3000788 Intro to Comp Molec Biol

Lecture 12: Functional enrichment analysis

September 22, 2022

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Overrepresentation analysis

Enrichment fold

Gene group	Kinase	Not kinase	Total
Differentially expressed	50	350	400
Not differentially expressed	150	5450	5600
Total	200	5800	6000

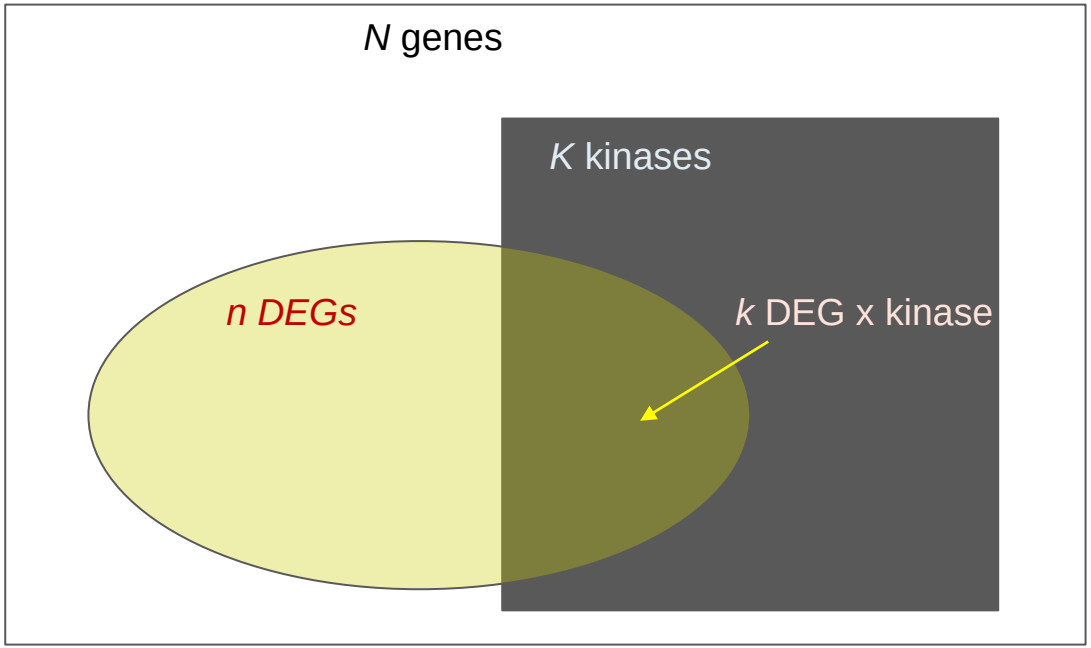
- There are 200 kinases among 6000 genes
- Expected 400 x 200 / 6000 = 13 kinases to be differentially expressed
- Enrichment = 50 / 13 = 3.85 folds

Fisher's Exact Test

Gene group	Kinase	Not kinase	Total
Differentially expressed	$k \geq 50$	$400 - k$	400
Not differentially expressed	$200 - k$	$5400 + k$	5600
Total	200	5800	6000

- P-value for this observation = P(Kinase & DE ≥ 50)
- P(Kinase & DE = k) = Hypergeometric(N = 6000, K = 200, n = 400, k)

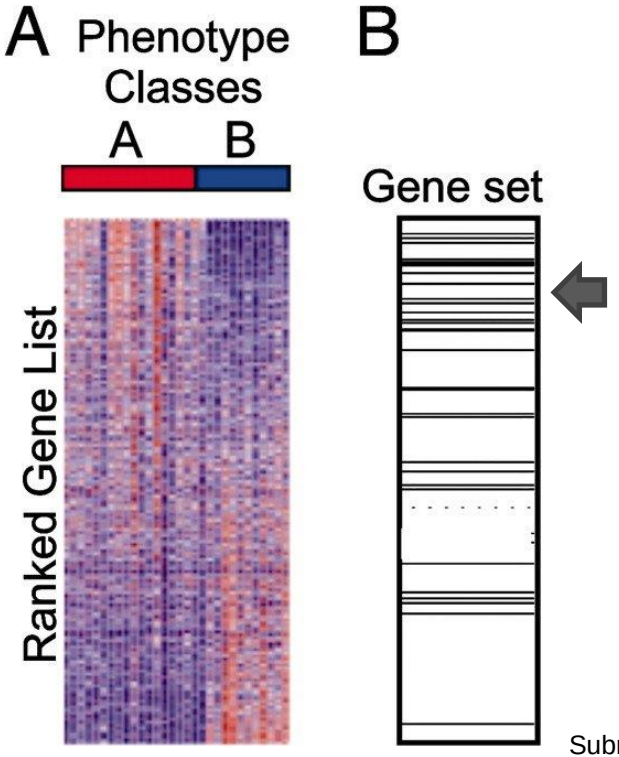
Hypergeometric distribution



- $\binom{K}{k}$ ways to select the intersected k genes
- $\binom{N-K}{n-k}$ ways to select the remaining $n-k$ non-kinase genes
- Total of $\binom{N}{n}$ ways
- Probability = $\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Gene Set Enrichment Analysis (GSEA)

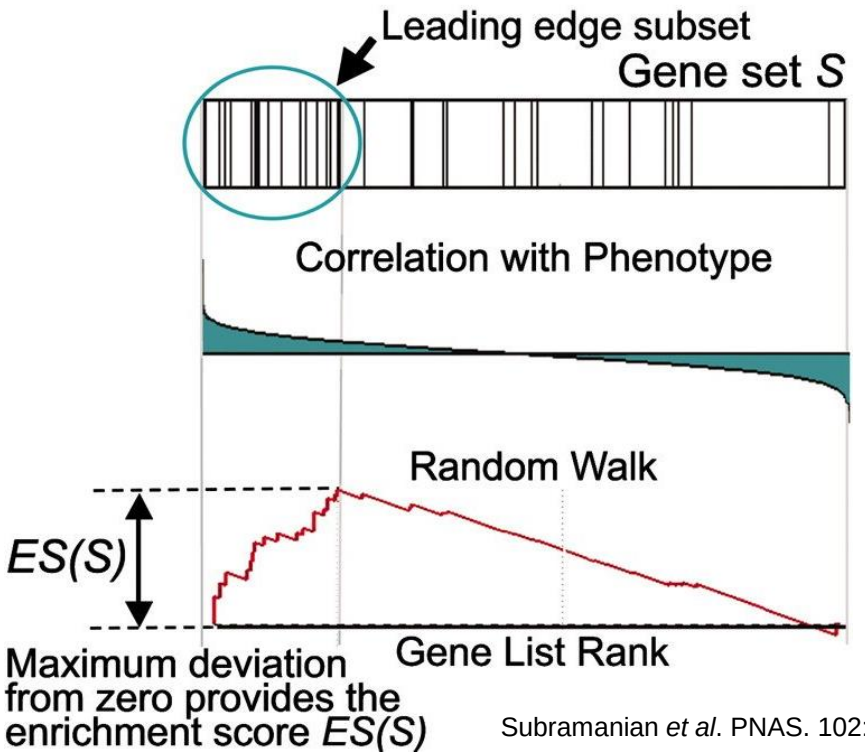
GSEA algorithm sketch



- Sort genes by the extent of up-/down-regulation across conditions
- Label genes annotated with a function
- If these genes are clustered together at the top, then this function is up-regulated
- If these genes are clustered together at the bottom, then this function is down-regulated

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

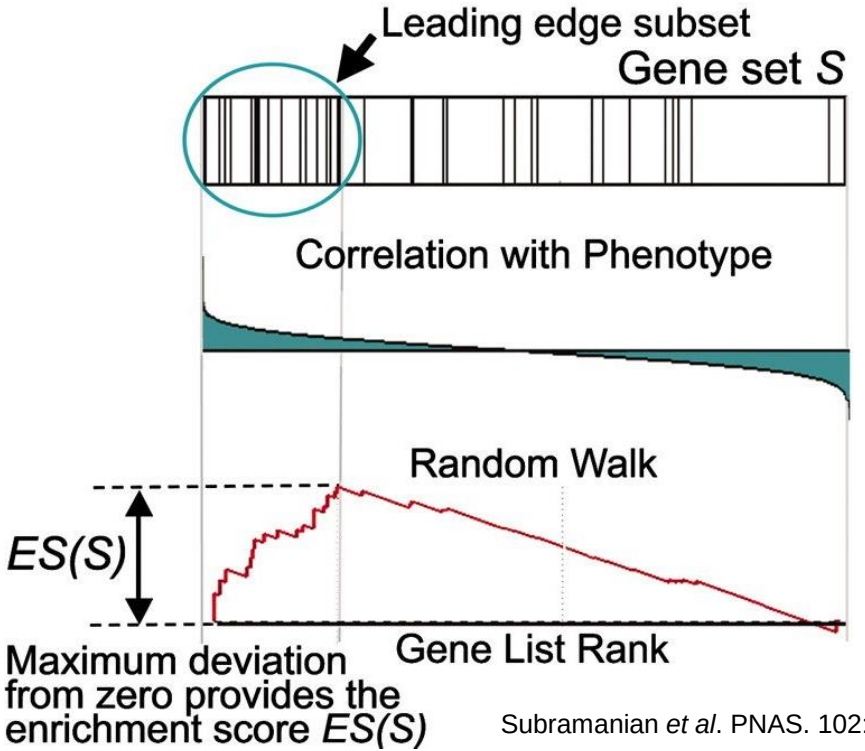
GSEA scoring



- Starting at score = 0 from the top of the sorted gene list
- If encounter gene from *S*, +score
- Otherwise, –score
- Score indicates the extent of up-/down-regulation
  - Correlation with conditions
  - Log fold-change

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

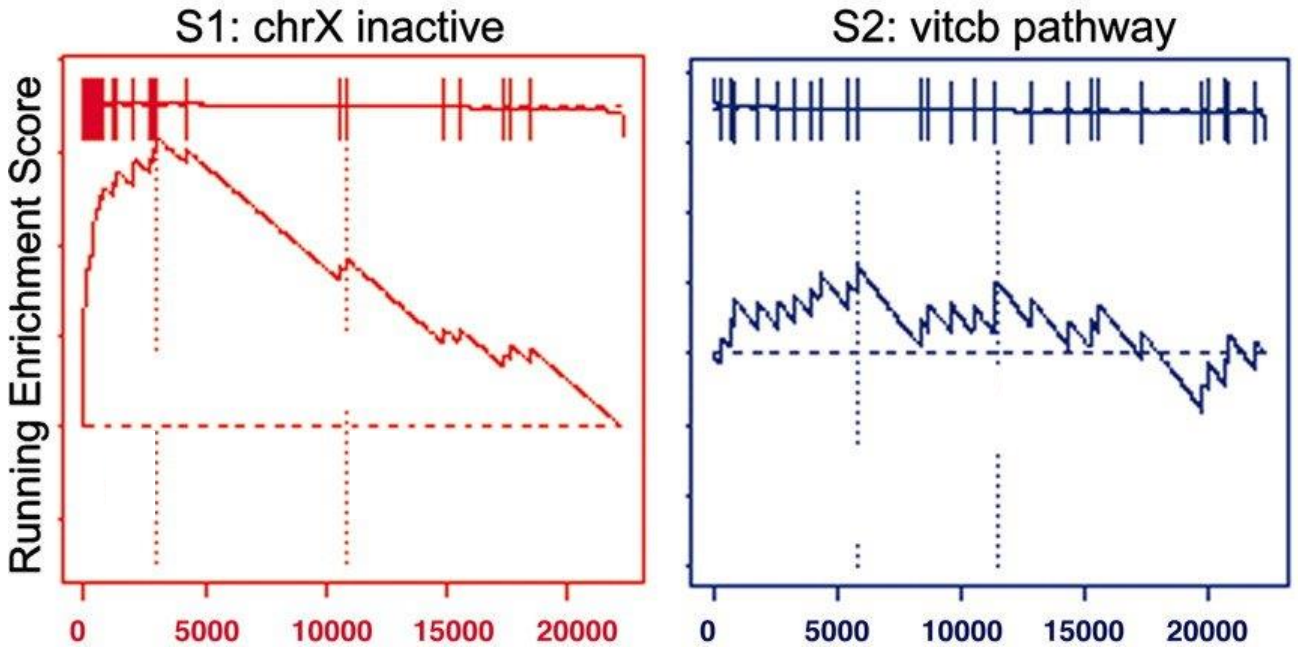
Interpretation of GSEA score



- High enrichment score indicates cluster of genes from function *S* at the top of the gene list
  - Up-regulation
- Low (negative) enrichment score indicates cluster of genes from function *S* at the bottom of the list
  - Down-regulation

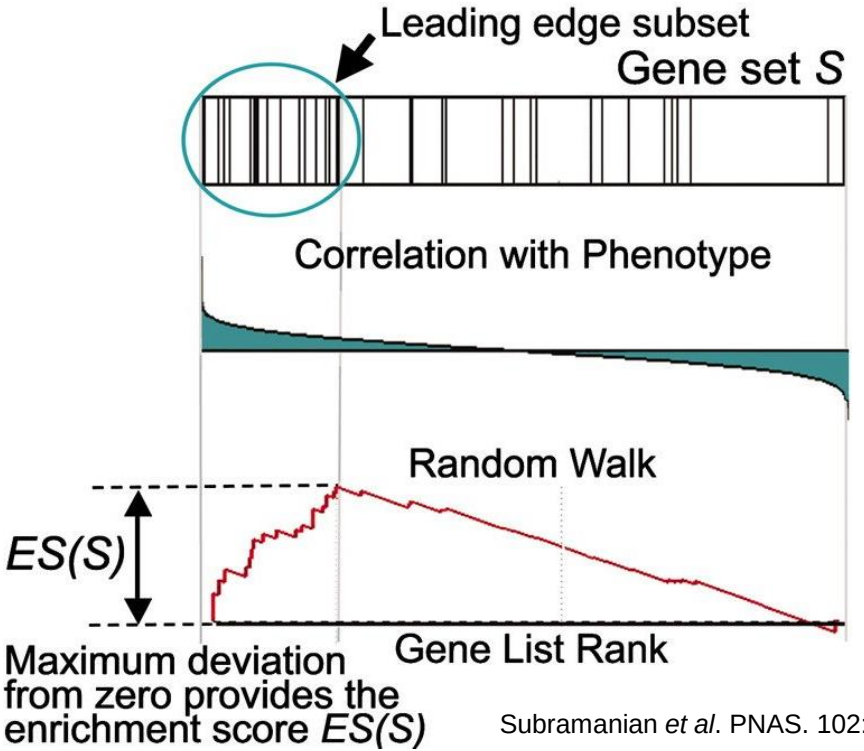
Subramanian *et al*. PNAS. 102:15545-15550 (2005)

Up-regulated versus unchanged pathways



Subramanian *et al*. PNAS. 102:15545-15550 (2005)

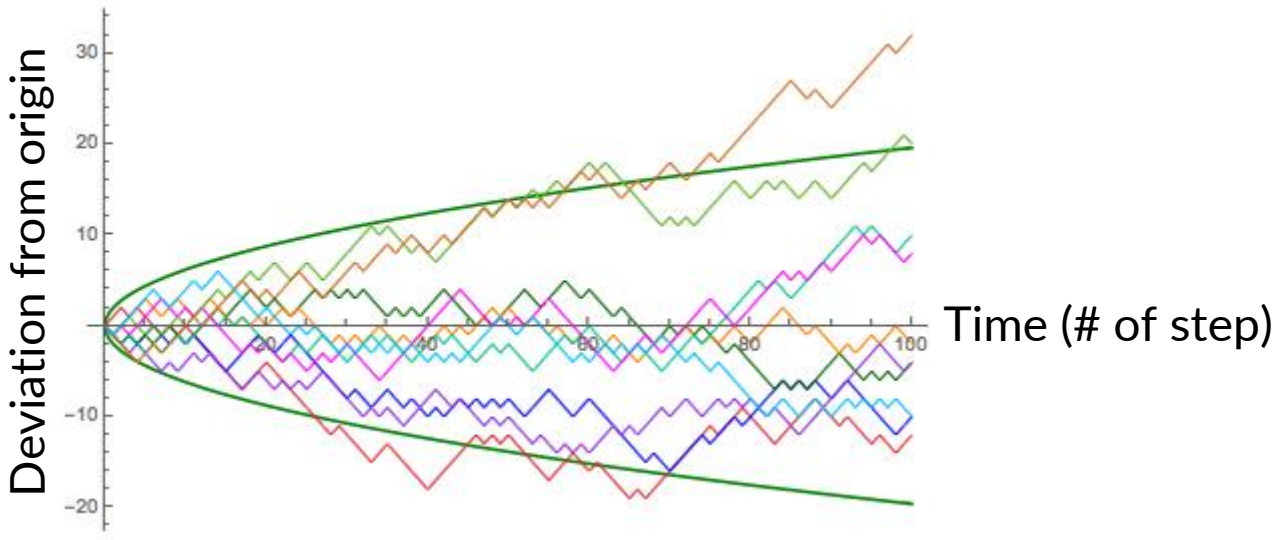
Null hypothesis for GSEA



- **Null hypothesis**: Genes from *S* are uniformly distributed in the list
- +score and –score are uniformly distributed in the list
- This is a Random Walk

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

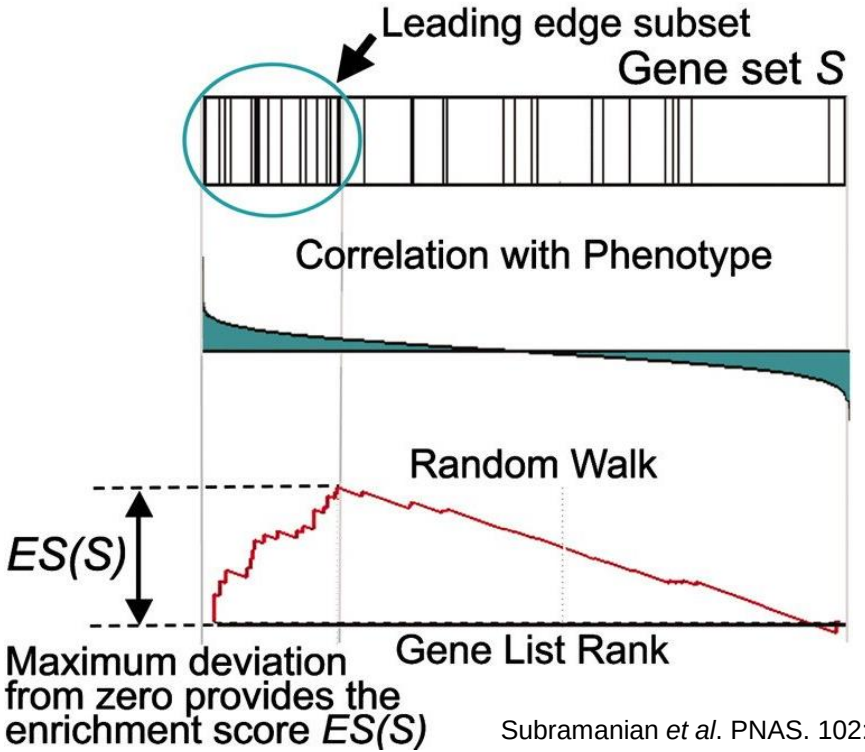
Statistical behaviors of random walks



https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/

- P(maximal deviation > d) $\approx 2\sum_{k=1}^{\infty}(-1)^{k-1}\, e^{-2(kd)^2}$

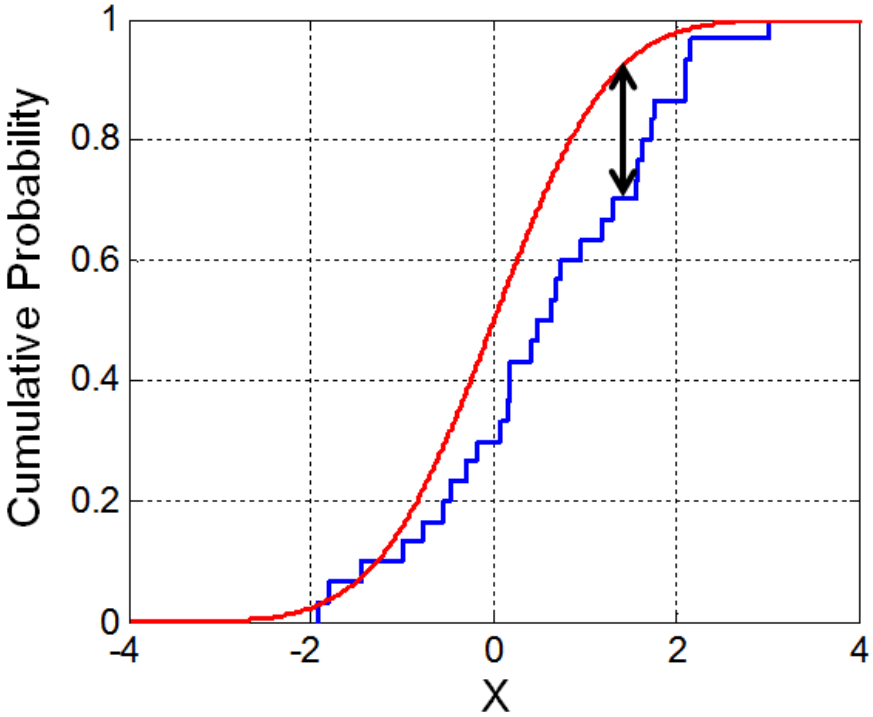
Statistical testing for GSEA



- Score ES(*S*) for a function *S*
- P-value = P(score ≥ ES(*S*)) under the random walk model
- Test of deviation from ES = 0
- Kolmogorov-Smirnov test

Subramanian *et al*. PNAS. 102:15545-15550 (2005)

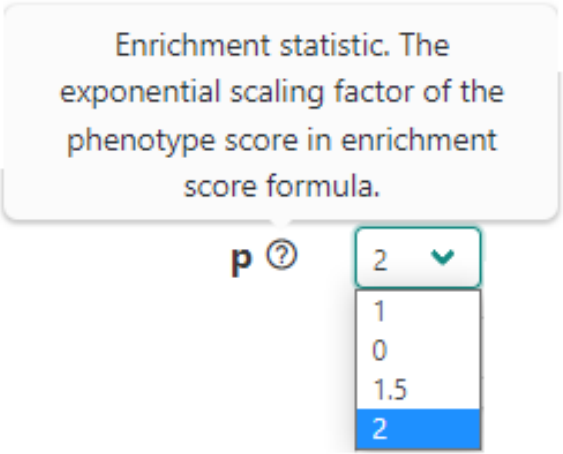
Kolmogorov-Smirnov test



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

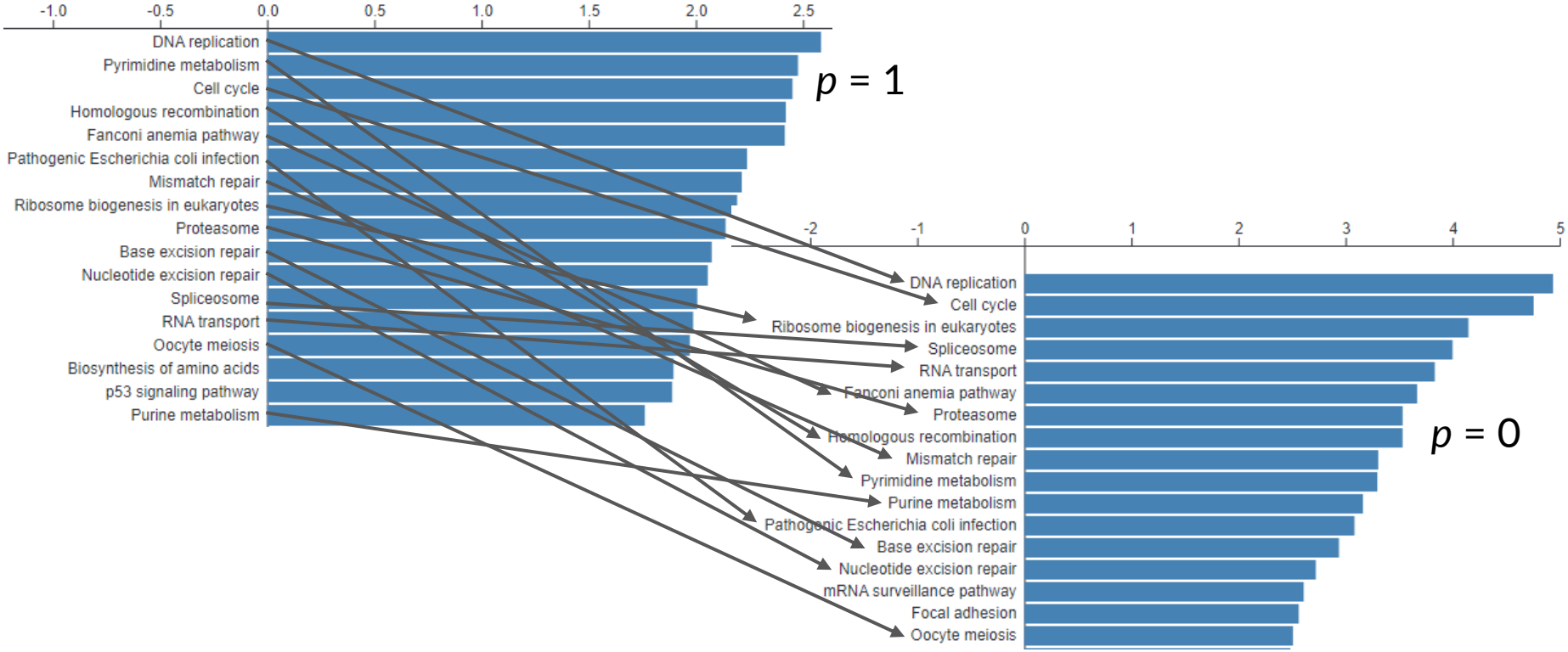
- Test whether two probability distribution are equal
- Compare cumulative density (red and blue trends)
- If they are equal, the two curves should stay close to each other
- **Null hypothesis**: random walk

Setting the score for GSEA



- Originally developed for microarray data
- Adapted to RNA-seq
  - Log fold-change
  - No score (simply rank genes)
- Weighted score = $(\text{score})^p$
  - Default: $p = 1$
  - No score: $p = 0$
  - More weights for top genes: $p > 1$

Comparing the impact of p = 0 and 1

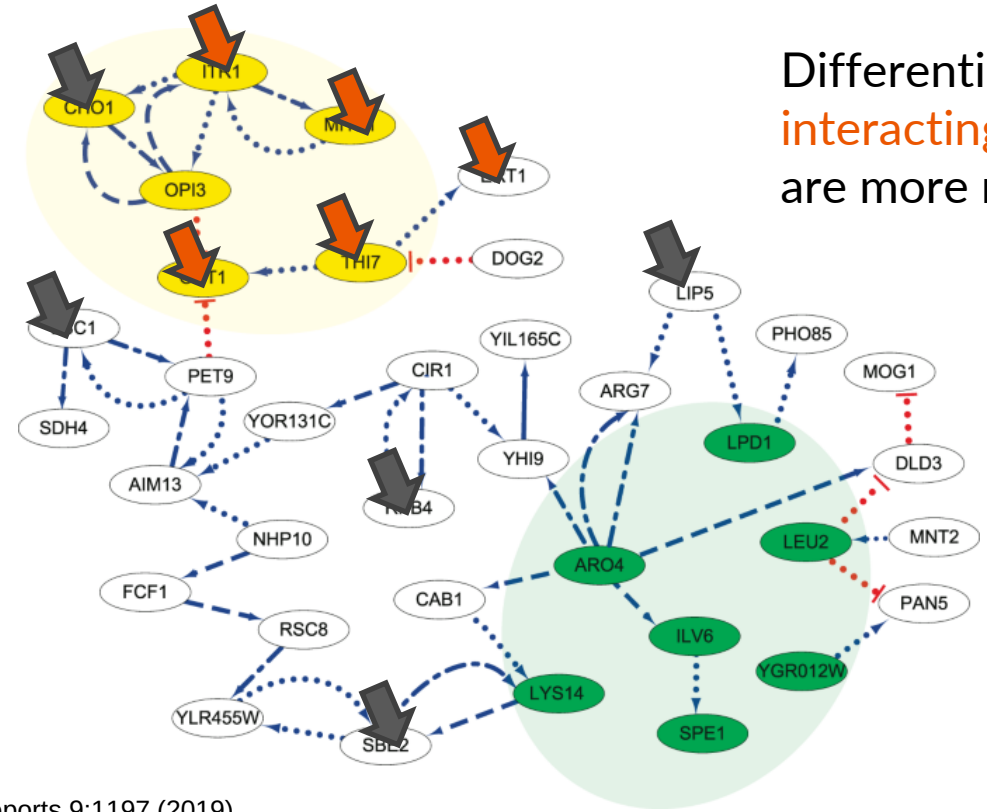


Pros and cons of GSEA

- No need to define p-value cutoff
- Identify both up- and down-regulated functions at once

Network topology-based analysis

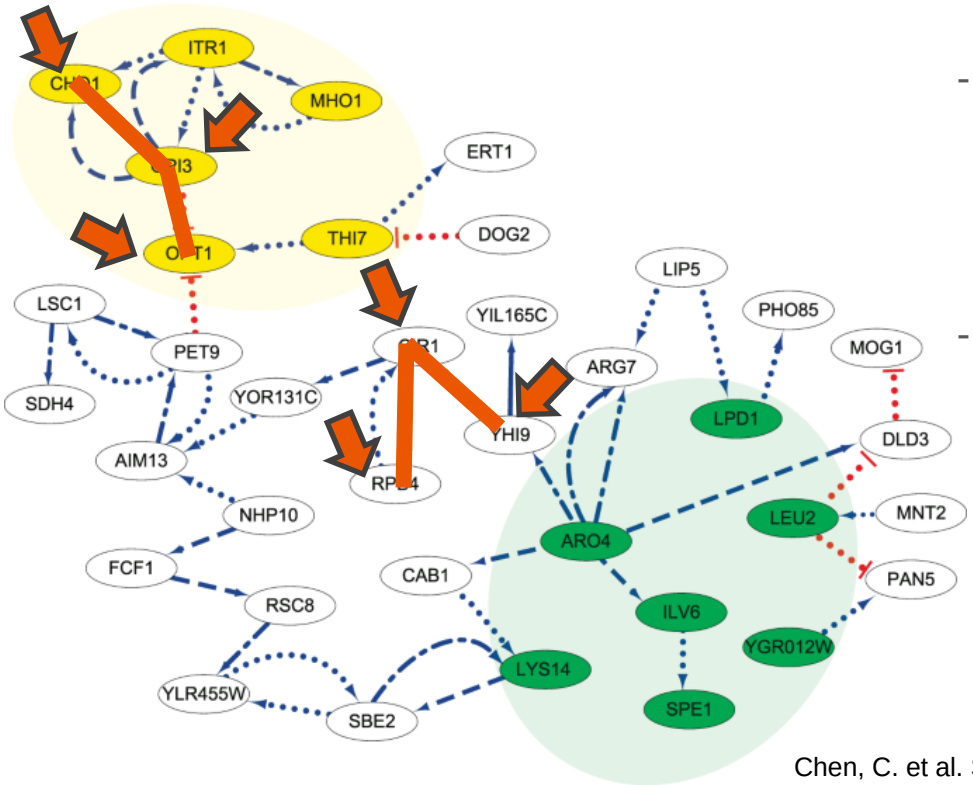
Gene and protein interaction networks



Differential expression of interacting genes/proteins are more meaningful

Chen, C. et al. Scientific Reports 9:1197 (2019)

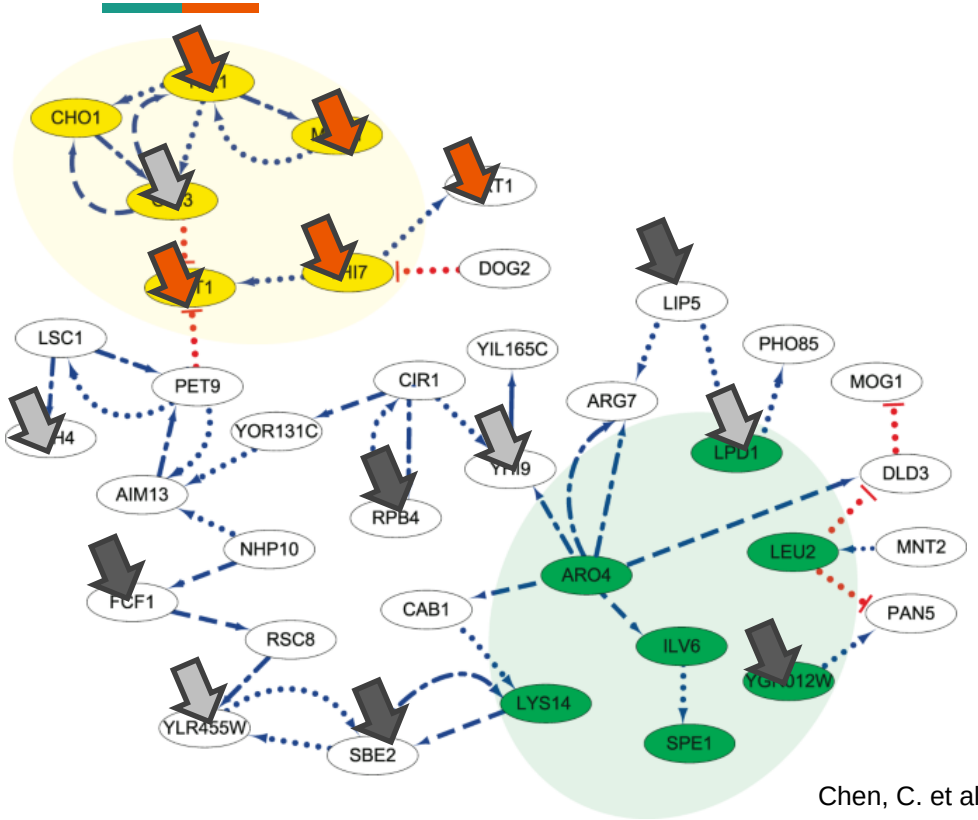
Network coherence scores



- Connectedness
  - Number of components
  - Number of edges
- Path length between genes
  - Unweighted
  - Weighted by fold changes

Chen, C. et al. Scientific Reports 9:1197 (2019)

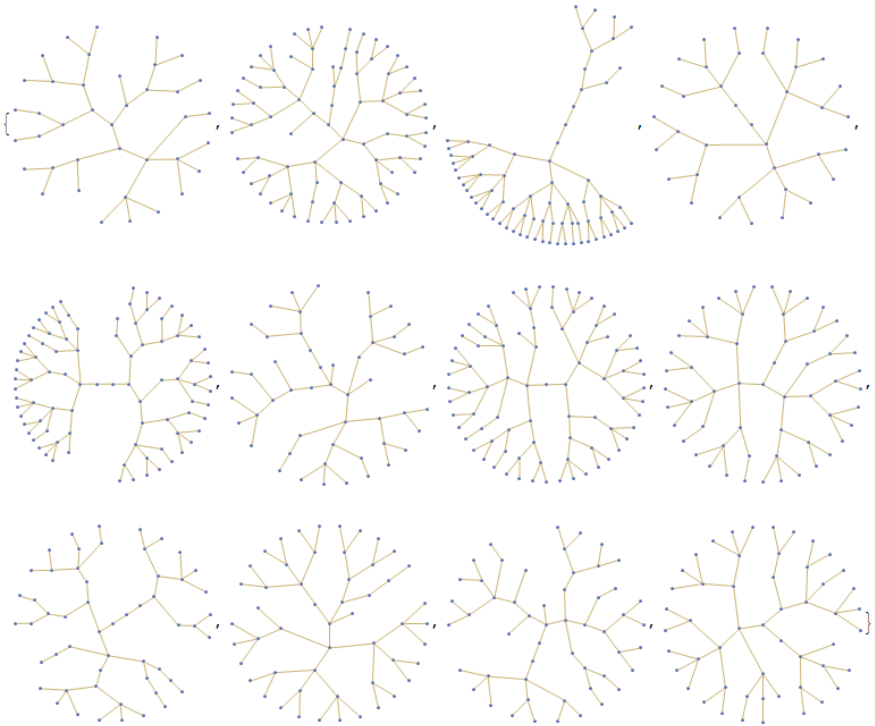
Permutation test: Gene set



- Randomly select the same number of genes
- Recalculate network coherence scores
- P-value = fraction of samplings that the score is ≥ the original

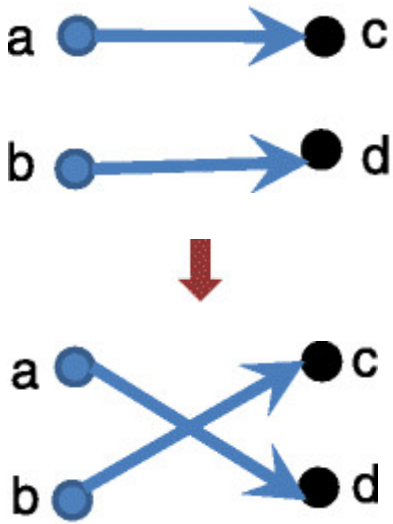
Chen, C. et al. Scientific Reports 9:1197 (2019)

Permutation test: Network



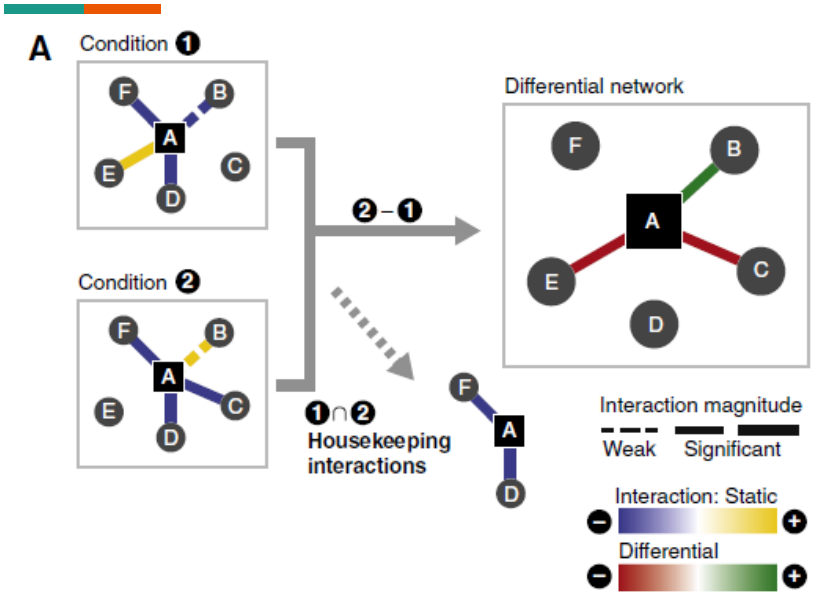
Source: https://mathematica.stackexchange.com/questions/11632/how-to-generate-a-random-tree

Edge switching

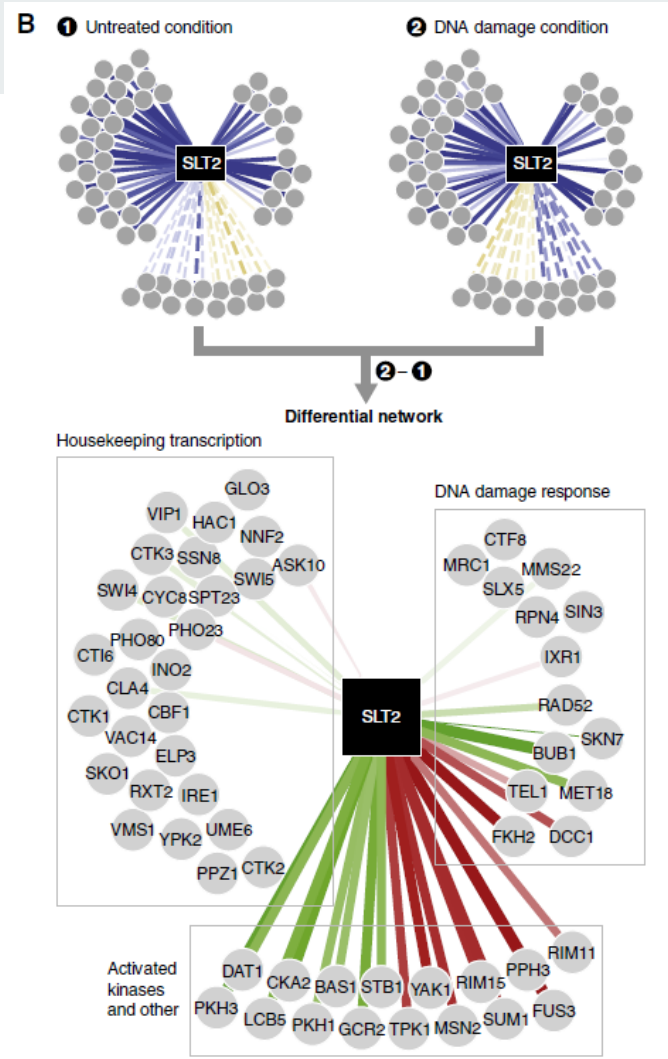


Temate-Tiageru *et al*. BMC Genomics, 17:542 (2016)

Differential network



Ideker *et al*. Mol Syst Biol, 8:565 (2012)



- Detect gain/loss gene co-expression
- Unrelated interaction remains the same

Pros and cons

- Overrepresentation
  - Easy and fast to calculate
  - Depend on p-value cutoff
- GSEA
  - No p-value cutoff
  - Distinguish up- and down-regulated functions
- Network-based
  - Most biologically meaningful
  - Network data is incomplete

Any question?

- Coming up:
  - Functional enrichment analysis demo