Meeting report

Integrative functional genomics

Martha L Bulyk

Address: Brigham and Women's Hospital and Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.
E-mail: mlbulyk@receptor.med.harvard.edu

Published: 24 June 2004

Genome ***Biology*** 2004, **5:**331

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/7/331

A report on the Keystone Symposium 'Biological Discovery Using Diverse High-Throughput Data', Steamboat Springs, USA, 30 March-4 April 2004.

The 2004 Keystone Symposium meeting 'Biological Discovery Using Diverse High-Throughput Data' was organized by David Gifford, Edward Rubin and Richard Young. As the title suggests, the talks at this meeting spanned a wide range of research efforts, many of which combined various types of computational and/or experimental data sources and approaches. It was an outstanding meeting, with many presentations describing new developments and findings. Eric Lander (Whitehead Institute and Massachusetts Institute of Technology, Cambridge, USA) kicked off the meeting with a keynote talk in which he outlined the "audacious goals" of the genomics community: first, to sequence the entire human genome (the only one of these goals that has been achieved so far); second, to identify all the functional elements in the human genome; third, to identify all signatures of cellular responses; and fourth, to identify all common human genetic variation. All the talks at the meeting presented work aimed at achieving some aspect of these goals, either in model organisms or in humans; the work being undertaken ranged from technological to biological.

Genome-sequence analysis and the control of gene expression

Svante Pääbo (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany) presented some interesting findings on human-specific traits. By studying 50 randomly picked olfactory receptor genes, his group found evidence that humans are still losing functional genes. Looking at transcription levels in different parts of human and chimpanzee brains, Pääbo's group has estimated that around 10% of genes exhibit significant transcriptional differences between human and chimpanzee, with greater expression differences being observed in those genes with greater sequence divergence.

The talk that aroused perhaps the most discussion was the presentation by David Haussler (University of California, Santa Cruz, USA) of the identification of 481 ultra-conserved non-protein-coding regions, which are conserved across the human, mouse and rat genomes at a level of 100% sequence identity over at least 200 base-pairs (bp), with the largest region spanning 779 bp. Haussler noted that most of these ultra-conserved regions do not overlap protein-coding regions, and those that do overlap do not extend significantly into the protein-coding regions. Over half of the ultra-conserved regions were found in gene deserts, with many being over 100 kilobases (kb) away from a gene. Interestingly, many of the genes flanking ultra-conserved regions were enriched for annotation with the Gene Ontology (GO) terms 'development' and 'DNA binding'. Many of the genes encoding exonic ultra-conserved regions were involved in DNA or RNA binding, or were ribosomal genes. There were many questions and much off-line discussion after this session regarding what roles the noncoding ultra-conserved regions might serve.

A large number of research groups are currently using phylogenetic footprinting to find non-protein-coding regions of DNA that are most likely to correspond to *cis*-regulatory elements. Edward Rubin (Lawrence Berkeley National Laboratory, Berkeley, USA) presented intriguing results on non-protein-coding regions that are conserved between human and mouse. His group has deleted two large genomic regions, totaling almost 3 million bp, from mice; they found no observable phenotype in mice carrying these deletions and only minor expression differences in the genes surrounding the deleted conserved noncoding regions. It remains to be seen whether these deleted conserved regions either exert an effect on global gene expression or serve some other role in tissues, various other settings or timeframes, or genetic backgrounds that have not yet been assayed.

A number of groups are now using *in vivo* genome-wide location analysis to infer transcriptional regulatory networks; this technique is also known as 'ChIP-chip' and involves chromatin immunoprecipitation (ChIP), followed by hybridization to DNA microarrays to identify the immunoprecipitated DNA. One limitation has been that thus far essentially all yeast ChIP-chip experiments studying a given transcription factor have been performed under just one culture condition. Richard Young (Whitehead Institute and Massachusetts Institute of Technology, Cambridge, USA) announced that his group is currently performing ChIP-chip experiments on 85 transcription factors in at least one of 12 culture conditions in addition to rich medium. These conditions have been selected to correspond to the known roles of the 85 transcription factors in metabolism, stress and development. Young noted that from analysis of the binding of transcription factors to DNA *in vivo*, in rich medium compared to conditions of amino-acid starvation, the transcription factors can be classified into four categories according to their binding properties: condition-invariant, condition-enabled, condition-expanded and condition-altered. Transcription factors that are condition invariant occupy the same set of DNA binding sites independent of the culture condition; condition-enabled transcription factors occupy their sites only in a given culture condition, condition-expanded bind a broader set of sites under particular conditions and condition-altered bind different sets of sites under different conditions. In a later session, David Gifford (Massachusetts Institute of Technology, Cambridge, USA), a close collaborator of Young's, presented progress his group has made in the development of algorithms for discovering regulatory networks of gene modules, using both Young's ChIP-chip data and available gene-expression data. ChIP-chip datasets from analyses of cells from various environmental conditions and also of various kinds of cells in multicellular organisms will help to understand the dynamic nature of interactions between transcription factors and DNA. My own talk followed Young's and presented the *in vitro* protein-binding microarray (PBM) technology that my lab has developed for the highly parallel, rapid characterization of the binding specificities of transcription factors. Comparison of PBM data with ChIP-chip data and analysis of the cross-species sequence conservation of transcription-factor binding sites derived from PBM analysis has allowed the identification of many new putative targets for regulation by yeast transcription factors. We hope that the PBM technology will contribute to the identification of the regulatory targets of transcription factors in various genomes.

Bing Ren (Ludwig Institute for Cancer Research and University of California, San Diego, USA), whose group is part of the Encyclopedia of DNA Elements (ENCODE) Consortium, presented new results of his group's ChIP-chip analysis of RNA polymerase II, TATA-binding protein (TBP) associated factor II 250 ($TAF_{II}250$), and various modified histones in human tissue-cultured cells. The results indicate that the binding of these factors is extremely well correlated with the transcription start sites of genes. Ren's presentation stimulated much discussion about the measurement of transcription levels throughout the genome using various microarray platforms.

Applications of genomics and proteomics

A number of the talks on proteomics focused on technological developments. Ruedi Aebersold (Institute for Systems Biology, Seattle, USA) presented the exciting progress that his group is making towards the quantitative measurement of proteins using mass spectrometry. To achieve this goal, his group is producing ordered peptide arrays, which, when combined with synthetic peptide standards, will allow the absolute quantification of peptide levels. Stephen Burley (Structural GenomiX Inc., San Diego, USA) talked about the significant progress that Structural GenomiX has achieved in high-throughput protein production, crystallization and structure discovery, focusing on kinases as drug targets. Burley stated that their pipeline, which combines structure discovery with combinatorial chemistry, allows them to generate a candidate drug for a particular target in roughly 6-8 months.

Moving to other methods of 'functional genomics', Thijn Brummelkamp (Netherlands Cancer Institute, Amsterdam, The Netherlands) discussed exciting work using short synthetic small interfering RNAs (siRNAs) to perform RNA interference (RNAi) screens on mammalian cells. His group is treating cells with an RNAi library that has been bar-coded with 59-mers, selecting cells that survive a particular stress, using PCR to amplify DNA from survivors, and then hybridizing the amplicons to DNA microarrays to identify which siRNAs allowed survival. They are now screening for siRNAs that are lethal in tumor cells but not in normal cells.

In work that is similarly directed at human health, Kelly Frazer (Perlegen Sciences, Mountain View, USA) presented work that Perlegen has done, using 223 high-density DNA microarrays that cover the entire human genome, to identify 1.6 million single nucleotide polymorphisms (SNPs). Importantly, half of these SNPs were not found in the dbSNP database at the National Center for Biotechnology Information (NCBI [http://www.ncbi.nlm.nih.gov/SNP/]). Frazer stated that Perlegen's capacity is currently 30 million genotypes in just one week, and that many SNPs fall outside of the 10 kb upstream or downstream of known genes. She described Perlegen's work to identify SNPs that differ in groups of individuals exhibiting low or high levels of high density lipoprotein (HDL) cholesterol, and she noted that the company is also examining metabolic syndromes. Kathleen Giacomini (University of California, San Francisco, USA) described exciting discoveries her group has made in the pharmacogenetics of membrane transporters. She and her colleagues found around 680 SNPs in 24 membrane transporter genes; about half were coding SNPs and half of those

were non-synonymous. Out of 80 variants that Giacomini's group followed up by individually expressing the corresponding synthetic variants in cells, 14 were non-functional or had significantly decreased function. Of the 14 corresponding SNPs, 11 were population-specific. Studies like these will be important in understanding the variable clinical responses that different populations have to various drugs.

In addition to SNP analyses, many groups are performing gene-expression analysis in normal and affected individuals with the aim of understanding a wide range of disease states, including cancer and infection by various pathogens. Ron Davis (Stanford University, Stanford, USA) presented important findings his group has made concerning the significant effects that the mode of blood collection from patients can have on the outcomes of subsequent gene-expression analysis. The results indicated that certain methods for blood collection are much more reproducible than others, including some that are currently considered standard methods for blood collection. Davis noted that at times nurses can be resistant to changes in the typical procedures they follow in blood collection, but that with proper training and further technological developments currently underway, more sensitive, reproducible results could be attained.

It was apparent from this meeting that the many high-throughput genomic and proteomic approaches that are now available are generating complementary datasets that are frequently being integrated into analyses aimed at understanding the functions of various portions of the genome and of genomic and proteomic networks. Appropriately, towards the end of the final presentation of the meeting, Leroy Hood (Institute for Systems Biology, Seattle, USA) noted that "data space is infinite", and that "hypothesis-driven perturbations must illuminate those dimensions of data space that are biologically relevant". Altogether, the work presented in this meeting will help to attain the "audacious goals" that Lander outlined in his introductory talk.