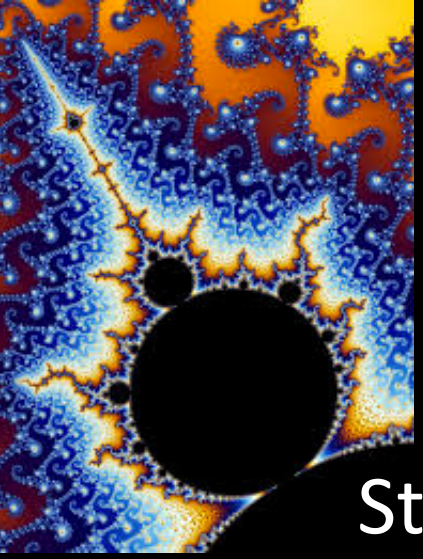
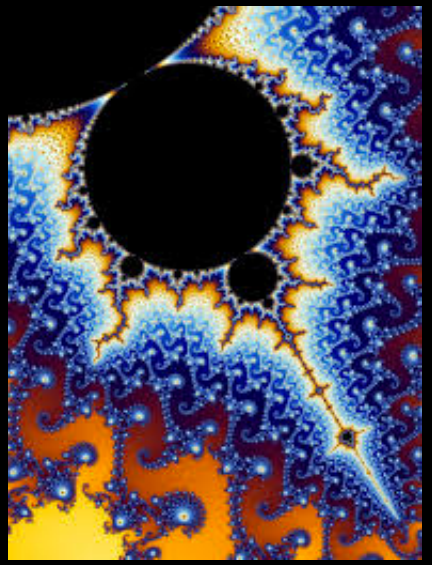


Statistical Interpretation of Inherently Structured Data

Michelle N. Archuleta Ph.D.

Previous Eisai & Broad Institute of MIT

Present MapQuest



Feature	Observation								
	N1	N2	N3	N4	N5	N6	N7	N8	N9
Color	yellow	yellow	green	yellow		green		yellow	green
Roundness	flat	round	flat		flat	round		flat	round

Yellow balls are flat!!

Nice problem:

- Features are **interpretable**.
  We understand color and roundness
- More **observations** than features.
- **No** inherent **correlation** btw features

Feature	Observation								
	N1	N2	N3	N4	N5	N6	N7	N8	N9
zz1									
zz2									
.									
.									
.									
.									
zz20000									

Not so nice problem:

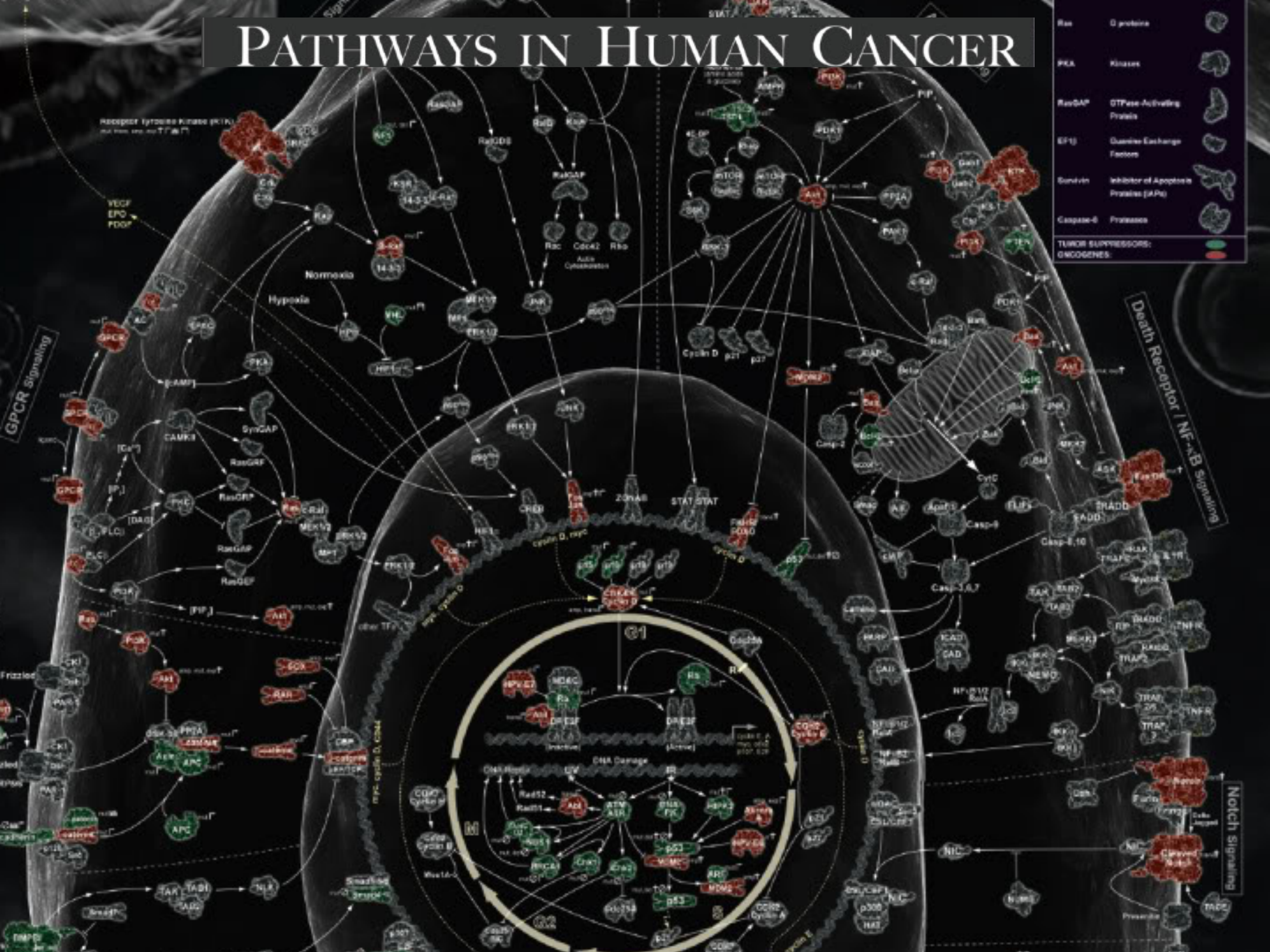
- Features are *un***interpretable**.
  No idea zz8 means to zz93
- More **features** than observations.
- And what if there are deep relationships within this feature set

What relationships could be underlying the data?

Human Embryonic Stem Cells

Image Ref www.wisegeek.com

Pathways in Human Cancer



When the experts don’t know?

Feature

p53

Met

.
.
.
.

NFkB

zz20000

Take this to research oncologist or immunologist..
Response this makes no sense

The data scientist is an artist and must provide interpretable context for the data

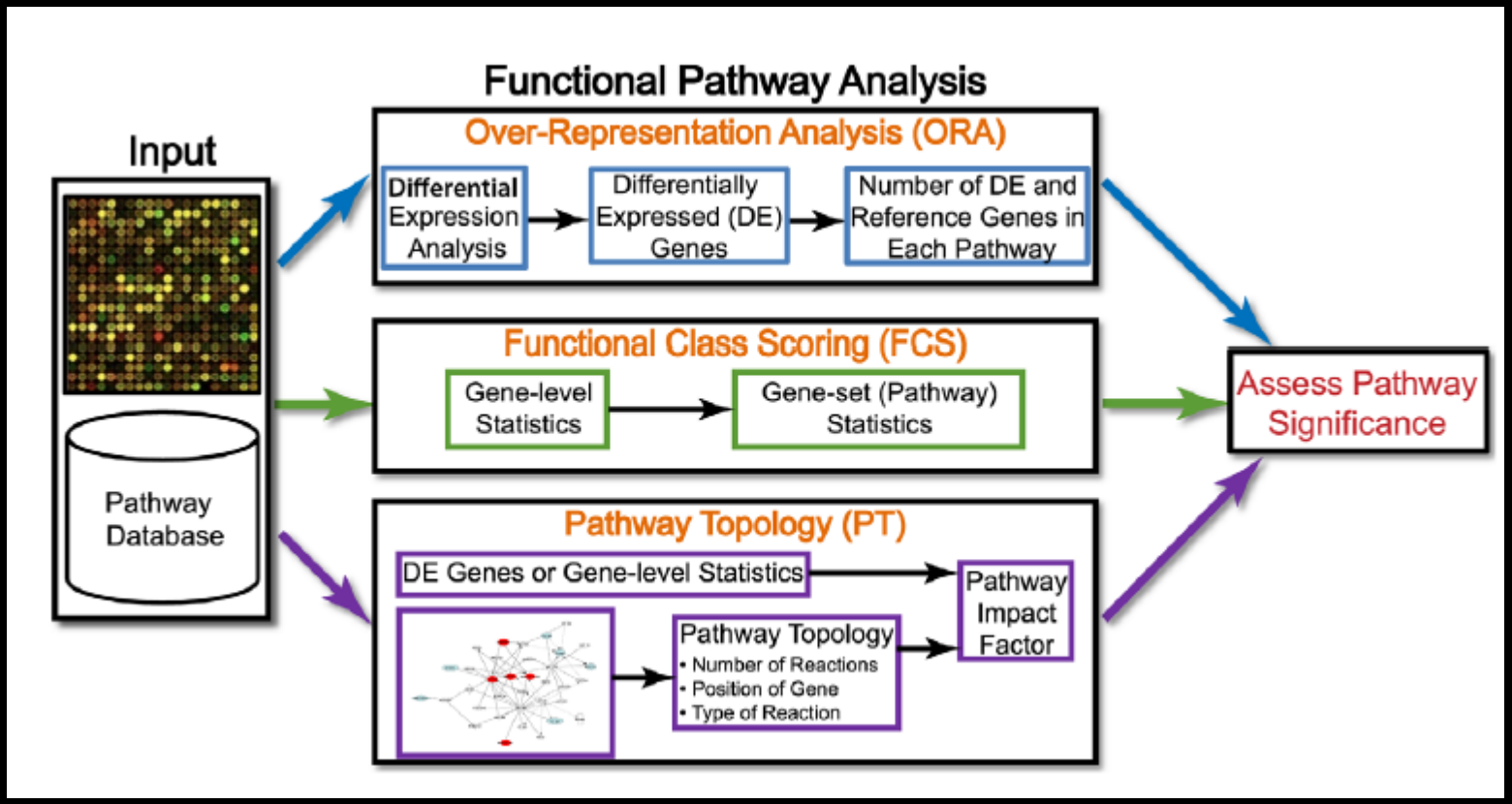
Pathway analysis methods

Review

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

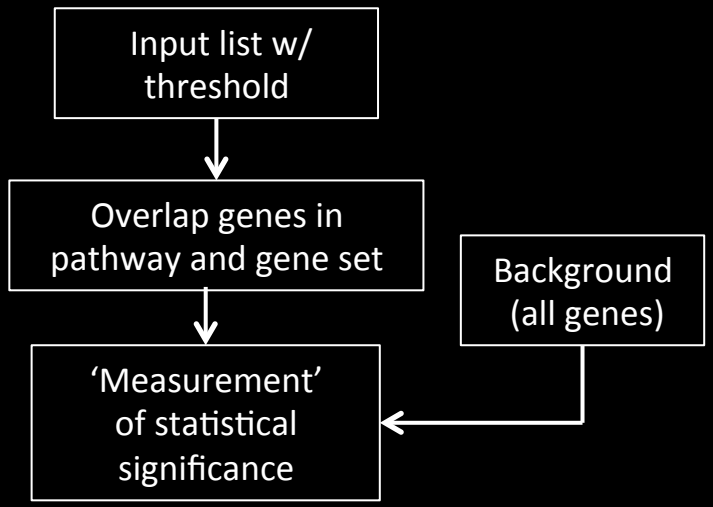
Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States of America, 2 Lucile Packard Children's Hospital, Palo Alto, California, United States of America



Khatri et al. PLoS Comp Bio 2012

Over-Representation Analysis (ORA)



What does it do? Evaluates the fraction of genes in a particular pathway

What measurements are used?: Hypergeometric, chi-square, or binomial distribution

What are the limitations?

1. The 'measurement' of **significance is independent of the measured changes.** Ignores probe intensities.

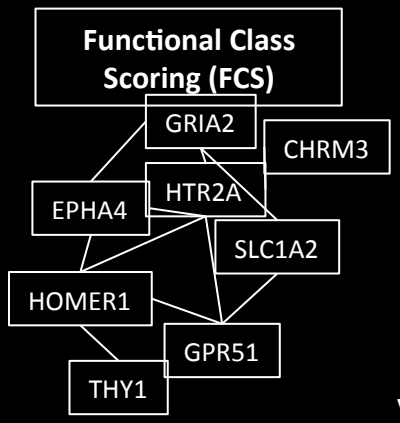
2. Uses **only the most significant genes** and discards all others**.** Marginally less significant genes fold change = 1.999 or p-value = 0.51 disregarded.

Few “Mountains” many “hills”
1st to 2nd generation pathway analysis

Hypothesis: Although large changes in individual genes can have significant effects on pathways**, weaker but coordinated changes in** sets of functionally **related genes** (i.e. , pathways) can also **have significant effects**

Over-Representation Analysis (ORA)

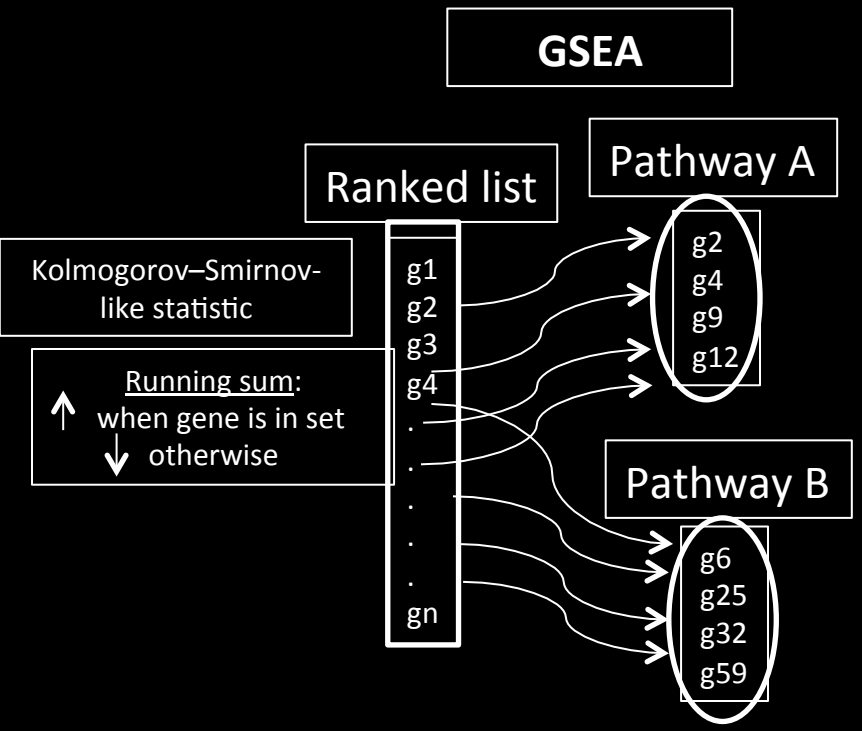
CHRM3
GRIA2
NRGN
SLC1A2
HOMER1
EPHA4



Khatri et al. PLoS Comp Bio 2012

Volgelstein et al. Science 2013

Gene Set Enrichment Analysis



Null distribution: Randomly shuffle phenotype labels

1. **Calculate an Enrichment Score** Increasing a running-sum statistic when encounter a gene in S and decreasing it otherwise

2. **Estimate the significance** Level of ES. Pathway distribution vs. Null distribution (self-contained)

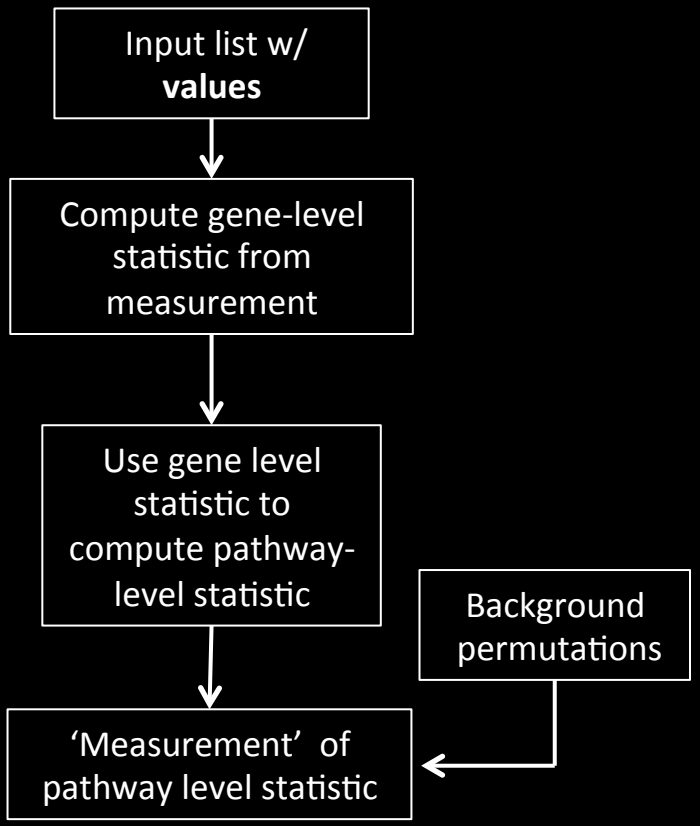
3**. Adjustment for Multiple Hypothesis Testing** Compute normalized ES (NES) for each gene set then determine FDR for NES.

Novelty of GSEA method comes from **thinking of genes in terms of sets** or distribution instead of lists.

Mootha et al. Nat Genet 2003

Subramanian et al. PNAS 2005

Functional Class Scoring (FCS)



What does it do? Evaluates the distribution of genes in a pathway that are differentially expressed

What measurements are used?:

Gene Level statistic: 1) Univariate: ANOVA, Q-statistic, signal-to-noise ratio, t-test, and Z-score. 2) Multivariate: GlobalANOVA, and Hotelling T^2.

Pathway Level statistic: Kolmogorov-Smirnov statistic, sum, mean, or median of gene level statistic, the Wilcoxon rank sum, and the maxmean statistic.

How is statistical significance determined? Compute the null distribution:

1) Competitive null hypothesis permutes **gene labels** for each pathway, and compares the set of genes in the pathway with a set of genes not in the pathway.

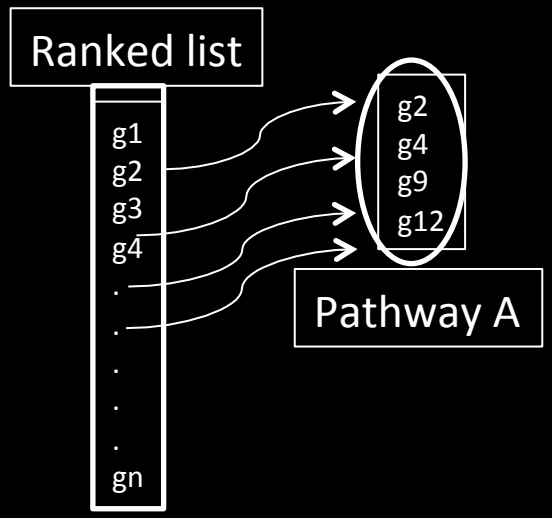
2) Self-contained null hypothesis permutes **class labels** for each sample and compares the set of genes in a given pathway with itself.

Khatri et al. PLoS Comp Bio 2012

Univariate vs. Multivariate FCS

GSEA

Univariate



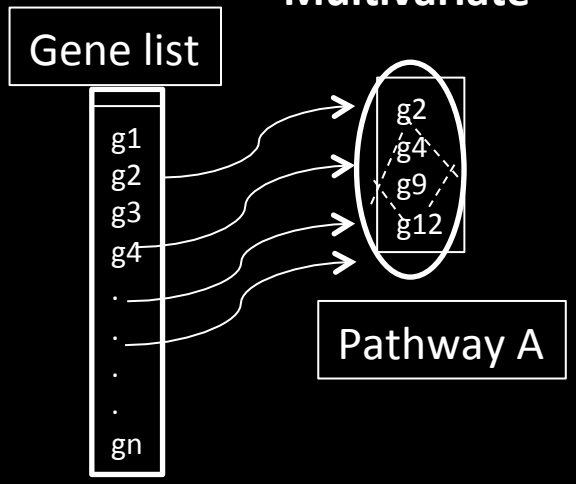
↑ Running sum: ↓when gene is in set otherwise

Doesn't consider gene set correlations

Subramanian et al. PNAS 2005

PCOT2

Multivariate



Hotelling's T^2 statistic

$$T^2 = \frac{n_1 n_2}{n}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^t \mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2),$$

w/ pooled covariance matrix S

$$\mathbf{S}_i = 1/(n_i - 1)\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^t$$

$H_0 : \bar{\mathbf{X}}_1 = \bar{\mathbf{X}}_2,$ Sampling distribution of T^2 will follow

$$(n-2)q/(n-q-1)F_{q,n-q-1}.$$

Kong et al. Bioinformatics 2006

Benefits and Limitations of FCS

- **What are the benefits over ORA?**
  1. They **do not require an arbitrary threshold** for dividing expression data into significant and non-significant pools**.**
  2. ORA completely ignores measurements when identifying significant pathways.
  3. Considering the coordinated changes in gene expression, FCS methods **account for dependence between genes** in a pathway. ORA does not

- **What are the limitations?**
  1. FCS analyzes each **pathway independently.**
  2. Many FCS methods use changes in gene expression to rank genes in a given pathway and **discard changes from further analysis.**

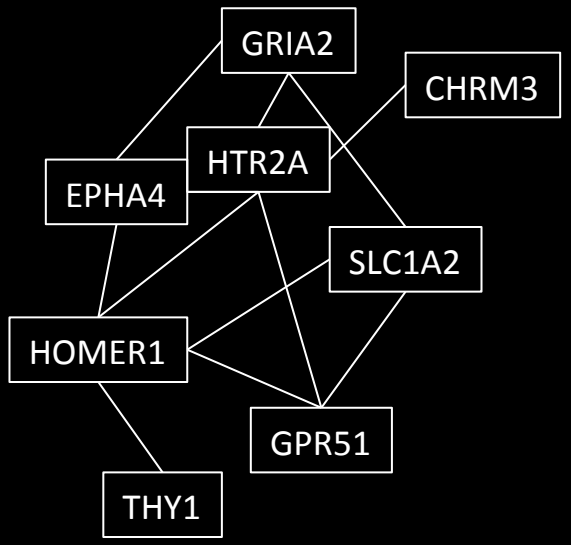
Khatri et al. PLoS Comp Bio 2012

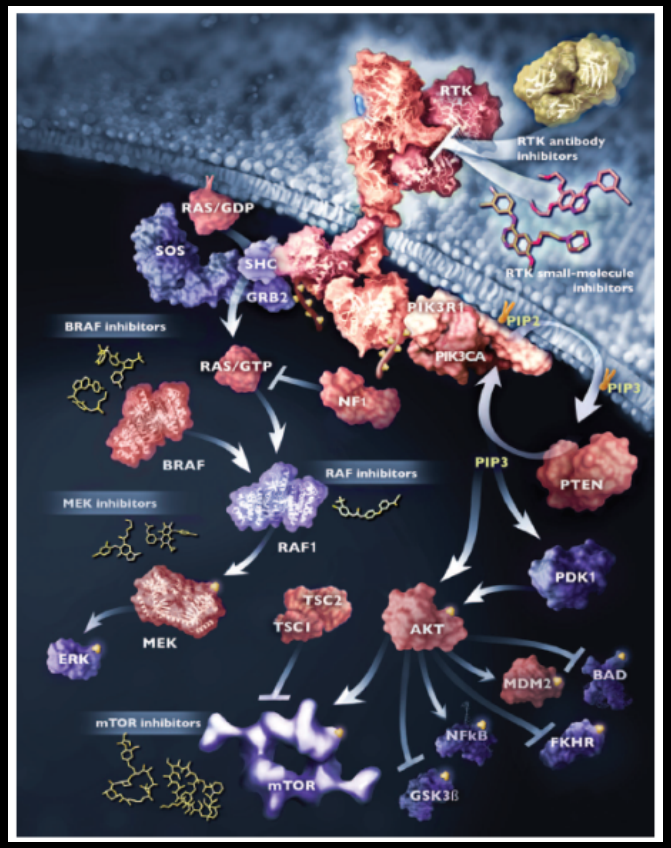
Leveraging Pathway Structure

Functional Class Scoring (FCS)

↓

Coordinated Gene Interactions

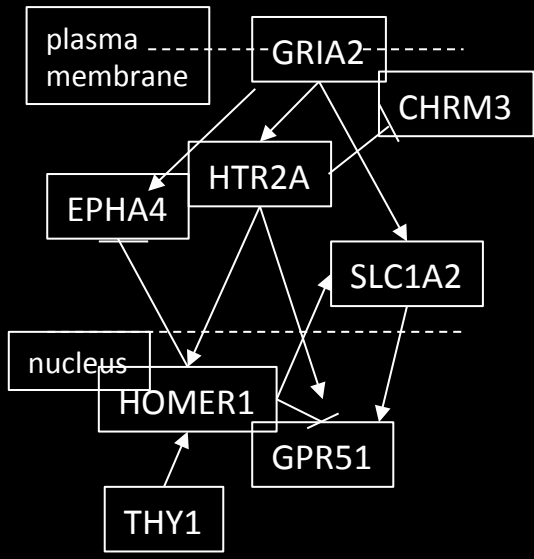




Pathway Topology (PT)

↓

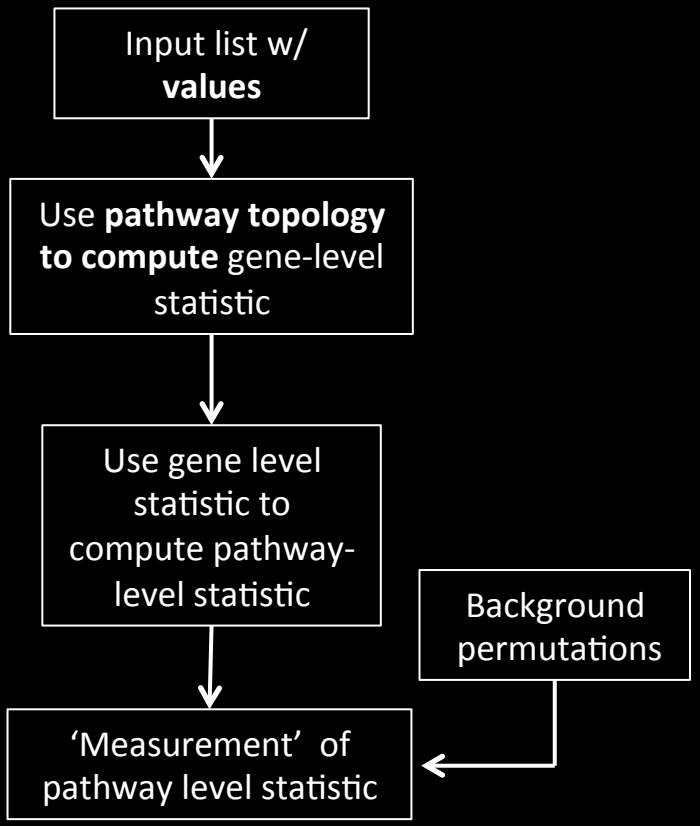
Type and Localization Interaction



Hypothesis: Knowledge bases providing information about **gene product interactions, type of interaction** (e.g., activation, inhibition), and **where they interact** (e.g., cytoplasm, nucleus) could **be leveraged** in pathway analysis.

Khatri et al. PLoS Comp Bio 2012

Pathway Topology (PT)



What does it do? Measures significance of gene level interactions with respect to pathway topology

What measurements are used?:

Gene Level statistic: 1) ANOVA, Q-statitic, signal-to-noise ratio, t-test, Z-score,

Pathway Level statistic: Univariate/Multivariate, disregards/ considers gene dependences Univariate: sum, mean, or median of gene level Multivariate: Global ANOVA, Hotelling T2, Kolmorgorov-Smirnov statistic

How is statistical significance determined? Compute the null distribution:

1) Competitive null hypothesis permutes **gene labels**.

2) Self-contained null hypothesis permutes **class labels**.

Khatri et al. PLoS Comp Bio 2012

Signaling pathway impact analysis (SPIA)

Combines two metrics:

1) Overpresentation of DE features in pathway
2) Abnormal perturbation of pathway

$$P_{NDE} = P(X \geq N_{de} \mid H_0)$$

Signed normalized Change of expression

Sum of perturbation factors Upstream of target gene

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

Strength of interaction Btw feature I and feature j

Number of downstream Features

$$Acc(g_i) = PF(g_i) - \Delta E(g_i)$$

Total net accumulated perturbation

$t_A = \sum_i Acc(g_i)$

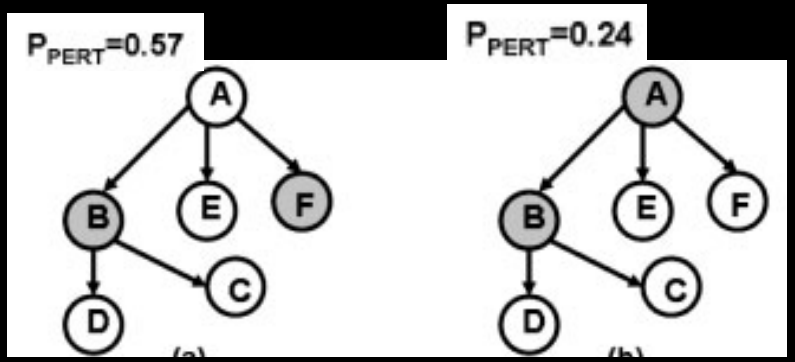
Total accumulated perturbation

T_A,

$$P_{PERT} = P(T_A \geq t_A \mid H_0)$$

Probability to observe total accumulated probability

Probability: bootstraping same number of features are allowed to occupy any position in pathway



Benefits & Limitations of PT

- What are the benefits over FCS?
    1. The **structure of the network and types of interactions** in the network are included in the pathway analysis**.**
        - FCS methods **only consider the number of genes** in a pathway and ignore additional information.

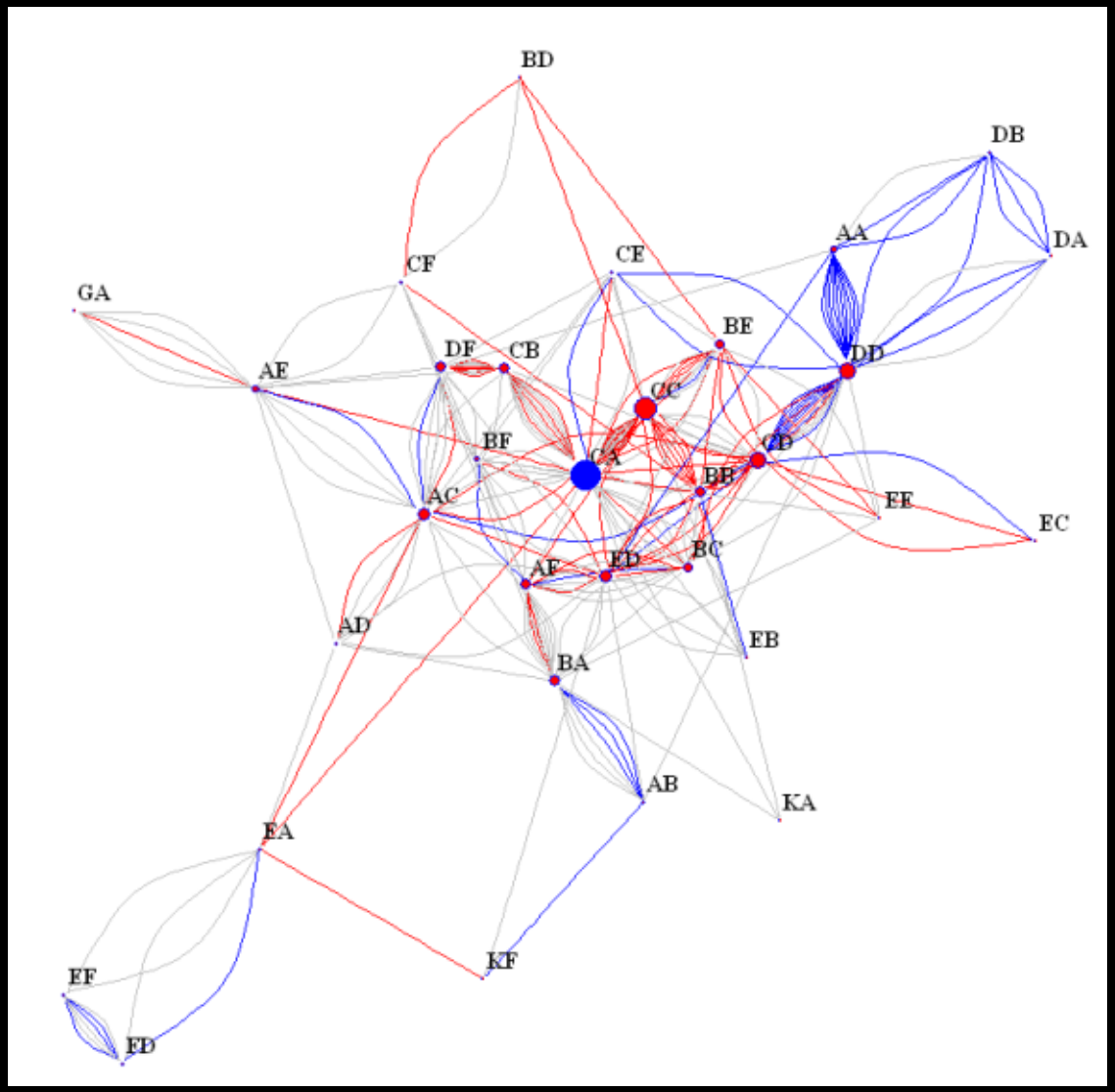
- What are the limitations of PT?
    1. True pathway topology is dependent on the type of cell.
        - Knowledge with regard to **cell type** and conditions being studied are typically **unavailable.**
    2. **Inability to model dynamic states** of the system and inability to consider the interactions between pathways.

Outstanding Challenges in Pathway Analysis

- Annotation Challenges
  1. Low resolution knowledge bases
  2. Incomplete and inaccurate annotations
  3. Missing condition and cell-specific information

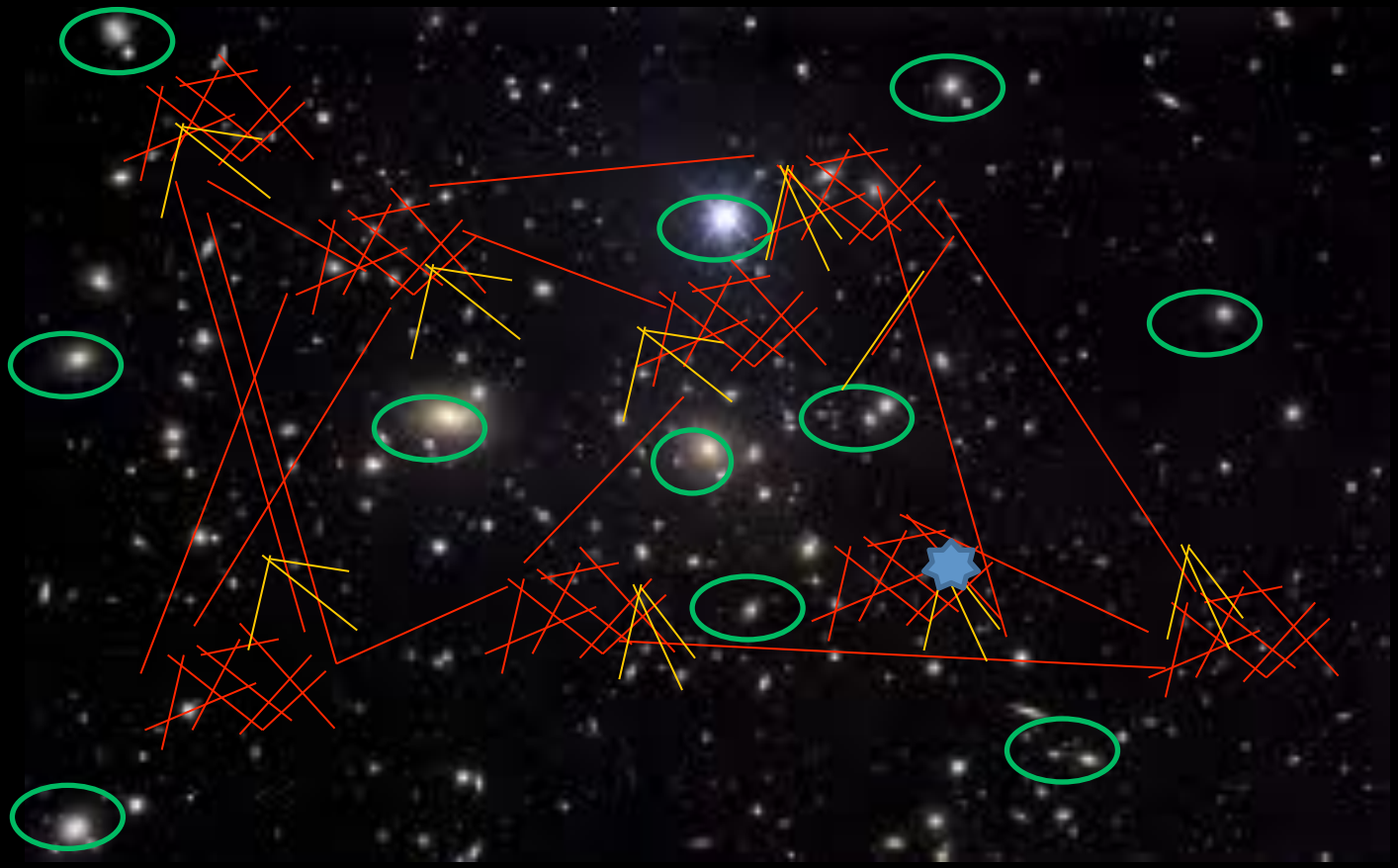
- Methodological Challenges
  1. **Benchmark data sets for comparing different methods**
  2. Inability to model and analyze dynamic response.
  3. Inability to model effects of an external stimuli

Visualizations of pathway analysis



Bioconductor pkg Igraph

Galaxy of Differential Expression



ORA P-value 0.12

FCS P-value 0.04

PT P-value 0.0016

Biomarker