Pathway and Gene Set Analysis Part 2

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gov

Goals

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

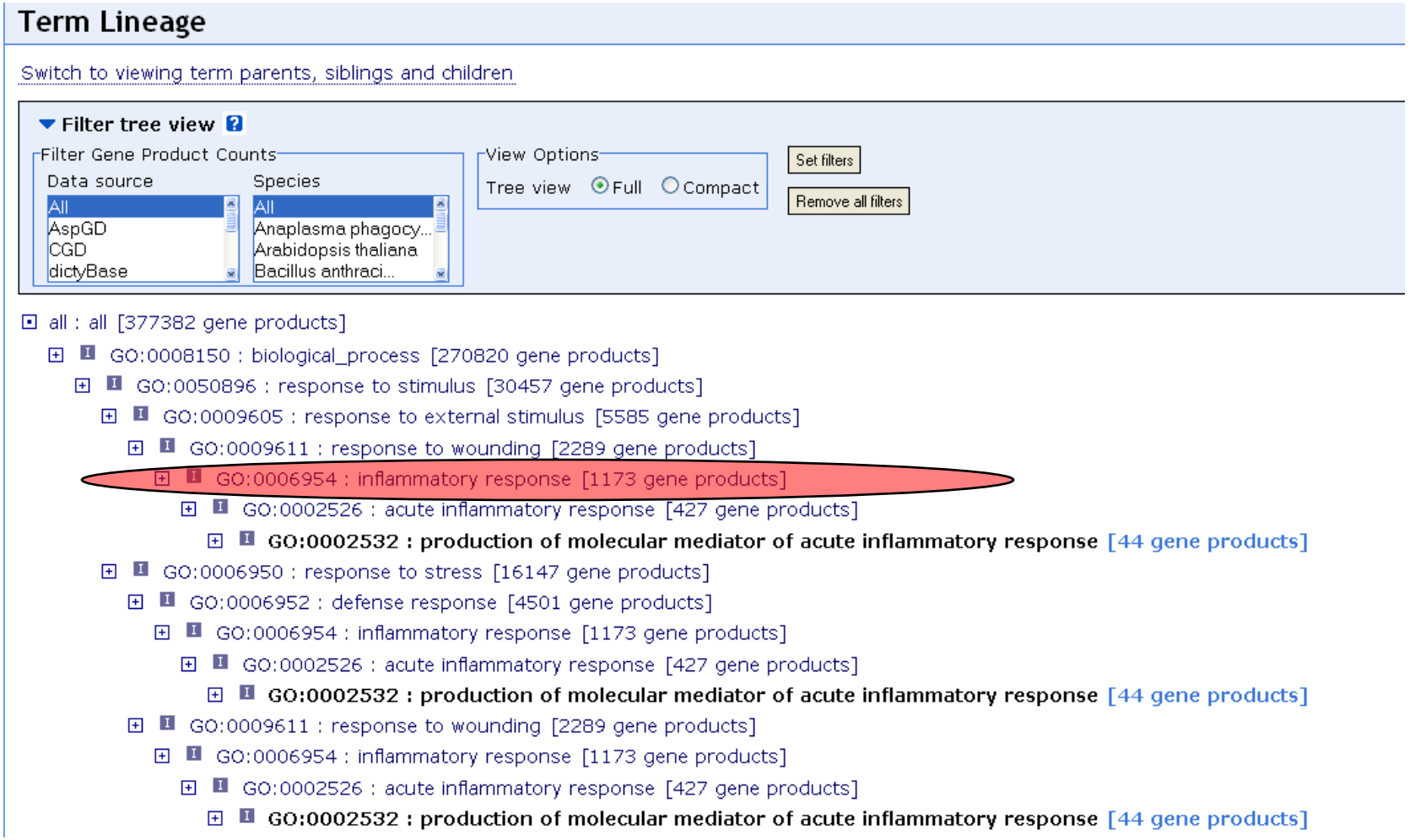
Aims of Analysis

- Reminder: The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)?
  - Can we give a number (p-value) to the probability of observing these changes just by chance?
  - Similar to single gene analysis statistical hypothesis testing plays an important role

General differences between analysis tools

- Self contained vs competitive test
  - The distinction between “self-contained” and “competitive” methods goes back to Goeman and Buehlman (2007)
  - A self-contained method only uses the values for the genes of a gene set
    - The null hypothesis here is: H = {“No genes in the Gene Set are differentially expressed”}
  - A competitive method compares the genes within the gene set with the other genes on the arrays
    - Here we test against H: {“The genes in the Gene Set are not more differentially expressed than other genes”}

Example: Analysis for the GO-Term “inflammatory response” (GO:0006954)



Back to the Real Data Example

- Using Bioconductor software we can find 96 probesets on the array corresponding to this term
- 8 out of these have a p-value < 5%
- How many significant genes would we expect by chance?
- Depends on how we define “by chance”

The “self-contained” version

- By chance (i.e. if it is NOT differentially expressed) a gene should be significant with a probability of 5%
- We would expect 96 x 5% = 4.8 significant genes
- Using the binomial distribution we can calculate the probability of observing 8 or more significant genes as p = 0.108, i.e. not quite significant

The “competitive” version

- Overall 1272 out of 12639 genes are significant in this data set (10.1%)
- If we randomly pick 96 genes we would expect 96 x 10.1% = 9.7 genes to be significant “by chance”
- A p-value can be calculated based on the 2x2 table
- Tests for association: Chi-Square-Test or Fisher’s exact test

	In GS	Not in GS
sig	8	1264
non-sig	88	11 279

P-value from Fisher’s exact test (one-sided): 0.733, i.e very far from being significant

Competitive Tests

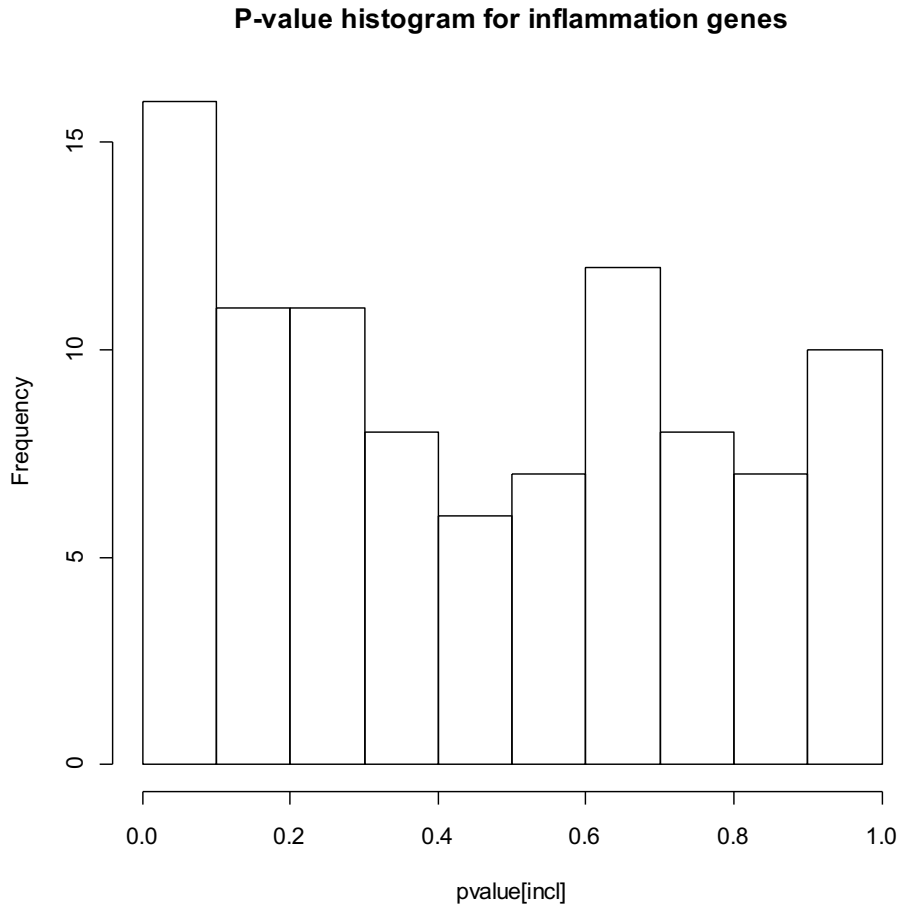
- Competitive results depend highly on how many genes are on the array and previous filtering
  - On a small targeted array where all genes are changed, a competitive method might detect no differential Gene Sets at all
- Competitive tests can also be used with small sample sizes, even for n=1
  - BUT: The result gives no indication of whether it holds for a wider population of subjects, the p-value concerns a population of genes!
- Competitive tests typically give less significant results than self-contained (as seen with the example)
- Fisher's exact test (competitive) is probably the most widely used method!

Cut-off methods vs whole gene list methods

- A problem with both tests discussed so far is, that they rely on an arbitrary cut-off

- If we call a gene significant for 10% alpha threshold the results will change
  - In our example the binomial test yields p= 0.022, i.e. for this cut-off the result is significant!

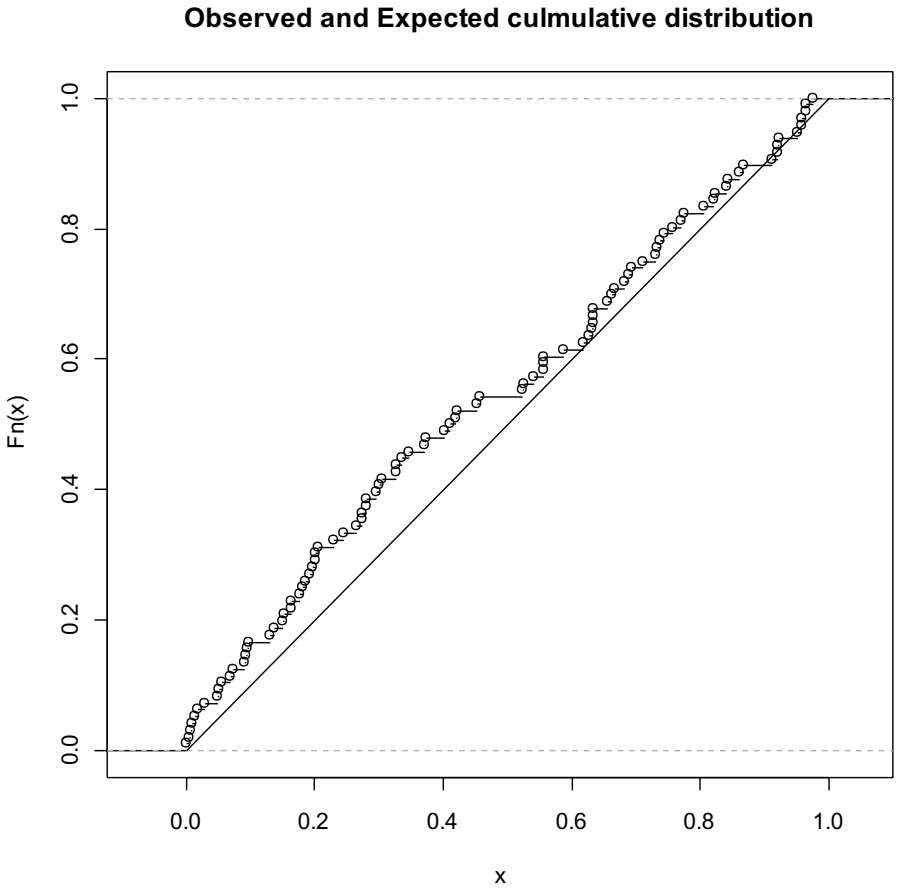
- We also lose information by reducing a p-value to a binary (“significant”, “non-significant”) variable
  - It should make a difference, whether the non-significant genes in the set are nearly significant or completely unsignificant

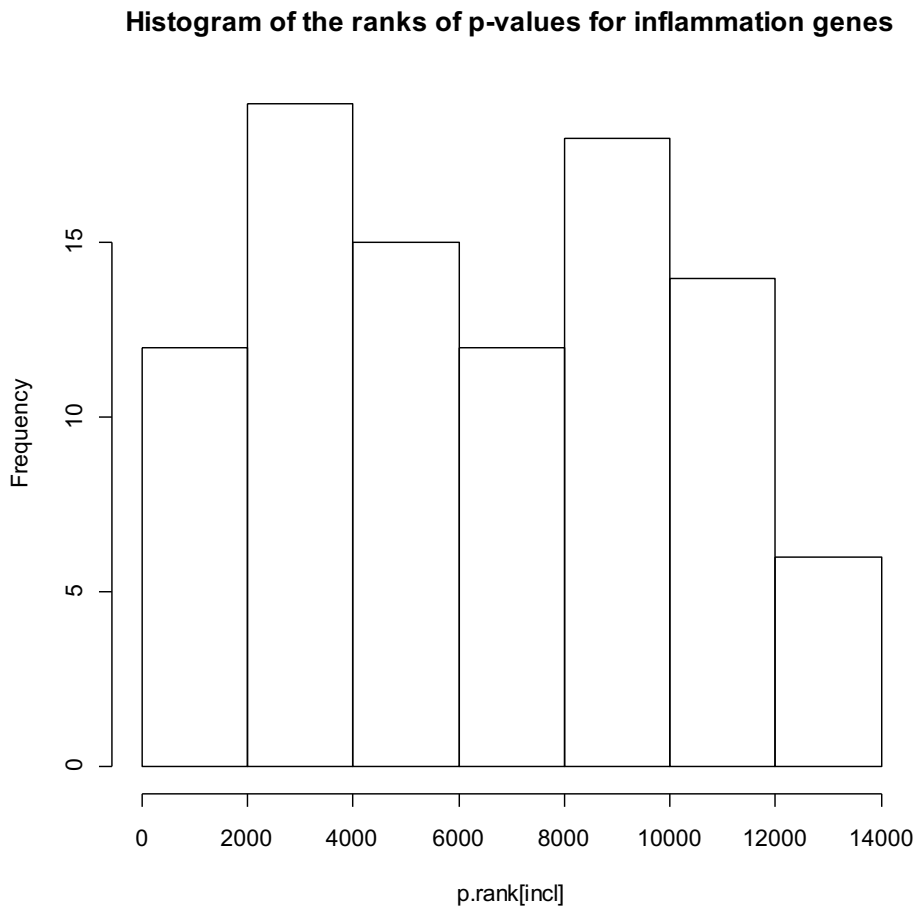


- We can study the distribution of the p-values in the gene set
- If no genes are differentially expressed this should be a uniform distribution
- A peak on the left indicates, that some genes are differentially expressed
- We can test this for example by using the Kolmogorov-Smirnov-Test
- Here p = 0.082, i.e. not quite significant
- This would be a “self-contained” test, as only the genes in the gene set are being used

Kolmogorov-Smirnov Test

- The KS-test compares an observed with an expected cumulative distribution
- The KS-statistic is given by the maximum deviation between the two





- Alternatively we could look at the distribution of the RANKS of the p-values in our gene set
- This would be a competitive method, i.e we compare our gene set with the other genes
- Again one can use the Kolmogorov-Smirnov test to test for uniformity
- Here: p= 0.851, i.e. very far from significance

Other general issues

- Direction of change
  - In our example we didn't differentiate between up or down-regulated genes
  - That can be achieved by repeating the analysis for p-values from one-sided test
    - Eg. we could find GO-Terms that are significantly up-regulated
  - With most software both approaches are possible

- Multiple Testing
  - As we are testing many Gene Sets, we expect some significant findings "by chance" (false positives)
  - Controlling the false discovery rate is tricky: The gene sets do overlap, so they will not be independent!
    - Even more tricky in GO analysis where certain GO terms are subset of others
  - The Bonferroni-Method is most conservative, but always works!

Multiple Testing for Pathways

- Resampling strategies (dependence between genes)
  - The methods we used so far in our example assume that genes are independent of each other…if this is violated the p-values are incorrect
  - Resampling of group/phenotype labels can correct for this
  - We give an example for our data set

Example Resampling Approach

1. Calculate the test statistic, e.g. the percentage of significant genes in the Gene Set
2. Randomly re-shuffle the group labels (lean, obese) between the samples
3. Repeat the analysis for the re-shuffled data set and calculate a re-shuffled version of the test statistic
4. Repeat 2 and 3 many times (thousands…)
5. We obtain a distribution of re-shuffled % of significant genes: the percentage of re-shuffled values that are larger than the one observed in 1 is our p-value

Resampling Approach

- The reshuffling takes gene to gene correlations into account
- Many programs also offer to resample the genes: This does NOT take correlations into account
- Roughly speaking:
  - Resampling phenotypes: corresponds to self-contained test
  - Resampling genes: corresponds to competitive test

Resampling Approaches

- Genes being present more than once
  - Common approaches
    - Combine duplicates (average, median, maximum,…)
    - Ignore (i.e treat duplicates like different genes)

- Using summary statistics vs using all data
  - Our examples used p-values as data summaries
  - Other approaches use fold-changes, signal to noise ratios, etc…
  - Some methods are based on the original data for the genes in the gene set rather than on a summary statistic

Resampling Approaches

- The resampling approaches are highly computationally intensive

- New methods are being developed to speed this up
  - Empirical approximations of permutations
  - Empirical pathway analysis, without permutation.
    - Zhou YH, Barry WT, Wright FA.Biostatistics. 2013 Jul;14(3):573-85. doi: 10.1093/biostatistics/kxt004. Epub 2013 Feb 20.

Summary

- Databases
- Choice makes a difference
- Not all use the same IDs – watch out ☺
- Major differences between methods
- Issues with multiple testing

- Next lecture, will go into more detail on a few methods

Questions?