Module 7: Differential Gene Analysis (Part 2)

Announcements

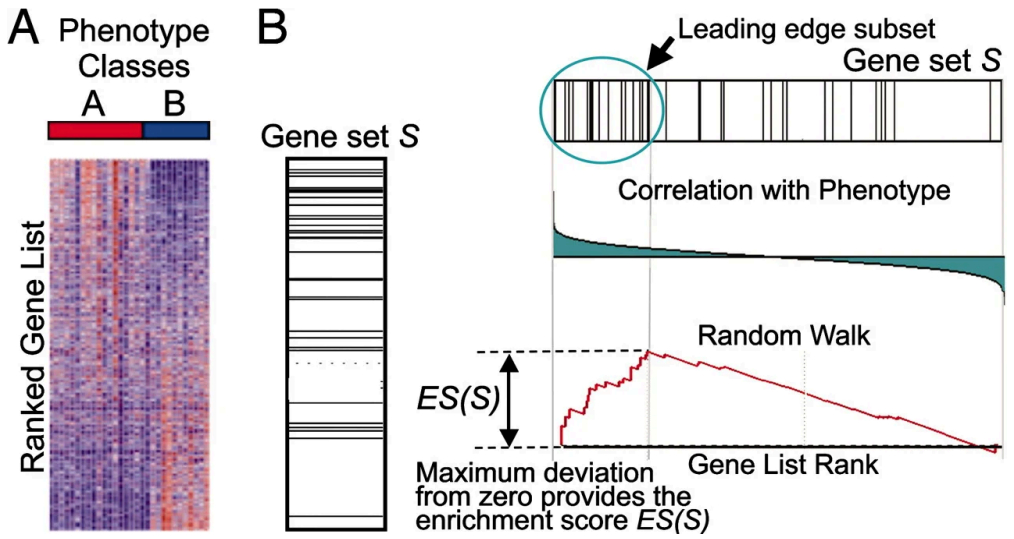
- Lab 0 and 1 marks
- Module 5 and 6 asyncronous class polls on BB
- Module 8: Group Assignment 2

Key Concepts

- Gene Set Enrichment Analysis
- Statistical Metrics
- Over-Representation Analysis
- Limitations

Gene Set Enrichment Analysis [3]

- The **gene sets** are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments.
- Goal: Determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction.



Tool	GSEA
Function	Tests whether predefined sets of genes (e.g., pathways) show statistically significant, coordinated differences between two biological states.
Key Features	- Uses a ranked list of all genes (no initial cutoff) - Outputs metrics like **NES**, **Nominal p-value**, and **FDR** - Permutation-based approach for robust p-value estimation
Use Case	- Integrating RNA-seq (or microarray) data to identify enriched biological processes - Evaluating curated gene signatures (e.g., hallmark pathways)

Gene Set Enrichment Analysis

Question: Which biological pathways or processes are altered under specific conditions?

- **Example**: A group of genes involved in a specific pathway (e.g., oxidative phosphorylation, immune response, cell cycle) is collectively up- or down-regulated in your samples.
- Identifying enriched pathways can help us understand the mechanisms underlying the biological response to a perturbation (e.g., drug treatment, disease state, environmental stimulus).

Gene Set Enrichment Analysis

Question: Do known disease-related gene signatures correlate with a sample phenotype?

- **Example**: Previously characterized gene signatures (e.g., for certain cancers, metabolic diseases, or immune disorders) show coordinated changes in your samples.
- Validates (or not) existing clinical or biological hypotheses (e.g., whether a new set of patient samples displays a “classic” cancer signature or a novel subtype).

Gene Set Enrichment Analysis

Question: How can we prioritize functional follow-up experiments?

- Knowing the the most relevant (enriched) processes or pathways associated with a treatment response can help guide focused functional studies (e.g., targeted knockdowns/knockouts).

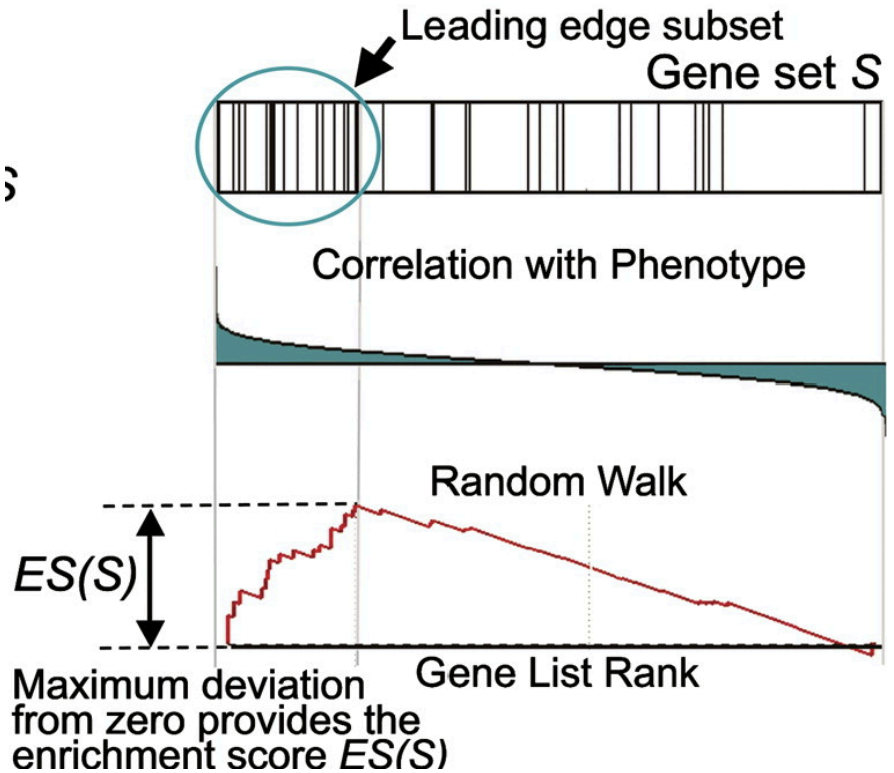
Gene Set Enrichment Analysis

Question: (Exploratory) What new hypotheses can we generate about disease mechanisms or drug actions?

- **Example**: A drug designed to target one pathway may be found to also enrich gene sets related to metabolic processes.
- Discovery of unexpected connections between conditions and pathways can lead to new lines of research into off-target drug effects, polypharmacology, or novel disease mechanisms.

Gene Set Enrichment Analysis

- **Enrichment Score (ES)** measures how overrepresented (positively) or underrepresented (negatively) a gene set is at the top or bottom of a ranked list of genes.
- **Normalized Enrichment Score (NES)** adjusts the raw ES to account for gene set size and other dataset-specific factors so it is more comparable across gene sets of different sizes and across different experiments.



Gene Set Enrichment Analysis

- **False Discovery Rate (FDR)**: the estimated probability that a set with a given NES represents a false positive finding
- An FDR q-value of 0.05 means that, among all gene sets called significant, approximately 5% are expected to be false positives.
    - We want to control the proportion of false positives
- When reporting results, NES with FDR < 0.05 (or a user-defined cutoff) is often the statistical threshold for calling a gene set “significantly enriched.”

Gene Set Enrichment Analysis

- The **nominal p-value** estimates how likely it is to observe an enrichment score for a gene set based on permutation tests.
- It does not account for multiple testing across many gene sets.
- Even if the nominal p-value is small, the FDR may not be significant if many gene sets are tested or the dataset is large.

Gene Set Database/Tools

	Description
MSigDB	A comprehensive collection of annotated gene sets designed for use with Gene Set Enrichment Analysis (GSEA).
mulea	An R package for multi-level enrichment analysis, providing flexible tools for gene set enrichment and annotation.
Enrichr	A user-friendly web-based tool for gene set enrichment analysis, offering a broad range of libraries and visualizations.

Over-representation Analysis

- Identifies whether certain biological categories (e.g., pathways, Gene Ontology terms, functional clusters) are statistically overrepresented in a user-defined list of “interesting” genes (commonly differentially expressed genes).
- Use it as a **hypothesis-generating tool** rather than definitive proof of pathway activation.

Usage Limitations

- Best applied to a well-defined gene list (e.g., significantly DE genes).
- Requires an appropriate background set of genes for statistical comparison (often all genes measured).
- Requires multiple testing correction (e.g., FDR, Bonferroni) when testing many gene sets.

Over-representation Analysis

Interpretation Limitations

- **Depends on annotation quality**. If databases (GO, KEGG, etc.) are incomplete or outdated, results may be misleading.
- Some enriched terms can be overly general (e.g., “metabolic process”)
- **Biological Complexity**. ORA treats genes independently; so it doesn’t capture pathway interactions or gene-gene dependencies.
- **Cutoff Bias**. The initial gene selection criteria (e.g., p-value thresholds) can influence which categories appear enriched.
- **No Directionality**. Traditional ORA doesn’t distinguish between up- or down-regulation within a gene set—just the presence or absence of genes.

Over-representation Analysis

Question: Which biological processes are most prominent in my gene list?
Context: You have identified a set of up- or down-regulated genes under a specific condition (e.g., disease state vs. control).
Interpretation: If immune-system-related terms (e.g., “inflammatory response,” “T cell activation”) are significantly overrepresented, it signals that immune processes could be key drivers of the phenotype.

Over-representation Analysis

Question: Do certain pathways show significant sverrepresentation?

- Many gene sets are curated as “pathways” (e.g., KEGG pathways, Reactome, WikiPathways).

Example: In a cancer study, ORA might flag the “p53 signaling pathway” when multiple p53 target genes show differential expression, implicating aberrant p53 regulation in tumor progression.

Over-representation Analysis

Question: Are there functional categories linking my genes of interest?

- ORA can also consider categories such as Cellular Components (e.g., “mitochondrial membrane”) or Molecular Functions (e.g., “ATPase activity”).

Example: A cluster of newly identified genes in the sample is enriched for “transcription factor activity.”
Interpretation: These genes are part of a transcriptional regulatory network driving the observed phenotype.

Over-representation Analysis

Question: How do known disease gene signatures map onto the data?

- Many gene sets are curated around specific diseases or traits (e.g., OMIM, DisGeNET, or published disease signature gene sets).

Example: The significant gene list from a new mouse model of neurodegeneration overlaps strongly with a known Alzheimer's disease genes.
Interpretation: The experimental model captures key disease mechanisms relevant to Alzheimer's disease.

Tools for functional analysis

	clusterProfiler
Puporse	An R package that performs over-representation analysis (ORA), GSEA, and other enrichment methods using multiple annotation databases (e.g., GO, KEGG).
Key Features	- Highly flexible for various enrichment methods (ORA, GSEA, etc.) - Supports multiple species and annotation sources - Generates publication-ready visualizations (dot plots, bar plots, network plots)
Usage	- Annotate and visualize enriched pathways in R - Customize analyses with user-defined gene lists and annotation sets