GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

4 September 2007

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. Let me disabuse you of that notion immediately. It is true that the biological sequences of the probes that were placed on the array do not change. However, our knowledge of the human genome continues to evolve, and thus our opinion about exactly what genes are targeted by those sequences must be continually updated.

For Affymetrix microarrays, the company maintains a web site that always contains their latest opinion on the nature and identity of the targeted genes.

Why Do We Care?

Recall from the last lecture that 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 a list of 610 probe sets that were differentially expressed.

It is considered bad form to just hand the biologists a list of 610 genes and wish them good luck as they go on their way. They typically want to know: do these genes reflect particular biological functions that are different betwen the two groups of samples, or do they identify specific biological pathways or networks that are perturbed?

List of Differentially Expressed Genes

Microsoft Excel - affyShortCourse compare result.xls

File Edit View Insert Format Tools Data Window Help

AW36 = -4.23

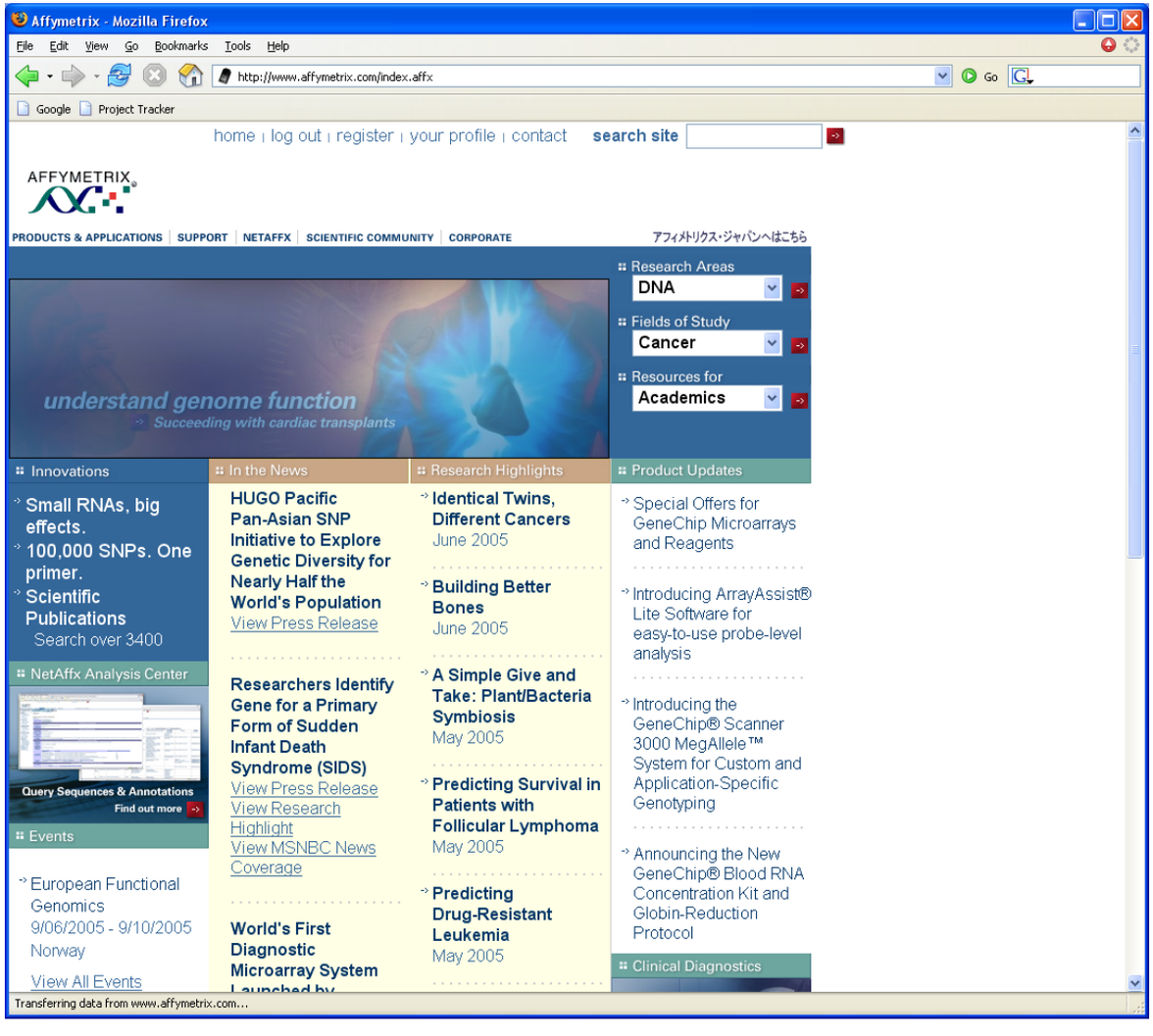
	A	B	AA	AB	AU	AV	AW	AX	AY
12	probe set	gene	baseline mean	baseline	experiment mean	experiment	fold change	lower bo	upper bo
13	37680_at	A kinase (PRKA) anchor protein (gravin) 12	2973.7	560.63	148.29	24.19	-20.05	-12.93	-30.28
14	1325_at	MAD, mothers against decapentaplegic homolog 1	7759.92	1390.4	595.18	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

affyShortCourse compare result

Ready NUM

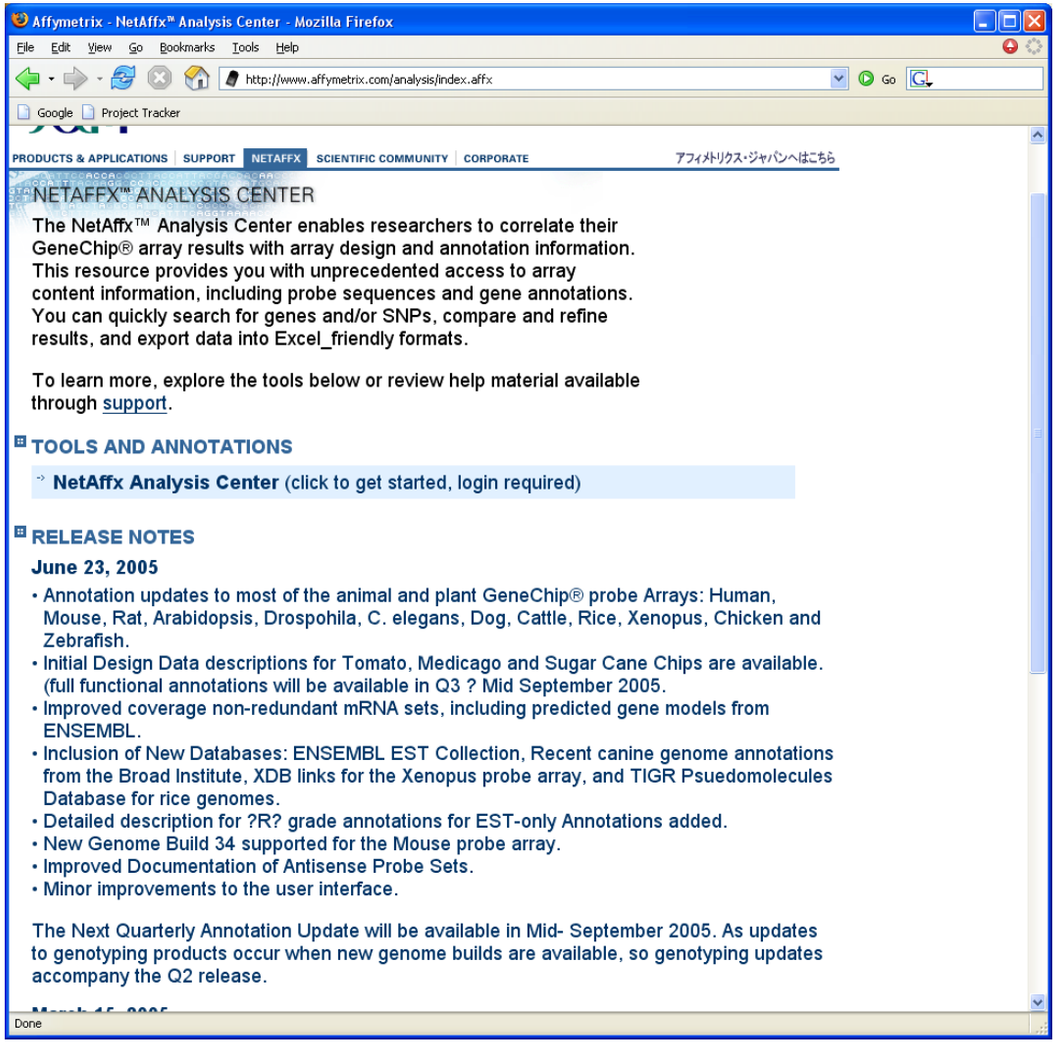
Affymetrix Web Site

http://www.affymetrix.com



NETAFFX

Annotations are updated quarterly...

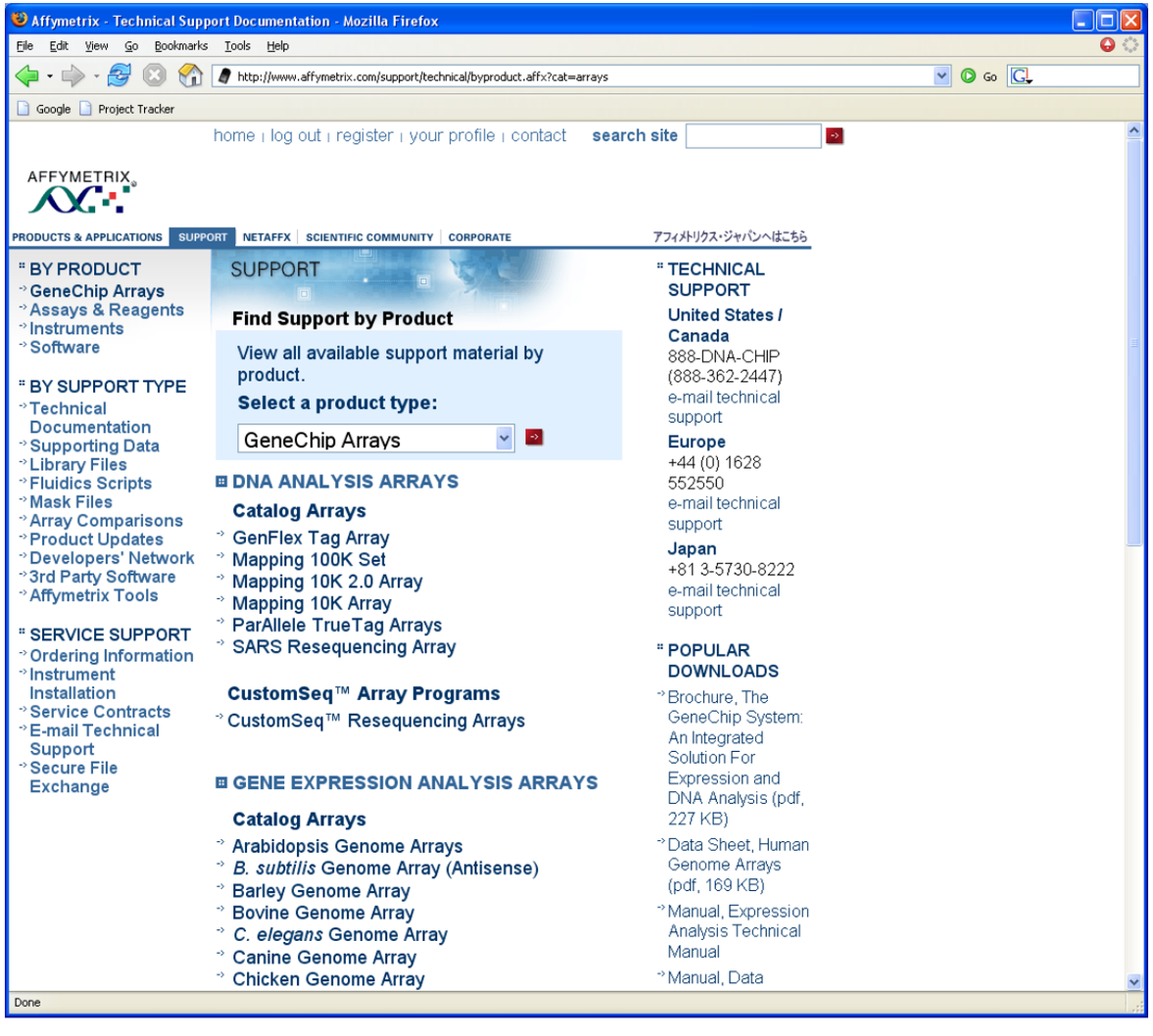


Affymetrix Support

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

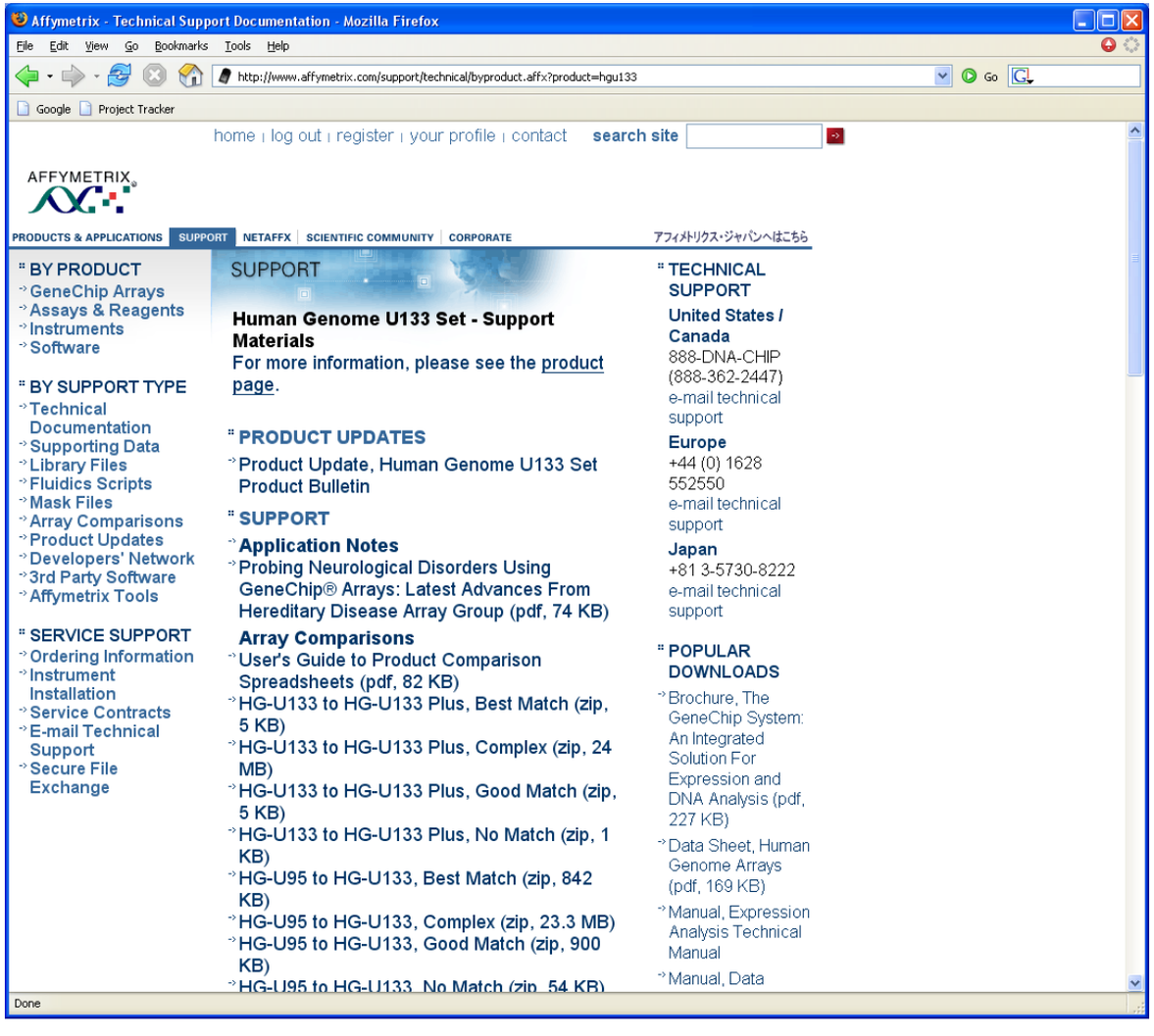
Support By Product

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



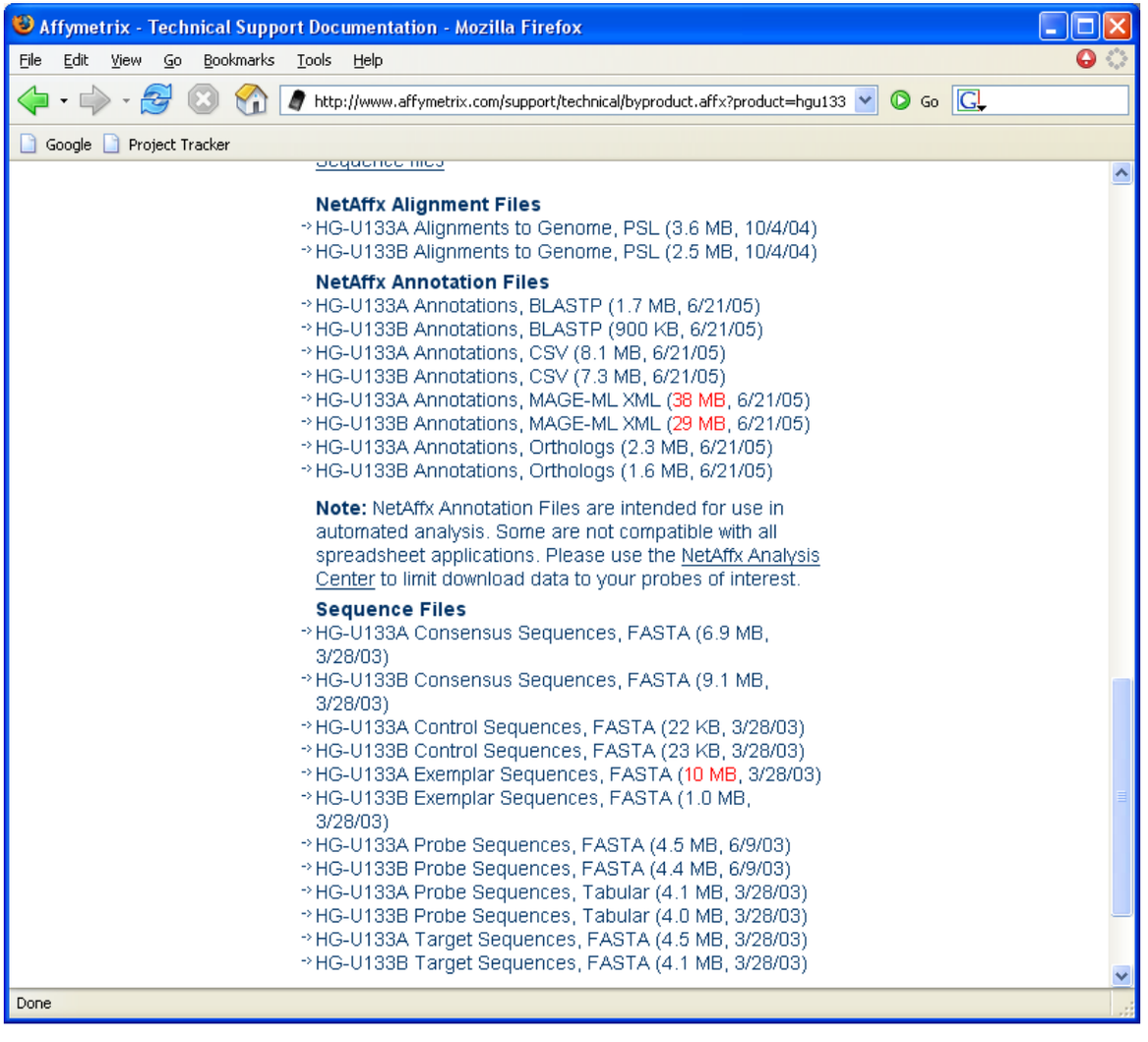
Affymetrix Annotations for HU133

Scroll down 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 are three primary annotation files:

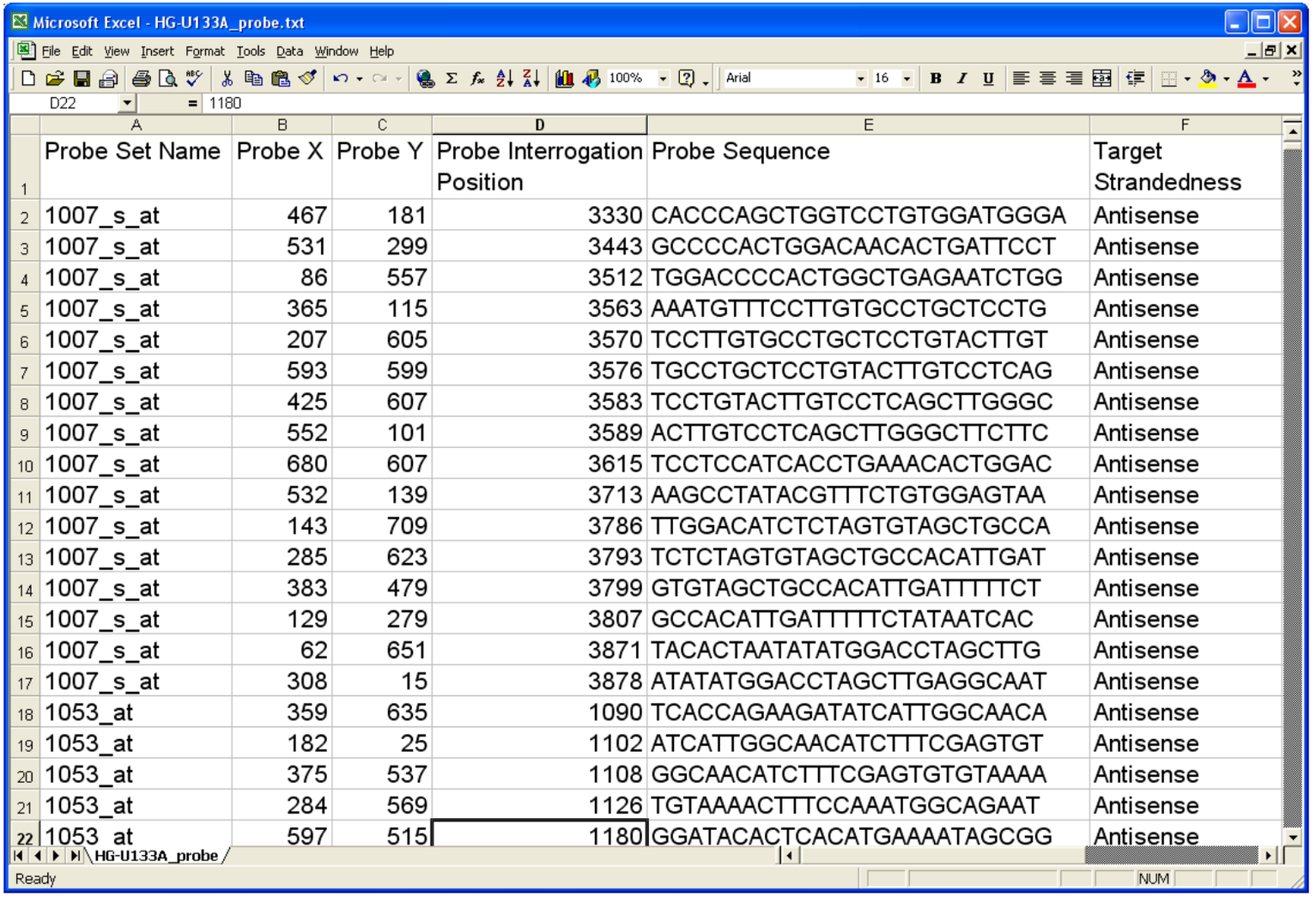
Probe Sequence File: Contains a complete listing of all the probes (25-mers) and probe sets on the microarray. (In tab-separated values format, the zipped file is 4.1MB; unzipped, it is 14.4MB.)

Alignment File: Contains mappings of targets and probes to the human genome. (In PSL format, the zipped file is 3.6MB; unzipped, it is 12.7MB.)

Annotation File: Contains the updated annotations of all the genes targeted by the microarray. (In comma-separated-value format, the zipped file is 6.1MB; unzipped, it is 47.9MB.)

Affymetrix Probe Sequences for HU133

The `HG-U133A_probe_tab` file lists the probe sequences.



Microsoft Excel - HG-U133A_probe.txt

File Edit View Insert Format Tools Data Window Help

D22 = 1180

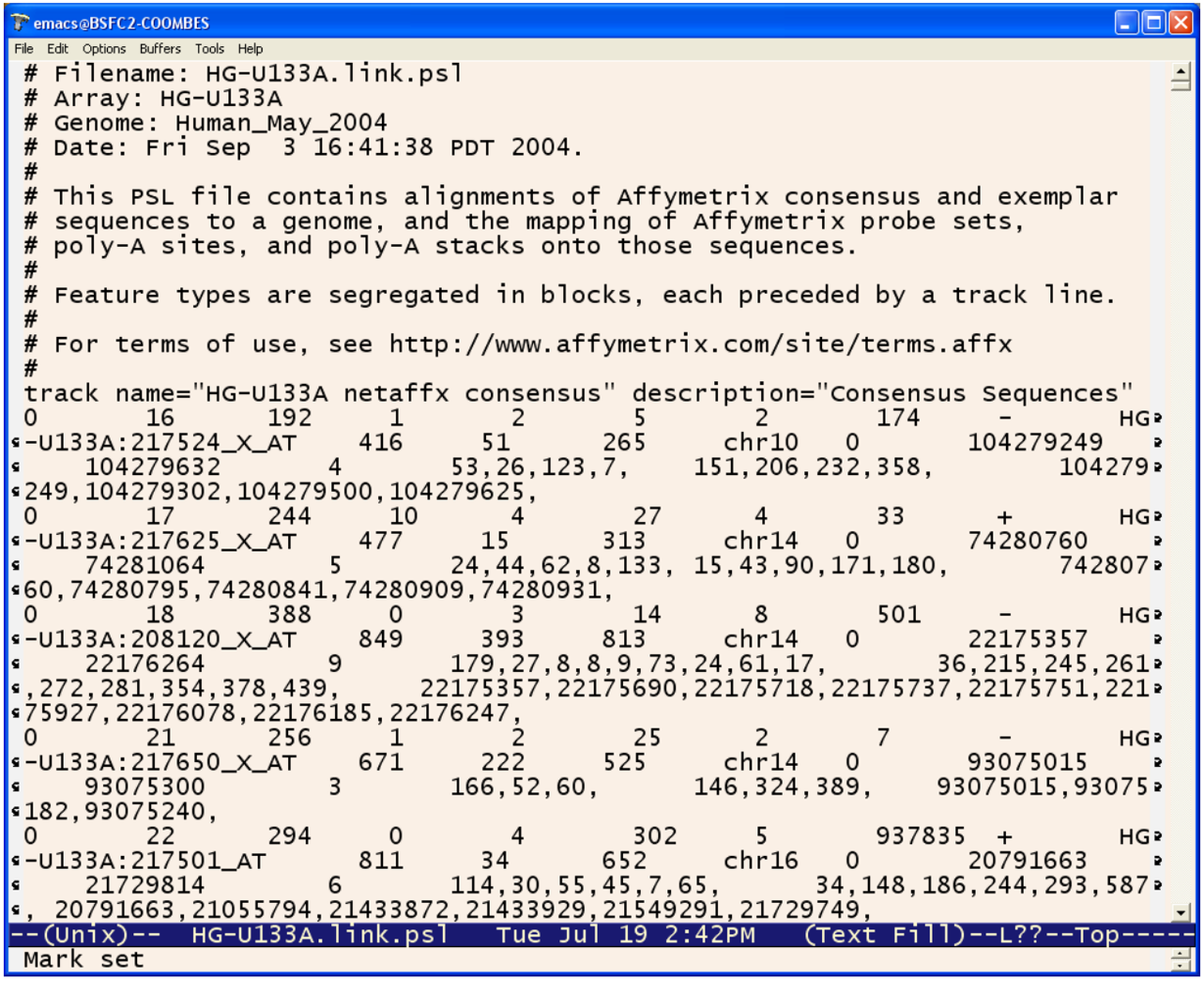
	A	B	C	D	E	F
1	Probe Set Name	Probe X	Probe Y	Probe Interrogation Position	Probe Sequence	Target Strandedness
2	1007_s_at	467	181	3330	CACCCAGCTGGTCCTGTGGATGGGA	Antisense
3	1007_s_at	531	299	3443	GCCCCACTGGACAACACTGATTCCT	Antisense
4	1007_s_at	86	557	3512	TGGACCCCACTGGCTGAGAATCTGG	Antisense
5	1007_s_at	365	115	3563	AAATGTTTCCTTGTGCCTGCTCCTG	Antisense
6	1007_s_at	207	605	3570	TCCTTGTGCCTGCTCCTGTACTTGT	Antisense
7	1007_s_at	593	599	3576	TGCCTGCTCCTGTACTTGTCCTCAG	Antisense
8	1007_s_at	425	607	3583	TCCTGTACTTGTCCTCAGCTTGGGC	Antisense
9	1007_s_at	552	101	3589	ACTTGTCCTCAGCTTGGGCTTCTTC	Antisense
10	1007_s_at	680	607	3615	TCCTCCATCACCTGAAACACTGGAC	Antisense
11	1007_s_at	532	139	3713	AAGCCTATACGTTTCTGTGGAGTAA	Antisense
12	1007_s_at	143	709	3786	TTGGACATCTCTAGTGTAGCTGCCA	Antisense
13	1007_s_at	285	623	3793	TCTCTAGTGTAGCTGCCACATTGAT	Antisense
14	1007_s_at	383	479	3799	GTGTAGCTGCCACATTGATTTTTCT	Antisense
15	1007_s_at	129	279	3807	GCCACATTGATTTTTCTATAATCAC	Antisense
16	1007_s_at	62	651	3871	TACACTAATATATGGACCTAGCTTG	Antisense
17	1007_s_at	308	15	3878	ATATATGGACCTAGCTTGAGGCAAT	Antisense
18	1053_at	359	635	1090	TCACCAGAAGATATCATTGGCAACA	Antisense
19	1053_at	182	25	1102	ATCATTGGCAACATCTTTCGAGTGT	Antisense
20	1053_at	375	537	1108	GGCAACATCTTTCGAGTGTGTAAAA	Antisense
21	1053_at	284	569	1126	TGTAAAACTTTCCAAATGGCAGAAT	Antisense
22	1053_at	597	515	1180	GGATACACTCACATGAAAATAGCGG	Antisense

HG-U133A_probe

Ready NUM

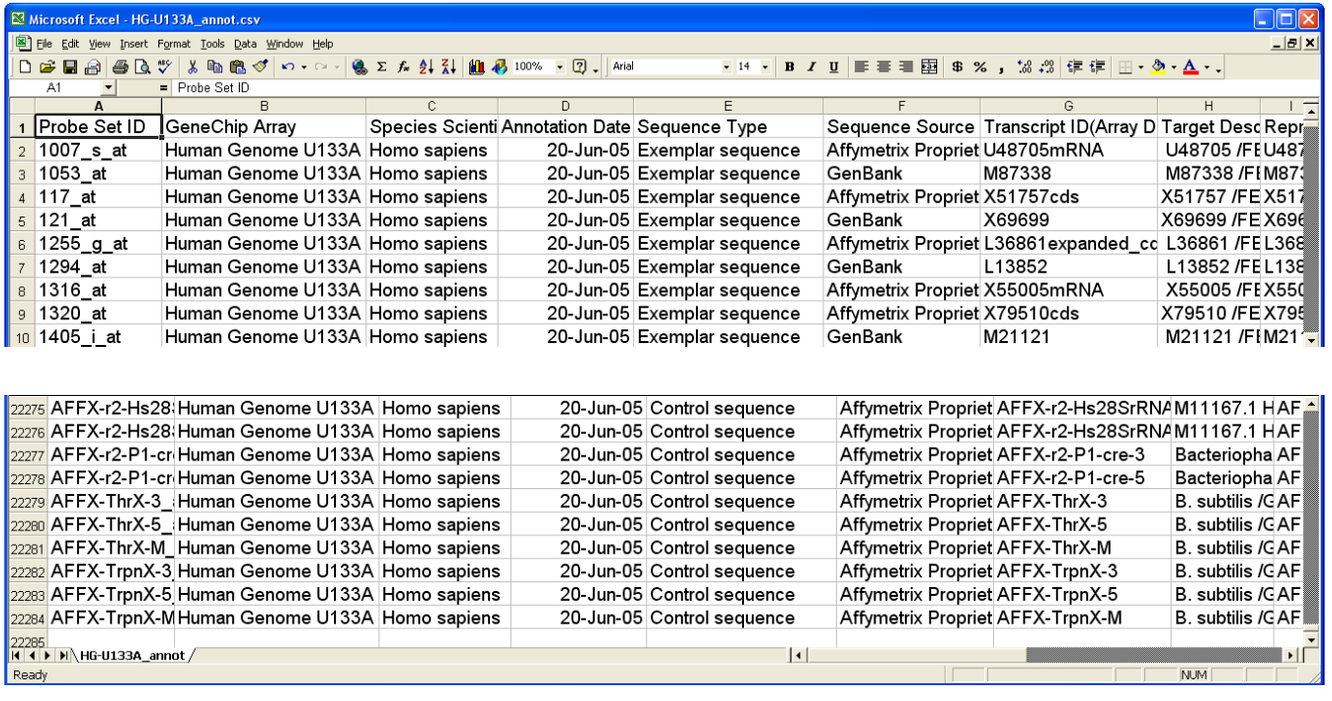
Affymetrix Alignment Information for HU133

The `HG-U133A.link.psl` file aligns the probes to the genome.



What annotations does Affymetrix supply?

As noted earlier, HG-U133A_annot.csv contains 47.9MB worth of annotation information. What occupies all that space?



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

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									

First, we note that the file seems to contain redundant copies of lots of information. Second, it has information on 22,283 probe sets, one per line, in 43 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. The 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 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 (either 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.

Protein Families. Families to which the protein belongs.

Protein Domains. Domains included in the protein.

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.

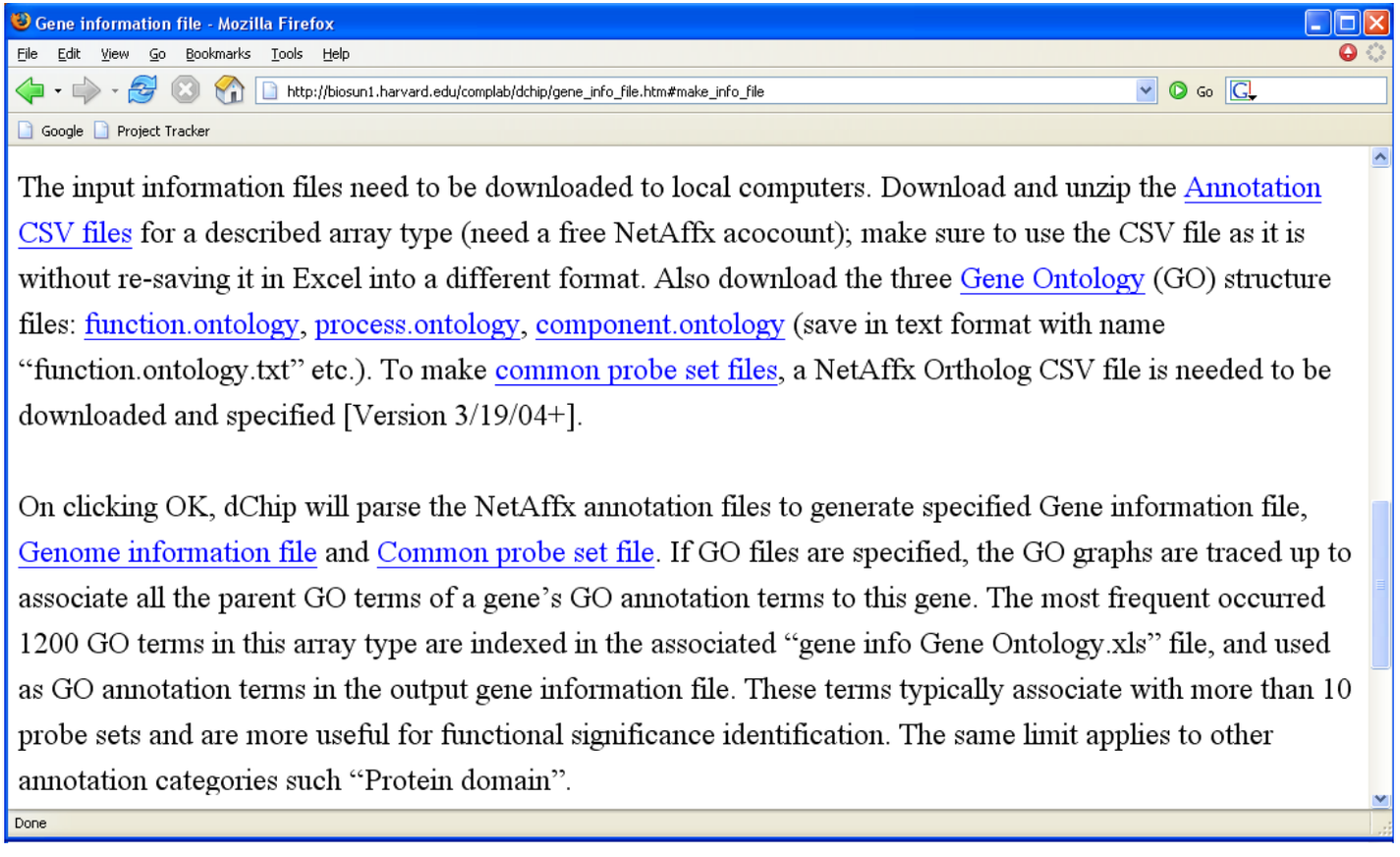
Updating annotations in dChip

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.

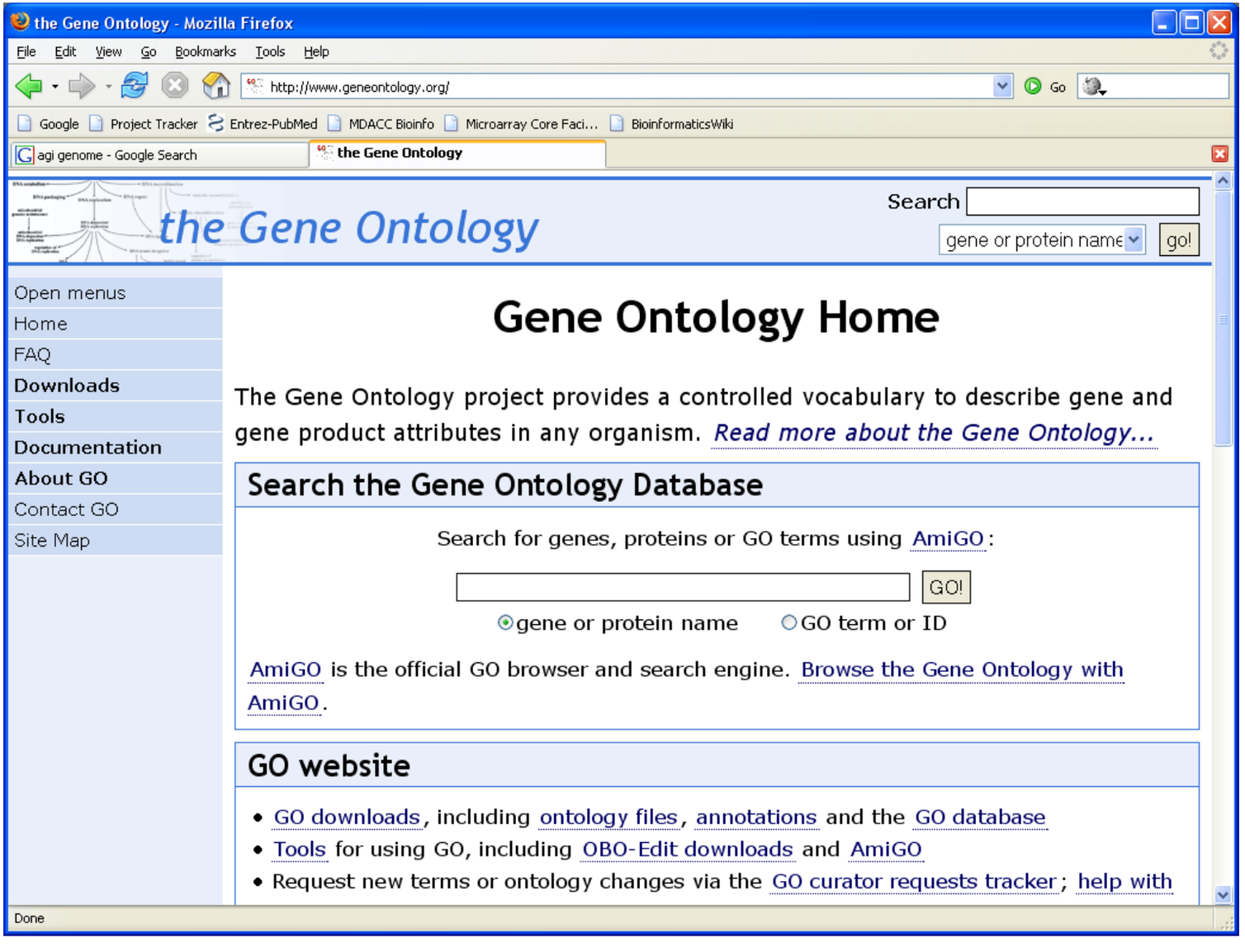
For many common chip types, the `dChip` web site contains up-to-date copies of these files. It's still useful to see where the data comes from how and how you can update your own versions.

dChip Manual on Gene Information



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



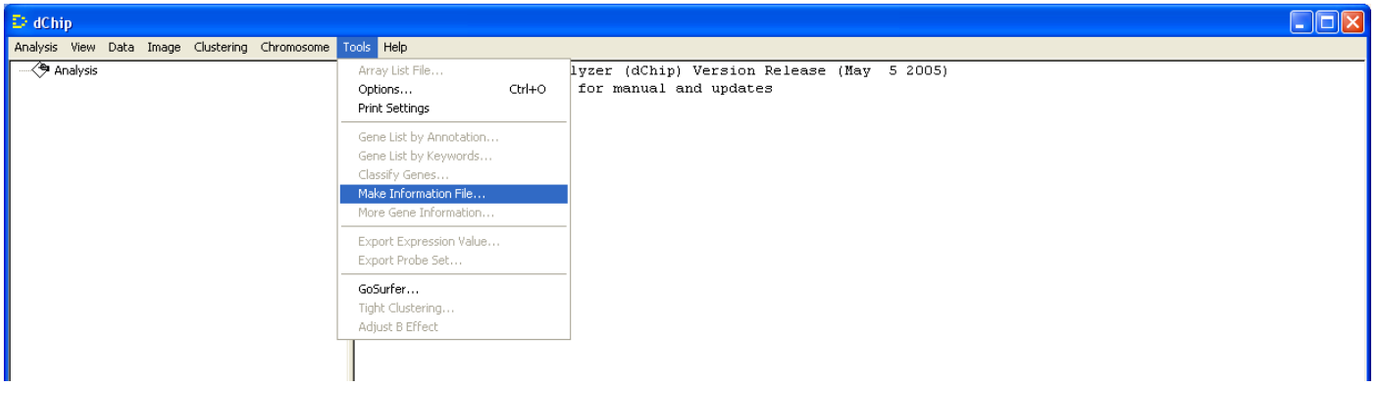
Making the Gene Information file

1. Get and unzip the file containing the updated annotation CSV file from Affymetrix.

2. Get the three updated text files from GeneOntology.

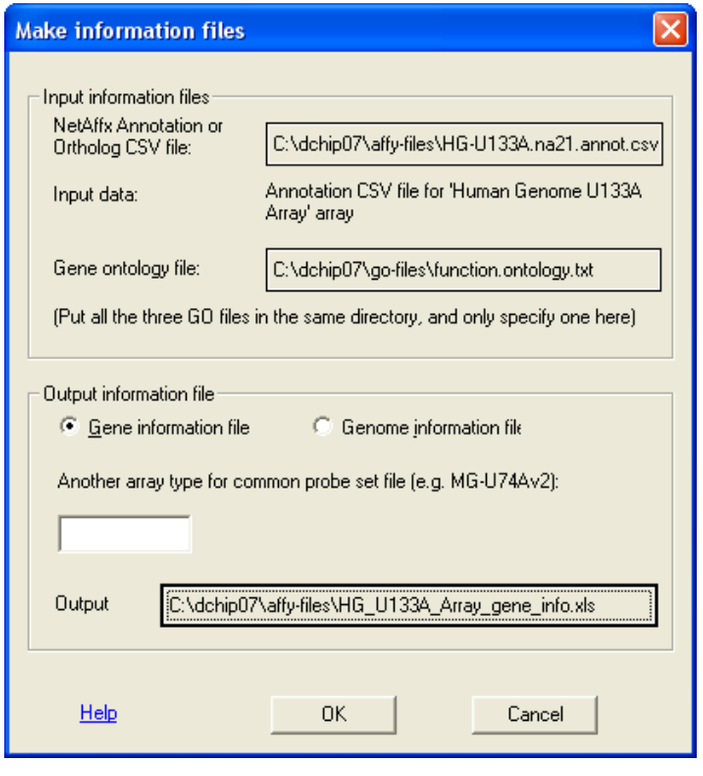
3. Rename the three GeneOntology files, adding the ".txt" extension.

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

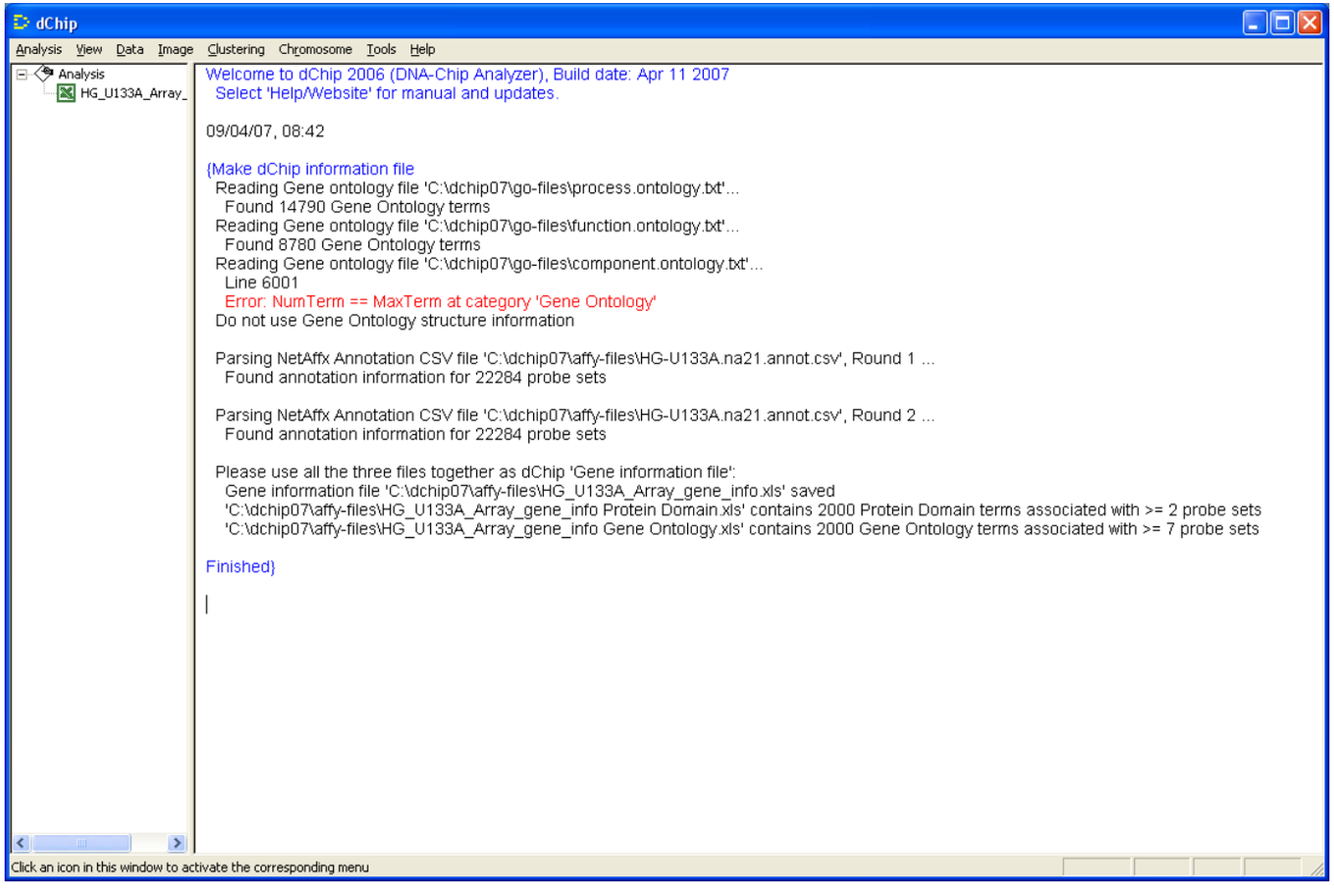


Making the Gene Information file

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



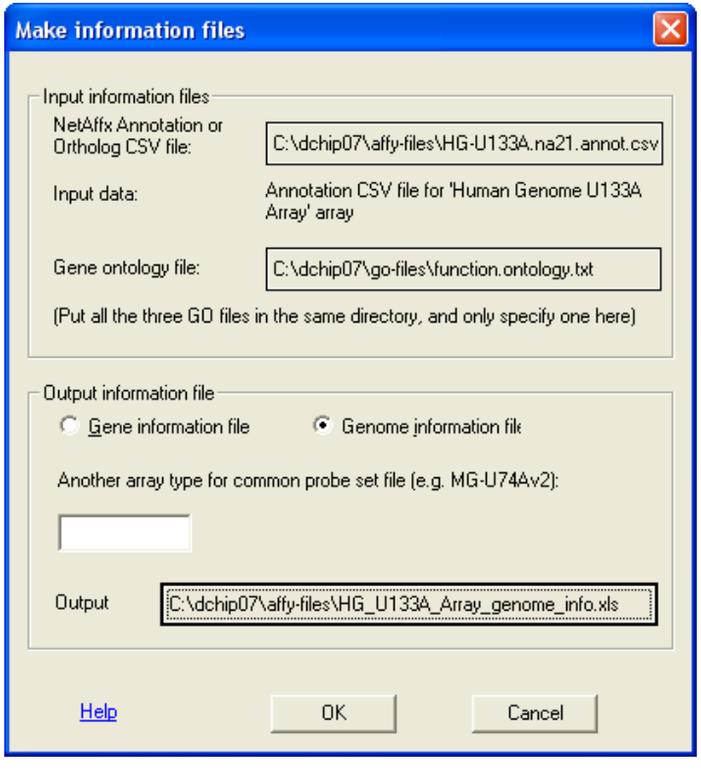
The Gene Information file



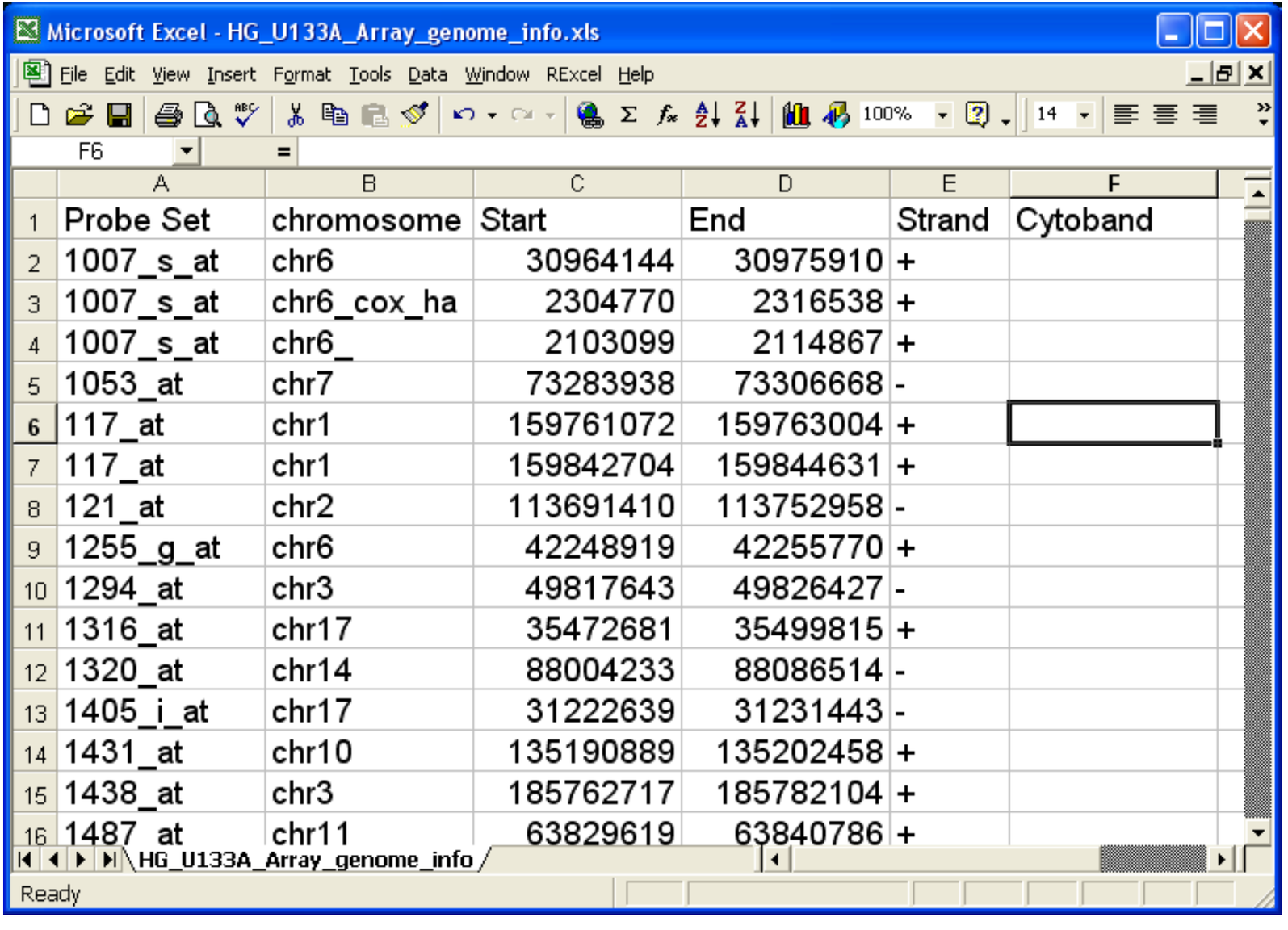
Note the error message! This should be fixed in the "latest" version. This step produces the three `dChip` annotation files that were described in Lecture 2.

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

File Edit View Insert Format Tools Data Window REXcel Help

F6 =

	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	+	

HG_U133A_Array_genome_info

Ready

What is GeneOntology?

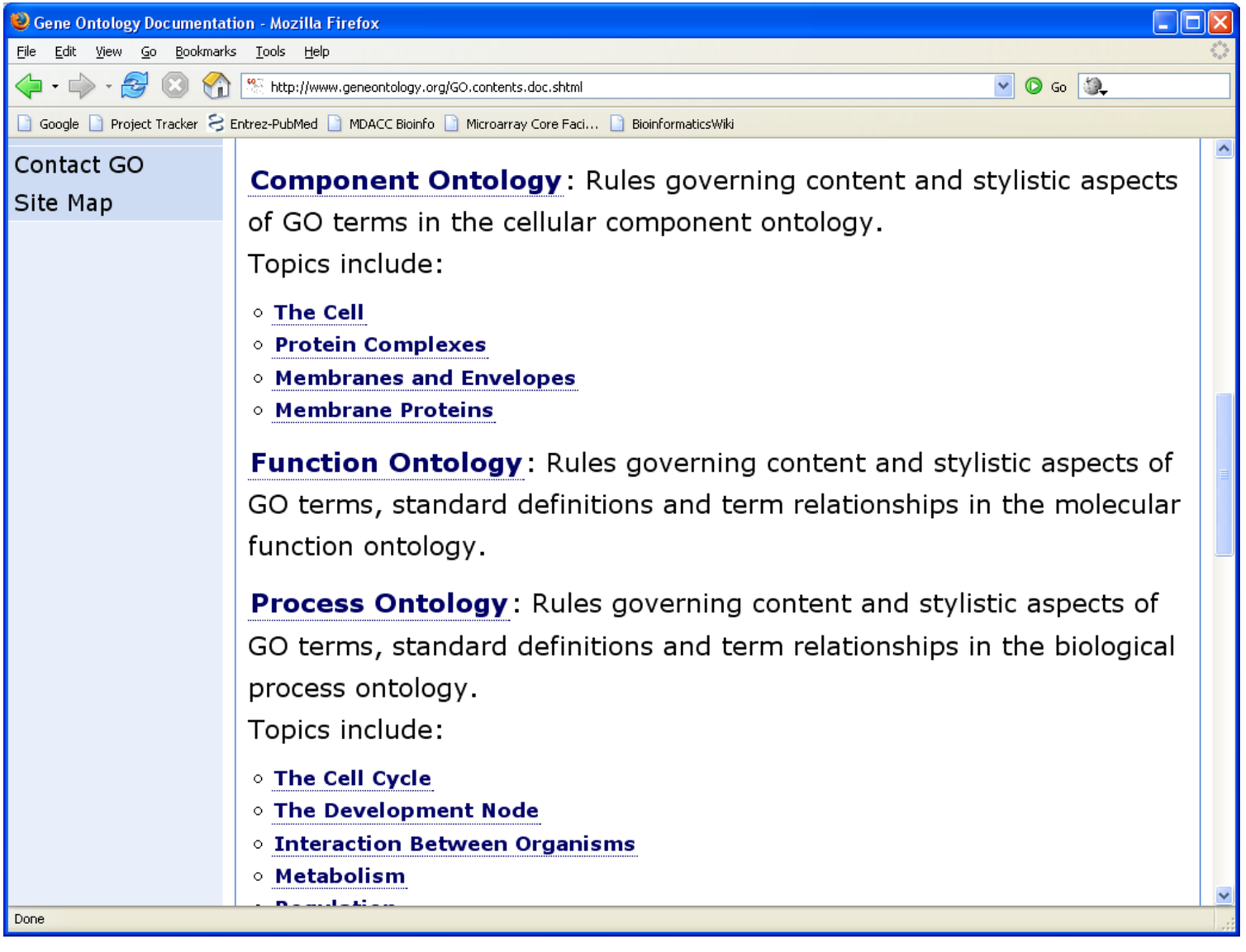
GeneOntology uses controlled vocabularies to create a directed acyclic graph (DAG; a generalized tree) that describes the kinds of functions or properties that a gene might have. There are two parts to GeneOntology:

- 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

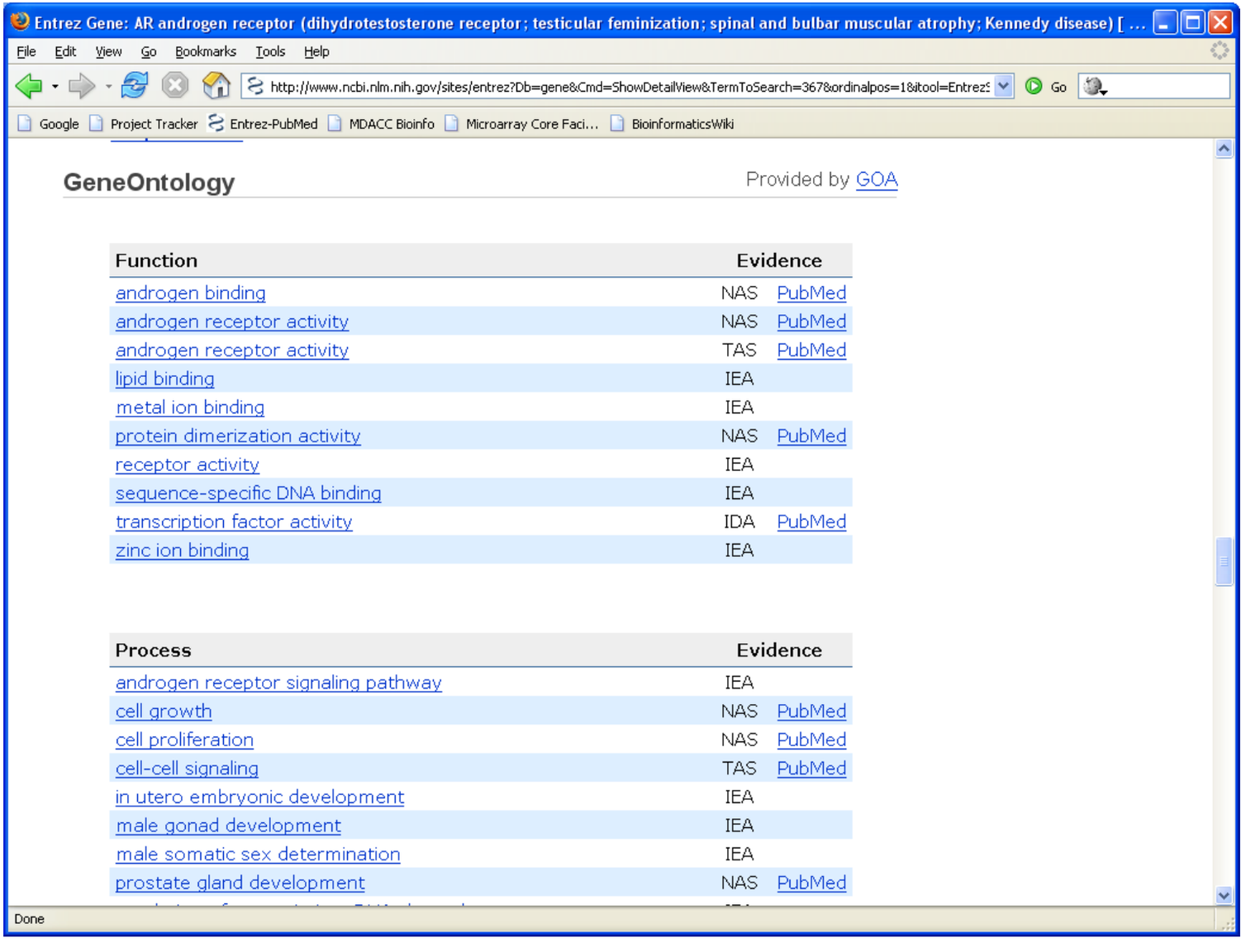


GeneOntology annotations in Entrez Gene

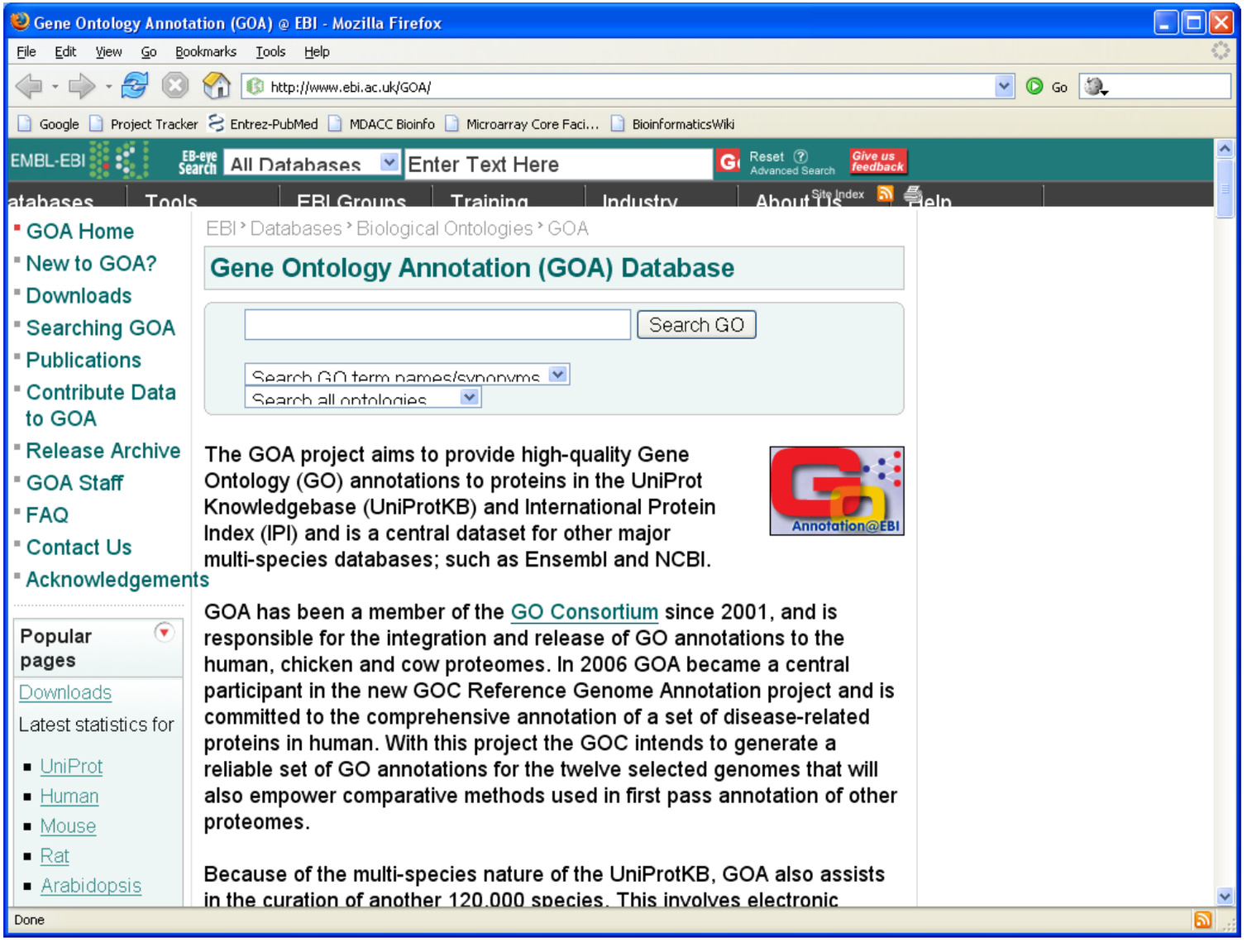
You can find the GeneOntology annotations for individual genes in Entrez Gene. For genes with known functions, the Entrez Gene page will contain a section titled “GeneOntology”, which contains a list of the known functions for that gene.

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.

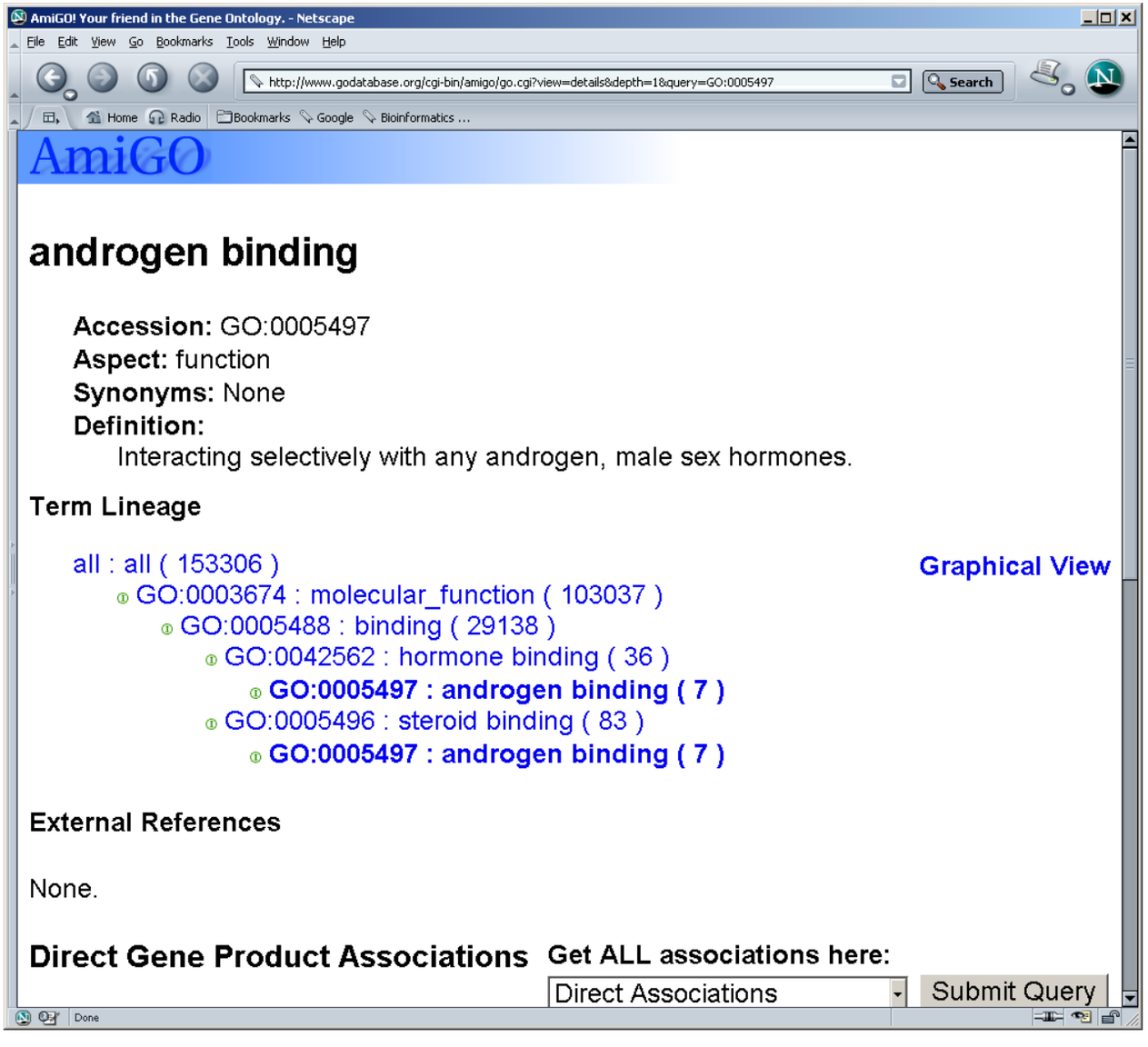
GO annotations of the androgen receptor



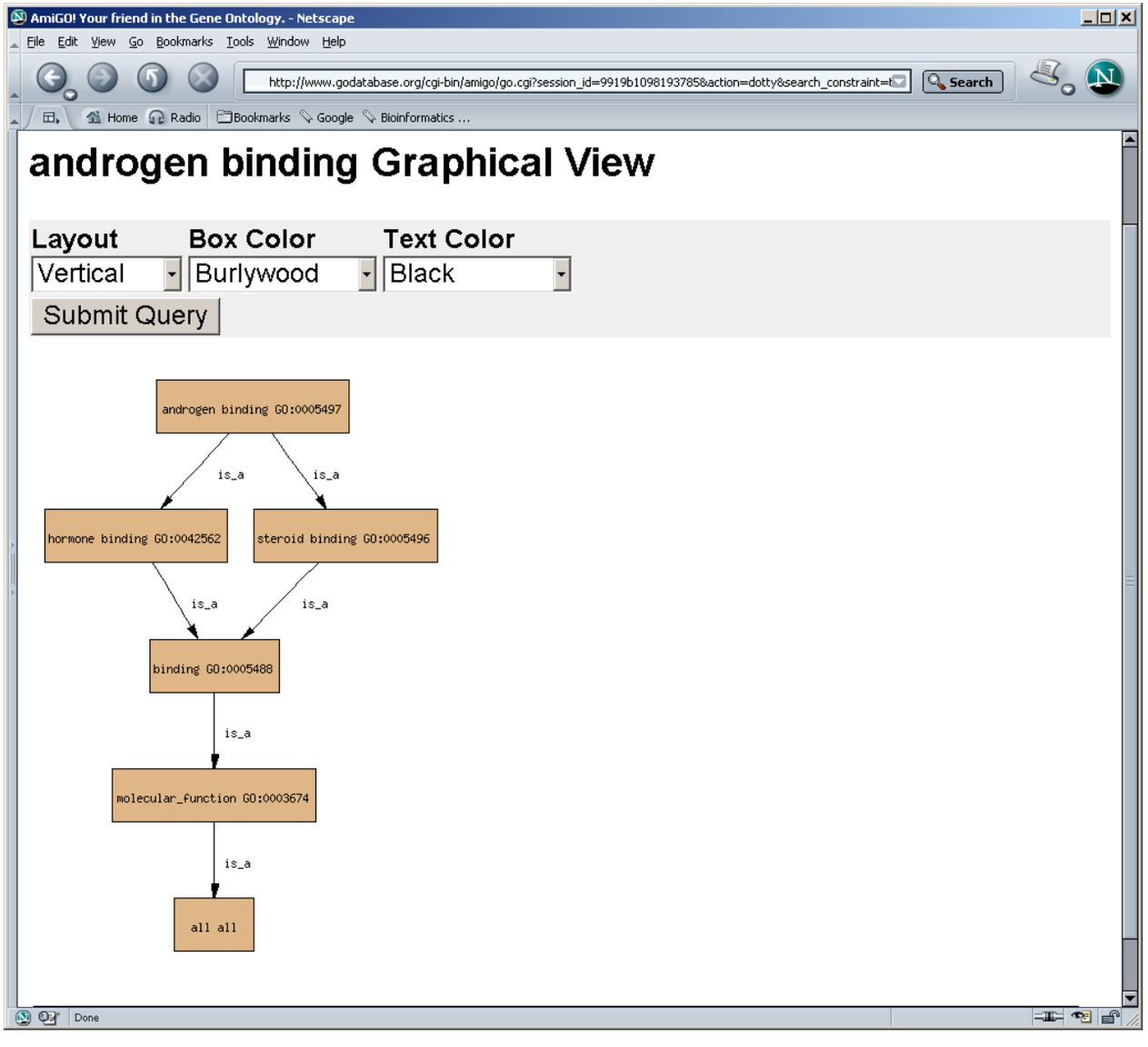
http://www.ebi.ac.uk/GOA/



GO browsing



GO browsing



Edges are relationships

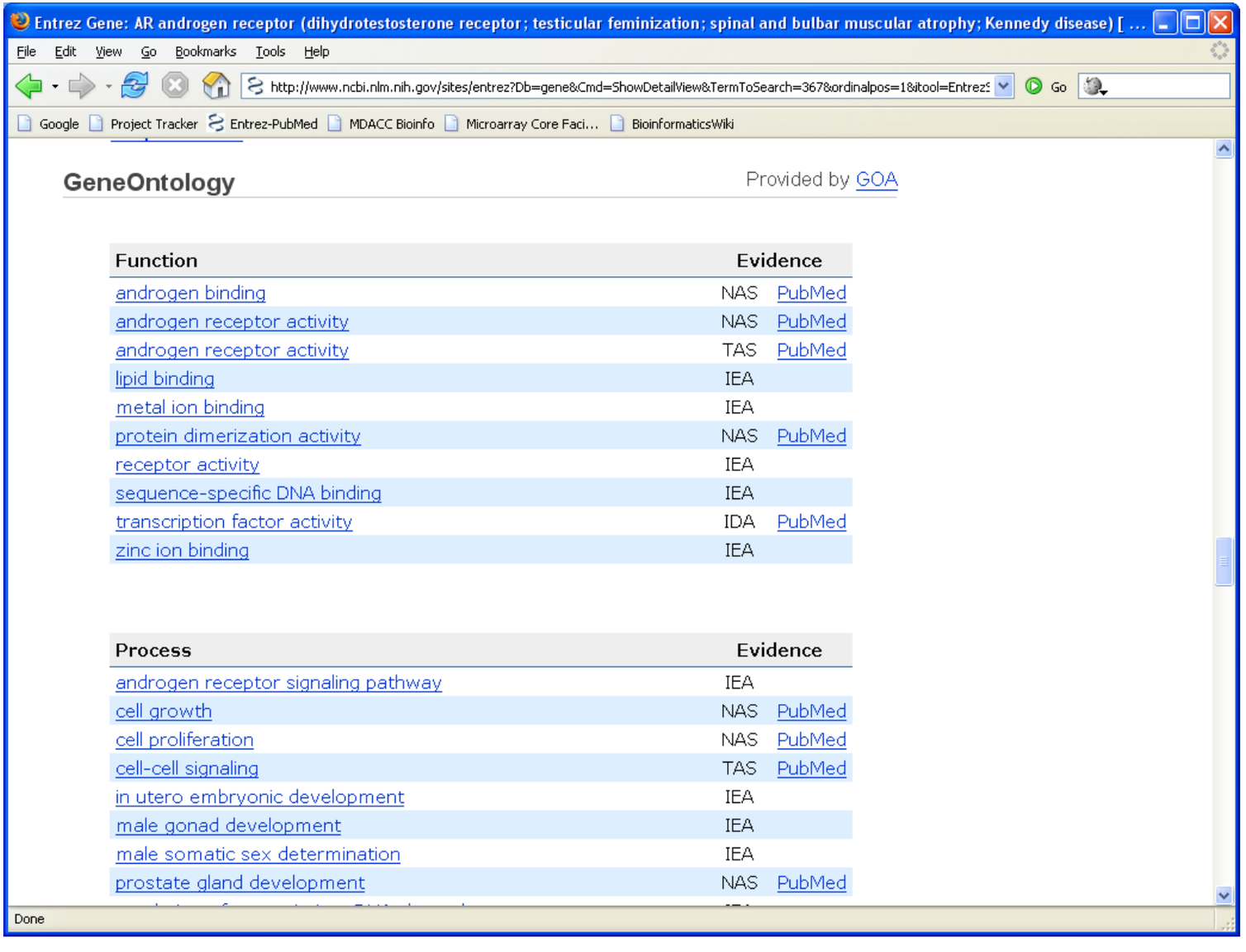
Edges in the DAG represent two kinds of relationships:

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: 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 that includes 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

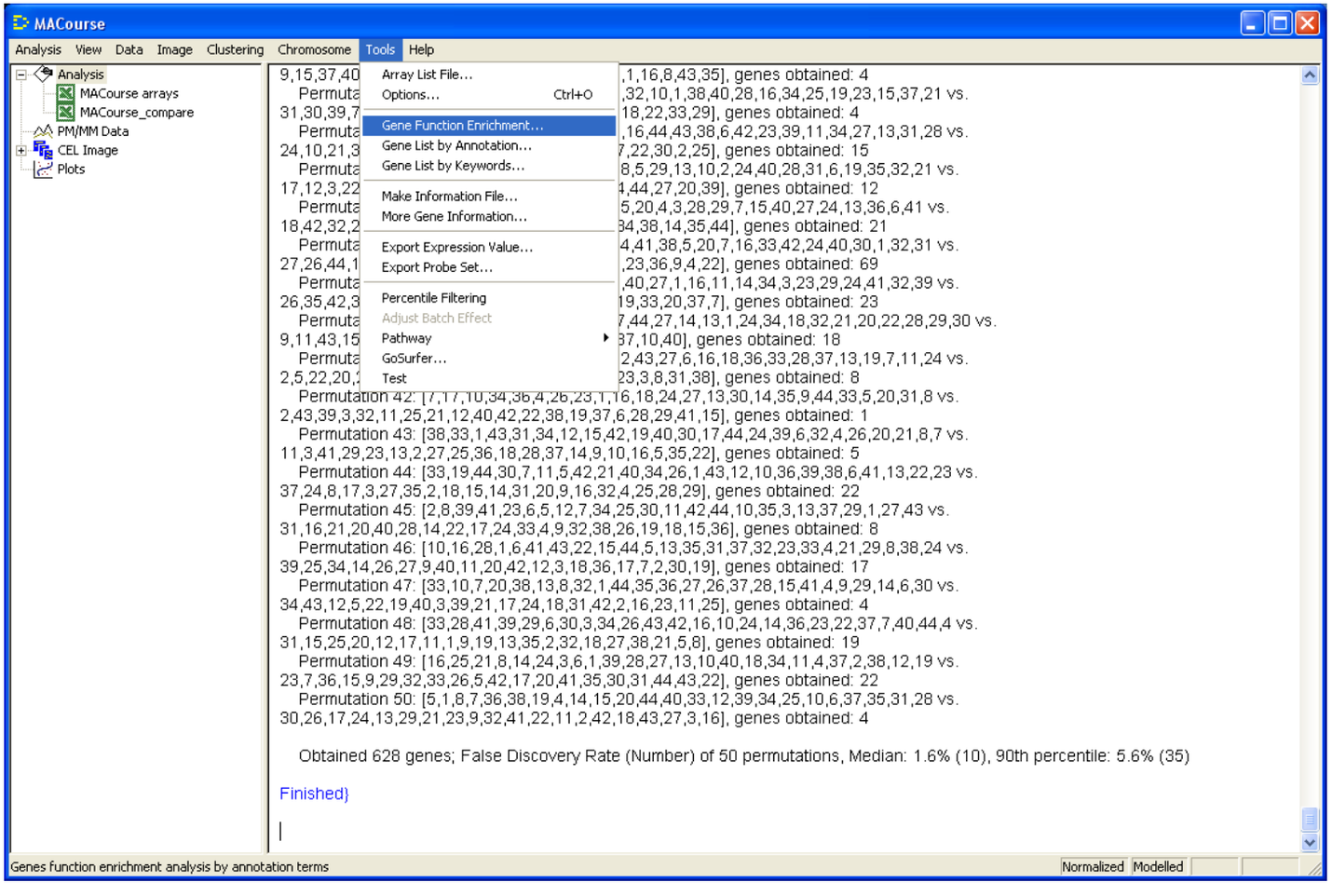
Quality of evidence

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

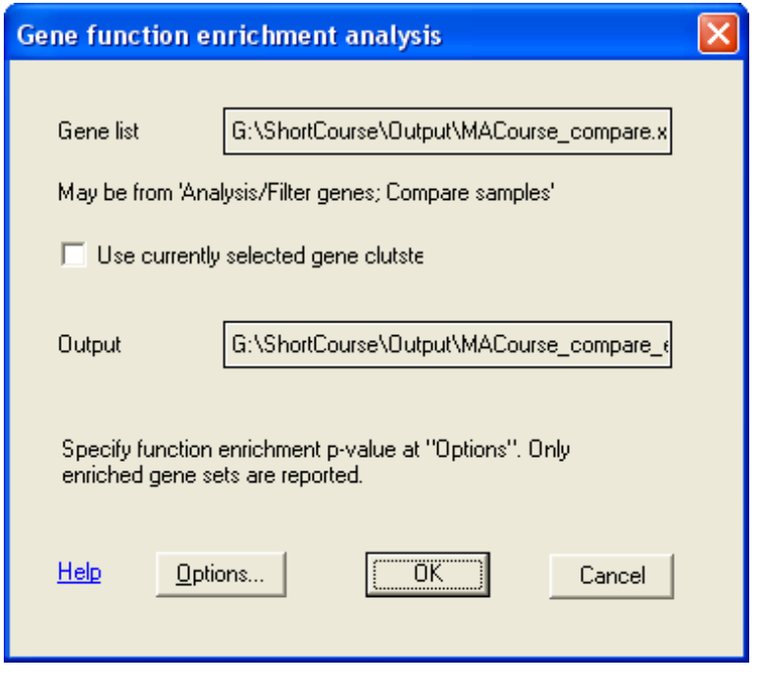
Using GeneOntology in dChip

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



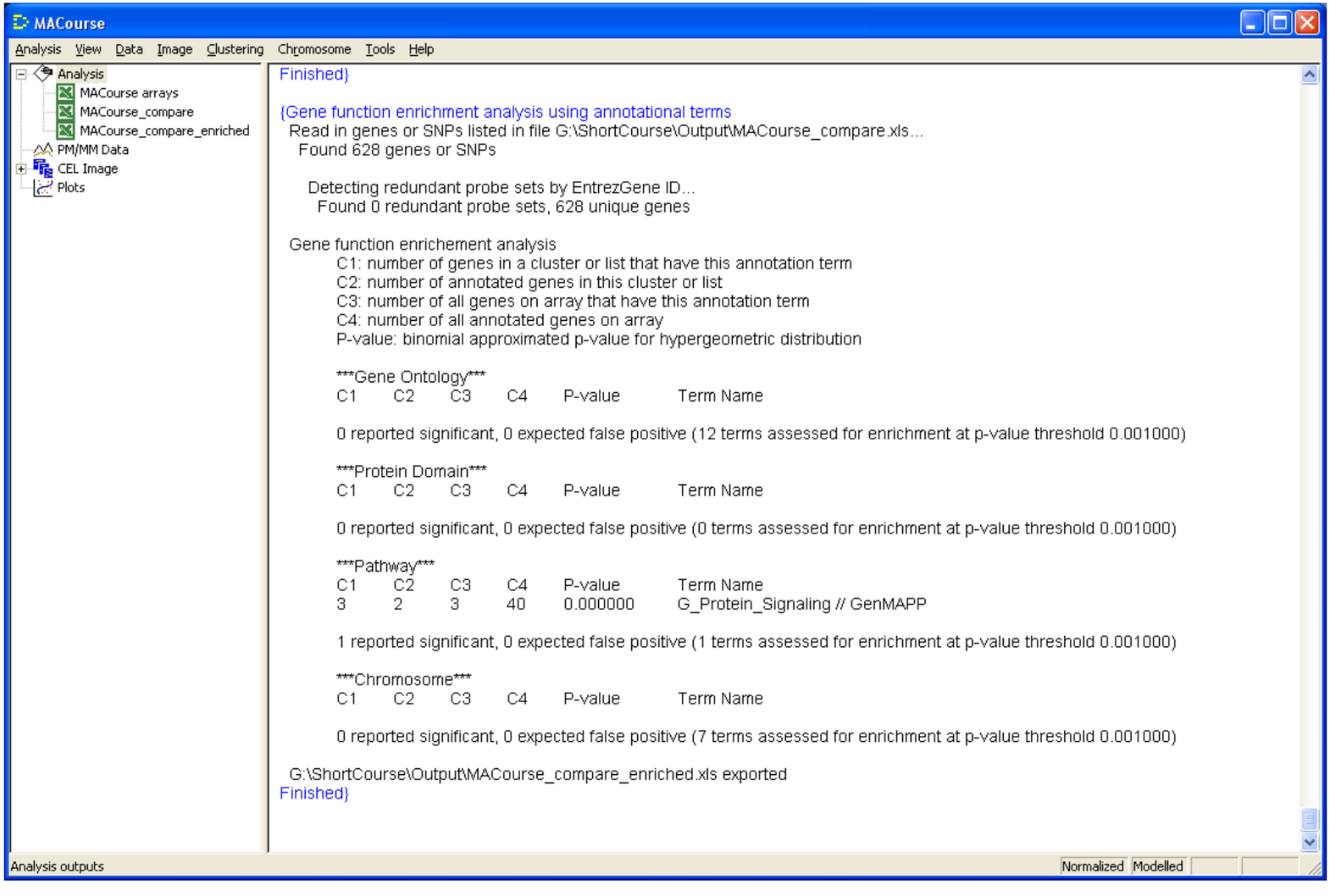
Using GeneOntology in dChip

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

File Edit View Insert Format Tools Data Window Help

A1 = probe set

	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

Note: No longer get these results; result of earlier error?

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.

The numbers tell us:

1. There were 7685 probe sets on the array with some kind of GeneOntology annotation.

2. There were 391 probe sets selected as differentially expressed that had some kind of GeneOntology annotation.

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

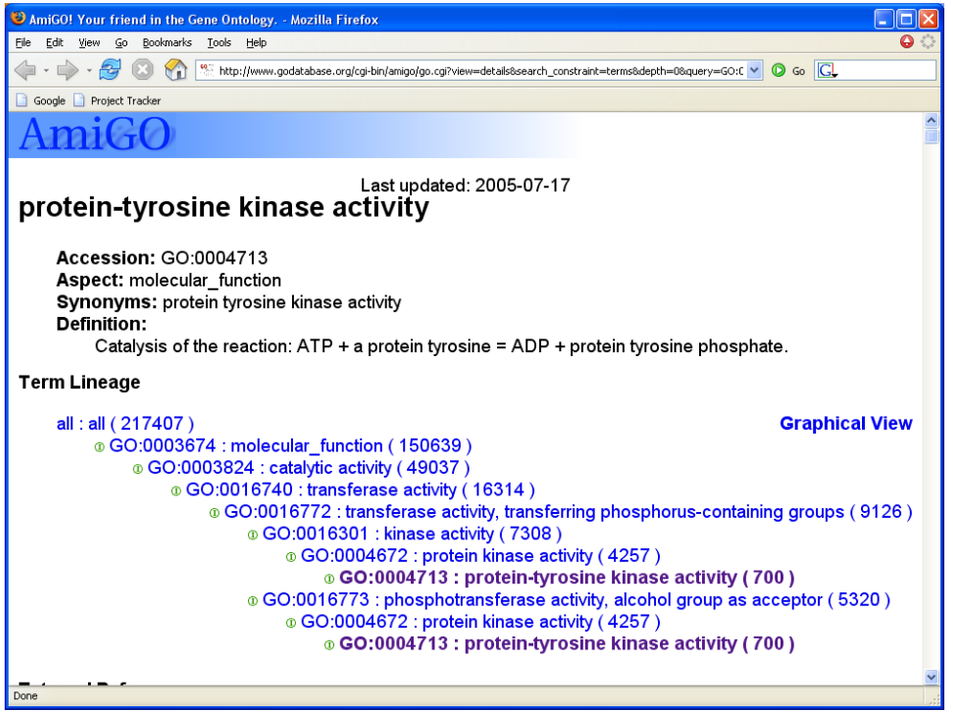
4. Of the selected annotated probe sets, 21 had the “protein tyrosine kinase” function.

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?

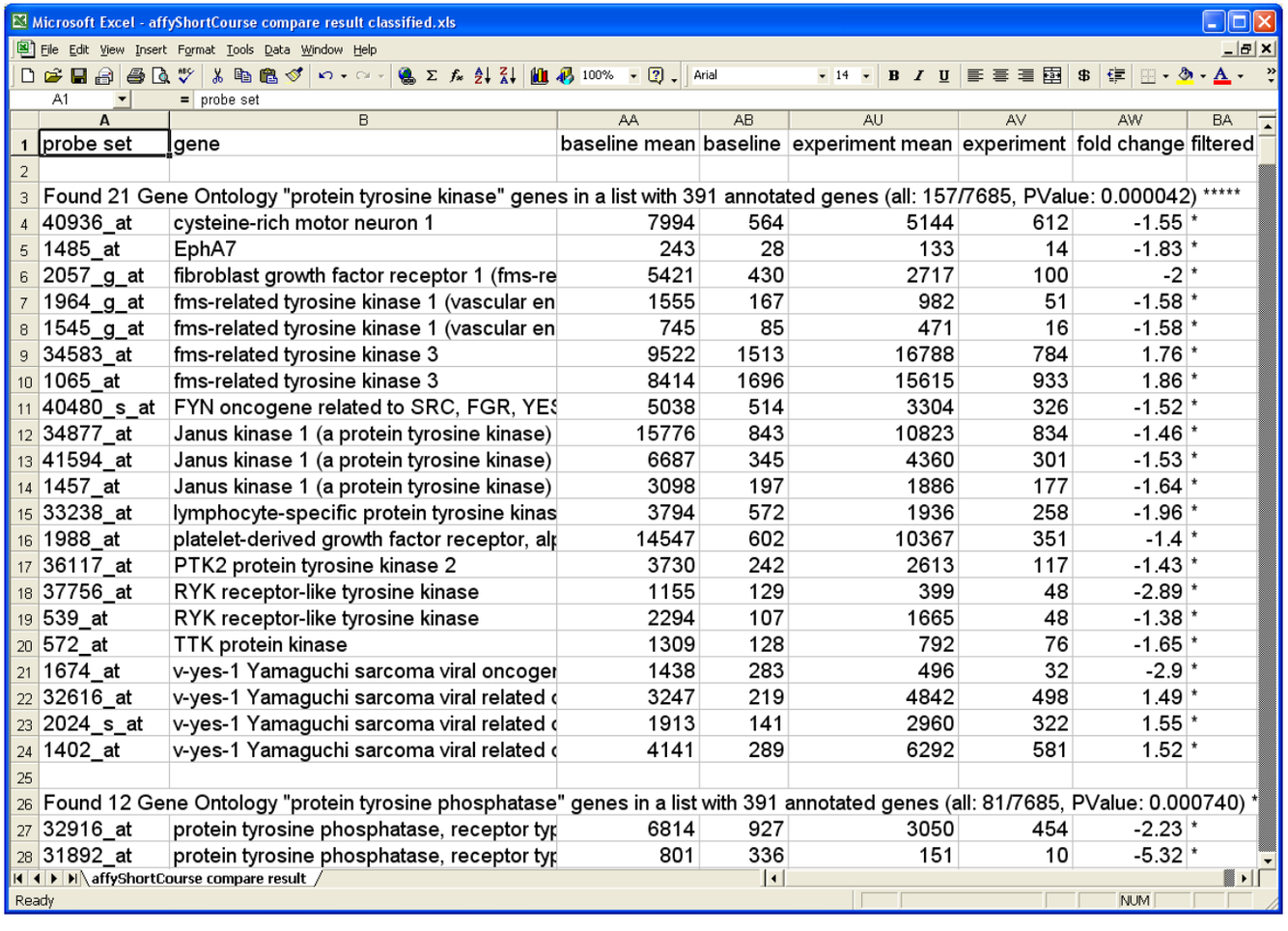
First, the p-values have not been adjusted for multiple testing.

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

File Edit View Insert Format Tools Data Window Help

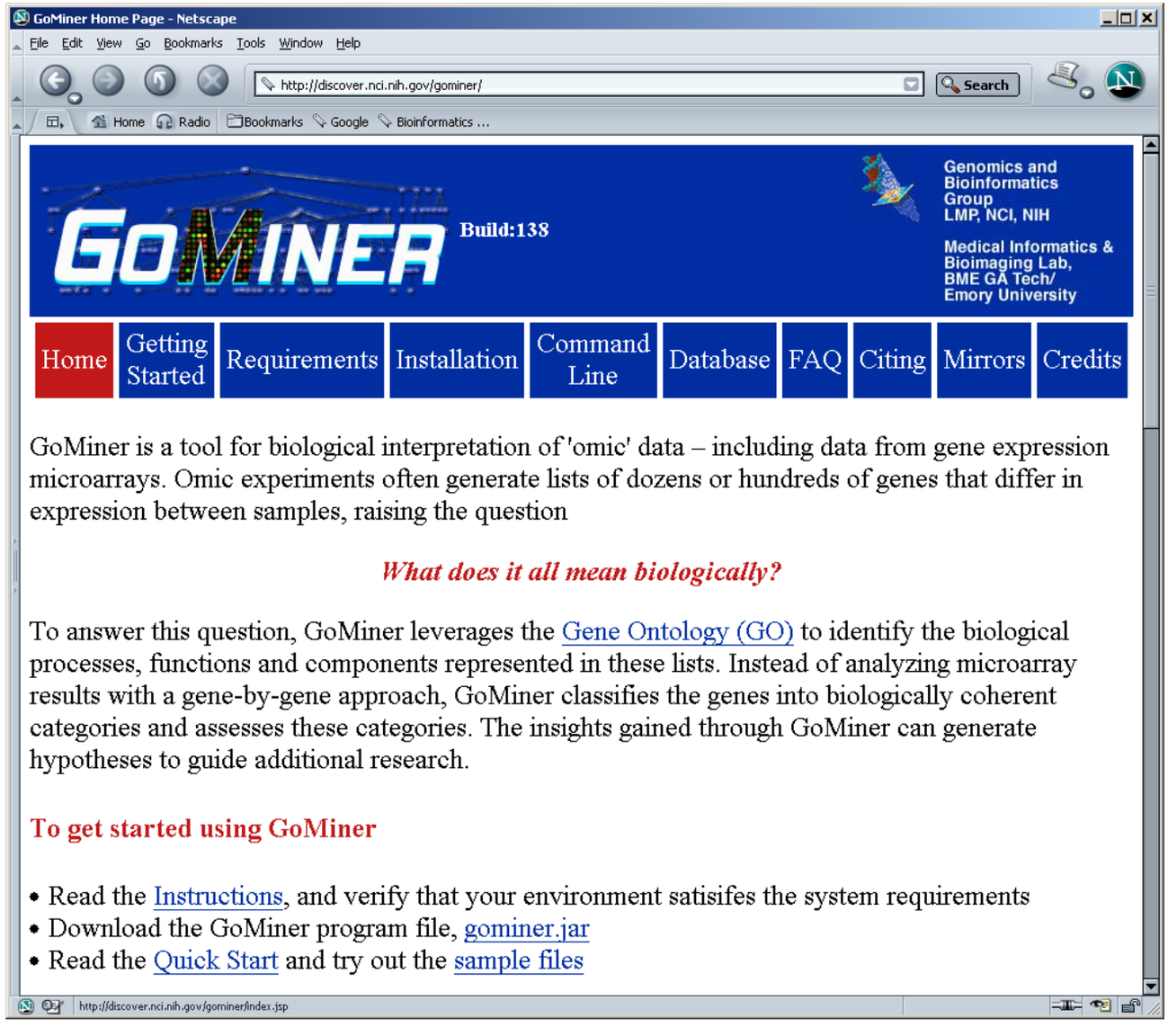
A1 = probe set

	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?



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

GoMiner: Getting Started

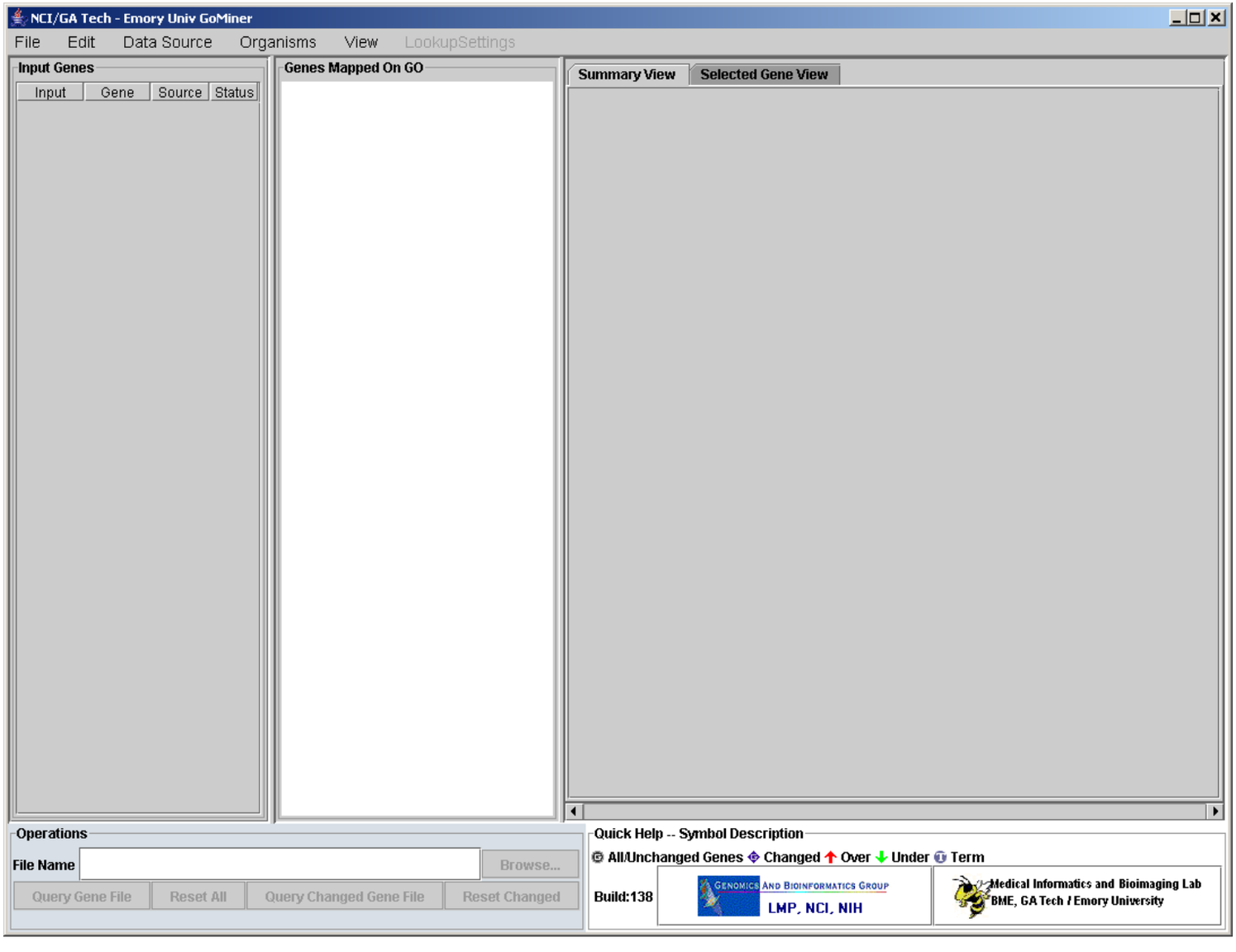
You need a machine with

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

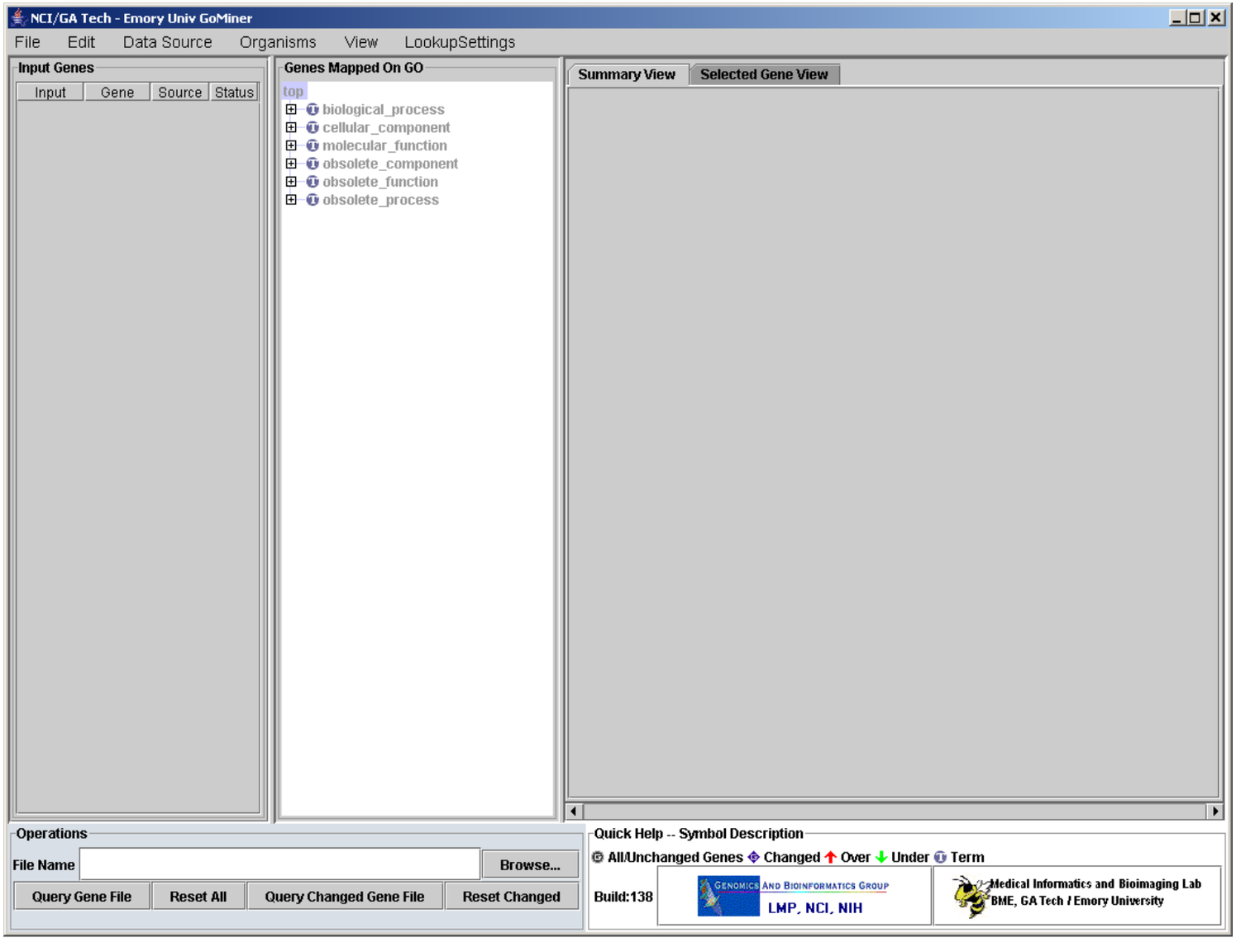
Download the GoMiner Java code, install it, and double-click on it to start the program.

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

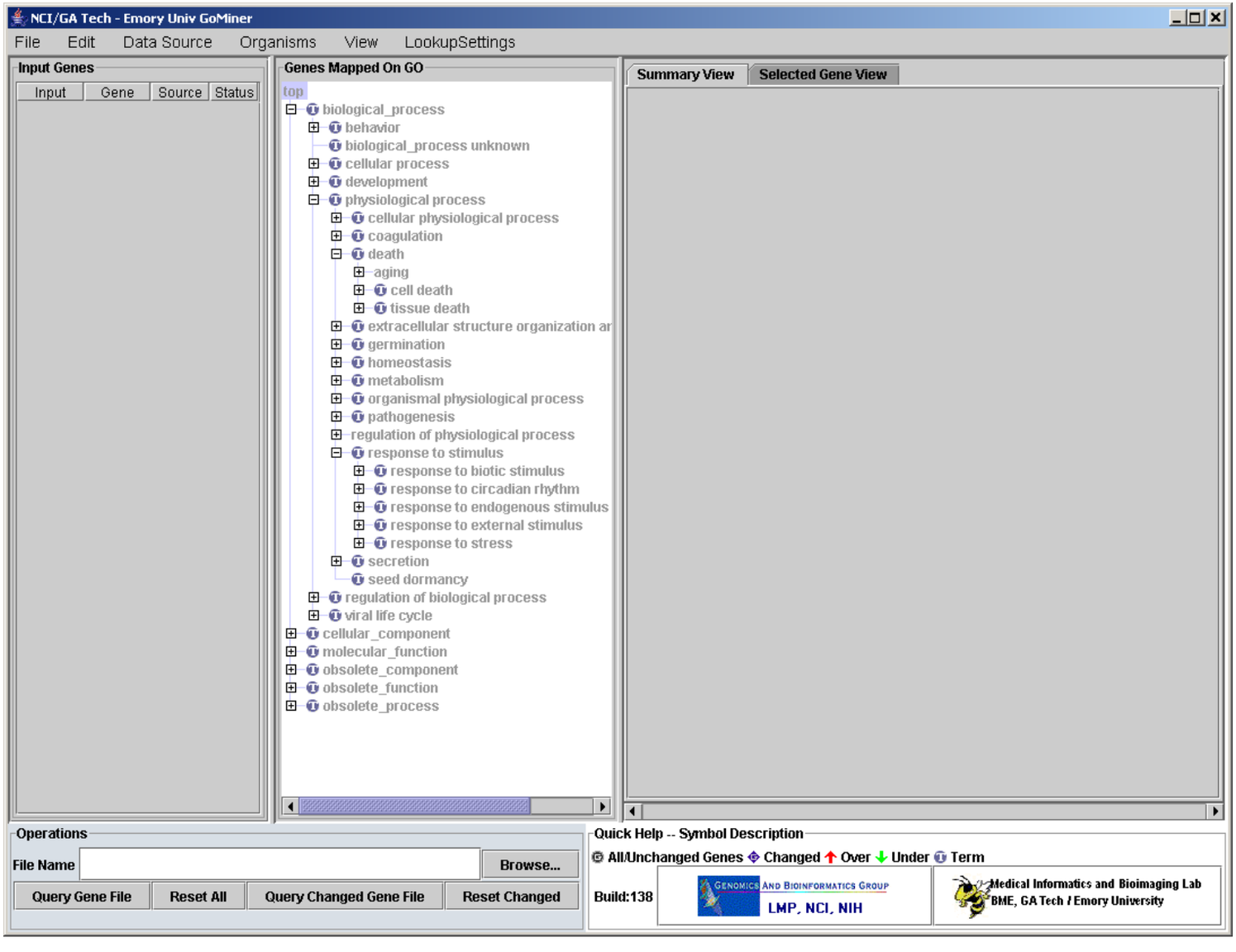
GoMiner Start



GoMiner: GO terms loaded



GoMiner as GO browser



Getting array data into GoMiner

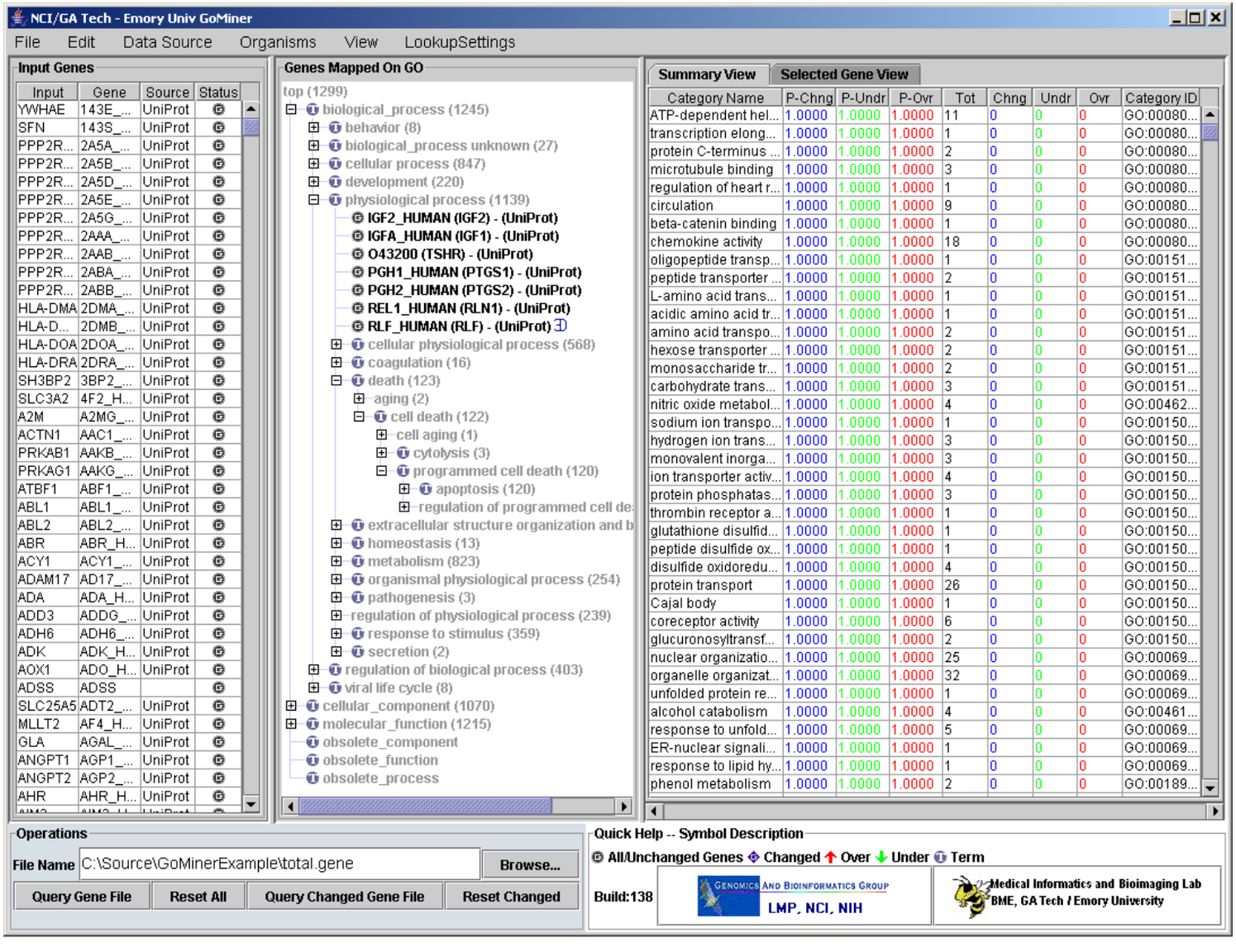
1. Go to “Data Source” and select “UniProt (Hs)” to restrict to human gene annotations

2. Need a file containing a list of 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 may take some time...

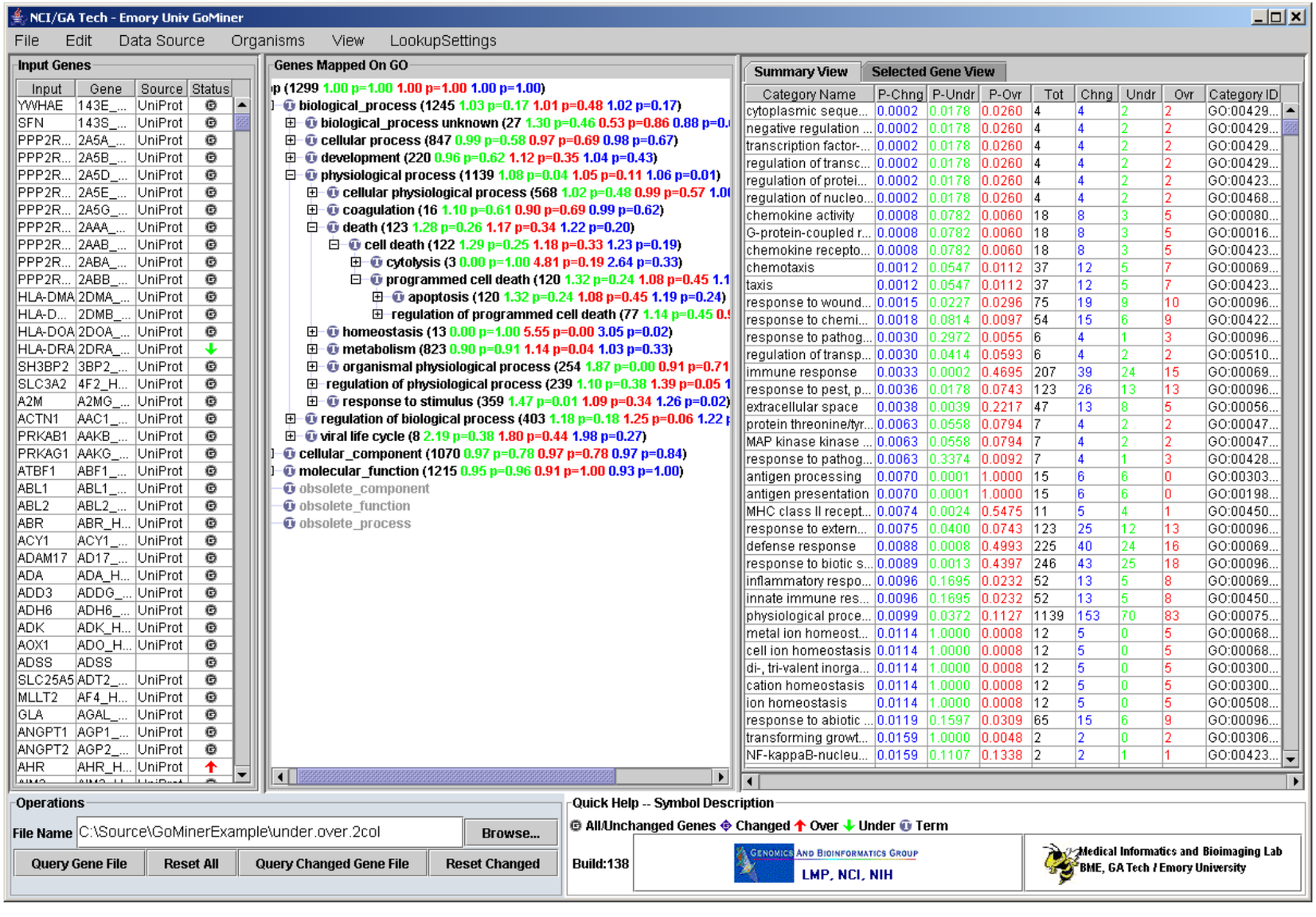
3. Need a file containing a list of genes that changed. Can be one HUGO symbol per line. Optionally, you can include a second column with 1 (overexpressed) or -1 (underexpressed). Use “Browse” and “Query Changed Gene File” to load this data.

Note: GeneLink or Source can convert from various gene ids to HUGO symbols.

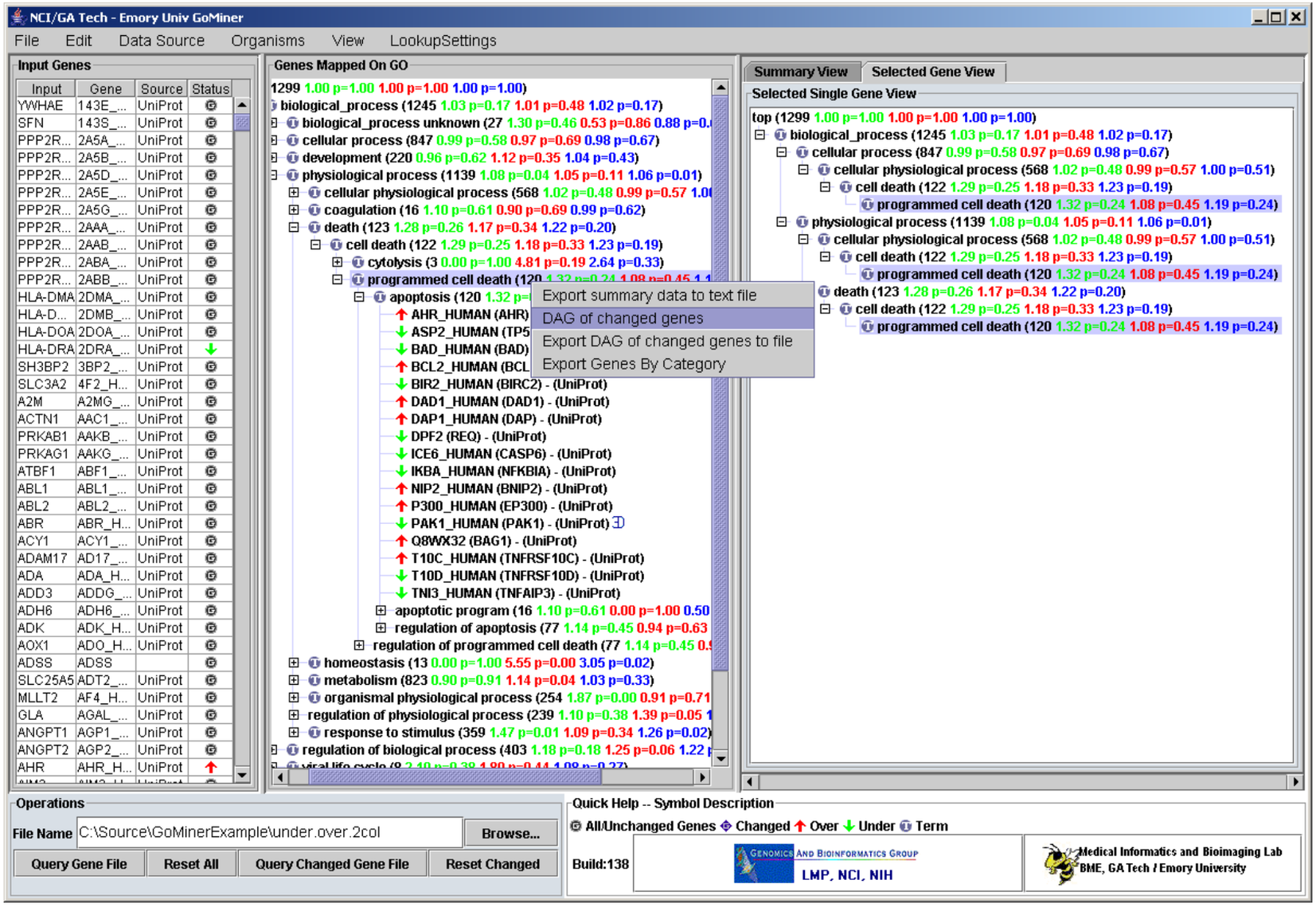
GoMiner with array gene list loaded



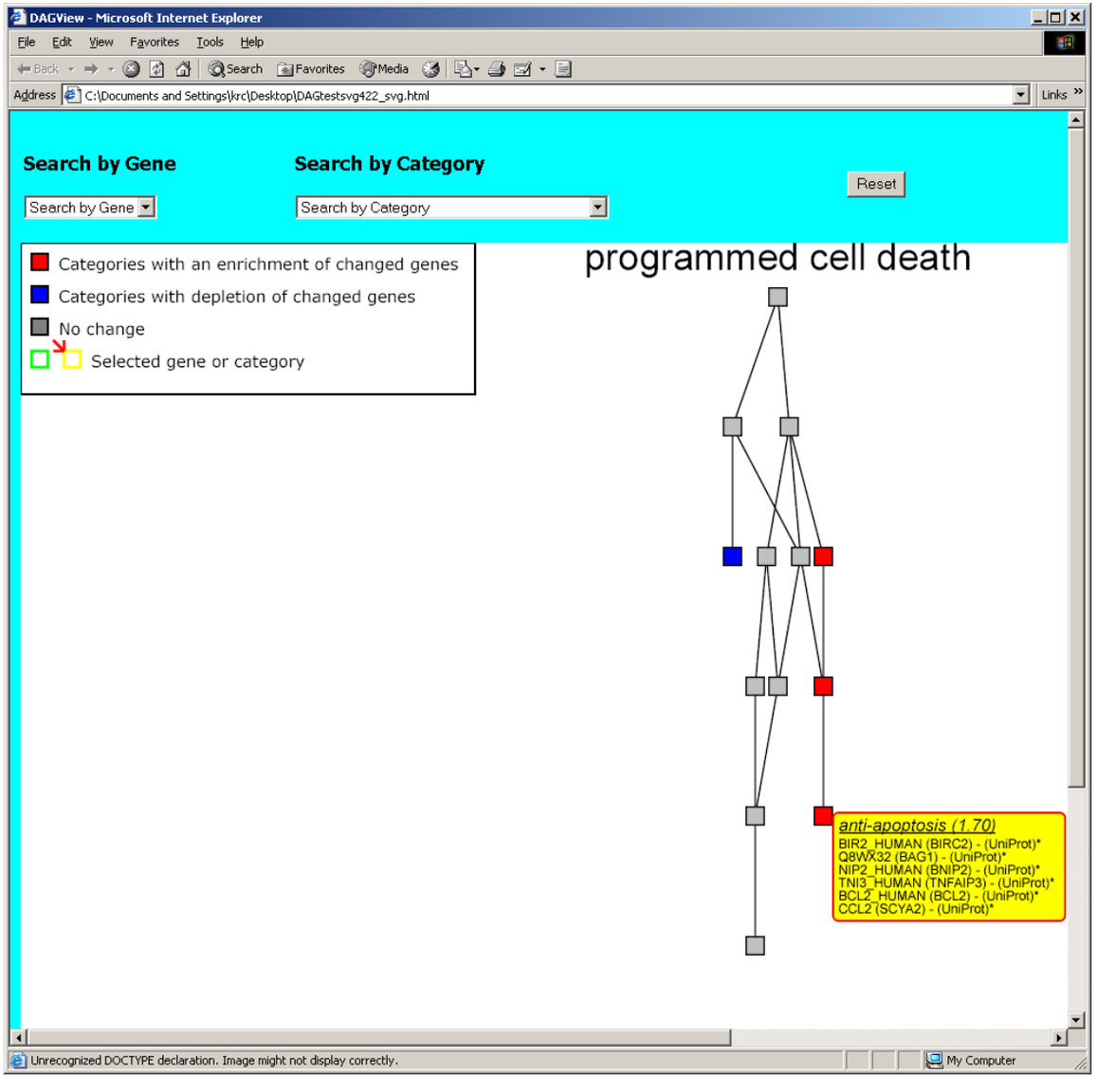
GoMiner with changed gene list loaded



GoMiner subgraphs



GoMiner subgraphs



Intepreting GoMiner results

Enrichment is computed as

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

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 one filters the gene list from the array before testing differential expression (for example, 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 is not completely appropriate, since genes can have multiple overlapping annotations into the GO DAG.

No existing test exploits the quality of evidence for the GO annotations.