R for Gene Ontology Analysis

Genome-wide studies

- Sanger sequencing
- Next generation sequencing
- Microarray
- ChIP-chip and ChIP-seq
- RNA-seq
- Tandem mass spectrometry
- ……

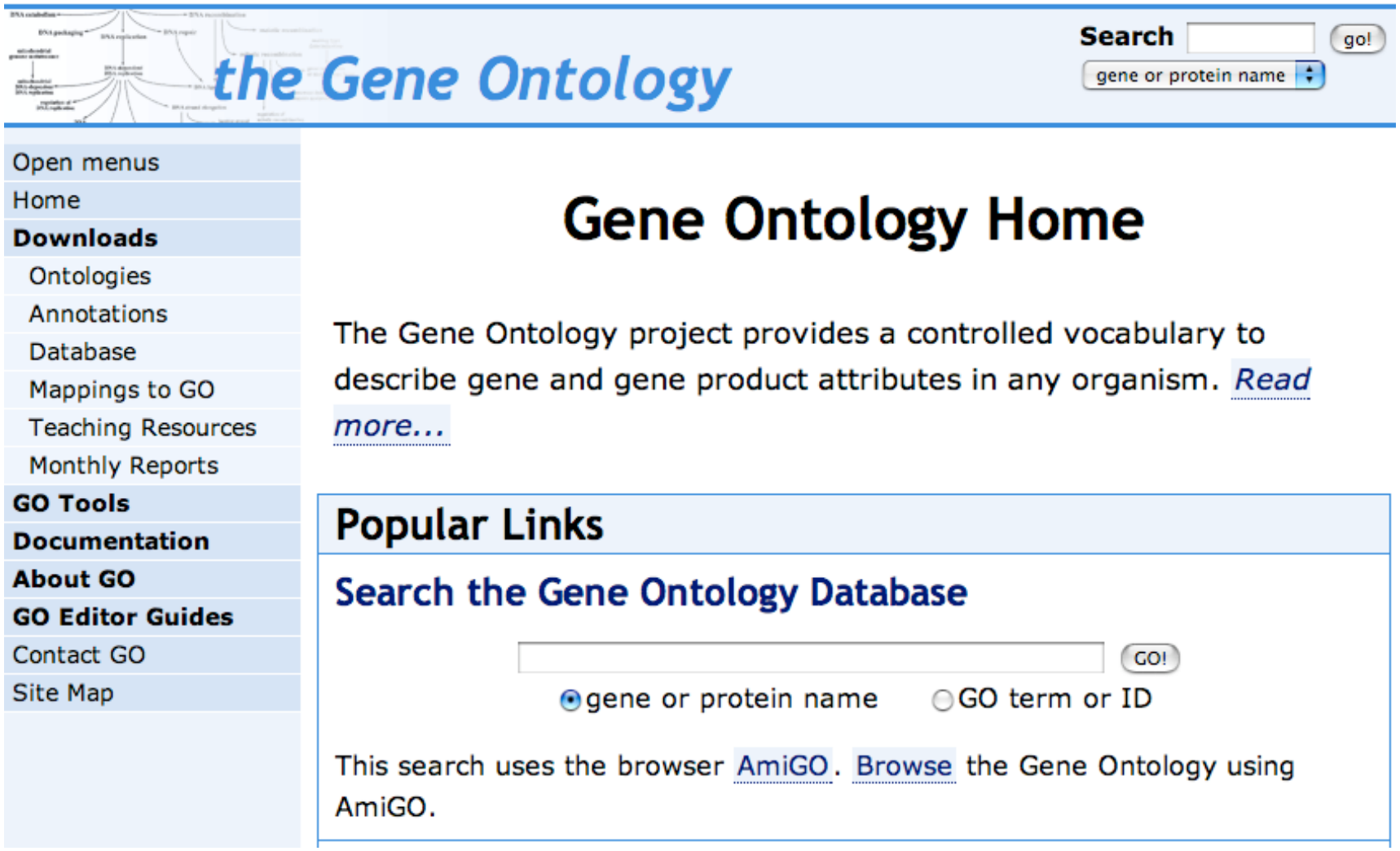
An Example

5 normal sample and 9 myeloma (MM) samples 12558 genes (rows)

probe set	gene	Normal m4	Normal m4	Normal m4	Normal m4	Normal m43	MM m282	MM m331a	MM m332a	MM m333a	MM m334a	MM m353a	MM m408a	MM m423a
31307_at	pre-T/NK c	28.53	32.61	29.56	36.55	33.19	25.1	32.79	34.3	35.44	28.48	29.55	22.28	28.77
31308_at	pre-T/NK c	69.14	53.69	52.78	62.07	58.74	67.88	85.82	83.54	85.91	60.93	62.82	47.17	77.07
31309_r_a	Human bre	16.9	67.7	27.61	46.16	51.46	45.62	35.57	32.62	35.14	96.18	45.94	63.2	38.27
31310_at	glycine rec	67.42	49.55	55.51	59.57	68.42	91.06	91.23	83.66	76.37	71.23	74.95	74.04	100.77
31311_at	Homo sapi	78.73	62.91	60.84	72.98	72.9	79.39	85.52	82.57	69.69	63.72	64.29	62.85	67.58
31312_at	potassium	66.65	59.46	55.47	61.75	69.92	75.28	85.53	97.91	69.92	74.77	71.83	58.17	72.15
31313_at	mannosyl (	115.33	95.51	84.48	94.99	109.04	105.05	118.68	106.76	142.88	103.72	106.19	98.58	104.13
31314_at	bone morp	71.89	36.24	41.86	46.99	45.94	46.67	67.56	66.14	53.95	40.97	47.96	43.63	53.54
31315_at	immunoglo	103.99	88.27	83.81	81.81	254.63	87.12	99.11	109.56	86.37	75.03	74.97	69.02	97.22
31316_at	Human vac	16.79	10.08	9.53	16.48	11.98	12.8	16.7	18.76	11.25	12.09	18.89	10.81	19.49
31317_r_a	Human unp	316.75	269.61	254.92	352.61	342.4	327.12	366.39	346	308.43	279.81	312.4	318.06	334.27
31318_at	Stem cell f	32.68	19.79	27.45	29.56	28.34	26.55	38.04	41.05	31.91	22.76	23.58	28.29	22.61
31319_at	Cluster Inc	252.78	441.07	143.32	400.01	373.4	105.06	105.72	87.02	110.75	161.69	84.88	240.91	210.54
31320_at	Cluster Inc	101.42	89.07	79.51	100.69	120.06	116.74	121.41	134.74	131.36	137.4	114.15	119.89	126.74
31321_at	Cluster Inc	112.27	62.17	62.44	80.17	110.97	53.89	55.04	55.16	63.37	54.35	57.79	48.07	47.89
31322_at	Cluster Inc	44.15	52.5	44.8	46.25	55.96	50.01	53.2	52.24	62.16	49.94	47.24	40.64	50.1
31323_r_a	Glutamate	141.44	177.7	138.58	142.61	167.28	169.49	199.64	185.22	218.79	196.56	150.14	185.24	226.37
31324_at	Cluster Inc	70.87	57.8	61.61	65.93	84.05	106.41	106.73	87.01	112.12	78.47	111	89.08	100.53
31325_at	Cluster Inc	68.63	167.66	69.04	112.84	120.46	126.72	107.04	100	116.83	207.5	125.65	155.19	102.55
31326_at	Cluster Inc	157.67	127.49	123.37	146.18	150.95	159.46	184.08	206.02	182.95	139.01	154.57	143.09	175.27
31327_at	Cluster Inc	35.57	28.17	33.64	32.36	38.76	43.13	40.16	46.8	34.47	33.71	25.74	29.45	37.37
31328_at	solute carr	61.61	48.23	50.76	58	57.58	57.91	69.91	72.34	70.29	54.98	59.74	45.55	63.02
31329_at	Human put	12.23	18.91	15.36	19.99	21.15	15.9	20.76	22.26	16.15	28.86	13.59	16.06	23.85
31330_at	ribosomal	108.87	133.3	89.84	113.02	147.61	169.87	156.81	136.47	153.07	220.54	220.96	332.11	183
31331_at	surfactant	28.21	17.99	23.56	26.37	30.35	28.84	31.54	35.06	22.53	24.45	23	21.37	30.86
31332_at	RIG-like 1	20.77	18.58	19.03	18.29	20.86	23.56	25.11	24.43	19.3	28.72	17.27	23.18	25.35
31333_at	tolloid-like	22.97	52.9	26.95	41.22	48.38	48.85	42.09	40.13	40.73	89.86	46.96	59.74	40.87
31334_at	G protein-c	98.57	100.16	78.76	119.09	118.58	97.42	110.67	104.95	143.47	111.28	102.88	115.9	133.17
31335_at	clone 1900	65.79	54.3	54.79	57.23	60.75	66.98	72.89	86.97	76.34	57.65	59.83	49.21	70.83
31336_at	Cluster Inc	40.97	26.15	32.55	26.26	33.73	36.15	36.03	34.93	24.12	26.26	22.55	23.64	28.11

34 taste genes

- NM_199153
- NM_207011
- NM_020501
- NM_207016
- NM_199154
- NM_020502
- NM_207017
- NM_199155
- NM_207018
- NM_207019
- NM_207020
- NM_053212
- NM_207021
- NM_207022
- NM_020503
- NM_207023
- NM_207024
- NM_001039128
- NM_207025
- NM_207026
- NM_207027
- NM_207028
- NM_207029
- NM_199156
- NM_207030
- NM_199158
- NM_199159
- NM_181276
- NM_001025385
- NM_001001451
- NM_181275
- NM_021562
- NM_001001452
- NM_001001453



http://www.geneontology.org/

cellular component, biological process and molecular function.

A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions.

For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.

Cellular component

A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer). See the documentation on the cellular component ontology for more details.

Biological process

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolism or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Further information can be found in the process ontology documentation.

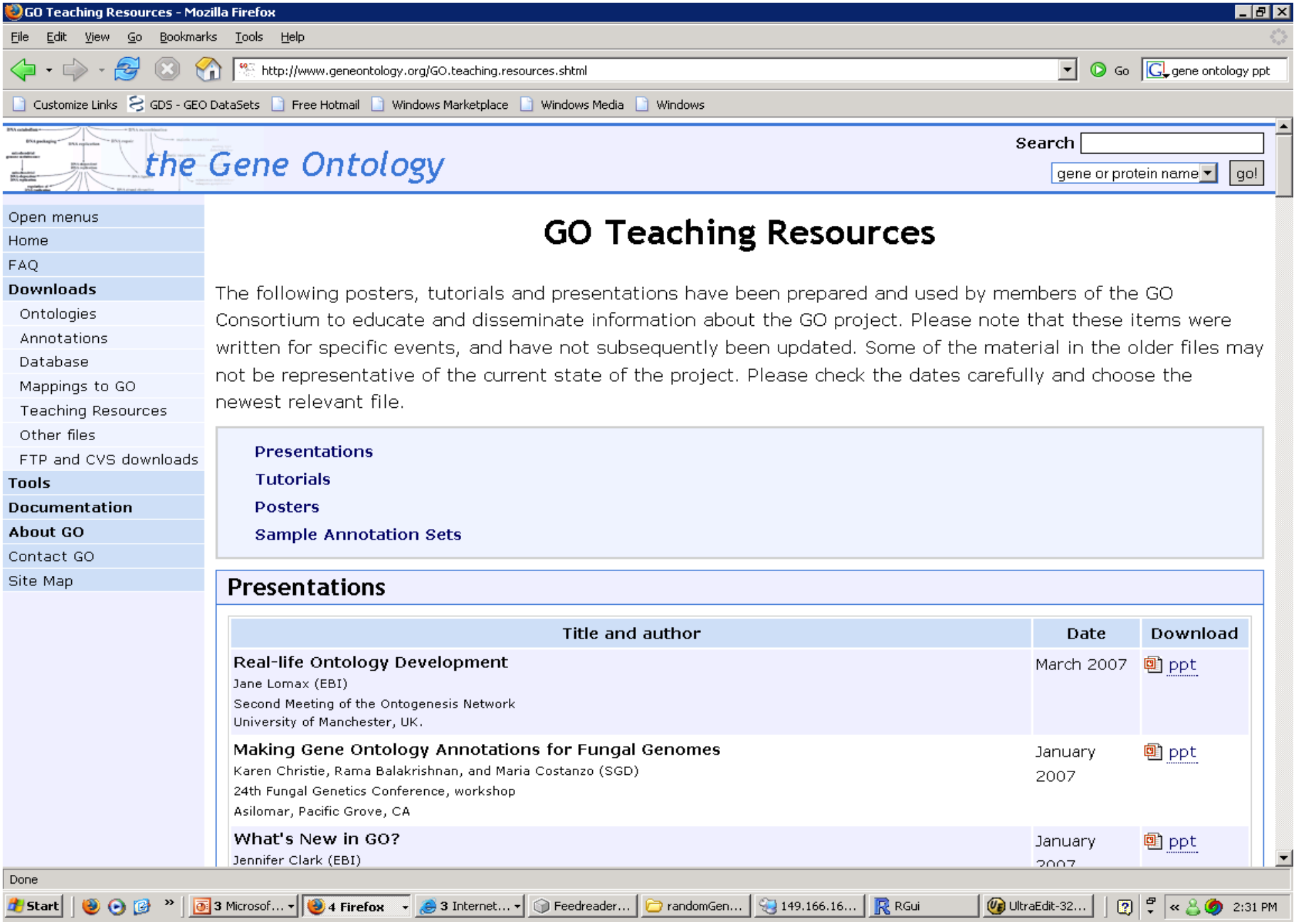
Molecular function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

http://www.godatabase.org/

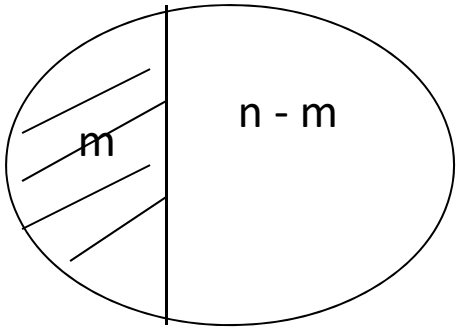
Species, Database	Gene Products Annotated	Annotations	Submission date *MM/DD/YYYY*	Download filtered files
Anaplasma phagocytophilum HZ TIGR	1292	3559 (3559 non-IEA)	1/13/2007	annotations [38.7 kb] README
Arabidopsis thaliana TAIR/TIGR	35026	105826 (83959 non-IEA)	4/28/2007	annotations [2.3 mb] README
Bacillus anthracis Ames TIGR	5289	13393 (13393 non-IEA)	3/19/2007	annotations [148.6 kb] README
Bos taurus GO Annotations @ EBI	22780	84862 (3010 non-IEA)	4/28/2007	annotations [1.1 mb] README
Carboxydothermus hydrogenoformans Z-2901 TIGR	2616	6610 (6610 non-IEA)	4/18/2007	annotations [79.7 kb] README
Caenorhabditis elegans WormBase	14011	93427 (33378 non-IEA)	4/28/2007	annotations [745.9 kb] README
Campylobacter jejuni RM1221 TIGR	1834	4791 (4791 non-IEA)	1/13/2007	annotations [60.0 kb] README
Candida albicans CGD	1280	5425 (5214 non-IEA)	4/11/2007	annotations [68.2 kb] README
Coxiella burnetii RSA 493 TIGR	2038	5294 (5294 non-IEA)	1/13/2007	annotations [58.0 kb] README
Danio rerio ZFIN	12961	70855 (20485 non-IEA)	4/28/2007	annotations [1.1 mb] README
Dehalococcoides ethenogenes 195 TIGR	1584	4068 (4068 non-IEA)	3/19/2007	annotations [47.4 kb] README
Dictyostelium discoideum DictyBase	6581	27438 (15179 non-IEA)	4/22/2007	annotations [330.7 kb] README
Drosophila melanogaster FlyBase	11605	70438 (47870 non-IEA)	4/7/2007	annotations [862.4 kb] README

http://www.geneontology.org/GO.teaching.resources.shtml

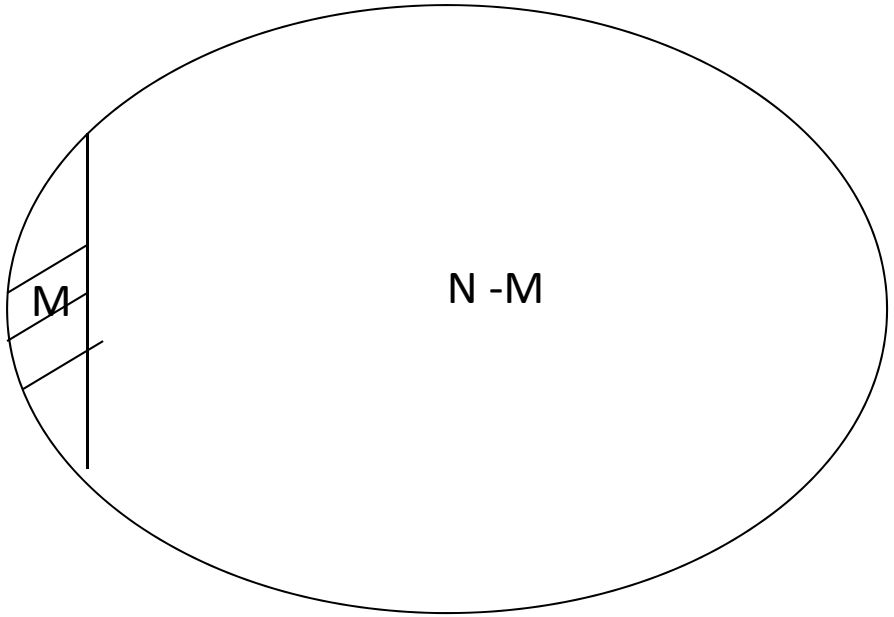


Common function

What do we mean by common function? 90% genes have the function?

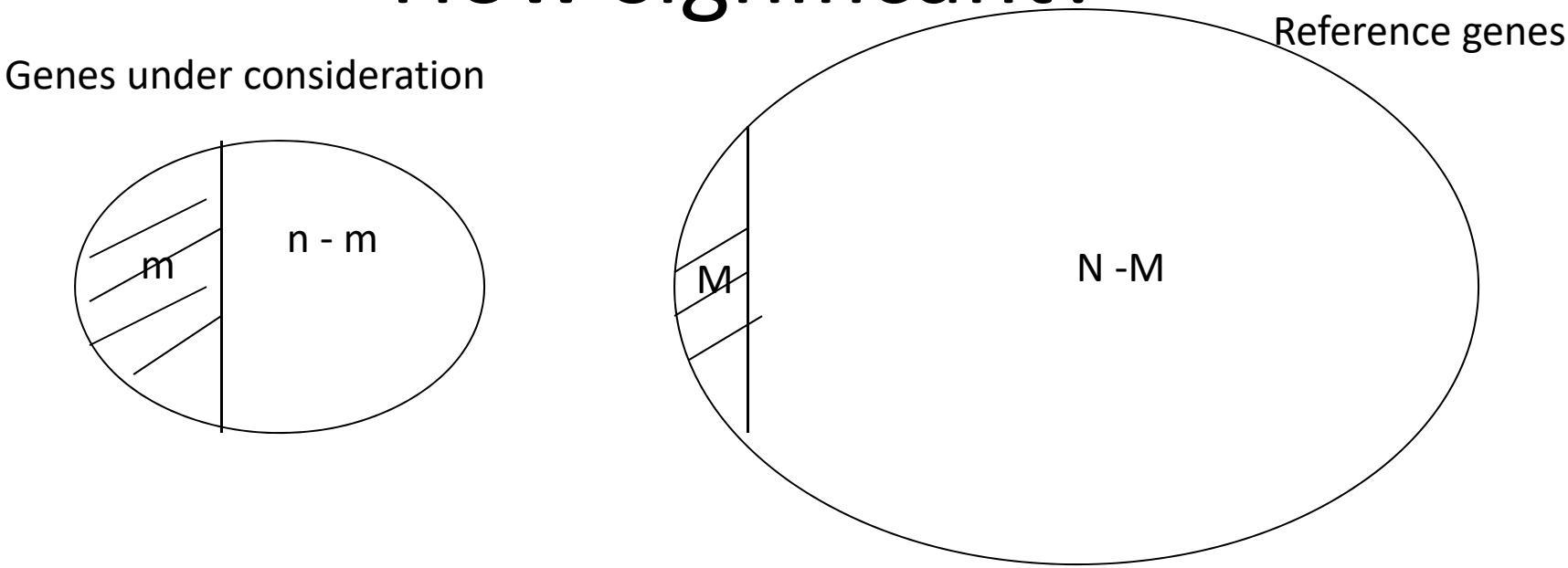


Genes under consideration



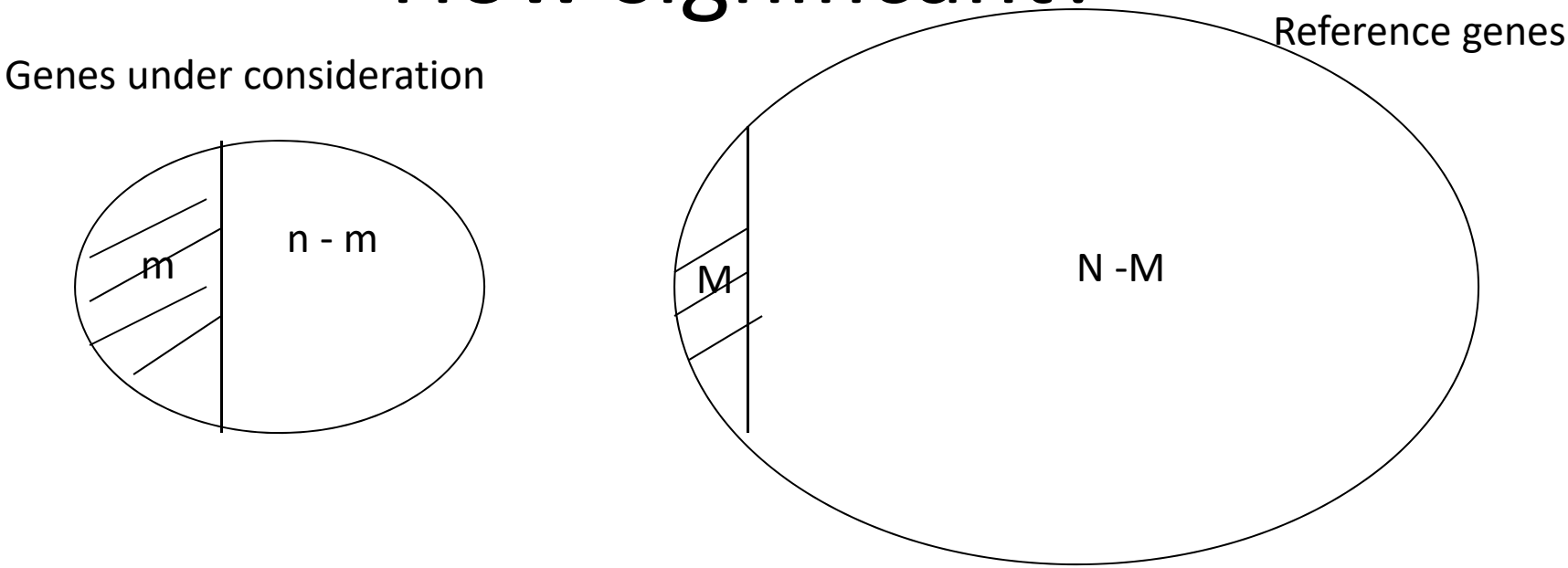
Reference genes

How significant?



n	m	N	M
34	10	25000	2000
34	10	25000	200
34	2	25000	400

How significant?



n	m	N	M	significance
34	10	25000	2000	0.0002304588
34	10	25000	200	9.503509e-14
34	2	25000	400	0.1025492

P-value

In statistical hypothesis testing, the **p-value** is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The fact that p-values are based on this assumption is crucial to their correct interpretation.

Hypergeometric testing

	Annotated with term A	Annotated w/o term A
Reference gene set	m	n - m
Target gene set	M	N-M

$$p - value = \sum_{i=m}^{\min(n,M)} \frac{C_M^i C_{N-M}^{n-i}}{C_N^n}, \qquad C_i^j = \frac{j(j-1)\cdots 1}{i(i-1)\cdots 1}.$$

GO enrichment analysis

http://genemerge.cbcb.umd.edu/

http://david.abcc.ncifcrf.gov/home.jsp

http://www.informatics.jax.org/gotools/MGI_Term_Finder.html

http://go.princeton.edu/cgi-bin/GOTermFinder

http://gostat.wehi.edu.au/cgi-bin/goStat.pl

Multiple comparisons and false discovery rate

multiple comparisons problem occurs when one considers a set, or family, of statistical inferences simultaneously.

Dealing with Multiple Comparison

- **Bonferroni inequality**: To control the **family-wise error rate** for testing m hypotheses at level α, we need to control the FPR for each individual test at α/m
- Then P(false rejection at least one hypothesis) < α

  or P(no false rejection) > 1- α
- This is appropriate for some applications (e.g. testing a new drug versus several existing ones), but is too conservative for our task of gene selection.

GOstats

- if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
- #install Bioconductor if not installed

- BiocManager::install("GOstats")
- #any Bioconductor packages should be install by such a command

- Google “how to use Gostats”, you will have “*How To Use GOstats* and Category to do Hypergeometric testing with unsupported model organisms” by M. Carlson. October 26, 2021 in the pdf format.

GOstats

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GOstats")

## Fake up some data
BiocManager::install("hgu95av2.db")
library("hgu95av2.db")
library("annotate")
prbs <- ls(hgu95av2GO)[1:300]
## Only those with GO ids
hasGO <- lengths(lapply(mget(prbs, hgu95av2GO), names)) != 0
prbs <- prbs[hasGO]
prbs <- getEG(prbs, "hgu95av2")
## remove duplicates, but keep named vector
prbs <- prbs[!duplicated(prbs)]
## do the same for universe
univ <- ls(hgu95av2GO)[1:5000]
hasUnivGO <- lengths(lapply(mget(univ, hgu95av2GO), names)) != 0
univ <- univ[hasUnivGO]
univ <- unique(getEG(univ, "hgu95av2"))
p <- new("GOHyperGParams", geneIds=prbs, universeGeneIds=univ,
ontology="BP", annotation="hgu95av2", conditional=TRUE)
## this part takes time...
if(interactive()){
hyp <- hyperGTest(p)
ps <- probeSetSummary(hyp, 0.05, 10)
}
```

GOstats

```
#The following is for non-model organism.

library("AnnotationForge")
available.dbschemas()


library("org.Hs.eg.db")
#if the above line does not work, install this package in Bioconduction with the
#command BiocManager::install("org.Hs.eg.db")

#the following 3 lines is to make your own GO database
frame = toTable(org.Hs.egGO)
goframeData = data.frame(frame$go_id, frame$Evidence, frame$gene_id)
head(goframeData)

#The following is to convert your own GO database into GSEA accepting database
goFrame=GOFrame(goframeData,organism="Homo sapiens")
goAllFrame=GOAllFrame(goFrame)
library("GSEABase")
gsc <- GeneSetCollection(goAllFrame, setType = GOCollection())
```

GOstats

```
#the following section is for the GSEA GO analysis. It outputs the significant
#GO ID and terms and p-value and gene count. But not the gene IDs as in other
#tools. You need to parse the above goAllFrame and these files to obtain the
#similar output with gene names.
library("GOstats")
universe = Lkeys(org.Hs.egGO)
genes = universe[1:500]
params <- GSEAGOHyperGParams(name="My Custom GSEA based annot Params",
  geneSetCollection=gsc,
  geneIds = genes,
  universeGeneIds = universe,
  ontology = "MF",
  pvalueCutoff = 0.05,
  conditional = FALSE,
  testDirection = "over")

Over <- hyperGTest(params)
head(summary(Over))
```

GOstats

```
#the following section is for the KEGG GO analysis, similar to the aboe.
frame = toTable(org.Hs.egPATH)
keggframeData = data.frame(frame$path_id, frame$gene_id)
head(keggframeData)

keggFrame=KEGGFrame(keggframeData,organism="Homo sapiens")
gsc <- GeneSetCollection(keggFrame, setType = KEGGCollection())
universe = Lkeys(org.Hs.egGO)
genes = universe[1:500]
kparams <- GSEAKEGGHyperGParams(name="My Custom GSEA based annot Params",
  geneSetCollection=gsc,
  geneIds = genes,
  universeGeneIds = universe,
  pvalueCutoff = 0.05,
  testDirection = "over")
kOver <- hyperGTest(params)
head(summary(kOver))

#output the tools and versions used.
toLatex(sessionInfo())
```

references

- https://bioconductor.org/packages/devel/bioc/manuals/GOstats/man/GOstats.pdf

- *How To Use GOstats* and Category to do Hypergeometric testing with unsupported model organisms" by M. Carlson. October 26, 2021 in the pdf format.