Beyond the Identification of Transcribed Sequences: Functional and Expression Analysis

9th Annual Workshop, October 28-31, 1999

Co-sponsored by the U.S. Department of Energy

Contents

Speaker Abstracts

DURET: Codon usage in animals and plants: new data, new paradox
Laurent Duret and Dominique Mouchiroud

THIESEN: Genomic sequences of KOX zinc finger gene clusters are used to determine (auto) regulatory networks of gene regulation
Dirk Koczan, Christian Sina, Peter Lorenz, Tom Hearn, Saleh Ibrahim, Andreas Rump, Andre Rosenthal, Michael Jackson, M.-F. Rousseau-Merck, I. Legrand, and Hans-Juergen Thiesen

EISEN: Phylogenomics and the benefits of an evolutionary perspective in genome analysis
Jonathan A. Eisen

PELTZ: Understanding the role of mRNA turnover in regulating gene expression and in disease states
Stuart Peltz

EMANUELSSON: Identifying localization signals in proteins
Olof Emanuelsson

KORN: Tools for functional analysis using resources of RZPD
B. Korn[1, 2], J. Boer, P. Kioschis[1,2], A. Vente, H. Lehrach[3,4] and A. Poustka[1,2]

WASSERMAN: De novo discovery of transcription factor binding sites in co-regulated sets of human genes
Wyeth W. Wasserman

FONTES: Using large cloned genomic fragments as expression vectors to create or correct mutant phenotypes
E. Passage and M. Fontés

SOMOGYI: Expression profiling and genetic networks
R. Somogyi, S. Fuhrman, X. Wen, P. D'haeseleer*, S. Liang, and J.F. Loring

ZHOU: Gene expression profiling in molecular pharmacology
Y. Zhou, U. Scherf, M. Waltham, W.C. Reinhold, L.H. Smith, J.K. Lee, D.A. Scudiero, E.A. Sausville, D. Ross, M. Eisen, D. Botstein, P.O. Brown, L. Miller, E. Liu and J.N. Weinstein

HOWELL: Annotation of the human X chromosome sequence
Gareth R. Howell, Alison J. Coffey, Susan Rhodes, Robert A. Brooksbank, Shirin S. Joseph, Jackie M. Bye, Andrew King, Laurens Wilming, David R. Bentley and Mark T. Ross

SOARES: Development of non-redundant arrays of full-length cDNAs
Maria F. Bonaldo, Brian Berger, Todd Scheetz[1, 2], Kyle Munn, Chad Roberts, Kang Liu, Thomas Casavant and Marcelo Bento Soares[1, 3]

WERNER: Promoter prediction in large genomic sequences: compromise between sensitivity and specificity
Scherf, M., Klingenhoff, A., and Werner, T.

HESKETH: Perinuclear mRNA localisation by 3'untranslated sequences
John Hesketh, Marilyne Levadoux, Gillian Dalgleish, John Beattie, Heather Wallace, and Jean-Marie Blanchard

MORSE: Substrate specificity of adenosine deaminases that act on RNA (ADARs) in C. elegans
Daniel P. Morse, Leath A. Tonkin, P. Joe Aruscavage, and Brenda L. Bass

BODE: Scaffold/matrix attached regions (S/MAR Elements): Detection and activities in vivo
Jürgen Bode, Alexandra Baer, Angela Knopp, Dirk Schübeler, Jost Seibler, Craig Benham Armin Baiker , and Hans-Joachim Lipps

GOJOBORI: A gene expression profile obtained by determining over 1000 cDNA clones from a single Planarian eye cell
Takashi Gojobori, Kazuho Ikeo, Silvana Gaudieri, Akira Tazaki, Masumi Nakazawa, Masafumi Shimoda, Yuzuru Tanaka, and Kiyokazu Agata

SVERDLOV: Towards the understanding of the possible impact of human endogenous retroviruses on evolutionary changes in genome transcription
Eugene D. Sverdlov

PESOLE: Structural and evolutionary analysis of eukaryotic mRNA untranslated regions
Pesole G.[1,4], Liuni S.[2,4], Larizza A., Makalowski W.[5] and Saccone C.[3,4]

SIMONNEAU: Analysis of new transcribed sequences possibly involved in same functional pathways of neuronal development and identified by differential display
Michel Simonneau, Christophe Mas, Francine Bourgeois, and Fabien Guimiot

WINGENDER: TRANSFAC: an integrated system for gene expression regulation
E. Wingender, X. Chen, H. Karas, A. Kel, I. Liebich, V. Matys, T. Meinhardt, M. Pruess, I. Reuter, and F. Schacherer

LIN: Identification and characterization of tissue-specific genes using transgenic zebrafish
Shuo Lin

EBERWINE: Molecular biology of single cells: Insights into human disease
Jim Eberwine

WIEMANN: Sequencing and analysis of full length cDNAs in the course of the German Genome Project
Stefan Wiemann, Wilhelm Ansorge, Helmut Blöcker, Helmut Blum, Andreas Düsterhöft, Karl Köhrer, Werner Mewes, Brigitte Obermaier, Rolf Wambutt, and Annemarie Poustka

VON MELCHNER: Disruption of a gene induced in early mouse development results in severe emphysema and Adenocarcinoma
Irmgard S. Thorey, Anja Sterner-Kock, Jürgen Otte, and Harald von Melchner

BRAZMA: Mining the yeast genome expression and sequence data
Alvis Brazma

WEISSMAN: Transcript profiling of hematopoietic cell development and activation
Yamaga, S., Yeramilli, S.[5], Lian, Z., Prashar, Y.[5], Lee, H., Berliner, N., Liu, Y-C., Goguen, J., Newburger, P., and Weissman, S.

BARBAZUK: Construction of a whole genome transcribed sequence map for the zebrafish Danio Rerio

W. Brad Barbazuk, Ian Korf, Frank Li, John McPherson and Steve Johnson

HICKS: Development of an embryonic stem cell library of defined mutations
Geoffrey G. Hicks

WORLEY: Search services to improve the identification of expressed sequences and their functions
J. Bouck, M. McLeod, T. McNeill, G. Weinstock[1,3], R. A. Gibbs, and K. C. Worley

FICKETT: Finding gene boundaries on large contigs
James Fickett

JENNINGS: Dissecting transcriptional circuitry and mechanisms in yeast
Ezra Jennings

WELSH: The RAP-array approach to cDNA array hybridization
John Welsh

MURAL: Extracting meaningful information from draft sequence
R. J. Mural, F. W. Larimer, M. B. Shah and E. C. Uberbacher

PAN: Approaches towards global profiling of cDNA mutations
Pan, X. and Weissman, S.

NELSON: Functional dissection of the RANTES promoter: Insights into mechanisms of tissue specific regulation of transcription
Peter J. Nelson, Sabine Böhlk, Sabine Fessele and Thomas Werner

BORSANI: A tale of two diseases
Giuseppe Borsani

ARONOW: Stepwise chromatin restructuring and unexpected rules for interaction of distributed cis-elements that form a locus control region in vivo
Catherine Ley-Ebert, Carolyn Florio, John Maier, Chris Cost and Bruce Aronow

OVERTON: (ABSTRACT NOT AVAILABLE)

Poster Abstracts

DE BEKKE: Experimental approach towards identification of small non-messenger RNAs in the genome of Caenorhabditis elegans
Anja op de Bekke, Alexander Huttenhofer, Martin Kiefmann, John O'Brien, Hans Lehrach and Jurgen Brosius

HEUERMANN: Stable expression of epitope-tagged proteins in mammalian cells
Kenneth E. Heuermann and Bill L. Brizzard

KAMP: Structure and function of the spermatogenesis genes located in AZFa, a region of the human Y chromosome deleted in men with complete germ cell aplasia

Kamp, Christine, Kirsch, S, Hirschmann, P, Ditton, HJ, Brede, G, Tyler-Smith, C, Rappold, GA, and Vogt, PH

BULL: Analysis of gene expression data generated by oligonucleotide fingerprinting
Christof Bull, John O'Brien, Uwe Radelof, Ralf Herwig, Steffen Hennig, Axel Nagel and Hans Lehrach

SLAVOV: Analysis of novel genes from human chromosome 21: determination and characterization of complete protein sequences and examples of overlapping genes
Dobromir Slavov, Roger Lucas, Andrew Fortna and Katheleen Gardiner

SPENGLER: ASDB: Novel database of alternatively spliced genes
I. Dubchak, M. S. Gelfand, I. Dralyuk, M. Zorn, S. Spengler

KIKUNO: Functional and molecular evolutional analysis of predicted gene products of human long cDNAs
Reiko Kikuno, Takahiro Nagase, Ken-Ichi Ishikawa, Mikita Suyama, Mina Waki, Makoto Hirosawa, Nobuo Nomura, and Osamu Ohara

BORISSEVITCH: Large scale cloning, sequencing and expression profiling of genes expressed in transcription factor CREM dependant manner during mouse spermatogenesis
Igor Borissevitch, Tim Beissbardth, Andreas Hoerlein and Guenter Schuetz

MONTPETIT: Genomic comparative analysis of the Fugu rubripes homologue of ETV6, a gene frequently rearranged in human leukemias
Alexandre Montpetit and Daniel Sinnett

CLIFTON: Comparison of completed genomes to sample sequences of related genomes
Sandra W. Clifton, Michael McClelland, Webb Miller, William R. Pearson, Aaron J. Mackey, and Richard K. Wilson

SCHMIDT-KITTLER: Genomic characterisation of early disseminated tumor cells isolated from bone marrow of breast cancer patients
Oleg Schmidt-Kittler, Julian Schardt, Günter Schlimok, Gert Riethmüller and Christoph Klein

ROTTIERS: Oxidative metabolism and gene expression: Gene discovery array analysis
Pieter Rottiers, Vera Goossens and Johan Grooten

RUUSKANEN: Divergent 2-adrenoceptor subtypes in the zebrafish (Danio rerio)
Jori Ruuskanen, Minna Varis, Erik Salaneck, Tiina Salminen, Tommi Nyronen, Mark S. Johnson, Dan Larhammar and Mika Scheinin

TEMPLE: GATEWAY cloning: A high-throughput gene transfer technology for rapid functional analysis and protein expression
James Hartley, Gary Temple, Michael Brasch, et al.

KAPANADZE: Sequence homology between human and mouse genomic regions to identify the tumor suppressor gene involved in B cell chronic lymphocytic leukemia
Bagrat Kapanadze, Nataliy Makeeva, Olle Sangfelt, Martin Corcoran, Anna Baranova, Eugene Zabarovsky, Nick Yankovsky, Dan Grander and Stefan Einhorn

CLARK: Comparative mapping in the Japanese pufferfish (Fugu rubripes)

Clark, M. S; Edwards, Y.J.K; Shaw, L; Snell, P.; Smith, S; and Elgar, G.

FROLOVA: Transcriptional regulation of the collagen a1(IX) gene during eye development
Elena I. Frolova and David C. Beebe

KREFT: Identification of a novel cellular protein that binds to the HBV RNA pregenome
S. Kreft and M. Nassal

Beyond the Identification of Transcribed Sequences:
Functional and Expression Analysis

9th Annual Workshop
October 28-31, 1999

Agenda

All sessions will be held at the Sheraton Reston Hotel in Reston, Virginia, just west of Washington, DC.

Thursday, October 28, 1999

5:30 pm 7:30 pm Wine and cheese reception, pick up abstract book

Friday, October 29, 1999

8:30 am 12:00 pm SESSION 1

Chairmen: Graziano Pesole and John Welsh

8:30 am Opening Remarks

8:30 am 9:50 am

- W.J. (Joe) Dickinson "Evolutionary perspectives on functional analysis: some cautionary tales"
- Martin Ringwald "Gene expression database for the laboratory mouse"
- Osamu Ohara "Analysis of uncharacterized human cDNAs which encode large proteins in brain: The Kazusa Approach"
- Nina Heiss "Functional characterisation of genes in Xq28"

9:50 am 10:20 am BREAK

10:20 am 12:00 pm

- Giorgio Bernardi "The formation and maintenance of the isochores from the genomes of warm-blooded vertebrates is due to natural selection"
- Yuji Kohara "Post-genomics strategies in *C. elegans* toward understanding of the molecular mechanisms of development"
- Katheleen Gardiner "Genomic sequence analysis and novel gene characterization on human chromosome 21"
- Tom Freeman "Development of expression profiling technologies for the rat"
- Laurent Duret "Codon usage in animals and plants: New data, new paradox"

12:00 pm 1:00 pm LUNCH

1:00pm 2:00 pm POSTER VIEWING I (odd numbers)

2:00 pm 5:50 pm SESSION 2

Chairmen: John Hesketh and Kim Worley

2:00 pm 3:40 pm

- Hans-Juergen Thiesen "Genomic sequences of KOX zinc finger gene clusters are used to determine (auto) regulatory networks of gene regulation"
- Jonathan Eisen "Phylogenomics and the benefits of an evolutionary perspective in genome analysis"
- Stuart Peltz "Understanding the role of mRNA turnover in regulating gene expression and in disease states"
- Olof Emanuelsson "Identifying localization signals in proteins"
- Bernhard Korn "Tools for functional analysis using resources of RZPD"

3:40 pm 4:10 pm BREAK

4:10 pm 5:50 pm

- Wyeth Wasserman "De novo discovery of transcription factor binding sites in co-regulated sets of human genes"
- Michel Fontes "Using large cloned genomic fragments as expression vectors to create or correct mutant phenotypes"
- Roland Somogyi "Expression profiling and genetic networks"
- Yi Zhou "Gene expression profiling in molecular pharmacology"
- Gareth Howell "Annotation of the human X chromosome sequence"

7:30 pm 9:30 pm RECEPTION

Saturday, October 30, 1999

8:30 am 12:00 SESSION 3

Chairmen: James Fickett and Bernhard Korn

8:30 am 9:50 am

- Bento Soares "Development of non-redundant arrays of full-length cDNAs"
- Thomas Werner "Promoter prediction in large genomic sequences: compromise between sensitivity and specificity"
- John Hesketh "Perinuclear mRNA localisation by 3' untranslated sequences"
- Daniel Morse "Substrate specificity of adenosine deaminases that act on RNA (ADARs) in *C. elegans*"

9:50 am 10:10 am BREAK

10:10 am 12:00 am

- Jurgen Bode "Scaffold/matrix attached regions (S/MAR elements): Detection and activities in vivo"
- Takashi Gojobori "A gene expression profile obtained by determining over 1000 cDNA clones from a single Planarian eye cell"
- Eugene Sverdlov "Towards the understanding of the possible impact of human endogenous retroviruses on evolutionary changes in genome transcription"
- Graziano Pesole "Structural and evolutionary analysis of eukaryotic mRNA untranslated regions"
- Michel Simonneau "Analysis of new transcribed sequences possibly involved in same functional pathways of neuronal development and identified by differential display"

12:00 pm 1:00 pm LUNCH

1:00 pm 2:00 pm POSTER VIEWING II (even numbers)

2:00 pm 5:50 pm SESSION 4

Chairmen: Jurgen Bode and Yuji Kohara

2:00 pm 3:40 pm

- Edgar Wingender "TRANSFAC: An integrated system for gene expression regulation"
- Shuo Lin "Identification and characterization of tissue-specific genes using transgenic zebrafish"
- James Eberwine "Molecular biology of single cells: Insights into human disease"
- Stefan Wiemann "Sequencing and analysis of full length cDNAs in the course of the German Genome Project"
- Harald von Melchner "Disruption of a gene induced in early mouse development results in severe emphysema and Adenocarcinoma"

3:40 pm 4:10 pm BREAK

4:10 pm 5:50 pm

- Alvis Brazma "Mining the yeast genome expression and sequence data"
- Sherman Weissman "Transcript profiling of hematopoietic cell development and activation"
- W. Brad Barbazuk "Construction of a whole genome transcribed sequence map for the zebrafish Danio Rerio"
- Geoffrey Hicks "Development of an embryonic stem cell library of defined mutations"
- Kim Worley "Search services to improve the identification of expressed sequences and their functions"

5:50 pm 8:00 pm DINNER (on your own)

8:00 pm 10:00 pm DISCUSSION SECTION

Sunday, October 31, 1999

8:30 am 12:00 pm SESSION 5

Chairmen: Thomas Werner and Sherman Weissman

8:30 am 9:40 am

- James Fickett "Finding gene boundaries on large contigs"
- Ezra Jennings "Dissecting transcriptional circuitry and mechanisms in yeast"
- John Welsh "The RAP-array approach to cDNA array hybridization"
- Richard Mural "Extracting meaningful information from draft sequence"

9:40 am 10:10 am BREAK

10:10 am 12:00 pm

- Xing Hua Pan "Approaches towards global profiling of cDNA mutations"
- Peter Nelson "Functional dissection of the RANTES promoter: Insights into mechanisms of tissue specific regulation of transcription"
- Giuseppe Borsani "A tale of two diseases"
- Bruce Aronow "Stepwise chromatin restructuring and unexpected rules for interaction of distributed cis-elements that form a locus control region in vivo"
- Christian Overton (Abstract not available)

11:45 am CLOSING REMARKS

Beyond the Identification of Transcribed Sequences:
Functional and Expression Analysis

9th Annual Workshop
October 28-31, 1999

1999 Invited Speakers

(*Accepted)

David Baillee*
Simon Fraser University
British Columbia,Canada

Georgio Bernardi*
Stazione Zoologica Anton Dohrn
Naples, Italy

Juergen Bode*
Genetik von Eukaryonten
Braunschweig, Germany

Mark Boguski
National Center for Biotech Information
Bethesda, Maryland, USA

Guiseppe Borsani*
TIGEM
Milan, Italy

Alwis Brasma*
EBI
Cambridge, UK

Anthony Brookes
Karolinska Institute
Stockholm, Sweden

W. J. Dickinson*
University of Utah
Salt Lake City, Utah, USA

Johan T. den Dunnen

Michael McClelland*
Sidney Kimmell Cancer Center
San Diego, CA, USA

Danny Morse*
University of Utah
Salt Lake City, UT, USA

Richard Mural*
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Peter Nelson*
Poliklinik LMU
Munich, Germany

Osamu Ohara*
Kazusa DNA Research Institute
Chiba, Japan

Christian Overton*
Philadelphia, PA, USA

Stuart Peltz*
Robert Wood Johnson Medical School
New Jersey, USA

Graziano Pesole*
Dipartimento di Fisiologia
e Biochimica Generali
Milan, Italy

Guenter Plickert*

Leiden University
Leiden, The Netherlands

Laurent Duret*
Universite Claude Bernard
Lyon, France

James Eberwine*
Univ of Pennsylvania
Philadelphia, PA, USA

Jonathan A. Eisen*
TIGR
Bethesda, Maryland, USA

James Fickett*
SmithKline Beecham Pharmaceuticals
King of Prussia, PA, USA

Thomas C. Freeman*
The Sanger Centre
Hinston, Cambridge, UK

Katheleen Gardiner*
Eleanor Roosevelt Institute
Denver, Colorado, USA

Takaski Gojobori*
National Institute of Genetics
Mishima, Japan

Nina Heiss*
German Cancer Research Institute
Heidelberg, Germany

John E. Hesketh*
Rowett Research Institute
Aberdeen, Scotland, UK

Geoffrey G. Hicks
Manitoba Inst of Cell Biology
Winnipeg, Canada

Takashi Ito*
University of Tokyo
Tokyo, Japan

Ezra Jennings
MIT/Whitehead Institute

Barbara Knowles
The Jackson Laboratory

Universitat Koln
Koln, Germany

John H. Postlethwait*
University of Oregon
Eugene, Oregon, USA

Martin Ringwald*
The Jackson Laboratory
Bar Harbor, Maine, USA

Michel Simonneau*
INSERM
Paris, France

Marcelo Bento Soares*
The University of Iowa
Iowa City, Iowa, USA

Roland Somogyi*
NIH
Bethesda, Maryland, USA

Eugene Sverdlov*
Russian Academy of Science
Moscow, Russia

Gunnar von Heijne*
Arrhenius Laboratory
Stockholm Sweden

Harald von Melchner*
Univ of Frankfurt
Medical School
Frankfurt am Main, Germany

Wyeth Wasserman*
Karolinska Institute
Stockholm, Sweden

Sherman Weissman*
Yale Univ Schl of Med
New Haven, CT, USA

Thomas Werner*
GSF-National Research Center for Environment and Health
Neuherberg, Germany

Stefan Wieman*
German Cancer Research Center
Heidelberg, Germany

Bar Harbor, Maine, USA

Yuji Kohara*
National Inst of Genetics
Mishima, Japan

Benjamin Koop
University of Victoria
Victoria, British Columbia
Canada

Bernhard Korn*
German Cancer Research Center
Heidelberg, Germany

Shuo Lin*
Medical College of Georgia
Augusta, Georgia, USA

Edgar Wingender*
GBF Braunschweig
Germany

Kim Worley*
Baylor University School of Medicine
Houston, Texas, USA

Yi Zhou
National Institutes of Health

DICKINSON

Evolutionary perspectives on functional analysis: some cautionary tales

W. J. (Joe) Dickinson

University of Utah, Department of Biology, Salt Lake City, Utah, USA

Three studies on gene regulation and function conducted in explicitly evolutionary contexts provide cautionary lessons potentially relevant to large-scale functional analyses. 1) Closely related species can have dramatically different patterns of expression of homologous genes. It is not clear how (or even whether) most of these regulatory differences relate to significant biological differences. One interpretation is that patterns of regulation are not always tightly related to essential (or even important) sites of action. 2) Conversely, patterns of regulation of genes that control development can be highly conserved even when the morphological "output" differs significantly between species. Together, these results suggest that, in some cases, analysis of gene expression may not be very informative. 3) It is well established that many genes (probably a majority) have no essential function. However, most such genes seem to make small ("marginal") contributions to fitness under routine conditions, presumably by improving efficiency and/or reliability. It is possible that mutations in some of these genes will reveal more dramatic phenotypes under specific environmental conditions and/or in appropriate genetic backgrounds (e.g., mutant in "redundant" loci), but this need not be so. In other words, even "knockouts" of many genes may never produce a phenotype detectable by any means short of laborious measurements of long-term fitness consequences.

Return to Table of Contents

RINGWALD

Gene expression database for the laboratory mouse

M. Ringwald, Jon Beal, D. Begley, G. Davis, J. T. Eppig, K. Frazer, D. Hill, J. Kadin, J. Richardson, M. Sasner, L. Trepanier

The Jackson Laboratory, Bar Harbor, Maine, USA

The Gene Expression Database (GXD) is a community resource of gene expression information for the laboratory mouse. The database is designed as an open-ended system that can integrate many different types of expression data. Thus, as data accumulate, GXD can provide increasingly complete information about what transcripts and proteins are produced by what genes; where, when and in what amounts these gene products are expressed; and how their expression is affected in different mouse strains and mutants. GXD is available at http://www.informatics.jax.org/. It is integrated with the Mouse Genome Database (MGD). Interconnections with sequence databases and with databases from other species place the gene expression information in the larger biological and analytical context.

Since the release of GXD 1.0 in June 1998, new expression data are made available on a daily basis. Data are acquired from the literature by editorial staff and via electronic submission from laboratories. We continue to work with 'large scale expression' groups that maintain their own laboratory databases from which we can download data in bulk. In addition, we have developed a flexible and customizable tool, the 'Gene Expression Notebook', that will enable conventional laboratories to manage expression data locally and to submit data to GXD. GXD currently includes RNA in situ hybridization and immunohistochemistry data, Northern and Western blot data, RT-PCR data, RNAase protection data, and mouse cDNA/EST data. Future releases will include additional types of expression data, in particular those derived from the analysis of high density cDNA arrays.

Through collaborations with the SwissProt database and the Unigene project, we are establishing new links between sequences and genes. Interconnections with the Swissport database will provide access to structural and biochemical classification schemes. Further, we began, together with MGD, Flybase, and the Saccharomyces Genome Database, to build shared controlled vocabularies to describe biological processes, and cellular function and location of gene products. These links and classification schemes, combined with skilled data curation, will provide important new search parameters for expression data.

The Gene Expression Database is supported by NIH grant HD33745.

Return to Table of Contents

OHARA

Analysis of uncharacterized human cDNAs which encode large proteins in brain: The Kazusa Approach

Osamu Ohara

Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

We have conducted a cDNA sequencing project for these five years. To date, more than 1200 cDNA sequences have been determined and deposited to the public databases. Since the average size of our cDNAs is about 5 kb, the total number of the sequenced nucleotide residues exceeds 6 Mb. These sequence data greatly help interpretation of the human genome sequence and allow us to discover many interesting genes. These achievements are highly important in human genomics, but they are only early fruits of our project; our cDNA project has been also looking beyond the identification of unknown human transcripts. This is why we have taken quite a unique approach for human cDNA analysis, i.e., analysis of long cDNAs (>4 kb) which encode large proteins (>50 kDa). This approach makes our cDNA project clearly distinctive from others.

The reasons why we decided to take this approach are as follows: (1) Long cDNAs are missing pieces in human cDNA analyses by others currently going on worldwide; (2) a significant number of large proteins are specific to mammals, if we define proteins larger than 100 kDa as large proteins; (3) in most cases (>75 %), functions of the large proteins in mammals are classified into either cell structure/motility (e.g., cytoskeletal proteins, membrane skeletal proteins, and motor proteins), or cell communication/signaling (e.g., ion channels, receptors, adhesion molecules, and regulators of small-G proteins), or nucleic acid management (e.g., transcription factors, RNA binding proteins, and DNA binding proteins); (4) many positionally cloned gene products are large proteins consisting of multiple domains. Because we have been particularly interested in brain functions and genetic causes of neurological disorders, all these reasons strongly motivated us to implement analysis of long cDNAs encoding large proteins.

In my presentation, the data obtained so far will be overviewed through introduction of our database for Human Unidentified Gene-Encoded large proteins (the HUGE protein database, http://www.kazusa.or.jp/huge). In addition, I would like to describe several new technical developments for achieving the analysis of large cDNAs because many serious problems have been overlooked or only vaguely anticipated in current cDNA analysis.

Return to Table of Contents

HEISS

Functional characterisation of genes in Xq28

Nina S. Heiss, Zdenek Sedlacek, Annemarie Poustka

Deutsches Krebsforschungszentrum (DKFZ), Department of Molecular Genome Analysis, Heidelberg, Germany

The systematic generation and integration of high-coverage physical and transcript maps in the chromosomal region Xq28 has facilitated the isolation of most genes in this region, as well as the identification of genes causally associated with diseases. We are currently analysing the developmental and tissue-specific expression of a selection of genes in Xq28 by RNA in situ hybridisation. Functional analyses include determination of the intracellular localisation of the corresponding proteins by using the enhanced green fluorescent protein (EGFP) as a tag, as well as generating antibodies and working on murine models in some cases. To complement the functional analyses on genes in Xq28 we are also carrying out evolutionary studies in this region. The DKC1 gene is responsible for causing the bone marrow failure syndrome dyskeratosis congenita (DKC) and is one example of a gene which we are characterising at the RNA expression and protein level. The DKC1 transcript is expressed ubiquitously and very early in murine embryological development. The protein dyskerin is highly conserved and appears to be involved in the pseudouridylation and cleavage of pre-rRNA. Dyskerin fused to the EGFP and expressed in mammalian cell lines localises to the nucleoplasm, the nucleoli, and the coiled bodies which are functionally associated with the nucleoli. Mutant constructs permitted a delineation of the sequences responsible for the nuclear targeting and indicated that mislocalisation of dyskerin is not a mechanism involved in the pathogenesis of DKC. Examples of genes in Xq28 being studied at the evolutionary level are rab GDI and XAP5. The autosomal paralogues of genes related to rab GDI and to XAP5 were isolated and mapped, and the amphioxus and fugu orthologues of the rab GDI genes were isolated. A comparison of the tissue-specific expression indicates that the various homologues have specialised functions.

Return to Table of Contents

BERNARDI

The formation and maintenance of the isochores from the genomes of warm-blooded vertebrates is due to natural selection

Giorgio Bernardi

Laboratorio di Evoluzione Molecolare, Naples, Italy

Ten years after the discovery of isochores (Macaya et al., 1976), it was proposed (Bernardi and Bernardi, 1986) that their formation and maintenance were due natural selection. This stirred a debate which is still going on. The most recent contribution along this line (which certainly will no be the last) was a correspondence with the peremptory title: "Isochores result from mutation not selection" (Francino and Ochman, 1999). In my presentation I will summarize the arguments, old and new, in favour of natural selection.

[1] Macaya, G., Thiery, J.P., Bernardi, G., (1976) J. Mol. Biol. 108, 237-254.
[2] Bernardi, G., Bernardi, G., (1986) J. Mol. Evol. 24, 1-11
[3] Francino, M.P., Ochman, H., (1999) Nature 400, 30-31

Return to Table of Contents

KOHARA

Post-genomics strategies in C. elegans toward understanding of the molecular mechanisms of development

Yuji Kohara

Genome Biology Lab, National Institute of Genetics, Mishima, Japan

We think the main targets in the post genomics era are as follows: (1) to integrate the information on the expression and the function of all genes of the genome in the context of development, (2) to extract the rules that govern the molecular mechanisms of development that are carried out by a finite number of genes, (3) ultimately to reconstruct the developmental process in the computer.

We have identified about 10,000 cDNA species (more than a half of all genes) of the nematode C.elegans through our EST project. About 4,000 cDNA species mostly from chromosome 3 (autosome) and X (sex chromosome) have been analyzed by using of whole mount in situ hybridization for mRNA distribution throughout the life of worm. The mRNA patterns were classified into several categories; maternal expression, zygotic expression, expression in a specific cell lineage but at different time, expression in specific cell(s) and so on. Based on the information, we are performing clustering analyses and finer analysis on subsets of the genes to elucidate the network of gene regulation. As a trial, focusing on the early embyrogenesis, we selected a subset of 100 genes on the criteria that the mRNA was expressed maternally and disappeared quickly before gastrulation, which seemed to play important roles in this period. For these genes, we are analyzing (i) the phenotypes caused by the systematic RNAi (dsRNA mediated interference) experiments, (ii) the protein distribution through the systematic raising of antibodies, and (iii) interacting genes by the yeast two hybrid technology. To integrate and analyse the information, first we have established a WWW-based database NEXTDB, and recently we are constructing a computer graphics based 4-dimensional (3D + time course) database that covers the early embryogenesis.

Return to Table of Contents

GARDINER

Genomic sequence analysis and novel gene characterization on human chromosome 21

Katheleen Gardiner, Dobromir Slavov, Roger Lucas, Andrew Fortna and Alla Rynditch

Eleanor Roosevelt Institute, Denver, Colorado, USA

Sequencing centers in Japan and Germany anticipate the completion of the entire sequence of human chromosome 21 by early 2000. Currently, >25 Mb of the total of 40Mb have been deposited in public databases. We are analyzing this sequence for gene and repeat sequence content, and for interesting/important features of genome organization. Novel genes and gene models discovered in this analysis are also being analyzed for expression patterns and for alternative processing. At present, we have detailed information on >7.5 Mb, deriving from various isochore classes and including ~1 Mb with a GC level >50%. Results in the following areas will be presented:

i) Gene identification: 45 genes have been found that either have previously described protein sequences or that represent new homologies or new members of gene families. For each of these, the number and size of all introns and exons (coding and noncoding) have been determined. Associations with CpG islands have also been assessed. A further 45 novel gene models are predicted based on exon predictions, CpG island identification and EST matches. In addition to models comprised of excellent exon predictions associated with dbEST matches, we also find excellent exon predictions lacking ESTs and EST matches lacking good exon predictions. This suggests a fourth class of novel gene may be lurking in the human genome: those lacking both ESTs and reliable exon prediction.

ii) Genome organization: More than 3/4 of "known" genes are associated with CpG islands; more than 1/2 of novel models likely are. Within 6.5 Mb, 14 of 41 known genes contain 22 introns >20 kb in size. In total, these large introns comprise >1 Mb, more than 15% of the DNA analyzed. In contrast to large amounts of intronic DNA, there are numerous examples of short intergenic distances. Of 24 well defined intergenic distances, 16 are less than 5 kb and 8 are less than 2 kb. In addition, two examples of overlapping/interdigitated genes are predicted.

iii) Expression analysis: Of 19 gene models tested, all but one were negative by Northern analysis. RT-PCR has been used to define expression in 11 tissues, including 8 and 10 week fetus and adult brain regions. The majority of genes show restricted expression and interesting patterns of alternative processing. Of seven models for which complete protein sequences have been deduced, five have no discernible protein homologies and two show only weak homologies.

In spite of significant successes, these analyses serve to emphasize several problems associated with attempts to interpret human genomic sequence information. These problems and potential solutions will also be discussed.

Return to Table of Contents

FREEMAN

Development of expression profiling technologies for the rat

Tom Freeman, David Vetrie, Clare East, Adam Butler, Liz Campbell and Peter Wooding

Gene Expression Group, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, United Kingdom

The rat is a widely used animal model for biological investigation including many pharmacological, neurobiological and physiological studies. Over the last few years there has been a significant increase in the amount of information available on the rat genome and genetics. We have therefore begun to develop the tools to allow us to exploit this new information resource in order to perform large-scale expression studies of rat systems. We have developed a cDNA sequence-based database named RATACE, based on the widely used platform of ACeDB. It currently contains all the publicly available rat sequence, with the exception of ESTs. Sequence comparison has allowed sequences originating from the same gene to be identified and clustered, and sequence alignments can be viewed graphically from within the database. We are currently developing it as a tool for handling all the reagents and data required for, and generated by, large-scale expression studies. Expression studies are being carried out using RT-PCR and high-density microarrays. To date, we have over 1,000 working primer pairs for rat genes and have used these to profile the expression of each gene over a range of 22 rat tissues using RT-PCR. We have also been optimising the conditions necessary for spotting these PCR products on glass slides to act as probes for microarrays, as well as determining how they perform during hybridisation. We are currently scaling up the production of probes for the rat microarrays. It is our intention to use these tools on rat pharmacological models to further our understanding of the mechanism of drug action, as well as to identify novel therapeutic targets.

Return to Table of Contents

DURET

Codon usage in animals and plants: new data, new paradox

Laurent Duret and Dominique Mouchiroud

Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard, Villeurbanne Cedex, France

We analyzed the expression pattern and codon usage in 8133, 1550, 2917 and 760 genes respectively from Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana and human. In human, codon usage is correlated to the base composition of the genomic context (isochore) in which the gene is embedded, but not to its expression pattern. Variations of codon usage among human genes probably result of neutral evolution due to mutational biases rather than of selection for optimal gene expression.

Interestingly, in the three other species, we observed a clear correlation between codon usage and gene expression levels, and showed that this correlation is not due to a mutational bias. This provides direct evidence for selection on silent sites in those three distantly related multicellular eukaryotes. Surprisingly, there is a strong negative correlation between codon usage and protein length. This effect is not due to a smaller size of highly expressed proteins. Thus, for a same expression pattern, the selective pressure on codon usage appears to be lower in genes encoding long rather than short proteins. This puzzling observation is not predicted by any of the current models of selection on codon usage and thus raises the question of how translation efficiency affects fitness in multicellular organisms.

Return to Table of Contents

THIESEN

Genomic sequences of KOX zinc finger gene clusters are used to determine (auto) regulatory networks of gene regulation

Dirk Koczan[1], Christian Sina[1], Peter Lorenz[1], Tom Hearn[2], Saleh Ibrahim[1], Andreas Rump[3], Andre Rosenthal[3], Michael Jackson[2], M.-F. Rousseau-Merck[4], I. Legrand[4], and Hans-Juergen Thiesen[1]

[1]Institute of Immunology, University of Rostock, Rostock, Germany
[2]Human Molecular Genetics Unit, Univ Newcastle upon Tyne, United Kingdom
[3]IMB, Jena, Germany
[4]Institute Curie, Paris, France

The evolutionarily conserved KRAB domain is encoded in about one-third of all human Krueppel-type zinc finger genes. Initially, the KRAB domain of KOX1 has been identified to display the strongest repression activity identified in mammalian organisms. To study regulatory networks of gene regulation, high density PAC zinc finger gene filters have been generated to map Krueppel-type cDNAs to the human genome. 364 PAC zinc finger clones initially identified by a mixture of KOX cDNAs (KOX3, 4, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, 22 and 28) were mapped by FISH to individual chromosomes. In particular, a contig of PAC clones encoding human KRAB zinc finger genes has been generated representing the zinc finger gene cluster on chromosome 10p11.2 encoding the genes of KOX19 (ZNF25), of KOX21 (ZNF37) and of KOX31 (ZNF33) extended by two novel ZNF genes designated ZNF248 and ZNF249 physically linked to ZNF25. To determine whether these genes are coregulated within their clusters Northern blots complemented by TaqMan analysis were performed. Interestingly, differential splice forms were detected for KOX19 and their distribution are currently quantitated by quantitative real time RT-PCR (TaqMan) analysis. In addition, more than 300.000 bp describing the zinc finger gene cluster of chromosome 10p11.2 are currently instrumentalized to determine DNA binding sites for KOX1 and KOX19 proteins by making use of our TDA selection system. Our ongoing analysis indicates the presence of autoregulatory networks within and between zinc finger gene clusters mediated by target sequences of repetitive nature.

Return to Table of Contents

EISEN

Phylogenomics and the benefits of an evolutionary perspective in genome analysis

Jonathan A. Eisen

The Institute for Genomic Research, Rockville, Maryland, USA

It is generally accepted that comparative studies can greatly benefit the interpretation and analysis of genome sequences. Comparative molecular biological studies usually focus on quantifying and qualifying the levels and types of similarity or differences in molecular features within and between species. I discuss here how and evolutionary perspective is beneficial in comparative studies because it allows one to focus not just on the similarities and differences but in how and even why they came to be. In particular I will discuss the combination of evolutionary inference and genome analysis, which I refer to as phylogenomics. I will discuss four different aspects of phylogenomics: 1) prediction of gene function; 2) inferring evolutionary events such as gene loss, duplication and gene transfer and 3) inferring mutational processes and 4) confirming and predicting protein structures. In addition, I will discuss the technique of evolutionary genome scanning which allows one to identify genes that may contribute to specialized features of a particular organism. The discussion of phylogenomics and evolutionary genome scanning will focus on some of the genomes and chromosomes that have been recently completed at TIGR.

Return to Table of Contents

PELTZ

Understanding the role of mRNA turnover in regulating gene expression and in disease states

Stuart Peltz

Robert Wood Johnson Medical School-UMDNJ, Piscataway, New Jersey, USA

It is now clear the regulation of mRNA decay provides a powerful means of controlling gene expression. The mechanisms and structural features that dictate mRNA decay rates, and their regulation, are now being defined. These studies have revealed that the processes of translation and mRNA turnover are intimately linked. A clear example of this link is the observation that premature translation termination enhances mRNA decay rates, a phenomena called nonsense-mediated mRNA decay (NMD). Transcripts containing premature nonsense codons are rapidly degraded, thus preventing synthesis of incomplete and potentially deleterious proteins. There are well over two hundred genetic disorders which can result from premature translation termination. Understanding how this process affects translation termination and mRNA degradation can lead to rational approaches for the treatment of these disorders. We have been characterizing the sequence and factors involved in NMD. These results have identified a "surveillance complex" that monitors the translation process and determines whether translation termination has occurred at the correct position within the mRNA. These results have led to a model that in which the surveillance complex assesses translation termination by monitoring the transition of an RNP as it is converted from a nuclear to a cytoplasmic form during the initial rounds of translation.

Return to Table of Contents

EMANUELSSON

Identifying localization signals in proteins

Olof Emanuelsson

Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

The subcellular localization of a protein is an important characteristic with implications for protein function. Various methods have been developed to predict protein localization, both based on the detection of targeting signals and on differences in overall amino acid composition between proteins from different compartments. In particular, we have developed the widely used SignalP program (for the identification of signal peptides in secretory proteins), as well as ChloroP and MitoP for chloroplast and mitochondrial targeting peptides, respectively. We are now combining these prediction methods into a global protein localization prediction scheme that should be useful for, e.g., automatic database annotation.

Return to Table of Contents

KORN

Tools for functional analysis using resources of RZPD

B. Korn[1, 2], J. Boer[1], P. Kioschis[1,2], A. Vente[3], H. Lehrach[3,4] and A. Poustka[1,2]

[1] Abteilung für Molekulare Genomanalyse, DKFZ, Heidelberg, Germany
[2] RessourcenZentrum am DKFZ, Heidelberg, Germany
[3] RessourcenZentrum am Max-Planck-Institut für Molekulare Genetik, Berlin-Dahlem, Germany
[4] Max-Planck-Institut für Molekulare Genetik, Berlin-Dahlem, Germany

Since the launch of the German Human Genome Project (DHGP) by the Federal Ministry of Education, Science Research and Technology (BMBF) and the Deutsche Forschungsgemeinschaft (DFG) in 1995, the central unit (Resource Center) does provide standardised material and technology for genomics. In order to mimic the development in the field, we shifted our focus of product development towards functional genomics.

Collections of minimal gene sets for human, mouse and rat as well as indication-specific sets for oncology, immunology, haematology, and toxicology have been collected and characterised. These clone sets are made available as high density filter arrays, together with accompanying database information. Furthermore, we plan to provide courses to introduce the users of RZPD to different kinds of analysis tools for expression profiling using these gene sets. A genome wide project is underway to further improve the target material for expression profiling. The use of whole genome expression profiling in cancer research will be shown. Cancer cells are genetically different from normal cells, and exhibit aberrant gene expression. Knowledge of the changes in gene expression typical for certain types and stages of tumours, could give insight into the molecular changes involved in tumour development and progression, and provide molecular markers for tumour diagnosis and prognosis.

We generated gene expression profiles for kidney, breast and brain tumours and normal tissues through the complex hybridisation of high-density Nylon arrays with radioactively labelled whole tissue cDNA. Array expression data for renal clear cell carcinomas confirmed overexpression of several genes known to be upregulated in renal carcinomas, e.g. VEGF, vimentin, and haptoglobin. Other known genes that have not previously been implicated in kidney cancer were found overexpressed in the tumours, including beta-2-microglobulin, thymosin beta-4, and DNAX activation protein 12. In addition, ESTs similar to angiopoietin, and ESTs without homologies were found. Underexpressed genes in renal tumours compared to normal tissue included five members of the metallothionin family and several unknown genes (ESTs). We will compare hybridizations using whole tissue tumour and normal cDNAs with suppression subtraction cDNAs, with the aim to increase the sensitivity.

Linking the tumour hybridisation data with histopathological and clinical data in a queryable relational database will allow correlation of gene expression and tumour characteristics. By comparing at least 10 to 20 well-characterised tumour/normal pairs of the same type and stage, we expect to find significant tumour-specific expression patterns and co-ordinated changes in the expression of multiple genes.

Return to Table of Contents

WASSERMAN

De novo discovery of transcription factor binding sites in co-regulated sets of human genes

Wyeth W. Wasserman

Center for Genomics Research, Karolinska Institute, Stockholm, Sweden

Patterns of gene expression change in response to developmental, physiological, and environmental signals. Individual gene expression may range from selective (observed under specific-conditions) to broad (observed in most cells). While several biochemical mechanisms exist for the regulation of gene expression, the most significant switch is at the level of gene transcription. For any group of genes selectively expressed in a common subset of tissues, transcriptional regulation appears to be directed by a tissue-related handful of interacting transcription factors. These factors bind to locally dense clusters of sites (modules) which can be situated in close proximity to or at great distances from the transcription start site of each gene in the set. Progress has been made in generating predictive models for the detection of modules which selectively direct expression to well-studied tissues like skeletal muscle or liver, but discovery of modules in the absence of extensive experimental information remains a challenge. By fusing the comparative analysis of genomic sequence with modeling of regulatory modules, progress has been made in developing methods for the de novo discovery of classes of regulatory control elements.

Return to Table of Contents

FONTES

Using large cloned genomic fragments as expression vectors to create or correct mutant phenotypes

E. Passage and M. Fontés

INSERM U491, Marseille Cedex 5, France

The development of genomic resources (YAC, BAC, PAC), which have been largely used for gene hunting and systematic sequencing, ask the question of using these clones as expression vectors. These clones will have the following advantages: - One would expect there to be stable expression which would last for the whole of the transfected cell's life - Expression levels are similar to those found for endogenous copies. - The control of transgenic expression should perfectly reproduce that of the endogene. This removes the risk of adverse phenotypic effects resulting from expression of the transgene where it is not normally produced. Another advantage of this approach is that expression is not dependent on the site of integration. - This method does not produce any inflammatory, allergic or immune responses as it has been described in viral based gene therapy approaches.

We have thus developed this approach to create mouse models of human inherited disease, by injection of YAC or PAC in the murine oocyte. Two models will be presented: Charcot-Marie-Tooth disease type 1A (YAC injection) and Creutzfeld-Jacob/Prion Diseases (PAC injection). We may note that this approach can be used to check expression and provide information on the role of genes in a given genomic fragment.

Moreover, we are now developing this technology in a gene therapy approach, using ex vivo or in vivo approaches, the strategy we will used for Polykystic Kidney Disease type 1 gene therapy, will be presented.

Return to Table of Contents

SOMOGYI

Expression profiling and genetic networks

R. Somogyi, S. Fuhrman, X. Wen, P. D'haeseleer*, S. Liang, and J.F. Loring

Incyte Pharmaceuticals Inc., Palo Alto, California, USA
*University of New Mexico, Department of Computer Science, Albuquerque, New Mexico, USA

Biological information is transmitted from gene sequence to gene activity patterns. This information feeds back to the regulation of gene expression through a system of inter- and intracellular signaling functions. We have conducted extensive surveys of the dynamics of gene expression to capture essential information flow in genetic signaling networks. Cluster analysis of these data suggests conserved modules of genetic programs and interlinked pathways. Ultimately we seek to build truly predictive models by applying reverse engineering approaches to these data. Using a linear method we have identified several plausible causal links between genes. The knowledge gained through these integrated experimental/computational methods will be critical to therapeutic target discovery and validation, and bioengineering in general.

(http://psb.stanford.edu/psb99/genetutorial.pdf).

Return to Table of Contents

ZHOU

Gene expression profiling in molecular pharmacology

Y. Zhou, U. Scherf, M. Waltham, W.C. Reinhold, L.H. Smith, J.K. Lee, D.A. Scudiero, E.A. Sausville, D. Ross[1], M. Eisen[1], D. Botstein[1], P.O. Brown[1], L. Miller, E. Liu and J.N. Weinstein

National Cancer Institute, Bethesda, Maryland, USA
[1]Stanford University Medical Center, Stanford, California, USA

High-density cDNA microarrays and oligonucleotide chips provide an opportunity to study gene expression in a parallel fashion, often thousands of genes at a time. In the context of molecular pharmacology, large scale gene expression profiling can be used to identify genes whose expression patterns predict the activity patterns of drugs. Since 1990, the National Cancer Institute's drug discovery program has used 60 cancer cell lines to screen more than 70,000 chemical compounds for anticancer activity. The resulting patterns of activity encode incisive information about mechanisms of drug action and resistance (Paull, et al., JNCI 81:1088, 1989; Weinstein, et al., Science 258:343, 1992 and 275:343, 1997). We have recently studied gene expression profiles of the 60 cancer cell lines using two-color fluorescence detection on glass slide cDNA microarrays containing approximately 9,000 genes (Ross, et al. and Scherf, et al., submitted). The results are, or will soon be, displayed at our web site (http://discover.nci.nih.gov).

Those studies delineated sensitivity to treatment. We are also analyzing in detail the molecular consequences of therapy as a function of dose and time after treatment for a selected drug-cell pair. Both types of experiments provide information potentially useful for delineation of molecular pathways, clinical selection of therapy, and discovery of new anticancer agents. Funded in part by a grant from the NCI Breast Cancer Think Tank.

Return to Table of Contents

HOWELL

Annotation of the human X chromosome sequence

Gareth R. Howell, Alison J. Coffey, Susan Rhodes, Robert A. Brooksbank, Shirin S. Joseph, Jackie M. Bye, Andrew King, Laurens Wilming, David R. Bentley and Mark T. Ross

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

The Sanger Centre is sequencing approximately 110 Mb of the human X chromosome between DMD (Xp21.1) and DXS532 (Xq27.3). We have so far produced 35 Mb of finished sequence from this region, together with a further 11 Mb of unfinished. Annotation of transcribed regions begins with the analysis of finished sequence for nucleic acid and protein homologies and for the predicted presence of exons, genes and CpG islands. Analysis results are displayed graphically in acedb. Primer pairs are designed to strongly predicted exons and used to screen PCR pools of cDNA clones from a collection of tissues. Verification and extension of partial gene structures is effected using single-side specificity (SS)PCR within the positive cDNA pools. The sequences of the SSPCR products are added to the acedb display to illustrate the experimental confirmation of the genomic sequence predictions.

Most of our effort so far has been in Xq23-q27, where genomic sequencing is most advanced. Here we have so far identified and established at least partial structures for 45 genes, including the SH2D1A gene, mutations of which result in X-linked lymphoproliferative disease (XLP). The sequence annotated with these gene structures is made available using webace, a web-based version of the acedb display. Further details can be obtained from our X chromosome WWW page at http://www.sanger.ac.uk/HGP/ChrX/.

For part of this region (Xq25) we have initiated a small scale comparative sequencing project in the syntenic region of the mouse X chromosome. The region contains at least four genes in human (CXorf3, XIAP, SH2D1A and Tenascin-M). The availability of the mouse sequence should provide us with valuable information on the power of comparative sequencing for confirmation of complete gene structures, detection of genes not found using the approach described above, and identification of potential gene regulatory elements.

Return to Table of Contents

SOARES

Development of non-redundant arrays of full-length cDNAs

Maria F. Bonaldo[1], Brian Berger[1], Todd Scheetz[1, 2], Kyle Munn[2], Chad Roberts[2], Kang Liu[2], Thomas Casavant[2] and Marcelo Bento Soares[1, 3]

Departments of [1]Pediatrics, [2]Electrical and Computer Engineering, and [3]Physiology and Biophysics, University of Iowa, Iowa City, Iowa, USA

Large-scale programs are soon going to be initiated aimed at the generation of complete and accurate sequence of large numbers of full-length cDNAs. It is anticipated that the resulting information will not only expedite the identification of human disease-causing genes but also will facilitate the process of annotation of the genomic sequences currently being generated. Central to this effort, however, is the existence of non- redundant sets of full-length cDNAs. Although several methods have been developed for construction of libraries enriched for full-length cDNAs, there aren't any publicly available arrayed sets of full-length cDNA clones. That will require en masse identification of full-length clones in libraries enriched for full-length cDNAs, clustering for identification of a non-redundant set of cDNAs, and their rearraying into 96 or 384 well plates. Towards this goal, we have generated and thoroughly characterized four cDNA libraries enriched for full-length cDNAs that we constructed from size fractionated human germinal center B cell cytoplasmic mRNA. Our strategy to generate non- redundant arrayed collections of full-length cDNAs involves the following steps:

(1) Generation of 5'ESTs from a large number of clones,

(2) Clustering for identification (and sequence assembly) of a non-redundant collection of cDNAs/ESTs,

(3) Blast analysis for identification of cDNAs corresponding to mRNAs for which sequence information is not yet available in Genbank,

(4) Informatics analysis with a "5' classifier program" for identification of 5'ESTs likely to encompass the start codon,

(5) re-arraying of a non- redundant set of full-length cDNAs.

To date we have generated in excess of 13,000 5'ESTs and we have started to build a non-redundant collection of full-length cDNAs.

Return to Table of Contents

WERNER

Promoter prediction in large genomic sequences: compromise between sensitivity and specificity

Scherf, M., Klingenhoff, A., and Werner, T.

Institute for Mammalian Genetics, GSF-National Research Center for Environment and Health, Neuherberg, Germany

The availability of large anonymous DNA sequences produced by the current genome sequencing projects often runs far ahead of functional analyses and cDNA projects. This necessitates an initial analysis that is based solely on computer prediction of genes and other units within the sequences. Both the identification of coding exons and regulatory sequences (e.g. promoters) face very similar problems connected with false positive (specificity) as well as false negative (sensitivity) predictions. A new method for the detection of polymerase II core promoters developed in our group will be introduced which is very specific but is of limited sensitivity. The advantages and disadvantages of several systems for promoter prediction will be compared on several examples. During this study it became clear that strategies mixing existing knowledge and experimental results with in silico predictions are most promising in the quest for promoter annotation of anonymous sequences. This finding parallels results already known from exon predictions. There is also apparently an inevitable trade-off between the specificity and the sensitivity of in silico predictions which cannot be completely overcome by combining different methods. It became very clear that individual methods have different defined ranges of sequence length where they are most effective.

Return to Table of Contents

HESKETH

Perinuclear mRNA localisation by 3'untranslated sequences

John Hesketh[1], Marilyne Levadoux[1], Gillian Dalgleish[1], John Beattie[1], Heather Wallace[2], and Jean-Marie Blanchard[3]

[1]Rowett Research Institute, Buckburn, Aberdeen, Scotland
[2]Dept of Medicine and Therapeutics, University of Aberdeen, United Kingdom
[3]IGMM-CNRS, Montpellier, France

There is increasing evidence that mRNAs can be found localised in different regions of the cytoplasm. Such mRNA localisation occurs in a variety of cells from yeast to mammals, as does the association of mRNAs with the cytoskeleton. Both association of mRNAs with the cytoskeleton and mRNA localisation depend in the majority of cases on signals within the 3' untranslated sequences (3'UTRs) of the mRNAs. Using cell fractionation techniques the mRNAs encoding c-myc, c-fos, ribosomal proteins L1 and S6, cyclin A and metallothionein (MT1) have been found associated with the cytoskeleton. These mRNAs all code for proteins which under all or some circumstances are imported into the nucleus after synthesis; e.g. MYC and FOS are transcription factors and MT1 can be found in the nucleus during

S-phase of the cell cycle. In situ hybridisation shows that c-myc and MT1mRNAs are localised around the nucleus. In addition analysis of mammalian cell lines expressing chimaeric gene constructs in which coding or reporter sequences are linked to different 3'UTRs shows that c-myc, c-fos and MT1 3'UTRs all contain sequences capable of targeting mRNAs to the perinuclear cytoplasm and the cytoskeleton.

Recently, we have developed clonal cell lines transfected with constructs which produce either normal localisation of MT1-encoding transcripts or delocalised transcripts. Using these cell lines it has been found that loss of localisation of MT1transcripts is associated with a lack of MT1 in the nucleus at the G1-S transition of the cell cycle.

These data lead to the hypothesis that 1] certain mRNAs are targeted to the perinuclear cytoplasm by signals in their 3'UTRs and that this functions to promote the efficient imported of the newly synthesised proteins into the nucleus, and 2] this mechanism is a general one utilised by a range of nuclear proteins and proteins which shuttle from cytoplasm to nucleus during the cell cycle. To identify other mRNAs which may be localised in this way the strategy is now to define the localisation signal required for c-myc and metallothionein mRNA localisation, to predict a concensus structure and then search databases to identify other mRNAs containing a similar signal.

Return to Table of Contents

MORSE

Substrate specificity of adenosine deaminases that act on RNA (ADARs) in C. elegans

Daniel P. Morse, Leath A. Tonkin, P. Joe Aruscavage, and Brenda L. Bass

Department of Biochemistry/HHMI, University of Utah, Salt Lake City, Utah, USA

Adenosine deaminases that act on RNA (ADARs) comprise a family of RNA editing enzymes that convert adenosines to inosines within double-stranded regions of RNA. What are the biological functions for this type of RNA editing? In mammals, one function is to alter codons allowing the production of more than one form of a protein from a single mRNA.

To expand our knowledge of ADAR function, we have exploited the tools available in the nematode C. elegans. We developed a method to identify new ADAR substrates (Morse and Bass, PNAS 96:6048-53) and, with the aid of the completed C. elegans genome sequence, have now found seven edited transcripts in the worm. Unlike mammals, editing of these C. elegans RNAs does not alter codons. Instead, five of the substrates are mRNAs edited in untranslated regions, and two are non-coding RNAs.

Examination of the C. elegans genome sequence revealed two predicted open reading frames that potentially encode active ADARs, H15N14.1 and T20H4.4. We have obtained mutant worms containing deletions in one or the other of these genes. Extracts from the H15N14.1 deletion strain exhibited reduced, but measurable, ADAR activity. There was no detectable activity in extracts from the T20H4.4 deletion strain. Consistent with these in vitro results, editing of the seven ADAR substrates was severely reduced or lacking in the T20H4.4 deletion strain, while there were more subtle effects on editing in the H15N14.1 deletion strain. These results suggest that the two C. elegans ADARs cooperate to produce the wild-type editing pattern. We hope that correlation of mutant phenotypes with in vivo editing patterns will reveal specific functions for C. elegans ADARs.

Return to Table of Contents

Beyond the Identification of Transcribed Sequences: Functional and Expression Analysis

9th Annual Workshop, October 28-31, 1999

Co-sponsored by the U.S. Department of Energy

BODE

Scaffold/matrix attached regions (S/MAR Elements): Detection and activities in vivo

Jürgen Bode, Alexandra Baer, Angela Knopp, Dirk Schübeler, Jost Seibler, Craig Benham Armin Baiker , and Hans-Joachim Lipps

GBF - National Institute for Biotechnological Research, Braunschweig, Germany

S/MARs have been discovered more than a decade ago and have been defined as DNA-elements staying attached to or associating with the nuclear skeleton after the extraction of the histones and soluble factors from eukaryotic nuclei. The macroscopic binding properties have been carefully characterized and are now mainly ascribed to the lamins as well as scaffold attachment factor A (SAF-A = hnRNP-U). While S/MARs do not conform to any obvious sequence consensus, their recognition appears to be governed by structural features, most significantly their propensity to expose single strands under negative superhelical tension. This property has been used successfully to localize S/MARs in SIDD (stress-induced duplex destabilization) profiles, to predict their activity and and to guide the design of artificial S/MARs (Review: Gene Therapy Mol. Biol. 1, 551, 1998).

The first generally accepted biological activity of S/MARs was the augmentation of transcription initiation rates which operates by a mechanism distinct from enhancement. Since then the number of potential biological functions has increased and can be classified as follows:

A. Transcriptional Level
    - Augmentation (parallels binding strength in vitro)
    - Long term Stabilization (interference with formation of a 30 nm fiber?)
B. Transcriptional Competence
    - Insulation (position-independence not copy-number dependence)
    - Domain Opening (de-methylation, acetylation, enhancer support)
    - Platform for the Formation of Transcriptosome/Spliceosome complexes
C. Others
    - Origin of Replication (-support);
    - Amplification Promoting Element
    - Recombination Hotspot (genome instability, retrovirus integration target)
    - Targeting Element

This overlap of functions has made difficult the unambiguous demonstration of any of these components. Standard transfection techniques cannot, for instance, discriminate between a targeting action of a S/MARs and a cis-activity of the same element. It was originally for this reason that we have developed a variety of techniques based on site-specific recombination systems like Flp/FRT and Cre/loxP. With these techniques, complete chromatin domains can be decomposed, inverted or elaborated at a predefined chromosomal locus - a concept with obvious relevance for the rational construction of cell lines with a high and consistent expression and the efficient generation of transgenic animals with a predictable regulation of the transgene.. Our most advanced system, the recombinase-mediated cassette exchange (RMCE) permits the exchange of an integrated cassette which is flanked by an FRT-site and a FRT-mutant. During

RMCE, any pre-existing positive/negative marker is removed resulting in an integration event that is not perturbed by either a co-expressed selection marker or prokaryotic vector sequences. It will allow the efficient stepwise, stable (but reversible) introduction of insulator/bordering elements to support the autonomous expression of a transgene at a given genomic site.

So far the stable modification of target cells is mostly achieved by integrating vectors although their expression is rapidly silenced and may give raise to insertional mutagenesis or recombination. While episomal vectors exist, the function of their replication origins usually relies on virally encoded transacting factors which often lead to cellular transformation. Major recent efforts have therefore been devoted to the use of S/MAR-ori sequences from the human genome to obtain vectors which replicate autonomously and thereby provide a stable and high-level expression. We have demonstrated that the function of the large- T oncogen on the SV40 origin can be substituted for by a human S/MAR element and that this combination of S/MAR and ori sequences efficiently prevents integration. There are indications that a S/MAR directs the episome to replication sites on the nuclear matrix, recruiting all the endogenous cellular factors which are required for its propagation and stable extrachromosomal maintenance.

Return to Table of Contents

GOJOBORI

A gene expression profile obtained by determining over 1000 cDNA clones from a single Planarian eye cell

Takashi Gojobori[1], Kazuho Ikeo, Silvana Gaudieri[1], Akira Tazaki[2], Masumi Nakazawa[1], Masafumi Shimoda[3], Yuzuru Tanaka[4], and Kiyokazu Agata[2]

[1]Center for Information Biology, National Institute of Genetics
[2]Department of Life Science, Himeji Institute of Technology
[3]Hitachi Software Engineering Co., Ltd.
[4]Electronic and Information Engineering, Hokkaido University, Japan

Planarians are considered to be the most primitive animal that have acquired a central nervous system (CNS) with longitudinal nerve and sense organs, and their anterior-posterior axis is very similar to vertebrates and higher invertebrates. It has been suggested that the brains of deuterosomes and proteosomes descended from a common ancestor and that common regulatory and important genes have been conserved between vertebrates and invertebrates. Therefore, Planarians are an important model in understanding the brain and its related processes. Here we present the gene expression profile of the Planarian eye obtained by determining over 1000 cDNA clones from a single eye cell. We then discuss functional and evolutionary significance of the gene expression profile obtained, along with the perspective of this kind of analysis.

Return to Table of Contents

SVERDLOV

Towards the understanding of the possible impact of human endogenous retroviruses on evolutionary changes in genome transcription

Eugene D. Sverdlov

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia

Human endogenous retroviruses (HERVs) likely representing footprints of ancient germ-cell retroviral infections occupy about 1% of the human genome. HERVs can influence genome regulation through expression of retroviral genes, genomic rearrangements following HERV integrations, or through the involvement of HERV LTRs in the regulation of gene expression.

Newly integrated HERVs might change the pattern of gene expression and therefore play a significant role in the evolution and divergence of primates. Comparative analysis of HERVs, HERV LTRs, neighboring genes, and their regulatory interplay in the human and ape genomes will help us to understand the impact of HERVs on the evolution and genome regulation. This report will present the data on:

1. Relative positions of HERV LTRs and genes on human chromosomes 7, 19, and 21.
2. The sequence of some individual LTRs' appearance in the primate genomes.
3. The rate of LTR mutations and evolutionary ages of individual LTRs.
4. The occurrence of human specific LTR integrations.
5. The potential of individual LTRs in transcription regulation.

These data as well as the results reported by other authors suggest the evolutionary significance of retroviral invasions in the primate genomes.

The research was partially supported by Russian National Human Genome Project and INTAS 96-1710 grants.

Return to Table of Contents

PESOLE

Structural and evolutionary analysis of eukaryotic mRNA untranslated regions

Pesole G.[1,4], Liuni S.[2,4], Larizza A.[3], Makalowski W.[5] and Saccone C.[3,4]

[1] Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, Milano, Italy
[2] Centro di Studio sui Mitocondri e Metabolismo Energetico, C.N.R., Bari, Italy
[3] Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Bari, Italy
[4] Area di Ricerca del CNR, Bari, Italy
[5] National Center for Biotechnology Information, NLM-NIH, Bethesda, Maryland, USA

The 5' and 3' untranslated regions of eukaryotic mRNAs may play a crucial role in the regulation of gene expression controlling mRNA localization, stability and translation efficiency. In order to study the general structural and compositional features of these sequences we have previously developed UTRdb, a specialized database of 5' and 3' UTR sequences of eukaryotic mRNAs cleaned from redundancy (Pesole et al., 1999).

Utrdb (release 10.0) contains 75,448 entries (26,145,985 nucleotides) which are also annotated for the presence of functional sequence patterns whose biological activity has been experimentally demonstrated. All these patterns have been collected in the UTRsite database where for each functional pattern, corresponding to a specific entry, the consensus structure is reported with a short description of its biological activity and the relevant bibliography. Furthermore, UTRdb entries have been annotated for the presence of repeated elements present in the Repbase database (Jurka, 1998). A total of 5,818 functional elements and 54,975 repetitive elements are annotated in UTRdb. All Web resources we implemented for the retrieval and the analysis of UTR sequences are available at the UTR home page (Pesole and Liuni, 1999) we recently implemented. UTRdb entries can be retrieved through the SRS system where crosslinks to UTRsite as well as to the nucleotide or aminoacid primary database are also established. Through the Web facility UTRscan any input sequence can be searched for the presence of a functional pattern annotated in UTRsite and UTRfasta allows to assess sequence sililarity between a query sequence and UTRdb entries.

The analysis of complete UTR sequences contained in this database showed that 5'-UTR sequences, on the average 187 nucleotides long, were 1,2 to 4,3 times shorter than the corresponding 3'-UTR sequences in the various taxonomic groups considered. As far as the compositional properties were concerned, on average 5'-UTR sequences resulted in all cases GC richer than 3'-UTR sequences and significant correlation was found between the GC content of 5' and 3'-UTR sequences and the GC content of the third silent codon positions of the corresponding protein coding genes (Pesole et al., 1997). Some structural features of 5'UTRs were investigated, such as presence of upstream ORFs and context of initiator ATG, which are known to affect the mRNA translation efficiency. In order to assess the level of functional constraint of UTR sequences we have studied their evolutionary dynamics also in comparison with the corresponding coding regions. With suitable evolutionary models we have calculated the nucleotide substitution rate of 5'-UTR, 3'-UTR, synonymous and asynonymous positions by comparing complete human, murid (rat and mouse) and artiodactyl mRNAs, for which a suitable number of orthologous sequences was available.

Bibliography

- Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. Curr. Op. Struct. Biol., 8, 333-337.
- Pesole, G. and Liuni, S. (1999) Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. Trends Genet., 15, 379-380.
- Pesole, G., Liuni, S., et al. (1999) UTRdb: a specialized database of 5' and 3' untranslated regions of eukaryotic mRNAs. Nucleic Acids Res, 27, 188-191.
- Pesole, G., Liuni, S., et al. (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene, 205, 95-102.

Acknowledgements

This work was partially financed by EC grant ERB-BIO4-CT96-0030.

Return to Table of Contents

SIMONNEAU

Analysis of new transcribed sequences possibly involved in same functional pathways of neuronal development and identified by differential display

Michel Simonneau, Christophe Mas, Francine Bourgeois, and Fabien Guimiot

Neurogenetique/ INSERM E9935, Hopital Robert Debre, Paris, France

We identified new transcribed sequences using different differential display paradigms in order to select genes expressed in proliferating neuroblasts from mouse embryonic telencephalon. Differentially expressed genes were validated by radioactive relative RT-PCR. Their expression in proliferating telencephalon stem cells was tested by in situ hybridization on mouse embryo sections. From a set of 50 candidates, 58 % are unknown genes, whereas the other 42 % correspond to known genes. For these known genes, both their sequences and their pattern of expression suggest that they are involved in telencephalon stem cell proliferation.

We used similar approaches to isolate and validate differential expression of new genes either up or down regulated from in the developing brain of mouse mutant or knockout mouse embryos. Submicrogram of total RNA is sufficient to isolate and validate differential expression of up to 5 candidates.

As example, we found possible to identify new gene transcripts whose level varies in the brainstem as a function of a physiological variable (here, a ventilatory response), from Mash-1 +/- animals, suggesting that this approach is instrumental to identify new transcripts involved in a functional pathway.

Return to Table of Contents

WINGENDER

TRANSFAC: an integrated system for gene expression regulation

E. Wingender, X. Chen, H. Karas, A. Kel, I. Liebich, V. Matys, T. Meinhardt, M. Pruess, I. Reuter, and F. Schacherer

GBF, Molecular Bioinformatics of Gene Regulation, Braunschweig, Germany

The "post-genomic" era of research will focus largely on so-called "functional genomics", assigning functional properties to all genes of an organism. Part of these efforts is the mass generation of gene expression data due to new experimental technologies such as the application of DNA microarrays. Standardized storage of these data is required to avoid unnecessary duplication of cost-intensive work and to draw an integrated view of the overall gene expression profile of, e. g., the human organism. We therefore started to develop an integrated system for a unifying view on "expression states", defined by organ / cell types, their developmental stages, and conditions (e. g. hormonal induction). The system integrates five databases which we have developed: On the first level, the database TRANSFAC which provides data on transcriptional control, is integrated with a newly developed database on pathological mutations in transcription control components (PathoDB) and with a database on scaffold / matrix attachment regions (S/MARt DB). At a second level, the systems incorporates a data resource on signal transduction components and reactions (TRANSPATH) and another one on (human) organs, cell types, physiological systems and their developmental stages (CYTOMER). As an example, we started to apply our ideas on "mapping" the expression patterns of transcription factors as they are represented in the TRANSFAC database.

Return to Table of Contents

LIN

Identification and characterization of tissue-specific genes using transgenic zebrafish

Shuo Lin

Institute of Molecular Medicine and Genetics, Medical College of Georgia, Augusta, Georgia, USA

It is important that an Expressed Sequence Tag (EST) database produced for a model organism represents mRNA transcripts from all possible tissues. Zebrafish has become an important model organism for genetic studies of vertebrate development and an EST database is currently being generated. Tissue-specific cDNA libraries are usually made from surgically excised organs of larger animals but this technique is not applicable for the construction of tissue- or cell lineage-specific cDNA libraries from zebrafish embryos due to their extremely small size. To overcome this problem, we developed transgenic zebrafish that differentially express the green fluorescent protein (GFP) reporter gene. These fish allow the isolation and purification of lineage-specific embryonic cells using fluorescence activated cell sorting (FACS). To date, we have generated transgenic zebrafish that express GFP in embryonic erythroid, lymphoid, olfactory, pancreas, brain and neuronal progenitor cell lineages. RNA has been isolated from the FACS-purified GFP positive cells and used to construct cDNA libraries for the EST production. Using the ESTs from these cDNA libraries, we have also initiated a large-scale whole mount RNA in situ screen to identify novel genes that are expressed in specific tissues. Finally, we have utilized transgenic zebrafish to characterize function of tissue-specific genes and the study of a novel hematopoietic TNF receptor will be presented.

Return to Table of Contents

EBERWINE

Molecular biology of single cells: Insights into human disease

Jim Eberwine

Department of Pharmacology, University of Pennsylvania Medical School, Philadelphia, Pennsylvania, USA

Pathophysiology associated with disease encompasses dysregulation of many aspects of normal cellular biology. Even in the simplest of diseases where an individual protein may bear mutations that initiate a disease process, usually the coordinated dysregulation of multiple downstream genes and proteins results in the disease phenotype. Methods for analysis and comparison of the expression of multiple genes and proteins from normal and disease tissue are required to establish the set of genes that are coordinately regulated in the disease process. An added complication to this analysis is that often individual diseased cells are mixed in a heterogeneous population of cell types making it difficult to analyze the diseased cells independently of other cells. A further complication in the analysis of human disease is the limited amount of well-characterized human tissue for molecular analysis of disease state. To address these issues, over the last several years we have developed methods that allow the analysis of complex mRNA populations from discreet amounts of tissue, down to single cells and subregions of single cells. Additionally these methods can be applied to fixed human tissue specimens making it possible to analyze the mRNA compliment of pathological tissue samples. This presentation will highlight the methods utilized for analysis of complex mRNA populations from small amounts of tissue using single neuronal dendrites, schizophrenia and Alzheimer's Disease as examples.

Return to Table of Contents

WIEMANN

Sequencing and analysis of full length cDNAs in the course of the German Genome Project

Stefan Wiemann[1], Wilhelm Ansorge, Helmut Blöcker, Helmut Blum, Andreas Düsterhöft, Karl Köhrer, Werner Mewes, Brigitte Obermaier, Rolf Wambutt, and Annemarie Poustka

[1]Molecular Genome Analysis, German Cancer Research Center and the German cDNA Sequencing Consortium, Heidelberg, Germany

A consortium of eight sequencing laboratories and Germany's leading bioinformatics institute has formed in the frame of the German Genome Project. We aim at the sequence analysis of 3,000 to 4,000 complete novel cDNAs, comprising eight megabases of finished sequence. Sequencing started in September 1997 and a progress report of the consortium will be presented. The libraries generated in the course of the grant "Generation of full length cDNAs in the course of the German Genome Project" are the primary source for sequencing. EST sequences of 12,000 independent clones are generated to identify novel genes. The EST sequences are analyzed for the likelihood of the clones to be full length (e.g. by the presence of CpG clusters) in order to obtain a minimal set of full length clones for efficient complete sequence analysis. Clones identified to be full length are sequenced and further analyzed by members of the consortium. The sequences are analyzed for possible function in silico. Functional analysis projects have started using the clones analyzed by the consortium as resource. All clones and data generated in the project are made publicly available via the Resource Centre of the German Genome Project (RZPD).

Return to Table of Contents

VON MELCHNER

Disruption of a gene induced in early mouse development results in severe emphysema and Adenocarcinoma

Irmgard S. Thorey, Anja Sterner-Kock, Jürgen Otte, and Harald von Melchner

Laboratory for Molecular Hematology, University of Frankfurt Medical School Frankfurt, Germany

Several strategies involving gene traps have been developed to identify genes that are regulated during mouse development. However, none specifically selects for mutations into genes that are only transiently ex-pressed. Since these genes are key regulators of many important biological processes and are generally difficult to isolate by cDNA based methodology, we have developed a strategy based on gene trap mutagenesis and site specific recombination (Cre/loxP) to isolate short lived transcripts during early mouse development. Five ES cell clones isolated by this method were passaged to the germ line and mice heterozygous for the transgene were mated to obtain null mutations. One clone (3C7) generated an overt phenotype in F2 homozygous offspring. By the age of 2-3 weeks, mice with a prevalent C57BL6 background (4 backcrossings) develop a rectal prolapse associated with invasive adenocarcinoma, severe pulmonary emphysema and die around 6 months of age. Molecular analysis revealed that the gene trap integration disrupted an exon of a yet unknown gene just several nucleotides downstream of a 3' splice consensus sequence. When hybridized to mRNA from wild type mice, the cloned exon sequences identified a 4.4 kb transcript which was missing from the mRNA of mutant mice. This indicated that the phenotypic abnormalities in mice homozygous for the gene trap integration are caused by a null mutation of a cellular gene. Using a combination of 5' and 3' RACE, we are presently cloning the full length cDNA of the disrupted cellular gene.

Return to Table of Contents

BRAZMA

Mining the yeast genome expression and sequence data

Alvis Brazma

EMBL Outstation -- Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

The rapid increase in the amounts and complexity of the bioinformatics data is creating new challenges of finding ways to transform this data into knowledge, and also opening new possibilities of pure in silico studies of genome functioning. First genomic scale data about gene expression have recently started to become available in addition to complete genome sequences and annotations. Among the first such public data are experiments by DeRisi et al (Science, Vol 278, 1997) regarding the diauxic shift in the complete yeast genome. We have used these data in combination with genome sequence and annotation data from the database MIPS as a case study of data mining in bioinformatics. Among other approaches we used a clustering algorithm based on discretizing the time-series of the expression measurement space to cluster potentially coregulated genes. We extracted the genome sequences upstream from genes for each cluster and used a specifically designed sequence pattern discovery algorithm SPEXS to look for common patterns in each cluster. The algorithm was able to discover sequence patterns that are potential transcription factor binding sites that can be expected to participate in the regulation of diauxic shift. For details see "Predicting Gene Regulatory Elements in Silico on a Genomic Scale" (A.Brazma, I.Jonassen, J.Vilo, E.Ukkonen), Genome Research, Vol. 8, Issue 11, 1202-1215, November 1998. One of the general conclusions from this research has been that using the existing public gene expression data we can mostly "rediscover" and explain previously known facts, while it seems that finer data are needed for real in silico discoveries. Therefore European Bioinformatics Institute is currently looking into the feasibility of establishing a public repository of DNA microarray-based gene expression data. We are interested in discussing the opinions of people involved in using these technologies.

Return to Table of Contents

Beyond the Identification of Transcribed Sequences: Functional and Expression Analysis

9th Annual Workshop, October 28-31, 1999

Co-sponsored by the U.S. Department of Energy

WEISSMAN

Transcript profiling of hematopoietic cell development and activation

Yamaga, S.[1], Yeramilli, S.[5], Lian, Z.[1], Prashar, Y.[5], Lee, H.[1], Berliner, N.[2], Liu, Y-C.[1], Goguen, J.[3], Newburger, P.[4], and Weissman, S.[1]

[1]Yale University School of Medicine, Department of Genetics, New Haven, Connecticut, USA
[2]Yale University School of Medicine, Department of Internal Medicine, New Haven, Connecticut, USA
[3]University of Massachusetts Medical Center, Department of Molecular Genetics and Microbiology, Worcester, Massachusetts, USA
[4]University of Massachusetts Medical Center, Pediatrics, Worcester, MA, USA
[5]Gene Logic Inc., Gaithersburg, Maryland, USA

The hematopoietic system is a favorable mammalian system for studying changes in gene expression during differentiation and activation. At least twelve lineages of cells are derived from a common precursor. Several of these cell types can be isolated in convenient numbers and purity for analysis. Model cellular systems exist for certain stages of in-vitro differentiation. Cells from specific types of hematopoietic neoplasia can also be obtained in pure form without extensive manipulation. We have relied principally, but not solely, on 3' end restriction fragment gel display methods to profile gene expression in several stages of differentiation of myeloid cells and, to a lesser extent, in cells of several other lineages. Neutrophil activation has been studied intensively. Combinatorial use of genes in different lineages is striking. For example, the Lander group (Cell 97:227, 1999) has studied the effects of activation of murine 3T3 cells by signaling through the cytoplasmic domain of the platelet derived growth factor receptor. They found 66 genes that were actively up-regulated within 4 hours after the onset of stimulation. Of these genes, at least 45% are actively up-regulated in mature neutrophils within 2 hours after addition of non-pathogenic bacteria, although the neutrophils are post-mitotic, highly specialized cells. The overall short term response of the neutrophils, however, involved changes in the levels of more than 600 mRNAs, achieved by changes in transcription but also by both up- and down-regulation of the stability of specific mRNAs. Biologic effects that are at least partially interpretable include the development of new autocrine loops, complex anti-apoptotic changes, and changes favoring cell survival and increased membrane trafficking. In the process of this work certain complications became apparent that would probably be shared by all approaches to cDNA profiling that involve an oligo-dT priming step. These affect only a minor subset of the products. In certain cases investigation is needed to determine the possible biologic significance of particular truncated mRNAs.

Return to Table of Contents

Beyond the Identification of Transcribed Sequences: Functional and Expression Analysis

9th Annual Workshop, October 28-31, 1999

Co-sponsored by the U.S. Department of Energy

BARBAZUK

Construction of a whole genome transcribed sequence map for the zebrafish Danio Rerio

W. Brad Barbazuk[1], Ian Korf[1], Frank Li[2], John McPherson[1] and Steve Johnson[2]

[1]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA
[2]Dept of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

The zebrafish is an important vertebrate model for mutational analysis of developmental processes. Development of zebrafish genomic resources serves to expedite gene identification, high resolution physical map construction and the molecular localization and identification of mutated genes. The construction of a whole genome radiation hybrid map provides an invaluable resource for both the initiation of positional cloning projects and the production of physical maps to support genomic sequencing. We are developing a radiation hybrid map of the zebrafish genome map using the LN54 whole genome radiation hybrid panel constructed by Marc Ekker. In collaboration with Niel Hukreide, over 1000 markers derived from genetically mapped CA repeat sequences, zebrafish EST sequences and mapped genes have been typed across the LN54 panel to produce our framework map. We are now attempting to increase the marker density of this map by typing an additional 5000 STSs, and are providing a further 5000 STSs to RH mapping efforts undertaken by other members of the zebrafish community.

This effort is synchronized with the zebrafish EST project being conducted at Genome Sequencing Center in St. Louis. ESTs generated by this project serve as substrates from which STS markers are being developed. We have currently placed over 500 zebrafish ESTs onto this RH map. Comparison of the map positions of zebrafish genes with their human orthologues helps reveal the correspondence between the zebrafish and human genomes. We anticipate that such a comparison will aid gene identification and function in both human and zebrafish, and assist in reconstructing the evolutionary history of the vertebrate genome.

Return to Table of Contents

HICKS

Development of an embryonic stem cell library of defined mutations

Geoffrey G. Hicks

Manitoba Institute of Cell Biology Centre for Mammalian Functional Genomics, Winnipeg, Canada

Most mammalian genes will soon be characterized as cDNA sequences with little information as to their function. To utilize this sequence information for large-scale functional studies, we developed a process of tagged-sequence mutagenesis to disrupt genes expressed in mouse embryo-derived stem (ES) cells and to characterize each mutation by direct DNA sequencing. Comparison of these sequence tags (PSTs) with the existing databases identifies disruptions of known genes or genes which may be related by homology or functional domains. The process will generate large numbers of insertion mutations that will be available for transmission into the mouse germline where mammalian gene function can be directly analyzed in vivo.

We have recently reported the results from over 400 such mutations (Nature Genetics 16:338). Analysis of this group has substantiated two important assumptions about this sequenced-based approach: 1) The number of target genes is large and approaches the total number of expressed genes; and 2) The PSTs provide enough sequence information to identify disruptions in known genes when compared to the rapidly expanding nucleotide databases. In addition, the analysis has revealed new insights into the mechanisms of gene entrapment and new vector designs.

The ability to induce, characterize and maintain mutations in ES cells circumvents many limitations associated with conventional mammalian genetics, and will greatly increase the number of mutant alleles (typically loss of function mutations) by which gene functions can be studied in mice and in cell lines derived from such mice. The PSTs allow sequence-based screening of the library of mutations, thereby bridging the gap between gene sequence and mammalian gene function. The process will facilitate a functional analysis of a mammalian genome and will provide animal models for human genetic diseases.

Return to Table of Contents

WORLEY

Search services to improve the identification of expressed sequences and their functions

J. Bouck[1], M. McLeod[1], T. McNeill[2], G. Weinstock[1,3], R. A. Gibbs[1], and K. C. Worley[1]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine

[2]Department of Biochemistry, University of Houston

[3]Department of Microbiology and Molecular Genetics, University of Texas - Houston Medical School, Houston, Texas, USA

Two of our approaches taken to identify and characterize transcribed sequences will be presented here. Existing transcript data that is publicly available is often not accessible due to the form that it is stored in. One such set of data is the inconsistently annotated collection of cDNA sequences in the GenBank database. The Human Transcript Database isolates these sequences and improves the access to them. Identifying the putative function of an expressed sequence may not be a straightforward process, regardless of the degree of sequence similarity identified in a sequence similarity search. Our second tool, BEAUTY, provides a collection of information about functional annotation in combination with sequence similarity to aid in functional assessment.

Improving access to cDNA sequences by isolating existing transcribed sequences from GenBank. Transcript sequences that are full-length mRNAs or cDNAs are available in the GenBank NR DNA database, but these sequences may not be readily accessible in that database due to the volume of other sequences. The Human Transcript Database (HTDB) is a curated collection of expressed sequences isolated from GenBank. The HTDB is a valuable resource for studying human genes and targeting cDNA sequencing projects. This collection can be searched using keywords or sequence similarity from the web site at http://hgsc.bcm.tmc.edu/HTDB.

Identifying the function of expressed sequences is the focus of the BEAUTY programs. BEAUTY presents information from the Annotated Domains database, a collection of protein sequence annotations from Prosite, Blocks, Prints, Entrez, and Pfam, in combination with reports of local BLAST similarities. This combination provides a better assessment of functional domain conservation than BLAST searches alone. BEAUTY searches are available for protein and DNA queries and searches of protein databases from the BCM Search Launcher http://searchlauncher.bcm.tmc.edu/.

Return to Table of Contents

FICKETT

Finding gene boundaries on large contigs

James Fickett

SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania, USA

Sequence similarity is the most reliable tool available for gene structure prediction, but it is not infallible and, more importantly, is only available when a close homolog to the new gene is known. De novo gene prediction will play an especially important role in the next year, as the newly sequenced genome is scrutinized for medically important genes. The greatest limitation of the current generation of gene finders is in locating gene boundaries. We will discuss the extent of the problem and some progress towards a solution.

Return to Table of Contents

JENNINGS

Dissecting transcriptional circuitry and mechanisms in yeast

Ezra Jennings

Whitehead Institute, Cambridge, Massachusetts, USA

Genome-wide expression analysis is being used to obtain clues to the roles played by transcriptional regulators, components of the transcription initiation apparatus, histones and chromatin modifying enzymes in gene regulation. Previous genetic and biochemical studies identified much of the transcriptional machinery of eukaryotic cells and provided mechanistic insights into the functions of various components at specific promoters. However, knowledge of the contributions of regulators and the transcriptional machinery to the complete transcriptional circuitry is lacking. Our data describes how expression of the genome depends on over 50 components critical to transcriptional regulation in yeast cells. This information is providing the foundation for understanding the molecular mechanisms involved in genome-wide expression, and new insights into transcriptional regulation will be described.

Return to Table of Contents

WELSH

The RAP-array approach to cDNA array hybridization

John Welsh

Sidney Kimmel Cancer Center, San Diego, California, USA

Transcript abundance in biological samples vary typically from many thousands to fewer than one copy per cell. An important practical concern with current microarray technology is the reliable detection of signals from rare transcripts. To address this problem, optimizations in dyes, detection systems, hybridization conditions, attachment strategies, and other variables are on-going in many laboratories. An alternative solution involves methods for the construction of labeled cDNAs with distorted ratios of individual sequences such that their representation is decoupled from their original abundances. In one such approach RNA arbitrarily primed-PCR fingerprinting or Differential Display is used to create probes with reproducibly altered abundances. Due to the selectivity of arbitrary priming, sequences in the low abundance-high complexity class are more highly represented in the resulting cDNA probe relative to transcripts in the higher abundance classes. Within a single sample, sequence abundances are highly distorted relative to the corresponding abundances in the original RNA population. However, this distortion is reproducible on a sequence-by-sequence basis, so that original ratios of transcript abundances are preserved in comparisons between two or more RNA sources.

In our expression profiling studies, we have encountered several interesting new phenomena. A new EGF and TGF-beta regulated gene, VAV3, and a natural 5'-trunkated variant, VAV3.1, will be discussed. Also, several phenomena have emerged from expression profiling of combinatoric treatment experiments, including a vector/vector complement pattern, a pattern characteristic of translation inhibition, and a particularly puzzling pattern that implies that UVC irradiation overrides the well-documented repressive effect that cycloheximide has on RNA turnover.

Return to Table of Contents

MURAL

Extracting meaningful information from draft sequence

R. J. Mural, F. W. Larimer, M. B. Shah and E. C. Uberbacher

Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
(http://compbio.ornl.gov)

The plan to complete a 90% draft of the sequence of the human genome by next spring poses special problems for the annotation process. It is clear that a 3 to 5 X coverage of genomic DNA can yield large amounts of biologically useful information if the appropriate analysis methods can be applied. There are a number of features that can be located and annotated in draft-sequence which are useful for further analysis, these include: STS's that allow some clones to be assigned to chromosomes or chromosomal locations, BAC end-sequences (STC's) that will help to identify neighboring clones and help to build framework maps, and EST's that can provide gene identification information and in some cases map information. As catalogs of full length cDNAs become available they will be even more useful than EST's in helping to define biological content of draft sequence. These features can be annotated by standard similarity methods given sufficient computational resources.

Using various gene identification programs, particularly those that incorporate similarity data such as Grail-Exp, can provide another level to the analysis of draft data. Results from such analysis allows not only gene identification but it can also provide some internal ordering information for the sequence contigs that make up the clone being analyzed. Also recall that essentially all of the genes that can be found in finished sequence can be identified in draft sequence at about 3X coverage.

We have begun to modify the analysis pipeline that we have developed for finished sequence, the results of which can be viewed in the Genome Channel and the Genome Catalog, to provide analysis of draft data. The initial annotation of draft sequence is a catalog of the clone contents (STS's, STC's, genes models predicted by Grail-Exp and Genscan, as will as Blast searches of their translations against the NR protein database). Further analysis of this information will help to define relationships among draft clones and will allow ordering, within and between clones. To date we have analyzed over 3500 draft clones from human chromosomes 5,16 and 19 and we are building the data structures to handle other draft data.

Supported by the Office of Biological and Environmental Research, US DoE, contract DE-Ac05-84OR21400 with Lockheed Martin Energy Systems, Inc.

Return to Table of Contents

PAN

Approaches towards global profiling of cDNA mutations

Pan, X. and Weissman, S.

Yale University School of Medicine, Department of Genetics, New Haven, Connecticut, USA

The ability to detect a large fraction of sequence variations between two complex genomes or pools of genomic DNA from two sources would have a wide range of applications. The approach we are taking towards this is to initially focus on comparing variations in cDNA sequence between two sources. Several types of chemical or enzymatic approaches have been used to detect sequence variation. Comparison of cDNA sequences, in principle, could be particularly valuable for identifying mutations associated with specific malignancies, or for more rapidly identifying mutations in simple Mendelian disorders including severe sporadic dominant lesions that could not be mapped by genetic means. We find that certain DNA-glycosylases are quite promising for our purposes. The glycosylases recognize specific types of lesions and have the great advantage that they bind relatively tightly to the abasic DNA that they produce enzymatically. Also, in at least some cases, borohydride reduction can be used to generate covalent links between the enzyme and its product. These features are useful to specifically separate mismatch and perfect match DNA duplexes from the same reaction. Glycosylases exist that would, in combination, detect a large fraction of single base sequence variation; although in certain cases it may be necessary or preferable to use specific nucleases.

Gel display of mismatch fragments in principle is an attractive approach because prior knowledge of the existence of sequence variation is not needed. Various array hybridization procedures would also require considerable development for this purpose because mutations might occur anywhere within a cDNA and also it would be desirable to the extent that it is possible to distinguish new mutations from common polymorphisms in the same mRNA. To achieve the above goal, several technical issues must be dealt with. For accurate comparison, it would be desirable to be able to form heteroduplexes in a single denaturation-renaturation step, then use protocols that permit selective examination of either homoduplexes re-formed from one of the two input cDNAs, or only heteroduplexes in which one strand is derived from each of the two input pools. For display purposes it would be necessary to divide cDNA fragments into non-overlapping, exhaustive subsets, such that each subset would contain a limited number of fragments in the display range. Display conditions should require only stringent PCR steps, and the conditions for PCR should not strongly bias for sequences of a specific range of GC content. Annealing conditions and fragment complexity should be optimized for efficient but stringent annealing of low abundance sequences. We will describe our experimental designs and progress along the above lines.

Return to Table of Contents

NELSON

Functional dissection of the RANTES promoter: Insights into mechanisms of tissue specific regulation of transcription

Peter J. Nelson, Sabine Böhlk, Sabine Fessele and Thomas Werner

Medizinische Poliklinik, LMU Munich, Germany,
Institute of Mammalian Genetics, GSF-National Research Center for Environment and Health, Neuherberg, Germany and
Genomatix Software GmbH, Munich, Germany

The chemokine RANTES is produced by a variety of cell types in response to diverse stimuli. The molecular mechanisms involved in transcriptional control of RANTES can vary significantly between the various cells that express the gene and the specific activating stimuli used. For example, T cells strongly induce RANTES "late" (i.e. 3 to 7 days) after activation through their T cell receptor. Monocytes do not upregulate RANTES in response to TNF-alpha, IL-1beta or gamma-IFN, but quickly and transiently induce RANTES following stimulation with lipopolysaccaride (LPS) (maximal expression by 6 to 9 hours. By contrast, fibroblasts and astrocytes produce RANTES in response to TNF-alpha, IL-1 beta and gamma-IFN with the initial expression seen by 6 hours and maximal expression by 48 hours. The promoter region that regulates these diverse modes of expression may act as an enhancesome. It is compact, highly conserved over evolution, and can be efficiently studied using computer models of transcriptional control as well as by conventional methods for functional promoter analysis. The results of an analysis of transcriptional regulation of RANTES in T cells and monocytes will be discussed.

Return to Table of Contents

BORSANI

A tale of two diseases

Giuseppe Borsani

Tigem - Telethon Institute of Genetics and Medicine, Milano - Italy

EST database mining led to the identification of four novel members of a novel amino acid transporter family (Solute Carrier Family 7, SLC7). The chromosomal assignment of the genes revealed that they represent positional candidates for lysinuric protein intolerance (LPI) and cystinuria type III (CSNU3) two human inherited disorders due to impaired amino acid transport.

LPI is an autosomal recessive multisystem disorder, caused by a defective cationic amino acid (CAA) transport at the basolateral membrane of epithelial cells in kidney and intestine. After weaning, LPI patients present poor feeding, vomiting and failure to thrive. A severe pulmonary complication and episodes of metabolic imbalance may lead to death. Two of the new transcripts we have identified, SLC7A7 and SLC7A8 map within the LPI critical region on chromosome 14q11.2, and are located on the same YAC. Analysis of both transcripts identified mutations in the SLC7A7 gene only.

CSNU3 is a heritable disorder of amino acid transport, transmitted as an autosomal recessive trait, with an overall prevalence of approximately 1 in 7,000. It is due to the defective transport of cystine and dibasic amino acids through the epithelial cells of the renal tubule and intestinal tract. Cystine has a low solubility, and its precipitation results in the formation of calculi in the urinary tract, which leads to obstruction, infections, and ultimately renal insufficiency. Again, two transcripts belonging to the SLC7 gene family were found to map within a small critical region on chromosome 19q13.1. Sequence analysis in CSNU3 revealed that the disease is due to mutations in the SLC7A9 gene only.

Preliminary data provide evidences of additional members of this gene family and suggest the existence of two homologous genes clusters on chromosome 14 and 19.

Return to Table of Contents

ARONOW

Stepwise chromatin restructuring and unexpected rules for interaction of distributed cis-elements that form a locus control region in vivo

Catherine Ley-Ebert, Carolyn Florio, John Maier, Chris Cost and Bruce Aronow

Children's Hospital Research Foundation and the University of Cincinnati, Cincinnati, Ohio, USA

We have performed a mutational analysis of the human adenosine deaminase (ADA) gene thymic locus control region in transgenic mice using CAT reporter genes. We examined in detail gene expression, chromatin structure based on low and high resolution MNase and DNAse hypersensitivity, restriction enzyme accessibility, and histone acetylation state analyses using chromatin immune precipitations. Mutations and perturbations of LCR cis-elements led to variably compromised LCR function and the formation of several common distinguishable chromatin structures.

Our results suggest that the ADA thymic LCR is formed by the interaction of distributed cis-elements that consist of a 200 97bp enhancer region which is bilaterally flanked by two 1 kb regions that contain specific cis-elements distinct from those of the enhancer. The flanking sequences facilitate the stepwise formation of an intricate chromatin architecture at the enhancer core. The central enhancer elements are partially redundant with each other and, at a stage prior to hypersensitive site formation, are required for the nucleosomal array dephasing and histone acetylation of an extended nucleosomal array associated with the inactive LCR. Unexpected rules govern the interactions of the flanking facilitative cis-elements with the enhancer core and suggest that a supranucleosomal organization, beyond the phased nucleosomes, is critical for proper LCR function in developmental gene activation.

Return to Table of Contents

DE BEKKE

Experimental approach towards identification of small non-messenger RNAs in the genome of Caenorhabditis elegans

Anja op de Bekke[1], Alexander Huttenhofer[1], Martin Kiefmann[1], John O'Brien[2], Hans Lehrach[2] and Jurgen Brosius[1]

[1]Institute of Experimental Pathology/Molecular Neurobiology, ZMBE, University of Munster, Munster, Germany
[2]Resource Centre, Max-Planck-Institute for Molecular Genetics, Berlin, Germany

Genome projects allow identification of complete sets of genes in a given organism. This is a prerequisite for completely understanding its biology including gene expression, function of its products and evolutionary relationships. Current approaches primarily focus on protein coding genes. Those expressing transcripts that contain short (< 300 nt) open reading frames (ORFs) or non-messenger RNAs are currently difficult to identify with biocomputational methods only. Non-messenger RNAs play a wide range of roles in the cell. It is expected that eukaryotic cells contain a significant number of unknown non-messenger RNAs with interesting functions. In the recently completed sequence of the C. elegans genome only genes encoding ribosomal RNAs, tRNAs, five small nuclear RNA, two snoRNAs, 7SL RNA, SL1 and SL2 splice leader RNA, Y RNA and lin-4 RNA were annotated. Many RNA species that are expected to be present in the C. elegans genome were not detected. Therefore, we took an experimental approach akin to the EST projects in order to identify the majority of expressed RNA sequences (ERNs) transcribed from the genome of C. elegans.

Return to Table of Contents

HEUERMANN

Stable expression of epitope-tagged proteins in mammalian cells

Kenneth E. Heuermann and Bill L. Brizzard

SIGMA Chemical Company, St. Louis, Missouri, USA

Analysis of gene function often requires stable expression of the recombinant gene in a mammalian cell line. This can be facilitated by incorporating an epitope tag, such as the FLAG® peptide (AspTyrLysAspAspAspAspLys), commonly used for the isolation, purification, and detection of recombinant proteins expressed in E.coli. pFLAG-CMV-3 and pFLAG-CMV-4 vectors stably express secreted or intracellular N-terminal FLAG-fusion proteins, respectively, in mammalian cell lines. Initially, COS7 cells transiently transfected with pFLAG-CMV-3-BAP or pFLAG-CMV-4-BAP were shown to express bacterial alkaline phosphatase (E. coli phoA) by immunostaining using the M2 anti-FLAG monoclonal antibody. Western analyses of cell extracts and media confirmed these results. COS7 and CHO-K1 cells were then transfected with pFLAG-CMV-3, pFLAG-CMV-3-BAP, pFLAG-CMV-4, or pFLAG-CMV-4-BAP, to obtain stably transformed cell lines. Transient expression of BAP by cells transfected with pFLAG-CMV-3-BAP or pFLAG-CMV4-BAP, but not pFLAG-CMV-3 or pFLAG-CMV-4, was demonstrated by immunostaining in parallel test plates. Transfected cells were then selected by treatment with 500 ug/ml G418 sulfate. At 20 days post-transfection after three changes of medium, surviving COS7 and CHO cells transfected with pFLAG-CMV-3-BAP or pFLAG-CMV-4-BAP continued to express BAP, as shown by detection of FLAG-tagged BAP by Western analysis. This result indicates stable integration of the neomycin resistance gene and FLAG-tagged BAP construct.

Return to Table of Contents

KAMP

Structure and function of the spermatogenesis genes located in AZFa, a region of the human Y chromosome deleted in men with complete germ cell aplasia

Kamp, Christine[1], Kirsch, S[1], Hirschmann, P[1], Ditton, HJ[1], Brede, G[2], Tyler-Smith, C[2], Rappold, GA[1], and Vogt, PH[1]

[1]Institute of Human Genetics, Heidelberg, Germany
[2]Department of Biochemistry, Oxford, United Kingdom

In mammalian species X and Y, the so-called sex chromosomes, evolved from an extant pair of ordinary autosomes. One of these autosomes was elected to become the Y chromosome most likely because of a male specific selection of SA (Sexually Antagonistic) alleles, i.e., favoured in one sex, but disfavoured in the other. This resulted in a continuous reduction of crossing-over events and an accumulation of Y-specific DNA loci functional for male sex determination and male germ cell development. We focussed our research on a Y region in proximal Yq11 which was mapped by STS content analyses to be essential for male germ cell proliferation (AZFa region; 1). Men with deletions of AZFa suffer from a complete aplasia of germ cells in their testis tubules.

The molecular extension of the AZFa region is not known. We therefore established first a physical restriction map along the AZFa region with the aid of a complete YAC contig and estimated a molecular AZFa extension of at least one 1 megabase (Keil, R et al. in prep.). To analyse the gene content of AZFa we performed systematically organised exon trapping experiments with a series of PAC clones mapped in a contig by Alu-vector PCR, cross hybridizations of Y-specific end fragments and overlapping STS contents. 11 PAC clones were sufficient to cover the complete AZFa region. Multiple exons were isolated in each exon trapping experiment. Sequence analyses and homology searches in the EST and genomic databases identified some of them as exons of the DFFRY gene and DBY gene isolated recently as complete cDNA clones by Lahn and Page (2). Some of them hit other ESTs expressed in multiple tissues, some of them hit no data bank entry. This suggests that the AZFa region contains multiple Y genes expressed not only in testis tissue. This view got support by subsequent analysis of each novel exon clone on RNA-dot-blots and their identification in GeneFinder cDNA pools (Resource Center of German Human Genome project).

References

1. Vogt, PH et al. (1996) Human Y chromosome azoospermia factors mapped to different subregions in Yq11, Hum. Mol. Gen. 5: 933-943.
2. Lahn, BT and Page, DC (1997) Functional Coherence of the Human Y Chromosome, Science 278: 675-680.

Return to Table of Contents

BULL

Analysis of gene expression data generated by oligonucleotide fingerprinting

Christof Bull, John O'Brien, Uwe Radelof, Ralf Herwig, Steffen Hennig, Axel Nagel and Hans Lehrach

Abteilung Lehrach, Max-Planck-Institut fuer Molekulare Genetik, Berlin, Germany

Oligonucleotide fingerprinting (OFP) is a powerful method for genome-wide expression analysis and gene finding. It is based on the analysis of arrayed cDNA libraries by sequential hybridisation of 200 oligonucleotides 10 bp in length. Clones are grouped into clusters according to their hybridisation fingerprints. The number and the size of the clusters provide information about the spectrum of expressed genes and their relative expression levels respectively, whereas the fingerprint itself is used for database matching of the cDNA clones. We can therefore identify the corresponding gene of a cDNA clone and get information about expression rates at the same time. We have performed OFP on cDNA libraries from human monocytes and dendritc cells with 100,000 clones each. The clones were grouped into 11,897 clusters plus 25,582 singletons (clusters with just one member). This would correspond to a variety of 37,479 different genes that are found to be expressed in either of the cell types. However, due to technical reasons we observed a 1,57 fold overestimation of expressed genes in previous experiments so that we would estimate the real number to be around 24,000 expressed genes. Of the genes that are differentially expressed between monocytes and dendritic cells, we selected 260 genes that are of particular interest to us for further studies. We will re-array these and other selected clones from the libraries to a non-redundant set. This clone set will be further evaluated in expression studies using cDNA arrays and complex probes derived from hematopoetic cell types including monocytes and dendritic cells from various differentiation stages. There were also approximately 1,000 potentially new genes which are currently being tag-sequenced at the MPI-MG. The massive fingerprinting and sequencing data that we have obtained are analysed by highly automated computer tools. The sequence data are compared to the following databases: dbEST, GenEMBL, human UniGene, SWISSPROT and our cDNA sequence databases from sea urchin, amphioxus and zebrafish. Following the database searches (BLAST) a series of further analysis steps is performed, including filtering of blast output files, clustering of related matches and tabulating the results in web-pages, which allow easy access to the analysis details. We will integrate all our data and think that especially the comparision of gene expression patterns from homologue genes in model organisms will be very useful to determine the function of new human genes.

Return to Table of Contents

SLAVOV

Analysis of novel genes from human chromosome 21: determination and characterization of complete protein sequences and examples of overlapping genes

Dobromir Slavov, Roger Lucas, Andrew Fortna and Katheleen Gardiner

Eleanor Roosevelt Institute
Denver, Colorado, USA

We are currently using both computer based and experimental approaches to identify and characterize novel genes within the genomic sequence of human chromosome 21. Exon prediction, EST database matches and CpG island identification together are highly efficient at demonstrating the presence of a gene within a segment of DNA. Determining the complete structure of a novel gene and verifying its expression, however, is often more challenging, in particular for genes lacking any significant protein similarities. Problems are compounded when low or restricted expression precludes obtaining information from Northerns or cDNA libraries.

Our preliminary gene identification is based on consistent exon prediction by at least Genscan and Grail programs and/or EST matches that show evidence of exon splicing. CpG island identification, information from RT-PCR and RACE experiments are then added to these data. By these means, we have deduced complete protein sequences for seven novel genes. Protein sizes range from ~250 amino acids to >1500 amino acids. Five proteins contain no discernible protein homologies or motifs; two show only distant homologies that provide no functional clues. None is positive by Northern analysis. Protein sequences have been examined for biochemical and structural features such as amino acid content, hydrophobicity, polarity, and presence of beta sheets and alpha helices. Rarely have such data shown unusual features.

In two other cases, we have evidence of potentially overlapping genes. In both cases, the gene on one strand is represented by consistent exon prediction but no ESTs, and the gene on the opposite strand is represented by one or more ESTs but by no convincing exon predictions. Consensus splice sites are found only on the appropriate strand in all cases. In one case, the exon prediction gene is located within an intron of the EST gene; in the other case, EST exons interdigitate with consistent exon prediction. Expression levels in all cases are low and/or restricted. Analysis of such gene models would be facilitated by corresponding mouse genomic sequence, by adding more coding sequence data to EST sequences, and by more comprehensive information on exon prediction false positive rates.

Return to Table of Contents

SPENGLER

ASDB: Novel database of alternatively spliced genes

I. Dubchak[1], M. S. Gelfand[2], I. Dralyuk[1], M. Zorn[1], S. Spengler[1]

[1]National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[2]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia

Alternative splicing is an important regulatory mechanism in higher eukaryotes. By recent estimates, at least 30% of human genes are spliced alternatively (1). Alternative splicing plays a major role in sex determination in Drosophila, antibody response in humans and other tissue or developmental stage specific processes. The database of alternatively spliced genes can be of potential use for molecular biologists studying splicing, developmental biologists, geneticists, and cell biologists. We have created a public Alternative Splicing Database (ASDB) (2) for the biological community as a repository of data on alternatively spliced genes. ASDB is currently available at the URL http://hazelton.lbl.gov/~teplitski/alt/. The administrator of the database can be contacted by Email: asdb@lbl.gov.

Our original set of 1663 proteins was generated by selecting all SwissProt entries containing the words "alternative splicing". Clusters of proteins that could arise by alternative splicing of the same gene were created by a string comparison procedure. Two proteins from the same species were considered belonging to a cluster if they have common fragments not shorter than 20 amino acids. Each cluster is represented in the database by the multiple global alignment of its members, allowing for easy identification of regions produced by alternative splicing. The database contains 241 clusters with more than one member.

The database can be searched using Medline, SwissProt, and GenBank identifiers and accession numbers. Standard context search can be performed over SwissProt keyword, description, taxonomy, and comment fields and feature tables. ASDB contains internal links between entries and/or clusters, as well as external links to Medline, GenBank and SwissProt entries.

The next step in ASDB development will involve building a nucleotide division of ASDB by incorporating DNA data from GenBank and other sources, classification of the main types of alternative splicing, and adding data on aberrant splicing and splicing mutations.

References

1. Mironov, A.A. and Gelfand, M.S. Proc. 1st Int. Conf. on Bioinformatics of Genome Regulation, 1998. v. 2, p. 249.
2. Gelfand, M., Dubchak, I., Dralyuk, I., M.Zorn. Nucl.Acids.Res. 1999, 27, 301-302.

Return to Table of Contents

KIKUNO

Functional and molecular evolutional analysis of predicted gene products of human long cDNAs

Reiko Kikuno, Takahiro Nagase, Ken-Ichi Ishikawa, Mikita Suyama, Mina Waki, Makoto Hirosawa, Nobuo Nomura, and Osamu Ohara

Kazusa DNA Research Institute
Kisarazu, Chiba, Japan

Our cDNA sequencing project has focused on human cDNAs that have a potential to code for large proteins expressed in brain and so far published more than 1200 new cDNA sequences. We have assigned a KIAA number to each sequence as a name of the gene. Among them, the gene products of 1117 cDNAs were predicted and studied in detail from functional and molecular evolutional viewpoints by computer analysis at amino acid sequence level. Database search of the protein sequences revealed that functions of 682 proteins (61.1%) were classified into 7 classes such as Structure/Motility (8.7%), Metabolism (3.2%), Cell division (1.1%), Signal/Communication (25.3 %), Nucleic acids managing (16.7%), Protein managing (4.1%), and others (1.9%), while the functions of 139 protein could not be predicted although the sequences indicated significant similarity to other sequences and/or motifs in public databases. The number of the protein sequences that showed no significant similarity neither to known sequences nor to motifs was 296 (26.5%).

The detail results of our analyses for each KIAA gene can be browsed in our HUGE protein database, which is accessible via World Wide Web at http://www.kazusa.or.jp/huge. In addition, the HUGE database contains experimental results such as expression profiling and RH-mapping together with their experimental conditions. Functions of keyword search and homology search were also prepared to retrieve KIAA entries of user's interest.

To examine the evolutionary origin of KIAA genes, we compared the protein sequences with those deduced from the genomes of yeast (S. cerevisiae) and nematode (C. elegans). It was shown that 1) only 3% and 10% of KIAA genes had homologous genes (i.e., homologous along the entire coding regions) in yeast and nematode, respectively; 2) 35% and 14 % of KIAA gene products were lack of any homologous regions to gene products encoded by the genomes of yeast and nematode, respectively. The remaining KIAA gene products were found to have some homologous region(s) in these lower eukaryotes. Possible evolutionary processes of these genes will be also discussed.

Return to Table of Contents

BORISSEVITCH

Large scale cloning, sequencing and expression profiling of genes expressed in transcription factor CREM dependant manner during mouse spermatogenesis

Igor Borissevitch, Tim Beissbardth, Andreas Hoerlein and Guenter Schuetz

German Cancer Research Center, Heidelberg, Germany

CREM belongs to CREB (cAMP Responsive Element Binding protein) family of transcription factors. CREMtau, an activator splice isoform of the CREM protein is highly expressed after meiosis in round spermatides at stages from 1 to 5. According to the high level and time restriction of expression, CREM seems to be the major trigger of the expression at the late stages of spermatogenesis.

This work represents large scale cloning, sequencing and expression profiling of RNA messages expressed in a CREM dependant manner. We have used gene targeting to selectively eliminate the transcription factor CREM (Nature, 1996, vol.380, pp.162-165). CREM -/- mice display a normal phenotype but males are sterile due to an arrest of spermatogenesis. Spermatide development is blocked at stages 2-5 and results in the absence of cells of all further stages including spermatozoa. By use of subtractive suppression hybridisation we have cloned messages expressed in wild type but not in a CREM -/- mutant mouse. 12000 clones were analysed by sequencing and hybridisation. Redundancy of this library has been reduced by high density filter hybridisation with the most abundant clones. 950 clusters of sequences were obtained. They represent 79 known mouse genes, 81 homologs to known genes (mostly rat and human) 170 different mouse ESTs, 91 ESTs from other spices, 21 homologs to genomic sequences, 139 novel sequences.

These data compile new extensive information about gene expression during spermatogenesis. In addition, these data provide selection of genes to search for direct CREM target genes (for instance, known CREM target gene ACE presents in our library). Based on our results application may be found for diagnostic and therapeutic intervention in infertile patients with spermatogenetic abnormalities.

Return to Table of Contents

MONTPETIT

Genomic comparative analysis of the Fugu rubripes homologue of ETV6, a gene frequently rearranged in human leukemias

Alexandre Montpetit and Daniel Sinnett

Division of Hemato-Oncology, Charles-Bruneau Cancer Center, Sainte-Justine Hospital, Montreal, Canada and Department of Biochemistry, University of Montreal, Montreal, Canada

Acute lymphoblastic leukemia (ALL) is the most frequent pediatric cancer. Little is known on the molecular pathogenesis of this disease. Recently, loss of heterozygosity (LOH) studies have reported deletions of the chromosome 12p12.3 in 15-47% of pre-B ALL patients. This region was also found deleted in several other hematological malignancies as well as a variety of solid tumors suggesting the presence of a putative tumor suppressor gene. The chromosomal region containing this tumor suppressor locus was restricted to a ~750kb interval that includes the gene ETV6, encoding an ets-like transcription factor required for hematopoiesis in bone marrow. Accumulating evidences suggest that ETV6 is not the tumor suppressor gene targeted by the deletions. However this gene is frequently found translocated with different partners leading to the creation of chimeric products in hematological disorders suggesting that this region might be intrinsically unstable. The search for genes in large mammalian genomic region is the limiting step in positional cloning. We propose that this task could be facilitated by genomic comparative studies of the region of interest in the compact genome of Fugu rubripes that is 7.5 times smaller than that of human but contains a similar repertoire of genes. Here we report the characterization of the Fugu homologue of ETV6 (frETV6) that constitutes the initial step toward the comparative mapping of the human chromosome 12p12.3 tumor suppressor locus. Two Fugu genomic libraries were screened with a human ETV6 cDNA leading to the identification of 6 genomic clones that are part of a contig covering more than 175 kb. The 8 exons of frETV6 were identified and located within a 15kb region. Compared to the human homologue, which is spread over 350 kb, this represents a substantial 23-fold compaction. At the protein level, we observed an overall 57% sequence identity between both species. In particular, the conserved ETS and PNT domains showed 95% and 69% amino acid identity, respectively. A phylogenetic analysis conducted on ETV6 sequences from human, mouse, chicken, Fugu and zebrafish revealed a close relationship between the two teleost fishes. The comparative mapping of the suppressor locus will be extended to the region flanking frETV6 in order to identify candidate genes that could be involved in ALL and/or other cancers as well as to get clues about the cause of the chromosomal instability affecting this region.

Return to Table of Contents

Beyond the Identification of Transcribed Sequences: Functional and Expression Analysis

9th Annual Workshop, October 28-31, 1999

Co-sponsored by the U.S. Department of Energy

CLIFTON

Comparison of completed genomes to sample sequences of related genomes

Sandra W. Clifton[(1)], Michael McClelland[(2)], Webb Miller[(3)], William R. Pearson[(4)], Aaron J. Mackey[(4)], and Richard K. Wilson[(1)]

[1]Genome Sequencing Center, Wash Univ School of Medicine, St. Louis, MO, USA
[2]SKCC, San Diego, California, USA
[3]Dept Computer Science & Engineering, Penn State Univ, University Park, PA, USA
[4]Dept of Biochemistry, University of Virginia, Charlottesville, Virginia, USA

Complete genomes provide a useful framework for organizing and analyzing partial sequences from related genomes. A sample consisting of 2X or 3X genome equivalents gives coverage of over 90% of the genome in which more than 99% of all ORFs over 500 bases in length should be represented by a fragment of at least 100 bases. Information on the presence of shared ORFs and partial identities of unique ORFs can be obtained at a fraction of the cost of complete sequencing.

To determine the utility of sample sequences, we have collected data from two Enterobacteria, Salmonella paratyphi A (SPA), and a clinical isolate of Klebsiella pneumoniae (KPN). These strains are of interest as human pathogens and for understanding enterobacterial evolution. SPA is very closely related to the completed Salmonella genomes, whereas, KPN is a sister clade of Salmonella and Escherichia. Over 10 million bases of raw sequence, representing between 2X and 3X genome equivalents, were collected from both SPA and KPN, which melded to 4,384 kb and 5,084 kb, respectively.

For Enterobacteriaceae, the E. coli K12 genome (ECO) is completely sequenced [U.Wisconsin] and the genomes of Yersinia pestis (YPE), Salmonella typhi (STY) [Sanger Center] and S. typhimurium LT2 (STM)[Wash. U., http://genome.wustl.edu/gsc/bacterial/salmonella.shtml] are soon to be completed. The ECO sequence has been aligned to the available sequence from each of STM, STY, SPA, KPN, YPE, and Vibrio cholera. These alignments can be viewed as a "percent identity plot" or PIP, in which percent identities of ungapped matches are shown in the Y-axis for each pairwise comparison. Deletions in the sampled genomes and the sites of rearrangements and of significant insertions are visualized in color. The alignments can be queried with any named ECO gene and the corresponding region is visualized in multiple genomes, simultaneously. Matching sequences in each aligned genome, associated with the reference gene and flanking regions, are automatically made available.

Unique portions of the complete and sampled genomes were identified with the FASTX and TFASTX programs. To search for unique regions and potential rearrangements in the sampled genomes, each sampled sequence is compared to the E. coli proteome using FASTX and the complete E. coli proteome is compared against partially sequenced genome databases using the TFASTX program. We present lists of (a) all ORFs found in ECO for which orthologues are apparently absent in the STM, SPA, or KPN samples,(b) sequences over 400 bp in length that are found in one or more of STM, SPA or KPN, but are absent in ECO. The best homologues of these "unique" regions are determined from other sequence databases, including incomplete genomes deposited at NCBI.

Return to Table of Contents

SCHMIDT-KITTLER

Genomic characterisation of early disseminated tumor cells isolated from bone marrow of breast cancer patients

Oleg Schmidt-Kittler[1], Julian Schardt[1], Günter Schlimok[2], Gert Riethmüller[1] and Christoph Klein[1]

[1]Institut für Immunologie der LMU, München, Germany
[2]Zentralklinikum Augsburg

Because genomic changes constantly accumulate during tumor progression the link between structural changes within the genome and the malignant behaviour of a cell is hard to establish at a later stage of the disease. To identify genes involved in processes of early systemic disease, such as dissemination and ectopic survival, we analyzed single disseminated tumor cells from the bone marrow of breast cancer patients. The genomic aberrations of these cells should be the result of selection pressures.

Disseminating tumor cells can be detected at a frequency of about one tumor cell per one million bone marrow cells and isolated by micromanipulation. We then amplified the genome of the single tumor cells using a recently developed PCR technique. Subsequent comparative genomic hybridization (CGH) revealed gains and losses of specific genomic regions. With more genomic profiles of single disseminated tumor cells now becoming available and with the use of high resolution techniques such as matrix CGH one should be able to identify genotypes and genes that may be characteristic for dissemination and ectopic survival.

Return to Table of Contents

ROTTIERS

Oxidative metabolism and gene expression: Gene discovery array analysis

Pieter Rottiers, Vera Goossens and Johan Grooten

Laboratory of Molecular Biology, Flanders Interuniversity Institute for Biotechnology (VIB) and University of Ghent, Ghent, Belgium

Treatment of the mouse fibrosarcoma cell line L929 with tumor necrosis factor (TNF) induces necrotic cell death by a mechanism that depends on production of reactive oxygen intermediates (ROI) by mitochondria (Goossens et al., 1995). Besides oxygen, bioenergetic pathways characteristic of tumor cells (usage of glutamine instead of glucose as oxidative substrate) markedly enchanced ROI production and thus influenced the sensitivity of the cell to TNF-induced necrosis (Goossens et al., 1996). Besides resistance to TNF cytotoxicity L929 cells that have been adapted to use glucose (normal metabolism) instead of glutamine (tumor-specific metabolism) as oxidative substrate exhibit, a differentiated morphology and decreased rate of cell division (Goossens et al., 1996). Apparently, the oxidative metabolism of the cell affect also features characteristic of transformed cells namely uncontrolled growth and dedifferentiation. To verify whether altered gene expression underlies this differential behaviour, we performed PCR-based suppression subtraction hybridization to identify genes, which are differentially expressed between L929 cells dependent on glutamine (TNF-sensitive; dedifferentiated morphology) and L929 cells dependent on glucose (TNF-resistant; differentiated morphology). The subtracted PCR products were hybridized on GDA filters (Genome Systems), spotted with 18,000 non-redundant mouse cDNA clones. Subsequent analysis identified genes known to be involved in the oxidative metabolism of the cell or to contribute to signal transduction pathways, and resistance/sensitivity to apoptosis. In addition a large number of unknown genes were revealed. This result establishes a firm link between oxidative metabolism and gene expression.

Goossens, V., Grooten, J., De Vos, K. & Fiers, W. (1995) Direct evidence for Tumor Necrosis Factor-induced mitochondrial reactive oxygen intermediates and their involvement in cytotoxicity. Proc. Natl. Acad. Sci. U.S.A. 92: 8115-8119.

Goossens, V., Grooten, J. & Fiers, W. (1996) The oxidative metabolism of glutamine - A modulator of reactive oxygen intermediate-mediated cytotoxicity of tumor necrosis factor in L929 fibrosarcoma cells. J. Biol. Chem. 271: 192-196.

Return to Table of Contents

RUUSKANEN

Divergent 2-adrenoceptor subtypes in the zebrafish (Danio rerio)

Jori Ruuskanen[1], Minna Varis[2], Erik Salaneck[4], Tiina Salminen[2], Tommi Nyronen[3], Mark S. Johnson[2], Dan Larhammar[4] and Mika Scheinin[1]

[1]Department of Pharmacology and Clinical Pharmacology, Univ of Turku, Finland
[2]Department of Biochemistry and Pharmacy, Akademi University, Turku, Finland
[3]Center for Scientific Computing (CSC), Espoo, Finland
[4]Department of Neuroscience, Unit of Pharmacology, Uppsala University, Sweden

2-Adrenergic receptors (2-AR:s) belong to a large family of G-protein coupled receptors. A common feature of these receptors is seven a-helical transmembrane domains (TM) thought to form the binding pocket for ligands. They mediate many of the physiological effects of adrenaline and noradrenaline and are target molecules for several drugs. Three human 2-AR subtype genes (2A, 2B and 2C) have been cloned to date. The number of 2-ARs in fish has remained unclear. Only one 2-AR in the fish Cuckoo wrasse (Labrus ossifagus) has been cloned. This receptor, named 2F, has been thought to represent an ancestral 2-AR subtype. Its ligand binding properties are intermediate between 2A and 2C. However, it shows greatest sequence similarity to the 2C and from an evolutionary point of view it is more likely that fish also have three 2-AR subtypes. To study the structure and evolution of 2-AR:s and their possible importance in developmental biology, we have turned to another species of fish, the zebrafish (Danio rerio), which is highly amenable for developmental and genetic studies.

We have cloned the genes coding for the zebrafish 2A-AR and 2C-AR. At protein level, both of these show around 60 % sequence identity when compared to their mammalian counterparts and 50 % identity when compared to other 2-AR subtypes. Analysis of the TM regions is based on a frog rhodopsin based model of the human 2A-AR. In the predicted TM regions, identity of the zebrafish 2A-AR with the human orthologue is 83.3 %, with the rat 2A 82.3 and with the mouse 82.3%. For the zebrafish 2C identities are: human 2C 80.3 %, rat and mouse 2C 86.9 %. These values are much lower than the percentage identities between human and rat or mouse orthologues; 97.5 % for the 2A and 98.5 for the 2C. The rat and mouse orthologues are 99-100 % identical. The greater diversity of the zebrafish receptors is expected to reveal certain residues important for a typical 2-adrenergic ligand binding, which in turn could help in designing subtype selective 2-drugs. Screening for additional subtype(s) using a probe corresponding to an 2B-AR like unpublished EST (GenBank acc. no. AI461341) has been carried out and further analysis of the resulting clones is in progress. Chromosomal mapping of the cloned receptor genes has been started in collaboration with Prof. John H. Postlethwait's group at the University of Oregon, Eugene, USA. Expression and pharmacological testing of the cloned receptor genes has also been started.

Return to Table of Contents

TEMPLE

GATEWAY cloning: A high-throughput gene transfer technology for rapid functional analysis and protein expression

James Hartley, Gary Temple, Michael Brasch, et al.

Life Technologies, Inc., Rockville, Maryland, USA

Each step of characterization of new ORFs requires subcloning into specialized vectors that impart functional properties to the cloned segment. We describe a new method, called Recombinational Cloning (RC), that uses in vitro site-specific recombination to speed the accurate transfer of DNA segments between vector backbones. DNA segments flanked by recombination sites in an Entry Clone can be "automatically" transferred into new vector backgrounds simply by adding the desired "Destination" vector and recombinase, incubating for 1 hour, and transforming any standard E. coli strain. Strong selections ensure that the desired subclones are recovered at high efficiency (typically >90%), reducing or eliminating downstream analysis of candidate clones. The recombination is conservative (no net addition or loss of nucleotides) and transfer occurs without affecting the cloned DNA segment (in contrast to PCR-based approaches). Thus once Entry Clones are created, these clones can be validated and then transferred, unchanged, into any number of vectors. This permits the generation of large collections of validated ORFs (e.g., from model organisms) that can serve as a common source of clones for research. Collaborations to build such collections are in progress.

By incorporating 25 bp attB recombination sites into the 5' end of PCR primers, RC also permits efficient, directional cloning of PCR products (as Entry Clones). The resulting Entry Clones then can be rapidly transferred into any number of Destination Vectors for further analysis.

The RC method is fast, convenient, and can be automated, allowing numerous DNA segments to be cloned and then transferred in parallel into many different vector backgrounds. The resulting subclones maintain reading frame, allowing amino- and carboxy-fusions. Essentially any vector can be readily converted to a Destination vector. Approaches for optimization of protein expression, rapid functional analysis, and the integration of technology platforms will be discussed.

Return to Table of Contents

KAPANADZE

Sequence homology between human and mouse genomic regions to identify the tumor suppressor gene involved in B cell chronic lymphocytic leukemia

Bagrat Kapanadze, Nataliy Makeeva, Olle Sangfelt, Martin Corcoran, Anna Baranova, Eugene Zabarovsky, Nick Yankovsky, Dan Grander and Stefan Einhorn

CCK, Research Laboratory of Radiumhemmet, Karolinska Hospital, Stockholm, Sweden

Previous studies have indicated the presence of a putative tumor suppressor gene on human chromosome 13q14, commonly deleted in patients with B-cell chronic lymphocytic leukemia (B-CLL). We have recently identified a minimally deleted region (MDR) of less than 10 kb, encompassing parts of two adjacent genes, termed Leu1 and Leu2 (leukemia-associated gene 1 and 2). Mutational analysis of Leu1 and Leu2 in 170 CLL samples revealed no small intragenic deletions or point mutations. Subsequent examination of the genomic sequence around the MDR revealed several additional expressed transcripts (ESTs). In addition 50kb centromeric to this region another gene, Leu5 has been identified. This gene encodes a zing-finger protein and shares homology to known genes involved in tumorogenesis. In order to further understand the genomic organisation of such a complex gene rich region, we decided to directly compare the human sequence with that of the mouse, as this may indicate the most important genes in the region, since critical genes tend to be highly conserved between mouse and human. A mouse genomic PAC library was screened with a number of probes covering a 100 kb distance in the human 13q14.3 region, including the MDR. Southern hybridization of subcloned fragments covering the MDR revealed several highly conserved areas, including exon1 and exon 2 of Leu2. Interestingly, human Leu1 does not seem to be conserved in mouse, by sequence analysis, whereas Leu2 sequence was found to be highly conserved. In addition, the Leu5 protein encoding exon has > 95% homology with mouse sequence. In conclusion, following this analysis, the strongest candidates for a conserved tumor suppressor gene in this region are Leu5 and Leu2. Further work is required to elucidate their mechanism of action in this disease and to further identify which gene or genes is the real tumor suppressor gene in BCLL.

Return to Table of Contents

CLARK

Comparative mapping in the Japanese pufferfish (Fugu rubripes)

Clark, M. S; Edwards, Y.J.K; Shaw, L; Snell, P.; Smith, S; and Elgar, G.

Fugu Genomics HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

The project within this laboratory to randomly sequence scan the Fugu genome finished a year ago. Since then we have concentrated on analysing three regions in depth, equivalent to the human chromosomal regions 11p, 20q and MHC. These regions are currently being mapped and, to a limited extent, sequenced. Data currently available indicates the utility of the Fugu genome for identifying conserved regulatory elements, novel genes, confirming predicted genes in human and evolutionary conservation of gene blocks. Examples of these will be illustrated from current unpublished research. There is also an ongoing project to develop ESTs in Fugu. A number of cDNA libraries have been constructed and so far 3,000 sequences have been randomly sequenced from these. Clustered data will be presented on these sequences.

Return to Table of Contents

FROLOVA

Transcriptional regulation of the collagen a1(IX) gene during eye development

Elena I. Frolova and David C. Beebe

Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, St. Louis, Missouri, USA

We have shown that differentiation of the ciliary epithelium at stage 18 of chicken embryo development was accompanied by increased expression of mRNA for the long-isoform of collagen a1(IX), LI-col(IX). Since the expression of this gene is very selective for the ciliary epithelium and occurs early in differentiation of this tissue study of its transcriptional regulation could provide information on transcriptional mechanisms responsible for targeting gene expression to the ciliary epithelium.

To determine the sites in the proximal promoter that were occupied by transcription factors in the differentiating ciliary epithelium we used an in vivo footprinting assay. The main conclusion from this experiment was that one or more DNA binding proteins in ciliary epithelium and retina occupies the proximal promoter of LI-col(IX). In addition, differences in the in vivo DMS footprinting patterns allowed us to conclude that there are different sets of proteins bound to this promoter in the ciliary epithelium and retina. Because LI-col(IX) is expressed in the ciliary epithelium, but not in the retina, these different complexes may be responsible for activating and repressing transcription in these two tissues. To map in more detail the fragments of the LI-col(IX) promoter that bind tissue-specific transcription factors, we used electrophoretic gel mobility retardation assays. Nuclear extracts from ciliary epithelium and neural retina of E7 embryos were analyzed for DNA binding with double-stranded fragme

Two fragments identified by mobility shift assay were used in a one-hybrid screen to identify prospective DNA-binding proteins. A cDNA library fused to the GAL4 activation domain (AD) in pACT2/Asc vector was synthesized from total RNA of the ciliary epithelium of day 7 chicken embryos. The size of the library was approximately 1.3x107 clones. At least 90% of clones contained inserts of 500 to 3,000 base pairs. Two potential DNA binding proteins were identified. One of the proteins (cZic2) showed high homology to mouse ZIC2 mRNA, a sequence recently shown to be expressed in the ciliary epithelium during mouse development.

To determine temporal and spatial pattern of cZic mRNA expression during development, we performed whole-mount in situ hybridization using chicken embryos. Overall expression in the chicken embryo was similar to expression of ZIC2 in mouse. In the eye, cZic is expressed throughout the inner layer of the optic cup at a low level at stage 14. At stage 18 expression is increased at the margin of optic cup, the differentiating ciliary epithelium. By stage 30 there is strong expression of cZic in the non-pigmented ciliary epithelium but not in the retina.

Return to Table of Contents

KREFT

Identification of a novel cellular protein that binds to the HBV RNA pregenome

S. Kreft and M. Nassal

Department of Internal Medicine II, University Hospital, Freiburg, Germany

Due to its small genome of only 3,2 kb, the replication of the hepatitis B virus (HBV) is expected to be tightly linked to the exploitation of host cell functions. The viral RNA pregenome (pgRNA) serves several functions in the viral life-cycle. It functions as mRNA for the capsid and the polymerase (P protein), and it is the substrate for encapsidation and reverse transcription. On its 5' region the structured encapsidation signal is present, which upon binding of P protein, mediates specific RNA packaging into capsids and initiation of reverse transcription. Previous UV-crosslinking data provided direct evidence for the existence of cellular factors that bind close to .

A North-Western (NW) screening procedure with DIG-labeled RNA encompassing HBV as a probe was employed to identify cellular binding proteins. Thereby, a 2.1 kb cDNA from a human liver cDNA expression library was isolated, whose gene product, NIII, consistently bound to RNA in the presence of excess nonspecific nucleic acid competitors. It contained a 3' terminally incomplete ORF encoding a protein of 666 aa with no strong homology to any known protein or RNA-binding motif in the database. Using deletion variants of a bacterially expressed MBP fusion protein in NW-experiments, we mapped the RNA-binding domain to a lysine-rich region close to the C terminus of NIII. In a search for a full-length cDNA, several additional libraries were screened for homology to the central part of NIII (service provided by RZPD, Heidelberg, Germany). Four independent clones with differing 3'-ends were isolated. Two of them encode proteins with 95 and 542 additional amino acids at their C terminus compared to NIII, and hence contain the putative RNA-binding domain as an integral part of the peptide chain. The longest form was termed RBP138 for RNA-binding protein of 138kD. The two other clones both encode the same "truncated" form of RBP138, corresponding to the first 303aa of RBP138, consequently lacking the RNA-binding domain. Both carry a 14nt insertion that generates a STOP codon shortly thereafter. Although the genomic sequence of RBP138 is not yet characterized, this insertion is most likely due to an alternative splicing event. Taken all that together, different isoforms of RBP138 protein seem to exist. Currently, we study the RNA-binding specificity as well as the subcellular localization of some isoforms to unravel the biological role of these new proteins.

Return to Table of Contents