Meeting report

Multi-genome biology

Thomas A Down

Address: Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. Email: td2@sanger.ac.uk

Published: 10 February 2006

Genome ***Biology*** 2006, **7:**305 (doi:10.1186/gb-2006-7-2-305)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/2/305

A report on the Genome Informatics meeting held at Cold Spring Harbor Laboratory, Cold Spring Harbor, USA, 28 October-1 November 2005.

Halloween 2005 saw the fifth of the annual genome informatics meetings organized jointly by the Wellcome Trust and Cold Spring Harbor Laboratory. These meetings began in the rush of excitement surrounding the publication of the draft human genome sequence in 2001. Since then, bioinformaticians around the world have developed a wide range of high-throughput techniques and pipelines. With new methods and improved computer hardware in place, gigabase-scale genome sequences no longer seem quite as intimidating as they once did. The production of large-scale biological data continues apace, however. The rate of genome sequencing continues to increase, while at the same time more and more biological assays - capturing variations in the genome sequence, and information about biological function - are scaled up to offer genome-wide information. An important aspect of this meeting was the range of different data types that could be associated with the genome sequence. A lot of progress has been made in this direction in the past five years, but exciting challenges remain, both in the management of high-throughput datasets, and in integrating them to give a coherent picture of how cells and organisms work.

Genomes and sequences

Genomes, sequencing, and sequence analysis still took center stage for many - but not all - of the sessions. New methods for sequencing and assembling genomes are still being actively developed. In the genome-assembly field, many researchers are focusing on methods based on efficient data structures called string graphs, which efficiently record the set of overlaps between short sequence fragments. Michael Brudno (University of Toronto, Canada) described an experimental assembler based on these methods, and we also heard from David Messina (Washington University, St. Louis, USA) about techniques to automate the practice of using hand-finished clones of genomic sequence to evaluate whole-genome assemblies, including about a method called Semblance, which checks assembled shotgun sequence for consistency with finished clones.

There are many scientifically or commercially important organisms that have yet to be sequenced. Starting a new genome project often presents new challenges and demands new approaches: there has been significant interest in sequencing the maize genome (*Zea mays*), but its large size - due primarily to large repetitive regions - makes obtaining a whole genome sequence both challenging and costly. Brad Barbazuk (Donald Danforth Plant Science Center, St Louis, USA) presented two different strategies for selectively cloning gene-rich regions while excluding the repetitive and highly methylated portions. Combining these approaches could give a useful view of the maize gene set for a fraction of the cost of a complete genome sequence.

The proliferation of genome sequences has brought comparative genomics to the fore. The keynote address by David Haussler (University of California, Santa Cruz, USA) described a project that aims to reconstruct complete genome sequences from ancestral species that were branch points at the base of the tree of mammalian evolution, a project made possible by the large set of low-coverage mammalian genome projects currently underway. This kind of effort hints that future comparative genomics methods - both sequence aligners and analysis tools - will need to handle large sets of sequences but also need to deal gracefully with the kind of missing-data problems posed by fragmentary assemblies resulting from low-depth shotgun sequencing.

The core problem for most workers in comparative genomics remains the identification of sequences - especially noncoding sequences - that are under selection. Gerton Lunter (University of Oxford, UK) presented a promising new technique, which focuses on the pattern of gaps in alignments rather than the pattern of substitutions. While any sequence under selection is a potential subject of interest, it is perhaps understandable that a lot of attention is focused on the extreme end of the conservation spectrum: the highly conserved vertebrate sequence elements. Two such sets were discussed: Haussler described a set of elements that are perfectly conserved in human, mouse and rat, and Gayle McEwan (MRC Biostatistics Unit, Cambridge, UK) a set that is strongly conserved between human and fish. In both sets many of the elements are in non-protein-coding regions, and at least some of them might be enhancers or other regulatory elements. Both groups have now found paralogous elements - cases where one of these elements has been duplicated in the genome and both copies have remained under strong selection.

There are still interesting unexplored areas of single-genome sequence analysis. Splice sites and other signals influencing gene splicing have been a key interest of computational biologists for many years, not least because of their importance when developing eukaryotic gene-prediction methods. In recent years, the focus has switched from the splice sites themselves to splice enhancers and suppressors in the surrounding sequence. Rodger Voelker (University of Oregon, Eugene, USA) looked at the co-location of short DNA motifs around annotated splice sites in the human genome and found that they fell into two distinct clusters, raising the interesting possibility that there might be two distinct major classes of intron.

Interest in epigenetics is also increasing, and although assays for epigenetic information are still being scaled up, a large amount of data is available. Paul Flicek (EMBL European Bioinformatics Institute, Cambridge, UK) talked about methods for analyzing histone modification data across the 1% of the genome chosen for study by the ENCODE consortium. Several consortium members have produced data for a number of cell lines, and at some sites the chromatin state seems to differ significantly between lines. Matthew Vaughn (Cold Spring Harbor Laboratory, Cold Spring Harbor, USA) presented large-scale DNA methylation data from *Arabidopsis*. A key approach to understanding epigenetics is to link it back to the DNA sequence, and John Greally (Albert Einstein College of Medicine, New York, USA) presented some promising results using the SOMBRERO motif-finder to detect motifs associated with the promoters of imprinted genes.

In search of a function

Although a huge amount of information remains to be mined from genomes, sequence analysis was far from the sole theme of the meeting. Many different high-throughput assays are coming of age, with interaction networks being a notable focus. The first challenge to informatics when a new high-throughput assay appears is how to handle and store the data effectively. A number of public databases exist for storing interactions, and Francis Ouellette (University of British Columbia, Vancouver, Canada) introduced a data warehouse system called Atlas [http://bioinformatics.ubc.ca/atlas/] for collecting interactions from multiple databases and presenting them with a consistent interface. There is also much effort going into making network data searchable. Mona Singh (Princeton University, USA) presented ProteinPathGrep [http://www.cs.princeton.edu/~ebanks/pgrep/index.html], a method that allows users to search large interaction databases for novel candidate pathways by finding patterns of interactions similar to known pathways. With large amounts of data safely stored and made accessible, the future of high-throughput biology has to lie in integrating different types of data into a consistent picture: in the introduction to the Panther (Protein analysis through evolutionary relationships) Pathways database [http://www.pantherdb.org/pathway] by Paul Thomas (Applied Biosystems, Foster City, USA), we saw integration between a curated set of interaction data and results from microarray gene-expression experiments.

Imaging has been an important part of biological science for many years. Complete cells and tissues are the end product of the genome, and looking at them directly is an important counterpoint to the analysis of DNA sequence and biochemical data. The talk by David Hall (Albert Einstein College of Medicine) on WormImage [http://www.wormimage.org], a database of *Caenorhabditis elegans* images, highlighted the history of biological imaging: the project involved scanning electron micrographs up to 30 years old. Imaging techniques continue to develop, as shown by David Knowle (Lawrence Berkeley National Laboratory, Berkeley, USA) who described the three-dimensional imaging of gene expression in the *Drosophila* blastoderm. These data are stored electronically from the start, but are also subject to computational analysis, with automatic image-segmentation techniques used to identify nuclei in the developing embryos.

Overall, the meeting showed up many challenges: all areas of bioinformatics are becoming large-scale, high-throughput science, and we are seeing an explosion not just in the volume of data but also in the diversity of data types. But there are also plenty of promising results, especially where a range of data sources can be integrated to give a coherent picture. The next five years of multi-genome bioinformatics are sure to be exciting.

Acknowledgements

I am supported by the Wellcome Trust.