Clinical Validation of Targeted Next-Generation Sequencing for Inherited Disorders

Sophia Yohe, MD; Adam Hauge, MS; Kari Bunjer, BS; Teresa Kemmer, MS; Matthew Bower, MS, CGC; Matthew Schomaker, BS; Getitia Onsongo, PhD; Jon Wilson, MD; Jesse Erdmann, MS; Yi Zhou, MD, PhD; Archana Deshpande, BS; Michael D. Spears, MD; Kenneth Beckman, PhD; Kevin A. T. Silverstein, PhD; Bharat Thyagarajan, MD, PhD

- **Context.**—Although next-generation sequencing (NGS) can revolutionize molecular diagnostics, several hurdles remain in the implementation of this technology in clinical laboratories.

- **Objectives.**—To validate and implement an NGS panel for genetic diagnosis of more than 100 inherited diseases, such as neurologic conditions, congenital heart disease, eye disorders, developmental disorders, nonmalignant diseases treated by hematopoietic cell transplantation, familial cancers, connective tissue disorders, metabolic disorders, disorders of sexual development, and cardiac disorders. The diagnostic gene panels ranged from 1 to 54 genes with most of panels containing 10 genes or fewer.

- **Design.**—We used a liquid hybridization–based, target-enrichment strategy to enrich 10 067 exons in 568 genes, followed by NGS with a HiSeq 2000 sequencing system (Illumina, San Diego, California).

- **Results.**—We successfully sequenced 97.6% (9825 of 10 067) of the targeted exons to obtain a minimum coverage of 20× at all bases. We demonstrated 100% concordance in detecting 19 pathogenic single-nucleotide variations and 11 pathogenic insertion-deletion mutations ranging in size from 1 to 18 base pairs across 18 samples that were previously characterized by Sanger sequencing. Using 4 pairs of blinded, duplicate samples, we demonstrated a high degree of concordance (>99%) among the blinded, duplicate pairs.

- **Conclusions.**—We have successfully demonstrated the feasibility of using the NGS platform to multiplex genetic tests for several rare diseases and the use of cloud computing for bioinformatics analysis as a relatively low-cost solution for implementing NGS in clinical laboratories.

*Arch Pathol Lab Med. 2015;139:204–210; doi: 10.5858/arpa.2013-0625-OA*
MATERIALS AND METHODS

Sample and Testing Selection

We used target enrichment, followed by NGS with the HiSeq 2000 (Illumina, San Diego, California) to develop clinical test panels for 568 genes (1.84 MB genomic region) that are associated with more than 100 distinct, inherited disorders. The inherited disorders included neurologic conditions, congenital hearing loss, congenital eye disorders, developmental disorders, nonmalignant pediatric diseases treated by hematopoietic cell transplantation, familial cancers, connective tissue disorders, metabolic disorders, nonmalignant hematologic conditions, disorders of sexual development, and cardiac disorders. Those 568 genes comprised 10,067 exons and can be ordered as single gene tests, multiple gene tests, or as a predefined gene panel. Eighteen patients in whom pathogenic mutations were previously identified by Sanger sequencing were evaluated for the 568 genes for validation of the sensitivity and specificity of the NGS assay. Four samples were run in duplicate to compare intertechnologist and intratechnologist reproducibility. A list of all 568 genes and the gene panels included in our assay is shown in supplementary table 1 (see supplemental digital material file at www.archivesofpathology.org in the February 2015 table of contents).

Sample Preparation

DNA was extracted from a 1 ml blood sample with the QiaGen QIAamp DNA Blood Midi extraction kits (Qiagen, Valencia, California) in the Molecular Diagnostics Laboratory (University of Minnesota Medical Center, Fairview, Minneapolis) and was then submitted to the University of Minnesota Genomics Center (Saint Paul campus) for library preparation and sequencing. Before submission to the University of Minnesota Genomics Center, each patient’s DNA sample was assigned a unique laboratory number. Patient’s DNA sample was assigned a unique laboratory number and added to 2-dimensional, bar-coded tubes that were spiked with synthetic, nonhuman (Arabidopsis) DNA oligo spikes to track DNA specimens throughout the assay process, to be able to detect any cross-contamination, and to otherwise de-identify the specimen.

Library Preparation and Sequencing

DNA was quantitated with Quant-IT PicoGreen reagents (catalogue no. P7581, Life Technologies, Grand Island, New York) to ensure a minimum starting quantity of 3000 ng of genomic DNA. The DNA was then sheared with a Covaris machine (Covaris, Woburn, Massachusetts), according to the Agilent Sure-Select (Agilent Technologies Inc., Santa Clara, California) protocol, to achieve fragments ranging from approximately 150 to 200 base pairs (bp). The sheared DNA was used for the subsequent library preparation, which was performed according to the Agilent SureSelect protocol. The library creation involved repairing blunt ends, adenylation of the 3’ ends, and ligation of adapters. The adapter-ligated library was amplified using 6 cycles of polymerase chain reaction. Quality control was performed for sizing and concentration, and 1000 ng of amplified library was used for enrichment.

Enrichment was performed using a custom SureSelect library (Agilent Technologies) for the coding regions and adjacent intronic areas for the 568 genes. SureSelect hybridization and capture was performed according to the manufacturer’s instructions. The efficiency of the capture is determined by quantitative polymerase chain reaction with 4 primer-probe sets—2 of which are designed for targeted regions (target products), and 2 of which are designed for nontargeted regions (control product). The increased amplification of the target products in the quantitative polymerase chain reaction, when compared with control products (which show delayed or no amplification) on the postcapture library, indicates successful sequence capture.

Following capture, the samples were index tagged during postcapture amplification for 16 cycles. That amplified, postcapture library was rechecked for sizing and concentration. Samples with different index tags were pooled, and the sample pools were divided across 2 or 3 lanes (depending on the number of samples in the pools) to minimize lane-to-lane variation. Sequencing was performed on the HiSeq 2000, obtaining paired-end sequencing reads 30 bp long.

Sequencing Analysis and Variant Detection

Data Processing—The raw image files were analyzed and converted to base calls by real-time analysis (using the default settings recommended by the manufacturer) on the HiSeq 2000. The real-time analysis outputs base call files (*.bcl), which were converted to FASTQ files with consensus assessment of sequence and variation (CASAVA pipeline, version 1.8, Illumina) on Minnesota Supercomputing Institute servers, and CASAVA also demultiplexed the data to obtain FASTQ files for individual samples. Initial-sample quality control involved evaluating the synthetic, spiked oligonucleotides from each sample to ensure that there was no crossover contamination that would interfere with variant calling and data interpretation. Sequences that passed Illumina’s initial quality filter were uploaded to a locked-down instance of Galaxy (open source, Web-based platform for data intensive biomedical research), developed at Minnesota Supercomputing Institute, which is run on an Amazon (Seattle, Washington) cloud server and analyzed at the Molecular Diagnostics Laboratory. An in-house script was used to verify that forward and reverse reads were synchronized. FastQC (Sibraham Institute, Cambridge, Cambridgeshire, England) reports were generated and manually inspected to ensure cycle uniformity and run quality.

Processing of all data followed the Genome Analysis Tool Kit (GATK) (version 1.6, Broad Institute, Cambridge, Massachusetts) best practices pipeline. Specifically, Burrows-Wheeler Aligner’s software was used to map reads in paired-end mode to the University of California, Santa Cruz, canonical-ordering build of the human reference genome (hg19 version, Genome Reference Consortium GRCh37). To speed up the analysis and reduce the computational costs on the Amazon cloud, we created a reduced genome, which contained a subset of the human genome, and mapped the reads against that reduced genome. The reduced genome consisted of the regions targeted for capture and all homologous regions that matched at least one of the 120-bp capture baits with a minimum of 75% identity for at least 48-bp in length. We merged all overlapping, targeted and homologous regions and expanded them by 1000 bp, both upstream and downstream, to ensure paired-end sequences with expected fragment lengths would still map to the final, reduced genome. The BAM files (*.bam) binary version of sequence alignment/map or SAM (*.sam) files) that referenced the reduced genome coordinate system were mapped back to the reference hg19 with and automap scripted behavior with subsequent steps in the analysis pipeline. Picard (SourceForge, Mountain View, California)’s software was used to sort mapped reads into coordinate order and to ensure all mate-pair information was properly updated among mate pairs. If reads were sequenced in multiple lanes, BAM files containing mapped reads for the different lanes were merged using SAMtools (SourceForge) while preserving separate read-group information for each lane. Picard and SAMtools were used to remove duplicate reads and low-quality reads, respectively. Low-quality reads were defined as reads that were ambiguously mapped (mapping quality score [MAQ, SourceForge] = 0), that were unpaired or unmapped, that failed a platform/vendor quality check, or that had polymerase chain reaction or optical duplicates. The GATK was then used for recalibration of the base quality score and for insertion/deletion (INDEL) realignment, with Picard used again for duplicate removal, to obtain analysis-ready BAM files.

Coverage Analysis

Before performing single nucleotide polymorphism (SNP) and INDEL discovery and before genotyping, we analyzed coverage in the targeted exons to assess target enrichment. Galaxy workflows were used to generate 2 coverage analyses: (1) a coverage plot with coverage on the x-axis and percentage of bases covered on the y-axis, and (2) an 8-column table containing exon coverage. Supplemen-
A. The distribution of discrepant calls on blinded, duplicate reproducibility testing. Abbreviations: INDEL, insertion/deletion; SNP, single nucleotide polymorphism.

B. Example of a discrepancy in a G-C–rich region. The G-C–rich region is followed by XXZXX, which in this case is “CACC.” There are many low-level miscalled C nucleotides with low-Phred (quality) scores (circles) in the area, and A is miscalled as C. Such regions are often characterized by calls that deviate significantly from the expected 50:50 ratio.


Clinical Validation

Sensitivity and specificity were calculated and compared with results obtained by standard testing by Sanger sequencing. Reproducibility was determined by comparison of the blinded, duplicate, intertechnologist and intratechnologist samples. The discrepant variant calls were divided by the total number of variant calls averaged for the duplicate samples to give a concordance value for each sample. The concordance values were then averaged for the 4 samples.

RESULTS

On average, for the 18 samples, we detected 1102 variants across the 568 genes (range, 1035–1185). Based on available literature, we determined the minimum coverage required to accurately genotype a base at any position was 20×. We were successfully able to sequence 97.6% (n = 9825) of the targeted 10 067 exons to a minimum coverage depth of 20× at every base-pair position of interest (supplementary figure 1). The remaining 2.4% of exons (n = 242) (supplementary table 1) could not be sequenced with adequate coverage using the NGS platform, and we have developed Sanger sequencing assays for those exons. In a limited, 18-sample validation cohort with 30 known mutations, we showed 100% analytic sensitivity in detecting the 19 pathogenic single-nucleotide variations and 11 pathogenic INDELs, ranging in size from 1 to 18 bp across 18 samples. We showed 100% analytic specificity for 10 exons that were within reference range by Sanger sequencing (supplementary table 1).

Within the 4 blinded duplicates we used for reproducibility testing, more than 900 variants were identified across all exons that had a 20× minimum coverage. Mean (SD) intratechnologist reproducibility (technologist A) on the 4 duplicate samples was 99.22% (0.25%), and intertechnologist reproducibility (technologist A versus technologist B) was 99.28% (0.05%). Discordant variant calls in the coding exons or in the adjoining intron-exon boundaries (±20 bp) fell into 5 categories that include (1) G-C rich areas (43%), (2) homologous sequences (26%), (3) mismapping of INDELs (16%), (4) repetitive sequences (12%), and (5) trinucleotide SNPs (3%) (Figure).

Subsequently, 91 clinical samples, with 689 genes, had clinical testing completed (average, 7.6 genes/sample; range, 1–29 genes). The most common gene panels ordered were related to neurogenetics and hearing loss. Visual inspection

Analysis of Variants

All identified variants were processed using ANNOVAR. All additional variants were also used in silico software, such as SIFT (scale-invariant feature transform), PolyPhen2, and Provean (protein variation effect analyzer), to evaluate the functional importance of novel variants. We also evaluated data from the Exome Sequencing Project (ESP; Heart, Lung, and Blood Institute, Bethesda, Maryland), the 1000 Genomes Project, and the dbSNP (National Center for Biotechnology Information [NCBI], Bethesda, Maryland) database to evaluate the prevalence of the identified variants in human populations. In addition to those efforts, we also evaluated locus/disease-specific databases, performed PubMed searches, and searched the Online Mendelian Inheritance in Man (OMIM, NCBI) database to evaluate the functional significance of the identified variants.

All potentially pathogenic variant calls were reviewed manually, looking at quality scores and visual inspection of the data in the Integrative Genome Viewer (IGV; Broad Institute), and were then confirmed by Sanger sequencing. Additionally, because large deletions may cause hemizygous mutations leading to an apparent implausible haplotype, all implausible haplotypes were manually reviewed in the Integrated Genome Viewer.

Variant Discovery, Genotyping, and Annotation

The Unified Genotyper (Broad Institute) was used for SNP and INDEL discovery and genotyping. The SNP variant recalibration and INDEL filtration were both done following guidelines from GATK’s best practices recommendations. For each patient sample, a browser extensible data (BED) file containing the requested genes was uploaded to the Amazon cloud. That information was used to generate a variant call file (VCF) containing only the variants in genes requested by the clinician. Variants were annotated using ANNOVAR, a software tool that uses up-to-date information to functionally annotate genomes.


c, Example of a discrepancy at a polymorphic repetitive sequence. Problems occur at mononucleotide, dinucleotide, and trinucleotide repeat sequences. The example shows a polymorphic CAG trinucleotide repeat, where the reference sequence has 10 repeats and the patient is heterozygous for 9 and 10 repeats. The shorter 9-repeat allele is only identified in reads that completely span the repetitive sequence. Most reads in the targeted 10 067 exons to a minimum coverage depth of 20× at every base-pair position of interest (supplementary figure 1).

d, Example of mismapping of an INDEL. Accurate mapping of INDELs larger than 7 nucleotides is a known limitation of the GATK genotyping pipeline. Nine variants at 3 unique sites were discrepant because of inaccurate mapping of large INDEL variations. The figure shows a homozygous 18-bp deletion mutation where the C, all at the end of several reads, could reasonably be mapped without the 18-bp deletion. The presence of these mismapped C calls was sufficient for a heterozygous call in from GATK, rather than the true homozygous genotype.

f, Example of a discrepancy because of a known homologous region/pseudogene. There is a highly homologous sequence, which differed from the target gene FLNC at the highlighted position. The FLNC gene has a C at that position, whereas the homologous sequence has an A. In this example, the A reads accounted for 18% of the reads at that position. However, A reads accounted for a higher percentage of reads in some duplicates and were sufficient in those duplicates for a heterozygous G/A call. Sanger sequencing verified a normal G/G genotype at that position.

of variant calls in the IGV was performed looking for the issues listed above (G-C rich areas, pseudogene or INDEL mismapping, and hemizygous mutations). All variant calls identified with the bioinformatics pipeline were visually inspected in the IGV and have also been confirmed by Sanger sequencing.

**COMMENT**

We described the successful validation and implementation of NGS for diagnosis of a variety of inherited disorders. Although several laboratories offer NGS for inherited disorders, the unique feature of our process includes multiplexing of many targeted disease panels, collaboration with a physically separate genomics center, and the use of cloud computing. Those strategies provide relatively low-cost solutions for implementation of NGS-based testing for clinical laboratories that do not have access to the capital-intensive resources that are needed during the initial implementation of NGS-based clinical testing. These features were necessary for us because the University of Minnesota does not have a dedicated clinical genomics center or the upfront capital needed for development of a separate server for data processing and storage and the necessary test volume to justify developing test panels for individual disorders.

For validation purposes, DNA sequencing for inherited disorders is equivalent to multiple qualitative tests in that it looks to see whether a mutation is present or absent at each base pair position. Our validation included measures of analytic sensitivity, analytic specificity, and reproducibility. **Analytic sensitivity**, defined as the ability to detect a positive result compared with the reference method, reflects the accurate identification of mutations. **Analytic specificity**, defined as the ability to obtain negative results in accordance with the reference method, reflects the confidence that a mutation is absent. Our analytic sensitivity and specificity were 100%, compared with Sanger sequencing, although that was based on a limited set of 18 samples. Our reproducibility was greater than 99%, both by the same technologist (intra technologist) and by a second technologist (inter technologist). The discrepant variant calls identified by analysis of duplicate samples indicated poor reliability for identifying variants in certain genomic regions. Those genomic regions need to be identified a priori, and alternative testing strategies to identify variants in those regions should be implemented. In addition, the lower confidence scores of variants in those regions can be used as an indicator to identify variants that would need to be confirmed by Sanger sequencing. We categorized the discrepancies found on reproducibility testing into one of 5 causes: GC-rich areas, homologous regions, repetitive sequences, INDEL mapping, and triallelic SNPs (Figure, A).

The sequence in GC-rich areas that resulted in discordance was usually an A or T surrounded by several G's or C's (eg, XXZZX, XZXX, or XXZX, where X is either G or C at all positions, and Z is an A or T). The result is an X call at the Z position. The GC-rich area leads to frequent dyssynchrony of strands, and any strand that is leading by 1 or 2 bp gave an X read at the Z position. The GC-rich areas showed many poor-quality base pair reads that differed from the reference within the area, which can be identified by viewing the area in IGV (Figure, B). Homologous regions, including pseudogenes, are known issues in sequencing, and repetitive sequences are known to cause problems in mapping short-sequence reads obtained from NGS (Figure, C and D). The final 2 issues (mismatching of INDELs and triallelic SNPs) are known limitations of the GATK (version 1.6) (Figure, E and F). Insertion/deletion mismapping occurs with GATK; other tools, such as PINDEL and the GATK 2, perform better with INDEL mapping. We plan to incorporate one of these other tools into our bioinformatics pipeline in the future. The GATK (version 1.6) also lacks the capacity to make a variant 1/variant 2 call at positions with 3 or even 4 possible common variants. These later issues were identified by viewing the sequence data in the IGV and by determining a priori the homologous areas, repetitive sequences, and triallelic SNPs in the ordered gene panels. We use alternative methodologies to assess repeat regions and do not report repeat sizes using NGS data. We perform Sanger sequencing confirmation of all pathogenic variants identified by NGS and have confirmed 100% of the pathogenic variants in regions that do not fall into any of the 5 categories. These data strongly suggest that confirmation of variants identified by NGS can be restricted to regions that are known to be problematic for NGS.

Despite the promise of NGS to deliver large amounts of sequence data, the technology remains cost effective clinically only if many samples are evaluated simultaneously. Although we can get more than 200 million reads in each sequencing run, the costs of sequencing are fixed, and NGS is most cost effective if several samples are simultaneously evaluated. Because individual inherited genetic diseases are extremely rare, clinical laboratories, outside of large reference laboratories or referral centers, are likely to have very low test volumes for individual genetic tests where a single test or gene panel might only be ordered a few times a year. However, as demonstrated at the University of Minnesota, when several rare disease gene panels are aggregated together an acceptable annual test volume can be achieved. Although most genomic regions (97.6%; 9825 of 10 067) in our panel meet the requirement of 20× minimum coverage at every base position, a small percentage of genes (2.4%; 242 of 10 067) have regions with less than 20× coverage (low coverage). Those areas can often be predicted because coverage patterns are consistent, and we have implemented Sanger sequencing for the areas that have low coverage. To achieve that level of coverage with our 568 gene panel, we can multiplex up to 32 samples on one HiSeq 2000 run. A major limitation of this approach is the long turnaround time because of the need to batch dozens of samples together, the long run times of the HiSeq 2000, the Sanger confirmation of pathogenic variants, and the Sanger sequencing of regions not adequately covered by NGS. Our current turnaround times are approximately 12 weeks. We have taken several steps to reduce turnaround times that include optimizing our laboratory workflow to Sanger sequence regions known to have low coverage before sending the sample for NGS, and we plan to change our laboratory protocol to confirm pathogenic variants only in regions that have been identified to be problematic for NGS. For example, during assay development for the current project, the costs of the SureSelect reagents was around $400/sample when reagents were ordered for target enrichment of 100 samples, but that costs decreases to approximately $150/sample when sufficient reagents were ordered for 500 samples. We are currently evaluating strategies that will reduce turnaround times while maintaining the advantages of multiplexing by using alternative...
sequencing strategies, such as the HiSeq 2500 that provides substantially more sequencing reads, as compared with the desktop sequencers, within 1 to 2 days. Our preliminary analysis suggests that the cost will be comparable, and turnaround time will be greatly improved. All our NGS tests are coded as 81479, and preauthorization is obtained in some cases. Exact reimbursement rates for our NGS test are difficult to determine because we receive a lump reimbursement for all services on a particular day. However, based on analysis of a subset of patients with a few of the other services performed the same day, private insurers reimburse 53% (range, 27%–90%) of the total NGS and non-NGS services charges. For those samples, the NGS charges accounted for an average of 80% (range, 42%–98%) of the total charges, which limited the effect of the non-NGS charges.

An alternate approach could be the use of small, targeted NGS panels, followed by sequencing on desktop sequencers that have limited sequencing capacity. Although these options can reduce the turnaround times, the sequencing costs on the desktop sequencers will be higher. In our institution, a smaller, targeted panel would have lower annual test volume and testing for the same number of diseases with smaller, targeted panels would require multiple different sets of sequence capture, which would increase the cost.

Based on available literature, we had determined that the minimum coverage required to accurately call a base at any position was 20× [22,25]. However, the liquid-hybridization approach to target enrichment results in uneven enrichment of various targets. Hence, it is necessary to sequence to high depths to get the adequate coverage of 20× at the maximum number of nucleotides within a given gene panel. Alternate approaches, such as amplicon-based strategies, for enrichment and to optimize loading densities based on coverage characteristics of the ordered gene panel can reduce the coverage depth and still allow us to maintain a minimum of 20× coverage at individual nucleotide positions.

Although best practice guidelines have been published for bioinformatics analysis of NGS data, the data analysis steps that are required to convert raw NGS sequence data into a list of genetic variants typically requires access to high-performance computers and the ability to store large amounts of data. We have implemented the GATK (version 1.6) analysis pipeline in the Galaxy workflow and have created a virtual instance of the Galaxy workflow that uses cloud computing to perform all the analysis required to reproducibly and accurately identify genetic variants by NGS data. Limited availability of servers for processing and storage of clinical data and sequencers validated for clinical use were major limitations in clinical implementation of NGS. The use of cloud computing resources obviates the need to make capital investments in purchase of powerful computers that may be required to process NGS data quickly and reliably. Because of concerns about data privacy, we remove patient identifiers from the data (using only the unique laboratory-generated number and Arabidopsis data to track), and we do not store any genetic information on the cloud. However, with the advent of the Health Insurance Portability and Accountability Act—certified cloud computing services, the large amounts of data could be created on the cloud, which would reduce the need to purchase a large amount of storage.

Reporting of incidental genetic findings is a common issue encountered by laboratories using NGS-based tests, and we minimize that problem by analyzing the data only for gene panels that have been specifically ordered by the clinicians. We have created a bioinformatics workflow that reports variants only in the genes that have been ordered, and so, we do not have any information on other genetic variants that may be incidentally detected. However, given the pleiotropic effects of several genes, for some of our larger gene panels, we have identified variants within the ordered gene panel that have important medical consequences for reasons unrelated to the clinical condition being evaluated. In those cases, we have worked closely with the clinicians and genetic counselors to report those results back to the patients with appropriate counseling and follow-up.

In summary, we have demonstrated a low-cost alternative for implementing clinical NGS testing for inherited disorders. Future studies will need to address the possibility of incorporating copy number variation analyses into bioinformatics analyses to simultaneously provide both copy number and mutation analysis using NGS data and to consider alternative workflows and sequencing platforms to reduce turnaround times while reducing the costs of sequencing tests.

This work was supported by the Institute of Translational Neuroscience at the University of Minnesota. We thank Anne-Francoise Lamblin, PhD, the Research Informatics Support Systems program director and project lead for Galaxy Bioinformatics workbench; Shawn Houston, BS, manager for Infrastructure Operations and the User Support group; and Dan Debertin, BS, Galaxy System Administrator, all at the Minnesota Supercomputing Institute (MSI). We also thank MSI at large for its support on this project, without which, this project would not have been possible.

References


---

Prepare Now for the CAP ’15 Abstract Program

Don’t Wait! Plan now to submit abstracts and case studies for the College of American Pathologists (CAP) 2015 meeting, which will be held October 4th Nashville, Tenn. Submissions for the CAP ’15 Abstract Program will be accepted from:

Monday, February 9, 2015 through 6 p.m. Central time
Friday, April 10, 2015

Accepted submissions will appear online on the Archives of Pathology & Laboratory Medicine Web site as a supplement to the October 2015 issue. Visit the CAP ’15 Web site at http://www.cap.org/cap15 for additional abstract program information as it becomes available.