# Clustering and Cluster Marker Identification

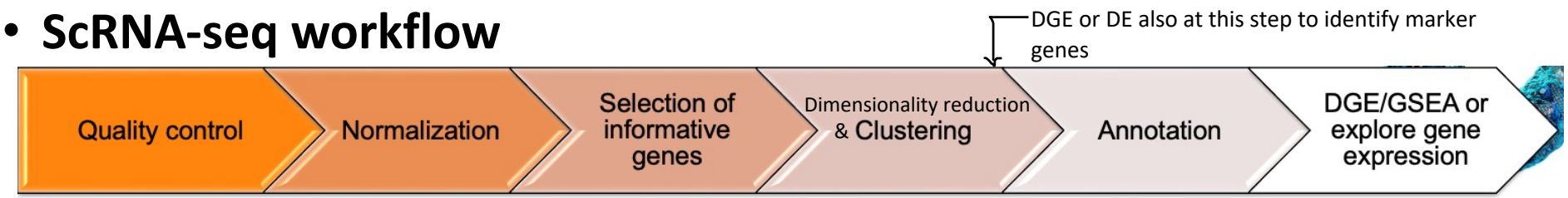
Sadiksha Adhikari

06.10.2022

Nemhesys single cell RNA sequencing workshop

# Recap?

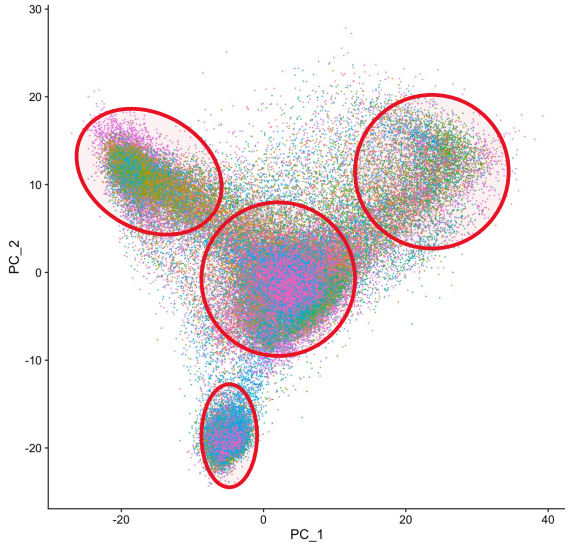
- **ScRNA-seq workflow**




- **PCA**

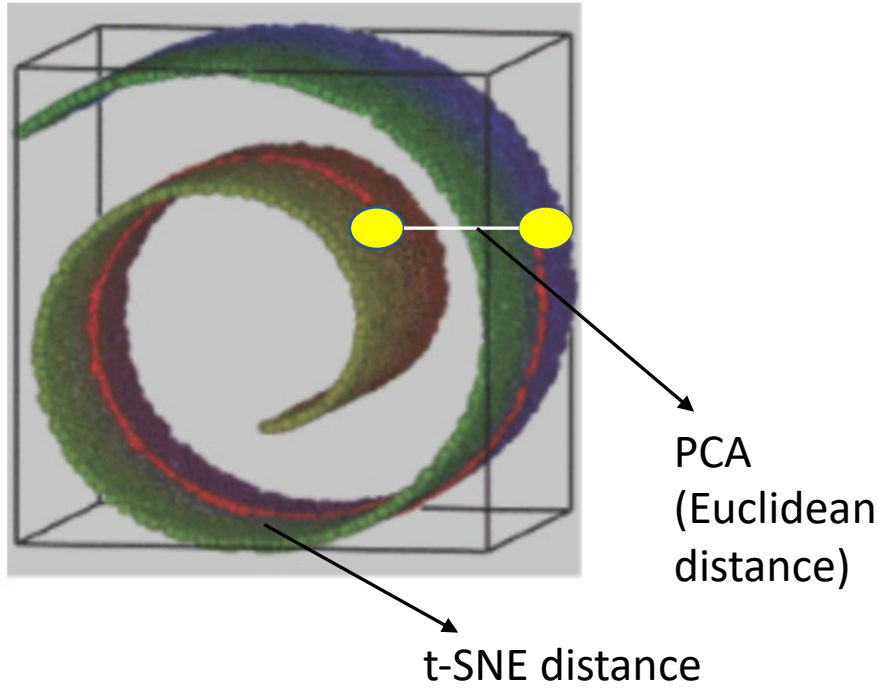
➢ Preparation for clustering and visualization

➢ Separation of cell type visible also using just PCA, but highly affected by outliers, only works if first two PCA explains most of the data.



# Clustering and visualization

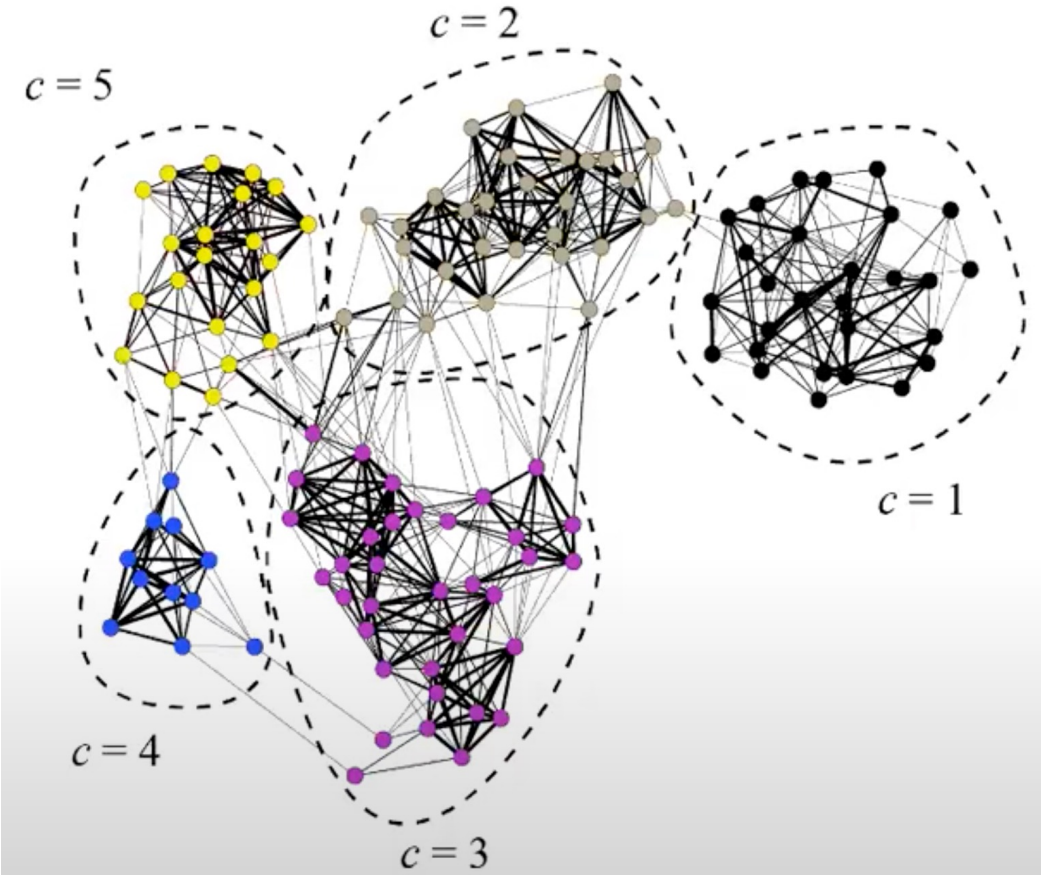
- Many possible methods for clustering: hierarchical clustering, K-means clustering, Graph based clustering
- For visualization, methods that use graph based, non linear clustering is popular in scRNA-seq analysis (UMAP and tSNE)
- These methods embed cells in a graph structure eg.K-nearest neighbor (KNN) graph
- Run on top of PCA in Seurat workflow




(Takahashi et al. 2009)

# Clustering

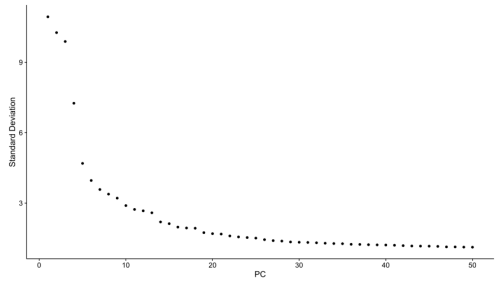
Seurat uses modularity optimization techniques are used such as Louvain algorithm for clustering

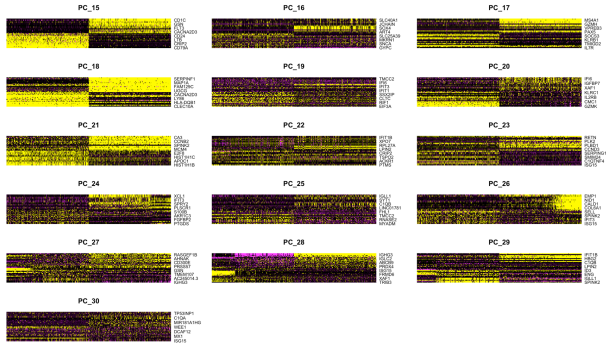


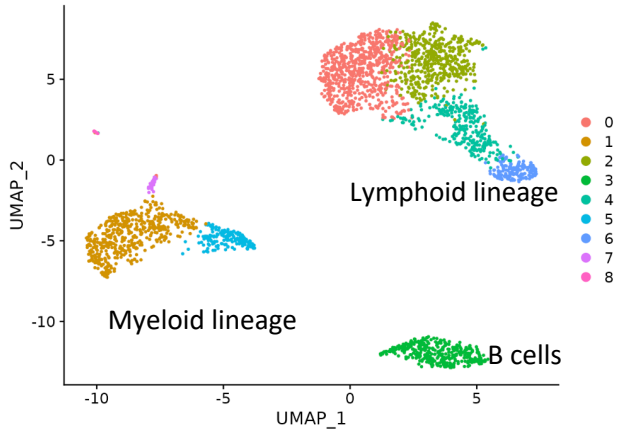

Source unknown

# Preparation for for clustering

- Choose significant principal components: elbow plot, PCA heatmap
  -Repeat analysis with different number of PCAs
  -Choose higher cutoff instead of lower
- Set a resolution parameter (higher resolution = more clusters)



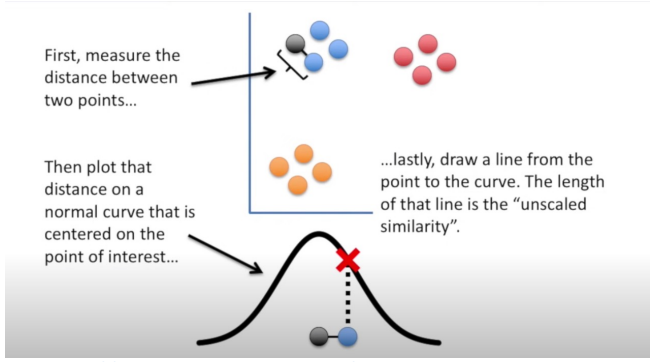
# Visualization

- t-SNE and UMAP
- Principles are same: Create a graph to represent the data in a high-dimensional space, and then attempt to recreate the graph as accurately as possible in a lower-dimensional environment.
- Similar steps: (1) calculate the “distance” between a cell and a controlled number of neighbors in the high dimensional space, (2) make sure that these “distances” are similar when data is moved to a lower dimensional space.
- Difference lies in mathematics: The graph is moved point-to-point from a high- to a low-dimensional space using t-SNE. UMAP compresses the graph into a lower dimension while maintaining its fuzziness and topological similarity.
- Similar problem: The cells might be all connected or isolated unless the distance is constrained (Curse of Dimensionality). The constraint is put on the number of neighbors that a cell can pick. UMAP: number of neighbour, t-SNE: perplexity

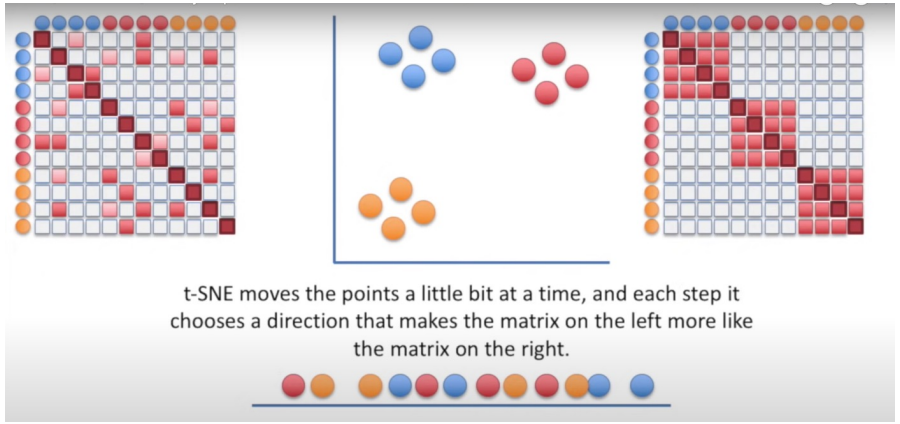
# Visualization

T-sne : t–Stochastic Neighbourhood Embedding (Van der Maaten and Hinton, 2008)

- The higher dimensional dataset is transformed into a lower dimension using the "T-distribution." In essence, it determines how likely a cell is to be drawn toward a neighboring cell (similarity measure), and then it repeats the process for each cell. Cells are reconfigured in the low dimension space in accordance with these criteria.
- Focus on distance between similar points
- Preserves the local structure, but not global structure
  -distance within the cluster is meaningful but that between clusters are not meaningful

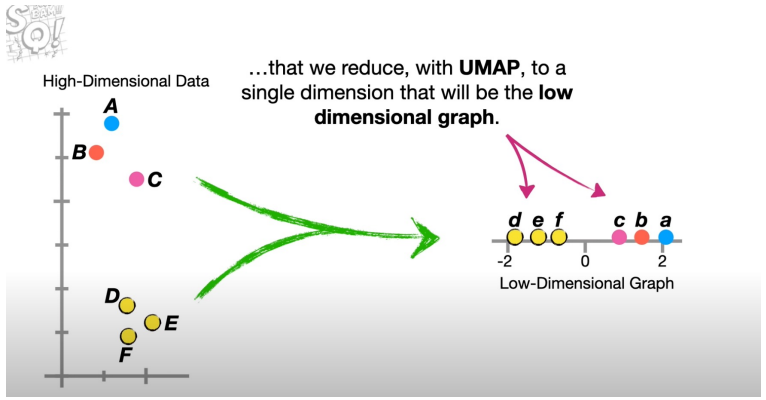



https://www.youtube.com/watch?v=NEaUSP4YerM

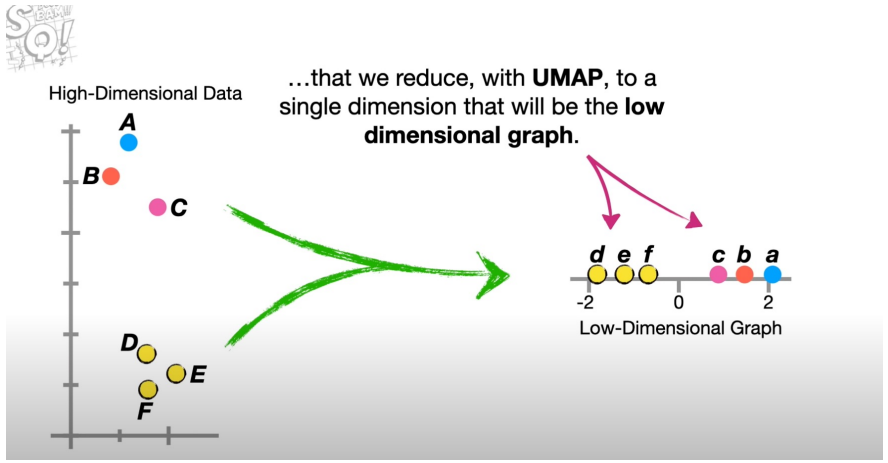


# Visualization

- UMAP : Uniform Manifold Approximation and Projection (McInnes et al., 2018)
- Based on topological structures in multidimensional space:
- It first develops a fuzzy graph that reproduces the topology (shape) of the real high-dimensional graph, determines the weights assigned to its edges, and then constructs a low-dimensional graph that closely resembles the fuzzy graph.
- Preserve global structure better than t-SNE
- Faster (In case of large datasets)



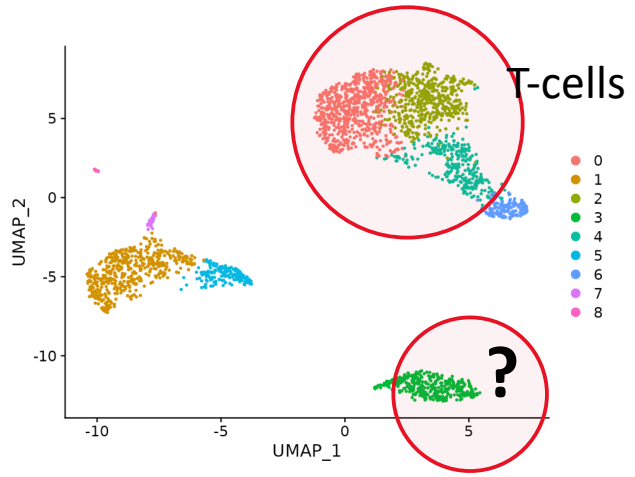

https://www.youtube.com/watch?v=eN0wFzBA4Sc&t=649s



# Cluster biomarker identification

- Difference between clusters
- Marker genes are the genes that are differentially expressed (DE) between the cluster
- Eg CD14 is a marker gene for classical monocytes, MPO for neutrophils, CD3E CD3D CD3G for T cells, CD8A for CD8 T cells, GNLY NKG7 for NK cells.

  (Markers genes can not always be useful for annotation)

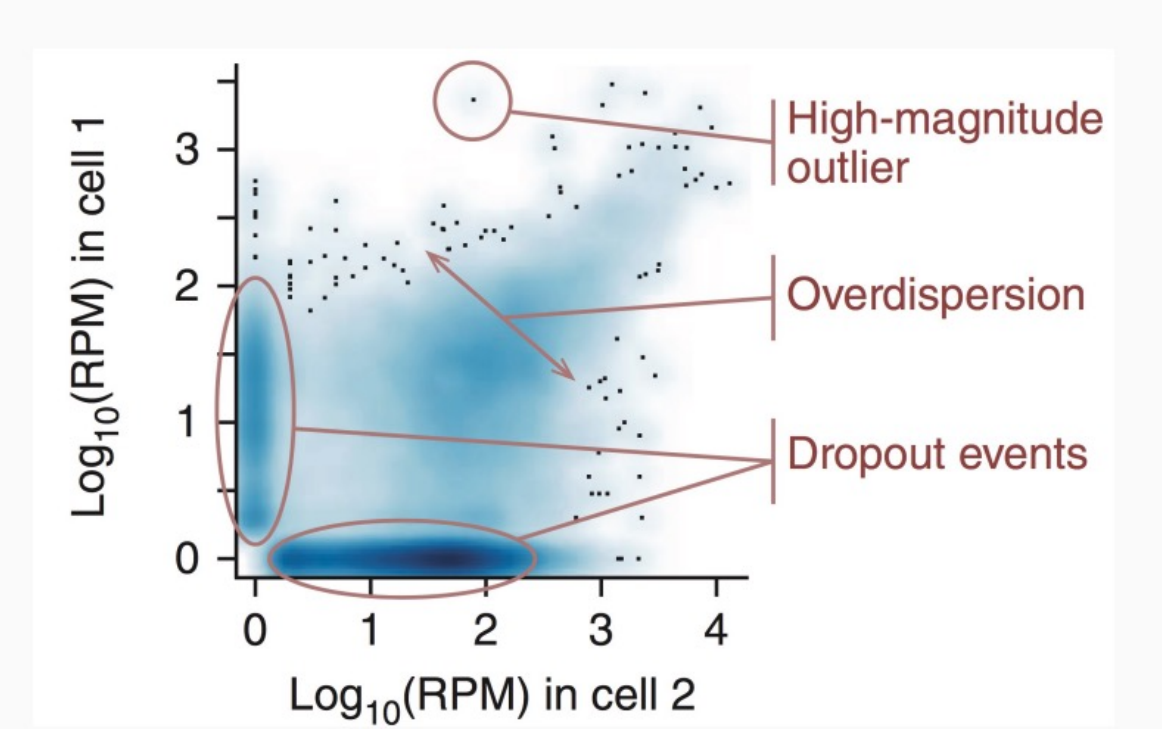



Seurat (PBMC)

- ScRNA seq DE is challenging compared to bulk RNA-seq due to noise such as dropouts, amplification biases (low mRNA) and also uneven sequencing depth

# Dropouts in scRNA-seq

A dropout event is when a transcript is expressed in a cell but is entirely undetected in its mRNA profile




(Kharchenko et al., 2014)

# DE analysis for marker identification

- MAST: works good for large number of samples (note that samples = cells in scRNA-seq analysis )
- Wilcox rank sum test (non-parametric test): Affected by dropouts (zeros) but works fine.
- Bulk RNA-seq method such as DESeq2
  - No prior gene filtering
  - DE between two cluster at a time, slow
- Multiple testing correction of p-values (bonferroni correction) to avoid false positives: raw p-value * number of genes tested, bonferroni is more strict -> make it less strict by testing less number of genes
- Limit testing to genes which show minimum of X% fold difference, also speeds up the analysis.

# Links

**References:**


- van der Maaten, Laurens & Hinton, Geoffrey. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605.
- McInnes, Leland & Healy, John & Saul, Nathaniel & Grossberger, Lukas. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 3. 861. 10.21105/joss.00861.
- Kharchenko, Peter & Silberstein, Lev & Scadden, David. (2014). Bayesian approach to single-cell differential expression analysis. Nature methods. 11. 10.1038/nmeth.2967.
- Codes are mostly from here: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html


**Other reading (from the symposium):**


Sini Junttila, Johannes Smolander, Laura L Elo, Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data, *Briefings in Bioinformatics*, Volume 23, Issue 5, September 2022, bbac286, https://doi.org/10.1093/bib/bbac286

Let’s go through the codes now!