Gene set enrichment analysis in R

Kim Dill-McFarland

2021.07.14

Gene set

two or more genes with shared function, structure, localization, or any other defining similarity

Gene set features

- Defined by individual study to general knowledge

- Vary in size

- Not exclusive (1 gene in many sets)

- Redundant and overlapping

Broad Molecular Signatures Database

Hallmark

- summarize and represent specific well-defined biological processes
- generated by computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression

- Total: 50
- Examples: glycolysis, inflammatory response, apoptosis

Broad Molecular Signatures Database

Curated (C2) Canonical pathways

- from pathway databases including KEGG, REACTOME, etc
- Canonical representations of a biological process compiled by domain experts

- Total: 2922
- Examples: Caspase pathway, signaling by NOTCH1, DNA repair

Broad Molecular Signatures Database

Gene ontology (C5) Biological process

- molecular-level activities performed by gene products

- Total: 7481
- Examples: viral life cycle, vitamin D biosynthetic process, mitochondrial calcium-ion transmembrane transport

Broad Molecular Signatures Database

- too specific (individual studies)
  - C2 chemical and genetic perturbations
  - C7 immunologic signatures

- areas not relevant to our experimental design
  - C1 chromosome position
  - C3 gene regulation
  - C4/C6 cancer-oriented

Hypergeometric enrichment (aka Fisher’s Exact test)

Probability that the number of significant genes in a gene set occurred by chance

Hypergeometric enrichment

1. Define genes in a gene set

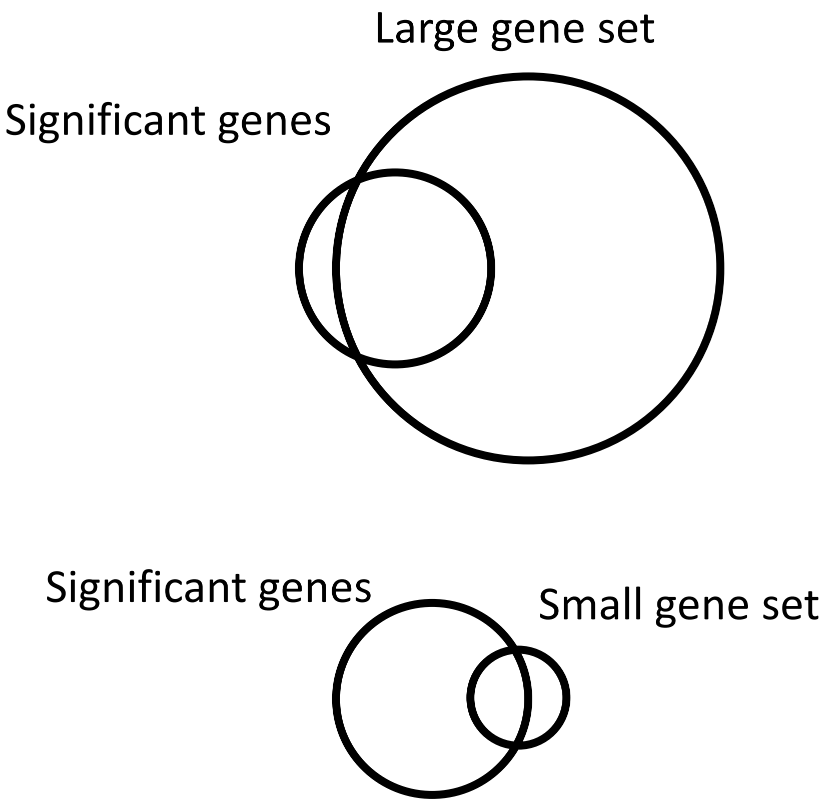
2. Define genes significant for your variable of interest (in our case, media vs Mtb-infected)

3. Calculate proportion of significant genes in gene set

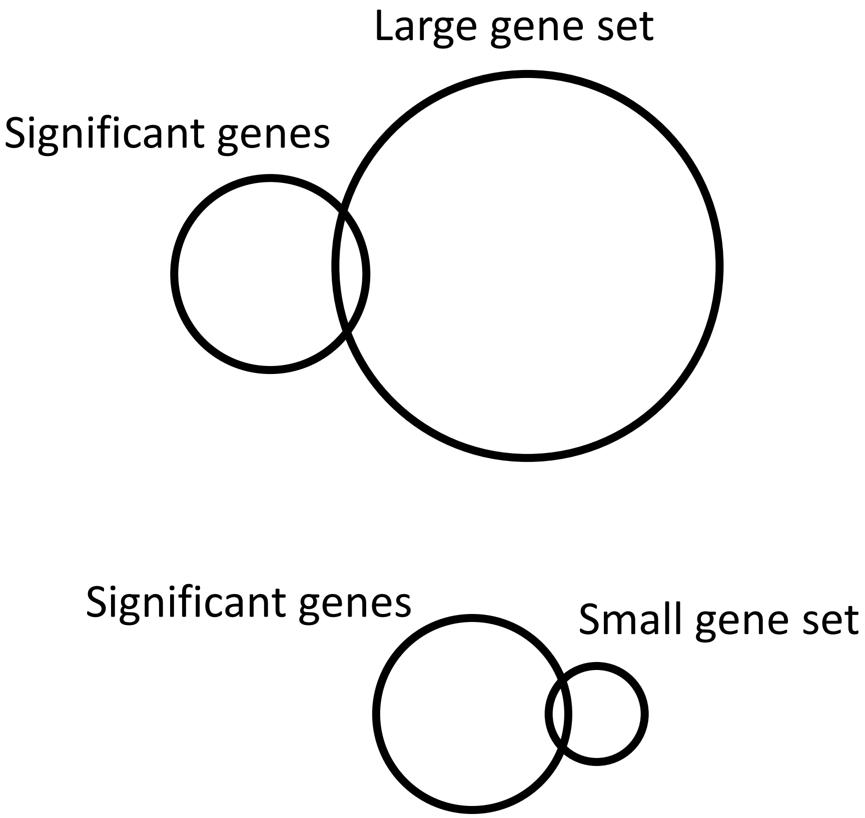
4. Estimate probability and significance (P-value)

Hypergeometric enrichment

Likely enriched



Not likely enriched



Gene set enrichment analysis (GSEA)

Compares expression in two biological states and determines if gene sets show significant, concordant change

aka

Determines if genes in a set generally increase or decrease in expression

Gene set enrichment analysis (GSEA)

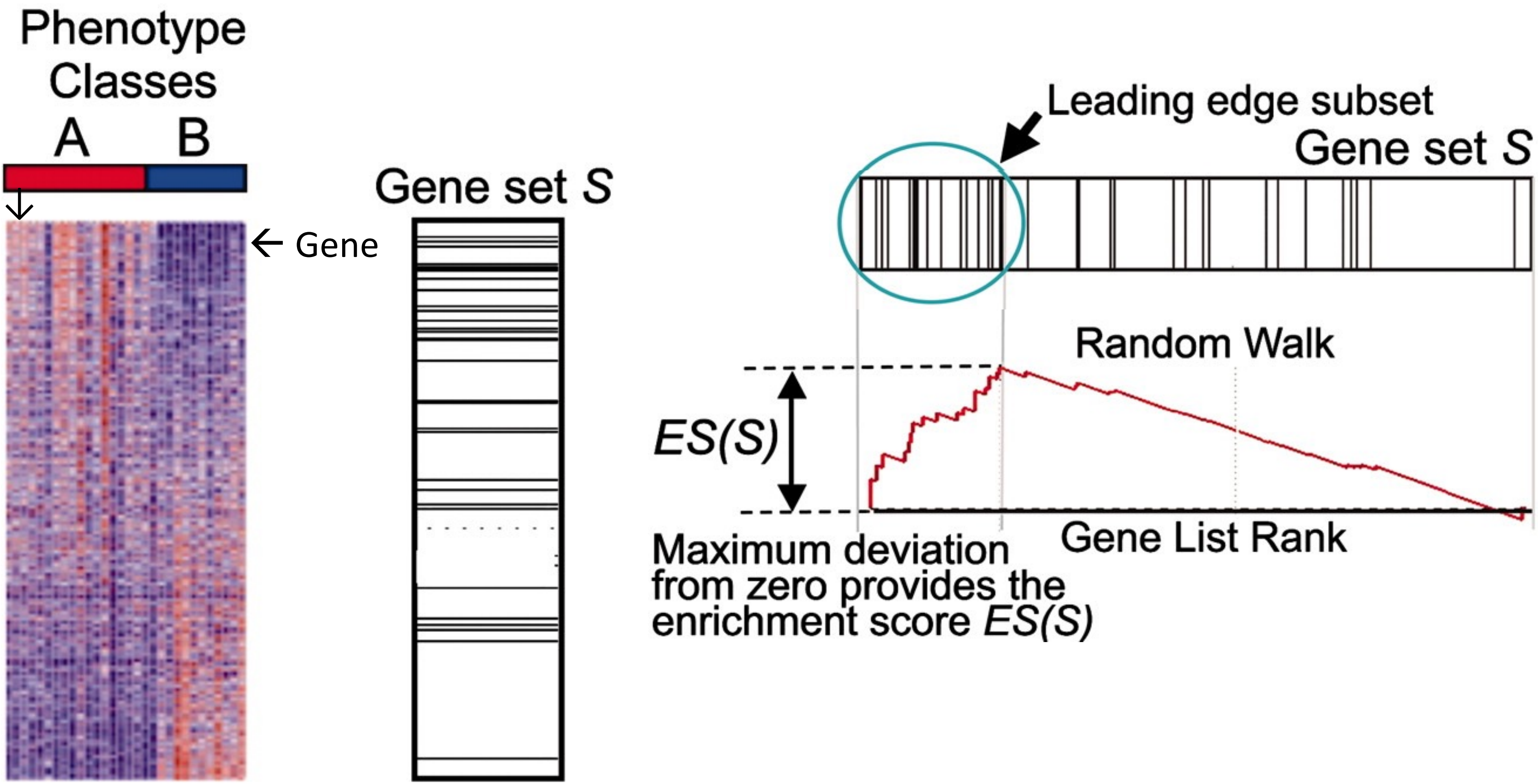
1. Define genes in a gene set

2. Calculate fold change

3. Order genes by fold change and map to gene set

4. Estimate enrichment score and significance (P-value)

Gene set enrichment analysis (GSEA)



Sample ↓

← Gene

Most positive correlation with phenotype

Most negative correction with phenotype

Subramanian et al PNAS

doi.org/10.1073/pnas.0506580102

Hypergeometric enrichment	Gene set enrichment analysis
• Significant genes	• All genes
• Binary significant vs not	• Numeric fold change values
• Depends on how you define significance	• Mean fold change per gene
• Compare 2+ states	• Compare only 2 states