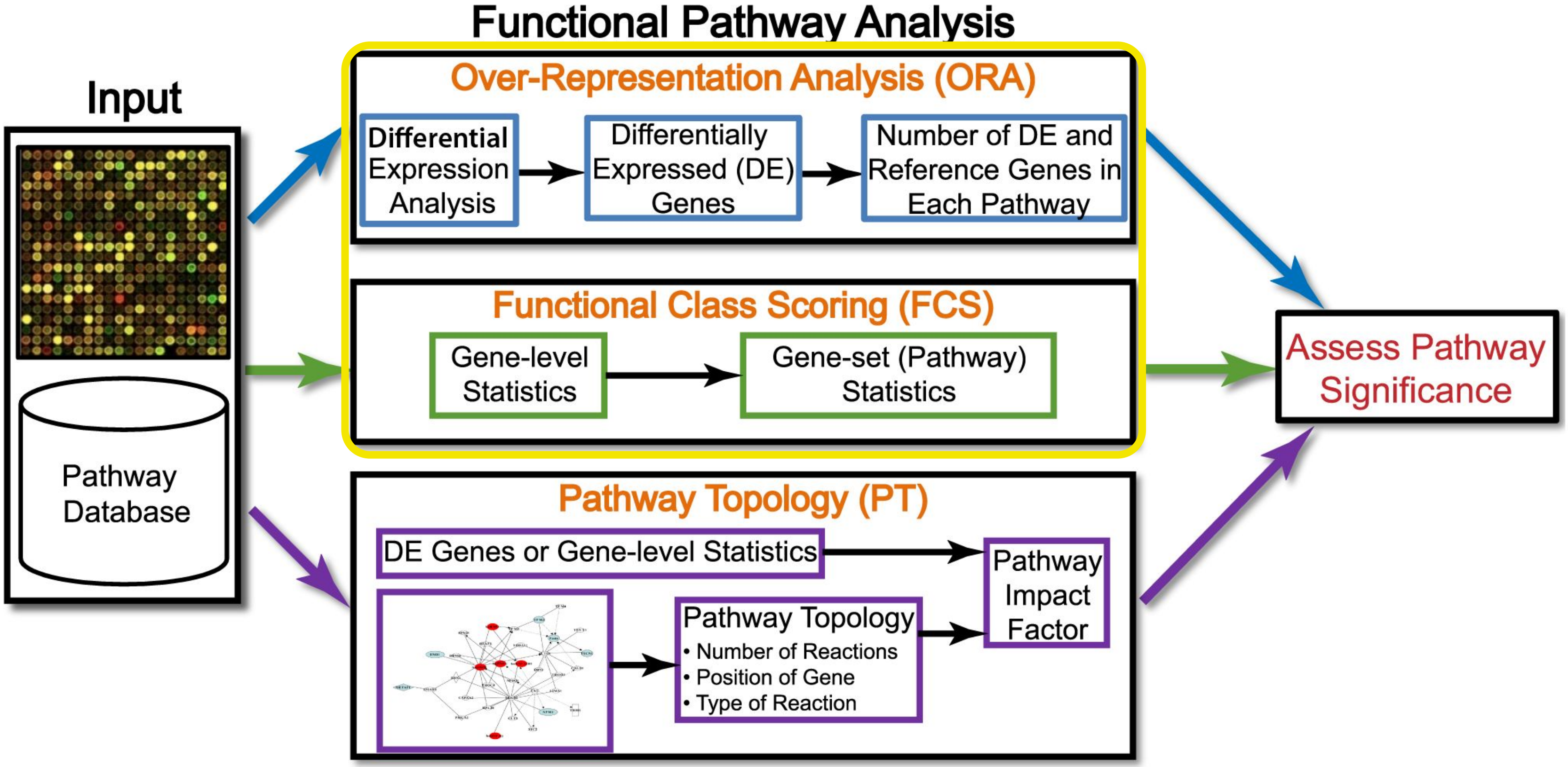
Introduction to Pathway Analysis

The CCDL

Why pathway analysis?

“...one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.”

- Subramanian et al. *PNAS.* 2005.



Functional Pathway Analysis

Input

Pathway Database

Over-Representation Analysis (ORA)

Differential Expression Analysis

Differentially Expressed (DE) Genes

Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)

Gene-level Statistics

Gene-set (Pathway) Statistics

Pathway Topology (PT)

DE Genes or Gene-level Statistics

Pathway Topology

- Number of Reactions
- Position of Gene
- Type of Reaction

Pathway Impact Factor

Assess Pathway Significance

Khatri, Sirota, and Butte. *PLoS Comp Bio*. 2012.

Today we'll cover three types of pathway analysis

Over-representation analysis (ORA)

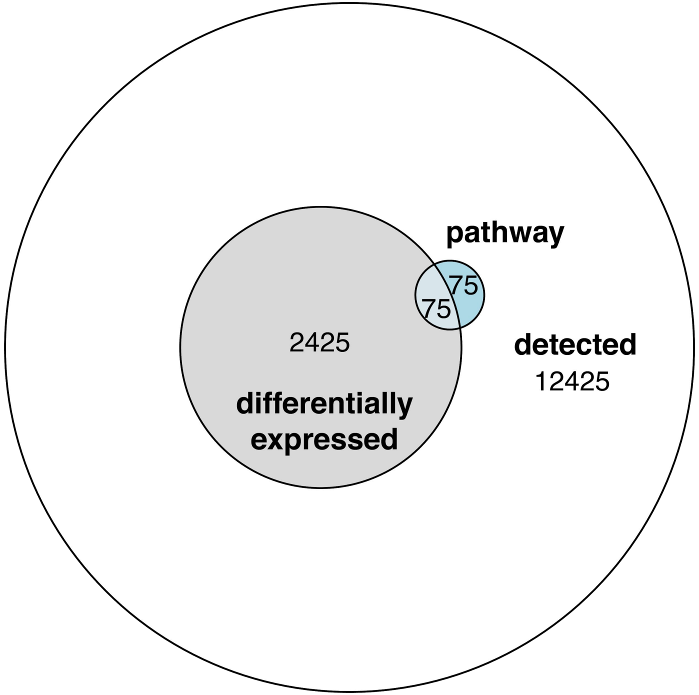
I have a list of genes from my analysis and I'm interested in if genes from a pathway are represented in that list more than I would expect by chance.

✓ **Pros**

- Simple
- Computationally inexpensive to compute p-values

⚠ **Cons**

- Requires arbitrary thresholds and ignores any statistics associated with a gene
- Assumes independence of genes and pathways



Today we'll cover three types of pathway analysis

Gene Set Enrichment Analysis (GSEA)

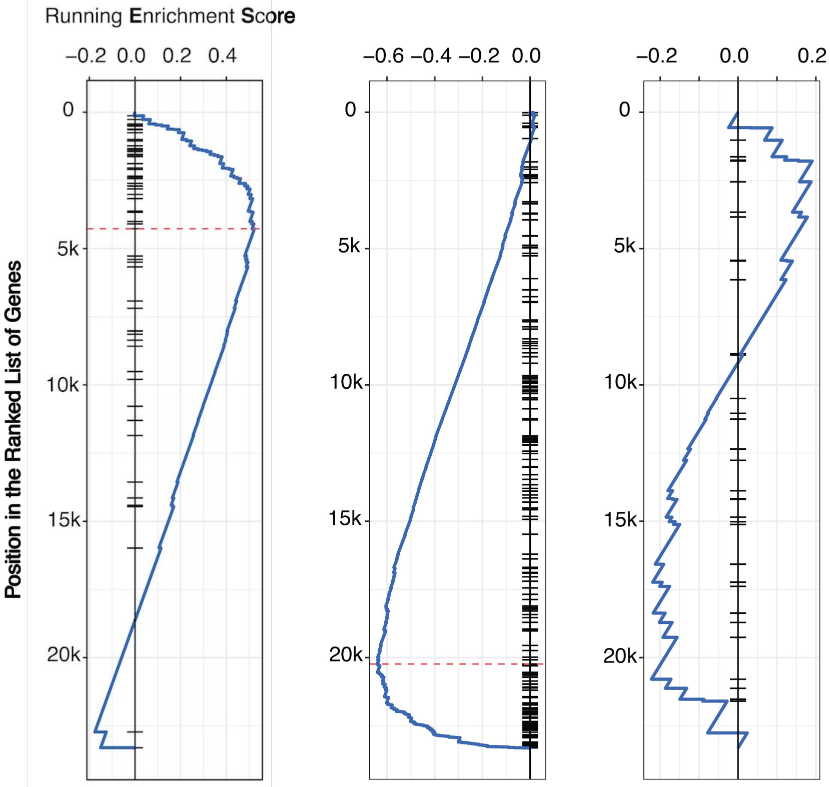
I have a gene-level statistics from a two-group comparison, and I would like to know if there are coordinated changes in pathway that are unlikely to be detected by looking at differentially expressed genes alone.

✓ **Pros**

- Includes *all* genes (no arbitrary threshold!)
- Attempts to measure coordination of genes

⚠ **Cons**

- Permutations can be expensive
- Does not account for pathway overlap
- Two-group comparisons not always appropriate/feasible

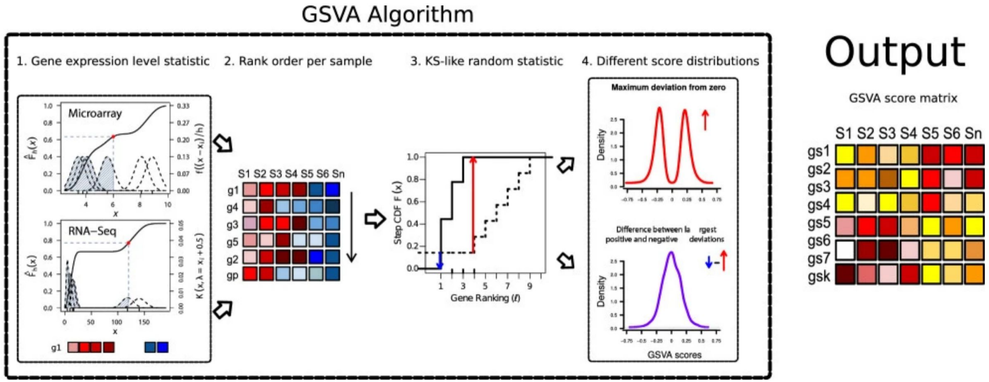


Subramanian et al. *PNAS.* 2005.

Today we’ll cover three types of pathway analysis

Gene Set Variation Analysis (GSVA)

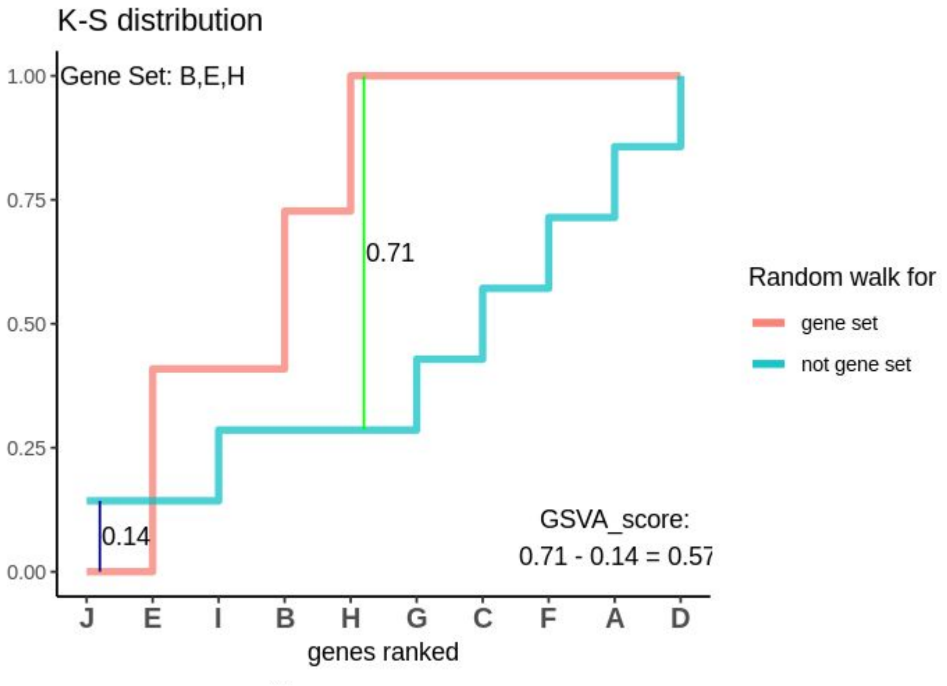
I don’t have two groups to compare, so I want pathway-level scores on a per-sample basis that tell me if genes in the pathway are over- or under-expressed relative to the overall population.

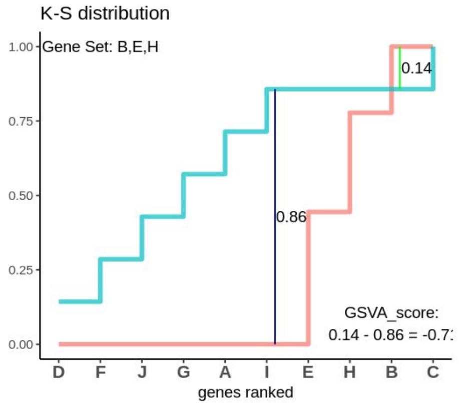


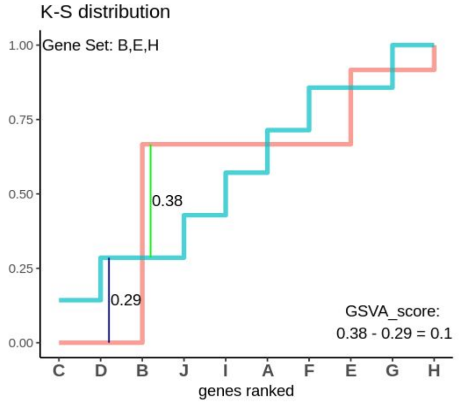
Hänzelmann, Castelo, and Guinney. *BMC Bioinformatics*. 2013.

Today we’ll cover three types of pathway analysis

Gene Set Variation Analysis (GSVA)





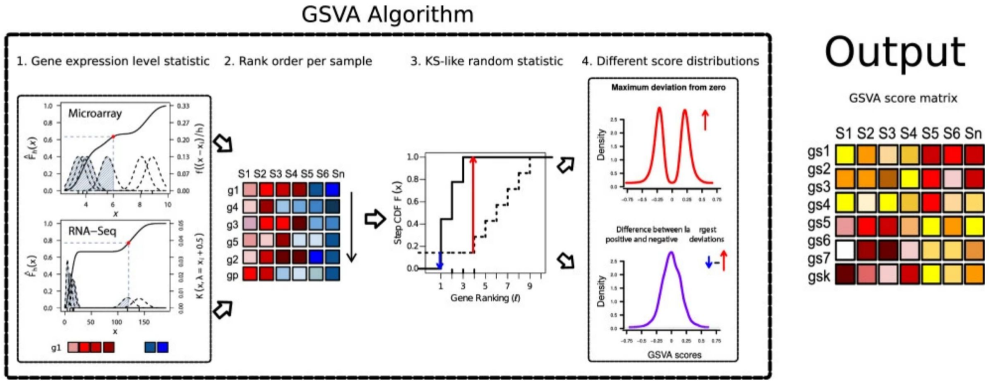


All images from Malhotra. 2018.

Today we'll cover three types of pathway analysis

Gene Set Variation Analysis (GSVA)

I don't have two groups to compare, so I want pathway-level scores on a per-sample basis that tell me if genes in the pathway are over- or under-expressed relative to the overall population.



✓ Pros

- Does not require two groups to compare upfront
- Normally distributed scores

⚠️ Cons

- Scores are not a good fit for gene sets that contain genes that go up AND down
- Method doesn't assign statistical significance itself
- Recommended sample size n > 10

Hänzelmann, Castelo, and Guinney. *BMC Bioinformatics*. 2013.

Resources

Guangchuang Yu. *clusterProfiler: universal enrichment tool for functional and comparative study*.

Harvard Chan Bioinformatics Core Training. *Intro to DGE: Functional analysis*.

Saksham Malhotra. *Decoding Gene Set Variation Analysis*.

refine.bio examples on pathway analysis

Molecular Signatures Database (MSigDB)