Use of Next-Generation Sequencing in the regulated domain of drug development

Next-Generation Sequencing is moving quickly from early research into the regulated domains of drug development, diagnostic development, and clinical decision-making. This article summarises some of the technical and regulatory challenges posed by these technologies and the efforts being made to address them.

By Keith Nangle and Mike Furness

Next-generation sequencing (NGS) has moved from the realm of research into those of clinical development, drug approval and clinical diagnostics, as the cost has decreased and the reliability of the underlying technologies has increased. However, the process of translating raw reads into reliable genotypes is still subject to much variability. This variability presents a challenge when using NGS in the regulated domain of the drug development process.

Unlike more traditional biomedical assay techniques, or even other genomic technologies such as microarrays, interpreting NGS data depends on a long chain of data-processing steps after the raw data are generated. Each of these steps is the subject of many competing algorithms, with more being developed and improved all the time. Furthermore, most algorithms have parameters designed to allow the algorithm to be 'tuned' to accommodate data generated under different experimental conditions. The result is that two independent analyses of the same underlying sequence data can lead to divergent conclusions. However, in a regulated environment such as a clinical trial or a treatment clinic, the goal is to have results that are robust and reproducible, and both analytically and clinically valid. New technologies are being developed to help manage this problem and regulators are grappling with the nature of these algorithms in the context of their regulatory requirements.

Algorithms

A typical NGS data processing pipeline includes the following steps:

1. The sequencing platform conducts image processing and the generation of raw reads (so-called 'primary analysis').
2. The next three steps shown in **Table 1** (Read QC, Alignment or mapping and Variant calling) are often referred to as 'secondary analysis' of NGS data.
3. The last step (Variant annotation) is part of 'tertiary analysis' in which the detected variants are annotated and interpreted as to their likely biological or clinical impact.

Each step in the pipeline introduces its own opportunities for variability and generates quality metrics to help the analysts judge the usability of the pipeline's outputs. Each individual nucleotide in a raw short read has an associated quality score that represents the likelihood that the base was identified correctly. Each aligned read also has a score that represents the likelihood that the read has been uniquely positioned within the reference genome. Each variant, and then each individual genotype, has a score that quantifies the uncertainty of the corresponding determination. Taken together, a sample analysis can follow any number of paths from a set of raw reads to a set of genotypes for the sample, with each path delivering differing results.

In addition to these quality scores, algorithms

Table 1

Processing step	Read QC	Alignment	Variant calling	Variant annotation
Input files	FASTQ or SAM/BAM	FASTQ (for sample) and FASTA (for reference)	SAM/BAM	VCF/BCF
Output files	FASTQ or SAM/BAM	SAM/BAM/CRAM	VCF/gVCF/BCF	VCF/BCF, BED or TXT
Common algorithms and toolkits	FASTqc, GATK ClipReads, Trimmomatic	BWA and a host of others	SAMtools, GATK-UnifiedGenotyper, a host of others	SIFT, PolyPhen, SNPeff, Annovar, VEP, Varant, ClinVar, a host of others

typically provide parameters that help tune the algorithm to the quality and depth of the data, the types and frequencies of variants expected, the characteristics of the genome, or indeed for computational efficiency. These parameters have default values that work well for most cases, but significant experience is often required to know when and how to adjust the parameters for less straightforward data. Finally, some algorithms (especially the computationally-intensive alignment algorithms) introduce stochastic effects by their design. They may use heuristics for the sake of computational efficiency, or depend on the order of execution when executing parallel threads.

By way of example, the PrecisionFDA Consistency Challenge evaluated the reproducibility of secondary analyses of the same known input across multiple executions of the same pipeline. Of 18 pipelines that participated in the challenge, eight were denoted as 'Deterministic', giving the same set of variants each time. The remaining 10 had inter-run differences ranging from 0.01% to 2.6% of the total number of variants detected. These discrepancies may seem small in numerical terms, but the actual number of clinically-relevant variants in any analysis is often small, and one must be sure that these variants are not the ones subject to much variability.

Interpretation

If one has a secondary analysis pipeline that is robust and reproducible, ie analytically valid, one then faces the next challenge: to interpret properly the meaning of the variants that are found in terms of their clinical impact, and to make sound decisions based on that interpretation. Public annotation databases such as ClinVar and The Cancer Genome Atlas offer curated sources of information about variants which have reasonable evidence of clinical effect. For variants that have not yet reached this level of certainty, tools such as SIFT, PolyPhen, Variant Effect Predictor, etc, can use other methods to assess the likely biological (if not clinical) impact of variants. These tools provide qualitative assessments such as 'benign', 'possibly damaging', or 'likely damaging' to convey the predicted impact of a variant on the associated protein. The FDA has issued draft guidance for assessing whether a public annotation database provides valid scientific evidence that might support claims of clinical validity of NGS-derived variants.

In addition, studies have investigated the variability in variant data interpretation between different locations, such as the nine-lab study run by the Clinical Sequencing Exploratory Research (CSER) Consortium. This study demonstrates that consistent interpretation of the clinical impact of variants remains a challenge, even when the same guidelines are being followed by different organisations.

Enabling technologies

In addition to ongoing development of new and better NGS algorithms, there have been many efforts to develop higher-level technologies that can help address these issues. The Common Workflow Language (CWL) is an open-source language for specifying the exact steps and parameters used in a lengthy analysis pipeline. There are many examples of reproducible analysis platforms, including Galaxy and Taverna, that can record and replay an analysis flow. The combination of a common workflow language with new container technologies such as Docker, mean that these frameworks can be implemented in a way that scales within and across computing environments and cloud configurations (for example, Rabix from Seven Bridges Genomics).

The FDA has been working on the specification of a BioCompute Object (BCO). The goal is to define a single research object that combines all of the computational steps and their parameters (using the CWL), as well as all of the input, intermediate and output data, into a single object that can be referenced by a unique accession identifier. Reproducible analysis frameworks could create these BCOs for submission to regulatory agencies, and they could be re-executed in a different environment to reproduce the analysis from start to finish.

Technology service providers

There is now a thriving market for solutions to the technical infrastructure needs of NGS practitioners, specifically including those working in regulated environments. The volumes of data generated are huge and the computational burdens are similarly large, and often transient. NGS analysis environments must accommodate large-scale data transfer and storage, metadata management, workflow management, data provenance, data archiving and access to public databases. This must all be provided in an environment that supports data encryption and security, access control and auditing and data centre management according to standards such as ISO27001/27002. Furthermore, it must be easily scalable in order to handle the next project or batch of samples, but without the high overhead of a fixed infrastructure that sits idle between projects.

All of these characteristics point naturally to hosted Platform-as-a-Service (PaaS) or Software-as-a-Service (SaaS) solutions. Companies such as DNAnexus, Bluebee and Seven Bridges Genomics provide NGS-tailored PaaS/SaaS environments built on existing infrastructure providers such as Amazon Web Services, Microsoft Azure or Google Cloud. They provide ready access to the tools and databases typically used for NGS analyses and allow customisation, sharing and reuse of genomic analysis pipelines. These providers can also help manage data localisation, where particular countries or

Table 2: Genome screening projects

Project	(Target) Cohort size*
AstraZeneca 2M Genomes Project	2,000,000
Ancestry.com	1,400,000
23andMe	1,000,000
Million Veteran Program	1,000,000
Precision Medicine Initiative	1,000,000
Korea Biobank Project	618,958
European Network for Genetic and Genomic Epidemiology (ENGAGE)	600,000
Resilience Project	589,306
China Kadoorie Biobank Repository	512,000
Kaiser Permanente: Genes, Environment, and Health (RPGEH) Repository,	500,000
UK Biobank Repository, Consortium	500,000
deCode Genetics	500,000
Regneron/Kaiser Permanente MyCode® Community Health Initiative Repository	250,000
French Genome Project	235,000
Vanderbilt's BioVU Repository	215,000
BioBank Japan Repository Specimens	200,000
Leiden Open Variation Database (LOVD) Repository	170,000
Psychiatric Genomics Consortium (PGC)	170,000
100K Wellness Project	100,000
Turkish Genome Project	100,000
Genomics England	100,000
Actionable Cancer Genome Initiative (ACGI) Data-Sharing Project	100,000
Genome Asia 100K Consortium	100,000
Saudi Human Genome Program	100,000
East London Genes & Health	100,000
Exome Aggregation Consortium (ExAC)	60,706
Electronic Medical Records and Genomics (eMERGE) Network Repository	55,028
Estonian Genome Project, Estonian Biobank and the Estonian Genome Center (EGCUT)	52,000
International Multiple Sclerosis Genetics (IMSG)	50,000
International Genomics of Alzheimer's Project (IGAP)	40,000
Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium	20,000
Cancer Moonshot 2020 Consortium Phase I	20,000
DECIPHER Repository	19,014
GENIE/AACR Data-Sharing Project	17,000
International Cancer Genome Consortium (ICGC)	16,000
CIMBA Consortium	15,000
Tohoku Medical Megabank Project (ToMMO)	15,000
Sequencing Initiative Suomi (SISu)	10,000
Genome Korea in Ulsan	10,000
UK10K Research Project	10,000
T2D-GENES Consortium	10,000
PopGen (Germany)	10,000
SardiNIA Study	7,000
Qatar Genome	6,500
Personal Genome Project	5,015
Clinical Sequencing Exploratory Research (CSER) Consortium	4,000
Scottish Genomes Partnership (SGP)	3,000
TBResist	2,600
Faroe Genome Project (FarGen)	1,500
African Genome Variation Project	1,481
Human Genome Diversity Project	1,050
Genome of the Netherlands (GoNL)	750
Singapore Genome Variation Program	268
GenomeDenmark	150

** these figures are for samples and are taken from published information on the projects*

regions may require data derived from their citizens to reside within the country or region.

For those organisations that utilise external laboratories to perform the sequencing itself, another option is to rely on the laboratory to also provide the necessary storage and computational services, such as Illumina BaseSpace or BGI Online. There are many laboratories that provide NGS services in a CLIA-certified environment.

Clinical diagnostic laboratories work directly with hospitals and physicians to provide diagnosis and treatment options for individual patients. They may use NGS technologies, but deliver clinical reports and advice rather than just a set of variants.

Companion and complementary diagnostics

Perhaps the most visible applications of NGS in the clinical realm are companion and complementary diagnostics. A 'companion diagnostic' is a medical device, often an *in vitro* device, which provides information that is essential for the safe and effective use of a corresponding drug or biological product. A 'complementary diagnostic' is a device which is essential for the safe and effective use of a corresponding medicinal product to identify, before and/or during treatment:

- Patients who are most likely to benefit from the corresponding medicinal product, or
- patients likely to be at increased risk of serious adverse reactions as a result of treatment with the corresponding medicinal product.

The oncology therapy area is especially active in its use of NGS technologies, which are well-suited to characterising tumours based on their genomic variants, often mutations specific to the individual tumour. One of the first examples of a drug with an associated companion diagnostic (though not NGS-based) is trastuzumab (HERCEPTIN®, 1998). A more recent example is the Foundation-One™ from Foundation Medicine, Inc, which interrogates 324 genes in tumour tissue for a variety of variant types and total mutational burden in order to select the most appropriate therapy for the individual across a range of cancer types.

There are a sizeable number of large-scale biobanking and genomic sequencing initiatives currently under way, as shown in **Table 2**. An outstanding question is how can the industry make use of these data for both discovery and diagnostic development purposes? For example, can it use sequence information to stratify biobank subjects for enrolment in clinical trials? For this to happen, data generated for research purposes must be collected and managed to the same regulatory standards as clinical trials data.

Government, academic and industry initiatives

Government, academic and industry players are heavily involved in efforts to provide the infrastructure, databases, standards and regulatory oversight necessary to support use of NGS in clinical development. Among the examples are:

ELIXIR: An EU-sponsored programme to develop interoperable data, computational and training resources for life science research. ELIXIR is a leader in the promotion of the FAIR principles, that biomedical data must be Findable, Accessible, Interoperable and Reusable.

Precision Medicine Initiative: An initiative of the US National Institutes of Health, established in 2015. The goal of the PMI is to develop the scientific evidence needed to move the concept of precision medicine into clinical practice.

The **Association of Molecular Pathology** and the **College of American Pathologists** have recently published a joint set of standards and guidelines for validating NGS bioinformatics pipelines.

Global Alliance for Genomics and Health (GA4GH): An alliance whose goal is to enable the interoperability of systems and processes used to process clinical and genomic data, and thereby enhance the sharing process. One of the primary deliverables is a set of standard application programming interfaces (APIs) that support the discovery and interchange of genomic data. The alliance has more than 500 institutional members and individuals can be members as well.

Pistoia Alliance: The mission of the Pistoia Alliance is to lower the barriers to innovation in life sciences R&D through pre-competitive collaboration. Among the projects in its portfolio is Faster CDx by Aligning Discovery & Clinical Data in the Regulatory Domain, which aims to address many of the issues described in this article.

Regulatory guidance

In its presentations and workshops on regulatory oversight of NGS-based tests, the FDA recognises some key differences between 'conventional' and 'precision' diagnostics (**Table 3**.)

The agency clearly recognises that these technologies require a different, more adaptive approach to regulation as compared with earlier technologies, and they are keen not to stifle innovation that will lead to real benefit for patients.

Both EU and US regulatory agencies have issued draft guidelines to address the challenges of using NGS technologies in drug development. Examples include:

- 'Use of Standards in FDA Regulatory Oversight of Next Generation Sequencing (NGS)-Based In Vitro Diagnostics (IVDs) Used for Diagnosing Germline Diseases'.
- 'Use of Public Human Genetic Variant Databases to Support Clinical Validity for Next Generation Sequencing (NGS)-Based In Vitro Diagnostics'.
- Guideline on good pharmacogenomic practice (Draft).
- Guidelines for diagnostic next-generation sequencing.

What next ?

In this article we have reviewed some of the challenges posed by NGS technology in clinical development, including:

- Rapidly changing sequencing and analysis technology.
- Complex data processing pipelines and infrastructure requirements.
- Non-deterministic algorithms, with variable performance across the genome.
- Use of a variety of public annotation sources to help establish clinical validity of results.

Each of the topics discussed above is a 'work in progress', and the examples shown for each category are by no means exhaustive. Neither regulators

Table 3

Conventional diagnostic	Precision diagnostic
Low/medium resolution technology	High resolution technology ('omics')
Detect a finite number of analytes (usually one)	Undefined (millions?)
One test – one disease	One test – many diseases
Clinical evidence from clinical studies – research separate from practice	Clinical evidence from learning health systems – merging of research and practice

nor industry have completely settled on a single approach to the use and validation of NGS-sourced data. The questions for those involved in the regulated domains of drug and diagnostic development then become: where to begin and which technologies, standards and initiatives are worth following?

It may be neither possible nor desirable to develop a single, standard approach that works for all of the ways that NGS data can be applied in clinical development and practice, but that should not stop the industry from developing best practices that can serve as a template for practitioners (while recognising and adapting to the reality of fast-moving technological change). Pre-competitive consortia such as the Pistoia Alliance can serve to bring together stakeholders from the pharma and diagnostic industries, the technology service providers and the regulators, to share experience and to develop those best practices in real-world development. In this way the industry can work with the regulators to develop appropriate approaches to the use and validation of these new technologies and avoid each company individually having to learn the same lessons on its own. DDW

***Keith Nangle** was formerly Head of Genetic Data Sciences at GlaxoSmithKline Inc. He also served as European Community Manager for the tranSMART Foundation before working with the Pistoia Alliance on the use of NGS in regulated areas of drug development.*

***Mike Furness** is Founder and MD of TheFirstNuomics and has worked in genomics and bioinformatics for the last 30 years providing expert technical and business development support to a range of start-up companies and investors. He has previously worked at Congenica, DNAnexus, Incyte Genomics, Pfizer, Cancer Research UK and Life Technologies, and is working with the Pistoia Alliance on its Companion Diagnostics NGS project.*

ADVERTISEMENT INDEX