Evaluation of Oligonucleotide Sequence Capture Arrays and Comparison of Next-Generation Sequencing Platforms for Use in Molecular Diagnostics

Nicole Hoppman-Chaney,1* Lisa M. Peterson,1 Eric W. Klee,2 Sumit Middha,2 Laura K. Courteau,1 and Matthew J. Ferber1

BACKGROUND: Next-generation DNA sequencing (NGS) techniques have the potential to revolutionize molecular diagnostics; however, a thorough evaluation of these technologies is necessary to ensure their performance meets or exceeds that of current clinical sequencing methods.

METHODS: We evaluated the NimbleGen Sequence Capture 385K Human Custom Arrays for enrichment of 22 genes. We sequenced each sample on both the Roche 454 Genome Sequencer FLX (GS-FLX) and the Illumina Genome Analyzer II (GAII) to compare platform performance.

RESULTS: Although the sequence capture method allowed us to rapidly develop a large number of sequencing assays, we encountered difficulty enriching G+C-rich regions. Although a high proportion of reads consistently mapped outside of the targeted regions, >80% of targeted bases for the GAII and >30% of bases for the GS-FLX were covered by a read depth of ≥20, and >90% of bases for the GAII and >80% of bases for the GS-FLX were covered by a read depth of ≥5. We observed discrepancies among sequence variants identified by the different platforms.

CONCLUSIONS: Although oligonucleotide arrays are quick and easy to develop, some problematic regions may evade capture, necessitating sequential redesigning for complete optimization. Neither sequencing technology was able to detect every variant identified by Sanger sequencing because of well-known drawbacks of the NGS technologies. The rapidly decreasing error rates and costs of these technologies, however, coupled with advancing bioinformatic capabilities, make them an attractive option for molecular diagnostics in the very near future.

* Address correspondence to this author at: 971 Hilton, 200 First St. S.W., Rochester, MN 55905. Fax 507-284-0043; e-mail hoppmanchaney.nicole@mayo.edu.

Received February 18, 2010; accepted May 26, 2010. Previously published online at DOI: 10.1373/clinchem.2010.145441

© 2010 American Association for Clinical Chemistry

1 Departments of Laboratory Medicine and Pathology, and 2 Health Science Research, Mayo Clinic, Rochester, MN.

1297–1306 (2010) Molecular Diagnostics and Genetics

Clinical Chemistry 56:8
We evaluated the ability of NimbleGen 385K Custom Sequence Capture Arrays (3, 4) to enrich patient samples for every exon in a panel of 22 genes, most of which are associated with hereditary colorectal cancer. Five patient samples previously tested for germline mutations in one of 4 colorectal cancer–related genes [MLH1,4 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli); MSH2, mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli); MSH6, mutS homolog 6 (E. coli); or APC, adenosomatous polyposis coli] were selected for library generation. Three samples were repeated on different arrays to study chip reproducibility. Captured sequence libraries were analyzed with both the GS-FLX and the GAII, and a subset of targeted exons was sequenced by the Sanger method to validate both the GS-FLX and the GAII, and a subset of targeted exons was sequenced by the Sanger method to validate the NGS-identified variants.

Materials and Methods

SAMPLE SELECTION AND DNA EXTRACTION
Deidentified clinical samples with documented germline mutations in genes associated with hereditary nonpolyposis colorectal cancer (HNPCC) or polyposis syndromes (MLH1, MSH2, MSH6, or APC) were selected for analysis. DNA was isolated from peripheral blood lymphocytes on the Autopure LS with Puregene chemistry (Qiagen; http://www.qiagen.com) according to the manufacturer’s protocol. In brief, DNA samples were brought to a final concentration of 250 ng/μL in 1× Tris-EDTA (10 mmol/L Tris-HCl, 1 mmol/L EDTA, pH 7.0).

LIBRARY PREPARATION AND HYBRIDIZATION
Twenty micrograms of DNA from each sample was sonicated on the Sonics Vibra-Cell™ sonicator (Sonics & Materials; http://www.sonicsandmaterials.com). Each sample was evaluated by agarose gel electrophoresis and with an Agilent Bioanalyzer (Agilent Technologies; http://www.agilent.com) to ensure that at least 5 μg of total DNA was available for hybridization.

QUANTITATIVE FLUORESCENCE PCR
Quantitative fluorescence PCR with the LightCycler® 480 Real-Time PCR System (Roche Applied Science; http://www.roche-applied-science.com) was performed on pre- and postenriched libraries to calculate the -fold enrichment at several loci. Exon-specific forward and reverse oligonucleotides (see Table 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol56/issue8), Universal ProbeLibrary probes, and LightCycler 480 Probes Master mix (Roche Applied Science) were used according to the manufacturer’s protocol. In addition, the housekeeping gene ACTB (actin, beta) was assessed as a negative selection control. At least 500-fold enrichment of the subset of targeted loci was required before proceeding with NGS.

ROCHE 454 SEQUENCING
Sequencing on the GS-FLX was performed off site by 454 Life Sciences with 5 μg of the enriched libraries. Generation of single-stranded template and emulsion PCR were performed according to the manufacturer’s protocol (5) with the standard sequencing kit (LR70), and 4 samples were applied to each picotiter plate.

ILLUMINA GAII SEQUENCING
Because of the short read length generated by the GAII (36-base read protocol), each library was digested with BamHI, which removed all but 7 bases of the adapter

4 Human genes: MLH1, mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli); MSH2, mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli); MSH6, mutS homolog 6 (E. coli); APC, adenomatous polyposis coli; AXIN2, axin 2; KRAS, v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog; PIK3CA, phosphoinositide-3-kinase, catalytic, alpha polypeptide; BRAF, v-raf murine sarcoma viral oncogene homolog B1.

1298 Clinical Chemistry 56:8 (2010)
sequences. Each library (3 µg) then underwent library preparation for GAII sequencing, with the manufacturer’s protocol slightly modified to start with a 500-bp gel-excised fragment instead of a 200-bp fragment so that NimbleGen-captured libraries could transition directly into GAII library preparation. We also eliminated the nebulization step, because our NimbleGen libraries were originally generated from sonicated genomic DNA. To each flow cell, 3.5 pmol of 1 sample was added, and cluster generation was performed according to the manufacturer’s protocol. Sequencing was performed for 36 cycles with the sequencing-by-synthesis Illumina Sequencing Kit v1.

VALIDATION OF MUTATIONS AND POLYMORPHISMS BY SANGER SEQUENCING

Primers were designed to amplify each exon, including 20–50 bp of flanking intronic sequence, of MLH1, MSH2, MSH6, and the coding region of the APC gene. All primers were 5′ tagged with universal primer sequences (sequences available upon request). Samples were prepared for fluorescence sequencing on the ABI 3730 DNA Analyzer with BigDye® Terminator v1.1 chemistry and the BigDye XTerminator™ Purification Kit (Applied Biosystems; http://www.appliedbiosystems.com). Sequence data were analyzed with Mutation Surveyor® v2.41 software (SoftGenetics; http://www.softgenetics.com).

DATA ANALYSIS

To calculate the -fold enrichment for each locus, we used the expression $E^{\Delta Cp}$, where $E$ is the efficiency of amplification and $Cp$ is the point at which the generated fluorescence signal rises above that of the background. $\Delta Cp$ was calculated by subtracting the $Cp$ of the captured library from that of the noncaptured library. We used an $E$ value of 2 (maximum PCR efficiency) for all calculations of -fold enrichment.

GS-FLX sequence reads were aligned to the human genome reference sequence (Build 36, Release 49) with the Roche runMapping v.2.0.1 tool (the runMapping script is part of the Roche 454 software package; 454 Life Sciences; http://www.454.com), with the -nrm flags and -reg option to restrict output to regions defined by the NimbleGen capture *.gff file. Run statistics were extracted from the NewblerMetrics.txt file, and the mean read length was computed by dividing the total number of bases sequenced by the total number of reads sequenced. Events were extracted from the HC-Diffs file, and the positional depth of coverage was obtained from the AlignmentInfo.tsv file. The Integrative Genomics Viewer version 1.3 (Broad Institute; http://www.broad.mit.edu/igv) was used with the custom track feature to visualize sequence depth of coverage and positional event mapping. The percent G+C track was computed by means of a moving average over a 1000-bp window.

GAII sequences were extracted from .tiff images with the Firecrest and Bustard tools from the Illumina Genome Analyzer Pipeline (version 1.0; http://www.illumina.com). We aligned these sequence reads to the human genome reference sequence (Build 36, Release 49) with MAQ version 0.7.1 (6) and discarded any reads that had multiple best-match alignments. The first 7 bases of each read were trimmed before the alignment to account for residual adapter sequences remaining after the restriction digest. Base substitutions were called by MAQ with a quality value threshold of 20. SOAP version 1.05 (7) was used on reads not mapped by MAQ to detect small insertions and deletions. Perl and shell scripts were used for converting formats, parsing files, and comparing results across sequencing platforms.

All statistical analyses were performed with R. The ANOVA for on-target read mapping used a linear model with sample as a factor and was computed across the 3-sample replicate pairs analyzed on both platforms. Biological variation was associated with the linear model $R^2$ value. Exon depth-of-coverage statistics were based on a positional depth average across the length of the exon. Interplatform Pearson product-moment correlations in exon depth of coverage were calculated with the 7 samples run in common on the 2 platforms. GC content was calculated with EMBOSS explorer’s geecee tool (http://emboss.sourceforge.net/apps/cvs/emboss/apps/geeece.html). Sensitivity and positive predictive values were computed by assuming all Sanger-detected events were true positives. Events were totaled across samples, with Sanger-identified events not detected by the GS-FLX or the GAII considered as false negatives. Sensitivity = true positives/ (true positives + false negatives); positive predictive value = true positives/(true positives + false positives).

Results

The samples selected for library preparation and sequencing on the NGS platforms were chosen to have different types of HNPCC-causing point mutations (samples 1–4), including point mutations, small insertions, and small deletions (Table 1). We chose samples with these particular types of mutations to evaluate the ability of each sequencing platform to identify different types of sequence variation. Duplicate libraries were prepared for all samples except sample 1 (no point mutations), for which there was insufficient DNA to prepare 2 libraries. Therefore, an additional sample with no HNPCC-related point mutations was included in the second set of library preparations and hybridizations (samples 5–8).
A custom NimbleGen 385K array was designed to target all exons, including 200 bp of the flanking upstream and downstream intronic sequence, for 22 genes, 21 of which are associated with colorectal cancer. These genes are distributed over a wide size range (6–106 exons). Probes are approximately 60–80 bp in length and tiled across all target regions except repetitive sequences, with significant redundancy between probes. Our custom array targeted 410 exons, incorporating just over 272 000 bases (see Tables 2 and 3 in the online Data Supplement).

To determine the degree of enrichment of our targeted loci before NGS, we calculated the -fold enrichment for each sample by quantitative fluorescence PCR with aliquots of pre-enriched and postenriched libraries. We observed enrichment levels that varied depending on the locus tested. Data from a positive-control locus included on every NimbleGen custom array indicated 917- to 5793-fold enrichment. Probes specific to a subset of our custom targeted loci also exhibited a variable -fold enrichment for 1 exon in different samples (3848- to 14 412-fold for MSH2 exon 15) as well as for different exons in the same sample (sample 2: 3293-fold for APC exon 1 and 11 347-fold for MSH2 exon 15). These results, although variable, indicated that target sequences from certain exons were consistently captured more efficiently than sequences from other exons (Table 2). Despite this variation, these data met minimum enrichment requirements and therefore indicated successful enrichment of at least a subset of the targeted exons.

From each captured sample library, we submitted 5 µg of sample for GS-FLX sequencing and 3 µg for GAII sequencing. As expected, a single run on the GAII produced approximately 80 times more reads per sample than the GS-FLX, and a high proportion of reads were successfully mapped to the human genome. We mapped 18.5%–62.3% of the GS-FLX reads and 21%–58% of the GAII reads to the target region. An ANOVA revealed that 67% of the GS-FLX variation and 77% of the GAII variation in on-target read percentage was attributable to the biological differences modeled (sample-to-sample differences). The remaining 33% (GS-FLX) and 23% (GAII) of the variation can be thought of as technical variation (i.e., between sample replicates) that arose during the experiment. The off-target reads may be explained by nonspecific sequence binding to the oligonucleotides on the sequence capture arrays. The total coverage of all targeted bases at 5x was 89% for the GS-FLX and 82.4% for the GAII; at 20x, the total coverage was 54.6% for the GS-FLX and 86.2% for the GAII (Table 3). The mean (SD) exon depth of coverage was 30 (21.7) for the GS-FLX, with an intrasample correlation coefficient of 0.71. The mean exon depth of coverage for the GAII was 266
with an intrasample correlation coefficient of 0.96. Despite the high mean exon read depth and target region coverage, 4 exons [AXIN2 (axin 2) exon 1, KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog) exon 1, and PIK3CA (phosphoinositide-3-kinase, catalytic, alpha polypeptide) exons 13 and 14] had no mapped reads. Another 6 exons had mean exon depths of coverage of <5 in all samples on both platforms.

The interplatform correlation coefficient in read depth of coverage across all sequenced exons was 0.14; however, sample pair–specific correlation coefficients for individual genes varied from −0.12 to 0.98, with a median value of 0.69. Variations in depth of coverage may arise from both the capture process and the sequencing methodologies and may reflect such factors as G+C content, target sequence uniqueness, overlap with repeat regions, and the presence of long homopolymers. As an example, Fig. 1A shows platform-specific coverage for BRAF (v-raf murine sarcoma viral oncogene homolog B1). The correlation between the GS-FLX and GAI1 platforms in mean exon read depth was 0.45 for one sample. A closer examination shows that exon 1 was poorly covered, with a mean coverage of <2 for both the GS-FLX and the GAI1. Low coverage in the first exon of genes was common. GS-FLX data had a mean depth of coverage of 9 in first exons and 32 in all other exons, GAII data had a mean depth of coverage of 42 in first exons and 280 in the other exons. High G+C content may explain the low coverage in the first exons of genes, given that the mean G+C content of exon 1 is 0.69, compared with 0.51 in the other exons. A high G+C content may hinder elution from the capture array or LMPCR amplification in later steps; both mechanisms could explain the absence of sequence data. G+C content does not account for all of the coverage variation in exons, however, as illustrated by PIK3CA exons 13 and 14. These exons have low G+C contents of 0.31 and 0.42, respectively, yet no mapped reads were obtained from either the GS-FLX or GAI1 platform (data not shown). The absence of mapped reads in these 2 exons may be due to the lack of target region uniqueness. Although probes were designed to capture sequences from these regions, both exons have 100% identical matches to off-target regions on chromosome 22, which caused the read-alignment programs to not map reads to these exons.

### Table 3. Sequencing run metrics.

<table>
<thead>
<tr>
<th>Sample</th>
<th>High-quality reads, n</th>
<th>Mapped to human genome, %</th>
<th>On target, %</th>
<th>Mean read length, bases*</th>
<th>Bases covered (5x), %</th>
<th>Bases covered (20x), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-FLX</td>
<td>1b</td>
<td>218 292b</td>
<td>96.6</td>
<td>18.5</td>
<td>240</td>
<td>93.0</td>
</tr>
<tr>
<td>2</td>
<td>51 912</td>
<td>88.9</td>
<td>49.2</td>
<td>209</td>
<td>82.3</td>
<td>33.8</td>
</tr>
<tr>
<td>3b</td>
<td>171 083b</td>
<td>96.5</td>
<td>63.2</td>
<td>243</td>
<td>95.3</td>
<td>88.0</td>
</tr>
<tr>
<td>4</td>
<td>53 787</td>
<td>88.1</td>
<td>50.7</td>
<td>202</td>
<td>84.6</td>
<td>34.5</td>
</tr>
<tr>
<td>5</td>
<td>79 459</td>
<td>93.9</td>
<td>49.8</td>
<td>220</td>
<td>92.5</td>
<td>63.1</td>
</tr>
<tr>
<td>6</td>
<td>58 225</td>
<td>93.2</td>
<td>47.5</td>
<td>220</td>
<td>85.9</td>
<td>40.3</td>
</tr>
<tr>
<td>7</td>
<td>81 880</td>
<td>93.0</td>
<td>54.0</td>
<td>218</td>
<td>90.9</td>
<td>62.3</td>
</tr>
<tr>
<td>8</td>
<td>80 541</td>
<td>92.5</td>
<td>41.5</td>
<td>215</td>
<td>88.0</td>
<td>46.9</td>
</tr>
<tr>
<td>GAI1</td>
<td>1</td>
<td>6 482 719</td>
<td>93.7</td>
<td>21.1</td>
<td>29</td>
<td>92.3</td>
</tr>
<tr>
<td>2</td>
<td>6 955 700</td>
<td>92.6</td>
<td>45.2</td>
<td>29</td>
<td>92.8</td>
<td>87.9</td>
</tr>
<tr>
<td>3</td>
<td>6 532 133</td>
<td>92.7</td>
<td>57.6</td>
<td>29</td>
<td>92.7</td>
<td>86.9</td>
</tr>
<tr>
<td>4</td>
<td>6 548 201</td>
<td>92.9</td>
<td>48.9</td>
<td>29</td>
<td>92.6</td>
<td>87.0</td>
</tr>
<tr>
<td>5</td>
<td>N/Ac</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>6 991 435</td>
<td>90.9</td>
<td>45.2</td>
<td>29</td>
<td>93.1</td>
<td>87.9</td>
</tr>
<tr>
<td>7</td>
<td>7 233 645</td>
<td>88.8</td>
<td>58.2</td>
<td>29</td>
<td>92.9</td>
<td>87.7</td>
</tr>
<tr>
<td>8</td>
<td>6 775 382</td>
<td>91.1</td>
<td>37.4</td>
<td>29</td>
<td>92.1</td>
<td>85.4</td>
</tr>
</tbody>
</table>

* *The mean read length on GAI1 is 36 − 7 = 29. Seven bases of residual adapter sequence were trimmed prior to data analysis.

b Samples 1 and 3 were run 3 times each on the GS-FLX because of insufficient reads. Data shown are the sums of all 3 runs.

c N/A, not applicable.
These examples illustrate the importance of reviewing multiple factors when determining the likelihood of sufficient on-target read coverage in an NGS capture experiment.

The 454 and the GAII identified a large number of sequence variants within our targeted regions with 100% concordance, even between technical replicates (Fig. 1). This result could be due to a number of variables, including efficiency of capture, capture of homologous sequences that cause false-positive findings upon alignment, or artifacts of the sequencing methods themselves. To assess the validity of variants identified by NGS, we sequenced 4 well-characterized genes (MLH1, MSH2, MSH6, and APC) in each sample by the Sanger method. Variants detected by the GS-FLX and GAII methods were then compared with the Sanger sequence variants. If we assume all variants detected by Sanger sequencing are true events, the GS-FLX had a 65% sensitivity and a 62% positive predictive value. We did not calculate specificity because it would be incorrect to assume Sanger sequencing detects all true events; thus, we could not quantify the true-negative rate.

False-negative variants detected by the GS-FLX were often in regions containing poor sequence coverage (insufficient reads) and homopolymers. This fact led to failure of the GS-FLX to identify a germline mutation in MSH2 known to cause Lynch syndrome (IVS5 +3A>T, samples 2 and 6) (Table 4). Poor capture, inferred from the low read depth at this position, could have contributed to the absence of detection because this mutation was also missed by the GAII in one of the replicates; however, the presence of the homopolymer likely contributed to this result. The remaining 2 germline mutations were identified by GS-FLX sequencing. In another instance, a polymorphism was not called as a high-quality difference because only reverse reads aligned to the reference sequence at that position.
The GAII correctly identified 2 of the 3 germline mutations but, as mentioned above, failed to identify 1 germline mutation in *MSH2* owing to poor read coverage. The GAII was also unable to detect several polymorphisms correctly identified by the GS-FLX and the Sanger method. Most of these variants were not reported in our GAII data because of sequences with low quality scores at that position or poor read coverage. We also considered that nearby *Bam*HI restriction sites might have hindered the detection of polymorphisms; however, no such sites were identified near the undetected variants (data not shown). We likewise investigated variants that were found by Sanger sequencing but were not identified by either the GS-FLX or the GAII. Insufficient read depth, in both platforms, was the primary reason these variants were not detected. This result almost certainly reflects the oligonucleotide array capture and not the sequencing platforms (Table 4).

Finally, several variants were identified by NGS technologies and not by the Sanger method; however, these variants all shared several characteristics in common, suggesting they are false-positive mutation calls. First, all of these variants were identified by the GS-FLX and not the GAII. Second, with 1 exception in a homopolymer region, they are not known germline mutations or polymorphisms. Third, none were confirmed in the corresponding replicate sample, and, fourth, most were found at low frequencies, although the frequency of 1 variant did approach 36% in homopolymer regions of suitable read depth (Table 5). These observations led us to hypothesize that these variants were likely false-positive calls. The possibility of low-level somatic mosaicism cannot be completely ruled out; however, the lack of reproducibility between replicates and the failure of the GAII to detect these variants suggest false-positive calls as a more likely explanation. To be inclusive, we used a 10% minimum variant read threshold for event detection. A low threshold can lead to increased false-positive detection and effect performance metrics. If a 30% minimum variant read threshold had been required to detect heterozygous alleles, only 1 false-positive call, attributable to the presence of a homopolymer, would have been made. This requirement would have reduced the total true positives detected by 3 but would have led to an overall improvement in GS-FLX performance compared with Sanger sequencing.

**Discussion**

The NGS technologies, as demonstrated by numerous research laboratories and publications, have tremendous laboratory benefits because of their output capabilities in terms of bases sequenced per run; however, because >99% of a patient’s genome currently is not targeted

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene</th>
<th>Position</th>
<th>Reference base</th>
<th>Modified base</th>
<th>Detected by</th>
<th>db SNP</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>MSH6</td>
<td>IVS4–101</td>
<td>G</td>
<td>C</td>
<td>GS-FLX</td>
<td>rs2072447</td>
<td>Present; low quality score&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>8</td>
<td>MSH6</td>
<td>IVS2 + 52</td>
<td>T</td>
<td>A</td>
<td>GS-FLX</td>
<td>rs3136282</td>
<td>Present; low quality score&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>2</td>
<td>MSH2</td>
<td>IVS5 + 3</td>
<td>A</td>
<td>T</td>
<td>GAII</td>
<td>Germline mutation</td>
<td>Homopolymer</td>
</tr>
<tr>
<td>2</td>
<td>MSH2</td>
<td>IVS9–91</td>
<td>G</td>
<td>T</td>
<td>GAII</td>
<td>rs3732182</td>
<td>Homopolymer</td>
</tr>
<tr>
<td>7</td>
<td>MSH2</td>
<td>IVS10 + 12</td>
<td>G</td>
<td>A</td>
<td>GAII</td>
<td>rs3732183</td>
<td>Reverse reads only&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>7</td>
<td>MSH2</td>
<td>IVS9–91</td>
<td>G</td>
<td>T</td>
<td>GAII</td>
<td>rs3732182</td>
<td>Homopolymer</td>
</tr>
<tr>
<td>8</td>
<td>MSH2</td>
<td>IVS1 + 9</td>
<td>C</td>
<td>G</td>
<td>GAII</td>
<td>rs2303426</td>
<td>Low depth of coverage (2)</td>
</tr>
<tr>
<td>8</td>
<td>MSH2</td>
<td>IVS1 + 9</td>
<td>C</td>
<td>G</td>
<td>GAII</td>
<td>rs2303426</td>
<td>Low depth of coverage (3)</td>
</tr>
<tr>
<td>1</td>
<td>MSH6</td>
<td>116</td>
<td>G</td>
<td>A</td>
<td>Neither</td>
<td>rs1042821</td>
<td>Low read depth</td>
</tr>
<tr>
<td>4</td>
<td>MSH6</td>
<td>IVS1 + 22</td>
<td>C</td>
<td>G</td>
<td>Neither</td>
<td>rs55927047</td>
<td>Few/0 reads</td>
</tr>
<tr>
<td>4</td>
<td>MSH6</td>
<td>186</td>
<td>C</td>
<td>A</td>
<td>Neither</td>
<td>rs1042820</td>
<td>Few/0 reads</td>
</tr>
<tr>
<td>6</td>
<td>MSH2</td>
<td>IVS5 + 3</td>
<td>A</td>
<td>T</td>
<td>Neither</td>
<td>Germline mutation</td>
<td>Homopolymer; low GAII quality score&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>8</td>
<td>MSH6</td>
<td>IVS1 + 22</td>
<td>C</td>
<td>G</td>
<td>Neither</td>
<td>rs55927047</td>
<td>Few/0 reads</td>
</tr>
<tr>
<td>8</td>
<td>MSH6</td>
<td>186</td>
<td>C</td>
<td>A</td>
<td>Neither</td>
<td>rs1042820</td>
<td>Few/0 reads</td>
</tr>
</tbody>
</table>

<sup>a</sup> db SNP, single-nucleotide polymorphism database.
<sup>b</sup> Was present in multiple reads, but the Q_phred quality score was <20.
<sup>c</sup> Only reads from the reverse direction (not the forward direction) aligned at this position; hence, this variant was identified but as a low-quality variant reported in alldiffs and not HCdiffs.

Table 4. Sanger sequencing–validated variants missed by one or both NGS methods.
by most clinical laboratory tests, ways of isolating only
the desired material are necessary for the use of such
methods.

PCR amplification, our current method of choice,
has drawbacks in that many reactions are required to
sequence an entire gene. Likewise, it is well known that
sequence variants located beneath annealed primers
can cause allele dropout. The oligonucleotide capture
arrays are advantageous because they allow the simul-
taneous development of a large number of assays in a
matter of weeks, compared with the months to possibly
years required to design and develop the assays with
traditional PCR amplicon–based methods. In addi-
tion, this technology does not rely on the use of primers
complementary to genomic sequences, thereby elimi-
ating the potential for polymorphisms and rare vari-
ants to cause allele dropout.

The oligonucleotide arrays were unable to capture
certain exons, however, particularly those high in G+C
content and/or those corresponding to repetitive ele-
ments. New elution protocols that may aid in the cov-
erage of such regions have been released since these
experiments were performed; however, postelution
amplification steps may still present an opportunity for
G+C bias. Although we observed a reproducible cap-
ture profile across multiple samples for most exons,

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene</th>
<th>Position</th>
<th>Reference base</th>
<th>Modified base</th>
<th>Identified by</th>
<th>Reads, n</th>
<th>Percent with variant</th>
<th>db SNP,* comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MLH1</td>
<td>IVS16–12</td>
<td>T</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>32</td>
<td>25.0</td>
<td>T&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>2</td>
<td>APC</td>
<td>IVS3+9_+15</td>
<td>TAAAAG</td>
<td>AAA</td>
<td>GS-FLX</td>
<td>14</td>
<td>35.7</td>
<td>(T&lt;sub&gt;A&lt;/sub&gt;G&lt;sub&gt;C&lt;/sub&gt;), no SNP</td>
</tr>
<tr>
<td>2</td>
<td>APC</td>
<td>5068</td>
<td>GTAGG</td>
<td>INS</td>
<td>GS-FLX</td>
<td>38</td>
<td>10.5</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>2</td>
<td>APC</td>
<td>7063</td>
<td>T</td>
<td>A</td>
<td>GS-FLX</td>
<td>27</td>
<td>11.1</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>2</td>
<td>MLH1</td>
<td>2111–2112</td>
<td>TG</td>
<td>GAT</td>
<td>GS-FLX</td>
<td>20</td>
<td>15.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>2</td>
<td>MSH6</td>
<td>2469</td>
<td>TAA</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>32</td>
<td>12.5</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>3</td>
<td>APC</td>
<td>IVS3+9_+10</td>
<td>TA</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>37</td>
<td>13.5</td>
<td>T&lt;sub&gt;P&lt;/sub&gt;A&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>3</td>
<td>APC</td>
<td>5234</td>
<td>A</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>101</td>
<td>12.9</td>
<td>A&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>3</td>
<td>MLH1</td>
<td>588</td>
<td>A</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>86</td>
<td>12.8</td>
<td>A&lt;sub&gt;P&lt;/sub&gt;, SNP</td>
</tr>
<tr>
<td>3</td>
<td>MLH1</td>
<td>2019</td>
<td>T</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>82</td>
<td>11.0</td>
<td>T&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>3</td>
<td>MSH6</td>
<td>404</td>
<td>T</td>
<td>INS</td>
<td>GS-FLX</td>
<td>70</td>
<td>15.7</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>3</td>
<td>MSH6</td>
<td>3828</td>
<td>A</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>100</td>
<td>10.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>4</td>
<td>APC</td>
<td>2982</td>
<td>T</td>
<td>C</td>
<td>GS-FLX</td>
<td>49</td>
<td>10.2</td>
<td>A&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>4</td>
<td>APC</td>
<td>6847</td>
<td>A</td>
<td>G</td>
<td>GS-FLX</td>
<td>28</td>
<td>14.3</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>4</td>
<td>APC</td>
<td>6903</td>
<td>ATCA</td>
<td>GATCC</td>
<td>GS-FLX</td>
<td>27</td>
<td>11.1</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>4</td>
<td>MSH6</td>
<td>2443</td>
<td>C</td>
<td>G</td>
<td>GS-FLX</td>
<td>39</td>
<td>10.3</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>6</td>
<td>APC</td>
<td>331</td>
<td>A</td>
<td>G</td>
<td>GS-FLX</td>
<td>27</td>
<td>11.1</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>6</td>
<td>APC</td>
<td>4279</td>
<td>CC</td>
<td>A</td>
<td>GS-FLX</td>
<td>28</td>
<td>10.7</td>
<td>C&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>7</td>
<td>APC</td>
<td>8042</td>
<td>C</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>50</td>
<td>14.0</td>
<td>C&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>7</td>
<td>APC</td>
<td>8108</td>
<td>A</td>
<td>G</td>
<td>GS-FLX</td>
<td>45</td>
<td>11.1</td>
<td>A&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
<tr>
<td>7</td>
<td>APC</td>
<td>8341</td>
<td>G</td>
<td>A</td>
<td>GS-FLX</td>
<td>33</td>
<td>12.1</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>7</td>
<td>MSH2</td>
<td>1546</td>
<td>AG</td>
<td>DEL</td>
<td>GS-FLX</td>
<td>19</td>
<td>15.8</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>APC</td>
<td>2524</td>
<td>GAT</td>
<td>TGGATCC</td>
<td>GS-FLX</td>
<td>37</td>
<td>10.8</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>APC</td>
<td>3652</td>
<td>A</td>
<td>T</td>
<td>GS-FLX</td>
<td>34</td>
<td>11.8</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>APC</td>
<td>3719</td>
<td>GT</td>
<td>TTC</td>
<td>GS-FLX</td>
<td>30</td>
<td>10.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>APC</td>
<td>7032</td>
<td>AC</td>
<td>GAA</td>
<td>GS-FLX</td>
<td>43</td>
<td>14.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>MSH2</td>
<td>2587</td>
<td>T</td>
<td>A</td>
<td>GS-FLX</td>
<td>25</td>
<td>12.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>MSH6</td>
<td>2202</td>
<td>G</td>
<td>C</td>
<td>GS-FLX</td>
<td>30</td>
<td>10.0</td>
<td>No homopolymer, no SNP</td>
</tr>
<tr>
<td>8</td>
<td>MSH6</td>
<td>2542</td>
<td>A</td>
<td>TT</td>
<td>GS-FLX</td>
<td>25</td>
<td>12.0</td>
<td>A&lt;sub&gt;P&lt;/sub&gt;, no SNP</td>
</tr>
</tbody>
</table>

*a db SNP, single-nucleotide polymorphism database; DEL, deletion; INS, insertion.*
some displayed high variation in capture efficiencies, even between technical replicates of the same sample. In a clinical laboratory, such variation may necessitate the use of backup methods, such as PCR amplification followed by traditional Sanger sequencing, for regions that evade capture.

Likewise, reliance on hybridization technologies also leads to the capture of homologization and/or repetitive sequences, such as pseudogenes. Alignment of these nontarget sequences may lead to the identification of artificial mutations (false positives). Traditional PCR is still the method of choice for dealing with pseudogenes because primers can be designed to take advantage of sequence differences and to amplify only the functional copy of interest. Improved design algorithms that have the potential to improve capture efficiency and specificity are available; we are currently testing such arrays and how their results compare with those generated by this first generation of sequence capture arrays. Likewise, customized design assistance is also available in cases of difficult genomic regions.

Aside from array-based technologies, several capture methods have become commercially available within the past few years. Among these methods are Agilent’s SureSelect in-solution hybrid capture method (8), RainDance Technologies’ microdroplet PCR-based enrichment (9), and Febit’s microfluidics-based HybSelect technology (10). The Agilent and Febit methodologies for sequence capture are similar to NimbleGen arrays in that they rely on hybrid selection, whereas the RainDance methodology is a highly multiplexed PCR reaction. All of these technologies could potentially suffer from one or more of the drawbacks observed with hybrid chip selection; however, all are likely to evolve over time and may transition from useful research tools to clinically useful techniques as they become cheaper and further refined.

With respect to the NGS technologies we evaluated, each has a unique chemistry and therefore different strengths and weaknesses. We confirmed a well-known drawback of the GS-FLX platform with respect to a difficulty in and around homopolymer sequences, but as the output of GS-FLX increases, which has already been observed with the Titanium technology, many read depth–related errors should be corrected. The difficulty with sequencing homopolymers is not likely to be resolved in the near future because of the chemistry used in this process; however, improvements to the mutation detection software in these regions could help.

In general, the GAII overcame poor read depth merely because of the higher number of reads generated by the instrument; however, this platform was not without its own limitations. The short lengths of the reads generated by the GAII makes the detection of insertions and deletions difficult because only a few mismatches can be tolerated when aligning to the reference sequence. With the rapidly increasing read lengths and output, however, along with improved bioinformatics methods for short-read data, these limitations will become less of an issue.

In conclusion, we have demonstrated that oligonucleotide arrays for reducing genomic complexity are successful at enriching multiple genomic loci simultaneously but that some exons resist capture. The GS-FLX and GAII platforms have differences in their ability to detect certain types of variants, which must be taken into account; however, as the cost and error rates of these technologies continue to decrease while the output continues to increase, we foresee their implementation for clinical diagnostics in the near future. More experience with all of the platforms is necessary, however, to address the shortcomings we have identified.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures of Potential Conflicts of Interest: Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: M.J. Ferber, Mayo Clinic.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played a direct role in the final approval of the manuscript.

Acknowledgments: We acknowledge Bruce Eckloff and Yan Asmann for assistance with GAII sequencing and data analysis, respectively. We are also grateful to Luke Dannenberg and Dan Burgess at Roche/NimbleGen for their help with the sequence capture experiments and for a critical reading of the manuscript. Finally, we thank the Mayo Clinic’s Research Computing Facility for their computing support.

References


