Meeting report

Biology of Genomes: making sense of sequence

Daniel G MacArthur

Address: Human Evolution, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
Email: dm8@sanger.ac.uk

Published: 12 June 2009

Genome Medicine 2009, **1**:61 (doi:10.1186/gm61)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/6/61

Abstract

A report on the Biology of Genomes meeting held at Cold Spring Harbor Laboratory, NY, USA, 5-9 May 2009.

Introduction

The Cold Spring Harbor Laboratory Biology of Genomes meeting is one of the most eagerly awaited events in the genomics calendar, and this year's meeting [1] did not disappoint: participants were treated to four days of cutting-edge research on a diverse array of topics. This report focuses on the major themes of the meeting relevant to the field of medical genomics.

An explosion of sequence data

The single dominant message emerging from this year's meeting was simple: advances in DNA sequencing technology are now enabling the generation of biological data at a frightening (and accelerating) pace. Increasing sequencing capacity promises rapid advances in biological understanding, but it also brings tremendous challenges in terms of storing, disseminating and analyzing vast quantities of data.

Both the power and the challenges of large-scale sequencing data were evident in the first session of the meeting, which focused on cancer genomics. Several speakers in this session, including Mike Stratton (Wellcome Trust Sanger Institute, Hinxton, UK), Elaine Mardis (Washington University, St Louis, USA) and Gad Getz (Broad Institute, Cambridge, USA), discussed progress in the use of large-scale sequencing to develop comprehensive catalogs of the genetic changes underlying cancer progression.

The general strategy is to generate sequence data from both tumor samples and normal tissue from the same patient; genetic differences between the two samples represent candidates for somatic changes occurring during cancer progression. Stratton presented results from low-coverage sequencing of 24 breast cancer genomes, illustrating the power of this approach for the detection of structural variants (SVs); Mardis and Getz both presented high-coverage sequencing of smaller numbers of cancer samples for combined analysis of SVs and smaller-scale genetic variation.

These approaches have successfully generated high-resolution snapshots of genetic variation in tumors, but many challenges remain. For instance, there are many different sources of sequencing artifacts, and Getz emphasized the need for very careful control of false positive rates to ensure that the list of candidate somatic changes is as reliable as possible. In addition, interpreting the functional effects of variants that fall outside protein-coding regions remains very difficult, and discriminating between mutations underlying cancer progression ('drivers') and changes resulting from a general decrease in genomic stability ('passengers') requires variants in many cancers to be assayed to look for those present multiple times. This task will be eased by the generation of whole-genome sequence data on hundreds of tumors and matched normal samples; Mardis noted that Washington University plans to sequence samples from 150 cancer patients over the next 12 months.

Although the number of cancer genome sequencing projects currently underway is impressive, when it comes to sheer scale it is hard to compete with the 1000 Genomes (1KG) Project [2]. This massive international collaboration aims to

generate a near-comprehensive catalog of human genetic variants with a frequency above 1% by performing whole-genome sequencing on some 1,500 individuals from Europe, Asia and Africa.

Early results from three pilot projects conducted by the 1KG consortium were presented by Gonçalo Abecasis (University of Michigan, Ann Arbor, USA). The pilot projects are ambitious undertakings in their own right: low-coverage (approximately 4X) whole-genome sequencing of 180 individuals, very high-coverage (over 30X) of six individuals, and targeted resequencing of 1,000 randomly selected genes in several hundred individuals. These analyses have already contributed substantially to the catalog of human genetic variation, identifying 21.7 million single nucleotide polymorphisms (SNPs; 11.2 million novel), 400,000 short insertion/deletion variants and over 4,000 larger SVs. These numbers will only increase as the project enters its main phase; the participants have committed to sequencing 1,200 low-coverage genomes by the end of 2009.

The immediate utility of 1KG data for researchers was neatly illustrated by several other presentations at the meeting. Gil McVean (University of Oxford, UK) demonstrated that 1KG sequence data could be used to increase the power of existing genome-wide association study data through the use of genotype imputation, while Michael Snyder (Yale University, New Haven, USA) and Tony Kwan (McGill University, Montreal, Canada) have already used early release 1KG data to look for genetic variants associated with variation in transcription factor binding and gene expression, respectively.

The data being generated by new sequencing technology extend well beyond human genomic DNA. To provide just two examples from the meeting: Stephen Montgomery (Wellcome Trust Sanger Institute, Hinxton, UK) presented the use of RNA sequencing for identifying variants associated with variation in gene expression and alternative splicing, while Claire Fraser-Liggett (University of Maryland, USA) described sequence-based exploration of microbial communities living on and in the human body.

Converting sequence data into meaningful information

The second major theme from the meeting was the need for diverse approaches for generating biological meaning from sequence data. This need grows ever more urgent as mountains of data generated by new sequencing technology begin to accumulate, and as we move into the era of personal genome sequencing.

One important task is to determine precisely which regions of the human genome are actually functional, allowing variants found in those regions to be prioritized for follow-up. Several approaches to functional annotation were presented at the meeting. Michele Clamp (Broad Institute, Cambridge, USA) and Adam Siepel (Cornell University, Ithaca, USA) both presented on the use of comparative genomic data from nearly 30 mammalian species to highlight regions of strong evolutionary conservation. Rick Myers (HudsonAlpha Institute, Huntsville, USA) described the integrated and collaborative approach taken by the ENCODE Project Consortium, which ultimately aims to characterize all of the functional elements in the human genome. The extension of detailed functional annotation into non-coding regions is particularly crucial; David Goode (Stanford University, Stanford, USA) argued on the basis of evolutionary constraint that over 90% of the functional variation in any individual human genome lies outside protein-coding regions.

Another important goal is to characterize the genetic architecture of human diseases and complex traits, moving beyond the common SNPs that have formed the backbone of recent genome-wide association studies. Peter Donnelly (University of Oxford, UK) presented an analysis of large-scale copy number variations (CNVs) in the large Wellcome Trust Case Control Consortium set of common disease and control cohorts, laying out the daunting technical challenges of genotyping CNVs and the risk of false positive associations resulting from artifacts. More positively, he also suggested that (contrary to previous reports) most common easily assayable CNVs are actually well captured by existing SNP chips.

Rare and *de novo* genetic variation is another currently poorly surveyed region of the human genetic landscape. Jonathan Cohen (University of Texas Southwestern Medical Center, Dallas, USA) described rare variants associated with lipid levels and noted the benefits of surveying multiple populations for rare variant discovery. Philip Awadalla (University of Montreal, Canada) described a resequencing study of 401 synaptic genes, revealing an excess of deleterious *de novo* mutations in these genes in schizophrenia and autism patients. In addition to emphasizing the power of large-scale resequencing of patients and controls to identify rare disease-associated variants, Awadalla sounded a cautionary note for researchers analyzing cell-line samples: 13 of 28 putative *de novo* variants were determined to be cell-line artifacts, highlighting the need for storing original blood-derived DNA from patients for validation.

Conclusions

The presentations at the meeting illustrated the growing power of new sequencing technologies to uncover disease-related genetic variants, as well as highlighting several important challenges: the management of very large datasets, the careful analysis required to avoid systematic artifacts, and the need for the integration of multiple data sources to guide biological interpretation. In particular, although considerable attention has been focused on the

potential of structural variants and rare variants for disease association, it is clear that the detection, validation and clinical interpretation of both of these classes of variants remain problematic. We clearly have much more work to do before our ability to make biological sense of sequence data - and design clinical interventions accordingly - advances to match our ability to generate such data.

Finally, the breadth and quality of the research presented at the Biology of Genomes meeting continues to impress, and it is likely that the already fierce competition for the limited places at the meeting will continue to intensify; those interested in the 2010 meeting would be well advised to register early! As per CSHL meeting policy all presenters have granted permission for the description of their work in this article.

Abbreviations

1KG, 1000 Genomes; CNV, copy number variation; SNP, single nucleotide polymorphism; SV, structural variation.

Competing interests

The author declares that he has no competing interests.

Author information

Daniel MacArthur is a postdoctoral scientist at the Wellcome Trust Sanger Institute, and writes about genetics and genomics at Genetic Future [3].

References

1. **2009 Cold Spring Harbor Laboratory Biology of Genomes meeting** [http://meetings.cshl.edu/meetings/genome09.shtml]
2. **1000 Genomes Project** [http://www.1000genomes.org]
3. **Genetic Future** [http://scienceblogs.com/geneticfuture/]