Gene List Enrichment Analysis

George Bell, Ph.D.

BaRC Hot Topics

March 16, 2010

Outline

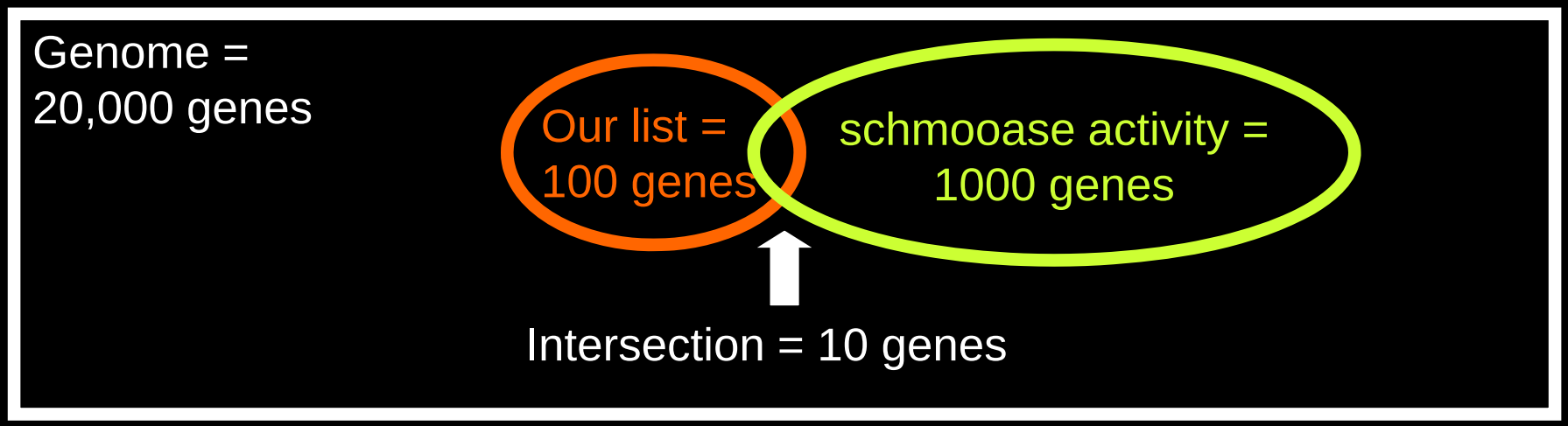
- Why do enrichment analysis?
- Main types
- Selecting or ranking genes
- Annotation sources
- Statistics
- Remaining issues
- Presenting findings
- Recommended tools

Why do enrichment analysis?

- Most array, sequencing, and screens produce
  - A measurement for most or all genes
  - List(s) of “interesting” genes
- Most cellular processes involve sets of genes.
- Can we compare the above two datasets?
- Is the overlap different than expected?
- Does this tell us something about cellular mechanisms?

Why not just link genes to physiology?

- Too many genes to examine in detail.
- Are we biased?
- How do we know that what we're seeing is surprising?

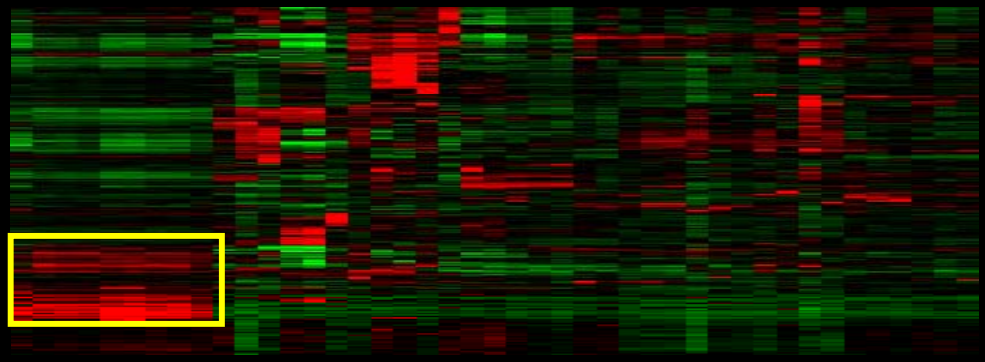


Main types of enrichment analysis

- List-based: inputs are
  - A subset of all genes chosen by some relevant method
  - A list of annotations, each linked to genes
- Rank-based: inputs are
  - A set of all genes ranked by some metric (ratio, fold change, etc.)
  - A list of annotations, each linked to genes
- List-based with relationships: inputs are
  - A subset of all genes
  - A list of annotations, each linked to genes, organized in some relationship (e.g., a hierarchy)

Getting your list

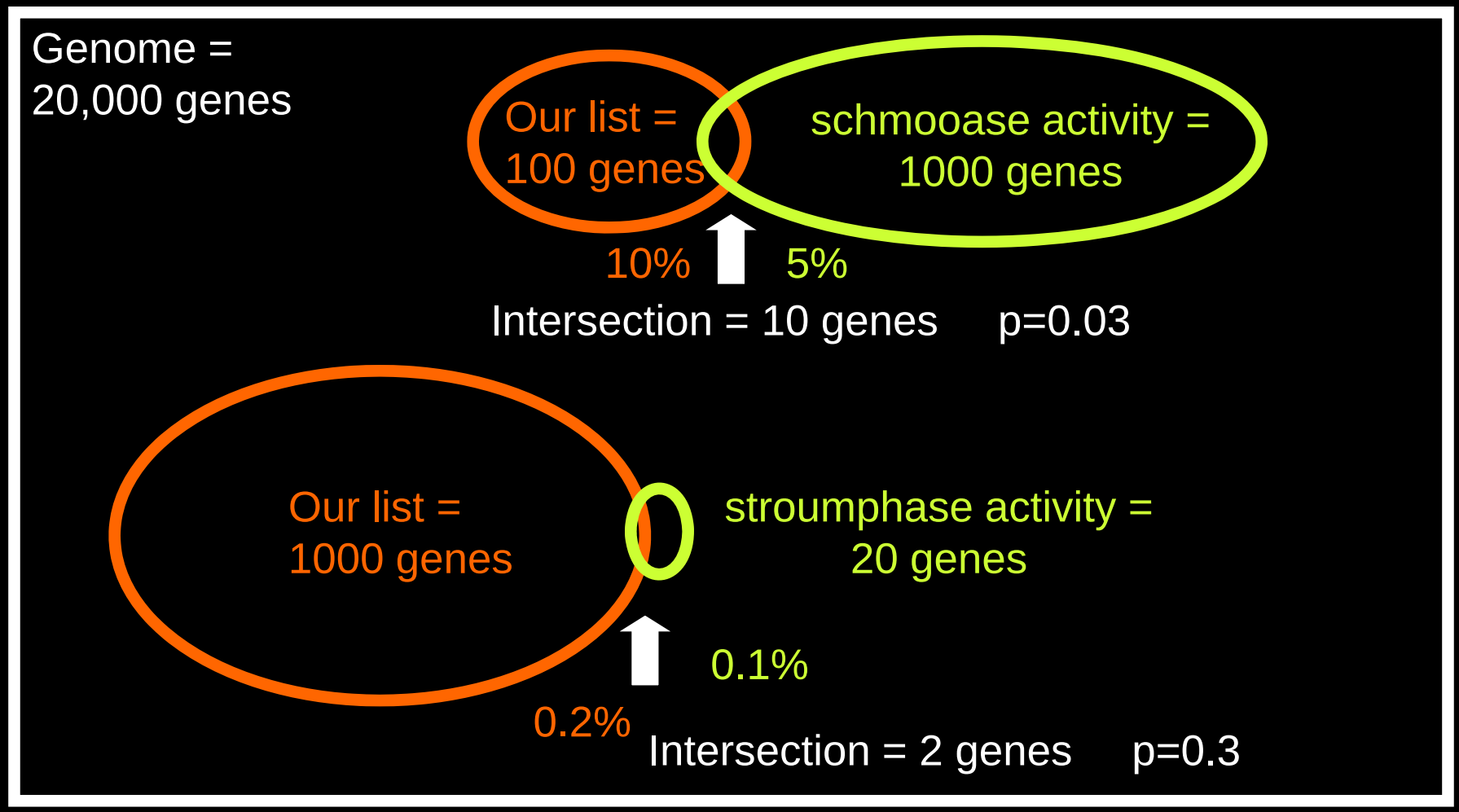
- Goal: Identify a list of genes (or probes) that appear to be working together in some way.
- What identifiers to use?
- Most common method: Get a list of differentially expressed genes
  - P-value and/or fold change?
  - Threshold?
- Alternatives:
  - Define a cluster
  - Sort data and/or apply a model to rank genes
- Recommendations:
  - Try lists of varying length
  - Try to maximize signal / noise (What produces the smallest p-values for enrichment?)



Annotation sources

- Gene Ontology (most popular)
  - biological process, molecular function, cellular component
  - Terms may have >1 “parent” (more general term)
  - GO Slim: includes only general categories
- KEGG; REACTOME pathways
- Genes sharing a motif of regulated by the same protein/miRNA
- Genes found on the same chromosome
- Also … see Broad’s Molecular Signatures Database (MSigDB)
- [any grouping that is biologically sensible]

Statistics to test for enrichment



Tests for enrichment

- Fisher's exact
- Hypergeometric
- Binomial
- Chi-squared
- Z
- Kolmogorov-Smirnov
- Permutation
- …..

Statistics to test for enrichment

- What is the chance of observing enrichment at least this extreme due to chance?
- Different tests produce very different ranges of p-values
- All look for over-enrichment; some look for under-enrichment
- Recommendation: Use p-values as a tool to rank genes but don't take them literally
- Most methods correct for multiple testing (e.g., with FDR), which is necessary

Other statistical issues

- Goal: Identifying theme(s) of maximal biological significance
  - but this is not perfectly correlated with statistical significance
- What is your background gene set?
  - All genes that could appear in your list
- What about sparse annotation groups?
- Some annotation terms may be subsets of other terms.

Practicalities

- Choose a tool that
  - Includes your species
  - Includes your gene / probe identifiers
  - Has up-to-date annotation
  - Lets you define your background (if possible)
- Get recommendations from the usual sources.
- Try at least a few tools.
- Try lists of varying length.

Presenting results

- Generally ignore enriched categories which
  - Contain very few genes
  - Show high overlap with other categories
- When in doubt, select more general category.
- Simplify complex results.
- Graphical or text summary?
- Plan to share your gene lists when you publish.

Enrichment tools

- See http://www.geneontology.org/GO.tools.shtml

Some recommended tools

- DAVID
- GSEA
- BIOBASE (Whitehead has license)
- BiNGO (uses Cytoscape)
- GoMiner: http://discover.nci.nih.gov/gominer
- GOstat: http://gostat.wehi.edu.au

DAVID

- Database for Annotation, Visualization and Integrated Discovery (NIAID)
- List-based
- http://david.abcc.ncifcrf.gov/
- Lots of identifiers; lots of species
- Allows background definition
- Statistic is a modified Fisher exact test

Welcome to the new, temporary home of DAVID2008. We have extended the retirement of this version until 3/31/2010. Please complete any analysis using this version by this date as it will no longer be available. Thanks for using and supporting DAVID

Functional Annotation Chart

Help and Manual

Current Gene List: Testes enriched
Current Background: HOMO SAPIENS
74 DAVID IDs

Options

Rerun Using Options | Create Sublist

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	FDR
☐	GOTERM_MF_ALL	catalytic activity	RT		58	78.4	2.6E-14	4.6E-11
☐	GOTERM_MF_ALL	transmembrane transporter activity	RT		18	24.3	9.2E-7	1.6E-3
☐	SP_PIR_KEYWORDS	oxidoreductase	RT		12	16.2	1.2E-6	1.9E-3
☐	GOTERM_MF_ALL	transporter activity	RT		21	28.4	1.4E-6	2.5E-3
☐	GOTERM_MF_ALL	cation transmembrane transporter activity	RT		13	17.6	3.3E-6	9.4E-3
☐	GOTERM_MF_ALL	ion transmembrane transporter activity	RT		15	20.3	3.6E-6	1.0E-2
☐	GOTERM_MF_ALL	substrate-specific transmembrane transporter activity	RT		16	21.6	3.9E-6	1.1E-2
☐	GOTERM_BP_ALL	cellular carbohydrate catabolic process	RT		7	9.5	6.9E-6	1.3E-2
☐	GOTERM_BP_ALL	alcohol metabolic process	RT		10	13.5	8.0E-6	1.5E-2
☐	GOTERM_BP_ALL	carbohydrate catabolic process	RT		7	9.5	9.8E-6	1.9E-2
☐	GOTERM_CC_ALL	flagellum	RT		5	6.8	1.0E-5	1.6E-2
☐	SP_PIR_KEYWORDS	glycolysis	RT		5	6.8	1.3E-5	2.0E-2
☐	GOTERM_BP_ALL	glucose catabolic process	RT		6	8.1	1.3E-5	2.5E-2
☐	GOTERM_BP_ALL	carbohydrate metabolic process	RT		12	16.2	1.5E-5	3.0E-2

GSEA

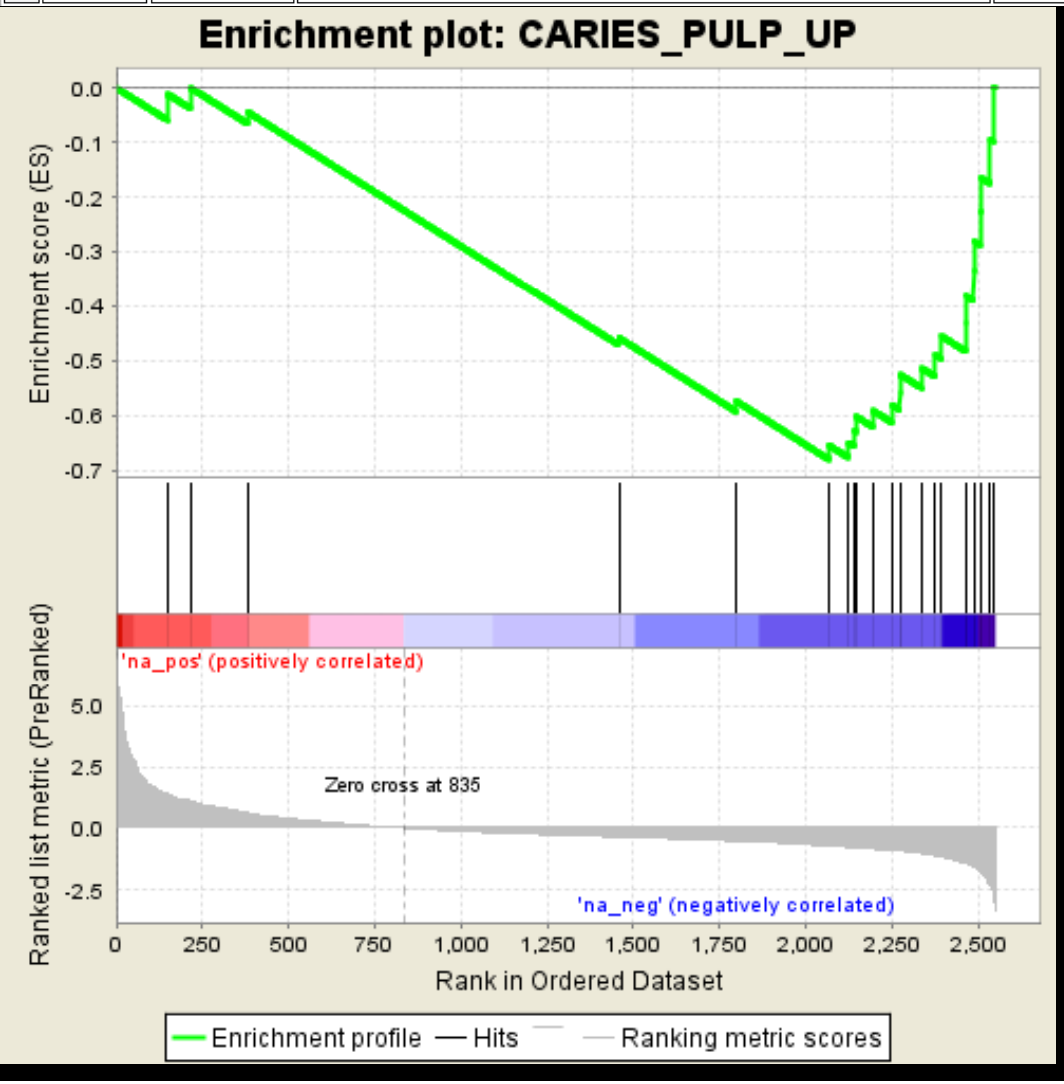
- Gene Set Enrichment Analysis
- Rank-based
- http://www.broadinstitute.org/gsea/
- As a Java Web Start or desktop application
- Linked to MSigDB (annotated gene lists)
- Also permits custom annotation

Fig 1: Enrichment plot: CARIES_PULP_UP
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [plain text format]

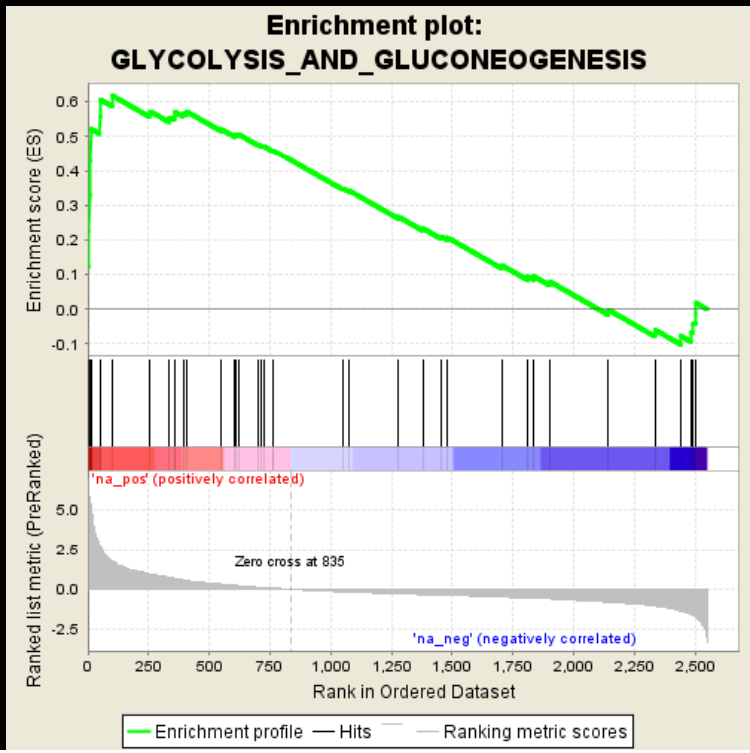
	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	MTHFD2	MTHFD2 Entrez, Source	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	149	1.379	-0.0116	No
2	KYNU	KYNU Entrez, Source	kynureninase (L-kynurenine hydrolase)	215	1.092	0.0001	No
3	SOD2	SOD2 Entrez, Source	superoxide dismutase 2, mitochondrial	382	0.609	-0.0447	No

Input: pre-ranked gene list



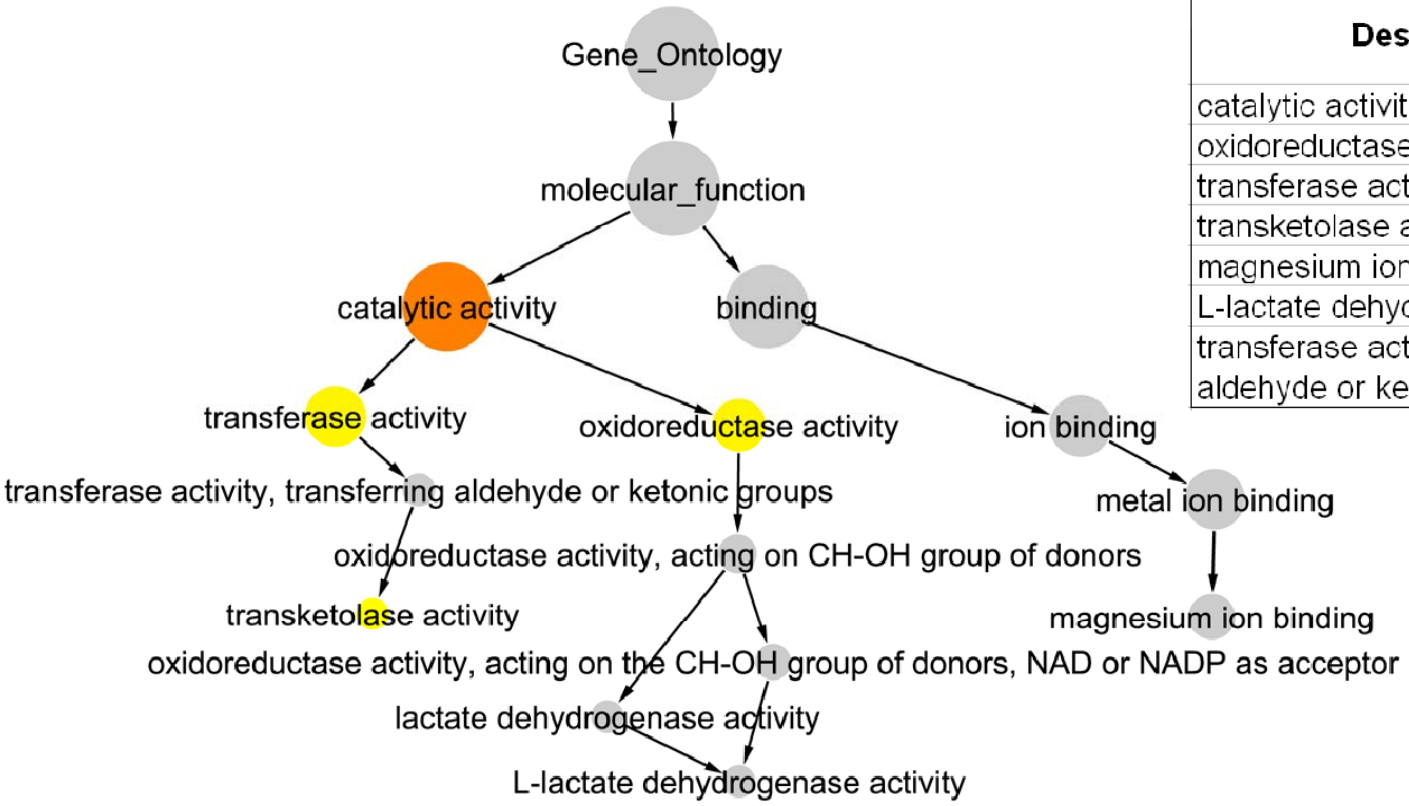
Enrichment at bottom of list

Enrichment at top of list



BiNGO

- BiNGO: A Biological Network Gene Ontology tool
- http://www.psb.ugent.be/cbd/papers/BiNGO/
- Works with Cytoscape network visualization tool
- Also permits custom annotation

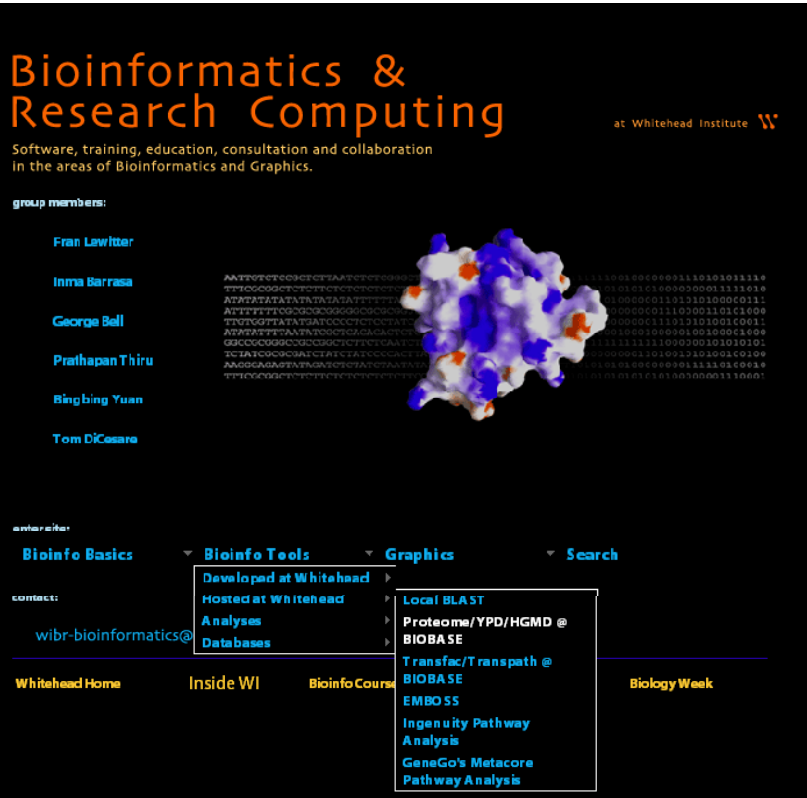


Description	corrected p-value	Number of genes in list
catalytic activity	2.0E-13	48
oxidoreductase activity	1.0E-03	12
transferase activity	1.2E-03	18
transketolase activity	2.4E-03	2
magnesium ion binding	4.5E-03	8
L-lactate dehydrogenase activity	4.5E-03	2
transferase activity, transferring aldehyde or ketonic groups	4.5E-03	2

Shows relationship between annotation categories

BIOBASE

- BIOBASE Knowledge Library
- Use Internet Explorer
- Go to “Gene Set Analysis”



Legend:

BP = GO biological process	GF = gene family	PP = plant phenotype
CC = GO cellular component	IN = protein interaction	PT = canonical pathway
DI = disease	MD = protein modification	PX = expression in plants
DG = pharmaceutical	MF = GO molecular function	RE = regulators of fungal genes
DO = protein domain	PH = mouse phenotype	SP = species & chromosomes
EX = expression in mammals	PM = yeast or worm phenotype	WX = expression in worms

P-value	Term	Protein count	Expected Protein count
5.03e-45	SP Human	58	10.7
8.62e-21	SP Mammal	58	28.3
2.3e-12	EX testis	27	7.73
3.09e-11	MF catalytic activity	45	21.3
1.03e-10	SP Human chromosome 17	6	0.594

References

- Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. (PMID: 19033363) *Review*
- Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. (PMID: 19131956) *DAVID*
- Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. (PMID: 16199517) *GSEA*

Statistics – supplementary info

Fisher's test by hand in R

- counts = (matrix(data = c(3, 297, 40, 19960), nrow = 2))
- counts
- fisher.test(counts)
- # is better than
- chisq.test(counts)

	Gene list	Genome
In anno group	3	40
Not in anno group	297	19960

```
        Fisher's Exact Test for Count Data

data:  counts
p-value = 0.02552
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9918169 15.9604612
sample estimates:
odds ratio
  5.039206
```

(3/297) / (40/19960)

Binomial test by hand in R

- binom.test(3, 300, p=40/20000)

	Gene list	Genome
In anno group	3	40
Not in anno group	297	19960

```
        Exact binomial test

data:  3 and 300
number of successes = 3, number of trials = 300,
  p-value = 0.02298
alternative hypothesis: true probability of success
  is not equal to 0.002
95 percent confidence interval:
 0.002067007 0.028944511
sample estimates:
probability of success
                  0.01
```

40 / 20,000

Hypergeometric test by hand in R

- min(1 - cumsum(dhyper(0:(3-1), 40, 19960, 300)))
- 0.02193491

	Gene list	Genome
In anno group	3	40
Not in anno group	297	19960

- Equation above tests only for over-enrichment