Report on the
Twelfth International Workshop on the Identification of Transcribed Sequences: Functional, Expression and Evolutionary Analysis

Held at Washington DC (USA)
October 26–28, 2002

Organized by
T. Werner
S. Weissman
S.B. Hedges
K. Gardiner
Report on the twelfth international workshop on the identification of transcribed sequences: functional, expression and evolutionary analysis

T. Werner, S. Weissman, S.B. Hedges and K. Gardiner

Institute for Biotechnology, Okhranodeschtema (Germany); Yale University School of Medicine, New Haven CT; Pennsylvania State University, University Park PA; Eleanor Roosevelt Institute, Denver CO (USA)

The twelfth workshop in the series "Beyond the Identification of Transcribed Sequences - Functional, Expression and Evolutionary Analysis" was held October 26-28, 2002, in Washington DC and was again sponsored by the US Department of Energy. The seventy attendees discussed a broad range of topics relevant to genomics, gene identification, RNA analysis, and post-genome-sequencing questions of functional assessments. Consistent with developments in the field, several topics received increased attention and some new topics were introduced. In functional analyses, large-scale mutagenesis projects in mouse and fish are expanding rapidly, protein chips are merging on practicality and determining network connections and interactions among genes is being approached from multiple avenues. In expression studies, cDNA arrays remain popular but new applications are addressing additional mRNA characteristics, including stability and translation, detection of large scale basal transcription of apparent intergenic and intron sequences creates new questions, and nonsense mediated decay provides potential new insights into functions of alternative splice variants. Evolutionary topics included new estimates of the timing of increased organismal complexity, patterns of genomic and gene duplication in vertebrates, and detection among primates of human-specific sequences. Summaries of oral presentations follow.

Mutation and SNP analysis

Geoffrey Hicks (Manitoba Institute of Cell Biology) discussed the procedure of tagged sequence mutagenesis using retroviral insertions in ES cells. Currently >4000 lines with single insertion sites are available, with ~30% disrupting known genes. It is anticipated that 10,000-20,000 ES cell lines bearing disruptions of specific genes eventually will be available. Data on these genes can be obtained through a web-based database (http://www.iscelba.ca). This effort is part of a large international consortium on mouse mutagenesis that offers the promise in the near future of providing an array of mouse mutations with coverage comparable to the knockout mutations available for the yeast genome.

Harald von Meckmuller (University of Frankfurt Medical School) discussed progress of the German Gene Trap Consortium in executing similar large-scale mutagenesis of ES cells. Integration sites have been identified in over 7000 clones and include ~2500 known genes. Over 70 germ line chimeras have been generated with over 60% showing a phenotype. By using several gene trap vectors it appears that most genes can be targeted. Data is being deposited in a public database. This and the related efforts discussed above are moving forward at a very impressive rate towards saturation mutagenesis of the mouse genome with its consequent enlightenment at least for the earliest essential functions of most mouse genes.

Hebe de Angelis (GSF Center for Environment and Health, Neuhosterberg) discussed a massive effort to identify interesting dominant and recessive mouse mutants generated by ENU mutagenesis. Currently, 405 confirmed lines have been developed, of which 39 are recessive mutations and 12 show behavioral phenotypes. This group has focussed analysis in part on studies of the notch/jagged signalling system in mice.
Recent findings include the knockout of a novel Delta1 gene allele and its effects on development of the endocrine pancreas as well as a role of the Notch system for inner ear development.

Adam Amsterdam (Massachusetts Institute of Technology, Cambridge, MA) discussed work from Nancy Hopkins group on large-scale mutagenesis in zebrafish. Their protocol uses mouse retroviral vectors as intentional mutagens and results in founder fish with multiple insertion sites. Several subsequent crosses are required to generate homozygous mutants. To date, 550 mutations have been bred to homozygosity and corresponding insertion sites have been cloned for 410 of these. This has identified 350 candidate genes, estimated to be 15–20% of all genes that can be mutated to give a visible phenotype during the first five days of zebrafish life.

Complete SNP analysis for individuals would cause a combinatorial explosion in data (for example, all SNPs from just 1000 individuals would produce 20 billion possible phenotypes). Therefore, Sequenom (San Diego, California) is concentrating on haplotypes rather than individual SNPs and using DNA pooling to reduce the number of assays required. Michael Shi described how SNPs were detected by primer extension followed by MALDI TOF analysis. The integration of array data with normal biological variability allows for a quantitative measure of the proportions of alleles in pooled samples. Common variants may be associated with diseases like diabetes, Alzheimer, HIV resistance, or cardiovascular diseases.

Sherman Weissman reviewed progress in the development of methods using DNA glycosylases to recover most or all mismatch-containing fragments from a pool of DNA duplexes. He further reviewed the Yale Center for Evolutionary and Genomic Studies preliminary applications of genomic tiling arrays to detection of these mismatch-containing fragments and potentially to other issues in mammalian genomics, such as global detection of transcription factor binding sites (Chip-Chip experiments).

Protein analysis

Thomas Joos (Nihon Natural and Medical Science Institute, Tubingen) reviewed the principles of protein microarray technology and discussed applications including the parallel detection of antibodies directed against a number of potential autoantigens, for detection and monitoring the progress of autoimmune disorders in man. This type of array is only in the initial stages of development and application for biomedical purposes but offers similar promise to that obtained with cDNA microarrays.

Michael Snyder (Yale University) discussed his elegant work preparing and analysing arrays containing almost all proteins encoded by 25,000 genes. These arrays have permitted detection of new activities including a number of tyrosine kinases. Other cDNA arrays include detection of proteins binding to specific phosphorylated forms of histone and direct detection of protein-protein interactions. Because the mammalian genome encodes fewer than ten fold more proteins than does yeast, these results offer hope that in the future extensive or nearly complete arrays of mammalian proteins might become available.

Jeremy Simpson (European Molecular Biology Laboratory, Heidelberg, Germany) described work of the Verona group and large-scale localization of protein localization in intracellular structures. OF 1000 cDNAs targeted for analysis, >270 have been successfully localized to one or more of 12 compartments of the cell, with a further ~70 failing to localize due to interference from the green fluorescent protein used as a tag. This system also allows direct monitoring of protein transport through use of a temperature-sensitive mutant of the VSV-G protein, which is blocked in the endoplasmic reticulum at 38°C. Lowering of the temperature to 33°C releases the ER block, and proteins that affect the VSV-G protein transport can be determined by knockdown studies.

Olof Emanuelsson (Stockholm Bitinformatics Center, Stockholm, Sweden) presented a study aiming at the in silico detection of proteins in the peptidome. All peroxisomal proteins containing the Peroxosomal Targeting Signal (PTS1) sequence were identified, with a total of 1552 being collected from SwissProt, from which a total of 33 different PTS1 motifs were defined. Nonperoxisomal proteins containing the PTS1 signal were also collected as a negative control set. In this algorithm, the nine amino acids immediately N-terminal to the PTS1 were analyzed by a combination of neural networks and support vector machines, and the most accepted only if confirmed by one of the 33 known motifs. Further filtering by other predictors (like TargetP) to remove other subcellular compartment specific proteins (e.g., transmembrane proteins) resulted in prediction of only 60 peroxisomal proteins in eight higher eukaryotic genomes, with 80–90% of all matching proteins correctly predicted as peroxisomal proteins.

Mark Gerstein (Yale University) focused on computational aspects of proteomics, in particular on surveying the dead-protein fossil pseudogenes in a number of large eukaryotic genomes. He identified pseudogenes in yeast, worm, fly and human, finding thousands of new ones. Some of the interesting findings were: (1) The population of pseudogenes is not uniform amongst protein families and tends to favor environmental response genes. (2) There is a large number of processed pseudogenes in the human genome mostly stemming from ribosomal proteins. (3) The fly genome has a very small number of pseudogenes, especially compared to its cousin the worm. However, this can be explained through a process of genome decay and one can find numerous broken up pseudogenes in this genome.

Large scale cDNA analysis

Ruth Wolberer from the German Cancer Center (DKFZ) in Heidelberg, Germany, presented the progress of the German cDNA consortium. They focused on site-selected RNAs for amplification and cDNA cloning to avoid length bias. As of October 2002 there were 160,126 clones available, representing 66,27 different genes, all of which are being evaluated in a series of experimental and silico assays. Also included are functional protein assays based on protein chips.
and antibody arrays. The database is called LIFExdb and is accessible at www.elsa-bieleberg.de/ab10840/gfp.

Takashi Gujojori (National Institute of Genetics, Japan) described his research on the evolution of the central nervous system. He and his group have sequenced many expressed sequence tags (ESTs) from the brain of a primitive animal, the planarian (platyhelminth), and have conducted microarray analyses. From this work they have identified and characterized genes such as "ndv-2" (ndv) that play important roles in brain development.

Osamu Ohara (Kanazawa DNA Research Institute, Riken, Japan) reported that a total of >2000 KIAA cDNAs, representing large cDNAs, and encoding large (>100 kDa) proteins expressed in brain, have been sequenced. Because this is predicted to represent close to a complete set of such genes, efforts are now being directed towards isolation and sequencing of homologous mouse cDNAs, preparation of antibodies to mouse proteins, and analysis by yeast two-hybrid screening.

Thomas Gingeras (Affymetrix, Santa Clara, USA) showed data from non-repetitive sequence, genomic nucleotide arrays of human chromosomes 21 and 22, screened with cytoplasmic cDNA from 11 cell lines (10 of which were cancer cell lines). Surprisingly, 24% and 65% of positive probes were found in intron and intergenic regions (in at least five of the 11 cell lines tested), respectively, with a subset verified by RT-PCR of Northern hybridization. A number of regulatory factor binding sites (for e-myc, SPI and p53) were defined by chromatin immunoprecipitation, within upstream regions of the putative new transcripts, as well as at the 5' ends of known genes. These data suggest much higher transcriptional activity, of unknown function, than anticipated.

Martin Ringwald (The Jackson Laboratory, Bar Harbor, Maine) reported on the ongoing development of the Mouse Expression Database (GXD). This database has now completed literature analysis for gene expression during the mouse embryonic development from the year of 1993 to present. Ringwald also proposed to use orthogonal vocabularies to describe phenomena, such as functional terms (e.g., induce or cause) together with anatomical terms (e.g., bones or liver) and medical terms (e.g., cancer or diabetes). This results in an almost natural language description that abides strictly by the rules of controlled vocabularies and is thus computer processible.

cDNA arrays

Dr. Yoav Arava (Stanford University) studied translational efficiency in S. cerevisiae, by experiments measuring the distribution of mRNAs associated with polysomes. This was coupled with a method for specific cleavage in the central portion of an mRNA. Overall conclusions were that elongation is processive and that termination is not a slow step in protein generation, but that initiation is rate limiting for most mRNAs and, in general, longer mRNAs, in particular those with long 3'UTRs, are translated less efficiently because of slower initiation rates.

Proteins encoding RNA Recognition Motifs (RRMs) represent the seventh most abundant protein family and are likely to play diverse roles in regulation of many features of RNA processing and metabolism. Scott Tenenbaum (Duke University Medical School) reported on efforts to correlate individual RNA binding proteins with their target mRNAs. Protein-RNA complexes were precipitated using an antibody to the specific RNA binding protein (e.g., PABP, FMRP) and associated RNAs are then labeled and used to screen a Clontech cDNA array. Unique subsets of RNAs have been identified, allowing further analysis of recognition sequences.

Aravind Raghavan (University of Minnesota) discussed analysis of mRNA decay rates, using actinomycin D to prevent new mRNA synthesis, in T cells before and after activation of the cells. He observed many short lived transcripts and, importantly, many transcripts that were either stabilized or destabilized by cell activation. Many of the transcripts regulated at the level of mRNA stability contained no previously characterized stability determinants, emphasizing how much remains to be learned in this field.

Dr. Michael Ilyes (Raytest GmbH, Straubenhardt, Germany) presented data on the variability of results in array experiments caused by different software evaluation strategies for spot detection and background subtraction. Comparison of three spot detection methods (variable spot area with autocontouring of spot shape, variable spot diameter and background subtraction, and constant spot diameter and background subtraction) showed no significant differences for spots with high signal intensities. However, for low intensity signals, the spot detection with variable spot area with autocontouring of the spot shape, as well as the variable spot diameter demonstrated a clear tendency to underestimate or ignore spot data.

Michael Hulten (National Institute for Genome Research, Bethesda, MD) described insightful interpretation of cDNA array data for the identification of genes relevant to melanoma. Mutations causing coat color and eye defects in mouse map near the Robbl gene, implicated as a candidate gene for human melanoma.

Expression networks

Thomas Werner (Genomatix Software GmbH, Munich, Germany) presented studies with the PAX8 gene as an example of in silico determination of potential targets for a transcription factor (TF). For most promoters, multiple TF binding sites are required and transcriptional response is only elicited by two or more factors binding simultaneously. For PAX8, this computational approach helped to reduce the number of candidate target genes — 100-fold, leading to the identification of a functional PAX8 network in signal transduction.

Sriram Ittiraj (Celera Genomics) described the Davidson model of transcription regulation developed in sea urchin, and focusing on the endo16 gene and its associated network of interactions. Transcription factor binding sites and promoter modules were defined experimentally through exhaustive mutational analysis. Mathematical models of the network and algorithms related to it were discussed as an illustration of techniques that may be useful in analysis of diverse regulatory networks in this and other organisms.
Alternative splicing and pre-mRNA processing

Lynne Maquat (University of Rochester Medical Center, Rochester, NY) discussed aspects of Nonsense Mediated Decay and its relationship to alternative splicing events that introduce premature stop codons. Stefan Stamm (Institute of Biochemistry, Erlangen, Germany) discussed the role of phosphorylation in alternative splicing. Specific examples include the CDC2-like kinases, CLK1-4, that phosphorylate serine-arginine-rich proteins, that in turn regulate alternative splicing, abo-mediated phosphorylation of SLM2 which results in exon skipping in the tau protein mRNA, and insulin treatment altering phosphorylation of SRp40 which in turn alters alternative splicing of PKG.

Winston Hide (South African National Bioinformatics Institute, Cape Town, South Africa) analyzed RNASeq and dbEST sequences to describe alternative splicing events caused by exon skipping. Because 90% of such events occur in coding regions, they affect protein coding, with ~50% resulting in reading frame truncation. Currently, the observed frequency of exon skipping in mouse transcripts is lower than in human, which may simply be due to the lower quality of the mouse genomic assembly. To correlate exon skipping events with disease (e.g. cancer), Hide proposed a controlled vocabulary to describe ontologies for cell type, developmental stage and pathology, thus permitting their linkage to transcript accession numbers.

Kirk Jensen (Rockefeller University, New York) also reported on efforts to identify targets of the RNA binding proteins, NOVA and Hu, relevant to nervous system development. A novel protocol, CLIP, has been developed in which RNA-protein complexes are photo-crosslinked in situ directly in neurons. The RNA is then “pulled” to the 30–50 nucleotides protected by the protein, followed by immune precipitation with appropriate antibody, and purification and cloning of the RNAs. Analysis of 130 RNAs has shown the following distribution: 75% are located within introns, 20% are adjacent to alternatively spliced exons, 20% are within 3’UTRs and 2% are located within coding exons. These data suggest potentially diverse functions relevant to processing and stability.

The ADAR1 protein functions in the deamination of specific adenosine residues in specific pre-mRNAs. Although well-characterized substrates include serotonin and glutamate receptors, Jing-Hua Yang (Yale University) has shown that ADAR1 activity is broader than that. ADAR1 activity, as monitored by using the direct product of deamination) production, is increased in spleen, thymus and lung during inflammation. RNA and protein isoform patterns are also altered, most often affecting the nuclear localization signal, the Z-DNA binding domain or one of the RNA binding domains of the protein. These novel isoforms appear to be responsible for a shift in localization from the nucleus to the nucleolus and the cytoplasm, suggesting possible novel functions for ADAR1.
Gene organizational analyses

Lee Rowen (Institute for Systems Biology, Seattle) described the identification and organization of 211 human genes each spanning >500 kb in genomic DNA, and thus together accounting for ~5% of the human genome. Such genes tend to be located within AT-rich regions, exhibit highly complex alternative splicing, although mature mRNAs average only 2.8 kb, play critical roles in development and physiology of the central nervous system, and are frequently located at sites of chromosomal instability. Lastly, the large sizes appear to be conserved in mice.

Yoram Groner (The Weizmann Institute, Rehovot, Israel) discussed the family of three mammalian Runx transcription factors. Runx1 (see AML1) is often disrupted by chromosomal translocations in acute leukemias and is essential for definitive hematopoiesis. The gene is transcribed from either of two widely separated promoters, one of which produces a 5' untranslatable region of over 400 bases and includes an IRES. Runx3 is also transcribed from two alternative promoters. However, unlike Runx1, knockout of Runx3 leads to a specific neurological deficit (severe ataxia) as a consequence of apoptosis of a subset of dorsal root ganglion neurons.

Tom Blumenthal (University of Colorado Health Sciences Center, Denver, CO) described the identification and characterization of operons in C. elegans, where a total of 2377 genes are found to be transcribed as 896 operons. Frequently included in operons are mitochondrial genes, and genes with roles in transcription and translation, excluded from operons are transcription factors and collagen genes. Interestingly additional features of operons are that they are not defined by their promoters (switching an operon promoter for a non-operon promoter still results in transcription as an operon), component genes do not necessarily appear in equal levels, and component genes are not necessarily obviously functionally related. Lastly, where operon components are functionally related, they may provide a useful tool for analysis of orthologous human disease related genes.

Evolutionary and computational approaches

Xun Gu (Iowa State University) discussed his recent work testing two theories for the evolution of the vertebrate genome. One model suggests that there were two complete genome duplications (polyploidizations) in the history of vertebrates whereas another model postulates continuous small-scale duplications. His dating of 1.739 gene duplication events provided support for inclusion of both models, involving at least one complete genome duplication followed by a wave of smaller-scale duplications.

Blair Hodges (Pennsylvania State University) discussed some recent bioinformatic studies in his laboratory concerning the evolution of complex multicellular life. The times of divergence between selected protists, plants, animals, and fungi were estimated using sequences from 50-150 genes each. The results show that organismal complexity increased much earlier (1.5 billion years ago) than previously believed and about a billion years before the Cambrian explosion of animals.

Naruya Saijo (National Institute of Genetics, Japan) reported on work from his group, with the goal of identifying genetic changes responsible for morphological traits unique to humans. As a pilot project, they targeted the HOXA gene cluster in humans and apes and sequenced 21 kilobases (100 kilobases for gorillas and chimpanzees), including the HOXA4, HOXA5, HOXA6, and HOXA7 genes. With these data they were able to identify several short DNA regions of potential value in addressing these evolutionary questions.

2002 Roster of Participants

Adam Amsterdam
MIT Center for Cancer Research
MIT
77 Massachusetts Avenue, E17-360
Cambridge, MA 02139 USA
Telephone: 617-253-5609
Fax: 617-253-0258
Email: aadam@mit.edu

Yasuyuki Arai
Biochemistry
Stanford Medical School
455 Regeneron Center
Palo Alto, CA 94305 USA
Telephone: 650-724-6444
Fax: 650-724-6444
Email: aarai@cnmg.stanford.edu

Ian Barrett
Department of Human Genetics
Wellcome Trust Sanger Institute
Hinxton, Cambridge CB10 1SA
United Kingdom
Telephone: 44-1223-494-732
Fax: 44-1223-494-919
Email: i.barrett@sanger.ac.uk

Stephanie Becktel
Molecular Genome Analysis
German Cancer Research Center
Im Neuenheimer Feld 240
Heidelberg, D-69120 Germany
Telephone: 49-6221-56-7785
Fax: 49-6221-65-6782
Email: s.becktel@dkfz.de

Michael Bittner
NIH/NHGRI
25 South Drive, Bldg 50, Rm 6124
Bethesda, MD 20892 USA
Telephone: 301-480-7980
Fax: 301-402-2244
Email: mbittner@nih.gov

Thorsten Buehler
New Technologies
deCODE Genetics, Inc.
Stoughton 02072
Reykjavik, 101 Iceland
Telephone: 354-570-1839
Fax: 354-570-1803
Email: thb@decode.is
Nancy Hopkins
MIT Center for Cancer Research
MIT
77 Massachusetts Avenue, E17-335
Cambridge, MA 02139 USA
Telephone: 617-253-6414
Fax: 617-353-0258
Email: nhopkins@mit.edu

Marion Hruby de Angella
GIFP National Research Center for
Environmental and Health
Institute of Environmental Genetics
Institutsteichstrasse 1, D-89764
Wuerzburg, D-89764 Germany
Telephone: 49-931-5187-3502
Fax: 49-931-3187-2500
Email: mhrabe@gifp.de

Yasumasa Isatani
Informatics Research Center
Cybrics Genomics
45 West Guide Drive
Rockville, MD 20850 USA
Telephone: 240-443-3668
Fax: 240-443-1624
Email: yisatani@cybrics.com

Shinichi Inatomi
Mitsubishi Kagake Institute of Life Sciences
Minamiosumi 1-11, Machida,
Tokyo, 194-8511 Japan
Telephone: 81-42-724-6265
Fax: 81-42-724-6316
Email: shinichi.inatomi@mkls.co.jp

Luigi Ioffe
Developmental and Reproductive Biology
Biomedical Research Institute
Haarlemmalaan 9
Amsterdam, NL-1105 GG Netherlands
Telephone: 31-20-563-9000
Fax: 31-20-563-9000
Email: luigiioffe@hotmail.com

Kirk Jensen
Laboratory of Molecular Neuronal Function
The Rockefeller University
1230 York Avenue, Box 226
New York, NY 10021 USA
Telephone: 212-898-7471
Fax: 212-898-7109
Email: jensenk@rockefeller.edu

Thomas Joos
Natural and Medical Sciences Institute
University of Tbingen
Marxwegenstrasse 51
Reutlingen, D-72770 Germany
Telephone: 49-7071-2303-48
Fax: 49-7071-2303-48
Email: joos@uni-tuebingen.de

Jace Keene
Molecular Genetics and Microbiology
Duke University Medical Center
415 Medical Building, Research Drive
Durham, NC 27710 USA
Telephone: 919-684-5133
Fax: 919-684-5136
Email: keenej@duke.edu

Anja Kohl
Molecular Genome Analysis
DKFZ
In Neuenheimer Feld 280
Heidelberg, D-69120 Germany
Telephone: 49-6221-427474
Fax: 49-6221-424354
Email: a.kohl@dkfz.de

Bernhard Korn
RZPD Research Center For Genomics
Im Neuenheimer Feld 280
Heidelberg, D-69120 Germany
Telephone: 49-6221-426700
Fax: 49-6221-427004
Email: korn@rzpd.de

Sascha Kromer
NCBI & EMBL
Wayne State University
175 E. Hancock, 250 E. Mott Center
Detroit, MI 48201 USA
Telephone: 313-577-6970
Fax: 313-577-8350
Email: kromer@ncbi.nlm.nih.gov

Suehli Kueh
Center for Evolutionary and Functional Genomics
Arizona State University
Life Sciences A351
Tempe, AZ 85287 USA
Telephone: 480-727-6949
Fax: 480-965-2518
Email: kueh@asu.arizona.edu

Robert Levens
Mathematics and Statistics
Arizona State University
752 S. Kirkland Drive
Tempe, AZ 85287 USA
Telephone: 480-965-8550
Fax: 480-965-8650
Email: levensr@asu.arizona.edu

Michael Liebler
Raybiotech GmbH
Kamistädtstr 19-1
Aichstett, D-72601 Germany
Telephone: 49-7021-71964
Fax: 49-7021-71964
Email: liebler@raybiotech.de

Chun-Ye Liu
Genome Institute of Singapore
1 Science Park Road, The Capriqum #03-01
Singapore, 117626 Singapore
Telephone: 65-8227-9229
Fax: 65-8227-9204
Email: gil@genomes.gov.sg

Christian Mieczka
Research Center for Genome Research
In Neuenheimer Feld 280
Heidelberg, D-69120 Germany
Telephone: 49-6221-424741
Fax: 49-6221-424354
Email: c.mieczka@dkfz.de

Lyrene Macasur
Department of Biochemistry and Biophysics
University of Rochester
602 Elmwood Avenue, Box 712
Rochester, NY 14642 USA
Telephone: 716-273-0640
Fax: 716-273-0683
Email: lyrene-macasur@vms.rochester.edu

Caroline Madeisen
Hematology-oncology
Sainte-Justine Hospital
3175 Cote-St-Charles, B-541
Montreal, Quebec, 1J3 Canada
Telephone: 514-315-4991
Fax: 514-315-4871
Email: madeisen@umontreal.ca

Daniel Morse
Department of Chemistry
U.S. Naval Academy
MS22, Rm. 314
Annapolis, MD 21402 USA
Telephone: 410-545-9612
Fax: 410-545-2318
Email: dmorse@usna.com

Oriana Obara
Kazusa DNA Research Institute
2-4-7 Kazusa-Kamatsuka
Kamatsuka, 312-8584 Japan
Telephone: 81-438-57-3913
Fax: 81-438-57-3914
Email: obara@kazusa.or.jp

Avinoal Ragol
Department of Microbiology
University of Minnesota
420 Delaware Street SE, NMC 106
Minneapolis, MN 55455 USA
Telephone: 612-624-2315
Fax: 612-624-2313
Email: ragolav@umn.edu

Francisco Reali
Biologia Cellule, Molecular e Neurale
Istituto Italiano di Tecnologia
Dr. A. Aguzzi 80
Barcelona, 08003 Spain
Telephone: 34-3-221-3353
Fax: 34-3-221-3353
Email: reali@tiramisu.es

Martin Ringwald
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609 USA
Telephone: 207-288-4456
Fax: 207-288-6132
Email: ringwald@informatics.jax.org

Lee Rowen
Institute for Systems Biology
1441 N. 34th Street
Seattle, WA 98103 USA
Telephone: 206-612-1877
Fax: 206-612-1899
Email: lrrowen@systemsbiology.org
Chromosome Workshop

Cytogenetic and Genome Research 103:11-17 (2004)
DOI: 10.1138/00078000

Report on the thirteenth international workshop on the identification and functional, evolutionary and expression analysis of transcribed sequences: comparative and functional genomics workshop

A. Brazma, T. Freeman, K. Gardiner, S. Weissman, T. Werner, and B. Korn

*European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton (UK); †The University of Denver, Denver CO, ‡Yale University, New Haven CT (USA); §Genomatica Software GmbH, Munich, and ‡RZPD German Genome Research Center, Heidelberg (Germany)

This thirteenth workshop in the series “Beyond the Identification of Transcribed Sequences—Functional, Expression and Evolutionary Analysis” was renamed the Comparative and Functional Genomics Workshop and held at the Sanger Center, Hinxton, England, November 2–5, 2003. The meeting was sponsored again by the United States Department of Energy, and for the first time also by the Wellcome Trust. The digly attendees discussed a broad range of topics relevant to gene structure, and functional and regulation analysis in the era of complete and draft genomic sequences from multiple organisms.

Arrays and network analysis

Michael Snyder (Yale University) discussed the status and utility of arrays of yeast proteins with applications to the identification of kinase substrates, interacting proteins, and protein binding to specific phosphorylated forms of insulin, and also reviewed the use of arrays to demonstrate antibody specificity or the lack thereof. These examples showed that protein arrays provide a practical approach to obtain data in a variety of biochemical situations, including those otherwise requiring laborious one-by-one biochemical analysis. Nick Luscombe (from the Gurdon laboratory, Yale University) pooled a variety of data from the literature concerning yeast protein-protein interactions and gene co-expression studies. While the initial graph of all interactions appeared uninterpretable dense, use of general network parameters and identification of specific network “motifs” helped to dissect out responses to specific stimuli, and rapid, relatively confined, responses could be distinguished from more complex interactive ones. This clear presentation offered the prospect of eventual interpretation of gene interaction patterns from complex, diverse and extensive data sets.

Sherman Weissman (Yale University) reviewed mRNA expression patterns in various lineages of the hematopoietic system, emphasizing that each lineage was characterized by a highly distinctive quantitative pattern of gene expression that, notably, includes even differences in expression levels for “housekeeping” genes. This work included description of a cell line that continuously propagated two different forms of hematopoietic precursor cells as well as throwing off more differentiated cells. Lastly, Dr. Weissman briefly reviewed published work from the Yale Center of Genomics relating to analysis of transcribed sequences using genomic tiling arrays for chromosome 22, and of transcription factor binding within chromosome 22 as analyzed by chromatin immunoprecipitation plus genomic tiling array hybridization. Approximately half of all transcripts in placenta derived from regions of the genome lacking annotated gene features, and a large fraction of transcription factor binding sites occurred elsewhere than in the
promoters of known genes, often in fragments that did not have recognizable binding sites for the factor in question. In similar experiments with genomic tile arrays and transcription factor binding experiments, D. Gingeras (Affymetrix) described progress in preparing sets of oligonucleotide arrays that probe genomic sequences in a manner that could provide a remarkable resolution, particularly for mapping titers of transcribed sequences. An impressive number of these “novel” transcribed sequences have already been cloned by RACE methodology, and their further analysis is awaited with great interest. Together, these data show the need for additional transcribed sequence identification and verification projects, as well as novel interpretations.

Regulation of gene expression

A series of talks focused on aspects of gene regulation, ranging from direct identification of promoters and chromosomal domains to physical clustering of functionally related genes and adapted evolutionary rates of highly expressed genes. Retroviruses, especially HIV that integrate into the host genome obviously are also subject to such forces and seem to adapt to cellular mechanisms.

Thomas Werner (Genomix, Munich, Germany) reported on recent developments in the EIDorado system, which now includes comparative genomics (human, rat and mouse) at the promoter level. This allows prediction of new transcripts based on evidence from multiple species as well as development of organizational fingerprints that can successfully identify promoters of genes across species with 100% specificity. Such comparative promoter analysis is expected to aid both in annotation of new genomes using gene predictions and in the functional understanding of mammalian promoters. Gene regulation, and thus transcription, is significantly influenced by scaffold or matrix attachment regions (SMARs), as described by Juergen Bode (GBF Braunschweig, Germany). SMARs are regions of DNA that are always the first to become single stranded upon torsional stress on superhelical DNA. SMARs can now be visualized in situ, facilitating observation of SMAR-related events such as a tighter association with the nuclear matrix if genes are highly expressed. Tools have been developed that will allow a more detailed functional dissection of SMARs.

Laurence Hurst (University of Bath, England) examined the total distribution of essential and also widely expressed genes in yeast. He reported the preferential clustering of such genes independent of direct gene duplication or coexpression patterns. Maintenance of gene order is also supported by the observation that clusters of essential genes, at least in yeast, tend to have a relatively low recombination rate. Sudhir Kumar (Arizona State University, Tempe, USA) approached the same topic based on protein sequence conservation. Although the idea of molecular clocks that will be most useful for such analyses is still debated, Dr. Kumar nevertheless found an inverse correlation between protein sequence evolution and expression level, at least based on EST databases, in genomes ranging from invertebrates, such as Droso phila melanogaster to fugu (puller fish) to mammals. High expression levels may actually exert some selective pressure against mutation fixation in the corresponding genes.

Regulation of transcription by gene organization has been demonstrated in C. elegans, where many genes were recently shown by Tom Blumenthal (University of Colorado Health Sciences Center, Denver, USA) to be organized into bacteriophage-like operons. Primary polycistronic transcripts are processed into monocistronic mRNAs with tran-splicing into the downstream units of a uniform SL2 leader sequence. Microarrays were probed with the SL2 sequence, identifying approximately 15% of C. elegans genes that are expressed within approximately 1000 operons, each containing 2-8 genes. Functional annotation demonstrates both inclusion and exclusion within operons of particular classes of genes. The finding that some individual operons contain functionally related genes suggests that this feature may be useful for predicting functions of novel operon-associated genes.

Daniel Finley (Harvard Medical School) described regulation of gene expression at the proteome level, specifically by the ubiquitination system. Large-scale mass spectroscopic analysis of ubiquitinated peptides from S. cerevisiae identified >500 substrates for ubiquitination. Genomic studies have shown a truly remarkable number of potential ubiquitination receptor proteins, potentially rivaling kinase genes in number in mammalian. Exploration of the role of this system in regulating both abundance and subcellular localization of specific proteins is only in its beginning stages.

Forces dictating gene location, chromosomal and transcriptional regulation may well extend to retroviruses that lodge in the host genome. Alla Ryndich (Institute of Molecular Biology, Kiev, Ukraine) reported that the GC content of retroviruses usually matches that of the isochores of the host genome where they integrate. Thus, HIV, a GC-poor retrovirus, integrates into GC-poor regions of the host. In contrast, most expressed retroviruses are located in GC-rich regions because these are preferentially the open and active parts of the chromatin. HIV appears to change its preferred chromosomal environment during progression to AIDS, with the frequency of integration in GC-rich regions rising. High virion of HIV seems to be correlated with isocitric integration. HIV also exploits nucleo-cytoplasmic trafficking to tune its expression, as reported by Ruth Brack-Werner (GSF Research Center, Munich, Germany). Using a live-cell imaging system to follow the movement of protein, an essential posttranscriptional regulator of HIV expression, it was determined that the import and export signals of this protein are intimately coupled composite elements and cannot be clearly separated. It was found that this event is differently regulated in HeLa cells and astrocytes. Therefore, even a basic physiological process such as nucleo-cytoplasmic transport has at least some cell-type-specific regulatory aspects despite being active in every cell.

RNA processing

Jing Hua Yang (Yale University) previously described the remarkably high levels of adenine conversion to inosine in systemic inflammation or during activation of T cells or macro-
phages in vitro. New studies indicate that the level of conversion can be markedly elevated by overexpression of ADAR1 and by adding leptomycin to inhibit nuclear export of the deaminating enzyme, suggesting that demethylation may be an intranuclear event. Under appropriate protocols as much as 50% of adenosines can be converted to inosine. This suggests that gene expression, and perhaps even the nature of proteins produced, may be markedly altered in inflammations secondary to post-transcriptional modifications of the mRNA. Thomas Preiss (Victor Chang Cardiac Research Institute, Australia) described how transcription and mRNA turnover determine the quantitative composition of the cellular transcriptome. He noted that the transcriptome in turn templates the proteins via translation. Treatment of Saccharomyces cerevisiae with the TOR kinase inhibitor rapamycin causes increases and decreases in the mRNA levels of hundreds of genes. In his work he used genome-wide DNA microarray analysis to simultaneously monitor transcriptome and translational changes for all detectable yeast mRNAs.

Knockouts for gene functional analysis

Michael Bougro (Harvard Medical School) discussed a genome-wide approach for gene knockouts in Drosophila. Addition of RNAi gave highly penetrant and reproducible phenotypes, making it possible to undertake a high throughput screen of gene knockouts based on cell phenotypes. Assays have been developed based on fluorescent or luminescent reporters and adapted to plate readers as well as to phenotypes that can be detected by automatic microscopy.

Nancy Hopkins and Adam Amsterdam (Massachusetts Institute of Technology, Cambridge, USA) described their largescale insertional mutagenesis project in zebrafish, in which they seek to identify genes with functions critical to the first five days of development. They estimate that they have identified approximately 15% of the genes which result in a visible phenotype when mutated. Of these, 30% produce a unique and specific phenotype. To characterize these mutations, 70% of associated genes have been cloned. In adult fish, screening for cancer and early mortality identified a number of candidate genes, among them many encoding components of the translational machinery, including, intriguingly, some ribosomal proteins.

Gene-specific and disease-specific studies

Projects applying large and small scale proteomics and network analyses to several neurological diseases and individual genes were discussed. Ulrich Steinhart (Max Delbrueck Center for Molecular Medicine, Berlin) reported results of protein interaction studies with the Huntington's Disease gene, HD. Using fragments of HD and known interaction partners, a network comprising 188 interactions was defined. Although possibly not all of the 167 novel interactions are biologically relevant, 32 were verified in pulldown and overfly experiments, and functional assignments were made for 18. Cellular processes of vesicle trafficking, cytoskeletal organization and transcription regulation are implicated in the network. In a smaller scale network analysis, Kathleen Gardner (University of Denver) compiled the known interactions of chromosome 21 encoded genes with components of the MAP kinase and the calcium signaling pathways. These interactions predict the perturbation of these pathways due to gene dosage and overexpression of chromosome 21 genes in Down syndrome. Perturbations in both pathways were demonstrated in brains of the segmental trisomy mouse model of Down syndrome, for the first time correlating functions of chromosome 21 genes with a molecular phenotype relevant to behavior and learning.

In a single gene interaction study, Stefan Stadtm (University of Erlangen, Germany) described aberrant splicing of the Tau gene in frontotemporal dementia. Mutations have been found within a splicing enhancer sequence located within exon 18 of the Tau gene. These mutations alter interaction with the alternative splicing regulator, Tra2B and lead to abnormal patterns of exon 18 inclusion. Bento Soares (University of Iowa) described comparative analysis of the genomic structures of the metastasis suppressor gene, Kist, in mouse. This gene shows six alternative splice variants that encode three distinct proteins. In human, the orthologs for two of the three proteins are encoded by two separate but physically closely related genes.

Bioinformatics

Alvis Brazins (European Bioinformatics Institute, Hinxton, England) addressed the elucidation of the network topology using a graph-based approach, an important step towards modeling and ultimately understanding cellular control systems. Gene regulation networks for the yeast Saccharomyces cerevisiae were compared based on genome wide data sets from three different experimental techniques: chromatin immunoprecipitation experiments, computational analysis of transcription factor binding sites and microarray experiments on single gene deletion mutants.

Winston Hide (South African National Bioinformatics Institute, Belleville, South Africa) addressed the problem of descriptive expression of known gene transcript isoforms to generate an atlas of normal (non-disease associated) products. His group uses a transcript-based cross-platform integration of gene expression across categories of human anatomy. This work has been done in collaboration with Biological Information Research Center in Japan, CNRS in France, and the Ludwig Institute for Cancer Research in Brazil.

Graziano Pugliole (University of Milano, Italy) addressed the identification of conserved sequence tags (CSTs) through comparative genomic analysis. A simple algorithm is proposed that, based on the observation of the specific evolutionary dynamics of coding sequences, efficiently discriminates between coding and non-coding CST. The application of this method may help the validation of predicted genes, the prediction of alternative splicing patterns in known and unknown genes, and the definition of a dictionary of non-coding regulatory elements.
Martin Vighorn (Max Planck Institute for Molecular Genetics and Computational Molecular Biology, Berlin, Germany) introduced correspondence analysis as a tool for visualizing associations between genes and conditions in DNA microarray data. The same technique was also applied to establish associations between gene expression data and transcription factor binding sites. While for yeast this can be done based on published transcription factor binding data, for human data comparative analysis with mouse data in a search for binding sites has not been used.

Janet Trautten (European Bioinformatics Institute, Hinxton, England) described calculation of the physicochemical properties of 745 metabolites from E. coli based on their two-dimensional structures obtained from the EcoCyc and REGG databases. When used in conjunction with existing classifications of the proteome, the binding preferences and promiscuity of proteins and their cognate substrates can be elucidated.

Resources

Bernhard Korn (RZPD, Heidelberg, Germany) and Uwe Radelfo (RZPD, Berlin, Germany) described the range of products and services available from the German Genomics Resource Center. New resources include a growing set of shuttle vectors and RNAi knock down constructs. Osamu Ohara (Kazusa DNA Research Institute, Chiba, Japan) has supervised the development and analysis of the RIKEN human and mouse cDNA sets, because of the apparent relevance of these long transcripts (> 4 kb) in the brain transcriptome. Current efforts are now directed to the functional analysis of these genes, to this end, the cDNA sequences are being cloned into expression vectors and polyclonal antibodies are being produced. Dolores Cahill (Royal College of Surgeons, Dublin, Ireland) has constructed a human brain cDNA expression library for production of a protein array. Currently more than 11,000 different protein sequences have been verified. Preliminary screens with antibodies to well characterized genes have revealed novel array reactivities that suggest lack of strict specificities.

Geoff Hicks (University of Manitoba, Winnipeg, Canada) and Harald von Meckner (University of Frankfurt Medical School, Germany) described the progress of the International Genetics Consortium (IGC) in generating a collection of mouse embryonic stem cell lines carrying gene trap sequence tags (GTST) from intentional mutagenesis. Use of different vectors with different mutational mechanisms has resulted in thousands of ES cell lines each with a disruption in a specific, known, gene. Cell lines are described in a database and are available to investigators for construction of knockout mice. In a resource of related utility, Martin Hrabě de Angelis (GSF Institute of Experimental Genetics, Neuherberg, Germany) described the German Mouse Clinic where more than 5000 mice per year can be phenotyped for more than 160 parameters.

Martin Ringwald (The Jackson Laboratory, Maine, USA) united many of these resource issues in an update on the data and search tools available in the Mouse Expression Database, which continues with its collection and integration of all forms of expression data relevant to the laboratory mouse.

Participants

Isabel Alonso 
Unigene-IBMC
Rua de Campo Alegre 623
Porto 4150-180 Portugal
Telephone: +351 22 657 9492
Fax: +351 22 659 9157
E-mail: i.alonso@ibmce.up.pt

Adam Amsterdam
Massachusetts Institute of Technology Center for Cancer Research 77 Massachusetts Avenue, E17-325 Cambridge, MA 02139 USA
Telephone: +1-617-253-0609
Fax: +1-617-253-0518
E-mail: a.amsterdam@mit.edu

Angélique Anthony
University of Edinburgh
Institute of Cell, Animal and Population Biology
King's Buildings Campus
Edinburgh EH9 3JU UK
Telephone: +44 (0) 31 650 7403
Fax: +44 (0) 31 650 7403
E-mail: cian@ed.ac.uk

Alan Archbold
Riken Institute
Department of Genomics and Bioinformatics
Riken, Wako-shi, Saitama 351-0198 Japan
Telephone: +81 4 (0) 44 340 1134
Fax: +81 4 (0) 44 340 1135
E-mail: alan.archbold@riken.ac.jp

Jennifer Asaithurai
The Wellcome Trust Sanger Institute
The Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SA UK
Telephone: +44 (0) 1223 494949
Fax: +44 (0) 1223 494919
E-mail: j.asaithurai@sanger.ac.uk

Tom Blumenthal
University of Colorado School of Medicine
Biochemistry and Molecular Genetics Department
4200 E 9th Avenue
Denver, CO 80262 USA
Telephone: +1-303-275-8500
Fax: +1-303-275-6215
E-mail: tom.blumenthal@uchsc.edu

Jürgen Bode
GSF Institute of Experimental Genetics, Neuherberg, Germany
E-mail: j.bode@gsf.de

Alexander Bolshoy
University of Haifa
Genomic Divinity Centre
Mount Carmel
Haifa 34105 Israel
Telephone: +972 4 824 3832
Fax: +972 4 824 3832
E-mail: abolshoy@research.haifa.ac.il

Hidetaka Ikeda
University of Washington
Research Centre for Gene Expression (L397-1 Yamanote
Hiratsuka 350-1241 Japan
Telephone: +81 468 65 7318
Fax: +81 468 63 7529
E-mail: bionet@u.washington.edu
