Analysis of Microarray Data

Keith Baggerly and Brad Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

8 September 2009
Lecture 3: Linking Numbers to Biology

- So, why are we here?
- Why do we care?
- Affymetrix source for annotations
- List of Affymetrix annotations
- Updating the annotations in dChip
- What is GeneOntology?
- Using GeneOntology in dChip
- GoMiner
So, why are we here?

We want to learn about Gene Annotations.

Microarrays are designed, which means that someone first chooses a set of genes of interest, selects probe sequences to target those genes, and then places those sequences on a microarray. In order to interpret (and possibly to analyze) the data produced from a microarray experiment, you need to refer to the accompanying annotations, which describe both the probes and the targeted genes.
Things Change

One might naively think that gene annotations are static; meaning that they are produced when the microarray is designed and never change again. *Wrong.*

The base pair sequences of probes placed on the array do not change. However, our knowledge of the human genome is evolving, and thus our opinion about which genes are targeted by those sequences may need to be updated.

For Affymetrix microarrays, the company maintains annotation files (updated quarterly) that contain their latest opinion on the nature and identity of the targeted genes.
Why Do We Care?

Earlier, we compared microarray data from samples of acute lymphocytic leukemia (ALL) patients and mixed-lineage leukemia (MLL) patients. Using the criteria that the lower bound of fold change (LBFC) should be at least 1.2-fold and the mean difference in expression should be greater than 100, **we found about 600 probesets to be differentially expressed.**

It is considered bad form to just hand the biologists a list of 600 genes.

They typically want to know: (a) do these genes reflect particular biological functions that are different between the two groups of samples, or (b) do they identify specific biological pathways or networks that are perturbed?
## List of Differentially Expressed Genes

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene Description</th>
<th>Baseline Mean</th>
<th>Baseline SD</th>
<th>Experiment Mean</th>
<th>Experiment SD</th>
<th>Fold Change</th>
<th>Lower BC</th>
<th>Upper BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>37880_at</td>
<td>A kinase (PRKA) anchor protein (gravin)</td>
<td>2973.7</td>
<td>560.63</td>
<td>148.29</td>
<td>24.19</td>
<td>-20.05</td>
<td>-12.93</td>
<td>-30.28</td>
</tr>
<tr>
<td>1325_at</td>
<td>MAD, mothers against decapentaplegic homolog 1</td>
<td>7759.92</td>
<td>1390.4</td>
<td>595.18</td>
<td>64.06</td>
<td>-13.04</td>
<td>-8.89</td>
<td>-18.03</td>
</tr>
<tr>
<td>37280_at</td>
<td>MAD, mothers against decapentaplegic homolog 1</td>
<td>9124.17</td>
<td>1538.9</td>
<td>702.89</td>
<td>37.85</td>
<td>-12.98</td>
<td>-9.29</td>
<td>-16.88</td>
</tr>
<tr>
<td>37908_at</td>
<td>guanine nucleotide binding protein 11</td>
<td>2160.91</td>
<td>565.93</td>
<td>226.99</td>
<td>58.16</td>
<td>-9.52</td>
<td>-4.92</td>
<td>-18.23</td>
</tr>
<tr>
<td>34194_at</td>
<td>Homo sapiens mRNA; cDNA DKFZp564B076 (from unknown)</td>
<td>962.11</td>
<td>296.29</td>
<td>107.48</td>
<td>34.97</td>
<td>-8.95</td>
<td>-3.95</td>
<td>-21.14</td>
</tr>
<tr>
<td>753_at</td>
<td>nidogen 2 (osteonidogen)</td>
<td>2558.48</td>
<td>890.45</td>
<td>304.16</td>
<td>22.09</td>
<td>-8.41</td>
<td>-3.58</td>
<td>-13.49</td>
</tr>
<tr>
<td>1992_at</td>
<td>fragile histidine triad gene</td>
<td>1742.98</td>
<td>252.98</td>
<td>209.02</td>
<td>29.64</td>
<td>-8.34</td>
<td>-5.92</td>
<td>-11.72</td>
</tr>
<tr>
<td>1488_at</td>
<td>protein tyrosine phosphatase, receptor type, K</td>
<td>4128.67</td>
<td>1140</td>
<td>572.2</td>
<td>38.89</td>
<td>-7.22</td>
<td>-3.91</td>
<td>-10.70</td>
</tr>
<tr>
<td>1077_at</td>
<td>recombination activating gene 1</td>
<td>6827.92</td>
<td>1443.9</td>
<td>1021.43</td>
<td>204.86</td>
<td>-6.78</td>
<td>-4.09</td>
<td>-11.13</td>
</tr>
<tr>
<td>33910_at</td>
<td>Homo sapiens mRNA; cDNA DKFZp564P116 (from unknown)</td>
<td>460.85</td>
<td>209.6</td>
<td>72.66</td>
<td>7.64</td>
<td>-6.34</td>
<td>-1.59</td>
<td>-11.49</td>
</tr>
<tr>
<td>34800_at</td>
<td>leucine-rich repeats and immunoglobulin-like domain</td>
<td>5255.48</td>
<td>907</td>
<td>899.41</td>
<td>189.08</td>
<td>-5.84</td>
<td>-3.74</td>
<td>-9.53</td>
</tr>
<tr>
<td>35614_at</td>
<td>transcription factor-like 5 (basic helix-loop-helix)</td>
<td>7264.11</td>
<td>1378.1</td>
<td>1248.25</td>
<td>122.02</td>
<td>-5.82</td>
<td>-3.9</td>
<td>-8.05</td>
</tr>
<tr>
<td>41266_at</td>
<td>integrin, alpha 6</td>
<td>7523.59</td>
<td>1222.5</td>
<td>1445.79</td>
<td>200.87</td>
<td>-5.48</td>
<td>-3.84</td>
<td>-7.73</td>
</tr>
<tr>
<td>37343_at</td>
<td>inositol 1,4,5-triphosphate receptor, type 3</td>
<td>5231.99</td>
<td>747.28</td>
<td>966.99</td>
<td>97.72</td>
<td>-5.41</td>
<td>-3.98</td>
<td>-7.15</td>
</tr>
<tr>
<td>31892_at</td>
<td>protein tyrosine phosphatase, receptor type, M</td>
<td>801.09</td>
<td>336.26</td>
<td>150.51</td>
<td>9.57</td>
<td>-5.32</td>
<td>-1.64</td>
<td>-9.12</td>
</tr>
<tr>
<td>35669_at</td>
<td>KIAA0633 protein</td>
<td>1738.34</td>
<td>360.27</td>
<td>343.94</td>
<td>22.32</td>
<td>-5.05</td>
<td>-3.3</td>
<td>-6.93</td>
</tr>
<tr>
<td>38578_at</td>
<td>tumor necrosis factor receptor superfamily, membe</td>
<td>4038.17</td>
<td>674.75</td>
<td>847.39</td>
<td>129.09</td>
<td>-4.77</td>
<td>-3.23</td>
<td>-6.94</td>
</tr>
<tr>
<td>37780_at</td>
<td>piccolo (presynaptic cytomatrix protein)</td>
<td>2856.4</td>
<td>830.13</td>
<td>601.58</td>
<td>40.43</td>
<td>-4.75</td>
<td>-2.46</td>
<td>-7.15</td>
</tr>
<tr>
<td>40570_at</td>
<td>forkhead box O1A (rhabdomyosarcoma)</td>
<td>10218.69</td>
<td>1178.1</td>
<td>2227.99</td>
<td>482.41</td>
<td>-4.59</td>
<td>-3.16</td>
<td>-7.34</td>
</tr>
<tr>
<td>39878_at</td>
<td>protocadherin 9</td>
<td>12518.61</td>
<td>2120.5</td>
<td>2816.54</td>
<td>552.51</td>
<td>-4.44</td>
<td>-2.89</td>
<td>-7.03</td>
</tr>
<tr>
<td>307_at</td>
<td>arachidonate 5-lipoxygenase</td>
<td>6743.7</td>
<td>992.9</td>
<td>1521.71</td>
<td>136.37</td>
<td>-4.43</td>
<td>-3.26</td>
<td>-5.80</td>
</tr>
<tr>
<td>38408_at</td>
<td>transmembrane 4 superfamily member 2</td>
<td>6543.7</td>
<td>1009.8</td>
<td>1489.02</td>
<td>230.77</td>
<td>-4.39</td>
<td>-3.04</td>
<td>-6.36</td>
</tr>
</tbody>
</table>
Affymetrix Web Site

http://www.affymetrix.com
Annotations are updated quarterly...
Affymetrix Support

Go to the Affymetrix support page to get the full annotations.
Support By Product

Follow the “support by product” link to “GeneChip Arrays”.

[Image of a webpage with links to support by product for GeneChip Arrays]
Affymetrix Annotations for HU133

Scroll to “Human Genome Arrays”; select “HG-U133 Set”
Affymetrix Annotations for HU133

Scroll to get a list of available files.
Affymetrix Main Annotation Files

There is one primary annotation file:

**Annotation File:** HG-U133A_2.na29.annot.csv contains the updated annotations of all the genes targeted by the microarray. (the zipped file is 11.7MB; unzipped, it is 74.5MB.)
What annotations does Affymetrix supply?

As noted earlier, HG-U133A_2.na29.annot.csv is 74.5MB. What occupies all that space?

The file contains lots of redundant information. It has information on 22,283 probesets, one per line, in 41 columns.
Description of annotation columns

**Probe Set ID.** The unique identifier that describes an Affymetrix probe set. Also used in CEL files and CDF files.

**GeneChip Array.** The chip type on which the probe set appears. The same entry is repeated for all probe sets.

**Species Scientific Name.** The scientific name of the species whose gene sequences are on the array. The same information is repeated for all probe sets.

**Annotation Date.** When the annotations were last updated. The same information is repeated for all probe sets.

**Sequence Type.** The kind of sequence used in the design of the array: can be “Consensus”, “Control”, or “Exemplar”.
Sequence Source. Where did the design sequence come from? Usually “GenBank”, but rarely (only 81 times on the HG-U133A) from “Affymetrix Proprietary Database”.

Transcript ID(Array Design). An identifier into one of several unspecified databases indicating the designed target sequence.

Target Description. Long text string describing the target, formed by combining several other fields.

Representative Public ID. For non-control sequences, a GenBank/RefSeq identifier.

Archival UniGene Cluster. The UniGene cluster identifier from the sequence at the time the array was designed (in
this case, from UniGene build 133).

**UniGene ID.** UniGene cluster identifier from the build of UniGene current at the time the annotations were updated.

**Genome Version.** The build of the human genome used for sequence alignments. The same information is repeated for all probe sets.

**Alignments.** Location of the target sequence along the human genome, in base pairs along the chromosome.

**Gene Title.** Official gene title (from UniGene or Entrez Gene).

**Gene Symbol.** Official gene symbol (either from UniGene or Entrez Gene).
**Chromosomal Location.** Location of the gene in terms of cytogenetic bands; e.g., 16p12.

**Unigene Cluster Type.** Either absent if not present in this build of UniGene (indicated by “—”), “est”, “full length”, or “est /// full length”.

**Ensembl.** The unique identifier of the target sequence in the Ensembl database.

**Entrez Gene.** The unique identifier of the target sequence in Entrez Gene (formerly LocusLink). Sequences with these identifiers tend to be better understood and more reliable than genes without them. The identifiers refer to genetic loci that have been mapped explicitly because of their connection to specific diseases or biological processes.
**SwissProt.** The SwissProt identifier of the protein product produced by the gene corresponding to the target sequence.

**EC.** Yet another database identifier.

**OMIM.** The unique identifier associated to the target sequence gene in the Online Mendelian Inheritance in Man (OMIM) database, describing the ways in which the gene is known to be associated with genetic diseases.

**RefSeq Protein ID.** The GenBank identifier of the consensus sequence for the protein produced by the target sequence.

**RefSeq Transcript ID.** The GenBank identifiers of the consensus sequences for the mRNA’s produced by the
target gene. (Alternative splicing accounts for multiples.) In many cases, this coincides with the “Representative Public ID”.

**FlyBase.** Corresponding identifier in the drosophila database.

**AGI.** Arabidopsis genome identifier.

**WormBase.** Corresponding identifier in the *C. elegans* database.

**MGI Name.** Probably the identifier in the mouse database.

**RGD Name.** Probably the identifier in the rat database.
SGD accession number. The identifier in the saccharomyces database.

Gene Ontology Biological Process. List of identifiers for annotations of the target gene into the “biological process” section of GeneOntology. More about this later.

Gene Ontology Cellular Component. Similar.

Gene Ontology Molecular Function. Similar.

Pathway. List of pathways that the target sequence is involved in.

InterPro. Another protein database.
**Trans Membrane.** Description of trans-membrane part of the protein, if known or if applicable.

**QTL.** Unknown.

**Annotation Description.** Text description of how the probe set was annotated.

**Annotation Transcript Cluster.** Unclear.

**Transcript Assignments.** Very long description of the annotations.

**Annotation Notes.** Additional comments.
Updating annotations in dChip

In order for dChip (or any other Affymetrix microarray analysis package) to use the updated annotations, you have to tell the software package where to get the information.

In the case of dChip, their online manual page tells you how to build new gene information and genome information files.

For many common chip types, the dChip web site contains up-to-date copies of these files. It’s still useful to see where the data comes from how and how you can update your own versions.
The input information files need to be downloaded to local computers. Download and unzip the Annotation CSV files for a described array type (need a free NetAffx account); make sure to use the CSV file as it is without re-saving it in Excel into a different format. Also download the three Gene Ontology (GO) structure files: function ontology, process ontology, component ontology (save in text format with name “function ontology.txt” etc.). To make common probe set files, a NetAffx Ortholog CSV file is needed to be downloaded and specified [Version 3/19/04+].

On clicking OK, dChip will parse the NetAffx annotation files to generate specified Gene information file, Genome information file and Common probe set file. If GO files are specified, the GO graphs are traced up to associate all the parent GO terms of a gene’s GO annotation terms to this gene. The most frequent occurred 1200 GO terms in this array type are indexed in the associated “gene info Gene Ontology.xls” file, and used as GO annotation terms in the output gene information file. These terms typically associate with more than 10 probe sets and are more useful for functional significance identification. The same limit applies to other annotation categories such “Protein domain”.

Requires the annotation CSV files from Affymetrix, along with three Gene Ontology files, which you can get from dChip or from the primary source.
Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Read more about the Gene Ontology...

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

Gene or protein name
GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

GO website

- GO downloads, including ontology files, annotations and the GO database.
- Tools for using GO, including OBO-Edit downloads and AmiGO
- Request new terms or ontology changes via the GO curator requests tracker; help with
Making the Gene Information file

1. Get the updated annotation CSV file from Affymetrix.

2. Get function.ontology, process.ontology, and component.ontology from GeneOntology.

3. Rename the three GeneOntology files by adding “.txt”.

Making the Gene Information file

Specify the locations of the CSV file, the GeneOntology files, and where you want the output sent. I edited the default output file name to (i) start with the standard chip name and (2) use the underscore character as a separator.
The Gene Information file

This step produces the three dChip annotation files that were described in Lecture 2.
Making the Genome Information file

Using the same input files, you can also use dChip to create a “Genome information file”, which maps genes to specific positions along the genome.

![Make information files dialog box](image)
The Genome Information file

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>chromosome</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Cytoband</th>
</tr>
</thead>
<tbody>
<tr>
<td>1007_s_at</td>
<td>chr6</td>
<td>30964144</td>
<td>30975910</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1007_s_at</td>
<td>chr6_coxHa</td>
<td>2304770</td>
<td>2316538</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1007_s_at</td>
<td>chr6</td>
<td>2103099</td>
<td>2114867</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1053_at</td>
<td>chr7</td>
<td>73283938</td>
<td>73306668</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>117_at</td>
<td>chr1</td>
<td>159761072</td>
<td>159763004</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>117_at</td>
<td>chr1</td>
<td>159842704</td>
<td>159844631</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>121_at</td>
<td>chr2</td>
<td>113691410</td>
<td>113752958</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1255_g_at</td>
<td>chr6</td>
<td>42248919</td>
<td>42255770</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1294_at</td>
<td>chr3</td>
<td>49817643</td>
<td>49826427</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1316_at</td>
<td>chr17</td>
<td>35472681</td>
<td>35499815</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1320_at</td>
<td>chr14</td>
<td>88004233</td>
<td>88086514</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1405_i_at</td>
<td>chr17</td>
<td>31222639</td>
<td>31231443</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1431_at</td>
<td>chr10</td>
<td>135190889</td>
<td>135202458</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1438_at</td>
<td>chr3</td>
<td>185762717</td>
<td>185782104</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1487_at</td>
<td>chr11</td>
<td>63829619</td>
<td>63840786</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom

GS01 0163: Analysis of Microarray Data
What is GeneOntology?

GeneOntology uses controlled vocabularies to create a directed acyclic graph (DAG; a generalized tree) that describes the kinds of functions or properties that a gene might have. There are two parts to GeneOntology:

- Annotations, maintained in databases like Entrez Gene, that describe which genes actually have which functions.

- The DAG, maintained by the GeneOntology Consortium, that describes functions and relations between them:
  1. Biological process (what)
  2. Molecular function (how)
  3. Cellular component (where)
GeneOntology: The top level

Component Ontology: Rules governing content and stylistic aspects of GO terms in the cellular component ontology.

Topics include:
- The Cell
- Protein Complexes
- Membranes and Envelopes
- Membrane Proteins

Function Ontology: Rules governing content and stylistic aspects of GO terms, standard definitions and term relationships in the molecular function ontology.

Process Ontology: Rules governing content and stylistic aspects of GO terms, standard definitions and term relationships in the biological process ontology.

Topics include:
- The Cell Cycle
- The Development Node
- Interaction Between Organisms
- Metabolism
- Regulation
GeneOntology annotations in Entrez Gene

You can find the GeneOntology annotations for individual genes in Entrez Gene. For genes with known functions, the Entrez Gene page will contain a section titled “GeneOntology”, which contains a list of the known functions for that gene.

Every GO annotation asserts that a specific gene has a specific function. As part of the design of GO, each assertion is itself annotated to explain the kinds of evidence the assertion is based on, as well as the organization or individual that supplied the annotation.
GO annotations of the androgen receptor

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>androgen binding</td>
<td>NAS</td>
</tr>
<tr>
<td>androgen receptor activity</td>
<td>NAS</td>
</tr>
<tr>
<td>androgen receptor activity</td>
<td>TAS</td>
</tr>
<tr>
<td>lipid binding</td>
<td>IEA</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>IEA</td>
</tr>
<tr>
<td>protein dimerization activity</td>
<td>NAS</td>
</tr>
<tr>
<td>receptor activity</td>
<td>IEA</td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>IEA</td>
</tr>
<tr>
<td>transcription factor activity</td>
<td>IDA</td>
</tr>
<tr>
<td>zinc ion binding</td>
<td>IEA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>androgen receptor signaling pathway</td>
<td>IEA</td>
</tr>
<tr>
<td>cell growth</td>
<td>NAS</td>
</tr>
<tr>
<td>cell proliferation</td>
<td>NAS</td>
</tr>
<tr>
<td>cell-cell signaling</td>
<td>TAS</td>
</tr>
<tr>
<td>in utero embryonic development</td>
<td>IEA</td>
</tr>
<tr>
<td>male gonad development</td>
<td>IEA</td>
</tr>
<tr>
<td>male somatic sex determination</td>
<td>IEA</td>
</tr>
<tr>
<td>prostate gland development</td>
<td>NAS</td>
</tr>
</tbody>
</table>
Gene Ontology Annotation (GOA) Database

The GOA project aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB) and International Protein Index (IPI) and is a central dataset for other major multi-species databases, such as Ensembl and NCBI.

GOA has been a member of the GO Consortium since 2001, and is responsible for the integration and release of GO annotations to the human, chicken and cow proteomes. In 2006 GOA became a central participant in the new GOC Reference Genome Annotation project and is committed to the comprehensive annotation of a set of disease-related proteins in human. With this project the GOC intends to generate a reliable set of GO annotations for the twelve selected genomes that will also empower comparative methods used in first pass annotation of other proteomes.

Because of the multi-species nature of the UniProtKB, GOA also assists in the curation of another 120,000 species. This involves electronic
**Gene Annotations: Linking Numbers to Biology**

**GO browsing**

**androgen binding**

*Accession:* GO:0005497  
*Aspect:* function  
*Synonyms:* None  
*Definition:* Interacting selectively with any androgen, male sex hormones.

**Term Lineage**

- all : all (153306)  
  - GO:0003674 : molecular_function (103037)  
    - GO:0005488 : binding (29138)  
    - GO:0042562 : hormone binding (36)  
    - GO:0005497 : androgen binding (7)  
    - GO:0005496 : steroid binding (83)  
    - GO:0005497 : androgen binding (7)

**External References**

None.

**Direct Gene Product Associations**

Get ALL associations here:

Direct Associations

Submit Query

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom  
GS01 0163: Analysis of Microarray Data
GO browsing

androgen binding Graphical View

- Layout: Vertical
- Box Color: Burlywood
- Text Color: Black

Submit Query

Diagram showing the hierarchy of gene annotations related to androgen binding.
Edges are relationships

Edges in the DAG represent two kinds of relationships:

**is_a**: Used when the child node is a special case of the parent node. For example, *hormone binding* is_a kind of *binding*.

**part_of**: Used when the child node is a component of the parent node. For example, *a membrane* is part_of a *cell*.

Genes may be annotated into different levels of the hierarchy, depending on how detailed the evidence is. In general, a gene not only has the function corresponding to the node with direct annotation, but also has every property at parent nodes up through the hierarchy.
GO annotations of the androgen receptor

GeneOntology

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>androgen binding</td>
<td>NAS, PubMed</td>
</tr>
<tr>
<td>androgen receptor activity</td>
<td>NAS, PubMed</td>
</tr>
<tr>
<td>androgen receptor activity</td>
<td>TAS, PubMed</td>
</tr>
<tr>
<td>lipid binding</td>
<td>IEA</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>IEA</td>
</tr>
<tr>
<td>protein dimerization activity</td>
<td>NAS, PubMed</td>
</tr>
<tr>
<td>receptor activity</td>
<td>IEA</td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>IEA</td>
</tr>
<tr>
<td>transcription factor activity</td>
<td>IEA</td>
</tr>
<tr>
<td>zinc ion binding</td>
<td>IEA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>androgen receptor signaling pathway</td>
<td>IEA</td>
</tr>
<tr>
<td>cell growth</td>
<td>NAS, PubMed</td>
</tr>
<tr>
<td>cell proliferation</td>
<td>NAS, PubMed</td>
</tr>
<tr>
<td>cell-cell signaling</td>
<td>TAS, PubMed</td>
</tr>
<tr>
<td>in utero embryonic development</td>
<td>IEA</td>
</tr>
<tr>
<td>male gonad development</td>
<td>IEA</td>
</tr>
<tr>
<td>male somatic sex determination</td>
<td>IEA</td>
</tr>
<tr>
<td>prostate gland development</td>
<td>NAS, PubMed</td>
</tr>
</tbody>
</table>

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom

GS01 0163: Analysis of Microarray Data
GeneOntology: Evidence Codes

**IDA**: inferred from direct assay; indicates that the annotation is based on a paper describing an experiment that directly tested this function for this gene.

**TAS**: traceable author statement; based on a review article or textbook including references to the original experiments.

**IMP**: inferred from mutant phenotype; based on experiments involving mutations, knockouts, antisense, etc.

**IPI**: inferred from physical interaction; based on assays (like co-immunoprecipitation) that demonstrate physical interactions between the gene in question and other gene products.
IGI : inferred from genetic interaction; based on experiments (such as synthetic lethals, suppressors, functional complementation) that show a genetic interaction between the gene in question and another gene

ISS : inferred from sequence or structure similarity; based on BLAST results that have been reviewed for accuracy by a curator

IEP : inferred from expression pattern; based on Northern, Western, or microarray experiments that reveal information about the timing or location of expression

NAS : non-traceable author statement; statements in papers (abstract, introduction, discussion) that a curator cannot trace to another publication
IEA : inferred from electronic annotation; based on sequence similarity searches or database records that have not been reviewed by a curator

IC : inferred by curator; even though no direct evidence is available, the property can reasonably be inferred by the curator. For example, it is reasonable to infer from direct evidence of “transcription factor activity” that the gene product is found in the nucleus

ND : no biological data available; only used for annotations to “unknown”

NR : not recorded; used only for annotations created before curators started adding evidence codes
Quality of evidence

The evidence codes fall into a rough hierarchy indicating how strongly the annotation of function should be believed.

1. IDA, TAS
2. IMP, IPI, IGI
3. ISS, IEP
4. NAS
5. IEA
6. IC
Using GeneOntology in dChip

After running a sample comparison to find interesting genes, use the menu item “Tools” → “Gene Function Enrichment”.

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom
Using GeneOntology in dChip

For the gene list file, select the “compare result” file produced previously. It may be a good idea to use the “Options” to set the cutoff for significant p-values.
Using GeneOntology in dChip

The results are available in a few seconds.

Gene function enrichment analysis:
- C1: number of genes in a cluster or list that have this annotation term
- C2: number of annotated genes in this cluster or list
- C3: number of all genes on array that have this annotation term
- C4: number of all annotated genes on array
- P-value: binomial approximated p-value for hypergeometric distribution

Gene Ontology:
- 0 reported significant, 0 expected false positive (12 terms assessed for enrichment at p-value threshold 0.01000)

Protein Domain:
- 0 reported significant, 0 expected false positive (10 terms assessed for enrichment at p-value threshold 0.01000)

Pathway:
- 1 reported significant, 0 expected false positive (1 terms assessed for enrichment at p-value threshold 0.01000)

Chromosomes:
- 0 reported significant, 0 expected false positive (7 terms assessed for enrichment at p-value threshold 0.01000)
What do the results look like?

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene Description</th>
<th>Baseline Mean</th>
<th>Baseline Exp Mean</th>
<th>Experiment Mean</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>40936_at</td>
<td>cysteine-rich motor neuron 1</td>
<td>7994</td>
<td>564</td>
<td>5144</td>
<td>612</td>
</tr>
<tr>
<td>1485_at</td>
<td>EphA7</td>
<td>243</td>
<td>28</td>
<td>133</td>
<td>14</td>
</tr>
<tr>
<td>2057_g_at</td>
<td>fibroblast growth factor receptor 1 (fms-related)</td>
<td>5421</td>
<td>430</td>
<td>2717</td>
<td>100</td>
</tr>
<tr>
<td>1964_g_at</td>
<td>fms-related tyrosine kinase 1 (vascular end)</td>
<td>1555</td>
<td>167</td>
<td>982</td>
<td>51</td>
</tr>
<tr>
<td>1545_g_at</td>
<td>fms-related tyrosine kinase 1 (vascular end)</td>
<td>745</td>
<td>85</td>
<td>471</td>
<td>16</td>
</tr>
<tr>
<td>34583_at</td>
<td>fms-related tyrosine kinase 3</td>
<td>9522</td>
<td>1513</td>
<td>16788</td>
<td>784</td>
</tr>
<tr>
<td>1065_at</td>
<td>fms-related tyrosine kinase 3</td>
<td>8414</td>
<td>1696</td>
<td>15615</td>
<td>933</td>
</tr>
<tr>
<td>40480_s_at</td>
<td>FYN oncogene related to SRC, FGR, YES</td>
<td>5038</td>
<td>514</td>
<td>3304</td>
<td>326</td>
</tr>
<tr>
<td>34877_at</td>
<td>Janus kinase 1 (a protein tyrosine kinase)</td>
<td>15776</td>
<td>843</td>
<td>10823</td>
<td>834</td>
</tr>
<tr>
<td>41594_at</td>
<td>Janus kinase 1 (a protein tyrosine kinase)</td>
<td>6687</td>
<td>345</td>
<td>4360</td>
<td>301</td>
</tr>
<tr>
<td>1457_at</td>
<td>Janus kinase 1 (a protein tyrosine kinase)</td>
<td>3098</td>
<td>197</td>
<td>1886</td>
<td>177</td>
</tr>
<tr>
<td>33238_at</td>
<td>lymphocyte-specific protein tyrosine kinase</td>
<td>3794</td>
<td>572</td>
<td>1938</td>
<td>258</td>
</tr>
<tr>
<td>1988_at</td>
<td>platelet-derived growth factor receptor, alpha</td>
<td>14547</td>
<td>602</td>
<td>10367</td>
<td>351</td>
</tr>
<tr>
<td>36117_at</td>
<td>PTK2 protein tyrosine kinase 2</td>
<td>3730</td>
<td>242</td>
<td>2613</td>
<td>117</td>
</tr>
<tr>
<td>37756_at</td>
<td>RYK receptor-like tyrosine kinase</td>
<td>1155</td>
<td>129</td>
<td>399</td>
<td>48</td>
</tr>
<tr>
<td>639_at</td>
<td>RYK receptor-like tyrosine kinase</td>
<td>2284</td>
<td>107</td>
<td>1865</td>
<td>48</td>
</tr>
<tr>
<td>572_at</td>
<td>TTK protein kinase</td>
<td>1309</td>
<td>128</td>
<td>792</td>
<td>76</td>
</tr>
<tr>
<td>1874_at</td>
<td>v-yes-1 Yamaguchi sarcoma viral oncogene</td>
<td>1438</td>
<td>283</td>
<td>496</td>
<td>32</td>
</tr>
<tr>
<td>32616_at</td>
<td>v-yes-1 Yamaguchi sarcoma viral oncogene</td>
<td>3247</td>
<td>219</td>
<td>4842</td>
<td>498</td>
</tr>
<tr>
<td>2024_s_at</td>
<td>v-yes-1 Yamaguchi sarcoma viral oncogene</td>
<td>1913</td>
<td>141</td>
<td>2960</td>
<td>322</td>
</tr>
<tr>
<td>1402_at</td>
<td>v-yes-1 Yamaguchi sarcoma viral oncogene</td>
<td>4141</td>
<td>289</td>
<td>6292</td>
<td>581</td>
</tr>
</tbody>
</table>

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom  GS01 0163: Analysis of Microarray Data
Interpreting the Results

Each group of entries in the results file is introduced by a line like:

```
Found 21 Gene Ontology "protein tyrosine kinase" genes in a list with 391 annotated genes (all: 157/7685, PValue: 0.000042)
```  

The part within quotation marks is the name of the Gene Ontology category that was found to be significantly overrepresented among the differentially expressed genes.

What do the numbers tell us?
1. There were 7685 probesets on the array with some kind of GeneOntology annotation.

2. There were 391 differentially expressed probesets that had some kind of GeneOntology annotation.

3. Of all the annotated probe sets, 157 had the “protein tyrosine kinase” function.

4. Of the selected annotated probe sets, 21 had the “protein tyrosine kinase” function.

The p-value comes from modeling the data using a hypergeometric distribution, which means it is the same value produced by Fisher’s Exact Test on a $2 \times 2$ contingency table.
What's wrong with the results?

First, the p-values haven't been adjusted for multiple testing. Second, we cannot tell if the software has accounted for the fact that the GeneOntology categories form a DAG. In particular, a gene with “protein tyrosine kinase” activity also inherits every annotation above it in the DAG.
What’s wrong with the results?

Third, by working with probe sets instead of genes, the counts are wrong.
What alternatives are there?

GoMiner is a tool for biological interpretation of 'omic' data – including data from gene expression microarrays. Omic experiments often generate lists of dozens or hundreds of genes that differ in expression between samples, raising the question

What does it all mean biologically?

To answer this question, GoMiner leverages the Gene Ontology (GO) to identify the biological processes, functions and components represented in these lists. Instead of analyzing microarray results with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories. The insights gained through GoMiner can generate hypotheses to guide additional research.

To get started using GoMiner

- Read the Instructions, and verify that your environment satisfies the system requirements
- Download the GoMiner program file, gominer.jar
- Read the Quick Start and try out the sample files
**Core Application**

- Verify that your machine satisfies the system requirements
- Download the GoMiner program file, `gominer.jar`
- Install the optional components (if any) listed below you want to use.
- Either double-click the jar file, or run "java -jar gominer.jar" from the command line
- For large input files (>10,000 genes) run "java -Xms256M -Xmx256M -jar gominer.jar"
  This will allocate more memory to run the application.
- For optimal performance, install a local copy of the server and database.

**Install Helper Applications**

The first two helper applications are web browser components, and provide additional visualization features.

**Adobe SVG Viewer**
GoMiner: Getting Started

You need a machine with

- Java 1.3 or higher
- Windows 98 or higher, Mac OS X or higher, Solaris, Linux, or FreeBSD
- High-speed internet access

Download the GoMiner Java code, install it, and double-click on it to start the program.

Then go to “File” → “Load GO Terms” and click “OK”. Wait a few minutes while the program loads the GeneOntology information from the NCI.
GoMiner Start
GoMiner: GO terms loaded
GoMiner as GO browser
Getting array data into GoMiner

1. Go to “Data Source” and select “UniProt (Hs)” to restrict to human gene annotations.

2. Need a file listing all genes in the experiment, one HUGO symbol per line. Use the “Browse” button, and then click “Query Gene File” to load this information. This takes some time.

3. Need a file containing a list of genes that changed. Can be one HUGO symbol per line. Optionally, you can include a second column with 1 (overexpressed) or -1 (under). Use “Browse” and “Query Changed Gene File” to load this data.

Note: GeneLink or Source can convert from various gene ids.
to HUGO symbols.
GoMiner with array gene list loaded

![GoMiner interface screenshot](image-url)

### Gene Annotations: Linking Numbers to Biology

**GoMiner with array gene list loaded**

#### Summary View
- **Category Name**: ATP-dependent hel.
- **P-Chng**: 0.0005
- **P-Undr**: 1.0000
- **P-Ovr**: 1.0000
- **Tet**: 11
- **Chng**: 0
- **Undr**: 0
- **Ovr**: 0
- **Datagene ID**: GO:000880

**Selected Gene View**
- **Category Name**: ATP-dependent hel.
- **P-Chng**: 0.0005
- **P-Undr**: 1.0000
- **P-Ovr**: 1.0000
- **Tet**: 11
- **Chng**: 0
- **Undr**: 0
- **Ovr**: 0
- **Datagene ID**: GO:000880

**Operations**
- **File Name**: C:\Source\GoMinerExample\total_gene
- **Operations**: Query Genes File, Reset All, Query Changed Genes File, Reset Changed

© Copyright 2004–2009, KR Coombes, KA Baggerly and BM Broom

GS01 0163: Analysis of Microarray Data
GoMiner with changed gene list loaded
GoMiner subgraphs
GoMiner subgraphs
Interpreting GoMiner results

Enrichment is computed as

\[
\frac{\text{changed genes in category}}{\text{total genes in category}} = \frac{\text{changed genes on array}}{\text{all genes on array}}
\]

Statistical evidence of enrichment is based on a Fisher exact test.
Interpreting GoMiner results

The p-values from the Fisher test are not corrected for multiple testing, but they should be since one is potentially looking at all GO categories. The categories are not independent, so it is not clear exactly how one should correct for multiple testing.

If we filter genes before testing differential expression (e.g., by removing low expressing or low variance genes), should those genes be included in the “query gene file” for the experiment?

The Fisher exact test isn’t completely appropriate, since genes can have overlapping annotations into the GO DAG.

No existing test exploits the GO evidence codes.