# Gene Set Enrichment Analysis

Martin Morgan
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

28 April 2009

# Motivation

Many analyses:

- Exploratory, even in designed experiments: which of 1000's of probes are differentially expressed?

But often...

- *A priori* understanding of relevant biological processes
- Interested in signal from collection of probes (e.g., genes in a pathway)

Original idea applied to expresion data

- Mootha et al. (2003, Nat Genet 34, 267-273) – permutation-based GSEA.

# Overall approach

1. Identify *a priori* biologically interesting sets for analysis.
2. Pre-process and quality assess as usual.
3. Non-specific filtering – remove probes that cannot possibly be interesting.
4. Identify 'interesting' probes based on differential expression.
5. Ask whether genes corresponding to interesting probes are over-represented in each category.

# 1. *A priori* sets

- Biologically motivated.
- Combining 'signal' from several probe sets.
- Examples: KEGG or Gene Ontology (GO) pathways, chromosome bands, ...
- This lab: GO pathways.

# 2. Pre-processing and sample selection

- Use entire data set for background correction, normalization, probe set summary.

```
> library("ALL")
> data("ALL")
```

...(see `HyperG_Lecture.R` for details)

```
> dim(bcrneg)
```

```
Features  Samples
   12625       79
```

# 3. Non-specific filtering: invariant and un-annotated genes

- Exclude genes that cannot be interesting
- *Must not* use criteria to be used in analysis, e.g., *must not* filter on expression in biological pathway of interest.
- Criteria: exclude genes with limited variation across *all* samples, or that are un-annotated.

```
> library("genefilter")
> bcrneg_filt = nsFilter(bcrneg, var.cutoff = 0.5,
+     require.GOBP = TRUE)$eset
> dim(bcrneg_filt)
```

```
Features  Samples
    3751       79
```

# 4. Identify 'interesting' probes

- Many statistics possible; idea is to calculate a statistic that meaningfully contrasts expression levels between groups.
- We'll use a simple $t$-test, with $t_k$ being the statistic associated with the $k$th probe set.
- Discretize (!) the statistic. Two types of genes: 'selected' or 'not selected'.

```
> rtt <- rowttests(bcrneg_filt, "mol.biol")
> rttPrb <- rtt$p.value
> names(rttPrb) <- featureNames(bcrneg_filt)
> tThresh <- rttPrb < 0.05
```

# 5. Are interesting features over-represented? I

- 'Universe' divided into selected, not selected

```
> ids <- featureNames(bcrneg_filt)
> map <- hgu95av2ENTREZID
> universe <- unlist(mget(ids, map))
> selected <- unlist(mget(ids[tThresh],
+     map))
```

- Two possible categories: in GO, not in go. E.g., GO term GO:0006955

```
> library(GO.db)
> GOTERM[["GO:0006468"]]
```

```
GOID: GO:0006468
Term: protein amino acid phosphorylation
Ontology: BP
Definition: The process of introducing a
    phosphate group on to a protein.
```

# 5. Are interesting features over-represented? II

- E.g., for GO term `GO:0006468`...

| | Selected | Not selected |
|---|---|---|
| In GO | 37 | 610 |
| Not in GO | 132 | 2972 |

- Test (e.g., one-tailed): are selected genes more often in the GO category than expected by chance? *Hypergeometric* or one-tailed Fisher exact test

# The test: formulate and perform

```
> library(Category)
> library(GOstats)
> params = new("GOHyperGParams", geneIds = selected,
+     universeGeneIds = universe, annotation = annotation(h
+     ontology = "BP", pvalueCutoff = 0.001,
+     conditional = FALSE, testDirection = "over")
> (overRepresented = hyperGTest(params))

Gene to GO BP  test for over-representation
2682 GO BP ids tested (11 have p < 0.001)
Selected gene set size: 647
    Gene universe size: 3751
    Annotation package: hgu95av2
```

# The test: interpretting

```
> head(summary(overRepresented), n = 3)
      GOBPID  Pvalue OddsRatio ExpCount Count
1 GO:0007154 6.5e-07       1.6      189   241
2 GO:0007165 8.4e-07       1.6      177   228
3 GO:0010646 9.5e-06       2.0       42    68
  Size                                  Term
1 1094                 cell communication
2 1027               signal transduction
3  242 regulation of cell communication

> fl <- tempfile()
> htmlReport(overRepresented, file = fl)
> browseURL(fl)
```

# Hazards and issues

- What is the 'universe' of genes? Answer: all those passing non-specific filtering.
- GO categories are hierarchical, so not independent.
  - p-values misleading.
  - *Conditional* tests often appropriate.

# Conditional hypergeometric tests

- GO is a hierarchy, parent and child nodes.
- Naive application of hypergeometric reuses information from children to evaluate significance of parents.
- Philosophy: more general statements require evidence beyond that provided by children.
- Solution: remove genes significant in children before testing parents.