NGS bioinformatics pipeline—worries and wish lists: A look at the preanalytic, analytic, and postanalytic phases

William Check, PhD

October 2014—Last month the molecular and genomic pathology laboratory of the University of Pittsburgh Medical Center posted on the AMP listserv its requirements for a bioinformatics scientist to support next-generation sequencing for clinical testing. The requirements consisted of, but were not limited to:

- PhD in bioinformatics, computational biology, computer science, or biology with significant computational experiences; MS with the right combination of background and experiences also considered.
- Basic knowledge about molecular biology and genomics.
- Proficiency in Python and Linux/Unix/Mac environment.
- Experience in analyzing NGS sequencing data strongly preferred.
- Familiarity with commonly used databases and bioinformatics tools for NGS data analysis.

It would be hard to imagine a better illustration than this posting to highlight the importance of bioinformatics to the successful execution of NGS in a clinical setting and the need for trained and experienced bioinformaticists to support clinical NGS.

Marina N. Nikiforova, MD, associate professor and director of the molecular and genomic pathology lab at UPMC, explained the stringent requirements in an email response to a CAP TODAY inquiry. “While technical issues of NGS can be handled by trained technologists, interpretation of NGS data involves highly specialized knowledge in both bioinformatics and biology. Therefore,” she wrote, “it is crucial in every NGS-based laboratory to have a bioinformatician on staff.”

Based on the experience in her laboratory, she added, having such a person is essential to building NGS pipelines and providing routine help with data interpretation and maintaining quality assurance in the NGS area.

These and other insights into bioinformatics for clinical NGS were the focus of a session at last year’s Association for Molecular Pathology annual meeting. Franklin R. Cockerill III, MD, of Mayo Clinic, who moderated, called it “intuitive” to have a session on bioinformatics for NGS at that time, which wouldn’t have been the case a few years earlier. The pace at which complex genomic analysis has entered the laboratory has outrun expectations.

One of the speakers, Federico A. Monzon, MD, then of Baylor College of Medicine and now medical director of oncology for Invitae, said, “A tsunami of genomic information is coming to us [pathologists]” from NGS. “It has taken me by surprise how fast it came to the clinical laboratory.”

Dr. Carter

Most health care institutions don't develop their own laboratory software, another speaker, Alexis B. Carter, MD, of Emory University School of Medicine, said in a recent interview. "Because of the human resources needed to develop, test, and maintain software, most institutions prefer to purchase vendor-developed and vendor-supported software, but the analysis of NGS still requires support from trained people." In her AMP talk, Dr. Carter, director of pathology informatics in the Department of Pathology and Laboratory Medicine and the Department of Biomedical Informatics, defined informatics as a science at the intersection of information, technology, and people.

"There are many kinds of informatics," she tells CAP TODAY, "and NGS analysis involves both bioinformatics—information science at the molecular biology level—and clinical informatics—information science used in health care. A well-known informaticist said that informatics in general is 80 percent sociology, 10 percent information, and 10 percent technology." This means, she says, that people are what you have to study to do informatics well.

"Managing people and implementing systems that enable humans to accurately and efficiently use computer systems to acquire, analyze, manage, and store information to improve patient care are central to informatics."

Dr. Routbort

Mark Routbort, MD, PhD, of the University of Texas MD Anderson Cancer Center, another presenter, made a similar point in a recent interview. Three years ago, he said, his group's typical approach to a new method was to put in a vendor system and validate it and create simple reports. "However, this approach would not scale to the complex data coming from NGS. We needed to annotate that data with clinical significance." Dr. Routbort, director of computational and integrational pathology in the Division of Pathology and Laboratory Medicine, says the number of observable findings of known and unknown significance in NGS platforms "demands a quantum leap in terms of managing information."

To handle this complexity, his department has one specialized bioinformatics person, and he himself has expertise in bioinformatics. "Basically, I got hooked on it when the laboratory was initially setting up clinical NGS and asked for my input," he says. "It was quite illuminating—this is an area where pathology and informatics converge in a visceral way. To efficiently perform and report NGS testing for clinical and molecular diagnostics, you need a solid grounding in bioinformatics. You don't have to program the pipelines, and I don't. But I do program some downstream annotation and interpretation toolsets."

Given the complexity of NGS and the need for expert bioinformaticists and a laboratory director who has a basic grasp of the software, is it reasonable for most laboratories to think about bringing in this technology?

"That depends a lot on your focus and what type of testing you are going to be doing," Dr. Monzon tells CAP TODAY. Laboratories performing one or a few panels need bioinformatics resources but perhaps not full-time people. "However, you do need somebody in your department fully cognizant of the nuances of the process." On the other end, laboratories doing many types of NGS need a larger informatics group.

There are already many off-the-shelf tools, he says, but users of those tools need to understand the complexity of the process and the algorithms that go into those tools. "When things go wrong or don't work the way you expect, you need to know how to troubleshoot."

At the AMP session, NGS bioinformatics was divided into the preanalytic, analytic, and postanalytic phases.

Dr. Carter defined the preanalytic phase as the right patient, the right test order, the right specimen, accessioning,

aliquotting, and getting orders to instruments.

Most institutions use traditional methods of identifying the patient to whom a specimen belongs. However, some have suggested recently that DNA identification could be used to help the laboratory make sure it has the right patient. It is not rapid, with a turnaround time of 85 to 90 minutes, but its advantage over other biometrics is that it can be used to verify specimen identity when doing molecular testing. But “electronic health records have nowhere to put this data currently,” Dr. Carter said.

Getting the right order faces similar problems. In most EHRs, clinical decision support in computerized provider order-entry systems is not adequate. There is typically no place for informed genetic consent or genetic counseling, for example, or for avoiding duplicate ordering of germline genetic tests. Nor do CPOE systems help a provider choose the most appropriate test from complex tests like FISH, karyotyping, molecular, or an NGS panel.

What about getting orders to instruments? Dr. Carter introduced a survey she had taken among institutions with molecular information systems. There were 83 respondents, mostly academic and commercial reference laboratories. One question was: When your laboratory receives a specimen and logs it into your LIS, how do patient information and the test order get to the instrument that performs nucleic acid extraction? In 56 percent of responding laboratories, information is handwritten on a paper worksheet and manually typed into the instrument. In 26 percent, the barcode label on the specimen is scanned into the instrument. In 11 percent, the barcode label on a paper worksheet is scanned into the instrument, and in eight percent the LIS automatically sends the patient’s information and test order to the instrument.

“How we manage information in molecular labs right now is not great,” Dr. Carter said. “We have computer systems that we could use to tailor workflow, yet the vast majority of labs are still walking pieces of paper around. People are carrying flash drives from an instrument to the LIS or from one instrument to another for NGS analysis.”

Answers to another question—how are you recording nucleic acid quantities when measured?—revealed a similar pattern. For 58 percent of respondents, the choice selected was “We write the information down on a paper worksheet.” Only 23 percent said, “We record it in our LIS in a specific field meant for that purpose.”

“Core labs use automation lines and HL7 interfaces to transport information between pieces of equipment involved in the analysis of the specimen. In the clinical NGS lab, we are just starting to move in that direction but it is really slow getting there,” Dr. Carter says. Because of the low volume relative to chemistry or the core lab, many molecular laboratorians have not been pushing for higher automation and support for HL7 interfaces from their instrument or LIS vendors, she says.

“I’m not aware of any molecular/genomic instruments that have a true real-time HL7 interface to the LIS. Given the complexity of the technology we are using, manual transcription of data between instruments and the LIS creates a real and sustained risk of error.”

In the future, Dr. Carter would like to see positive patient identification at the biometric level for all laboratory specimens prior to analysis and reporting, real-time HL7 interfaces to communicate molecular data between instruments and the LIS, electronic orders, no manual entry, and robust clinical decision support.

Information security, too, is critical, not only for patient privacy, but for another pressing reason: Under the federal HIPAA final security rule and HITECH, unauthorized disclosure, loss, or theft of protected health information can be prosecuted. Security breaches now require mandatory reporting by the institution or provider. Institutions with breaches involving more than 500 patients are now listed on what Dr. Carter calls the HHS “Wall of Shame” (http://j.mp/breachnotificationrule). The problem is that, by law, health data privacy is the responsibility of the provider and the institution and not the vendor who sold the LIS or instrument software, Dr. Carter points out. “This gets really interesting because some of the security requirements call for the software to be built with certain features. If those features are absent, the institution or provider cannot get them added unless the vendor agrees to incorporate them. If the vendor refuses, you could be stuck.”

For example, part of the HIPAA final security rule requires that the software have an audit trail so that the laboratory knows which users have looked at or manipulated a patient's data in any electronic system containing these data. This includes instrument sofware, ancillary programs, and the LIS. "For any data since 2006, any patient can walk into any health care site and ask to see who has looked at their data. If the data are electronic, you have to be able to give them this information. Without an audit trail, you can't."

In the survey, 55 percent of institutions said that at some time they had needed to know which employees added, deleted, or even just viewed a specific patient record. "Partner with your vendor to make sure you are getting what you need and that the software you are purchasing meets all of the federal requirements for health data security and privacy," Dr. Carter advises.

With the massive amounts of data generated by sequencing, people are starting to look to cloud storage as a solution, and some cloud storage vendors are advertising themselves as HIPAA-compliant. Dr. Carter cautioned that three categories of security are required to make storage of patient health information HIPAA-compliant: administrative safeguards (policies and procedures), physical safeguards (safeguarding the hardware), and technical safeguards (what will you build into your software to keep unauthorized people from getting in? who has looked at which patient records?).

"Amazon Cloud started advertising itself as HIPAA-compliant," and she has heard that some have started putting NGS data on the Amazon Cloud. "But Amazon only had their servers set up so that there were physical safeguards on the hardware," she says. "To some extent, access to hardware via remote service met HIPAA rules." But technical safeguards also require unique user identification, passwords, an audit trail, and permissions to ensure people can access only the part of the software they need to get their jobs done (so-called minimum necessary rule). "Cloud storage systems that advertise as being compliant with HIPAA may be compliant with only some of the requirements. Laboratories should verify that all security requirements are met before placing health data on the cloud."

Dr. Routbort began his AMP talk with the impact of the tissue sample on analysis and interpretation in cancer, for which tissue is often limited. Combined with tumor purity and heterogeneity and consideration of allelic frequency, these factors must be considered in determining optimal read depth.

Dr. Routbort's take-home messages were, first, "Tissue qualification is key." You can't control the biopsy or eliminate contamination by normal tissue, such as vascular stromal cells and inflammatory cells. "So somatic assays have to be capable of making variant calls at far lower than the ideal 100 percent tumor/50 percent heterozygous mutation level," he said.

Defining standards for assay performance, such as the required depth of sequencing coverage, is not simple. "There is no platform that is magically better," he tells CAP TODAY. "All machines have a certain amount of room on the chip." Competing for this space are the number of samples per chip, how many genes you want to investigate, and the desired depth of coverage. "Specifying an absolute minimal depth of coverage is like saying, What is the minimum size of a bone marrow biopsy for a patient with Hodgkin's lymphoma," he says. "You may in some cases be able to identify reliable and actionable evidence of disease in a very tiny bone marrow sample, or at a relatively low coverage depth for NGS. That being said, we aim for a minimum of 250 reads per nucleotide."

While the basic principles of NGS are the same for cancer and medical genetics, one big difference in application is that genetic samples are much closer to 50 percent allelic frequency. So depth of coverage doesn't need to be as great. And in genetics, affected family members' DNA can be used in many cases to find linkage between variants and a clinical condition.

In the transition from Sanger sequencing to NGS, several key differences emerge. For instance, Sanger data can be viewed directly as a set of peak height traces. With NGS, there is no directly interpretable analog signal. Output cannot be visualized until you get a highly processed file. "The informatics pipeline is essentially part of the lab test," Dr. Routbort said. Despite the rapid influx of NGS, "Sanger is still considered the gold standard."

Sequencers provide a raw signal for each nucleotide position in the nucleic acid analyzed. In the first steps in the informatics pipeline, signal processing and base calling, those raw data are converted into an actual sequence of nucleotides in the form of a fastq file, which contains both sequence and quality information.

Alignment algorithms map each individual read against a reference genome for "best fit." The output of alignment algorithms is a bam (binary alignment map) file, which is viewable in genomic viewing software, such as the Integrative Genomic Viewer.

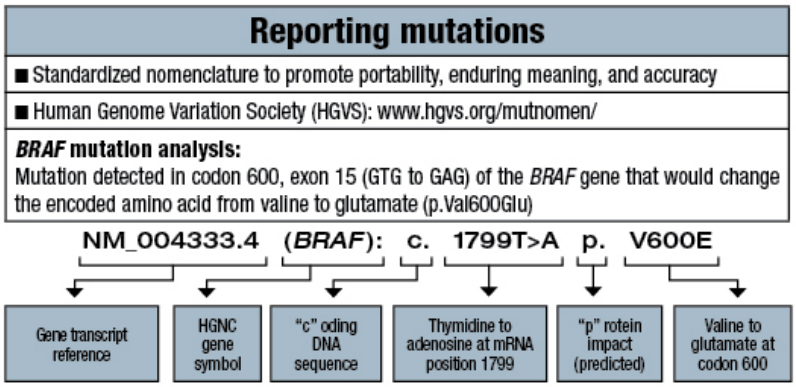
Of bam files Dr. Routbort said, "Trust but verify. Aligners are not aware of genes. They are just trying to map a sequence for best fit against a reference genome. Aligners don't know that the As, Cs, Gs, and Ts make up genes that can be read in a specific direction." Knowing about the underlying gene can help predict the impact of the variant on its protein.

Annotating the variants in the bam file yields a variant caller format (vcf) file. Once you have a list of possible variants, Dr. Routbort said, two questions should be asked. First, do you believe the computer when it tells you there is a difference at a position? "There are a variety of problems in pipelines that can yield false-positives," he cautioned.

Second, and more complicated, what does the difference mean in clinical terms? "At the very least, you have to translate the data in the vcf file, which is given in genomic coordinates, into something meaningful in terms of genes," he said. In the vcf file, four columns give the chromosome number (e.g. 3), nucleotide position (e.g. 1111333333), expected nucleotide at that position (e.g. A), and nucleotide found in the patient's genome at that position (e.g. T). In the standardized form proposed by the Human Genome Variation Society, the above information would read: g.1111333333A>T, but there are also guidelines for gene and protein level annotations that make the reporting of a given sequence change more informative and meaningful (see "Reporting mutations").

Once a variant is mapped against known databases, you can ask whether this variant has been reported before and whether it is a known somatic mutation. "In cancer samples we don't routinely test a patient's germline DNA," Dr. Routbort said, "so we always need to keep in mind a given variant may be a germline finding seen in all cells and not likely related to cancer."

In the first postanalytical step, aligning the sequence data in the fastq file to reference genomes, Dr. Monzon said the version of the reference genome is critical. Current standards are GRCh37 or hg19 (UCSC). "One issue in the alignment arena is that existing alignment algorithms can make mistakes," he said. "They were not originally designed to be clinical tools so you have to understand their limitations." However, most clinical laboratories using NGS are aware of this problem and take it into account when building clinical pipelines. A vcf file can also include a quality score for each position and information about whether the sequencing data passes pre-established quality filters.



Once a list of variants is generated, it is usually annotated with publicly available information. Multiple databases are used as annotation sources, including dbSNP (clinically relevant variants), HGMD (inherited disease), COSMIC (oriented toward cancer), My Cancer Genome

(therapeutic implications), BIC (breast cancer), OMIM (Online Mendelian Inheritance in Man), ClinVar (germline variant classification), and others.

ClinVar, which is maintained by the National Center for Biotechnology Information, aggregates information about sequence variation and its implications for human health. “ClinVar has been accessible for several years,” Dr. Monzon says. “It is meant to be the main resource in this field. It’s in constant development, but it is already very useful.” Laboratories can submit their findings to ClinVar, although not all of them do. “Its main drawback is that you don’t have the ability to review the evidence that a lab used to come to its conclusions about a specific variant. A lot of effort is going into improving it and making it a more robust clinical resource, but it is already a very useful tool,” Dr. Monzon says.

A number of factors affect the significance of a variant, such as quality of sequencing data, frequency of variant call, patient’s phenotype, family history, allele frequency in the population, and location of the variant in the protein, among others. Detailed guidelines for interpreting the clinical significance of germline variants are in the final stages of review under the joint sponsorship of the CAP, AMP, and American College of Medical Genetics and Genomics. (A webinar on these guidelines, “Interpretation of Sequence Variants,” was presented on April 24, www.amp.org.) A similar guideline effort for interpretation of somatic variants is needed.

Dr. Monzon

Dr. Monzon noted three difficulties in interpreting variants, even in panels. First, there is limited evidence of the clinical utility of specific mutations. “Few mutations are listed in consensus management guidelines,” he said, referring to the NCCN and ASCO in the somatic cancer field. “There have been few institutional efforts to gather and curate evidence.”

Second, the clinical significance of well-studied somatic mutations in different tumor types is unclear. One example is the clinical significance of the *BRAF* V600E mutation in malignant melanoma versus breast cancer. “Pathologists may have to research the evidence of clinical utility to issue a clinically relevant interpretation,” Dr. Monzon said.

Third, clinical evidence for novel mutations in targetable genes is often lacking. “For example,” he says, “if you find a novel mutation in *KRAS, BRAF, PIK3CA*, or other genes for which gain-of-function is the disease mechanism, how do you know that this novel sequence change has an effect in protein function and that it can be modulated by a targeted agent?” (A webinar on “Viewing and Interpreting Sequencing Data” was presented on Sept. 11, www.amp.org.)

After interpretation, variants considered clinically relevant are usually confirmed with a different method in most labs; most often used is Sanger sequencing, so this process has to be considered in a lab’s informatics support needs.

Working through an NGS informatics pipeline demands more from a pathologist than prior forms of testing. “We were working with a discrete result most of the time and relatively straightforward algorithms to analyze those results. With NGS,” Dr. Monzon says, “there is a lot more data and a higher degree of complexity, and we are measuring multiple different things—large swaths of the genome.”

This experience is not entirely new, however. “We have been employing genomic technologies in the clinical lab that are increasingly complex, such as in the cytogenetics arena, where we have been using microarrays for karyotyping for several years now,” he says. He notes, too, the use of mass spectrometry in the microbiology lab as another example of high-complexity data-sets, “which can be used to detect pathogenic microorganisms within

a mix of bacteria."

Like Dr. Carter, Dr. Monzon is not happy about how these reams of data are being moved around. "For tracking handoffs of tests and analysis within the laboratory, email is only a temporary solution for low volume. But it is not scalable." He favors use of Web technology within the laboratory to enable the lab to use tools that facilitate the transfer process. "Most of our systems at Invitae are Web-based, such as accessioning, interpretive tools, and reporting." Dr. Monzon sees the same obstacle Dr. Carter cited: "None of the widely used LISs or electronic health record systems are ready to deal with genomic information." One challenge is to deal with the multiparametric nature of NGS data, as opposed to typical single-analyte assays. For a specific reported mutation, sequence information, sequencing depth and quality, location, genome build, gene transcript evaluated, and technology used all need to be stored, among other things.

Another drawback of current EHR systems: "We have information communication standards that do not support data formatting and metadata—data associated to the result—and thus we need to 'dumb down' the result into text files and/or tables in order to be reported," Dr. Monzon said. Text-only reporting of NGS-based assay results is suboptimal, he said, adding, "Graphical or interactive presentation would be better."

That EHR systems are not able to deal with flexible reports creates a big problem, he adds. "We get different requests from treating physicians. Some want more detail, others want a more streamlined report." One way to cope would be an interactive report, which gives clear, concise information in the initial part, then provides an opportunity to delve into the details of the result and method. "Some consumer genetic companies have pioneered that area. But we are not there yet in terms of routine use of interactive clinical reports," he says.

In sum, Dr. Monzon's NGS reporting wish list includes the ability to report metadata, such as what was covered and how well, what could be missed, and what the evidence is for the interpretation of a specific variant. In medical genetics especially, it would be desirable to report whether there is new information on pathogenicity. On a broader scale, he would like to have a hospital information system that can handle molecular and genetic data.

Two special problems in reporting of NGS data are reporting variants with unknown clinical significance and raw data release. Of the first, Dr. Monzon says, "It is our obligation to report variants of unknown clinical significance and to make it clear why we can't come to a definitive conclusion."

For raw data release, there is a "clash of models," he says. In one model, "physician knows best," the patient has access to interpreted results only through a physician. In the other, "patient knows best," patients own their data and can decide how best to use them. "There are different degrees of genetic literacy among patients," Dr. Monzon notes. "Some people are very literate in genetics and want access to their own genome to explore. The idea that the patient owns their own data is widespread." But what is the best way for labs to release that information if the patient requests it? "I don't think we have adequate tools to release that information in a protected way," he cautions.

Dr. Routbort mentions one additional resource, internal databases, which can enable population-based frequency classification and filtering not only of common polymorphisms but also recurrent platform-specific sequencing artifacts. "It is extremely helpful to be able to know if you have seen something before and its frequency in your sample population. Part of our software puts vcf file data into a centralized internal database," he says, "so we can calculate on the fly from the thousands of samples we have run how many times we have seen a particular finding." Even if you don't have external databases, if you have an internal database, he says, you can discard a lot of noise. He calls his laboratory's internal database "our own COSMIC."

Cost remains a barrier to widespread use of NGS in routine clinical laboratory practice, and the informatics pipeline is a major reason. Dr. Routbort refers to a commentary titled "The $1,000 genome, the $100,000 analysis?" by genome scientist Elaine Mardis, PhD, of Washington University (*Genome Med.* 2010;2:84). "She brings up many good points," Dr. Routbort says. "It is important to keep perspective. Next-generation sequencing is likely to show some similarities with some of the other major techniques introduced into pathology historically, such as flow cytometry, immunohistochemistry, and cytogenetics. All had great impact and produce good information for

clinicians, but none has necessarily decreased cost or complexity. There is no magical box that has eliminated the need for other, complementary technologies."

In many cases, NGS will produce only incremental information, while for some patients it will be transformative. "It's fair to say that the feasibility of using next-generation sequencing to routinely determine personalized therapy for cancer is very much an open question," Dr. Routbort said. "We are still waiting to see how much it will contribute. Like microarrays and proteomics, these new toolsets can generate a huge amount of data. Now it is up to the community of pathologists and clinicians to find the best way to translate laboratory findings to clinical actions."

William Check is a writer in Ft. Lauderdale, Fla.