TRANSCRIPTOME 2002: From Functional Genomics to Systems Biology

March 10-13, 2002
Seattle, Washington, USA

Speaker Abstracts

Plenary Session 1: The Role of Transcriptome

Chair: **John Quakenbush,** The Institute for Genomic Research

Keynote Speakers:

- **Systems Biology: Integrating Genomics, Proteomics, Computation and Biology**
  Leroy Hood
- **Bioinformatics and Genome Sciences**
  Mark S. Boguski

Plenary Session 2: Transcription and Processing I

Chair: **Stefan Wiemann,** German Cancer Research Center

- **Globin Gene Activation: A Stochastic Process**
  Frank Grosveld, Mariken de Krom, Mariette van de Corput, and John Strouboulis
- **Molecular Mechanisms of Pre-mRNA Splicing Regulation**
  Juan Valcárcel
- **Alternate Splicing and Gene Expression States: Using cDNAs to Explore the Human Genome**
  Winston Hide, Vladimir N. Babenko and Janet F. Kelso
- **Size-Dependent Pareto-like Distributions in Genomics and Prediction of the Number of Protein-coding Genes in Human Genome**
  Vladimir A. Kuznetsov

Plenary Session 3: Transcription and Processing II

Chair: **Bento Soares,** University of Iowa

- **Multiple Determinants of RNA Editing in Physarum Mitochondria**
  Jonatha M. Gott
- **Employing DNA Microarrays to Study Eukaryotic Translation**
  Thomas Preiss, EMBL, Heidelberg, GERMANY
  Abstract not available for publication
- **RNA Transcription Detected on Chromosomes 21 and 22 Using High Density Oligonucleotide Arrays**
  Thomas R. Gingeras
- **The Genetics of Gene Expression**
  Eric E. Schadt

Plenary Session 4: CDNA Based Gene Discovery I

Chair: **Charles Auffray**, CNRS

- **The NIH Mammalian Gene Collection**
  Robert Strausberg, National Cancer Institute, Bethesda, MD
  Abstract not received in time for publication
- **Kazusa cDNA Project 2002: From Transcript to Protein**
  Osamu Ohara, Reiko Kikuno, and Takahiro Nagase

  FLJ cDNA Project in Japan
  Sumio Sugano
- **The German cDNA Network – cDNAs for Functional Genomics and Proteomics**
  Stefan Wiemann, Jeremy Simpson, Ruth Wellenreuther, Petra Heidrich, Regina Albert, Ingo Schupp, Vladimir Kuryshev, Detlev Bannasch, Rainer Pepperkok, Annemarie Poustka, and the German cDNA Consortium
  Slides: PDF file. **Addendum added May 2, 2002**

Plenary Session 5: CDNA Based Gene Discovery II

Chair: **Winston Hide,** South African National Bioinformatics Institute

- **RIKEN Mouse Genome Encyclopedia**
  Yoshihide Hayashizaki
- **Gene Catalog of Mouse Stem Cells and Early Embryos**
  Minuro Ko
- **Definition of Human Transcriptomes Using ORESTES**
  Andrew Simpson, Ludwig Institute, Sao Paulo, BRAZIL
  Abstract not available for publication
- **The University of Iowa Mammalian Gene Discovery and Expression Program**
  Bento Soares, University of Iowa, Iowa City, IA
  Abstract not available for publication

Plenary Session 6: Transcription and Processing III

Chair: **Yoshihide Hayashizaki,** RIKEN Genomic Sciences Center

- **Microarray Technology for Transcript Profiling and Analysis of Genetic Variability**
  Joakim Lundeberg
- cDNA-Microarray Profiling of Ovarian Cancer
  Claudio Schneider, LNCIB, Trieste, ITALY
  Abstract not available for publication
- **Zhu Chen,** Chinese Academy of Sciences, Shanghai, PEOPLE'S REPUBLIC OF CHINA
  Abstract not available for publication
- Clone Selection and Analysis Methods for Full-Insert cDNA Sequencing
  Lukas Wagner, NIH, Bethesda, MD
  Abstract not available for publication

Plenary Session 7: Sequenced-Based Tools for Gene Discovery I

Chair: **Andrew Simpson,** Ludwig Institute for Cancer Research

- **Comparative Genomics Tools: TIGR Gene Indices and TIGR Orthologous Gene Alignments (TOGA)**
  Yuandan Lee, Jennifer Tsai, Foo Chung, Svetlana Karamycheva, Babak Parvizi, Geo Pertea, Razvan Sultana, Valentine Antonescu, Joseph White, and John Quackenbush
- **UTRDB and UTRSITE: Specialized Databases of Sequences and Functional Elements of 5' And 3' Untranslated Regions of Eukaryotic mRNAs**
  Pesole, Graziano, Gissi C., Grillo G., Licciulli F., Mignone F., Iacono M., Liuni S.
- **I.M.A.G.E. Data Mining Tools: Introducing the IQ**
  Peg Folta
  Slides: PDF version
- **Using NCBI Resources for Gene Discovery**
  Kim D. Pruitt, Richa Agarwala, Hsiu-Chuan Chen, Slava Chetvernin, Deanna Church, Olga Ermolaeva, Wonhee Jang, Paul Kitts, Jonathan Kans, David Lipman, Jim Ostell, Sergey Resenchuk, Greg Schuler, Lynn Schrim, Steve Sherry, Tatiana Tatusova, Lukas Wagner, Sarah Wheelan
  Slides: PDF

Plenary Session 8: Sequenced-Based Tools for Gene Discovery II

Chair: **Greg Lennon,** Cell Logic, Inc.

- **Tim Hubbard,** The Sanger Centre, Cambridge, UK
  Abstract not available for publication
- **Extraction of Biological Meaning from Array Data: What is Required?**
  Roger Bumgarner, University of Washington, Seattle, WA
  Abstract not available for publication
- **Comparative and Functional Analysis of Cardiovascular-Related Genes**
  Len A. Pennacchio, Michael Olivier, Jaroslav A. Hubacek, Jonathan C. Cohen, Ronald M. Krauss, and Edward M. Rubin
- **The Detection of Env-Transcripts in Gossypium**
  Zaki, E.A.

Plenary Session 9: Gene Expression – Transcript Abundance and Disease

Chair: **John Quackenbush,** The Institute for Genomic Research

- **Delineation of transcriptional Pathways in Vivo: Temporal Profiling in Muscle Regeneration**
  Eric Hoffman, Children's National Medical Center, Washington, DC
  Abstract not available for publication
- **Application of Arrays to the Study of Colon Cancer**
  Timothy J. Yeatman, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL
  Abstract not available for publication
- **Using SAGE to Explore the Genome**
  Saurabh Saha, Alberto Bardelli, Kenneth W. Kinzler, & Bert Vogelstein
- **System Biology: From Genome to Physiology**
  Howard Jacob, Medical College of Wisconsin, Milwaukee, WI
  Abstract not available for publication

Plenary Session 10: Alternate Measures of Expression

Chair: **Winston Hide,** South African National Bioinformatics Institute

- **Gels, Mass Spectrometry and Chips: Large Scale Acquisition of Protein Abundance Fingerprints**
  N. Leigh Anderson, Large Scale Biology Corporation, Germantown, MD
  Abstract not available for publication
- **Toward a Proteome Atlas for C-elegans**
  Marc Vidal, Harvard Medical School, Boston, MA
  Abstract not available for publication
- **Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data**
  Frank Holstege, University Medical Center-Utrecht, AB Utrecht, THE NETHERLANDS
  Abstract not available for publication
- **Mihael Polymeropoulos,** Novartis Pharmaceuticals Corporation, Gaithersburg, MD
  Abstract not available for publication
- **Breaking the Bottle Neck of Genomics and Proteomics**
  Moncef Jendoubi
- **A Novel 22,000 Feature In Situ-Synthesized 60-Mer Oligonucleotide Microarray: Expression Profiling Of E12.5 Mouse Embryo and Placenta**
  Mark G. Carter, Condie E. Carmack, Yong Qian, Toshio Hamatani, Pius Brzoska, S. Stuart Hwang, and Minoru S.H. Ko

Plenary Session 11: The Future

Chair: **Zhu Chen,** Chinese Academy of Sciences

- **Back to the Future: Integrating Expression with Genomic, Genetic, and Metabolic Data**

- **John Quackenbush**
  Bioinformatics and Genomics to Gain Insight into the Etiology of Cancer
  Ken Buetow, National Cancer Institute, Bethesda, MD
- Abstract not available for publication
  ArrayExpress - A Public Database for Microarray Data
  Helen Parkinson
- **Slides: PDF version**
  A Novel Platform for High-Sensitivity Analysis of Cell-Specific Gene Expression using Microarrays
  Rajiv Raja, Jim Stanchfield, Mark Erlander and Steve Kunitake

Systems Biology: Integrating Genomics, Proteomics, Computation and Biology

Leroy Hood, Institute for Systems Biology, Seattle, WA

The Human Genome Project has altered the view and practice of biology and has led to several paradigm changes—systems biology, and predictive and preventive medicine. I will discuss these changes and consider the analysis of two biological systems: galactose metabolism in yeast, and sea urchin development, using integrative systems approaches. I will also discuss the Institute for Systems Biology and the approaches we are taking to create an optimal environment for carrying out systems biology.

Bioinformatics and Genome Sciences

Mark S. Boguski, Fred Hutchinson Cancer Research Center, Seattle, WA

Never before in the history of biology have we had so much information yet, proportionately, so little knowledge. New technologies, along with the new genome scale mindset, will be generating even more prodigious amounts of data, not only DNA sequence data but also more complex data from functional genomics and proteomics approaches. Optimizing interpretation of the combined outputs from "discovery-driven" and hypothesis-driven research represents one of the key challenges and opportunities at this turning point in the biomedical research.

Globin Gene Activation: A Stochastic Process

Frank Grosveld, Mariken de Krom, Mariette van de Corput, and John Strouboulis, Dept. of Cell Biology, Erasmus University, Rotterdam, THE NETHERLANDS

Models for stochastic gene activation have been invoked to explain mono-allelic expression patterns of genes involved in immune and cellular responses, dosage compensation and cell commitment, but have also been proposed for gene activation in general. We have tested the generality of a stochastic model at the allelic level using the alpha and beta globin genes, as they play no role in any cellular selection processes. Allelic transcription patterns and mRNA levels in single cells showed an imbalance of alpha versus beta globin expression in a significant proportion of cells and different probabilities in the activation of alpha and beta globin genes.Interestingly once a cell chooses a particular pattern of expression, that pattern becomes fixed in the cell. These data provide strong evidence for a stochastic basis of gene activation in general, which has broad implications for a variety of processes from the evolution of different layers of gene control to phenotypic differences between genetically identical cells and organisms.

Molecular Mechanisms of Pre-mRNA Splicing Regulation

Juan Valcárcel, Gene Expression Programme, European Molecular Biology Laboratory, Heidelberg, GERMANY

At least 45-60 % of human genes produce primary transcripts that can be alternatively spliced to generate up to thousands of potential mRNAs and protein isoforms, thus significantly expanding the informational content of the genome. Alternative splicing can originate proteins with diverse, even opposite functions, and has the capacity to spawn combinatorial complexity that can serve to establish cellular networks. The process is often subverted in disease, and up to 40% of genetic defects can be associated with incorrect RNA processing. Despite the prevalence of alternative splicing and the versatility it provides for the regulation of gene expression, the molecular mechanisms involved in the use of different splice sites during cell differentiation, development and disease are poorly understood.

Work in my lab focuses on three areas: To study in detail the mechanisms used by tissue-specific factors to promote or repress particular splice sites, with the goal of understanding which steps of the splicing process are targets of regulation. Work has focused on a Drosophila RNA binding protein, Sex-lethal, that is expressed exclusively in female flies and induces female-specific patterns of alternative splicing. While some alternative splicing events are regulated by Sex-lethal at the earliest steps of splice site recognition, recent results indicate that it can also regulate the very last step of the splicing process: the joining of the exons and intron release. This observation opens novel mechanistic possibilities for splicing regulation. Similar molecular events are also at the basis of cryptic splice site activation in at least one form of human beta-thalassemia. To understand how ubiquitous splicing factors can be activated to modulate alternative splicing during the process of programmed cell death. Work focuses on a signal transduction cascade that targets the splicing factor TIA-1 and modulates alternative splicing of the Fas receptor in T cell immune responses. Some elements of this cascade appear to be altered in Autoimmune Lymphoproliferative Syndromes. To gain insights into the molecular logic behind the establishment of post-transcriptional regulatory programs. We are using bioinformatic and microarray analyses to identify novel genes whose alternative splicing is controlled by specific splicing regulators.

Alternate Splicing and Gene Expression States: Using cDNAs to Explore the Human Genome

Winston Hide, Vladimir N. Babenko and Janet F. Kelso, South African National Bioinformatics Institute, University of the Western Cape, Bellville, SOUTH AFRICA

Recent studies utilizing ESTs and reference mRNA data have revealed that alternate transcription, including alternative splicing, polyadenylation and transcription start sites, occurs within at least 60 % of human genes. It is likely that this is an underestimate as EST sampling has been used to derive the figures. Transcript form surveys have yet to integrate the genomic context, expression, and contribution to protein diversity of isoform variation. Exhaustive manual confirmation of genome sequence annotation, coupled with comparison to available expressed sequence data has been used here to accurately associate isoforms showing exon skipping with genomic sequence to reveal potential protein coding alteration. In addition, relative expression levels of transcripts have been estimated from EST representation in the public databases. Our rigorous method has been applied to 545 described genes in the first intensive study of exon skipping based on chromosome 22 genome sequence and matched human transcripts. The study has led to the discovery of 62 exon skipping events in 52 genes, with 57 exon skips altering the protein coding region. A single gene, (FBXO7) expresses an exon repetition. EST sampling analysis indicates that 58.8% of highly represented multi-exon genes are likely to express exon-skipped isoforms in ratios that vary from 1:1 to 1:>100. Comparisons with mouse show a similar overall level of skipping, although not at the same exon boundaries/genes. Analysis of cancer genes show that aberrant forms of skipping may segregate with cancer expression libraries.

Size-Dependent Pareto-like Distributions in Genomics and Prediction of the Number of Protein-coding Genes in Human Genome

Vladimir A. Kuznetsov, National Institute of Child Health and Human Development, NIH, The Laboratory of Integrative and Medical Biophysics, Bethesda, MD

Recently, we described the class of skew size-dependent probability distributions that appear in samples provided by many large-scale gene expression experiments and by proteome and genome data sets /1-3/. The observed distributions have the following characteristic in common: there are few frequent and many rare classes. The form of the distribution systematically depends on size of the sample. We developed a stochastic model of population growth that leads us to a size-dependent Pareto-like probability distribution of classes by their frequencies of occurrences in multi-classes finite population. In this work, using SAGE data (www.sagenet.org), we statistically identified such distribution for the transcript abundance values in human colon cancer transcriptome presented by ~600,000 SAGE tags. We developed a new computational methodology to remove major experimental errors from SAGE database. The corrected probability distribution of transcript abundance values in the human transcriptome was obtained. A new method to estimate the number of genes was also developed. About 10,500 expressed protein-coding genes in a singe colon cancer human cell, and ~31,000 expressed protein-coding genes in a population of human cells were estimated as the conservative numbers. Our Pareto-like model also fits to empirical frequency distributions of the protein domain occurrence values for distinct protein domains in 10 archaean, 25 bacterial and 6 eukaryotic proteomes of fully-sequenced genomes (www.ebi.ac.uk/interpro/). About 36,500 evolutionary conserve protein-coding genes were predicted for the entire human genome based on our extrapolation analysis of the relationships between the number of genes in the fully-sequenced genome organisms and the number of protein domain occurrences per proteome, and the numerical characteristics of the Pareto-like distribution of the protein domain occurrences in the proteome.1. V. A. Kuznetsov, R.F. Bonner (1999) Statistical tools for analysis of gene expression distributions! with missing data. In: 3rd Annual Conference on Computational Genomics. Nov.18-21. Baltimore, MD:The Institute for Genomic Research, p.26.2. V. A. Kuznetsov (2001) Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. EURASIP J. on Applied Signal Processing, 4, 285-296. 3. V. A. Kuznetsov (2002) Statistics of the numbers of transcripts and protein sequences encoded in the genome. In: Computational and Statistical Methods to Genomics. Kluwer: Dordrecht etc. pp.125-171.

Multiple Determinants of RNA Editing in Physarum Mitochondria

Jonatha M. Gott, Center for RNA Molecular Biology, Case Western Reserve University, Cleveland, OH

The term RNA editing is used to describe a number of mechanistically distinct processes which result in the production of an RNA whose sequence differs in specific ways from that of its template. Editing can involve alteration of either bases (substitution) or nucleotides (insertion/deletion). Substitutional editing involves changing one base into another at the RNA level, usually via deamination of cytosine to uridine (C to U) or adenosine to inosine (A to I). Other forms of editing entail the deletion of encoded nucleotides and/or the insertion of additional, non-encoded nucleotides, either post-transcriptionally or as the RNA is being synthesized. These changes lead to programmed alterations in gene expression, ranging from amino acid substitutions to the creation of entire open reading frames.

The organism that carries out the broadest range of editing events known to date is the myxomycete Physarum polycephalum. The majority of the RNAs encoded in the Physarum mitochondrial genome (including mRNAs, tRNAs, and rRNAs) are altered through the frequent insertion of non-encoded C, U, UA, CU, GU, UU, AA, or GC residues at specific sites and rare C to U changes. These sequence alterations occur co-transcriptionally through a process that is both precise and extremely efficient. We are currently investigating the mechanism by which the insertional editing machinery specifies both the sites of insertion and the identity of the added nucleotides. Using a series of hybrid templates that consist of editing-competent transcription elongation complexes (mtTEC) linked to other DNA fragments, we show that editing is not strictly sequence-dependent and that the site of nucleotide insertion and the identity of the added nucleotide are likely to be specified independently.

RNA Transcription Detected on Chromosomes 21 and 22 Using High Density Oligonucleotide Arrays

Thomas R. Gingeras, Affymetrix Inc., Santa Clara, CA

The first drafts of complete human genome sequence have brought with them the opportunities to map the RNA transcription patterns that are characteristic of each differentiated and undifferentiated cell type and characterize the sequence variations that underlie the phenotypic differences observed in the human population. By using the very high information content inherent in high-density oligonucleotide arrays it will be possible to map the locations of RNA transcription along the length of the entire human genome. Such a transcriptome map will provide information concerning: 1) the identification of novel transcription domains of the genome, 2) the predominant utilization of exon sequences during differentially spliced gene expression and 3) a empirically derived set of results which can be compared to the sequence annotation now being assembled for the human genome. Initial experiments have been focused on mapping the transcription domains originating from chromosomes 21 and 22. Oligonucleotide probes of 25 nucleotides in length have been selected to interrogate the non-repetitive nucleotide sequences within these chromosomes at a resolution of approximately 35 base pairs (measured from the central positions of each adjoining probe). A total of three arrays each of which contain ~400,000 probe pairs (probes designed to be perfect match and mismatch) are needed to interrogate ~35 Mb of non-repetitive sequences from these chromosomes. The map results when overlaid on to the sequence annotations available for these two chromosomes reveal that as much as an order of magnitude more of the genomic sequences are used for transcription than envisioned by the predicted and characterized exons. These transcripts represent a hidden transcriptome not accounted for in current annotated maps.

The Genetics of Gene Expression

Eric E. Schadt, Rosetta Inpharmatics, Kirkland, WA

In 1980 Botstein et al. proposed that sequence differences be treated as markers in order to map genes involved in inherited traits. Since that time, the number of genes mapped to positions in the human genome has grown exponentially. Mapping these genes to inherited traits has been extremely successful for simple Mendelian diseases; however, finding such genes for diseases, and their associated risk traits, that are of public health interest has proven difficult. Reasons for this difficulty include disease heterogenity (disease sub-types with some or no overlapping genetic causes), missclassification (from using discrete classifications of disease from thresholds and combinations of thresholds), and cumulative environmental influences. With the advent of technology to measure changes in gene expression, i.e., changes in mRNA transcript abundance, it should be possible to unravel some of the complexity existing for these common diseases. We will show that studying the genetic component of mRNA expression is possible through the use of a sample of CEPH (Centre d'Etude du Polymorphisme Humain) families and other experimental, segregating populations. Discussion will include assessing differential expression of genes for the population, determining whether variation in expression is under genetic control and establishing loci that control variation in expression. Time permitting, it will be shown that combination of information across a number of genes can be used to establish, support and/or confirm a biological pathway.

Kazusa cDNA Project 2002: From Transcript to Protein

Osamu Ohara [1,2], Reiko Kikuno[1], and Takahiro Nagase[1], [1]Kazusa DNA Research Institute, Kisarazu, JAPAN, [2]RIKEN Research Center for Allergy and Immunology, Kisarazu, JAPAN

We have conducted a human cDNA sequencing project for the identification of unknown human transcripts in the past seven years. A distinctive characteristic that separates our project from other cDNA sequencing projects is that we have focused our sequencing efforts on long cDNA clones (>4 kb), particularly those encoding large proteins (>1000 amino acid residues). This approach has resulted in a unique cDNA resource consisting of nearly 2000 large cDNA clones for newly identified human genes, which are known as KIAA genes. The detailed information regarding KIAA genes can be accessible through HUGE protein database at http://www.kazusa.or.jp/huge. As we enter the end game of identification of human transcripts coding for unknown large proteins, we are gradually shifting our aim toward characterization of proteins encoded by KIAA genes. Besides functional characterization of KIAA gene products (*e.g.*, systematic search for binding partners of KIAA gene products *in vitro*), our important aim is to link the KIAA transcripts with their protein products *in vivo*. To achieve this, we are planning to extensively use antibodies as specific detection reagents for the KIAA gene products. The set of expression-ready KIAA cDNA clones and antibodies against the KIAA gene products will provide us with a powerful tool for filling the gap between the mammalian transcriptome and proteome.

FLJ cDNA Project in Japan

Sumio Sugano, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, JAPAN

Although a huge data of draft and finished sequences of the human genome are now available, it is still not a trivial task to identify genes from genome sequences. Thus, a full-length complementary DNA (cDNA), which is a complete DNA copy of mRNA, is required for the identification of a gene and the determination of its structure. We are now participating FLJ cDNA project to collect and sequence full-length cDNA clones under support of ministry of economy, industry and international trade. In this project, we extensively used cDNA libraries made by Oligo-capping method, a cap targeted selection procedure for full-length cDNAs. So far, we have accumulated more than 800,000 5' end sequences through random sequencing of cDNA clones from about 100 kinds of human cDNA libraries. Among them, 12340 clones have been fully sequenced. The accuracy of the sequences was more than 99.95%. The average length of cDNAs whose sequence were determined was about 2200bp, which distributed from 1kb to 5kb. Furthermore we could cluster 5' end of many mRNAs (more than 5000 genes), which could be mapped to onto the human draft genomic sequence. The detail of this project will be presented.

The German cDNA Network – cDNAs for Functional Genomics and Proteomics

Slides: PDF file

Stefan Wiemann[1], Jeremy Simpson[2], Ruth Wellenreuther[1], Petra Heidrich[1], Regina Albert[1], Ingo Schupp[1], Vladimir Kuryshev[1], Detlev Bannasch[1], Rainer Pepperkok[2], Annemarie Poustka[1], and the German cDNA Consortium, [1]Molecular Genome Analysis,German Cancer Research Center, Heidelberg, GERMANY, [2]Cell Biology Program, European Molecular Biology Laboratory, Heidelberg, GERMANY

We have formed a network within the German Genome Project aiming at the generation and sequencing of novel full-length cDNAs, and the comprehensive functional analysis the encoded proteins. Over 5,000 cDNAs (> 13.7 Mb) have been sequenced since. This set and greater 83.000 EST-sequenced clones is used to generate a minimal set of full-length cDNAs for employment in subsequent functional analysis. In order to study complex biological systems like the regulation of the cell cycle, apoptosis or protein secretion, a multitude of complementary approaches need to be followed that combine genomics and proteomics strategies, but also cell biology and computational biology. The availability of full-length cDNAs is elementary for most of these. Only the integration of data from many sources will help to eventually understand protein function and interaction in protein networks and complex biological systems.

ADDENDUM (May 2, 2002):

We have formed a network within the German Genome Project aiming at the generation and sequencing of novel full-length cDNAs, and the comprehensive functional analysis the encoded proteins. Over 5,100 cDNAs (> 14 Mb) have been sequenced since. This set and greater 83.000 EST-sequenced clones is used to generate a minimal set of full-length cDNAs for employment in subsequent functional analysis. In order to study complex biological systems like the regulation of the cell cycle, apoptosis or protein secretion, a multitude of complementary approaches need to be followed that combine genomics and proteomics strategies, but also cell biology and computational biology. The availability of full-length cDNAs is elementary for most of these. Only the integration of data from many sources will help to eventually understand protein function and interaction in protein networks and complex biological systems.

References:

1. Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R. and Wiemann, S. (2000). Systematic subcellular localization of novel proteins identified by large scale cDNA sequencing. EMBO Rep 1, 287-292.
2. Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., Lauber, J., Dusterhoft, A., Beyer, A., Kohrer, K., Strack, N., Mewes, H. W., Ottenwalder, B., Obermaier, B., Tampe, J., Heubner, D., Wambutt, R., Korn, B., Klein, M. and Poustka, A. (2001). Toward a Catalog of Human Genes and Proteins: Sequencing and Analysis of 500 Novel Complete Protein Coding Human cDNAs. Genome Res 11, 422-435.
3. Pepperkok, R., Simpson, J. and Wiemann, S. (2001). Being in the right location at the right time. Genome Biol 2, REVIEWS 1024.
4. Simpson, J. C., Neubrand, V. E., Wiemann, S. and Pepperkok, R. (2001). Illuminating the human genome. Histochemistry and Cell Biology 115, 23-29.

Links:

1. www.dkfz.de/en/mga/ related to Wiemann et al., Genome Res. 2001
2. www.dkfz.de/LIFEdb related to Simpson et al., EMBO Rep. 2000

RIKEN Mouse Genome Encyclopedia

Yoshihide Hayashizaki, Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, JAPAN

We have been working to establish the comprehensive mouse full-length cDNA collection and sequence database to cover as many genes as we can, named Riken mouse genome encyclopedia. Recently we are constructing higher-level annotation (Functional ANnoTation Of Mouse cDNA; FANTOM) not only with homology search based annotation but also with expression data profile, mapping information and protein-protein database. More than 1,000,000 clones prepared from 163 tissues were end-sequenced to classify into 128,000 clusters and 60,000 representative clones were fully sequenced. As a conclusion, the 60,000 sequences contained 35,000 unique sequences with more than 24,000 clear protein-encoding genes. The next generation of life science is clearly based on all of the genome information and resources. Based on our cDNA clones we developed the additional system to explore gene function. We developed cDNA microarray system to print all of these cDNA clones, protein-protein interaction screening system, protein-DNA interaction screening system and so on. The integrated data of all of the information are very useful not only for analysis of gene transcriptional network and for the connection of gene to phenotype to facilitate positional candidate approach. In this talk, the prospect of the application of these genome resourced should be discussed. More information is available at the web page: http://genome.gsc.riken.go.jp/

Gene catalog of mouse stem cells and early embryos

Minoru S. H. Ko[1], Yulan Piao[1], Carole A. Stagg[1], Patrick Martin[1], Dawood B. Dudekula[1], Yong Qian[1], Lakshmi Kosuru[1], Vincent VanBuren[1], Tetsuya S. Tanaka[1], Saied A. Jaradat[1], Mark G. Carter[1], Wendy L. Kimber[1], Kazuhiro Aiba[1], Toshio Hamatani[1], Toshiyuki Yoshikawa[1], Janet Kelso[2], Winston Hide[2].

1Developmental Genomics and Aging Section, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD 21224, USA

2South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

We report here our continued efforts to assemble a mouse gene catalog from early embryonic stages as well as embryonic and adult stem cells. To overcome the scarcity of materials in these stages, we recently developed a method to construct long-insert enriched cDNA libraries from submicrogram amounts of total RNA. This new method allowed us to add ~80,000 new ESTs, which may represent full-length cDNAs, to our previous collection of ~80,000 ESTs. These ESTs were clustered into ~26,000 unique gene collections, which added ~11,000 new cDNA clones to the NIA Mouse 15K cDNA clone set. The new 11K clone set is now being single-colony isolated and sequence-verified. These ESTs and cDNA clone collections will provide essential tools to study the mammalian development as well as the fundamentals of stem cells, and provide initial hints about the differential nature of embryonic and adult stem cells.

NIA Mouse cDNA Project Home Page: http://lgsun.grc.nia.nih.gov/cDNA/cDNA.html

Laboratory Home Page: http://www.grc.nia.nih.gov/branches/lg/dgas/dgas.htm

Microarray Technology for Transcript Profiling and Analysis of Genetic Variability

Joakim Lundeberg, Department of Biotechnology, Royal Institute of Technology, Stockholm, SWEDEN

Microarray systems for analysis of transcript profiles and genetic variability have been established to elucidate and map gene function in different biological systems. Transcript profiles can be achieved by subtractive or by global methods such as SAGE, EST sequencing or array based technologies. Here we will give an example of the combined use of subtractive methods and global cDNA arrays to identify low abundant transcripts. Furthermore systems for achieving high quality cDNA arrays and procedures to facilitate analysis of minute amounts of samples will be described. The former relates to that it has been shown that relatively high concentrations of probe nucleic acids are required to yield a high dynamic range for the subsequent hybridisation assays. Here we present a novel approach to capture and isolate of PCR products in conditions suitable for printing cDNA slides. We have also developed a microarray based allele specific extension method for the accurate typing of SNPs involving fluorescent nucleotides. The use of allele-specific primers has previously been employed to identify single base variations but it is acknowledged that certain mismatches are not refractory to extension thereby reducing the discriminatory power of these extension assays. Here we have overcome this limitation by introducing apyrase, a nucleotide degrading enzyme, to the extension reaction. Examples will given including analysis of cardiovascular SNP markers as well scanning analysis of whole gene sequences.

Comparative Genomics Tools: TIGR Gene Indices and TIGR Orthologous Gene Alignments (TOGA)

Yuandan Lee, Jennifer Tsai, Foo Chung, Svetlana Karamycheva, Babak Parvizi, Geo Pertea, Razvan Sultana, Valentine Antonescu, Joseph White, and John Quackenbush. The Institute for Genomic Research, Rockville, MD

While drafts of human genome sequences have been published recently and other eukaryotic genome sequencing projects are advancing rapidly, identification and classification of gene sequences remains a significant challenge because of the lack of experimental evidence and the shortcomings of the available gene prediction programs. EST sequencing and analysis remains a primary research tool for the identification and categorization of transcribed genes in a wide variety of species and an important resource for the annotation of genomic sequences. The ESTs, known genes, predicted gene transcripts, and available mapping and sequencing data have been integrated together to build a cross-genome reference database, the TIGR Gene Indices (TGI; <http://www.tigr.org/tdb/tgi.shtml>). TIGR gene indices are a collection of species-specific databases that use a highly refined protocol to analyze EST sequences in an attempt to identify the genes represented within the EST data and to provide structural and functional information regarding those genes. Gene Indices are constructed by first comparing, clustering, then assembling EST and annotated gene sequences (ET) from GenBank for the targeted species. This process produces a set of unique, high-fidelity virtual transcripts, or Tentative Consensus (TC) sequences. There are 49 gene indices with over 500,000 TCs available now for major economically and biologically important eukaryotic organisms including human, mouse, rat, cattle, pig, *Arabidopsis*, rice, soybean, yeast, *Drosophila*, *Xenopus*, zebrafish, *C. elegans*, *Plasmodium falciparum* etc. The TC sequences can be used to provide putative genes with functional annotation, to link the transcripts to mapping and genomic sequence data, to provide links between orthologous and paralogous genes, and as a resource for comparative sequence analysis. In particular, we have used comparisons between TC sequences from 49 gene indices to construct the TIGR Orthologous Gene Alignment (TOGA) database. TOGA uses a transitive association algorithm to link together tentative orthologues and has allowed the identification of more than 30,000 potential orthologous groups in eukaryotes. This data provides a unique opportunity for phylogenetic and functional analysis of genes, and for the annotation of genes used in microarray analysis.

UTRDB and UTRSITE: Specialized Databases of Sequences and Functional Elements of 5' and 3' Untranslated Regions of Eukaryotic mRNAs

Pesole, Graziano[1], Gissi C.[1], Grillo G.[2], Licciulli F.[2], Mignone F.[1], Iacono M.[1], Liuni S.[2], [1]Università di Milano, Milan, ITALY, [2] CSMME-C.N.R., Bari, ITALY

The 5' and 3' untranslated regions of eukaryotic mRNAs may play a crucial role in the regulation of gene expression controlling mRNA nucleo-cytoplasmic transport, subcellular localization, stability and translation efficiency. In order to study the general structural and compositional features of these sequences we have developed UTRdb, a specialized database of 5' and 3' UTR sequences of eukaryotic mRNAs cleaned from redundancy (Pesole, Liuni et al. 2002).UTRdb (release 15.0) contains about 250,000 entries (>65,000,000 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. All Web resources we implemented for the retrieval and the analysis of UTR sequences are available at the UTR home page (http://bighost.area.ba.cnr.-it/BIG/UTRHome/) 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 UTRblast allows to assess sequence similarity between a query sequence and UTRdb entries.The analysis of complete UTR sequences contained in this database allowed us to define specific structural and compositional features of UTRs from mRNAs belonging to various eukaryotic taxa (Pesole, Grillo et al. 2000). Pesole, G., G. Grillo, et al. (2000). 'The untranslated regions of eukaryotic mRNAs: structure, function and bioinformatic tools for their analysis.' Briefings in Bioinformatics 1(3): 236-249.Pesole, G., S. Liuni, et al. (2002). 'UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002.' Nucleic Acids Res 30(1): 335-40.

I.M.A.G.E. Data Mining Tools: Introducing the IQ

Slides: PDF version

Peg Folta, Lawrence Livermore National Laboratory, CA

The I.M.A.G.E. Consortium has maintained the largest cDNA clone collection in the world since its creation in 1993. U.S. and European distributors provide the current collection of over 5.5 million clones, from six species worldwide. The I.M.A.G.E. collection has been the basis of several key genomic projects, such as the current NIH Mammalian Gene Collection (MGC), The NIH Cancer Genome Anatomy Project (C-GAP), and the Merck Gene Index projects. 64% of the human ESTs in GenBank have been obtained from I.M.A.G.E. clones. Based on the collections impact to the industry, one of our major focuses has been to quantify and track the make-up of the collection and provide public access to all data with intelligent data mining techniques.

IMAGEne is a mature clustering software product that provides java-based query and display of each of its gene clusters. Clustering is based primarially on sequence overlaps and clone membership. The number of gene clusters with full-lenth clone representatives has risen significantly over the past year, as a direct result of the MGC project. The number of clusters without an I.M.A.G.E. clone has decreased, due to improved library creation techniques. Detailed statistics on the current build will be presented, as will trends in the collection over the last few years.

I.M.A.G.E. tracks all information associated with the clones within the collection, from the originating cDNA library data through the resulting EST and full insert sequence in an Oracle database. A new intelligent query tool, IQ, has been developed and released to the public for mining of this data. The IQ tool allows the user to specify the attributes used to define the query and the content, format, and destination of the results.

Problem clones identified by I.M.A.G.E. consortium members, the distributors, and users are also tracked in the database. These database tables and the associated web-based forms have been enhanced this year for ease of entry and query. Reported problems are now used directly in IMAGEne clustering and have effected the singleton analysis results.

This work was partially funded by the NIH and was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

Using NCBI Resources for Gene Discovery

Slides: PDF

Kim D. Pruitt, Richa Agarwala, Hsiu-Chuan Chen, Slava Chetvernin, Deanna Church, Olga Ermolaeva, Wonhee Jang, Paul Kitts, Jonathan Kans, David Lipman, Jim Ostell, Sergey Resenchuk, Greg Schuler, Lynn Schriml, Steve Sherry, Tatiana Tatusova, Lukas Wagner, Sarah Wheelan, NCBI, Bethesda, MD

1. RefSeq Information:
   http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
2. Human Genome Resources:
   http://www.ncbi.nlm.nih.gov/genome/guide/human/
3. Genome Assembly and Annotation Process:
   http://www.ncbi.nlm.nih.gov/genome/guide/build.html

NCBI provides several resources that support gene discovery including RefSeq sequence standards, the LocusLink gene-centered description database, UniGene clustering of related transcripts, HomoloGene computed homologous clusters, and comparative maps to the assembled and annotated human genome sequence displayed in the Map Viewer. The RefSeq project provides a non-redundant database of sequence standards through several methods including automated pipelines, manual curation, collaboration, and ab initio gene prediction on genome assemblies. Predicted model RefSeqs represent a data set which includes putative novel genes. Companion resources provide additional information about related sequences, expression, function, and other attributes. Similarly, UniGene and HomoloGene data, which provide information about related sequences, can be mined computationally or manually, using the query interface, to identify novel genes. The Map Viewer resource displays multiple sequence-based types of information including known genes, predicted genes, transcribed regions, markers, and variation features; these displays integrate several sources of information and provide a very powerful approach to identify genomic sequence data with novel coding potential. When used together, these resources provide a wealth of information about known and predicted genes and greatly facilitate identification of putative novel genes for further investigation.

Comparative and Functional Analysis of Cardiovascular-Related Genes

Len A. Pennacchio[1], Michael Olivier[2], Jaroslav A. Hubacek[3], Jonathan C. Cohen[3], Ronald M. Krauss[1], and Edward M. Rubin[1], [1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, [2]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, [3]Center for Human Nutrition and McDermott, Center for Human Growth and Development, UT Southwestern Medical Center, Dallas, TX

Cardiovascular diseases represent the leading cause of death in Western society. It has been well established that both genetic and environmental factors contribute to the pathophysiology of these disorders. To better understand genes important in cardiovascular biology, we are performing comparative sequence analysis of ~200 cardiovascular-related genes in numerous vertebrate species. We will describe numerous computational tools and databases we have constructed to exploit this information. As an example of this analysis, we will discuss the apolipoprotein gene cluster on human chromosome 11q23 (APOA1/C3/A4) which is known to influence a variety of plasma lipid parameters and atherosclerosis susceptibility in humans. To facilitate the identification of evolutionarily conserved sequences with potential function near this cluster, we determined the sequence of ~200 kilobasepairs (kbp) of orthologous mouse, rabbit, and chicken DNA and compared these sequences to human. The presence of a stretch of inter-species sequence conservation approximately 30 kbp proximal to the APOA1/C3/A4 gene cluster, led us to an interval that upon further analysis was shown to encode a new member (APOA5) of the chromosome 11 apolipoprotein gene cluster. We find that APOA5 is expressed primarily in liver tissue and encodes a secreted protein that dramatically impacts plasma triglyceride levels in humans and mice. Specifically, mice over-expressing a human APOA5 transgene display a 70% decrease in plasma triglyceride concentrations, while oppositely, mice lacking Apoa5 have a 400% increase in this lipid parameter. These findings in mice suggested that alterations in APOA5 could also influence human plasma lipid levels. To explore this possibility, we identified several single nucleotide polymorphisms (SNPs) in the human APOA5 gene and determined their distribution in two independent patient populations. Through this analysis, we found a significant association between several polymorphisms and abnormal triglyceride levels in both independent studies. These findings in humans and mice illustrate the utility of comparative sequence analysis to prioritize regions of the genome for further study and suggest an important physiological role for apoAV in affecting plasma levels of triglyceride, a major risk factor for heart disease in humans.

The Detection of Env-Transcripts in Gossypium

Zaki, E.A., Abdel Ghany, Research Area Borg El Arab, Alexandria, EGYPT

Eukaryotic genomes harbor mobile genetic elements known as long terminal repeat (LTR) retrotransposons. LTR retrotransposons are closely related to the infectious and endogenous retroviruses, and they are collectively referred as LTR retroelements. The envelope (env) gene of the retroviruses, which is being responsible for their infective properties, distinguishes them from the LTR retrotransposons. We have previously reported plant retroelements with env-like genes in Gossypium. Here, we report the detection of env-encoded transcripts. The detection of env-encoded transcripts in the cotton genome promotes the initiative to employ these retroelements as a DNA vector for engineering novel cotton genotypes. A vector based on gypsy-like retrotransposons could be an important additional tool for the production of transgenic cotton plants with well-defined, foreign DNA inserts required for biosafety approval and commercialization.

Using SAGE to Explore the Genome

Saurabh Saha, Alberto Bardelli, Kenneth W. Kinzler, & Bert Vogelstein, Johns Hopkins Medical Institutions, The Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD

With the recent completion of the draft human genome sequence, the SAGE method is now uniquely poised to provide a bridge between the transcriptome and the genome. New modifications to the SAGE method have permitted the analysis of gene expression in cell subpopulations or microanatomic structures, providing access to unexplored transcriptomes of normal and disease biology. To gain insights into the molecular basis for colorectal cancer, we have used SAGE to compare the global gene expression profile of metastatic colorectal cancers, primary cancers, benign colorectal tumors, normal colorectal epithelium, and endothelium from normal and tumor tissues. Genes differentially expressed in these tissues may be useful as diagnostic or therapeutic targets and provide insights into the pathogenesis of colorectal cancer. Among the genes identified, the PRL-3 protein tyrosine phosphatase gene was of particular interest as it was expressed at high levels in each of 50 cancer metastases studied but at lower levels in non-metastatic tumors and normal colorectal epithelium. In a subset of metastases examined, multiple copies of the PRL-3 gene were found within a small amplicon located at chromosome 8q24.3. These data suggest that the PRL-3 gene is important for colorectal cancer metastasis and provides a new therapeutic target for these intractable lesions.

Breaking the Bottle Neck of Genomics and Proteomics

Moncef Jendoubi, Milagen, Inc., Richmond, CA

While the genomics and proteomics fields are actively expanding, there is a need for the biotech industry to find new ways to detect, diagnose, understand and treat human diseases. Milagen has developed a unique antibody-based approach to directly correlate gene products to disease. Based on proprietary technologies globally referred to as ANTIBIOMIX TM, Milagen has generated up to 61,000 high affinity specific polyclonal antibodies to known and unknown human gene products, and is set up to generate antibodies to all human proteins. Within our oncology program, we have used antibody libraries to screen hundreds of matching normal and tumor protein samples, derived from cancer patient tissues and biological fluids, using our proprietary matrix protein array technology. When applied to colon, prostate and breast cancer, dozens of antibodies tested showed a differential protein expression profile between normal and cancer samples. Several validated targets have been identified so far, including secreted molecules in sera. With ANTIBIOMIX TM, antibodies are serving as tools to bridge the gap between the genotype (gene sequence) and the phenotype (disease state). This approach will lead to the rapid identification of a variety of disease specific proteins, and to the validation of novel targets for diagnostic and therapeutic applications.

A Novel 22,000 Feature In Situ-Synthesized 60-Mer Oligonucleotide Microarray: Expression Profiling Of E12.5 Mouse Embryo and Placenta

Mark G. Carter[1], Condie E. Carmack[2], Yong Qian[1], Toshio Hamatani[1], Pius Brzoska[2], S. Stuart Hwang[2], and Minoru S.H. Ko[1], [1]Developmental Genomics and Aging Section, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health (NIH), Baltimore, MD [2]Agilent Technologies, Palo Alto, CA

Inkjet-based in situ 60-mer oligo synthesis technology has opened the door for more rapid and flexible microarray design without the labor-intensive clone handling required by cDNA array production. However, for downstream experimental work on genes implicated by microarray experiments, it is essential to have the collection of readily available cDNA clones corresponding to these oligos. To this end, we have designed and constructed an in situ-synthesized microarray representing approximately 21,000 unique genes, at least 98% of which correspond to clones in the unique cDNA collections of NIA (see the abstract by Ko et al. in this meeting). As an initial performance assessment of this microarray, we have generated expression profiles of mouse E12.5 embryo and placenta and compared them to similar profiles previously published using the NIA 15K microarray platform. For over 11,000 unique genes represented on both platforms, the oligo platform measured over six times more statistically significant expression changes than the cDNA array, with more stringent statistical criteria. Twenty percent of genes which gave little or no signal in one or more channels using the cDNA array produced reliable measurements in the oligo system. Similarities and differences between the two data sets are being examined in detail.

Back to the Future: Integrating Expression with Genomic, Genetic, and Metabolic Data

John Quackenbush, The Institute for Genomic Research, Rockville, MD

The advent of genome sequencing has produced us with a "parts list" comprised of the predicted genes for humans and other organisms. The ultimate goal of a genome project, however, is the elucidation of genes, their functions, and the metabolic pathways that underlie cellular metabolism and this remains a significant challenge. Most predicted genes cannot even be accurately assigned to functions. DNA microarray technology, developed in recent years, allows the generation of gene expression data on a genome-wide scale. While this technology promises information that can be used for deducing gene function, to date its application has been limited. However, when expression profiling is combined with genomic sequence data and traditional genetic approaches, the result can be a powerful new tool for assigning function to novel genes and identifying those that may be involved in human disease or other important biological processes.

ArrayExpress - A Public Database for Microarray Data

Slides: PDF version

Helen Parkinson, European Bioinformatics Institute, Cambridge, UNITED KINGDOM

The handling and analysis of the huge amounts of microarray data is becoming the major bottleneck in microarray analysis. Storing and annotation of microarray data is a non trivial problem due to the size, complexity and of the datasets. The raw image data presents particular storage problems and the annotation and integration of sample and array annotation with existing genomic data is essential for maximum interpretability of the data.

The EBI as is a member MGED (Microarray Gene Expression Database Group) which is involved in an international effort to standardise the way that microarray data are represented with guidelines for which information should be supplied with the microarray data (www.mged.org). This standard is called MIAME (Minimum Information About a Microarray Experiment (Brazma et al., 2001).

The EBI is establishing a database - ArrayExpress, to store MIAME compliant data (Brazma et al., 2000). An object model -MAGE-OM, for ArrayExpress has been developed and submitted to the OMG (Object Management Group) for acceptance as an specification for expression data. An XML data exchange format, MAGE-ML has been generated from the model. Datasubmissions to ArrayExpress can be made in MAGE-ML format or via the submission tool MIAMExpress, a web based tool under development at the EBI (http://www.ebi.ac.uk/miamexpress/) for submission of MIAME compliant data.

A Novel Platform for High-Sensitivity Analysis of Cell-Specific Gene Expression using Microarrays

Rajiv Raja, Jim Stanchfield, Mark Erlander and Steve Kunitake, Arcturus, Inc., Mountain View, CA

Gene expression profiles of thousands of genes can be monitored in parallel using microarrays. Microarray technology has proven to be a valuable tool in studying normal and induced variations in gene expression, thereby helping to understand the molecular mechanism of diseases, and aiding in disease diagnoses and drug discovery. However, whole tissue biopsies are typically used for these studies because microgram amounts of RNA are required for performing microarray hybridizations. Hence, measurements of molecular signals are averaged across several dissimilar cell types, resulting in a decrease in assay sensitivity. A new platform has been recently developed by Arcturus that attains improved sensitivity at 3 stages: (1) capturing pure cell populations through laser capture microdissection (LCM) while maintaining integrity of cellular RNA, (2) efficiently recovering good quality RNA from very small samples through an optimized RNA isolation system, and (3) faithfully amplifying messages from cellular RNA to provide adequate amounts of amplified antisense RNA (aRNA) for microarray analysis. First, thin tissue sections are prepared from biopsies and are fixed on microscope slides using a protocol optimized for preserving RNA quality during the process. Using laser capture microdissection (LCM), small populations of specific cell types are quickly isolated from these sections. Miniaturized devices are then used to extract and purify RNA from the captured cells in both high quality and yield. Nanogram amounts of the recovered RNA are then amplified to produce microgram quantities of aRNA using a novel linear amplification process. aRNA is then labeled and hybridized to microarrays. When the whole platform is used for gene expression analysis, considerably higher sensitivity in quantifying differential expression is attained compared to traditional approaches. Moreover, the system is capable of revealing signatures that are not seen when whole tissues are assayed. When differential expression profiles revealed by unamplified samples are compared to those from amplified samples, we see a very high correlation of $r = 0.91$. Labeled cDNA prepared from aRNA is routinely used in our laboratories to probe microarrays for identification of differentially expressed genes. Using this platform, molecular signatures have been identified using as few as 250 cells from frozen breast cancer biopsies, suggesting direct application to clinical chemical analyses.