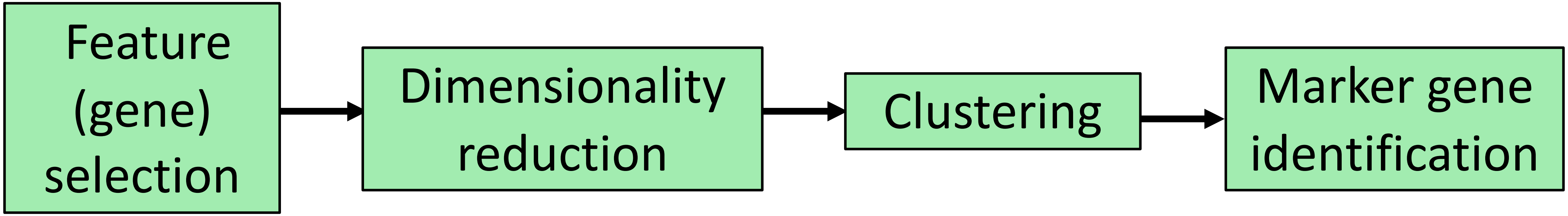
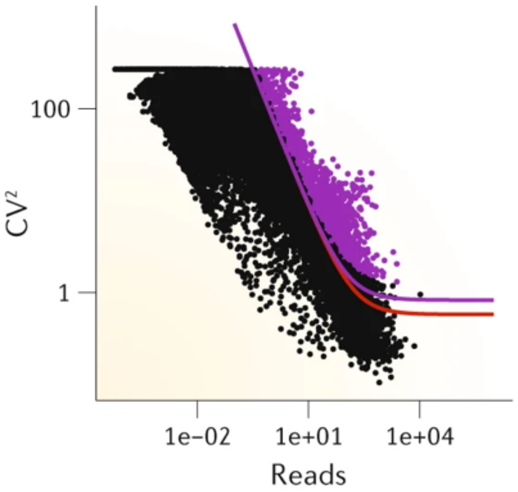
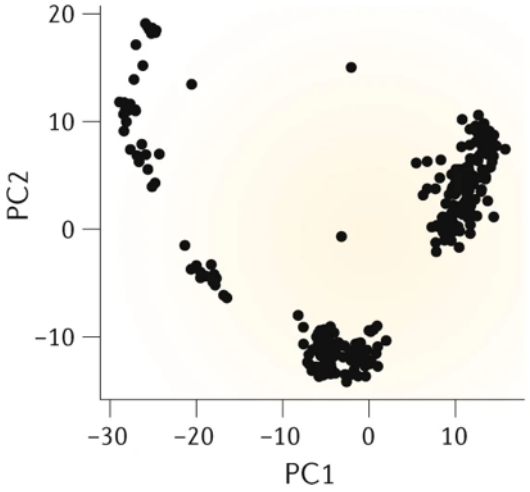
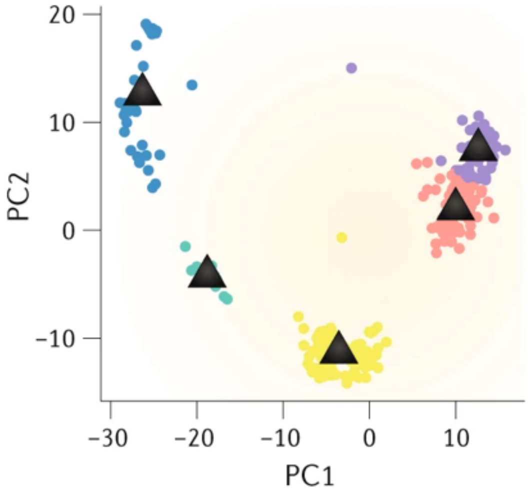
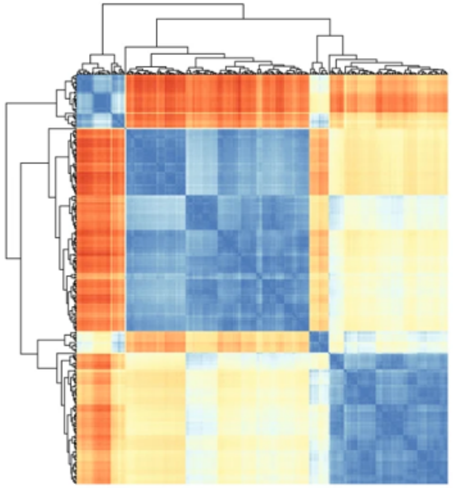


Feature (gene) selection → Dimensionality reduction → Clustering → Marker gene identification

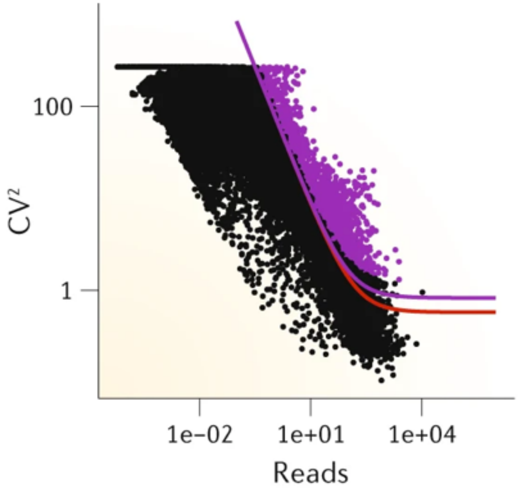


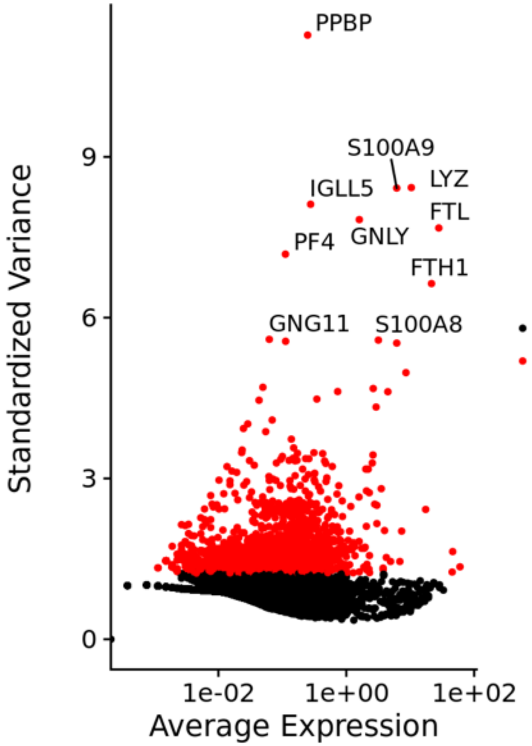






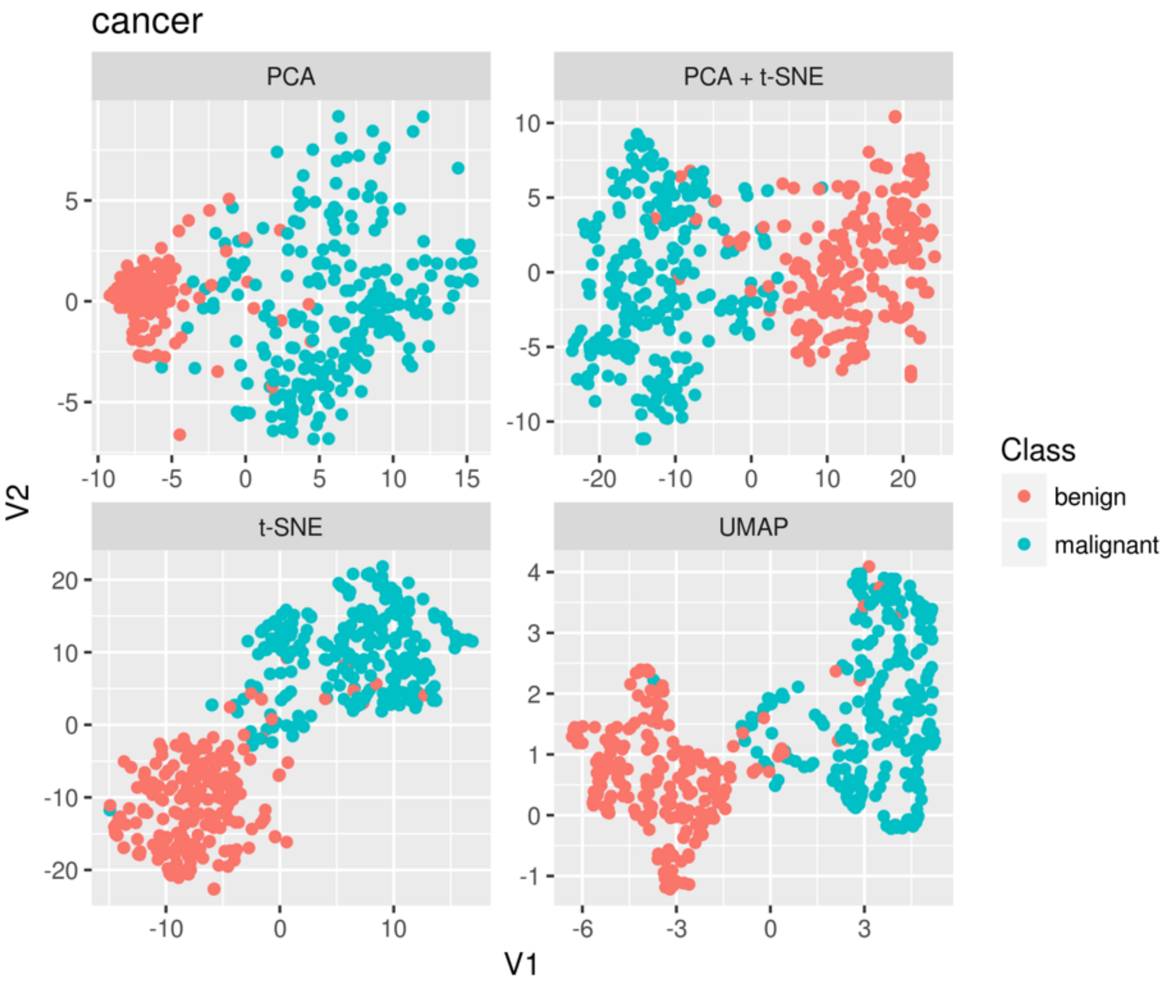
Feature (gene) selection





- Prior to dimensionality reduction, genes with highest expression variability are identified.
- Typically, 1000-5000 genes with the highest expression variability are selected.
- In robust workflows (e.g., Seurat and Scanpy), downstream analysis is not sensitive to the exact number of selected genes.
- Ideally, gene selection is done after batch correction.
- The goal is making sure genes variable only among batches (rather than cell groups within batches) do not dominate downstream results.

Dimensionality reduction of scRNA-Seq data



- scRNA-Seq data is **<u>inherently low-dimensional</u>**.

- Information in the data (expression variability among genes/cells) can be reduced from the number of total genes (1000s) to a much lower number of dimensions (10s).

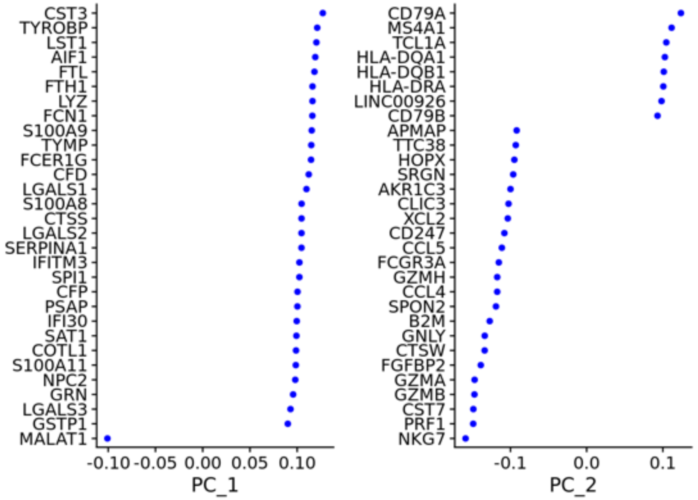
- Dimensionality reduction generates linear/non-linear combinations of gene expression vectors for clustering & visualization.

- Major dimensionality reduction techniques for scRNA-Seq:
  - Principal component analysis (PCA)
  - Most commonly used ones: UMAP and t-SNE (inputs: PCA results)
  - UMAPs typically preserve more of global structure with shorter run times
  - Other alternatives: Diffusion Maps & force-directed layout with k-nearest neighbors

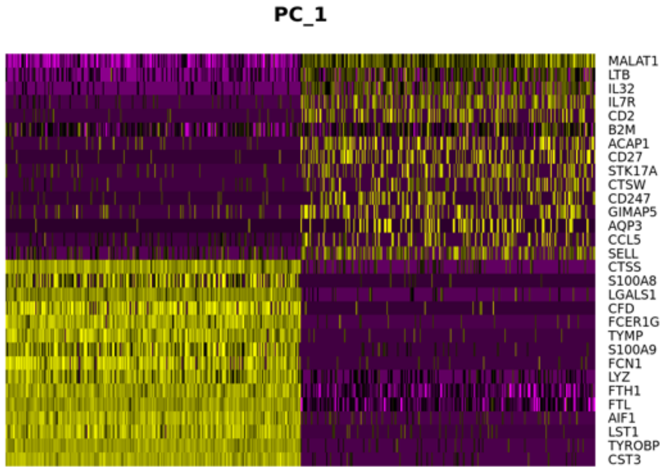
Scaling normalized data & performing PCA

Frederick National Laboratory for Cancer Research

- PCA is performed on the scaled data.
- Scaled data represented as z-scores.
- Mean=0 & variance=1 for each gene.
- z-scoring makes sure that highly-expressed genes do not dominate.

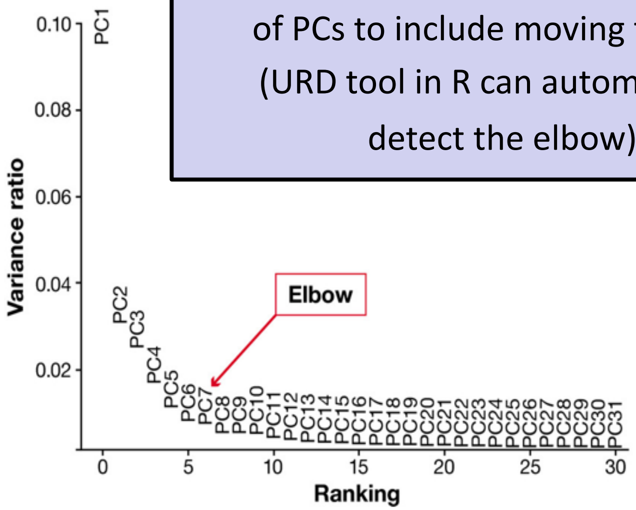


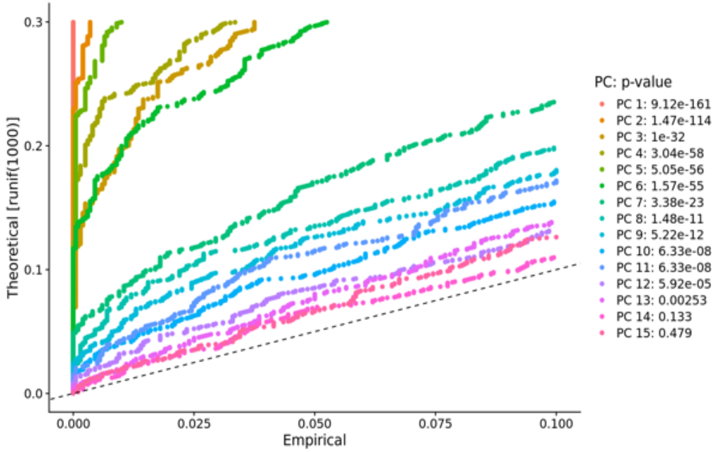
PC score plots show genes that dominate each PC



PC heatmaps visualize anti-correlated gene sets (yellow: higher expression)

Elbow plots show the number of PCs to include moving forward (URD tool in R can automatically detect the elbow).

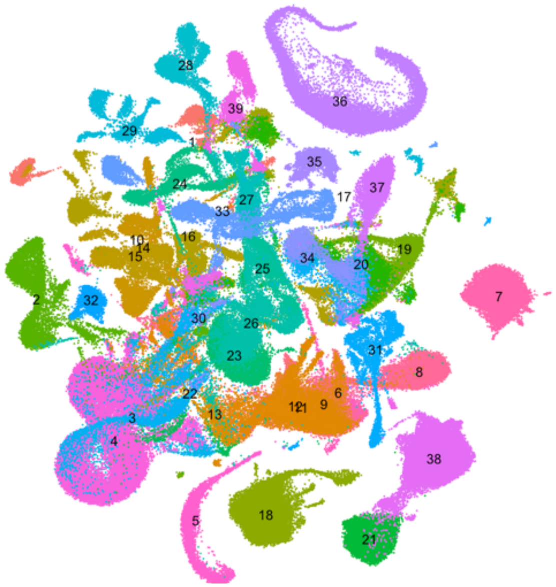


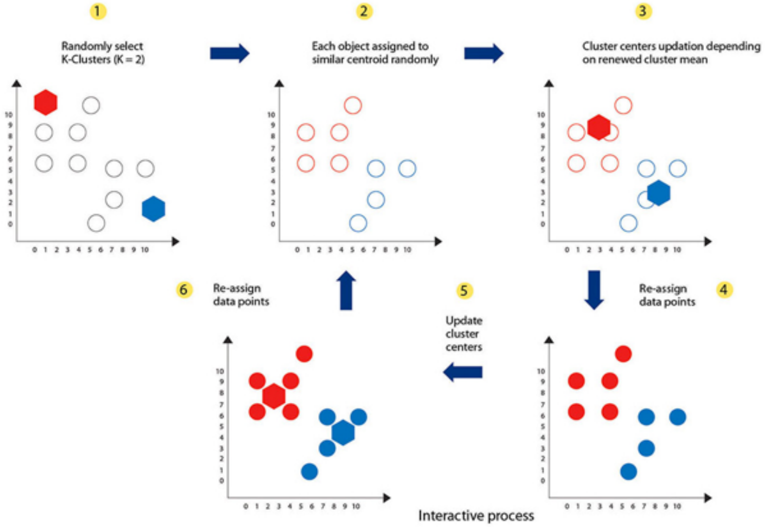


Jackstraw analysis generates a p-value (significance) of each PC 1% of the data is randomly permuted, PCA is rerun, ‘null distribution’ of gene scores constructed (these steps repeated many times).

‘Significant’ PCs have a strong enrichment of low p-value genes.

Clustering cells with similar expression profiles together




1
Randomly select K-Clusters (K = 2)
2
Each object assigned to similar centroid randomly
3
Cluster centers updation depending on renewed cluster mean
4
Re-assign data points
5
Update cluster centers
6
Re-assign data points
Interactive process
0 1 2 3 4 5 6 7 8 9 10

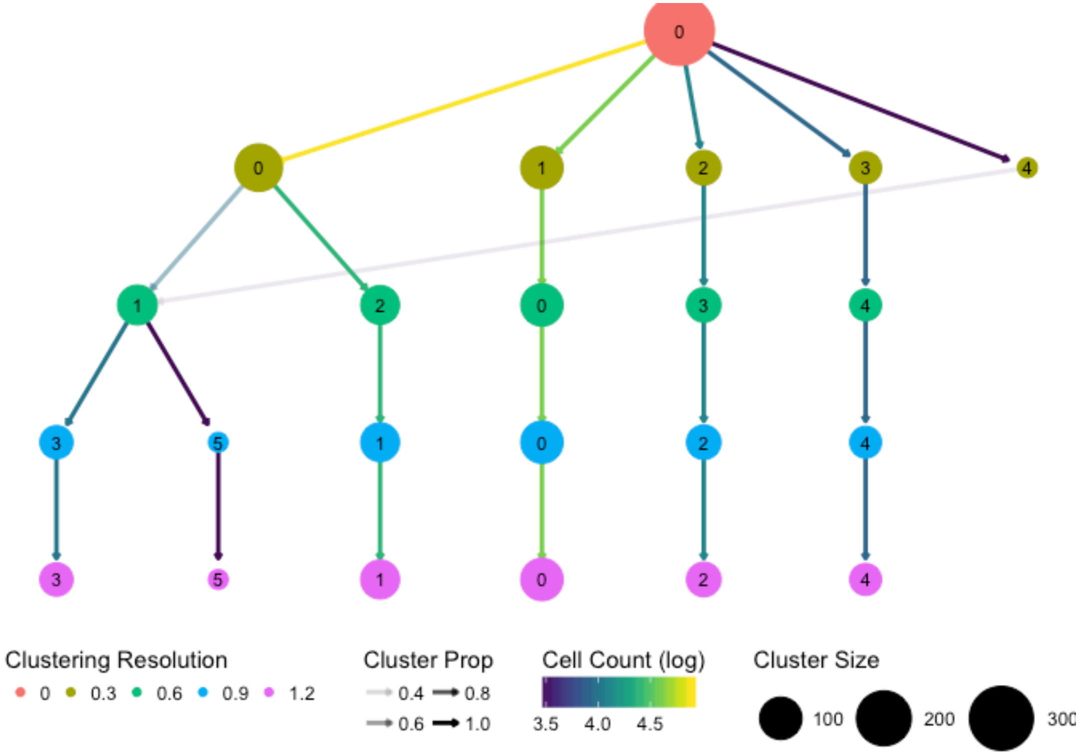
- Unsupervised machine learning problem
  - Input: distance matrix (cell-cell distances)
  - Output: Cluster membership of cells

- Cells grouped based on the similarity of their gene expression profiles
  - Distance measured in dimensionality-reduced gene expression space (scaled data)

- k-means clustering divides cells into k clusters
  - Determines cluster centroids
  - Assigns cells to the nearest cluster centroid
  - Centroid positions iteratively optimized (MacQueen, 1967).
  - Input: number of expected clusters (heuristically calibrated)

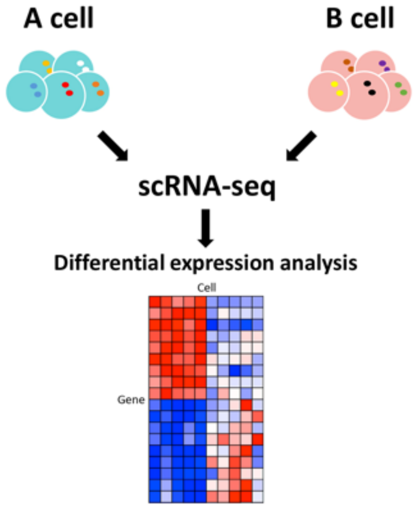
- k-means can be utilized with different distance metrics
- Alternatives to standard Euclidean distance:
  - Cosine similarity (Haghverdi et al, 2018)
  - Correlation-based distance metrics (Kim et al, 2018)
  - SIMLR method learns a distance metric using Gaussian kernels (Wang et al, 2017)

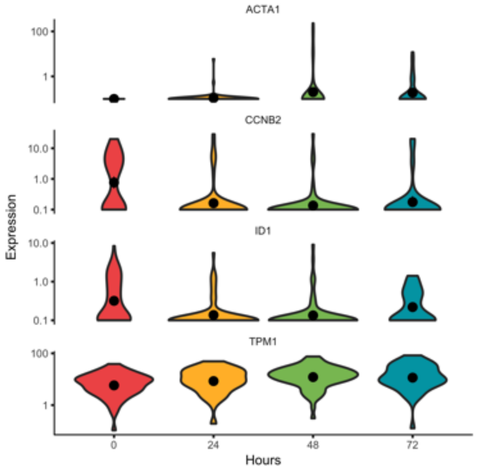
Number of clusters and biological context

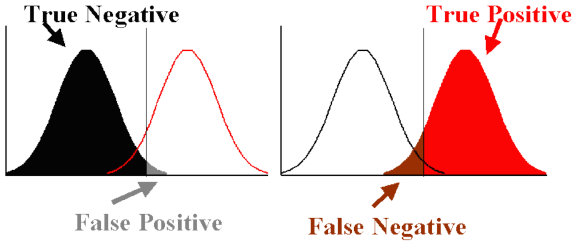


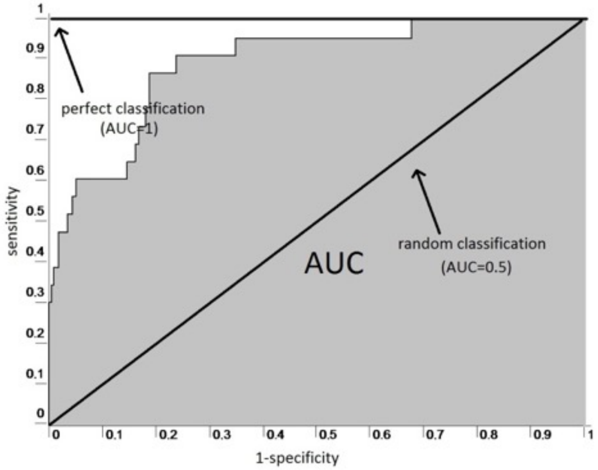
- Number of clusters is a function of the resolution parameter.
- Multiple resolution values can be explored to see the interplay between resolution and UMAP or t-SNE plots for a given data set.
- Biological context can be used for guidance.
- Examples: Expected number of major cell types or subtypes.
- Isolating a cluster to identify sub-clusters can generate useful biological insights (e.g., differential expression between cellular subtypes in a cluster).
- If cluster-specific markers for multiple clusters overlap (e.g., ribosomal genes), these clusters can be merged without losing much information regarding cell subtypes.

Marker gene identification









Differential expression approaches for marker identification:

- Wilcoxon rank sum test and student’s t-test
- Logistic regression
- DESeq2: Negative binomial generalized linear models (read counts) & Wald test for significance.
- MAST : GLMs in which cellular detection rate is treated as a covariate
- GLMs are flexible and do not make assumptions (homogenous distributions of residuals/fitting errors or normally distributed variances).

Classifier based approach for marker identification:

- Classifiers built with normalized expression levels (one classifier per gene).
- Genes ranked with respect to their ability of each gene to distinguish between two groups of cells (e.g. KO vs WT, cluster 1 vs 2, or cluster 1 vs all clusters).
- Area under each ROC curve represents the predictive power of the gene.

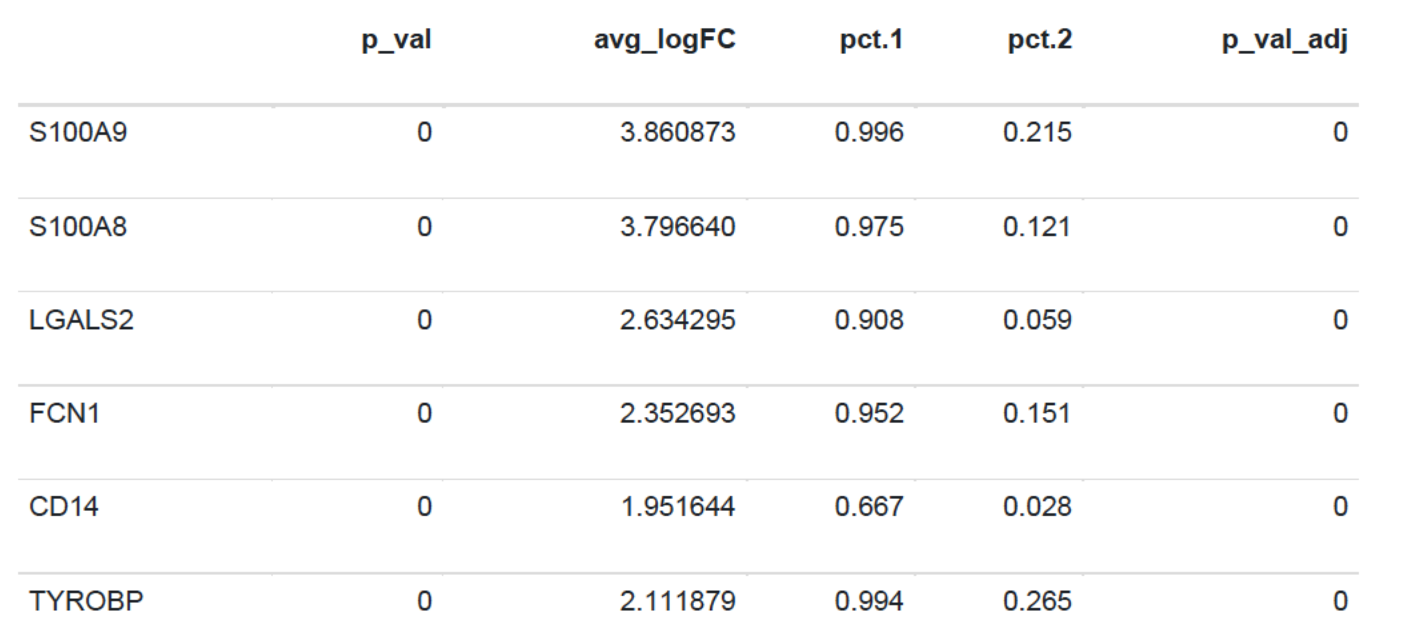
Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology. 2018 May;36(5):411.

A typical differential expression analysis output

Frederick National Laboratory for Cancer Research

Comparing gene expression in different cell groups:

Cluster 1 vs Cluster 2
Cluster 1 vs all other clusters
Cluster1_KO vs Cluster1_WT



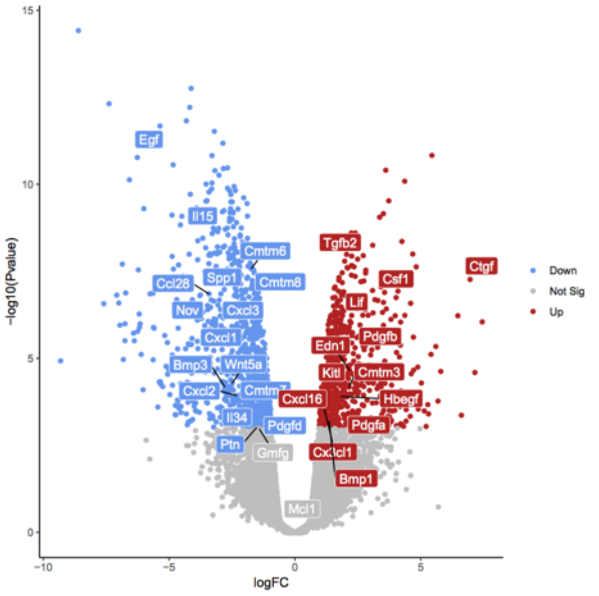
	p_val	avg_logFC	pct.1	pct.2	p_val_adj
S100A9	0	3.860873	0.996	0.215	0
S100A8	0	3.796640	0.975	0.121	0
LGALS2	0	2.634295	0.908	0.059	0
FCN1	0	2.352693	0.952	0.151	0
CD14	0	1.951644	0.667	0.028	0
TYROBP	0	2.111879	0.994	0.265	0

Average log FC
Ratio of expression in log-space

pct.1= percent of cells in Cluster 1 in which the gene is detected

pct.2=percent of cells in Cluster 2

p_val_adj=FDR



Tips for marker identification

Classifier based marker identification: AUC values replace the p-values

- Marker identification can take time with thousands of cells and genes
- Prefiltering cells and genes can reduce the computational time significantly
- Genes rarely detected in either group of cells, are not likely to be differentially expressed
- Genes with small fold-change can also be excluded
- Typically, only upregulated genes (>1 FC) are relevant for cluster-specific marker discovery