Introduction to Pathway Analysis for scRNA-seq

The CCDL

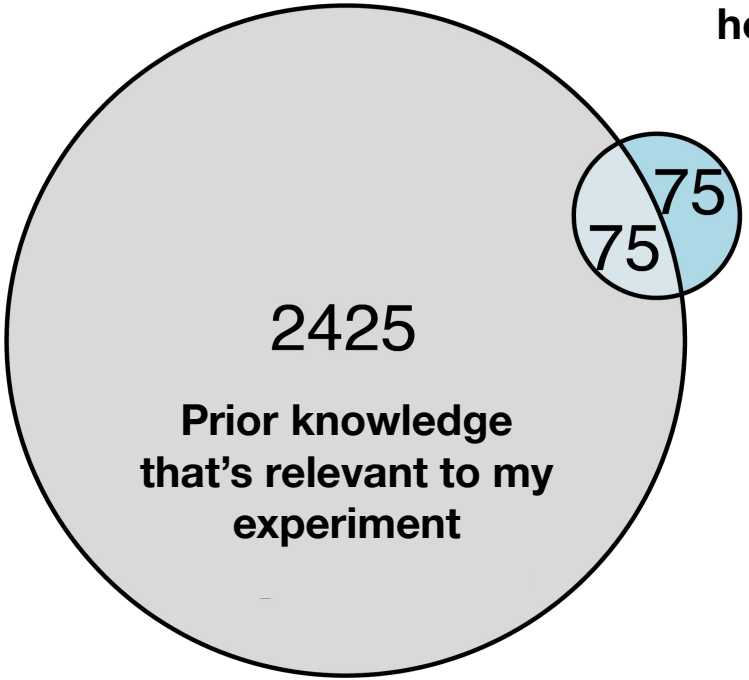
Why pathway analysis?

"...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.

Our choice of method and gene sets for pathway analysis will depend on our analytical goals!

Genes differentially expressed in a cluster that hopefully tell me about cell type or state



A(n incomplete) list of available gene sets

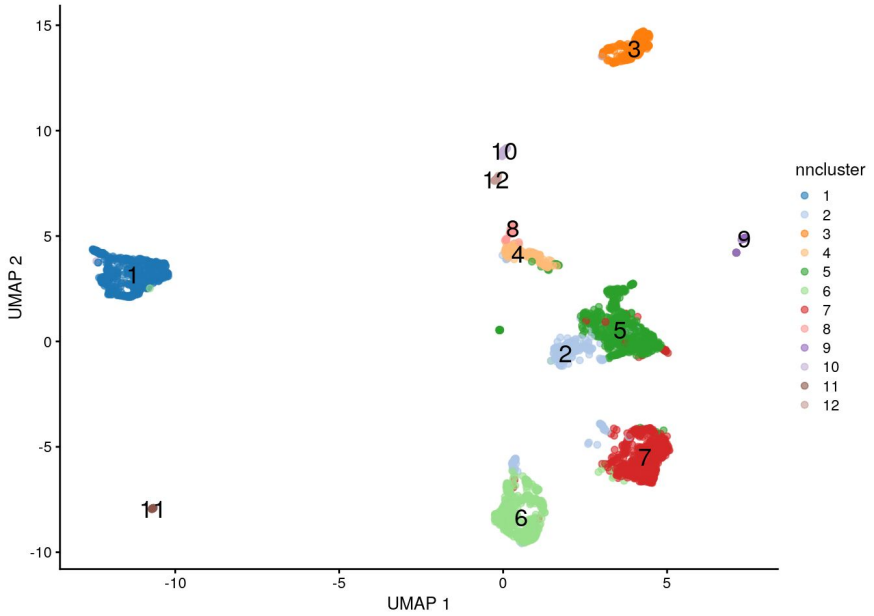
Molecular Signatures Database - Multiple collections including curated sets like KEGG that capture processes like signaling pathways or sets derived from gene expression experiments of specific perturbations.

The Gene Ontology - An ontology that describes our knowledge of the biological domain; comprised of 3 parts: Molecular Function, Biological Process, and Cellular Component.

CellMarker - Curated resource of cell markers; includes genes and proteins.

Marker genes

Pathway analysis for marker genes



For this module, we will use pathway analysis to gain insight into the clusters in the Hodgkin's lymphoma data. (Our pathway analysis results should not be used as justification for our clustering results!)

We will use the marker genes from the clustering we performed yesterday as *input*.

Marker genes: limitations & caveats

- p-values associated with marker genes are unreliable because we identified clusters based on gene expression and then tested for differential expression.

- When we're comparing multiple clusters, we may want genes to be "significant" in a comparison between any comparisons, some comparisons, or all comparisons. Yesterday, we picked *all comparisons*.

- We will also get a statistic for every pairwise comparison; we may want to summarize the effect into a single value.

Invalidity of p-values. Orchestrating Single-Cell Analysis.

Marker genes: limitations & caveats

gene <chr>	p.value <dbl>	FDR <dbl>	summary.logFC <dbl>	logFC.2 <dbl>	logFC.3 <dbl>
ENSG00000247982	1.171597e-96	8.509311e-93	1.3965975	1.3965975	1.4120335
ENSG00000224137	4.804290e-65	1.744678e-61	1.2418416	1.2601307	1.2484506
ENSG00000159958	1.444776e-50	3.497802e-47	2.0844021	2.0262768	2.1174520
ENSG00000177455	2.161339e-43	3.924451e-40	1.0255806	1.0496091	1.0478330
ENSG00000153064	5.183487e-39	7.529533e-36	1.9040733	1.9222008	1.9505774
ENSG00000105369	4.437805e-31	5.371963e-28	3.2070922	3.0830074	3.2597209
ENSG00000196092	1.804723e-28	1.872529e-25	0.7265076	0.7505360	0.7517113
ENSG00000156738	4.680927e-28	4.249696e-25	3.6501497	3.7194994	3.7156639
ENSG00000211898	8.614940e-28	6.952257e-25	1.6655425	1.7572380	1.7692441
ENSG00000211679	5.714563e-24	4.150487e-21	0.9026229	0.9168582	0.9442454

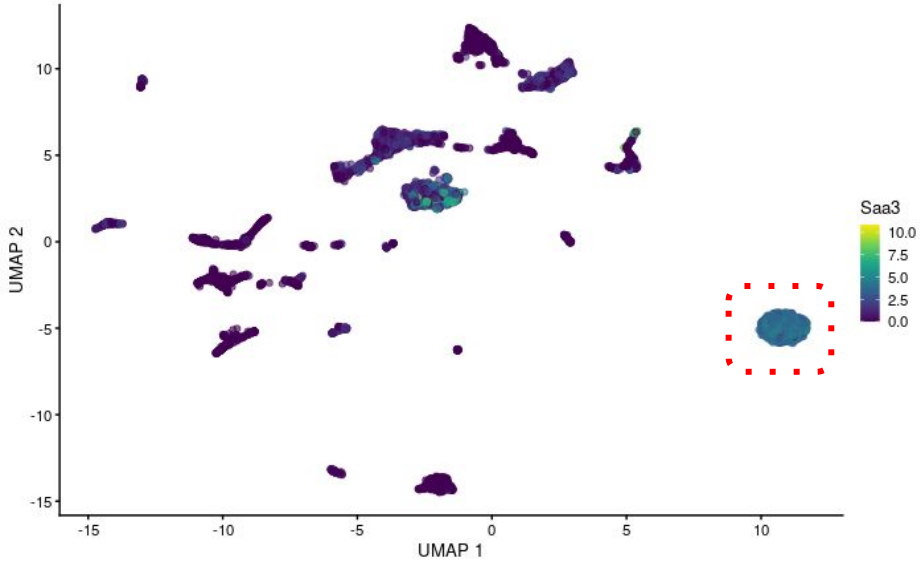
1-10 of 100 rows | 1-6 of 15 columns Previous 1 2 3 4 5 6 ... 10 Next

1 vs 2. 1 vs 3.

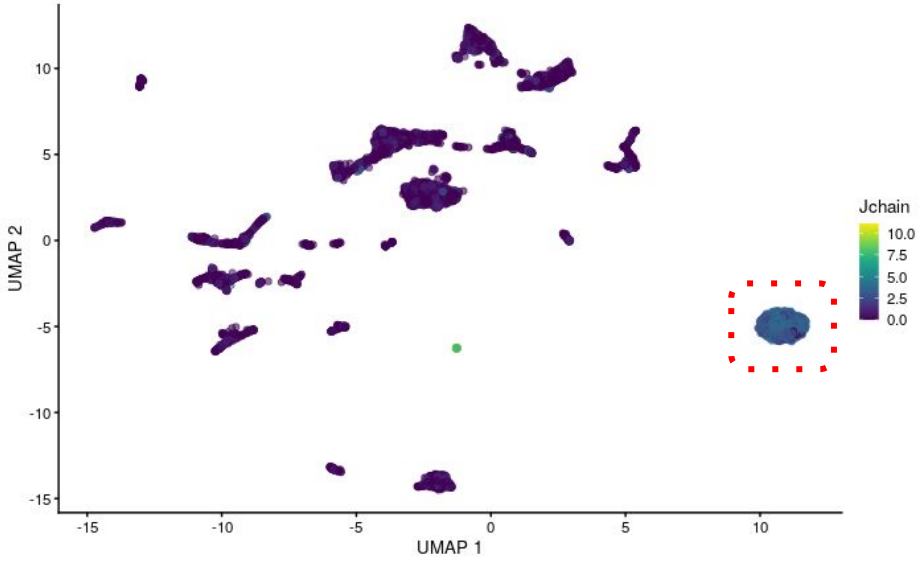
The summary log fold change here is the log fold change for the pairwise comparison with the largest p-value (e.g., weakest comparison). This is the choice the package makes, which may or may not be the right choice!

`scran::combineMarkers()` docs

What direction is the summary log FC for these genes?

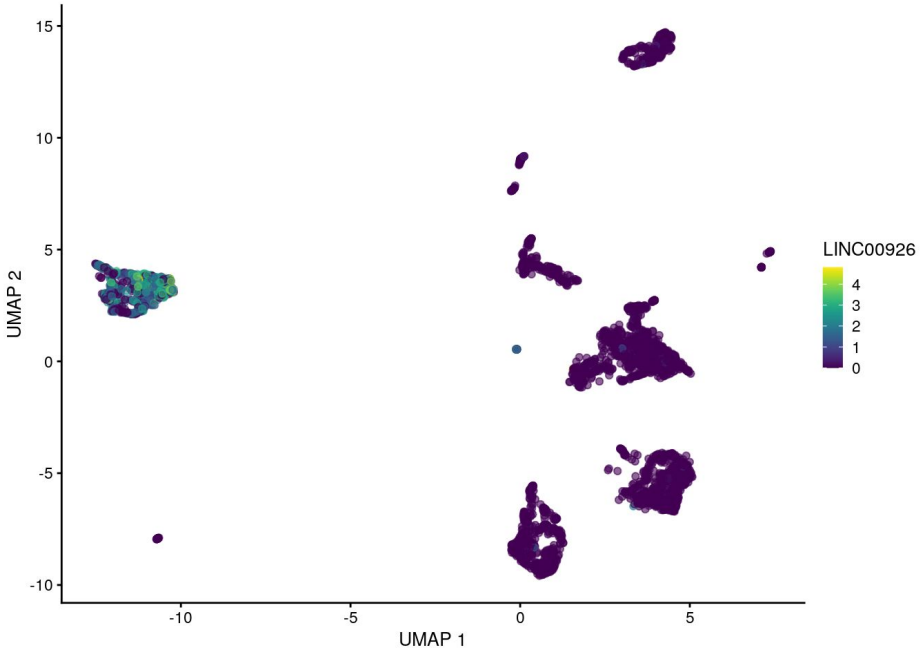


Positive



Negative

Marker genes: limitations & caveats



In practice, this means:

- Our marker genes table is sensitive to the underlying cluster assignments
- If overclustering occurs (e.g., many small clusters), we might “miss” genes because they may not uniquely define a single cluster when we set the p-value type to “all”

`scran::combineMarkers()` docs

Pathway analysis methods

Today we'll cover two types of pathway analysis

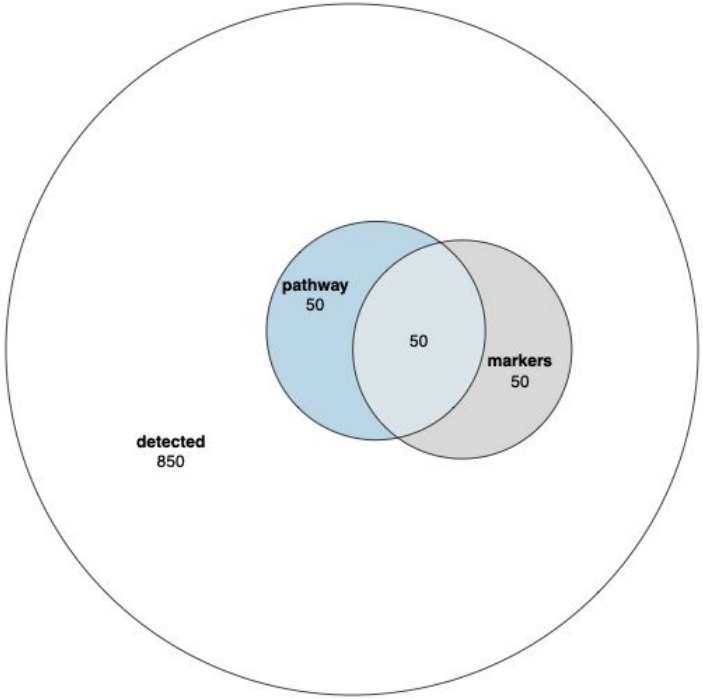
Over-representation analysis (ORA)

✅ **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 two types of pathway analysis

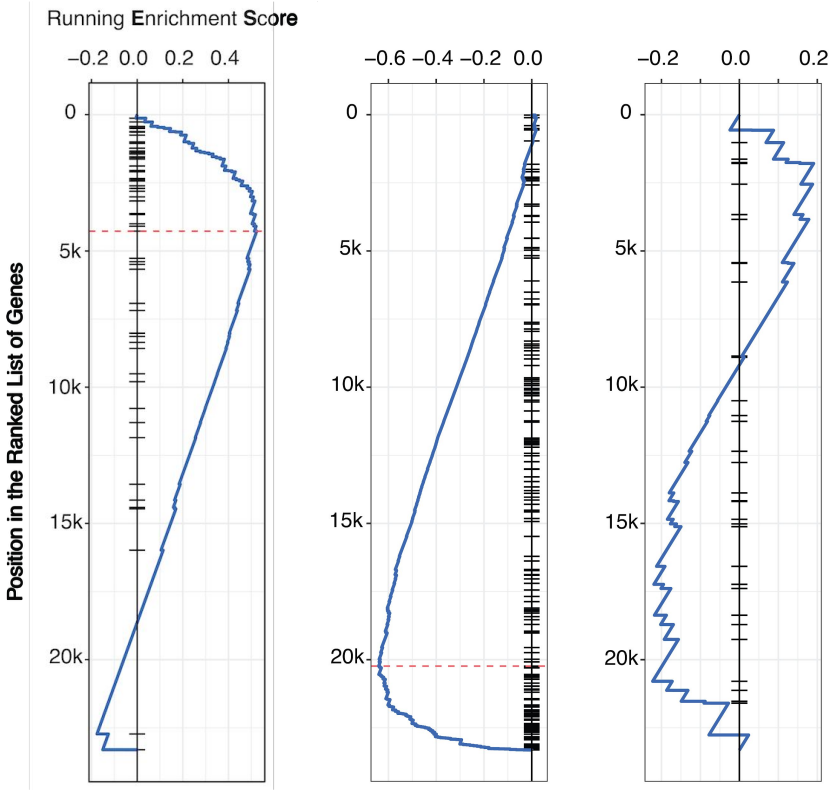
Gene Set Enrichment Analysis (GSEA)

✅ **Pros**

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

⚠️ **Cons**

- Gene-level metrics may be noisy for single-cell, making it difficult to assess small coordinated changes
- What gene-level metric to use is a bit of an open question
- May be more appropriate for comparing the same cell type across different samples



Subramanian et al. *PNAS*. 2005.