Applied Information

Delivering clinical applications of genome sequencing faces a number of informatics challenges relating to processing the vast amounts of data obtained. Annotation businesses could be the answer to making information available at the clinician's desktop

Rodrigo Barnes
at Aridhia Informatics

The pursuit of personalised medicine, driven by whole genome sequencing, has reached an interesting point. The core technologies involved are perceived to be coming in at the right price, and useful results are coming out of the laboratory. However, there are mixed signals as to whether this is ready for mainstream application.

In some health systems, doctors can already order a companion diagnostic that determines whether a patient has a specific genetic make-up to respond to a particular drug or not. But more generally, the challenges of making this work in a clinic are evident. As a result, corporations and governments – both local and national – are investing heavily to achieve health and monetary benefits.

Stratification Studies

Arguably, the biggest challenge is informatics. In order for personalised, precision medicine to work, there needs to be more than just efficient sequence data generation; there has to be evidence from a stratification study. This involves the classic double-cycle of analysis and application, in which individuals are sequenced and tracked over time for clinical outcomes. Typically, the study may run as part of a clinical trial for a drug – for example, lung cancer patients with a set of mutations in the gene EGFR were shown to respond to Gefitinib (Iressa®). Once results are repeated, drug approval or clinical guidelines can be obtained, and other trials can be implemented to look for genetic signals or biomarkers that might explain the variations in responses – leading to research into potential drug targets.

A number of technologies can be utilised to determine a patient's genetic information. Until relatively recently, genetic testing has been selective, and limited to processes such as polymerase

chain reaction amplification. However, as sequencing has become more affordable in the lab and the clinic, targeted sequencing has become popular – for instance, with the use of exome sequencing targets to drive protein coding.

There are approximately 20,000 genes in human DNA, so the magnitude of information available goes beyond that which can be obtained from a simple genetic test. As a result, new developments in sequencing can raise the data bar even further, well past what is known about genome function.

Stakeholder Engagement

Compared with targeted tests, whole genome sequencing is a simpler procedure to order and execute, as it can potentially be standardised involving reduced human contact. A lab's service catalogue can be consolidated, leaving the variation of outputs to algorithms and business rules.

At the same time, significant stakeholders are investing in the potential benefits of sequencing. An emerging network of centres demonstrates this across the world: the UK has Genomics England, Stratified Medicine Scotland and Stratified Medicine Northern Ireland; in the US, a notable example is the New York Genome Center, which combines a pooling of resources across New York hospitals and universities. Meanwhile, industrial investment from Life Technologies/Thermo Fisher Scientific, Illumina and Beijing Genomics is driving technology and scalability of data generation.

Another form of consolidation exists in the form of the Global Alliance for Genomics and Health which, in part, aims to federate the data gathered at different sites to maximise the potential stratification. Moving from here to the clinic requires informatics and operational models, which remain a start-up arena of great uncertainty.

Workforce Training

When it comes to precision medicine, the intellectual intensity of genomics and bioinformatics should be put into the context of a patient's journey of care. Medical professionals could come to rely on genomics to inform their decision-making, as well as looking to patient medical records and lifestyle choices.

As a result, the UK Human Genomics Strategy Group has highlighted the importance of workforce training, as well as the need for robust bioinformatics. Currently, informatics – which covers stratification, data and bioinformatics – remains largely off the curriculum for medical students and, even if there is a level of awareness, there is no specific means for a clinician to check what is available for their patient. In the UK, there are web resources with pathways and lab directories. However, in the NHS, if therapies do not appear on approved IT, then they are unlikely to be seen on a doctor's desktop. Whereas, in the US, it is advertising that drives awareness.

If a patient qualifies for whole genome sequencing, there is currently an implicit trade-off. The system creates more information than can be processed on the clinical pathway, in return for creating a research asset. Therefore, patients will have to authorise their consent. The analysis is then driven by the initial consultation – such as suspicion of lung cancer – and the system sifts through the sequence and variation data to find a match with, for example, known EGFR mutations.

This may highlight incidental findings, and the systematic application of previous research could expose more factors affecting the patient's health, leading to the need for ethical frameworks to be established as a form of operational guidelines. The simplest model provides an answer to the doctor's question, while retaining the additional data for anonymous research. However, in practice, this might be subject to exception when handling life-changing or life-threatening findings.

Data Interpretation

The specific process of taking a whole genome sequence and extracting clinically useful information has come to be known as an "annotation service". The result is an informatics product, with software applied to provide research and clinical decision support. The thought of taking such a product to market may not be what research groups aim for, preferring to consider themselves as breaking new ground in the knowledge of disease. In some markets, such as the US, there are companies competing with transactional services, while in the NHS-dominated UK, it is not clear what role private companies have in providing sequencing and annotation services.

"The specific process of taking a whole genome sequence and extracting clinically useful information has come to be known as an "annotation service". The result is an informatics product, with software applied to provide research and clinical decision support"

The translation path for informatics is not the same as for pharmaceutical products, and to understand this, it is important to look upstream, where the majority of the computing effort in annotation occurs.

This happens at the stages of interpreting the signal data from the sequence (either as images or pH), aligning fragments of data to reconstruct the whole sequence, and "calling variants" between a result and a reference human genome. During these stages, data sizes drop from terabytes of signal data to hundreds of gigabytes for sequences, and hundreds of megabytes for variant call files. For an individual in lifelong care with a whole genome sequence, some of these stages should be repeated as new information can update the reference human genome, or the quality of algorithms for one of the stages.

Software Pipelines

Bioinformatics core facilities and research teams have developed software pipelines based on the type of sequencing equipment that happens to be in their lab, using a blend of software components that are enough to perform their research but cannot be operated by a non-specialist, for example. A key distinction between the pharma translation path and that for bioinformatics is the difference between a patentable biomarker or compound, and an amorphous software pipeline. What is handed over to the company exploiting the technology?

Annotation itself is driven only in part by abstract algorithms, while variant calling is driven largely by knowledge – based on data from previous sequencing, encapsulated in the academic consensus of the reference human genome, currently referred to as GRCh38 from the Genome Reference Consortium. Classifying these 'deltas' remains a combination of biochemistry and linkage with the best-known research. In some cases, there is prediction of effect; in others, an indication of effect is observed elsewhere. In either case, access to knowledge is key.

In practice, most of this knowledge is in the public domain already, but it can be combined with research or corporate databases accumulated through many years of experiment or service. Annotation providers can also use commercially scalable information services. Often, these include manually curated content that makes sense of the enormous volume of research published each year.

An annotation pipeline needs to integrate with many reference databases, predictive tools and web services in order to distil the sequence to the actionable signal. As a result, their value is often tied quite specifically to clinically actionable statements about a disease type or a particular type of analysis. The bulk of the whole genome is not exploited, but stored for future breakthroughs.

Scalable Services

In the case of a start-up, when launching an innovation directly, a number of questions remain about what business model to develop, including what user community to serve. There are potentially many consumers for annotation services. These range from patients and disease specialists with knowledge of genomics, to clinical genetics labs integrated within the health system that have the task of designing scalable services. In all cases, it is important to understand users' needs and limitations. Some consumers, such as patients themselves, have treatment options as a priority, while clinical researchers want potential signals in their pursuit of insight into disease.

Providing a clinically valid service will eventually require regulation, approval and scale. Standards are only now emerging for upstream analysis – the US National Institute of Standards and Technology is leading a standardisation effort from sample to variant calling – and software will fall into this category, as it emerges as a medical device providing test results and decision support. This imposes degrees of traceability and changes how the software itself is managed in a way that might not have been important in the research project it came from.

Path To Success

A number of annotation services will co-exist in any given healthcare setting, raising questions as to how clinicians will interact with them. Much has been made of interactive tools, using visualisation to allow exploration of the depth of evidence behind a result, or just the volume of data available. This is useful to a researcher building up a hypothesis or searching for possibilities.

For the working clinician who referred the patient, a simple confirmation of diagnosis, or determination of the best pathway to follow, is critical. Meanwhile, where electronic medical record systems are in place, results can be delivered using standard messaging protocols (including HL7) or document repositories. As a result, annotation service providers should plan to deliver their data in partnership with the IT departments of the health system.

Once the research and clinical trials of stratification have been completed, results can be embedded in annotation services. A clinician ordering a test may not care whether a whole genome was sequenced, but increasingly this will underpin the service they request. For now, the success of the annotation business will rest with labs and how they deliver their information to the clinicians' desktop.

About the author

Rodrigo Barnes is Chief Technology Officer at Aridhia Informatics. He has been instrumental in designing the company's approach to big data within healthcare, as well as developing its research analytics and stratified medicine services. Rodrigo's current focus is on working with technology partners to develop innovative services.
Email: rodrigo.barnes@aridhia.com