Feasibility of Low-Throughput Next Generation Sequencing for Germline DNA Screening

Nur Sabrina Sapari,1 Eiram Elahi,1 Mengchu Wu,1 Marie Loh,2 Hong Kiat Ng,1 Xiao Han,1 Hui Ling Yap,3 Thomas P. Klemm,4 Brendan Pang,5 Touati Benoukraf,1 Yik Ying Teo,6 Barry Jacopetta,2 Soo Chin Lee,3 and Richie Soong1,5*

BACKGROUND: Next generation sequencing (NGS) promises many benefits for clinical diagnostics. However, current barriers to its adoption include suboptimal amenability for low clinical throughputs and uncertainty over data accuracy and analytical procedures. We assessed the feasibility and performance of low-throughput NGS for detecting germline mutations for Lynch syndrome (LS).

METHODS: Sequencing depth, time, and cost of 6 formats on the MiSeq and Personal Genome Machine platforms at 1–12 samples/run were calculated. Analytical performance was assessed from 3 runs of 3 DNA samples annotated for 7500 nucleotides by BeadChip arrays. The clinical performance of low-throughput NGS and 9 analytical processes were assessed through blinded analysis of DNA samples from 12 LS cases confirmed by Sanger sequencing, and 3 control cases.

RESULTS: The feasibility analysis revealed different formats were optimal at different throughputs. Detection was reproducible for 2619/2635 (99.39%) replicate variants, and sensitivity and specificity to array annotation were 99.42% and 99.99% respectively. Eleven of 16 inconsistently detected variants could be specifically identified by having allele frequencies ≤0.15, strand biases >−35, or genotype quality scores ≤80. Positive selection for variants in the Human Genome Mutation Database (colorectal cancer, nonpolyposis) and variants with ≤5% frequency in the Asian population gave the best clinical performance (92% sensitivity, 67% specificity).

CONCLUSIONS: Low-throughput NGS can be a cost-efficient and reliable approach for screening germline variants; however, its clinical utility is subject to the quality of annotation of clinically relevant variants.

Next generation sequencing (NGS),8 entailing massively parallel sequencing of DNA, has emerged as a potentially useful alternative to current diagnostic platforms for DNA analysis, such as Sanger sequencing and real-time PCR (1, 2). The simplicity and scalability of NGS have been demonstrated for various applications, including genetics (3), cancer (4, 5), infectious diseases (6), and forensic science (7). However, these demonstrations have been performed primarily on high-throughput systems such as the Roche 454 Genome Sequencer FLX, Illumina Genome Analyzer Ix, Illumina HiSeq 2000/2500, and Applied Biosystems Solid 5500/5500xl, which typically generate 1–1000 Gb of data per run. Assuming an output of 200 Gb and given that average gene panels target approximately 200 kb per sample, the output of such sequencing platforms implies runs of 10000 samples when aiming for an mean depth of 100 reads. This throughput is inconsistent with that of clinical diagnostic laboratories; hence many questions remain about the feasibility and reliability of NGS in the routine setting.

Other issues in the application of NGS to clinical diagnostics include the procedures used to assess sample quality and the accuracy of sequencing results. Currently, Sanger sequencing and real-time PCR rely on sample assessment and the inclusion of controls before analysis. The identification of DNA variants with these methods is often based on the positive selection of variants from established databases and/or manual selec-
tion based on previous experience. The data-rich output of NGS provides many sample-specific metrics for assessing the run and sample quality and for objectively selecting the variants of interest. These include sequencing depth and coverage, allele frequency, strand bias, and combinations thereof. Moreover, bioinformatics approaches allow prediction of variant function that can help determine the significance of unrecognized variants. Despite this potential, the optimal procedures for routine processing of NGS data in the setting of clinical diagnostics remain to be established.

Lynch syndrome (LS) is an inherited cancer syndrome characterized by a very high incidence of colorectal cancer at an early age (8). It is caused by germline mutations in genes coding for DNA mismatch repair enzymes, most commonly mutL homolog 1 (MLH1), mutS homolog 2 (MSH2), mutS homolog 6 (MSH6), and postmeiotic segregation increased 2 (PMS2) (9, 10). Because of the relatively large size of these genes (0.4 Mb in total) and the widespread distribution of mutations, current approaches for mutation screening are both tedious and costly. This can involve weeks of stepwise Sanger sequencing at specialist centers (11) and can cost upwards of US$2750. The need for sequencing is relatively infrequent, however, because the prevalence of LS among colorectal cancer patients is approximately 1% (12–14). For large cancer centers that see around 1000 new cases of colorectal cancer per year, this means that only 10 cases of LS can be expected from among the approximately 25 to 40 cases that will require germline testing.

Recent developments in sequencing technologies have led to the introduction of new “low-throughput” systems and formats, including the Nano, Micro, and Standard flow cell formats for the MiSeq and the 314, 316, and 318 chips for the Ion Torrent Personal Genome Machine (15). These formats have been developed for outputs between 0.3 and 5.0 Gb that are more compatible with routine throughputs in clinical diagnostics; however, their efficiency and amenability to routine testing in the clinical setting remains untested. Here, we assess the cost and time efficiency, assay reliability, and accuracy of low-throughput NGS using screening for germline mutations associated with LS as a relevant model.

Materials and Methods

DNA SAMPLES

DNA samples from 3 volunteers of Chinese, Malay, and Indian ethnicity previously characterized for DNA variants using Omni2.5 and Exome BeadChip arrays (16) were obtained from the School of Public Health, National University of Singapore. Fifteen DNA samples from blood analyzed previously (17) were obtained from the cancer genetics clinic at the National University Cancer Institute of Singapore. Twelve of these were from LS individuals with germline variants in MLH1 and MSH2 confirmed by Sanger sequencing (17, 18). The remaining 3 samples were from individuals without cancer. All samples were obtained according to a protocol approved by the National Healthcare Group Domain Specific Review Board.

TARGETED NGS

DNA was prepared using the TruSight Cancer Enrichment Kit (Illumina) and TruSight Cancer Content (Illumina) and sequenced on the MiSeq System (Illumina) according to the manufacturer’s instructions. The protocol enables sequencing of 255 kb DNA from 94 genes including MLH1, MSH2, MSH6, and PMS2, as well as 284 single nucleotide polymorphisms using 4000 80-mer probes designed from the human NCBI2137/hg19 reference genome. Briefly, 50 ng DNA was fragmented and tagged before sequencing adapters and indices were appended by PCR. Sample libraries were denatured into single-stranded DNA and pooled into groups of 3 samples. Pooled samples were hybridized to biotin-labeled probes specific to the targeted region and biotinylated DNA fragments were then enriched from the solution using streptavidin beads. Enriched DNA fragments were eluted from the beads, hybridized, and enriched a second time before undergoing PCR amplification. After amplification, the library underwent cluster generation and 2 × 150-bp paired-end sequencing using the MiSeq Reagent Nano Kit V2 (Illumina) on the MiSeq System (Illumina).

DATA ANALYSIS

A summary of the analytical processing used in this study is shown in Fig. 1. Sequencing data from the Miseq were processed using Miseq Reporter software version 2.2.29 (Illumina) and the preset Enrichment work flow. This aligns reads against the whole genome reference and performs variant analysis using Somatic Variant Caller version 2.1.12 for regions specified in the TruSight Cancer manifest file (Illumina). Additional sequence quality analysis was performed using Avadis NGS software version 1.4.5 (Strand Life Sciences). Cross-matching of variants with the dbSNP (v137), 1000 Ge-

---

* Human genes: MLH1, mutL homolog 1; MSH2, mutS homolog 2; MSH6, mutS homolog 6; PMS2, postmeiotic segregation increased 2; BRCA1, breast cancer 1, early onset; BRCA2, breast cancer 2, early onset; RB1, retinoblastoma 1; TP53, tumor protein p53; VHL, von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase.
nomes (March 2013 v3), COSMIC (v64), and Human Gene Mutation Databases (HGMD) were performed using VariantStudio version 1.0 software (Illumina). Variants called in MLH1, MSH2, MSH6, and PMS2 genes were cross-referenced to the International Society for Gastrointestinal Hereditary Tumors Incorporated (InSiGHT) variant database (http://www.insight-group.org/variants/database/). Functional prediction analysis by SIFT, Polyphen2 HDIV, Polyphen2 HVAR, LRT, and MutationTaster was performed using the dbNSFP database (19).

**Results**

**FEASIBILITY ANALYSIS**

As the first step in feasibility analysis, we calculated expected sequencing depths of the TruSight Cancer Content (255 kb DNA) and Ion AmpliSeq Inherited Disease Panel v2 (315 kb DNA) when analyzed on the Nano, Micro, and Standard flow cell and 314, 316, and 318 chip formats, respectively. Calculations were based on commercially reported sequencing outputs for the formats at 1–12 samples per run and on the nucleotide
target volume of the respective kits (see Table 1 in the Data Supplement that accompanies the online version of this report at http://www.clinchem.org/content/vol60/issue12). For germline DNA analysis, we estimated that a mean 100-fold depth of sequencing is required for each nucleotide. Seven sequencing reads are required to call a variant with confidence (20), equating to 14 reads for a heterozygous state. Given that the minimum read depth is approximately 7-fold less than the mean read depth, a mean depth of 98 (or 100) reads is therefore desirable to achieve a minimum read depth of 14. Our analysis showed that Standard and Micro formats could provide more than 100 reads at all throughputs (Fig. 2A). The 316, Nano, and 318 formats could achieve a depth of 100 reads at the maximum throughputs of 2, 3, and 4 samples, respectively. A 100-fold depth could not be achieved by the 314 format at any throughput.

Turnaround time from the receipt of a DNA sample to generation of the report was assessed for each format on the basis of a single library preparation and a single sequencing step. These estimations included the time for manual preparation and instrument running. To adjust for increased sample size, an additional 10 min for library preparation and 30 min for data analysis were added for each sample. The Personal Genome Machine formats were faster than MiSeq formats, requiring as little as 1.2 days compared to 4.4 days for the latter (Fig. 2B). The turnaround time for the Standard format was longer than for other MiSeq formats, primarily due to shorter sequencing run times for the Nano and Micro formats.

The cost analysis was based on reagent costs (prorated per sample), staff labor, and instrument amortization and maintenance (Fig. 2C). As expected, costs decreased with increased sample throughput due to the cost efficiency of running multiple samples. A lack of proportionality for MiSeq analysis was observed, due to the use of TruSight Enrichment kits with 1, 2, 4, and 24 indices. Formats with lower sequencing outputs were the cheapest. However, at certain throughputs these formats were not valid, due to inadequate sequencing depth (Fig. 2A). At the maximum throughputs for the 316, 318, Nano, Micro, and Standard formats (2, 4, 3, 12, and 12 samples/run, respectively), the costs per sample in Singapore dollars (SGD) were 1356, 1074, 1105, 455, and 488, respectively.

**TECHNICAL PERFORMANCE**

**MSH6** and **PMS2** are not included in the Ion AmpliSeq Inherited Disease Panel v2. Hence the MiSeq Nano format with an optimal throughput of 3 samples/run was chosen for further investigation. DNA samples from 3 healthy individuals were analyzed together in 3 separate runs. These were from volunteers of Chinese, Malay, and Indian ethnicity and had been previously characterized for DNA variants by Omni2.5 and Exome BeadChip array analysis (16). Across the 9 analyses, the mean (±CV) amount of total sequencing reads, of aligned reads, and of sequencing depth reads was, respectively, 870053 (21%), 566473 (21%) (65% of total reads), and 182 (21%). Only 3 of 1737 (0.17%) target sequence amplicons were covered by <20 reads [rhomboid 5 homolog 2 (Drosophila) (RHBDF2), chr17:74473261–74473381, 15 reads; RHBDF2, chr17:74467714–74468135, 16 reads; runt-related transcription factor 1 (RUNX1), chr21:36164431–36164908, 18 reads]. The number of variants detected in each replicate for sample 1 was 292, 292, and 295; for sample 2 it was 297, 298, 299; and for sample 3 it was 289, 285, and 288. This amounted to 2635 individual replicate variants and 431 unique variants after common variants between the 3 individuals were considered. A total of 2619/2635 (99.39%) replicate variants were consistently detected in all 3 replicates. Of the 16 replicate variants that were not consistently detected, 6
were observed in 2 of 3 replicates and 10 in 1 of 3 replicates.

We assessed various sequencing characteristics in an attempt to find the basis for poor-quality variants (i.e., those that were inconsistently detected). Variant allele frequency, strand bias, genotype quality, and sequencing depth were all identified as potential factors. Using the criteria of allele frequency $\leq 0.15$, strand bias $\leq -35$, and genotype quality score of $>80$, we determined that 11 of 16 (69%) poor quality variants could be filtered out without eliminating any replicate variants that were consistently detected (Fig. 3; also see online Supplemental Table 2). The remaining 5 variants had distinctly lower sequencing depths ($\leq 51$ reads) compared to other replicates, but application of this criterion ($\leq 51$ reads) would have disqualified 96/2635 (3.6%) high-quality variants.

We next assessed the concordance for variant detection between the NGS and array platforms. According to the manufacturer’s specifications, 2504 nucleotides were common to both platforms. Twelve nucleotides were noninformative in the array analysis, leaving 7500 informative nucleotide interrogations that could be assessed for concordance in the 3 samples. Of these, 695 (9.26%) were variant in both NGS and array analysis, 6800 (90.66%) were wild type in both, 1 (0.01%) was variant in NGS but wild type in array analysis, and 4 (0.05%) were wild type in NGS but variant in array analysis. Using array detection of variants as the benchmark, NGS therefore had a sensitivity of 99.42% (695/699), specificity of 99.99% (6800/6801), positive predictive value (PPV) of 99.86% (695/696) and negative predictive value (NPV) of 99.94% (6800/6804).

**CLINICAL PERFORMANCE**

We tested the ability of the NGS assay format to identify 12 cases of LS found earlier by Sanger sequencing to have germline mutations in *MLH1* and *MSH2*. These contained 14 reported variants, with 2 cases each having 2 reported variants. Three samples from noncancer individuals aged 53, 59, and 65 years were included as negative controls. An essential component of this exercise was that numerous data analysis processes and filters were devised and compared for performance (Fig. 1).

Table 1 provides a summary of the classification of cases and variants according to NGS analysis and Sanger reporting, with individual data shown in online Supplemental Table 3. In the absence of filtering, NGS detected all cases and variants reported by Sanger sequencing analysis, but the results lacked specificity.
This was improved by filtering out variants with >5% frequency in the Asian population, without loss of sensitivity for all approaches. Additional improvements in specificity and PPV without compromising sensitivity were obtained by using positive selection of variants in the InSiGHT [hereditary nonpolyposis colorectal cancer (HNPCC)] and HGMD (colorectal cancer, nonpolyposis) databases. This gave superior results compared to bioinformatics prediction algorithms, for which sensitivity, specificity, PPV, and NPV were at most 42%, 33%, 71%, and 13%, respectively. HGMD performed better than InSiGHT. Three variants reported by Sanger sequencing were identified using the HGMD but not InSiGHT database (MLH1 ins 12 bp codon 292; MLH1 380 + 2T>A; MSH2 G816term), and 1 was identified using the InSiGHT but not the HGMD database (MLH1 616delK) (see online Supplemental Table 3).

Discussion

In this study, we systematically evaluated sequencing depths, turnaround times, and running costs for the analysis of 1–12 samples/run on all 6 currently available low-throughput NGS formats. Our results showed no single format was optimal for all throughputs, and numerous factors needed to be considered in identifying optimal formats (Fig. 2). Running costs for the analysis of 2–4 samples/run ranged between SGD$1074 and SGD$1356, and results could be obtained within 1.2–4.4 days. This compares favorably to the current costs of SGD$1600 and turnaround time of 6 weeks for screening MLH1, MSH2, and MSH6 by Sanger sequencing at external specialist centers. Sequencing of PMS2 costs an additional SGD$2000. Further cost efficiency can be achieved by using multigene panels such as those investigated in this study to consolidate requests for tests such as breast cancer 1, early onset (BRCA1), breast cancer 2, early onset (BRCA2), retinoblastoma 1 (RB1), tumor protein p53 (TP53), and von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase (VHL).

To assess reliability, we examined the reproducibility of TruSight Cancer Content analysis in 3 MiSeq Nano runs of 3 samples/run. Overall, 99.39% of 2635 replicate variants were detected in all 3 replicates at a sensitivity of 99.42%, specificity of 99.99%, PPV of 99.86%, and NPV) of 99.94% from 7500 shared interrogations between NGS and array analysis. These frequencies are
similar to reproducibility rates of 95.6%–98.8% (21, 22), sensitivities of 95.9%–99.4% (21, 23, 24), and specificity of 100% (21) reported in studies using higher-throughput NGS formats, indicating an equivalence in analytical performance between modes.

Notably, 11/16 (69%) replicate variants that were detected inconsistently had allele frequencies ≤0.15, strand biases >−35, or genotype quality scores ≤80 (Fig. 3; also see online Supplemental Table 2), suggesting these metrics may be useful for identifying poor-quality variants. The remaining 5 variants detected inconsistently had mean sequencing depths ≤51. However, 3.6% of consistently detected variants were also below this read threshold, suggesting that sequencing depth may better serve as a quality indicator rather than filter. Others have identified allele frequencies of ≤0.15 (25) and ≤0.25 (5, 26) to be useful quality metrics in studies using different NGS formats and target sequencing depths. Given that quality metrics are a key component of diagnostic protocols (27), further data on indicators for monitoring NGS quality in diagnostic settings are eagerly awaited.

To assess clinical performance, we simulated an analysis of 12 LS cases and 3 controls at a throughput of 3 samples per run. We also compared 9 different analytical approaches that are being considered for diagnostic reporting, including 2 methods [InSiGHT (HNPPC) and HGMD (CRC, nonpolyposis)] in Table 1 which simulated standard filtering procedures currently used in Sanger sequencing assessment. The study of Asian participants allowed the issue of population diversity to be addressed. Previous reports have established that NGS has a high sensitivity (98%–100%) for detecting germline DNA variants in numerous clinical conditions such as cancer, heart disease, and polycystic kidney disease (21, 24, 25, 28–30). However, these studies used high-throughput NGS formats on predominantly white populations for which genetic variability and clinical databases are better established. Specificity and analytical processing have also not been a focus of these studies.

Similar to earlier reports, we found NGS could detect all variants identified by Sanger sequencing of LS cases (Table 1). However, this approach lacked specificity, with a mean of 284 (4253 total variants/15 participants) additional variants detected in each case. Excluding germline polymorphisms by removing variants with >5% allele frequency in the Asian population markedly improved PPV. Specifically, the MSH6 G39E and PMS2 P470S variants were excluded, leading to the correct identification of 2 of 3 controls. The filter also correctly eliminated PMS2 R20Q and PMS2 P470S when using InSiGHT and HGMD databases. However, the filter did not exclude MLH1 R217C or MLH1 790 + 1G>A, thus contributing to the observed imperfect specificity and PPV rates.

Positive selection of variants from the InSiGHT and HGMD databases was found to be a useful strategy for improving specificity, PPV, and NPV, while retaining high sensitivity (Table 1). Better sensitivity was obtained using the HGMD database than using InSiGHT, which could be attributed to differences in their curation. HGMD is based on published reports of absence in healthy controls, segregation in families, multiple occurrence in affected individuals, functional data, disruption of protein domains, and loss of conservation (31). InSiGHT has been based on public contributions, although an effort was made recently to determine the clinical relevance of variants in this database (32). Given the difference was only 3 variants (MLH1 ins 12 bp codon 292; MLH1 380 + 2T>A; MSH2 Q816X) detected using HGMD and 1 (MLH1 616delK) using InSiGHT, the superiority could also be specific to our series. Interestingly, the combined use of the 2 databases leads to 100% sensitivity without a reduction in specificity or PPV, thus representing another analytical approach worthy of further consideration.

Compared to the use of InSiGHT and HGMD databases, positive selection of variants using bioinformatics predictions performed poorly, even after filtering out those present at >5% in Asians and restricting to LS genes. Several reasons could underlie this performance, including inadequacy of algorithms, combinatorial effects of variants, lack of direct genotype-to-phenotype relationships, and the evolutionary dynamics of variant pathogenicity. Our results caution against the sole use of bioinformatics for predicting the clinical relevance of variants until better modelling of the factors involved in pathogenesis can be achieved.

Perfect sensitivity, specificity, PPV, and NPV could have been achieved by selecting for variants in HGMD and with ≥5% frequency in Asians if not for 3 variants. One variant, MLH1 618delE, was not in HGMD but was annotated in a single entry in InSiGHT. A second variant, MLH1 R217C, was identified in a control using HGMD, thus reducing specificity to 67% (2 of 3). This variant was also identified in an LS case (no. 19). This case also had the third variant considered to be falsely identified (MLH1 790 + 1G>A). MLH1 R217C is annotated in HGMD as a variant on the basis of functional and phenotypic evidence (33), and 5/41 entries in InSiGHT also annotated this variant as being present in LS individuals. MLH1 790 + 1G>A is an established functional variant (34, 35) and is in HGMD on the basis of functional and phenotypic evidence. The benchmark for performance in this study were variants that had previously been reported from Sanger sequencing analysis. Hence, the MLH1...
R217C and/or MLH1 790 + 1G>A variants detected in case 19 by NGS may actually have been present and clinically relevant, but were not previously detected or selected for reporting on the basis of Sanger sequencing. Unfortunately, information from the original report and additional DNA sample for this case were unavailable for further testing. The uncertainty surrounding such variants is typical of the challenges faced in establishing their clinical significance. However, it is important to recognize these challenges are common to both Sanger sequencing and NGS.

Pseudogenes can also be a source of inaccuracy in Sanger sequencing and NGS, as they can be concurrently sequenced and compromise variant detection. The probe design of the TruSight Cancer Content suggests that the assay does not discriminate PMS2 from its 15 pseudogenes (36), and this would affect results interpretation. Previous studies have attempted to enrich for PMS2 using careful probe design (25, 37, 38); however, the high sequence homology has made absolute enrichment elusive. This emphasizes the need for careful probe design for accurate diagnosis, whether by NGS, Sanger sequencing, or alternative methods.

In conclusion, this study found that low-throughput NGS can be cost and time efficient with high analytical reproducibility and potentially high accuracy for clinical samples. We have determined the optimal format for analysis and provided quality assessment of identified variants. Comparison of analytical processing strategies revealed that the use of variant databases may be more reliable than bioinformatics predictions, highlighting the critical importance of well-curated databases with clinical relevance. The rapidly diminishing costs of NGS (39) and the recent release of new desktop and “factory-scale” NGS platforms (40) mean that such variant databases can now be quickly expanded. However, their standardization and consolidation looms as a major challenge. Although our results indicate that germline DNA screening by NGS in low- to midvolume testing centers is feasible, this capability must be partnered with considerable expertise to conduct responsible analysis. This includes specialized reporting, informatics, genetic counselling, and regulatory and legal staff, as well as dedicated computing hardware and protocols. Nonetheless, the development of such infrastructure promises cheaper, faster, and more accurate health outcomes in the near future.

References


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 or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: X. Han, Cancer Science Institute; T.P. Klemm, Illumina.
Consultant or Advisory Role: T.P. Klemm, Illumina.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: R. Soong, Illumina.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: The authors thank Ana Carrera, Mei Ling Chong, Chee Seng Ku, Arjuna Kumarasuriyah, Chee Koon Lee, Kim Foong Lee, Gretchen Weightman and Kenneth Yew for their contributions to this project.