Genome **Biology**

MEETING REPORT

'Next-generation' sequencing becomes 'now-generation'

Daniel E Neafsey* and Brian J Haas

Abstract

A report on the Advances in Genome Biology & Technology conference, Marco Island, USA, 2-5 February 2011.

In the last *Genome Biology* report on this annual conference, Manolis Kellis noted in 2004 that sequencing had reached a mature state, in which the research community 'took for granted that we can sequence, assemble, and align complete genomes' (*Genome Biol* 2004, **5**:324). The technological innovations of the next few years, which shifted sequencing from the Sanger-style long reads used to complete the first human genome to short 'next-generation' reads, required a second wave of innovation to prepare samples for sequencing and deal with the new format of the resulting data. Finally, as attendees of the Advances in Genome Biology & Technology conference learned in February, the field can once again begin to take for granted our ability to effectively produce and process the copious sequencing data that contemporary sequencing technologies more affordably provide. The maturation of these technologies creates interesting new analytical applications for sequencing.

Upstream: innovation in sequencing sample preparation

Many presenters highlighted innovative approaches to simplify sequencing library preparation, expand the range of samples eligible for sequencing, or limit sequencing to specific genomic regions. In describing the Wellcome Trust Sanger Centre's sequencing pipeline, Harold Swerdlow (Sanger Centre, Hinxton, UK) detailed the degree to which amplification-free Illumina library construction reduces the effect of GC composition on sequencing coverage. Andi Gnirke (Broad Institute, Cambridge, USA) described approaches by which base composition coverage bias can be minimized during the amplification phase of Illumina library construction, for instances where circumstances do not permit amplification-free libraries.

Hybrid selection has become a widely adopted means by which to selectively sequence just the exonic portion of the human genome, or other specific regions of interest. By designing oligonucleotides complementary to targeted regions and then hybridizing those oligonucleotides with genomic DNA on a chip or in solution, significant enrichment in sequencing coverage of the resulting captured DNA may be achieved. Many talks reported use of this technology, such as those from Obi Griffith (Lawrence Berkeley National Laboratory, Berkeley, USA) on breast cancer pharmacogenomics and Donna Muzny (Baylor College of Medicine, Houston, USA) on the use of exon or regional capture at Baylor to characterize mutations associated with tumors, autism, and the 1000 Genomes Project, using Illumina or SOLiD sequencing. Hybrid selection is also beginning to be used for sequencing disease genomes from clinical samples - for example, hepatitis C virus (Reinhold Pollner, Gen-Probe, San Diego, USA). Despite the falling cost of sequencing, selective sequencing of genomic regions of interest will probably remain a key application into the future.

Downstream: innovation in data handling and analysis

The utility of cheap, abundant sequencing data is amplified by the research community's growing ability to effectively use and analyze the data. The short read nature of current generation sequencing necessitates significantly higher sequencing coverage than long reads for most applications, and so analysis algorithms and hardware must be capable of dealing not only with the shortness of the reads, but also their extreme abundance. Just as early adopters of the Pacific Biosciences RS sequencer have begun to report increasingly long reads from that new sequencing platform, the community seems to have become adept at dealing with the challenges of short reads. Steven Salzberg (University of Maryland, College Park, USA) described a collection of software tools for short-read alignment and analysis

*Correspondence: neafsey@broadinstitute.org
Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

BioMed Central

(Bowtie, TopHat, and Cufflinks). Bowtie belongs to a new breed of alignment tools that use the Burrows-Wheeler transform, which can compact a human reference genome assembly into as little as 1.1 GB of memory such that it allows ultra-fast mapping of short reads to the reference.

Although mapping of short reads can benefit from reduced hardware requirements, the hardware needs for short read assembly continue to grow, especially as short read assembly algorithms begin to tackle larger genomes. David Jaffe (Broad Institute) described the algorithmic and computational challenges of generating 'good cheap genome assemblies' as implemented in the new ALLPATHS-LG assembly software. Through a combination of this new assembly software, a large memory server (512 GB RAM), and a specialized laboratory recipe for genome sequencing involving Illumina paired fragments, Jaffe reported on being able to generate draft assemblies for 15 vertebrate genomes with quality approaching that derived from capillary-based sequencing.

Not having access to high performance computing resources can be a serious impediment to working with next-generation sequencing data, and for many, cloud computing is becoming an attractive solution. Toby Bloom (Broad Institute) described her experiences in migrating the Broad Institute's next-generation sequence analysis pipeline to the Amazon cloud. Cloud-based analysis presented certain difficulties, most notably a need to keep moving data around within the cloud to match disk storage and performance with targeted computing resources, but as cloud computing services continue to evolve and improve it will become a viable and effective analysis solution for small sequencing centers.

Whether one computes locally or uses the cloud, data processing and analysis workflows can be complex to navigate and maintain. James Taylor (Emory University, Atlanta, USA) described how the Galaxy workflow system can help. The new development of the Galaxy Tool Shed, akin to an app store for Galaxy, is destined to further popularize the system among the growing community of users.

New sequencing-based discoveries and applications

Now that genome sequencing data are not only inexpensive but intelligible, the question many seemed to be wrestling with at this conference was how to make the data useful outside a research context. Sequencing instrument manufacturers are doing their part to expand the territory of their technology to outside of the research laboratory. The Ion Torrent Personal Genome Machine from Life Technologies, which debuted at last year's conference, has now been deployed at a number of sites and boasts an extremely short run time (only 2 hours) for rapid data generation. Illumina debuted their MiSeq

machine, which costs significantly less than their other sequencing instruments and is also capable of producing sequencing data in hours rather than days.

Several cancer-themed talks discussed the potential for sequencing to have an impact on disease treatment and prognosis. David Craig (TGEN, Phoenix, USA) discussed data from a clinical trial designed to discover clinically actionable features of breast cancer patients using sequencing, but noted that sequencing and analysis of a patient's genome and transcriptome data required 6 weeks of time. Richard Weinshilboum (Mayo Clinic, Rochester, USA) described functional validation of genome-wide association study (GWAS) signals, in this case for breast cancer. He highlighted the utility of such studies for identifying markers associated with treatment response or tumor sensitivity to drugs such as aromatase inhibitors and selective estrogen-receptor modulators, but noted that the research investment to discover such markers can be massive (30,000 subjects studied over many years) and might not be broadly replicable for diverse diseases and drugs. Eric Boerwinkle (University of Texas School of Public Health, Houston, USA) also described the challenges imposed by follow-up investigations of GWAS hits. Boerwinkle and collaborators investigated a single nucleotide variant that conferred a relative risk of 1.3 for atherosclerosis. Although small, this relative risk would boost the 10-year risk of coronary heart disease from 15% to 21% in a typical patient, and could lead to a different treatment regimen for approximately 10% of atherosclerosis patients if physicians were to commonly have access to genotype data for this locus. Carlos Bustamante (Stanford University, Stanford, USA) pointed out that current GWASs are biased towards European populations, and the ancestry of a candidate disease-correlated marker in the genome must be taken into consideration before the data can be considered clinically relevant.

Several talks addressed the potential benefits and pitfalls of personalized genome data. James Lupski (Baylor College of Medicine) described his efforts to use whole genome sequencing to successfully hunt down the genetic variants responsible for a rare neuropathy that has afflicted him and his family, and points out that genomic data from one's relatives can be much more useful in such applications than data from the population at large. Joe Beery (Life Technologies, Carlsbad, USA) described a frustrating 14-year journey of working with medical professionals to diagnose and treat a mysterious illness in his twin children. Whole-genome sequencing of his twins using SOLiD technology identified mutations responsible for dopa-responsive dystonia, a pharmacologically treatable disease.

Though these examples plainly illustrate the potential value of personalized genome sequencing, clear policies

on access and use of this information by both patients and clinicians is essential. Ellen Wright Clayton (Vanderbilt University, Nashville, USA) advanced the notion that it is imperative to develop a policy framework and decide when genetic data are ready for prime-time in medicine. As sequencing data become increasingly cheaper, ubiquitous, and informative, clinicians and the public will need to be made aware that, in addition to the genome, a multitude of other factors, such as the physical environment, the microbiome, epigenome, and pleiotropy, can have complex roles in sculpting phenotype.