Gene Set Testing

Simon Anders

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Setting

Setting:

- study with expression microarrays
- samples from two conditions (e.g., control vs. treatment, or tissue or phenotype vs. another)
- several biological replicates for each condition
- we are interested in differentially expressed (DE) genes

Standard gene-centred analysis

- Do the usual pre-processing.
- Assess statistical significance of any difference in expression between the two conditions with a suitable test procedure, e.g. Student's *t* test or with Limma.
- Correct for multiple testing, e.g., with Benjamini-Hochberg
- Look at the list of genes with significant differential expression.
- If the list is long, just look at the genes at the top.

- Find a biological meaning in the list.

Gene sets

Gene sets are sets of genes that have something in common, e. g., that they are

- part of the same pathway
- coding for proteins that are part of the same cellular component
- co-expressed under certain conditions
- putative targets of the same regulatory factor
- on the same cytogenetic band
- have come up as hits in some published assay
- …

Sources of gene sets

Sources of gene sets:

- Gene Ontology (GO) Annotation (GOA)
  - cellular components (CC)
  - biological processes (BP)
  - molecular functions (MF)
- Pathway data bases:
  - KEGG
  - GenMAPP
  - Biocarta
- Gene set collections
  - MSigDB
  - GAzer

Any published assay result can also be a very useful gene list.

Gene set testing

The standard question in gene set testing is:

“It there an association between the expression of the genes in the given gene set and the studied condition?”

This may mean: “Are unusually many (or: unusually few) of the genes in *S* differentially expressed?”

To make the meaning of “unusually many” more precise:
“If I picked n genes at random (with n being the size of *S*), how probable is it that among these genes, there are at least as many differentially expressed ones as in *S*?”

Even though it sound natural, we shall see that this is a problematic way of formalizing the question.

R. A. Fisher and the lady tasting tea

EMBL-EBI

Hypergeometric testing

A simple example:

5 patients with the disease *D* and 5 healthy control subjects have been checked for elevated levels of the blood constituent *C*. 4 of the patients, but only 2 of the healthy subjects show an elevated level of *C*.

May we infer that the concentration of *C* is elevated in patients with disease *D* more often than in healthy subjects?

Or could our result have been mere coincidence?

2×2 contigency table

	Patient with disease *D*	Healthy control subject	Total
Elevated level of compound *C*	4	2	6
Normal level of compound *C*	1	3	4
Total	5	5	10

Expected value for top left corner
from null model (no association): $5 \times 6 / 10 = 3$

EMBL-EBI

Hypergeometric distribution

Probability to get this 2×2 table without an association between *D* and *C*:

$$\frac{\text{Number of ways to choose 4 out of the 5 patients to have elevated } C \times \text{Number of ways to choose 2 out of the 5 controls to have elevated } C}{\text{Number of ways to choose 6 out of the 10 persons to have elevated } C} = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}}$$

in R:

```
> dhyper( 4, 5, 5, 6 )
[1] 0.2380952
```

Hypergeometric distribution

Under the null hypothesis, i.e., the assumption that there is no association between elevated levels of compound *C* and presence of disease *D, the* probability that 4 or even more of the patients have elevated levels of C, is

$$p = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}} + \frac{\binom{5}{5}\binom{5}{1}}{\binom{10}{6}} = 0.26$$

in R:

```
> 1 - phyper( 3, 5, 5, 6 )
[1] 0.2619048
```

Hypergeometric testing in R

```
> contigency.matrix <- rbind( c(4, 2), c(1, 3) )
> contigency.matrix
     [,1] [,2]
[1,]    4    2
[2,]    1    3
> fisher.test( contigency.matrix, alternative="greater" )

Fisher's Exact Test for Count Data

data:  contigency.matrix
p-value = 0.2619
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3152217       Inf
sample estimates:
odds ratio
  4.918388
```

Hypergeometric testing of gene sets

Given a list of differentially expressed genes and a collection of gene sets, the following strategy is often employed:

- For each gene set, fill a 2x2 contigency table:

	Differentially expressed	Not differentially expressed	Total
in gene set	.	.	.
not in geneset	.	.	.
Total	.	.	.

- Calculate *p* value by hypergeometric testing (Fisher′s exact test)

Hypergeometric testing with R

All this can be done conveniently with the hyperGTest function in the Category package.

Try it in the labs.

[Gentleman et al., Category package]

Universe

It is important to chose the universe correctly.

Case 1: Universe is all genes in the genome

	Differentially expressed	Not differentially expressed	Total
in gene set	10	30	40
not in geneset	390	3570	3960
Total	400	3600	4000

$p = 0.049$

Case 2: Universe is only the expressed genes

	Differentially expressed	Not differentially expressed	Total
in gene set	10	30	40
not in geneset	390	570	960
Total	400	600	1000

$p = 0.0048$

Cut-off

So far, we have divided the list of all genes in differentially expressed and not differentially expressed ones.

This is not optimal

- The choice of cut-off is always somewhat arbitrary.
  - Nevertheless, it may influence the result drastically [Pan et al., 2005]
- The ranking of the genes (or the strength of their DE) is not used.
- A concerted but small effect on many genes will be missed.

We shall discuss alternative approaches in the second half of the talk.

Sampling over genes

- Hypergeometric testing for gene sets has been critizised on the ground of it sampling over genes (observation) instead of over microarrays (subjects)
- Hence, the meaning of the *p* values is quite unclear.
- Especially: Correlations between genes inflate the apparent sample size, causing potentially severe over-estimation of significance.
- Increasing the number of replicates influences significance only indirectly.

[Goeman and Bühlmann, 2007]

Sampling over subjects

Goeman and Bühlmann's suggestion:

Instead of using the hypergeometric distribution to get a *p* value from out statistic, we should better use subject permutation:

- Let L_0 be the list of differential expressed genes and $m = |L_0|$ its size.
- For *N* permutations σ_i ($i=1,\dots,N$) of the *subject* labels, calculate the DE statistic and let L_i be the list of the *m* top ranking genes.
- Let k_i be the number of differentially expressed genes in the gene set, i.e., the size of the intersetion $L_i \cap S$.
- The *p* value for gene set *S* is now the fraction of permutation that had a larger gene set than the correct sample assignment, i.e.,

$$p = \frac{|\{i | k_i > k_0\}|}{N}$$

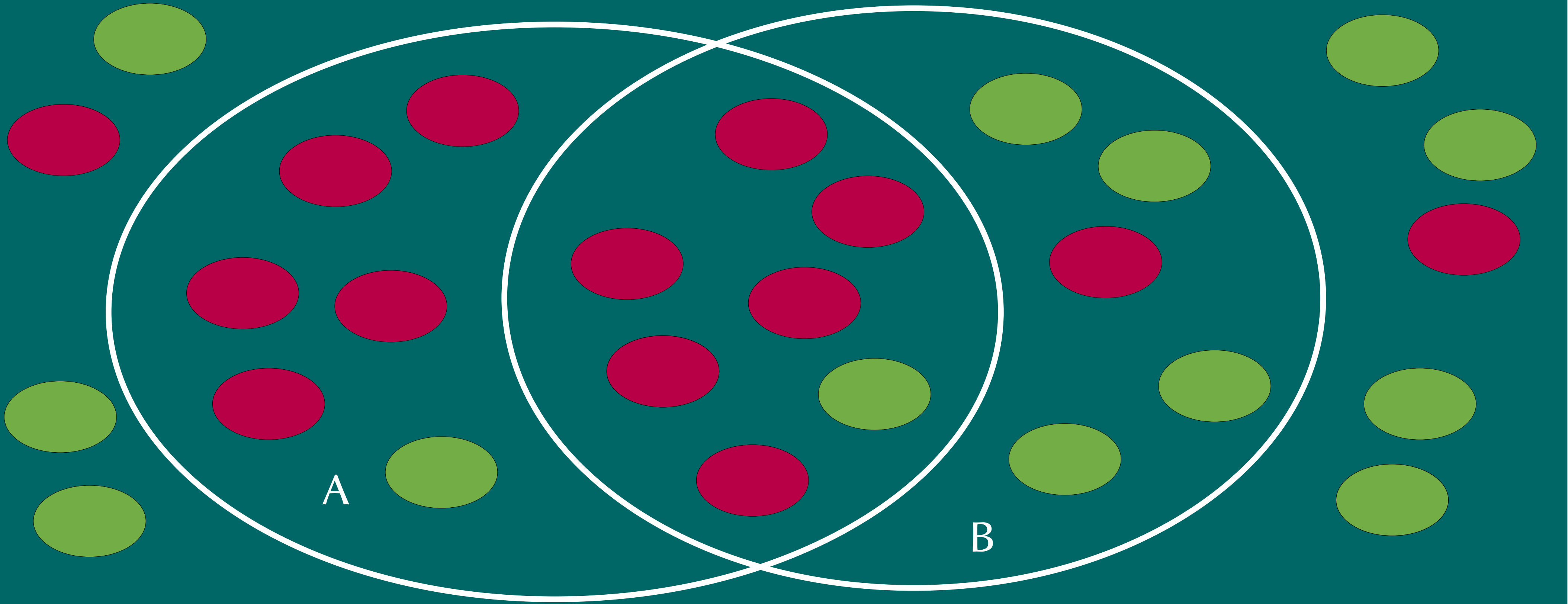
EMBL-EBI

Problems with subject sampling

- Enough replicates are required to have something to permute.
- The calculation is time consuming.

Hence, it may still make sense to use hypergeometric testing and live with the disagreement on whether it is statistically sound.

Overlap between gene sets



Both gene sets, A and B, are enriched. However, B seems to be enriched only because of its overlap with A.

GO is a directed acyclic graph (DAG)

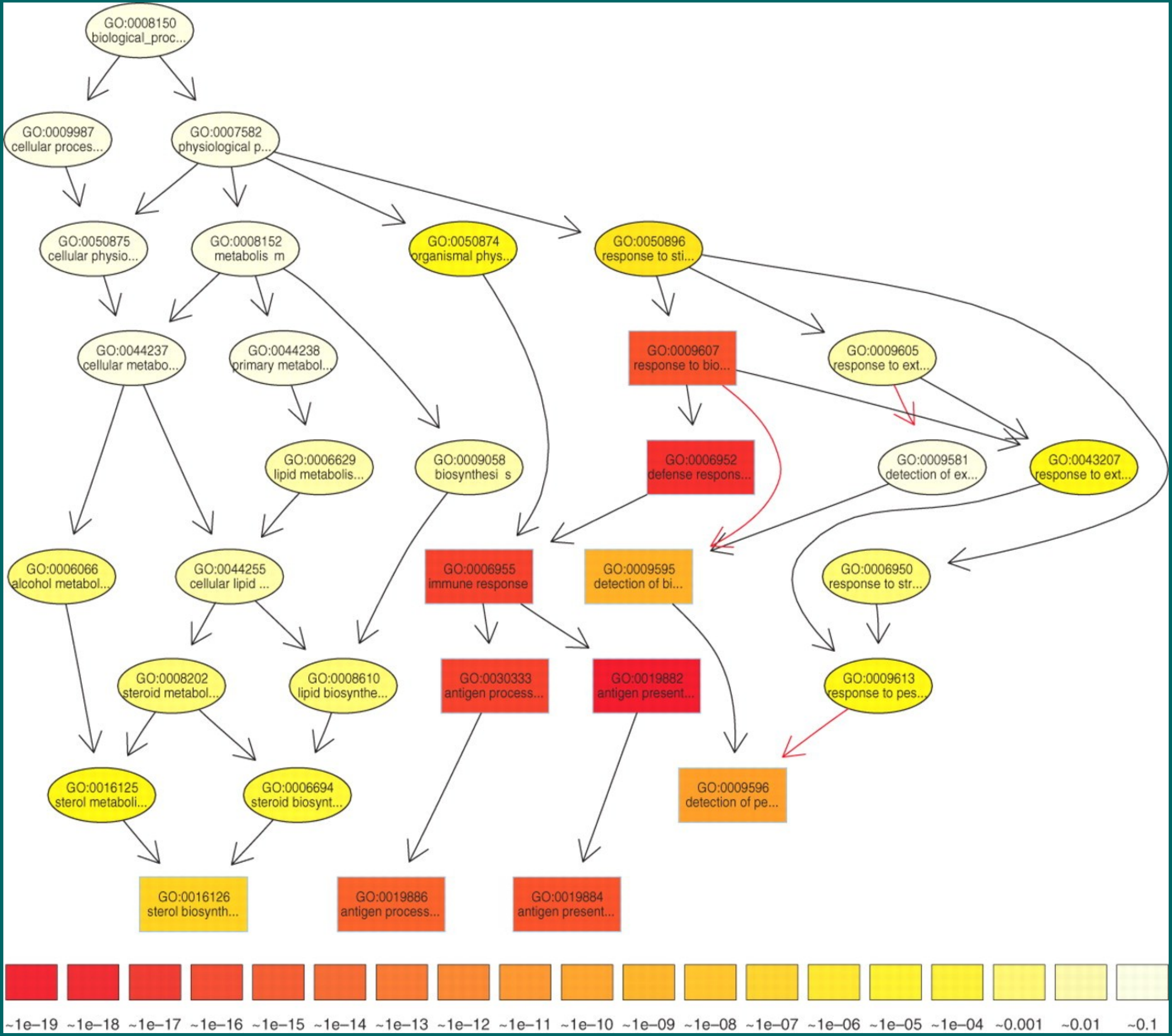

GO:0008150 biological_proc...
GO:0009987 cellular proces...
GO:0007582 physiological p...
GO:0050875 cellular physio...
GO:0008152 metabolis m
GO:0050874 organismal phys...
GO:0050896 response to sti...
GO:0044237 cellular metabo...
GO:0044238 primary metabol...
GO:0009607 response to bio...
GO:0009605 response to ext...
GO:0006629 lipid metabolis...
GO:0009058 biosynthesi s
GO:0006952 defense respons...
GO:0009581 detection of ex...
GO:0043207 response to ext...
GO:0006066 alcohol metabol...
GO:0044255 cellular lipid ...
GO:0006955 immune response
GO:0009595 detection of bi...
GO:0006950 response to str...
GO:0008202 steroid metabol...
GO:0008610 lipid biosynthe...
GO:0030333 antigen process...
GO:0019882 antigen present...
GO:0009613 response to pes...
GO:0016125 sterol metaboli...
GO:0006694 steroid biosynt...
GO:0009596 detection of pe...
GO:0016126 sterol biosynth...
GO:0019886 antigen process...
GO:0019884 antigen present...
~1e-19
~1e-18
~1e-17
~1e-16
~1e-15
~1e-14
~1e-13
~1e-12
~1e-11
~1e-10
~1e-09
~1e-08
~1e-07
~1e-06
~1e-05
~1e-04
~0.001
~0.01
~0.1

Figure taken from Alexa et al., 2006

EMBL-EBI

TopGO's elimination algorithm

The Bioconductor package "TopGO" [Alexa et al., Bioinformatics **22** (2006) 1600] offers this solution to the overlap problem in gene set collection that ar DAGs:

- Test the leaf gene sets first.
- If a gene set is significant, remove its genes from its ancestor sets before testing these.
  Alternatively, downweight them.

- The Category/GOstat package's [Falcon and Gentleman (2006)] hyperGTest offers a similar mechanism, called conditional testing.
- Goeman and Mansmann [2008] offer an alternative approach, namely to take the DAG into account when correcting for multiple testing.

Cut-off

So far, we have divided the list of all genes in differentially expressed and not differentially expressed ones.

This is not optimal

- The choice of cut-off is always somewhat arbitrary.
  - Nevertheless, it may influence the result drastically [Pan et al. (2005)]
- The ranking of the genes (or the strength of their DE) is not used.
- A concerted but small effect on many genes will be missed.

We shall now discuss alternative approaches.

A simple approach

- Calculate the log fold changes (LFCs) between conditions.
- Compare the LFCs in the gene set with those of the other genes with a two-sample *t* test.
- To get a *p* value, use Student's *t* distribution, or, better, get a null distribution from subject permutation.

Is this a good statistic?

Gene Set Enrichment Analysis (GSEA)

Mootha *et al.* [2003] suggest to use the Kolmogorov-Smirnov statistic:

- Sort all genes by LFC.
- Go through the list, increasing a running sum for each gene in the gene set by (*N*–*n*), and decreasing it for each gene not in the gene set by *n*.
  [*N*: number of genes, *n*: size of gene set]
- The maximum value of the running sum is the enrichment score (ES).

GSEA

Assessing significance

- To get *p* values, we do not use the KS distribution but rather estimate the null by subject permutation.

Improved enrichment score

- The KS statistic tests whether distributions are different, but this difference may not have a clear direction, making biological interpretation difficult.
- The updated GSEA algorithm [Subramanian *et al.*, PNAS **102** (2005) 15545] weights the in-/decrements of the running sum by the LFC.

What is the null hypothesis?

What does it mean to look for gene sets with “enrichment”?

It is important to distinguish:

- *Competitive null hypothesis*: The genes in the gene set do not have stronger association with the subject condition than the other genes.
- *Self-contained null hypothesis:* The genes in the gene set do not have any association with the subject condition (i.e., no gene in the set is differentially expressed). We do not care about what the genes outside the set do.

Tian et al. (2005),
Goemann and Bühlmann (2007),
Nam and Kim (2008)

EMBL-EBI

Competitive and self-contained null

Nam and Kim (2008) illustrate the difference with a simulation with 20 treatment and 20 control subjects: 30% of 2000 genes are DE. 100 gene sets were build by chosing 20 genes for each, at random and indipendently of the DE.

- The *p* values for the competitive null are uniform
- The *p* values for the self-contained null are below 5% for 83% of the gene sets.

Of course, if we want to test against the null hypothesis that the treatment does not cause any DE, this is fine.

Tian *et al.*'s self-contained test

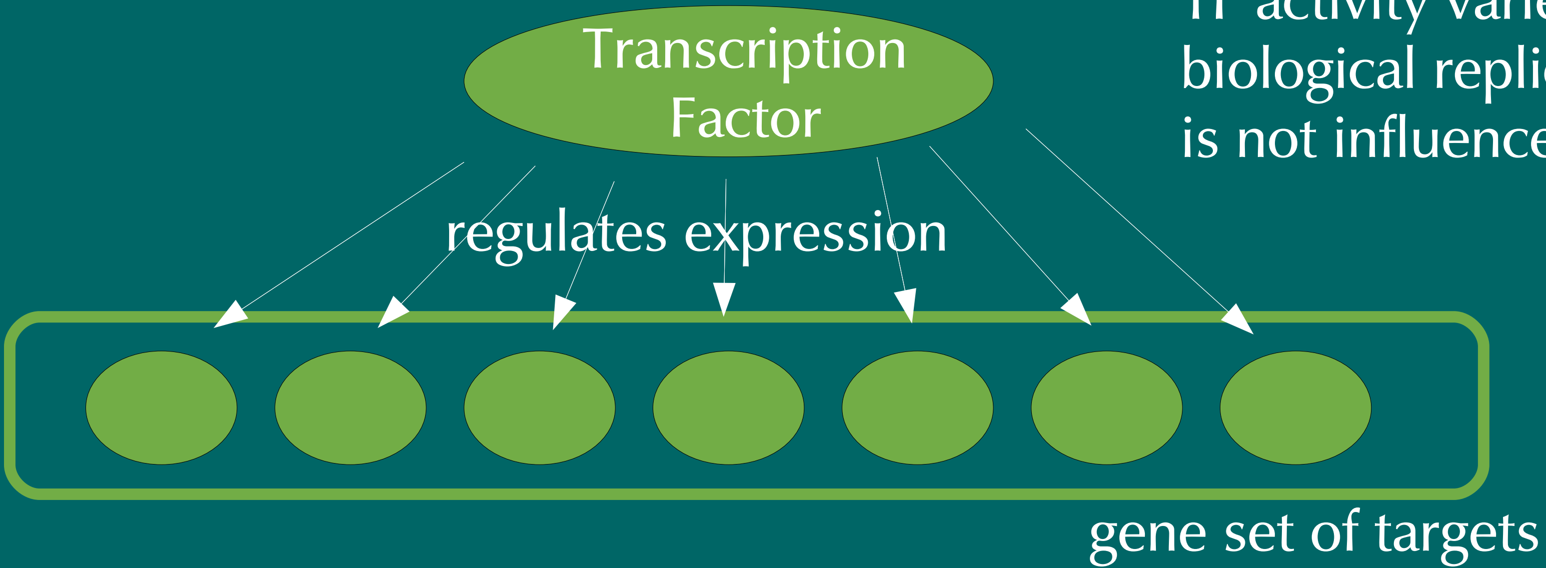
Tian *et al.* [2005] suggest to test each gene for differential expression and take the average of the t score of all the genes in a set as statistic. Subject permutation then yields a p value.

This is a test against the self-contained null, while GSEA [Mootha *et al.* (2003), Subramanian *et al.* (2005)] tests against the competitive null.

Other test statistics have been suggested, too, e.g. the sum of the squared t values [Dinu *et al.* (2007)] or Hotellings T^2 test [Kong *et al.* (2006)].

Now, subject permutation is crucial

Note that in tests without DE cut-off, it is even more important to use subject permutation to get *p* values.



TF activity varies between biological replicates but is not influenced by conditions

Neither the TF, not its targets may show a DE that is called significant, but the gene set might become significant nevertheless because of the correlation between the target genes.

Linear models for gene set testing

Hummel *et al.* (2007) suggest to use linear models to check gene sets for DE (self-contained null). This is especially useful in the presence of covariates.

Their example: Colorectal tumors may have good (“stage II”) or bad (“stage III”) prognosis. Which pathways show different activity in the two stages?

Differential expression may also be caused by these covariates:

- sex of the patient
- location of the tumor (colon or rectum)

GlobalANCOVA

The expression value of a gene in all the subject can now be regressed on the covariates tumor stage, sex and location and all their interaction ("full model") or only on sex and location ("reduced model")

For each gene set (pathway), we build a large model to regress all the genes in the set together and calculate how much variance the inclusion of the stage explains.

With subject permutations (not: with the *F* distribution), we check whether the reduction of the RSS is significant.

This tells us, for each gene set, whether there is association between its expression and the tumor stage.

GlobalANCOVA

As the permutation analysis may take time, an asymptotic calculation of the test statistic has been derived, too.

The R package “GlobalANCOVA” performs all this.

It also allows for intersting visualization by showing in plots, which genes and which subjects contribute how much to the reduction of the RSS of a gene set.

Mansmann and Meister (2005)
Hummel *et al.* (2007)

Summary

Gene set testing methods differ in these points:

- whether they employ a hard cut-off between differentially expressed and not diffentially expressed genes
- whether they calculate the *p* values from gene or subject sampling
- whether they test against the self-contained or the competitive null hypothesis (or against a hybrid of these)
- how the test statistic is calculated
- how they deal with overlapping gene sets

Summary

- whether they are available as stan-alone application, web tool or R package
- what gene set collections they can use

The review by Nam and Kim (2008) classifies all cut-off-free testing methods by these criteria.

Summary: R packages

- Category
- GOstats
- topGO
- GlobalANCOVA
- GSEAbase
- PGSEA
- SigPathways
- GSEAlm

References

- Alexa *et al.* (2006): Bioinf **22** 1600
- Dinu *et al.* (2007): BMC Bioinf **8** 242
- Falcon and Gentleman (2006): Bioinf
- Goeman and Bühlmann (2007): Bioinf **23** 980
- Goeman and Mansmann (2008): Bioinf **24** 537
- Hummel *et al.* (2007): Bioinf **23** 78
- Kong *et al.* (2006): Bioinf
- Mansmann and Meister (2005): Methods Inf Med **44** 449
- Mootha *et al.* (2003): Nat Genet **34** 267
- Nam and Kim (2008): Briefings in Bioinf
- Pan *et al.* (2005): PNAS **102** 8961
- Subramanian *et al.* (2005): PNAS **102** 15545
- Tian *et al.* (2007): PNAS **102** 13544