HealthManagement.org
The Journal
LEADERSHIP • CROSS-COLLABORATION • WINNING PRACTICES
VOLUME 19 • ISSUE 3 • 2019 • € 22
ISSN = 1377-7629

Top Target Treatments

A primer on next-generation sequencing data analytics

Next Generation Sequencing (NGS) is rapidly becoming more and more standardised in terms of sequencing techniques.

With Precision Medicine coming into the clinical forefront, it is important to know the steps and possibilities associated with Next-Generation Sequencing techniques.

Shelly Mahajan

Clinical Lead Genomics,
CARINGdx, Mahajan Imaging
New Delhi, India

drshelly@caring-research.com

@CARINGresearch

Vidur Mahajan

Head of Research
CARING, Mahajan Imaging
New Delhi, India

caring-research.com

vidur@mahajanimaging.com

@VidurMahajan1

Introduction

Next Generation Sequencing (NGS) is rapidly becoming more and more standardised in terms of sequencing techniques, library preparations and assay development. The main challenges arising in running a good NGS lab focus around data management and making sense out of that data. The amount of data produced in a single NGS run is magnanimous thanks to the high throughput associated with sequencing, reduced cost of sequencing per base pair and high quality library preparation kits which have enabled the creation of larger multigene panels (example – TSO500, which is a 523 gene panel).

The nature of this data, makes its analysis and interpretation time consuming and thus increases the turnaround time and cost in clinical settings. Hence data analysis and interpretation have become the problem areas and differentiating factor between good and average clinical NGS labs. Nowadays, several companies and research groups have developed tools for easier and much more efficient analysis and interpretation of the sea of data produced by sequencing. One would think that in today's day and age, the analysis of such data would be trivial- yet, NGS data analysis is different from other forms of data because of multiple reasons: the amount of data produced, the quality of the data produced and the impact of this data in the broader clinical perspective. For example, even a simple clinical exome identifies around 20,000-30,000 variants from around 6,000 genes. The main hurdle, and therefore opportunity, in NGS data interpretation is to zero in on a single or few variants responsible for the patient's phenotype.

Advanced bioinformatics solutions have significantly improved the data analysis and clinical interpretation of genetic variants. These user-friendly tools can augment the efficiency of a bioinformatics expert, data analyst and clinical interpreter. These programmes use a series of filters which can be manually selected and help to identify the causal variants. These filters can help answer questions such as "which of the mutations found in this patient have pathogenic findings for breast cancer and have an FDA approved therapeutic option" – consequently paving the way for precision medicine in oncology.

Data quality check

NGS data analysis software consists of various steps including quality assessment of the data, alignment, variant calling, annotation and visualisation. After the sequencing run is complete, the data is evaluated based on quality of raw reads- low quality reads are trimmed or removed to avoid wrong clinical interpretation. Various tools used for this quality check include FastQC, NGSQC, ContEst etc with each of these having specific roles. Integrated tools have also been developed which provide summary statistics as well as filtering and trimming functions. Now, platform specific tools have also been developed.

Alignment

After the reads have passed specific quality checks, they are aligned to a reference genome to check for 'deviations' from the reference genomes. There are two main sources of human genome assembly: University of Santa Cruz (UCSC) and Genome Reference Consortium

(GRC). Out of these two, UCSC offers 'hg19', which is currently used as a reference. The most commonly used alignment programs are Bowtie, Novoalign, BWA, MAQ, mrFAST. Reads which present with multiple mismatches are discarded from further analysis and after alignment the software removes PCR duplicates to avoid errors in variant calling.

Variant identification

The next important step is variant identification. It is influenced by all the steps of the test- from test design and coverage, to the bioinformatics tools used for data alignment and analysis. Tools to identify these 'variants' are called 'Variant Calling' tools, and their choice is determined by the kind of variants one is searching for. Major types of variants are Germline, Somatic and Structural. Germline variants are typically found in hereditary and rare diseases, somatic generally signify cancer related mutations and structural variants include Copy Number Variants (CNVs), insertions and deletions (INDELs), translocations etc.

Variant annotation

After identification, the variants are 'annotated' to filter those variants which the phenotype can be attributed using a computational tool. The phenotype in this case refers to the 'reason' why the test is ordered– for example, it can be the presence of a particular type of cancer, or a child with a suspected hereditary condition. There are different annotation tools focussing on SNPs (most common) and INDELs. The limitation of these tools is their use in structural variants as currently they are well developed only for CNVs. These tools mostly provide links to public databases for functional classification of the variant into accepted or deleterious mutations. There are now both web-based and offline applications for annotation. Although the web-based applications are easy to use and do not require physical hardware, they are dependent on service availability and require analysis of single variants entered manually. While the offline tools solve these issues, they require good user technical skills.

Data visualisation

Visualisation of generated data is very helpful in data interpretation. Examples include genome browsers to compare data with different annotations and viewers to compare sequences between different organisms. Like annotation tools, genome browsers can be web-based or offline, and can be accessed on standard platforms like Windows, MacOS and Linux. Apart from being extremely user friendly, web-based browsers provide access to a variety of annotations from various databases. An obvious risk is in the form of potential security and legal issues that might arise with uploading of patient data to external servers – this aspect needs to be dealt with differently according to each country's regulatory restrictions and mandates. Offline browsers, naturally, are safer in terms of data security but require highly skilled personnel who have to download annotation files, update the annotations regularly and perform complex calculations which are automatically done by the web-based browsers.

Summary

In summary, NGS data analysis is extremely complex and requires a series of computational steps which must be conducted and completed sequentially and accurately. Historically, NGS has remained a tool for research, but with the advent of precision medicine and personalised therapeutics, it is coming into mainstream clinical work, making it important for doctors and managers to know how to handle the data generated by these machines. To make things faster and more efficient, it is generally advised to create and establish end-to-end pipelines for managing the data. These pipelines have algorithms which require expertise to build, but once created are extremely helpful in data analysis and interpretation. There are now several companies that provide bioinformatics solutions taking care of all the steps of analysis of NGS data making the use of NGS in clinical scenarios a much more viable option in terms of time needed for analysis and cost effectiveness. ■

KEY POINTS

- ✓ NGS data analysis is extremely complex and requires a series of computational steps which must be conducted & completed sequentially and accurately.
- ✓ Main challenges in running a good NGS lab focus around data management and making sense out of that data.
- ✓ The amount of data produced in a single NGS run is magnanimous thanks to the high throughput associated with sequencing, reduced cost of sequencing per base pair and high quality library preparation kits.
- ✓ With the advent of precision medicine and personalised therapeutics, NGS is coming into mainstream clinical work, making it important for doctors & managers to know how to handle data generated by these machines.