DOE/ER/61490.1

✓
Conference - Transcribed Sequences in the Human Genome
to be held in San Francisco, November 7 & 8, 1992

✓
Final Report

for Period September 1, 1992 - August 31, 1993

Katheleen Gardiner

Eleanor Roosevelt Institute
1899 Gaylord Street
Denver, CO 80206

November 1993

Prepared for
THE U.S. DEPARTMENT OF ENERGY
AGREEMENT NO. DE-FG02-92ER61490

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED 875

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product or process disclosed or represents that its use would not infringe privately-owned rights.

Grant Title: Conference - Transcribed Sequences in the Human Genome to be held
in San Francisco, November 7 & 8, 1992

Abstract

The systematic isolation of transcribed sequences is critically important for solving disease related problems, and for understanding human genome organization as a whole. Although great progress has been, and continues to be made in adding detail to genetic and physical maps, most disease gene isolation efforts still require scanning hundreds of kilobases for expressed sequences and candidate genes. Many comprehensive gene searches have been launched in large chromosomal regions, and are resulting in the development of new strategies and approaches, and in the large scale application of both new and old techniques. For the most rapid and efficient construction of transcriptional maps, and the identification of important disease genes, discussions of the relative merits of the various approaches, by the practitioners of these approaches, is required.

Final Report

The Second International Workshop on the Identification of Transcribed Sequences was held in San Francisco on November 7-8, 1992. The purpose of the workshop, which was attended by 44 scientists from 7 countries, was to discuss and evaluate techniques for developing a complete transcriptional map of the human genome.

Such a map requires the positions, sequences, and expression patterns of all genes. This goal is being approached from two different directions, each with strengths and weaknesses. One method is to identify the transcribed sequences from genomic DNA of a given region; the other is to systematically sequence and map cDNAs. The cDNA approach yields sequence information rapidly, but mapping each cDNA is a technical challenge. In the first approach, the map locations of genomic sequences are known at the outset, and the challenge is to identify exons. The efficient construction of a transcriptional map will require a diverse array of techniques.

All participants agreed that this annual workshop should continue, maintaining the informal workshop style of oral presentations and discussions. A report of the workshop was published in Human Genome News, May 1993 (copy enclosed). Also enclosed is the workshop notebook, containing the agenda, abstracts presented and list of attendees.

Human Genome news

National Center
for Human
Genome Research

This newsletter is intended to facilitate communication among genome researchers and to inform persons interested in genome research. Suggestions are invited.

Human Genome Management Information System:

Managing Editor
Betty K. Mansfield

Editors/Writers
Anne E. Adamson
Denise K. Casey
Kathleen H. Mavournin

Production Manager/Editor
Judy M. Wyrick

Production Assistants
K. Alicia Davidson
Larry W. Davis
Sheryl A. Martin
Laura N. Yust

**Correspondence
Address:**
Betty K. Mansfield
ORNL
P.O. Box 2008
Oak Ridge, TN 37831-6050
Phone: 615/576-6669
Fax: 615/574-9888
BITNET: bkq@ornlsc
Internet: bkq@ornl.gov

Sponsor Contacts:

Daniel W. Drell
DOE Program Office
Germantown, MD 20545
301/903-6488
Fax: -8521
Internet:

#genome%er@mailgw.er.doe.gov

Leslie Fink
NIH National Center for
Human Genome Research
Bethesda, MD 20892
301/402-0911
Fax: 301/480-1950
Internet: lsf@cu.nih.gov

Second Transcribed Sequences Workshop

The DOE-sponsored Second International Workshop on the Identification of Transcribed Sequences was held in San Francisco on November 7–8, 1992. The purpose of the workshop, which was attended by 44 scientists from 7 countries, was to discuss and evaluate techniques for developing a complete transcriptional map of the human genome.

Such a map requires the positions, sequences, and expression patterns of all genes. This goal is being approached from two different directions, each with strengths and weaknesses. One method is to identify the transcribed sequences from genomic DNA of a given region; the other is to systematically sequence and map cDNAs. The cDNA approach yields sequence information rapidly, but mapping each cDNA is a technical challenge. In the first approach, the map locations of genomic sequences are known at the outset, and the challenge is to identify exons. The efficient construction of a transcriptional map will require a diverse array of techniques.

cDNA Sequencing and Libraries

Charles Auffray (Genethon, France), James Sikela (University of Colorado), and Mihael Polymeropoulos [NIH National Institute of Mental Health (NIMH)] presented large-scale, partial cDNA sequencing results. A number of methods are being used to integrate these sequences with physical or genetic maps. Minoru Ko (Wayne State University) used polymorphic sequences in 3' untranslated regions of mouse cDNAs as genetic markers in interspecific mouse crosses. Polymeropoulos employed polymerase chain reaction (PCR) primers with a panel of hamster-human cell hybrids to assign human cDNAs to chromosomes. Sikela reported use of fluorescence in situ hybridization to map some cDNA sequences.

The unequal representation of clones in cDNA libraries poses another problem for the cDNA sequencing approach. With 10,000 to 20,000 genes expressed in a given tissue, their unequal representation would require the sequencing of 1 million clones to approach complete representation for a tissue. An important goal is to achieve equal representation of both abundant and rare species in cDNA libraries (i.e., a "normalized" library). Bento Soares (Columbia University) has produced normalized cDNA libraries by reannealing cDNA at a high C_{ot} value and

selectively cloning the remaining single-stranded fraction. Cheng Chi Lee [Baylor College of Medicine (BCM)] proposed normalizing arrayed cDNA libraries by rescreening with pools of clones already sequenced to eliminate redundancies.

Another approach to defining a minimal set of cDNA clones is the identification of each clone according to its hybridization "fingerprint" against a panel of oligonucleotides. Radoje Drmanac (Argonne National Laboratory) and Sebastian Meirer-Ewert (Imperial Cancer Research Fund) reported progress in rapidly screening cDNA libraries with short oligonucleotide probes.

cDNA libraries from rare cell types are necessary for identifying the genes expressed in these cells. Barbara Knowles (Wistar Institute) has produced cDNA libraries from specific stages of early mouse embryos. James Eberwine (University of Pennsylvania) has developed techniques for producing cDNA libraries from single cells. His method relies on amplification by prokaryotic RNA polymerases rather than on PCR amplification, which can severely skew representation in heterogeneous mixtures.

Genomic Reference Libraries and Exon Identification

While mapping and normalization are challenges in the cDNA sequencing approach, the genomic approach requires identification of transcribed sequences among the surrounding untranscribed sequences. Genomic DNA for this strategy is made available by advances in physical mapping, especially large arrayed cosmid and lambda libraries from specific chromosomal regions and yeast artificial chromosome (YAC) clones. Ute Hochgeschwender (NIMH), Anne Marie Poustka (German Cancer Research Center), and Geoffrey Falk (Scripps Research Institute) reported using labeled cDNA for probing large arrayed genomic libraries to identify transcribed sequences. This technique is suited for very large regions covered by arrayed libraries; about one-half the genes in a given tissue are expressed at levels high enough to be detected with these probes. Furthermore, expression patterns can be determined by repeated cDNA probe hybridizations of different tissues at different developmental stages.

Meeting Reports

Exon Identification by Computer

Richard Mural [Oak Ridge National Laboratory (ORNL)] and Gordon Hutchinson (Canadian Genetic Disease Network) have developed software that can identify coding exons from genomic sequences with over 80% success and a low false-positive rate. An additional feature of the ORNL GRAIL program is a gene-assembly program (GAP) that produces complete gene sequences from the GRAIL output, resulting in an even lower false-positive rate. Susan Berget (BCM) discussed her work on splice-site selection, which is based on defining the nucleic acid "signals" associated with splicing. David Searls (University of Pennsylvania) presented a linguistic approach to sequence analysis based on conceptual similarities between meaningful language and expressed sequences.

Direct Exon Identification

Exons can also be identified by procedures in which a genomic fragment is cloned into an intron of a mammalian expression vector. A fragment containing an exon will be included in the mRNA expressed from the vector. Alan Buckler [Massachusetts General Hospital (MGH)] reported improvements to the splicing vector pSPL1. Paul Nisson (Gibco BRL/Life Technologies, Inc.) discussed a potential problem in transforming pools of genomic fragments cloned in splicing vectors: unequal exon-amplification rates in RT/PCR reactions can lead to the loss of all but the most efficiently amplified exons. This problem is also encountered in amplifying cDNA libraries. Geoffrey Duyk (Harvard Medical School) and Nicole Datson (Leiden University, Netherlands) reported on modifications of exon-trapping systems. While most exon-identification protocols are designed to detect internal exons, David Krizman (BCM) described a new system for detecting 3' terminal exons in which polymorphic sequences in the 3' untranslated regions may facilitate exon mapping.

David Kurnit (University of Michigan) described his system for identifying exons cloned in a plasmid vector by their ability to recombine homologously with cDNA clones in lambda. The system detects the exon, isolates the cDNA, and (when tested with multiple cDNA libraries) supplies information on expression patterns.

Another general approach to identifying transcribed sequences from genomic clones involves hybridizing cDNA to genomic DNA and eluting the specifically bound "selected"

cDNA. Michael Lovett (University of Texas), Poustka, Danilo Tagle (University of Michigan), and Sherman Weissman (Yale University) presented results using variations of this procedure with pools of cosmid clones and with whole YAC DNA. To map these genes within the YAC, Ruchira DasGupta (Albert Einstein College of Medicine) is using cDNAs isolated by this method and cloned in a yeast vector to truncate the YAC at the region of homology.

Perspectives

The usefulness of the different approaches can best be assessed by applying them to large regions. Gail Bruns (Children's Hospital, Boston) reported searching through the 15-Mb WAGR region for conserved sequences associated with HTF (*Hpa* II twin fragment) islands. Bernhard Weber (University of British Columbia) used a cDNA selection strategy and single-strand-conformation polymorphism technique to isolate new genes from the Huntington's disease region and to search these for mutations.

Giorgio Bernardi (Institut Jacques Monod, France) and Katheleen Gardiner (Eleanor Roosevelt Institute) discussed long-range differences in genome structure. While gene density is highest in regions richest in GC (in particular, most telomeres), one-third of genes are in AT-rich regions. Gene content as measured by CpG-island density does not appear to vary with GC content in a predictable manner. The applicability of techniques to the gene-poor regions was raised, and Mural and Hutchinson requested sequences of verified AT-rich exons for teaching neural networks.

The functional analysis of products encoded by novel genes was discussed. Roger Brent (MGH) described a yeast interaction trap system for identifying protein regions that interact in vivo. Miles Brennan (NIMH) reported on the specific integration of a mammalian cDNA at a homologous gene in yeast, a system that may allow direct selection and functional analyses of such cDNAs. The need for more functional assays was noted by Weissman.

*Reported by Miles B. Brennan
National Institute of Mental Health
and
Katheleen Gardiner
Eleanor Roosevelt Institute*

The Third International Transcribed Sequences Workshop is planned for October 2-4 in New Orleans. Contact: Ute Hochgeschwender; Unit on Genomics, NIMH; Bldg 10, Room 4N 320; Bethesda, MD 20892 (301/402-1769, Fax: -2140).

2nd International Workshop on the Identification of Transcribed Sequences

**San Francisco, CA USA
November 7-8, 1992**

Organizers

**Katheleen Gardiner, Ph.D.
Eleanor Roosevelt Institute**

**Miles Brennan, Ph.D.
NIMH**

**Sponsored by the
Department of Energy**

CONTENTS

Schedule	p. 1
Abstracts	p. 3
Roster of Participants	p.38

SCHEDULE

SATURDAY

8:00-8:30 **Introduction** Katheleen Gardiner

8:30-10:30 **Informatics** Richard Mural, Chair

Speakers

 Susan Berget p. 5

 Jean-Michel Claverie p. 11

 Radoje Drmanac p. 14

 Gordon Hutchinson p. 21

 Richard Mural

 David Searls p. 32

10:30-11:00 Break

11:00-12:30 **Techniques I** Ute Hochgeschwender, Chair

Speakers

 Roger Brent p. 7

 Alan Buckler

 Nicole Datson p. 13

 Geoffrey Duyk p. 15

 James Eberwine p. 16

 Barbara Knowles p. 22

 David Krizman p. 24

 Paul Nisson p. 29

12:30-1:30 Lunch

1:30-3:30 **Techniques I continued**

Techniques II Annemarie Poustka, Chair

Speakers

 Kuniya Abe p. 3

 Miles Brennan p. 6

 Anthony Brookes p. 8

 Ruchira Das Gupta p. 12

 Ute Hochgeschwender p. 20

 David Kurnit p. 25

 Marcelo Bento Soares p. 34

 Danilo Tagle p. 35

 Sherman Weissman p. 37

3:30-4:00 Break

4:00-6:00 **Techniques II continued**

SUNDAY

8:00-10:30	Regions	Sherman Weissman, Chair
	<u>Speakers</u>	
	Giorgio Bernardi	
	Gail Bruns	p. 9, 10
	Jeffery Falk	p. 17
	Katheleen Gardiner	p. 18
	Michael Higgins	p. 19
	Greg Lennon	
	Michael Lovett	p. 28
	Annemarie Poustka	p. 31
	Bernhard Weber	p. 36
10:30-11:00	Break	
11:00-12:30	cDNA Sequencing	Miles Brennan, Chair
	<u>Speakers</u>	
	Charles Auffray	p. 4
	Minoru Ko	p. 23
	Ching Chi Lee	p. 25
	Hans Lehrach	p. 27
	Kousaku Okubo	
	Michael Polymeropoulos	p. 30
	James Sikela	p. 33
12:30-1:30	Lunch	
1:30-2:00	cDNA continued	
2:00-3:00	Session Discussions	
3:00-4:30	Open Discussion	

THE DISTRIBUTION OF GENES IN THE HUMAN GENOME

Girogio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod,
2, Place Jussieu, 75005 Paris, France.

Vertebrate genomes are mosaics of isochores, namely of long (>300 Kb), compositionally homogeneous DNA segments that can be subdivided into a small number of families characterized by different GC levels. In the human genome, the compositional range of isochores is 30% to 60% GC, and five families of isochores have been identified : two GC-poor families, L1 and L2, representing together 62% of the genome, and three GC-rich families, H1, H2 and H3, representing 22%, 9% and 3%, respectively; (the remaining 4% of the genome are formed by satellite and ribosomal DNA).

The compositional distributions of large genome fragments, of exons (and their codon positions), and introns are correlated with each other. Knowing the compositional correlation between coding sequences and isochores, allows one to reconstruct gene distribution in the genome. This has shown that gene concentration is strikingly non-uniform in the human genome, being highest in the H3 isochore family, lowest in the L1 + L2 families, and intermediate in the H1 + H2 families. Moreover, different isochore families have different chromosomal locations. The H3 family corresponds to T(elomeric) bands of metaphase chromosomes, and the L1 + L2 families mainly to G(iemsa) bands, whereas R(everse) bands comprise both GC-poor and GC-rich isochores.

Bernardi G. The isochore organization of the human genome.

Ann. Rev. Genet., 23 637-661 (1989)

Gardiner K., Aïssani B., Bernardi G. A compositional map of human chromosome 21. EMBO J., 9 1853-1858 (1990)

Aïssani B., Bernardi G. CpG islands : features and distribution in the genome of vertebrates. Gene, 106 173-183(1991)

Aïssani B., D'Onofrio G., Mouchiroud D., Gardiner K., Gautier C., Bernardi G. The compositional properties of human genes. J. Mol. Evol., 32 497-503 (1991)

D'Onofrio G., Mouchiroud D., Aïssani B., Gautier C., Bernardi G. Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. J. Mol. Evol., 32 504-510 (1991)

Mouchiroud D., D'Onofrio G., Aïssani B., Macaya G., Gautier C., Bernardi G. The distribution of genes in the human genome. Gene, 100 181-187 (1991)

De Sario A., Aïssani B., Bernardi G. Compositional properties of telomeric regions from human chromosomes. FEBS Letters, 295 22-26 (1991)

Saccone, S., De Sario, A., De Valle, G., Bernardi, G. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA 89 4913-4917 (1992)

THE DISTRIBUTION OF GENES IN THE HUMAN GENOME

Girogio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod,
2, Place Jussieu, 75005 Paris, France.

Vertebrate genomes are mosaics of isochores, namely of long (>300 Kb), compositionally homogeneous DNA segments that can be subdivided into a small number of families characterized by different GC levels. In the human genome, the compositional range of isochores is 30% to 60% GC, and five families of isochores have been identified : two GC-poor families, L1 and L2, representing together 62% of the genome, and three GC-rich families, H1, H2 and H3, representing 22%, 9% and 3%, respectively; (the remaining 4% of the genome are formed by satellite and ribosomal DNA).

The compositional distributions of large genome fragments, of exons (and their codon positions), and introns are correlated with each other. Knowing the compositional correlation between coding sequences and isochores, allows one to reconstruct gene distribution in the genome. This has shown that gene concentration is strikingly non-uniform in the human genome, being highest in the H3 isochore family, lowest in the L1 + L2 families, and intermediate in the H1 + H2 families. Moreover, different isochore families have different chromosomal locations. The H3 family corresponds to T(elomeric) bands of metaphase chromosomes, and the L1 + L2 families mainly to G(iemsa) bands, whereas R(everse) bands comprise both GC-poor and GC-rich isochores.

Bernardi G. The isochore organization of the human genome.

Ann. Rev. Genet., 23 637-661 (1989)

Gardiner K., Aïssani B., Bernardi G. A compositional map of human chromosome 21. EMBO J., 9 1853-1858 (1990)

Aïssani B., Bernardi G. CpG islands : features and distribution in the genome of vertebrates. Gene, 106 173-183(1991)

Aïssani B., D'Onofrio G., Mouchiroud D., Gardiner K., Gautier C., Bernardi G. The compositional properties of human genes. J. Mol. Evol., 32 497-503 (1991)

D'Onofrio G., Mouchiroud D., Aïssani B., Gautier C., Bernardi G. Correlations between the compositional properties of human genes, codon usage and aminoacid composition of proteins. J. Mol. Evol., 32 504-510 (1991)

Mouchiroud D., D'Onofrio G., Aïssani B., Macaya G., Gautier C., Bernardi G. The distribution of genes in the human genome. Gene, 100 181-187 (1991)

De Sario A., Aïssani B., Bernardi G. Compositional properties of telomeric regions from human chromosomes. FEBS Letters, 295 22-26 (1991)

Saccone, S., De Sario, A., Della Valle, G., Bernardi, G. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA 89 4913-4917 (1992)

Identification of gene sequences using exon amplification.

¹Deanna M. Church, ¹Marcy E. MacDonald, ²David E. Housman, ¹James F. Gusella, and ¹Alan J. Buckler.

¹Molecular Neurogenetics Unit, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, MA 02129, and ²Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139

Characterization of the human and other model genomes requires the ability to establish the functional significance of DNA sequences being cloned and organized during the intensive genetic and physical mapping studies that are ongoing. One approach is to identify the structural genes contained within this complex collection of information. The development of exon amplification allows for rapid and efficient identification and cloning of coding sequences from complex sources of mammalian genomic DNA. These sequences are isolated by virtue of selection for both functional 5' and 3' splice sites in the genomic sequence being analyzed.

Isolation and characterization of a large number of products derived from human genomic DNA was required in order to determine the overall effectiveness of exon amplification. To this end, we have analyzed cosmids derived from a variety of human chromosomal sources. The results of these studies indicate that, after elimination of an easily identifiable class of false positives, the vast majority of cloned products derived from exon amplification experiments are authentic coding sequences and all but a small fraction are extremely useful as hybridization probes.

To further enhance the efficacy of the method we have constructed modifications of the vector, pSPL1, that increase its sensitivity, eliminate the major class of false positives, and further increase the sensitivity of the approach. Our studies indicate that the modifications to exon amplification are enabling for a rapid, highly reliable approach to gene identification and the method is presently adaptable to large scale analyses.

COMBINING NEURAL NETWORKS AND EXPERT SYSTEMS TO IDENTIFY FEATURES IN DNA SEQUENCES

Richard J. Mural¹, Xiaojun Guan², J. Ralph Einstein² and Edward C. Uberbacher². ¹Biology and ²Engineering Physics and Mathematics Divisions, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077.

We have developed an on-line e-mail service, GRAIL, which locates the protein coding regions of DNA sequences. This interface utilizes a multiple sensor-neural network (Uberbacher and Mural, 1991, PNAS 88:11261-11265) to find coding regions and a rule based interpreter to reduce this output to a table which includes other useful information such the identity of the coding strand and the preferred reading frame of the hypothetical coding exon. GRAIL can analyze up to 100 kbp of sequence at a time and several sequences may be included in a single e-mail message. The analysis is performed on both strands of the input sequence and the interpreter assigns the putative coding region to a preferred strand and reading frame. In our experience the strand and reading-frame assignments are accurate about 95% of the time. GRAIL continues to find 90% of coding exons over 100 bases in length with a very low false positive rate. Currently GRAIL has about 500 users world-wide. Additional modules, which determine the position and quality of potential splice junctions, and the positions of potential translation-initiation sites, are being added to GRAIL. A Gene Assembly Program (GAP) is being developed which combines the outputs from the various GRAIL modules and attempts to predict the sequence of the spliced mRNA from the genomic DNA sequence. If error conditions are detected by the GAP rule-base, then the program attempts to correct them by the insertion and/or deletion of one or more coding exons. These actions result in a net improvement in gene-model prediction, particularly in the recognition and characterization of very short coding regions. (This research was supported by the Office of Health and Environmental research, United States Department of Energy, under contract DE-AC05-84OR21400 with the Martin Marietta Energy Systems, Inc.)

RAPID RETRIEVAL OF DESIRED SEQUENCES FROM LONE LINKER PCR-AMPLIFIED
cDNA MIXTURE: APPLICATION TO RECOVERY OF EXPRESSED SEQUENCES FROM
CLONED GENOMIC DNA MAPPED TO MOUSE T/t COMPLEX

KUNIYA ABE : INSTITUTE OF MOLECULAR EMBRYOLOGY & GENETICS,
KUMAMOTO UNIVERSITY SCHOOL OF MEDICINE, KUMAMOTO 862, JAPAN

As part of a long-term project on the developmental mutations mapped on the *T/t* complex, we have been conducting a detailed characterization of cloned genomic DNA ultimately to find mutated genes. The *T/t* complex is a large genomic region located in mouse chromosome 17, which contains a set of mutations affecting germ cell and early embryos. Molecular genetic analysis of the mutations thus requires construction of cDNA libraries corresponding to early stages of embryogenesis or germ cell development. Such an attempt has been hampered by the difficulty of obtaining sufficient material: for instance, theoretically, at least 7,000 blastocysts should be dissected out to make one representative library. Searching for embryo-expressed genes in a large genomic region thus remains a cumbersome task, requiring repeated, large-scale screening of libraries with numerous probes derived at every walking step.

Recently the techniques that amplify DNA sequences tagged by "universal" primer sequences through PCR have been reported. We found that this modified PCR method could amplify each constituent of a highly complex mixture of DNA fragments such as mouse genomic DNA or embryonic cDNA, and termed it lone linker PCR (LL-PCR). A novel method coupling LL-PCR and the biotin-avidin technique for hybrid capture after solution hybridization was also devised. This "screening in solution" procedure, combining the efficiency of regular PCR and the versatility of conventional screening, made it possible to enrich for genes of interest directly from cDNA mixtures, skipping large scale screening steps. Here it has been applied to recovering expressed sequences from cloned genomic DNA mapped to the K-region of mouse major histocompatibility complex.

Towards a catalogue of human gene transcripts: a pilot study on a T lymphocyte cDNA library.

Charles Auffray°, Marie-France Bachelot°, Ellen Bahrend°, Ghislaine Behar*, Florence Bois°, Christine Bouchier°, Lydie Bougueleret°, Philippe Callulier°, Corinne DaSilva°, Anne Degavre°, Marie-Dominique Desvignes*, Didier Duchemin°, Simone Duprat°, Muriel Durand°, Alexandra Kobetz°, Marie-Noelle Labrousse-Jumeau°, Bernard Lamy°, Anne Manoukian°, Regine Mariage-Samson*, Bernard Prudhon°, Geraldine Richier-Zella°, Dominique Robert°, Corinne Sebastiani-Kabatchis°, Esther Toselli*.
The Genexpress cDNA Program. CNRS* UPR 420 - VILLEJUIF and GENETHON° - EVRY (France).

A subset of 8,592 clones from a human T lymphoblastoid cell line CEM cDNA library in Bluescript have been analyzed by hybridization with a variety of unique (actin, tubulin, ribosomal, mitochondrial) and complex probes (poly AC, CEM cDNA, muscle cDNA). Double-stranded DNA templates were prepared from 4,271 clones and subjected to cycle sequencing with the fluorescent forward primer. Validated sequence files were generated by a set of automated routines that cut the vector sequence and the part of the sequence containing too many indeterminations and classify the clones in redundancy classes. Additional partial sequences were derived from the other end of the unique clones using the reverse primers. From the 5,886 partial sequences obtained, 2,987 were derived from 1951 unique clones. Of these, 2,336 derived from 1,529 clones which were either related to known sequences or unknown as revealed by similarity searches in the nucleic acid and protein databases, have been deposited in the EMBL database with immediate public access, as a working resource to the community. Data related to these clones (insert size, hybridization pattern with complex and unique probes) and maintained in a local database. Our efforts to obtain complete sequences and integrate results of cDNA studies together with the physical and genetic mapping programs of Genethon will be discussed.

Supported by AFM, CNRS, MRE and the EEC.

HOW TO FIND AN EXON. M. Niwa, D. Sterner, D. Stolow, M. Talerico, and S. Berget. Verna and Marrs McClean Dept. of Biochemistry, Baylor College of Medicine, Houston, Texas USA 77030.

Vertebrate exons are short. Internal exons are on average less than 150 nucleotides. 3' terminal exons are longer with an average between 600 and 700 nucleotides (Hawkins, 1988). Few internal exons exist which are larger than 300 nucleotides. Attempts to engineer large internal exons fail. Such exons are either completely excluded from processed mRNA or are rendered smaller through the activation of internal cryptic splice sites. In in vitro processing extracts, large exons cannot associate with splicing factors to form initial spliceosome complexes. These restraints suggest that the splicing machinery has a mechanism to determine exon length, and ignore those splice sites that would define an exon of inappropriate length.

Interestingly, there may not be a minimum size for internal exons, because exons of 3 nucleotides have been identified. Mini-exons, however, require special activating mechanisms. Two classes of mini-exons seem to exist: those that reside close to a neighboring exon and those that do not. We have examined the recognition of the former class and observe that inclusion of the mini-exon in the final mRNA requires both the nearby exon and its close proximity. We suggest that the mini-exon is initially recognized as part of an exon-intron-exon unit from which the intervening intron is subsequently recognized. Mini-exons also may contain special enabling sequences within the exon. A compilation of mini-exon sequences reveals the presence of an AG-rich motif that has been found to promote recognition of exons in a variety of systems (Shimura, Cooper, personal communication).

We have recently suggested that the exon is the initial unit of vertebrate spliceosome assembly and splice site selection. From this perspective, splicing factors concertedly recognize 5' and 3' splice sites across exons during initial precursor RNA assembly. Our present working hypothesis is that recognition of an exon is influenced by four factors: the strength of the 3' splice site, the strength of the 5' splice site, the length of the exon, and the presence of additional sequences within or adjacent to the exon. When one or more of these is sub-optimal, the others must be better than average to compensate.

This view of exon recognition is especially useful when considering differentially-utilized exons. Such exons are typically sub-optimal for one or more of the four features. Improving the splice site strength, exon length, or internal sequence frequently causes constitutive recognition of a differential exon. The present goal of our research is to understand the relationship between the four features in determining exon recognition.

SELECTION OF HOMOLOGOUS GENES IN YEAST AND MAN USING IN VIVO RECOMBINATION

Miles B. Brennan, Unit on Genomics, NIMH, Bethesda, MD 20892, USA

The goal of this project is the identification of pairs of genes in man and yeast (*S. cerevisiae*) which have retained significant sequence identity. Not only will these define families of evolutionarily conserved genes, determination of their functions will be facilitated by experiments in yeast.

Despite significant stretches of amino acid identities between pairs of yeast and human proteins, the corresponding stretches of nucleic acid sequences are frequently interrupted by mismatches. *In vitro* hybridization procedures, both cross-hybridization and PCR, are limited by the sensitivity of short hybrids to mismatch.

To test whether *in vivo* recombination in yeast is less sensitive to mismatch, I transformed yeast with an integrative plasmid carrying a rat cyclophilin cDNA clone. While the parental plasmid gave less than 1 transformant per 100 ng, the cyclophilin plasmid yielded ~100 transformants per 100 ng.

Current work includes the use of truncation vectors in place of integrative vectors and definition of the length and homology dependence of recombination.

This approach holds the promise for directly selecting yeast homologs of mammalian genes by transformation of yeast with pools of human cDNA clones.

Assigning function to cDNAs of unknown function with the interaction trap

Roger Brent, Jenö Gyuris, Russ Finley, Helen Chertkov

Department of Genetics, Harvard Medical School and Department of Molecular Biology,
50 Blossom Street, Massachusetts General Hospital, Boston, Massachusetts 02114

We have developed a number of transcription-based assays for protein protein interactions. The most effective, which is an implementation of a suggestion made by Stan Fields in 1989, is called the interaction trap.

In this technique, transcription of specially rigged reporter genes in yeast is used as a synthetic phenotype to detect, interactions between a DNA bound LexA fusion "bait" protein and other proteins, for example members of specially engineered cDNA expression libraries. We are using this technique with much success to identify new proteins involved in cell growth decisions in higher eukaryotes.

Unexpectedly, we have also found it possible to use interaction technology as a tool to study the mechanisms by which newly identified proteins may function. I will describe three such uses.

1) We use panels of closely-related protein baits in simple genetic assays to identify residues in the bait proteins that are contacted by interacting proteins. This information is often sufficient to make testable predictions about the function of the interacting proteins.

2) We use libraries of expressed cDNAs to identify proteins that are upstream and downstream of known proteins in genetic and metabolic pathways. For example, using substrates as baits, in many cases we can identify protein kinases, and we can use protein kinases to identify substrates.

3) Perhaps most important for a genome effort, we are attempting to use genetic assays to assign function to large numbers of newly-identified proteins by determining their ability to interact with members of very large panels of baits. Knowledge that a protein contacts a known bait, for example an protein phosphatase or DNA polymerase subunit, might give immediate clues to its function, and suggest that the cDNA encoding it should be sequenced rapidly; knowledge that a given unknown has a pattern of interactions that suggests it is the homolog of a known structural protein might, for example, dictate that determination of its sequence should be given a lower priority.

I will describe other applications if these are successful.

The Use of Coincident Sequence Cloning to Identify Transcribed Sequences.

A. J. Brookes, E. Storch, B. Stevenson, B. Coyle and D. J. Porteous. MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh.

Coincident Sequence Cloning (CSC) is an approach to genome analysis by which partially overlapping DNA resources are integrated so as to recover only their common fraction. The application of this strategy to genomic and cDNA mixture pairs should allow the rapid identification of transcribed sequences within extensive genomic contigs. To this end we are developing effective CSC methodologies and evaluating their use in gene identification studies.

The experimental basis of CSC is the formation of heteroduplex DNA between the input DNAs followed by its selective recovery. One approach we have developed involves heteroduplex directed end-ligation of synthetic oligonucleotides to coincident DNAs so enabling the subsequent amplification of these sequences by the polymerase chain reaction. This method has now been successfully applied to the identification of genes using genomic sequences from human chromosome region 11p15.

A second approach to the direct cloning of genes by integrating genomic and cDNAs has been described by several groups. The variations of this method all attempt to purify heteroduplex species on some type of physical support. We have now combined this principle with the ideas underlying our CSC method. Based upon model experiments we believe that the method we have produced can achieve a greater degree of specific enrichment of coincident DNAs than could any previous method. In collaboration with others we are now testing this strategy as a means for transcribed sequence identification. These studies are focussed upon genes expressed in the developing brain and shall utilise genomic contigs from regions that are known to be rearranged or mutated in various neurogenic disorders.

A Gene and Conserved Sequence Region Map of the WAGR Deletion Area of Chromosome 11p.

G.A.P. Bruns (1), A. Pöge (1), R. Eisenman (1), F. Schwartz (1), R. Neve (2), and M. Gessler (3). (1) Genetics Division, Children's Hospital, Boston, MA; (2) Molecular Neurogenetics Laboratory, McLean Hospital, Belmont, MA; (3) Institut für Humangenetik, Philipps-Universität, Marburg, Germany.

The chromosome deletions associated with WAGR syndrome (Wilms tumor, aniridia, genitourinary anomalies and mental retardation) frequently encompass the entire 11p13 Giemsa light band and extend into the adjacent 11p12 and/or 11p14 Giemsa dark bands. Although deletion of less than 1 megabase of 11p13 DNA is sufficient for the phenotype of Wilms tumor, aniridia and the characteristic genitourinary anomalies, many of the constitutional deletions encompass 4 to more than 10MB of DNA. The chromosomal region involved in these deletions exceeds 15MB in size. To provide a framework for analysis of the genes, rearrangements and organization of this area, the order and distance relationships of more than 60 WAGR region specific probes were defined on a detailed long range restriction map of 15.4MB extending from distal 11p12 to mid 11p14. Seventeen HFT islands, identifiable in genomic DNA, constitute well localized anchor points. Eight islands have been cloned (5 by sequential HTF island jumping in rare cutter jumping libraries, 2 from an island targeted library, and 1 by walking) with conserved sequences identified adjacent to 6 of them. One island demarcates the WT1 gene and a second, the human homolog of the murine PAX-6 gene. Seventeen other distinct highly conserved DNA sequence regions have been located within distal 11p12-mid 11p14 with our bank of probes. In addition to WT1 and AN2, 13 conserved sequence regions map between CAT and FSHB. Five of these are HTF island adjacent regions; one is near several deletion breakpoints; and another identifies a new gene with predominant expression in fetal brain. cDNAs have been isolated for at least 3 new genes. DNA sequencing of other areas with strong cross species homology suggests the WAGR deletion region may encode yet additional DNA binding proteins. Two loci specify 11p13 members of separate, as yet unidentified gene families. Only single conserved sequence regions were identified in distal 11p12 and proximal 11p14 with random probes and an HTF island targeted library. However, RAG1/RAG2 have recently been localized to a very proximal deletion region (likely within 11p12) and BDNF mapped distal to FSHB in 11p14 by others.

The multifaceted map of the WAGR deletion region developed in this study-based initially on an extensive DNA probe bank and a large fragment restriction map with subsequent integration of conserved sequence regions, cloned HTF islands, identified genes, and rearrangement breakpoints-provides a first look at the organization of a prototypic genomic region associated with a contiguous gene syndrome. Analysis of YACs for the CAT-FSHB region and particularly for the flanking 11p12 and 11p14 areas can be expected to detail the gene content and relationships in these areas.

Banks of Conserved Sequence Probes from Flow Sorted Chromosomes.

G.A.P. Bruns (1), R. Eisenman (1), A. Pöge (1), A. Flint (1,2) and M. Lalande (1,2). (1) Genetics Division and (2) Howard Hughes Medical Institute, Children's Hospital, Boston, MA.

Collections of conserved sequence probes from specific human chromosomes or chromosome regions would be useful resources for isolation of candidates for disease loci linked to the chromosome; for generation of chromosomal or regional transcription maps; perhaps for detection of genic regions within YAC or cosmid contigs; and when coupled with differential expression technologies, for identification of tissue or developmental stage specific genes on a chromosome. Building on the observation of Lüdecke et al (Nature 338:348-350,1989) that a high fraction of clones from libraries of microdissected chromosomal DNA digested with RsaI showed cross species homology, we have now generated small insert X-chromosome and chromosome 2 libraries that are enriched for conserved sequences.

RsaI digested DNA from flow sorted X chromosomes or chromosome 2 was ligated to EcoRV digested, CIP treated pBSIIKS+ DNA and aliquots used to transform XL1 Blue or TB1. More than 6000 clones were obtained for each library, at least 90% of which contain inserts. Insert sizes range up to 1.5kb but cluster between 200-700bp. Of the 25 RsaI clones analyzed from the chromosome 2 library, 80% hybridized as unique sequence probes without blocking repetitive elements. 44% of these evidenced high stringency cross species homology. Seventy percent of the 36 clones analyzed from the RsaI library of flow sorted X chromosomes hybridized as unique sequence probes. Forty percent of these identified homologous sequences in murine and/or bovine DNA. No duplicates have been identified among the 45 unique sequence probes analyzed from the two libraries. Open reading frames extending through all or much of the insert are present in most of the sequenced clones that exhibit strong cross species homology.

Gridding and hybridization techniques currently in use to rid the clone collections of low order and rare repeats should permit progressive enrichment of these RsaI libraries for conserved sequences.

Information enhancements methods for large scale sequence analysis.

Jean-Michel Claverie

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20894

The improved efficiency of similarity search programs and the affordability of ever faster computers allow studies where whole sequences databases can be the target of various comparisons with increasingly larger or numerous query sequences. However, the usefulness of those "brute force" methods now becomes limited by the time it takes an experienced scientist to sift the interesting biological results from overwhelming, however, "statistically significant" outputs. The discrepancy between statistically *versus* biologically significant results has different causes: errored database entries, the presence of repetitive elements in natural sequences, the ubiquity of *low complexity* segments with biased composition. I will introduce two masking methods (programs XNU and XBLAST) capable of eliminating most of the irrelevant outputs in the context of a variety of large scale sequence analysis projects: "all against all" database comparison, massive partial cDNA sequencing (EST), positional cloning and genomic data analysis. I will emphasize the application of those methods for the automatic annotation of EST data and the interpretation of large human genomic contigs from the Kallmann syndrome gene region (Xp22.3) and the HLA class III region (6p21.3). Both studies have led to the identification of new human genes.

AN EFFICIENT AND INTEGRATED METHOD FOR MAPPING HUMAN GENES

Ruchira DasGupta¹, Bernice Morrow¹, Ivonne Marondel¹, Satish Parimoo², Arthur Skoultschi³, Sherman Weissman² and Raju Kucherlapati¹. ¹Department of Molecular Genetics, ³Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY; ²Boyer Center for Molecular Medicine, Yale University School of Medicine, New Haven, CT.

During the past 25 years 3-4% of the human genes have been mapped to human chromosomes and a subset of them have been mapped with greater precision. It is necessary to have new methods to accelerate the pace of identifying and simultaneously mapping human genes. We have developed methods to isolate genes corresponding to large defined DNA segments and map them with very high precision to specific regions of that segment. The first step in our procedure involves construction of a highly representative human cDNA library. This library is expected to contain 75-80% of all genes encoded by the human genome. It is prepared by purification of mRNA from a large number of tissues and organs at different stages of development and construction of cDNA libraries from these RNAs by a random primed polymerase chain reaction (PCR). cDNAs corresponding to genes encoded by DNA cloned into yeast artificial chromosomes (YACs) are isolated by hybridization of the library of cDNA inserts to immobilized YAC DNA and amplification of these cDNAs which hybridize to the DNA. After cloning, the affinity purified cDNAs can be sequenced to generate expressed sequence tags and to construct cDNA contigs if necessary. Individual members of the cDNA packets or a mixture of the cDNAs corresponding to a single YAC are then cloned into yeast chromosome fragmentation vectors and used to transform the yeast bearing the specific YAC used for affinity purification. We have shown that individual exons within a cDNA are capable of recognizing their corresponding exons in the YAC and undergoing homologous recombination. The homologous recombination events result in truncations of the YAC at sites corresponding to individual exons. Since all of the truncation events end with the same vector sequence, it is possible to obtain DNA sequence from these ends which can serve as sequence tagged sites. This method permits gene isolation and deduction of exon-intron organization. Additionally the truncated YACs provide easy access to regulatory regions of the gene. If the YAC is derived from a defined region of the chromosome, all of the genes isolated by using that YAC as the substrate for hybridization can be readily assigned to a specific region of the chromosome.

SEARCHING FOR EXPRESSED SEQUENCES: EXON TRAPPING

Datson, N.A.¹, Duyk, den Dunnen, J.T.¹ and van Ommen, G.J.B.¹

¹Department of Human Genetics, Leiden University, Leiden, the Netherlands. ²Department of Genetics, Harvard Medical School, Boston, U.S.A.

One of our main interests is the p21-pter region of the human X-chromosome. In our efforts to identify genes in randomly cloned DNA from this region we have investigated the use of exon trapping. We have developed a new exon trapping method using the vector pETV-SD2. This vector contains a splice donor site upstream of a multiple cloning site with a LacZ α marker gene located in between. DNA from a region of interest is subcloned in this vector and the cloned DNA is then introduced into COS-1 cells, in which it is amplified, transcribed and spliced. RT-PCR with vector-specific primers on RNA isolated from the COS1 cells results in a series of PCR products which are circularized to create a functional plasmid and transformed into *E.coli* cells. Clones that contain an exon will have been spliced, resulting in the loss of the LacZ α -gene and giving rise to white colonies. This method is distinct from other methods published previously. Firstly, trapped exons are identified by blue/white selection instead of PCR. Secondly, trapped exons are cloned instantly by the inclusion of a circularization step of the RT-PCR products. This reduces much background from PCR artefacts and the tricky step of subsequent cloning of the RT-PCR products is no longer necessary. Initial experiments show that known exons from both a specifically cloned DNA fragment and a clone-library constructed from an entire cosmid are trapped successfully. A portion of the obtained clones are false positives and probably derived from a cryptic splice site in the vector. These clones can easily be distinguished by PCR with a primer pair of which one primer is located upstream of the cryptic splice site, and discarded. The obtained PCR products can be used for further analysis, e.g. screening cDNA libraries, probing 'zoo-blots' and Northern blots and direct sequencing. Currently we are subcloning interesting YAC and cosmid clones from Xp21-Xpter in pETV-SD2. One 176 bp exon, present in a cosmid derived from the human DMD-gene, could be detected in 15% of the screened white clones. After correction for the false positives by PCR this figure rose to 50%. Experiments are being performed to establish the nature of the other clones, which may belong to cryptic genes in the DMD-gene.

OBTAINING THE CATALOGUES, MAPS, AND EXPRESSION PATTERNS OF GENES USING 20-200 OLIGONUCLEOTIDES

R. Drmanac, Biological and Medical Research Division, Argonne National Laboratory, Argonne,
Illinois.

In the past five years, we have demonstrated the feasibility of zero-mismatch hybridization of oligonucleotides as short as 6-mers. Recently we have developed a hybridization data production line for scoring 8-16 probes per day on up to 100,000 densely arrayed cDNA or genomic clones, 1-2 kb in size. The technology has a variety of applications. By providing appropriate sets of clones and probes, three main issues concerning expressed sequences (inventory, positioning on the genomic DNA and tissue specificity) can be addressed for all genes in parallel. We are collecting data on human brain cDNA clones toward the first goal of fingerprinting 100,000 clones with 100 6- to 8-mer probes. Library requirements, optimization of the probe sets, and other aspects of the schemes for identification of genes and for revealing their expression patterns in the organismal development will be discussed.

Work supported by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract No. W-31-109-ENG-38.

Abstract

Transposon based Exon Trapping.

Geoffrey Duyk* and Helene Powers, Department of Genetics, Harvard Medical School and Howard Hughes Medical Institute. EQRF, 200 Longwood Avenue, Boston, MA 02115

Our laboratory has been interested in the development of exon trapping as a method for identifying candidate genes within cloned genomic DNA. Exon trapping embraces a number of strategies which utilize splicing signals as primary identifiers of potential exons. The focus of our recent work has been on methods to directly utilize large insert clones for exon trapping. Towards that end we have developed a number of transposon based systems for the introduction of exon trapping cassettes into novel cosmid vectors. This strategy can easily be adapted to include F plasmid based or P1 based systems as well. A potential advantage of this approach is that it generates resources facilitating construction of physical contigs and recovery of STRs. We have also been exploring the use of RNA amplification systems to help overcome some of the limitations of RT-PCR for the recovery of trapped exons. We anticipate that these technical improvements will help increase the applicability of exon trapping to positional cloning exercises.

Using Defined Anatomical Regions and Single Cells to Direct the Cloning of Expressed Sequences

Jim Eberwine
University of Pennsylvania

The characterization of expressed sequences usually requires information concerning site of synthesis. Indeed, site of expression can often be used to predict physiological function of novel sequences. Data will be presented describing *in situ* transcription and aRNA amplification technologies which have been used to clone mRNAs from defined anatomical regions of fixed tissue sections as well as from single live cells. This experimental approach facilitates the enrichment of tissue and cell type specific sequences. It is hoped that some of the novel sequences that have been isolated will yield cell-type specific promoter sequences for use in directing transgenic expression of heterologous RNAs. The use of this technology has permitted the characterization of a "molecular fingerprint" of expressed sequences in tissues and individual cells.

Identification of Expressed Sequences on Human Chromosome 9q32-34
Jeffrey D. Falk and J. Gregor Sutcliffe, Department of Molecular
Biology, The Scripps Research Institute, La Jolla, CA 92037.

We are currently involved in identifying genes on human chromosome 9q32-34 and elucidating their tissues of expression. We are particularly interested in identifying candidate genes for the disease idiopathic torsion dystonia (DYT1 9q32-34), as well as other neurologically important genes transcribed from this region. To accomplish this we are using a modification of the method of Hochgeschwender et al for detecting expressed sequences within chromosome-specific libraries. A collection of 3,000 human clones spanning the 9q32-34 region has been assembled from a library derived from hamster-human cell hybrid E6B, whose only human component is chromosome 9q32-34. Analysis of this collection has shown that it contains approximately three genome equivalents of 9q32-34, resulting in a 99% statistical probability that the DYT1 locus is contained within our collection. The collection is now being surveyed with cDNA probes from various brain regions and peripheral tissues to categorize the tissue expression patterns of the genes corresponding to each clone. Candidate genes for torsion dystonia are being assessed based on their expression patterns in dystonic target tissues.

Chromosome 21: YACs, genes and CpG islands.

18

**Katheleen Gardiner and Flora Tassone
Eleanor Roosevelt Institute for Cancer Research
Denver, Colorado**

An understanding of the distribution and organization of transcribed sequences within large chromosomal regions should aid in the construction of complete transcriptional maps. YAC clones and the physical map of chromosome 21 have been used to investigate the distribution of CpG islands relative to the known gene distribution. YACs covering 20% of the long arm of the chromosome have been mapped with rare restriction enzymes, revealing 46 potential CpG islands. Base composition and rare site frequencies imply that these islands are likely to be real (and therefore mark the 5' ends of genes), although many are methylated in the cell lines and tissues examined so far. While the distribution of these islands grossly parallels the observed, nonrandom, gene distribution, there are strong local clusterings, resulting in 70% of islands mapping to <15% of the gene-rich region.

Additional points to be discussed include:

- i) the distribution of newly identified genes that suggests that odd regions may harbor unusual classes of genes;**
- ii) possible patterns in physical proximity of related genes and;**
- iii) "problem YACs" and observations of deletions and duplications.**

SCREENING CHROMOSOME 11 YACs FOR CpG ISLANDS.

M. Higgins, N. Kuhn, C. Isaacs, S. Qin, J. Zhang,
S. Nagafuchi, and T. Shows, Roswell Park Cancer
Institute, Buffalo, NY.

We have constructed a chromosome 11 YAC library from a somatic cell hybrid containing chromosome 11 as its only human component. These YACs range in size from about 100kb to 600 kb with an average length of 330 kb. Since it has been estimated that up to 50% of mammalian genes are associated with CpG islands, we are identifying those YACs containing these landmarks as a first step in uncovering new genes on chromosome 11. YAC DNA prepared in agarose plugs was digested with BssHII, SacII, EagI, NarI, NaeI, and SmaI and then separated by pulsed field gel electrophoresis. Digestion products were detected by Southern hybridization using human Cot1 DNA. A YAC was considered to contain a CpG island if it showed a similar restriction pattern following digestion with at least two enzymes. On this basis, approximately 40% of the YACs tested to date were found to contain at least one CpG island. These YACs are, therefore, good candidates for further analysis in a search for transcribed sequences.

IDENTIFICATION OF TRANSCRIBED SEQUENCES BY DIRECT cDNA SCREENING OF GENOMIC LIBRARIES

Ute Hochgeschwender, Unit on Genomics, NIMH, Bethesda, MD 20892, USA

Direct cDNA screening of arrayed genomic libraries is a method we developed for systematically identifying transcribed sequences in genomic DNA (Hochgeschwender et al., PNAS 86:8482-8486, 1989).

The genomic libraries are permanently arrayed collections of either phage or cosmid clones. The material for the libraries can be total genomic DNA, DNA from specific chromosomes or from parts of chromosomes (YACs).

The cDNA probes are made from mRNA of any tissue at any developmental time; further, they can be made from the same species or from other related species. Highly repeated sequences are removed from the cDNA probes by prehybridizing the probes with a 1000-fold excess of genomic DNA coupled to finely divided cellulose.

To proceed from a phage or cosmid clone carrying a transcribed sequence to the corresponding cDNA clone, small fragments containing transcribed sequences are identified by probing Southern blots of the digested phage or cosmid clones again with cDNA probes. These fragments are then used to isolate cDNA clones by probing an appropriate cDNA library.

The application of this approach to the screening of arrayed genomic libraries allows the rapid and systematic identification and isolation of expressed sequences and the determination of their patterns of expression. The approach is amenable to large-scale applications in identifying transcribed sequences in the human genome.

New Computer Techniques used in the Analysis of Transcribed Sequences in the Region of the Huntington Disease Gene

Gordon Hutchinson and Michael R. Hayden

ABSTRACT

We have developed computational techniques to improve the search for exons in uncharacterized vertebrate genomic DNA sequence. Our goal is to create software to identify precise intron/exon boundaries, with the assignment of the correct reading frame. The proposed exons may then be verified by identification in appropriate cDNA libraries and the amino acid translation may be used for protein homology searches. There are trade-offs between achieving maximal sensitivity (i.e. missing no true exons) and maximal specificity (i.e. having no false positive candidate exons). We have approached this problem by assigning a confidence level to each prediction. The program first identifies SORFs (spliceable open reading frames), defined as open reading frames bracketed by acceptable splice consensus sequences, or in the case of first exons, by an acceptable ATG initiation context at the 5' end. It then filters the resulting candidate exons through an algorithmic sieve based upon statistics that have been generated from known human genes. Thresholds for each confidence level are progressively relaxed up to 5 confidence levels. Restricting the program to the first 3 confidence levels (the default) identifies over 60% of first and internal exons with fewer than 50% false positives. Extending to confidence level 5 achieves 89% sensitivity for internal exons greater than 60 bp in length, and 83% for first exons, but with an increase in false positives to 60%. We employ the following strategy: A confidence 1 or 2 SORF has an *a priori* probability of identifying a true exon of over 80%. Their possible relationship to other putative exons of lower confidence can be determined by examining the reading frame phase. An important question arises when a highly confident exon prediction has not been verified by analysis of an appropriate cDNA. We hypothesize that some of these "phantom exons" exist as an alternately spliced form in a specific tissue type or developmental stage. The software is available by contacting the authors, and will run on IBM PCs or Sun workstations.

Assessing transcriptional changes in early mammalian development

**Barbara B. Knowles, Ph.D. The Wistar Institute,
3601 Spruce Street, Philadelphia, PA 19104.**

Because of the paucity of material and difficulty of access, the changes in gene expression which accompany development of the early mammalian embryo are largely unknown. To fill this void, we have made large and representative cDNA libraries in plasmid vectors from the unfertilized mouse egg, 2-cell and 8-cell embryo, and the blastocyst. We are now attempting to prepare similar libraries from dissected endoderm, ectoderm, and mesoderm of the early post-implantation mouse embryo. These primary libraries are a resource to quantitatively assess changing expression of known genes as the embryo transits the first cleavage divisions, implants in the uterus, and enters gastrulation. For example, stage-specific changes in the level of expression of cytokines and several members of the epidermal growth factor family are found in this period. The libraries also serve as a source of previously undescribed mouse genes which are homologous to those which regulate transcription in other organisms. To identify genes uniquely expressed at specific stages of development, we have employed a cDNA library subtraction protocol. Subtraction libraries of genes expressed as maternal transcripts and of genes uniquely expressed at the 2-cell stage are currently under investigation. Rapid identification of state-specific genes in these subtraction libraries is imperative to this stage of the investigation.

THE 3'-END REGION OF cDNAs AS PCR-BASED
POLYMORPHIC MARKERS FOR AN EXPRESSION MAP OF
THE MOUSE GENOME.

N. Takahashi^{1,2}, Y. Maezaki² and M.S.H. Ko^{1,2}

¹Center for Molecular Biology, Wayne State University,
Detroit MI, 48202, USA, ²Furusawa MorphoGene Project,
ERATO, JRDC, Tsukuba, Japan

Ideal markers for a cDNA map of the mouse genome should be amplifiable by the PCR and they should be polymorphic for genetic mapping, as well as unique for physical mapping. In our search for such markers, we did comparative sequencing of PCR-amplified genomic DNAs derived from 15 inbred strains and found that about 250 bp of the 3'-end region (3'-ER) of cDNAs, which is the sequence immediately upstream from the polyA tail, was polymorphic enough to distinguish the allele of a laboratory strain from that of a wild strain (*Mus spretus*) in 14 out of 22 cDNAs tested. Specifically, in 9 out of these 14 cDNAs, each allele was identified by the RFLP. This data indicates that about 65% of the 3'-ER of cDNAs can be automatically transformed into PCR-based genetic markers named "biallelic polymorphic expressed sequence tags (bESTs)." These markers can be easily and precisely mapped by typing of the panels of interspecific backcrosses. Because a large number of bEST markers can be efficiently obtained by a single-run automated sequencing of randomly selected cDNAs from the "normalized cDNA library," these markers will greatly facilitate the construction of high resolution genetic and physical maps of expressed sequences of the mouse genome.

3' Terminal Exon Trapping

David B. Krizman and Susan M. Berget
Marrs and Verna McClean Department of Biochemistry
Baylor College of Medicine
One Baylor Plaza
Houston, Texas 77030

The identification of transcriptional units in cloned unsequenced DNA is an arduous task at best, thus new technologies aimed at quickly identifying genes in uncharacterized segments of DNA would be useful for the mapping and sequencing of all the genes in the human genome.

To address this need we have developed an exon trapping strategy for the selection of 3'-terminal exons of genes based on recent investigations suggesting that exons are the units of recognition during vertebrate pre-mRNA processing. A 3'-terminal exon trapping vector, pTAG4, was constructed to permit positive selection for inserted genomic 3'-terminal exons. Recombinant populations containing insert DNA are transiently transfected into cultured mammalian cells. Polyadenylated, chimeric RNA resulting from transcription and processing of vector-born RNA is amplified by RT/PCR utilizing the 3' RACE method.

Insertion into pTAG4 of either a single DNA fragment containing only the the terminal exon from several genes or a mixture of DNA fragments consisting of both internal and terminal exons demonstrate the efficiency of the selection system for 3'-terminal exons. Experiments designed to test the potential of false positive scoring by the insertion into pTAG4 of DNA fragments containing partial terminal exons or random prokaryotic DNA indicate a high degree of specificity for intact 3' terminal exons. This protocol has also been used to scan anonymous pieces of DNA that map to a chromosomal region thought to contain disease genes. The trapped sequences indicate the discovery of a novel, transcriptionally active gene. Current work is focused on the trapping of 3'-terminal exons from cosmid and YAC inserts.

We feel that 3'-terminal exon trapping has the following significant advantages over previously described exon trapping protocols: 1) Last exons are 627bp in average length, thus a trapped 3'-terminal exon contains enough sequence to be useful for gene expression studies, full length cDNA cloning, and physical mapping. 2) This trapping scheme involves positive selection requiring the proximity of 2 types of signals (splicing and polyadenylation), thus reducing the possibility of false positives. 3) The vast majority of vertebrate genes contain a single 3' terminal exon thus minimizing scoring of the same gene more than once.

SEARCHING FOR GENES WITH A RECOMBINATION-BASED ASSAY: APPLICATION TO THE FRAGILE X AND TO GENES ON CHROMOSOME 21.

D.M. Kurnit, A.J. Hanzlik, M.A. Hauser, M.M. Osemlak-Hanzlik,
M. Van Keuren and G.-H. Xiao.

Departments of Pediatrics and Human Genetics and Howard Hughes Medical Institute,
University of Michigan

We developed a novel and efficient methodology for the isolation of cDNAs (genes). A sequence whose transcription profile is to be determined is cloned into our 2.5-kb *supF* R6K-derived plasmid, pMAD3. The bacterial strain carrying the insert-bearing plasmid is lysed with λ cDNA libraries and the recombinant *supF* bacteriophages are selected in our *dnaBam*, P1 *ban*, *lacZam*, *E. coli* host, DM21. Using primers abutting the cloning site, we reverse the recombination reaction (counterselection) via the polymerase chain reaction (PCR) to obtain the cDNA preparatively absent the genomic probe sequence. The rapidity of the RBA allows the screening of multiple cDNA libraries prepared from RNA isolated from different tissues at different times of development.

To demonstrate that this methodology works, we inserted the (CGG)₁₀AGG(CGG)₉AGG(CGG)₉ sequence from a fragile X gene into pMAD3. Clones were retrieved from multiple cDNA libraries, including from an 11 week human fetus and placenta, from a variety of 20 week human fetal tissues, including brain, spinal cord, eye, liver, kidney and skeletal muscle and from human adult jejunum. Recombination and counterselection retrieved the correct sequences as confirmed in each case by hybridization. This demonstrates that the fragile X sequence is transcribed in each of these tissues. The technique was then extended to demonstrate that multiple genomic sequences derived from several YACs in the distal portion of chromosome 21 (responsible for the phenotypic effects of Down syndrome) were transcribed in 20 week fetal brain, spinal cord, kidney, liver, eye or muscle, in the 11 week total fetus or in HeLa cells. Following counterselection, achieved via PCR to delete the original genomic sequence on chromosome 21, transcribed inserts are subcloned and sequenced.

We are attempting to extend this technology to phasmids. The strain, DM21 (pBR322), prevents replication from the ColE1 origin in phasmids via incompatibility between the ColE1 origin in DM21 (pBR322) and that in the phasmid. Using a M13 helper to mobilize the ColE1-R6K chimeric plasmid from the phasmid, we can infect an F'-carrying strain (N953-F') to select for chimeric pseudovirions carrying *supF* and the antibiotic resistance markers of both the ColE1 and R6K plasmids. We can then counterselect against *supF* using this strain, resulting in the elimination of the R6K plasmid and its insert and yielding the desired ColE1 plasmid and its cDNA insert. This technique includes advantages of ease of performance and the isolation of the desired counterselected cDNA sequence as a molecule cloned in a high copy-number ColE1 plasmid.

This procedure can be performed in a stand alone fashion or in tandem with other methodologies (such as hybrid selection, exon trapping, computer analyses, zoo blotting and expression in somatic cell hybrids) to identify sequences that are transcribed in different tissues. In addition to this critical time and tissue profile, the technique also yields a large cDNA starting from a small (>25 bp) probe, as would be generated from the above methodologies. Rapid DNA sequencing then enables one to determine if the gene belongs to a previously described family. Thus, the recombination-based assay will enable us to append a "genic" initiative onto the larger "genomic" initiative.

Generation of Human cDNA libraries and identification of cDNA with CAG repeat elements.

**Cheng Chi Lee and C. Thomas Caskey
Institute for Molecular Genetics and Howard Hughes Medical
Institute, Baylor College of Medicine, One Baylor Plaza,
Houston TX 77030, USA**

Human cDNA libraries have been synthesized from mRNA obtained from human placental, brain, muscle, heart, liver, and kidney tissue. Analysis of these libraries have revealed that the average size of the cDNAs is about 2.5kb. The average size of the cDNA recombinant ranges from 0.5×10^6 to 2×10^6 in these libraries. Screening the primary placental library with placental alkaline phosphatase cDNA probe reveal a frequency of 1:1500 suggesting good representation of this highly abundant gene. Sequencing several of these placental alkaline phosphatase cDNA have revealed that they are full-length cDNA of 2.9 kb. These libraries should have the necessary coverage for most of the expressed genes in these tissues and that the frequency of genes with large insert would be high.

We have screened these libraries with an oligonucleotide containing a CAG repeat elements and we have identified over 40 cDNAs that contained such a repeat sequence. Sequencing of these clones have revealed CAG repeat elements of 5 to 70.

Construction and analysis of cDNA reference libraries

**Sebastian Meier-Ewert and Hans Lehrach
ICRF, Genome Analysis Laboratory, London**

A series of cDNA libraries from human and mouse embryonic and adult tissues have been constructed in plasmid vectors. We have begun picking these clones into high density microwell plates (4x96 well format). Using robotic devices developed in our lab these can be arrayed at high density on nylon membranes. These filters can be used for hybridization with single copy of complex probes and will be incorporated in the ICRF reference library system. In addition, using a high throughput PCR system we are able to generate PCR products from all clones in a cDNA library, which are suitable targets for hybridization with short oligonucleotides. Under conditions discriminating perfect match from single base mismatch hybridizations, short oligo nucleotides can generate sequence information of their targets, which is the basis of the sequencing by hybridization method. The partial sequence information obtained can be used as a sequence fingerprint, identifying each clone. This system will make it possible to compare all sequences in any number of cDNA libraries and facilitate comparison to database sequences. It should be possible to identify all the different transcripts represented in the cDNA libraries constructed from mRNA of various tissues and developmental stages from both man and mouse. Comparison of fingerprints from different libraries should allow the generation of comparative expression data. We estimate that approximately 250 oligo fingerprints will be required to generate sufficient fingerprint information.

Toward a Chromosome 5-Specific cDNA Library.

Michael Lovett, Linda M. Hinton and Cynthia J. Hilliard.

Department of Biochemistry and McDermott Center, University of Texas
Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas TX
75325.

We have adapted our previously described direct selection protocols for more efficient and rapid isolation of low abundance cDNAs encoded by large genomic regions. Among the modifications are the substitution of a streptavidin/biotin capture system for the hybrid DNAs and alterations to our repeat suppression protocols. The modified technique has been used successfully on YAC contigs of several megabases in length to isolate novel low abundance cDNAs.

Enrichments of over 100,000-fold have been achieved in many cases. We are now applying these methods to the derivation of chromosome-specific cDNAs. The starting reagents for this scheme have been a pool of DNAs from the LANL chromosome 5-specific cosmid library and human placental cDNAs. To overcome the higher noise associated with using such a complex genomic DNA source we have successfully employed further modifications, to yield selective enrichment of all but one of the chromosome 5-specific reporter cDNAs that we have monitored.

TRAPPING OF EXONS FROM THE HUMAN MICROTUBULE ASSOCIATED PROTEIN TAU P. Nisson¹, A. Andreadis², K. Kosik² and P. Watkins¹, GIBCO-BRL/Life Technologies, Inc., Gaithersburg, Maryland; ²Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, Massachusetts.

Exon trapping is a method designed to identify and recover coding sequences (exons) from complex mammalian genomic DNA (Buckler et al., PNAS 88: 4009, 1991). Exons are selected by subcloning genomic DNA into a specially designed plasmid (pSPL1) containing a cloning site in the intron of the HIV tat gene flanked by functional 5' and 3' splice sites. Individual subclones or pools of subclones are transfected into COS-7 cells, which efficiently produce transcripts from pSPL1. Spliced products (candidate exons) are amplified by RNA-based PCR. Exon trapping was studied using a cosmid clone containing a portion of the gene encoding the human microtubule associated protein (MAP) tau, a brain specific MAP. This cosmid, which was isolated with a fetal tau cDNA probe, contains eight tau exons (3 to 9) which were subcloned into the splicing vector, pSPL1, and transfected into COS-7 cells. RNA was isolated, and RNA-based PCR was used to identify spliced products. PCR products were rapidly cloned using a highly efficient method recently developed for cloning PCR products - uracil DNA glycosylase (UDG) cloning (Nisson et al. PCR Methods and Applications 1: 120, 1991). Of the 8 exons tested, all four constitutive exons (4, 5, 7 and 9) were amplified (56 to 266 bp), cloned and verified by DNA sequencing to be correctly spliced. The four alternatively spliced exons (3, 4A and 8) were not captured by the assay, but the behavior of exon 6 depended on the identity of its flanking exons. These results suggest that this assay system may be used not only to capture constitutive exons from genomic DNA, but also to decipher the mechanisms regulating alternative splicing.

Large scale mapping of brain ESTs using PCR and somatic cell hybrids.

**Michael H. Polymeropoulos, Hong Xiao, Rosa Torres,
Anna Glodek, Carl J. Merril**

Laboratory of Biochemical Genetics, NIMH, Washington, D.C.

We have established chromosomal assignment for 350 brain derived expressed ESTs using PCR and human-rodent somatic cell hybrids. We have developed a procedure for the rapid chromosomal assignment of these ESTs: EST sequences are first analyzed by a computer program to determine regions likely not to be interrupted by introns in the genomic DNA. A pair of oligonucleotide primers is then designed to amplify this region by the polymerase chain reaction (PCR) using DNA template from human-rodent somatic cell hybrid chromosomal panels. The chromosomal assignment of the EST is determined by studying the segregation of the amplified products in these panels.

The designed oligonucleotide primers amplified the desired product in 81% of the amplified ESTs. Interference from the rodent background accounted for 8% of the failures to assign, while the remaining 1% was contributed by a total of three genes with multiple gene members. In summary, our approach has provided chromosomal assignment for 75% of the sequences attempted. This approach can be applied for sublocalization efforts using deletion and radiation hybrids, as well as large genomic fragments. This large scale cDNA mapping will facilitate physical as well as disease gene mapping. In order to facilitate transfer of mapping information, we have developed a database for the cataloguing of cDNA mapping information including primer sequences chromosomal assignment, regional localization and polymorphism information when available.

STRATEGIES FOR THE IDENTIFICATION OF TRANSCRIBED SEQUENCES IN THE Xq28 REGION

31

**Annemarie Poustka*, Bernhard Korn, Zdenek Sedlacek,
Johannes Coy, Petra Kioschis and David Konecki. DKFZ
Heidelberg, Germany**

As an essential step towards an exhaustive analysis of the coding potential of large regions of the genome, we are developing a number of different strategies to identify potential transcripts within regions covered by overlapping cosmid libraries. One of these techniques (Korn et al., Hum. Mol. Gen. 1992) is based on the hybridisation of inserts from cDNA libraries, which had been amplified by PCR from cDNA libraries, to biotinylated DNA from cosmids or cosmid pools. Nonspecific hybrids are then removed, the selected cDNAs are eluted and reamplified by PCR. Using a cosmid containing part of the FMR-1 gene as test, we were able to demonstrate an eighty thousand fold enrichment of cDNAs for this gene after two rounds of selection-amplification. The technique was applied to the analysis of transcripts from two cosmid contigs, together encompassing a region of 900 kb in Xq28. These experiments have thus far resulted in the identification of 81 different cDNA clones, of which 54 clones were mapped back to the cosmid contigs. Of the 54 clones placed on the contig maps, 12 cDNA clones can be shown to belong to genes which have been previously reported. We plan to use this method to establish enriched cDNA libraries from whole chromosomes. In addition, techniques have been developed to allow the isolation of conserved sequences by a similar series of selection and amplification steps, and to identify transcribed fragments in cosmid digests by direct hybridisation of radiolabelled cDNAs from e.g. pig tissues to cosmid grids and to Southern blots of cosmid digests.

The Linguistics of Transcribed Sequences

David B. Searls dsearls@cbil.humgen.upenn.edu

Department of Genetics, University of Pennsylvania School of Medicine

"Linguistic" approaches to biological sequence analysis are increasingly in evidence, due in part to the popularity of the common metaphor of DNA as language—a language, that is, describing the structures and processes required (at a minimum) to establish and maintain life. Much of the work in this area deals with analyses of vocabularies of frequently -occurring "words" in DNA. Our approach, however, is more concerned with higher-order structures, such as are characteristic of transcribed sequences. Such structure can be conveniently described using *grammars* and then searched for and analyzed with general-purpose programs called *parsers*.

We have developed a grammar formalism and parser that are especially suited to biological sequences, called GENLANG. Implemented in the logic programming language Prolog, GENLANG allows for interactive construction of complex patterns with a hierarchical, modular framework, in a rapid-prototyping mode. Other algorithms can be incorporated into this framework as needed, via foreign function calls.

We have written grammars ranging in complexity from simple motifs, to more complex features such as tRNA genes, to entire protein-encoding eukaryotic genes. The tRNA grammar was able to parse yeast chromosome III, finding all ten known tRNA genes in under half an hour. The same grammar finds over 95% of annotated tRNA genes in a range of GenBank divisions, so that it is comparable in both speed and sensitivity to special-purpose procedural tRNA finders. A more complex grammar for autocatalytic Group I introns was also developed that used a mix of information on pseudoknotted secondary structure and conserved primary sequence. This was demonstrated on a fungal mitochondrial genome where, together with grammars for tRNAs and ORFs, we were able to duplicate most of the published annotations for the sequence.

The protein-encoding gene grammar we have developed is able to assemble complex genes based on syntactic rules describing their structure, but until now it has not taken advantage of compositional cues for recognizing coding sequences, and thus has not been useful in search except for relatively simple genes (such as globins). We now report our efforts to incorporate compositional heuristics into the grammar, thus achieving more generally applicable results. We believe that this approach will provide a very flexible framework for easily modifying and specializing descriptions of gene classes, and serve as a useful adjunct to other "gene-finding" tools which are gaining widespread use.

D. B. Searls. Representing genetic information with formal grammars. In *Proceedings of the National Conference on Artificial Intelligence*, pages 386–391. AAAI, 1988.

D. B. Searls. Investigating the linguistics of DNA with Definite Clause Grammars. In E. Lusk and R. Overbeek, editors, *Logic Programming: Proceedings of the North American Conference on Logic Programming*, pages 189–208. MIT Press, 1989.

D. B. Searls. The computational linguistics of biological sequences. In L. Hunter, editor, *Artificial Intelligence and Molecular Biology*, chapter 2, pages 47–120. AAAI Press, 1992.

D. B. Searls and S. Dong. A syntactic pattern recognition system for DNA sequences. In H. A. Lim, J. Fickett, C. R. Cantor, and R. J. Robbins, editors, *Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*. World Scientific Publishing Co., 1992 (in press).

SINGLE-PASS AND FULL-LENGTH SEQUENCING AND PHYSICAL AND GENETIC MAPPING OF HUMAN BRAIN cDNAS. J.M. Sikela^{*}, T.J. Stevens, A.S. Wilcox, M.H. Polymeropoulos[‡], A.S. Khan, J.A. Hopkins, M. Robinson and A.K. Orpana, Univ. of Colorado Health Sci. Ctr., Denver, CO 80262 and [‡]Laboratory of Biochemical Genetics, NIMH, Washington, DC 20032.

The large scale collection, sequencing and mapping of human brain cDNAs represents a productive approach to the identification of most human brain genes and to the development of a cDNA expression map of the genome. For these reasons we have refined strategies for automated single-pass and full-length sequencing of human brain cDNAs and for physical mapping of cDNAs to specific locations in the genome. We have also identified a subset of cDNAs that contain polymorphic microsatellite sequences and demonstrate how they can be converted to highly informative (PIC value > 0.7) gene-associated genetic markers. At present, most of the first thousand sequenced cDNAs correspond to potentially new human brain genes, while a significant number of cDNAs appear to represent human homologs of interesting genes found in other species. We have also explored strategies for identifying and rapidly sequencing anonymous cDNAs that contain complete protein coding regions. Currently in our laboratory automated single-pass sequencing is being carried out at a rate of several thousand cDNAs per year. Therefore, coordination of this effort with other laboratories doing similar work should permit the sequence identification of most of the genes expressed in the human brain within the next few years.

"Methods for normalization of cDNA libraries and for isolation of chromosome-specific cDNAs". M. Fatima Bonaldo¹, M-T Yu², S. Brown³, L. Su¹, A. Efstratiadis², D. Warburton² & M. Bento Soares¹
Departments of Psychiatry¹, Genetics & Development² and Obstetrics and Gynecology³, Columbia University.

We have developed a method for normalization of directionally cloned cDNA libraries constructed in phagemid vectors, which involves priming of the library in the form of single-stranded circles with a Not I-oligo (dT) primer and controlled extensions with Klenow in the presence of dNTPs and ddNTPs. After purification of the partial duplexes over HAP, melting and reannealing to a moderate Cot, unhybridized (normalized) single-stranded circles are purified by HAP and electroporated into bacteria, generating a normalized library. We have followed this protocol to normalize a human neonatal brain cDNA library. The extent of normalization was determined by a series of screenings with probes that represent mRNAs from the 3 frequency classes. We have also developed a method for selection of chromosome-specific cDNAs and successfully utilized it to identify a number of genes on human chromosome 13. A normalized cDNA library, in the form of single-stranded circles, is hybridized to a filter immobilized chromosome-specific phage library and cDNA/ λ pairs are visualized by subsequent hybridization with a vector probe. Both members of a cDNA/ λ pair are then simultaneously isolated: the λ clone is purified by standard procedures, and the cDNA circles are eluted off the filter and electroporated into bacteria. Verification of correct cDNA selection is done by back hybridization with the corresponding λ clone. While the λ clone is used as a probe to position the gene to a cytogenetic band by in situ hybridization, the corresponding cDNA is sequenced from both ends for tentative gene function identification.

Positional cloning approaches towards isolating the Huntington Disease gene

Danilo A. Tagle¹, John Valdes¹, Lucio Castilla¹, Gill Bates², Marcy MacDonald³, Kristina Blanchard-McQuate¹, Manju Swaroop¹, James F. Gusella³, Hans Lehrach³, and Francis S. Collins¹. Department of Human Genetics, University of Michigan, Ann Arbor, MI²; Genome Analysis Laboratory, ICRF, London, U.K.; ³ Neurogenetics Laboratory, Massachusetts General Hospital, Boston, MA.

Huntington disease (HD) is a progressive, late onset, autosomal dominant disorder characterized by choreiform movements, personality changes, psychiatric disturbances and dementia. The pathological changes include premature neuronal cell death which is especially prominent in the basal ganglia of the brain. HD has been mapped to chromosome 4p16.3 within a 2.2 megabase (Mb) segment that lies between markers D4S10 and D4S168. In the current absence of a better genetic definition of the 2.2 Mb candidate region, the search for transcribed sequences in this large interval using conventional methods can be a laborious and daunting task. To identify candidate transcripts, we have employed novel approaches that permit the rapid isolation of cDNAs encoded within a large genomic segment cloned into YACs or cosmids. In the first strategy, YAC or cosmid genomic material is digested and ligated to linkers. The linkered genomic segments are PCR-amplified using a biotinylated primer and then hybridized at high stringency to PCR-amplified inserts from brain cDNA libraries. The biotinylated genomic-cDNA complexes are captured using streptavidin-coated magnetic beads and the nonspecific cDNAs are washed off. The captured cDNAs are eluted, PCR-amplified and are subcloned into a sublibrary of YAC or cosmid-specific cDNAs. To increase selectivity, it is possible to recycle the captured material through a second selection process.

The second strategy takes advantage of the observed high density of rare-cutter sites within this region that is likely to be indicative of many CpG islands which usually define the 5' end of genes and gene clusters. Using this approach, total YAC DNA is digested with infrequent-cutting restriction enzymes (NotI, BssH II, SacII, or SmaI) whose sites commonly occur in CpG islands. Uni-directional linkers (a.k.a. bubble or vectorette linkers) are ligated to the restricted DNA. PCR amplification is carried out in the presence of vectorette and Alu primers, thus generating a human specific probe which extends from the CpG island to the nearest Alu repeat element. The probe, which is gel purified away from the Alu-Alu PCR products, is likely to include the 5' end of a gene associated with the island and can be used to directly screen cDNA libraries.

The third strategy arises from the improving automated technology to rapidly determine the sequence of large genomic regions. This involves large-scale sequencing of cosmids within the candidate region and using a neural net-based computer search (GRAIL) to identify exons from the uncharacterized genomic sequences arising from this sequencing effort. Predicted exons are PCR-amplified and used as probes in Northern blots to determine tissue-specific expression patterns and in isolating full-length transcripts. This strategy was used to analyze 58 kilobases (kb) of sequences from three cosmids around the marker (D4S113) and resulted in the successful identification of at least three genes and several new polymorphisms.

Cloning and assessment of candidate genes for Huntington's disease

Bernhard Weber and Michael R. Hayden

University of British Columbia, Department of Medical Genetics, 416-2125 East Mall,
Vancouver B.C. V6T 1Z4

The genetic defect in Huntington's disease (HD) has been linked to DNA markers at 4p16.3. Subsequent efforts to refine the location of the HD gene by analyzing nonrandom allelic association and recombinant HD chromosomes have resulted in conflicting and mutually exclusive candidate regions for the HD mutation. The uncertainty of the precise location has significantly slowed progress in the isolation of the defective gene.

The identification of chromosomal rearrangements associated with or causing the genetic alteration has greatly facilitated the isolation of many disease genes to date. Thus far, no such structural rearrangements involving 4p16.3 have been identified in HD patients.

The new mutation rate for HD is considered to be one of the lowest known for human disorders. However, each identified case of sporadic HD is of particular interest as a new mutational event could involve a structural rearrangement which could disrupt or delete a functional locus thus leading to the diseased phenotype. We have now reexamined a previously published and well characterized case of sporadic HD. Haplotype analysis of the patient and 11 siblings has revealed that the HD patient has inherited a recombinant chromosome 4 from one of the unaffected parents. The site of recombination has been determined and narrowed to a 50 kb segment approximately 700 kb from the 4p telomere. Independently, DNA markers from this location have also been shown to be in strong allelic association with the HD phenotype. Thus far, we have identified two genes associated with this 50 kb segment. Mutational analysis using the single stranded conformational polymorphism (SSCP) technique and/or direct sequencing of all the exons and including 5' and 3' flanking sequences has not provided yet any direct evidence for an alteration which could be responsible for the HD mutation.

We have also identified several recombinant HD chromosomes in families with established history of HD. Some of these recombinant chromosomes point to a proximal region between loci D4S10 and D4S98 as a possible location for the HD gene. This region has previously been cloned in yeast artificial chromosomes (YACs) and partially in cosmid clones. By using a cDNA selection scheme we have isolated 16 non-overlapping cDNA clones from this region. The isolation of full-length cDNAs and the assessment of some of these genes as candidates for the HD gene is presently in progress.

Abstract for International Transcribed Sequences Workshop.
San Francisco, CA. November 7 & 8, 1992.

Parimoo, S., Patanjali, S., Shukla, H., Fan, W., Kandpal, R., Nallur, G.,
Vasavada, H. and S. M. Weissman.

We have been working on approaches for constructing normalized short fragment cDNA libraries and for rapid isolation of cDNA fragments corresponding to genes encoded by YACs or other large genomic DNA fragments. Normalized libraries have been constructed from several sources. We are exploring their use in isolation of cDNAs encoding peptides that interact with a given protein, in cataloging cDNAs encoding proteins with specific structural motifs, and in modified cDNA subtraction procedures. For example, we have used a contingent replication assay to isolate from a normalized cDNA library the cDNA encoding the large subunit of TFIIF as well as other cDNAs that appear to encode peptides that react specifically with TFIID in vivo.

Selection techniques have been evolved that make it possible to isolate fragments of cDNAs encoded by a given YAC with little contamination from ribosomal sequences, very little contamination with sequences from interspersed repetitive DNA, and virtually no contamination with unique copy genomic sequences not present in the YAC. We are currently applying this technique to the human MHC. Tests in the Class II region indicated that most or all known genes can be conveniently recovered by this approach. In the Class I region the approach has generated a number of cDNAs encoding synonymous proteins. These exhibit a variety of patterns of tissue specific expression. Currently, we are attempting to adapt this approach for selection of cDNAs with whole DNA from yeast carrying YACs, and improving sensitivity to the point where dense arrays of dots of yeast DNA can be used for cDNA selection.

INTERNATIONAL WORKSHOP ON THE IDENTIFICATION OF TRANSCRIBED SEQUENCES

ROSTER

Kuniya Abe, Ph.D.
Associate Professor
Institute of Molecular
Embryology and Genetics
Kumamoto University
School of Medicine
4-24-1 Kuhonji, Kumamoto City
Kumamoto 862
JAPAN
TEL 81-96-344-2111, ext. 6733
FAX 81-96-362-8393

Charles Auffray, Ph.D.
Genethon
1 rue de l'Internationale
91000 Evry
France
TEL 33-1-6947-2965
FAX 33-1-6077-8698

Susan Berget, Ph.D.
Department of Biochemistry
Baylor College of Medicine
RM T8-17
One Baylor Plaza
Houston TX 77030
TEL 713-798-5758
FAX 713-796-9438

Giorgio Bernardi, Ph.D.
Laboratoire de
Genetique Moleculaire
Institute Jacques Monod
Tour 43 -2 Place
Jussieu, 75005, Paris
France
TEL 33-1-43-29-58-24
FAX 33-1-44-27-79-77
Bernardi@FRCITI51.bitnet

Christiane Bouchier, Ph.D.
Genethon
1 rue de l'Internationale
91000 Evry
France
TEL 33-1-6947-29 7
FAX 33-1-6077-8698

Miles B. Brennan, Ph.D.
Unit on Genomics
NIMH, IRP, CNG
Bldg. 10, Rm 4N320
Bethesda MD 20892
TEL 301-402-1769
FAX 301-402-2140
Genunit@helix.nih.gov

Roger Brent, Ph.D.
Associate Molecular Biologist, Massachusetts General Hospital
Associate Professor of Genetics, Harvard Medical School
Department of Molecular Biology
Wellman 8
Massachusetts General Hospital
Boston MA 02114
TEL 617-726-5925
FAX 617-726-6893
brent@frodo.mgh.harvard.edu

Anthony J. Brookes, Ph.D.
MRC, Human Genetics Unit
Western General Hospital
Crewe Road
Edinburgh EH4 2XU
Scotland
TEL 44-31-332-2471
FAX 44-31-343-2620
Tony@UK.ac.mrcvax.

Gail A.P. Bruns, M.D., Ph.D.
Associate Professor
Genetics Division
Children's Hospital
300 Longwood Avenue
Boston MA 02115-5737
TEL 617-735-7575/7572
FAX 617-735-7588

Alan Buckler, Ph.D.
Molecular Neurogenetics Laboratory
Massachusetts General Hospital-East
Building 149, 13th Street
Department of Neurogenetics
Charlestown MA 02129
TEL 617-726-5725
FAX 617-726-5735
Buckler@MGH-helix.harvard.edu

Jean-Michel Claverie, Ph.D.
NCBI, NLM, NIH
Building 38A, Room 8S804
8600 Rockville Pike
Bethesda MD 20894
TEL 301-496-2475
FAX 301-480-9241
JML@ncbi.nlm.nih.gov

Ruchira Das Gupta, Ph.D.
Department of Molecular Genetics
Albert Einstein College of Medicine
1925 East Chester Road
Bronx NY 10461
TEL 718-597-4369
FAX 212-822-0845
gasgupta@accom.yu.edu

Nicole A. Datson, M.Sc.
Department of Human Genetics
Leiden University
Wassenaarseweg 72, 2333 Al Leiden
PO Box 9503, 2300 R.A. Leiden
The Netherlands
TEL 31-71-276046
FAX 31-71-276075
WMBDUNNE@rulgl.LeidenUniv.nl

Radoje, Drmanac, Ph.D.
Integral Genetics Group
Biological and Medical Research Division
Argonne National Laboratory
9700 South Cass Avenue
Argonne IL 60439-4833
TEL 708-252-3175
FAX 708-252-3387
rade@everest.bim.anl.gov

Geoffrey M. Duyk, M.D., Ph.D.
Department of Genetics
Howard Hughes Medical Institute
Assistant Professor,
Assistant Investigator
Harvard Medical School
200 Longwood Avenue
Boston MA 02115
TEL 617-432-6072
FAX 617-432-7663
Duyk@rascal.med.harvard.edu

James Eberwine, Ph.D.
Assistant Professor
Department of Pharmacology
University of Pennsylvania
Medical Center
36th Street and Hamilton Walk
Philadelphia PA 19104-6084
TEL 215-898-0420
FAX 215-898-9927

Jeffrey D. Falk, Ph.D.
Department of Molecular Biology
The Scripps Research Institute
10666 North Torrey Pines Road
La Jolla CA 92037
TEL 619-554-4314
FAX 619-554-6112

Katheleen Gardiner, Ph.D.
Eleanor Roosevelt Institute
1899 Gaylord Street
Denver CO 80206
TEL 303-333-4515
FAX 303-333-8423

Michael Higgins, Ph.D.
Department of Human Genetics
Roswell Park Cancer Institute
Buffalo NY 14263
TEL 716-845-3574
FAX 716-845-8449

Ute Hochgeschwender, M.D.
Unit on Genomics
NIMH, IRP, CNG
Bldg. 10, Rm 4N320
9000 Rockville Pike
Bethesda MD 20892
TEL 301-402-1769
FAX 301-402-2140
Genunit@helix.nih.gov.

Gordon B. Hutchinson, M.D., C.M.
Canadian Genetic Disease Network
NCE Building, Room 424
2125 East Mall
Vancouver, British Columbia
Canada V6T 1Z4
TEL 604-822-7742
FAX 604-822-9238
hutch@ulam.generes.ca

Barbara B. Knowles, Ph.D.
Professor
Wistar Institute
3601 Spruce Street
Philadelphia PA 19104
TEL 215-898-3789
FAX 215-573-2077

Minoru S.H. Ko. M.D., Ph.D.
Assistant Professor (Research)
Center for Molecular Biology
Wayne State University
517 Biological Sciences Building
5047 Gullen Mall
Detroit MI 48202
TEL 313-577-0616
FAX 313-577-6200

David B. Krizman, Ph.D.
Department of Biochemistry, T315
Baylor College of Medicine
One Baylor Plaza
Houston TX 77030
TEL 713-798-4622
FAX 713-796-9438

David M. Kurnit, M.D., Ph.D.
Investigator
Howard Hughes Medical Institute
Professor, Pediatrics and Human Genetics
University of Michigan Medical School
1150 W. Medical Center Drive
3520 MSRB I, Box 0650
Ann Arbor MI 48109-0650
TEL 313-747-4747
FAX 313-936-9353
david.kurnit@um.cc.umich.edu

Cheng Chi Lee, Ph.D.
T836, Institute for Molecular Genetics
Baylor College of Medicine
One Baylor Plaza
Houston TX 77030
TEL 713-798-6548
FAX 713-798-7383

Sebastian Meier-Ewert, Ph.D.
ICRF
PO Box 123
44 Lincoln's Inn Fields
London, WC2A 3PX
United Kingdom
TEL 44-71-269-3308
FAX 44-71-405-4303

Greg Lennon, Ph.D.
Human Genome Center, L-452
Lawrence Livermore National Labs
Livermore CA 94550
TEL 510-422-5711
FAX 510-423-3608
greg@mendel.llnl.gov

Michael Lovett, Ph.D.
Assistant Professor of Biochemistry
Department of Biochemistry
University of Texas-
Southwestern Medical Center at Dallas
5323 Harry Hines Boulevard
Dallas TX 75235-9038
TEL 214-688-5436
FAX 214-688-8856/7579

Mary Kay McCormick, Ph.D.
M-880
Los Alamos National Laboratory
Los Alamos NM 87545
TEL 505-665-4438
FAX 505-665-3024
MKM@flovax.lanl.gov

Richard J. Mural, Ph.D.
Biology Division
Oak Ridge National Laboratory
P. O. Box 2009
Oak Ridge TN 37840
TEL 615-576-2938
FAX 615-574-1274
M9L@stc10.ctd.ornl.gov

Paul E. Nisson, Ph.D.
Staff Scientist
Genome Analysis Group
GIBCO BRL/Life Technologies, Inc.
8717 Grovemont Circle
Gaithersburg MD 20877
TEL 301-670-7774
FAX 301-921-2116

Kousaku Okubo, M.D., Ph.D.
Institute for Molecular
and Cell Biology
Osaka University
1-3 Yamada-oka
Suita, Osaka
565 JAPAN
TEL 81-6-877-5244, x3314
FAX 81-6-877-1922
kousaku@imcb.osaka-u.ac.jp

Mihael H. Polymeropoulos, M.D.
Laboratory of Biochemical Genetics
WAW Building
NIMH Neuroscience Center at
St. Elizabeth's Hospital
Room 131
2700 Martin Luther King Avenue
Washington, D.C. 20032
TEL 202-373-6077
FAX 202-373-6087
mph@bgsun1.nimh.nih.gov

Annemarie Poustka, Ph.D.
In Neuenheimer Feld 506
German Cancer Research Center
Institute for Virus Research
D-6900 Heidelberg
Germany
TEL 49-6221-423-409
FAX 49-6221-470-333

David Searls, Ph.D.
Research Associate Professor
Department of Genetics
Room 475, Clinical Research Building
University of Pennsylvania
School of Medicine
422 Curie Boulevard
Philadelphia PA 19104-6145
TEL 215-573-3107
FAX 215-573-5892
dsearls@cbil.humgen.upenn.edu

Yosef Shiloh, Ph.D.
Department of Human Genetics
The University of Michigan Medical Center
1150 W. Medical Center Drive
4570 MSRB II
Ann Arbor MI 48109-0650
TEL 313-747-3407
FAX 313-763-4692

James Sikela, Ph.D.
Assistant Professor
Department of Pharmacology, C-236
University of Colorado
Health Sciences Center
4200 East 9th Avenue
Denver CO 80262
TEL 303-270-8637
FAX 303-270-7097
Sikela_J%maui@Vaxf.colorado.edu

Marcelo Bento Soares, Ph.D.
Assistant Professor of Neurogenetics
Department of Psychiatry
Columbia University
722 West 168th Street
Box 41
New York NY 10032
TEL 212-960-2313
FAX 212-795-5886
cvc@cuccfa.ccc.columbia.edu

Danilo A. Tagle, Ph.D.
Department of Human Genetics
The University of Michigan Medical Center
4578 MSRB II
1150 W. Medical Center Drive
Ann Arbor MI 48109-0650
TEL 313-747-3407
FAX 313-763-4692
DANILO_TAGLE@UM.CC.UMICH.EDU.

Gert J.B. Van Ommen, Ph.D.
Head, Department of Human Genetics
Leiden University
Wassenaarseweg 72, 2333 Al Leiden
PO Box 9503, 2300 R.A.
Leiden, The Netherlands
TEL 31-71-276065
FAX 31-71-276075

Bernhard H.F. Weber, Ph.D.
University of British Columbia
Department of Medical Genetics
422-2125 East Mall
Vancouver, B.C. Canada V6T 1Z4
TEL 604-822-5539
FAX 604-822-9238
bweb@ulam.generes.ca

Sherman M. Weissman, M.D.
Sterling Professor of Genetics and Medicine
Yale University School of Medicine
336A BCMM
295 Congress Avenue
New Haven CT 06510
TEL 203-737-2282
FAX 203-737-2286
Weissma@yalemed.edu

DATE

FILMED

14 / 14 / 94

END