scRNA-seq analysis and cellular populations

Neo Christopher Chung

Lecture 8, 1000-719bMSB

Project proposal meeting - May 13

- Prepare 5 min presentation on your project idea

- Show and summarize the paper – which your project is based on
- Present the data – at minimum, load and reproduce some of the original analysis
- Describe how you are changing the data analysis
    a. The original paper did X, but I believe it's better to do Y
    b. The original paper forgot to do X, therefore I will try X
    c. The original paper used a method X. I can improve it by modifying it.
    d. The original paper had a goal of understanding X (primary). I would like to understand Y (secondary).
    e. I found 2 or more papers on the same topic. I will combine the data and do the same analysis.
    f. Many more!
- Remember - you will need to write a final report (publication style) on this

Previous year’s examples

- Predicting the formation of chromatin loops using genomic data - search for the best model
- Polycystic ovary syndrome and obesity: gene expression in granulosa cells
- Gene Expression–Based Prediction of Neoadjuvant Chemotherapy Response in Early Breast Cancer
- Accounting for technical noise in scRNA-seq experiments: Identifying highly variable genes

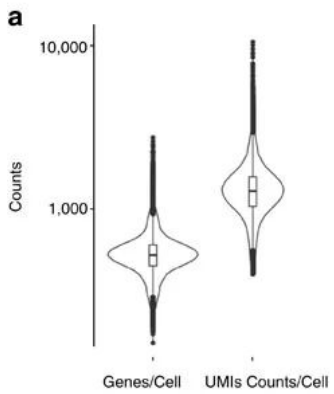
Heterogeneity in blood cells

- Zheng et al (2017) Nature Communications studies immune populations in peripheral blood mononuclear cells (PBMCs)
- Fresh PBMCs from a healthy donor (Donor A).
- 8–9k cells were captured from each of 8 channels and pooled to obtain ~68k cells.
- At ~20k reads per cell, the median number of genes and UMI counts detected per cell was ~525 and 1,300, respectively.

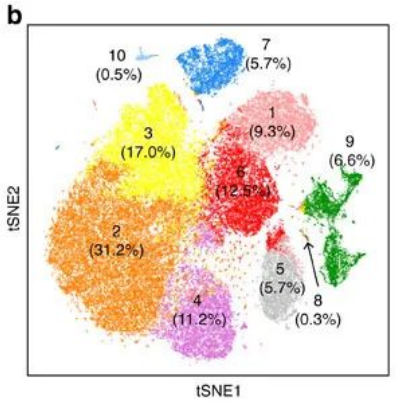
Dimension reduction and clustering

- PCA on the top 1,000 variable genes ranked by their normalized dispersion
- *K*-means clustering on the first 50 PCs identified 10 distinct cell clusters
- t-SNE (2D projection) are created based on the first 50 PCs

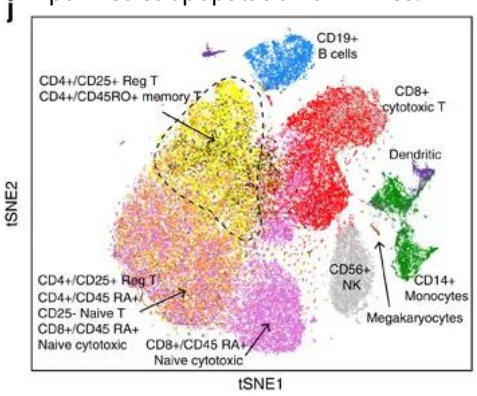
Distribution of number of genes (left) and UMI counts (right) detected per 68k PBMCs.



tSNE projection of 68k PBMCs, where each cell is grouped into one of the 10 clusters



tSNE projection of 68k PBMCs, with each cell coloured based on their correlation-based assignment to a purified subpopulation of PBMCs.

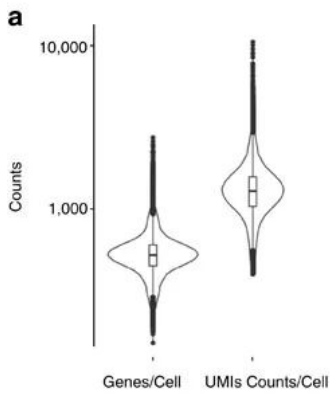


Zheng et al (2017) Nature Comm

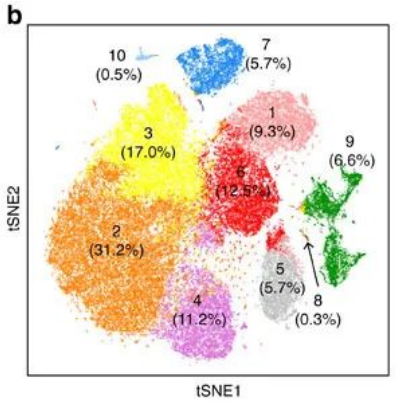
Looking at cell types in a single experiment

- >80% T cells (enrichment of *CD3D*, part of the T-cell receptor complex, in clusters 1–3 and 6)
- ~6% NK cells (enrichment of *NKG7* in cluster 5)
- ~6% B cells (enrichment of *CD79A* in cluster 7)
- ~7% myeloid cells (enrichment of *S100A8* and *S100A9* in cluster 9

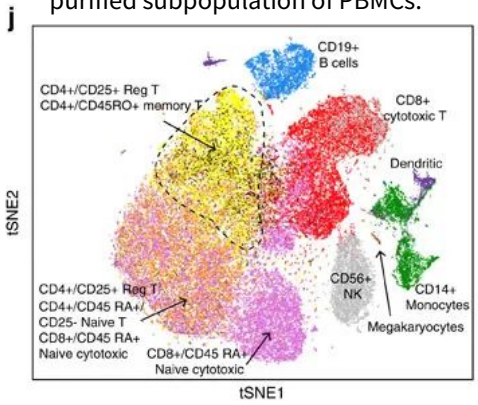
Distribution of number of genes (left) and UMI counts (right) detected per 68k PBMCs.



tSNE projection of 68k PBMCs, where each cell is grouped into one of the 10 clusters



tSNE projection of 68k PBMCs, with each cell coloured based on their correlation-based assignment to a purified subpopulation of PBMCs.



Zheng et al (2017) Nature Comm

Minor subpopulations or subsubpopulations

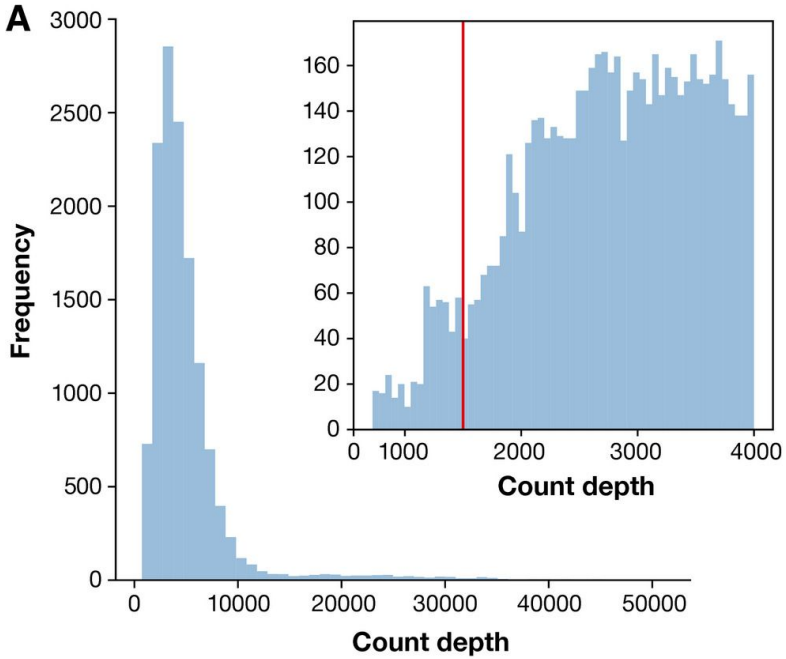
- Unsupervised clustering is very challenging
- The cells are developed in a branch process
- Cellular subpopulations may be in a hierarchy
- One may apply an algorithm to a large or a suspicious cluster
- Of course, there are hierarchical clustering and others that are computationally more expensive

- E.g., substructures were observed in CD34+ and CD14+ monocyte samples
- E.g., part of the inferred CD4+ naive T population was classified as CD8+ T cells.

Quality control: count depth per cell

Histograms of count depth per cell.

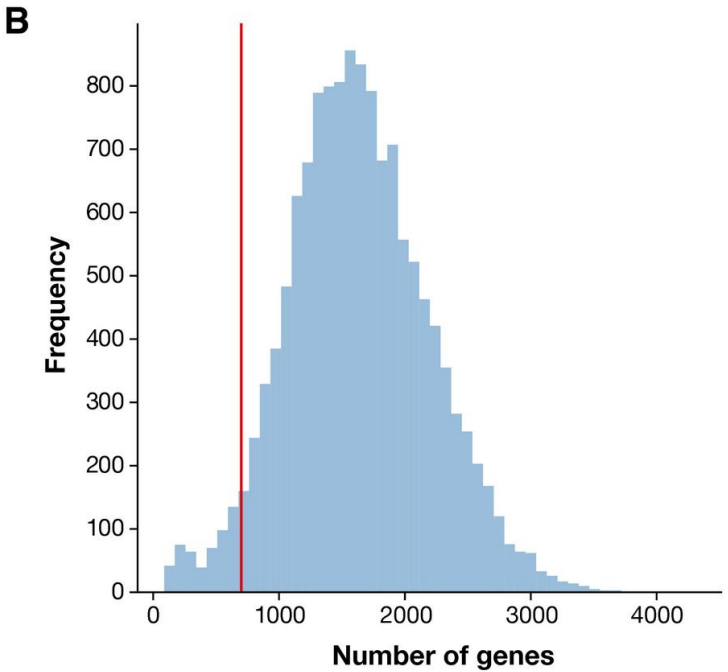
The smaller histogram is zoomed-in on count depths below 4,000. A threshold is applied here at 1,500 based on the peak detected at around 1,200 counts.



Quality control: number of genes detected per cell

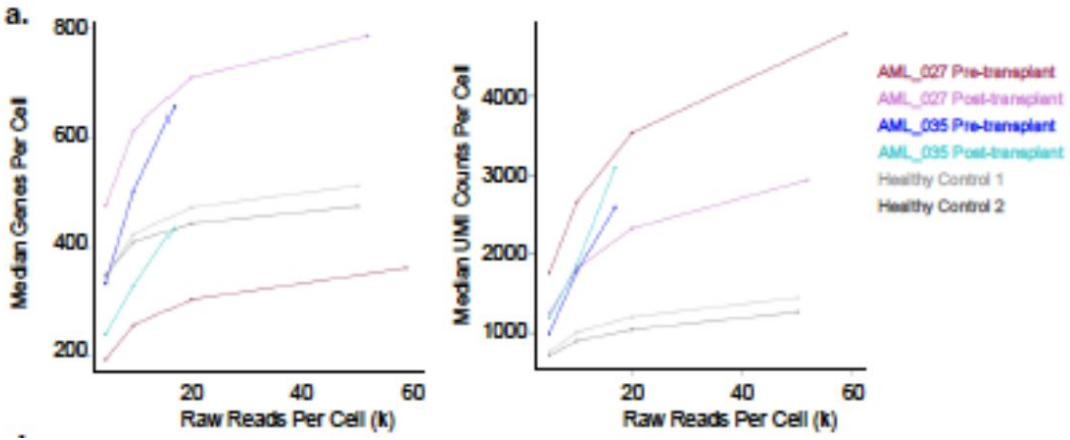
Histogram of the number of genes detected per cell.

A small noise peak is visible at approx. 400 genes. These cells are filtered out using the depicted threshold (red line) at 700 genes.



Luecken & Theis (2019) MSB

Quality control: # genes discovered get saturated

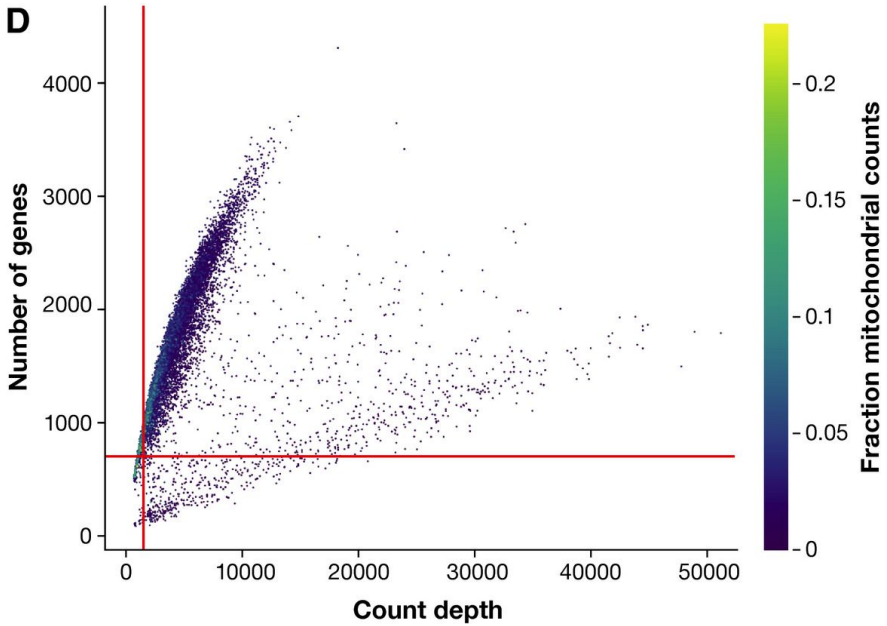


Supplementary Fig 10. Zheng et al.

Quality control: Count depth

Count depth distribution from high to low count depths.

This visualization is related to the log–log plot shown in Cell Ranger outputs that is used to filter out empty droplets. It shows an “elbow” where count depths start to decrease rapidly around 1,500 counts.

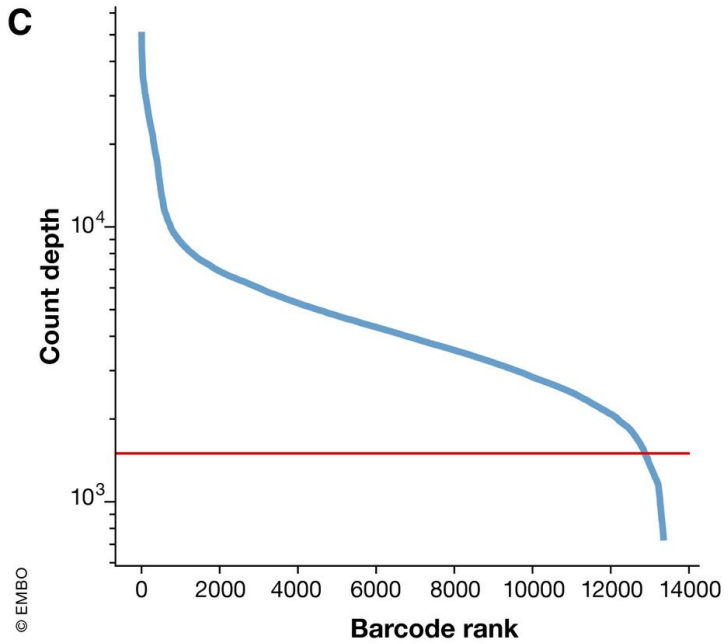


Luecken & Theis (2019) MSB

Quality control: Number of genes versus the count depth

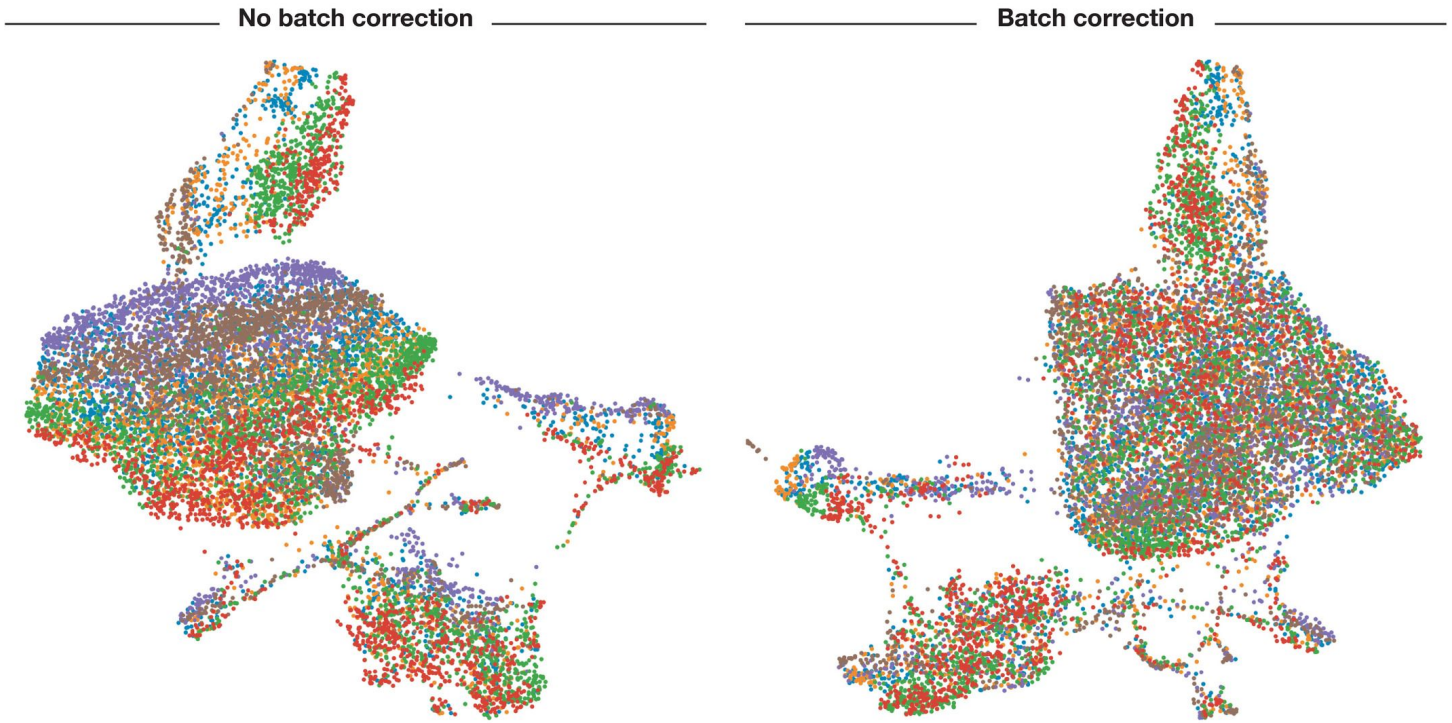
Number of genes versus the count depth coloured by the fraction of mitochondrial reads.

Mitochondrial read fractions are only high in particularly low count cells with few detected genes. These cells are filtered out by our count and gene number thresholds.



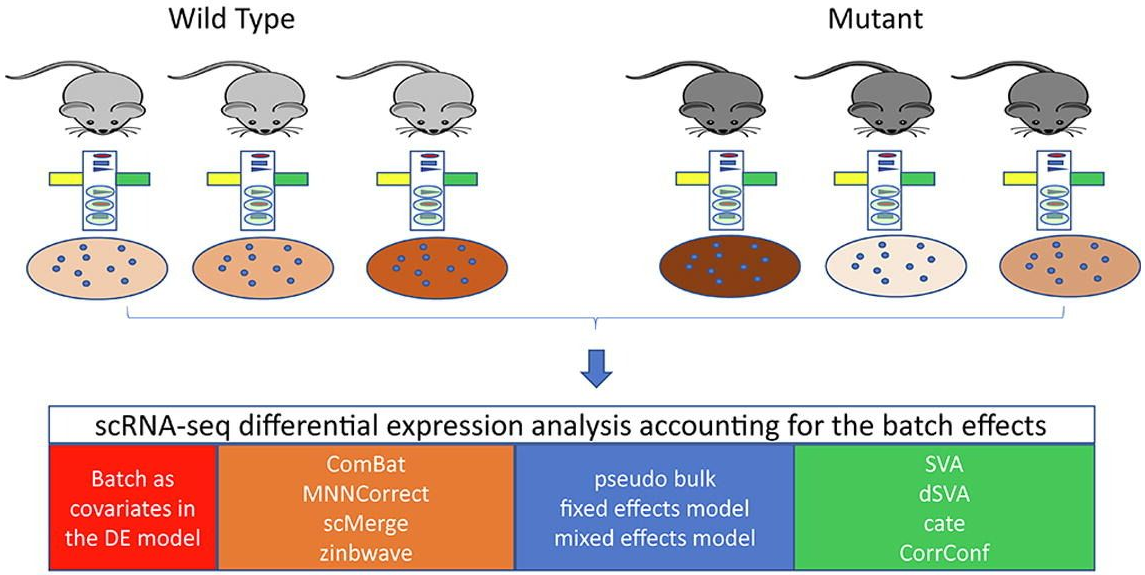
Luecken & Theis (2019) MSB

Batch effects, prevalent in scRAN-seq



Batch effects: we've been here before

Simulate two scenarios of batch effects: 1: matched batches, 2: independent batches and Consider: # of cells, purity of groups



Chen et al. 2020
10.1016/j.csbj.2020.03.026

Batch effects: we've been here before

Chen et al. 2020
A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing

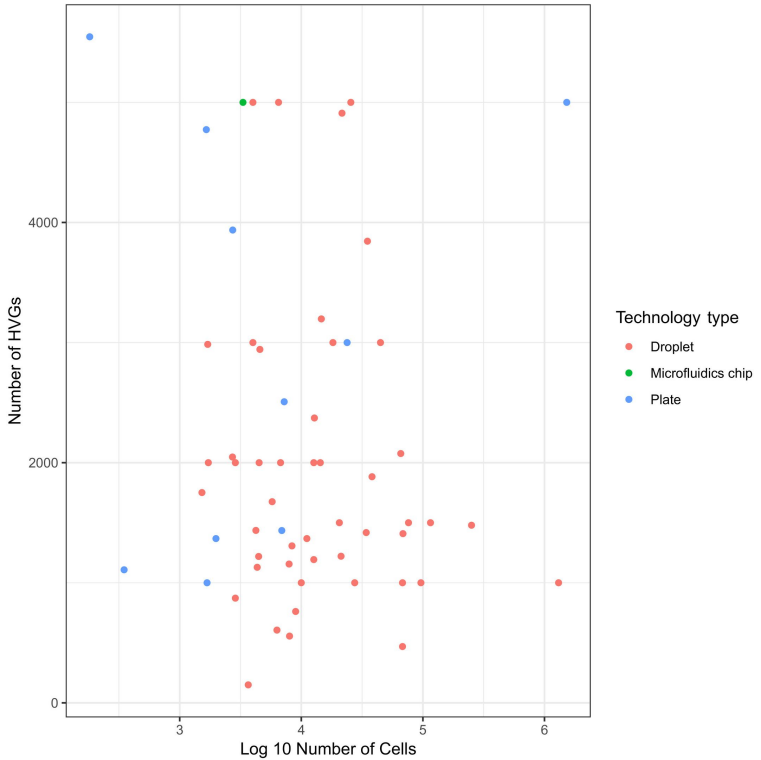
Evaluated 11 methods and recommendations for scRNA-seq DE analysis:
1) incorporate known batch variables instead of using batch-corrected data;
2) employ SVA for latent batch correction.

Chen et al. 2020
10.1016/j.csbj.2020.03.026

Feature Selection: variability

1. Keep only genes that are “informative” of the variability in the data.

2. Typically between 1,000 and 5,000 highly variable genes (HVGs) are selected for downstream analysis



Luecken & Theis (2019) MSB

Feature Selection: caveats

There are many different ways to measure variability

Downstream analysis may or may not be robust to the exact choice of the number of HVGs. Err on the side of higher numbers of HVGs.

HVGs should be selected after technical data correction to avoid selecting genes that are highly variable only due to batch effects.

Always check plots (volcano, histograms, heatmaps etc) before and after

Luecken & Theis (2019) MSB

Feature Selection: Normalized dispersion by Zheng et al.

```
.get_variable_gene<-function(m) {

  df<-data.frame(mean=colMeans(m),cv=apply(m,2,sd)/colMeans(m),var=apply(m,2,var))
  df$dispersion<-with(df,var/mean)
  df$mean_bin<-with(df,cut(mean,breaks=c(-Inf,quantile(mean,seq(0.1,1,0.05)),Inf)))
  var_by_bin<-ddply(df,"mean_bin",function(x) {
    data.frame(bin_median=median(x$dispersion),
               bin_mad=mad(x$dispersion))
  })
  df$bin_disp_median<-var_by_bin$bin_median[match(df$mean_bin,var_by_bin$mean_bin)]
  df$bin_disp_mad<-var_by_bin$bin_mad[match(df$mean_bin,var_by_bin$mean_bin)]
  df$dispersion_norm<-with(df,abs(dispersion-bin_disp_median)/bin_disp_mad)
  df
}
```

Dispersion = Variance/mean

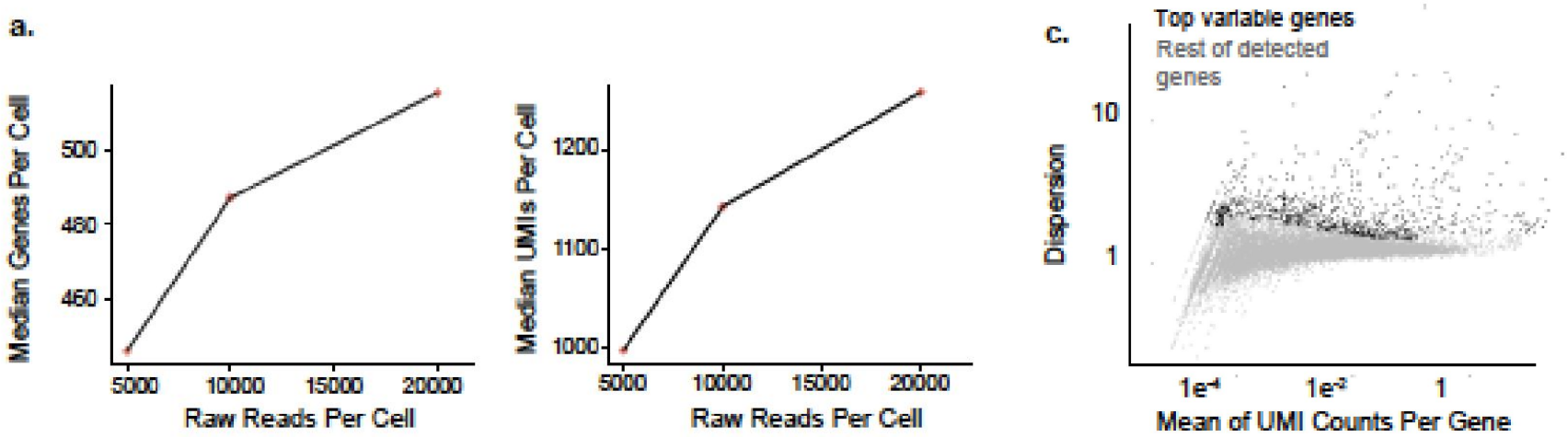
20 bins based on their mean expression

Bin_Median = median dispersion per bin

Bin_Mad = median absolute deviation per bin

Normalized dispersion = |Dispersion - Bin_Median| / Bin_Mad.

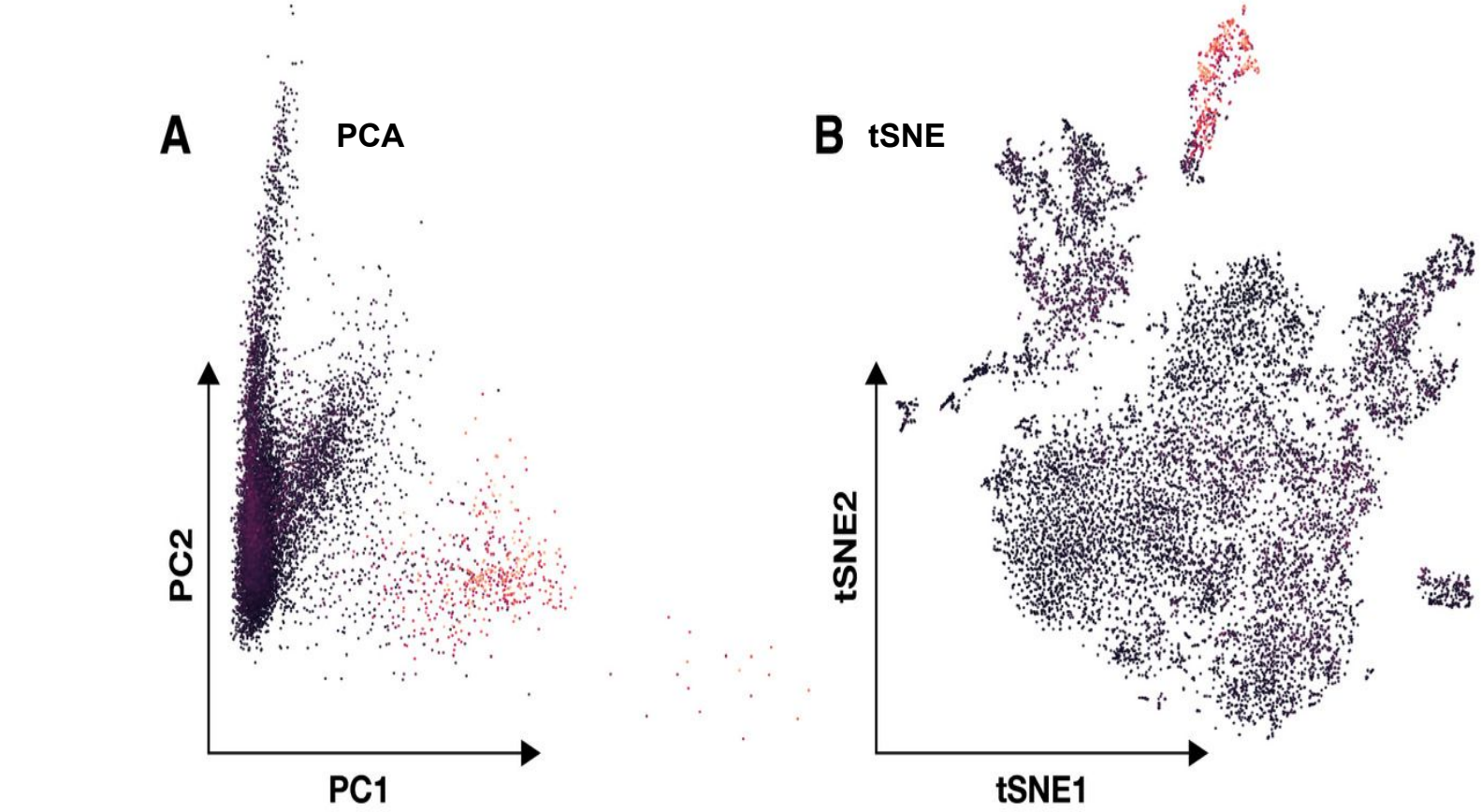
Feature Selection: Normalized dispersion by Zheng et al.



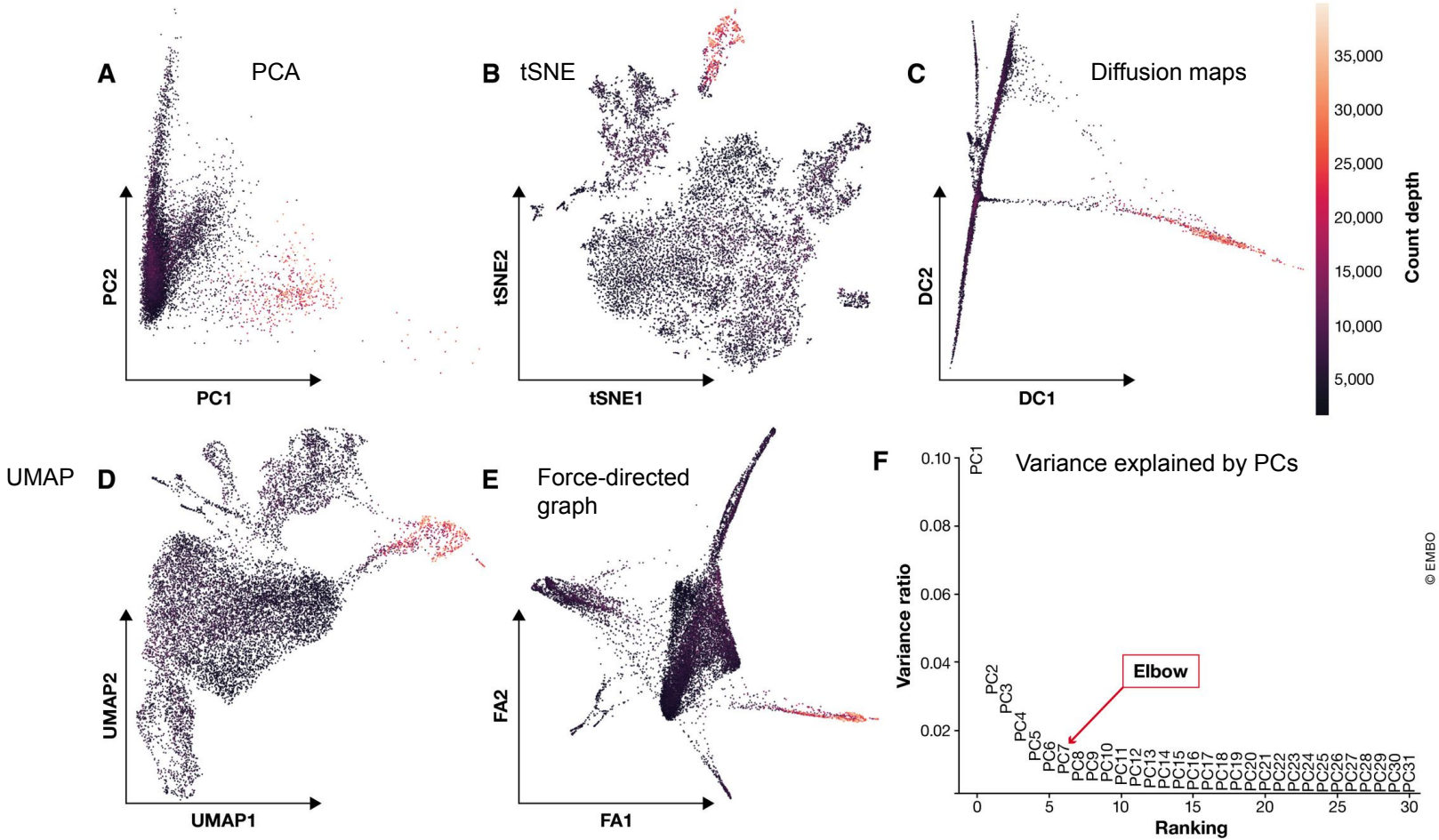
Supplementary Fig 5. Zheng et al.

Dimension reduction

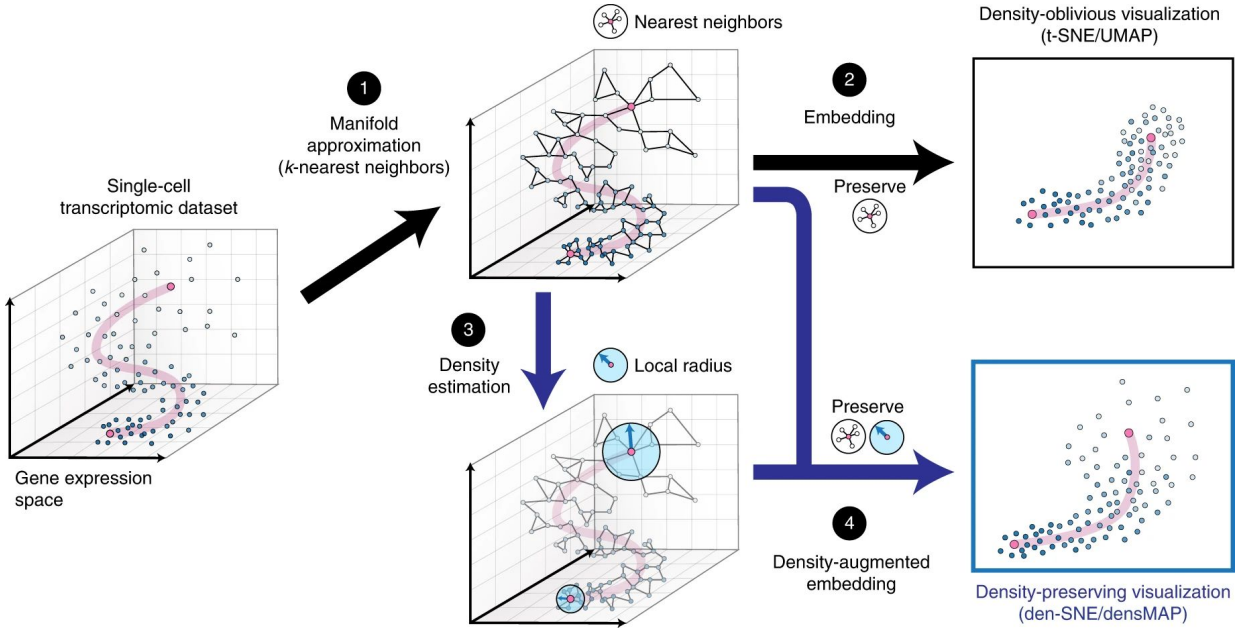
Mouse intestinal epithelium regions data from Haber *et al* (2017)



Dimension reduction



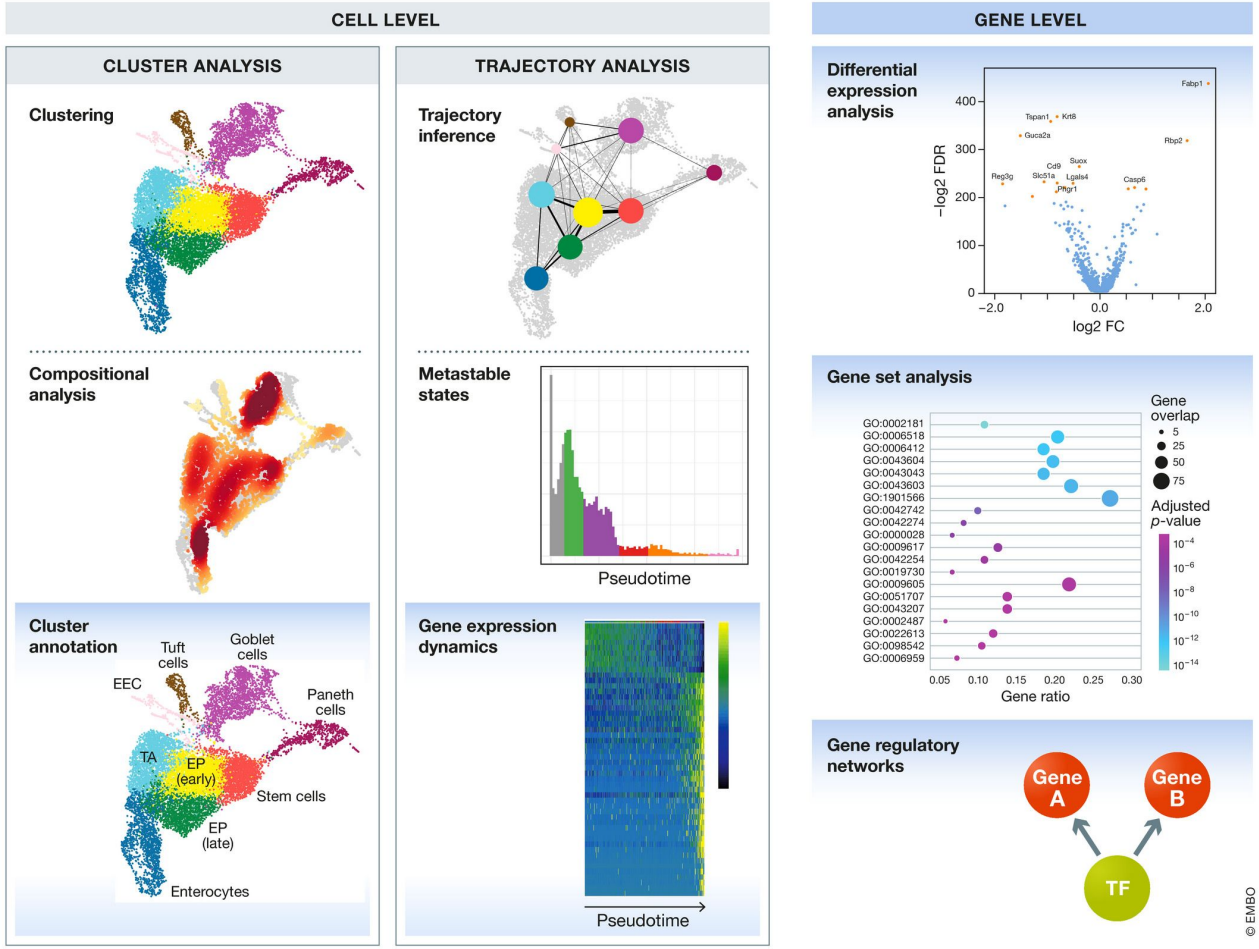
UMAP: Uniform Manifold Approximation and Projection



(1) *k*-nearest neighbor (KNN) graph is a compact summary of the data manifold
(2) Optimize the visualization coordinates of the points to maximally preserve local distances between neighbors in the graph
(3) A general, differentiable measure of density called the local radius on the KNN graphs that t-SNE and UMAP leverage
(4) By augmenting the original visualization objective with an additional term that encourages local radii to be consistent between the original space and the visualization, we transform both t-SNE and UMAP into density-preserving counterparts, den-SNE and densMAP

Ashwin Narayan, Bonnie Berger & Hyunghoon Cho. Nature Biotechnology (2021)

Downstream analysis overview



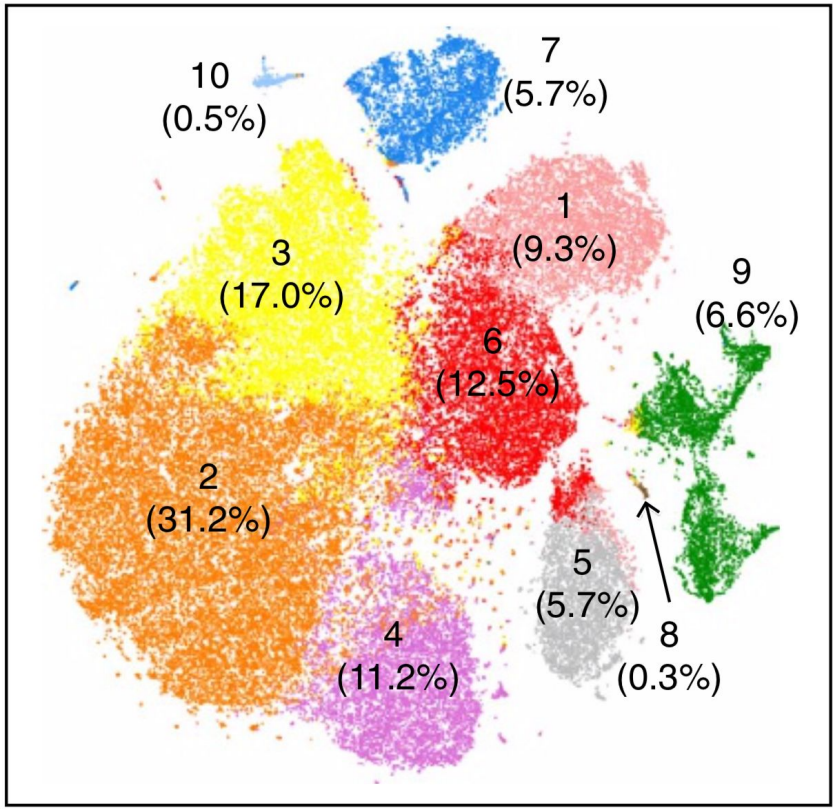
Major Challenges in scRNA-seq

1. Single droplet (or a wall) might contain two or more single cells. Known as **doublets or multiplets**, they may induce biologically irrelevant gene expression profiles in scRNA-seq studies.
2. Experiments that are done from different experimental and environmental conditions need to be **integrated**
3. Many cells are sequenced individually, without knowing which gene expression profiles correspond to (known and unknown) **cell identities**
4. Since cell identities are unknown externally, a run/lane of sequencing must be done on one “type” of cells

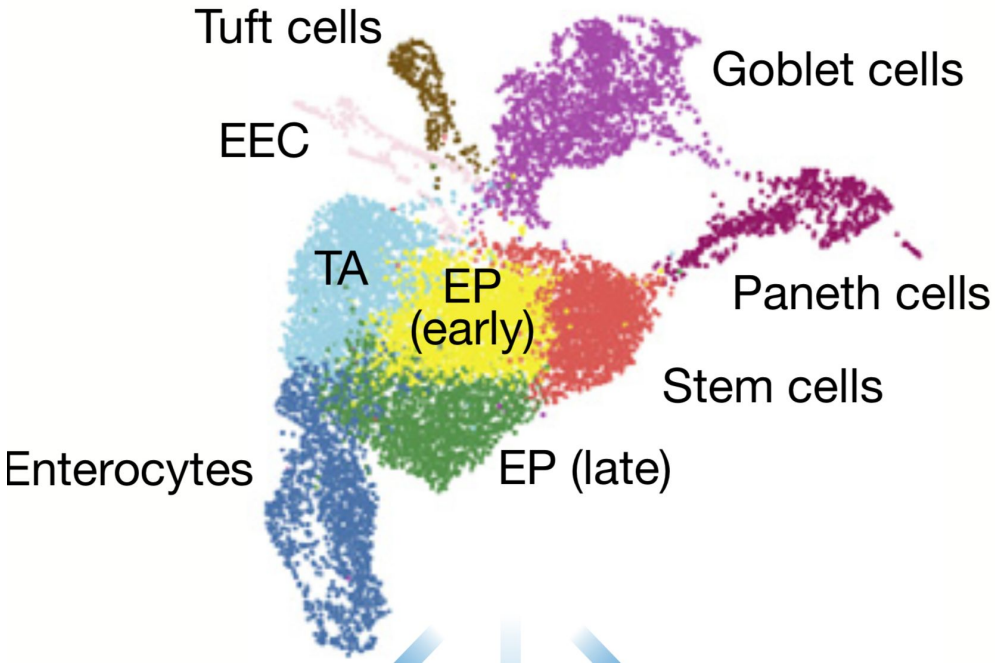
Solutions

1. Clustering is used to determine cell identities
2. Doublet/multiplet detection algorithms
3. Evaluation of clustering-based cellular identities
4. Identification of cellular subpopulations across multiple data sets
5. Multiplexing with barcoded antibodies or natural genetic variation
6. Estimation of cellular trajectories or developmental stages

Clusters ~ cellular populations



PBMCs from Zheng et al (2018)



Intestinal epithelium from Haber et al (2017)

Clustering algorithms

K means clustering (Macosko et al., 2015; Zheng et al., 2017)

K nearest neighbors in Seurat (Satija et al., 2015)

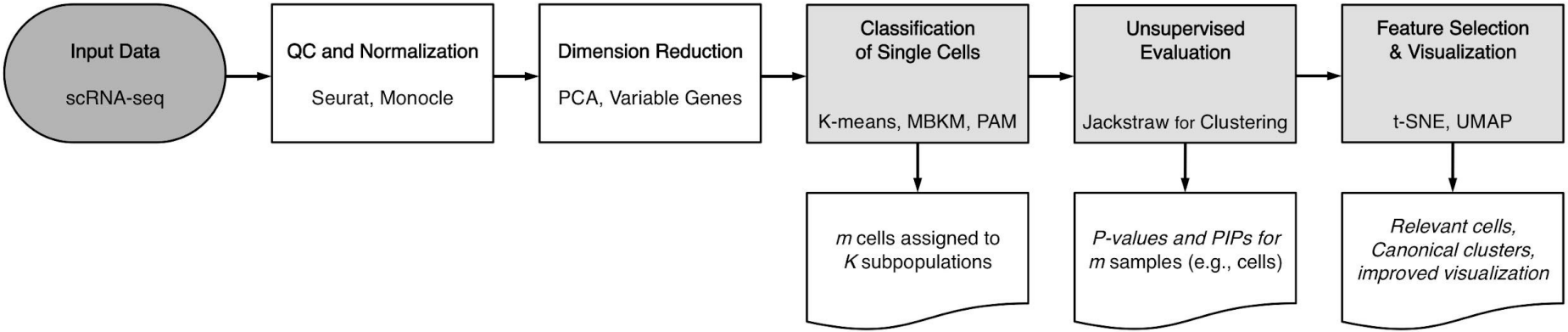
Hierarchical clustering in SINCERA (Guo et al., 2015)

Density peak clustering in Monocle (Qiu et al., 2017)

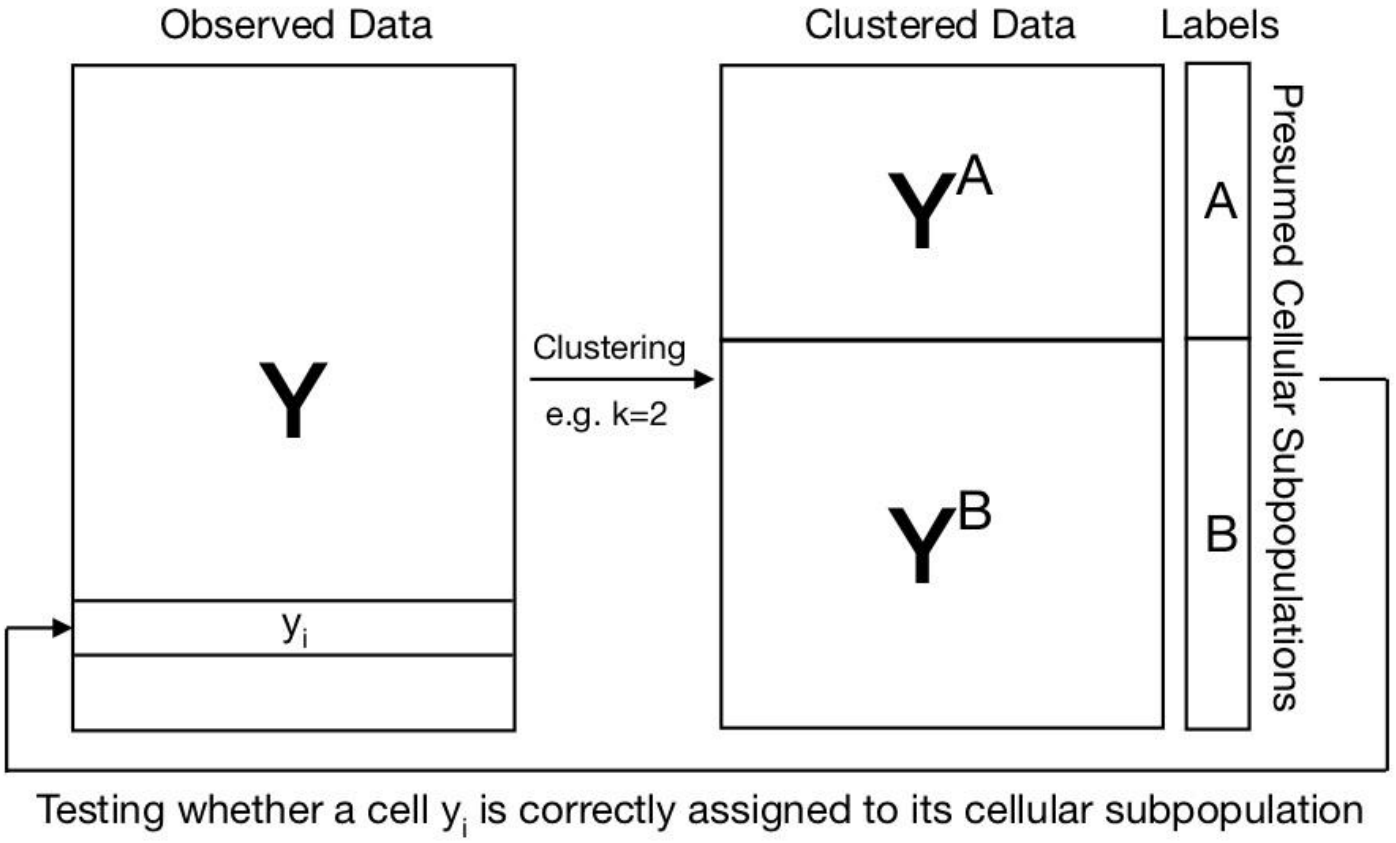
Algorithms for scRNA-seq data (Zeisel et al., 2015; Xu and Su, 2015; Buettner et al., 2015; Wang et al., 2017).

Consensus (ensemble) algorithms (Kiselev et al., 2017; Yang et al., 2018).

Evaluation of cell identities



Be aware of circular analysis



Testing whether a cell y_i is correctly assigned to its cellular subpopulation

Algorithm 1 Jackstraw Test for Cluster Membership

1. Apply the clustering algorithm to the observed data $\mathbf{Y}$, resulting in cluster centers $\mathbf{c}_k$ for $k = 1, ..., K$ and membership assignments $b_{i,K}$ for $i = 1, ..., m$ and $K = 1, ..., k$

2. Compute the observed statistics $F_1, ..., F_m$, where the full models include corresponding cluster centers $\mathbf{c}_k(\mathbf{Y})$

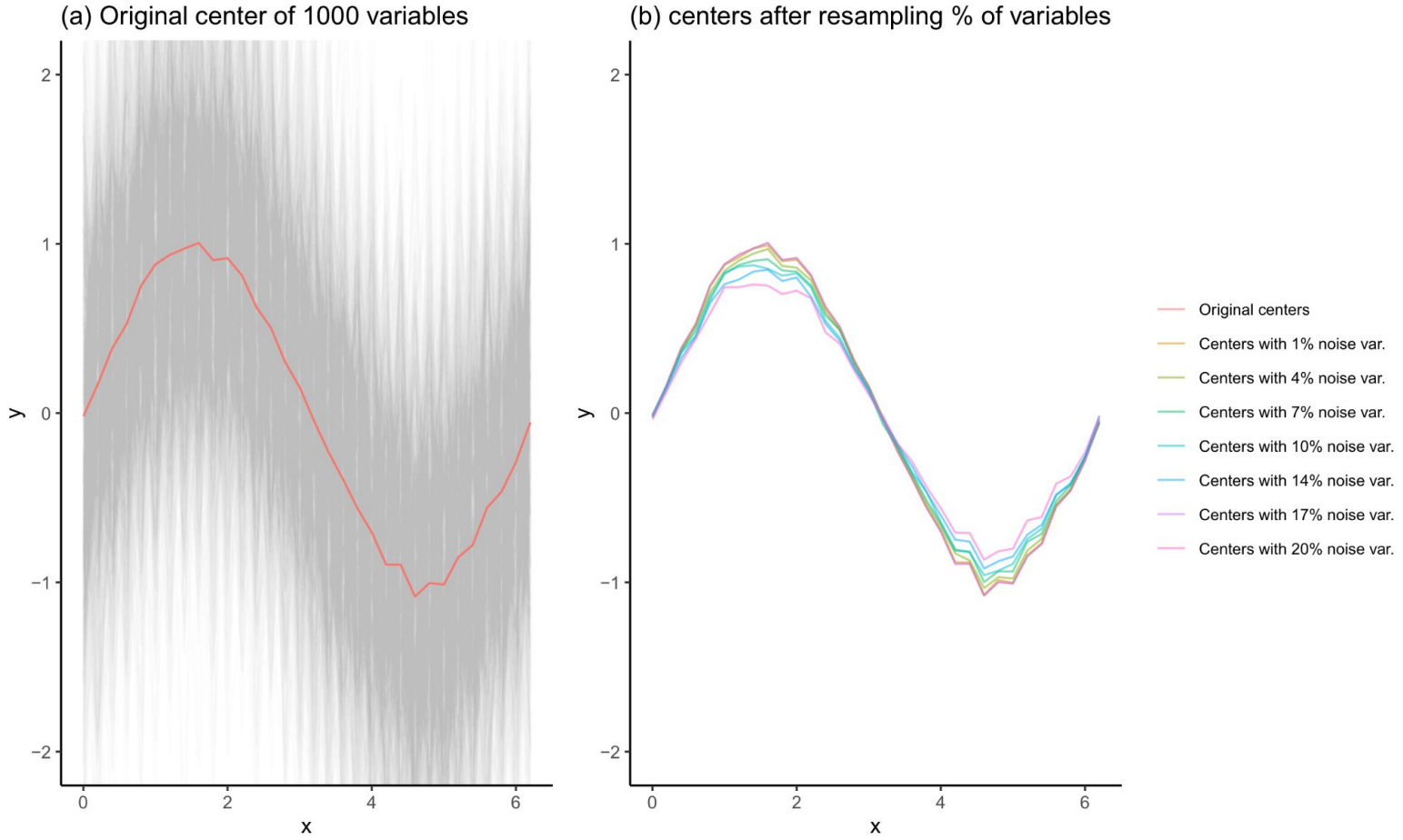
3. Create s synthetic null samples by resampling with replacement a small proportion of samples $s \ll m$, resulting in a jackstraw data $\mathbf{Y}^*$, with $m - s$ observed samples and s synthetic null samples

4. Apply the clustering algorithm to the jackstraw data $\mathbf{Y}^*$, resulting in cluster centers $\mathbf{c}_k^*(\mathbf{Y}^*)$ and membership assignments $b_{i,K}^*$

5. Compute the null statistics $F_1^*, ..., F_s^*$, where the full models include corresponding cluster centers $\mathbf{c}_k^*(\mathbf{Y}^*)$

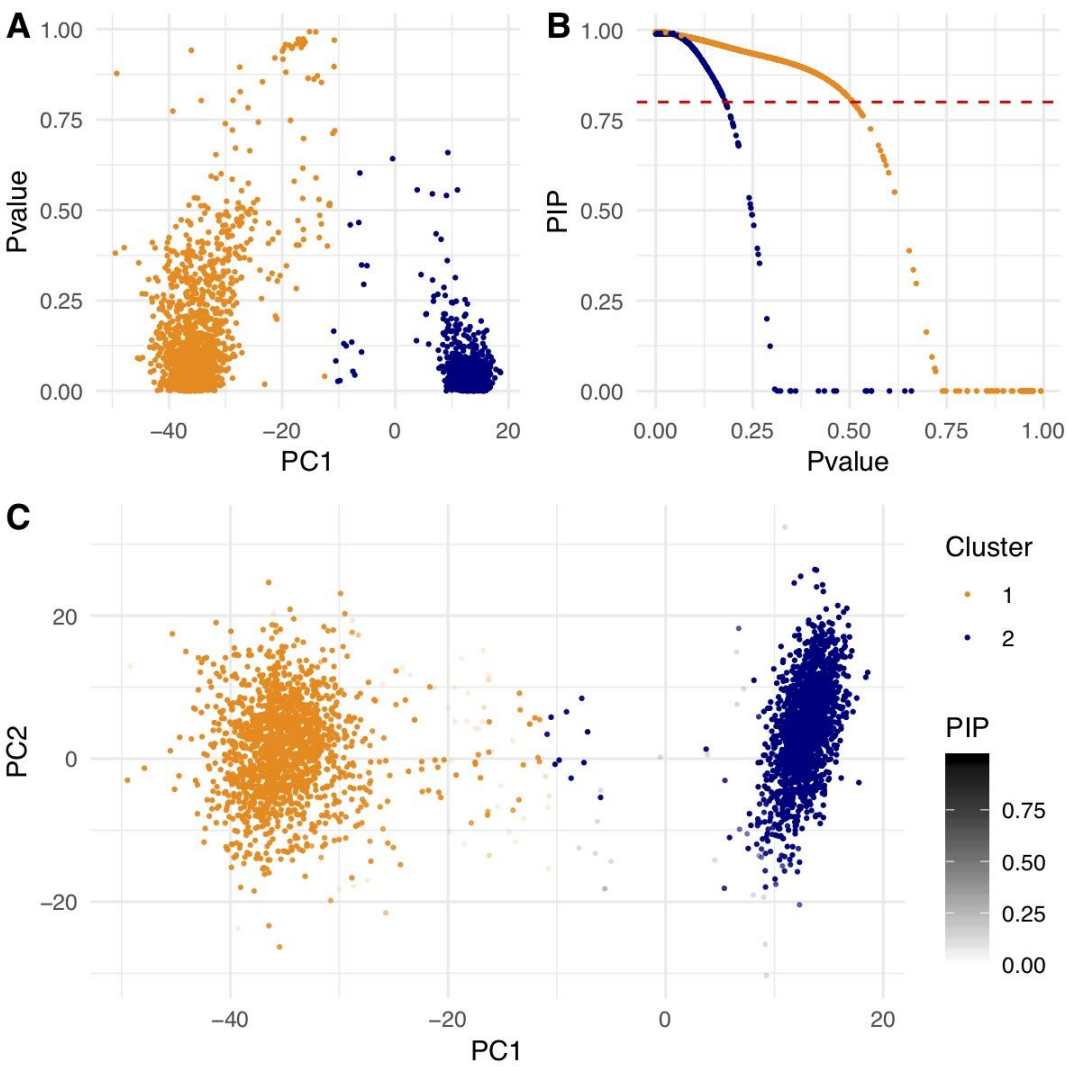
6. Repeat the above three steps $b = 1, ..., B$ times to obtain a total $s * B$ of null statistics

7. Compute the p-values by empirically ranking the observed statistics among the null statistics

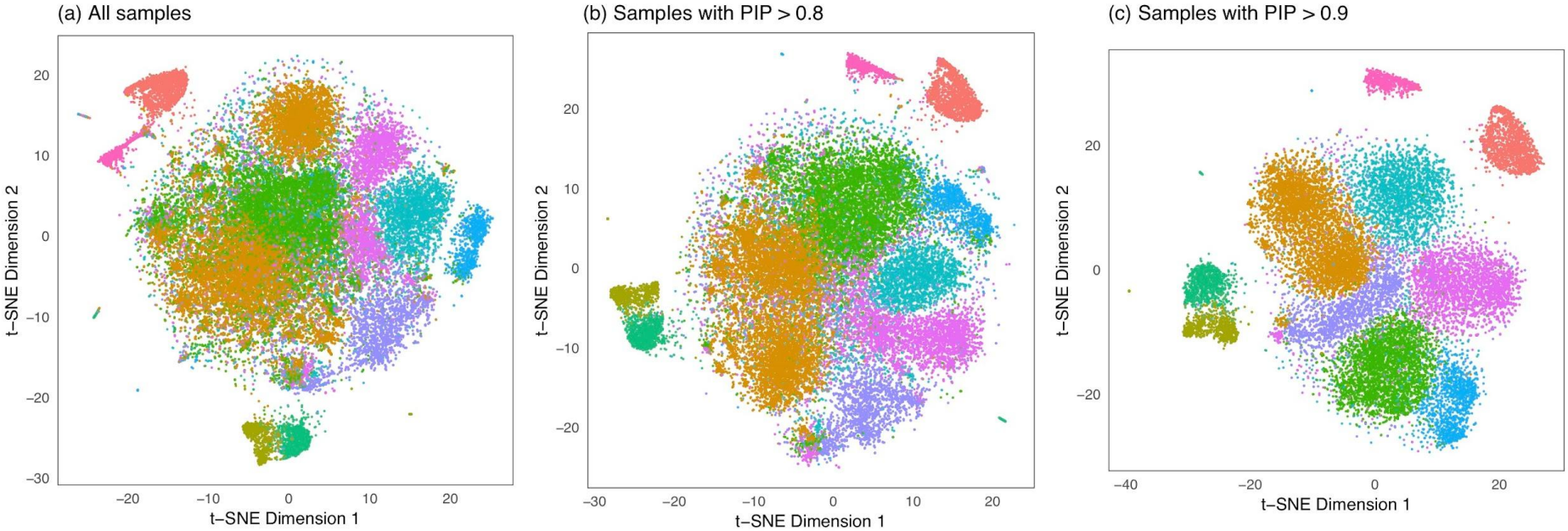


Impact of resampling a small proportion of variables on centroids

Mixture of Jurkat and 293T Cell Lines



Feature selection in PBMC data using the jackstraw



Doublet/multiplet detection

Several algorithms work similarly, by simulating doublets/multiplets：

1. DoubletDecon (DePasquale et al., 2018)
2. Scrublet (Wolock et al., 2019)
3. DoubletFinder (McGinnis et al., 2019)

DoubletFinder: Algorithm

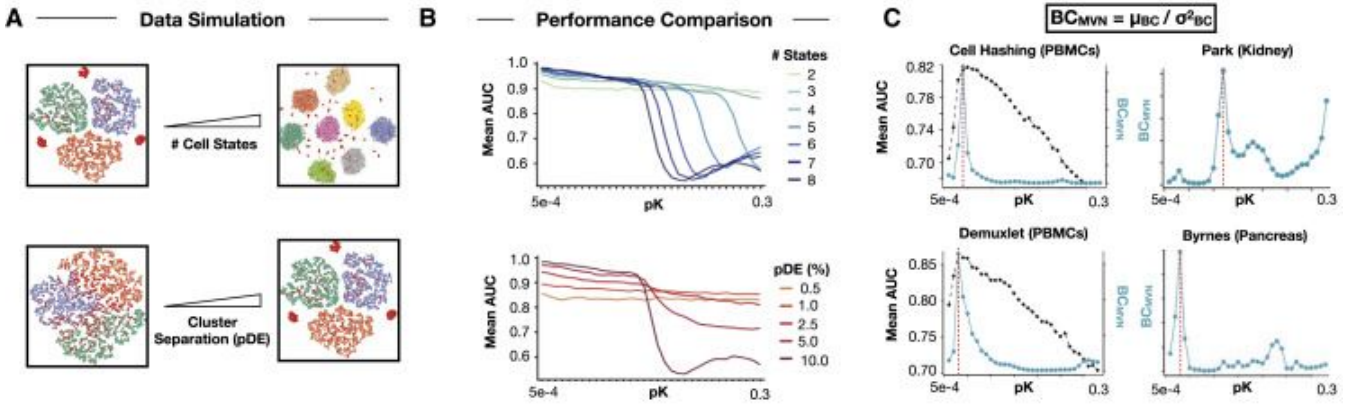
1. simulates **artificial doublets** from existing scRNA-seq data by averaging the gene expression profiles of random pairs of cells
2. merges and pre-processes real and artificial data using the “Seurat”
3. performs dimensionality reduction on the merged real-artificial data using PCA
4. detects the *k* nearest neighbors for every real cell in principal component (PC) space
5. compute each cell’s **proportion of artificial nearest neighbors (pANN)**
6. predicts doublets as cells with the top *n* pANN values, where *n* is set to the total number of expected doublets

Required parameters: the number of expected real doublets, the number of artificial doublets (pN) and the neighborhood size (pK) used to compute the number of artificial nearest neighbors

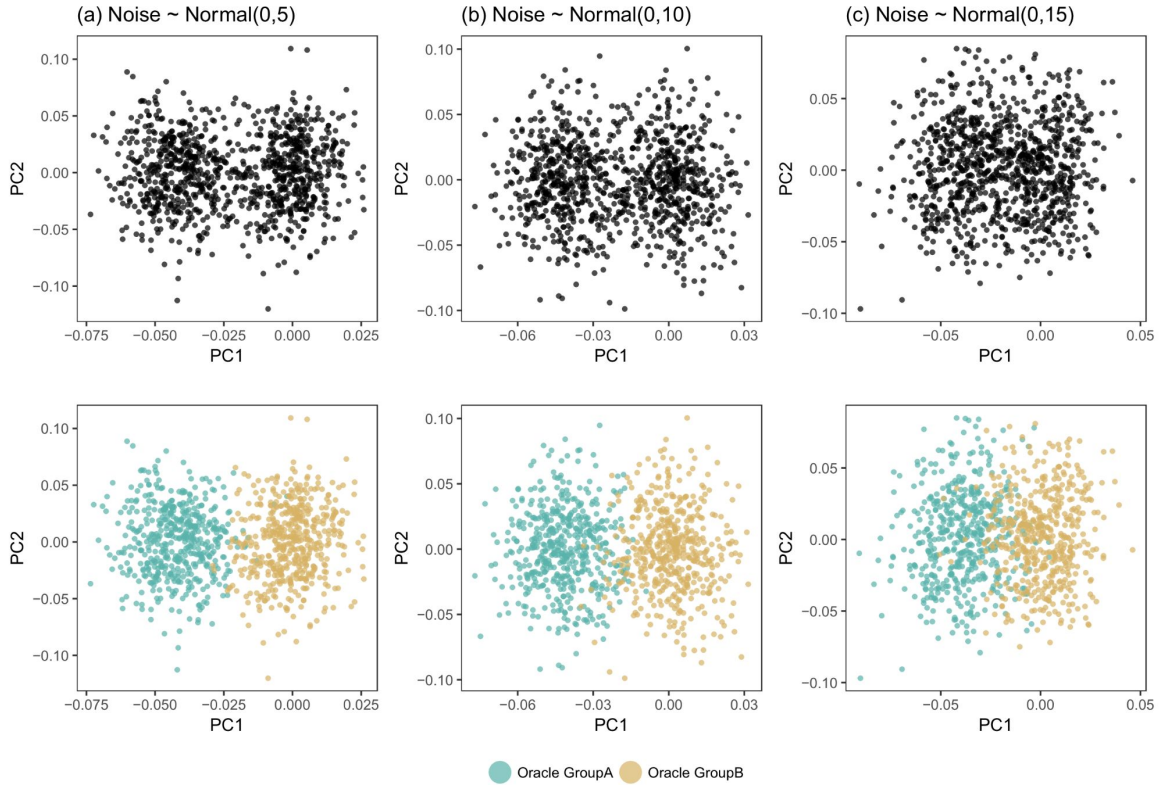
DoubletFinder: pK and structure

DoubletFinder is sensitive to changes in the input parameter specifying pK.
In an simulation study creating 3–8 distinct cell clusters (ie, real doublets were simulated by adding the gene expression profiles of randomly selected cells), calculate the mean AUC for each pK value across all pN

→ Mean AUC inflection point positions differed for simulations with variable numbers of cell states, suggesting that pK parameter selection is sensitive to the inherent diversity of scRNA-seq data.
→ mean AUC inflection points were only observed for simulations with well-separated clusters



Inherent ambiguity of clusters



Well-separated clusters will result in DoubletFinder to perform well

DoubletFinder and Jackstraw