

# GOSSIP: Biological Profiling of Gene Groups utilizing Gene Ontology A Statistical Framework and Software

## The Problem

Experiments produce groups of interesting genes  
Functional interpretation of the gene groups required

## Aim

Automatize this complex and laborious task

## Gene Ontology (GO)



About 19000 terms describing  
- Molecular functions  
- Biological processes  
- Cellular components

Hierarchical organised (DAG)  
Annotations available for most genes  
See <http://www.geneontology.org>

## The Algorithm

Tests for all terms in the Gene Ontology whether it is enriched in a **test group** when compared to a **reference group** using Fisher's exact test.

Major challenge: **Multiple testing.**

Analytical Expression for mean numbers of false discoveries at a certain single test p-value  $\alpha$ :

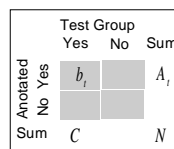
$$\langle NFD(\alpha) \rangle = \sum_{t \in T} \sum_j p_j(j, A_t, C, N) \leq \alpha$$

$$h(j, A_t, C, N)$$

$$\text{with } p_j(j, A_t, C, N) = \sum_{k=j}^{\min(A_t, C)} h(k, A_t, C, N)$$

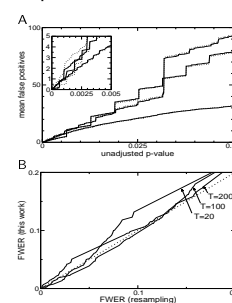
Analytical expression for FDR:

$$FDR(\alpha) = \langle NFD(\alpha) \rangle / FP(\alpha)$$



Also: Heuristic approach to calculate the **FWER** more exactly.

Comparison with Simulations



## GOSSIP Software

- Implemented in C++
  - Available for Windows and Linux
  - Reads Affymetrix annotations
  - One Profile takes 1-2 sec
- Free download:  
<http://www.microdiscovery.de/>

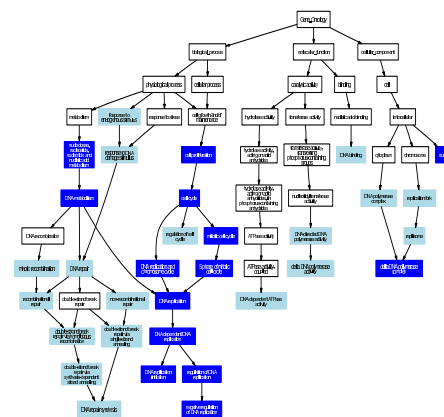
## Example

Whitfield et al. MBC 2002

- 14480 significantly expressed (Reference)
- 85 genes co-regulated (Test Group)
- GO annotations

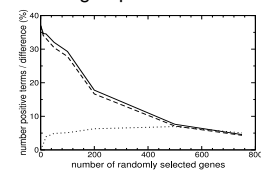
Functions of genes regulated in G1/S mytotic cell cycle

| Number | ECM  | Name   | p-value | FDR/GOSSIP | EWER/GOSSIP |
|--------|------|--|---------|------------|-------------|
| 1      | 6200 | DNA replication  | 1.2e-12 | 1.3e-08    | 1.3e-08     |
| 2      | 6011 | DNA dependent DNA replication                          | 2.2e-11 | 1.3e-08    | 1.4e-08     |
| 3      | 6012 | 3 phase of mitotic cell cycle                          | 7.7e-11 | 1.3e-08    | 1.4e-08     |
| 4      | 67   | DNA replication and chromosome cycle                   | 7.7e-10 | 1.2e-07    | 1.3e-07     |
| 5      | 613  | mitotic cell cycle                                     | 1.6e-9  | 1.2e-07    | 1.3e-07     |
| 6      | 5916 | DNA replication  | 1.4e-07 | 1.3e-07    | 1.3e-07     |
| 7      | 7040 | cell cycle, nucleoside, nucleotide and nucleic acid    | 8.4e-07 | 0.0006     | 0.0001      |
| 8      | 614  | transcription  | 7.2e-07 | 1.3e-06    | 0.00013     |
| 9      | 6142 | positive regulation of DNA replication                 | 6.4e-07 | 1.3e-06    | 0.00019     |
| 10     | 5633 | nucleus  | 1.1e-06 | 1.3e-06    | 0.00022     |
| 11     | 6270 | regulation of DNA replication                          | 1.7e-06 | 0.0001     | 0.00027     |
| 12     | 6271 | DNA replication initiation                             | 3.2e-6  | 0.0004     | 0.00043     |
| 13     | 5656 | alpha DNA polymerase complex                           | 3.9e-06 | 0.0005     | 0.0013      |
| 14     | 5917 | cell cycle   | 1.2e-05 | 0.0007     | 0.0008      |
| 15     | 4500 | double strand break repair via single-strand annealing | 2.2e-5  | 0.0016     | 0.0068      |
| 16     | 7311 | DNA repair synthesis                                   | 2.2e-5  | 0.0016     | 0.0068      |
| 17     | 4500 | strand annealing                                       | 2.2e-5  | 0.0016     | 0.0068      |
| 18     | 4501 | DNA polymerase complex                                 | 6.2e-06 | 0.0016     | 0.0072      |
| 19     | 5918 | DNA replication  | 1.2e-05 | 0.0024     | 0.01        |



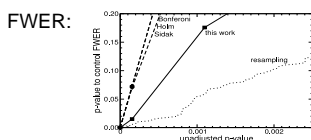
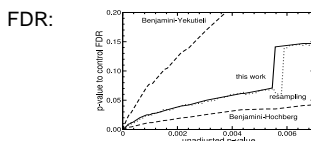
## Robustness

Functional profiles are insensitive to adding randomly selected genes to the test group.



solid: number of significant terms  
dashed: number of preserved terms  
dotted: percent new terms

## Validation of Method



## Applications

TFGOSSIP: Predicting functional targets of Transcription Factors (Kielbasa et al., 2004, Blüthgen et al., 2004, submitted)

Profiles of gene lists obtained from microarray studies:

- Groups of direct ERK targets (Jürchott et al., 2004, MS under prep.)
- Tumor invasion front (Brand et al., 2004, MS under prep.)
- LPS-stimulated Monocytes (Blüthgen et al, 2004, submitted)