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. Compute a test statistic, e.g., t-statisitic, for each probe.
5. Calculate an appropriate summary, call it z_k, of the test statistic in each set.
6. Compare the distribution of z_k across sets; by the *central limit theorem*, the distribution of z_k is approximately Normal.

1. *A priori* sets

- Biologically motivated.
- Combining 'signal' from several probe sets.
- Examples: KEGG or Gene Ontology pathways, chromosome bands, ...
- Here we'll use KEGG pathways.
- We'll also restrict attention to pathways represented by 10 or more probes.

2. Pre-processing

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

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

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

```
> dim(bcrneg)
```

```
Features  Samples
   12625       79
```

3. Non-specific filtering: invariant 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.
- Criterion: exclude genes with limited variation across *all* samples.

```
> library("genefilter")
> bcrneg_filt1 = nsFilter(bcrneg, var.cutoff = 0.5)$eset
> dim(bcrneg_filt1)
```

```
Features  Samples
    4487       79
```

3. Non-specific filtering: KEGG I

- Criterion: remove probes with no KEGG annotations, or participating in pathways with fewer than 10 probes represented.
- How? Create a *GeneSetCollection* from the expression set, identify relevant sets, then filter the expression set.

```
> library(GSEABase)
> gsc <- GeneSetCollection(bcrneg_filt1,
+     setType = KEGGCollection())
```

3. Non-specific filtering: KEGG II

```
> gsc

GeneSetCollection
  names: 00010, 00020, ..., 05340 (197 total)
  unique identifiers: 37707_i_at, 32747_at, ..., 33595_r_at
  types in collection:
    geneIdType: AnnotationIdentifier (1 total)
    collectionType: KEGGCollection (1 total)

> gsc[[2]]

setName: 00020
geneIds: 40881_at, 40077_at, ..., 40893_at (total: 21)
geneIdType: Annotation (hgu95av2)
collectionType: KEGG
  ids: 00020 (1 total)
details: use 'details(object)'
```

3. Non-specific filtering: KEGG III

```
> ok <- sapply(geneIds(gsc), length) > 10
> gsc <- gsc[ok]
> length(gsc)
```

```
[1] 117
```

```
> uids <- unique(unlist(geneIds(gsc)))
> bcrneg_filt2 <- bcrneg_filt1[uids, ]
> dim(bcrneg_filt2)
```

```
Features  Samples
    1539       79
```

4. Compute a test statistic

- Many statistics possible; idea is to calculate a statistic that meaningfully contrasts expression levels between groups.
- Statistic chosen should be scale- and sample-size independent.
- We'll use a simple t-test, with t_k being the statistic associated with the kth probe set.

```
> rtt <- rowttests(bcrneg_filt2, "mol.biol")
> rttStat <- rtt$statistic
> names(rttStat) <- featureNames(bcrneg_filt2)
> head(rttStat)
```

```
37707_i_at   32747_at   40685_at   33899_at
     -0.50       3.90      -1.23      -0.65
  40409_at   32336_at
      0.22      -1.33
```

5. Calculate an average for each set I

- t_k follows a t-distribution.
- Sum of *independent* t-statistics is approximately Normal.
- Sum standardized by the square root of the number of genes $|K|$ in a set K is approximately Normal with mean 0 and variance 1.

$$z_K = \frac{1}{\sqrt{|K|}} \sum_{k \in K} t_k$$

- Important that z_K is independent of the number of genes in the set.

5. Calculate an average for each set II

- Write a function to calculate z_K from a list of gene ids
- Apply that function to all gene ids in our gene set collection

```
> zCalc <- function(ids, tStat) {
+     sum(tStat[ids])/sqrt(length(ids))
+ }
> z <- sapply(geneIds(gsc), zCalc, tStat = rttStat)
> names(z) <- names(gsc)
> head(z)

00010 00020 00030 00051 00052 00071
-0.68 -1.94 -0.93 -0.25 -0.90 -0.93
```

6. Compare to Normal distribution

- We expect our z_K to have a Normal distribution. How to assess?
- *Quantile-quantile* plot: close agreement if points in plot lie on a diagonal.

```
> qqnorm(z)
> qqline(z)
```

- Select a distinct outlier!

```
> z[which.min(z)]
```

```
03010
 -8.3
```

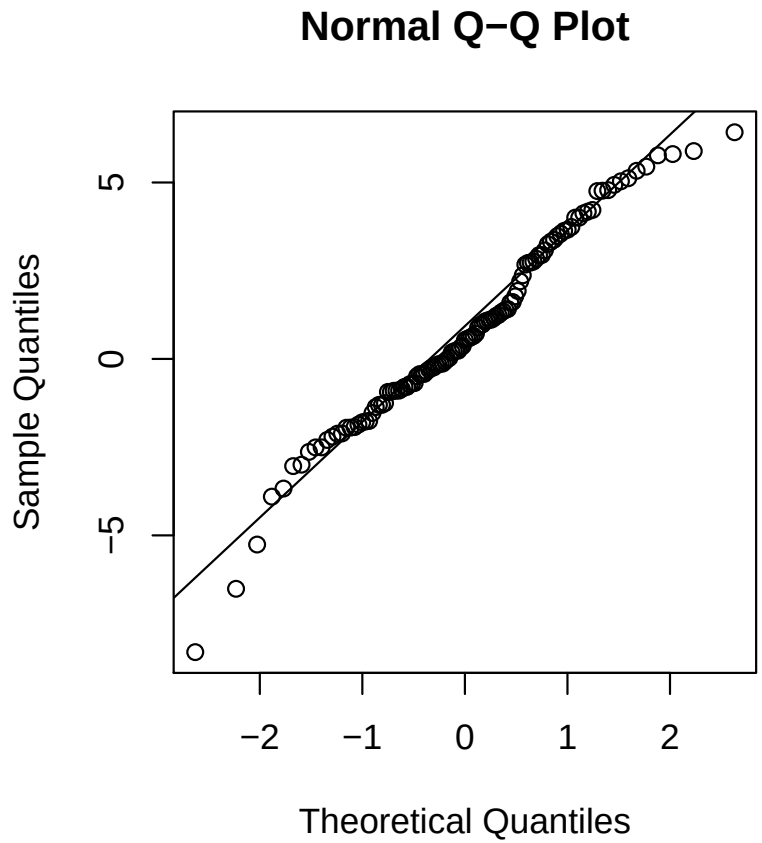

Figure: Gene set Q-Q plot

Investigating the outlier

```
> keggId <- names(z[which.min(z)])
> keggGS <- gsc[[keggId]]
> keggES <- bcrneg_filt2[keggGS, ]
> library(Category)
> getPathNames(keggId)

$`03010`
[1] "Ribosome"

> KEGGmnplot(keggId,
+      bcrneg_filt2,
+      annotation(bcrneg_filt2),
+      bcrneg_filt2$mol.biol,
+      pch=16, col="darkblue")
```

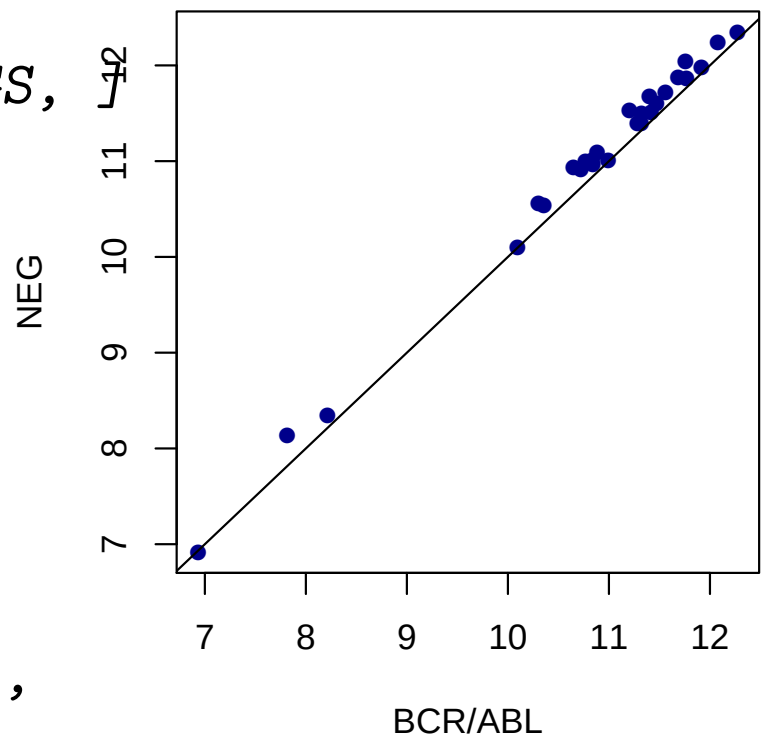

Figure: KEGG id 03010

More robust statistical assessment

Issues

- Strong assumptions, e.g., about independence of t statistics and normality of z_K.
- Very qualitative assessment; do other points deviate from Normal quantiles?

A solution

- More robust evaluation using permutation tests.
- Function `gseattperm` in Category package provides one implementation.
- Analysis in the lab leads to six significant pathways.

Other approaches possible...

Overlapping gene sets

Issues

- Two (or more) gene sets may share the same probes, e.g., 16 genes in common between sets 04512 and 04510.

```
> overlap <- gsc["04512"] & gsc["04510"]
> length(GSEABase::geneIds(overlap))
```

```
[1] 16
```

- If both gene sets are significant, is it because they share the same probes?

A solution

- Perform a series of linear models, e.g., models with (a) 04510, (b) 04512, (c) both sets, followed by a model with (d) probes only in 04510, only in 04512, and in both sets.
- Analyais in the lab suggests that 04512 is only interesting because of probes it shares with 04510.

Additional types of gene sets

- Chromosome bands
- Predefined sets, e.g., Broad Institute positional, curated, motif-based, or computed gene sets. See `?getBroadSets`, `BroadCollection`
- Gene Ontology (GO) and OBO collections.
- Pubmed IDs
- ...

Related approaches

- PGSEA: implements Kim and Volsky, 2005 (BMC Bioinformatics 6: 144).
- limma: `geneSetTest` performs like Mootha et al., but with different statistical tests.
- GOstats: gene ontology visualization, testing for statistical over-representation of probe sets in ontologies.
- GlobalAncova: Multivariate analysis suitable for assessing differential expression of specific gene sets.
- GSEAlm: flexible linear models to describe aggregate effects of probes in categories, rather than t-tests only.