Special Topics in Biostatistics and Bioinformatics
Week VIII

Ege Ülgen, MD, PhD

21 April 2022

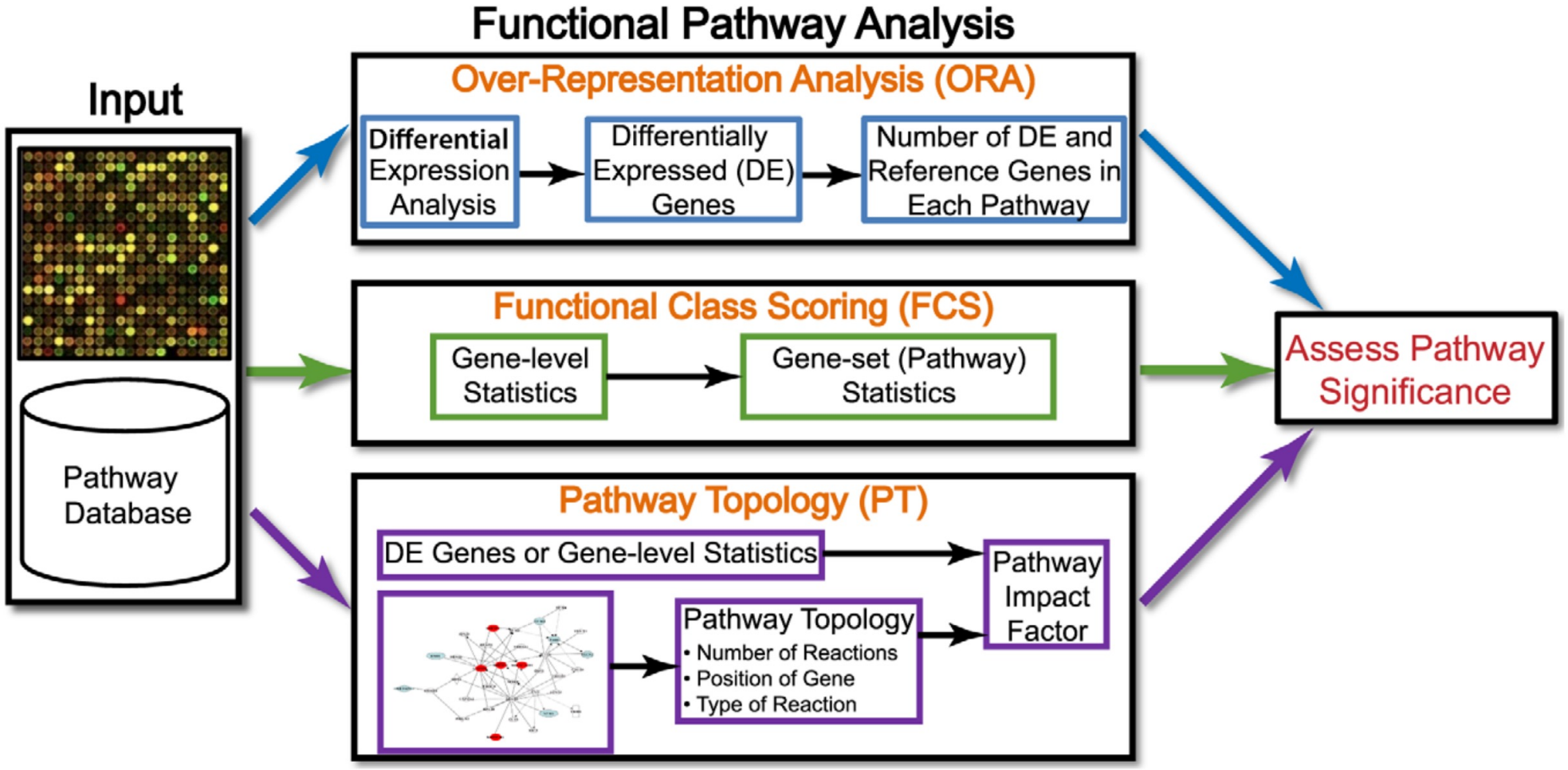
ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Background

- One of the most common use cases of NGS technologies is to perform experiments comparing two groups of samples (typically disease versus control) to identify **a list of significant (altered) genes**
- This list alone often falls short of providing mechanistic insights into the underlying biology of the disease being studied

- To **reduce the complexity of analysis** while **simultaneously providing great explanatory power**, one can investigate groups of genes that function in the same pathways/gene sets: **enrichment analysis**

Background



Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Gene Set Resources

- KEGG
- Reactome
- BioCarta
- WikiPathways

- Gene Ontology (GO)
  - GO-BP, GO-CC, GO-MF

- **MSigDB**

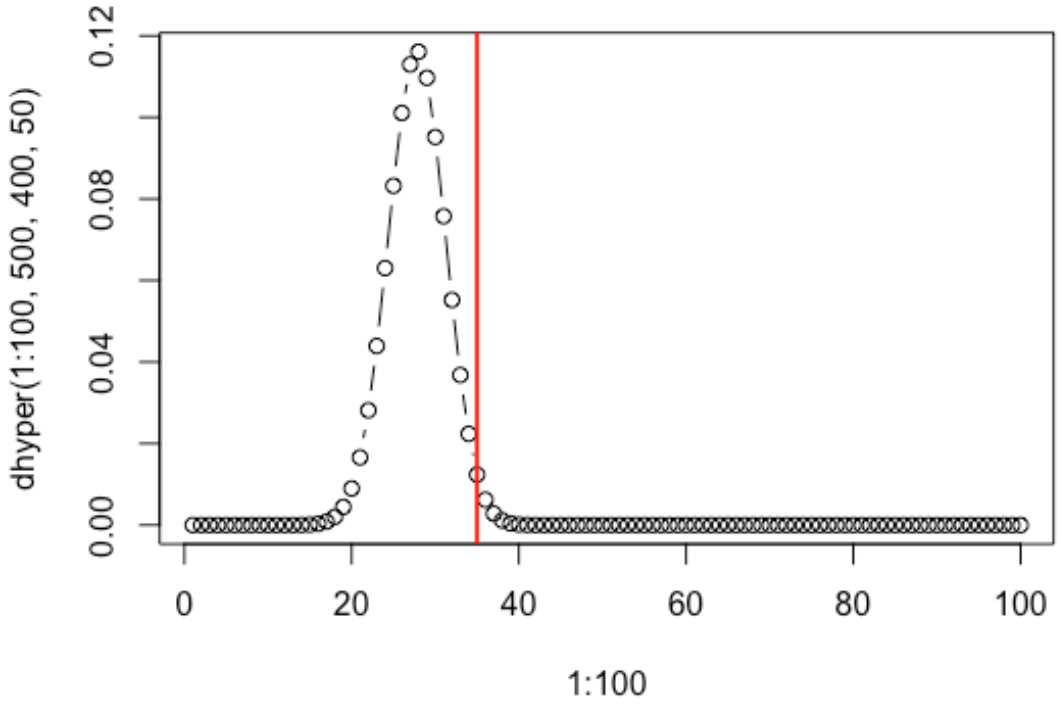
Over-Representation Analysis (ORA)

	In Gene Set	Not in Gene Set	Total
Selected	i	n – i	**n**
Not selected	M – i	N – M – n + i	**N – n**
Total	**M**	**N – M**	**N**

$$p = P(x \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

The test based on the hypergeometric distribution (hypergeometric test) is identical to the corresponding one-tailed version of Fisher's exact test

ORA



$$p = P(x \geq 35) = 1 - \sum_{i=0}^{34} \frac{\binom{500}{35}\binom{400}{15}}{\binom{900}{50}} = 0.010795$$

ORA Tools

- Enrichr
- DAVID
- ClueGO
- GenMAPP
- GoMiner
- …

Functional Class Scoring (FCS) Approaches

- Although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes can also have significant effects

1. Compute gene-level statistics
2. Aggregate gene-level statistics into a single pathway-level statistic
3. Assess the statistical significance of the pathway-level statistic (e.g., permutation testing)

FCS Tools

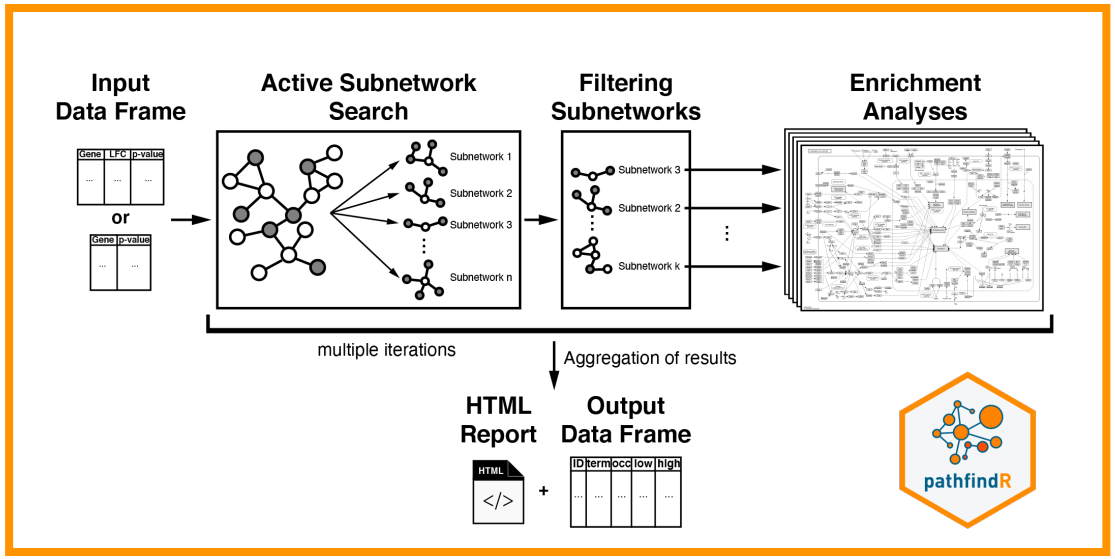
- GSEA
- SigPathway
- SAFE
- …

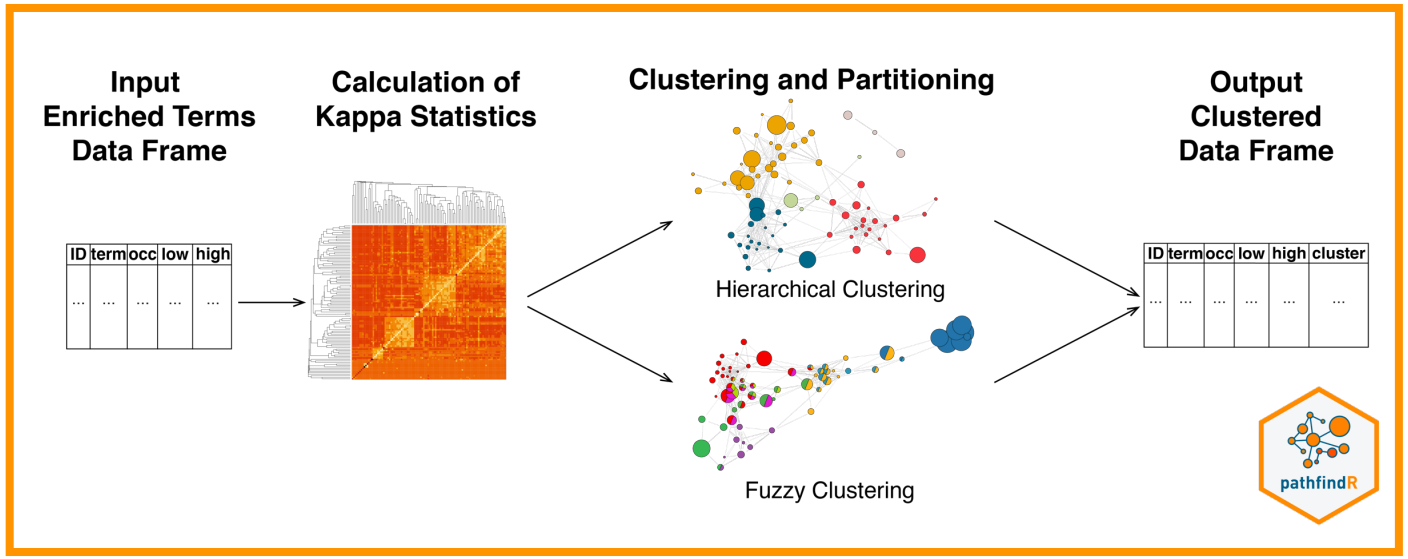
Pathway Topology (PT)-Based Approaches

- ORA and FCS methods consider only the number of genes in a pathway to identify significant pathways, and ignore the additional information available from interactions of genes
- Even if the pathways are completely redrawn with new links between the genes, as long as they contain the same set of genes, ORA and FCS will produce the same results

PT Tools

- pathfindR
- SPIA
- NetGSA
- Pathway-Express
- …





- Using input genes, pathfindR identifies sets of genes that form **active subnetworks** within a protein-protein interaction network

  An active subnetwork can be defined as a group of interconnected genes in a PIN that predominantly consists of significantly altered genes.

- It then performs **enrichment analyses** on the identified active subnetworks (see above diagram)
- Additionally, pathfindR provides functionality to:
  - **Cluster enriched terms** (see above diagram)
  - Calculate **agglomerated score per term activity per subject**
  - Combine and **compare** 2 pathfindR enrichment results
  - Create various **visualizations** of the analysis