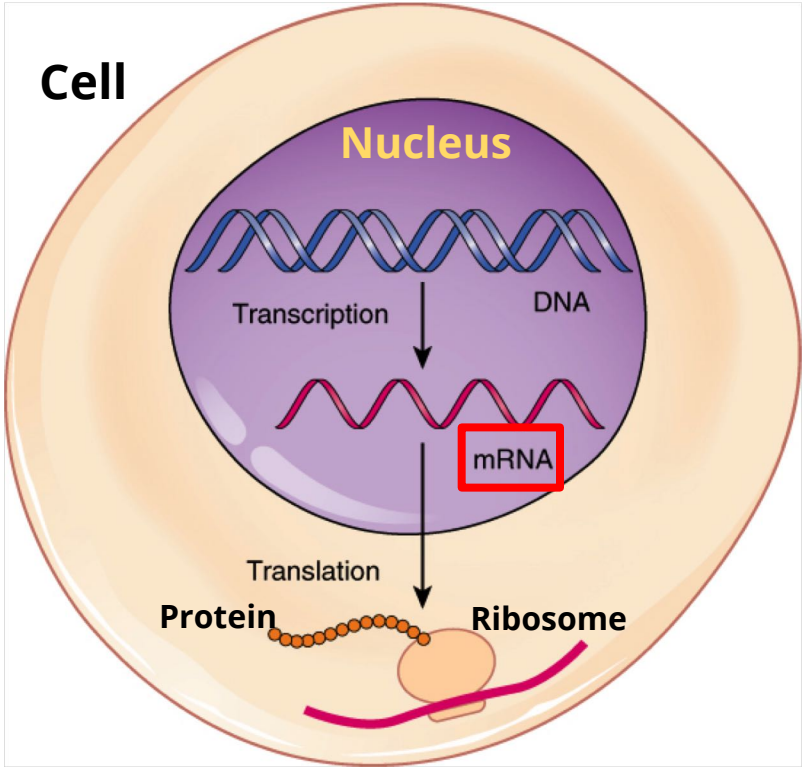
Comparing Methods For Analyzing Single-Cell RNA sequencing Data

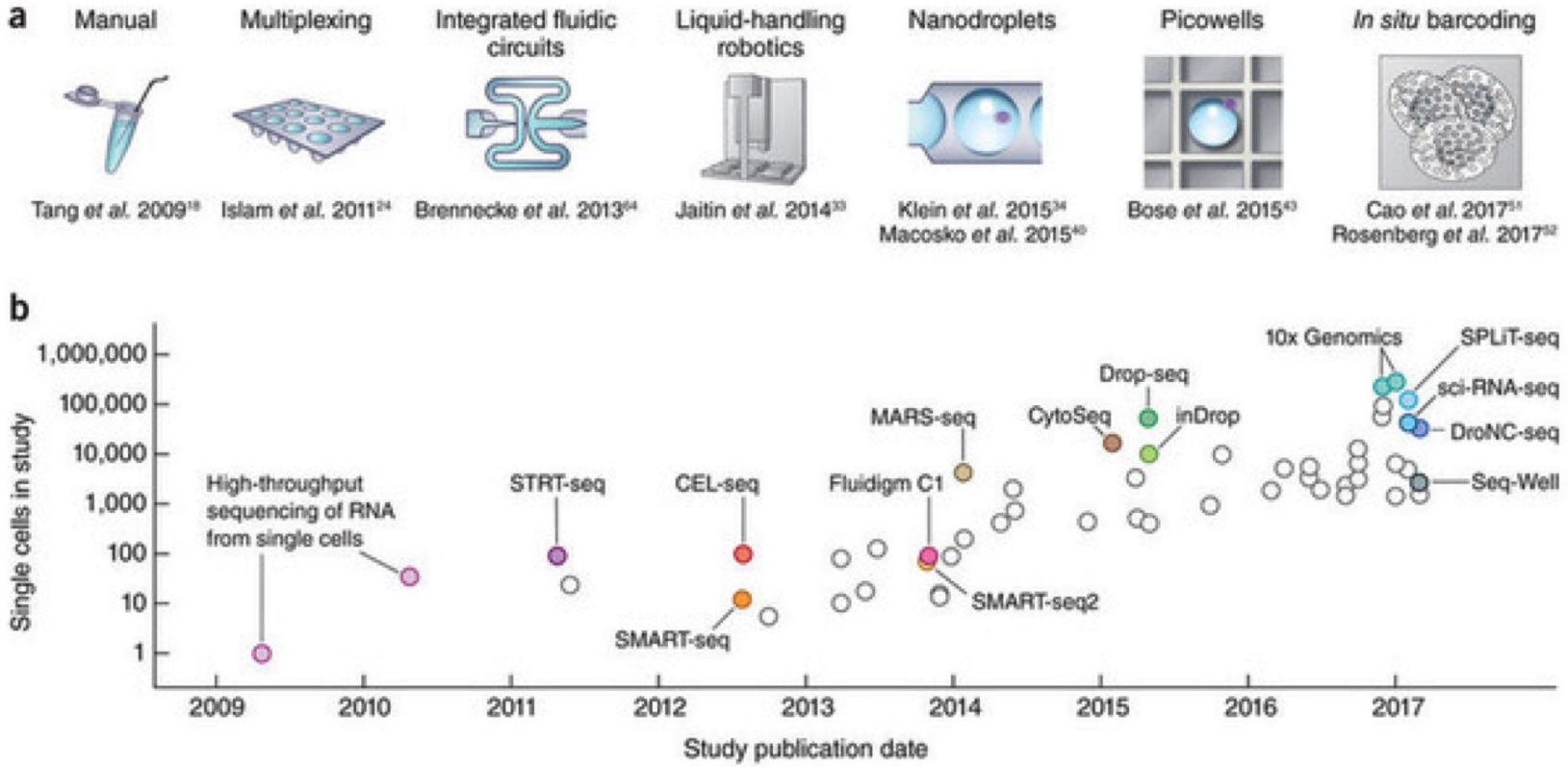
Yitong Wang (Client Liaison)
Ji-Eun Park (Project Manager)
Supervisor/Client : Dr. Davide Risso

What is RNA?



Unlike DNA, RNA differs in each cell

Single Cell RNA-seq Technology



Why do we want to cluster them?
Why in single cell level?

Noisy & Large

Bulk RNA-seq

Single Cell RNA-seq

Issue & Goal of the project:

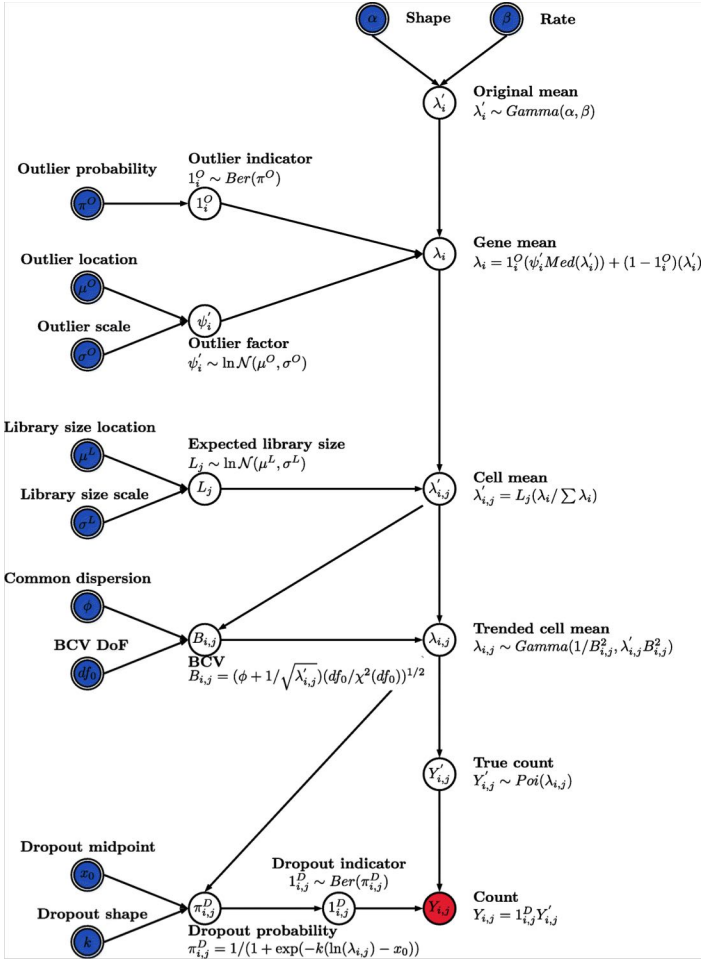
Issue: True Number of Clusters & Best Clustering Methods are unknown

> (scRNA-seq) technologies have rapidly developed. Many methods have been tested, developed, and validated using simulated datasets, current simulations are often poorly documented, their similarity to real data is not demonstrated, or reproducible code is not available.

Goal: Compare different clustering methods using simulation

Simulation - ‘Splatter’ package

Gamma-Poisson distribution is used to generate a gene-by-cell matrix of counts. Mean expression levels for each gene are simulated from a gamma distribution and the Biological Coefficient of Variation is used to enforce a mean-variance trend before counts are simulated from a Poisson distribution.



Clustering Methods

- **K-means (k : number of clusters)**

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i

  - number of clusters (k) and the center of the clusters(μ_i) are unknown
  - Even if it looks easy, it is impossible to solve as we don't know true values.
  - Use assumptions on number of clusters (k)
  - No closed formula to solve these problems

- **PAM** (similar to K-means but use medoids instead of centroids)

- **Sequential K-means** (similar but sequentially choose the cluster in each step until we run out of samples, advantage of being more robust to the specification of K)
- **Sequential PAM**

Evaluation Criteria : Adjusted rand index

- Rand Index : A measure of the similarity between two data clusterings.

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Given a set of n elements S=$\{o_1, .., o_n\}$ and two partitions of S to compare, X=$\{X_1, .., X_r\}$, a partition of S into r subsets, and Y=$\{Y_1, ..., Y_s\}$, a partition of S into s subsets, define the following:

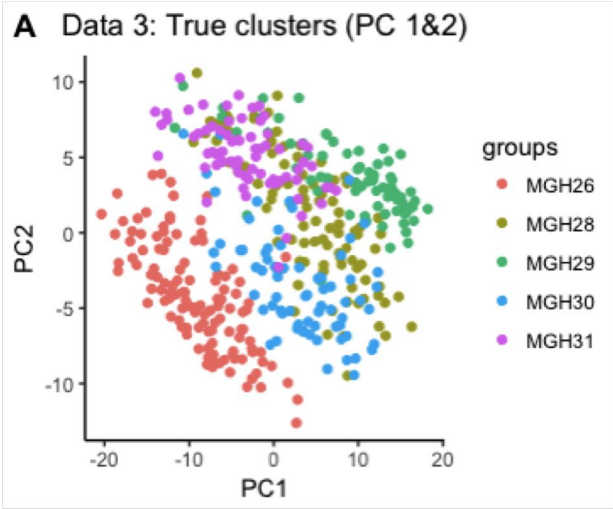
a: Number of pairs of elements that are same subset in X and in same subset in Y
b: Number of pairs of elements that are different subsets in X and in same subset in Y
c: Number of pairs of elements that are same subset in X and in different subsets in Y
d: Number of pairs of elements that are different subsets in X and in different subsets in Y

- **Adjusted Rand Index** is the corrected-for-chance version of the Rand index.

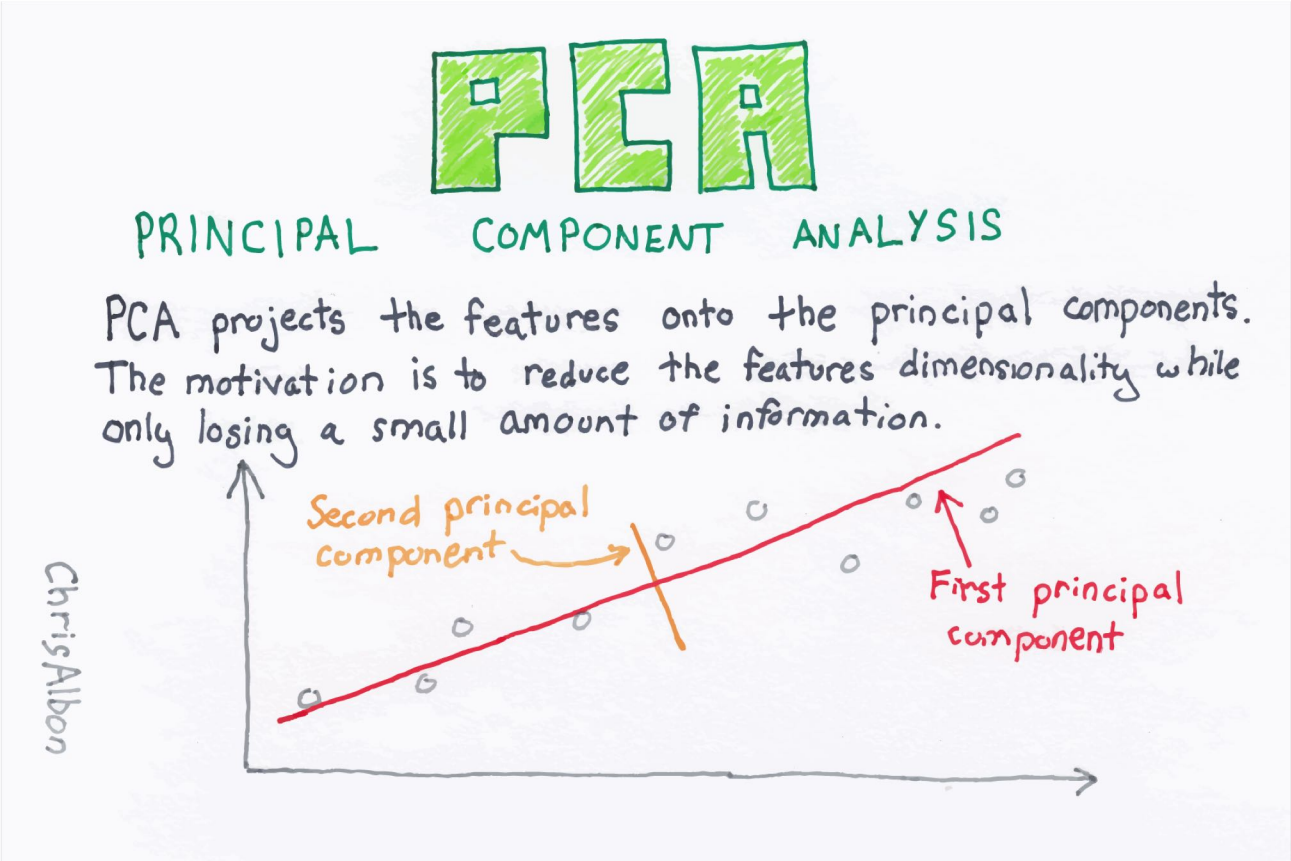
$$\overbrace{ARI}^{\text{Adjusted Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

Preprocessing : Principal Component Analysis (PCA)

- RNA-seq data is high dimensional - about 20,000 genes (variables)
- It has been shown that the dimensionality reduction step improves clustering
- Here we use Principal component analysis as a dimensionality reduction.



Principal Component Analysis (PCA) - detail



Understanding K-means

In real life, the true number of clusters
& what each clusters represent are unknown

Fruit Bowl = Set of Cells

True Clusters : 6

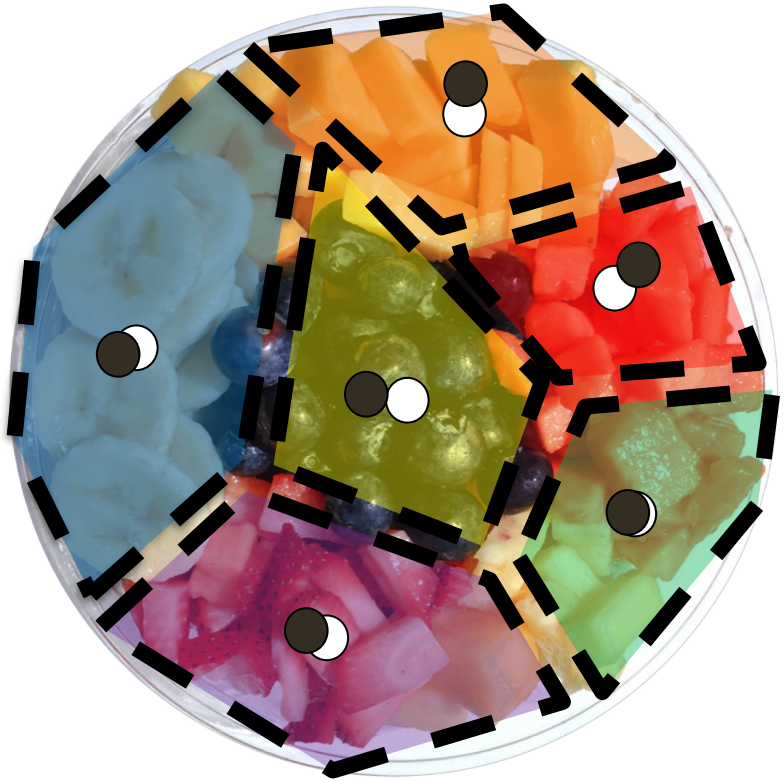
Blueberry
Banana
Cantaloupe
Watermelon
Strawberry
Pineapple

Iteration 1

6 centroids
selected randomly

Each cells are
assigned to the
nearest centroid

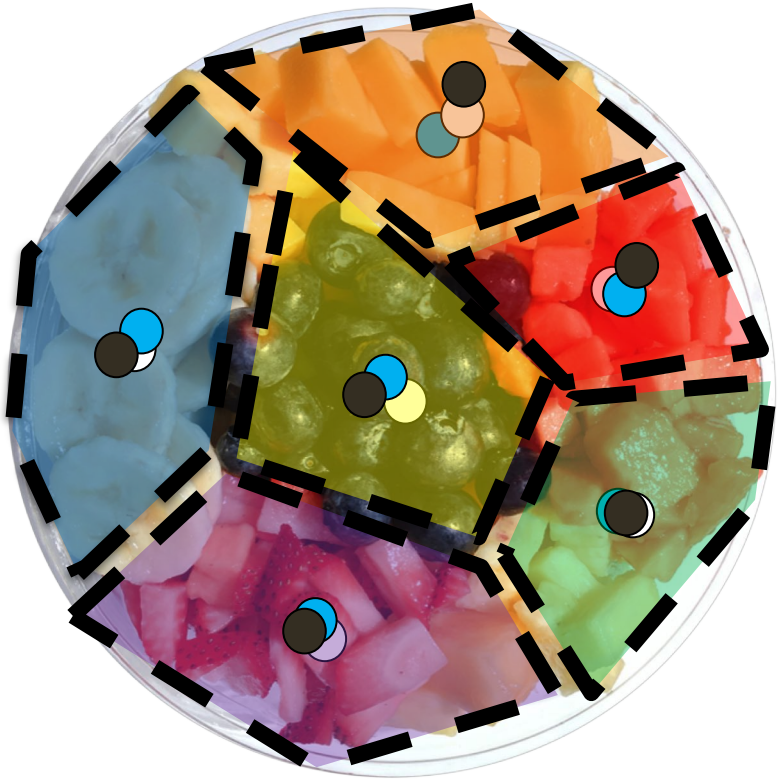
Iteration 1



Centroids are computed within each cluster

Centroids change : Black → White

Iteration 2



Centroids are recomputed within each cluster

Centroids change : Black → White → Blue

Final Iteration

Eventually, centroids & clusters converges and stop changing

→ Final cluster

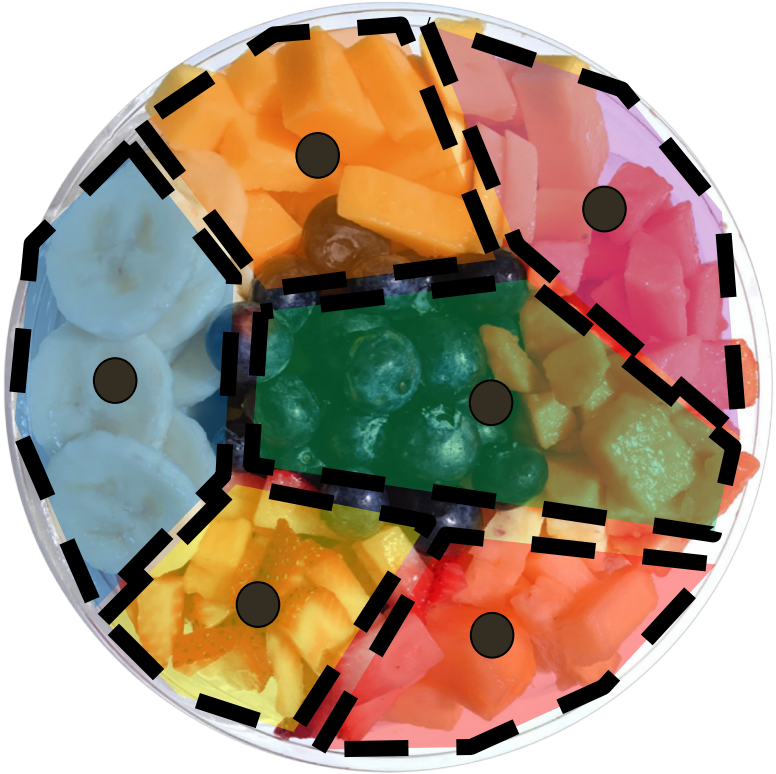
Sequential Methods

- Sequential method relies on a base clustering algorithm, like subsampling, which is iteratively re-applied after each removal of a cluster.
- For each clustering iteration, the sequential algorithm requires a method for specifying which is the “best” cluster so that it can be removed and then move on to the next iteration.
- Cluster which varies the least in its membership as the parameter K for the number of clusters is increased, as measured by the maximal percentage overlap of clusters from clusterings from K and K + 1

Sequential K-means : Iteration 1

Run k-means with k=6

Iteration 1



Run k-means with k=7

Iteration 1

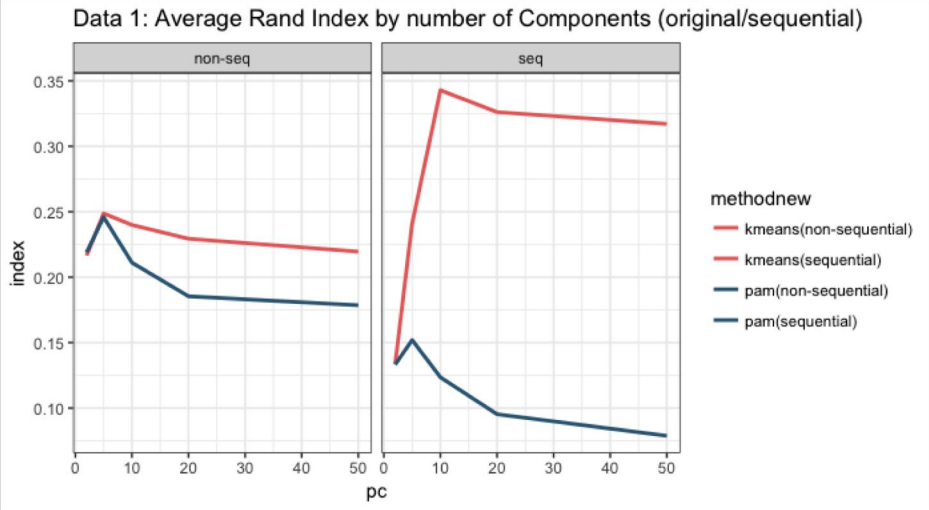
Select the most stable cluster

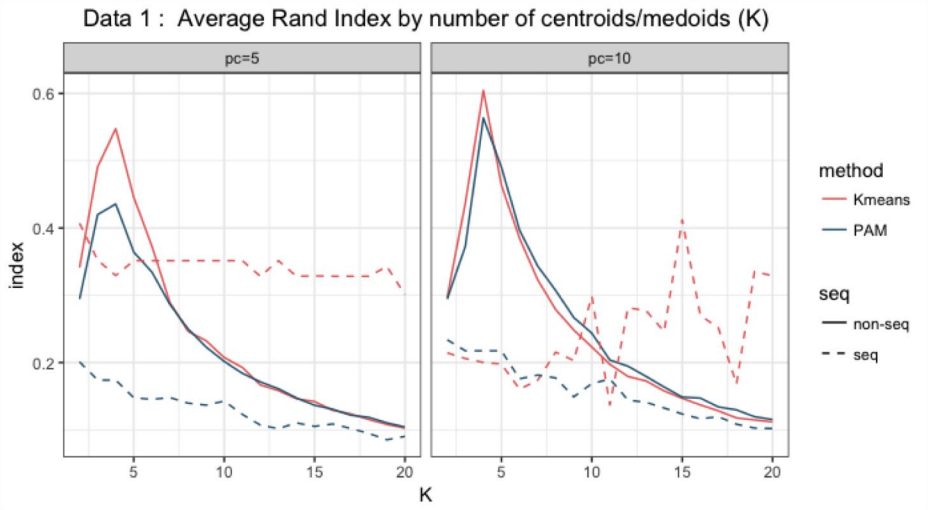
$$Max\left(\frac{\text{cardinality of intersection between the two cluster results}}{\text{cardinality of the union}}\right)$$

Iteration 2

Remove most stable cluster and start over iteration with k=5 & 6

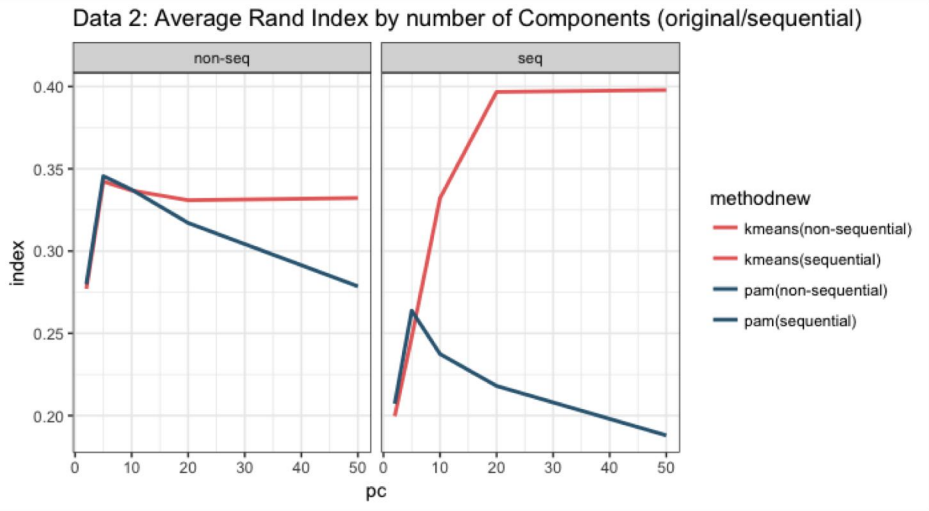
Results - Data 1

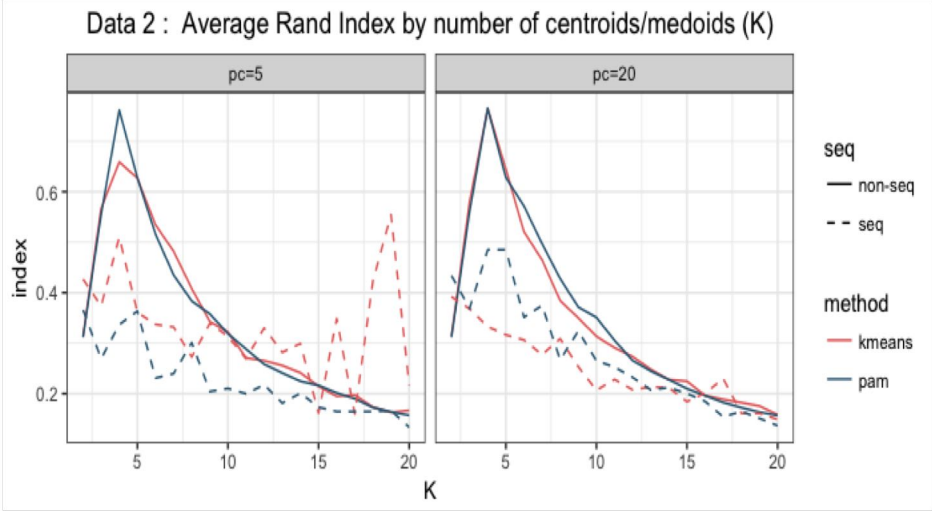




- K-means > PAM
- Sequential : stable
- When K is selected correctly, non-seq is better

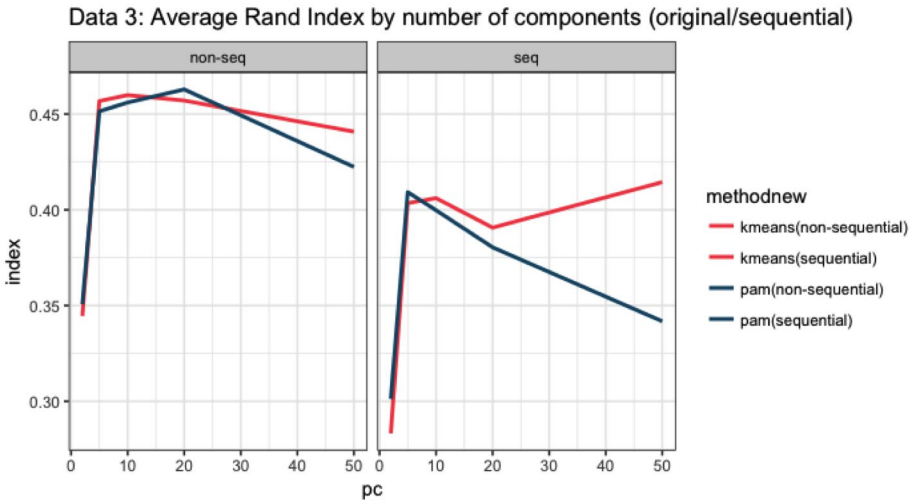
Results – Data 2

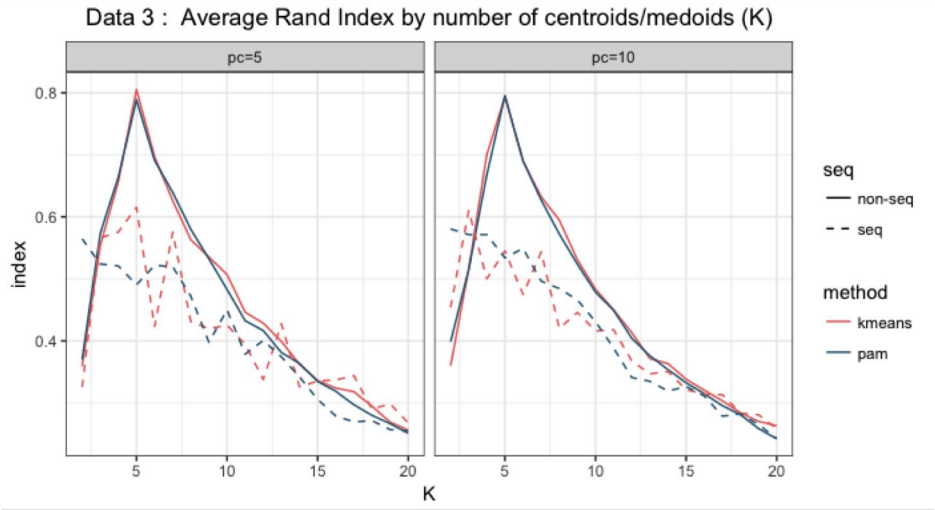




- K-means is more consistently better
- With the best value of PC and K, PAM is equally or even slightly better than K-means

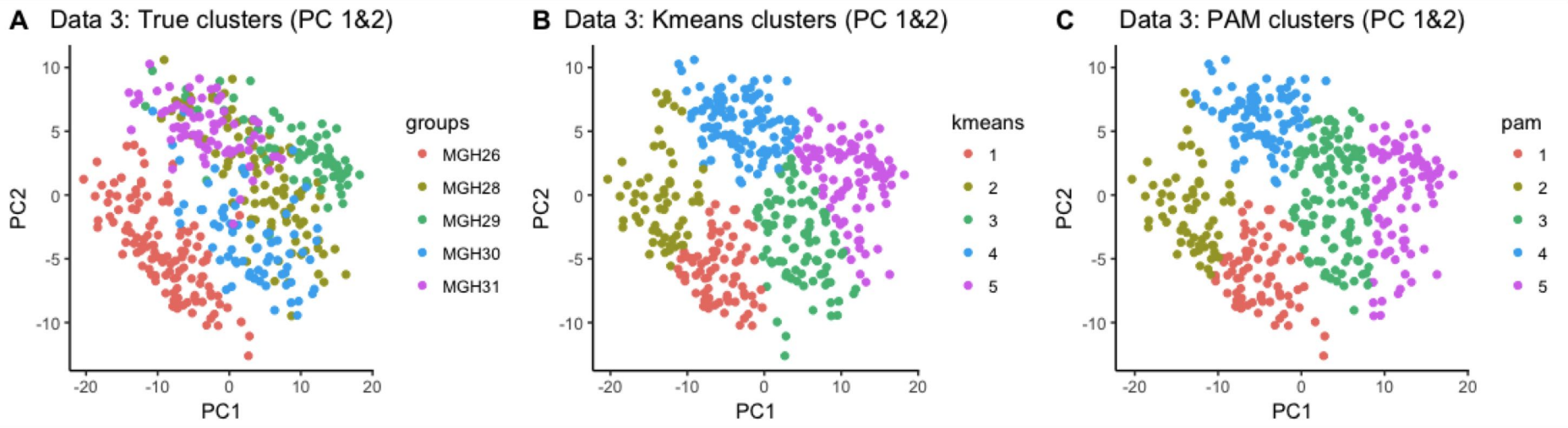
Results – Data 3 (No simulation)





- There is no significant difference between K-Means and PAM
- Sequential looks noisy

Result – Comparison



- K-means and PAM show similar results
- Does not perfectly match with the true clusters.

Conclusions

It is difficult to simulate realistic data.
For simulations, K-Means works better.
For the real data, neither K-Means nor PAM dominates as the “better” algorithm.

Non-sequential(traditional) methods gives the highest average random Index.

- Cons : Only when the number of clusters (k) is correctly assigned,

Sequential methods are robust to the choice of the number of clusters(k).

- Cons : Took twice as much time compared to traditional algorithms.

No method is perfect in our datasets

- Many possible proposed clustering methods can be attempted.
- I.e., Subsampling, Ensemble clustering, ...

Data

a) p63-HBC-diff dataset -> simulations

b)Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. (2017) **55186 features 561 samples -> simulations**

c) Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014) **5948 features 430 samples -> use real data, don't simulate**

Q&A