Review: Identification of cell types from single-cell transcriptomes using a novel clustering method

Chen Xu and Zhengchang Su

University of North Carolina at Charlotte

October 12, 2015

Problem Statement

Method

- Brief overview
- SNN graph construction
- Find quasi-cliques in the SNN graph
- Identify clusters by merging quasi-cliques
- Assign nodes to unique cluster
- Flowchart

Results

- Synthetic datasets
- Single-cell transcriptome datasets

Discussion

Motivation

- Single-cell measurements enable us to understand the cellular heterogeneity in homogenic populations and the underlying mechanisms
- The high variability in gene expression levels even between cells of the same type confounds straightforward clustering approach

Motivation

- Single-cell measurements enable us to understand the cellular heterogeneity in homogenic populations and the underlying mechanisms
- The high variability in gene expression levels even between cells of the same type confounds straightforward clustering approach
- Proposed a quasi-clique-based clustering approach
- Shared nearest neighbor (SNN) based similarity measure

Claims

- Automatically determine the number of clusters in the data
- Identify clusters of different densities and shapes
- Requires fewer parameters.

Brief overview

- Model dataset as SNN graph
- Nodes corresponds to data points e.g. vectors of gene expression levels of individual cells
- Weighted edges reflect similarity between data points
- The ultimate clustering solution is found by using graph-theoretic techniques to cluster the sparse SNN graph

SNN graph construction

- Compute pairwise distance between data points
- List k nearest neighbors for each data points
- Assign an edge $e(x_i, x_j)$ only if x_i and x_j have at least one shared KNN
- Weight on the edge

$$w(x_i, x_j) = max\{k - .5(rank(v, x_i) + rank(v, x_j) | v \in NN(x_i) \cap NN(x_j)\}$$

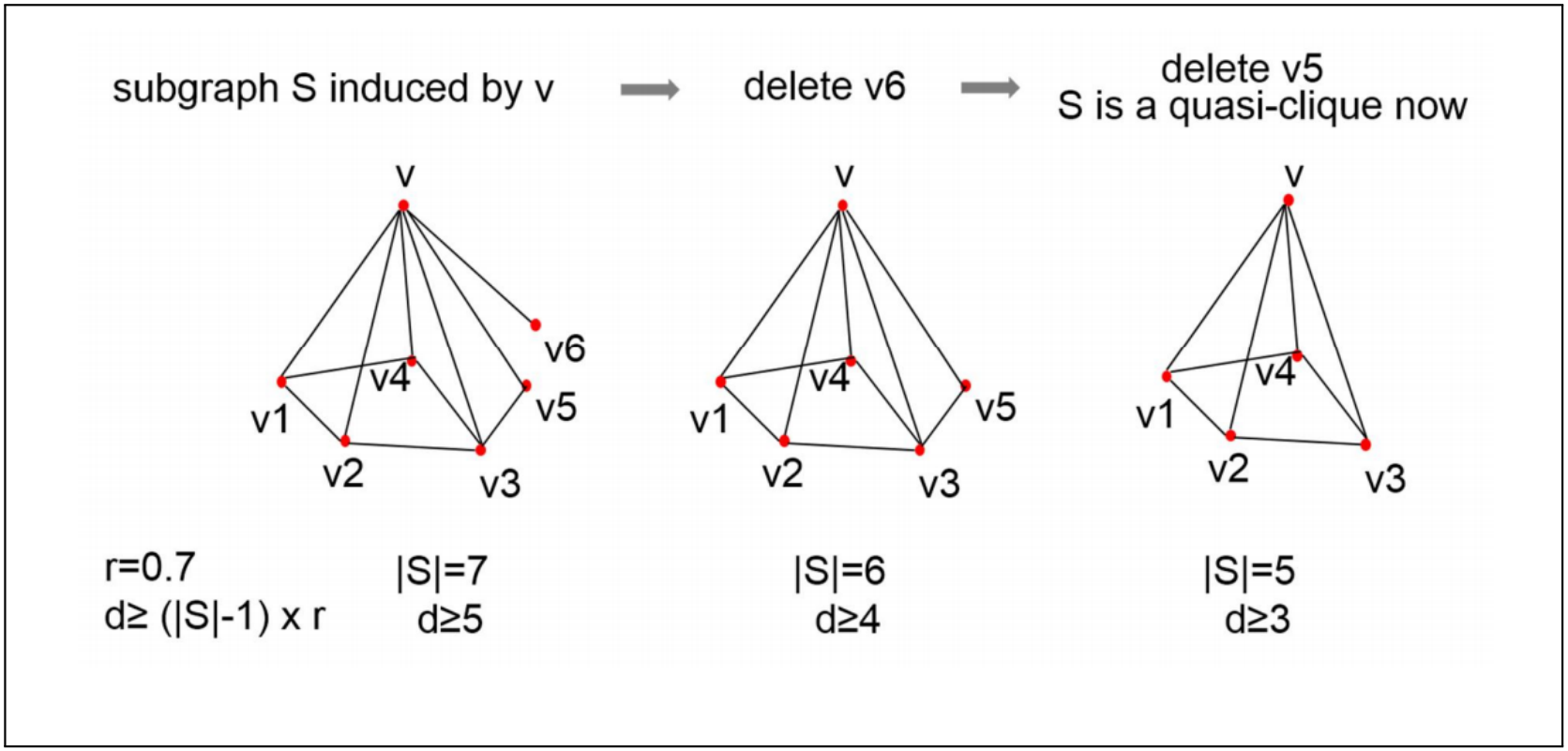
SNN graph construction

- Compute pairwise distance between data points
- List k nearest neighbors for each data points
- Assign an edge $e(x_i, x_j)$ only if x_i and x_j have at least one shared KNN
- Weight on the edge

$$w(x_i, x_j) = max\{k - .5(rank(v, x_i) + rank(v, x_j) | v \in NN(x_i) \cap NN(x_j)\}$$

- The ranking of shared neighbors of two nodes in a genuine cluster is expected to be high, thus leading to a highly weighed edge.
- The ranking of shared neighbors of two nodes from different clusters is expected to be low, resulting in a lowly weighted edge.

Find quasi-cliques in the SNN graph



Find quasi-cliques in the SNN graph

- r is a predefined threshold, defines connectivity of resulting cliques
- Eliminate redundancy by deleting quasi-cliques that are completely included in other quasi-cliques

Identify clusters by merging quasi-cliques

- Identify clusters in the SNN graph by iteratively combining significantly overlapping subgraphs starting with the quasi-cliques
- Overlapping rate

$$O_{i,j} = \frac{|S_i \cap S_j|}{min(|S_i|, |S_j|)}$$

- Merge if $O_{i,j}$ is larger than a predefined threshold, m
- Update the current set of subgraphs and recalculate pairwise overlapping rates.
- Repeat until no more merging can be done

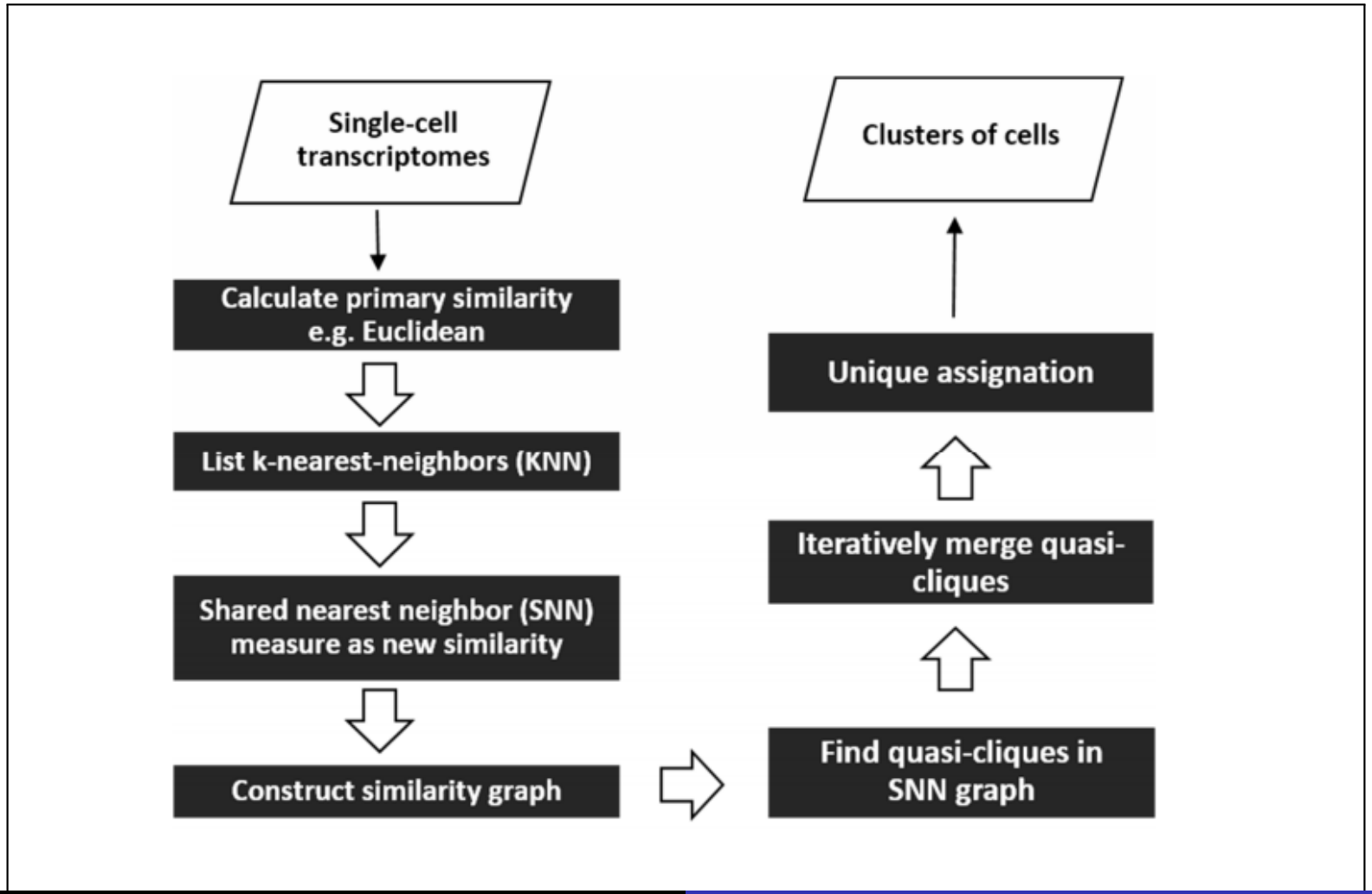
Assign nodes to unique cluster

- For each candidate cluster C that a target node v is in, calculate a score measuring the proximity between C and v,

$$Score(C, v) = \frac{1}{|C|} \sum_{i=1}^{|C|} w(c_i, v)$$

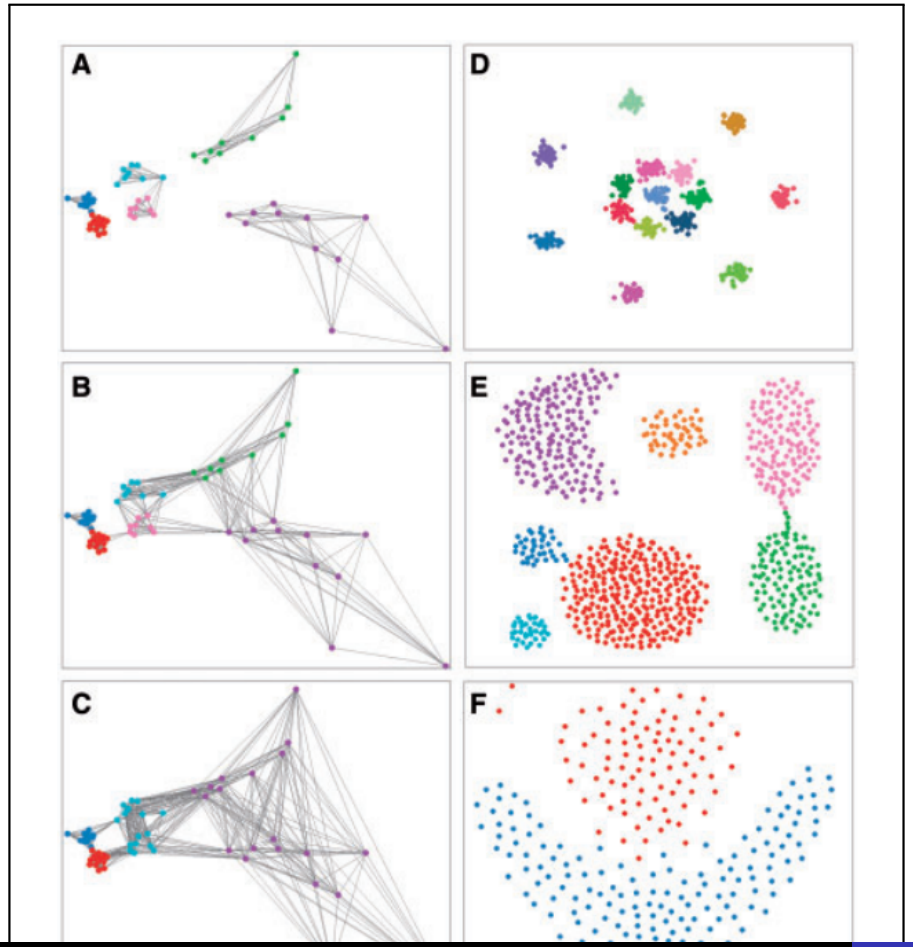
- Assign v to the cluster with the maximum score and eliminate v from all the other candidate clusters.

Flowchart

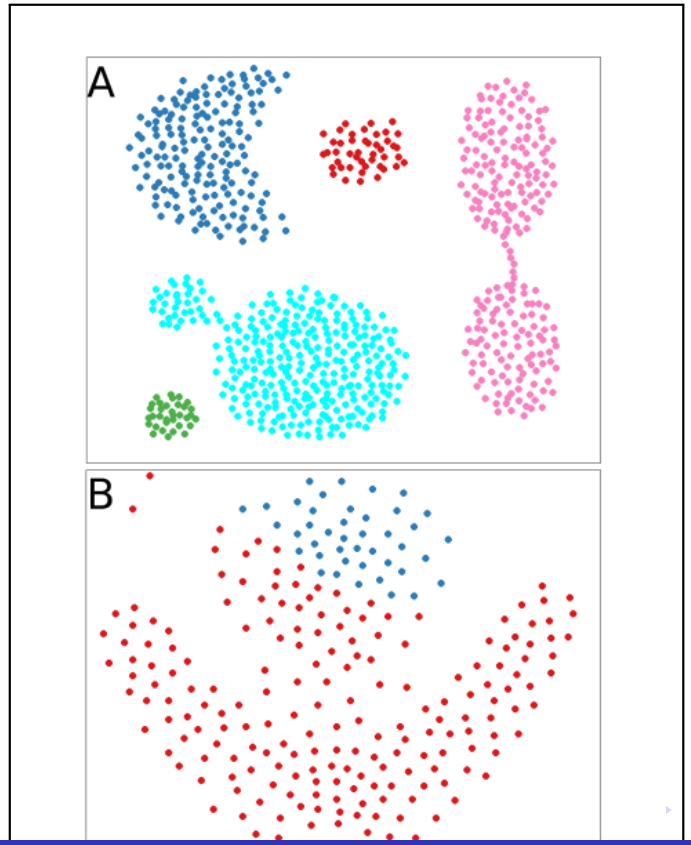

Single-cell transcriptomes
Calculate primary similarity e.g. Euclidean
List k-nearest-neighbors (KNN)
Shared nearest neighbor (SNN) measure as new similarity
Construct similarity graph
Find quasi-cliques in SNN graph
Iteratively merge quasi-cliques
Unique assignation
Clusters of cells

Synthetic datasets

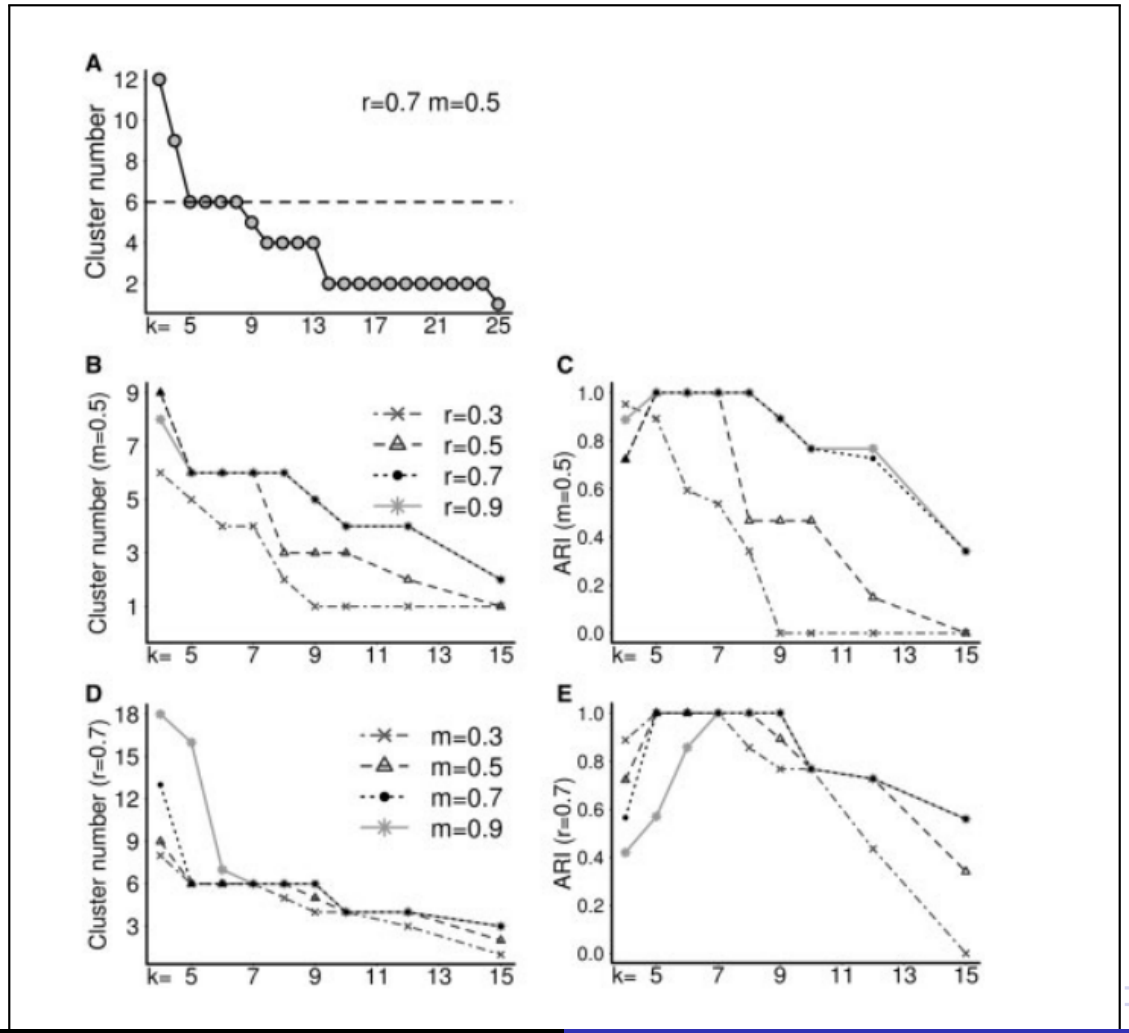
SNN-Cliq



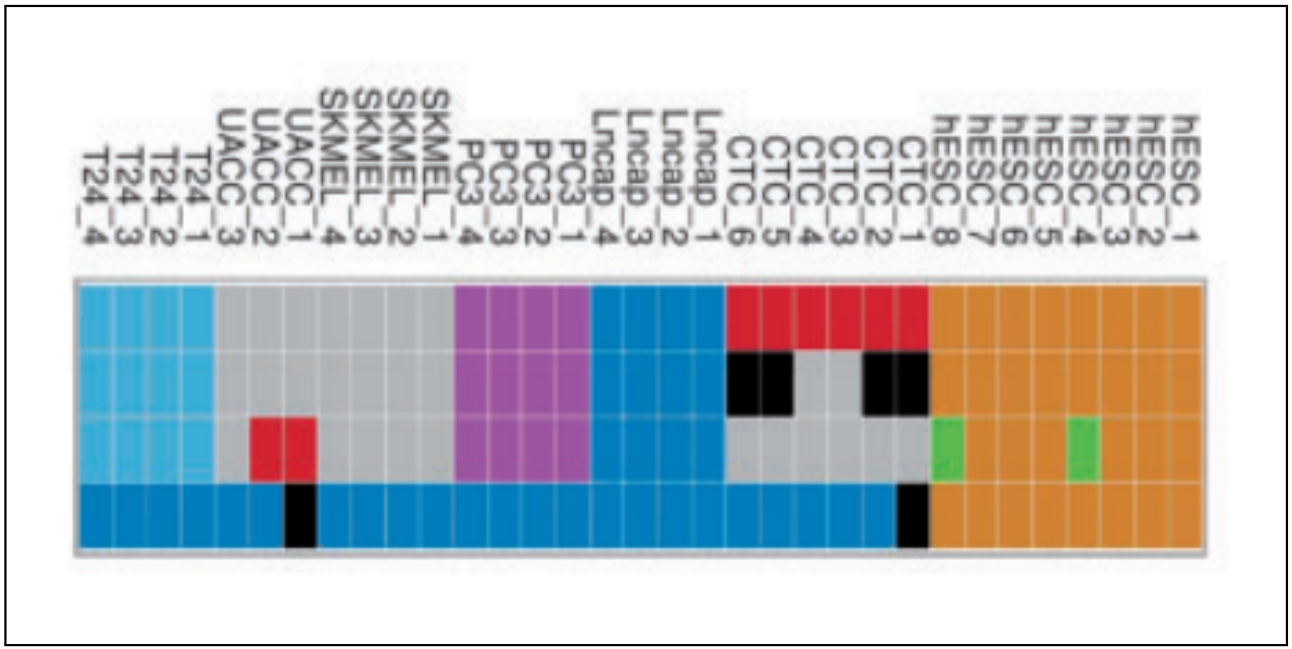
Highly Connected Subgraph Clustering



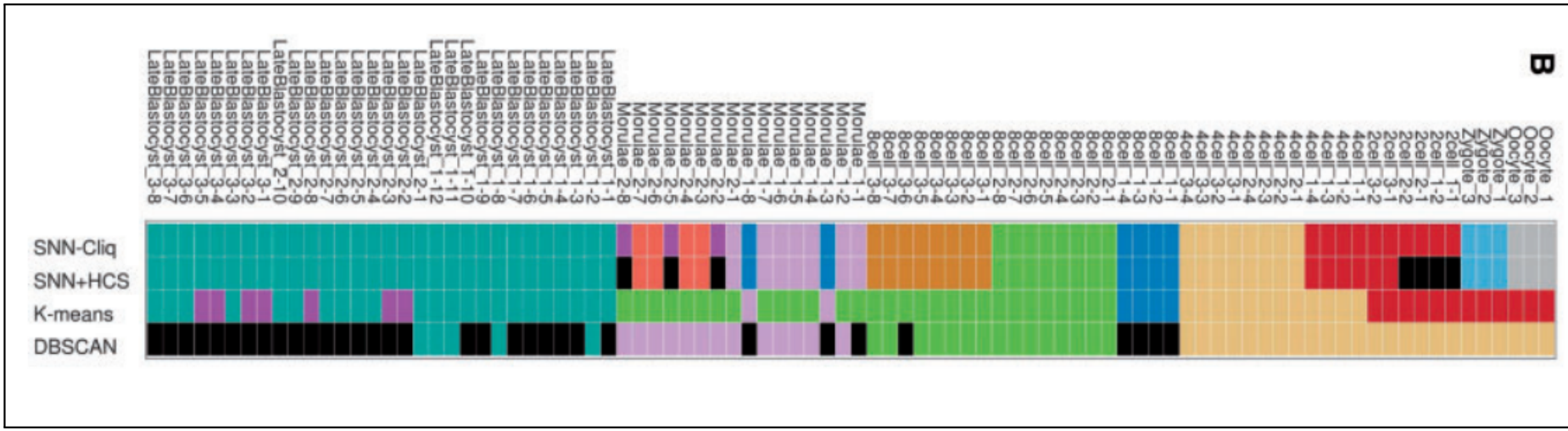
Effect of parameters on clustering



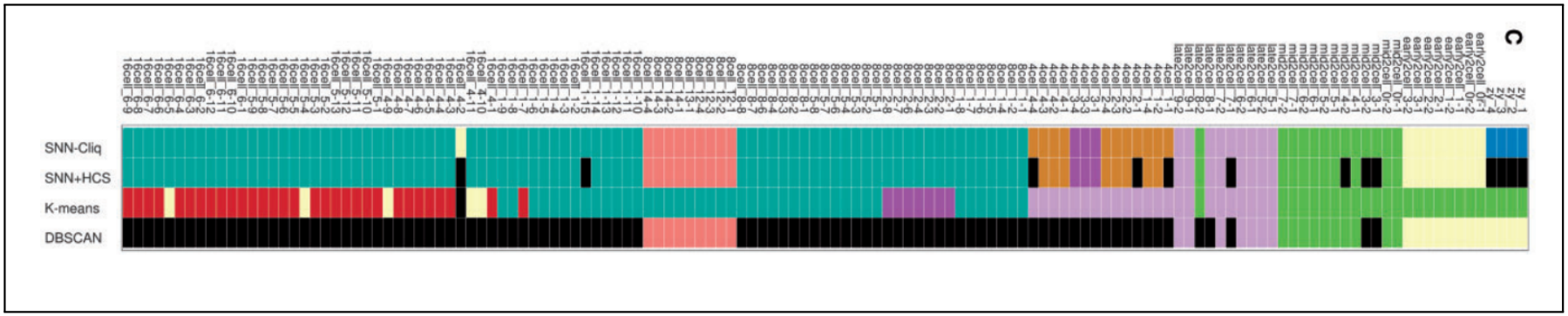
Human cancer cells



Human embryonic cells



Mouse embryonic cells



Evaluation of clustering techniques

- Purity

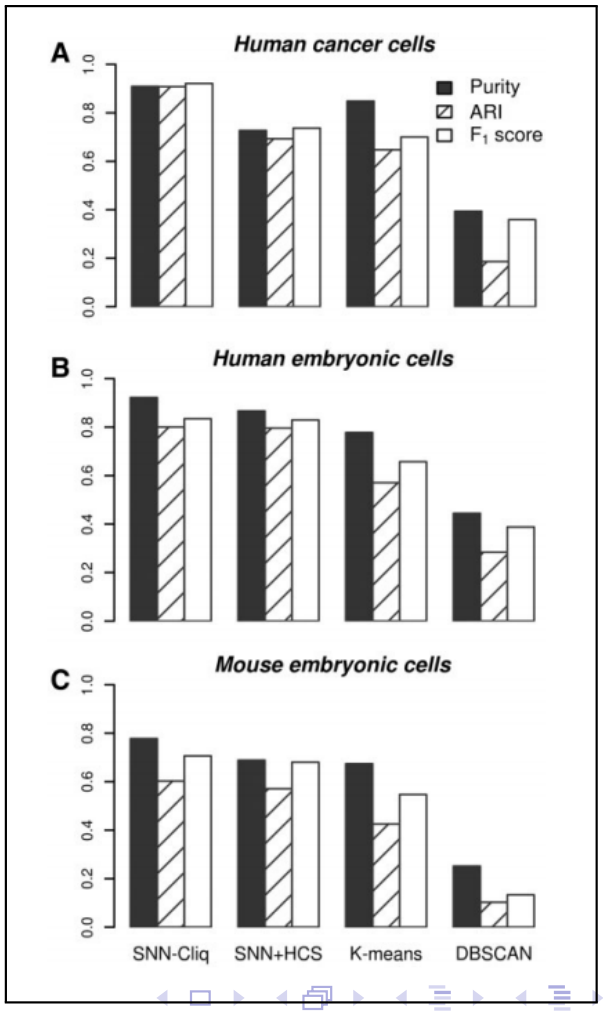
$$Purity = \frac{1}{N}\sum_{i}(v_i \cap u_j)$$

- Adjusted Rand Index

$$ARI = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)}$$

- F_1 score harmonic mean of precision and recall

$$F_1 = \frac{2a}{2a+b+c}$$



Key features

- First, it has low polynomial complexity [$O(n^2)$] and is efficient in practice.
- Since the algorithm does not make any assumptions on the structure of clusters, it can handle data with various shapes and densities.
- Easy parameter tuning.
- Performs better than existing clustering techniques.