Genome Biology

MEETING REPORT

Faster sequencers, larger datasets, new challenges

Christopher E Mason* and Olivier Elemento*

Abstract

A report on the Advances in Genome Biology and Technology (AGBT) meeting, Marco Island, Florida, USA, 15-18 February 2012.

The Advances in Genome Biology and Technology (AGBT) meeting this year maintained its reputation for being one of the few international scientific conferences that places revolutionary technology and groundbreaking science on an equal footing. Throughout the meeting, attendees had the opportunity to learn about the latest scientific research fueled by genome technologies while being exposed to many technological breakthroughs that may soon power the next generation of sequencers.

New technology

At this year's meeting we saw the emergence of several new sequencing platforms, and manufacturers of all of the sequencing platforms announced substantial upgrades for their instruments or chemistries, highlighting longer reads, more reads, higher accuracy, cheaper genomes, and faster assay times. Indeed, manufacturers of at least two different technologies are claiming that they will be able to generate a $1,000 genome by the end of 2012.

The most potentially disruptive and most talked about technology announced at AGBT was undoubtedly Oxford Nanopore's single-molecule sequencing technology. Clive Brown (Oxford Nanopore Technologies) explained that their technology uses a hemolysin nanopore attached to an application-specific integrated circuit built into a proprietary bilayer membrane; the electrical conductivity changes depending on the specific base that passes through the pore. Brown showed impressive data based on a preliminary version of their technology: reads over 5 kb long with low error rate (less than 4%) and fast sequencing speed (150 Mb per hour). The most striking moment of his talk was when he showed the MinION, a sequencer that can be powered by a USB port on a laptop. Although the MinION will be capable of sequencing only 1 to 5 Mb, the desktop version of the technology, called GridION, will be able to sequence a human genome at 30X coverage within 15 minutes for less than $1,000. The company plans to release these instruments to a few early-access sites in the middle of this year.

Not wanting to be overshadowed by Oxford Nanopore's announcement, several other groups introduced innovative products. John Healy (GNUbio) presented a new cartridge-based desktop sequencer, primarily targeted at amplicon sequencing. According to the company, the GNUbio instrument weighs only 16 kg and can sequence a 300 gene panel in less than 3 hours, thus making it a serious contender for clinical applications. During the Pacific Biosciences (PacBio) User Group Meeting, Kevin Jacobs (National Cancer Institute) presented the output from PacBio's new chemistry, called C2. Among several improvements, the C2 chemistry has a new polymerase with higher processivity and photodamage-resistance, which means that it can sequence more reads from different size ranges at a higher accuracy. Jacobs showed that C2 provides a two- to three-fold increase in sequencing capacity over the previous (C1) chemistry, meaning that each single molecule, real-time (SMRT) cell can now yield 100 to 140 Mb of sequence. In addition, the average read length has increased from 1.5 kb to 3.1 kb, and the maximum read length now reaches 13 to 15 kb. The long read lengths will continue to make Pacific Biosciences an attractive complementary platform for genome assembly projects.

Geoff Smith (Illumina) announced more details on the Illumina HiSeq 2500, which includes a 'fast output' option to generate a human genome in 27 hours with 40X coverage. In addition, data from a MiSeq run showed less than 1% error using 2 x 250-bp paired-end (PE) reads, and it is anticipated that 2 x 400-bp PE reads will be available this year. These improvements and fast turnaround time are pushing the MiSeq towards clinical applications and, to that end, Illumina also presented whole genome sequence data from a formalin-fixed, paraffin-embedded sample.

Finally, Life Technologies announced improvements to their 'post-light' semiconductor sequencer. The Personal

*Correspondence: chm2042@med.cornell.edu, ole2001@med.cornell.edu
Weill Cornell Medical College, Institute for Computational Biomedicine, Department of Physiology and Biophysics, 1,305 York Avenue, New York, NY 10021, USA

Genome Machine (PGM) can now support 2 x 200-bp PE reads, and their latest sequencer, the Ion Proton, can sequence a human exome for $500. Their upcoming Proton II is expected to bring the cost of exome sequencing down to $150 per sample by the end of the year, and that of whole genome sequencing to $1,000 per sample.

New biology

Besides the technology announcements, many talks highlighted how high-throughput sequencing keeps revealing previously unknown biological phenomena. Tom Gingeras (Cold Spring Harbor Laboratories) showed using data from the ENCODE project that genome transcription is more complex than currently appreciated. Using RNA sequencing (RNA-seq) profiling in 15 cell types and analyzing mRNAs purified from different cellular compartments, he identified up to 15,000 novel multi-exon transcripts in intergenic regions and up to 8,000 antisense transcripts. Altogether he estimated that up to 80% of the genome has evidence of transcription. Although a significant fraction of the transcripts he discovered are expressed at a very low level in a given cell type, these transcripts are frequently expressed much more highly in other cell types. Moreover, compared with more highly expressed transcripts, low-copy-number transcripts are frequently nuclear and localized to specific nuclear compartments.

Another striking talk was from Vivian Cheung (University of Pennsylvania), who used high-throughput sequencing to look for differences between DNA and RNA in human B cells. She identified robust differences between DNA and RNA at more than 10,000 exonic sites. Many of these sites were in coding exons (over 40%), and over 70% of these led to a predicted amino acid substitution; mass spectrometry showed that DNA-RNA differences are indeed present in some translated proteins. The functional relevance of these results is unclear, as are the underlying mechanisms. Cheung showed that RNA-specific adenosine deaminase (ADAR), an enzyme that can perform A-to-G RNA editing, is not responsible for most of the DNA-RNA differences, since knocking down the enzyme reduced A-to-G editing but had no effect on the majority of DNA-RNA differences. Nonetheless, this study demonstrates once more how assays with single nucleotide resolution can challenge dogmas, including ones about the faithful transmission of information between DNA template and messenger RNA.

In another captivating talk that echoed Barbara McClintock's work, Michel Georges (University of Liège) showed that color-sidedness in cows is probably caused by DNA segments translocating between chromosomes, and that circular DNA bubbles probably mediate the translocations. Other talks showed that scientists are inventing creative new ways to mine sequencing data; for example, Jesse Gray (Harvard University) discovered that steady-state RNA-seq read counts across gene bodies and introns can provide reliable estimates of intron processing times and other transcription kinetic parameters. As an example, he calculated that during intron processing, lariat formation takes 2 minutes on average, exon ligation takes 1 minute and intron degradation takes 30 seconds.

Several talks illustrated how use of high-throughput sequencing for clinical applications is maturing and expanding. Joseph Boland (National Cancer Institute) revealed that his laboratory can now sequence an exome in a day using five IonTorrent PGMs. Although the runtime does not include exome capture or library preparation, this is a definite step towards rapid clinical characterization of an exome. John McPherson (Ontario Institute for Cancer Research) reported results from a pilot study for cancer diagnostic testing using targeted resequencing of 19 actionable cancer genes (that is, potential drug targets), using PacBio and Sequenom technologies. Interestingly, PacBio identified several more actionable mutations than Sequenom, because of the former's improved ability to sequence entire genes and its random error profile. Geoff Smith also described the use of the Illumina MiSeq platform for clinical microbiology. Specifically, using case studies from neonatal intensive care and cesarean section infections, he showed how sequencing of bacterial isolates can quickly and accurately identify infections that are part of an outbreak and can predict antibiotic resistance.

However, several talks outlined challenges associated with clinical sequencing, with the accurate clinical-grade assessment of genetic variants being a recurrent theme. Through a comprehensive literature curation effort, Rong Chen (Stanford University) reported the identification of 12 replicated and independent single nucleotide polymorphisms that predict type 2 diabetes across multiple ethnic groups. Heidi Rehm (Harvard Medical School) provided an accredited clinical testing laboratory perspective for using sequencing in clinical applications. She commented on the necessity for a diagnostic laboratory to continuously add new variants as they are being discovered, illustrating this claim with the remark that 10 to 20 new BRCA1 variants are found each week.

Finally, several talks described novel computational techniques designed to address some of the analytical and interpretation challenges associated with the deluge of sequence data. Lior Pachter (University of California, Berkeley) described a promising new algorithm for streamed short read analysis. His algorithm, called eXpress, quantifies transcript expression levels by probabilistically assigning reads to transcripts on the fly, without the need to create and store extremely large alignment files. Samuel Levy (Scripps Translational

Research Institute) described a novel approach called DA-AM for identifying insertions and deletions (indels). Indel detection from short reads is a notoriously difficult task and current approaches such as the Genome Analysis Toolkit (GATK) can identify only short indels. Levy estimated that their algorithm identifies indels that are two- to eight-fold longer than indels detected by GATK. For genome assembly, Michael Schatz (Cold Spring Harbor Laboratories) presented the 'Metassembler' algorithm, which improves genome assembly using the output from multiple disparate assembly programs. Finally, for data storage, Eugene Yaschenko (National Center for Biotechnology Information) presented the CRAM format for storing reads, which simplifies the quality scores for increased storage efficiency.

In short, the meeting achieved the difficult goal of capturing a precise snapshot of a fast-moving and truly exciting field. The emerging picture is that genomics as a field is partly maturing - for example, with increasing use of sequencing in the clinic - and partly accelerating - for example, with the $1,000 genome just around the corner. Yet it seems that the tools to contextualize, understand, and predict the effects of the many molecular changes revealed by sequencing are only beginning to blossom.

Published: 27 March 2012

doi:10.1186/gb-2012-13-3-314
Cite this article as: Mason CE, Elemento O: Faster sequencers, larger datasets, new challenges. *Genome Biology* 2012, **13**:314.