A Variant Detection Pipeline for Inherited Cardiomyopathy—Associated Genes Using Next-Generation Sequencing

Théo G.M. Oliveira,* Miguel Mitne-Neto,† Louise T. Cerdeira,† Julia D.C. Marsiglia,* Edmund Arteaga-Fernandez,† José E. Krieger,* and Alexandre C. Pereira*

From the Laboratory of Genetics and Molecular Cardiology* and the Clinical Unit of Cardiomyopathies,† Heart Institute (InCor), University of São Paulo, São Paulo; and the Department of Research and Development,† Fleury Group, São Paulo, Brazil

Accepted for publication February 26, 2015.

Address correspondence to Théo G.M. Oliveira, B.Sc., Av. Dr. Enéas de Carvalho Aguiar 44, 05403-900 São Paulo, SP, Brazil. E-mail: theo.gremen@usp.br.

Inherited cardiomyopathies (ICs) are a group of cardiac diseases that include hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy, restrictive cardiomyopathy, left ventricular noncompaction, and arrhythmogenic right ventricular cardiomyopathy. These conditions have known genetic etiologies, and the diagnostic procedures may include genetic testing, which, in the context of ICs, has the potential to determine the inheritance pattern of a given condition as well as to support the clarification of overlaps that might exist among all of these cardiac conditions. Additionally, the sooner a mutation is identified, the faster it can be used for the guidance of genetic counseling in affected families.

Advances in DNA sequencing and target enrichment are allowing researchers and clinicians to assess genetic alterations in a more rapid and cost-effective manner, leading to an increasing trend in using molecular diagnostics tools for ICs. Although genome and exome sequencing are becoming accessible for genetic investigation, disease-targeted panels are seen as a more feasible alternative for molecular screening regarding inherited diseases. Advantages of delimited panels are not only primarily related to the objectivity of diagnostic interpretation but also rely on their technical advantages, including better sequencing coverage and reliable analytical sensitivity and specificity.

Supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Brazil Federal Government) and Fleury Group no. 12618 (FINEP).

Disclosures: Fleury Group (São Paulo, Brazil) provided development and research infrastructure for this study.

Copyright © 2015 American Society for Investigative Pathology and the Association for Molecular Pathology. Published by Elsevier Inc. All rights reserved.

http://dx.doi.org/10.1016/j.jmoldx.2015.02.003
Particularly for ICs, the indication of genetic panels is noticeable in the literature, with the number of interrogated genes ranging between 23 and >200. This variability can be attributed to an increasing number of genes indicated as related to cardiomyopathies and the comprehensiveness and objectivity designed for each panel. Attention is being given to dilated cardiomyopathies and HCM, the latter presenting the higher prevalence (1:500 population) and a more extensive evidence-based genetic background among all ICs. Both conditions are more commonly related to mutations in sarcomeric genes, although mutations in non-sarcomeric genes are also described.

Beyond its primary scope, cardiomyopathy panels can also include genes for differential diagnostics of other syndromes in which there is cardiac hypertrophy. In cases such as Noonan spectrum disorders, genetic investigation of known related genes along with sarcomeric genes is indicated, especially in cases of severe hypertrophy. Another issue concerns the sequencing of genes related to HCM phenocopies, such as glycogen-storage diseases (eg, amyloidosis, Fabry disease, and Pompe disease) as these conditions might present with cardiac hypertrophy but with a genetic etiology different from that of the HCM.

Efforts are being made to guide laboratories and researchers in the implementation of next-generation sequencing (NGS) assays into clinical routine, with best-practice use of these tools, and to validate the practicability of NGS assays. Of all of the steps involved in the analysis of NGS data, computational pipelines appear as a fundamental procedure, with crucial points that have influence in the final result. For an accurate analysis, the validation process is an important step in defining the use of filters and other parameters of data processing.

Specifically for NGS, guidelines suggest that a set of reference samples previously analyzed with a gold-standard method (ie, Sanger sequencing) be used in the validation process of the newly implemented technology. The confirmation of NGS findings via Sanger sequencing is a laborious and costly practice; the use of reference materials is becoming a widely adopted strategy for the validation of pipelines. A recently published article by the National Institute for Standards and Technology introduced a high-resolution list of variants from a HapMap sample (NA12878) that now can be used as a benchmark for the validation of NGS assays.

Here, we delineate a validation process to test the accuracy of a target panel that includes 74 genes accounting for several genetic classes, such as sarcomeric, Z-disc, desmosomal, cytoskeletal, and calcium homeostasis. Noonan spectrum disorder—related genes (PTPN11, KRAS, HRAS, SOS, RAF1, and SPRED1) were also included. Additionally, we evaluated genes from glycogen-storage diseases such as amyloidosis, Pompe disease, and Fabry disease (TTR, GAA, and GLA, respectively), for possible differential diagnostics of the HCM. An NGS pipeline was designed for clinical application in the ICs, using a set of molecular positive and negative samples from a previously genotyped casuistic of HCM patients and the reference sample (NA12878) from the National Institute for Standards and Technology as a benchmark.

### Materials and Methods

#### Panel Design

The enrichment of 74 target genes (Supplemental Table S1) was performed with the HaloPlex Target Enrichment System (Agilent Technologies, Santa Clara, CA). First, the probes were designed via a SureDesign web tool version 3.0.1.4 (Agilent Technologies), with genes being indicated by its respective RefSeq symbols. The genome reference used was University of California, Santa Cruz hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips), last accessed May
52 Samples from HCM-casistic (19 Positives and 33 Negatives) + NA12878 Triplet

Pre-enrichment Quantification

Enrichment Process 74 genes Haloplex System

Post-enrichment Quantification

Amplicon size distribution

Template Preparation

Ion OneTouch System
5 samples pool [10 pmol/L]

Ion PGM sequencing 318 Chip

FastQ

Sequencing QC

Trim Sequences
(Trim out: Phred <20; 5 bp of 3' end; short reads)

Map Reads to Reference (hg19)

Statistics for Target Regions (Exons ± 10 bp)

Quality-Based Variant Call

Neighborhood quality (5-bp radius)
Bidirectional Presence of Variants

Annotate with Overlap Information
(Gene names and accession numbers)

Annotate with Exon numbers
(GTF file)

Annotate from Known Variants (dbSNP 137)

Amino Acid Changes Predict Splice Site Effect

Graphical QC Report

Quality Metrics

BAM

Visual Inspection

Coverage Reports

Low-Coverage Regions

Minimum Coverage and VAF test:

SNVs
30x - 35%, 25%, and 20%
10x - 35% and 20%

Indels
30x - 35%, 25%, and 20%
20x - 35% and 20%

BED Filter

VCF Comparator

HCM-casistic variant list
NA12878 benchmark VCF

General Variant List

Inter- and Intra-assay Reproducibility

Sensitivity; Specificity;
PPV; FDR

CLC Genomics Workbench

Figure 1  Workflow of sample processing in the validation design. Pipeline was designed in CLC Genomics Workbench 6.5 (CLCbio, Boston, MA). Quality-based variant call was tested for five parameter sets of coverage and variant allele frequencies (VAFs) (30x of coverage with 35%, 25%, and 20% of VAF; 10x of coverage with 35% and 20% of VAF). BAM, .bam file; BED, .bed file; dbSNP, Single Nucleotide Polymorphism database; FDR, false-discovery rate; GTF, .gtf file; HCM, hypertrophic cardiomyopathy; PGM, Ion Torrent Personal Genome Machine (Life Technologies, Carlsbad, CA); PPV, positive predictive value; Q, Phred value (Q score); QC, quality control; SNV, single-nucleotide variant; VCF, .vcf file.
16, 2013). The final design consisted of 1754 target regions comprising 376,553 bp. Only 1.08% of all target bases were denoted as total or partially missed by the manufacturer. In total, coverage of regions was 98.92%, and amplicons covered coding exons with 10 bp flanking 3′ and 5′ ends. The recommended minimum sequencing per sample was 181.36 Mb for a 200× coverage depth.

Sample Selection and Experimental Design

Samples were selected from a previously studied casuistic of individuals with positive clinical diagnoses of HCM from the Heart Institute (University of São Paulo, São Paulo, Brazil). Clinical diagnosis was performed by experienced cardiologists and molecular diagnostic via Sanger sequencing of the three most HCM-associated genes: MYH7, MYBPC3, and TNNT2 (here referred to as the HCM-Sanger panel). An experimental set was composed of 52 samples, of which 19 (P1–P19) were positive for molecular diagnostics (ie, carrying at least one cardiomyopathy-related mutation in one of the three genes) and 33 were negative (N1–N33). Together, these samples presented 285 key variants. Reference sample NA12878 was purchased from the Coriell Institute for Medical Research (Camden, NJ).

The experimental design included 12 sequencing runs, each with five samples (Table 1 and Supplemental Table S2). Positive and negative samples were randomly assigned in runs 1 to 8. For reproducibility tests, NA12878 underwent a triplicate enrichment process, and replicates were assigned for inter- and intra-assay reproducibility tests. For the interassay test, each enriched reference sample was sequenced in parallel with four negative samples in runs 9, 10, and 11, thus treating each replicate as a different sample. The intra-assay test was performed in run 12, in which the same three replicates (referred as NA12878_1.1, _2.1, and _3.1) were sequenced along with two other repeated HCM-casuistic samples (P11.1 and P13.1). Thus, a total of 60 samples were sequenced: 54 from our HCM-casuistic samples (52 samples plus 2 repeats) and 6 replicates of the NA12878 reference sample (3 for inter- and 3 for intra-assay evaluation).

Target Enrichment Process

Enrichment protocol was performed according to the manufacturer’s instructions (HaloPlex Target Enrichment System version D3, March 2013, for Ion Torrent sequencing, catalog no. G9912B; Agilent Technologies). Before enrichment, sample quantification was performed using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA). An initial concentration of 5 ng/μL was used for optimal enrichment. After enzymatic fragmentation, DNA samples and barcodes were added to the probe mix and placed in a Veriti Thermal Cycler (Life Technologies) for 16 hours at 54°C. At the end of this step, probes were captured with the use of streptavidin beads to separate hybridized from nonhybridized DNA. A further elution with 50 mmol/L NaOH was performed to dehybridize enriched DNA from probes. The remaining material was then amplified using 20-cycle PCR and Hercule II polymerase (Agilent Technologies). PCR products were purified with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) and eluted with 10 mmol/L Tris-HCl (pH 8.0) for long-term storage at −20°C. Sample quantification and amplicon size distribution were assessed with bioanalyzer equipment (1000 DNA kit; Agilent Technologies).

Template Preparation

Template DNA was prepared with the Ion OneTouch System and Ion OneTouch 200 Template Kit version 2 (Life Technologies). After quantification, DNA was diluted to 10 pmol/L, and five samples were pooled in an equimolar manner for each template preparation. The total input for emulsion PCR was 25 μL. After emulsion PCR, template-positive ion sphere particles were enriched in the Ion enrichment system instrument (Life Technologies) to avoid the insertion of empty ion sphere particles in the sequencing reaction. Final templates were stored at 4°C for up to 3 days.

Sequencing

Samples were sequenced in Ion Torrent Personal Genome Machine (PGM) platform (Life Technologies), and all sequencing reactions were performed with the Ion PGM 200
sequencing kit. Five samples per run were sequenced in an Ion 318 Chip version 2 (Life Technologies) to respect the minimum recommended sequencing coverage per sample. Runs were planned with Ion Browser software version 4.4.2 (Life Technologies) and had a total time of 4 hours.

**Bioinformatics Analysis**

By the end of each run, reads were automatically processed for 3′ and 5′ adapter trimming and FastQ file generation. A visual quality assessment of reads was performed with FastQC (Supplemental Figures S1 and S2). After conversion, all FastQ files were imported into CLC Genomics Workbench 6.5 (CLCbio, Boston, MA) in which a pipeline was set up for downstream analysis (Figure 1). The main steps in the pipeline were:

i) In an initial trimming process, 5 bp from the reads’ 3′ end were trimmed out to avoid mapping of low-quality regions. Short reads and reads with a Phred score (Q value) < 20 were filtered out.

ii) Read mapping was performed against UCSC hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips, last accessed May 16, 2013) using default cost values for mismatches (2) and indels (3) and minimum reference similarity (80%). A mapping report and a .bam file were generated as outputs.

iii) Statistics for target regions were generated after read mapping. Target regions were defined as covered exons with 10 bp of intronic regions upstream and downstream. A coverage report was provided as outputs containing whole mapping and exon-specific information.

iv) Quality-based variant call (VC) uses the variant neighborhood quality (5-bp radius) as a measure for positive calls. Bidirectional presence of variant alleles was required in a concordance of at least 5% between forward and reverse reads. The pipeline was tested for a minimum coverage of 30× with variant allele frequencies (VAFs) of 35%, 25%, and 20%, and 10× coverage with VAFs of 35% and 20% for single-nucleotide variants (SNVs). For indels, the minimum tested coverage cutoff was 20× with VAFs of 35% and 20%.

v) The annotation process was divided into three substeps: a) annotation with overlap information (gene names and accession numbers), b) annotation with exon numbers using mRNA coordinates, and c) annotation with known variants using Single Nucleotide Polymorphism database (dbSNP; http://www.ncbi.nlm.nih.gov/snp) 137 as a reference. No variants were filtered out in this step.

vi) An impact-prediction step was used to predict splicing site variants and nonsynonymous variants. A general variant list including all variants detected was generated at the end of this step. All tracks were converted to .xls files for posterior data manipulation in Excel 2010 (Microsoft Corporation, Redmond, WA).

The time of Ion browser data processing after the end of a run was around 1 hour (adapter trimming, barcode sorting, and FastQ file generation). The designed pipeline in CLC Genomics Workbench version 6.5 presented a per-sample analysis time of 30 minutes (2 hours 30 minutes per run). Both software packages were installed on a server with 48 GB of random-access memory with dual six-core processors. We opted to pay more attention to the performance of the VC step, which has a great influence on the final analytical parameter values. The choices of values of coverage and VAF were based on their appearance in other studies as well as on the use of bidirectional presence of variants as a way to avoid false-positive results (FPs).

---

**Table 3 Validation Process with HCM-Sanger Panel Variants (N = 285)**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>30× Coverage</th>
<th>10× Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35% VAF</td>
<td>25% VAF</td>
</tr>
<tr>
<td>TP</td>
<td>263</td>
<td>268</td>
</tr>
<tr>
<td>FP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sensitivity, %</td>
<td>92.3</td>
<td>94</td>
</tr>
<tr>
<td>95% CI</td>
<td>89.55–96.45</td>
<td>91.41–97.48</td>
</tr>
<tr>
<td>Specificity, %</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>95% CI</td>
<td>99.0–100.0</td>
<td>99.0–100.0</td>
</tr>
</tbody>
</table>

FP, false positive; HCM, hypertrophic cardiomyopathy; TP, true positive.
Validation Process and Analytical Assessment

Target regions analyzed in the HCM-Sanger panel were annotated in a .bed file and were used as a filter for the general variant list so that only variants in the three previously sequenced genes (MYH7, MYBPC3, and TNNT2) were maintained in a filtered list. Therewith, only variants detectable in Sanger sequencing were used for checking the NGS accuracy. Sensitivity was calculated as

\[
\text{Sensitivity} = \frac{\text{True positives}}{\text{Total positives}}
\]

The NA12878 benchmark .vcf file was downloaded from FTP server of the National Institutes of Health’s National Center for Biotechnology Information (ftp://ftp-trace.ncbi.nih.gov/nuccore/NA12878/variant_calls/NIST, last accessed September 19, 2014). As indicated by the curators, the .vcf file NISTIntegratedCalls_14datasets_131103_allcall_UGHapMerge_HetHomVarPASS_VQSRv2.18_all_nouncert_excldesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs.vcf was used as the benchmark .vcf for containing highly confident hetero- and homozygous VCs (here referred to as key variants). A .bed file containing coordinates from well-covered regions was also downloaded (union13_callableMQonlymerged_addcert_nouncert_exclusimpler ep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_y2.18_2mindatsets_5minYesNoRatio.bed). Comparison between the benchmark and experimental .vcf files was performed with the VcfComparator program (Used software version 8.9.2; http://sourceforge.net/projects/useq, last accessed August 6, 2014).

One crucial point of comparison between .vcf files was the genotype match parameter, in which only variants called in the correct genotype were considered as a match (or true positive). True variants called within an incorrect genotype were considered as nonmatches, whereas variants detected solely in experimental .vcf files were considered as FPs. Thus, analytical parameters were calculated as follows:

\[
\text{Sensitivity} = \frac{\text{Match variants}}{\text{Total key variants}}
\]

\[
\text{Specificity} = \frac{\text{True-negative bases}}{\text{(True-negative bases + FPs)}}
\]

\[
\text{Positive predictive value (PPV)} = \frac{\text{Matches}}{\text{(Matches + Nonmatches + FPs)}}
\]

\[
\text{False-discovery rate (FDR)} = 1 - \text{PPV}
\]

Reproducibility was defined as the concordance (expressed as a percentage) of matches between samples in different runs in the same VC scenario.

**Results**

Sequencing

Sequencing in Ion Torrent PGM generated a mean sample throughput of 169.5 Mb, with a mean of 150.4 Mb Q > 20. Table 1 summarizes all of the raw and processed sequencing data. The alignment statistics showed that a mean of 94.27% (n = 60; 95% CI, 93.3–95.25) of reads were aligned with the reference genome and that the mean specificity to target regions was 94.12% (n = 60; 95% CI, 93.77–94.48). The template preparation showed acceptable values of polyclonal ion sphere particles, ranging between 7% and 26%.

### Table 1

<table>
<thead>
<tr>
<th>Samples</th>
<th>Known mutation, gene and amino acid change</th>
<th>NGS Coverage (VAF, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>MYH7 - p.Ser842Gly</td>
<td>Yes 249 × (37.7)</td>
</tr>
<tr>
<td>P2</td>
<td>MYBPC3 - c.3330+2T&gt;C</td>
<td>Yes 396 × (50.2)</td>
</tr>
<tr>
<td>P3</td>
<td>MYH7 - p.Ile263Thr</td>
<td>Yes 895 × (44.5)</td>
</tr>
<tr>
<td>P4</td>
<td>MYH7 - p.Ile702Val</td>
<td>Yes 242 × (46.7)</td>
</tr>
<tr>
<td>P5</td>
<td>MYH7 - p.Phe764Tyr</td>
<td>Yes 127 × (54.3)</td>
</tr>
<tr>
<td>P6</td>
<td>MYH7 - p.Ile702Val</td>
<td>Yes 329 × (50)</td>
</tr>
<tr>
<td>P7</td>
<td>MYH7 - p.Phe252Cys</td>
<td>Yes 340 × (41.1)</td>
</tr>
<tr>
<td>P8</td>
<td>MYH7 - p.Arg249Gln</td>
<td>Yes 599 × (45)</td>
</tr>
<tr>
<td>P9</td>
<td>MYH7 - p.Arg858Gys</td>
<td>Yes 139 × (43.1)</td>
</tr>
<tr>
<td>P10</td>
<td>MYH7 - p.Val320Met</td>
<td>Yes 227 × (48.4)</td>
</tr>
<tr>
<td>P11</td>
<td>MYH7 - p.Met493Ile</td>
<td>Yes 114 × (53.5)</td>
</tr>
<tr>
<td>P12</td>
<td>MYBPC3 - c.2905+1G&gt;A</td>
<td>Yes 155 × (52.2)</td>
</tr>
<tr>
<td>P13</td>
<td>MYH7 - p.Lys1459Asn</td>
<td>Yes 262 × (40)</td>
</tr>
<tr>
<td>P14</td>
<td>MYBPC3 - c.3628-16G&gt;A</td>
<td>Yes 136 × (43.4)</td>
</tr>
<tr>
<td>P15</td>
<td>MYBPC3 - Phe305 fs</td>
<td>No 88 × (80)*</td>
</tr>
<tr>
<td>P16</td>
<td>MYH7 - p.Val320Met</td>
<td>Yes 126 × (47.6)</td>
</tr>
<tr>
<td>P17</td>
<td>MYBPC3 - p.Glu619Lys/MYBPC3 - p.Leu1221Fs</td>
<td>No/Yes 9 × (89)/203 × (48.2)</td>
</tr>
<tr>
<td>P18</td>
<td>MYH7 - p.Ala797Thr</td>
<td>Yes 334 × (48.5)</td>
</tr>
<tr>
<td>P19</td>
<td>MYBPC3 - p.Asn850Fs</td>
<td>Yes 37 × (48)</td>
</tr>
</tbody>
</table>

*Variant with no bidirectional representation.

HCM, hypertrophic cardiomyopathy; NGS, next-generation sequencing; VAF, variant allele frequency.
Coverage of Target Regions

The overall mean coverage of target regions was 250×. As described by Meder et al,19 the enrichment factor was calculated, and the mean value was 7930-fold (n = 60; 95% CI, 7873.31–7985.86). The mean GC content of the target regions was 47.7%, and the minimum and maximum values were 20.7% and 74.8%, respectively. The coverage frequency was plotted in a histogram based on the mean coverage values of the target regions for the 60 samples so that the target regions could be analyzed in the coverage ranges (Figure 2). Twenty-eight target regions (1.6%) appeared within the range between 0 and 10× of coverage, and 63 target regions (3.5%) appeared within the range between 0 and 30×. The coverage of target regions was analyzed in a spectrum of seven cutoffs (Table 2) to observe the increment of analyzable percentages of the whole panel. The variability of coverage distribution among samples is shown in Figure 3.

Variant Call Sensitivity and Specificity

In the first assessment of VC sensitivity, all 285 variants previously found in MYH7, MYBPC3, and TNNT2 by Sanger sequencing (52 patients of HCM-casusistic) were considered. Of these, 20 were potentially pathogenic mutations (14 missense, 3 frameshift deletions, and 3 splicing variants) and 265 were SNPs. Except for the splicing variants, all of the other mutations and SNPs occurred in exons.

The quality-based VC was first tested for a fixed coverage cutoff of 30× and three VAF cutoffs (35%, 25%, and 20%). Sensitivity was 92.3% (263 of 285) in the 35% VAF and 94.0% (268 of 285) in the 25% and 20% VAFs. Subsequently, the coverage cutoff was decreased to 10× and was tested for VAF cutoffs of 35% and 20%. In this case, sensitivity values were 94.7% (270 of 285) in the 35% VAF and 96.5% (275 of 285) in the 20% VAF (Table 3). Of all pathogenic mutations, 18 (90%) were confirmed with a good coverage and were called with the correct genotype (Table 4). No FPs were found in the regions that were previously accessed by Sanger sequencing, indicating a target base specificity of 100%. Of all called variants with the 10× coverage/20% VAF parameters, 273 were SNVs and 2 were deletions. An extensive analytical measurement was performed with NA12878 benchmark data. Within the common regions of sequencing, 311,728 bp were interrogated and 134 key variants were to be confirmed in the designed pipeline (132 SNVs, 1 insertion, and 1 deletion). The same five parameter scenarios of VC sensitivity were tested for the NA12878 triplicate both for inter- and intra-assay validation. Tables 5 and 6 show the results of analytical validation performed with VcfComparator. In concordance with the HCM-Sanger panel, sensitivity increased as the stringency decreased, with 10× coverage/20% VAF being the maximum observed sensitivity point. The mean sensitivity value in inter- and intra-assay validation values were 92.7% and 91%, respectively, with a maximum value of sensitivity reaching 94.7% for NA12878_2 in run 10 (Table 5) and with a low occurrence of nonmatches. Specificity values were notably uniform along both processes, with a maximum value of 0.9999967, and this performance is directly related to a low occurrence of FPs. It is noteworthy that only SNVs were called, with a total absence of indels.
In a stratified analysis of all variants analyzed in both materials, 560 SNVs (282 in the HCM-Sanger panel, 132 in NA12878 from the interassay validation, 132 in NA12878 from the intra-assay validation, and 14 from the two repeated HCM-casuistic samples) and 7 indels (3 from HCM-casuistic, 2 from NA12878 interassay, and 2 from intra-assay experiments) were to be confirmed. Of all SNVs, 273 were called from the HCM-Sanger panel, 127 in NA12878 from the interassay validation, 126 in NA12878 from the intra-assay validation, and 14 from repeated P11.1 and P13.1. The maximum achieved SNV sensitivity was thus 96.4% (540/560). When adding nonmatches in the sensitivity calculation (as if they were matches), NA12878_2.1 would sum two variants to the final value, increasing the sensitivity to 96.7% (542/560). Of the 7 indels, only 2 were called, both presenting a length of 1 bp. Four indels (one 2-bp deletion in sample P15, one 7-bp insertion, and two 5-bp deletions in NA12878) were not called due to coverage problems, and one 7-bp insertion from NA12878_1.1, _2.1, and _3.1 was incorrectly called and not considered as a true positive. Therefore, indel sensitivity was 28.5% (2 of 7).

Whole-Panel PPV, FDR, and Reproducibility

PPV values were calculated for each triplicate of NA12878 in every VC scenario for both inter- and intra-assay experiments. These values remained above 0.959 in all scenarios, with a mean of 0.977. The FDR values (calculated as 1 – PPV) presented a mean of 0.021 and a maximum value of 0.041 (Table 6).

Reproducibility was evaluated from inter- and intra-assay perspectives and calculated as the concordant matches between each replicate in each VC scenario (Figure 4). Reproducibility evaluation with NA12878 (Tables 5 and 6) resulted in a maximum interassay reproducibility of 89.5% (120 concordant variants) (Figure 4), whereas the intra-assay reproducibility value was 87.3% (117 concordant variants) (Figure 4). Both reproducibility values had the influence of stringency decrease, with the maximum values occurring in the lower point of stringency (10/C2 coverage/20% VAF). The two HCM-casuistic repeated samples, P11.1 and P13.1, showed reproducibility values of 100%, with the 14 variants being identified in both runs.

Investigation of Noncalls and Nonmatches

With the decrease in VC stringency, the number of confirmed alterations in the HCM-Sanger panel increased, although 10 variants remained absent from the NGS results at the

Table 6 Results of Analytical Parameters from the Intra-Assay Validation Process with NA12878

<table>
<thead>
<tr>
<th>Coverage/Sample</th>
<th>Match (%)</th>
<th>NM</th>
<th>FP</th>
<th>Specificity</th>
<th>PPV</th>
<th>FDR</th>
<th>Rep (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAF (%)</td>
<td>30×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA12878_1.1</td>
<td>109 (81.3)</td>
<td>1</td>
<td>1</td>
<td>0.999967</td>
<td>0.981</td>
<td>0.019</td>
<td>105 (78.3)</td>
</tr>
<tr>
<td>NA12878_2.1</td>
<td>123 (91.7)</td>
<td>1</td>
<td>1</td>
<td>0.999967</td>
<td>0.984</td>
<td>0.016</td>
<td>126 (94)</td>
</tr>
<tr>
<td>NA12878_3.1</td>
<td>111 (82.8)</td>
<td>2</td>
<td>1</td>
<td>0.999967</td>
<td>0.973</td>
<td>0.027</td>
<td>118 (88)</td>
</tr>
<tr>
<td>95% CI</td>
<td>78.91-91.6</td>
<td></td>
<td></td>
<td>0.9998-1</td>
<td>0.97-0.01-0.03</td>
<td>0.99</td>
<td>93.92</td>
</tr>
<tr>
<td>VAF (%)</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA12878_1.1</td>
<td>112 (83.5)</td>
<td>0</td>
<td>2</td>
<td>0.999935</td>
<td>0.982</td>
<td>0.018</td>
<td>108 (80.5)</td>
</tr>
<tr>
<td>NA12878_2.1</td>
<td>123 (91.7)</td>
<td>1</td>
<td>1</td>
<td>0.999967</td>
<td>0.984</td>
<td>0.016</td>
<td>126 (94)</td>
</tr>
<tr>
<td>NA12878_3.1</td>
<td>111 (82.8)</td>
<td>2</td>
<td>1</td>
<td>0.999967</td>
<td>0.973</td>
<td>0.027</td>
<td>118 (88)</td>
</tr>
<tr>
<td>95% CI</td>
<td>80.4-91.6</td>
<td></td>
<td></td>
<td>0.9998-1</td>
<td>0.97-0.01-0.03</td>
<td>0.99</td>
<td>93.92</td>
</tr>
<tr>
<td>VAF (%)</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA12878_1.1</td>
<td>112 (83.5)</td>
<td>0</td>
<td>2</td>
<td>0.999935</td>
<td>0.982</td>
<td>0.018</td>
<td>108 (80.5)</td>
</tr>
<tr>
<td>NA12878_2.1</td>
<td>123 (91.7)</td>
<td>1</td>
<td>2</td>
<td>0.999935</td>
<td>0.976</td>
<td>0.024</td>
<td>126 (94)</td>
</tr>
<tr>
<td>NA12878_3.1</td>
<td>111 (82.8)</td>
<td>2</td>
<td>1</td>
<td>0.999967</td>
<td>0.973</td>
<td>0.027</td>
<td>119 (88.8)</td>
</tr>
<tr>
<td>95% CI</td>
<td>80.4-91.6</td>
<td></td>
<td></td>
<td>0.9998-1</td>
<td>0.97-0.02-0.03</td>
<td>0.98</td>
<td>94.04</td>
</tr>
</tbody>
</table>

FDR, false-discovery rate; FP, false positive; NM, nonmatch; PPV, positive predictive value; Rep, reproducibility; VAF, variant allele frequency.

Figure 4 Interassay (A) and intra-assay (B) reproducibility of NA12878 triplicates in 20× coverage/10% variant allele frequencies.
maximum point of sensitivity (10× coverage/20% VAF). Variants that were not detected in VC were visually inspected with Integrative Genomics Viewer software version 2.3 (Broad Institute, Cambridge, MA).20 Of the 10 absent variants, 2 were pathogenic alterations and 8 were polymorphisms. Table 7 lists all missing variants and the respective reasoning. The major reason for noncall was a lack of bidirectional coverage (7 variants), followed by low coverage due to high GC content (2 variants), and a case of total absence of a variant allele.

Both evaluations performed with NA12878 were investigated for noncalls and nonmatches. The intra-assay evaluation did not find 17 variants of the benchmark .vcf, and 7 in the experimental .vcfs were nonmatches (10× coverage/20% VAF). These numbers were decreased in the interassay evaluation, with 14 variants of the benchmark .vcf not being detected and 4 variants remaining as nonmatches in the experimental .vcf. Supplemental Tables S3 and S4 summarize all of the problematic variants and explain the reasons for these variants’ remaining absent. The major causes of missing variants were similar between the HCM-Sanger panel and the NA12878 validation test. Lack of coverage was majorly observed, frequently being associated with high or low GC content of target regions. Absence of bidirectional coverage was also observed.

Two issues may cause nonmatch of variants in experimental .vcfs: genotype nonmatch, which does not allow the variant to be recognized as a true positive although it is a genuine variant, and the variant being a FP and consequently not matching with the benchmark .vcf. These events were assessed, and their occurrence was of low frequency (Tables 5 and 6). With the 10× coverage/20% VAF cutoff, the maximum values were 3 both for nonmatches (NA12878_3.1) and FPs (NA12878_1.1). Of the seven positions implicated in the occurrence of FPs, four were homopolymeric regions.

Discussion

Here we explored two major steps involved in the application of NGS platforms and target panels in molecular diagnostics: the performance of an enrichment process and VC analytical parameters, with the purpose of identifying the balance between good sensitivity and specificity and a low FDR. Commercially available software was used, and we evaluated its capacity to analyze great amounts of data without the need for outer scripts in the main core of the analysis. The chosen NGS platform has been recently studied for molecular diagnostics applications, although for ICs it has been poorly explored. We also designed a validation process using previously Sanger-sequenced material and a reference sample from HapMap (NA12878) in accordance with the increasing trend of using this kind of material in the assessment of analytical parameters.13,21,22 Together, both sources provide a good amount of genetic variants to be analyzed.

In accordance with other applications,23,24 Ion Torrent PGM had a good sequencing performance, being capable of generating a satisfactory throughput with a high fraction of Q ≥ 20 bases (Table 1) and a good read quality (Supplemental Figure S1). The low rates of polyclonal ion sphere particles also contributed to a good throughput of usable reads. The use of 10 pmol/L as the initial concentration for template preparation was based on previous experience of HaloPlex and PGM users, posted on a virtual community (Ion Community; Life Technologies). Thus, even with the application of a trimming step in the pipeline, most of the generated reads (94.27%) could be used in the mapping process, and a good fraction of these (94.12%) were mapped in target regions.

These numbers are directly related to the good performance of the target enrichment. The mean coverage value in target regions was acceptable (250×) once these high values of sequencing coverage led to a more reliable VC process.7 All samples presented good coverage, although intersample coverage variability was observed (Figure 3). However, this kind of variability seemed not to have influence in coverage results, as a mean of 95.2% of interrogated base pairs were covered at least 10× and 97.7% were covered at least once. Thus, it might be related to pipetting and quantification bias as previously observed5 and to differences in sequencing run throughput.

In the same way, VC sensitivity presented reliable values for both HCM-Sanger panel samples (96.5%) and NA12878 samples (94.7% in NA12878_2), and the fact that it has been analyzed in five parameter settings gives robustness to the final results. It is noteworthy that changes in VAF and coverage cutoffs made a notable difference in the sensitivity results, and the cutoffs 10× coverage/20% VAF were defined...
as the maximum point of VC sensitivity while high values of specificity were maintained. Reproducibility values also increased along with the sensitivity, indicating that the match of alterations was occurring in a concordant manner for all NA12878 replicates. Inter- and intra-assay values of reproducibility were similar and outline a trend of uniformity between samples processing. Interestingly, our values of reproducibility with NA12878 were similar to previously reported values for SNVs in whole exome sequencing validation. These results reiterate the importance of a wide analysis of VC parameters such as presented in this work.

The correlation between decreased stringency and increasing sensitivity is expected, but it should be carefully adjusted for the avoidance of false discoveries. Our experimental design with NA12878 was a fundamental procedure to track the occurrence of FPs. Although the lower point of stringency presented the highest rate of FP occurrence, both inter- and intra-assay evaluations exhibited low values of FPs. Of the seven FP occurrences described in Supplemental Tables S3 and S4, four (chr2:179568916 T/G, chr6:7580724 A/G, chr6:7572262 A/G, and chr6:7584618 G/T) were within homopolymeric regions, confirming the problematic issues regarding the semiconductor sequencing chemistry and the correct identification of such regions. No FPs were found within the three most HCM-associated genes, both in the HCM-Sanger panel as in the whole-panel analysis. This fact is of major importance as, at least in Brazil, these three genes account for almost 50% of HCM cases with positive molecular diagnostics.

It is worthy to analyze nonmatches differently from FPs. Although both are not considered as true matches, their separate analysis allows for discerning between problems in VC. The occurrence of nonmatches indicates problems in the genotype call (which is not totally a FP) as the variant is true but appears with a wrong genotype. FP is more problematic as it is not a true variant and, depending on the localization and type of alteration, may cause problems in further analyses. In this work, nonmatch occurrence is not seen as a substantial problem, and more attention was given in relation to FP frequency, although the FDR was low and PPV remained higher than 0.959 during the whole validation process.

The minimum coverage cutoff used for SNVs was 10×, whereas for indels it was 20×. Our results show that the pipeline presents a tendency to favor the call of SNVs (maximum sensitivity, 96.7%), whereas for indels the resolution of results and the low number of representative variants do not allow us to state the same. As shown, with indels, sensitivity was 28.5%, with only 1-bp deletions being called. Thus, the pipeline performance for indels of >1 bp remains unclear. At first, this issue does not preclude the use of the designed pipeline for the molecular diagnostics of ICs, as the majority of observed mutations are point mutations and 1-bp indels.

A refined analysis of the noncalled variants allowed us to observe some possible problematic issues in the target enrichment. The main reason for missing variants in the HCM-Sanger panel (Table 7) was the lack of bidirectional coverage, which occurred in six variants, although the total coverage was sufficient for calling. Coverage problems appeared in two variants with <10× of coverage due to high–GC content regions, and one had no coverage for the variant allele. The same issues were present in noncalled variants from the NA12878 benchmark .vcf (Supplemental Tables S3 and S4). A lack of bidirectional representation and total coverage was observed, the latter being frequently associated with high or low–GC content regions.

Analyzing the distribution of missed variants among all runs performed, we did not observe a relationship between missing variants and low run throughput. For example, the run that presented the highest number of missed variants in the HCM-Sanger panel was run 2, even with a throughput of 1 Gb (the second greatest). The NA12878 interassay replicate that presented the highest number of noncalled variants (12 of 14) was NA12878_3 included in run 11. This run presented the greatest throughput among the three interassay runs. These data suggest that the lack of calling for the variants (described in Results) might be related to problematic genomic areas that failed to be enriched through the presented strategy more than to intervariability in the sample coverage and throughput. These areas, once identified, can be optimized by redesigned probes that might improve the coverage of such problematic regions.

Another strategy used for finding problematic regions was to calculate the mean coverage of all target regions and plot a histogram with a bin width of 10 (Figure 2). Thereby, all counts less than a given value can be identified as regions of low mean coverage and posteriorly optimized. We found 28 regions (1.6% of total) <10×, and of these, 8 (0.4%) were known as having been previously denoted as totally or partially missed by the kit manufacturer and are indicated for Sanger sequencing. So, actually, only 20 regions would represent blind spots due to poor enrichment.

Considering the presented data and our ability to identify problematic issues in the assay, we conclude that our enrichment process presents a good strategy for covering and sequencing IC-related genes, with an SNV sensitivity value of >96% and feasible values of specificity, PPV, and FDR. Additionally, the workflow evaluated is reliable, with a fast turnaround time for routine clinical screenings. Regarding the pipeline construction, CLC Genomics Workbench 6.5 exhibited an easy-to-use and flexible interface that allowed us to dispense any other kind of informatics effort in the main steps, presenting feasibility for clinical application. Coverage outliers seemed not to have influenced the whole coverage profile, but for future implementation (eg, copy number variant detection), these should be revisited.

Acknowledgment

We thank Fleury Group (São Paulo, Brazil) for development and research infrastructure.
Supplemental Data

Supplemental material for this article can be found at http://dx.doi.org/10.1016/j.jmoldx.2015.02.003.

References