GS01 0163
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 betwen the two groups of samples, or (b) do they identify specific biological pathways or networks that are perturbed?

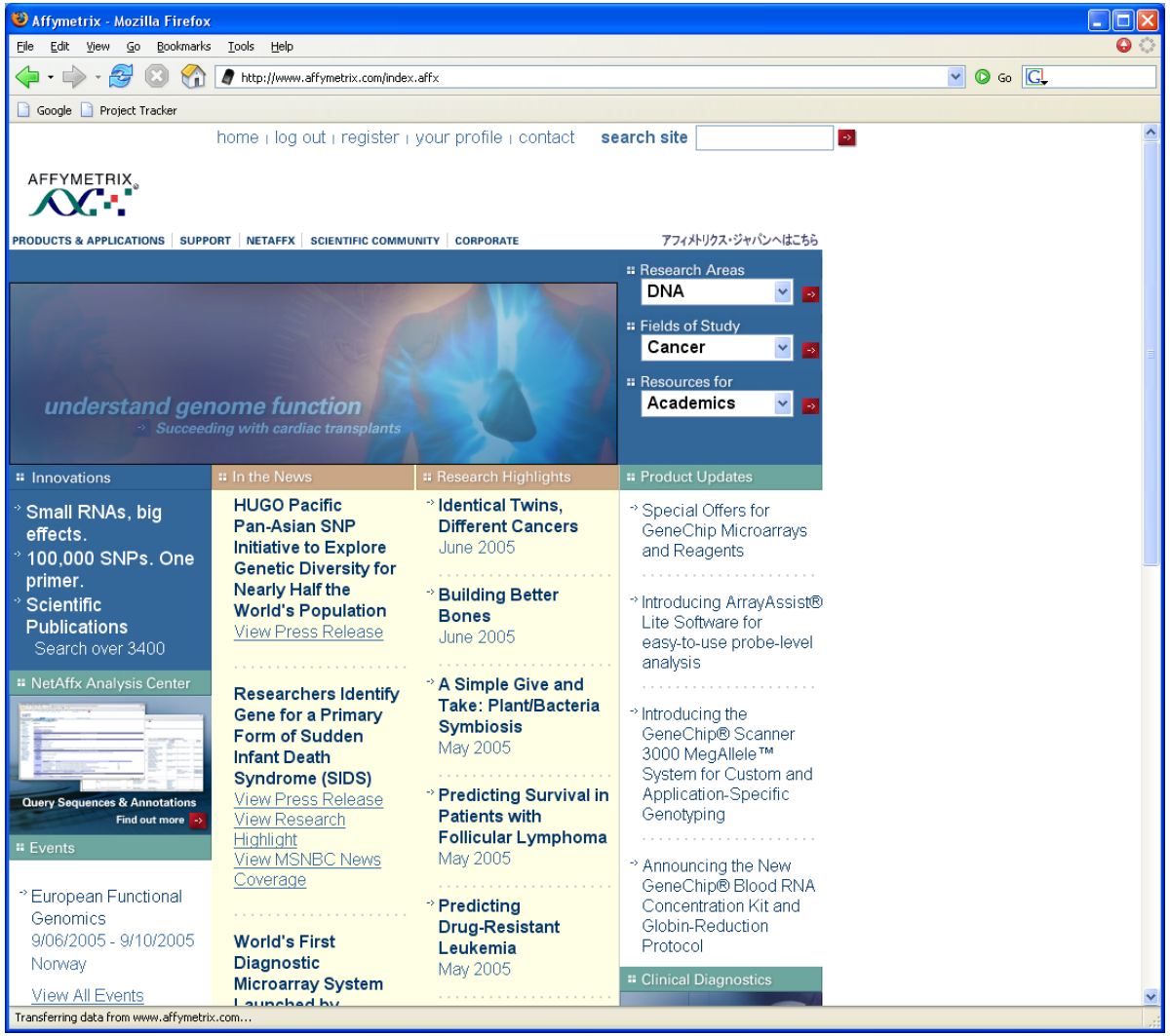
List of Differentially Expressed Genes

Microsoft Excel - affyShortCourse compare result.xls

	A	B	AA	AB	AU	AV	AW	AX	AY
12	probe set	gene	baseline mean	baseline	experiment mean	experiment	fold change	lower bo	upper bo
13	37680_at	A kinase (PRKA) anchor protein (gravin) 12	2973.7	560.63	148.29	24.19	-20.05	-12.93	-30.28
14	1325_at	MAD, mothers against decapentaplegic homolog 1	7759.92	1390.4	595.18	64.06	-13.04	-8.89	-18.03
15	37280_at	MAD, mothers against decapentaplegic homolog 1	9124.17	1538.9	702.89	37.85	-12.98	-9.29	-16.88
16	37908_at	guanine nucleotide binding protein 11	2160.91	565.93	226.99	58.16	-9.52	-4.92	-18.23
17	34194_at	Homo sapiens mRNA; cDNA DKFZp564B076 (fron	962.11	296.29	107.48	34.97	-8.95	-3.95	-21.14
18	753_at	nidogen 2 (osteonidogen)	2558.48	890.45	304.16	22.09	-8.41	-3.58	-13.49
19	1992_at	fragile histidine triad gene	1742.98	252.98	209.02	29.64	-8.34	-5.92	-11.72
20	1488_at	protein tyrosine phosphatase, receptor type, K	4128.67	1140	572.2	38.89	-7.22	-3.91	-10.70
21	1077_at	recombination activating gene 1	6927.92	1443.9	1021.43	204.85	-6.78	-4.09	-11.13
22	33910_at	Homo sapiens mRNA; cDNA DKFZp564P116 (fron	460.85	209.6	72.66	7.64	-6.34	-1.59	-11.49
23	34800_at	leucine-rich repeats and immunoglobulin-like domai	5255.48	907	899.41	189.08	-5.84	-3.74	-9.53
24	35614_at	transcription factor-like 5 (basic helix-loop-helix)	7264.11	1378.1	1248.25	122.02	-5.82	-3.9	-8.05
25	41266_at	integrin, alpha 6	7923.59	1222.5	1445.79	200.87	-5.48	-3.84	-7.73
26	37343_at	inositol 1,4,5-triphosphate receptor, type 3	5231.99	747.28	966.99	97.72	-5.41	-3.98	-7.15
27	31892_at	protein tyrosine phosphatase, receptor type, M	801.09	336.26	150.51	9.57	-5.32	-1.64	-9.12
28	35669_at	KIAA0633 protein	1738.34	360.27	343.94	22.32	-5.05	-3.3	-6.93
29	38578_at	tumor necrosis factor receptor superfamily, membe	4038.17	674.75	847.39	129.09	-4.77	-3.23	-6.94
30	37780_at	piccolo (presynaptic cytomatrix protein)	2856.4	830.13	601.56	40.43	-4.75	-2.46	-7.15
31	40570_at	forkhead box O1A (rhabdomyosarcoma)	10218.69	1178.1	2227.99	482.41	-4.59	-3.16	-7.34
32	39878_at	protocadherin 9	12518.61	2120.5	2816.54	552.51	-4.44	-2.89	-7.03
33	307_at	arachidonate 5-lipoxygenase	6743.7	992.9	1521.71	136.37	-4.43	-3.26	-5.80
34	38408_at	transmembrane 4 superfamily member 2	6543.7	1009.8	1489.02	230.77	-4.39	-3.04	-6.36

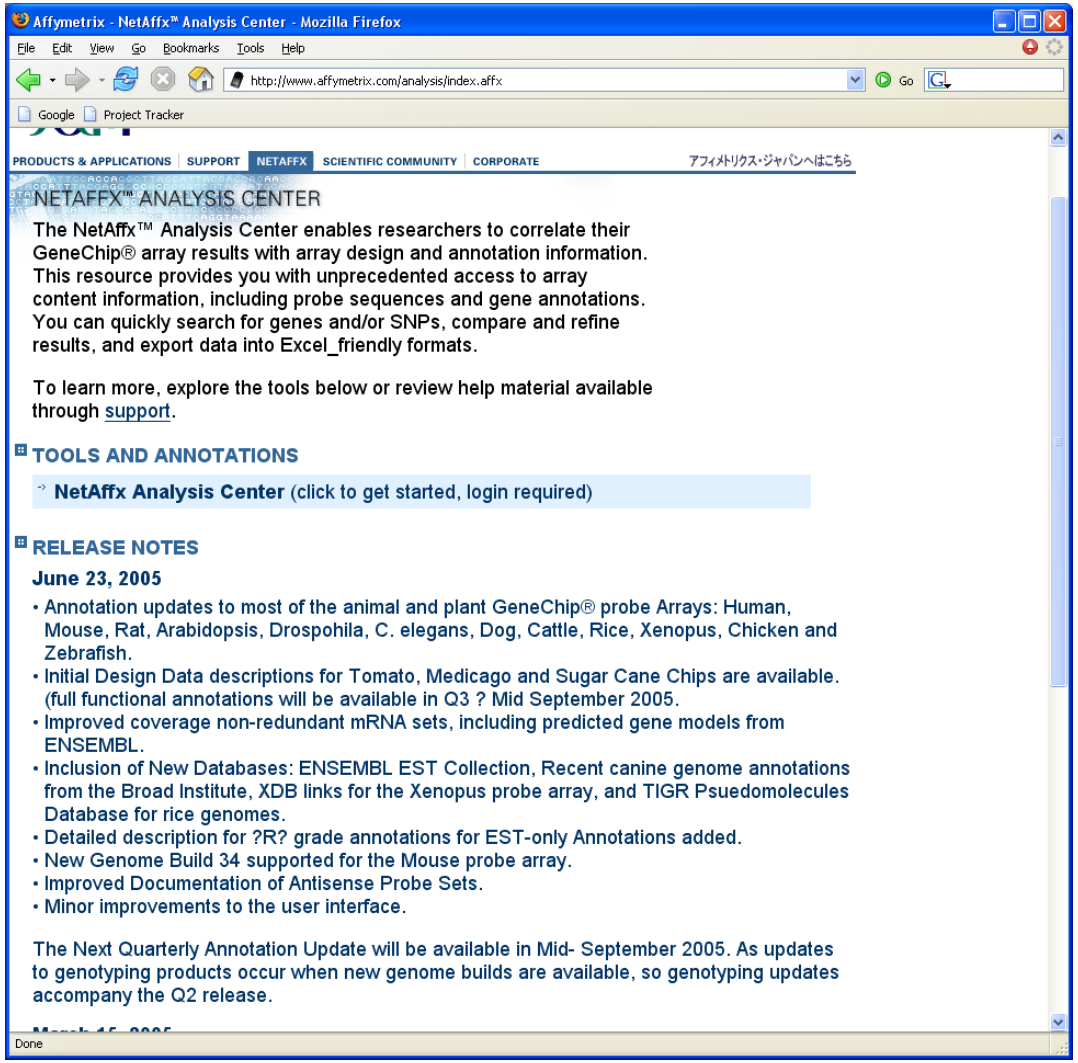
Affymetrix Web Site

http://www.affymetrix.com



NETAFFX

Annotations are updated quarterly...

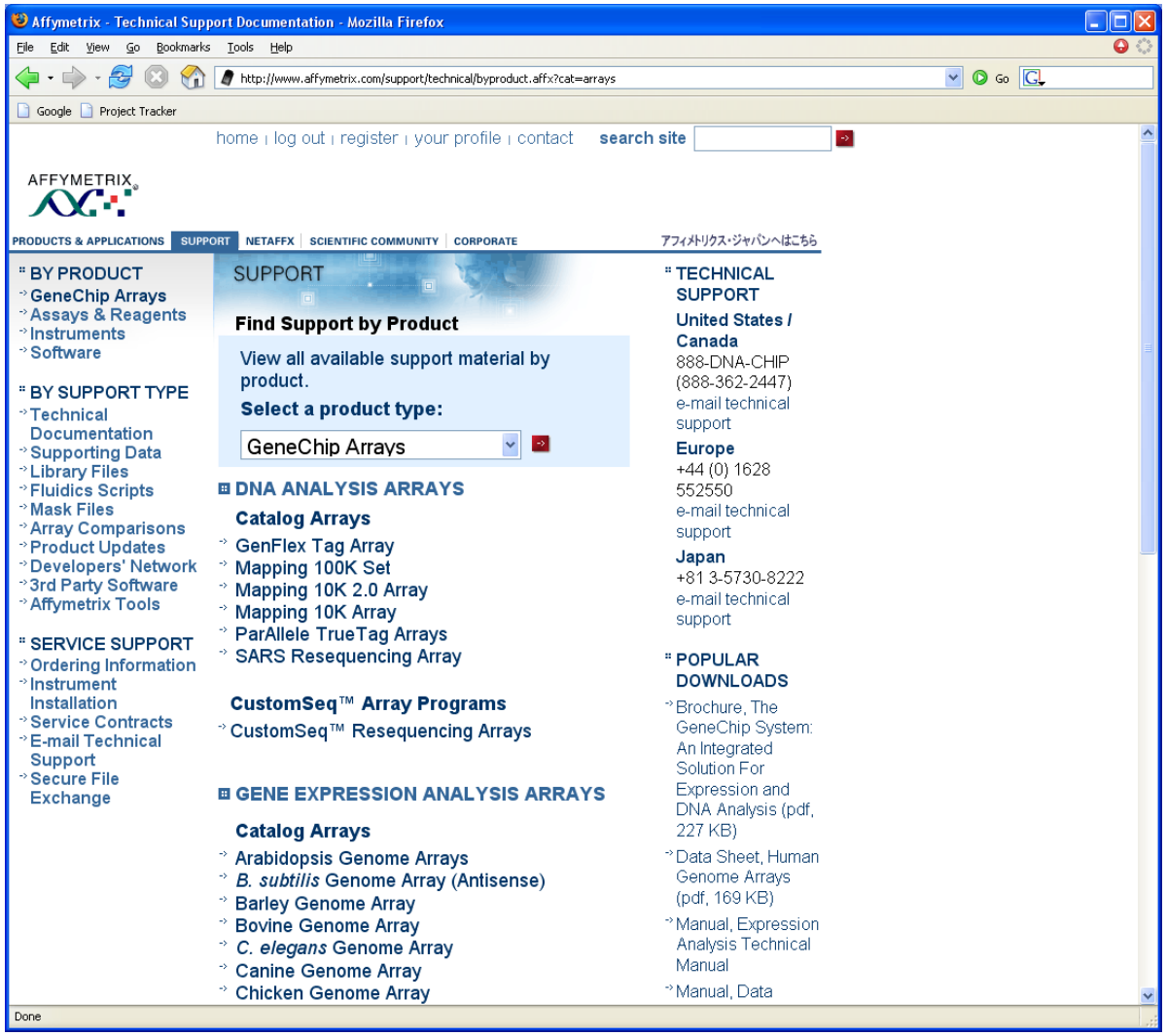


Affymetrix Support

Go to the Affymetrix support page to get the full annotations.

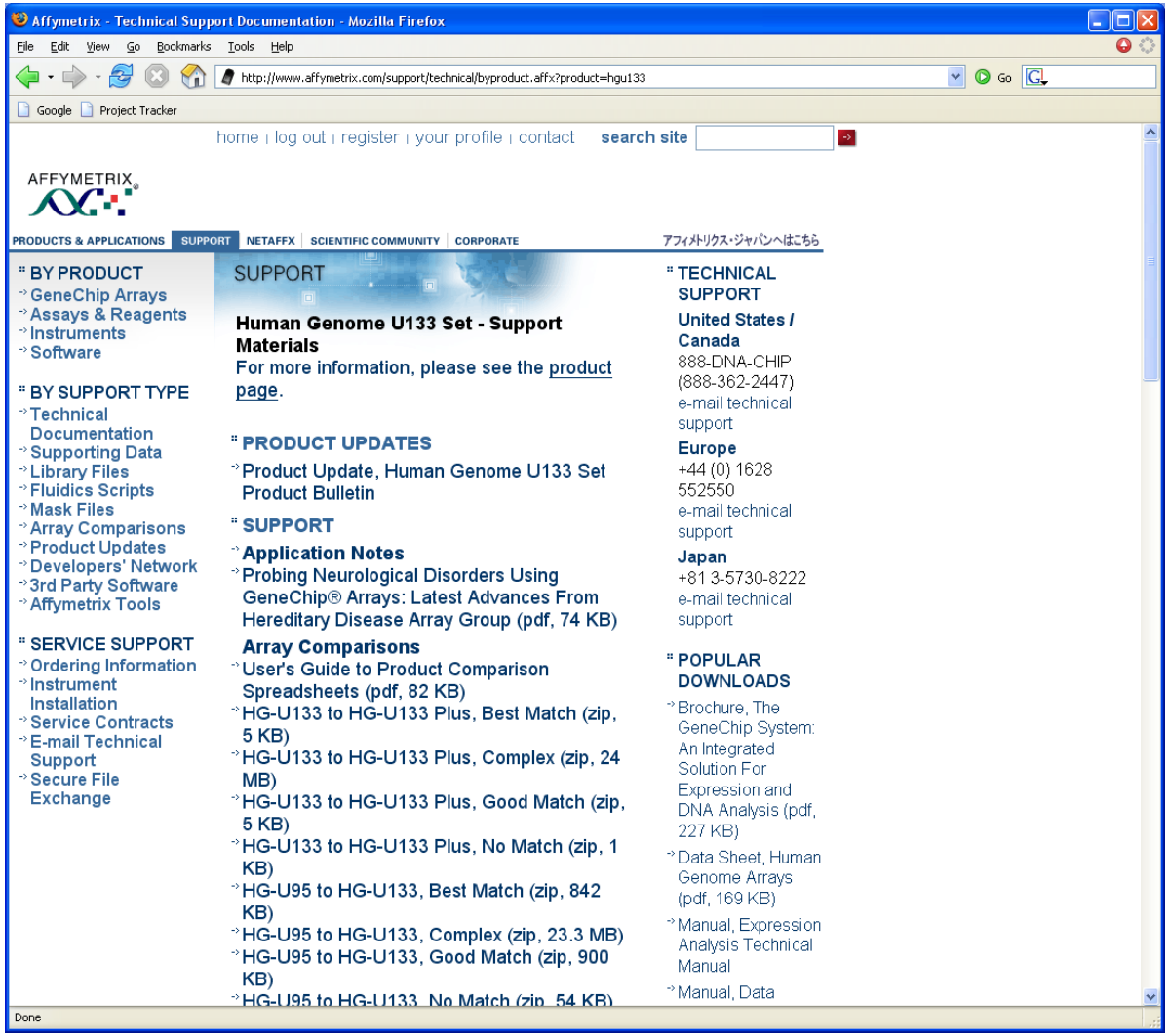
Support By Product

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



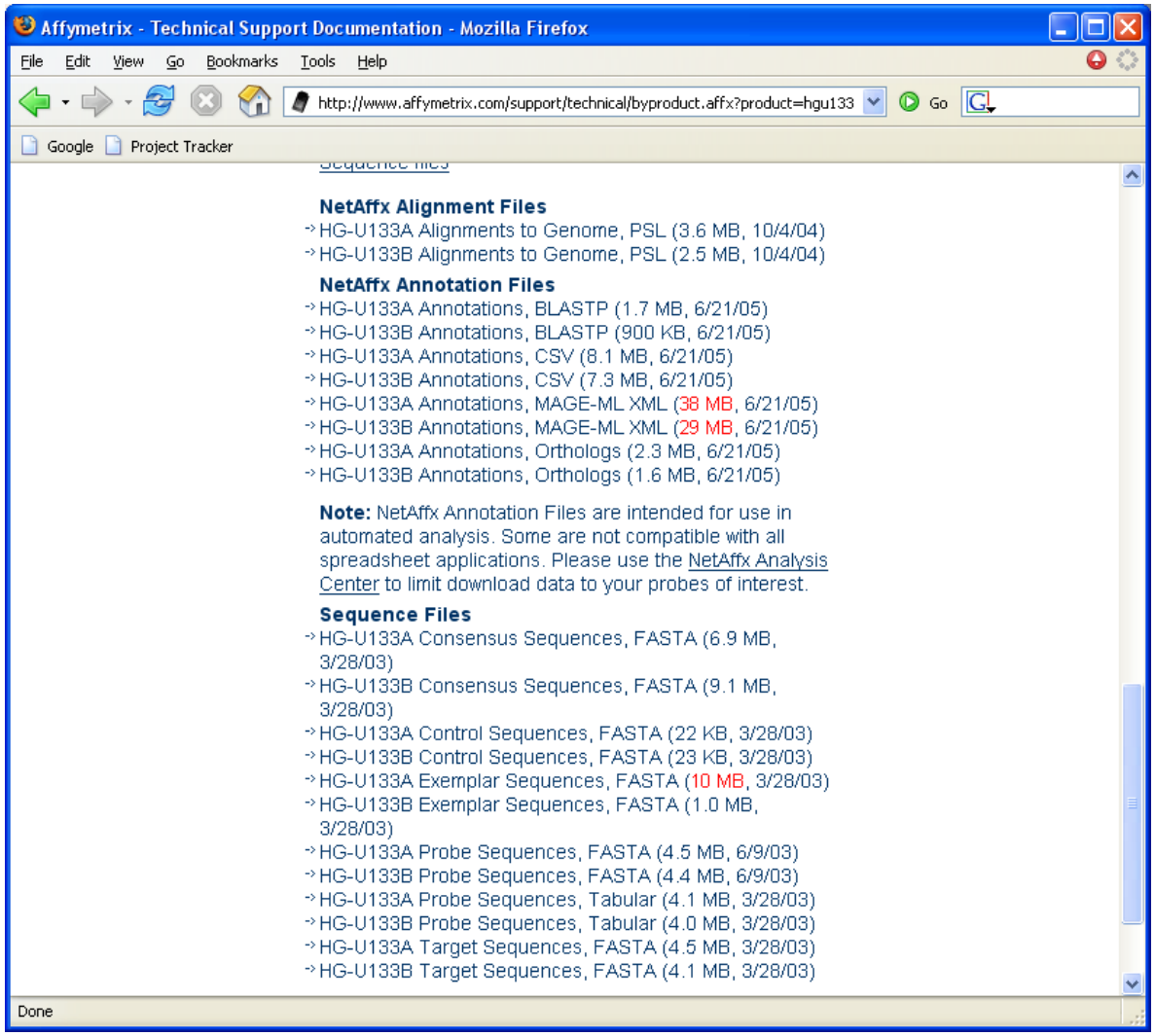
Affymetrix Annotations for HU133

Scroll to “Human Genome Arrays”; select “HG-U133 Set”



Affymetrix Annotations for HU133

Scroll to get a list of available files.



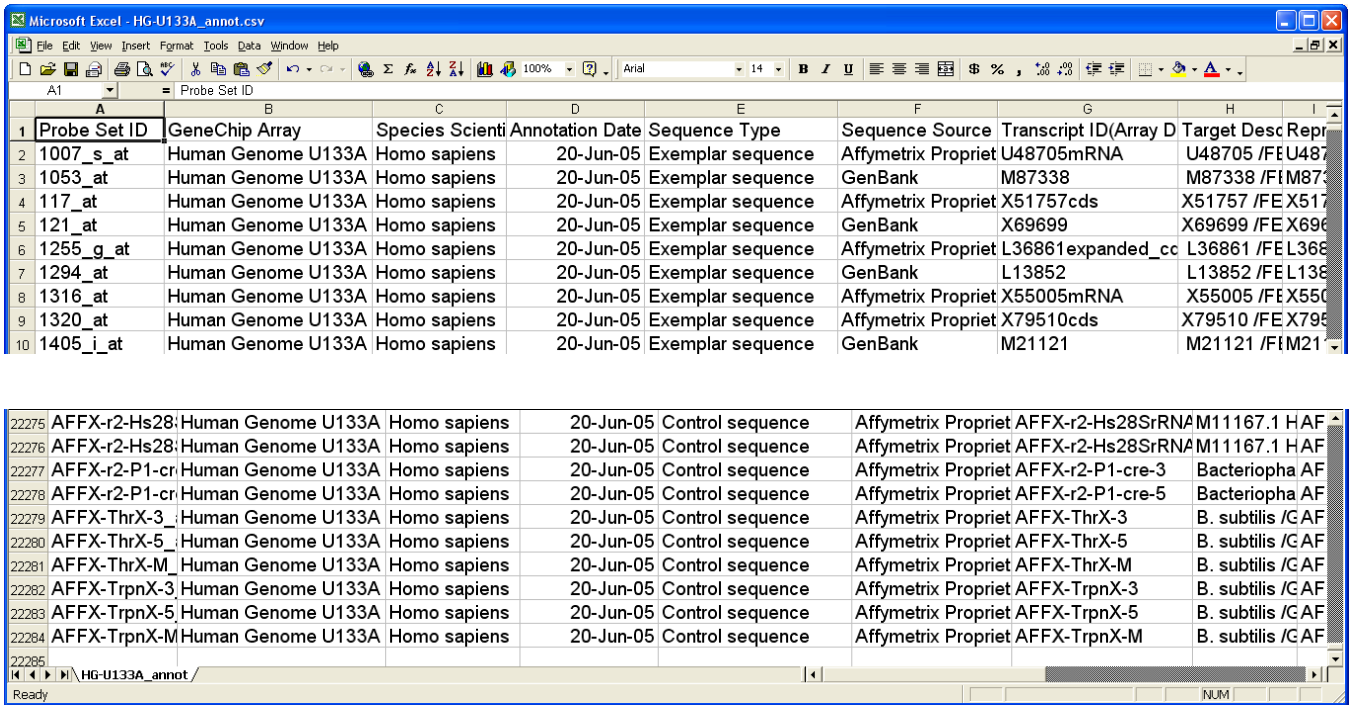
Affymetrix Main Annotation Files

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?



	A	B	C	D	E	F	G	H	I
1	Probe Set ID	GeneChip Array	Species Scienti	Annotation Date	Sequence Type	Sequence Source	Transcript ID(Array D	Target Desc	Repr
2	1007_s_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	Affymetrix Propriet	U48705mRNA	U48705 /FE	U487
3	1053_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	GenBank	M87338	M87338 /FE	M87
4	117_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	Affymetrix Propriet	X51757cds	X51757 /FE	X517
5	121_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	GenBank	X69699	X69699 /FE	X696
6	1255_g_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	Affymetrix Propriet	L36861expanded_cc	L36861 /FE	L368
7	1294_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	GenBank	L13852	L13852 /FE	L138
8	1316_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	Affymetrix Propriet	X55005mRNA	X55005 /FE	X550
9	1320_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	Affymetrix Propriet	X79510cds	X79510 /FE	X795
10	1405_i_at	Human Genome U133A	Homo sapiens	20-Jun-05	Exemplar sequence	GenBank	M21121	M21121 /FE	M21

22275	AFFX-r2-Hs28	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-r2-Hs28SrRNA	M11167.1 H	AF
22276	AFFX-r2-Hs28	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-r2-Hs28SrRNA	M11167.1 H	AF
22277	AFFX-r2-P1-cr	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-r2-P1-cre-3	Bacteriopha	AF
22278	AFFX-r2-P1-cr	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-r2-P1-cre-5	Bacteriopha	AF
22279	AFFX-ThrX-3_	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-ThrX-3	B. subtilis /G	AF
22280	AFFX-ThrX-5_	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-ThrX-5	B. subtilis /G	AF
22281	AFFX-ThrX-M_	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-ThrX-M	B. subtilis /G	AF
22282	AFFX-TrpnX-3	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-TrpnX-3	B. subtilis /G	AF
22283	AFFX-TrpnX-5	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-TrpnX-5	B. subtilis /G	AF
22284	AFFX-TrpnX-M	Human Genome U133A	Homo sapiens	20-Jun-05	Control sequence	Affymetrix Propriet	AFFX-TrpnX-M	B. subtilis /G	AF
22285									

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 asdsociated to the tartget 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 identifer 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.

dChip Manual on Gene Information

Gene information file - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://biosun1.harvard.edu/complab/dchip/gene_info_file.htm#make_info_file

Google Project Tracker

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 acocount); 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".

Done

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

http://www.geneontology.org

the Gene Ontology - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.geneontology.org/

Google | Project Tracker | Entrez-PubMed | MDACC Bioinfo | Microarray Core Faci... | BioinformaticsWiki

agi genome - Google Search | the Gene Ontology

the Gene Ontology

Search

gene or protein name go!

- Open menus
- Home
- FAQ
- Downloads
- Tools
- Documentation
- About GO
- Contact GO
- Site Map

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:

GO!

gene or protein name GO term or ID

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

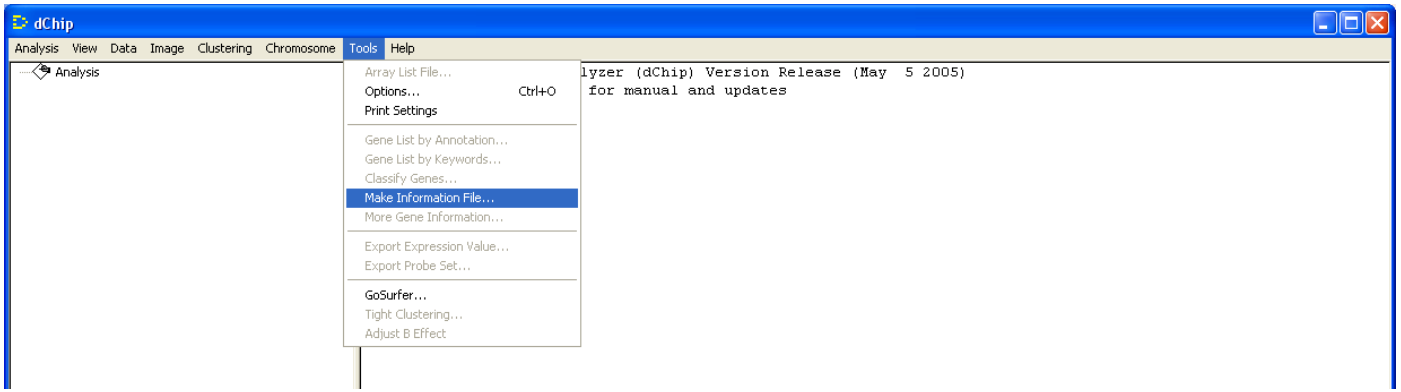
GO website

- 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

Done

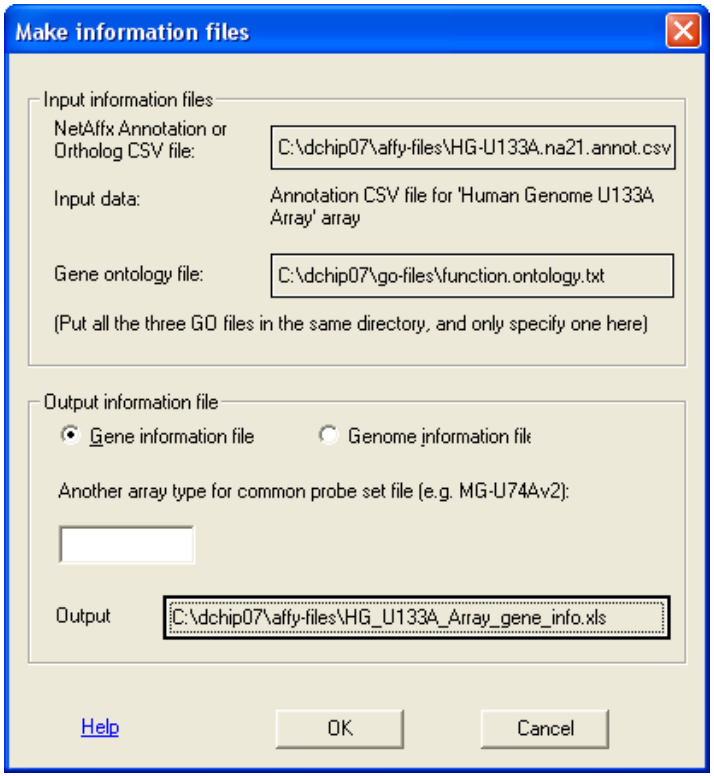
Making the Gene Information file

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".
4. Use "Tools" $-$ $>$ "Make information file" in `dChip`.

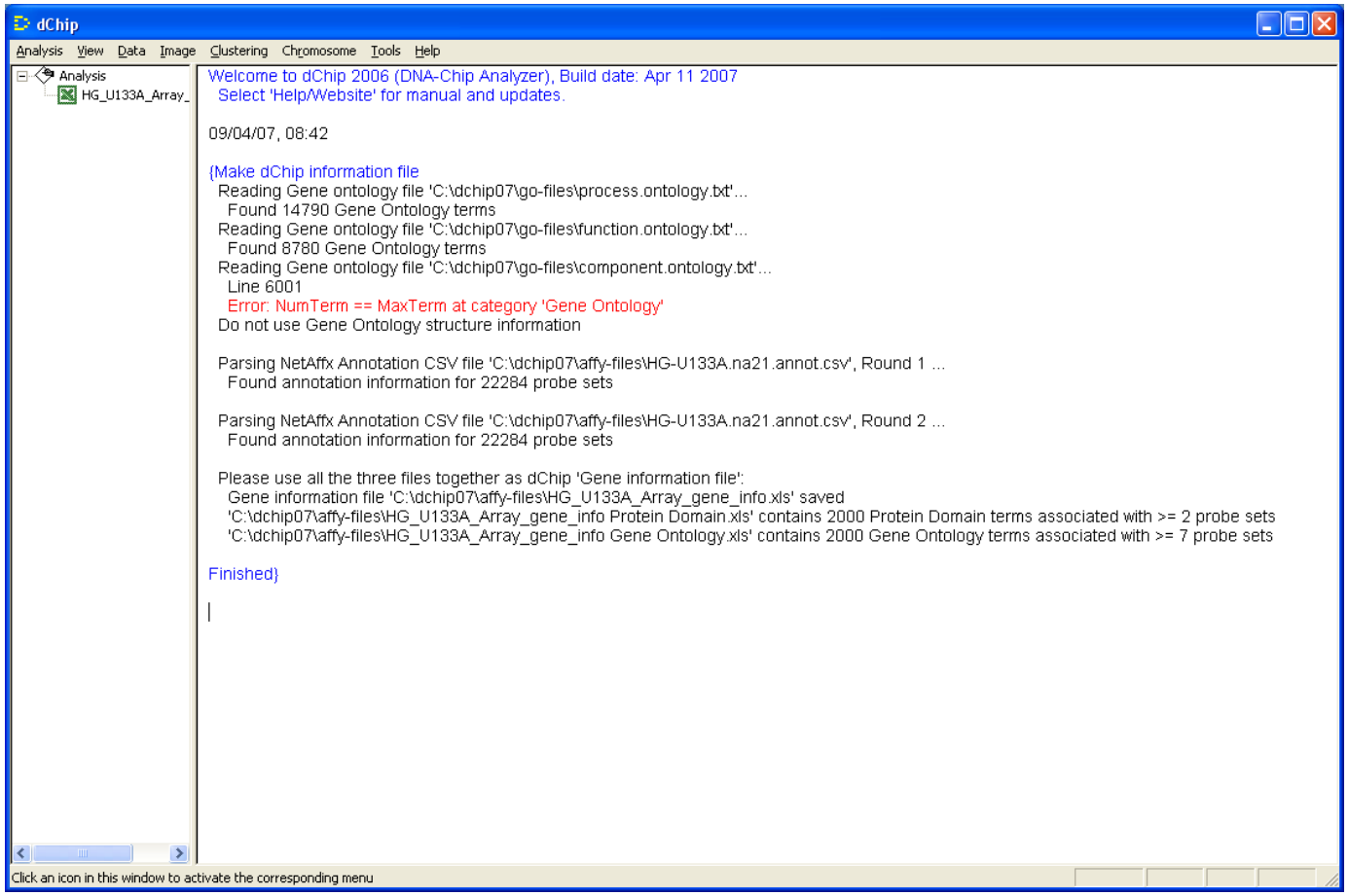


Making the Gene Information file

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.



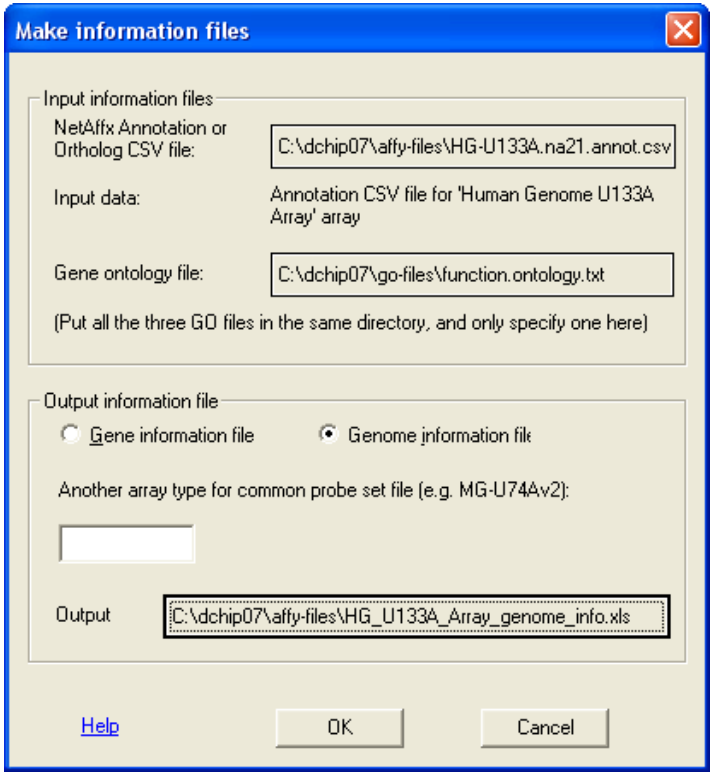
The Gene Information file



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.



The Genome Information file

Microsoft Excel - HG_U133A_Array_genome_info.xls

	A	B	C	D	E	F
1	Probe Set	chromosome	Start	End	Strand	Cytoband
2	1007_s_at	chr6	30964144	30975910	+	
3	1007_s_at	chr6_cox_ha	2304770	2316538	+	
4	1007_s_at	chr6_	2103099	2114867	+	
5	1053_at	chr7	73283938	73306668	-	
6	117_at	chr1	159761072	159763004	+	
7	117_at	chr1	159842704	159844631	+	
8	121_at	chr2	113691410	113752958	-	
9	1255_g_at	chr6	42248919	42255770	+	
10	1294_at	chr3	49817643	49826427	-	
11	1316_at	chr17	35472681	35499815	+	
12	1320_at	chr14	88004233	88086514	-	
13	1405_i_at	chr17	31222639	31231443	-	
14	1431_at	chr10	135190889	135202458	+	
15	1438_at	chr3	185762717	185782104	+	
16	1487_at	chr11	63829619	63840786	+	

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

Gene Ontology Documentation - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.geneontology.org/GO.contents.doc.shtml

Google | Project Tracker | Entrez-PubMed | MDACC Bioinfo | Microarray Core Faci... | BioinformaticsWiki

Contact GO
Site Map

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**

Done

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

Entrez Gene: AR androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) [...

File Edit View Go Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=367&ordinalpos=1&itool=Entrez

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... BioinformaticsWiki

GeneOntology

Provided by GOA

Function	Evidence	
androgen binding	NAS	PubMed
androgen receptor activity	NAS	PubMed
androgen receptor activity	TAS	PubMed
lipid binding	IEA	
metal ion binding	IEA	
protein dimerization activity	NAS	PubMed
receptor activity	IEA	
sequence-specific DNA binding	IEA	
transcription factor activity	IDA	PubMed
zinc ion binding	IEA	

Process	Evidence	
androgen receptor signaling pathway	IEA	
cell growth	NAS	PubMed
cell proliferation	NAS	PubMed
cell-cell signaling	TAS	PubMed
in utero embryonic development	IEA	
male gonad development	IEA	
male somatic sex determination	IEA	
prostate gland development	NAS	PubMed

Done

http://www.ebi.ac.uk/GOA/

Gene Ontology Annotation (GOA) @ EBI - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.ebi.ac.uk/GOA/

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... BioinformaticsWiki

EMBL-EBI EB-eye Search All Databases Enter Text Here Reset Advanced Search Give us feedback

Tools EBI Groups Training Industry About US Site Index Help

- GOA Home
- New to GOA?
- Downloads
- Searching GOA
- Publications
- Contribute Data to GOA
- Release Archive
- GOA Staff
- FAQ
- Contact Us
- Acknowledgements

Popular pages

Downloads

Latest statistics for

- UniProt
- Human
- Mouse
- Rat
- Arabidopsis

EBI › Databases › Biological Ontologies › GOA

Gene Ontology Annotation (GOA) Database

Search GO

Search GO term names/synonyms

Search all ontologies

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

Done

GO browsing

AmiGO! Your friend in the Gene Ontology. - Netscape

File Edit View Go Bookmarks Tools Window Help

http://www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&depth=1&query=GO:0005497

Search

Home Radio Bookmarks Google Bioinformatics ...

AmiGO

androgen binding

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

Term Lineage

Graphical View

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

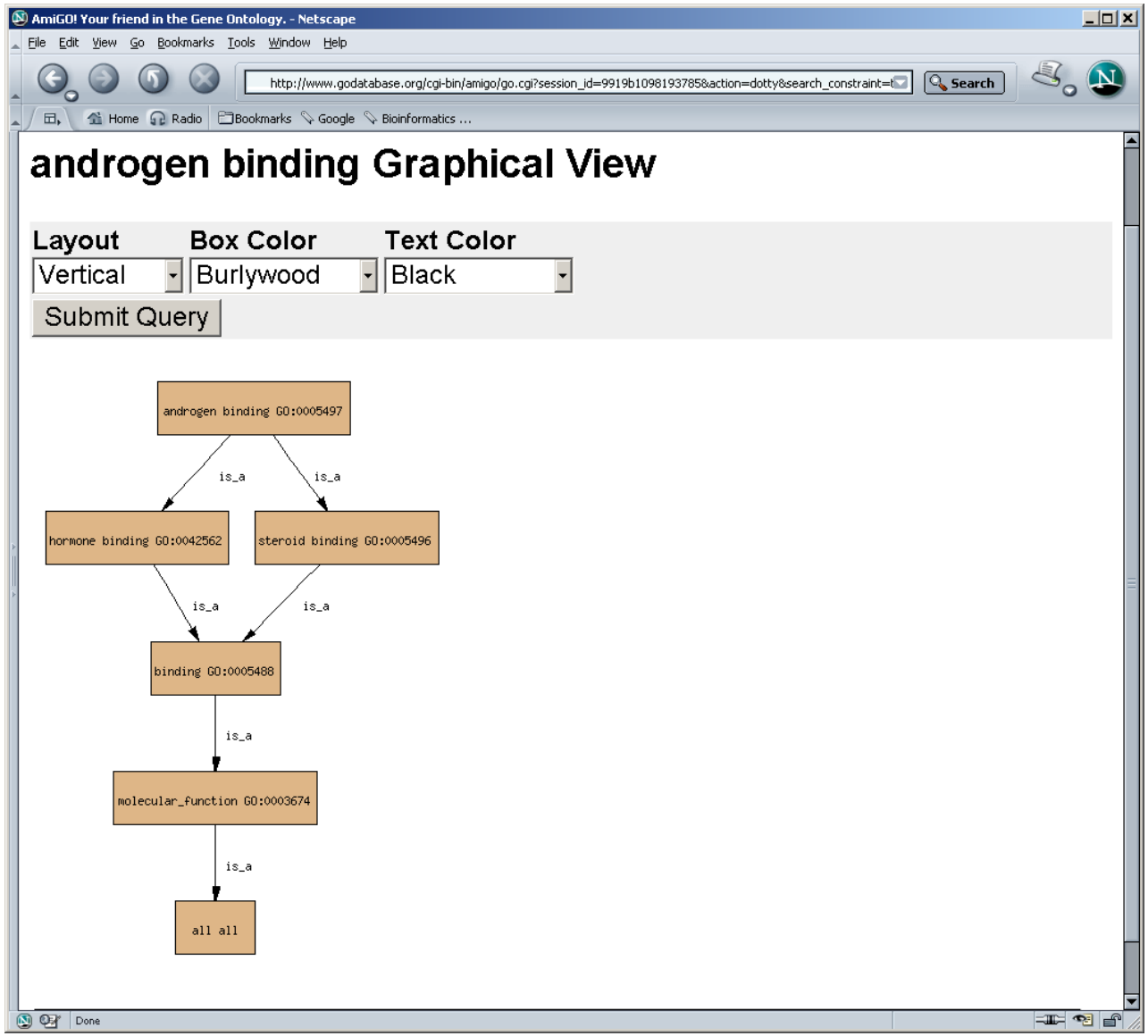
None.

Direct Gene Product Associations **Get ALL associations here:**

Direct Associations Submit Query

Done

GO browsing


AmiGO! Your friend in the Gene Ontology. - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.godatabase.org/cgi-bin/amigo/go.cgi?session_id=9919b1098193785&action=dotty&search_constraint=t
Search
Home Radio Bookmarks Google Bioinformatics ...
androgen binding Graphical View
Layout
Vertical
Box Color
Burlywood
Text Color
Black
Submit Query
androgen binding GO:0005497
is_a
is_a
hormone binding GO:0042562
steroid binding GO:0005496
is_a
is_a
binding GO:0005488
is_a
molecular_function GO:0003674
is_a
all all
Done

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

Entrez Gene: AR androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) [...

File Edit View Go Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=367&ordinalpos=1&itool=Entrez

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... BioinformaticsWiki

GeneOntology

Provided by GOA

Function	Evidence	
androgen binding	NAS	PubMed
androgen receptor activity	NAS	PubMed
androgen receptor activity	TAS	PubMed
lipid binding	IEA	
metal ion binding	IEA	
protein dimerization activity	NAS	PubMed
receptor activity	IEA	
sequence-specific DNA binding	IEA	
transcription factor activity	IDA	PubMed
zinc ion binding	IEA	

Process	Evidence	
androgen receptor signaling pathway	IEA	
cell growth	NAS	PubMed
cell proliferation	NAS	PubMed
cell-cell signaling	TAS	PubMed
in utero embryonic development	IEA	
male gonad development	IEA	
male somatic sex determination	IEA	
prostate gland development	NAS	PubMed

Done

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 interation; 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 Northerns, Westerns, 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

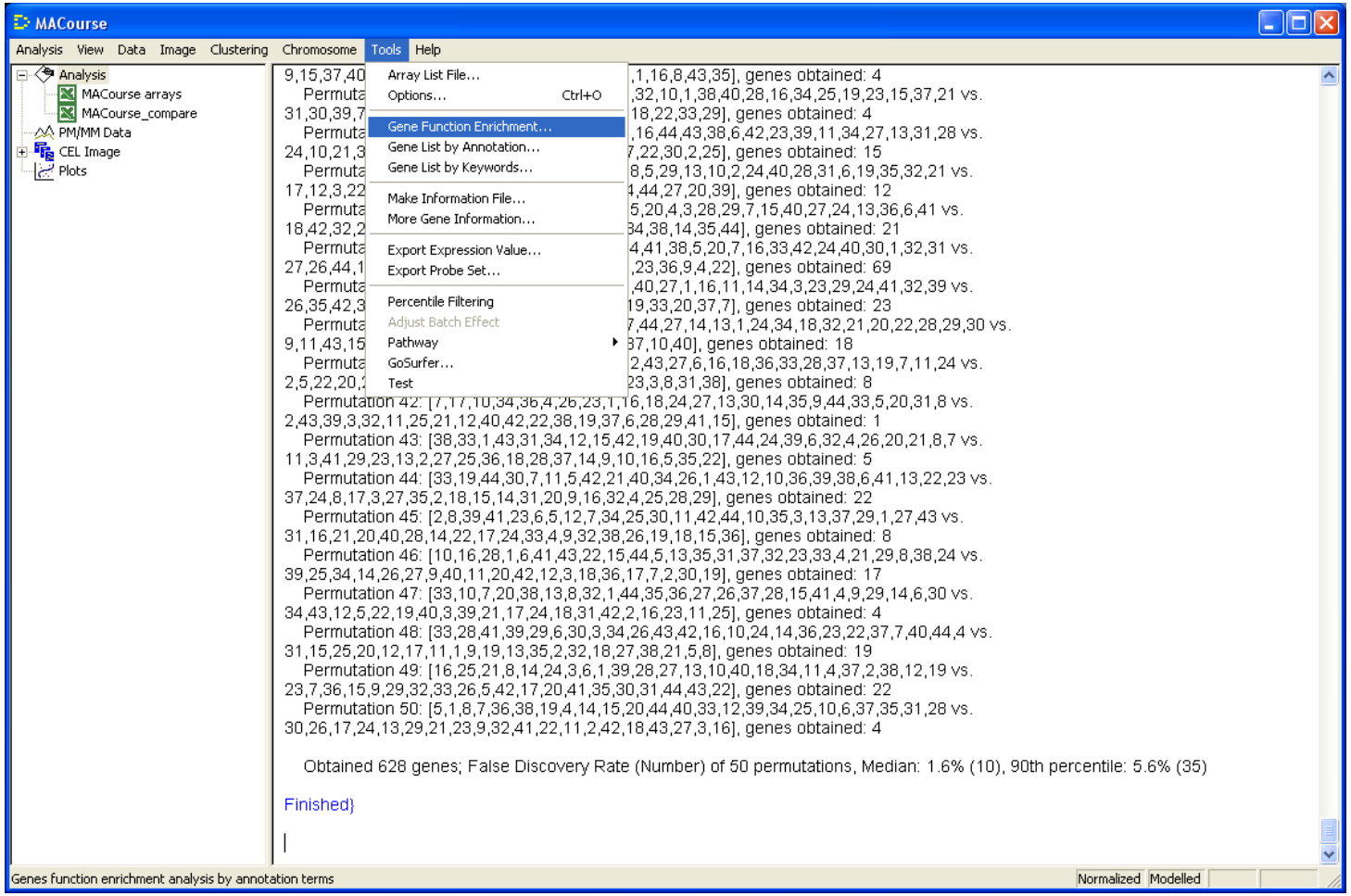
Quality of evidence

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

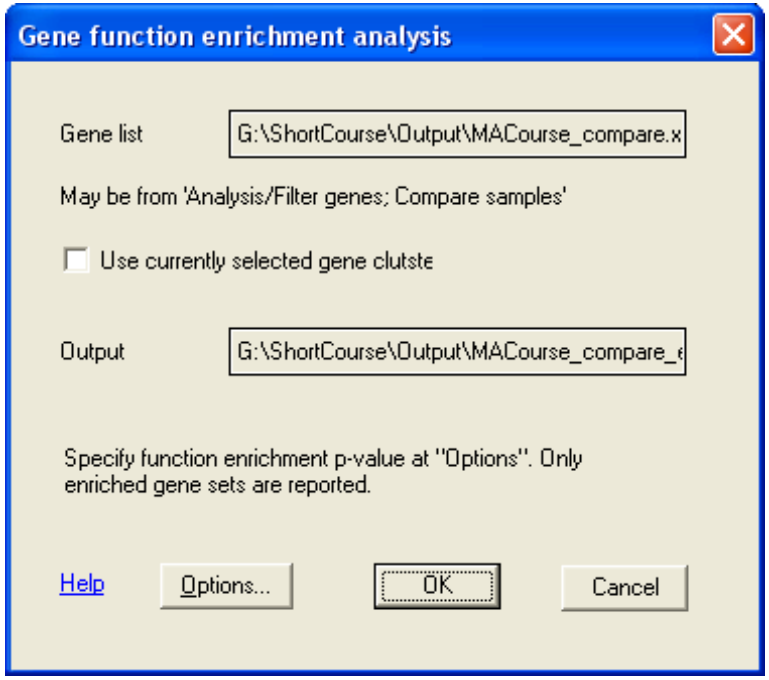
Using GeneOntology in dChip

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



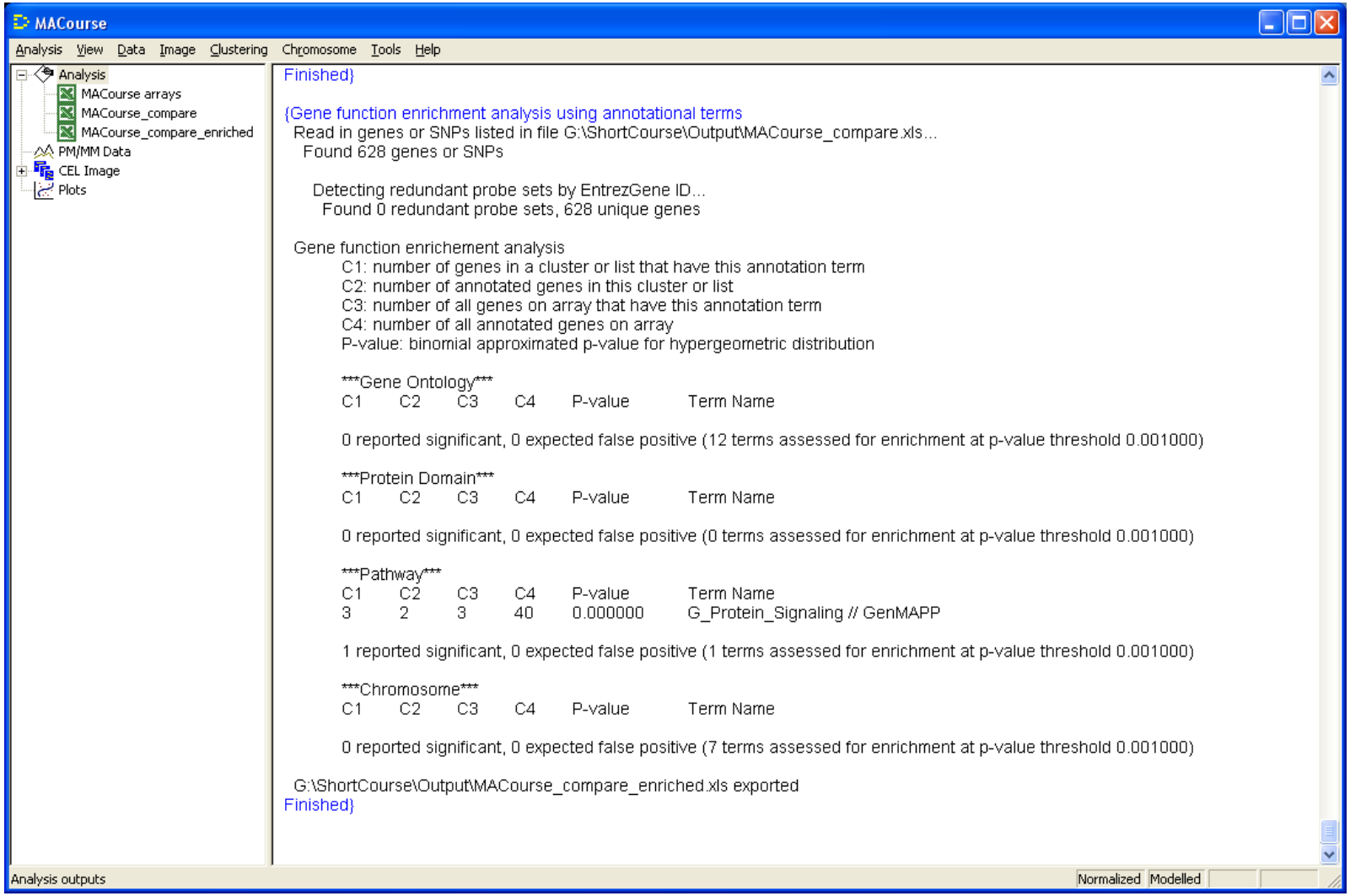
Using GeneOntology in dChip

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.



What do the results look like?

Microsoft Excel - affyShortCourse compare result classified.xls

	A	B	AA	AB	AU	AV	AW	BA
1	probe set	gene	baseline mean	baseline	experiment mean	experiment	fold change	filtered
2								
3	Found 21 Gene Ontology "protein tyrosine kinase" genes in a list with 391 annotated genes (all: 157/7685, PValue: 0.000042) *****							
4	40936_at	cysteine-rich motor neuron 1	7994	564	5144	612	-1.55	*
5	1485_at	EphA7	243	28	133	14	-1.83	*
6	2057_g_at	fibroblast growth factor receptor 1 (fms-re	5421	430	2717	100	-2	*
7	1964_g_at	fms-related tyrosine kinase 1 (vascular en	1555	167	982	51	-1.58	*
8	1545_g_at	fms-related tyrosine kinase 1 (vascular en	745	85	471	16	-1.58	*
9	34583_at	fms-related tyrosine kinase 3	9522	1513	16788	784	1.76	*
10	1065_at	fms-related tyrosine kinase 3	8414	1696	15615	933	1.86	*
11	40480_s_at	FYN oncogene related to SRC, FGR, YES	5038	514	3304	326	-1.52	*
12	34877_at	Janus kinase 1 (a protein tyrosine kinase)	15776	843	10823	834	-1.46	*
13	41594_at	Janus kinase 1 (a protein tyrosine kinase)	6687	345	4360	301	-1.53	*
14	1457_at	Janus kinase 1 (a protein tyrosine kinase)	3098	197	1886	177	-1.64	*
15	33238_at	lymphocyte-specific protein tyrosine kinas	3794	572	1936	258	-1.96	*
16	1988_at	platelet-derived growth factor receptor, alp	14547	602	10367	351	-1.4	*
17	36117_at	PTK2 protein tyrosine kinase 2	3730	242	2613	117	-1.43	*
18	37756_at	RYK receptor-like tyrosine kinase	1155	129	399	48	-2.89	*
19	539_at	RYK receptor-like tyrosine kinase	2294	107	1665	48	-1.38	*
20	572_at	TTK protein kinase	1309	128	792	76	-1.65	*
21	1674_at	v-yes-1 Yamaguchi sarcoma viral oncoger	1438	283	496	32	-2.9	*
22	32616_at	v-yes-1 Yamaguchi sarcoma viral related c	3247	219	4842	498	1.49	*
23	2024_s_at	v-yes-1 Yamaguchi sarcoma viral related c	1913	141	2960	322	1.55	*
24	1402_at	v-yes-1 Yamaguchi sarcoma viral related c	4141	289	6292	581	1.52	*
25								
26	Found 12 Gene Ontology "protein tyrosine phosphatase" genes in a list with 391 annotated genes (all: 81/7685, PValue: 0.000740) *							
27	32916_at	protein tyrosine phosphatase, receptor typ	6814	927	3050	454	-2.23	*
28	31892_at	protein tyrosine phosphatase, receptor typ	801	336	151	10	-5.32	*

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 GeneOntology category that was found to be signficantly 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.

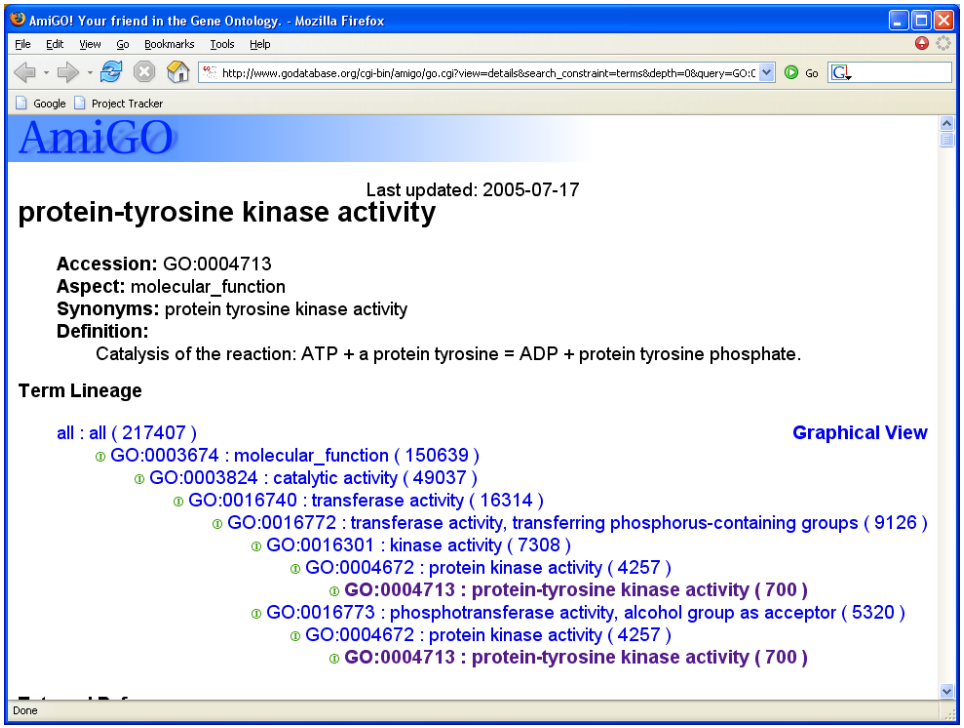
3. Of all the annotated probe sets, 157 had the “protein tyrosine kinase” function.

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 2×2 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.

Microsoft Excel - affyShortCourse compare result classified.xls

	A	B	AA	AB	AU	AV	AW	BA
1	probe set	gene	baseline mean	baseline	experiment mean	experiment	fold change	filtered
2								
3	Found 21 Gene Ontology "protein tyrosine kinase" genes in a list with 391 annotated genes (all: 157/7685, PValue: 0.000042) *****							
4	40936_at	cysteine-rich motor neuron 1	7994	564	5144	612	-1.55	*
5	1485_at	EphA7	243	28	133	14	-1.83	*
6	2057_g_at	fibroblast growth factor receptor 1 (fms-re	5421	430	2717	100	-2	*
7	1964_g_at	fms-related tyrosine kinase 1 (vascular en	1555	167	982	51	-1.58	*
8	1545_g_at	fms-related tyrosine kinase 1 (vascular en	745	85	471	16	-1.58	*
9	34583_at	fms-related tyrosine kinase 3	9522	1513	16788	784	1.76	*
10	1065_at	fms-related tyrosine kinase 3	8414	1696	15615	933	1.86	*
11	40480_s_at	FYN oncogene related to SRC, FGR, YES	5038	514	3304	326	-1.52	*
12	34877_at	Janus kinase 1 (a protein tyrosine kinase)	15776	843	10823	834	-1.46	*
13	41594_at	Janus kinase 1 (a protein tyrosine kinase)	6687	345	4360	301	-1.53	*
14	1457_at	Janus kinase 1 (a protein tyrosine kinase)	3098	197	1886	177	-1.64	*
15	33238_at	lymphocyte-specific protein tyrosine kinas	3794	572	1936	258	-1.96	*
16	1988_at	platelet-derived growth factor receptor, alp	14547	602	10367	351	-1.4	*
17	36117_at	PTK2 protein tyrosine kinase 2	3730	242	2613	117	-1.43	*
18	37756_at	RYK receptor-like tyrosine kinase	1155	129	399	48	-2.89	*
19	539_at	RYK receptor-like tyrosine kinase	2294	107	1665	48	-1.38	*
20	572_at	TTK protein kinase	1309	128	792	76	-1.65	*
21	1674_at	v-yes-1 Yamaguchi sarcoma viral oncoger	1438	283	496	32	-2.9	*
22	32616_at	v-yes-1 Yamaguchi sarcoma viral related c	3247	219	4842	498	1.49	*
23	2024_s_at	v-yes-1 Yamaguchi sarcoma viral related c	1913	141	2960	322	1.55	*
24	1402_at	v-yes-1 Yamaguchi sarcoma viral related c	4141	289	6292	581	1.52	*
25								
26	Found 12 Gene Ontology "protein tyrosine phosphatase" genes in a list with 391 annotated genes (all: 81/7685, PValue: 0.000740) *							
27	32916_at	protein tyrosine phosphatase, receptor typ	6814	927	3050	454	-2.23	*
28	31892_at	protein tyrosine phosphatase, receptor typ	801	336	151	10	-5.32	*

affyShortCourse compare result

Ready NUM

What alternatives are there?

GoMiner Home Page - Netscape

File Edit View Go Bookmarks Tools Window Help

http://discover.nci.nih.gov/gominer/

Search

Home Radio Bookmarks Google Bioinformatics ...

GoMiner Build:138

Genomics and Bioinformatics Group LMP, NCI, NIH

Medical Informatics & Bioimaging Lab, BME GA Tech/ Emory University

Home | Getting Started | Requirements | Installation | Command Line | Database | FAQ | Citing | Mirrors | Credits

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 satisifes the system requirements
- Download the GoMiner program file, gominer.jar
- Read the Quick Start and try out the sample files

http://discover.nci.nih.gov/gominer/index.jsp

http://discover.nci.nih.gov/gominer

GoMiner Installation - Netscape

File Edit View Go Bookmarks Tools Window Help

http://discover.nci.nih.gov/gominer/installation.jsp

Search

Home Radio Bookmarks Google Bioinformatics ...

GoMiner Build:138

Genomics and Bioinformatics Group LMP, NCI, NIH

Medical Informatics & Bioimaging Lab, BME GA Tech/ Emory University

Home | Getting Started | Requirements | Installation | Command Line | Database | FAQ | Citing | Mirrors | Credits

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

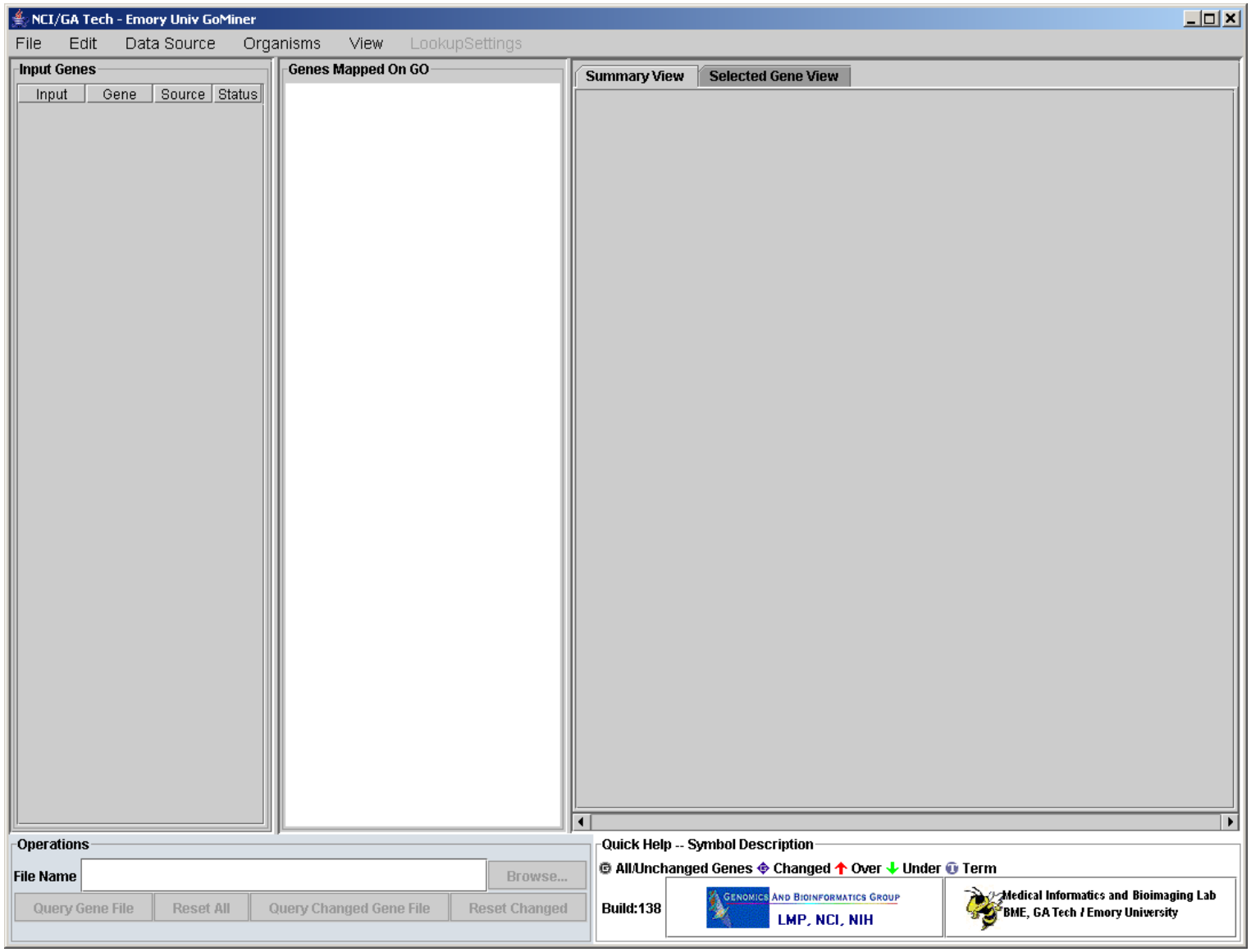
You need a machine with

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

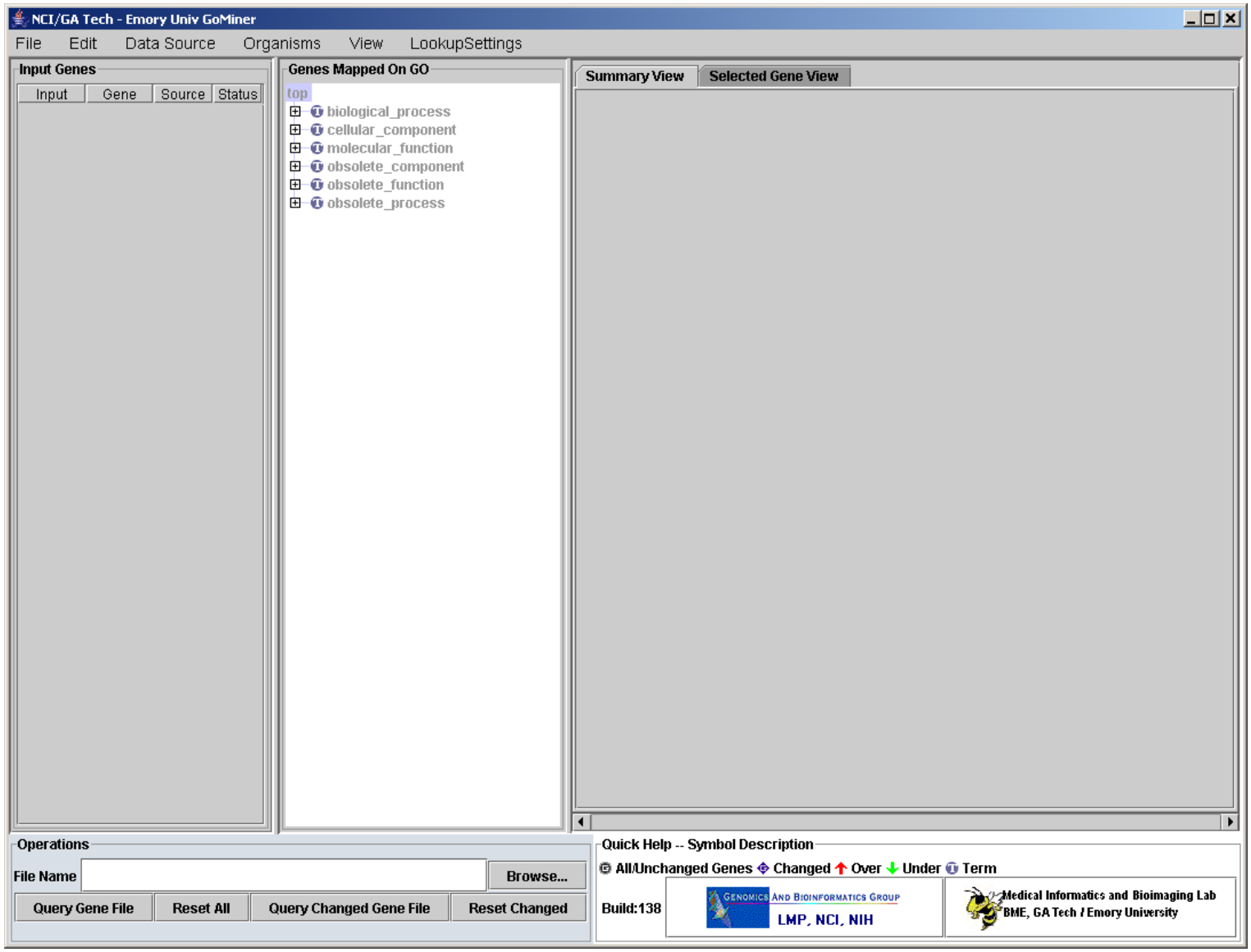
Download the GoMiner Java code, install it, and double-click on it to start the program.

Then go to “File” $-$ $>$ “Load GO Terms” and click “OK”. Wait a few minutes while the program loads the GeneOntology information from the NCI.

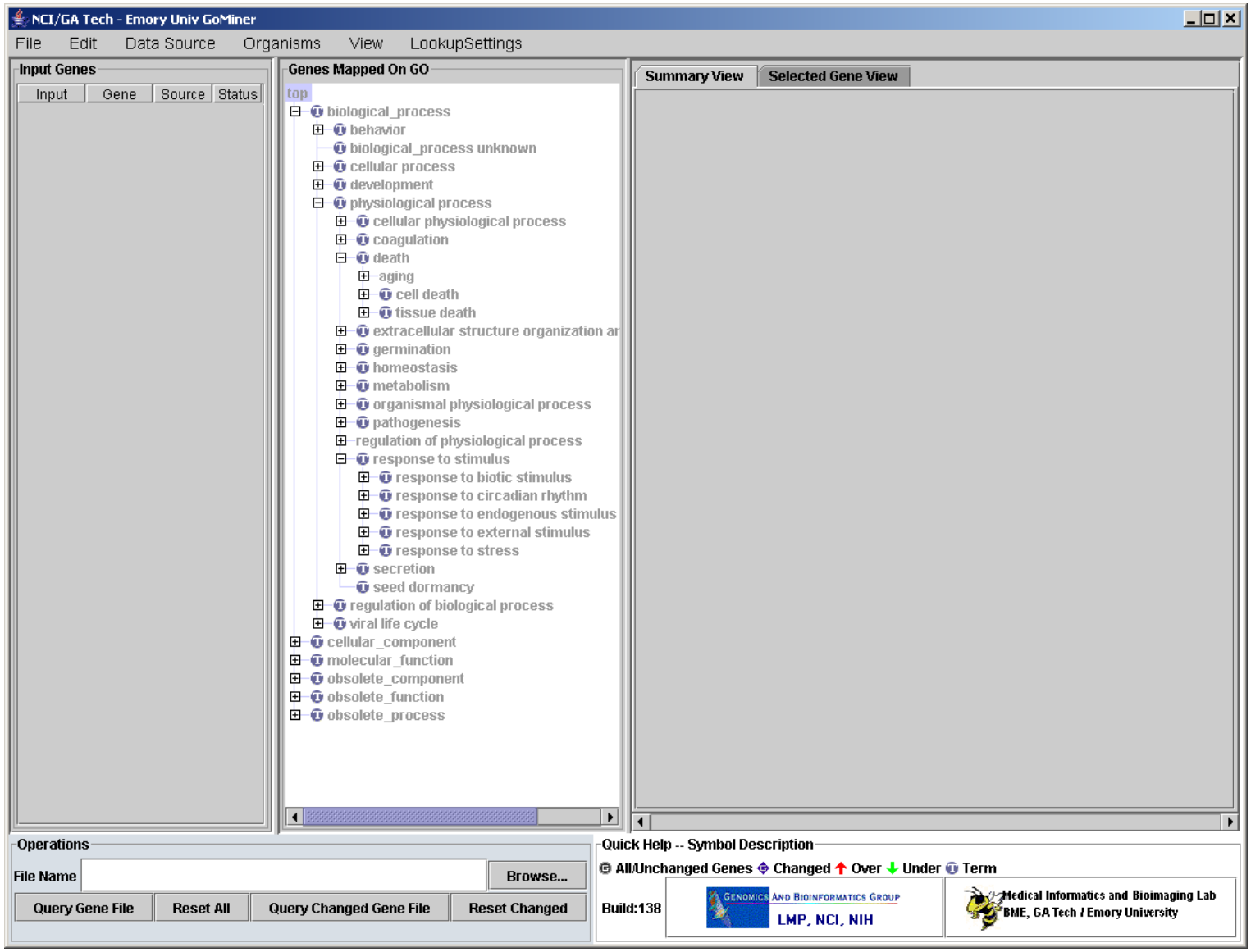
GoMiner Start



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. Time sink.

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

NCI/GA Tech - Emory Univ GoMiner

File Edit Data Source Organisms View LookupSettings

Input Genes

Input	Gene	Source	Status
YWHAE	143E_...	UniProt	
SFN	143S_...	UniProt	
PPP2R...	2A5A_...	UniProt	
PPP2R...	2A5B_...	UniProt	
PPP2R...	2A5D_...	UniProt	
PPP2R...	2A5E_...	UniProt	
PPP2R...	2A5G_...	UniProt	
PPP2R...	2AAA_...	UniProt	
PPP2R...	2AAB_...	UniProt	
PPP2R...	2ABA_...	UniProt	
PPP2R...	2ABB_...	UniProt	
HLA-DMA	2DMA_...	UniProt	
HLA-D...	2DMB_...	UniProt	
HLA-DOA	2DOA_...	UniProt	
HLA-DRA	2DRA_...	UniProt	
SH3BP2	3BP2_...	UniProt	
SLC3A2	4F2_H...	UniProt	
A2M	A2MG_...	UniProt	
ACTN1	AAC1_...	UniProt	
PRKAB1	AAKB_...	UniProt	
PRKAG1	AAKG_...	UniProt	
ATBF1	ABF1_...	UniProt	
ABL1	ABL1_...	UniProt	
ABL2	ABL2_...	UniProt	
ABR	ABR_H...	UniProt	
ACY1	ACY1_...	UniProt	
ADAM17	AD17_...	UniProt	
ADA	ADA_H...	UniProt	
ADD3	ADDG_...	UniProt	
ADH6	ADH6_...	UniProt	
ADK	ADK_H...	UniProt	
AOX1	ADO_H...	UniProt	
ADSS	ADSS		
SLC25A5	ADT2_...	UniProt	
MLLT2	AF4_H...	UniProt	
GLA	AGAL_...	UniProt	
ANGPT1	AGP1_...	UniProt	
ANGPT2	AGP2_...	UniProt	
AHR	AHR_H...	UniProt	

Genes Mapped On GO

- top (1299)
  - biological_process (1245)
    - behavior (8)
    - biological_process unknown (27)
    - cellular process (847)
    - development (220)
    - physiological process (1139)
      - IGF2_HUMAN (IGF2) - (UniProt)
      - IGFA_HUMAN (IGF1) - (UniProt)
      - O43200 (TSHR) - (UniProt)
      - PGH1_HUMAN (PTGS1) - (UniProt)
      - PGH2_HUMAN (PTGS2) - (UniProt)
      - REL1_HUMAN (RLN1) - (UniProt)
      - RLF_HUMAN (RLF) - (UniProt)
      - cellular physiological process (568)
      - coagulation (16)
      - death (123)
        - aging (2)
        - cell death (122)
          - cell aging (1)
          - cytolysis (3)
          - programmed cell death (120)
            - apoptosis (120)
            - regulation of programmed cell de
      - extracellular structure organization and b
      - homeostasis (13)
      - metabolism (823)
      - organismal physiological process (254)
      - pathogenesis (3)
      - regulation of physiological process (239)
      - response to stimulus (359)
      - secretion (2)
    - regulation of biological process (403)
    - viral life cycle (8)
  - cellular_component (1070)
  - molecular_function (1215)
  - obsolete_component
  - obsolete_function
  - obsolete_process

Summary View | **Selected Gene View**

Category Name	P-Chng	P-Undr	P-Ovr	Tot	Chng	Undr	Ovr	Category ID
ATP-dependent hel...	1.0000	1.0000	1.0000	11	0	0	0	GO:00080...
transcription elong...	1.0000	1.0000	1.0000	1	0	0	0	GO:00080...
protein C-terminus ...	1.0000	1.0000	1.0000	2	0	0	0	GO:00080...
microtubule binding	1.0000	1.0000	1.0000	3	0	0	0	GO:00080...
regulation of heart r...	1.0000	1.0000	1.0000	1	0	0	0	GO:00080...
circulation	1.0000	1.0000	1.0000	9	0	0	0	GO:00080...
beta-catenin binding	1.0000	1.0000	1.0000	1	0	0	0	GO:00080...
chemokine activity	1.0000	1.0000	1.0000	18	0	0	0	GO:00080...
oligopeptide transp...	1.0000	1.0000	1.0000	1	0	0	0	GO:00151...
peptide transporter ...	1.0000	1.0000	1.0000	2	0	0	0	GO:00151...
L-amino acid trans...	1.0000	1.0000	1.0000	1	0	0	0	GO:00151...
acidic amino acid tr...	1.0000	1.0000	1.0000	1	0	0	0	GO:00151...
amino acid transpo...	1.0000	1.0000	1.0000	2	0	0	0	GO:00151...
hexose transporter ...	1.0000	1.0000	1.0000	2	0	0	0	GO:00151...
monosaccharide tr...	1.0000	1.0000	1.0000	2	0	0	0	GO:00151...
carbohydrate trans...	1.0000	1.0000	1.0000	3	0	0	0	GO:00151...
nitric oxide metabol...	1.0000	1.0000	1.0000	4	0	0	0	GO:00462...
sodium ion transpo...	1.0000	1.0000	1.0000	1	0	0	0	GO:00150...
hydrogen ion trans...	1.0000	1.0000	1.0000	3	0	0	0	GO:00150...
monovalent inorga...	1.0000	1.0000	1.0000	3	0	0	0	GO:00150...
ion transporter activ...	1.0000	1.0000	1.0000	4	0	0	0	GO:00150...
protein phosphatas...	1.0000	1.0000	1.0000	3	0	0	0	GO:00150...
thrombin receptor a...	1.0000	1.0000	1.0000	1	0	0	0	GO:00150...
glutathione disulfid...	1.0000	1.0000	1.0000	1	0	0	0	GO:00150...
peptide disulfide ox...	1.0000	1.0000	1.0000	1	0	0	0	GO:00150...
disulfide oxidoredu...	1.0000	1.0000	1.0000	4	0	0	0	GO:00150...
protein transport	1.0000	1.0000	1.0000	26	0	0	0	GO:00150...
Cajal body	1.0000	1.0000	1.0000	1	0	0	0	GO:00150...
coreceptor activity	1.0000	1.0000	1.0000	6	0	0	0	GO:00150...
glucuronosyltransf...	1.0000	1.0000	1.0000	2	0	0	0	GO:00150...
nuclear organizatio...	1.0000	1.0000	1.0000	25	0	0	0	GO:00069...
organelle organizat...	1.0000	1.0000	1.0000	32	0	0	0	GO:00069...
unfolded protein re...	1.0000	1.0000	1.0000	1	0	0	0	GO:00069...
alcohol catabolism	1.0000	1.0000	1.0000	4	0	0	0	GO:00461...
response to unfold...	1.0000	1.0000	1.0000	5	0	0	0	GO:00069...
ER-nuclear signali...	1.0000	1.0000	1.0000	1	0	0	0	GO:00069...
response to lipid hy...	1.0000	1.0000	1.0000	1	0	0	0	GO:00069...
phenol metabolism	1.0000	1.0000	1.0000	2	0	0	0	GO:00189...

Operations

File Name C:\Source\GoMinerExample\total.gene Browse...

Query Gene File | Reset All | Query Changed Gene File | Reset Changed

Quick Help -- Symbol Description

All/Unchanged Genes Changed Over Under Term

Build:138

GENOMICS AND BIOINFORMATICS GROUP LMP, NCI, NIH

Medical Informatics and Bioimaging Lab BME, GA Tech / Emory University

GoMiner with changed gene list loaded

NCI/GA Tech - Emory Univ GoMiner

File Edit Data Source Organisms View LookupSettings

Input Genes

Input	Gene	Source	Status
YWHAE	143E_...	UniProt	
SFN	143S_...	UniProt	
PPP2R...	2A5A_...	UniProt	
PPP2R...	2A5B_...	UniProt	
PPP2R...	2A5D_...	UniProt	
PPP2R...	2A5E_...	UniProt	
PPP2R...	2A5G_...	UniProt	
PPP2R...	2AAA_...	UniProt	
PPP2R...	2AAB_...	UniProt	
PPP2R...	2ABA_...	UniProt	
PPP2R...	2ABB_...	UniProt	
HLA-DMA	2DMA_...	UniProt	
HLA-D...	2DMB_...	UniProt	
HLA-DOA	2DOA_...	UniProt	
HLA-DRA	2DRA_...	UniProt	↓
SH3BP2	3BP2_...	UniProt	
SLC3A2	4F2_H...	UniProt	
A2M	A2MG_...	UniProt	
ACTN1	AAC1_...	UniProt	
PRKAB1	AAKB_...	UniProt	
PRKAG1	AAKG_...	UniProt	
ATBF1	ABF1_...	UniProt	
ABL1	ABL1_...	UniProt	
ABL2	ABL2_...	UniProt	
ABR	ABR_H...	UniProt	
ACY1	ACY1_...	UniProt	
ADAM17	AD17_...	UniProt	
ADA	ADA_H...	UniProt	
ADD3	ADDG_...	UniProt	
ADH6	ADH6_...	UniProt	
ADK	ADK_H...	UniProt	
AOX1	ADO_H...	UniProt	
ADSS	ADSS		
SLC25A5	ADT2_...	UniProt	
MLLT2	AF4_H...	UniProt	
GLA	AGAL_...	UniProt	
ANGPT1	AGP1_...	UniProt	
ANGPT2	AGP2_...	UniProt	
AHR	AHR_H...	UniProt	↑

Genes Mapped On GO

```
p (1299 1.00 p=1.00 1.00 p=1.00 1.00 p=1.00)
- biological_process (1245 1.03 p=0.17 1.01 p=0.48 1.02 p=0.17)
  + biological_process unknown (27 1.30 p=0.46 0.53 p=0.86 0.88 p=0.
  + cellular process (847 0.99 p=0.58 0.97 p=0.69 0.98 p=0.67)
  + development (220 0.96 p=0.62 1.12 p=0.35 1.04 p=0.43)
  - physiological process (1139 1.08 p=0.04 1.05 p=0.11 1.06 p=0.01)
    + cellular physiological process (568 1.02 p=0.48 0.99 p=0.57 1.0
    + coagulation (16 1.10 p=0.61 0.90 p=0.69 0.99 p=0.62)
    - death (123 1.28 p=0.26 1.17 p=0.34 1.22 p=0.20)
      - cell death (122 1.29 p=0.25 1.18 p=0.33 1.23 p=0.19)
        + cytolysis (3 0.00 p=1.00 4.81 p=0.19 2.64 p=0.33)
        - programmed cell death (120 1.32 p=0.24 1.08 p=0.45 1.1
          + apoptosis (120 1.32 p=0.24 1.08 p=0.45 1.19 p=0.24)
          + regulation of programmed cell death (77 1.14 p=0.45 0.
    + homeostasis (13 0.00 p=1.00 5.55 p=0.00 3.05 p=0.02)
    + metabolism (823 0.90 p=0.91 1.14 p=0.04 1.03 p=0.33)
    + organismal physiological process (254 1.87 p=0.00 0.91 p=0.71
    + regulation of physiological process (239 1.10 p=0.38 1.39 p=0.05 1
    + response to stimulus (359 1.47 p=0.01 1.09 p=0.34 1.26 p=0.02)
  + regulation of biological process (403 1.18 p=0.18 1.25 p=0.06 1.22 p
  + viral life cycle (8 2.19 p=0.38 1.80 p=0.44 1.98 p=0.27)
- cellular_component (1070 0.97 p=0.78 0.97 p=0.78 0.97 p=0.84)
- molecular_function (1215 0.95 p=0.96 0.91 p=1.00 0.93 p=1.00)
  obsolete_component
  obsolete_function
  obsolete_process
```

Summary View | **Selected Gene View**

Category Name	P-Chng	P-Undr	P-Ovr	Tot	Chng	Undr	Ovr	Category ID
cytoplasmic seque...	0.0002	0.0178	0.0260	4	4	2	2	GO:00429...
negative regulation ...	0.0002	0.0178	0.0260	4	4	2	2	GO:00429...
transcription factor-...	0.0002	0.0178	0.0260	4	4	2	2	GO:00429...
regulation of transc...	0.0002	0.0178	0.0260	4	4	2	2	GO:00429...
regulation of protei...	0.0002	0.0178	0.0260	4	4	2	2	GO:00423...
regulation of nucleo...	0.0002	0.0178	0.0260	4	4	2	2	GO:00468...
chemokine activity	0.0008	0.0782	0.0060	18	8	3	5	GO:00080...
G-protein-coupled r...	0.0008	0.0782	0.0060	18	8	3	5	GO:00016...
chemokine recepto...	0.0008	0.0782	0.0060	18	8	3	5	GO:00423...
chemotaxis	0.0012	0.0547	0.0112	37	12	5	7	GO:00069...
taxis	0.0012	0.0547	0.0112	37	12	5	7	GO:00423...
response to wound...	0.0015	0.0227	0.0296	75	19	9	10	GO:00096...
response to chemi...	0.0018	0.0814	0.0097	54	15	6	9	GO:00422...
response to pathog...	0.0030	0.2972	0.0055	6	4	1	3	GO:00096...
regulation of transp...	0.0030	0.0414	0.0593	6	4	2	2	GO:00510...
immune response	0.0033	0.0002	0.4695	207	39	24	15	GO:00069...
response to pest, p...	0.0036	0.0178	0.0743	123	26	13	13	GO:00096...
extracellular space	0.0038	0.0039	0.2217	47	13	8	5	GO:00056...
protein threonine/tyr...	0.0063	0.0558	0.0794	7	4	2	2	GO:00047...
MAP kinase kinase ...	0.0063	0.0558	0.0794	7	4	2	2	GO:00047...
response to pathog...	0.0063	0.3374	0.0092	7	4	1	3	GO:00428...
antigen processing	0.0070	0.0001	1.0000	15	6	6	0	GO:00303...
antigen presentation	0.0070	0.0001	1.0000	15	6	6	0	GO:00198...
MHC class II recept...	0.0074	0.0024	0.5475	11	5	4	1	GO:00450...
response to extern...	0.0075	0.0400	0.0743	123	25	12	13	GO:00096...
defense response	0.0088	0.0008	0.4993	225	40	24	16	GO:00069...
response to biotic s...	0.0089	0.0013	0.4397	246	43	25	18	GO:00096...
inflammatory respo...	0.0096	0.1695	0.0232	52	13	5	8	GO:00069...
innate immune res...	0.0096	0.1695	0.0232	52	13	5	8	GO:00450...
physiological proce...	0.0099	0.0372	0.1127	1139	153	70	83	GO:00075...
metal ion homeost...	0.0114	1.0000	0.0008	12	5	0	5	GO:00068...
cell ion homeostasis	0.0114	1.0000	0.0008	12	5	0	5	GO:00068...
di-, tri-valent inorga...	0.0114	1.0000	0.0008	12	5	0	5	GO:00300...
cation homeostasis	0.0114	1.0000	0.0008	12	5	0	5	GO:00300...
ion homeostasis	0.0114	1.0000	0.0008	12	5	0	5	GO:00508...
response to abiotic ...	0.0119	0.1597	0.0309	65	15	6	9	GO:00096...
transforming growt...	0.0159	1.0000	0.0048	2	2	0	2	GO:00306...
NF-kappaB-nucleu...	0.0159	0.1107	0.1338	2	2	1	1	GO:00423...

Operations

File Name C:\Source\GoMinerExample\under.over.2col Browse...

Query Gene File | Reset All | Query Changed Gene File | Reset Changed

Quick Help -- Symbol Description

All/Unchanged Genes Changed ↑ Over ↓ Under Term

Build:138

GENOMICS AND BIOINFORMATICS GROUP LMP, NCI, NIH

Medical Informatics and Bioimaging Lab BME, GA Tech / Emory University

GoMiner subgraphs

NCI/GA Tech - Emory Univ GoMiner

File Edit Data Source Organisms View LookupSettings

Input Genes

Input	Gene	Source	Status
YWHAE	143E_...	UniProt	
SFN	143S_...	UniProt	
PPP2R...	2A5A_...	UniProt	
PPP2R...	2A5B_...	UniProt	
PPP2R...	2A5D_...	UniProt	
PPP2R...	2A5E_...	UniProt	
PPP2R...	2A5G_...	UniProt	
PPP2R...	2AAA_...	UniProt	
PPP2R...	2AAB_...	UniProt	
PPP2R...	2ABA_...	UniProt	
PPP2R...	2ABB_...	UniProt	
HLA-DMA	2DMA_...	UniProt	
HLA-D...	2DMB_...	UniProt	
HLA-DOA	2DOA_...	UniProt	
HLA-DRA	2DRA_...	UniProt	
SH3BP2	3BP2_...	UniProt	
SLC3A2	4F2_H...	UniProt	
A2M	A2MG_...	UniProt	
ACTN1	AAC1_...	UniProt	
PRKAB1	AAKB_...	UniProt	
PRKAG1	AAKG_...	UniProt	
ATBF1	ABF1_...	UniProt	
ABL1	ABL1_...	UniProt	
ABL2	ABL2_...	UniProt	
ABR	ABR_H...	UniProt	
ACY1	ACY1_...	UniProt	
ADAM17	AD17_...	UniProt	
ADA	ADA_H...	UniProt	
ADD3	ADDG_...	UniProt	
ADH6	ADH6_...	UniProt	
ADK	ADK_H...	UniProt	
AOX1	ADO_H...	UniProt	
ADSS	ADSS		
SLC25A5	ADT2_...	UniProt	
MLLT2	AF4_H...	UniProt	
GLA	AGAL_...	UniProt	
ANGPT1	AGP1_...	UniProt	
ANGPT2	AGP2_...	UniProt	
AHR	AHR_H...	UniProt	

Genes Mapped On GO

1299 1.00 p=1.00 1.00 p=1.00 1.00 p=1.00)
biological_process (1245 1.03 p=0.17 1.01 p=0.48 1.02 p=0.17)
biological_process unknown (27 1.30 p=0.46 0.53 p=0.86 0.88 p=0.
cellular process (847 0.99 p=0.58 0.97 p=0.69 0.98 p=0.67)
development (220 0.96 p=0.62 1.12 p=0.35 1.04 p=0.43)
physiological process (1139 1.08 p=0.04 1.05 p=0.11 1.06 p=0.01)
cellular physiological process (568 1.02 p=0.48 0.99 p=0.57 1.0
coagulation (16 1.10 p=0.61 0.90 p=0.69 0.99 p=0.62)
death (123 1.28 p=0.26 1.17 p=0.34 1.22 p=0.20)
cell death (122 1.29 p=0.25 1.18 p=0.33 1.23 p=0.19)
cytolysis (3 0.00 p=1.00 4.81 p=0.19 2.64 p=0.33)
programmed cell death (120
apoptosis (120 1.32 p=
AHR_HUMAN (AHR)
ASP2_HUMAN (TP5
BAD_HUMAN (BAD)
BCL2_HUMAN (BCL
BIR2_HUMAN (BIRC2) - (UniProt)
DAD1_HUMAN (DAD1) - (UniProt)
DAP1_HUMAN (DAP) - (UniProt)
DPF2 (REQ) - (UniProt)
ICE6_HUMAN (CASP6) - (UniProt)
IKBA_HUMAN (NFKBIA) - (UniProt)
NIP2_HUMAN (BNIP2) - (UniProt)
P300_HUMAN (EP300) - (UniProt)
PAK1_HUMAN (PAK1) - (UniProt)
Q8WX32 (BAG1) - (UniProt)
T10C_HUMAN (TNFRSF10C) - (UniProt)
T10D_HUMAN (TNFRSF10D) - (UniProt)
TNI3_HUMAN (TNFAIP3) - (UniProt)
apoptotic program (16 1.10 p=0.61 0.00 p=1.00 0.50
regulation of apoptosis (77 1.14 p=0.45 0.94 p=0.63
regulation of programmed cell death (77 1.14 p=0.45 0.
homeostasis (13 0.00 p=1.00 5.55 p=0.00 3.05 p=0.02)
metabolism (823 0.90 p=0.91 1.14 p=0.04 1.03 p=0.33)
organismal physiological process (254 1.87 p=0.00 0.91 p=0.71
regulation of physiological process (239 1.10 p=0.38 1.39 p=0.05
response to stimulus (359 1.47 p=0.01 1.09 p=0.34 1.26 p=0.02)
regulation of biological process (403 1.18 p=0.18 1.25 p=0.06 1.22

Export summary data to text file
DAG of changed genes
Export DAG of changed genes to file
Export Genes By Category

Summary View | Selected Gene View

Selected Single Gene View

top (1299 1.00 p=1.00 1.00 p=1.00 1.00 p=1.00)
biological_process (1245 1.03 p=0.17 1.01 p=0.48 1.02 p=0.17)
cellular process (847 0.99 p=0.58 0.97 p=0.69 0.98 p=0.67)
cellular physiological process (568 1.02 p=0.48 0.99 p=0.57 1.00 p=0.51)
cell death (122 1.29 p=0.25 1.18 p=0.33 1.23 p=0.19)
programmed cell death (120 1.32 p=0.24 1.08 p=0.45 1.19 p=0.24)
physiological process (1139 1.08 p=0.04 1.05 p=0.11 1.06 p=0.01)
cellular physiological process (568 1.02 p=0.48 0.99 p=0.57 1.00 p=0.51)
cell death (122 1.29 p=0.25 1.18 p=0.33 1.23 p=0.19)
programmed cell death (120 1.32 p=0.24 1.08 p=0.45 1.19 p=0.24)
death (123 1.28 p=0.26 1.17 p=0.34 1.22 p=0.20)
cell death (122 1.29 p=0.25 1.18 p=0.33 1.23 p=0.19)
programmed cell death (120 1.32 p=0.24 1.08 p=0.45 1.19 p=0.24)

Operations

File Name C:\Source\GoMinerExample\under.over.2col Browse...

Query Gene File | Reset All | Query Changed Gene File | Reset Changed

Quick Help -- Symbol Description

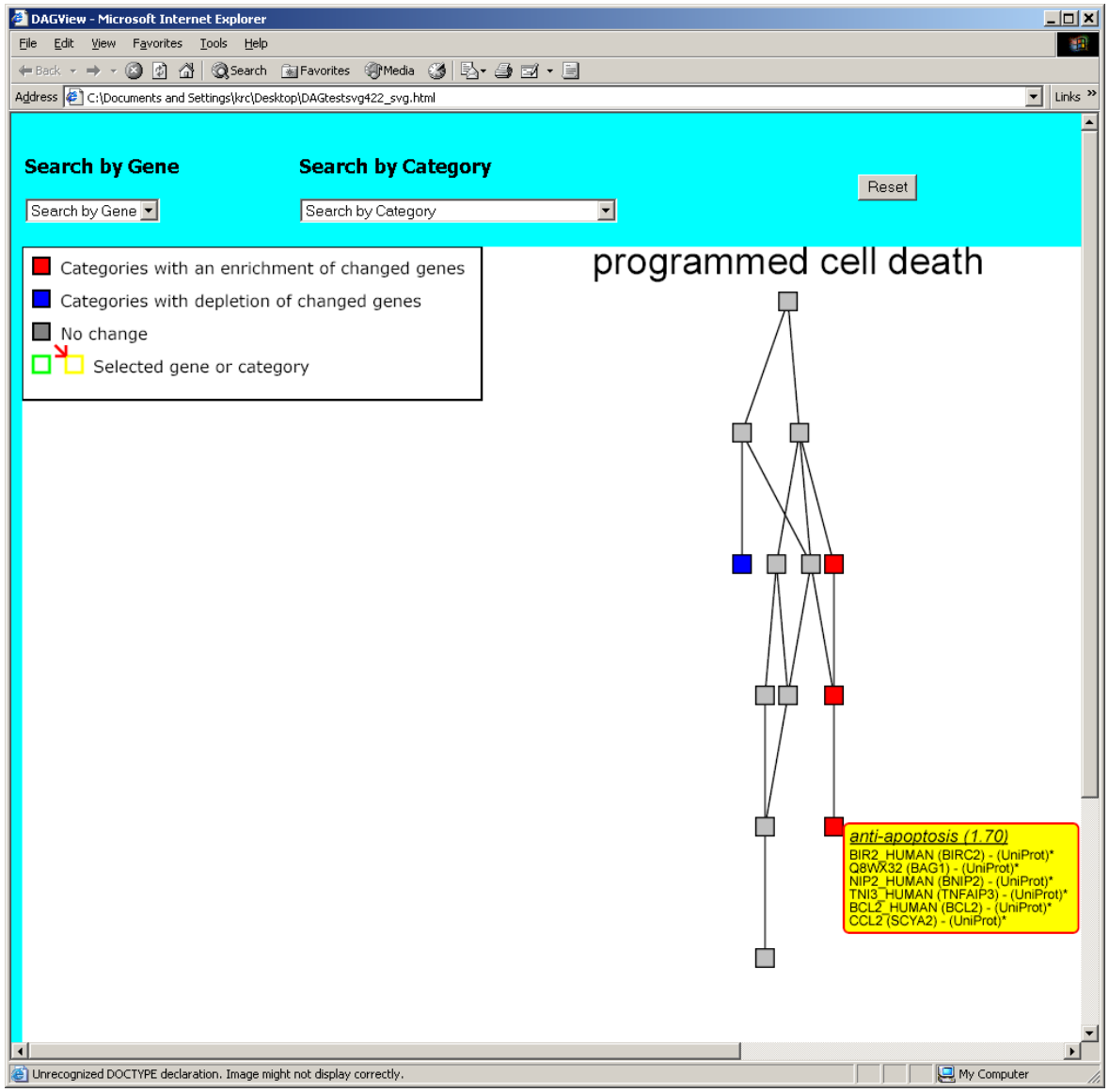
All/Unchanged Genes Changed Over Under Term

Build:138

GENOMICS AND BIOINFORMATICS GROUP LMP, NCI, NIH

Medical Informatics and Bioimaging Lab BME, GA Tech / Emory University

GoMiner subgraphs



Intepreting GoMiner results

Enrichment is computed as

$$\frac{\textit{changed genes in category} \big/ \textit{total genes in category}}{\textit{changed genes on array} \big/ \textit{all genes on array}}$$

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

Intepreting 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.