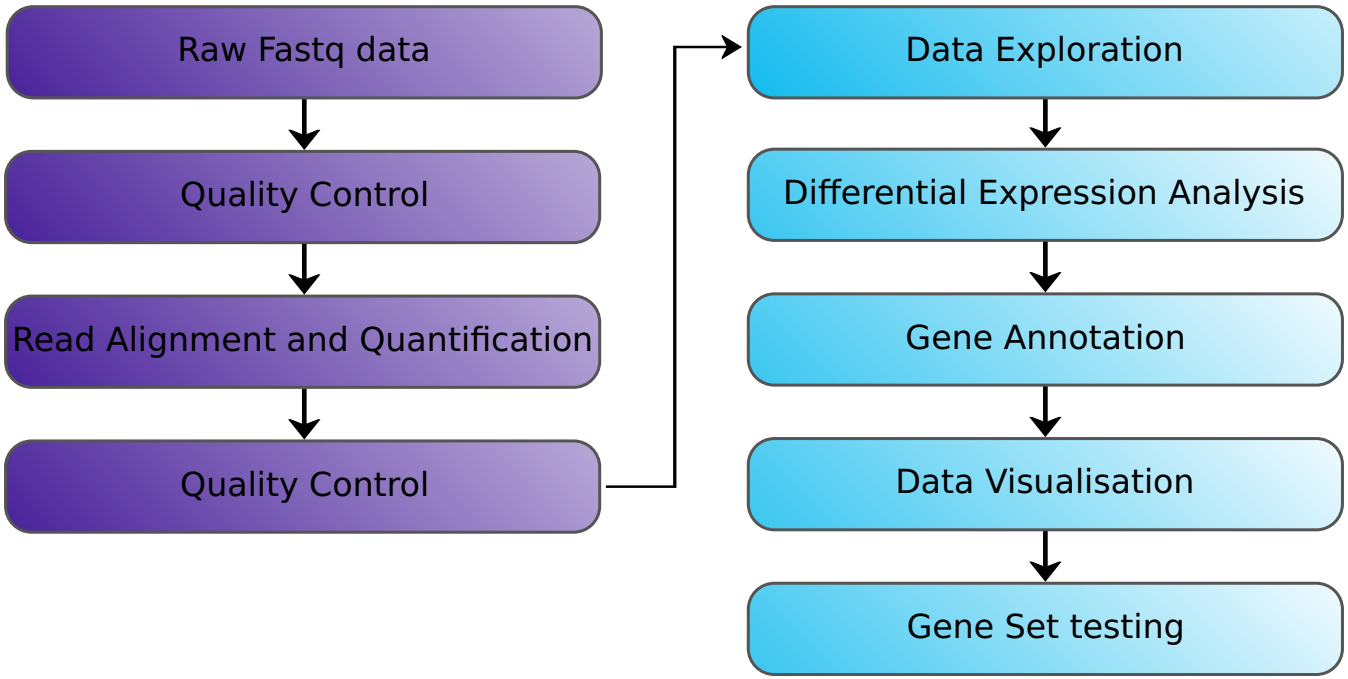
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Introduction to Gene Set Testing in R

March 2023

Differential Gene Expression Analysis Workflow



Gene Set Testing - Overview

The list of differentially expressed genes is sometimes:

- so long that its interpretation becomes cumbersome and time consuming,
- or very short while some genes have low p-value yet higher than the given threshold.

There are many approaches to searching for biological meaning in the results of differential expression analysis.

Commonly we assess whether the differentially expressed genes tend to relate to specific pathways or ontological groups of genes.

We will look at two methods:

- Over Representation Analysis (ORA)
- Gene Set Enrichment Analysis (GSEA)

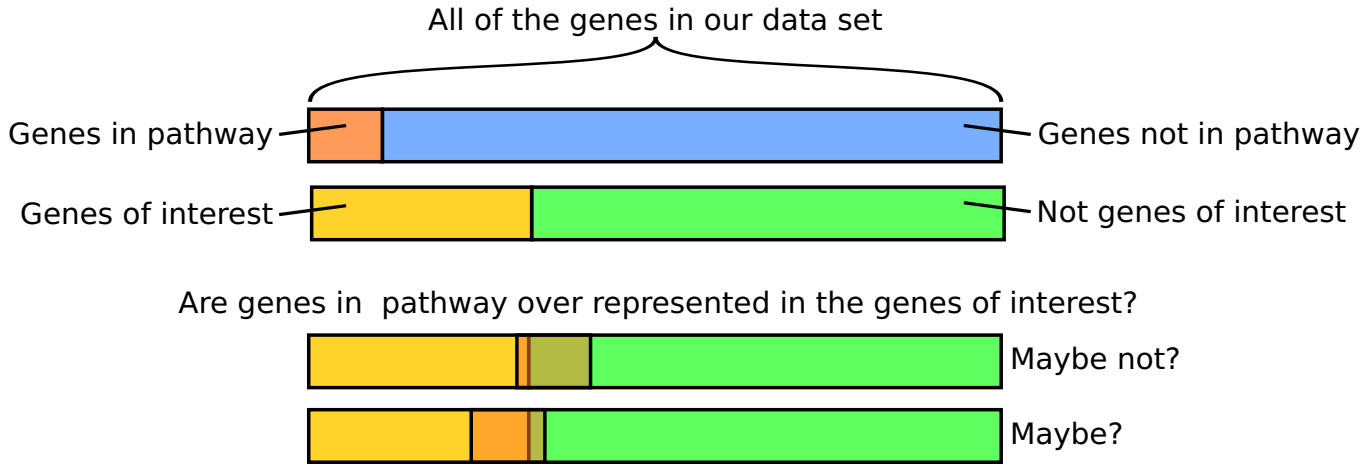
Gene Set Testing - Resources

Common sources of gene sets:

- KEGG pathways
- Gene Ontologies
- Reactome
- Molecular Signature Database, MSigDB (GSEA)
- Manually curated gene lists

Over Representation Analysis - Method

- This method tests whether genes in a specific pathway are present in a subset of genes of interest in our data more than expected.
- The genes of interest could be e.g. statistically significant genes or a cluster of genes from hierachical or k-means clustering.
- Given the ratio of genes in the pathway to genes not in the pathway, is the number of genes in the pathway and in our subset statistically unlikely by chance.



Over Representation Analysis - Example

Genes in the experiment are split in two ways:

- annotated to the pathway or not
- differentially expressed or not

Contingency table:

		Of interest: yes	Of interest: no	
In pathway	yes	20	300	320
	no	80	19600	19680
		100	19900	20000

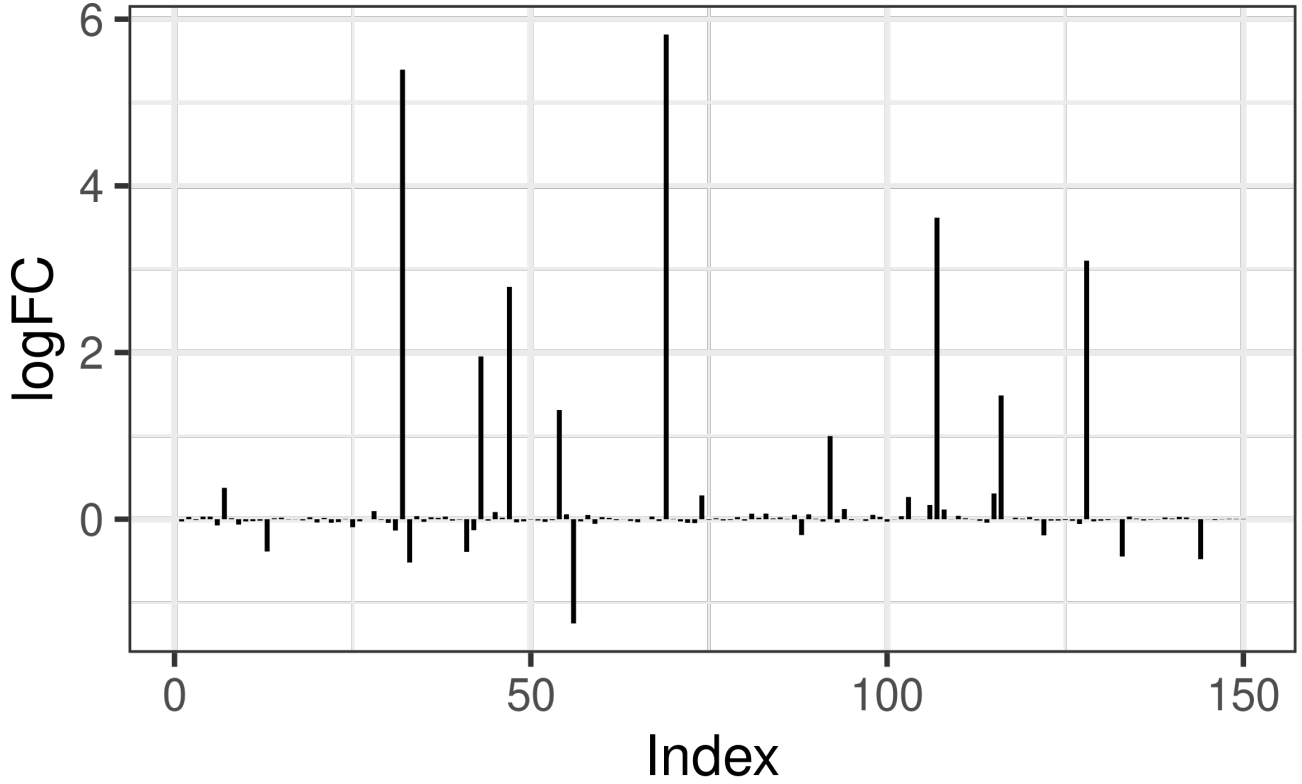
- Analysis with the hypergeometric/Fisher's exact test

Gene Set Enrichment Analysis (GSEA)

- This method is based on ranking all genes in our dataset
- If the gene set is significantly affected in our experiment, then the genes in the set should tend to be at one end or the other of our ranking.
- The ranking method is arbitrary, but p-value and fold change are common choices.
- GSEA calculates an enrichment score based on the ranking, and then uses permutation to calculate a p-value for how significant the enrichment score is.

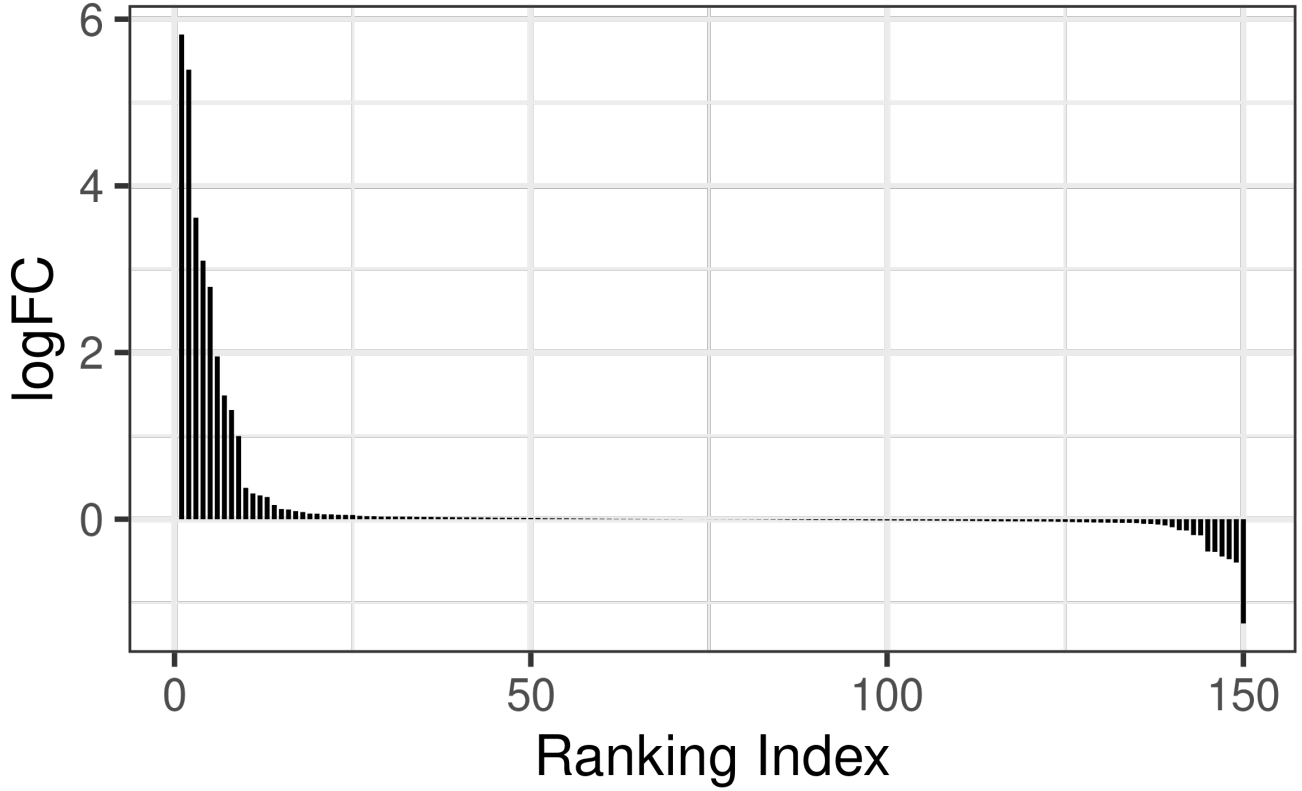
GSEA - Calculate the enrichment score

- Ranking by Fold Change - unsorted logFC



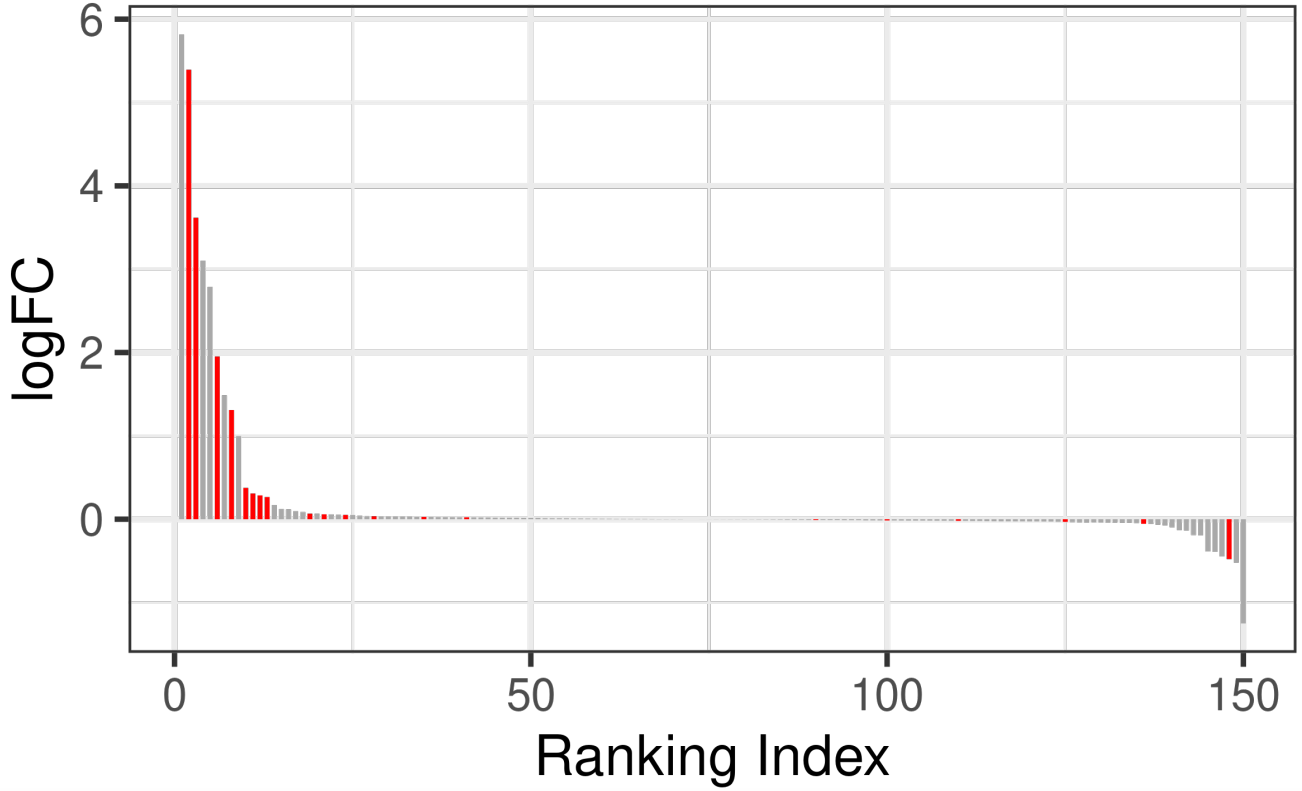
GSEA - Calculate the enrichment score

- Ranking by Fold Change - sorted logFC, in decreasing order



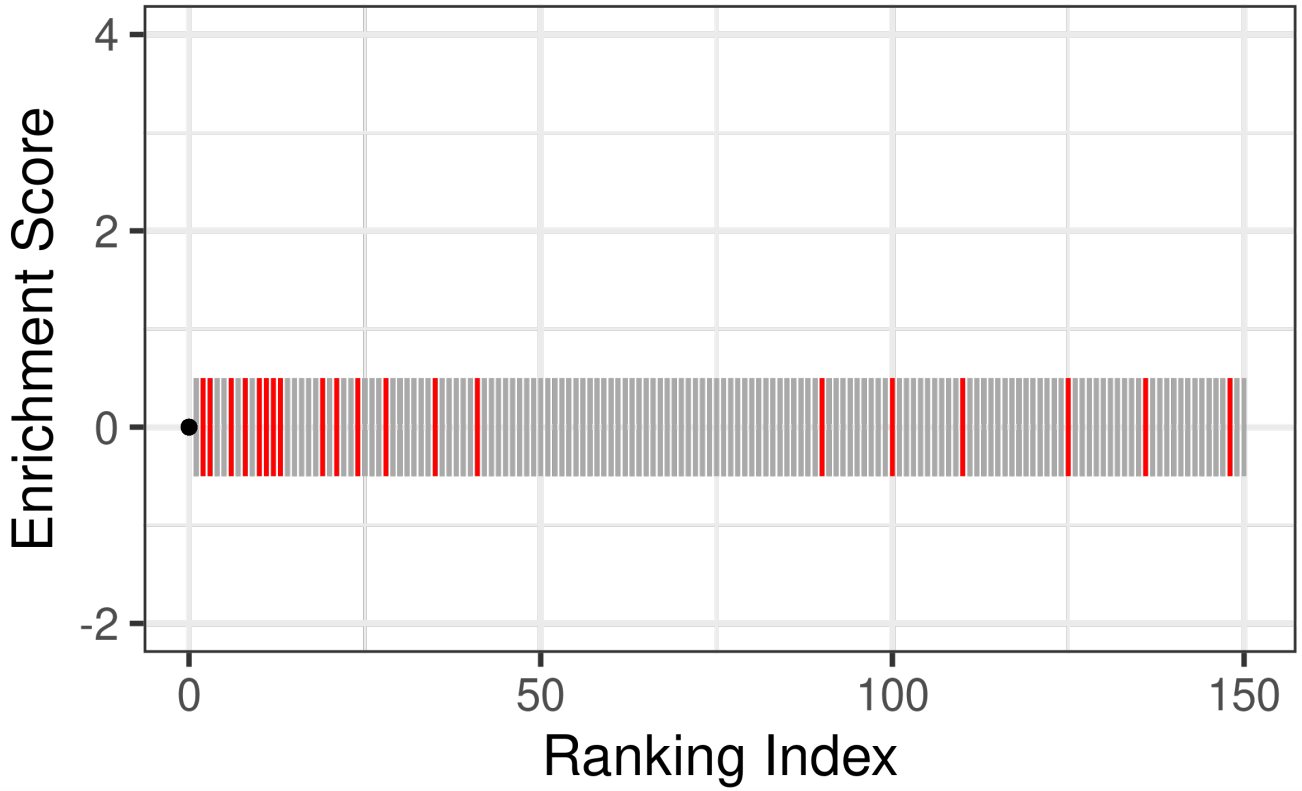
GSEA - Calculate the enrichment score

- Identify genes in set



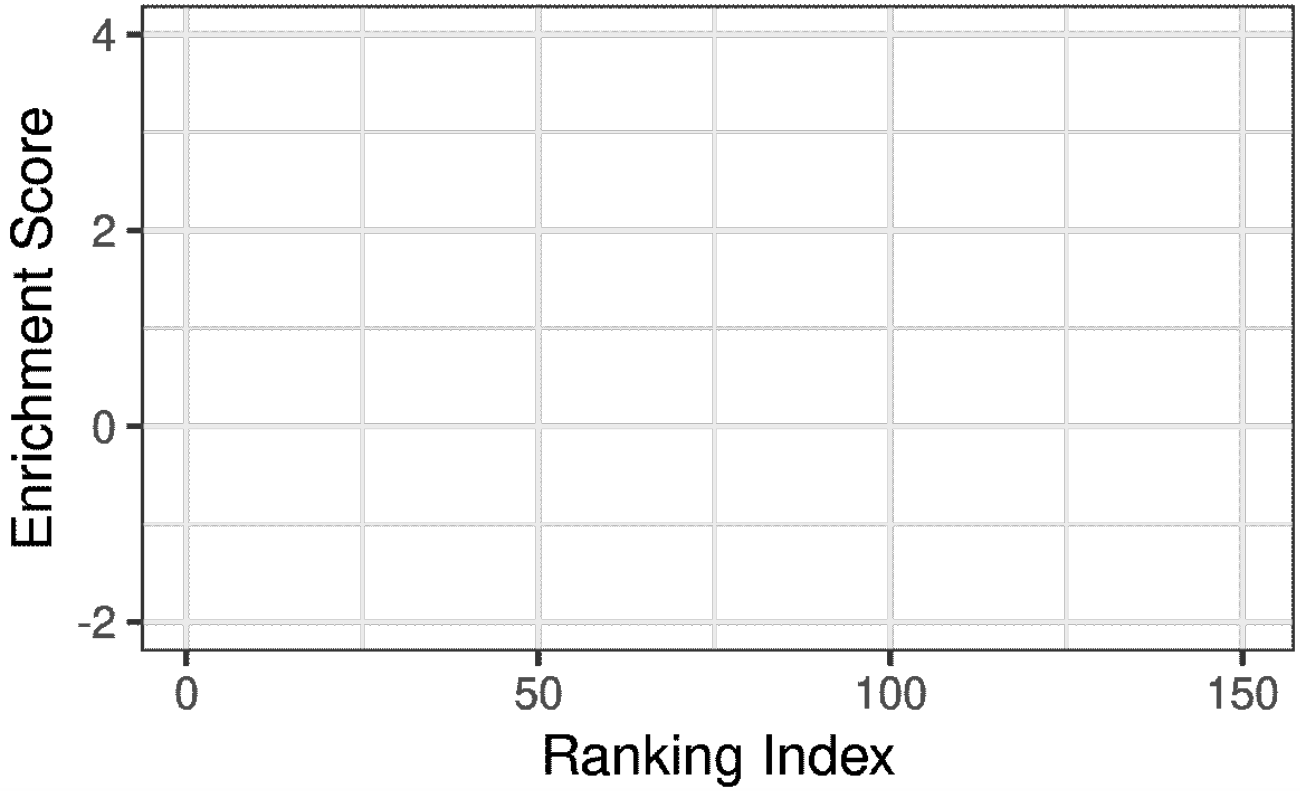
GSEA - Calculate the enrichment score

- Calculate the enrichment score … start at 0 and an enrichment score of 0



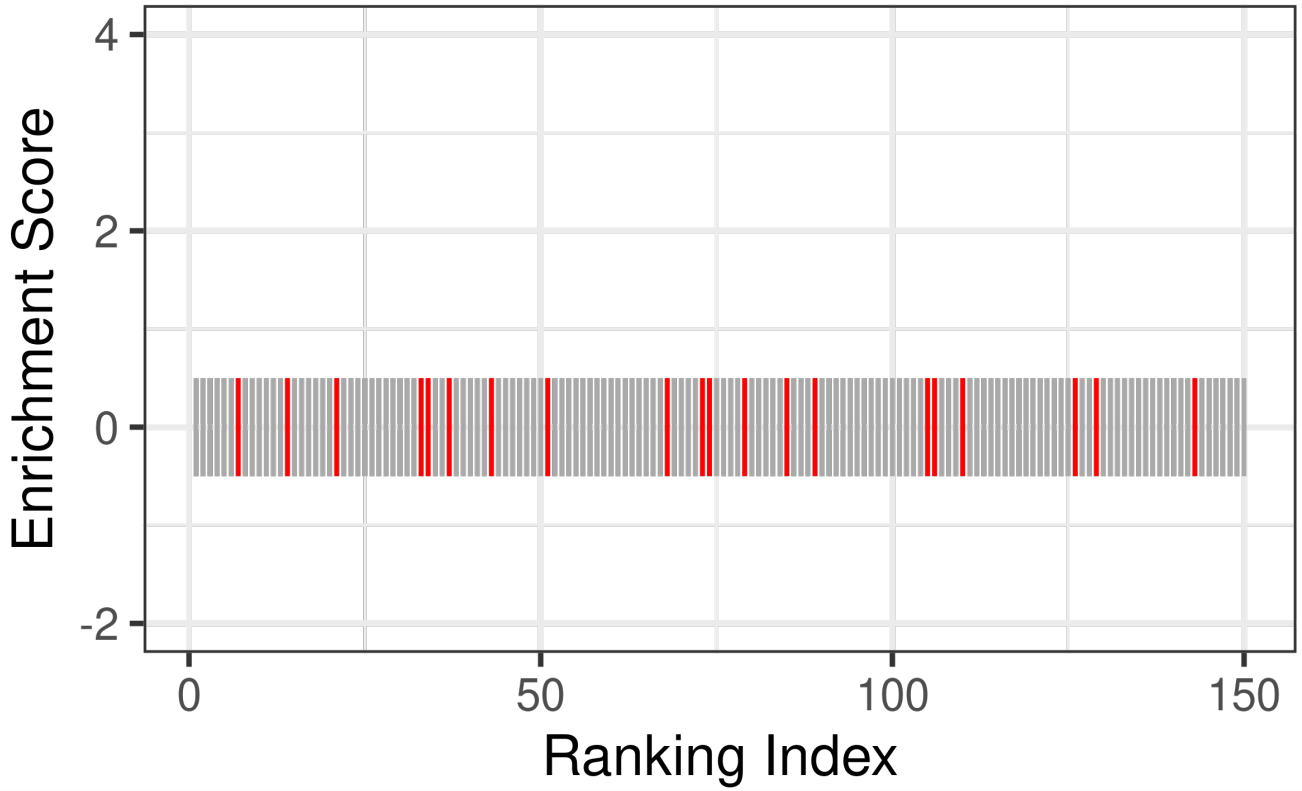
GSEA - Calculate the enrichment score

- Walk along genes and calculate a cumulative score



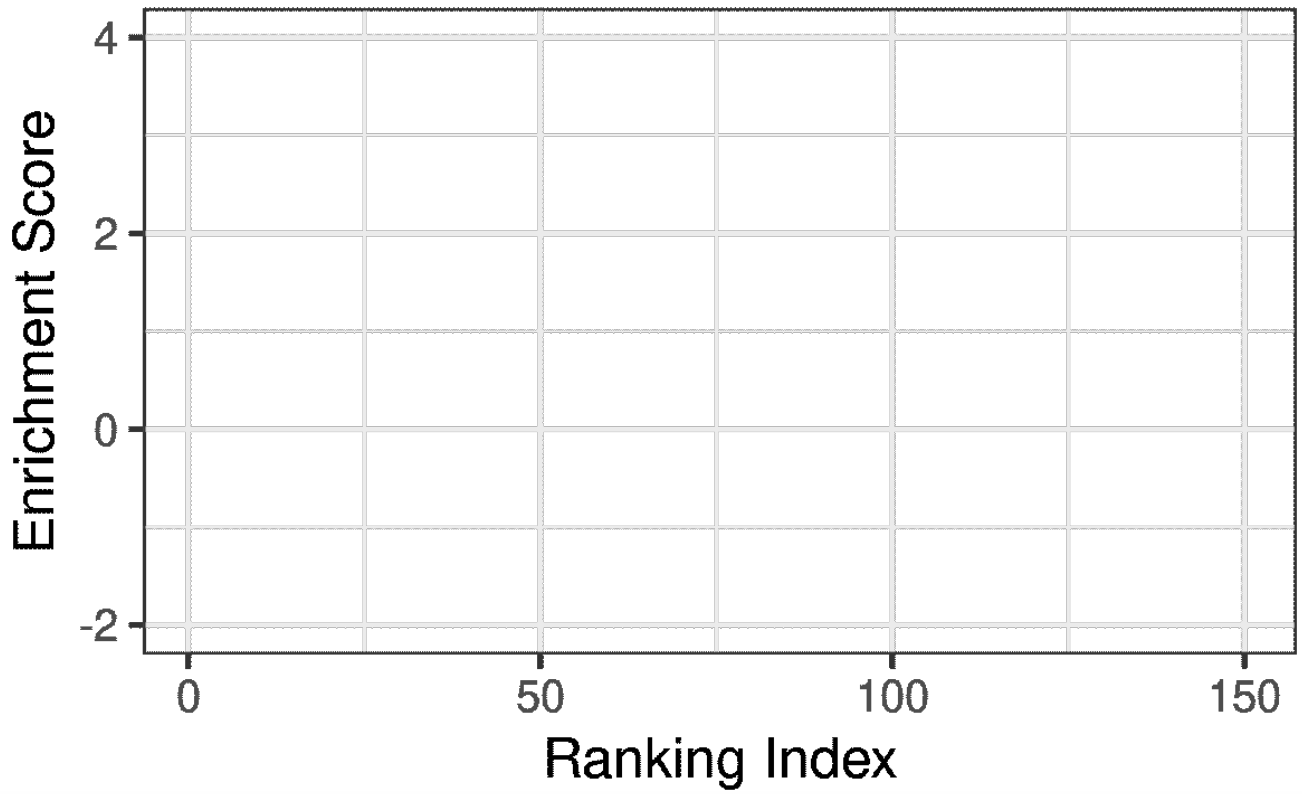
GSEA - Calculate the enrichment score

- A different gene set



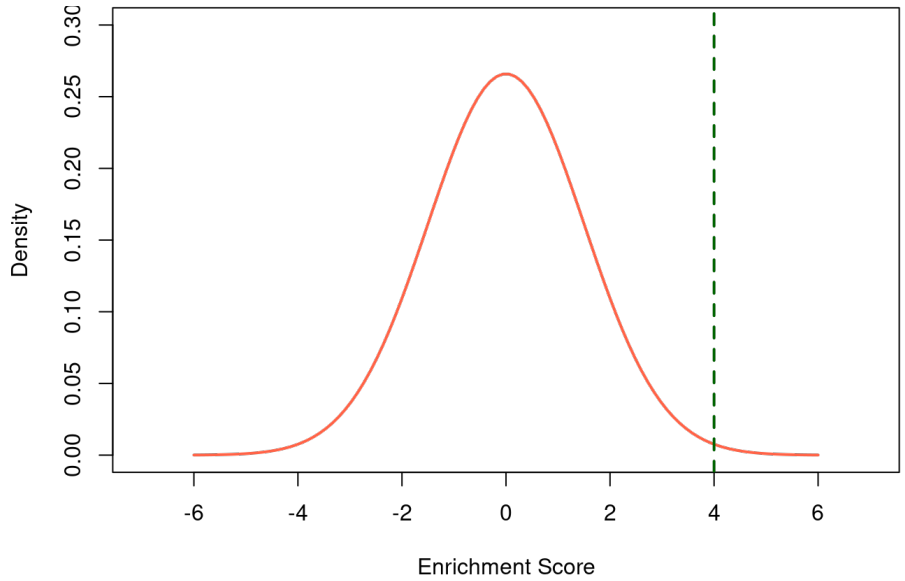
GSEA - Calculate the enrichment score

- A different gene set



GSEA - Estimate a p-value

- Randomly permute the ranking and recalculate the enrichment score, repeat many times.
- From a distribution of our permuted enrichment scores determine how likely our ES.



Recap

Question: Do the differentially expressed genes tend to relate to specific pathways or ontological groups of genes?

For a given contrast and a given gene set.

Two methods:

- Over Representation Analysis (ORA)
  - split genes two ways: in pathway or not, of interest or not
  - Fisher exact test for ratio of ‘pathway’ odds in the two ‘interest’ classes
- Gene Set Enrichment Analysis (GSEA)
  - rank all genes using significance and/or logFC
  - compute enrichment score
  - compute its significance

Both methods are applicable to series of gene sets