is to identify all potential uses of sequencing so that the cost of these assays can be amortized across multiple

applications, thereby decreasing the effective cost of the test. Prior studies with small sample sizes showed high ES genotype concordance rate with other platforms (99.6% with MiSeq and 98.9% with iPLEX ADME PGx panel)[6] and variable (60–80%) ES coverage of DMET Plus PGx variant positions depending on the capture kit used.[7] An extensive analysis with a larger data set was needed to assess the capability of ES in detecting clinically relevant PGx variants. We set out to assess MPS concordance and coverage of annotated PGx variants compared with a current genotyping platform to determine whether MPS could serve as a genotyping source for preemptive PGx testing.

MATERIALS AND METHODS

Participants

This study was performed at the NIH Clinical Center as part of the ClinSeq project and included 973 participants enrolled between 45 and 65 years of age who were consented for baseline clinical tests, ES and/or GS, return of genetic results, and iterative phenotyping based on an individual's genetic variants.[8,9] The National Human Genome Research Institute

Institutional Review Board reviewed and approved this study. See **Supplementary Methods** online.

Selection of clinically relevant pharmacogenetic variants for comparison

We identified 50 Pharmacogenomics Knowledgebase (Pharm-GKB) level 1A and 1B PGx variants (https://www.pharmgkb.org/) and 154 Clinical Pharmacogenetics Implementation Consortium (CPIC) (http://www.pharmgkb.org/page/cpic) variants from 40 gene–drug pairs with level A evidence (two promoter variants were located at the same genomic position) for a total of 203 PGx variant positions. We evaluated coverage of these 203 PGx variant positions from 973 exomes, 5 genomes, and 5 chip data. Three *HLA-B* variants (HLA-B*52:01:01, HLA-B*57:01:01, HLA-B*58:01:01) were excluded because they were not amenable to genotyping by the chip, ES, or GS. MPS genotype concordance was determined by comparing five individuals with ES, GS, and DMET Plus genotypes (hereafter referred to as the chip). The chip has been previously shown to have high genotype concordance (91 to 99%) compared with six orthogonal genotyping platforms.[10] We selected *CYP2D6* for copy-number variant (CNV) analysis because 1–2% of individuals carry more than two functional copies that may have an ultrarapid metabolizer phenotype that can lead to codeine toxicity.[11]

Laboratory methods

See **Supplementary Methods** online.

RESULTS

Detection of 203 CPIC/PharmGKB variant positions by exome versus genome versus chip

Five individuals were examined for 203 curated variants (132 coding, 71 noncoding positions) by ES, GS, and chip-based testing. One would ideally like to detect a total of 1,015 genotypes (203×5). A total genotype count regardless of genotype quality from five individuals is shown in **Figure 1**. GS detected 998/1,015 genotypes (657/660 coding and 341/351 noncoding). In the coding positions, 129/132 positions were covered in five individuals and 3/132 were covered in four individuals. In the noncoding positions, 63/71 positions were covered in five, 5/71 in four, and 3/71 in two individuals. For ES, 117/203 positions were targeted by two capture kits for five individuals (Agilent38Mb $n = 2$, TruSeqV2 $n = 3$), 12/203 were targeted only by Agilent38Mb, and 14/203 were targeted only by TruseqV2. The expected total genotype count is 651 ((117×5) + (12×2)) + (14×3)). The targeted genotype detection rate was 647/651. Of the positions targeted by both capture kits, 114/117 variant positions were covered in five individuals, 2/117 in four individuals, and 1/117 in three individuals. All 26 positions targeted by only one of the capture kits had complete coverage. The total ES genotype count was 849 (647 targeted and 202 off-target). The chip targeted 46/132 coding and 14/71 noncoding positions, and the targeted detection rate was 225/230 (coding) and 70/70 (noncoding) (**Figure 1**, **Supplementary Tables S1–S3** online). The in-house cost
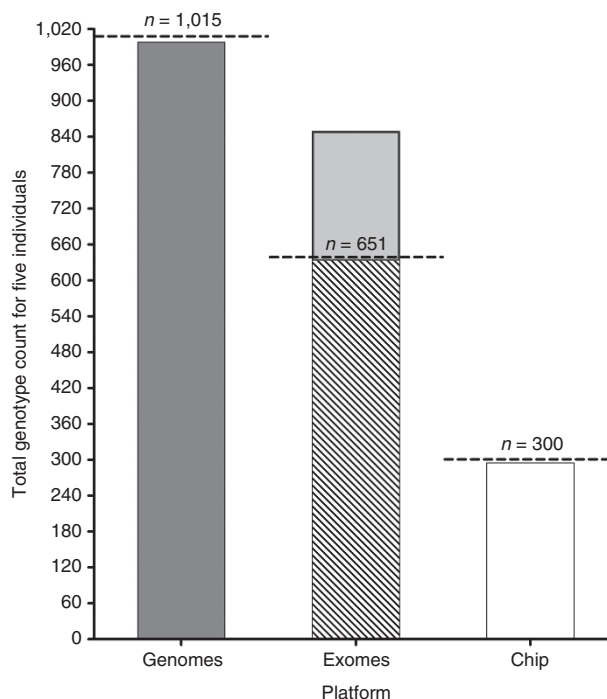

Figure 1 Genotype count of five genomes, five exomes, and five chip data at 203 pharmacogenetic variant positions in five individuals. Total genotype count for five individuals at 203 pharmacogenetic variant positions regardless of genotype quality from genome, exome, and chip. The total genotype count is 1,015 for genomes: 651 targeted genotypes for exomes and 300 targeted genotypes for the chip (represented by the horizontal dashed line). Genomes detected 998/1,015, exomes detected 849/1,015 (647 targeted represented by the diagonally striped area and 202 off-target represented by the light gray area), and the chip detected 295/1,015. Chip, Affymetrix DMET Plus (Drug Metabolizing and Transporters array).

per genotypable site was $43.77 ($8,710/199 positions) for genomes, $4.79 ($810/169 positions) for exomes, and $9.31 for the chip ($549/59 positions). These figures may not reflect clinical costs.

We next examined the detection rate of high-quality genotypes (GQ ≥50) per individual at 203 positions. We included 973 exomes captured with four kits (Agilent38Mb, Agilent50Mb, TruSeqV1, TruSeqV2). ES, GS, and chip data were grouped as coding and noncoding variants (intergenic, intronic, promoter, or 3′ untranslated region). GS and ES detected an average of 101 and 120 genotypes per individual at coding positions, respectively. At noncoding positions, GS and ES detected an average of 55 and 27 genotypes per individual, respectively (**Figure 2a,b**; ES average based on TruSeqV1 and V2 data). ES coverage was the highest in coding regions and the TruSeqV2 kit had the highest average (122); the chip captured 45 genotypes per individual (**Figure 2a**). ES coverage in noncoding regions was low. Among the 71 noncoding positions, TruSeqV1/V2 had the highest average (27) and the Agilent38Mb kit and the chip had the lowest average (14) of genotypes per individual (**Figure 2b**). GS coverage was outperformed by the Agilent50Mb and TruSeqV1/V2 kits in coding regions (**Figure 2a**; **Supplementary Table S4** online).
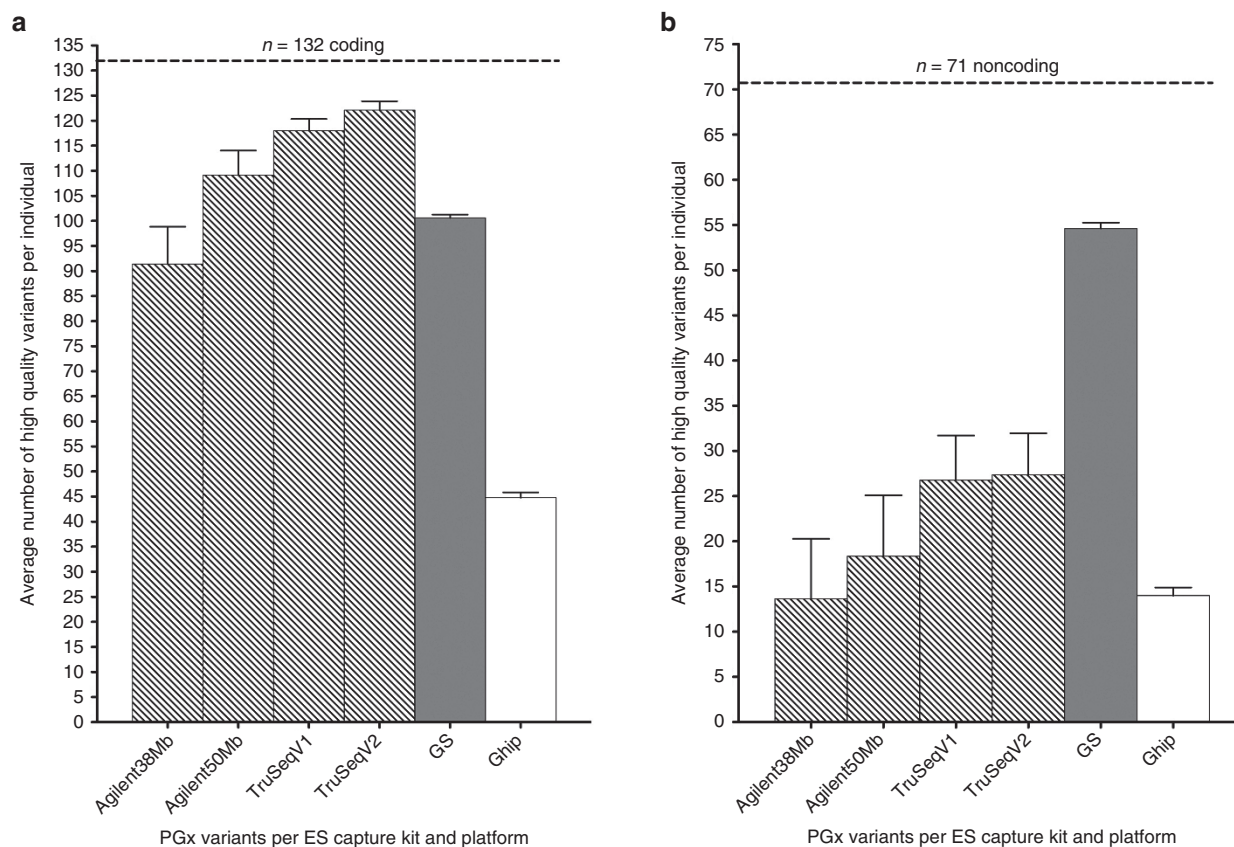
Figure 2 Comparison of variant coverage between exomes, genomes and chip. (**a**) Exome capture kits versus genome versus chip coverage of 132 coding pharmacogenetic variant positions. (**b**) Exome capture kits versus genome versus chip coverage of 71 noncoding pharmacogenetic variant positions. Total number of variant positions represented by the horizontal dashed line. Bar graphs shows the average number of high-quality variants per individual shown by four exome capture kits (Agilent 38Mb (*n* = 393), Agilent 50Mb (*n* = 318), Illumina TruSeqV1 (*n* = 147), Illumina TruSeqV2 (*n* = 115)) versus genome sequence (*n* = 5) versus chip data (*n* = 5). The tops of the bars indicate the average number of high-quality (GQ score ≥50) variant(s) detected per individual for exomes, genomes, and chip. The whiskers above the bars represent the standard error of the mean. See **Supplementary Table S4** online for mean, SEM, and N. 3′ UTR, 3 prime untranslated region; Chip, Affymetrix DMET Plus (Drug Metabolizing and Transporters array); CPIC, Clinical Pharmacogenetics Implementation Consortium; ES, exome sequence; GQ, genotype quality; GS, genome sequence; Mb, megabase; N, number of individuals tested per platform; PGx, pharmacogenetic; PharmGKB, Pharmacogenomics Knowledgebase; SEM, standard error of the mean.

Detection of CPIC and PharmGKB pharmacogenetic variants and rare loss-of-function variants in known pharmacogenes

ES identified 36 star (*) allele variants with CPIC recommendations for change in therapy, including individuals homozygous for CYP2C19 *2 (*n* = 18), TPMT *3B, TPMT *3C (*n* = 5),[12] SLCO1B1 *5 (*n* = 21),[13] and individuals heterozygous for DPYD*13 (*n* = 2) and rs67376798 (*n* = 6) (**Supplementary Table S5** online).[14] Twenty individuals with rare loss-of-function and eight with splice variants were identified in eight known pharmacogenes (**Supplementary Table S6** online).

Genotype concordance between exomes, genomes, and genotyping chip

The chip had 1,929 unique variant positions and identified 9,598 genotypes for the five samples tested.

Of 8,040 genotype calls made by chip–ES, 7,258 homozygous/hemizygous and 639 heterozygous calls were concordant and 143 (1.8%) calls were discordant. Of the chip–ES discordant calls, the chip called 89/143 heterozygous and 54/143

homozygous, 83/143 of the discordant calls had ES GQ <50, and 77/83 are noncoding. For discordant calls with ES GQ ≥50, 57/60 were concordant in ES–GS (12/57 heterozygous and 45/57 homozygous) (**Supplementary Tables S7 and S8** online).

Of 9,543 genotype calls made by chip–GS, 8,411 homozygous/hemizygous and 1,029 heterozygous calls were concordant and 103 (1.1%) were discordant. Of the discordant chip–GS calls, the chip called 19/103 heterozygous and 84/103 homozygous/hemizygous, 29/103 had GS GQ <50, and 74/103 had GS GQ ≥50. More than two-thirds (20/29) were discordant coding calls. Among the discordant calls with GS GQ ≥50, 52/74 were concordant between ES and GS (12/52 heterozygous, 40/52 homozygous) (**Supplementary Tables S7 and S8** online).

Of 8,013 genotype calls made by ES–GS, 7,267 homozygous/hemizygous and 649 heterozygous were concordant and 97 (1.2%) were discordant. Of the discordant ES–GS calls, the chip called 78/97 heterozygous and 19/97 homozygous, 80/97 had ES GQ <50, and 73/80 were noncoding. The majority (76/97) of the discordant ES–GS calls were concordant between chip and GS (75/76 heterozygous, GS GQ ≥50; 1/76 homozygous, GS

GQ <50). A few (17/97) of the discordant ES–GS calls had ES GQ ≥50 and 14/17 were concordant between chip and ES (2/14 heterozygous, 12/14 homozygous) (**Supplementary Tables S7 and S8** online).

Detection and validation of *CYP2D6* CNVs using eXome hidden Markov model

CYP2D6 CNVs were detected in 57/973 exomes (duplication *n* = 39, deletion *n* = 18) (**Supplementary Table S9** online). XHMM quality scores (QS) ranged from 38 to 99. Seven individuals with the highest XHMM QS of 99 (duplication *n* = 6, deletion *n* = 1) were selected for validation with real-time quantitative polymerase chain reaction (qPCR) and all samples were confirmed (**Supplementary Table S10** online). An additional 19 samples with XHMM QS ranging from 38 to 99 (duplication *n* = 17, deletion *n* = 2) were selected for a second round of validation with qPCR and all were confirmed (**Supplementary Table S10** online). Of the 26 samples tested, 11 samples showed agreement across all CNV regions, nine samples were inconclusive (XHMM does not make CNV calls in noncoding regions), and six samples (168397, 136439, 181872, 181608, 185076, 196659) showed breakpoint discrepancies between the XHMM predictions and the qPCR results. This was not a surprising finding because their XHMM Q_exact scores (confidence measure of the predicted CNV breakpoint) were low, ranging from 4 to 18 (data not shown). Nine samples (142307, 175100, 187383, 140601, 190031, 190871, 194883, 131340, 167715) with predicted whole-gene duplication showed a normal copy number with the 5′ probe (**Supplementary Table S10** online). This may be due to the *CYP2D6* duplication not extending into the 5′ region. High sequence identity (96.9%) of *CYP2D6* and the *CYP2D7P* pseudogene (NM_000106.5 and NR_002570.2, respectively) can result in *CYP2D6-2D7P* hybrid genes.[15] For these nine individuals, ES data analysis did not find paired-end reads mapping to *CYP2D6-CYP2D7P*. Our paired-end reads (89 bp) and inserts (180 bp) are short; therefore, the absence of detecting paired-end reads mapping to *CYP2D6-CYP2D7P* does not rule out the presence of fusion/hybrid genes.

DISCUSSION

Adoption of PGx-guided therapy has been limited by insufficient data to support clinical utility and cost-effectiveness, knowledge gaps in pharmacogenomics, and the inherent delay engendered by PGx testing. We propose leveraging existing MPS data by extracting PGx variants preemptively based on two premises. The first is that thousands of patients are currently undergoing clinical ES and GS, and these data comprise a valuable resource for pharmacogenomics. The second is that the extraction of PGx variants from ES/GS data is part of a larger effort to maximize the utility of ES/GS testing results. Studies have demonstrated how ES data can be used to extract variants for the secondary screening of susceptibility to cancer, malignant hyperthermia, cardiomyopathy, cardiac dysrhythmias, and aortic dissection.[16–19] We assessed the capability of MPS for preemptive PGx testing by comparing the coverage of 203 important PGx variants in 973 exomes with that of a widely used PGx chip.

ES and GS had several advantages over chip-based testing. The genome-wide coverage of ES and GS allowed coverage of more PharmGKB class 1A, 1B, and CPIC gene–drug level A variants than the genotyping chip and identified both known and yet to be discovered PGx variants in one test.

CYP2D6 is a good example for exploring the ability of ES to interrogate CNVs. XHMM detected complete and partial *CYP2D6* deletions and duplications, but the chip only detects deletions.

Limitations of this study include the fact that 399/973 of the ES sequences were generated with the Agilent38Mb capture kit, which accounted for the majority of the NC in the ES data, thus decreasing the coverage of some variant positions. The use of four capture kits provided us an opportunity to assess variance in capture-kit coverage.

Our results showed that high exome genotype concordance rates and higher coverage with the TruSeq capture kits (using GQ ≥50) are consistent with findings from recent studies evaluating exome capability for pharmacogenomics screening.[6,7] An updated array targeting PharmGKB level 1A and 1B and CPIC level A variants may be a more cost-efficient initial screen than exomes; however, panel testing and enhanced exome capture with additional targets in noncoding regions[20] will require periodic updating of the test platform and repeat testing of subjects for future discoveries. Although our results showed that exomes can be used to extract PGx variants, we are not advocating ordering an exome primarily for pharmacogenomics screening because our analyses did not answer the question of whether there is clinical utility and validity for using MPS for preemptive PGx screening for these variants.

We have demonstrated the utility of MPS data for the detection of single PGx variants and *CYP2D6* CNVs. Currently, no tools are available to extract and annotate PGx variants from MPS data. We conclude that tools should be developed to extract PGx variants from existing ES and GS data for research and potential future use.

BRIEF REPORT

REFERENCES

1. Weitzel KW, Elsey AR, Langaee TY, et al. Clinical pharmacogenetics implementation: approaches, successes, and challenges. *Am J Med Genet C Semin Med Genet* 2014;166C:56–67.
2. Janssens AC, Evans JP. Returning pharmacogenetic secondary findings from genome sequencing: let's not put the cart before the horse. *Genet Med* 2015;17:854–856.
3. Daly TM, Dumaual CM, Miao X, et al. Multiplex assay for comprehensive genotyping of genes involved in drug metabolism, excretion, and transport. *Clin Chem* 2007;53:1222–1230.
4. Sissung TM, English BC, Venzon D, Figg WD, Deeken JF. Clinical pharmacology and pharmacogenetics in a genomics era: the DMET platform. *Pharmacogenomics* 2010;11:89–103.
5. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014;370:2418–2425.
6. Chua EW, Cree SL, Ton KN, et al. Cross-comparison of exome analysis, next-generation sequencing of amplicons, and the iPLEX(®) ADME PGx panel for pharmacogenomic profiling. *Front Pharmacol* 2016;7:1.
7. Londin ER, Clark P, Sponziello M, Kricka LJ, Fortina P, Park JY. Performance of exome sequencing for pharmacogenomics. *Per Med* 2014;12:109–115.
8. Biesecker LG. Hypothesis-generating research and predictive medicine. *Genome Res* 2013;23:1051–1053.
9. Biesecker LG, Mullikin JC, Facio FM, et al.; NISC Comparative Sequencing Program. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 2009;19:1665–1674.
10. Fernandez CA, Smith C, Yang W, et al. Concordance of DMET plus genotyping results with those of orthogonal genotyping methods. *Clin Pharmacol Ther* 2012;92:360–365.
11. Crews KR, Gaedigk A, Dunnenberger HM, et al.; Clinical Pharmacogenetics Implementation Consortium. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. *Clin Pharmacol Ther* 2014;95:376–382.
12. Relling MV, Gardner EE, Sandborn WJ, et al.; Clinical Pharmacogenetics Implementation Consortium. Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin Pharmacol Ther* 2013;93:324–325.
13. Ramsey LB, Johnson SG, Caudle KE, et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1 and simvastatin-induced myopathy: 2014 update. *Clin Pharmacol Ther* 2014;96:423–428.
14. Caudle KE, Thorn CF, Klein TE, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for dihydropyrimidine dehydrogenase genotype and fluoropyrimidine dosing. *Clin Pharmacol Ther* 2013;94:640–645.
15. Black JL 3rd, Walker DL, O'Kane DJ, Harmandayan M. Frequency of undetected CYP2D6 hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metab Dispos* 2012;40:111–119.
16. Johnston JJ, Rubinstein WS, Facio FM, et al. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am J Hum Genet* 2012;91:97–108.
17. Gonsalves SG, Ng D, Johnston JJ, et al.; NISC Comparative Sequencing Program. Using exome data to identify malignant hyperthermia susceptibility mutations. *Anesthesiology* 2013;119:1043–1053.
18. Ng D, Johnston JJ, Teer JK, et al.; NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program. Interpreting secondary cardiac disease variants in an exome cohort. *Circ Cardiovasc Genet* 2013;6:337–346.
19. Dorschner MO, Amendola LM, Turner EH, et al.; National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet* 2013;93:631–640.
20. Patwardhan A, Harris J, Leng N, et al. Achieving high-sensitivity for clinical applications using augmented exome sequencing. *Genome Med* 2015;7:71.