Functional analysis of differentially expressed genes

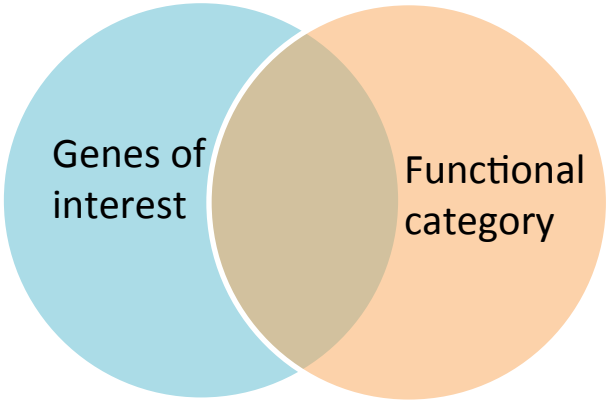
Yulia Mostovoy

CGRL workshop

4/30/12

Analysis of differentially expressed genes

- What do a given subset of genes have in common?
  - Functions?
  - Pathways?
- Approach: look for categories that have an overabundance of genes from your list



Gene Ontology (GO)

- Describes properties of gene products in a structured, standardized way
  - Biological process
  - Molecular function
  - Cellular component
- Hierarchical: broader terms lead to more specific terms
- Can be applied to any species
- www.geneontology.org

Other functional classifications

- Biochemical/metabolic pathways
- Transcription factor regulation
- Protein complexes
- Co-expression modules
- Single category, e.g. TATA-box-containing genes
- Many other possibilities!

Fisher's exact test

- Two-by-two contingency table:

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**k**	m-k	**m**
Not differentially expressed genes	n-k	N-m-n+k	N-m
Sums	**n**	N-n	**N**

k: # of DE genes are in category

m: # of total DE genes

n: # of total genes in category

N: # of genes with valid data in your study

Fisher's exact test

- Two-by-two contingency table:

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**k**	m-k	**m**
Not differentially expressed genes	n-k	N-m-n+k	N-m
Sums	**n**	N-n	**N**

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**14**	197	**211**
Not differentially expressed genes	25	5916	5941
Sums	**39**	6113	**6152**

Fisher's exact test

- Two-by-two contingency table:

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**14**	197	**211**
Not differentially expressed genes	25	5916	5941
Sums	**39**	6113	**6152**

- Null hypothesis: differentially expressed genes are as likely to belong to the category as any genes in the genome

Fisher's exact test

- Two-by-two contingency table:

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**14**	197	**211**
Not differentially expressed genes	25	5916	5941
Sums	**39**	6113	**6152**

```
> table <- matrix(c(14,25,197,5916), nrow=2, ncol=2)
> fisher.result <- fisher.test(table, alternative='g')
> p.value = fisher.result$p.value
> p.value
[1] 1.43551e-11
```

Multiple testing correction

- When testing multiple categories, don't forget multiple testing correction.

```
> pvals.example <- c(0.05, 0.00001, 0.4, 0.2, 0.002)
> pvals.adjusted <- p.adjust(pvals.example,
method='bonferroni')

> pvals.adjusted
[1] 0.25000 0.00005 1.00000 1.00000 0.01000
```

Analysis with multiple categories

functional_enrichment.py tests enrichment of your DE genes across any custom categories

- Reads three files:
  - List of DE genes
  - List of all genes with valid data in your study
  - File with category information
    - Category name [tab] list of genes (tab-separated)
    - Optionally: Category name | description [tab] list of genes

Analysis with multiple categories

- Sample run:
  - DE genes = upregulated_heat_shock.txt
  - All valid genes = all_valid_genes.txt
  - Group file = TF_target_file_S_cerevisiae.txt

```
$ python functional_enrichment.py upregulated_heat_shock.txt all_valid_genes.txt
TF_target_file_S_cerevisiae.txt > output.txt
```