Gene Set Analysis

Hubert Rehrauer

23.11.2015

From Differential Gene Expression to System-Level Differences

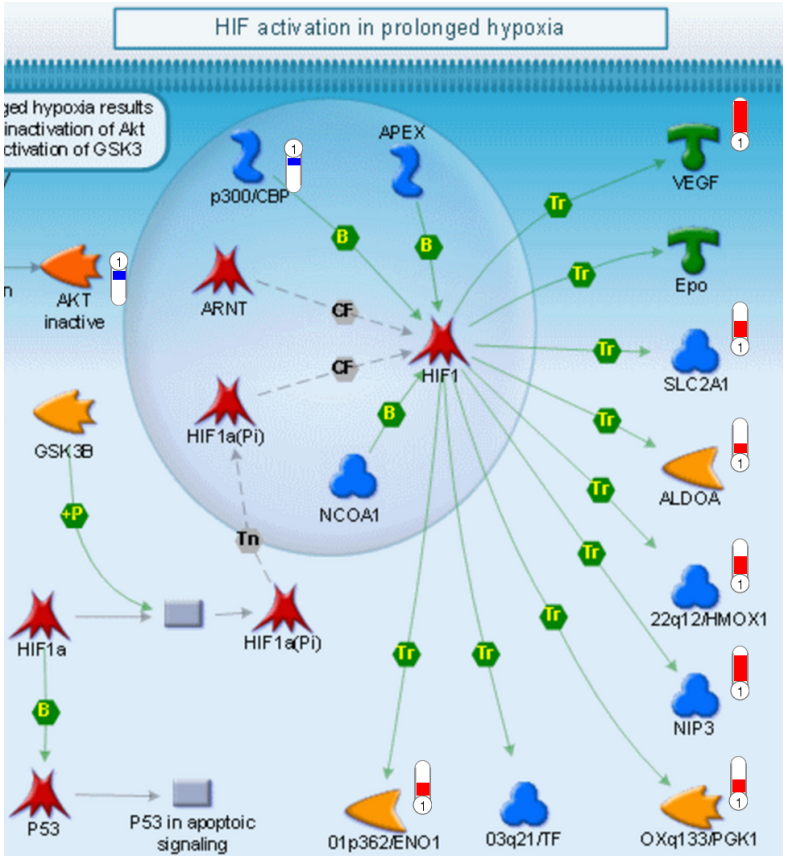
- Differential gene expression analyzes individual genes
- System-level functions are the product of multiple genes acting together
- Popular examples of **external knowledge** on system-level information of genes
  - pathway databases (KEGG, …)
  - Gene Ontology: categorization of individual genes according to
    - Biological Process
    - Cellular Component
    - Molecular Function
  - Chromosomal position
  - …

Functional Annotation

- Common characteristic:
  - Mappings of a **system-level property/function to a gene set**

- Note:
  - **external knowledge** ---- independent of my data
  - property → set of genes ---- NOT: {genes} → property

Example: Pathway Annotation



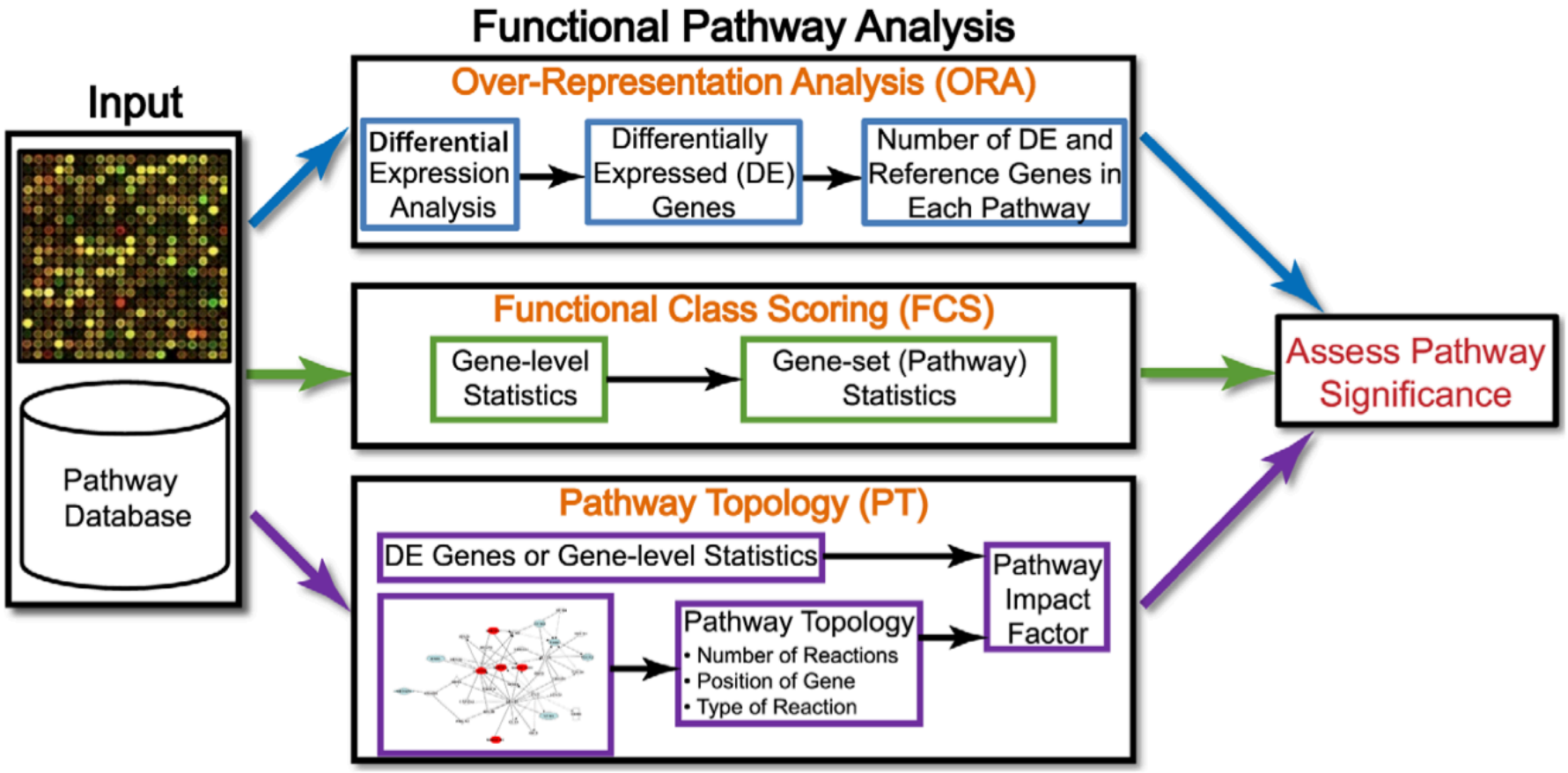
Hypoxia pathway
(response to lack of oxygen)

Thermometer for the diff. exp. genes
- red: up-regulated
- blue: down-regulated

For the central control element HIF1 no response has been measured, but the majority of the genes controlled by HIF1 go up

Is the hypoxia pathway activated? Is it significant?

Approaches



Khatri *etal.* **Ten years of pathway analysis.** *PLoS Comput Biol* 2012, **8**:e1002375.

Over-representation analysis (ORA)

- Find the differentially expressed genes with your method of choice
- Match the group of differentially expressed genes to the system-level gene sets
- Test of independence of differentially expressed genes with a system-level gene set
- Example for a contingency table:

	diff. expressed	not diff. Expressed	sum
in hypoxia pathway	107	673	780
not in hypoxia pathway	452	8673	9125
sum	559	9346	9905

Are there particularly many hypoxia genes differentially expressed?
Is there an association diff. expression – hypoxia? Is it significant?

Probabilities in the contingency table

Example:

- 780 genes of all 9905 genes considered are in the hypoxia pathway
- When drawing a random gene from the 9905, what is the probability that it is in the hypoxia pathway?

  780/ 9905

- When drawing 559 genes randomly, how many of them will be in the hypoxia pathway? Expectation value?

  559 * 780/9905 = 44.02

- We observed 107 genes. Is this significantly different from the expected 44?

Test of independence in the contingency table

- Approximate test: Chi-Square test
  - Is only precise if there is in every cell of the table a number that is “not small” (>6)

- Fisher’s Exact test:
  - hypergeometric test ---the probability of an observation in the contingency table is given by the hypergeometric distribution
  - Fisher’s Exact test sums up the exact probabilities of the contingency table
  - not easily computable by hand, therefore only widely used since the nineties

Hypergeometric distribution

Category A

Category B	In A	Not in A	
In B	**k**	**M-k**	**M**
Not in B	**K-k**	**N-K-M+k**	**N-M**
	K	**N-K**	**N**

Probability of observing k:

$$P[k \mid K, M, N] = \frac{\binom{K}{k}\binom{N-K}{M-k}}{\binom{N}{M}}$$

Expectation value:

$$E[k \mid K, M, N] = \frac{MK}{N}$$

Fisher's Exact Test

- Assume that the observed count is k'
- Null hypothesis: both categories are independent
  → the probability of k being equal or larger than the observed k' is:

$$P[k \geq k'] = \sum_{k=k'}^{K} P[k \mid K, M, N] = \sum_{k=k'}^{K} \frac{\binom{K}{k}\binom{N-K}{M-k}}{\binom{N}{M}}$$

- This is the p-value for the hypothesis test, that the two categorizations are positively associated with each other
- This formula is only for the one-sided test (positive association)
- The formula for the two-sided test can be obtained as an extension

Shortcoming of ORA

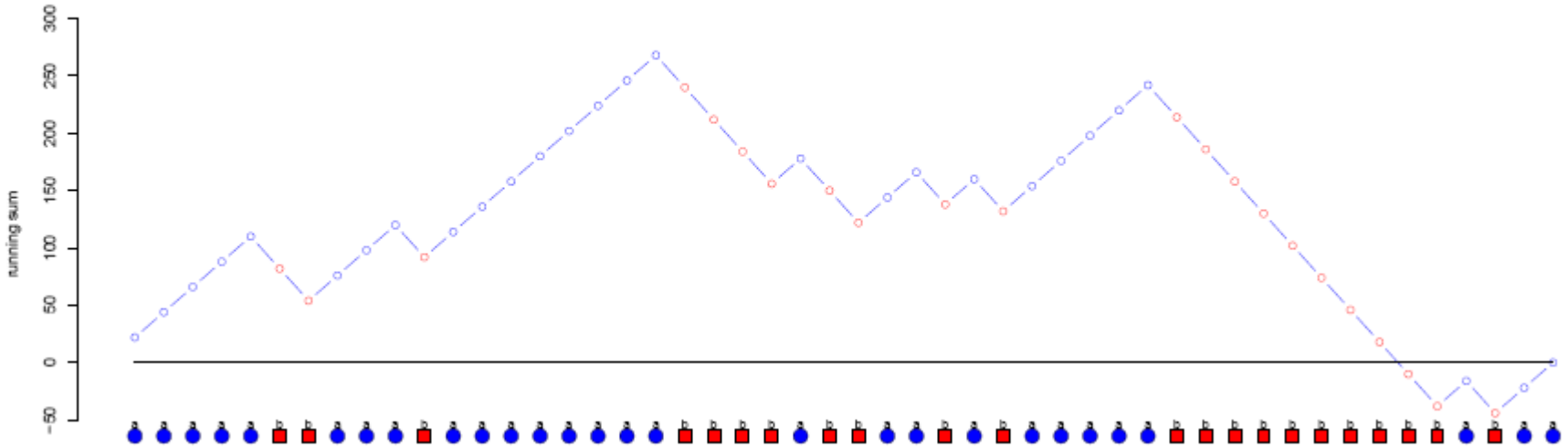
- The subdivision based on significance in “differentially expressed” and “not differentially expressed” is done by an arbitrary choice for a threshold

- In order to assess the association with an annotation category we count only the genes that are in the category and satisfy the threshold
- The ranking of the genes above/below the threshold is not used

- Idea:
  - Exploit the entire distribution of values from all genes in order to assess the association of differential expression with the annotation category
  - Functional Class Scoring (FCS)

Threshold-free Analysis

- Sort all genes according to differential expression (p-value, fold-change, ….)
- Draw a graph that shows the “cumulated overrepresentation of the genes in a category”
- Measure the maximal distance of the graph from the x-axis and compute the significance of the association of the category with differential expression (Lamb et al. 2003)
- Legend: red = gene is in the category blue = gene is not in the category



Lamb Algorithm

- Assume
  - n_A genes are in the category
  - n_B genes are not in the category
- Sort genes according to their differential expression score
- Iterate through the sorted genes and compute the "cumulated sum" S:
  - if the next gene is in the category, add n_B
  - if the next gene is not in the category, subtract n_A
- At the end the cumulated sum is: S = 0
  since $n_A\ n_B\ - n_B\ n_A = 0$
- Consider now the maximum of the sum $S_{max} = max(S)$
- Compare the observed maximum with the distribution of maxima S'_{max} that are obtained from random permutations of the rank
- Compute the p-value for positive association of the differential expression with the category

GSEA Algorithm

- Subramanian et al. (2005) improved the Lamb algorithm by introducing the summing rule:
- If gene i is in the category add:

$$\frac{r_i^p}{S}; \text{ mit } S = \sum_{j \subset A} r_j^p$$

- r_i: correlation coefficient of the expression values of gene i with the sample groups (only defined for 2 groups)
- p is a parameter
  - p=0 corresponds to the Kolmogorov-Smirnov Test
  - p=1 corresponds to the weighted sum of the correlations
- If gene i is not in the category subtract:
- Sum of all additions is 1 and sum of all subtractions is 1

$$\frac{1}{N - N_A}$$

GSEA Example

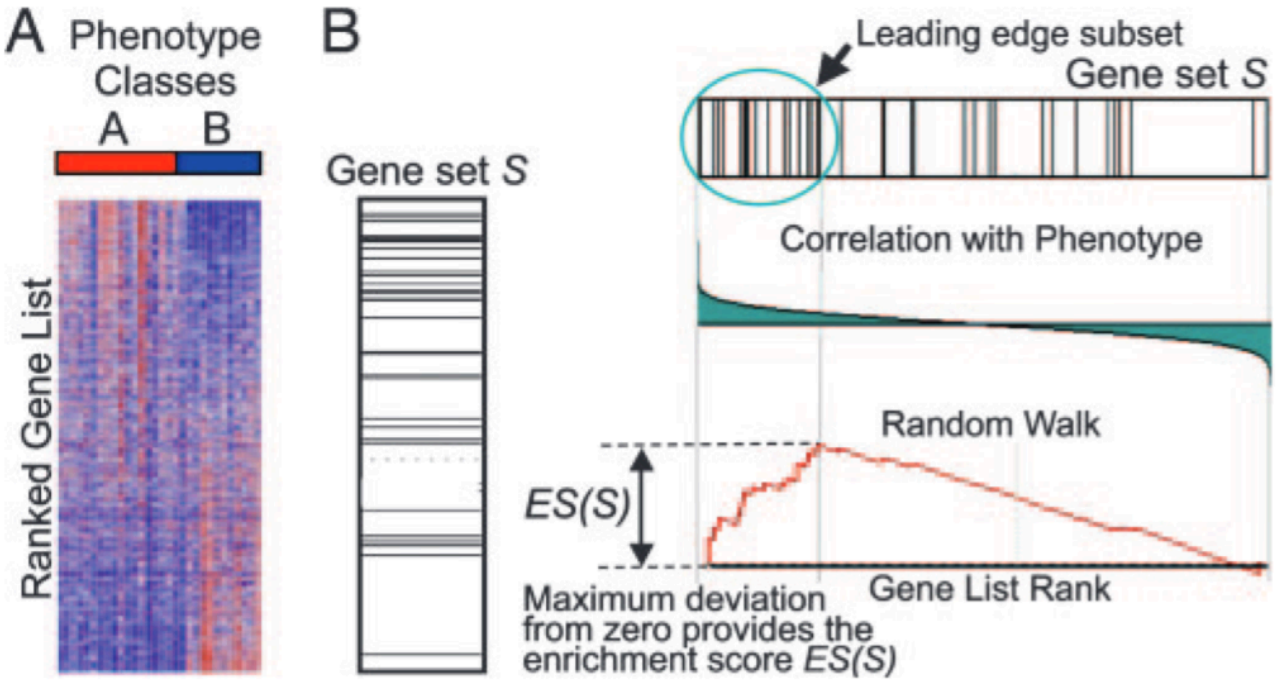


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Subramanian *etal*. 2005

Which hypothesis did we test?

- Assume two phenotypes A and B, for a given gene we have
  - the expression values: X_A and X_B and the difference score Δ

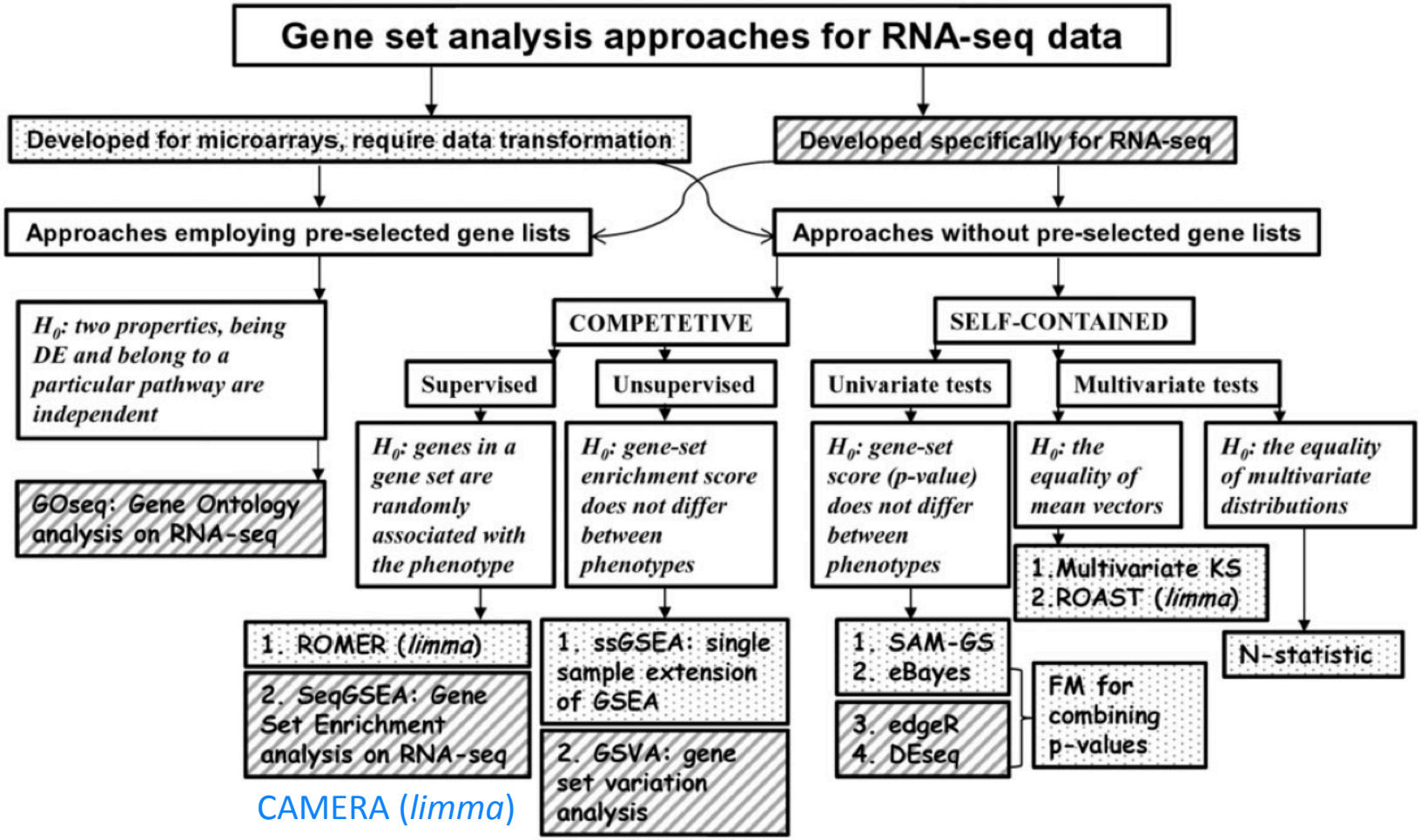
Competitive test:

- H0: For genes in the category the distribution of Δ is the same as for all other genes (genes not in the category)
- Both, Fisher-exact test based ORA and the GSEA algorithm test this hypothesis:

Self-contained test:

- H0: For genes in the category the distributions of X_A and X_B are the same
- No dependence on how the other genes behave

Approaches specific for RNA-seq

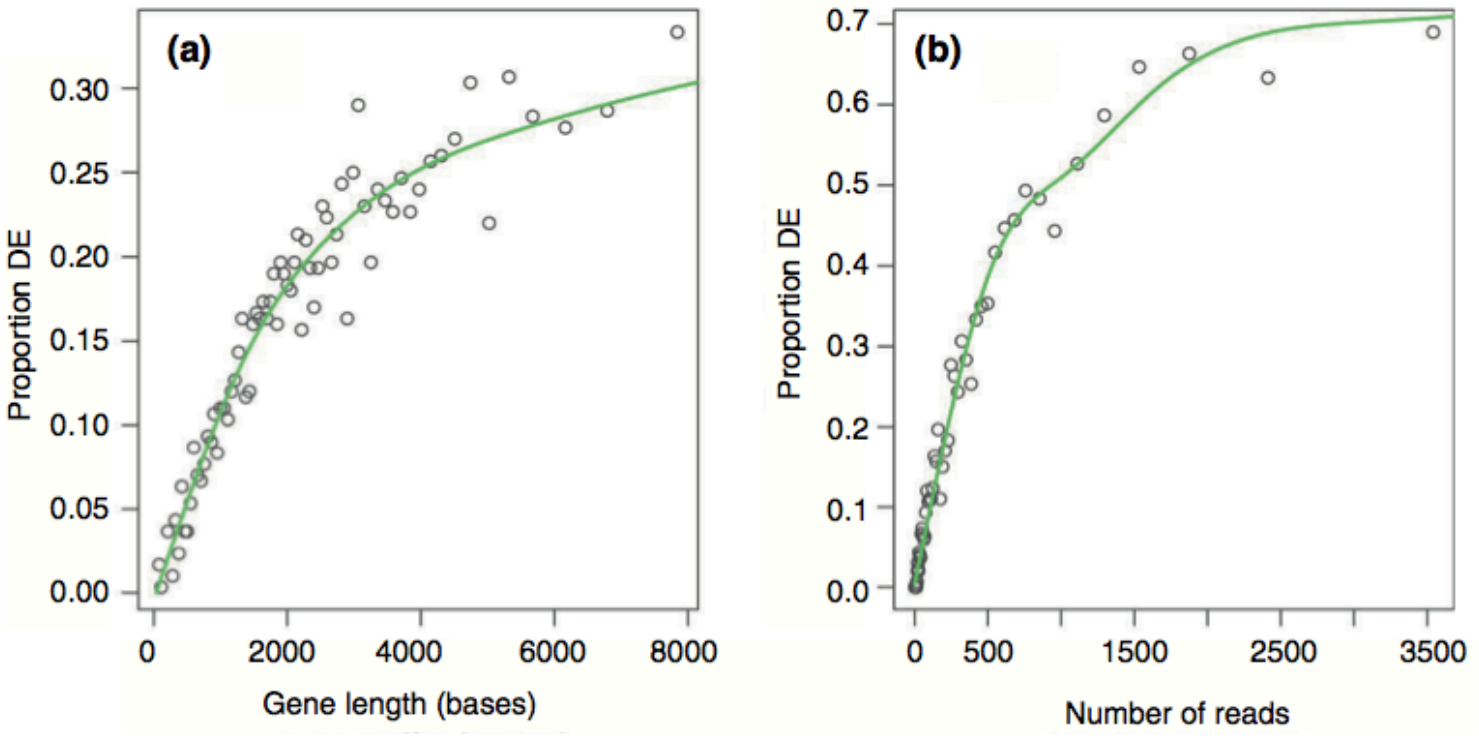


ROMER – GSEA for linear models

Rahmatallah *etal.* **Gene set analysis approaches for RNA-seq data** *Brief Bioinformatics* 2015.

goseq – considering biases in significance analyses

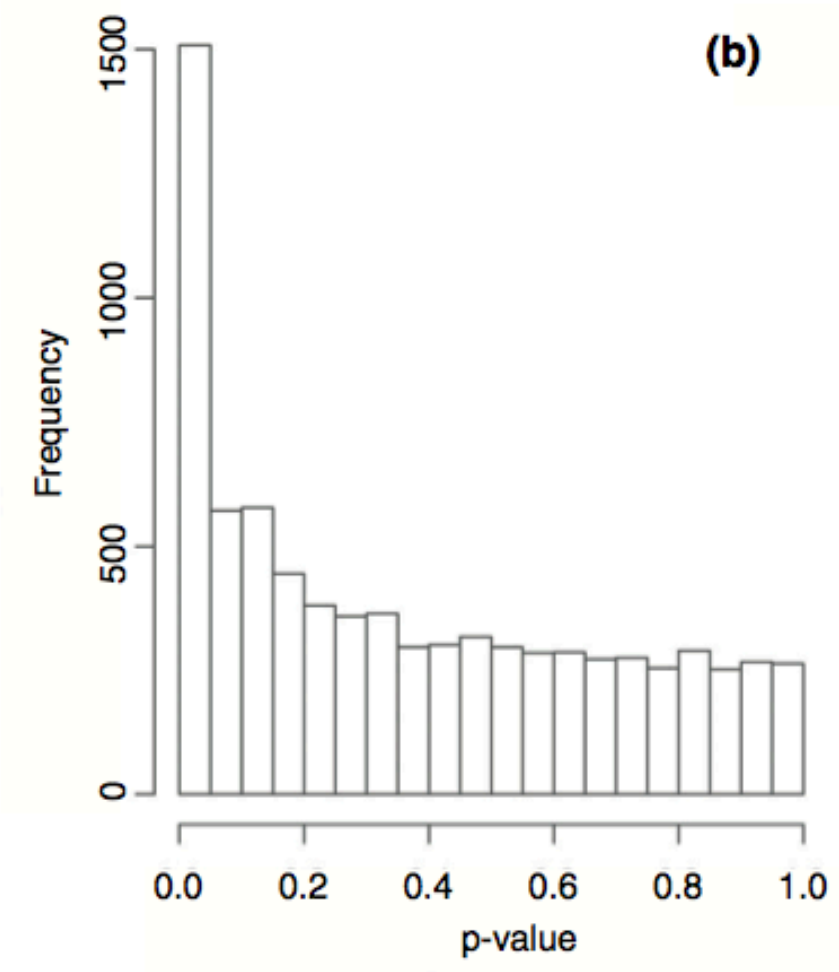
- Genes with many reads (long and or highly expressed) are more likely to be called significant if selected by p-value
- If gene selection is by fold-change the trend may be inversed



Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol* 2010, **11**:R14.

goseq – considering biases in significance analyses

- some GO categories show a significant overlap with short/ long genes



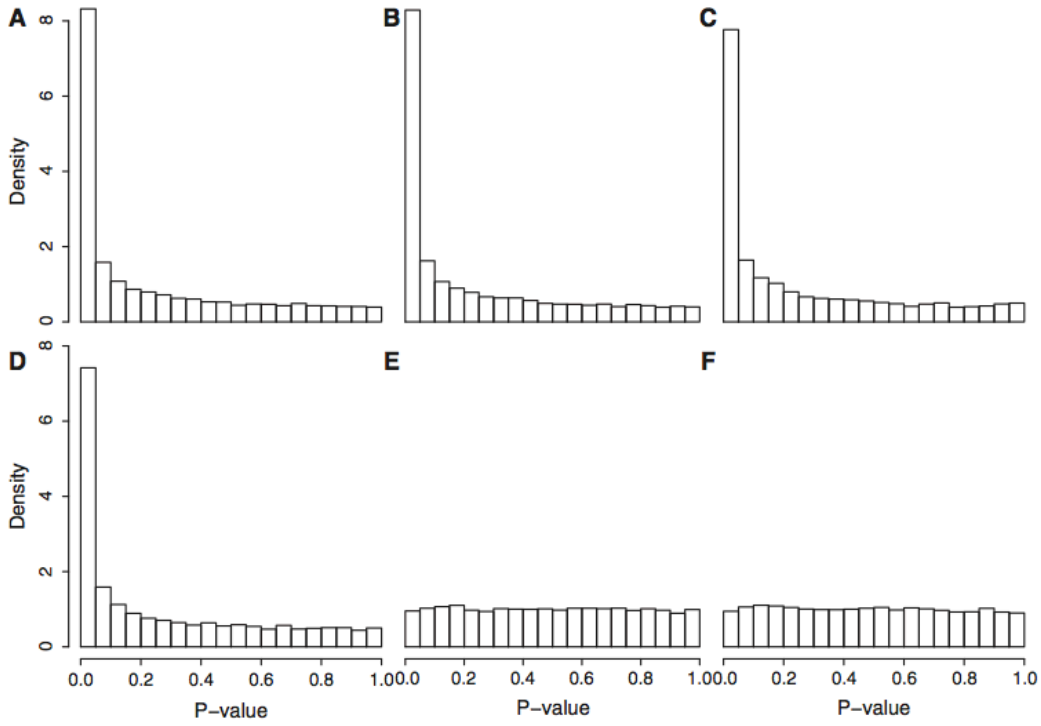
Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol* 2010, **11**:R14.

Construction of the NULL distribution

- NULL distribution can be obtained by permutations
- Competitive approach:
  - permute genes
- Self-contained approach:
  - permute samples

CAMERA – account for inter-gene correlation

- P-value distributions of gene set tests in the absence of differential expression



(A) geneSetTest (mod t),
(B) geneSetTest (ranks of mod t),
(C) sigPathway,
(D) PAGE,
(E) CAMERA (modt)
(F) CAMERA (ranks of modt)

Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation.** *Nucleic Acids Res* 2012, **40**:e133.

ROAST: rotation gene set tests

We instead adapt the idea of rotation tests from Langsrud (2005). Rotation tests use a type of simulation to generate *P*-values. The first step is to remove the nuisance parameters in the linear model, all the α_{gj} other than the contrast of interest β_g, by projecting the data for each probe onto the $d + 1$ dimensional residual space orthogonal to them. This yields a set of $d+1$ independent residuals, such that the *t*-statistic $\tilde{t}_g$ can be computed from the first residual. This step allows us to test $\beta_g = 0$ without making assumptions about the other coefficients in the linear model. The second step randomly rotates the residuals in $d+1$ dimensional space. For each rotation, the gene set statistic *T* is re-computed, and compared to the observed value. The final *P*-value is $p=(b+1)/(B+1)$, where *B* is the total number of rotations and *b* the number that yield a rotation statistic at least as extreme as that observed. This is an exact *P*-value (Barnard, 1963).

- Self contained
- Uses rotation instead of sample permutation
- Allows linear models with multiple factors

Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK: **ROAST: rotation gene set tests for complex microarray experiments.** *Bioinformatics* 2010, **26**:2176–2182.

Competitive tests: Choosing the gene universe

- Use all genes
  - simplest approach
- Use only present genes
  - considers that low expressed (at or below detection level) can not be differentially expressed
- Use only annotated genes
  - a gene may have no category assigned
  - this may have two reasons
    - it is in no category
    - it has not been studied yet

Considerations when testing for functional categories

- **Independence of measurements**
  - multiple correlated measurements for the same gene symbol
  - microarrays: multiple probes for the same gene
- Discordance of measurements for the same gene
- Many categories (>1000) → multiple testing
- Different categories may be highly correlated (example GO)

Independence of External Knowledge

- MSigDb (BROAD): Molecular Signature Databases
- Hallmark gene sets: *"These gene sets were generated … by identifying gene set overlaps and retaining genes that display coordinate expression. The hallmarks reduce noise and redundancy and provide a better delineated biological space for GSEA."* http://software.broadinstitute.org/gsea/msigdb/index.jsp
- **Risk of Circular Reasoning**
  - systematic bias in data can be misinterpreted

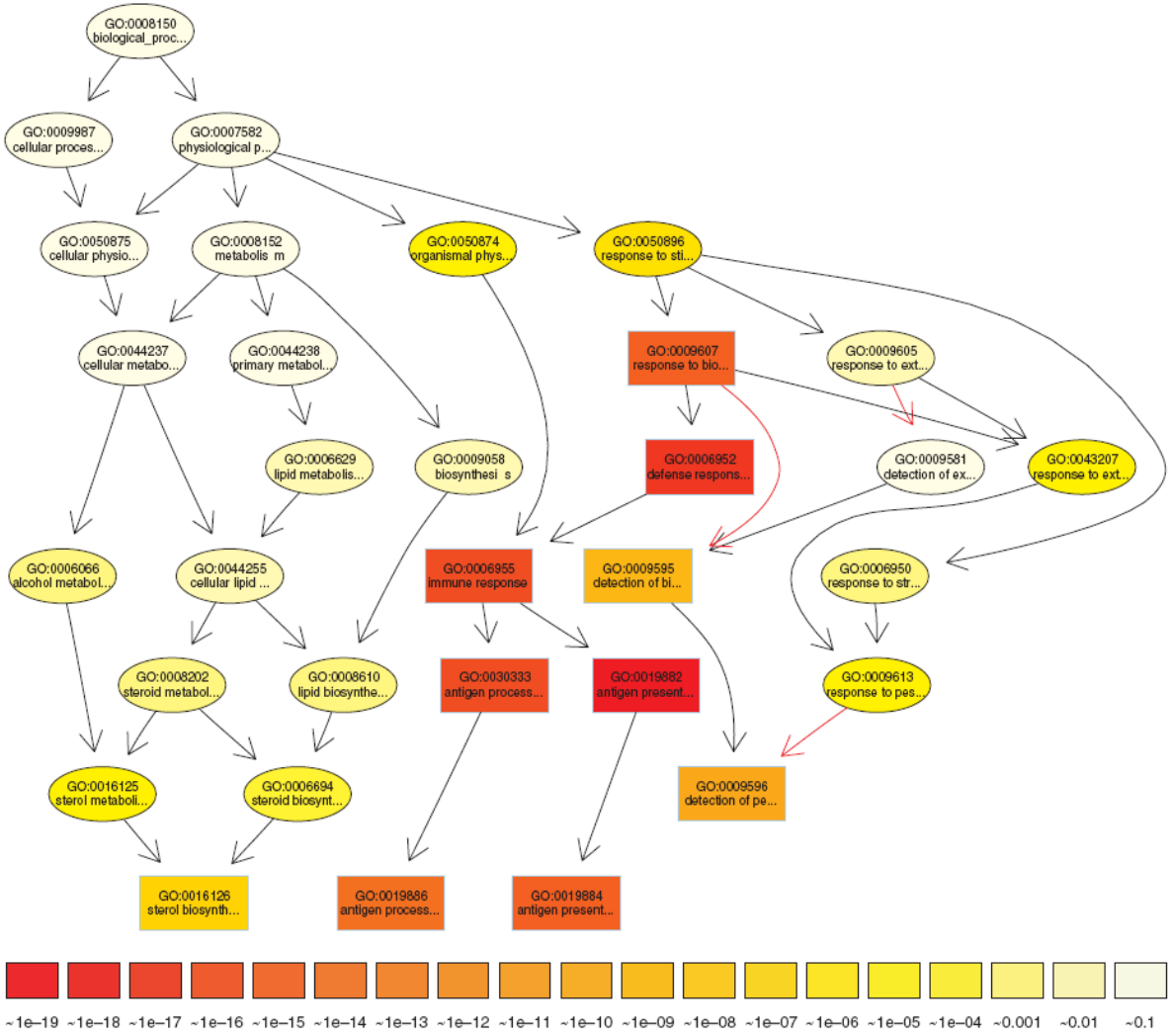
H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

Correlated terms in the gene ontology

- Tree-like structure implies that categories in the same branch have strongly correlated gene sets
- In hypothesis testing each category will be tested independently
- p-values of the categories of the same branch are correlated
- Multiple testing correction is only strictly valid for independent p-values

- Solution:
  - Alexa et al, 2006: Approach to decorrelation
  - Goeman et al 2007, Focus-Level Analysis

Decorrelation: Task



Standard analysis considers all "red" categories as significant

The information is redundant:

- response to bio. stimulus
- defense response
- immune response
- antigen presentation

Alexa et al. 2006

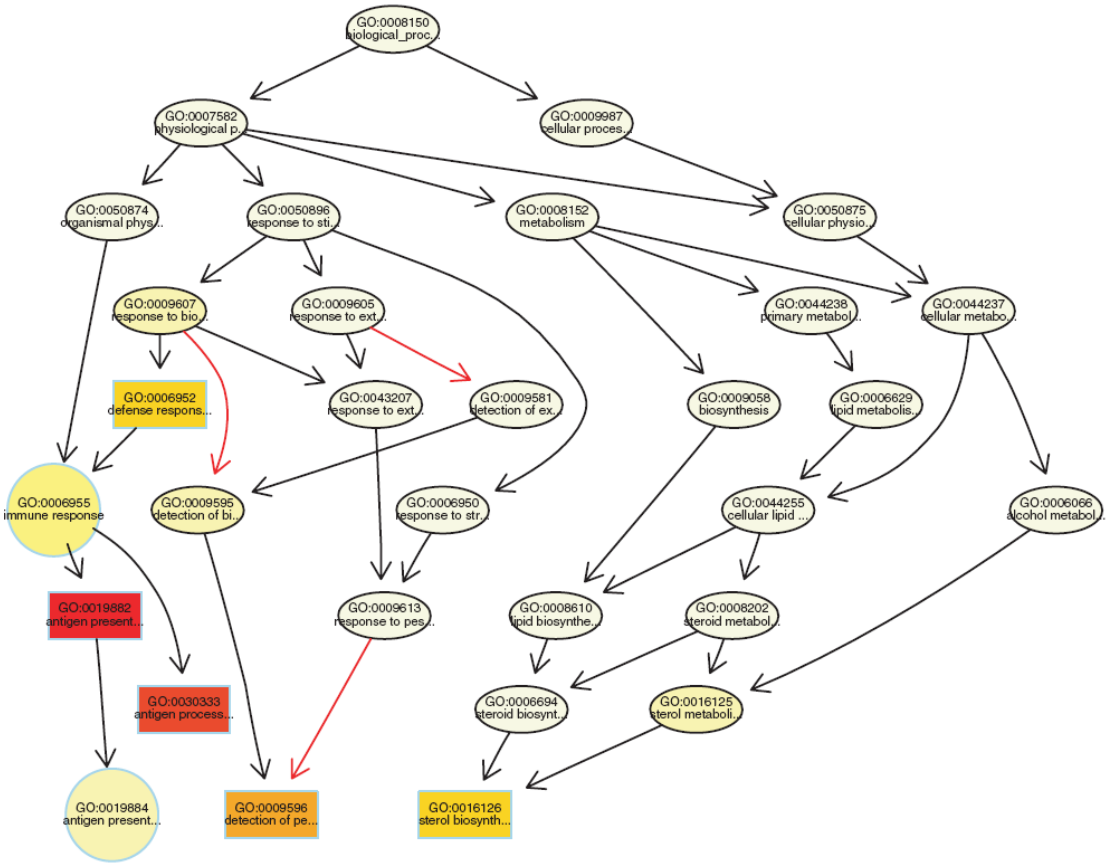
Decorrelation: Algorithm

- Correlation is caused by the fact that all genes of child node contribute also to the parent node

- Gene elimination:
  - When computing the significance of a parent, remove all those genes that belong already to a significant child-node

- Gene weighting:
  - Every gene provides a weighted contribution to each node where it occurs
  - The most significant node gets the highest weight
  - Iterative und approximate approach

Decorrelation: Result



Using the decorrelation approach the only significant categories are

- antigen processing …

Alexa et al. 2006