Meeting report

Functional genomics: strict tempo and hierarchical vocabularies

Thomas Preiss

Address: Molecular Genetics Program, Victor Chang Cardiac Research Institute, 384 Victoria Street, Darlinghurst, Sydney NSW 2010, Australia. E-mail: t.preiss@victorchang.unsw.edu.au

Published: 16 January 2004

Genome ***Biology*** 2004, **5:**307

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/2/307

A report on the Comparative and Functional Genomics Workshop, Hinxton, UK, 2-5 November 2003.

Approximately 80 scientists converged recently on the Wellcome Trust Genome Campus near Cambridge, UK, to discuss what lies 'Beyond the identification of transcribed sequences' (BITS). This phrase used to serve as the title of a successful meeting series, which was a forerunner to this current workshop. As in previous years, participants were exposed to a dense program of presentations, which discussed ongoing efforts to provide centralized genomics resources for the research community, as well as studies that have made use of these resources in innovative ways. Thus, resource providers and database curators could experience at first hand how their efforts bear fruit in the research community, and 'users' were given an update on the latest services that will become available to them.

Providing resource collections and data repositories

A core genomics service for wet-lab scientists continues to be the provision of verified cDNA clones and similar gene-based reagents. Bernhard Korn and Uwe Radelof (German Genomics Resource Centre, RZPD, Berlin/Heidelberg, Germany) presented the range of products and services available from the German Genomics Resource Centre [http://www.rzpd.de], which covers several model organisms and includes a growing set of shuttle vectors and RNAi knock-down constructs. Osamu Ohara (Kazusa DNA Research Institute, Chiba, Japan) maintains a set of long cDNAs (more than 4 kb) of human and mouse origin, enigmatically termed KIAA cDNAs. To aid the functional analysis of KIAA genes, the cDNA sequences have been cloned into expression vectors and a program to raise polyclonal antibodies has begun. Tom Freeman (MRC Rosalind Franklin Centre for Genomics Research, formerly the Human Genome Mapping Project-Resource Centre, Hinxton, UK) talked about ongoing efforts to generate an expression map of the mouse transcriptome and described the center's microarray production and distribution program: all arrays are available free of charge to any UK academic group from the center's website [http://www.hgmp.mrc.ac.uk].

Geoff Hicks (University of Manitoba, Winnipeg, Canada) and Harald von Melchner (University of Frankfurt Medical School, Germany) represented the International Genetrap Consortium (IGTC), which is building a reference library of gene-trap sequence tags from insertional mutations generated in mouse embryonic stem (ES) cells. Through this effort thousands of ES cell lines with disruptions in specific genes are available for making mutant mice. Many of these 'mouse patients' may end up being examined in the German Mouse Clinic headed by Martin Hrabé de Angelis (GSF Institute of Experimental Genetics, Neuherberg, Germany). More than 3,000 mice per year can be phenotyped for more than 160 parameters (characteristics) in this facility, which has all its state-of-the-art clinical equipment scaled down to suit mouse anatomy.

Setting up and curating searchable databases for various novel types of genomic information constitutes another valuable service to the research community. In addition to software and hardware issues, the development and popularization of standardized vocabularies for data description is an important issue in this area. ArrayExpress [http://www.ebi.ac.uk/arrayexpress/] is a public repository for microarray data that uses the MIAME annotation standard and is maintained by Alvis Brazma (European Bioinformatics Institute, Hinxton, UK) and colleagues. Winston Hide (South African National Bioinformatics Institute, Cape Town, South Africa) works towards establishing a non-diseased expression profile of the human genome and presented a set of orthogonal gene expression ontologies, termed eVOC [http://www.sanbi.ac.za/evoc/], which integrate anatomical

system, pathology, developmental stage and cell type. Martin Ringwald (Jackson Laboratory, Bar Harbor, USA) spoke on related efforts directed towards maintaining and enhancing the Mouse Gene Expression Database [http://www.informatics.jax.org/]. Greg Elgar (MRC Rosalind Franklin Centre for Genomics Research) and colleagues maintain a comprehensive website providing access to the annotated genome of the puffer fish, *Fugu rubripes* [http://fugu.hgmp.mrc.ac.uk].

Generating new insights

Access to genome-sequence and expression databases allows experts to mine the data to improve annotation and/or establish new relationships between datasets. The talks by Brazma, Nick Luscombe (Yale University, New Haven, USA) and Martin Vingron (Max Planck Institute for Molecular Genetics, Berlin, Germany) all dealt with new ways of relating yeast gene-expression data to information about transcription-factor-binding sites, the latter generated by a combination of chromatin immunoprecipitation with microarray analysis, termed ChIP-chip, and/or *in silico* predictions. Luscombe and colleagues merged genetic, biochemical, and ChIP-chip datasets to compile a regulation network comprising 180 transcription factors and 3,474 target genes. They integrated this network with a time-course of gene expression during the cell cycle and looked for connections. They found that on average, each transcription factor is linked to around 100 target genes, and genes targeted by the same factor(s) tend to be co-expressed. The relationships between the expression of transcription factors and their targets often exhibit more complicated time-shifted or inverted behavior. The resultant network features certain transcription factors as key regulatory hubs and may shift its weight to different hubs depending on cellular conditions. Brazma and colleagues found a good correspondence between ChIP-chip data and *in silico* binding-site predictions, but a much lower correlation between both of these datasets and expression changes in mutant strains, highlighting the complexity of mutant responses.

Vingron added to this picture a significant correlation between the co-occurrence of transcription-factor-binding sites and the vicinity of the corresponding factors in protein-protein interaction networks. His group is also involved in comparative analyses of available genomes to identify regulatory DNA sequences and make them available through the Comparative Regulatory Genomics website [http://corg.molgen.mpg.de]. A similar effort was presented by Thomas Werner (Genomatix Software GmbH, Munich, Germany), who developed a proprietary set of software tools to predict and analyze mammalian gene promoters. One of these tools is ElDorado [http://www.genomatix.de/software_services/software/ElDorado/ElDorado_stb.html], an extended genome annotation software kit that contains over 140,000 experimentally verified or predicted promoters and is tantalizingly provided as 'free access, to some extent'.

Drawing from yeast genome-sequence and expression data, Laurence Hurst (University of Bath, UK) showed that essential genes are clustered, independent of previously recognized clustering of co-expressed genes and of tandem duplications, and that these clusters tend to be in genomic regions of low recombination. Also, tracing the footprints of natural selection in available genomes, Sudhir Kumar (Arizona State University, Tempe, USA) presented evidence that the intensity of gene expression relates inversely to the rate of protein-sequence evolution.

Last, but not least, a number of studies were presented that tied genomics tools and resources with wet-lab approaches. In the worm *Caenorhabditis elegans* the tendency for similarly expressed genes to be linked is taken a step further, in that many are organized into bacteria-like operons. The primary polycistronic transcripts are processed into monocistronic mRNAs with the downstream units receiving a uniform leader sequence, called SL2, by trans-splicing. Tom Blumenthal and colleagues (University of Colorado, Denver, USA) used microarrays to probe for SL2-containing mRNAs and found that as many as 15% of *C. elegans* genes are expressed in at least 1,000 operons, each of which is two to eight genes long. Analysis of these operons demonstrates an enrichment of particular classes of genes, especially mitochondrial genes and the genes encoding the basic transcription, splicing and translation machinery. This suggests that previously unrecognized functional relationships could be found between genes within operons. In my own work, microarrays were used to simultaneously monitor transcriptional and translational changes in yeast. We observed that signal-induced changes in the transcriptome are amplified at the translational level. These results unveil a novel, higher level of coordinated gene regulation.

Again using the yeast model, the power of proteomic approaches was demonstrated by several talks. Michael Snyder (Yale University) reported on the expanding uses of his group's comprehensive yeast protein arrays, which include screens for interacting proteins, lipids, nucleic acids and small molecules, as well as for post-translational modifications and antibody specificity. Daniel Finley (Harvard Medical School, Boston, USA) reported work done in collaboration with Steven Gygi's group (also at Harvard Medical School); they are using a combination of affinity purification with mass spectrometry to further characterize the ubiquitin-proteasome pathway. The use of a strain expressing ubiquitin tagged with six His tags has allowed the purification and identification of 72 ubiquitinated proteins, in the first large-scale, systematic identification of ubiquitin conjugates. This set of proteins is likely to be incomplete, and experiments with proteasome inhibitors are now underway to capture rapidly degraded substrates of this pathway.

Nancy Hopkins and Adam Amsterdam (Massachusetts Institute of Technology, Cambridge, USA) gave talks on the

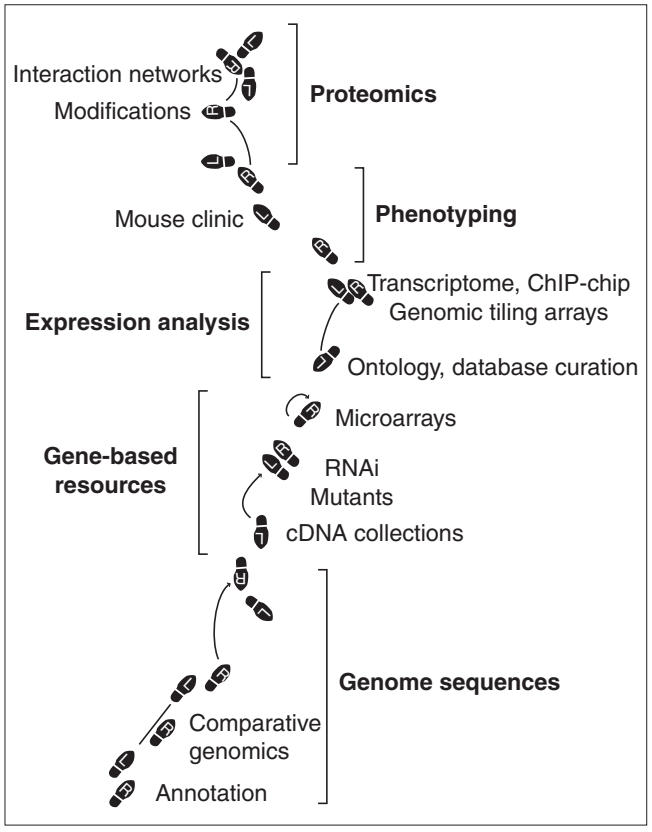

Figure 1
The 'Functional Genomics Waltz'. Some of the research areas that featured prominently at the workshop have been 'mapped' onto a foot diagram of the 'International Waltz Routine'.

ongoing analysis of the results of their large-scale insertional mutagenesis screen in zebrafish. They estimate that they have identified about 25% of the genes that, when mutated, affect fish development in the first five days; 30% of the genes that affect development when mutated give specific developmental phenotypes, while 70% have a relatively non-specific impact. A screen for cancer susceptibility in heterozygous adult fish has also been performed, and has identified about a dozen candidate genes, among them many encoding components of the translational machinery.

On the basis of the entire set of predicted transcripts in the *Drosophila* genome, Michael Boutros (Harvard Medical School and German Cancer Research Center, Heidelberg, Germany) has generated 21,000 double-stranded RNAs and established methods for cell-based high-throughput RNA interference (RNAi) screens that show great promise. Sherman Weissman (Yale University) and Tom Gingeras (Affymetrix Inc., Santa Clara, USA) both use DNA microarrays that interrogate the entire length of human chromosomes 21 and/or 22 to identify binding sites for several transcription factors, as well as to comprehensively identify transcribed sequences. In some ways these 'genomic tiling' studies refer us back to square one and genome annotation, in that large fractions of the detected transcripts derive from chromosomal regions outside known or predicted genes, or are expressed antisense to previously annotated regions. A substantial proportion of identified transcription-factor-binding sites also maps well away from known genes.

In summary, the presentations at this year's meeting demonstrate the maturation of a research area that is now often referred to as functional genomics (Figure 1). An earlier sense of urgency to get beyond merely cataloging transcribed sequences has perhaps been assuaged by an impressive body of studies reporting substantial progress in this regard. Nevertheless, new genome-sequencing data, innovation in high-throughput technologies, and the growing pace in generating complex functional genomic datasets will ensure lively activity across all levels of research in this area for years to come.