Clinical Chemistry Trainee Council
Pearls of Laboratory Medicine
www.traineecouncil.org

TITLE: Massively Parallel (Next-Generation) Sequencing

PRESENTER: Melody Caramins, B.Med, PhD, FFSc, FRCPA

Slide 1:
Hello, my name is Melody Caramins. I am a Genetic Pathologist, also known as a Clinical Molecular Geneticist, at New South Wales Health Pathology. Welcome to this Pearl of Laboratory Medicine on "Massively Parallel (Next-Generation) Sequencing." In this presentation, I'll start by giving a perspective on the development of sequencing and the enormous impact that massively parallel sequencing (MPS) has made. I'll then give an overview of the general laboratory steps involved in data generation and analysis. Finally I'll talk about some of the clinical applications for massively parallel sequencing.

Slide 2:
The definition of sequencing is determining the order of nucleotides in a given DNA molecule.
This process was first described by Maxim and Gilbert, and also independently by Fred Sanger, and resulted in their joint award of the Nobel Prize in Chemistry in 1980.

Until very recently, Sanger sequencing has been the method of choice for DNA sequencing in the diagnostic setting.

One of the aims of the Human Genome Project was to encourage the development of technologies which would enable faster and cheaper sequencing. The cost of sequencing when the human genome project was proposed in 1985 was ~$10/base. By the time they finished in 2003, the cost had fallen to ~$0.01/base; and now, with the introduction of massively parallel sequencing, the cost is in the order of ~millionth of $/base and still rapidly falling. For the cost of the human genome project you could probably get about 50,000 or more genomes today, so the aim of cheaper sequencing was certainly achieved.

Slide 3:
All sequencing platforms generate sequence data in the form of independent reads, which then need to be assembled back together to get the complete sequence. In medical resequencing, those reads are usually assembled by mapping them back to the already known reference genome.

To give you an idea of scale, with Sanger sequencing, read lengths are typically 800-1000 bp, and the instruments have 1-96 capillaries, and each capillary generates one single read. So you could generate a maximum of 96 reads of about 1000 bp per run. That's around 1,000,000 nucleotides or 1 Mb/day, which means the haploid human genome, that's one copy of each chromosome, at about ~ 3 billion bp or 3 gigabases, would take about 8.5 years.

Massively parallel sequencing methods produce much larger quantities of sequence as much larger numbers of reads, but typically in much shorter read lengths of 50-200 bp. The current work horse instruments of genomic sequencing are capable of generating up to 800 Gb of base call data per sequencing run.

One problem with these shorter reads is that any repetitive regions longer than the read length can't really be accurately mapped and assembled, so the technology isn't really great in regions with long repetitive DNA, but there are constant improvements.

Slide 4:
In this figure, you can clearly see the game-changing impact of the introduction of MPS on the cost of sequencing. The cost per genome on the Y-axis of this graph is logarithmic, and this demonstrates that introduction of MPS technology has had an impact on the cost of sequencing which outperforms Moore's Law. It's also worth mentioning that in calculating these costs, the genome was covered six-fold (or 6x) with Sanger sequencing, and about 30x with some of the newer technologies.

Slide 5:
Here you get a general overview of the laboratory workflow in massively parallel sequencing. The first decision is obviously one about the clinical application, and we'll talk about that in more detail a bit later.

A typical MPS workflow consists of a number of laboratory steps including library preparation and sequencing and three tiers of data analysis: primary data analysis, which refers to base calling from raw sequence machine output; secondary analysis, which involves the alignment of reads to the human reference sequence and the identification of sequence variants; and finally, tertiary analyses describe the annotation of sequence variants.

The two major bottlenecks in MPS workflows are the upfront sample and library preparation and the final annotation and interpretation of the resulting variant data. Both of these processes are labor-intensive, although some recent exciting advances in library preparation protocols and platforms for automation of library builds can now simplify the laboratory workflow.

Slide 6:
The next step is to prepare your library. To prepare the library, take the extracted genomic DNA and randomly fragment it. Some common methods to do this include sonication or nebulization. Then, the ends of the DNA fragments are repaired and adaptors are stuck on, or ligated, to the fragments so they can be physically bound to a surface such as a bead or a glass slide. The fragmented DNA molecules are then enriched, generally either by PCR, or by capture and amplification.

The purpose of enrichment PCR is to selectively enrich those DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library. It's usually a small number of PCR cycles (e.g. 10) to avoid skewing and amplification bias in the library.

Slide 7:

Finally, the prepared library is sequenced in parallel. There are several competing next-generation sequencing (NGS) technologies, but the main platforms currently on the market are from:

- Illumina
- Roche 454
- Life Technologies
- Pacific Biosciences
- Complete Genomics

Each has platform-specific methodologies for sequencing, such as sequencing by synthesis (Illumina, 454), ligation-based sequencing (SOLiD), or by measuring H+ ion release (Ion torrent, Ion Proton). Each of these technologies has advantages and disadvantages, and if you are interested, you can read the review articles that compare them. Ultimately, the decision on which platform is most suitable will be based on the intended applications, financial considerations, and available support. It's also worth remembering that bioinformatics analysis may need to be adapted to each of the technologies to achieve optimal results.

Slide 8:

As mentioned, primary data analysis refers to base calling from raw sequence machine output. Most of this occurs on-instrument these days using reasonably robust algorithms. Some institutions, such as the Broad Institute at Cambridge, Massachusetts, which is affiliated with MIT and Harvard, have actually performed their own primary analyses.

Secondary analysis involves the alignment of reads to the reference sequence and the identification of regions which are different from the reference – these are known as sequence variants. The choice of aligning and variant calling software can significantly affect the number and types of variants you get, so it's an important step.

Finally, tertiary analyses describe the annotation of sequence variants. This step is essentially about assigning biological meaning to variant discovery and might include things like annotation of location, amino acid change, and a prediction of biological significance.

So to recap, the basic steps from raw instrument output to variant calls include base calling, quality control, alignment to a reference genome, variant and genotype calling, and annotation and filtering of the resulting variants (reviewed in ALTMANN et al 2012).

For clinical applications, labs implementing these technologies need to consider the demand requirements of different platforms and their applications on compute power and storage. As an example, the final binary aligned file, or .bam file, from a bench top platform for a gene panel for a few pooled and bar-coded samples might be approximately 2GB. A whole exome file for one individual at ~40x coverage might typically be ~5 Gb, and a whole genome file at similar coverage might exceed 80 GB.

Another decision point will be whether to implement an in-house bioinformatics pipeline for variant calling, provided sufficient compute and bioinformatics resources are available, or whether to go with a commercial product.

Slide 9:

Let's talk about some of the applications of MPS now – and there are many, both in the research and diagnostic settings. Whole genome sequencing is still largely the province of research, mostly because its clinical utility requires further assessment, its interpretation in a clinical setting is challenging, and it's still relatively costly. However, clinical exome sequencing, or sequencing of the entire coding region (which is about ~2% of the genome), is more manageable and has been available as a clinical service from some accredited laboratories since late 2011, and is mostly used to identify the cause of rare and unknown genetic conditions in a family.

Some other current and immediate applications of massively parallel sequencing in the diagnostic setting have included:

- High throughput and low-cost testing for well-established high penetrance genes (such as BRCA1 and 2) for germline variants
- Pharmacogenomic testing to identify genomic variants that might influence drug absorption and metabolism and therefore, affect their biological availability, efficacy, and toxicity (such as variants affecting clopidogrel or tamoxifen usage)
- Tumor driven testing for somatic mutations in genes which might influence treatment decisions or outcomes, such as EGFR and KRAS variants in cancer
- Characterization of cancer genomes from tumor-derived DNA from plasma of cancer-affected patients (non-invasive cancer testing)
- Non-invasive prenatal diagnosis of prenatal aneuploidy by maternal screening of fetal DNA in maternal plasma

Slide 10:

In Mendelian disorders, exome sequencing has demonstrated clinical utility by enabling the diagnosis of previously uncharacterized rare genetic disorders, leading to gene discovery, characterization of genetic heterogeneity within known clinical disorders, and disease reclassification. This approach has been popular because it can be much more cost-effective than individually testing all possible candidate genes.

The main approaches for gene discovery include:

- Sequencing multiple affected individuals from the same family to identify a shared novel variant and thus diagnose rare disease
- Sequencing multiple unrelated, affected individuals for variants in the same gene or pathway
- Sequencing parent–child trios to identify de novo mutations
- Sequencing and comparison of extremes of a phenotype distribution to identify variants for quantitative traits

Additionally, massively parallel sequencing can be used to effectively screen for carrier status for common Mendelian disorders by providing the means to simultaneously interrogate all loci of interest.

Slide 11:
Cancer is a one of the major causes of mortality and morbidity worldwide, and using MPS to identify the complete DNA sequence of cancer genomes has the potential to provide major breakthroughs in our understanding of the origin and evolution of cancer.

MPS applications in cancer can be at multiple levels:

- Single patient studies can try to discover which somatic mutations are important in the pathogenesis of cancer in an individual patient and can help guide treatment and clinical practice, and possibly explore clonal evolution. These studies are positioned to be key components of personalized genomic medicine.
- A genome discovery cohort of the same type or subtype of cancer can, on the other hand, potentially uncover recurrent mutations in particular genes and pathways in different individuals. This can then implicate those genes and pathways in pathogenesis, and does this in an unbiased, hypothesis-free manner. The current recommendation from the International Cancer Genome Consortium (ICGC) for such a cohort is about 100 tumor normal pairs in a discovery cohort.
- Cohorts incorporating multiple "omics," such as whole genome/exome, transcriptome, and methylome, are very exciting as they provide the opportunity to detect all the types of abnormalities implicated in cancer pathogenesis by integrative analyses. This type of analysis integrates all the multiple "omics" to try to explore how the different types of mutation discovered might then all converge on a mutated locus, gene, or pathway.

Slide 12:
Implementing MPS in a clinical academic setting is a non-trivial exercise. Here I've classified some of the challenges as pre-analytical, analytical, and post-analytical.

The pre-analytical challenges include the actual technology and associated overheads, such as adequate space, power, tools, infrastructure, and staff training. Sample quality and quantity issues are gradually improving with better preparation protocols.

Analytical challenges include things like providing sample turnaround times which are clinically useful, assessing analytical validity, and building and validating a data pipeline for the acquisition and management of properly consented samples. Facilities that may not have access to an experienced bioinformatician may need to build a group of trained in-house staff. The core skills necessary will be in computing systems, programming, biology, and statistics. Commercial and open-source solutions can be considered, although most research laboratories prefer to create custom solutions. The challenge with locally-developed solutions in the clinical setting is the lack of standardization of data analysis for benchmarking.

Post-analytical challenges include making sense of the sheer volume of data generated by MPS applications, which still require much time, and multidisciplinary teams of clinicians, genetic pathologists, bioinformaticians, and scientists.

Slide 13:

In conclusion, the implementation of massively parallel sequencing means that sequencing has become a commodity technology: the exact technology used is not as important as what you want to do with it and whether it will do what you want.

An important consideration for clinical implementation is how much of the genome is sequenced, how much of the sequence is interrogated, and how much of that is reported. In disorders with low numbers of high penetrance genes – e.g. BRCA1/2, the technology will allow markedly increased throughput. For disorders with a moderate number of disease genes, panel testing will probably be important for a little while yet, but ultimately it may be more efficient to perform exome sequencing.

Whole genome sequencing is not a current widespread application but may become so in the medium to long term.

Slide 14: References

Slide 15: Disclosures

Slide 16: Thank You from www.TraineeCouncil.org

Thank you for joining me on this Pearl of Laboratory Medicine on "Massively Parallel Sequencing." I am Melody Caramins.