Clinical and Technical Aspects of Genomic Diagnostics for Precision Oncology

Yuri Sheikine, Frank C. Kuo, and Neal I. Lindeman

ABSTRACT

The emergence of precision medicine has been predicated on significant recent advances in diagnostic technology, particularly the advent of next-generation sequencing (NGS). Although the chemical technology underlying NGS is complex, and the computational biology expertise required to build systems to facilely interpret the results is highly specialized, the variables involved in designing and deploying a genomic testing program for cancer can be readily understood and applied by understanding several basic considerations. In this review, we present key strategic decisions required to optimize a genomic testing program and summarize the technical aspects of different technologies that render those methods more or less suitable for different types of programs.

INTRODUCTION

The advent of precision medicine has been fueled by advances in diagnostic modalities, largely based in molecular diagnostics, the pathology specialty focusing on analysis of nucleic acids. Over the past two decades, we have witnessed the introduction of quantitative polymerase chain reaction (PCR), pyrosequencing, microarrays, digital droplet PCR, and other technologies into the clinical diagnostic arena. Of these technical innovations, however, the largest contributions to precision medicine have come from massively parallel, or next-generation, sequencing (NGS).

NGS enables the generation of genome-scale sequencing information in relatively rapid time frames, such that it can impact clinical decision making. The fundamental principle of NGS is spatial separation of individual DNA molecules, which allows simultaneous analyses of millions of individual molecules. As each nucleotide in the sequences of each of the DNA strands is individually analyzed, the data are recorded and compiled computationally. The compiled data enable concurrent analysis of multiple genes from multiple samples. This is a disruptive technology in the true sense of the word—it is both more sensitive and more specific, faster and more efficient, compatible with more and smaller samples, and generates more data, with greater precision, at a cost that is rapidly declining.

DISRUPTIVE TECHNOLOGY BRINGS NEW CHALLENGES

Technical innovation has led to an explosive growth of biomedical knowledge, which has challenged the traditional framework for establishing standard of care. At the analytical level, established procedures for validating clinical assays are predicated on the one test, one analyte model, which is ill suited for massively multiplexed NGS assays. At the clinical level, demonstration of utility is based on evidence from large, randomized clinical trials. Because specific genetic variations are typically present in small subgroups of patients within each cancer type, traditional clinical trials on the basis of genetic characteristics can be challenging. A current model in academic centers is to perform initial broad-based genomic testing and direct each patient to an investigational agent appropriate for their genomic findings. The combination of these novel validation concepts led to the situation where most laboratories offering cancer genomic testing are unsure about the regulatory requirements for establishing assays and collecting reimbursement for testing. At the same time, public demand is increasing rapidly, fueled by episodic reports of dramatic responses, and laboratories adopting NGS-based testing are faced with several challenges: which patients and targets to test, and which methods to use?

At a first pass, there are clinical scenarios for which genomic markers have established need, such as *BCR-ABL1* rearrangement in chronic myeloid leukemia.[1] These applications have been established in the medical literature and incorporated into practice guidelines.[2]

A challenge arises for markers with a lesser level of established need: mutations that have been shown to respond to investigational therapeutics or that have shown clinical response in a different cancer type. For example, IDH1/2 inhibitors have shown great promise in early-phase clinical trials in patients with *IDH*-mutant acute myeloid leukemia[3] but have not yet gained Food and Drug Administration approval or inclusion in guidelines. Similarly, BRAF inhibitors are standard therapy in patients with *BRAF*-mutant melanoma[4] but not for *BRAF*-mutant Langerhans cell histiocytosis. Philadelphia-like acute lymphoblastic leukemia is a distinct subgroup of acute lymphoblastic leukemia with poor prognosis, defined by alterations in multiple kinase genes, some of which can be targeted by kinase inhibitors.[5] Total mutational burden is predictive of response to immunotherapy in some tumor types.[6] Should a laboratory test for these alterations? Although targeting these alterations in certain cancers may not have been investigated in randomized clinical studies, it may represent the only chance for patients who fail established therapies and are in need of experimental targeted therapies. A recent prospective study speaks to the value of this approach.[7]

WHAT SHOULD WE SEQUENCE?

The decision of which genes to test gets at the heart of each institution's mission. Those decisions are essential for selecting the appropriate test to offer and can be boiled down to answering three questions: how much of the genome should be tested, which types of alterations should be tested, and what are the needs for throughput and turnaround time?

Regarding the first question, the temptation is to sequence the whole genome. The wet laboratory process for sequencing the genome is simpler than sequencing specific targets, and for alterations involving regulatory regions and introns, such as translocations, this is the only way to be comprehensive.[8] However, analysis of an entire genome consumes sufficient sequencing capacity that the depth of coverage—the number of individual strands of DNA analyzed at any given position—is limited. When the coverage drops too low, sequencing is less sensitive and less accurate, which poses a significant problem for heterogeneous samples such as cancers. Another challenge presented by whole-genome analysis is the sheer amount of data generated and the time it takes to process these data. This limits the throughput, because analyzing a single sample can take weeks,[9] which is impractical for a patient with advanced-stage cancer requiring prompt treatment. Whole-genome sequencing of a cancer sample also requires concurrent sequencing of a paired normal (germline) sample from the same patient, to determine which alterations are present only in the cancer sample.[10]

At the other end of the spectrum are targeted sequencing panels that interrogate scores to low hundreds of genes. For oncogenes, often only specific mutational hotspots are sequenced, whereas for tumor suppressor genes, entire coding sequences are usually interrogated.[11] A well-designed clinical panel typically includes genes for which approved targeted therapies exist, genes that have therapeutics in clinical trials, and genes with diagnostic or prognostic value. Some panels also include genes implicated in cancer biology, which are currently not targetable with available agents or have potential therapeutics in early stages of development. Selected introns can also be targeted to detect specific rearrangements. In general, smaller panels afford deeper coverage and, therefore, superior sensitivity to characterize highly heterogeneous samples. In addition, analysis of such panels is faster and, accordingly, they are commonly used as clinical assays where results are needed within a few days. Using information gleaned from databases containing genomic sequences from thousands of normal individuals and pools of normal samples, cancer-specific somatic mutations can be distinguished from germline variations to the extent that allows testing cancer samples without paired normal samples, particularly for the well-studied oncogenic alterations. This simplifies laboratory operations and reduces cost and turnaround time.

Between these lies whole-exome sequencing, which involves sequencing the entire coding region of a sample but excluding most of the intronic sequences, thus rendering these methods generally unable to detect structural rearrangements.[8] The amount of data is still too large to rapidly analyze, however, and whole-exome tests often use a postanalytic computational filter to mask irrelevant genes and narrow the scope of genes analyzed to a manageable subset, particularly when turnaround time is important. Moreover, whole-exome sequencing of cancer samples, like whole-genome sequencing, requires a paired germline sample from each patient. Thus, the whole-exome analysis can resemble a larger targeted cancer panel, except that the coverage depth is lower (so the false-positive and false-negative rates are higher), and all of the genes have actually been sequenced and can be analyzed at a later time.

The second question concerns the type of alterations that need to be assayed. The simplest type of alteration, single nucleotide variant (SNV), is fairly straightforward and readily measured by NGS. Small insertions and deletions (indels) < 20 bp are more challenging, however, because NGS reads relatively short stretches of DNA (typically < 150 bp) and relies on software to align partially overlapping sequences to a reference genome. The presence of indels, especially near the end of a read, causes misalignment as software tries to find the best match by adding gaps. Larger indels (> 30 bp) are not well detected by the bioinformatics tools that are typically used in clinical applications, and the sensitivity of different computational pipelines can vary considerably.[12] The inability to reliably detect large indels can pose particular challenges for several clinically significant alterations, such as *FLT3* internal tandem duplication in acute myeloid leukemia,[13] *CALR* exon 9 deletion in myeloproliferative neoplasms,[14] and *KIT* exon 11 deletion in GI stromal tumors.[15]

Detection of copy number variations (CNVs) is also challenging with NGS, especially in cancer samples. Cancer samples can have multiple CNVs, ranging from small focal changes involving part of a gene to arm-level or chromosome gains and losses, and they often contain a combination of all types of CNVs. Heterogeneity within the sample is an important confounding factor that makes CNV detection difficult. Many laboratories rely on the visual assessment of data by a trained laboratory professional, although accuracy and reproducibility of this approach have been questioned. Other laboratories use computational algorithms that incorporate SNV data and other variables to estimate CNVs,[16] but these have been developed to work mostly with whole-exome sequencing. Ultimately, NGS on a tumor sample uses a pool of DNA from a population of cells, and the final readout of CNVs is based on the average of the pool and therefore is unable to assess CNVs in minor subpopulations.

Detection of structural variants (SVs), such as translocations, inversions, and large deletions or duplications, relies on the detection of breakpoints, where two sequences from noncontiguous genomic coordinates are joined together in the sample. Because the breakpoints are more frequently found in introns than exons, small targeted NGS panels and whole-exome sequencing do not generally cover them, and whole-genome sequencing provides the best sensitivity for detecting SVs.[17] Although introns can be added to targeted panels, the large size of introns necessitates a substantial investment in sequencing capacity. In essence, sequencing more introns means sequencing fewer exons. Some SVs fuse two genes and create a chimeric transcript encoding a fusion protein with oncogenic activities. RNA-based techniques remain the gold standard for detection of fusion transcripts, with the challenges inherent in the limited stability and variable expression of RNA.

The third question for laboratories is logistic: how quickly are the results needed, and how many samples need to be tested? The need for speed is difficult to establish clearly. A 2013 practice guideline from the College of American Pathologists, Association for Molecular Pathology, and International Association for the Study of Lung Cancer regarding molecular diagnostic testing in lung adenocarcinoma recommended a 2-week turnaround time for tests for *EGFR* and *ALK*.[18] This was an expert consensus opinion, however, and not an evidence-based recommendation. Still, it is a reasonable approximation of an appropriate time frame for most clinical cases.

The answers to the three questions above will dictate optimal assay design for the kind of tests an institution will offer. From there, essentially two decisions remain: which type of library preparation method to use and which platform to use for sequencing.

HOW SHOULD SEQUENCING BE DONE?

The library is the full set of DNA strands made from each sample. The library preparation starts with the DNA molecules isolated from the sample and creates a replica of these DNA molecules with attached adapters and barcodes. The barcodes are used to identify the sample from which they originated, and the adapters contain primer recognition sequences to initiate sequencing reactions. The main two library preparation methods in use today are ligation based and amplification based (Table 1).[19]

Ligation-based library preparation first fragments DNA through physical shearing or enzymatic cleavage. The adapters and barcodes are then added to the ends of the DNA fragments by template-independent ligation. Because the fragmentation and ligation processes are both sequence independent, theoretically, the entire genomic sequence is represented in the library. A library made this way can be used in whole-genome sequencing as is. For whole-exome or targeted sequencing, additional steps are needed to select only those fragments containing the sequence of interest.

The most commonly used method for selection today is the so-called hybrid capture method. Hybrid capture methods rely on a pool of many individual nucleic acid probes (baits), each of which is designed to be complementary to a sequence of interest within the genome. A multiplexed hybridization reaction is performed, in which the baits bind to their targets and are captured on a solid surface, such as a bead or chip. The unbound DNA from the library is washed away, and the captured fragments, which are enriched for the sequences of interest, are then minimally PCR amplified and sequenced. Hybridization reaction is relatively robust with regard to variations in conditions and number of concurrent reactions and is the method of choice when the number of targets is large or the intended target is the whole exome. Moreover, this method uses limited PCR amplification and better preserves the relative amounts of the different regions of the genome and is thus more consistent for copy number evaluation (compared with amplicon-based library preparation). However, even under best-case scenarios, about 30% to 50% of the DNA fragments in the sequenced library can be from off-target regions, leading to lower depth of coverage for the targeted regions. Also, this method involves multiple steps, most significantly a long hybridization reaction, which leads to longer turnaround times.

By contrast, amplicon-based library preparation uses PCR reactions to amplify regions for sequencing, which can be done with unique primers to each targeted region. Because of the amplification of targets earlier in the process, this method provides few off-target sequences and greater depth of coverage for the targeted regions and is the method of choice for scanty and heterogeneous samples. It is also a faster method, useful in certain clinical situations, such as acute

Table 1. Library Preparation Methods

Assay Characteristic	Hybrid Capture Based	Amplicon Based
Input DNA requirement	As low as 50 ng	As low as 10 ng
Maximum number of genes per run	Thousands	Hundreds
Hotspot SNV detection	Yes, but limited for subpopulations	Yes
Any SNV detection	Yes	Yes
Indel detection	Yes	May be suboptimal near the end of amplicons
CNV detection	Yes, but maybe affected by GC content	Yes, but may be subject to amplification bias
SV detection	Yes	No
Library preparation time	Days	Hours

Abbreviations: CNV, copy number variation; GC, proportion of nucleotides that are either guanine (G) or cytosine (C); Indel, insertion/deletion; SNV, single nucleotide variation; SV, structural variant.

Table 2. Sequencing Methodologies

Assay Characteristic	Emulsion Semiconductor	Bridge Dye Terminator	Nanopore
Sequencing time	Hours	1-2 days	Hours
Ability to handle high case volume	Moderate	High	High
Cost	Low	Medium	High
Accuracy	Moderate	High	Low

NOTE. Sequencing technologies undergo continuous improvement and development, which increases accuracy and productivity. Information in this table is only a current estimate of the capabilities of different technologies.

leukemia profiling. However, unlike hybridization capture, where the size of the bait set has little bearing on the outcome, competition and interference among primers in multiplexed PCR reactions can lead to unequal amplifications among individual targets. Thus, the amplicon-based approach is limited to panels with a relatively small number of targets and is less suitable for copy number assessment. In addition, errors that arise in early stages of PCR can be exponentially amplified and give rise to false-positive findings (so called jackpot errors), or, alternatively, low-level variants may not participate in early rounds of PCR and be diluted out and give rise to false-negative findings.

The final decision for each laboratory concerns the selection of a sequencing platform (Table 2). Without expressing a preference for any one specific manufacturer, the commonly used platforms are emulsion PCR with semiconductor sequencing, bridge amplification with reversible dye terminator sequencing, and nanopore sequencing.[20] These methods differ primarily in the manner in which they spatially separate the library DNA strands for sequencing and the technology used to generate the sequence.

Emulsion PCR involves hybridization of the DNA library to beads bearing capture sequences complementary to the library adapter sequences in a dilute molar ratio, such that each bead captures either zero or one template DNA strand, following Poisson distribution. Beads are then partitioned into aqueous droplets and stabilized by immersion in oil. Each droplet contains all the components for a PCR reaction, to enable independent amplifications of the singly bound DNAs on individual beads. Each bead-in-droplet becomes a separate reaction chamber, as amplicons generated each round are captured by the bead and serve as templates for the ensuing rounds of PCR. After PCR, the beads containing the amplicons are separated in space (usually in wells fabricated in silicon chips) and sequenced in parallel using a modified pyrosequencing reaction with electrochemical detection of incorporated nucleotides. This method is relatively quick and inexpensive, rendering it especially useful for amplicon-based methods. The method generates medium-length sequences, which is helpful for correct mapping of indels and SVs. However, the method is limited to a smaller number of targets compared with other methods and is subject to particular challenges in properly differentiating sequences with single-nucleotide repeats (homopolymers). In general, this is a method of choice for small targeted clinical panels.

Bridge amplification involves capturing the DNA library on a solid surface with attached probes complementary to the adapter sequences in the library. PCR amplification is then performed directly on the surface, and PCR products are captured by additional adapter recognition sequences, such that the surface becomes seeded with polymerized colonies of amplicons, each derived from a single DNA strand. The sequencing reaction then is also performed on the solid surface, using a dye-labeled deoxynucleotide that is protected from extension at the 3′-hydroxyl site, such that only one nucleotide can be added at a time until the protection is removed. This way, homopolymers can be sequenced unambiguously, but the sequences are shorter, which is more challenging for identifying indels and SVs. This method is most suitable for assays with a large number of targets and is more accurate than the pyrosequencing, but it is also relatively slow and costly.

Nanopore sequencing involves directing individual DNA strands through a small channel that can accommodate only one strand of DNA at a time. Within the channel, either a dye-based chemical sequencing reaction takes place or electrical conductivity is measured as each nucleotide traverses the pore. This technology is the most costly and least well developed at this time and is fairly error prone, which renders it less suitable for SNVs. However, it generates long reads, ideal for SVs, and with no PCR steps in library preparation it is theoretically better in assessment of CNVs. Currently, this method is largely used for investigational applications.

In conclusion, although technology will continue to evolve, and the current cutting edge will one day seem limited, NGS is well poised to support the development of precision oncology in the coming years. The myriad technical choices available to a laboratory today can be reduced to a limited number of options once a few simple variables are understood. Is the goal quick and efficient clinical management, selection for investigational trials, or large-scale basic research? Is the need a subset of well-characterized activating mutations, or should the analysis include copy number changes and rearrangements? Understanding those variables and, thereby, selecting appropriate techniques does not rely on detailed and complex understanding of arcane molecular biology but rather on traditional medical judgment and an understanding of the clinical and scientific mission.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at jco.org.

REFERENCE:

JOURNAL OF CLINICAL ONCOLOGY .Published at jco.org on February 2017 ,13.