GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

12 October 2004
Lecture 13: Biological Interpretation

- Introduction
- Primary probe identifiers
  - Sigma-Genosys
  - Agilent
  - Affymetrix
  - IMAGE
- GenBank
- UniGene
- LocusLink
- Batch Resources
Introduction

After analyzing a microarray experiment, you typically end up with a list of “interesting” genes. Today’s lecture deals with how to make biological sense of that list.

Keep in mind that the list may be quite long. For example, in our analysis of the prostate cancer data set, we found about 3500 genes that were differentially expressed (along with 3500 potential biomarkers, with an overlap of about 2500).

There are numerous databases (GenBank, UniGene, LocusLink, etc.) that include gene-related information. It can be difficult to keep track of exactly what kind of information each identifier describes. Each of the three kinds of microarrays has a different primary descriptor that tells you what biological material was placed on the array.
Primary Identifiers

synthesized oligonucleotide array: These are the Affymetrix arrays. The primary identifier is an Affymetrix probe set ID, which refers to the collection of 25-mers that form a probe set.

spotted cDNA array: These are the most common glass microarrays. The primary identifier is an IMAGE clone id, which refers to the actual cDNA clone attached to a vector and propagated in bacteria.

spotted oligonucleotide array: Newer glass arrays often spot commercially synthesized 60- or 70-mers instead of cDNA clones. The primary identifier is usually the commercial identifier that corresponds to the actual synthesized sequence.
Oligonucleotide Libraries

An accurate transcriptome should be the starting point for gene expression probe design. Sigma-Genosys and Compugen together offer Oligonucleotide Libraries designed using a superior transcriptome database generated by Compugen's LEADS™ technology. These genome-wide probe sets:

- Capture the maximum number of splice variants
- Avoid SNPs, repeats, chimeras, and intron contamination
- Are logically organized by Gene Ontology™ functional assignment
- Maximize sensitivity and specificity

Each library probe is a 65-mer (except human library probes, which are 60-mers) with a 5'-C6 amino modifier for covalent attachment. Oligos in this length range can be designed...
INTRODUCTION TO MICROARRAYS

Sigma-Genosys Human Oligo

NEW 29,000 - Gene Human Oligonucleotide Library

Designed using newly available genomic and expressed sequence data and an enhanced LEADS platform for transcriptome analysis, the Human Oligonucleotide Library Release 2.0 sets new standards for coverage and quality.

This new library represents an advance over Release 1.0 in the following ways:

- Wider genomic coverage
- New representation of predicted genes
- Better splice variant representation
- Lower cross homology
- Better sequence quality and fewer polymorphisms
- Fewer bases in repeat alignments
- Improved identification of chimeric sequences and antisense genes

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly
INTRODUCTION TO MICROARRAYS

Sigma-Genosys Product List

<table>
<thead>
<tr>
<th>Catalog #</th>
<th>Description</th>
<th>Downloads</th>
<th>Gene list</th>
<th>Product Info (PDF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACLIB96</td>
<td>B. Subtilis OligoLibrary: 96 well format, 43 plates (4,128 oligos)</td>
<td><img src="Image" alt="Image" /></td>
<td>Excel, 2 MB Zip, 262 KB</td>
<td>Info sheet Product profile</td>
</tr>
<tr>
<td>BACLIB384</td>
<td>B. Subtilis OligoLibrary: 384 well format, 11 plates (4,128 oligos)</td>
<td><img src="Image" alt="Image" /></td>
<td>Excel, 2 MB Zip, 268 KB</td>
<td>Info sheet Product profile</td>
</tr>
<tr>
<td>HUBLIB96</td>
<td>Human OligoLibrary Release 2.0: 96 well format, 304 plates (29,134 oligos)</td>
<td><img src="Image" alt="Image" /></td>
<td>Excel, 14 MB Zip, 2 MB</td>
<td>Info sheet Product profile</td>
</tr>
</tbody>
</table>
Sigma-Genosys Oligo List

<table>
<thead>
<tr>
<th>Plate</th>
<th>Location</th>
<th>GenBank_accession</th>
<th>UniGene_ID</th>
<th>Serial_number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>M33197</td>
<td>Hs.189476</td>
<td>CGEN_HUMCTRL_10001_1_0</td>
<td>Human gly</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>NM_005856</td>
<td>Hs.25681</td>
<td>CGEN_HUMAN_LIB2 6010300008 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>NM_012076</td>
<td>Hs.169745</td>
<td>CGEN_HUMAN_LIB2 6010300016 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>NM_001243</td>
<td>Hs.1314</td>
<td>CGEN_HUMAN_LIB2 6010300024 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>NM_006028</td>
<td>Hs.241377</td>
<td>CGEN_HUMAN_LIB2 6010300032 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>NM_000810</td>
<td>Hs.24969</td>
<td>CGEN_HUMAN_LIB2 6010300040 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>NM_004857</td>
<td>Hs.48714</td>
<td>CGEN_HUMAN_LIB2 6010300048 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>S76942</td>
<td>Hs.99922</td>
<td>CGEN_HUMAN_LIB2 6010300056 0</td>
<td>dopamine</td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>NM_000871</td>
<td>Hs.22180</td>
<td>CGEN_HUMAN_LIB2 6010300064 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>NM_014286</td>
<td>Hs.301760</td>
<td>CGEN_HUMAN_LIB2 6010300072 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A11</td>
<td>NM_002531</td>
<td>Hs.110642</td>
<td>CGEN_HUMAN_LIB2 6010300080 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>A12</td>
<td>U37028</td>
<td>Hs.381264</td>
<td>CGEN_HUMAN_LIB2 6010300088 0</td>
<td>Homo leu</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>NM_002924</td>
<td>Hs.79348</td>
<td>CGEN_HUMAN_LIB2 6010300001 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>NM_001458</td>
<td>Hs.58414</td>
<td>CGEN_HUMAN_LIB2 6010300009 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>NM_017589</td>
<td>Hs.128180</td>
<td>CGEN_HUMAN_LIB2 6010300017 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>NM_014207</td>
<td>Hs.58665</td>
<td>CGEN_HUMAN_LIB2 6010300025 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B5</td>
<td>NM_001702</td>
<td>Hs.194654</td>
<td>CGEN_HUMAN_LIB2 6010300033 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B6</td>
<td>NM_002246</td>
<td>Hs.24040</td>
<td>CGEN_HUMAN_LIB2 6010300041 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B7</td>
<td>NM_001302</td>
<td>Hs.374866</td>
<td>CGEN_HUMAN_LIB2 6010300049 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B8</td>
<td>NM_152219</td>
<td>Hs.135211</td>
<td>CGEN_HUMAN_LIB2 6010300057 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
<tr>
<td>B9</td>
<td>NM_019074</td>
<td>Hs.3736</td>
<td>CGEN_HUMAN_LIB2 6010300065 0</td>
<td>Homo sapi</td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCTION TO MICROARRAYS

http://www.chem.agilent.com

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GS01 0163: ANALYSIS OF MICROARRAY DATA
INTRODUCTION TO MICROARRAYS

Agilent Human Microarrays

Whole Human Genome Oligo Microarray Kit

Double Your Data!

This single microarray represents a compiled view of the human genome as it is understood today. The sequence information used to design this product was derived from a broad survey of well-known sources such as RefSeq, Genmap, Ensembl, Unigene and others. The resulting view of the human genome covers 41K unique genes and transcripts which have been verified and optimized by alignment to the human genome assembly and by Agilent's Empirical Validation process.

This product is intended for use in research only. Use of the product requires agreement to specific licenses.
**Agilent Product List**

<table>
<thead>
<tr>
<th>Microarray Kit</th>
<th>Gene List</th>
<th>ref Guide</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard file</td>
<td>Compressed File</td>
</tr>
<tr>
<td></td>
<td>html format</td>
<td>Excel format</td>
</tr>
<tr>
<td><strong>Oligo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human 1A (V2) G4110B</td>
<td>.xls 4.8 Mbytes</td>
<td>.xls .zip 1.3 Mbytes</td>
</tr>
<tr>
<td>Human 1B G4111A</td>
<td>.xls 3.2 Mbytes</td>
<td>. xls .zip 0.9 Mbytes</td>
</tr>
<tr>
<td>Human Genome, Whole G4112A</td>
<td>.xls 8.4 Mbytes</td>
<td>.xls .zip 2.2 Mbytes</td>
</tr>
<tr>
<td>Mouse (Dev) G4120A</td>
<td>.htm 10.4 Mbytes</td>
<td>.xls 3.8 Mbytes</td>
</tr>
<tr>
<td>Mouse G4121A</td>
<td>.htm 8.3 Mbytes</td>
<td>.xls 3.6 Mbytes</td>
</tr>
<tr>
<td>NEW Mouse Genome, Whole</td>
<td>.xls 10.8 Mbytes</td>
<td>.xls .zip 2.8 Mbytes</td>
</tr>
</tbody>
</table>
**Agilent Oligo List**

<table>
<thead>
<tr>
<th>Systematic Name</th>
<th>Probe Name</th>
<th>Gene Name</th>
<th>Additional Accessions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_000015</td>
<td>A_23_P31798</td>
<td>NAT2</td>
<td>ref</td>
<td>NM_000015.1</td>
</tr>
<tr>
<td>NM_000016</td>
<td>A_23_P96781</td>
<td>ACADM</td>
<td>ref</td>
<td>NM_000016.2</td>
</tr>
<tr>
<td>NM_000017</td>
<td>A_23_P65022</td>
<td>ACADS</td>
<td>ref</td>
<td>NM_000017.1</td>
</tr>
<tr>
<td>NM_000018</td>
<td>A_23_P207650</td>
<td>ACADVL</td>
<td>ref</td>
<td>NM_000018.1</td>
</tr>
<tr>
<td>NM_000019</td>
<td>A_23_P24515</td>
<td>ACAT1</td>
<td>ref</td>
<td>NM_000019.2</td>
</tr>
<tr>
<td>NM_000020</td>
<td>A_23_P324783</td>
<td>ACVR1</td>
<td>ref</td>
<td>NM_000020.1</td>
</tr>
<tr>
<td>NM_000021</td>
<td>A_23_P206586</td>
<td>PSEN1</td>
<td>ref</td>
<td>NM_000021.2</td>
</tr>
<tr>
<td>NM_000022</td>
<td>A_23_P340728</td>
<td>PSEN1</td>
<td>ref</td>
<td>NM_000021.2</td>
</tr>
<tr>
<td>NM_000023</td>
<td>A_23_P210482</td>
<td>ADA</td>
<td>ref</td>
<td>NM_000022.1</td>
</tr>
<tr>
<td>NM_000024</td>
<td>A_23_P66637</td>
<td>SGCA</td>
<td>ref</td>
<td>NM_000023.1</td>
</tr>
<tr>
<td>NM_000025</td>
<td>A_23_P145024</td>
<td>ADRB2</td>
<td>ref</td>
<td>NM_000024.2</td>
</tr>
<tr>
<td>NM_000026</td>
<td>A_23_P168993</td>
<td>ADRB3</td>
<td>ref</td>
<td>NM_000025.1</td>
</tr>
<tr>
<td>NM_000028</td>
<td>A_23_P155103</td>
<td>ADSS</td>
<td>ref</td>
<td>NM_000026.1</td>
</tr>
<tr>
<td>NM_000029</td>
<td>A_23_P7099</td>
<td>AGA</td>
<td>ref</td>
<td>NM_000027.2</td>
</tr>
<tr>
<td>NM_000030</td>
<td>A_23_P115261</td>
<td>AGT</td>
<td>ref</td>
<td>NM_000029.1</td>
</tr>
<tr>
<td>NM_000031</td>
<td>A_23_P83550</td>
<td>AGXT</td>
<td>ref</td>
<td>NM_000030.1</td>
</tr>
<tr>
<td>NM_000032</td>
<td>A_23_P324787</td>
<td>ALAD</td>
<td>ref</td>
<td>NM_000031.3</td>
</tr>
<tr>
<td>NM_000033</td>
<td>A_23_P385897</td>
<td>ALAS2</td>
<td>ref</td>
<td>NM_000032.1</td>
</tr>
<tr>
<td>NM_000034</td>
<td>A_23_P148556</td>
<td>ABCD1</td>
<td>ref</td>
<td>NM_000033.2</td>
</tr>
<tr>
<td>NM_000035</td>
<td>A_23_P88996</td>
<td>ALDOA</td>
<td>ref</td>
<td>NM_000034.2</td>
</tr>
<tr>
<td>NM_000036</td>
<td>A_23_P415974</td>
<td>ALDOB</td>
<td>ref</td>
<td>NM_000035.2</td>
</tr>
</tbody>
</table>
http://www.affymetrix.com
Main page for NetAffx

The NetAffx™ Analysis Center enables researchers to correlate their GeneChip® array results with array design and annotation information. This resource provides you with unprecedented access to array content information, including probe sequences and gene annotations. You can quickly search for genes and/or SNPs, compare and refine results, and export data into Excel-friendly formats.

To learn more, explore the tools below or review help material available through support.

Tools and Annotations

NetAffx Analysis Center (click to get started, login required)

Release Notes


August 9, 2004 - A new feature, "Save Current List" was released. Users can now save refined lists of target sequences for later reference.

June 23, 2004 - The quarterly annotation and downloadable file update was released and includes the following updates:

- Genomic alignments of consensus sequences were used as the primary means to assign genes to probe sets (instead of an accession-based method) for the human, mouse, and rat arrays.
- Pfam, SMART, ProfileScan, ScanRegExp and TM (transmembrane domain) annotations are now provided through InterPro
INTRODUCTION TO MICROARRAYS

Affymetrix Analysis Central

Expression
- Search all available information in the database for a particular term or identifier. This is recommended as a starting point for your searches. [Quick Query]
- Search specific fields in the database for a term or identifier [Standard Query]
- Retrieve annotations for a probe list [Batch Query]
- Find probe sets that align to your sequence(s) through BLAST [BLAST]
- Find probes that identically match your sequence(s) [Probe Match]
- Query the UCSC Browser for genomic alignment [UCSC Query]

Genotyping
- Search all available information in the database for a particular term or identifier. This is recommended as a starting point for your searches. [Quick Query]
- Search specific fields in the database for a term or identifier [Standard Query]
- Retrieve annotations for a probe list [Batch Query]
- Query the UCSC public genome by position [UCSC Query]
- Search for SNPs between microsatellites [SNP Finder]

Begin
Affymetrix Quick Search

1. Select a GeneChip Array:
(Use control-select to search up to three arrays simultaneously.)

- Human Genome U133 Plus 2.0 Array
- Mouse Genome 430 2.0 Array
- Mouse Genome 430A 2.0 Array
- Human Genome U133 Set
- Human Genome U95 Set

2. Enter search terms:
Query All Descriptions

Leave empty to get all probe sets for the selected array(s).
Use & for AND, | for OR, and ! for NOT between terms. See Query Examples
Quick query adds wildcards(%) at the front and end of each term.

search →
**INTRODUCTION TO MICROARRAYS**

**Affymetrix Batch Search**

- **QUERY**
  - **Expression**
  - **Standard**
  - **Query**
  - **Batch Query**
  - **BLAST**
  - **Probe Match**
  - **UCSC**
  - **Query**
  - **Genotyping**
  - **Quick Query**
  - **Standard**
  - **Query**
  - **Batch Query**
  - **UCSC**
  - **Query**
  - **SNP Finder**

- **CURRENT QUERY**
  - 60 probe sets
  - Annotations
  - Show
  - Orthologs
  - GO Browser
  - Export

- **QUERY HISTORY**
  - Annotation
  - Views
  - Expression

**QUERY**

**Batch Query**

Rapidly obtain results for up to 3,000 Affymetrix probe set accession numbers, gene names, or sequences ids at once.

1. **Select a GeneChip Array:**
   (Use control-select to search up to three arrays simultaneously.)

   - Human Genome U133 Plus 2.0 Array
   - Mouse Genome 430 2.0 Array
   - Mouse Genome 430A 2.0 Array
   - Human Genome U133 Set
   - Human Genome U95 Set

2. **Select a search option:**
   - Probe Set ID

3. **Upload a text (*.txt) file with the appropriate query values:**
   **Example**

   ![Image of the Affymetrix Batch Search interface]

   Note: Query values should be separated by a new line (\n) character. A Microsoft Excel file should be saved as "text only" file before uploading.

4. **Select a view:**
   - Annotation List

5. **Name Query:** (Optional)
   - (95 Character max)

---

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GS01 0163: Analysis of Microarray Data
INTRODUCTION TO MICROARRAYS

Affymetrix Search Results

<table>
<thead>
<tr>
<th>Details</th>
<th>Probe Set ID</th>
<th>Gene Title</th>
<th>Gene Symbol</th>
<th>GO Biological Process Description</th>
<th>GO Molecular Function Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1569449_at</td>
<td>cofactor required for Sp1 transcriptional activation, subunit 2, 150kDa</td>
<td>CRSP2</td>
<td>regulation of transcription from Pol II promoter</td>
<td>receptor activity, RNA polymerase II transcription mediator activity</td>
</tr>
<tr>
<td></td>
<td>1570186_at</td>
<td>nuclear receptor</td>
<td>NR113</td>
<td>initiation from Pol II promoter</td>
<td>transcription activity, activator activity, ligand-dependent nuclear receptor activity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>androgen receptor signaling pathway</td>
<td>transcription coactivator activity, vitamin D receptor binding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>thyroid hormone receptor binding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>regulation of transcription factor</td>
</tr>
</tbody>
</table>
Affymetrix Detailed Search Results

GeneChip Array Information
- Probe Set ID: 211621_at
- GeneChip Array: Human Genome U133 Plus 2.0 Array
- Organism Common Name: Human

Probe Design Information
- Transcript ID: g178655
- Sequence Type: Exemplar sequence
- Representative Public ID: M73069 NCBI
- Target Description: This probe set was annotated using the Genome based pipeline to a Locus Link identifier using 2 transcripts.
**INTRODUCTION TO MICROARRAYS**

Affymetrix Sequence Information

The image displays a screenshot of a computer interface for analyzing microarray data, specifically focusing on Affymetrix sequence information. The screenshot includes a table with the following columns:

- **Target Sequence**: The target sequence is presented in a sequence format.
- **BLASTn GenBank NR**: BLASTn results for GenBank NR.
- **Probe Sequence (5'-3')**: The probe sequence is listed with the corresponding probe X and probe Y information.
- **Probe Interrogation Position**: The position of the probe interrogation.
- **Strandedness**: Indication of whether the probe is antisense or sense.

The table contains several sequences and their corresponding probe information, illustrating the process of analyzing microarray data using Affymetrix technology.
INTRODUCTION TO MICROARRAYS

http://image.lnl.gov

THE I.M.A.G.E. CONSORTIUM

Welcome to the world’s largest public collection of genes

- Order an I.M.A.G.E. clone
- Resources: libraries, vectors, more
- Query our database
- Get information on projects
- Query IMAGEnc clustering tool
- Search the problematic clone list
- History and goals
- What’s new?
- Check the FAQ
- Current plate availability
- Quality control process and results
- Meet our staff

This page has been viewed 00355469 times, since 08/27/1999

Web page maintained by:
webmaster@image.lnl.gov

Enter your email to join the I.M.A.G.E. Consortium Announcements List

Biological Questions and Comments to:
images@image.lnl.gov
IMAGE is integrated into GenBank

**MGC verified full-length, human:** This is a subset of the IRAK/IRAL MGC predicted full-length set; these clones have been sequence-verified to match the EST already in dbEST, and have been fully sequenced to high accuracy (full-length sequences are submitted to the primate division of Genbank, NOT dbEST). This rearray is in 96-well format, with each plate containing clones from one species and vector only. Rarray names are IRAT (human, ampicillin-resistant vectors) and IRAU (human, chloramphenicol-resistant vectors). Periodic updates will be made as more full-length sequences are submitted. To access the full-length sequences, search [Genbank](not dbEST) using the following format: "MGC:MGC_id" to return individual sequences or MGC [kywd] for all sequences. The I.M.A.G.E. cloneID can be found in the FEATURES field of the Genbank entry.


Additional datafiles, including fasta files of all full-length sequences, can be found at the MGC [website](not dbEST).

Status: ongoing

**MGC verified full-length, mouse:** This is a subset of the IRAK/IRAL MGC predicted full-length set; these clones have been sequence-verified to match the EST already in dbEST, and have been fully sequenced to high accuracy (full-length sequences are submitted to the primate division of Genbank, NOT dbEST). This rearray is in 96-well format, with each plate containing clones from one species and vector only. Rarray names are IRAV (mouse, ampicillin-resistant vectors) and IRAW (mouse, chloramphenicol-resistant vectors). Periodic updates will be made as more full-length sequences are submitted. To access the full-length sequences, search [Genbank](not dbEST) using the following format: "MGC:MGC_id" to return individual sequences or MGC [kywd] for all sequences. The I.M.A.G.E. cloneID can be found.
GenBank: the common denominator

You may have noticed that all the commercial products map their proprietary identifiers to GenBank. GenBank is the primary repository for sequence information.

While the IMAGE ID refers to the actual clone, the corresponding GenBank entry (or entries) describes the actual mRNA sequence.

For Affymetrix, the GenBank entry describes the sequence from which the probe set was constructed.

For Sigma-Genosys or Agilent, the GenBank entry describes the sequence from which the long oligo was selected.
Entrez Nucleotide

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases grows at an exponential rate. As of April 2004, there are over 38,989,342,565 bases.

Building the human genome

The Human Genome Reference DNA Sequence was completed in April 2003. The current version is listed as a build number on the Genome View page and includes an accompanying set of statistics and release notes.
Entrez Tools

Advanced Entrez Tools

Web Tools:

- **Batch Entrez** - Upload a file of GI or accession numbers to retrieve sequences.
- **Batch Citation Matcher** - Send citation information to Entrez and retrieve PubMed IDs for linking, citation display, or other applications.
- **Advanced Entrez Searching** - Advanced searching techniques for Web Entrez.
- **NCBI Cubby** - Set user preferences, store search strategies and default e-mail address, and specify resources for LinkOut display.

Programming Tools:

- **E-Utilities** - Run Entrez queries and download data from your own scripts over the Web.
- **Linking to Entrez** - Link to specific Entrez pages from your own web pages or applications.
- **Entrez Client/Server** - C language library for embedding Entrez calls into your programs.
INTRODUCTION TO MICROARRAYS

Batch Entrez

Batch Entrez has changed!

You will need Batch Entrez now only to upload a file of GI or accession numbers for an Entrez search. You can do all other large searches directly within Entrez.

There are now two ways to retrieve large amounts of sequence data using Entrez:

1. Begin at one of the Entrez pages and perform a large search just like any other Entrez search:
   - Enter a query
   - Designate the database as Nucleotide or Protein, as appropriate
   - Press Go; you will see a list of sequence entries.
## Getting a Gene List

```r
> extreme <- abs(t.stat) > 6 & tr.stats > 20
> extreme.genes <- gene.info[extreme,
> dim(extreme.genes)
> [1] 1005 6
> extreme.genes[1:8,]

<table>
<thead>
<tr>
<th>Clone.ID</th>
<th>Accession</th>
<th>Gene.Symbol</th>
<th>Cluster.ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>X24</td>
<td>IMAGE:34849</td>
<td>R20379</td>
<td>EEF2</td>
</tr>
<tr>
<td>X27</td>
<td>IMAGE:45525</td>
<td>H08440</td>
<td>RFP</td>
</tr>
<tr>
<td>X63</td>
<td>IMAGE:295831</td>
<td>N74602</td>
<td>CGI-26</td>
</tr>
<tr>
<td>X82</td>
<td>IMAGE:143322</td>
<td>R74357</td>
<td></td>
</tr>
<tr>
<td>X100</td>
<td>IMAGE:753381</td>
<td>AA406332</td>
<td>SEC23A</td>
</tr>
<tr>
<td>X132</td>
<td>IMAGE:341083</td>
<td>W58562</td>
<td>C6orf56</td>
</tr>
<tr>
<td>X205</td>
<td>IMAGE:78946</td>
<td>T61792</td>
<td>PDK4</td>
</tr>
<tr>
<td>X225</td>
<td>IMAGE:586831</td>
<td>AA130866</td>
<td>TMLHE</td>
</tr>
</tbody>
</table>
```

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly  
GS01 0163: ANALYSIS OF MICROARRAY DATA
Just GenBank

> extreme.gb <- extreme.genes$Accession
> write.table(extreme.gb, 'extreme.txt',
+ quote=FALSE, row.names=FALSE, col.names=FALSE)
> write.table(extreme.genes, 'extreme-genes.txt',
+ quote=FALSE, row.names=FALSE, col.names=TRUE,
+ sep='\t')

We need the one-column list of GenBank identifiers for a batch search of GenBank. Other tools can use the more general list.
INTRODUCTION TO MICROARRAYS

Batch Entrez

Batch Entrez has changed!

You will need Batch Entrez now only to upload a file of GI or accession numbers for an Entrez search. You can do all other large searches directly within Entrez.

There are now two ways to retrieve large amounts of sequence data using Entrez:

1. Begin at one of the Entrez pages and perform a large search just like any other Entrez search:

   ○ Enter a query
   ○ Designate the database as Nucleotide or Protein, as appropriate
   ○ Press Go; you will see a list of associated sequences
Entrez Results

1: R20379
yg40f10.r1 Soares infant brain 1NIB Homo sapiens cDNA clone
IMAGE:34849 5' similar to gb:X51466 ELONGATION FACTOR 2 (HUMAN), MRNA sequence
gi|775013|gb|R20379.1||775013|

2: H08440
yl89c06.r1 Soares infant brain 1NIB Homo sapiens cDNA clone
IMAGE:45525 5' similar to contains Alu repetitive element;contains MSR1 repetitive element ;, MRNA sequence
gi|873262|gb|H08440.1||873262|

3: N74602
za48h04.s1 Soares fetal liver spleen 1NFLS Homo sapiens cDNA clone
IMAGE:295831 3', MRNA sequence
gi|1231887|gb|N74602.1||1231887|
Entrez Results

Why did we only get 1003 results from a batch query with 1005 entries? Probably because two of the GenBank identifiers have been “retired” since the microarray was constructed. This typically happens when someone finds out there was a problem (often vector contamination) with the original sequence entry.

Note: To save the results, click the “Send to” button after first changing the destination to “File”.

We can use the results of the search to compare the IMAGE clone IDs supplied with the original array with the current record in GenBank. It is certainly possible that someone (in the distant past) typed one of the numbers incorrectly.
UniGene

GenBank only refers to individual sequences. Because lots of people have sequenced lots of fragments of RNA and entered them into GenBank, a sequence entry is not the same thing as a gene.

UniGene is the NCBI’s attempt to organize sequences into coherent clusters that should represent genes.

**Critical Fact:** UniGene changes regularly. The current sequence data is reclustered about once a month. The presence of new sequence information can change the clusters. As of two or three years ago, as many as 25% of the UniGene cluster assignments changed over the course of a year. One hopes that the rate of change is decreasing.
UniGene Home Page

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

<table>
<thead>
<tr>
<th>Species</th>
<th>UniGene Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chordata</td>
<td></td>
</tr>
<tr>
<td>Mammalia</td>
<td></td>
</tr>
<tr>
<td><em>Bos taurus</em></td>
<td>25,713</td>
</tr>
<tr>
<td><em>Canis familiaris</em></td>
<td>15,694</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>54,560</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>46,544</td>
</tr>
<tr>
<td><em>Ovis aries</em></td>
<td>3,154</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>40,329</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>24,028</td>
</tr>
<tr>
<td>Aves</td>
<td></td>
</tr>
<tr>
<td><em>Gallus gallus</em></td>
<td>21,371</td>
</tr>
</tbody>
</table>

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly
UniGene Results

1: Hs.75309  
Eukaryotic translation elongation factor 2  
*Homo sapiens*, 1493 sequence(s)

2: At.47255  
Glycine-rich protein (GRP16)  
*Arabidopsis thaliana*, 8 sequence(s)
UNIGene Details

UniGene Cluster Hs.75309 Homo sapiens Eukaryotic translation elongation factor 2 (EEF2)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with proteins supported by a complete genome. The alignments can suggest function of a gene.

A. thaliana ref:NP_172112.1 39.59 % / 846 aa
- unknown (see ProtEST)
protein [Arabidopsis thaliana]

C. elegans ref:NP_492457.1 78.74 % / 858 aa
- Elongation factor Tu family (see ProtEST)

D. melanogaster pir:S05988 78.42 % / 858 aa
S05988 translation elongation factor (see ProtEST)
UniGene Expression Information

**GENE EXPRESSION**

- Tissues and development stages from this gene's sequences survey gene expression.
- Links to other NCBI expression resources.

**cDNA sources:**
- Bladder, Bone, Bone Marrow, Brain, Cervix, Colon, Eye, Heart, Kidney, Larynx, Liver, Lung, Lymph Node, Mammary Gland, Muscle, Ovary, Pancreas, Peripheral Nervous System, Placenta, Prostate, Skin, Small Intestine, Soft Tissue, Spleen, Stomach, Tongue, Testis, Uterus, Vascular, Blood, Embryo, Juvenile, Adult

**Restricted Expression:** Adult [Show more like this]

**Expression Profile:**
- View expression levels using UniGene's EST ProfileViewer

**GEO profiles:**
- Gene expression profiles in the NCBI Gene Expression Omnibus database

**NLAIII tags:**
- NLAIII anchored SAGE tags mapped to Hs.75309 in the NCBI SageMap database

**SAU3A tags:**
- SAU3A anchored SAGE tags mapped to Hs.75309 in the NCBI SageMap database
INTRODUCTION TO MICROARRAYS

Expression Profile

Expression profile suggested by analysis of EST counts.
Hs.75300- EEF2: Eukaryotic translation elongation factor 2

Expression by Tissue

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Count</th>
<th>EST</th>
<th>TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder</td>
<td>336</td>
<td>71</td>
<td>20873</td>
</tr>
<tr>
<td>Bone</td>
<td>71</td>
<td>42</td>
<td>599</td>
</tr>
<tr>
<td>Bone Marrow</td>
<td>109</td>
<td>34</td>
<td>36496</td>
</tr>
<tr>
<td>Brain</td>
<td>339</td>
<td>156</td>
<td>45673</td>
</tr>
<tr>
<td>Cervix</td>
<td>513</td>
<td>22</td>
<td>41263</td>
</tr>
<tr>
<td>Colon</td>
<td>565</td>
<td>184</td>
<td>16919</td>
</tr>
<tr>
<td>Eye</td>
<td>271</td>
<td>44</td>
<td>162029</td>
</tr>
<tr>
<td>Heart</td>
<td>71</td>
<td>47</td>
<td>55457</td>
</tr>
<tr>
<td>Kidney</td>
<td>194</td>
<td>26</td>
<td>133604</td>
</tr>
<tr>
<td>Larynx</td>
<td>502</td>
<td>12</td>
<td>23894</td>
</tr>
<tr>
<td>Liver</td>
<td>68</td>
<td>9</td>
<td>131488</td>
</tr>
<tr>
<td>Lung</td>
<td>340</td>
<td>96</td>
<td>282392</td>
</tr>
<tr>
<td>Lymph Node</td>
<td>207</td>
<td>33</td>
<td>128146</td>
</tr>
<tr>
<td>Mammary Gland</td>
<td>843</td>
<td>102</td>
<td>120986</td>
</tr>
<tr>
<td>Muscle</td>
<td>477</td>
<td>52</td>
<td>109859</td>
</tr>
</tbody>
</table>
UniGene Mapping Information

MAPING POSITION
Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

Genome View: Multiple Mappings
CytoGenetic map: 19pter-q12
UniSTS entry: Chr 19 RH80123
UniSTS entry: Chr 20 EST11A2
UniSTS entry: Chr 17 RH11500

SEQUENCES
Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

mRNA sequences (12)
M19997.1 Human elongation factor 2 (EF-2) mRNA, 3' end.
NM_001961.2 Homo sapiens eukaryotic translation elongation factor 2 (EEF2), mRNA.
BG088002.1 Homo sapiens cDNA clone MGC:76402 IMAGE:9052873, complete cds
CR616031.1 full-length cDNA clone CS0DF058YA20 of fetal brain of Homo sapiens (human).
CR608809.1 full-length cDNA clone CS0CAP08YN19 of Thymus of Homo sapiens (human).
CR802709.1 full-length cDNA clone CS0DG06YD23 of B cells (Ramos cell line) of Homo sapiens (human).

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly
INTRODUCTION TO MICROARRAYS

Chromosome Mapping

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

GS01 0163: ANALYSIS OF MICROARRAY DATA
UniGene Links

UniGene Cluster Hs.75309 *Homo sapiens*
Eukaryotic translation elongation factor 2 (EEF2)

SELECTED PROTEIN SIMILARITIES
Comparison of sequences in UniGene with proteins supported by a complete genome. Alignments can suggest function of a gene.

- *A. thaliana*  
  ref: NP_172112.1 39.59% / 846  
  - unknown protein  
  [Arabidopsis thaliana]

- *C. elegans*  
  ref: NP_492457.1 78.74% / 858 aa  
  - Elongation factor Tu family

- *D. melanogaster*  
  pir:S05988 78.42% / 858 aa  
  - Translation elongation factor  
  (see ProtEST)
LocusLink

The first link out from UniGene is to LocusLink. LocusLink provides a single query interface to curated sequence information and descriptive information about genetic loci. This includes

- official nomenclature (symbol, name)
- aliases
- sequence accession numbers
- phenotypes
- MIM numbers
- UniGene clusters
- homology
- map locations
LocusLink Home Page

Introduction

LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.

Sequence accessions include a subset of GenBank accessions for a locus, as well as a new type, the NCBI Reference Sequence (RefSeq). RefSeq records are built according to the process detailed here. See the About and FAQ pages for more information.
LocusLink Results

LocusLink will be replaced by Entrez Gene. Check Gene FAQ for current information.

Click to Display mRNA-Genomic Alignments (spanning 9391 bps)

Homo sapiens Official Gene Symbol and Name (HGNC)

EEF2: eukaryotic translation elongation factor 2
LocusID: 1938

Overview

RefSeq Summary: This gene encodes a member of the GTP-binding translation elongation factor family. This protein is
LocusLink Results

RefSeq Summary: This gene encodes a member of the GTP-binding translation elongation factor family. This protein is an essential factor for protein synthesis. It promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome. This protein is completely inactivated by EF-2 kinase phosphorylation.

Locus Type: gene with protein product, function known or inferred
Product: eukaryotic translation elongation factor 2
Alternate Symbols: EF2, EF-2
Alias: polypeptidyl-tRNA translocase

Function: Submit GeneRIF

Please note: As a consequence of our transition to Entrez Gene, GeneRIFs will only be displayed in the Entrez Gene record for this locus.

Gene Ontology™:
LocusLink Results

LocusLink Results

mRNA: NM_001961
Protein: NP_001952

Related Sequences

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Type</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC006547</td>
<td>m</td>
<td>AAH06547</td>
</tr>
<tr>
<td>BC024689</td>
<td>m</td>
<td>AAH24689</td>
</tr>
<tr>
<td>M19997</td>
<td>m</td>
<td>AAA50388</td>
</tr>
<tr>
<td>X51466</td>
<td>m</td>
<td>CAA35829</td>
</tr>
<tr>
<td>Z11692</td>
<td>m</td>
<td>CAA77750</td>
</tr>
<tr>
<td>None</td>
<td>p</td>
<td>P13639</td>
</tr>
</tbody>
</table>

Additional Links

- OMIM: 130610
- UniGene: Hs.75309

Questions or Comments?
Write to the NCBI Service Desk
Both UniGene and LocusLink create automatic links to OMIM. OMIM is a curated database of human genes and genetic disorders. It typically includes information about which diseases appear to be linked to specific genes, along with primary references that explain how the gene was sequenced and mapped to specific chromosomal regions.
OMIM Main Page

Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above.
OMIM Results

EUKARYOTIC TRANSLATION ELONGATION FACTOR 2; EEF2

Alternative titles; symbols

ELONGATION FACTOR 2; EF2
POLYPEPTIDYL-tRNA TRANSLOCASE

Gene map locus 19pter-q12

TEXT

Diphtheria toxin and Pseudomonas exotoxin A (PA toxin) inhibit protein synthesis by catalyzing covalent binding of the ADP-ribose moiety of NAD to elongation factor-2 (EF2). EF2 is required for the translocation step in protein...
OMIM References


REFERENCES

CONTRIBUTORS

Rebekah S. Rasooly - updated: 2/10/1999

CREATION DATE

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly
Batch Resources

You may have noticed that only GenBank (Entrez-Nucleotide) provided a “batch” option that allowed us to search for an entire list of genes at once. Every other example we have presented works with one gene at a time. That’s probably not a good way to deal with 1000 genes.

M.D. Anderson maintains a service called GeneLink to deal with this problem. You can try it at [http://bioinformatics.mdanderson.org/GeneLink.html](http://bioinformatics.mdanderson.org/GeneLink.html).

Stanford has a similar tool (that right now is faster) called Source. Check out [http://source.stanford.edu](http://source.stanford.edu).
GeneLink Description

Welcome to GeneLink

GeneLink is a cross reference tool that allows you to search for information about many genes simultaneously. It performs as a web-based service to conduct batch searches of standard bioinformatics databases such as GenBank, UniGene, LocusLink, etc. The advent of high throughput biological techniques, such as microarrays and SAGE, has led to an increasing need for simultaneously collecting the existing information on large numbers of genes. GeneLink is an ongoing project being developed by the Section of Bioinformatics to facilitate these simultaneous searches.

Because the actual location of the search page may change depending on the demand for various services, please bookmark the current page as your starting point for GeneLink searches.

Go to the GeneLink Search Page

GeneLink is also available as a web service. Applications can retrieve bioinformatics data via a web service call. Available functions and application instructions are available on the web page below.
GeneLink Search Page

GeneLink links gene information across databases in spreadsheet format that can be pasted into an existing spreadsheet. More information is available below.

GeneLink is currently using Homo Sapiens Unigene Build 175.

Select the desired output columns. Check the box next to "Include input" if you want your search terms included as the first column of the result table.

Search for Accession Numbers that match exactly the items listed either in the file or on separate lines of the following list:

Rows to display per page (default is 25)

Note: Enter 'ALL' to display all rows on one page

Output file delimiter (default is tab '\t')

Include links in output file

Output as an html table:
INTRODUCTION TO MICROARRAYS

GeneLink Search Page

[Image of GeneLink Search Page]

Search for Accession Numbers that match exactly the items listed either in the file C:\Rprojects\tailrank\extreme-genes.txt or on separate lines of the following list:

- [ ] Rows to display per page (default is 25)
- [ ] Note: Enter 'ALL' to display all rows on one page
- [ ] Output file delimiter (default is tab '\t')
- [ ] Include links in output file
- [ ] Output as an html table:
- [ ] Output to Email Address:
  (A tab-delimited text file will be emailed to you when it is ready. This option is good for large queries such as those that contain more than 1000 search items.)

Submit

Output Fields
Basic output
- Include
  - Search Term
  - Accession Number
  - Unigene Cluster
  - Gene Symbol
  - Chromosome
  - Cytoband
  - Gene Name
  - LocusLink
Advanced output
- Gene Ontology
- Biocarta
- Kegg
Source Description

SOURCE is a unification tool which dynamically collects and compiles data from many scientific databases, and thereby attempts to encapsulate the genetics and molecular biology of genes from the genomes of Homo sapiens, Mus musculus, Rattus norvegicus into easy to navigate GeneReports.

The mission of SOURCE is to provide a unique scientific resource that pools publicly available data commonly sought after for any clone, GenBank accession number, or gene. SOURCE is specifically designed to facilitate the analysis of large sets of data that biologists can now produce using genome-scale experimental approaches.

Five Most Recent Additions:

Read the SOURCE paper at NAR or as a PDF
SOURCE is now available at http://source.stanford.edu
Retrieval of upstream genomic sequences from human GeneReports
Link to HGNC GeneReports from human
Source Search Page

Choose organism: Homo sapiens

Choose search option:
Gene Name/Symbol

Enter a search term:

Use a wildcard character (*) at the end and/or beginning of the term to broaden your search.

Choose type of information to display:
- GeneReport: Gene Information (limited to those in UniGene)
- CloneReport: cDNA Clone Information (limited to those in dbEST)

Submit  Reset Form

Batch SOURCE: Extract data for multiple genes at once
Introduction to microarrays

Source Batch Search Page

Input File:
C:\R\projects\tailrank\extreme.txt  Browse...

Or enter a list of identifiers:

Select the type of input identifier: GenBank Accession

Choose organism: Homo sapiens

2) Choose field(s) for extraction:
- UniGene Cluster ID
- UniGene Name
- UniGene Symbol
- UniGene Representative Protein Acc.
- Chromosome Location
- Cytoband
- Other Ontology Annotations (full)
- Other Ontology Annotations (short)
- Enzymatic Function
Source Results

Successfully read input file: extreme.txt
File contains 1003 line(s) of data
Data type in input file: Acc
Disregarding choice of organism since it is intrinsic to Acc

Performing Acc to UniGene Cluster look-up...
500 accessions read from file
1000 accessions read from file
Extracting requested data...
500 clones processed
1000 clones processed

876 keys were found in UniGene (and were not chimeric)
127 keys were not found in UniGene
0 keys mapped to more than one UniGene cluster

Download the results file.
### Source Details

<table>
<thead>
<tr>
<th>Source ID</th>
<th>Name</th>
<th>Symbol</th>
<th>Chromosome</th>
<th>Cytoplasm</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.501493</td>
<td>Transcribed sequence with strong similarity to protein sp:1</td>
<td>R20379</td>
<td>Sec23 homolog A (S. cerevisiae)</td>
<td>14</td>
<td>14q11</td>
</tr>
<tr>
<td>Hs.400382</td>
<td>Ret finger protein RFP</td>
<td>6p22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.3727</td>
<td>Serine/threonine kinase receptor associated protein</td>
<td>STRAP 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.367688</td>
<td>Clone IMAGE:4794726, mRNA</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.528298</td>
<td>Sec23 homolog A (S. cerevisiae)</td>
<td>SEC23A 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.528298</td>
<td>Sec23 homolog A (S. cerevisiae)</td>
<td>SEC23A 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.102471</td>
<td>Phosphatase and actin regulator 2</td>
<td>PHACTR2 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.8364</td>
<td>Pyruvate dehydrogenase kinase, isoenzyme 4</td>
<td>PDK4 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.133321</td>
<td>Trimethyllysine hydroxylase, epsilon</td>
<td>TMLHE X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.58551</td>
<td>S protein-coupled receptor 87</td>
<td>GPR87 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.121576</td>
<td>Myosin IB</td>
<td>MYO1B 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.35380</td>
<td>Bobby sex homolog (Drosophila)</td>
<td>BXR 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.42173</td>
<td>Chromosome 6 open reading frame 107</td>
<td>C6orf107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.212787</td>
<td>KIAA0303 protein</td>
<td>KIAA0303 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.190477</td>
<td>Zinc finger CCH type domain containing 6</td>
<td>ZC3H12A 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.392874</td>
<td>Transforming growth factor, beta receptor III (beta TC1)</td>
<td>3q24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.32111</td>
<td>Hypothetical protein MGC3265</td>
<td>MGC3265 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.459470</td>
<td>WD repeat and SOCS box-containing 2</td>
<td>WSB2 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.404761</td>
<td>Nuclear factor (erythroid-derived 2)-like 3</td>
<td>NFE2L3 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.407934</td>
<td>Neuron navigator 2</td>
<td>NAV2 11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.419149</td>
<td>Junctional adhesion molecule 3</td>
<td>JAM3 11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.452398</td>
<td>CDNA FLJ30740 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.157211</td>
<td>Clq and tumor necrosis factor related protein 5</td>
<td>C1QTNF5 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.23180</td>
<td>Secreted protein, acidic and rich in cysteine</td>
<td>SREC1 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.598124</td>
<td>Dynamin, axonemal, heavy polypeptide 5</td>
<td>DNAH5 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.93112</td>
<td>Data not found</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.15108</td>
<td>Data not found</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.124578</td>
<td>Phosphatase and actin regulator 2</td>
<td>PHACTR2 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.102501</td>
<td>Ras homolog gene family, member B</td>
<td>RHOB 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.124578</td>
<td>Netrin 4</td>
<td>NTN4 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.90797</td>
<td>Hypothetical protein E0C06005</td>
<td>LOC129642 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.381089</td>
<td>ARAP-associated sperm protein</td>
<td>ASP 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.134859</td>
<td>V-maf musculoaponeurotic fibrosarcoma oncogene homolog</td>
<td>MAP1A 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reprise

- Primary identifiers describe the biological material used as probes
- GenBank accession numbers are a common denominator to start updating gene annotations
- UniGene clusters sequences that represent the same gene
- LocusLink contains curated sequence information about names, mappings, etc
- OMIM contains curated connections between genes and diseases, with references into the literature
- There are batch processing tools to go from a list of GenBank accession numbers through UniGene to LocusLink or OMIM.

Main Question: How do we learn about gene functions, networks, and pathways?
INTRODUCTION TO MICROARRAYS

Return to LocusLink

The role of the ribosome. This protein is completely inactivated by EF-2 kinase phosphorylation.

Locus Type: gene with protein product, function known or inferred

Product: eukaryotic translation elongation factor 2

Alternate: EF2, EEF-2

Symbols: polypeptidyl-tRNA translocase

Function: Submit GeneRIF

Please note: As a consequence of our transition to Entrez Gene, GeneRIFs will only be displayed in the Entrez Gene record for this locus.

Gene Ontology™:

- GTP binding
- protein biosynthesis
- translation elongation factor activity
- translational elongation

Evidence Source Pub

IEA GOA
IEA GOA
NR GOA
IEA GOA
INTRODUCTION TO MICROARRAYS

http://www.geneontology.org

GENE ONTOLOGY CONSORTIUM

What is the Gene Ontology?
Download the Ontologies

The goal of the Gene Ontology™ (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes. GO is one of the controlled vocabularies of the Open Biological Ontologies.

• Submit new GO term suggestions via the Curator Requests Tracker at SourceForge. Help with new term submission is available.

• Send comments and questions to go@geneontology.org

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly  GS01 0163: ANALYSIS OF MICROARRAY DATA
GeneOntology

GeneOntology uses controlled vocabularies to create a directed acyclic graph (a generalized tree) that describes the kinds of functions or properties that a gene might have.

The properties are divided into three categories:

1. Biological process (what)
2. Molecular function (how)
3. Cellular component (where)