Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for Resequencing of Hypertrophic Cardiomyopathy-Associated Genes

Shale Dames,[1,] Jacob Durtschi,[1] Katherine Geiersbach,[1,2] Jack Stephens,[1] and Karl V. Voelkerding[1,2]*

[1]ARUP Laboratories for Clinical and Experimental Pathology, Salt Lake City, Utah 84108, USA; and [2]University of Utah, Department of Pathology, Salt Lake City, Utah 84112, USA

Next-generation sequencing (NGS) is widely used in biomedical research, but its adoption has been limited in molecular diagnostics. One application of NGS is the targeted resequencing of genes whose mutations lead to an overlapping clinical phenotype. This study evaluated the comparative performance of the Illumina Genome Analyzer and Roche 454 GS FLX for the resequencing of 16 genes associated with hypertrophic cardiomyopathy (HCM). Using a single human genomic DNA sample enriched by long-range PCR (LR-PCR), 40 GS FLX and 31 Genome Analyzer exon variants were identified using ≥30-fold read-coverage and ≥20% read-percentage selection criteria. Twenty-seven platform concordant variants were Sanger-confirmed. The discordant variants segregated into two categories: variants with read coverages ≥30 on one platform but <30-fold on the alternate platform and variants with read percentages ≥20% on one platform but <20% on the alternate platform. All variants with <30-fold coverage were Sanger-confirmed, suggesting that the coverage criterion of ≥30-fold is too stringent for variant discovery. The variants with <20% read percentage were identified as reference sequence based on Sanger sequencing. These variants were found in homopolymer tracts and short-read misalignments, specifically in genes with high identity. The results of the current study demonstrate the feasibility of combining LR-PCR with the Genome Analyzer or GS FLX for targeted resequencing of HCM-associated genes.

Key Words: sequence analysis, DNA, next-generation sequencing

INTRODUCTION

The last 5 years have witnessed the emergence of next-generation sequencing (NGS) technologies that have provided a powerful, new approach to complex genetic studies. NGS has enabled sequencing of entire transcriptomes (RNA-Seq), characterization of regulatory and chromatin protein DNA-binding sites [chromatin immunoprecipitation (ChIP)-Seq], and whole genome sequencing, among others.[1–7] An emerging application of NGS is the targeted resequencing of genomic subregions. Prior to sequencing, these regions of interest are enriched by one of several methods, including long-range PCR (LR-PCR) or fragment-capture using solid surface arrays or in-solution oligonucleotide capture.[5,8–10]

One candidate for targeted enrichment is primary hypertrophic cardiomyopathy (HCM), which has an incidence of one in 500 and displays autosomal-dominant inheritance with variable penetrance, manifesting primarily in adulthood.[11,12] HCM is characterized by ventricular and septal-wall enlargement, with development of angina, arrhythmias, and in its severest form, sudden death.[13,14] At least 27 genes with over 450 mutations have been implicated in HCM, and those with the highest mutation frequency encode core sarcomere proteins, including myosin heavy and light chains, actin, troponins T and I, and tropomyosin.[15–18] Using HCM as a diagnostic model, we are developing a NGS-based approach for multi-gene resequencing in the clinical laboratory. In the current proof-of-concept study, we sequenced 16 genes implicated in HCM using a previously uncharacterized human genomic DNA sample. LR-PCR was used for gene enrichment, followed by comparative sequencing on the Illumina Genome Analyzer and Roche 454 GS FLX platforms.

MATERIALS AND METHODS

Sample

The human genomic DNA sample was residual and de-identified in accordance with University of Utah Institutional Review Board (Salt Lake City, UT, USA), Protocol Number 7275.

HCM Gene Enrichment by LR-PCR

The 16 genes analyzed for this study are shown in Table 1. A total of 344,082 bp was amplified using 67 primer pairs

*Address correspondence to: Shale Dames, ARUP Laboratories for Clinical and Experimental Pathology, Salt Lake City, UT 84108, USA (Phone: (801) 583-2787, ext. 3254; E-mail: shale.dames@aruplab.com).

TABLE 1

HCM-Associated Genes Sequenced by Illumina Genome Analyzer and Roche 454 GS FLX

Gene symbol	Gene	RefSeq	Bases sequenced
ACTC1	actin, α, cardiac muscle 1	NG_007553.1	11,525
CAV3	caveolin 3	NG_008797.1	20,705
CSRP3	cysteine and glycine-rich protein 3	NM_003476	23,617
MYBPC3	myosin-binding protein C, cardiac	NG_007667.1	30,122
MYH6	myosin, heavy-chain 6, cardiac muscle, α	NM_002471	34,302
MYH7	myosin, heavy-chain 7, cardiac muscle, β	NG_007884.1	30,273
MYL2	myosin, light-chain 2, regulatory, cardiac, slow	NG_007554.1	16,168
MYL3	myosin, light-chain 3, alkali, ventricular, skeletal, slow	NG_007555.1	7517
PLN	phospholamban	NG_009082.1	19,692
PRKAG2	protein kinase, adenosine monophosphate-activated, noncatalytic, γ-2	NG_007486.1	48,314
TCAP	titin-cap (telethonin)	NG_008892.1	8874
TNNC1	troponin C type 1 (slow)	NG_008963.1	10,085
TNNI3	troponin I type 3 (cardiac)	NG_007866.1	13,217
TNNT2	troponin T type 2 (cardiac)	NG_007556.1	26,731
TPM1	tropomyosin 1 (α)	NG_007557.1	37,844
TTN	titin	NG_011618.1	5096

RefSeq, Reference sequence gene; annotation from human genome build 36.3. Bases sequenced, Total number of unique bases sequenced for a given gene.

(Supplementary Table 1). Primers were designed to amplify up to 5 Kb upstream of the first exon and 1 Kb downstream from the stop codon for each gene, except for *PRKAG2* and *TTN*. For *PRKAG2*, only exons and flanking intronic sequences were amplified, and for *TTN*, only the exonic region that contained the p.R740L variant was amplified.[19]

Oligonucleotides were designed using Vector NTI (Invitrogen, Carlsbad, CA, USA; Version 7) and ordered with standard desalting (Integrated DNA Technologies, Coralville, IA, USA). Amplifications were performed in a GeneAmp® 9700 thermocycler (Applied Biosystems, Foster City, CA, USA). Each reaction consisted of 1× LA PCR Buffer II (proprietary, with 2.5 mM $MgCl_2$, TaKaRa Bio Inc., Madison, WI, USA), 0.4 μM forward primer, 0.4 μM reverse primer, 1.6 mM dNTPs (TaKaRa Bio Inc.), 1 U TaKaRa Bio Inc. *LA Taq*™ Hot Start DNA polymerase, and 50 ng (~23 zmoles) genomic DNA (50 μL total volume). PCR amplifications were performed as follows: $94°C^{(1:00)} + [98°C^{(0:10)}+65°C^{(0:10)}+68°C^{(7:00)}] \times 40$ cycles + $72°C^{(10:00)}$. Amplicons were electrophoresed in 1% agarose Tris-boric acid-EDTA (TBE) gels prior to gel purification using a QIAquick gel extraction kit (Qiagen, Valencia, CA, USA). Gel-purified amplicons were resuspended in nuclease-free water (Quality Biological, Gaithersburg, MD, USA), and DNA concentration was determined by averaging three OD_{260} measurements per amplicon (NanoDrop™ 1000, NanoDrop, Wilmington, DE, USA). Each amplicon (70 fmol) was added to a 1.5-mL microfuge tube, and the volume was reduced to 50 μL at room temperature using a Speed Vac DNA 120 (Savant Instruments, Holbrook, NY, USA). Seventy microliters 2× 20 mM Tris-HCl, 2 mM EDTA, pH 7.5, and 20 μL nuclease-free water were added to the amplicon solution to obtain a final concentration of 0.5 μM pooled amplicon.

Illumina Genome Analyzer Library Preparation and Sequencing

A single-end library was made following the manufacturer's protocol with reagents supplied in the Illumina DNA sample kit (FC-102-1001). Briefly, 20 μl of the 0.5-μM pooled amplicon solution (representing ~2.0 μg total input DNA) was nebulized at 32 pounds per square inch (PSI) for 6 min. The nebulized DNA was end-repaired using Klenow and T4 DNA polymerases, phosphorylated with T4 polynucleotide kinase, and adenylated using Klenow exo-DNA polymerase, and oligonucleotide adapters were added using DNA ligase. Ligated products were visualized in a 2% agarose TBE gel, and a 200- to 250-bp size range was excised and purified using a Qiagen gel extraction kit. The size-selected, adapter-modified DNA fragments were amplified using adapter-specific primers 1.1 and 2.1 with Phusion DNA polymerase using the following protocol: $98°C^{(0:30)} + [98°C^{(0:10)}+65°C^{(0:30)}+72°C^{(0:30)}] \times 18$ cycles + $72°C^{(5:00)}$. The library was purified prior to quantification using an Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). A single-flow cell lane was sequenced in on the Illumina Genome Analyzer at the University of Utah Huntsman Cancer Institute Core Sequencing Facility.

Roche 454 Genome GS FLX Library Preparation and Sequencing

A GS FLX Amplicon DNA Library was made following the manufacturer's protocol (USM-00032.A 12/07). Briefly, 4.0 μg of the 0.5-μM pooled amplicon solution was nebulized at 45 PSI for 1 min and purified using a MiniElute PCR purification kit (Qiagen). Small DNA fragments were removed using an AMPure PCR purification system (Agencourt Bioscience, Beverly, MA, USA) and analyzed in an Agilent 2100 bioanalyzer. The nebulized DNA was subsequently end-repaired and phosphorylated using T4 DNA polymerase and T4 polynucleotide kinase, and oligonucleotide adapters were ligated to the DNA fragments with DNA ligase. Adapter-modified fragments were diluted, annealed to capture beads, and clonally amplified by emulsion PCR. After emulsion PCR, beads with clonal amplicons were enriched and deposited in an entire picotiter plate flow cell and sequenced on the GS FLX platform at the University of Colorado Health Sciences Center (Denver, CO, USA).

Sanger Sequencing

For variant confirmation, new primers were designed and amplicons generated for the regions of interest and sequenced using 1× BigDye Ready Reaction Mix (Applied Biosystems) in an Applied Biosystems 3130 capillary sequencer. Four amplicons required cloningfor sequence verification (Tables 2–4). In these cases, amplicons were cloned into pCR®-II TOPO® (Invitrogen) following the

TABLE 2

Genome Analyzer and GS FLX Concordant Exon Variants in HCM-Associated Genes

			Genome analyzer		GS FLX		
Gene	rs number	Variant position	Read %	Coverage	Read %	Coverage	Variant effect
CAV3	rs1008642	g.8715660C>T	43.2	498	34.4	32	syn
MYH6	rs2277474	g.4874362C>T	94.2	497	100	372	syn
MYH6	rs2277473	g.4874346G>T	43.3	499	46.7	366	syn
MYH6	rs451794	g.4858071C>T	53.0	411	51.4	370	syn
MYH6	rs382872	g.4855159G>A	45.9	499	51.7	315	syn
MYH6	rs8004990	g.4853994G>A	25.9	499	50.8	313	syn
MYH7	rs2069540	g.4902592G>A	49.0	478	33.1	266	syn
MYH7	rs2069542	g.4900633G>A	20.6	499	47.6	145	syn
MYH7	rs735712	g.4898899G>A	41.1	499	45.0	109	syn
MYH7	rs735711	g.4898866C>T	59.0	488	54.6	108	syn
MYH7	**rs7157716**	**g.4892727A>G**	**91.9**	**479**	**100**	**222**	**syn**
MYH7	rs45523835	g.4887351G>A	45.6	331	42.7	75	syn
TCAP	rs1053651	g.1546606A>C	42.1	499	51.0	192	syn
TNNT2	rs3729547	g.51824735G>A	50.7	491	50.7	107	syn
TPM1	rs11558747	g.34142190T>C	42.0	88	47.4	72	syn
MHY6	rs365990	g.4861650A>G	41.6	466	55.6	99	p.A1101V
MHY6	rs28730771	g.4859449C>T	22.2	496	38.6	298	p.A1130T
ACTC1	rs604689	g.5871080T>C	53.5	499	47.8	92	3′ UTR
ACTC1	rs1370154	g.5872782C>T	42.9	261	60.7	84	3′ UTR
ACTC1	rs34323254	g.5872252delA	43.0	498	40.8	125	3′ UTR
ACTC1	**rs589759**	**g.5872131C>T**	**42.5**	**499**	**35.9**	**103**	**3′ UTR**
ACTC1	rs533021	g.5871488T>C	51.7	499	50.0	84	3′ UTR
PLN	rs12198461	g.23050069T>G	35.9	370	48.3	60	3′ UTR
PRKAG2	rs8961	g.11830191T>C	53.7	216	50.7	454	3′ UTR
PRKAG2	rs1051956	g.11829781C>A	33.3	84	51.0	429	3′ UTR
PRKAG2	rs7429	g.11829242C>T	61.4	44	42.3	336	3′ UTR
PRKAG2	**rs11329945**	**g.11829291delT**	**65.9**	**41**	**85.9**	**320**	**3′ UTR**

Twenty-seven platform-concordant exon variants were identified that met selection criteria of >30-fold and read coverage >20% read percentage. All 27 variants were Sanger-confirmed. syn, Synonymous amino acid change. Variants shown in bold required cloning for confirmation.

TABLE 3

Genome Analyzer and GS FLX Discordant Exon Variants in HCM-Associated Genes

			Genome analyzer		GS FLX		
Gene	rs number	Position	Read %	Coverage	Read %	Coverage	Variant effect
PRKAG2	rs11427106	g.11829277insC	66.6	12	44.0	298	3′ UTR
PRKAG2	rs7429	g.11829242C>T	35.3	17	45.4	302	3′ UTR
MYH6	rs178640	g.4855408A>G	80.0	10	100	206	syn
MYH7	rs17794387	g.4881790C>T	52.0	492	69.0	29	3′ UTR
PRKAG2	rs66628686	g.12149747C>T	48.6	231	55.6	9	5′ UTR

Five variants with >30-fold coverage on one platform but with <30-fold coverage on the alternate platform were identified. All five variants were confirmed by Sanger sequencing.

manufacturer's protocol, and plasmids were Sanger-sequenced using primer M13 forward (−20).

Bioinformatic Analysis

Data from Illumina Genome Analyzer and Roche GS FLX sequencing runs were processed using platform-specific pipeline software to generate sequencing reads, base-call quality scores, and filter/remove low-quality reads. The sequence files were aligned to reference sequences for the 16 HCM genes shown in Table 1, using SeqMan NGen, Version 1.2 (DNAstar, Madison, WI, USA). The alignment settings used for both platforms were a 94% minimum read match to reference sequence, which allowed for two mismatches per 36 base-read length. An initial 15-base match size, match spacing of 10 bases, and match, mismatch, and gap scores of 10, −20, and −30 were used. To reduce the alignment effects of low-quality read ends, ends were trimmed to obtain a quality score of ≥18 over a two-base average. Variant discovery was performed using SeqMan Pro (DNAstar). Criteria for variant identification were a read coverage of ≥30-fold and a read percentage of ≥20%. Read coverage was defined as the number of bases that aligned to a specific nucleotide position, and read percentage was defined as the fraction of total bases that differed from the reference sequence. Variants of interest were cross-queried using National Center for Biotechnology Information Single Nucleotide Polymorphism Database (http://www.ncbi.nlm.nih.gov/projects/SNP/) to determine if the variant was a known polymorphism.

RESULTS

Enrichment and NGS of HCM-Associated Genes

LR-PCR was used to enrich 16 genes implicated in HCM (Table 1). A total of 319,646 nonoverlapping base pairs of

TABLE 4

Genome Analyzer and GS FLX Discordant Exon Variants in HCM-Associated Genes

		Genome analyzer		GS FLX		
Gene	Position	Read %	Coverage	Read %	Coverage	Comments
MYH6	**g.4867881delA**	**1.4**	**499**	**25.7**	**276**	**Ref, poly A**
MYH7	g.3690509delG	0.0	450	30.3	33	Ref, poly G
MYH7	g.4892783C>T	21.8	293	0.0	39	Misalignment
MYH7	g.4896485insT	0.0	499	21.0	181	Ref, poly T
MYH7	g.4902793C>T	22.6	402	8.6	163	Misalignment
MYL2	g.1918468delC	0.2	499	21.8	78	Ref, poly C
MYL2	g.1918460delC	0.0	499	20.0	80	Ref, poly C
PLN	g.23050706delT	0.0	330	24.7	89	Ref, poly T
PRKAG2	g.11968221delC	1.8	278	32.4	37	Ref, poly C
PRKAG2	g.11848871delT	2.2	182	47.3	448	Ref, poly T
TPM1	g.34125403delG	0.0	499	30.9	181	Ref, poly G
TPM1	g.34125415delA	0.5	436	22.7	181	Ref, poly A

Twelve variants with read percentages ≥20% on one platform but with <20% read percentage on the alternate platform were identified. Ten of the 12 variants were a result of GS FLX homopolymer sequencing errors, and two were a result of misalignments of Genome Analyzer 36 base-length reads between highly homologous regions of *MYH6* and *MYH7*. Variant shown in bold required cloning for confirmation.

genomic DNA was amplified representing 35,399 bases of mRNA transcript. Libraries were generated from equimolar amplicon pools and sequenced on the Illumina Genome Analyzer and Roche 454 GS FLX platforms. The Genome Analyzer single-lane output was 7.88×10^6 single-end 36 base-length reads. Of these reads, 6.94×10^6 (88.1%) aligned to HCM gene reference sequences with an average read-coverage depth of 782 and average base quality of 31. The single GS FLX picotiter plate output was 281,831 reads with an average read length of 235 bases. Of these reads, 253,238 (89.9%) aligned to HCM gene-reference sequences with an average coverage depth of 186 and average base quality of 32. An example of coverage variation for the cysteine and glycine-rich protein 3 (*CSRP3*) gene is shown in Figure 1.

Identification of Exon Variants

For variant identification, we used selection criteria that required variants to: 1) have a base-quality score of ≥18 at the respective base position in each read; 2) be present in ≥20% of aligned reads; and 3) have a read-coverage depth of ≥30. As the majority of reported HCM mutations resides in exons, we focused analyses on exon variant identification.[20] Analysis of Genome Analyzer data identified 31 exon variants. For this set, the average read percentage for the 29 heterozygous variants (including substitutions and single-base deletions) was 43.2% (range 20.6–65.9%) and 93.1% (range 91.9–94.2%) for the two homozygous variants. Average read coverages for the heterozygous and homozygous variants were 385 (range 41–499) and 488 (range 479–497), respectively. Analysis of Roche 454 GS FLX data identified 40 exon variants. For this set, the average read percentage for the 37 heterozygous variants (including substitutions, single-base deletions, and insertions) was 42.7% (range 20.0–85.9%) and 100% for the three homozygous variants. Average read coverages for the heterozygous and homozygous variants were 193 (range 32–454) and 267 (range 32–454), respectively. Between both platforms, 27 variants were concordant, and they were confirmed by Sanger sequencing and comprised of 25 heterozygous and two homozygous variants (Table 2).

After accounting for the 27 concordant exon variants, 17 discordant exon variants remained. Five variants, two Genome Analyzer and three GS FLX, had coverages ≥30-fold on one platform but <30-fold on the alternate platform. All five variants were Sanger-confirmed, resulting in a total of 32 exon variants between both platforms (Tables 2 and 3). The remaining 12 variants had read percentages ≥20% on one platform but <20% on the alternate platform (Table 4). Ten variants with ≥20% read percentage were observed in GS FLX data and two in Genome Analyzer data. Analysis of the 10 GS FLX variants showed the presence of sequencing errors in homopolymer tracts, usually observed as erroneous single-base deletions. The two discordant variants observed in Genome Analyzer data resulted from short-read misalignments in areas of high homology between myosin, heavy-chain 6 and 7 (*MYH6* and *MYH7*, respectively).

Of the 32 Sanger-confirmed exon variants, 18 were in coding regions. Sixteen of the 18 variants are synonymous, and two were nonsynonymous. Both nonsynonymous variants, g.4861650A > G (p.A1101V) and g.4859449C > T (pA1130T), are located in *MYH6* and have been reported previously not to be associated with HCM.[21] The other 14 variants located in untranslated regions have been reported previously (http://www.ncbi.nlm.nih.gov/

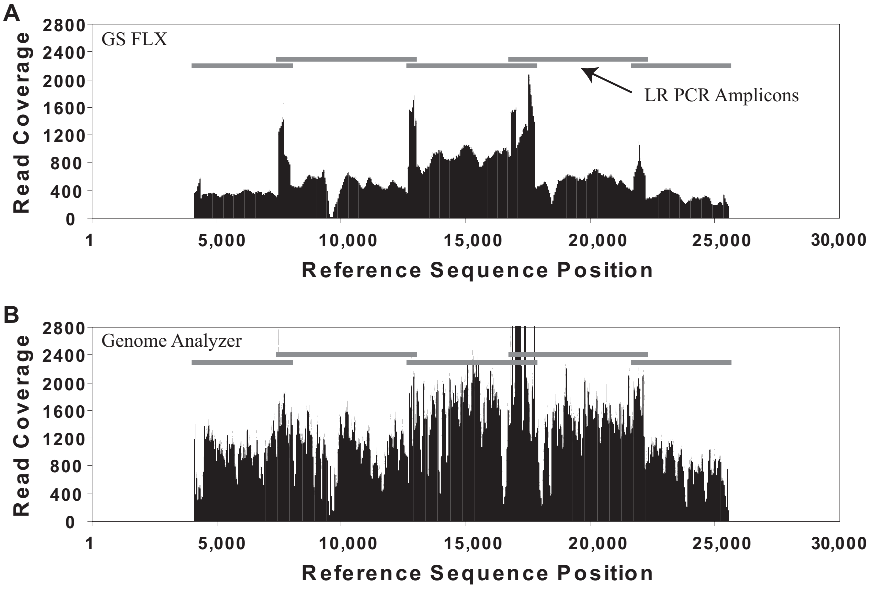

FIGURE 1

Read coverage versus reference position for the *CSRP3* gene for GS FLX (A) and Genome Analyzer (B). Overlapping LR-PCR amplicon positions are shown as horizontal gray bars above coverage plots. Significant features include coverage variability and coverage peaks at amplicon ends.

projects/SNP/snp_blastByOrg.cgi). Of the 23,016 bases Sanger-sequenced (7.2% of total unique bases analyzed by NGS), no false-negative variants were detected in data from either platform.

Identification of Intron Variants

The same variant identification criteria were applied in a preliminary analysis of intron sequences. A total of 965 intron sequence variants (419 Genome Analyzer and 546 GS FLX) was identified. Of these variants, 326 Genome Analyzer and 248 GS FLX variants were single-base variants with 224 concordant. Ninety-three Genome Analyzer and 298 GS FLX small insertions or deletions were identified, of which 19 were concordant. Cumulatively, 243 intron variants were concordant. Eleven of the concordant intron variants were Sanger-confirmed as a result of their proximity to exons. Preliminary analysis also indicated that homopolymer sequencing errors and misalignments in regions of cross-homology and low-complexity were primary causes for discordant results.

DISCUSSION

This study compared the performance of the Illumina Genome Analyzer and Roche 454 GS FLX NGS platforms for targeted resequencing of genes implicated in HCM. LR-PCR was chosen to enrich the HCM genes of interest as a result of the increased specificity associated with PCR versus hybridization capture methods and the lower costs associated with the technique. Prior to primer design, *MYH6* and *MYH7* were noted to have high degrees of sequence homology (51% for genomic DNA and 80% for mRNA), and the use of LR-PCR ensured that *MYH6* and *MYH7* were amplified independently. Of the 67 LR-PCR amplifications, 84% reproducibly yielded single bands of the correct length. As some amplicons contained minor products, we elected to gel-purify all amplicons. Despite efforts to ensure equimolar amplicon pooling, the coverage for one myosin-binding protein C3 (*MYBPC3*) amplicon (primer pair F1/R1) was under-represented in both platforms' data, which was attributed to a pooling error based on follow-up experiments. Redesign of suboptimal primer pairs and the elimination of the amplicon gel-purification step have been achieved subsequently. A series of protocol improvements has been implemented or is ongoing to optimize library preparation for clinical applications. These include the elimination of Speed Vac concentration prior to fragmentation (known to introduce cross-contamination between samples when dried in parallel); the use of sonication instead of nebulization for DNA fragmentation (reducing sample-to-sample fragment variation and the potential cross-contamination of samples as a result of volatilization); and the elimination of the LR-PCR quality-control gels.

As observed in our data, read-coverage patterns varied along the targeted genomic regions for both platforms (Fig. 1). Sources of variability may include enrichment methods, differential adapter ligation-to-target fragments, and unequal PCR amplification efficiencies during library generation.[1,9,22–24] Variability in coverage is a limitation of NGS that requires samples to be over-sequenced to achieve the minimum desired coverage per base. Coverage peaks corresponding to the ends of LR-PCR amplicons are a result of a relative excess of adapter ligation to amplicon ends. A recent report shows that this can be reduced by 5′ blocking of amplification primers.[23]

Different criteria for variant identification have been used in NGS studies depending on platform, software, and specific study goals. We required an initial read coverage of ≥30-fold for variant identification based on a yeast genomic resequencing study, in which a 10- to 15-fold coverage of the haploid yeast genome yielded accurate variant detection with the Genome Analyzer and GS FLX platforms.[25] With respect to read percentage required for variant identification, we relaxed our criterion to ≥20% to minimize false-negatives, albeit anticipating false-positives. Applying these criteria, 31 and 40 exon variants were observed in Genome Analyzer and GS FLX data, respectively. Of these variants, 27 were concordant between platforms and confirmed by Sanger sequencing. The 17 discordant exon variants segregated into two categories. Five of the 17 discordant variants had read coverages ≥30 on one platform but <30-fold (range 9–29) on the alternate platform. These five variants were Sanger-confirmed, suggesting that a coverage criterion of ≥30-fold is too stringent for variant discovery. Other groups have suggested a minimum coverage of 20-fold.[1,26] The second category comprised of 12 variants with read percentages ≥20% on one platform but <20% on the alternate platform. Ten of the 12 discordant variants resulted from GS FLX sequencing errors in homopolymer tracts. In most cases, this error type had one less base than present in the reference sequence homopolymer tract. Homopolymer sequencing errors have been reported previously for the Roche 454 technology and result from nonlinear luminescence relative to homopolymer length during pyrosequencing.[27,28] The other two discordant variants, observed in Genome Analyzer data, resulted from misalignment of 36 base-length reads in homologous regions of *MYH6* and *MYH7*. During the alignment process, the SeqMan NGen software generates a best-match consensus; however, a degree of randomness is introduced when the software attempts to align short reads to homologous reference sequences. This results in misplaced alignments and skewed variant-read percentages associated with short-read lengths. It should be noted that misalignment as a result of cross-homology also skewed read percentages for concordant *MYH6* and *MYH7* variants (Table 2). Three concordant *MYH6* and *MYH7* heterozygous variants (*MYH6* g.4853994G>A, *MYH7* g.4900633G>A, and *MYH6* g.4859449C>T) had Genome Analyzer read percentages between 21% and 26% compared with GS FLX read percentages between 38% and 51%. Comparison of the *MYH6* and *MYH7* genes showed a 95–100% sequence identity ±18 bases from these variant positions. The observation that the GS FLX read percentages for these variants were closer to the expected 50% is consistent with the expectation that longer reads improve alignments. A particular alignment challenge was posed by the *MYH6* variant g.4859449C > T, which resides in a 416-bp region of 100% identity with *MYH7*. This degree of identity skewed variant read percentages for Genome Analyzer and GS FLX reads (22.2% vs. 38.6%, respectively). Accurate alignment of short reads to regions of cross-homology is an inherent challenge in NGS data analysis and not unique to the SeqMan NGen software.[29,30]

The current proof-of-concept study for the implementation of a HCM multigene-resequencing assay in the clinical environment is limited in that only a single human genomic DNA was sequenced. However, our analysis of the comparative performance of the Genome Analyzer and GS FLX platforms has provided valuable insights into technical issues that need to be addressed. Of critical importance will be reducing the number of false-positive variants resulting from homopolymer sequencing errors and misalignments. Subsequent to performing the experiments described, relevant technical advances became commercially available for both platforms. For the GS FLX, ≥400 base-length reads may be achieved using Titanium chemistry. The GS FLX picotiter plate has been modified with metal-coated walls to confine luminescence to individual wells, reducing signal-to-noise ratio as well as well-to-well crosstalk. These improvements, as reported by Roche 454, increase the overall sequencing accuracy, especially in homopolymer tracts. For the Genome Analyzer, ≥72 base-length reads and the introduction of pair end sequencing provides new approaches to reduce misalignments.

Future experiments will include the analysis of a larger number of samples, including those with larger insertions and deletions, which will be required to further establish criteria for variant identification and to characterize more fully platform performance. A balance among sensitive variant identification, minimization of false-positives, and the amount of confirmatory Sanger sequencing required needs to be determined. Although LR-PCR is a robust enrichment technique, it becomes less practical as the target size increases. Capture arrays and the recently described parallel microdroplet-based PCR enrichment method by RainDance Technologies will facilitate scaling up to larger targets.[31–33] One limitation for arrays, however, is cocapture of highly homologous sequences present in related genes and pseudogene correlates.[34–36] Library preparation is a technical, multi-step process requiring approximately 1.5 days for the Genome Analyzer and 3–4 days for the GS FLX. Commercial efforts are ongoing to automate aspects of library preparation. Our material costs for LR-PCR enrichment, library preparation, and sequencing of a single human genomic DNA sample were approximately $1500 for the Genome Analyzer single lane and $5500 for the GS FLX picotiter plate. The GS FLX sample cost could have been reduced by using one-half of the picotiter plate, although this may have resulted in missing variants in areas of insufficient coverage. An alternative approach for cost reduction is pooling sample libraries after they have been prepared, and identifier sequences or molecular barcodes are incorporated into the oligonucleotide adapters used for library generation. Post-sequencing, the identifiers are used bioinformatically to assign reads to their sample of origin.[35,37]

In conclusion, targeted resequencing by combining LR-PCR and NGS offers a new approach for multigene analysis. In this study, we have shown initial feasibility of this approach for resequencing genes associated with HCM using the lllumina Genome Analyzer and the Roche 454 GS FLX platforms. We are currently incorporating platform improvements in an effort to improve sequencing accuracy and streamlining technical processes as next steps toward transitioning NGS into the clinical laboratory.

ACKNOWLEDGMENTS

Support for this study was provided by the ARUP Laboratories Institute for Clinical and Experimental Pathology. The authors thank Maria Erali for providing helpful comments in the preparation of this manuscript. We also thank David Nix, Brian Dalley, and Brad Cairns of the Huntsman Cancer Institute and Alex Poole, Todd Castoe, and David Pollack at the University of Colorado for access to the Illumina Genome Analyzer and Roche 454 GS FLX platforms, respectively.

REFERENCES

1. Hillier LW, Marth GT, Quinlan AR, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 2008; 5:183–188.
2. Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* 2009;48:240–248.
3. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10:669–680.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63.
5. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008; 92:255–264.
6. Wilhelm BT, Landry JR. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 2009;48:249–257.

7. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 2009;19:1124–1132.
8. Summerer D. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics* 2009;94:363–368.
9. Harismendy O, Ng PC, Strausberg RL, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009;10:R32.
10. Yeager M, Xiao N, Hayes RB, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum Genet* 2008;124:161–170.
11. Soor GS, Luk A, Ahn E, et al. Hypertrophic cardiomyopathy: current understanding and treatment objectives. *J Clin Pathol* 2009;62:226–235.
12. Roberts R, Sidhu J. Genetic basis for hypertrophic cardiomyopathy: implications for diagnosis and treatment. *Am Heart Hosp J Spring* 2003;1:128–134.
13. Wren C, O'Sullivan JJ, Wright C. Sudden death in children and adolescents. *Heart* 2000;83:410–413.
14. Taylor MR, Carniel E, Mestroni L. Familial hypertrophic cardiomyopathy: clinical features, molecular genetics and molecular genetic testing. *Expert Rev Mol Diagn* 2004;4:99–113.
15. McNally E, Dellefave L. Sarcomere mutations in cardiogenesis and ventricular noncompaction. *Trends Cardiovasc Med* 2009;19: 17–21.
16. Chang AN, Parvatiyar MS, Potter JD. Troponin and cardiomyopathy. *Biochem Biophys Res Commun* 2008;369:74–81.
17. Bos JM, Towbin JA, Ackerman MJ. Diagnostic, prognostic, and therapeutic implications of genetic testing for hypertrophic cardiomyopathy. *J Am Coll Cardiol* 2009;54:201–211.
18. Sehnert AJ, Huq A, Weinstein BM, Walker C, Fishman M, Stainier DY. Cardiac troponin T is essential in sarcomere assembly and cardiac contractility. *Nat Genet* 2002;31:106–110.
19. Gerull B, Gramlich M, Atherton J, et al. Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat Genet* 2002;30:201–204.
20. Rodriguez JE, McCudden CR, Willis MS. Familial hypertrophic cardiomyopathy: basic concepts and future molecular diagnostics. *Clin Biochem* 2009;42:755–765.
21. Carniel E, Taylor MR, Sinagra G, et al. α-Myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. *Circulation* 2005;112:54–59.
22. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 2008;5:1005–1010.
23. Harismendy O, Frazer K. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques* 2009;46:229–231.
24. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 2009;6:291–295.
25. Smith DR, Quinlan AR, Peckham HE, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* 2008;18:1638–1642.
26. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature* 2008;456:60–65.
27. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Res* 2001;11:3–11.
28. Alderborn A, Kristofferson A, Hammerling U. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Res* 2000;10:1249–1258.
29. Li H, Ruan J, Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;18:1851–1858.
30. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009;6:S6–S12.
31. Brouzes E, Medkova M, Savenelli N, et al. Droplet microfluidic technology for single-cell high-throughput screening. *Proc Natl Acad Sci USA* 2009;106:14195–14200.
32. Leamon JH, Link DR, Egholm M, Rothberg JM. Overview: methods and applications for droplet compartmentalization of biology. *Nat Methods* 2006;3:541–543.
33. Tewhey R, Warner JB, Nakano M, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* 2009;27:1025–1031.
34. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26:1135–1145.
35. Maeda N, Nishiyori H, Nakamura M, et al. Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra high-throughput sequencer. *Biotechniques* 2008;45:95–97.
36. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;27:182–89.
37. Binladen J, Gilbert MT, Bollback JP, et al. The use of; coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* 2007;2:e197.