UCLA QCBio Collaboratory

W20: Single Cell RNA-Sequencing Analysis using R

Giovanni Quinones Valdez, PhD

November 26th, 2024

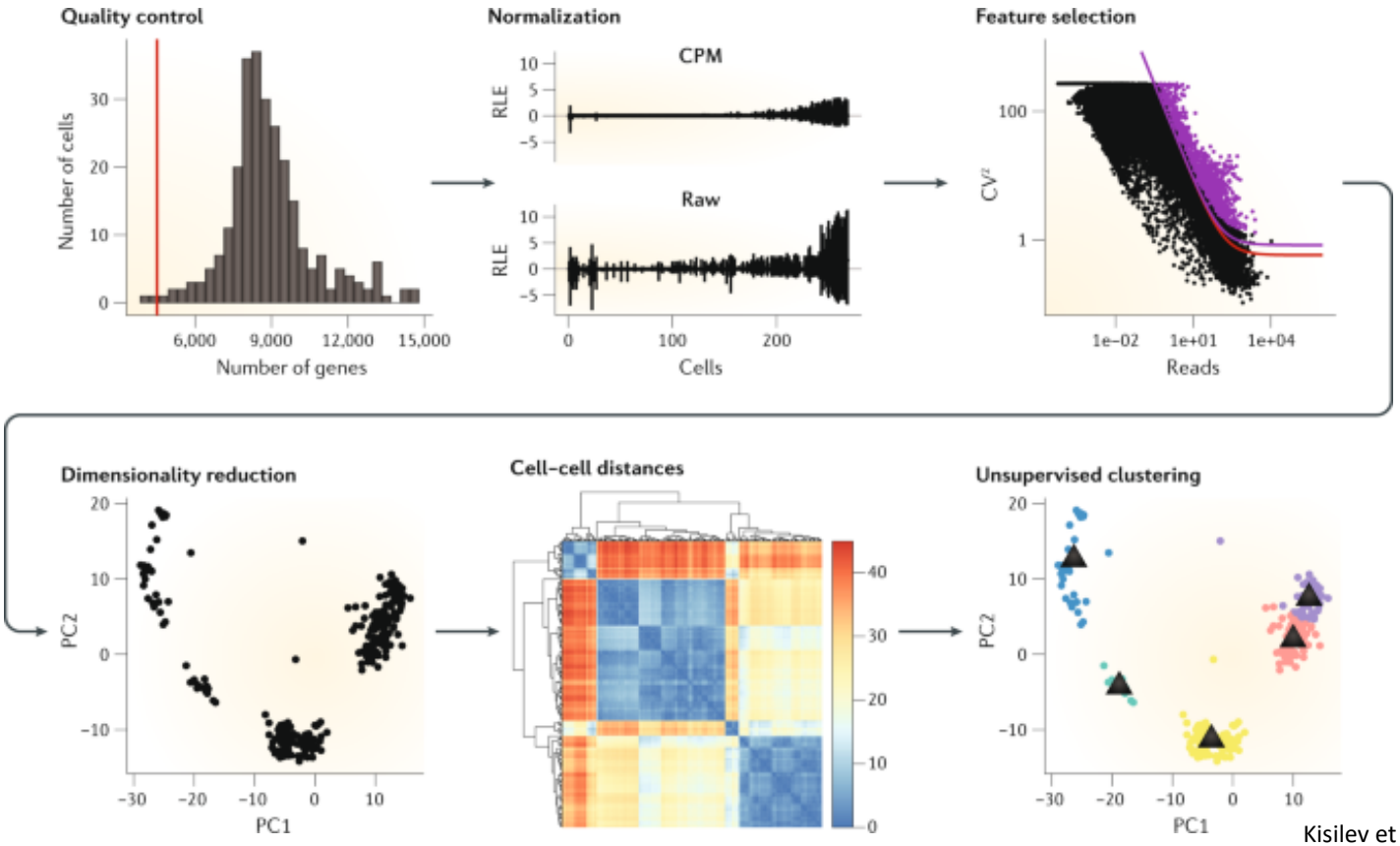
W20: Single Cell RNA-Sequencing Analysis using R

Day 2

- Data processing
- Clustering and visualization
- Cell annotation

UCLA QCBio Collaboratory

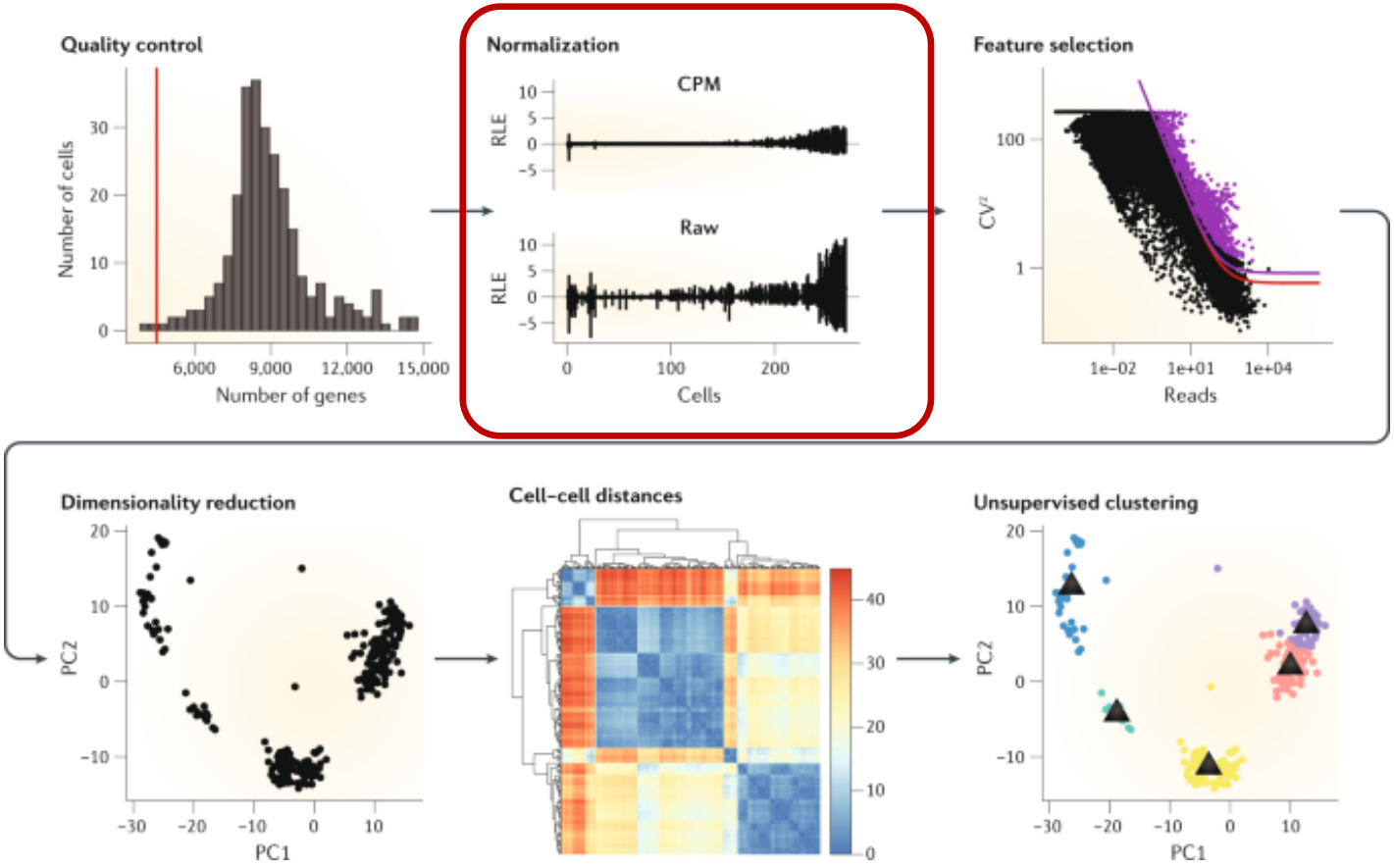
Data analysis



Kisilev et al 2019

UCLA QCBio Collaboratory

Normalization



UCLA QCBio Collaboratory

Normalization

- The goal is to make the data comparable across cells.
- The main assumption is that ***all cells have the same number of mRNAs***
- We use counts per million (CPM) as out new unit of expression.
- For gene *i* in cell *j:* r_{ij} = gene UMI count, R_j = total UMI count, sf = scale factor (1,000,000)

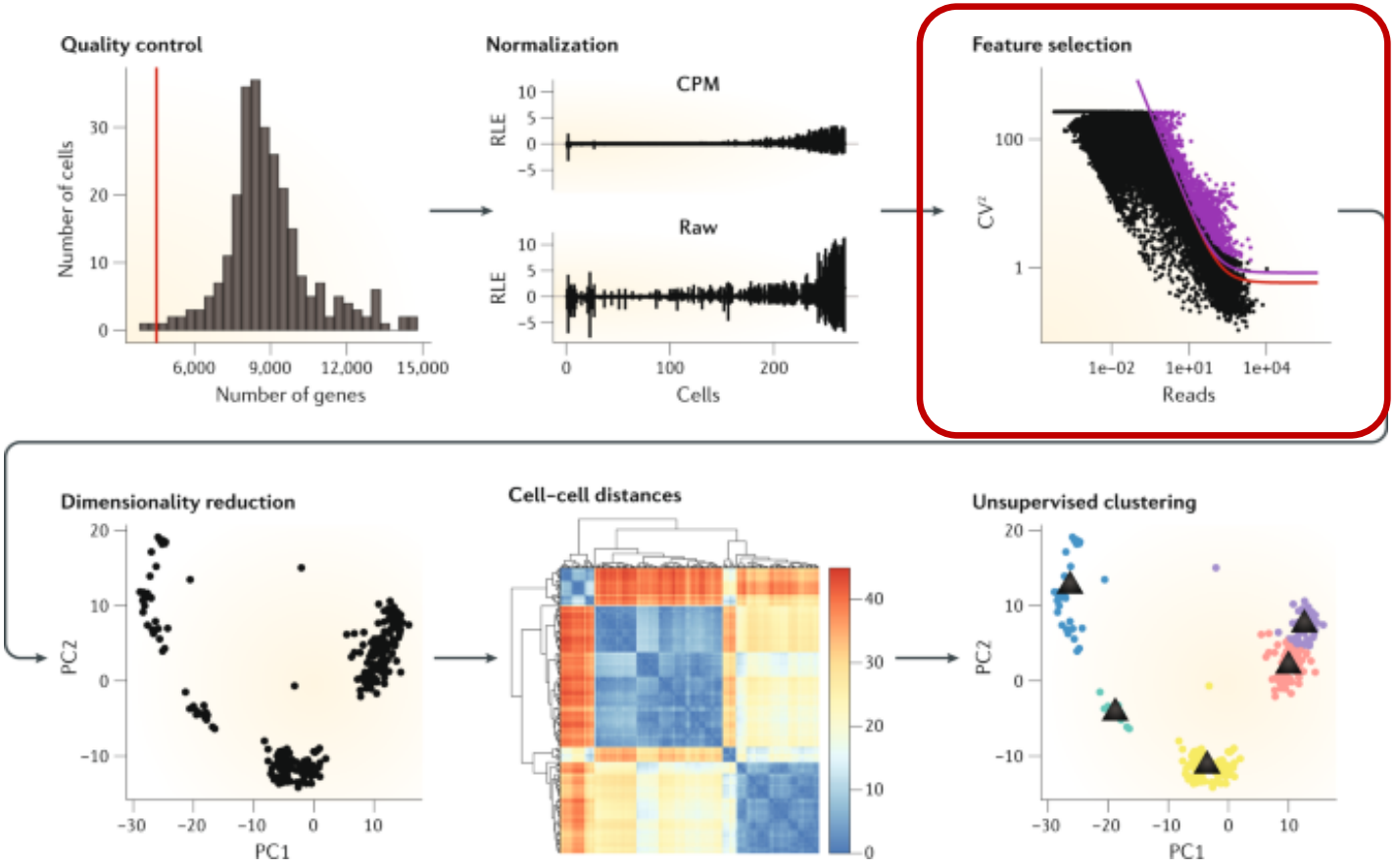
$$CPM_{ij} = \frac{r_{ij}}{R_j} \times sf$$

Normalization

- Another technique is to down-sample UMIs until all cells have the same count.
- Log transformation makes the data comparable (mean and variance) when it spans multiple orders of magnitude.
- Proper normalization is still an ongoing question.

$$\log(CPM_{ij} + 1)$$

Feature selection (HVG)

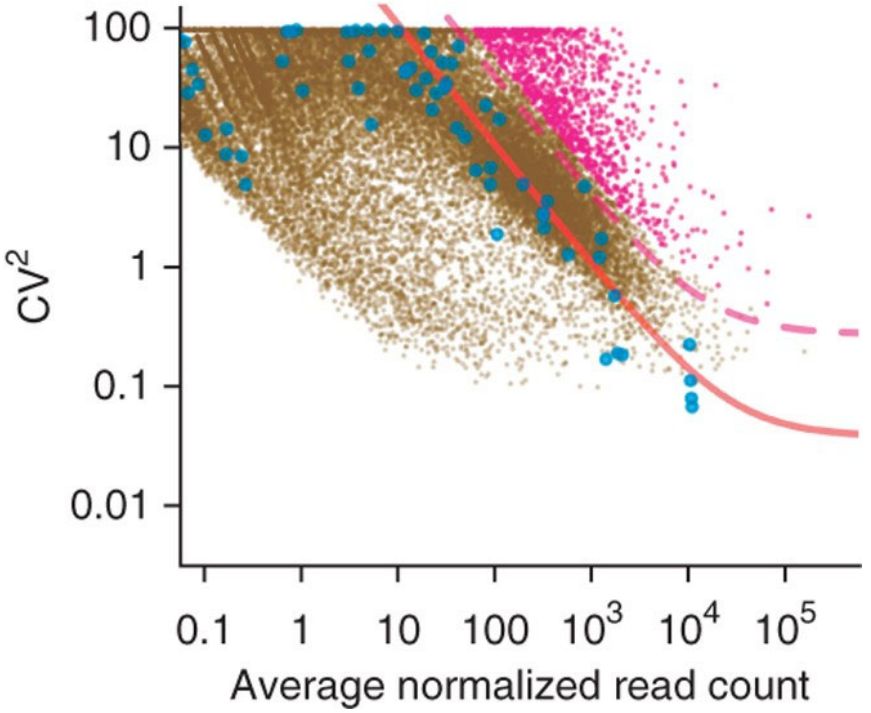


UCLA QCBio Collaboratory

Highly variable genes (HGV)

- Expression is less reproducible at lower values.
- Variance (CV^2) is modelled as a function of mean expression in order to identify genes that 'stand out' as highly variable

- $CV = \frac{\sigma}{\mu} = \frac{Std.dev.}{mean}$



Kiselev 2019

Highly variable genes (HGV)

- Recommendations:
  - Select 2k-5k most variable genes for downstream analysis.
  - This is an arbitrary cutoff.
  - Needs careful consideration.
  - A good practice is to try different cutoffs and evaluate consistency of results.

Normalization → Scaling

	C1	C2	C3
G1	40	4	2
G2	10	1	0
G3	20	2	1
G4	30	3	2
G5	0	0	0

Raw counts

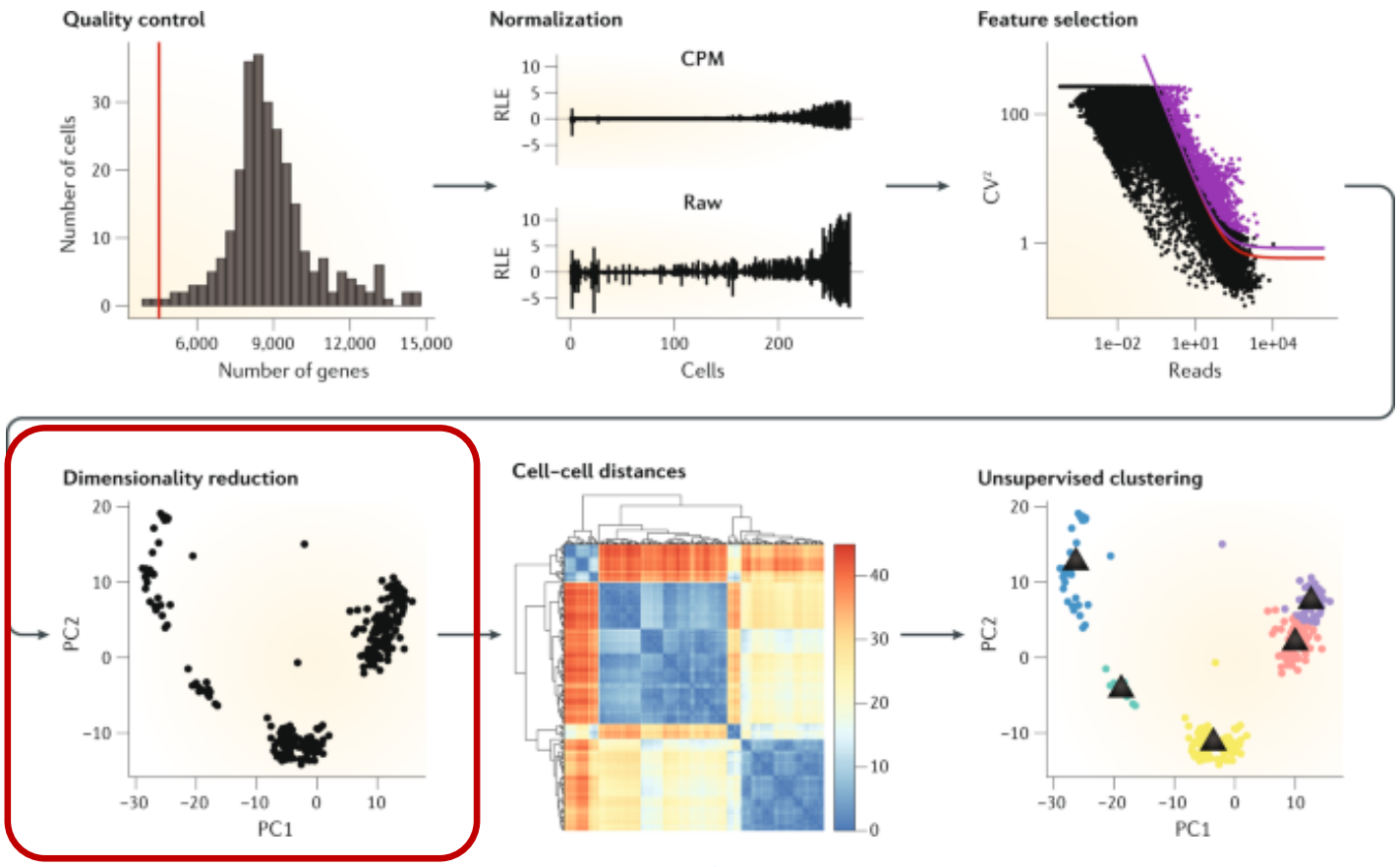
	C1	C2	C3
G1	0.4	0.4	0.4
G2	0.1	0.1	0.0
G3	0.2	0.2	0.2
G4	0.3	0.3	0.4
G5	0.0	0.0	0.0

Normalized

	C1	C2	C3
G1	0.0	0.0	0.0
G2	0.33	0.33	-0.67
G3	0.0	0.0	0.0
G4	-0.03	-0.03	0.07
G5	0.0	0.0	0.0

Scaled

Dimensionality reduction



UCLA QCBio Collaboratory

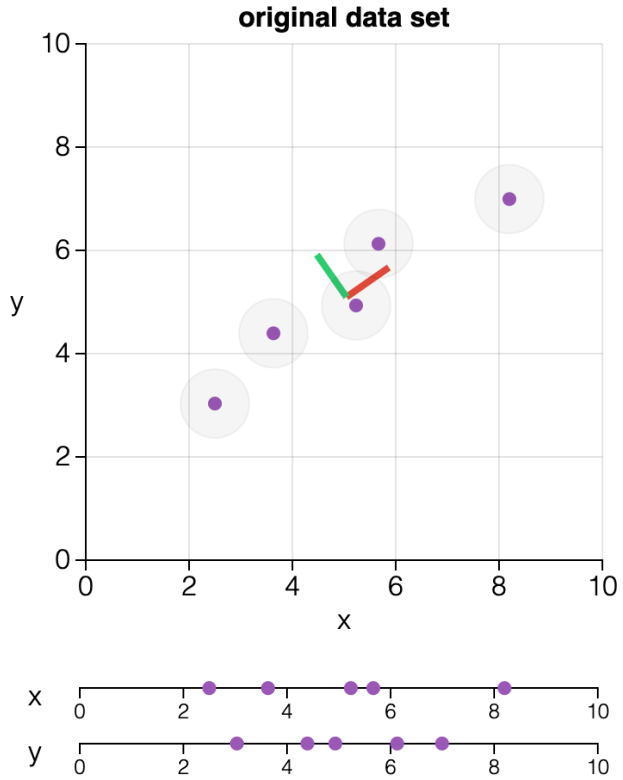
Dimensionality reduction: PCA

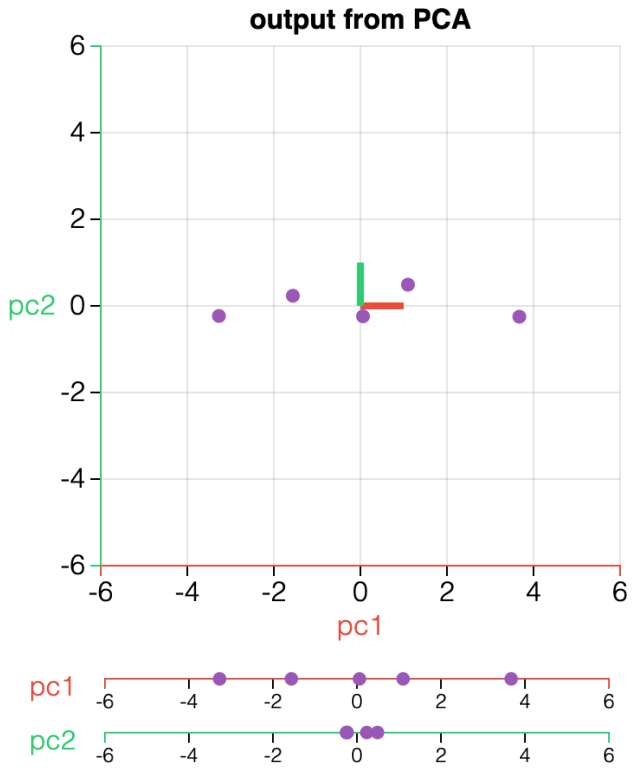
- Our data is multi-dimensional (many genes).
- Genes are expressed in “networks”.
- It is not practical/feasible to analyze data with so many dimensions.
- Instead, we can find a metagene which represents all genes.

	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0

Expression Matrix

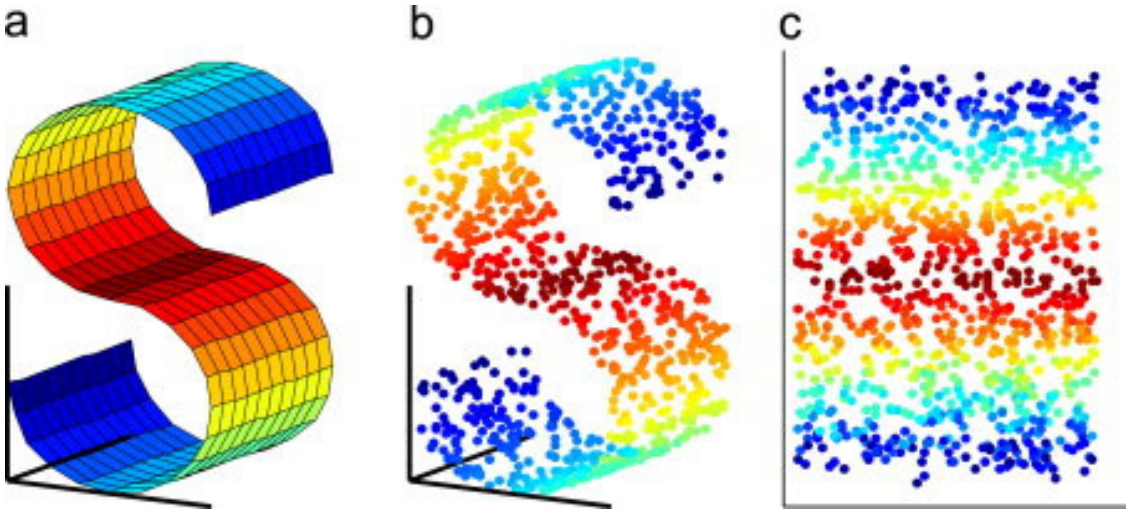
Dimensionality reduction: PCA





UCLA QCBio Collaboratory

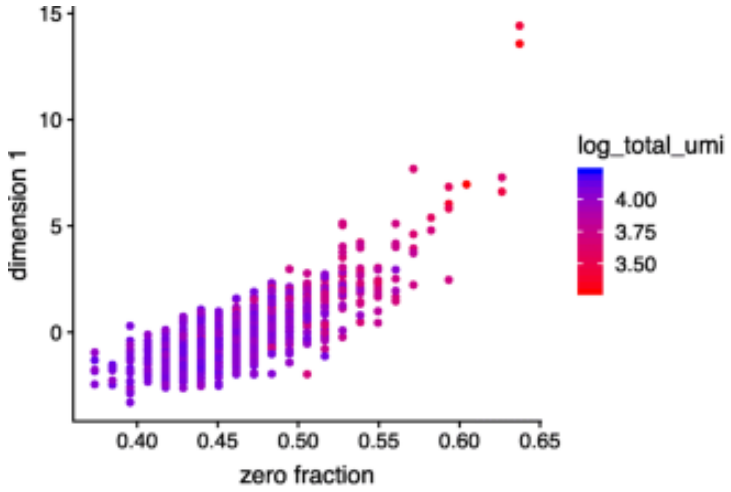
Limitations of PCA



Jeremy Kalfon, 2019

By reducing the number of dimensions, it is inevitable to lose some of the connectivity of points in the original structure (manifold)

PCA and drop out

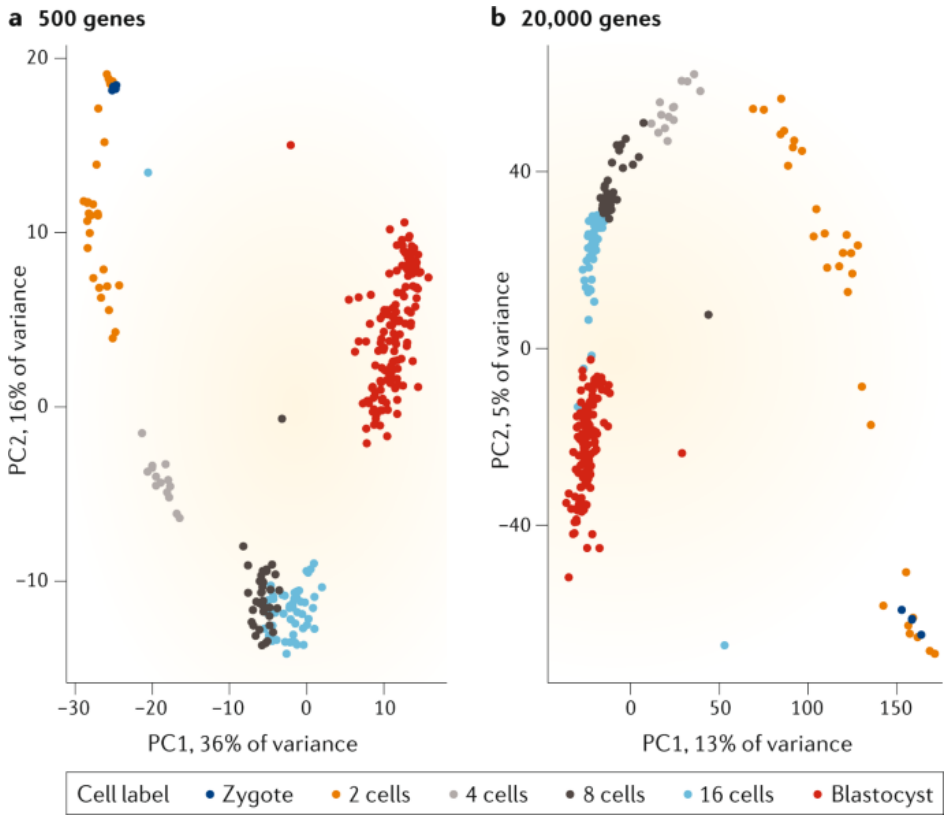


(a) Technical replicates, PCA

Townes et al 2019

- PC1 strongly correlates with the fraction of zeros.

Feature selection improves PCA



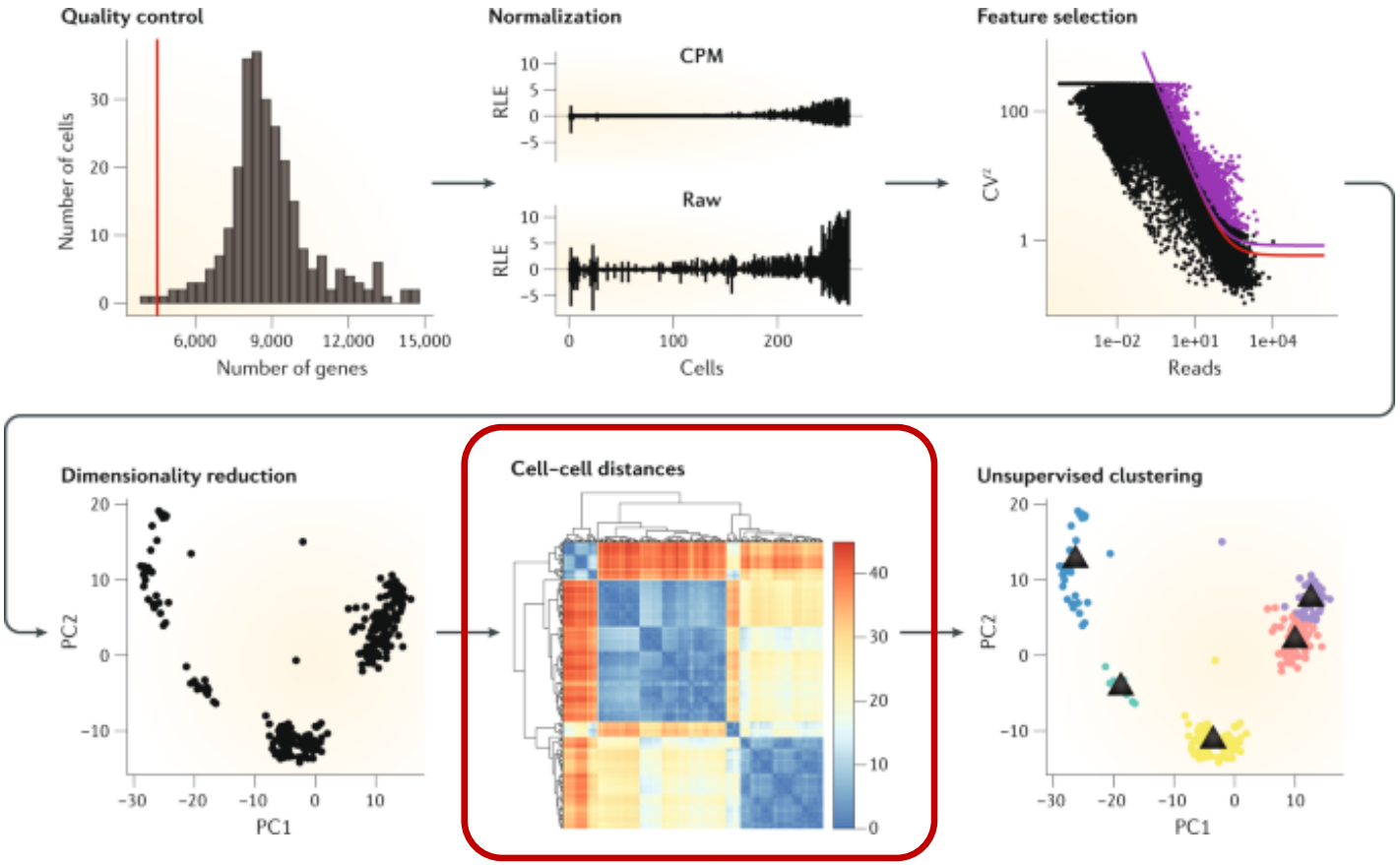
Kiselev 2019

UCLA QCBio Collaboratory

Dimensionality reduction: PCA

- How many PC's should I select?
  - Jackstraw
  - Elbow plot
- Similarly, as Feature Selection, the number of PC's should be carefully evaluated.

Cell-to-cell distances



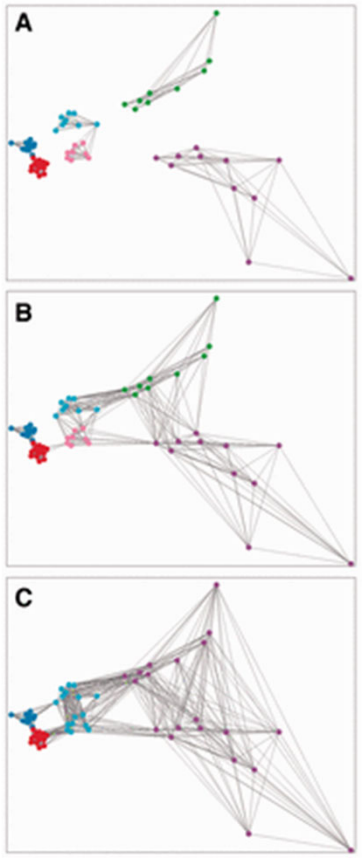
UCLA QCBio Collaboratory

Finding neighboring cells (K-NN)

- Identify and connect k nearest neighbors.
- The distance is defined as the Euclidian distance in the PC space.
- Find “neighbor similarity” using the Jacard Index.

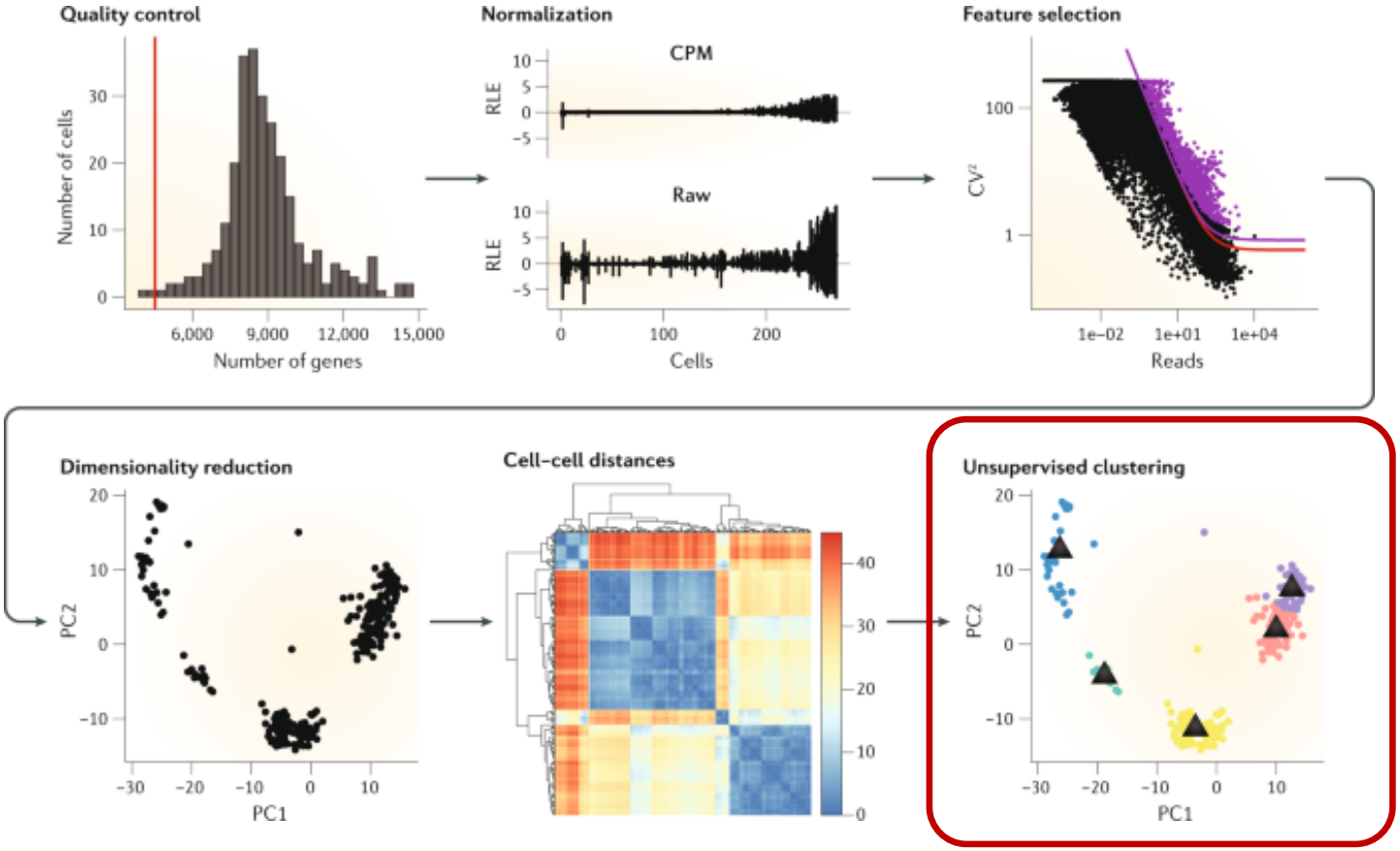
- $J = \frac{N(C_1) \cap N(C_2)}{N(C_1) \cup N(C_2)}$

- Louvain algorithm → optimal modularity
  - It maximizes the modularity based on the weight of the edges



Xu et al (2015)

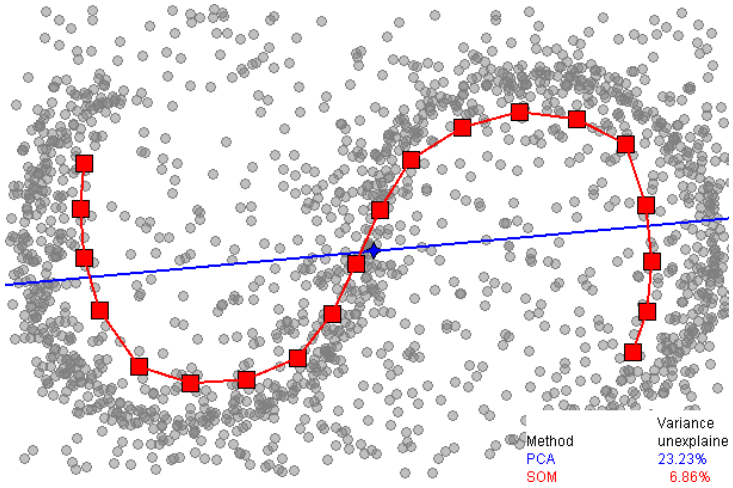
Clustering



UCLA QCBio Collaboratory

Uniform Manifold Approximation and Projection (UMAP)

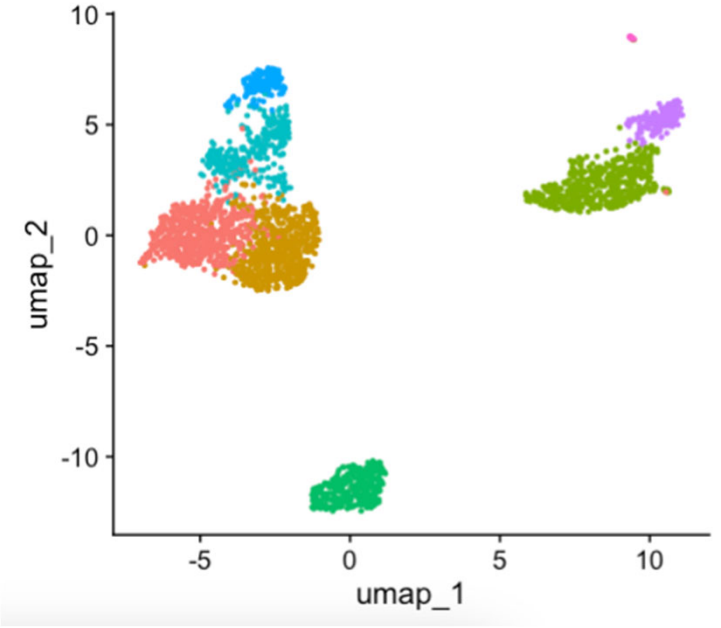
- Also a dimensionality reduction technique.
- Used predominantly for visualization.
- It aims to capture the manifold (topology/shape/structure) of the data organization in higher dimensions.
- It embeds the neighborhood of points to not lose significant connectivity.



E.M. Mirkes 2011

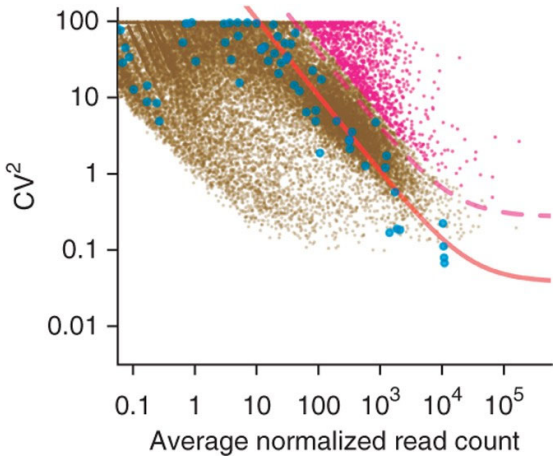
Uniform Manifold Approximation and Projection (UMAP)

- It aims to capture the manifold (topology/shape/structure) of the data organization in higher dimensions.



Identification of marker genes

- Which genes give identity to each cluster?
- Marker genes are genes that are differentially expressed in one cell population (or cell type) compared to all the rest.
- Seurat performs DGE using the Wilcox Rank Sum test.

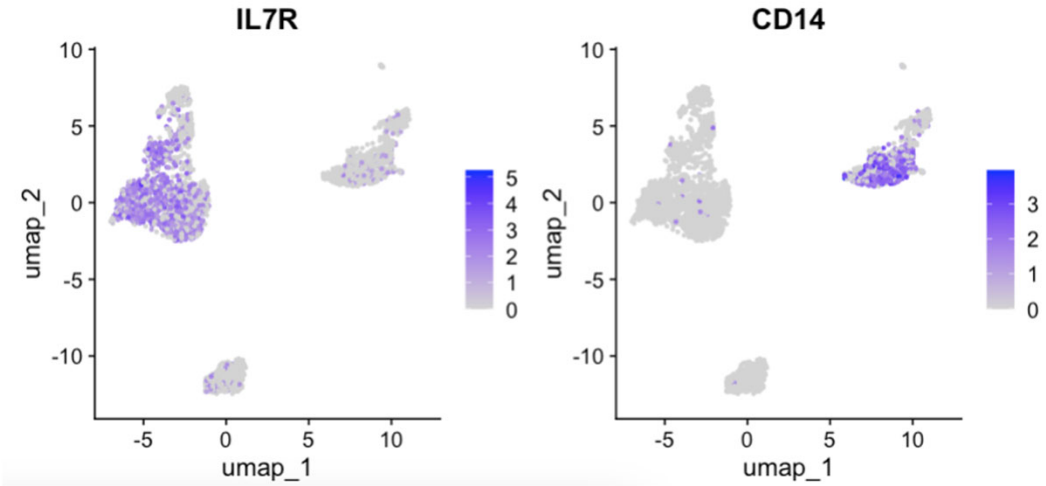


Kiselev 2019

What is a “differentially-expressed” gene?

Identification of marker genes

- Which genes give identity to each cluster?
- **Marker genes** are genes that are differentially expressed in one cell population (or cell type) compared to all the rest.
- Seurat performs DGE using the Wilcox Rank Sum test.



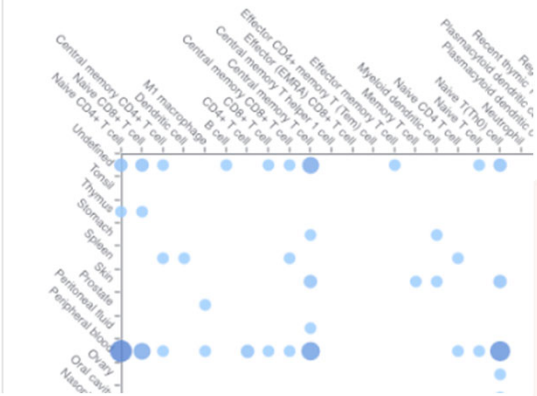
UCLA QCBio Collaboratory

Marker gene annotation

CellMarker 2.0

Statistical graph of cell markers

Key Word: CCR7

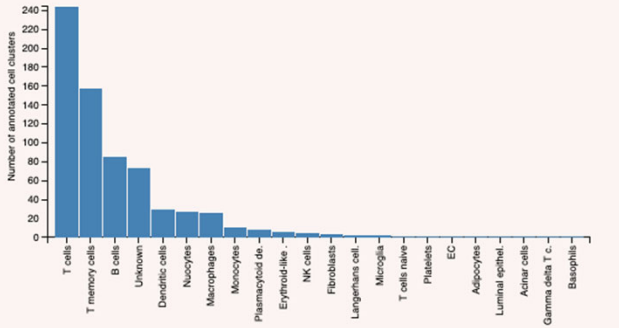


PanglaoDB.se

Summary of search results

Gene	Description	Type	No. samples	No. cell clusters
IL7R	interleukin 7 receptor	protein-coding gene	171	450
CCR7	C-C motif chemokine receptor 7	protein-coding gene	91	237

Barplot of cell clusters (Y-axis) and cell types (X-axis) where the gene is expressed.



ChatGPT 4

You

given markers IL7R, CCR7 from a peripheral blood sample (human), what is the most likelhy cell type?

ChatGPT

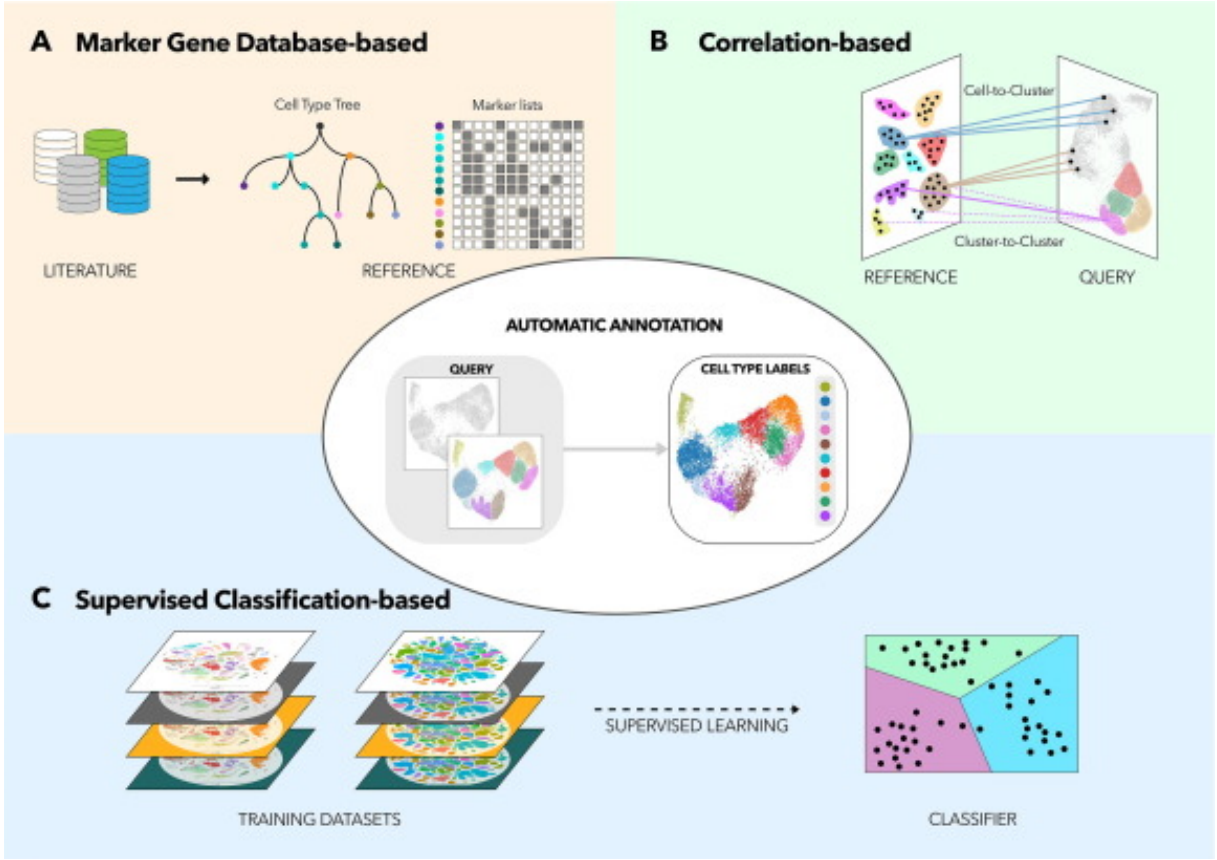
Based on the expression of the markers IL7R and CCR7 from a human peripheral blood sample, the most likely cell type would be naive T cells. Specifically, these markers are often associated with naive CD4+ and CD8+ T cells. Here's a bit more detail:

- **IL7R (CD127)**: This interleukin-7 receptor alpha chain is critical for the survival and development of T cells. It is prominently expressed on naive T cells and memory T cells but is typically low or absent on regulatory T cells (Tregs) and most differentiated effector T cells.
- **CCR7**: The C-C chemokine receptor type 7 is involved in homing cells to lymph nodes through its ligands CCL19 and CCL21. CCR7 is highly expressed on naive T cells and central memory T cells, facilitating their migration to secondary lymphoid tissues.

The co-expression of IL7R and CCR7 indicates a naive T cell phenotype, suggesting these cells are capable of responding to new antigens and maintaining long-term immune surveillance.

Marker gene annotation

Derive cell-cluster marker from a cell type hierarchical model



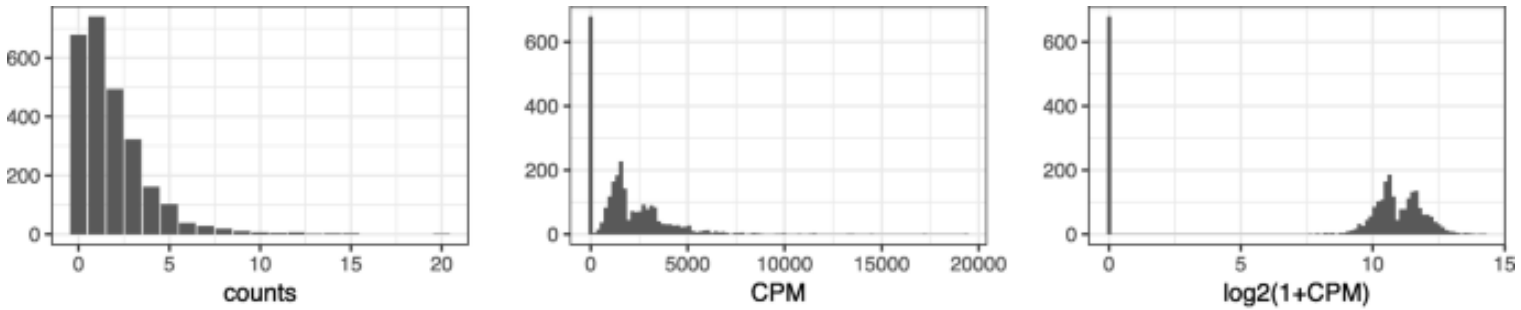
Correlate expression levels between query and annotated clusters

Machine-Learning based approaches to classify cell clusters and overcome batch variability

Pasquini et al, 2021

Normalization

- Normalization can dramatically change the shape of the distribution



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Townes et al (2019)

Check:

- Sina (2021)

Seurat's SCTransform

Load data and create Seurat object

```
pbmc_data <- Read10X(data.dir = "../data/pbmc3k/filtered_gene_bc_matrices/hg19/")
pbmc <- CreateSeuratObject(counts = pbmc_data)
```

Apply sctransform normalization

- Note that this single command replaces `NormalizeData()`, `ScaleData()`, and `FindVariableFeatures()`.
- Transformed data will be available in the SCT assay, which is set as the default after running sctransform
- During normalization, we can also remove confounding sources of variation, for example, mitochondrial mapping percentage

```
# store mitochondrial percentage in object meta data
pbmc <- PercentageFeatureSet(pbmc, pattern = "^MT-", col.name = "percent.mt")

# run sctransform
pbmc <- SCTransform(pbmc, vars.to.regress = "percent.mt", verbose = FALSE)
```

https://satijalab.org/seurat/articles/sctransform_vignette.html

SCTransform normalization

Negative Binomial modelling

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m,$$

Pearson residual's

$$\begin{aligned} z_{ij} &= \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}}, \\ \mu_{ij} &= \exp(\beta_{0_i} + \beta_{1_i} \log_{10} m_j), \\ \sigma_{ij} &= \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}}, \end{aligned}$$

- A NB GLM model is fit for every gene.
- The Pearson residuals is the difference between the modelled and observed value
- This substitutes the scaled and log normalized value

Hafemeister 2019

Further considerations

- **scDeeds** = single-cell dubious embeddings detector
  - https://github.com/JSB-UCLA/scDEED
- **ClusterDE =** The R package **ClusterDE** is a post-clustering DE method for controlling the false discovery rate (FDR)
  - https://github.com/JSB-UCLA/ClusterDE
- **scDesign3 =** The R package **scDesign3** is an all-in-one single-cell data simulation tool
  - https://github.com/JSB-UCLA/scDesign3

The Junction of
Statistics and Biology
UNIVERSITY OF CALIFORNIA, LOS ANGELES

UCLA QCBio
Collaboratory

References

- Jiang, R., Sun, T., Song, D. *et al.* Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol* **23**, 31 (2022). https://doi.org/10.1186/s13059-022-02601-5
- Kiselev, V.Y., Andrews, T.S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20**, 273–282 (2019). https://doi.org/10.1038/s41576-018-0088-9
- Erica A.K. DePasquale, Daniel J. Schnell, Pieter-Jan Van Camp, Íñigo Valiente-Alandí, Burns C. Blaxall, H. Leighton Grimes, Harinder Singh, Nathan Salomonis. DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Reports.* **29**, 1718-1727 (2019). https://doi.org/10.1016/j.celrep.2019.09.082.
- Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, Volker Busskamp. Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*. **19**, 961-969, (2021). https://doi.org/10.1016/j.csbj.2021.01.015.
- A Sina Booeshaghi , Lior Pachter, Normalization of single-cell RNA-seq counts by $\log(x + 1)$ or $\log(1 + x)$, *Bioinformatics*, Volume 37, Issue 15, 1 August 2021, Pages 2223–2224, https://doi.org/10.1093/bioinformatics/btab085
- Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019). https://doi.org/10.1186/s13059-019-1874-1