Meeting report

Genomics - from Neanderthals to high-throughput sequencing

Matthew John Wakefield

Address: ARC Centre for Kangaroo Genomics, Bioinformatics Division, The Walter and Eliza Hall Institute, 1G Royal Parade, Parkville 3050, Australia. Email: wakefield@wehi.edu.au

Published: 24 August 2006

Genome ***Biology*** 2006, **7:**326 (doi:10.1186/gb-2006-7-8-326)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/8/326

A report on 'The Biology of Genomes' meeting, Cold Spring Harbor, USA, 10-14 May 2006.

This year, the emphasis of 'The Biology of Genomes' meeting held in May at the Cold Spring Harbor Laboratory was on the plural, with population genomics and comparative genomics taking center stage. The initial challenge of genomics was to efficiently generate the vast amount of data required to assemble genome sequences, and to apply our knowledge to sensible annotation of this deluge of data. Although these tasks are continuing, the new challenges are not only to scale up traditional experiments to previously unimagined levels, but to leverage our knowledge to forge truly new discoveries.

Next-generation DNA sequencing

Genomics is still driven and directed by advancements and limitations of sequencing technology. After a long incubation, the first effects of the next-generation sequencing platforms are now being felt. Two new high-throughput platforms for ultrarapid DNA sequencing were in evidence. The GenomeSequencer 20 (GS20) platform has been developed by 454 Life Sciences (Branford, USA) and uses bead-attached DNA fragments as templates and pyrosequencing, a sequencing-by-synthesis technique in which the number of incorporations of a given base is detected by the intensity of a light signal. The Clonal Single Molecule Array platform from Solexa, based in Little Chesterford, UK, uses amplification of clusters of DNA fragments on glass slides and sequencing-by-synthesis, adding one base at a time using reversible terminators.

The driving force behind this next generation of sequencing technology is human health, with population studies and medical resequencing the intended applications. The demands of these applications require a 10-100-fold increase in throughput and a tenfold reduction in cost. Jane Rogers (The Wellcome Trust Sanger Institute, Hinxton, UK) presented results from a trial of the GS20 and Solexa technologies at the Sanger Centre, in which both performed successfully in sequencing bacterial genomes and in the sequencing of the human major histocompatibility complex genomic regions cloned in bacterial artificial chromosomes (BACs) - representing a medical resequencing application. Comparisons with clone-based sequencing by the conventional Sanger method showed that both the new platforms have difficulty with rRNA repeats, whereas small genes with strong promoters may be missed from the Sanger-type clone-based sequencing.

Future applications of these technologies were suggested in two talks on work that utilized the GS20 platform. In a study of idiopathic generalized epilepsy, John McPherson (Baylor College of Medicine, Houston, USA) reported on the resequencing for medical purposes of 250 ion-channel genes from 500 cases and 500 controls that identified nearly 900 nonsynonymous single-nucleotide polymorphisms (SNPs), 75% of which were novel. The sequence dataset was generated from pooled samples from 250 individuals that were PCR-amplified for 50 loci and sequenced on the GS20 platform in one run. This approach has uncovered functional variation in known epilepsy genes and it is hoped that it will help identify individuals who carry multiple rare alleles that may interact to cause the disease.

Ramy Arnaout (Brigham and Women's Hospital, Broad Institute and Harvard University, Boston, USA) presented a study using the GS20 platform to investigate the variability in the genomic rearrangements that generate complete immunoglobulin genes during B-cell development. By using multiplex PCR and primers, all rearrangements of V, D and J gene segments in an individual's B-cells could be identified by 200 bp reads generated on the GS20 sequencer. In healthy unimmunized people, variation between individuals was found to be similar to that within an individual over the

several months duration of the study, and this technique is being applied to identifying VDJ profiles in T-cells and B-cells associated with disease and vaccination.

Making sense of SNPs

The most recent inrush of genomic data has come from the HapMap project, which aims to identify and map all the SNPs present in the human population, thus providing an unrivaled set of markers for identifying disease-causing genes. Several talks discussed the analysis methods, applications and technical platforms that are needed to take advantage of these data.

The huge increase in the number of data points that can be used to map disease genes has created new statistical issues in analysis. One approach to mitigating the challenging multiple testing problems inherent in large SNP studies was presented by Kathryn Roeder (Carnegie Mellon University, Pennsylvania, USA), who advocated using weights both in analysis and incorporated into the experimental design. The applications of mapping disease genes by identifying shared haplotypes in case-control studies was presented by Ingileif Hallgrímsdóttir (University of Oxford, UK). This involves inferring co-inheritance alleles from a parent using haplotype data and identifying the haplotypes shared between affected individuals and absent in non affected controls, and provides another method for digging through the rising mountain of genotype data.

The new powerful tools for assessing variation will undoubtedly lead to a dramatic surge in the number of new associations between genes and diseases. The large amount of genotyping data generated by the HapMap project also featured in projects to identify inversion polymorphisms; to map gene-expression variation as quantitative traits; to analyze human demographic history; to detect regions under recent positive selection; and to analyze the effect of sequence variation on splicing.

Functional genomics: an extra dimension

The availability of genome sequences has triggered a wave of functional genomic studies aimed at experimentally verifying and enriching genome annotation and understanding how genome sequence relates to whole-organism biology. In order to deal with this large volume of functional data and integrate it with genomic data, data standards and sophisticated storage methods are required.

The most visually spectacular talk of the conference was a presentation by Angela DePace (Lawrence Berkeley National Laboratory, Berkeley, USA) on the three-dimensional atlas of gene expression in *Drosophila*. The three-dimensional atlas is based on the confocal imaging of many independent gene-expression experiments in whole-mount *Drosophila* embryos. The data are converted to a point cloud coordinate system, which enables virtual integration of the results of the different experiments, and the result is a three-dimensional display of the simultaneous expression patterns of multiple genes. A comparable system was presented by John Murray (University of Washington School of Medicine, Seattle, USA) for *Caenorhabditis,* which has the advantage of a complete cell-fate map on which to integrate the data. Although this approach is still in the early stages of development, similar methods will undoubtedly be able to provide the same sort of information-rich resources for developmental genomics in other model species.

Proteins were not neglected at the meeting. An attempt to develop antibodies to the products of all predicted human genes and use these to analyze tissue arrays of 48 normal and 20 cancer tissues was presented by Mathias Uhlén (Royal Institute of Technology, Stockholm, Sweden) as 'The Protein Atlas'. Annotation by expert pathologists and the availability of the affinity-purified polyclonal antibodies to researchers promise to make this a valuable resource as it grows to cover more of the genome.

Lessons from comparative genomics

The main contribution of comparative genomics has been to identify functional regions of the genome by their sequence conservation. With a multitude of sequenced genomes now available, the ability to analyze changes in these conserved elements among species has expanded. Adam Siepel (Cornell University, Ithaca, USA) presented an analysis of lineage-specific conserved elements in the 1% of the genome covered by the ENCODE project. Using a phylogenetic hidden Markov model program, Siepel has identified elements that cover about 5% of the ENCODE regions. A quarter of these elements vary among the mammalian lineages, with more losses than gains and a large number of gains on the lineage leading to eutherian mammals.

With the evolutionary history of the genome playing such a major role in its current function, easy access to the reconstructed ancestors of current genomes will be an important aid to interpretation. David Haussler (University of California, Santa Cruz, USA) introduced plans for a large multi-group effort to reconstruct the history of the eutherian genome, by modeling both large scale rearrangements and substitutions and computing over all possible evolutionary histories. This could rapidly evolve into a new and valuable resource.

The origin, evolution and function of ultraconserved elements is one of the enigmas to arise from large-scale comparative genomics. Gill Bejerano (University of California, Santa Cruz, USA), described how an ancient short interspersed nuclear repeat element (SINE) has been recruited as both an exon and a distal enhancer. The SINE that is found

as a highly amplified family in coelacanths acts as a distal enhancer of the *ISL1* gene in mammals. In reporter assays this element recapitulates the *ISL1* gene's neural expression pattern. The same SINE also acts as an alternatively spliced exon of another mammalian gene, *PCBP2*. This work provides an interesting example of possible 'exaptation', the situation whereby fortuitously distributed identical elements appear to have been co-opted to provide new multigene regulatory networks for existing genes. On a more practical front, it may also highlight the need for a more cautious approach to the common practice of masking and subsequent removal of large numbers of repeated elements from genomic analyses.

The usefulness of comparative genomics of domestic animals for mapping traits was apparent in several talks. Leif Andersson (Uppsala University, Uppsala, Sweden) presented a study in which melanocyte migration defects were rapidly mapped using dog breeds, highlighting the benefits of population structure, strong selection and closely monitored phenotypes of domestic animals in disease gene mapping. Michel Georges (University of Liège, Belgium) described work on quantitative traits affecting myogenesis in Texel sheep. His identification of a SNP in the 3' untranslated region of the *MSTN* gene that accounts for up to 20% of the muscle-trait variation is of particular interest as it joins an exclusive, but growing, group of mutations that produce a phenotype by disrupting the interactions of microRNAs (miRNAs) with genes. This particular miRNA interaction inhibits the translation of the protein but not the abundance of the mRNA transcript.

Additional species continue to be added to the comparative genomics repertoire. Proving that extinction need not be a barrier to inclusion, Svante Pääbo (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany) showed how the Pleistocene has contributed genomic information from mammoths, cave bears and Neanderthals. The small amount of data available so far indicate that Neanderthals diverged from modern humans (*Homo sapiens*) 400,000 years ago and that there was separation of the gene pools.

Progress on the sequencing and analysis of the opossum genome was presented by Kerstin Lindblad-Toh (Broad Institute, Boston, USA), and among the many meetings of small groups of researchers throughout the conference was one of the platypus genome sequencing consortium, indicating that next year will be an exciting one for comparative genomics.

The power of large-scale comparisons remains the core of genomics. This meeting emphasized that comparisons are beginning to reveal unexpected secrets of the genome and its evolutionary history. In the nearer future, however, there is the promise of using high-throughput capability to assess variation in the human population and gain fresh insights into disease.