Gene Set & Pathway Testing of RNA-Seq Expression Data

Ryan C. Thompson

Su Lab Meeting, February 6, 2014

Outline

- Why Gene set testing?
- Competitive & self-contained tests
- Why pathway testing?
- Available gene set and pathway databases
- Available gene set and pathway tests
- Test results on real data!

Introduction

Why gene set testing?

- We want to make inferences about biological systems from gene expression (or other) data
- Individual differentially expressed genes may not be important
- Interpreting biology from a list of 1000s of individual DE genes is difficult
- Some processes may be regulated by the cumulative effect of many genes with small changes not detectable at the gene level

Competitive & Self-contained gene set tests

- **Null Hypothesis Q1 (Competitive test)**: The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes.
  - Alternative: Genes in the set are *more* strongly associated with the phenotype compared to genes not in the set
  - Problem: Sensitive to inter-gene correlations & definition of "all genes"
- **Null Hypothesis Q2 (Self-contained test)**: The gene set does not contain any genes whose expression levels are associated with the phenotype of interest.
  - Alternative: Some genes in the set have a nonzero association with the phenotype
  - Problem: Not specific enough when many genes are DE

Why pathway testing?

- Some genes in a pathway are more important (greater effect on pathway output)
- Some genes are inhibitory and oppose the action of other genes
- More generally, pathway topology is important

Available Gene Set & Pathway Databases

What can we test against?

- Molecular Signatures Database (MSigDB) from the Broad
  - Organized into seven collections of gene sets
  - Some collections are further subdivided
- graphite: "GRAPH Interaction from pathway Topological Environment"
  - "Graph objects from pathway topology derived from Biocarta, HumanCyc, KEGG, NCI, Panther, Reactome and SPIKE databases."
  - All converted into a common representation, non-gene nodes reduced to edges between genes
  - Also provides routines for running many pathway tests on these pathways
- WikiPathways
  - You are probably familiar with this one
  - Not currently sure how to use topological information, so I'm treating them as gene sets

MSigDB: Curated Gene Sets from the Broad

- **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 (TF & miRNA targets)
- **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.

graphite: Reducing Non-gene Nodes to Edges

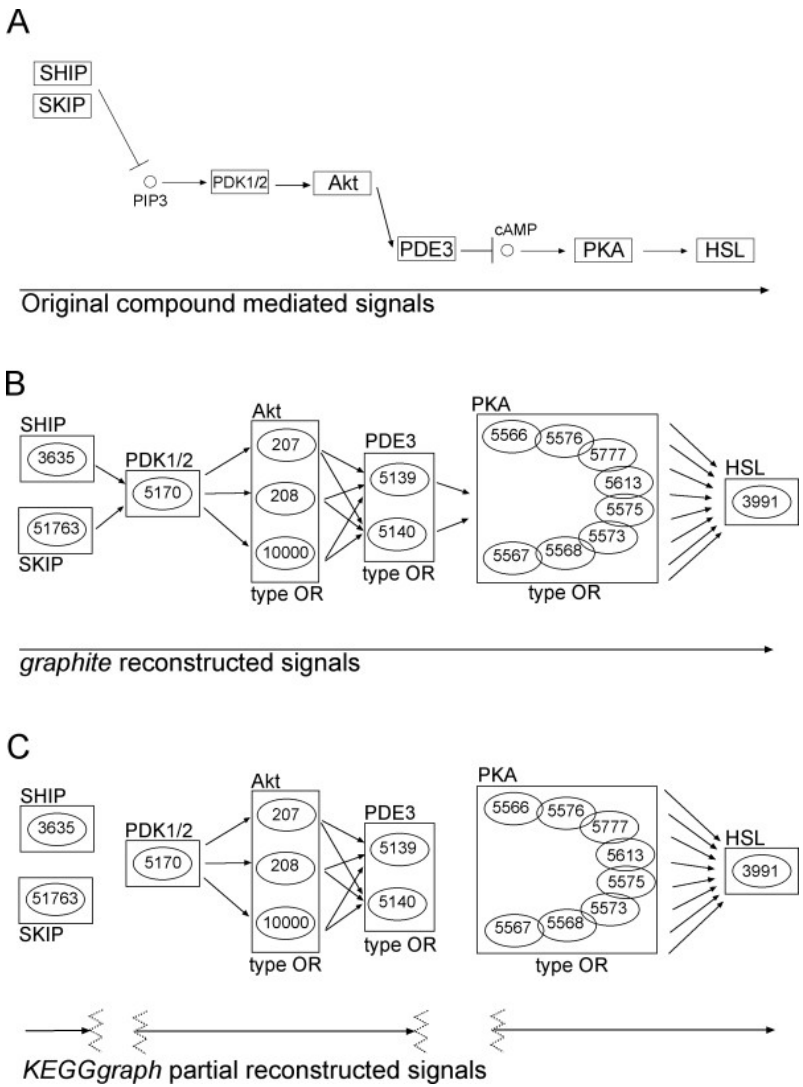

Figure : Signal pathway reconstruction

Gene Set & Pathway Tests

Naive approach: contingency tables

- Two categorical variables: gene in/out of gene set and significantly DE/NDE
- Form 2x2 contingency table and test for association using Fisher's exact test or hypergeometric test
- Very crude test based on DE threshold
- This is a competitive test (Q1) and very sensitive to inter-gene correlation
- Don't use this test if you can avoid it

Better appraoch: t-test & Wilcoxon rank-sum tests

- Sort all genes by fold change (or t-stat or other significance measure)
  - For a non-directional test, use absolute value of stat
- Test whether genes in set have a significantly different mean than the rest of the genes, parametrically or non-parmetrically
- Far better than contingency tables because DE is defined on a continuous spectrum
- Inter-gene correlation is still a large issue
- Even small correlations can result in lots of false postives

CAMERA: adjusting for correlation

- CAMERA stands for "Correlation-Adjusted MEan RAnk" test (also provides correlation-adjusted t-test as well)
- First, fit your gene-level linear model (or GLM) using limma/edgeR
- For each gene set, compute a variance inflation factor (VIF) based on correlation in the residuals of genes in that set
- Usually not enough samples available to estimate all pairwise correlations between genes
- CAMERA estimates a single VIF for each set representing the average correlation between all genes in the gene set, instead of estimating individual correlations and averaging them
- Define correlation-adjusted variations of t- and Wilcoxon tests incorporating the VIF as a penalty for positive inter-gene correlation

CAMERA: Variance Inflation Factors in MSigDB

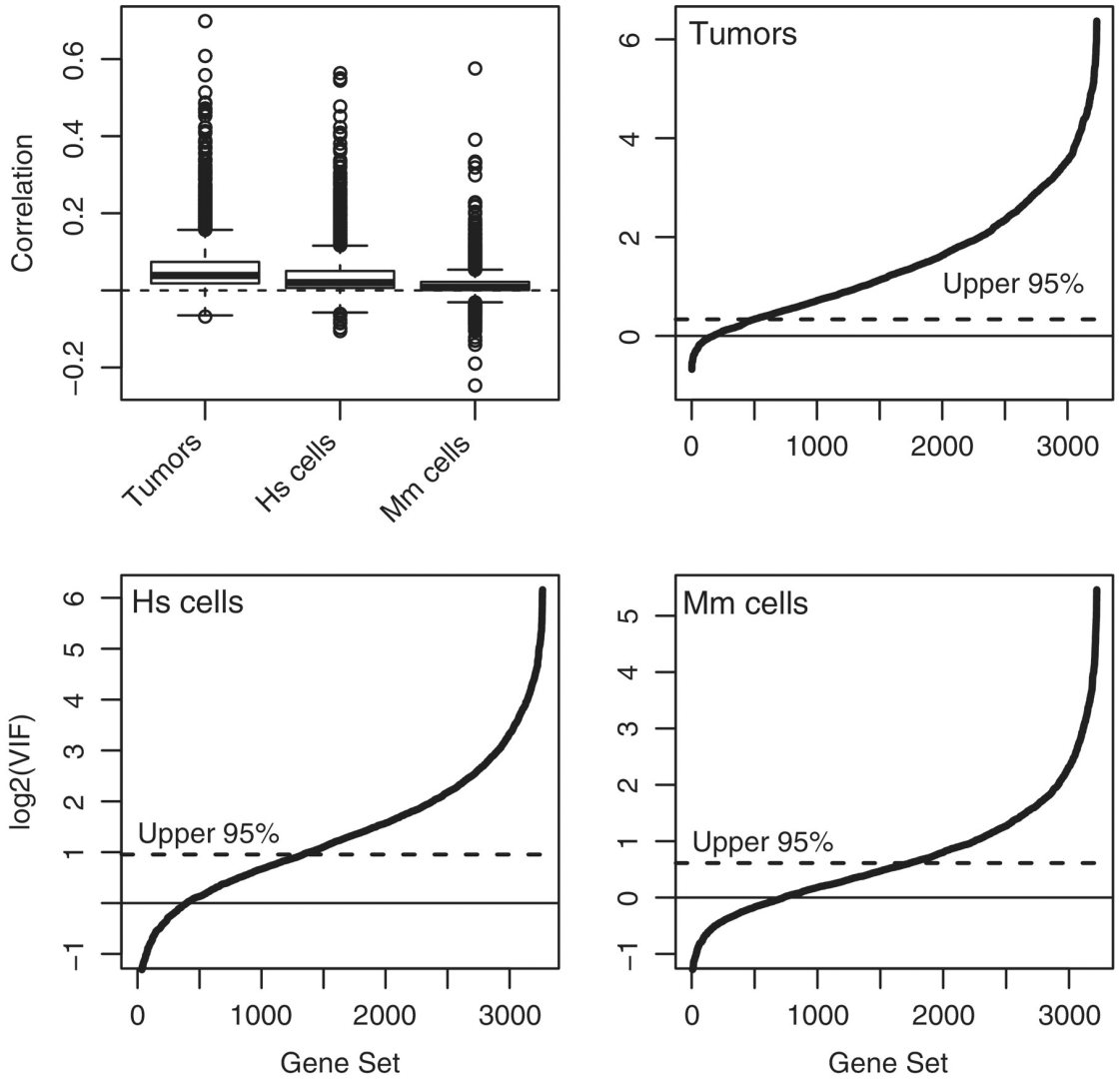

Figure : Distribution of VIFs in MSigDB sets

CAMERA: Proper False-positive Control

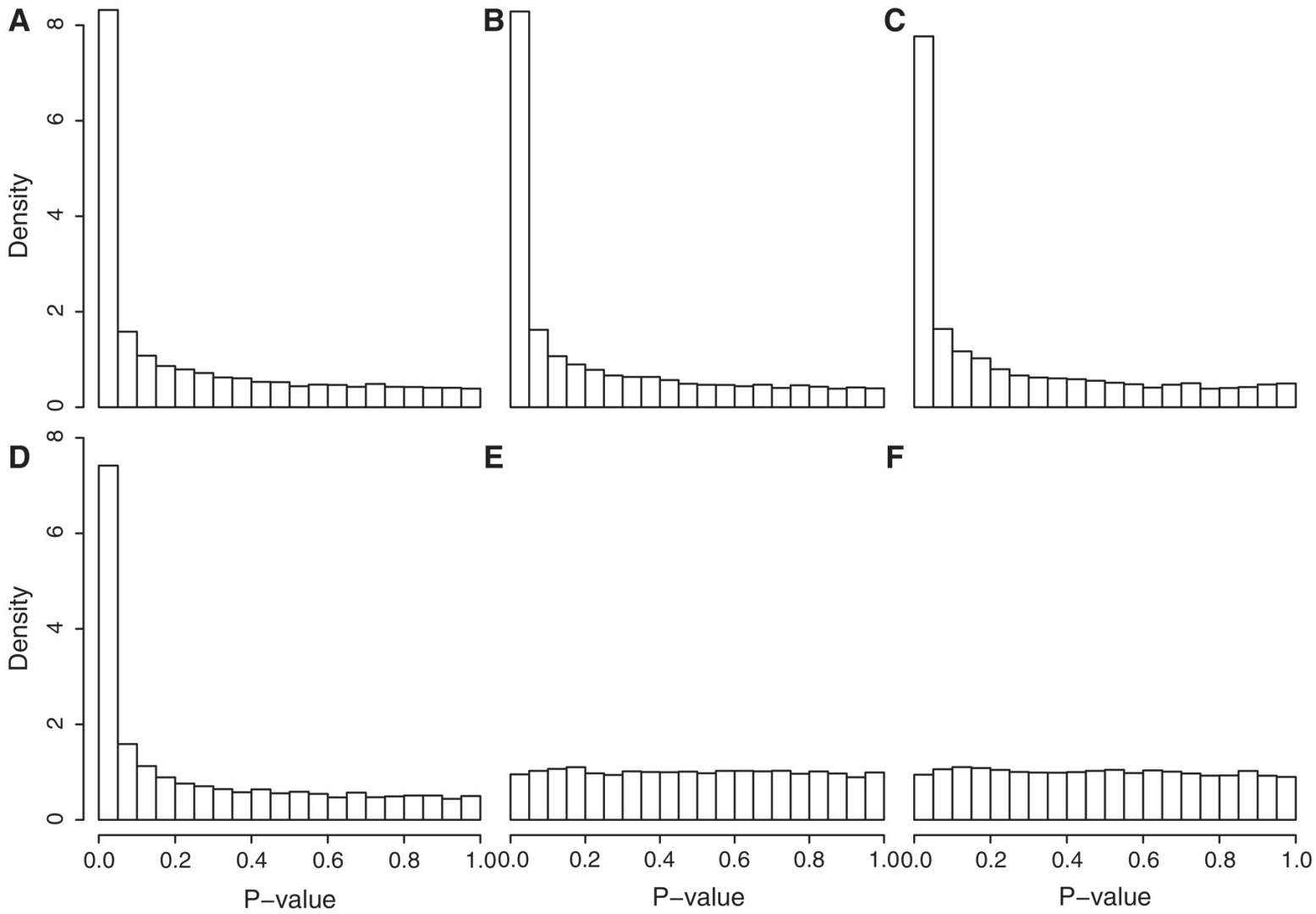

Figure : False positive control in CAMERA (A=t-test; B=Wilcox; C=sigPathway; D=PAGE; E=CAMERA t-test; F=CAMERA Wilcox)

SPIA: Signal Pathway Impact Analysis

- A pathway test that operates on a graph representation of a pathway
- Includes both activating and inhibitory interactions between genes
- Performs an ordinary test for enrichment (Q1, P_{NDE}) and a second test for "pathway perturbation", which tests for multiple gene effects accumulating along a path through the graph (P_{PERT})
- The two tests are independent, so we multiply the p-values to get combined p-value (P_G)
- This is the only pathway test worth running in my experience

SPIA: Testing "Pathway perturbation"

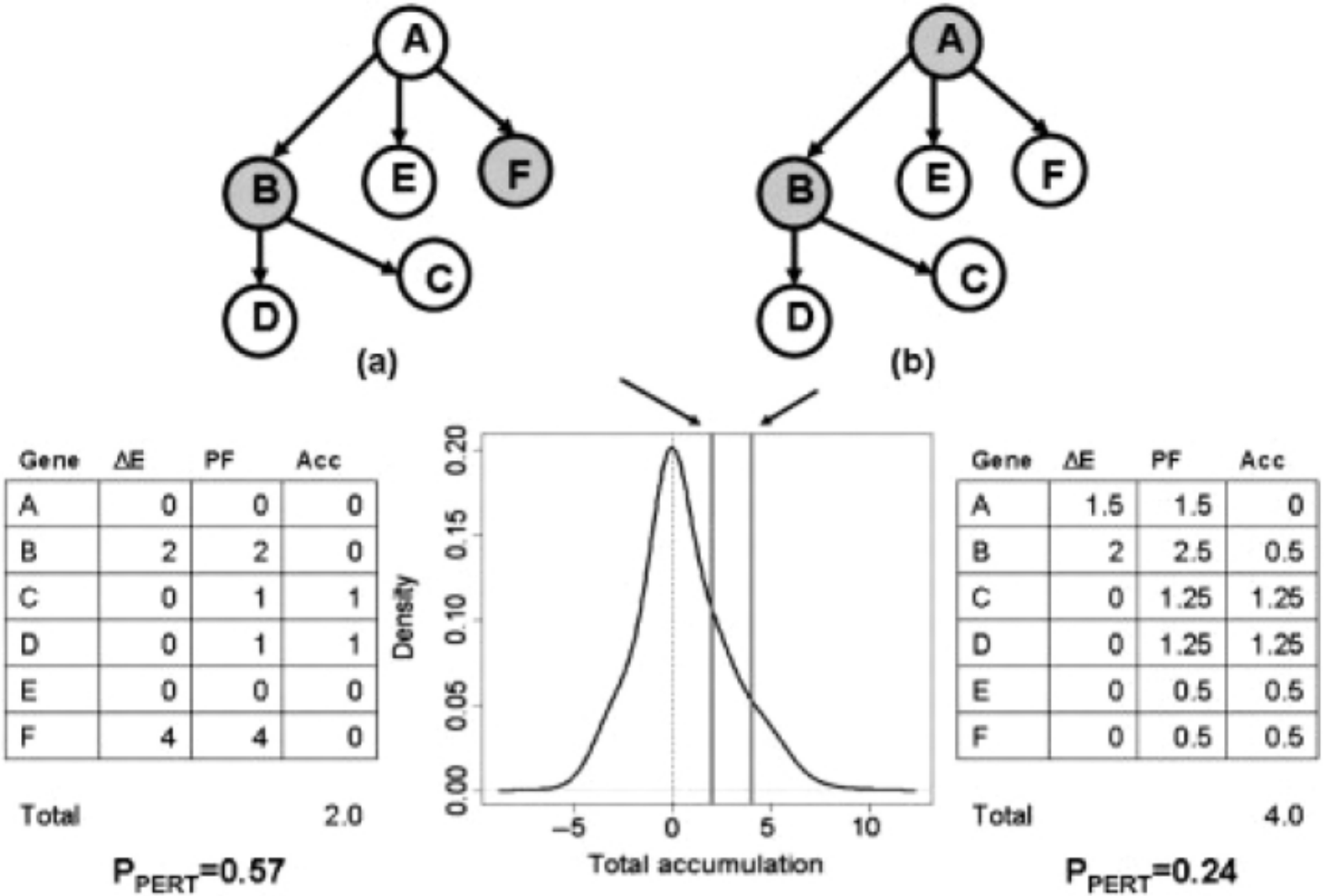

Figure : Using topological information

SPIA: Independence of the two tests

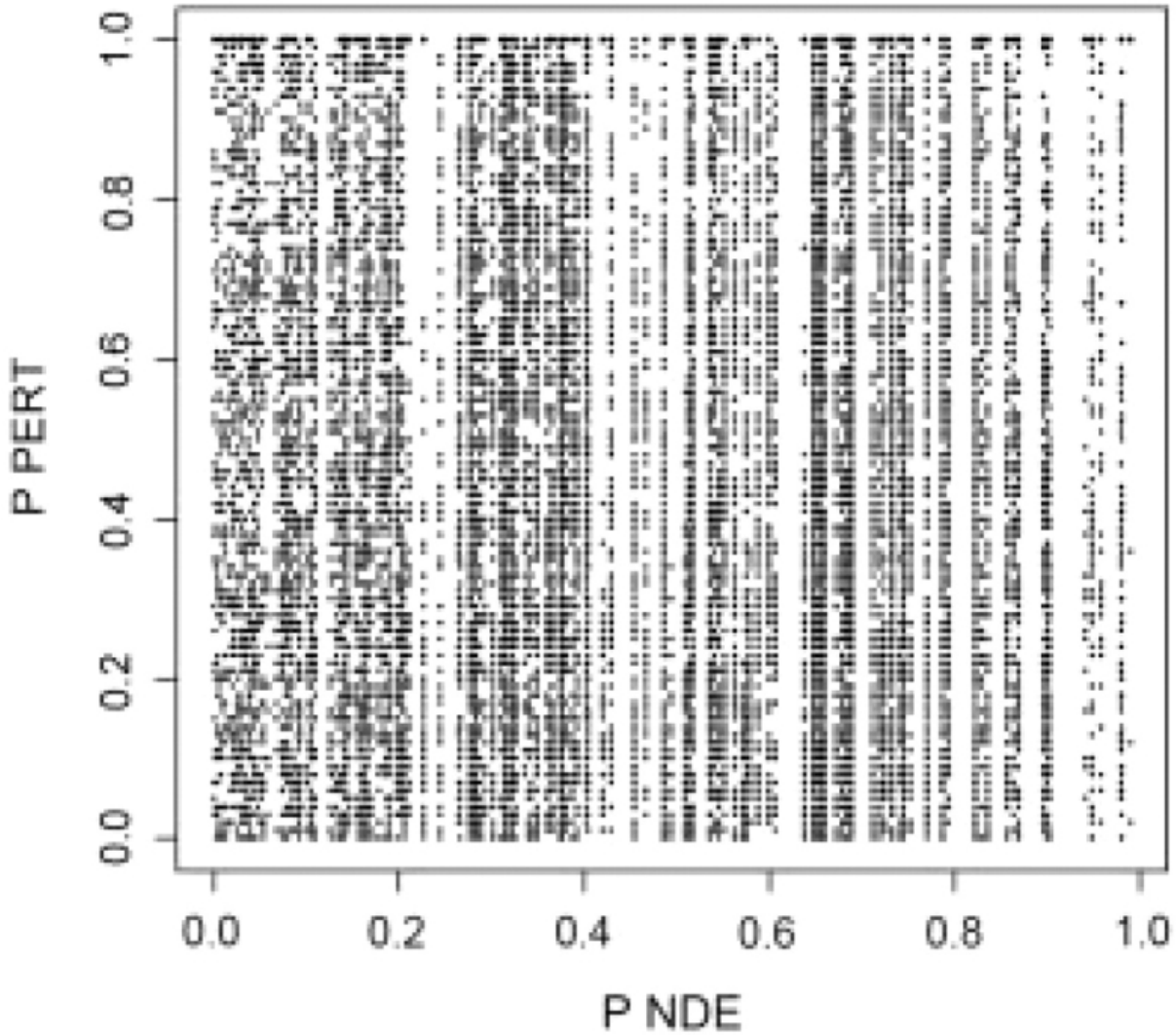

Figure : P_{NDE} vs P_{PERT}

SPIA: Combining p-values

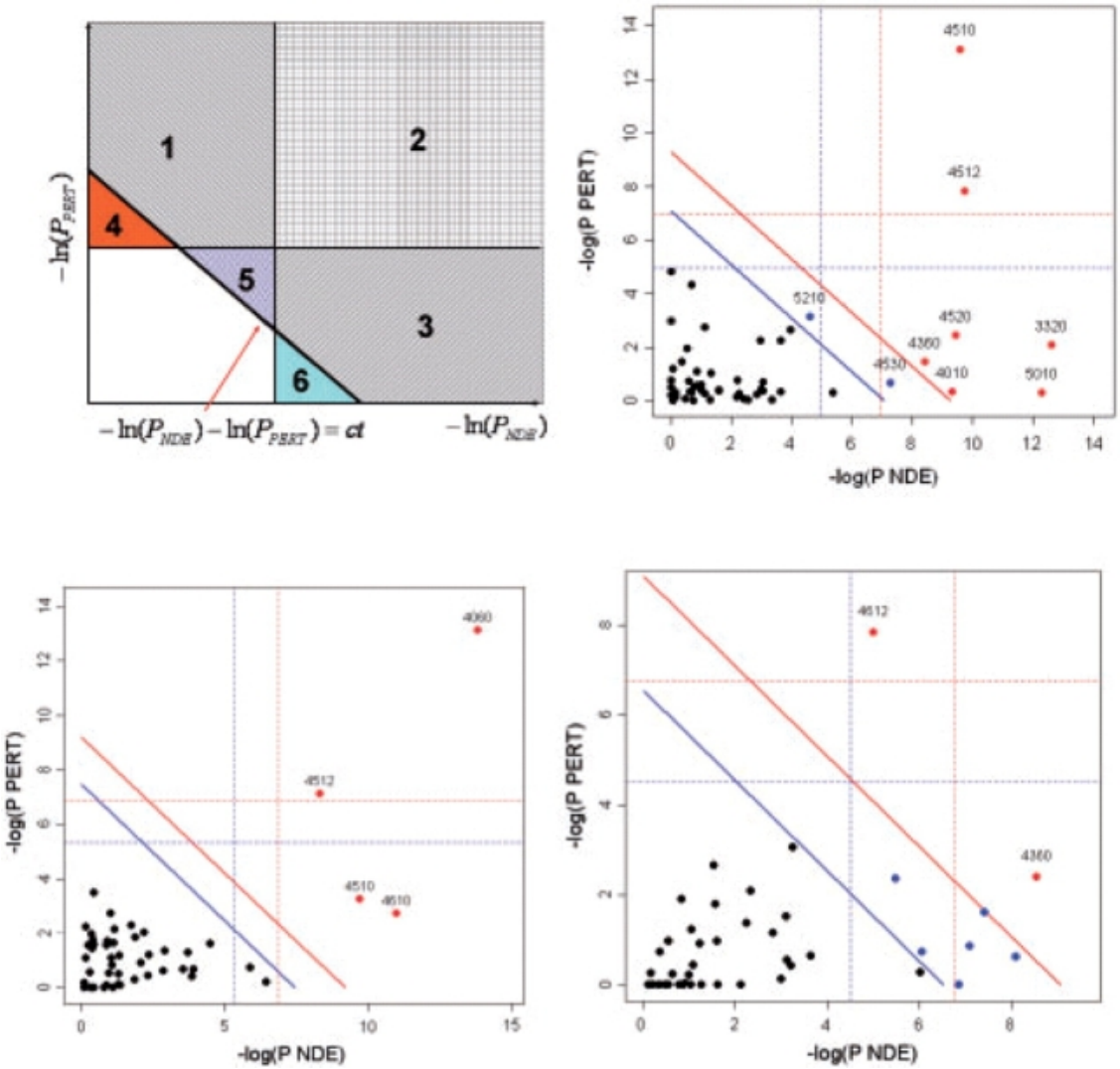

Figure : Combining P_{NDE} and P_{PERT}

SPIA: Test of false positive control

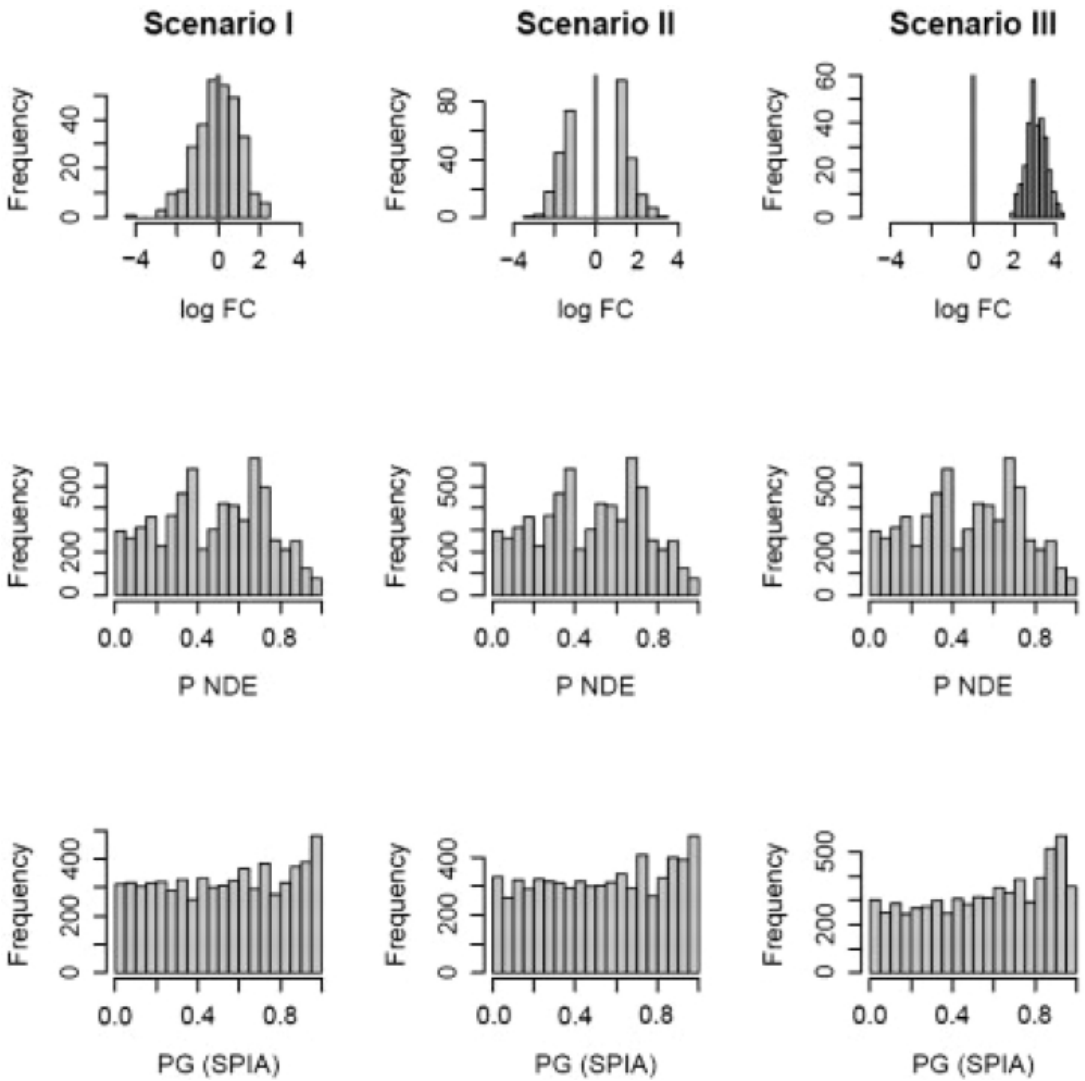

Figure : False positive control in SPIA

Results on Interferon gamma treatment data set

- IFNG is an immune signaling molecule
- It is known to upregulate its own receptor pathway, resulting in a positive feedback loop
- This data is from cynomolgus monkeys
- Genes were mapped to human orthologs in order to assign them to gene sets & pathways
- Let's open the results file

Acknowledgments

- Bioconductor, the source of all the packages implementing these tests
- Dan Salomon, my PI, for guidance & support

Any Questions?