Computer Science & Engineering Department

University of Connecticut

2017

Marmar Moussa
Ion Măndoiu

Clustering Single Cell RNA-Seq Data using TF-IDF based Methods

Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

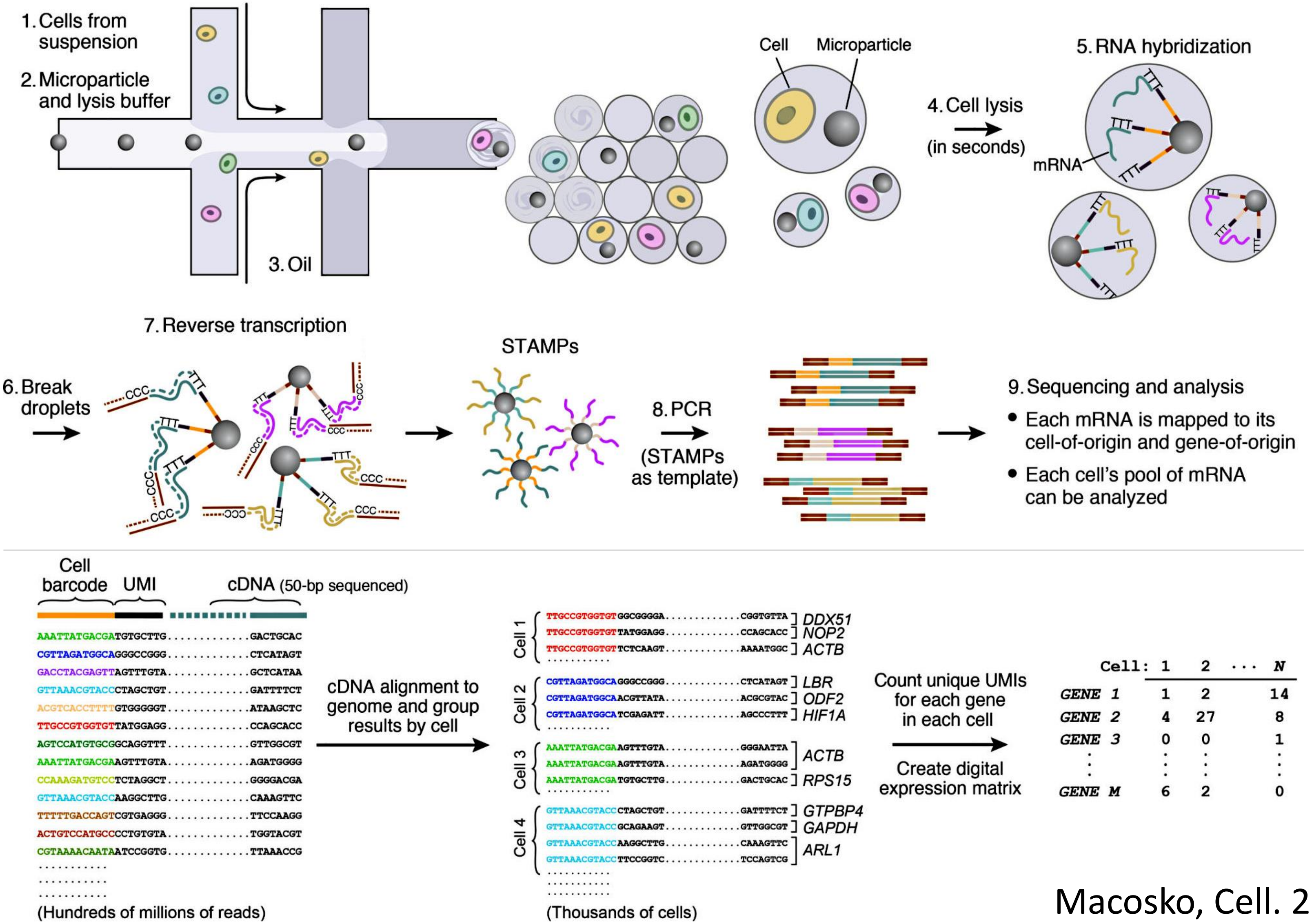
Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

Why Single Cell RNA-Seq?

- ***New***, first publication by [Tang et al. ***2009***], increased popularity by ~2014
- Measures ***distribution of expression levels for each gene across a population of cells*** (bulk RNA-seq measures average expression levels)
- Bulk→ useful for comparative transcriptomics, e.g. comparing samples of the same tissue from different species or quantifying expression signatures from ensembles, e.g. in disease studies.
- SC→ biological questions in which **cell-specific changes in transcriptome** are important (Applications?)

Droplet-based scRNA-seq technology



1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
TTT
6. Break droplets
7. Reverse transcription
CCC
STAMPs
8. PCR
(STAMPs as template)
9. Sequencing and analysis
• Each mRNA is mapped to its cell-of-origin and gene-of-origin
• Each cell's pool of mRNA can be analyzed
Cell barcode
UMI
cDNA (50-bp sequenced)
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
Cell 2
Cell 3
Cell 4
(Thousands of cells)
Count unique UMIs for each gene in each cell
Create digital expression matrix

```
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
GACCTACGAGTTAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
...........
...........
...........
```

```
Cell 1
TTGCCGTGGTGTGGCGGGGA.............CGGTGTTA ] DDX51
TTGCCGTGGTGTTATGGAGG.............CCAGCACC ] NOP2
TTGCCGTGGTGTTCTCAAGT.............AAAATGGC ] ACTB
...........
Cell 2
CGTTAGATGGCAGGGCCGGG.............CTCATAGT ] LBR
CGTTAGATGGCAACGTTATA.............ACGCGTAC ] ODF2
CGTTAGATGGCATCGAGATT.............AGCCCTTT ] HIF1A
...........
Cell 3
AAATTATGACGAAGTTTGTA.............GGGAATTA ]
AAATTATGACGAAGTTTGTA.............AGATGGGG ] ACTB
AAATTATGACGATGTGCTTG.............GACTGCAC ] RPS15
...........
Cell 4
GTTAAACGTACCCTAGCTGT.............GATTTTCT ] GTPBP4
GTTAAACGTACCGCAGAAGT.............GTTGGCGT ] GAPDH
GTTAAACGTACCAAGGCTTG.............CAAAGTTC ]
GTTAAACGTACCTTCCGGTC.............TCCAGTCG ] ARL1
...........
...........
...........
```

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Macosko, Cell. 2015

Challenges

- Noisy data: Low *RT efficiency* & *sequencing depth* causes ‘zero-inflated’ data, *cell quality*, *stochastic effects*, cell *capture bias*, gene ‘dropouts’ (a gene is observed at a moderate expression level in one cell but is not detected in another cell).
- Number of cells (thousands – million(s))

→requires adaptation of the existing methods or development of new ones.

Applications

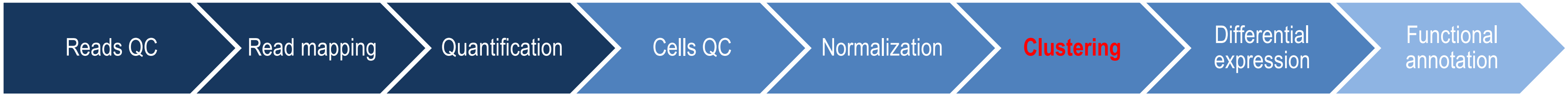
Studying heterogeneous systems:

- Cell Differentiation, e.g. early development studies, complex tissues (brain)
- Tumor Heterogeneity
- Cell Type Identification
- Stochasticity of gene expression
- Inference of gene regulatory networks across the cells.

≤

Typical scRNA-Seq Analysis Pipeline

- Primary analysis
  - Reads QC
  - Read mapping
  - Gene expression quantification
- Secondary analysis
  - Cells QC
  - Normalization
  - **Clustering**
  - Differential expression
- Tertiary analysis
  - Functional annotation



scRNA-Seq Clustering Algorithms

- Many methods available
  - K-means
  - Hierarchical clustering
  - Expectation-Maximization (GMM)
  - Graph based
  - …
- Active area of research
  - Reducing effect of confounders such as cell quality, detection rate & cell cycle phase
  - Discriminative similarity metrics
  - Scalability to millions of cells…

Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

TF-IDF Transformation

- Term Frequency x Inverse Document Frequency
  - Successfully employed in *information retrieval* field to prioritize search terms in documents
  - Considers *term frequency* (how many times a term occurs in a document)
  - Considers *document/collection frequency* (term specificity: rare terms in a collection are more *informative* than frequent terms; stop-words vs. keywords)

TF-IDF Transformation

- Term Frequency x Inverse Document Frequency for scRNA-Seq data:
  - For gene i in cell j with count f:

    $$TF_{ij} = f_{ij} / \max_{k} f_{kj}$$

  - If gene i is detected in n_i out of N cells:

    $$IDF_i = \log_2(N/n_i)$$

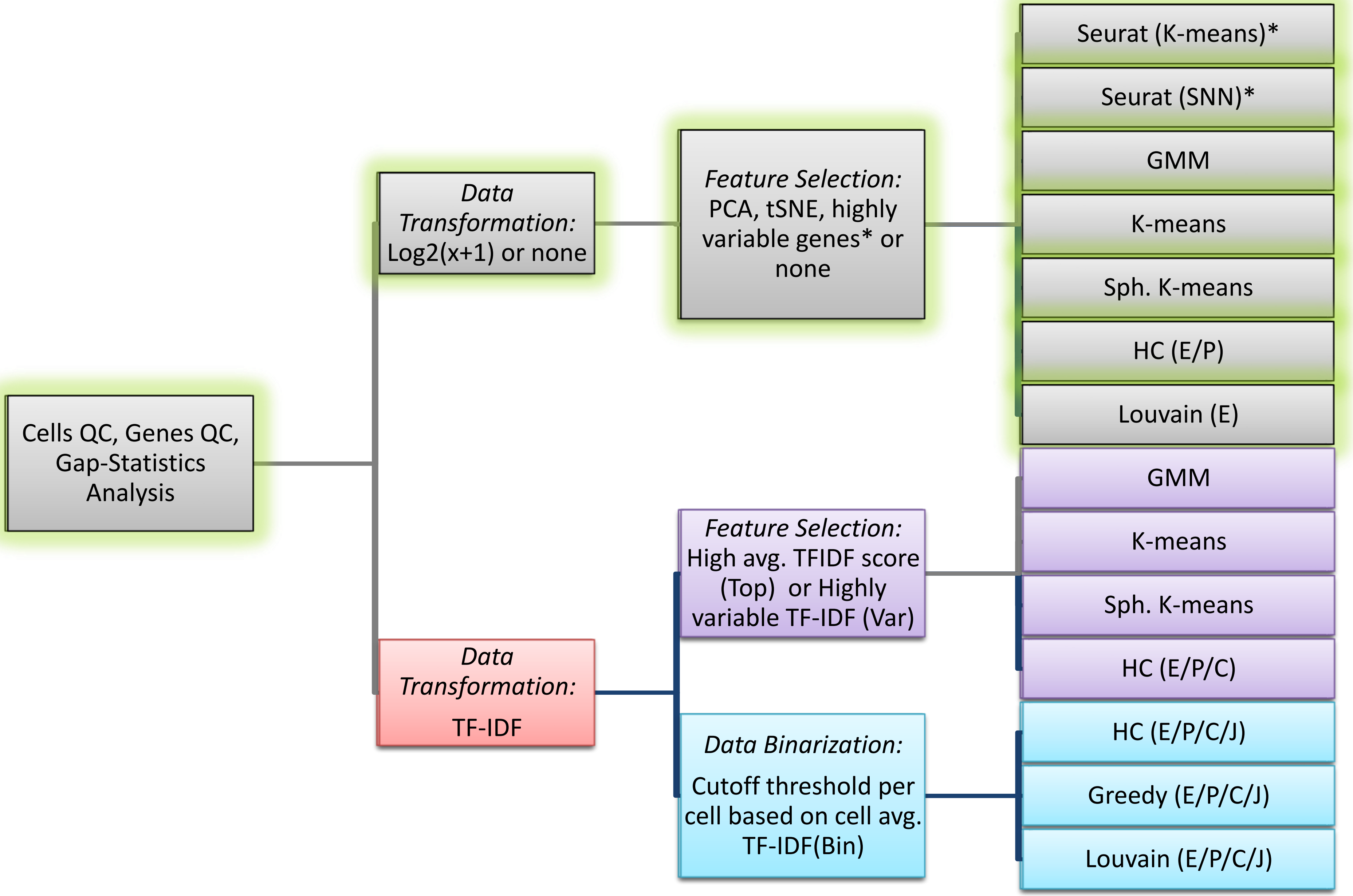
  - TF-IDF score:

    $$TF_{ij} \ * IDF_i$$

Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

scRNA-Seq Clustering Methods



Cells QC, Genes QC, Gap-Statistics Analysis

Data Transformation: Log2(x+1) or none

Feature Selection: PCA, tSNE, highly variable genes* or none

Seurat (K-means)*

Seurat (SNN)*

GMM

K-means

Sph. K-means

HC (E/P)

Louvain (E)

Data Transformation:

TF-IDF

Feature Selection: High avg. TFIDF score (Top) or Highly variable TF-IDF (Var)

GMM

K-means

Sph. K-means

HC (E/P/C)

Data Binarization:

Cutoff threshold per cell based on cell avg. TF-IDF(Bin)

HC (E/P/C/J)

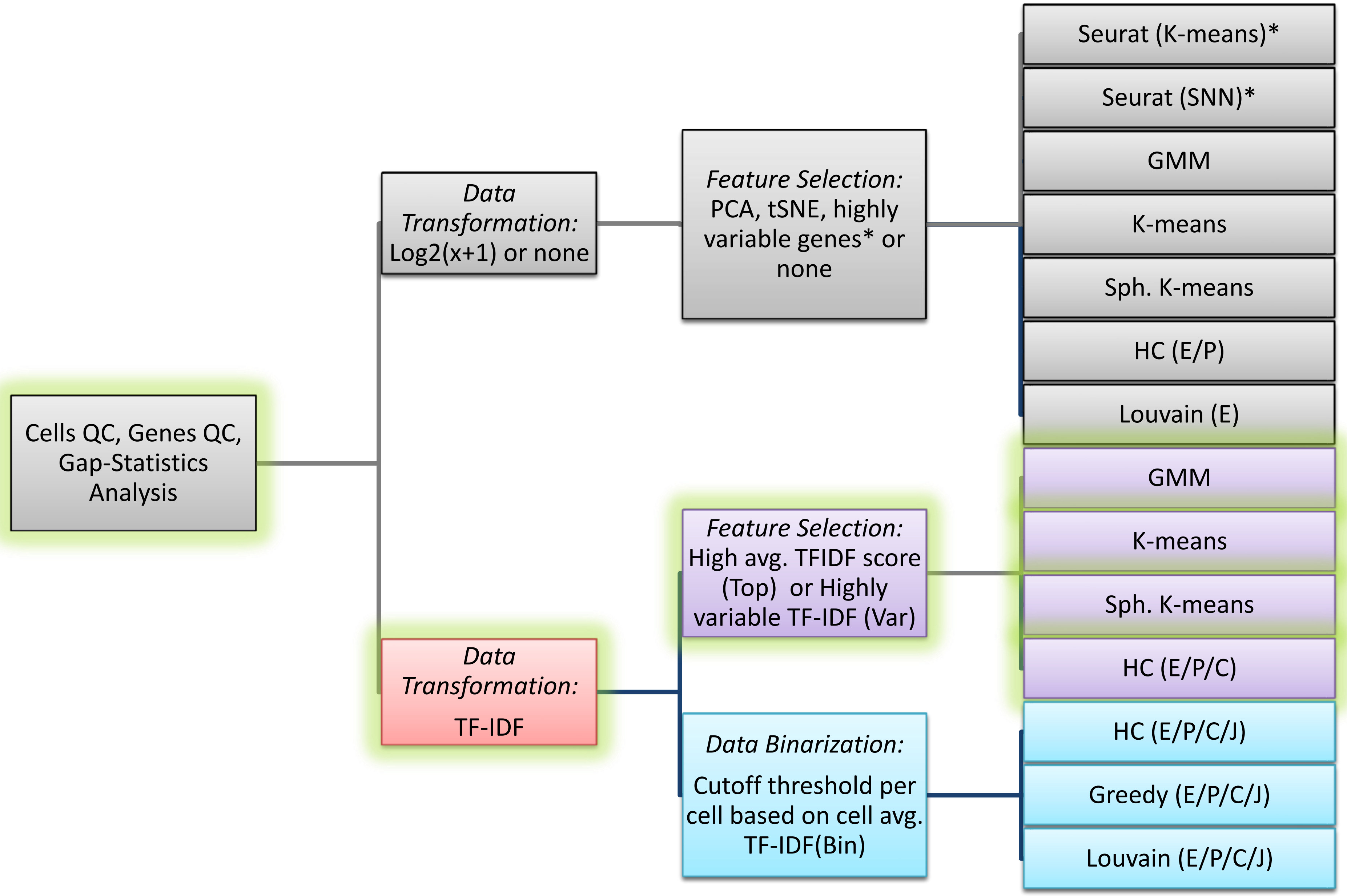
Greedy (E/P/C/J)

Louvain (E/P/C/J)

Existing scRNA-Seq clustering methods

- Seurat: DOKMeans()
- Seurat_SNN: FindClusters() shared nearest neighbor (SNN) clustering algorithm (SNN assigns objects to a cluster, which share a large number of their nearest neighbors).
- Log_PCA_GMM (Gaussian Mixture Model based clustering using mclust R package).
- K-means clustering variants:
  - Log_Kmeans (motivated by Granatum pipeline)
  - Log_PCA_Kmeans (motivated by CellRanger pipeline)
  - tSNE_Kmeans (Granatum).
- Log_PCA_sKmeans (Spherical K-means with log transform and PCA variants)
- Hierarchical Clustering variants:
  - Log_PCA_HC_E, Log_PCA_HC_P, tSNE_HC_E, tSNE_HC_P
- Log_Louvain_E (Graph based Louvain modularity optimization clustering algorithm, CellRanger)

scRNA-Seq Clustering Methods



Cells QC, Genes QC, Gap-Statistics Analysis

Data Transformation: Log2(x+1) or none

Feature Selection: PCA, tSNE, highly variable genes* or none

Seurat (K-means)*

Seurat (SNN)*

GMM

K-means

Sph. K-means

HC (E/P)

Louvain (E)

Data Transformation: TF-IDF

Feature Selection: High avg. TFIDF score (Top) or Highly variable TF-IDF (Var)

GMM

K-means

Sph. K-means

HC (E/P/C)

Data Binarization: Cutoff threshold per cell based on cell avg. TF-IDF(Bin)

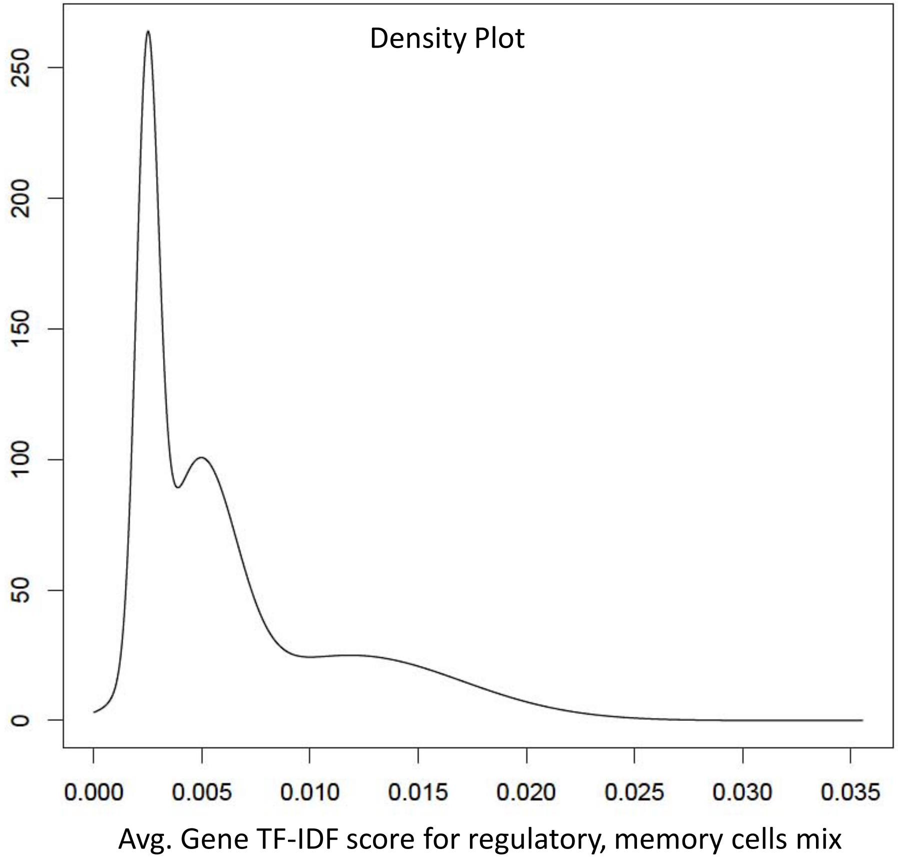
HC (E/P/C/J)

Greedy (E/P/C/J)

Louvain (E/P/C/J)

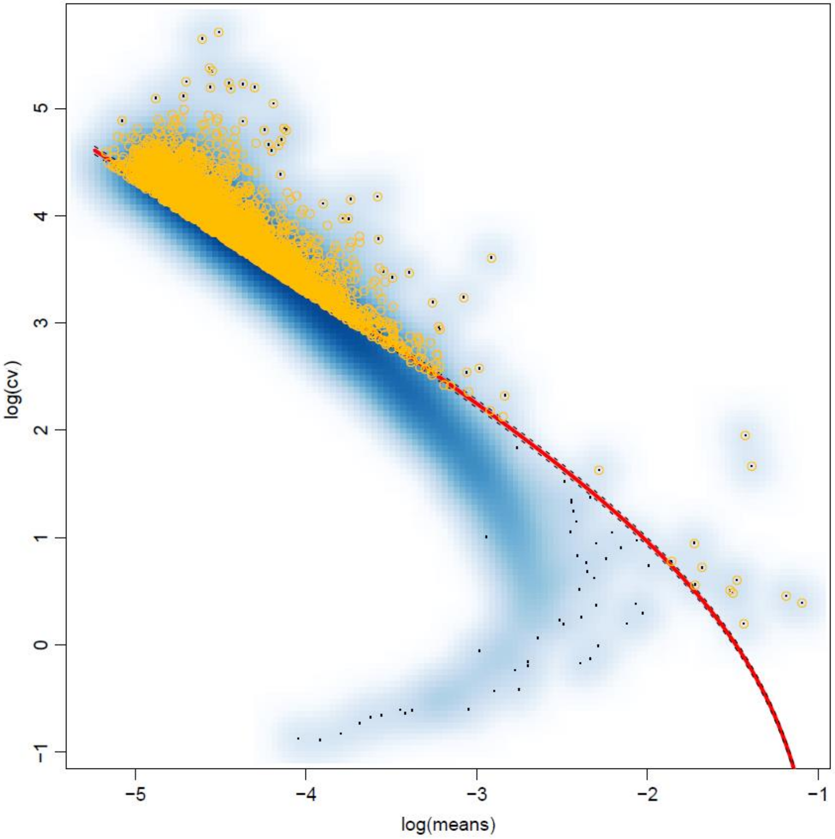
TF-IDF based gene selection

- genes with highest TF-IDF average (Top):
  1. fitted a 2-mixture GMM model to the distribution of TF-IDF gene averages
  2. selected the genes assigned to the mixture component with highest mean
  3. If more than k (3,000) genes , then rank by number of detecting cells.

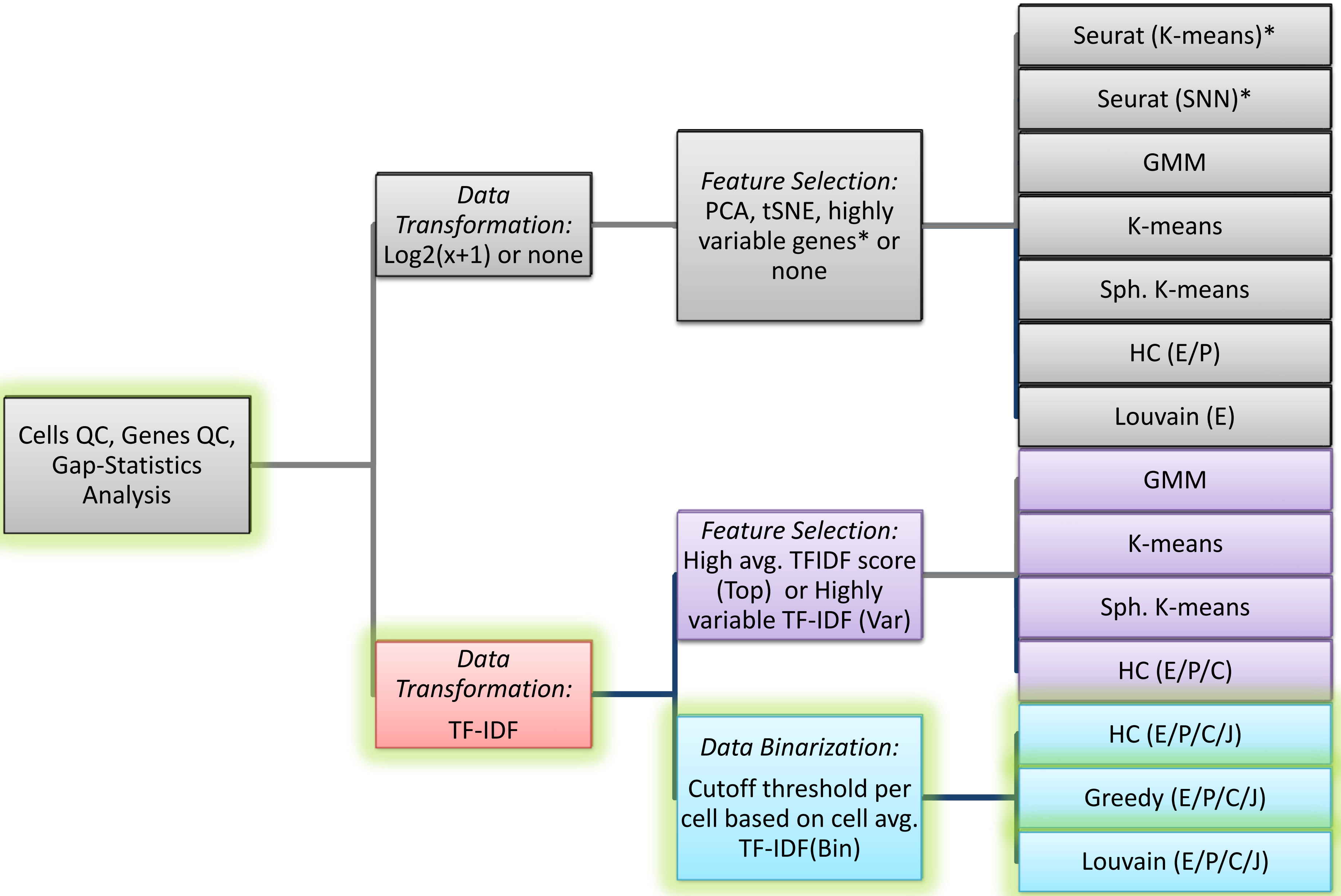


TF-IDF based gene selection

- genes with highest variability (Var) in TF-IDF values:
  - Variability decided by the relationship between the coefficient of variation (CV) and average expression levels.
  - CV (Dispersion) : ratio of the standard deviation to the mean. $CV = \frac{\sigma}{|\mu|}\ (*\ 100\%)$
  - Useful in comparison between data sets with different units or widely different means.
  - We pick the genes above the fitted line (fitted by linear regression) of CV vs. mean plot.



scRNA-Seq Clustering Methods



Cells QC, Genes QC, Gap-Statistics Analysis

Data Transformation: Log2(x+1) or none

Feature Selection: PCA, tSNE, highly variable genes* or none

Seurat (K-means)*

Seurat (SNN)*

GMM

K-means

Sph. K-means

HC (E/P)

Louvain (E)

Data Transformation: TF-IDF

Feature Selection: High avg. TFIDF score (Top) or Highly variable TF-IDF (Var)

GMM

K-means

Sph. K-means

HC (E/P/C)

Data Binarization: Cutoff threshold per cell based on cell avg. TF-IDF(Bin)

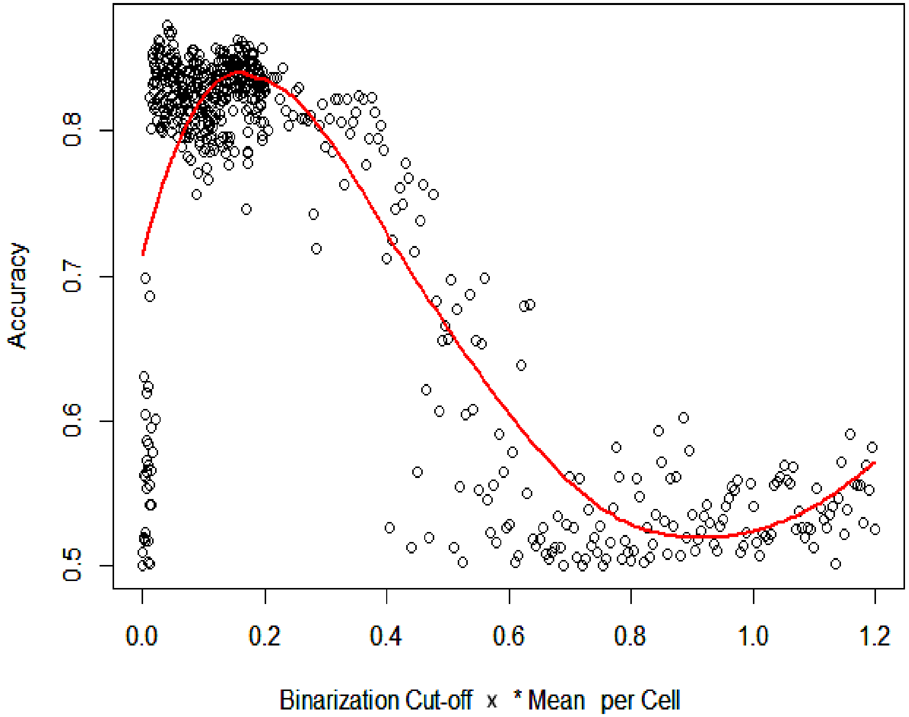
HC (E/P/C/J)

Greedy (E/P/C/J)

Louvain (E/P/C/J)

TF-IDF Binarization

- per-cell cutoff
- set the expression signature of all genes with a TF-IDF above the cutoff (‘informative’) to 1, and all remaining signatures to 0 (removing unnecessary ‘noise’).
- choice of TF-IDF cutoff can affect the clustering accuracy,
- near maximum accuracy is achieved by using a cutoff value equal to 0.1 the mean of the per-cell non-zero TF-IDF values.



(plotted for a mix of 1,000 memory and 1,000 regulatory T cells)

Graph based clustering

1. Undirected graph
   - cells : vertices,
   - edges: connecting pairs of cells for which the binarized TF-IDF transformed expression signature vectors have Euclidean, Pearson, Cosine, or Jaccard similarity above a certain cutoff value (low cutoff for dense graph)
   - Weights: edges weighted by the corresponding pairwise similarity measures
2. Clustering by greedy/Louvain modularity optimization (igraph R).
3. Keep on partitioning based on silhouette score for homogeneity and to force a minimum number of clusters when required.

>

Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- For scRNA-seq: $J = \frac{N_{11}}{N_{01}+N_{10}+N_{11}}$
  - N_{11} represents the total number of genes where cell A and cell B both express the gene.
  - N_{10} represents the total number of genes where cell A and expresses the gene and cell B not…etc.
  - 0 means no similarity, 1 means identical

- Generalized $J(x,y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$

≤

Cosine Similarity

- Given two vectors of attributes, *A* and *B*, the cosine similarity, *cos(ϑ):*

$$\cos(\theta) = \frac{\left(\sum_{i=1}^{n} A_i B_i\ \right)}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

- −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating decorrelation; 0 to 1 range for tf-idf.

≤

Modularity Optimization

- Modularity to *optimize* : value between -1 and 1 that measures the *density of links* inside communities compared to links between communities. For a weighted graph, modularity is defined as:

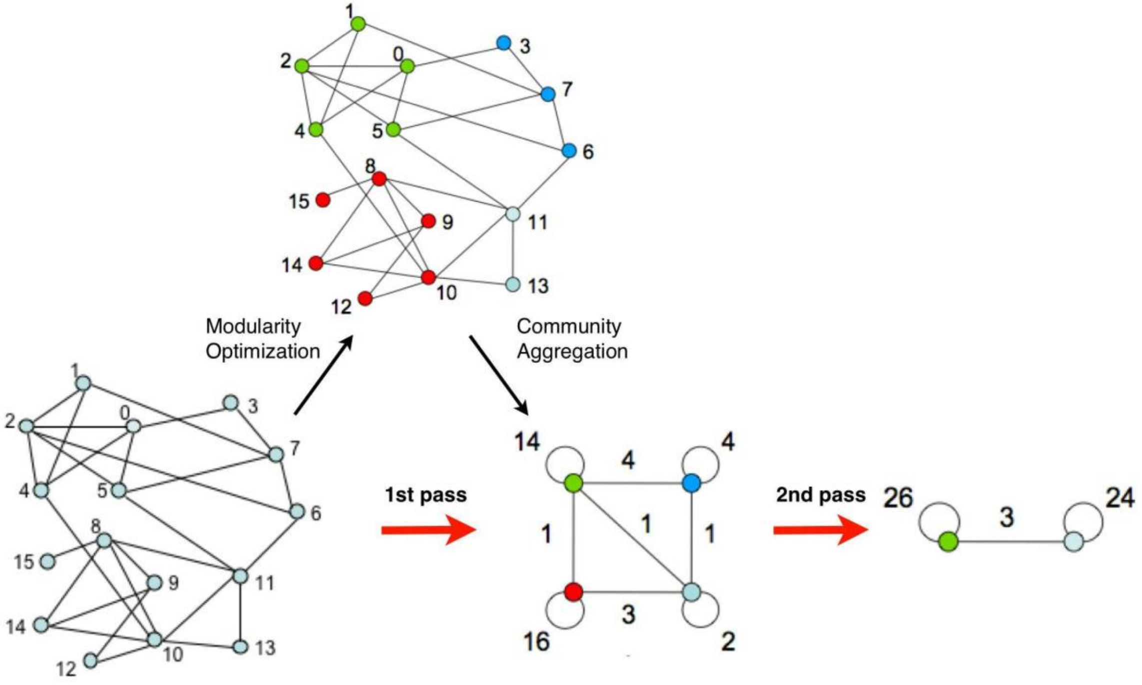
$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \ \delta(c_i, c_j)$$

- where
  - A_{ij}represents the *edge weight* between nodes i and j;
  - k_i and k_j are the *sum of the weights of the edges attached to nodes i and j* respectively;
  - m is the sum of all of the edge weights in the graph;
  - $c_i$and $c_j$are the *communities* of the nodes; and
  - δ is a simple *Kronecker* delta

≤

Louvain Method

- first small communities are found by optimizing modularity locally on all nodes (evaluates the change of modularity by removing i from its community and then by moving it into a neighboring community),
- then each small community is grouped into one node and the first step is repeated.



≤

Silhouette Score

- s(i) score between -1 & 1, average s(i) measures how well the data points are clustered.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- a(i) be the average dissimilarity of i with all other data within the same cluster
- b(i) be the lowest average dissimilarity of i to any other cluster, of which i is not a member.

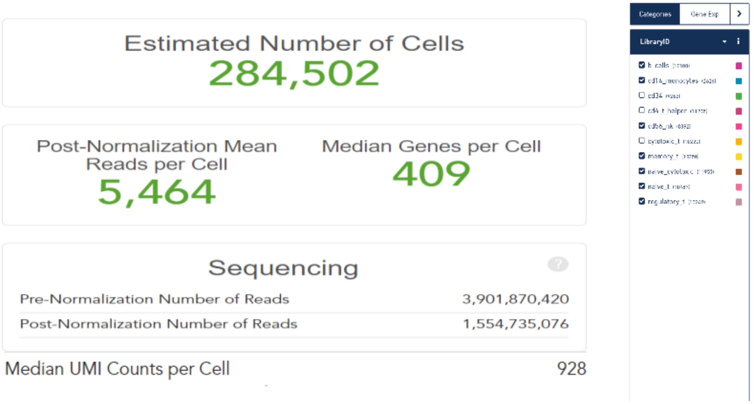
≤

Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

Experimental Setup: PBMC data set

- *FACS* sorted blood cells of 7 types [Zheng et al. 2017] using the 10x Genomics platform
  - CD14+ Monocytes
  - CD19+ B Cells
  - CD4+/CD25+ Regulatory T Cells
  - CD4+/CD45RA+/CD25- Naive T cells
  - CD4+/CD45RO+ Memory T Cells
  - CD56+ Natural Killer Cells
  - CD8+/CD45RA+ Naive Cytotoxic T Cells



Estimated Number of Cells
284,502

Post-Normalization Mean Reads per Cell
5,464

Median Genes per Cell
409

Sequencing

Pre-Normalization Number of Reads 3,901,870,420
Post-Normalization Number of Reads 1,554,735,076

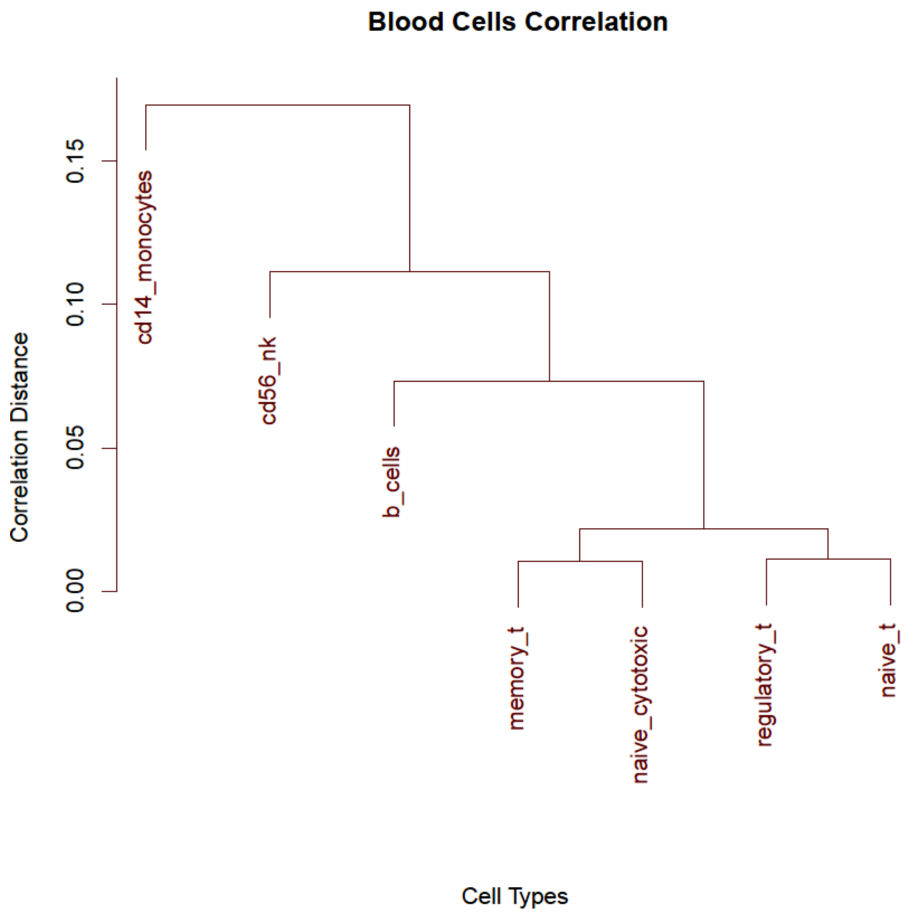
Median UMI Counts per Cell 928

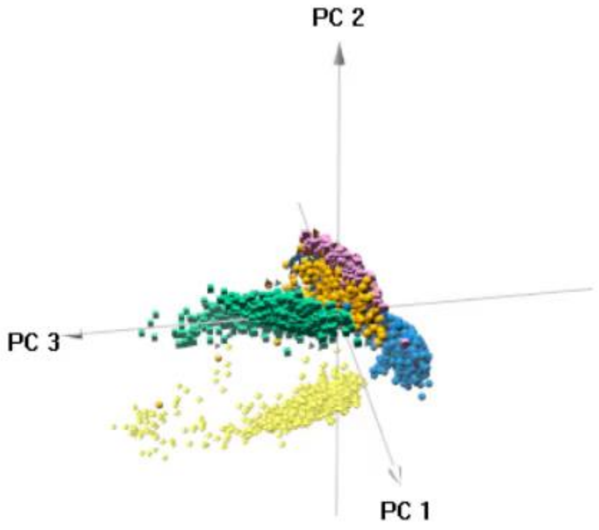
☑ b_cells [illegible]
☑ cd14_monocytes [illegible]
☐ cd34 [illegible]
☐ cd4_t_helper [illegible]
☑ cd56_nk [illegible]
☐ cytotoxic_t [illegible]
☑ memory_t [illegible]
☑ naive_cytotoxic (11953)
☑ naive_t [illegible]
☑ regulatory_t [illegible]

- 7:1, 3:1, 1:1, 1:3, and 1:7 mixtures of cell type pairs of varying dissimilarity, bootstrapping (5x sampling, 1000 cells/pair)
  - highly dissimilar: (b cells and cd14 monocytes) and (b cells and cd56 nk)
  - highly similar : (memory t and naive cytotoxic) and (regulatory t and naive t)
  - intermediate similarity: (memory t and naive t) and (regulatory t and naive cytotoxic)
- 7-way mixture, equal proportions (5x sampling, 7000 cells/mix)

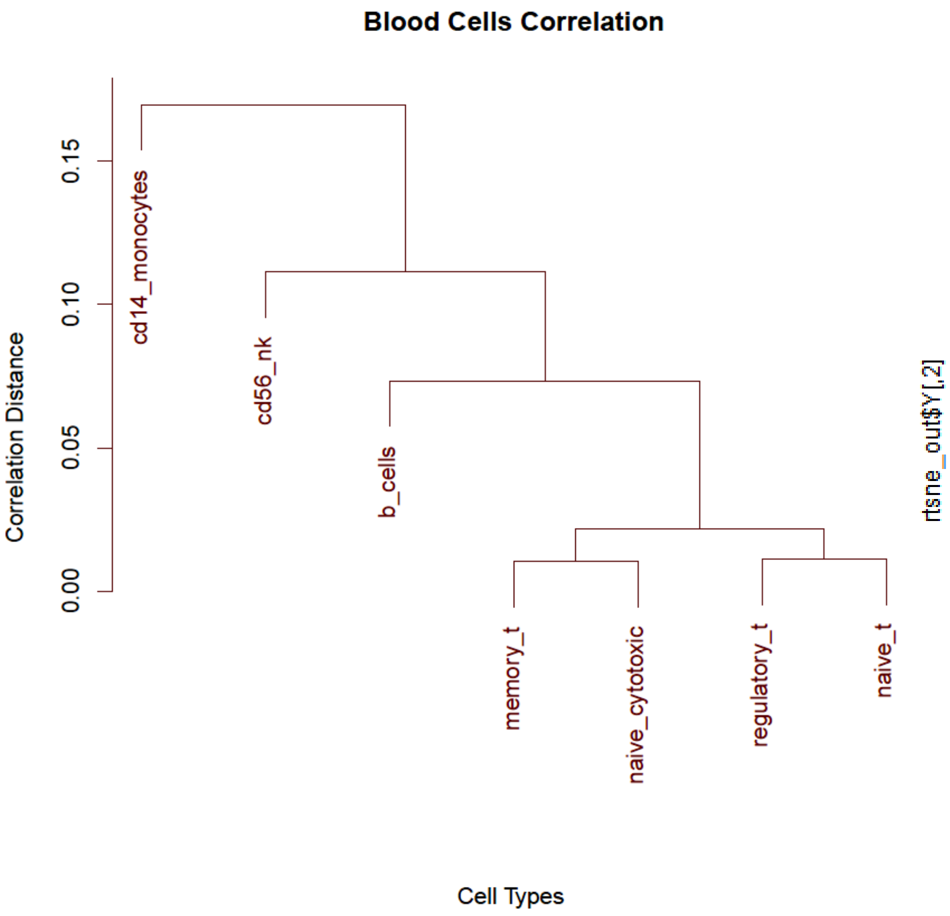
https://support.10xgenomics.com/single-cell-gene-expression/datasets
http://cnv1.engr.uconn.edu:3838/SCA/

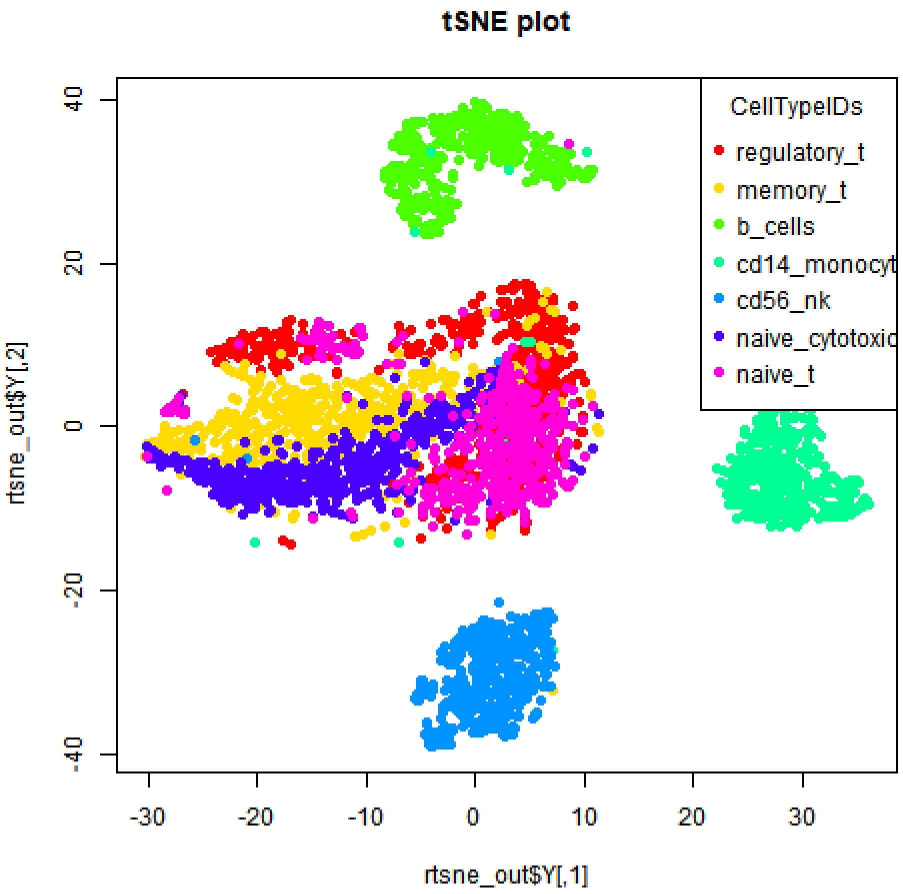
Experimental Setup: PBMC data set





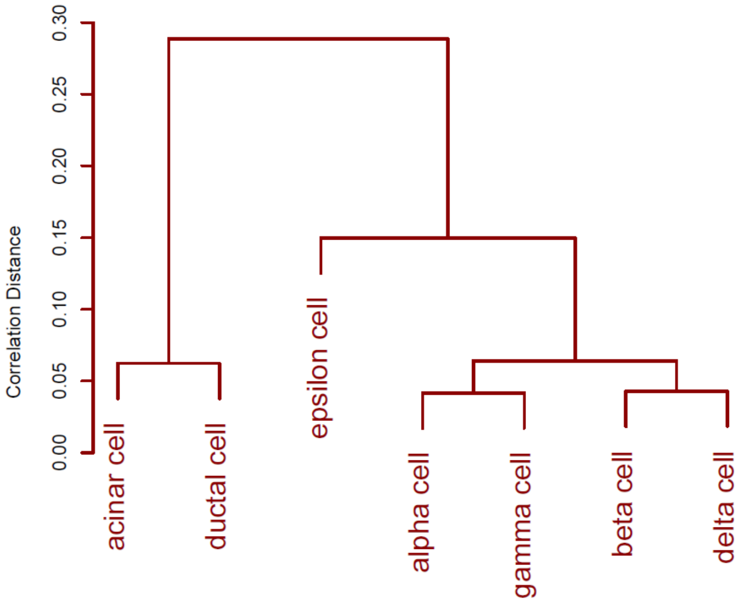
Experimental Setup: PBMC data set

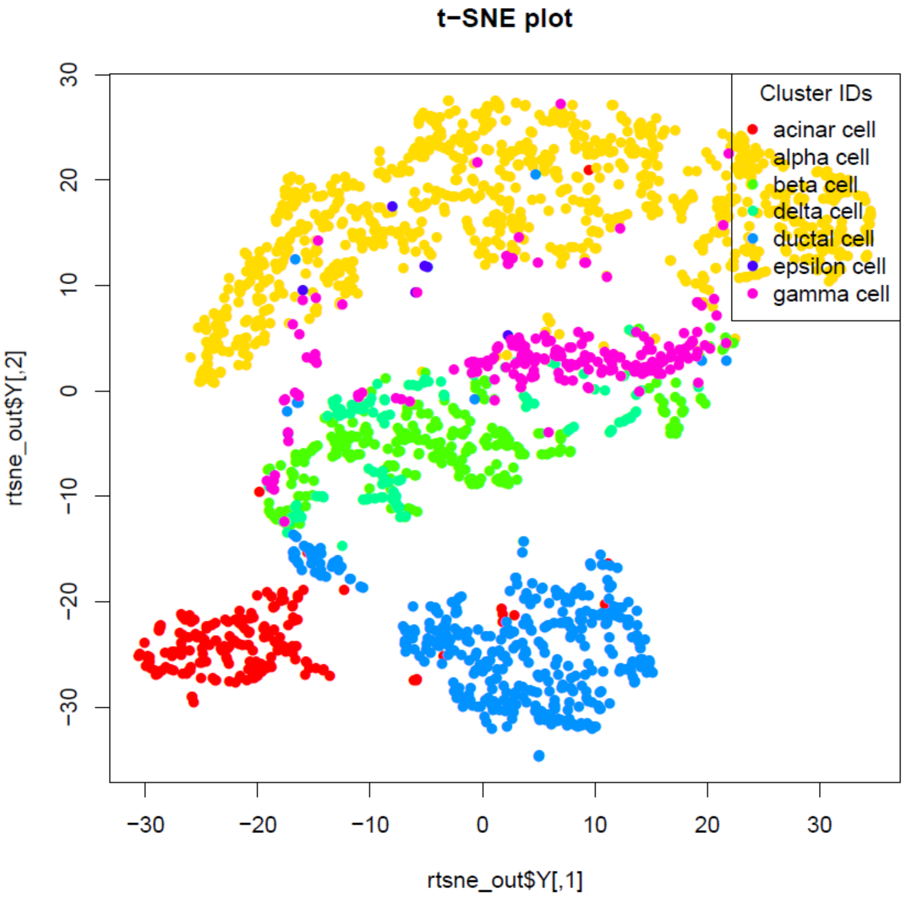




Experimental Setup: Pancreatic cells

- 2045 Pancreatic cells of 7 types [Segerstolpe et al. 2016]
  - Annotated based on known markers
  - Capture proportions: (185 acinar cells, 886 alpha cells, 270 beta cells, 197 gamma cells, 114 delta cells, 386 ductal cells, and 7 epsilon cells)





Cells’ & Genes’ QC

- For all 10x Genomics datasets:
  - filtered cells based on number of detected genes and total UMI count per cell.
  - removed outliers based on the median-absolute-deviation (MAD) of cell distances from the centroid of the corresponding cell type.

$$MAD = median(|x_i - median(x)|)$$

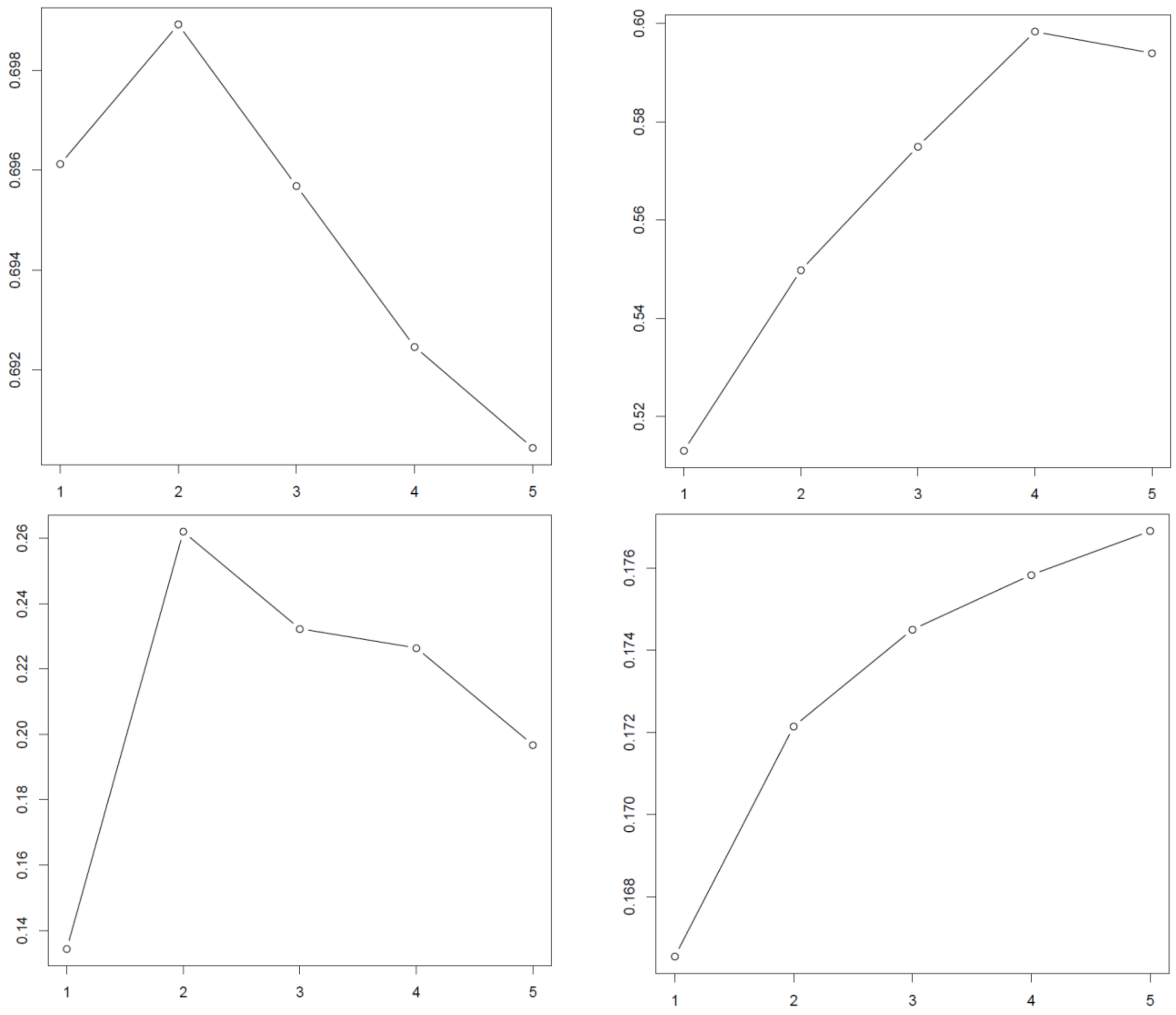
  - basic gene quality control by applying a cutoff on the minimum total UMI count per gene across all cells and removing outliers based on MAD. (outlier>5MAD)

- For Pancreatic cells:
  - No cell QC
  - marker genes with unusually high expression levels (INS for beta cells, GCG for alpha cells, SST for delta cells, PPY for PP/gamma cells, and GHRL for epsilon cells) were removed prior to clustering to eliminate thepossibility that they drive the clustering by themselves.

‘Optimal’ number of clusters

- the optimal number of clusters is selected as
$$argmax_k Gap_n(k)$$
- where the *Gap Statistic* [Tibshirani, 2001] for clustering n points into k clusters is given by
$$Gap_n(k) = E_n^*\left(\log(\mathrm{W}_\mathrm{k}^*)\right) - \log(\mathrm{W}_\mathrm{k})$$
- W_k is the normalized sum of pairwise distances in the k clusters
- W_k^* its expectation under a suitable null reference distribution (Monte Carlo sampling).

Example: Regulatory_t and naïve_t data set



Clockwise from top left: Gap statistics for log-transformed, log-transformed PCA, tSNE, and TF-IDF transformed and binarized expression levels of a 7:1 mixture of regulatory t and naive t cells.

The x-axis gives the number of clusters K and the y-axis gives the gap statistic.

Accuracy measures

- Overall Accuracy:

$$\sum_{i=1}^{K} C_i \,/ \sum_{i=1}^{K} N_i$$

- Average Cluster Accuracy:

$$\frac{1}{K}\sum_{i=1}^{K}\frac{C_i}{N_i}$$

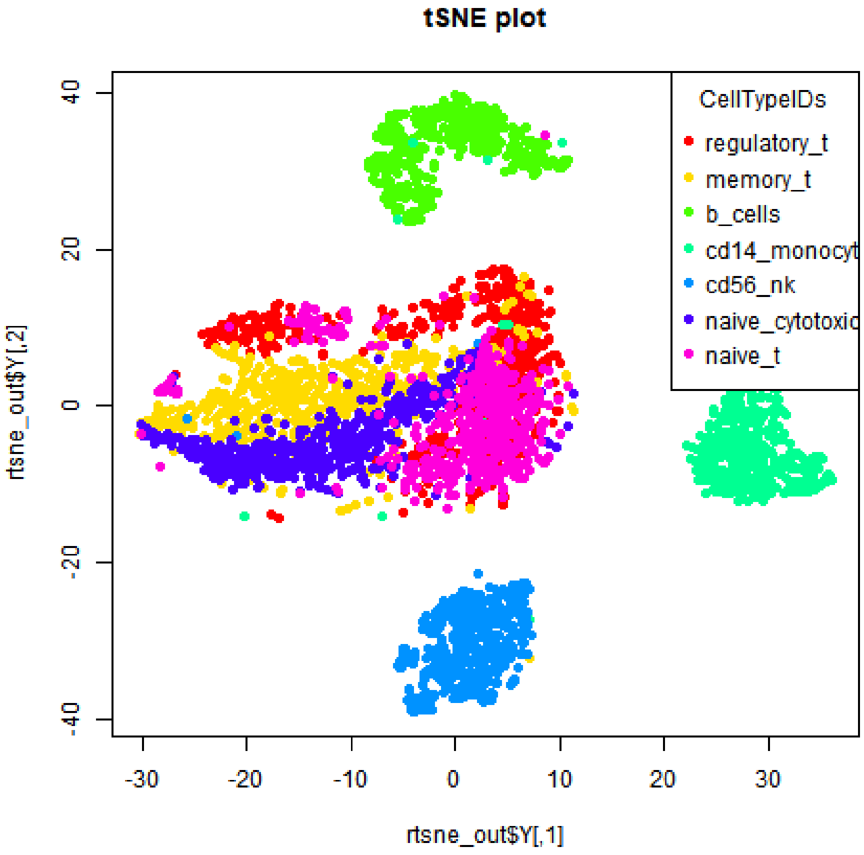
  - where K is the number of classes,
  - N_i is the number of samples in class i,
  - and C_i is the number of correctly labeled samples in class i.

- Note that both are identical for 1:1 mixtures, but may differ significantly for imbalanced datasets, as macro-averaging gives equal weight to the accuracy of each class, whereas micro-averaging gives equal weight to each cell classification decision.

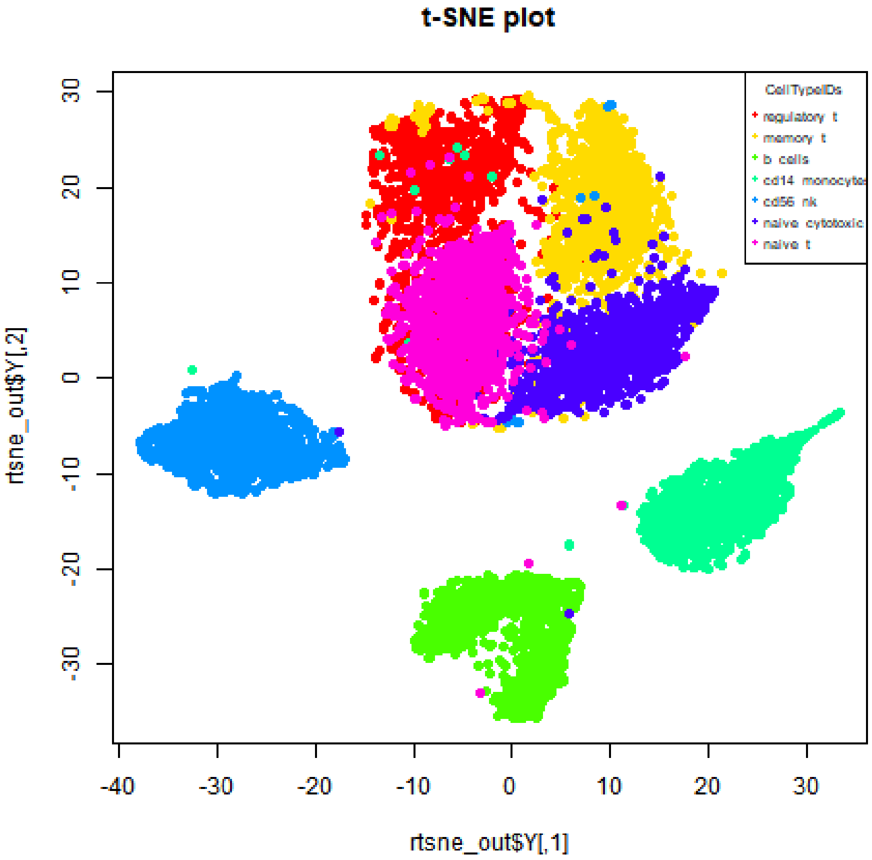
Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

t-SNE TF-IDF transformation

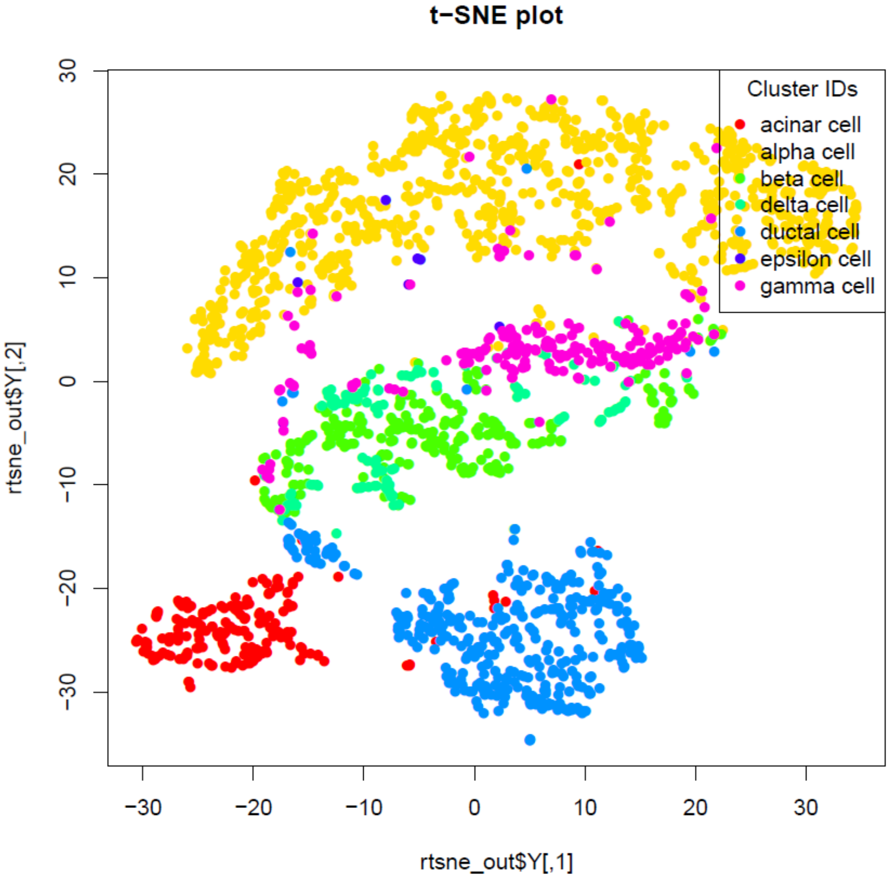


Raw PBMC data t-SNE plot

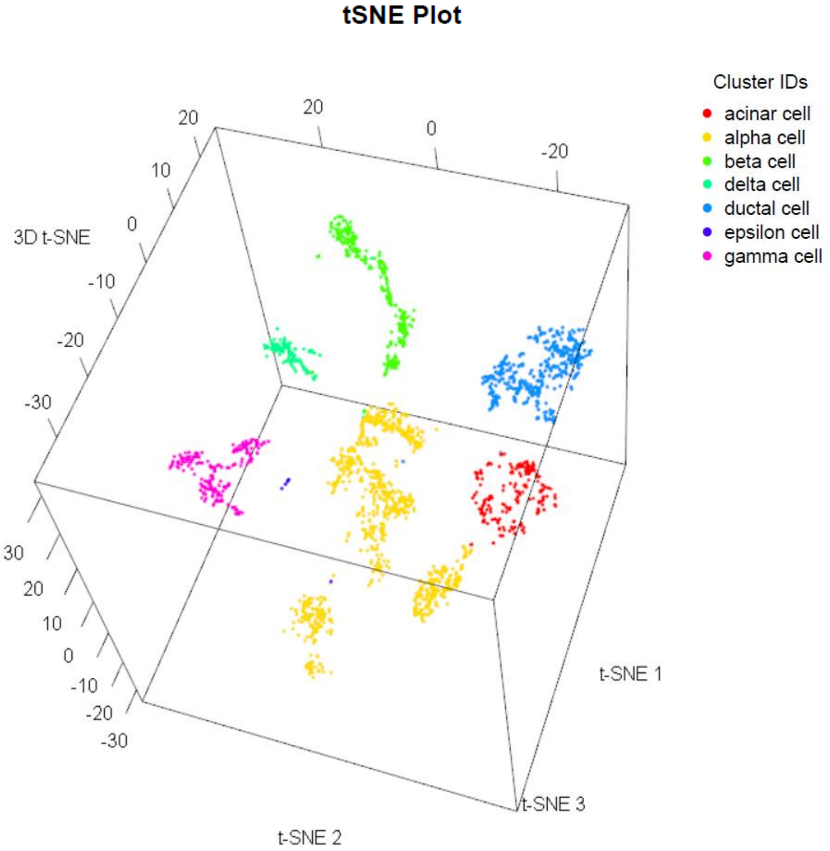


TF-IDF transformed data t-SNE plot

t-SNE TF-IDF transformation

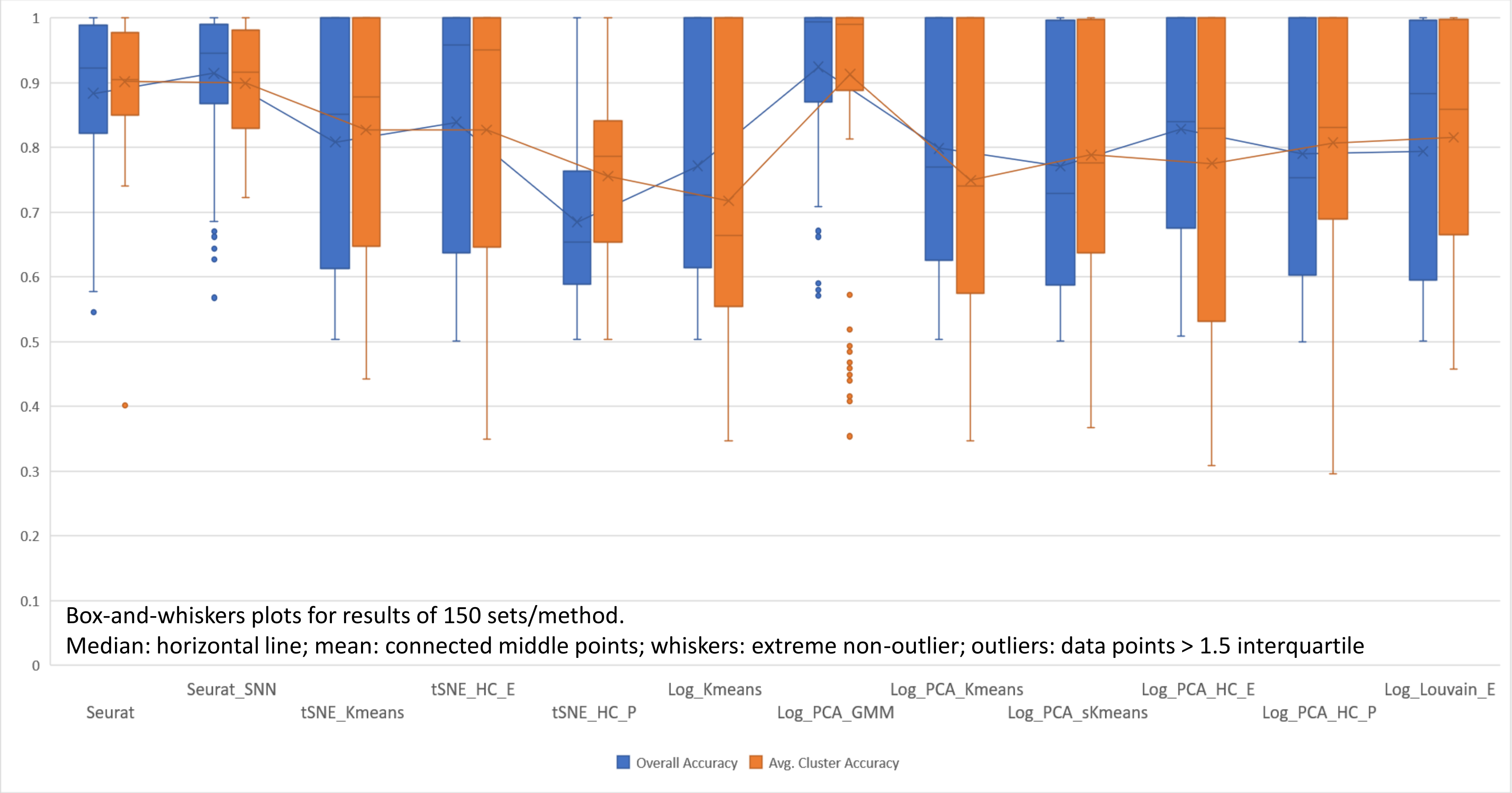


Raw Pancreas data t-SNE plot



TF-IDF transformed data t-SNE plot

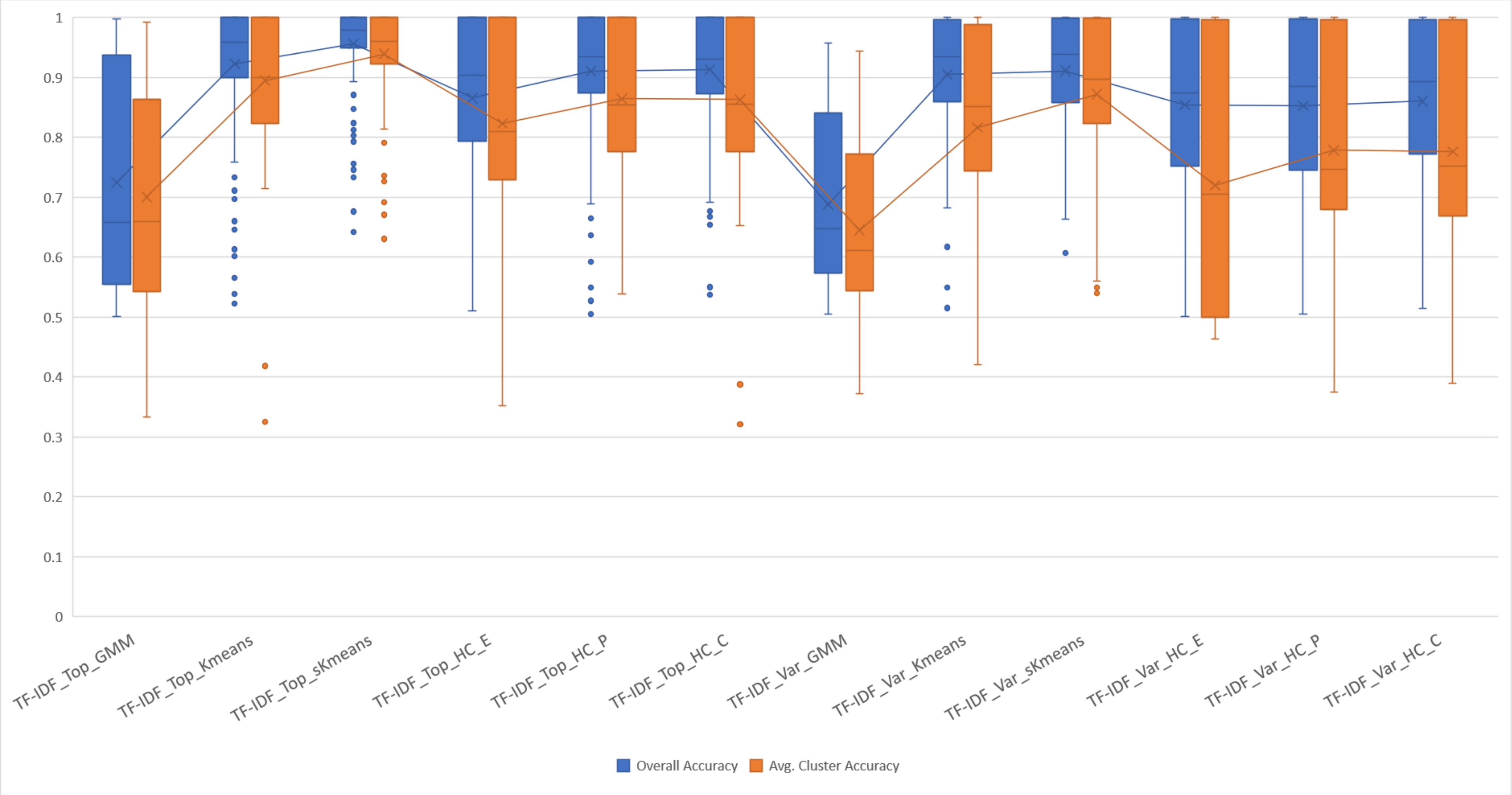
Pairs: Existing Methods



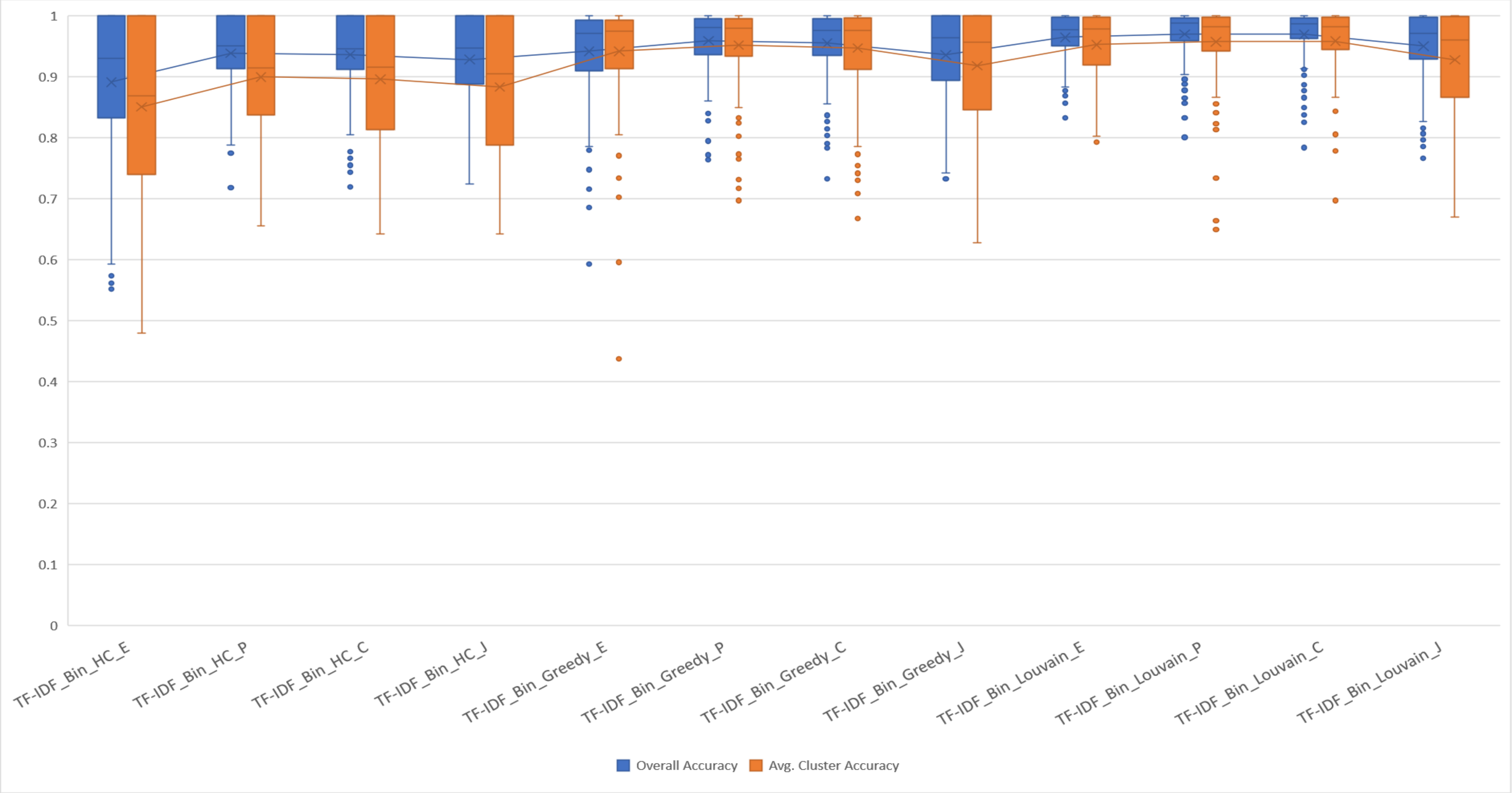
Box-and-whiskers plots for results of 150 sets/method.

Median: horizontal line; mean: connected middle points; whiskers: extreme non-outlier; outliers: data points > 1.5 interquartile

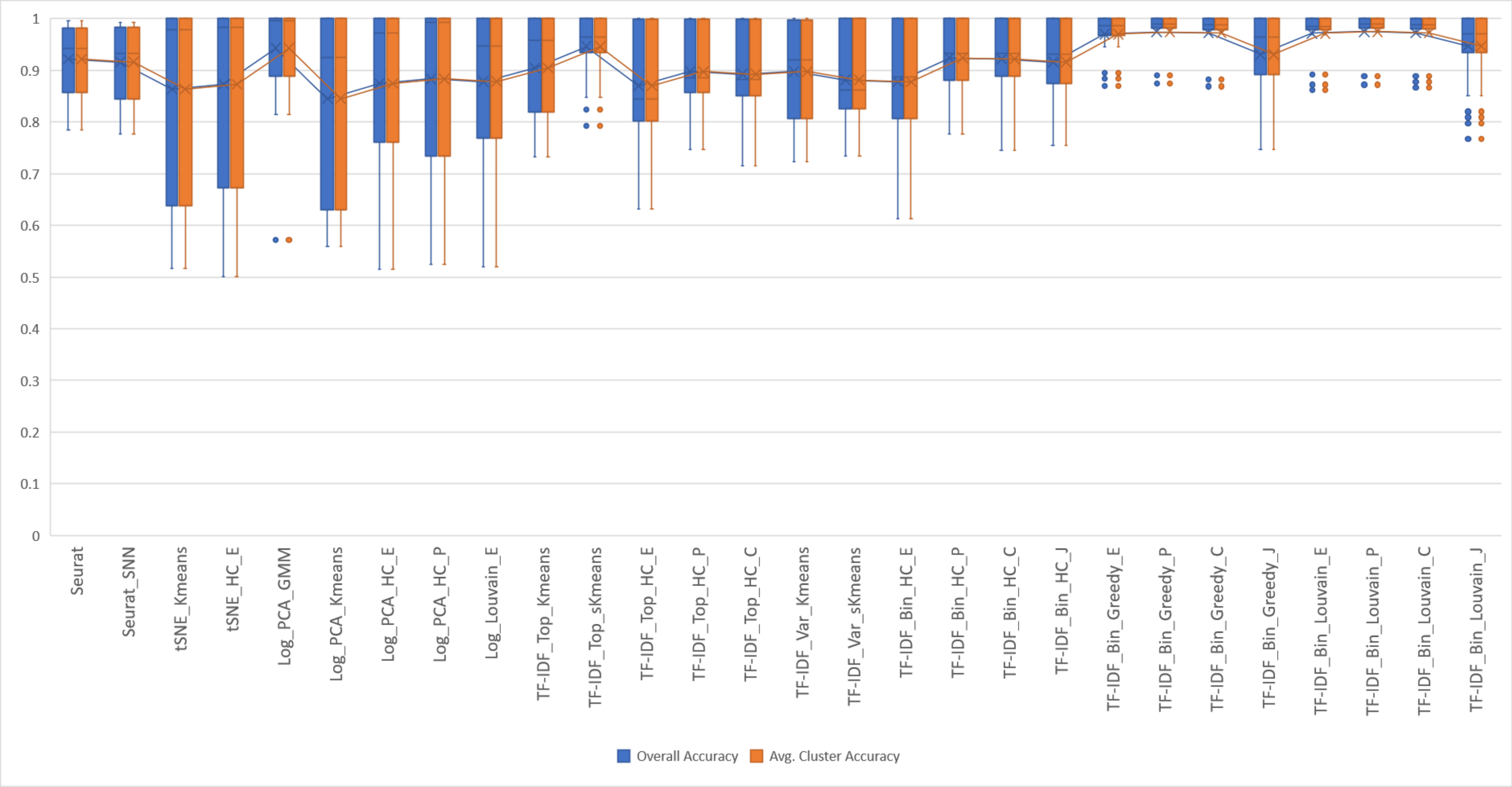
Pairs: Algorithms using TF-IDF gene selection



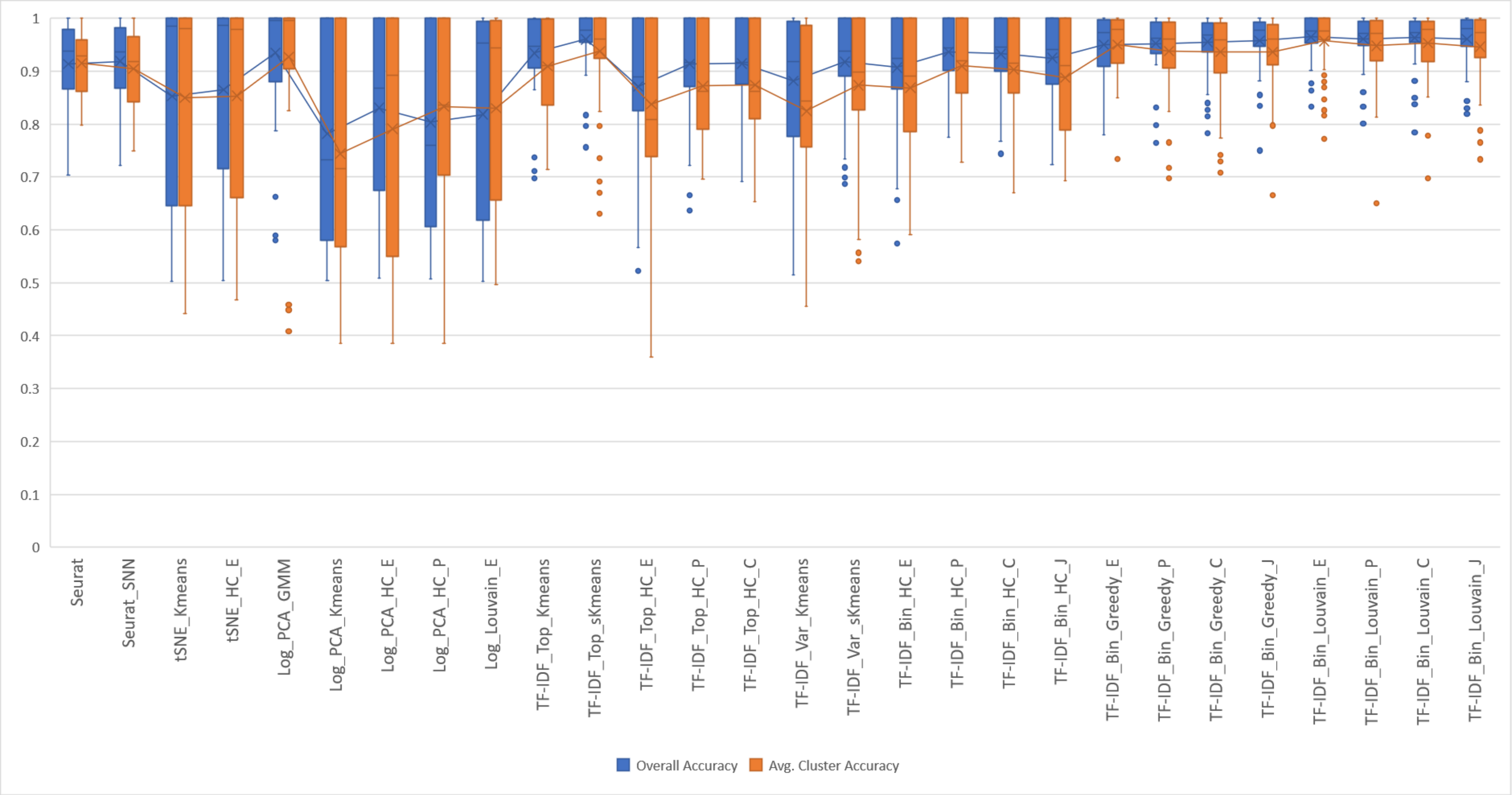
Pairs: Algorithms using TF-IDF binarization.



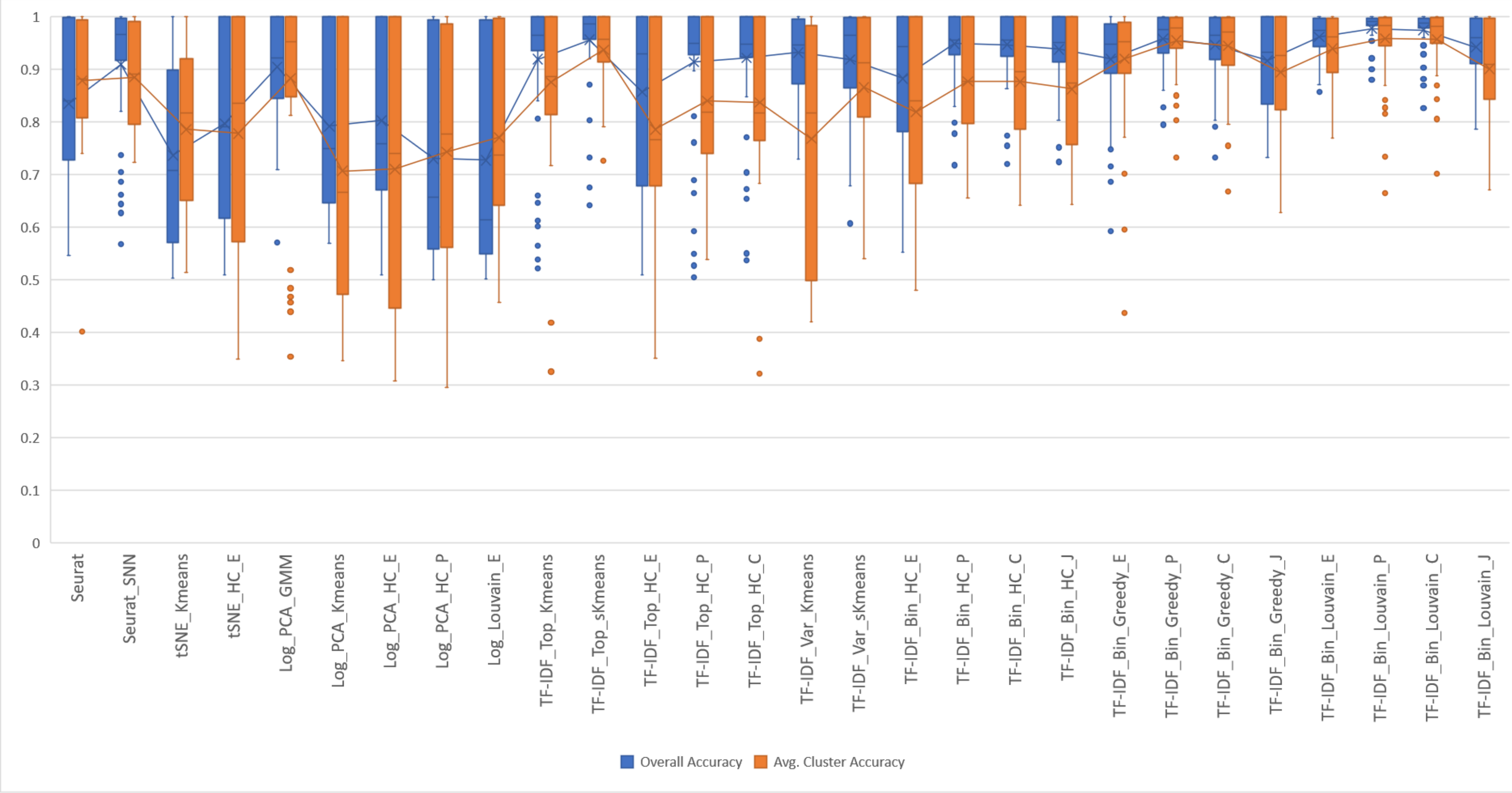
Pairs: 1:1 mixtures



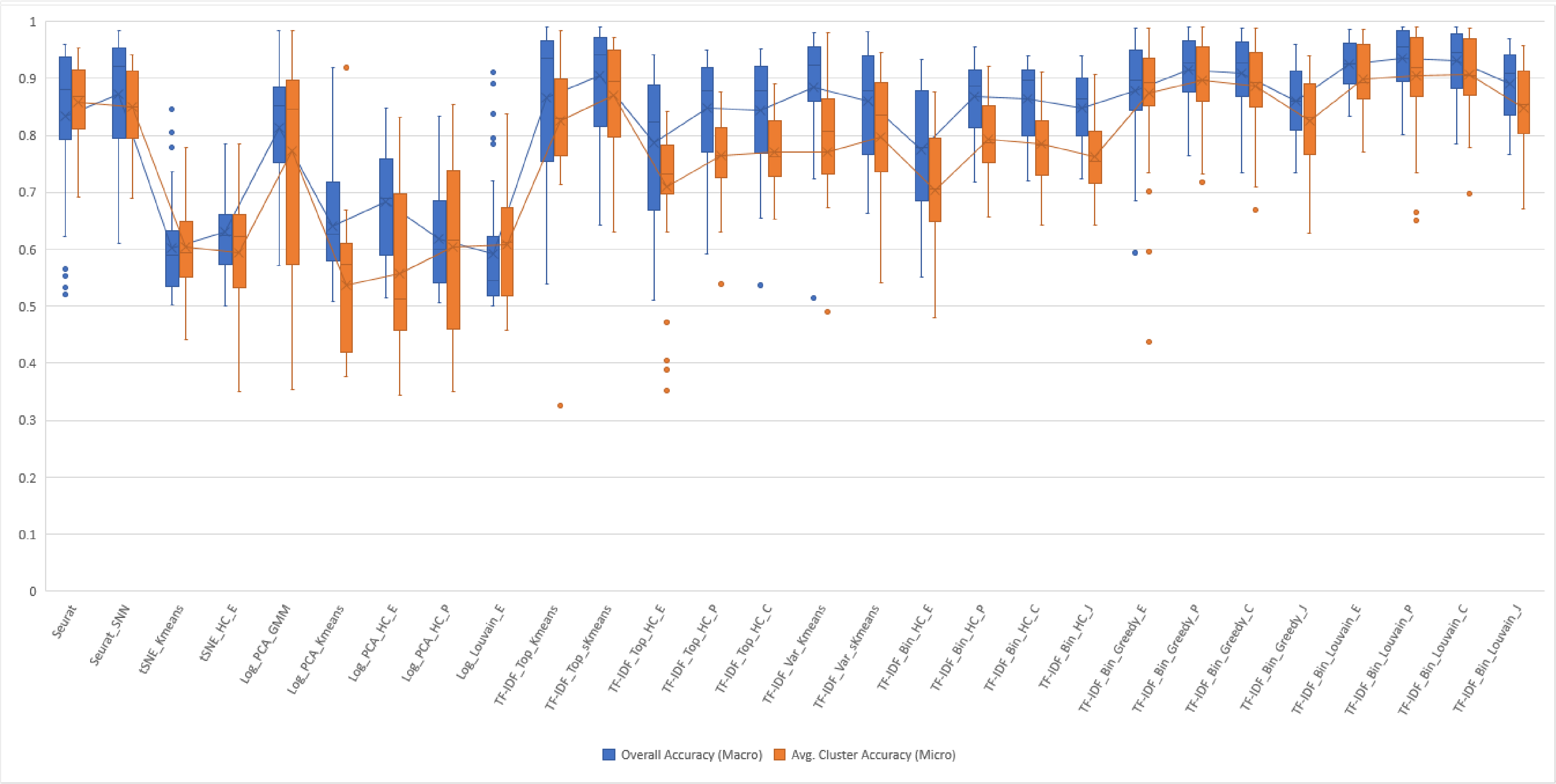
Pairs: 1:3/3:1 mixtures



Pairs: 1:7/7:1 mixtures

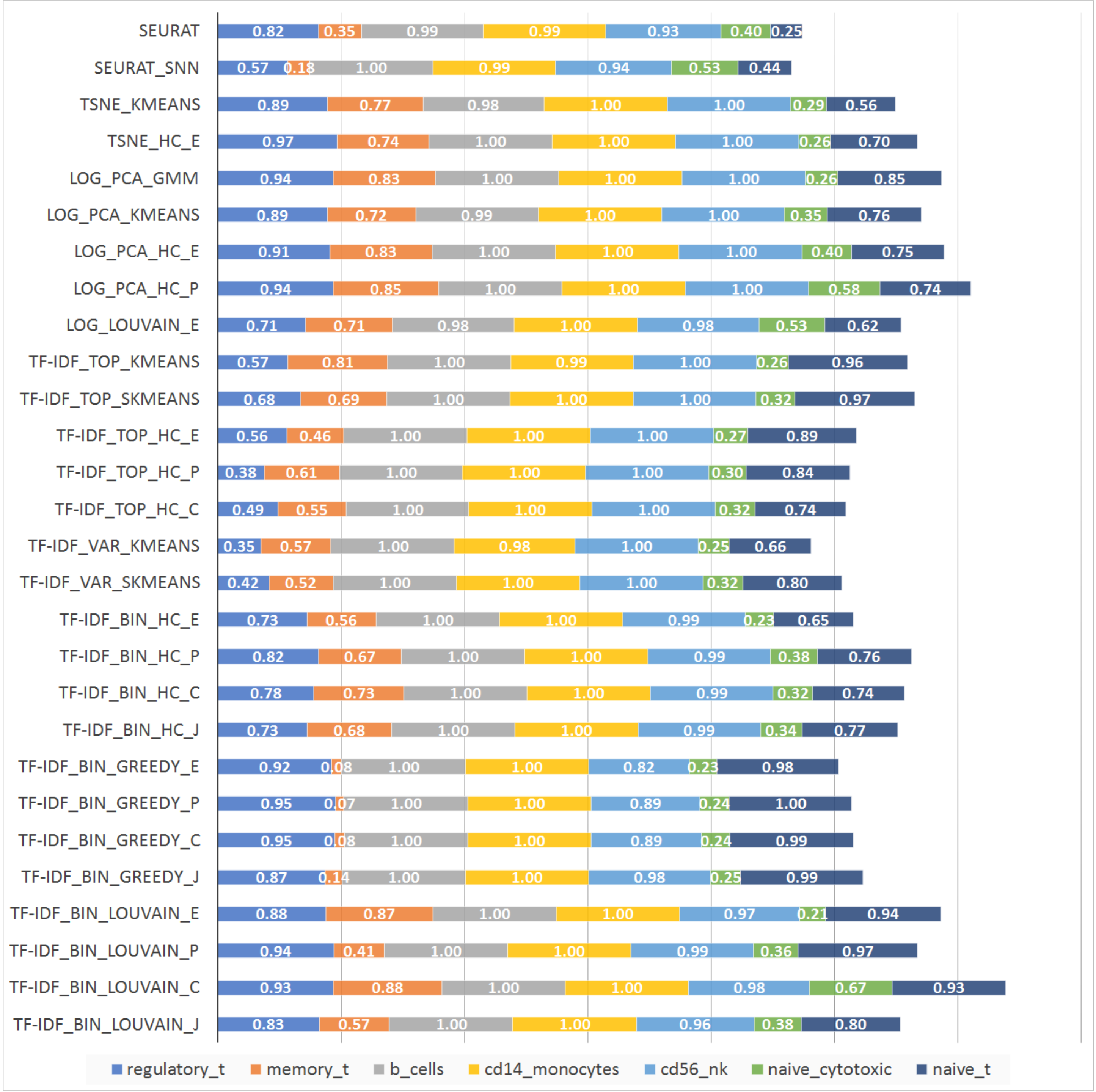


Pairs by ‘difficulty’

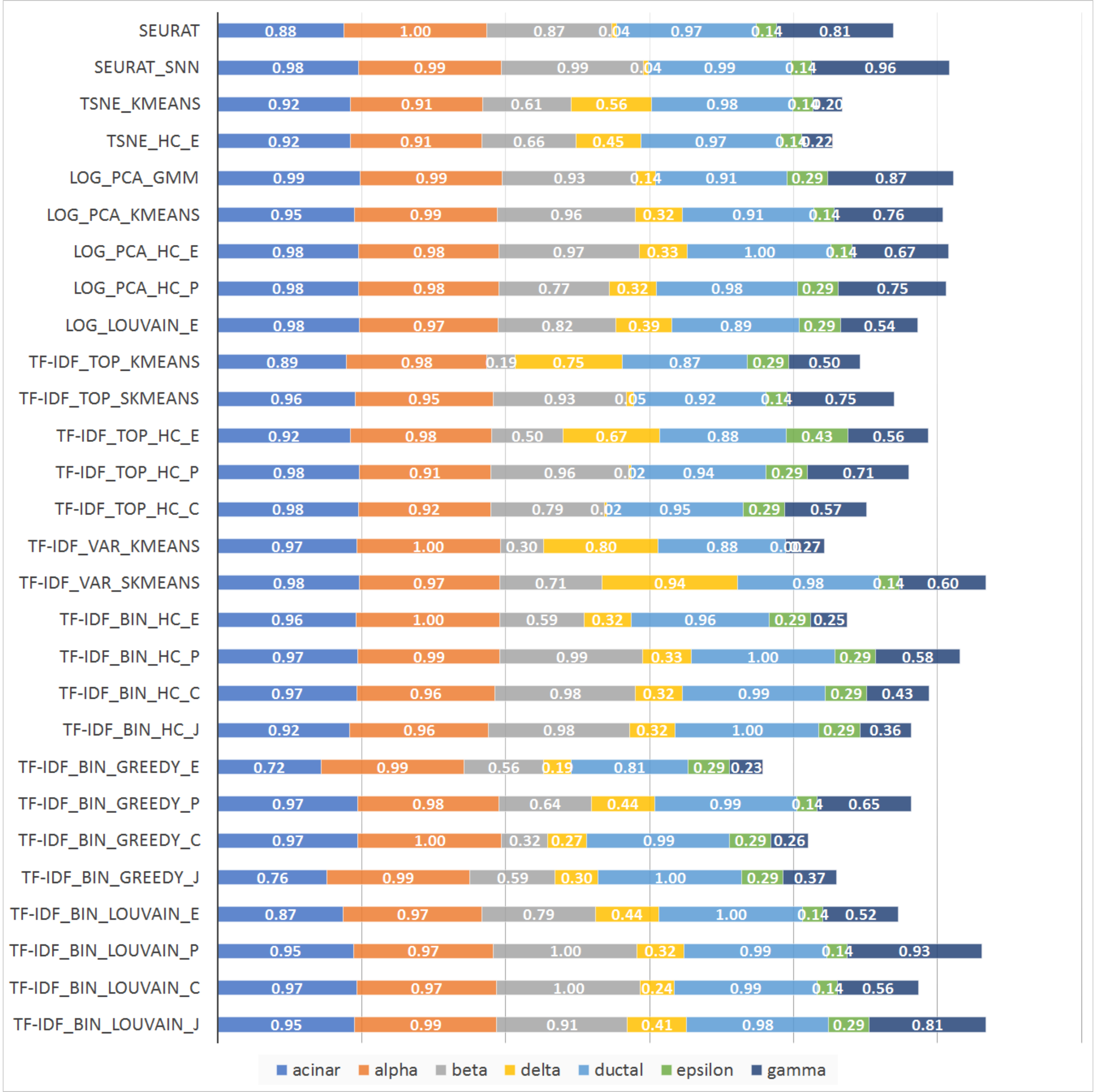


highly dissimilar: (b cells and cd14 monocytes) and (b cells and cd56 nk)
highly similar : (memory t and naive cytotoxic) and (regulatory t and naive t)
intermediate similarity: (memory t and naive t) and (regulatory t and naive cytotoxic)

Accuracy for PBMC Cells, 7-way mixture

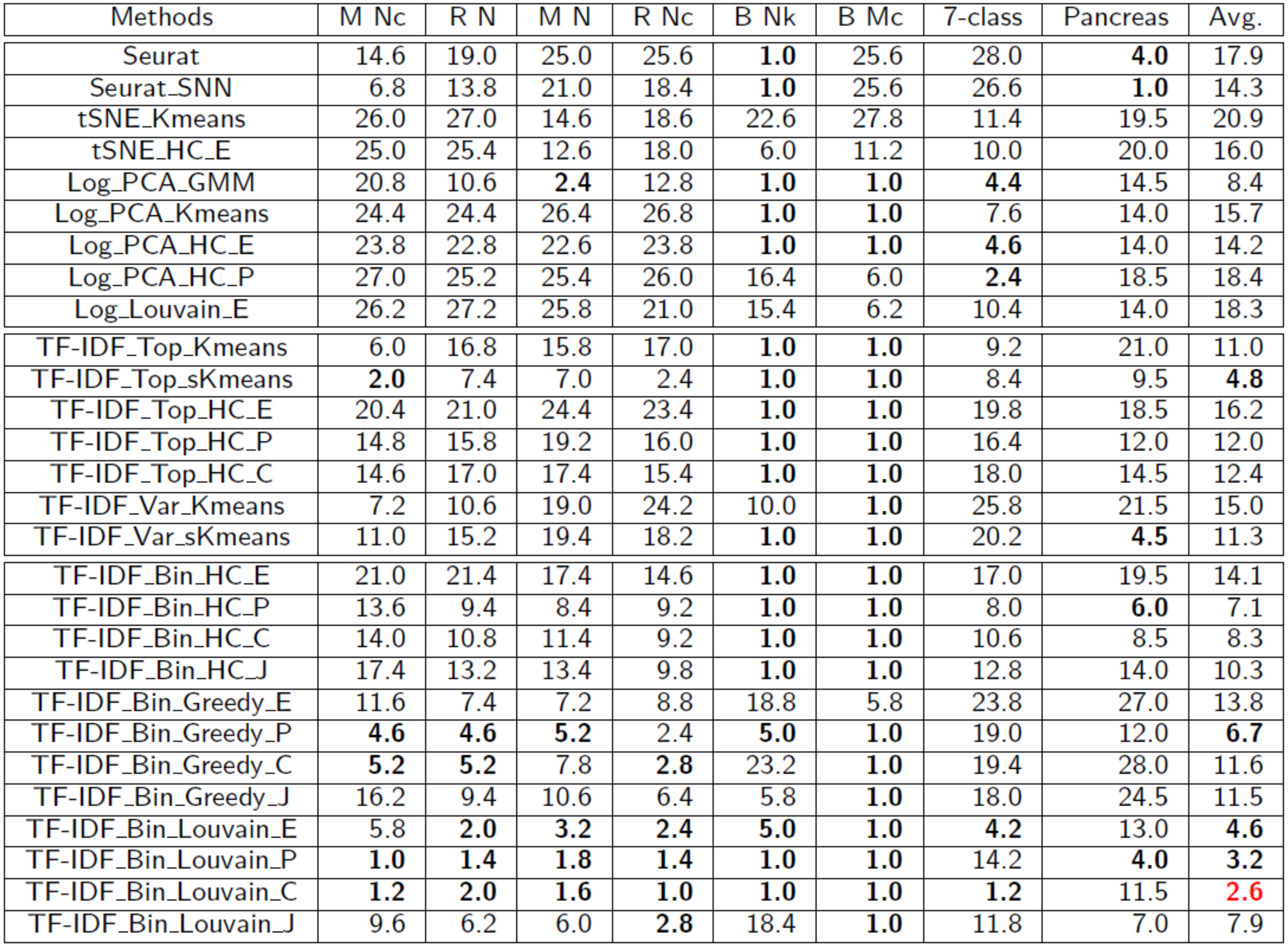


Accuracy for Pancreatic mixture



Average ranks based on overall accuracy.

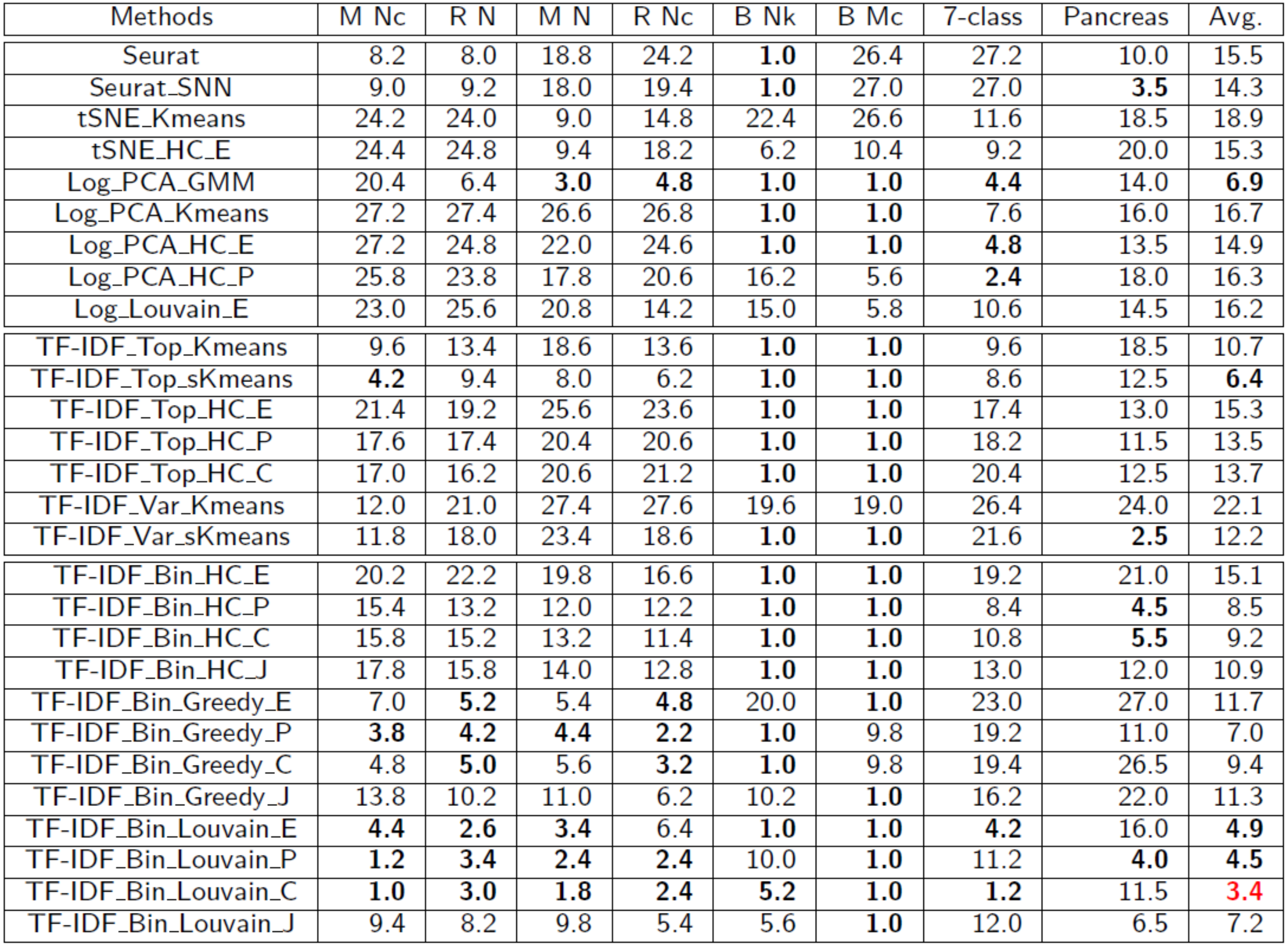
The lowest five average ranks (including ties) for each dataset are typeset in bold, and the best overall average rank is shown in red.



Methods	M Nc	R N	M N	R Nc	B Nk	B Mc	7-class	Pancreas	Avg.
Seurat	14.6	19.0	25.0	25.6	**1.0**	25.6	28.0	**4.0**	17.9
Seurat_SNN	6.8	13.8	21.0	18.4	**1.0**	25.6	26.6	**1.0**	14.3
tSNE_Kmeans	26.0	27.0	14.6	18.6	22.6	27.8	11.4	19.5	20.9
tSNE_HC_E	25.0	25.4	12.6	18.0	6.0	11.2	10.0	20.0	16.0
Log_PCA_GMM	20.8	10.6	**2.4**	12.8	**1.0**	**1.0**	**4.4**	14.5	8.4
Log_PCA_Kmeans	24.4	24.4	26.4	26.8	**1.0**	**1.0**	7.6	14.0	15.7
Log_PCA_HC_E	23.8	22.8	22.6	23.8	**1.0**	**1.0**	**4.6**	14.0	14.2
Log_PCA_HC_P	27.0	25.2	25.4	26.0	16.4	6.0	**2.4**	18.5	18.4
Log_Louvain_E	26.2	27.2	25.8	21.0	15.4	6.2	10.4	14.0	18.3
TF-IDF_Top_Kmeans	6.0	16.8	15.8	17.0	**1.0**	**1.0**	9.2	21.0	11.0
TF-IDF_Top_sKmeans	**2.0**	7.4	7.0	2.4	**1.0**	**1.0**	8.4	9.5	**4.8**
TF-IDF_Top_HC_E	20.4	21.0	24.4	23.4	**1.0**	**1.0**	19.8	18.5	16.2
TF-IDF_Top_HC_P	14.8	15.8	19.2	16.0	**1.0**	**1.0**	16.4	12.0	12.0
TF-IDF_Top_HC_C	14.6	17.0	17.4	15.4	**1.0**	**1.0**	18.0	14.5	12.4
TF-IDF_Var_Kmeans	7.2	10.6	19.0	24.2	10.0	**1.0**	25.8	21.5	15.0
TF-IDF_Var_sKmeans	11.0	15.2	19.4	18.2	**1.0**	**1.0**	20.2	**4.5**	11.3
TF-IDF_Bin_HC_E	21.0	21.4	17.4	14.6	**1.0**	**1.0**	17.0	19.5	14.1
TF-IDF_Bin_HC_P	13.6	9.4	8.4	9.2	**1.0**	**1.0**	8.0	**6.0**	7.1
TF-IDF_Bin_HC_C	14.0	10.8	11.4	9.2	**1.0**	**1.0**	10.6	8.5	8.3
TF-IDF_Bin_HC_J	17.4	13.2	13.4	9.8	**1.0**	**1.0**	12.8	14.0	10.3
TF-IDF_Bin_Greedy_E	11.6	7.4	7.2	8.8	18.8	5.8	23.8	27.0	13.8
TF-IDF_Bin_Greedy_P	**4.6**	**4.6**	**5.2**	2.4	**5.0**	**1.0**	19.0	12.0	**6.7**
TF-IDF_Bin_Greedy_C	**5.2**	**5.2**	7.8	**2.8**	23.2	**1.0**	19.4	28.0	11.6
TF-IDF_Bin_Greedy_J	16.2	9.4	10.6	6.4	5.8	**1.0**	18.0	24.5	11.5
TF-IDF_Bin_Louvain_E	5.8	**2.0**	**3.2**	**2.4**	**5.0**	**1.0**	**4.2**	13.0	**4.6**
TF-IDF_Bin_Louvain_P	**1.0**	**1.4**	**1.8**	**1.4**	**1.0**	**1.0**	14.2	**4.0**	**3.2**
TF-IDF_Bin_Louvain_C	**1.2**	**2.0**	**1.6**	**1.0**	**1.0**	**1.0**	**1.2**	11.5	**2.6**
TF-IDF_Bin_Louvain_J	9.6	6.2	6.0	**2.8**	18.4	**1.0**	11.8	7.0	7.9

Average ranks based on average cluster accuracy.

The lowest five average ranks (including ties) for each dataset are typeset in bold, and the best overall average rank is shown in red.



Methods	M Nc	R N	M N	R Nc	B Nk	B Mc	7-class	Pancreas	Avg.
Seurat	8.2	8.0	18.8	24.2	**1.0**	26.4	27.2	10.0	15.5
Seurat_SNN	9.0	9.2	18.0	19.4	**1.0**	27.0	27.0	**3.5**	14.3
tSNE_Kmeans	24.2	24.0	9.0	14.8	22.4	26.6	11.6	18.5	18.9
tSNE_HC_E	24.4	24.8	9.4	18.2	6.2	10.4	9.2	20.0	15.3
Log_PCA_GMM	20.4	6.4	**3.0**	**4.8**	**1.0**	**1.0**	**4.4**	14.0	**6.9**
Log_PCA_Kmeans	27.2	27.4	26.6	26.8	**1.0**	**1.0**	7.6	16.0	16.7
Log_PCA_HC_E	27.2	24.8	22.0	24.6	**1.0**	**1.0**	**4.8**	13.5	14.9
Log_PCA_HC_P	25.8	23.8	17.8	20.6	16.2	5.6	**2.4**	18.0	16.3
Log_Louvain_E	23.0	25.6	20.8	14.2	15.0	5.8	10.6	14.5	16.2
TF-IDF_Top_Kmeans	9.6	13.4	18.6	13.6	**1.0**	**1.0**	9.6	18.5	10.7
TF-IDF_Top_sKmeans	**4.2**	9.4	8.0	6.2	**1.0**	**1.0**	8.6	12.5	**6.4**
TF-IDF_Top_HC_E	21.4	19.2	25.6	23.6	**1.0**	**1.0**	17.4	13.0	15.3
TF-IDF_Top_HC_P	17.6	17.4	20.4	20.6	**1.0**	**1.0**	18.2	11.5	13.5
TF-IDF_Top_HC_C	17.0	16.2	20.6	21.2	**1.0**	**1.0**	20.4	12.5	13.7
TF-IDF_Var_Kmeans	12.0	21.0	27.4	27.6	19.6	19.0	26.4	24.0	22.1
TF-IDF_Var_sKmeans	11.8	18.0	23.4	18.6	**1.0**	**1.0**	21.6	**2.5**	12.2
TF-IDF_Bin_HC_E	20.2	22.2	19.8	16.6	**1.0**	**1.0**	19.2	21.0	15.1
TF-IDF_Bin_HC_P	15.4	13.2	12.0	12.2	**1.0**	**1.0**	8.4	**4.5**	8.5
TF-IDF_Bin_HC_C	15.8	15.2	13.2	11.4	**1.0**	**1.0**	10.8	**5.5**	9.2
TF-IDF_Bin_HC_J	17.8	15.8	14.0	12.8	**1.0**	**1.0**	13.0	12.0	10.9
TF-IDF_Bin_Greedy_E	7.0	**5.2**	5.4	**4.8**	20.0	**1.0**	23.0	27.0	11.7
TF-IDF_Bin_Greedy_P	**3.8**	**4.2**	**4.4**	**2.2**	**1.0**	9.8	19.2	11.0	7.0
TF-IDF_Bin_Greedy_C	4.8	**5.0**	5.6	**3.2**	**1.0**	9.8	19.4	26.5	9.4
TF-IDF_Bin_Greedy_J	13.8	10.2	11.0	6.2	10.2	**1.0**	16.2	22.0	11.3
TF-IDF_Bin_Louvain_E	**4.4**	**2.6**	**3.4**	6.4	**1.0**	**1.0**	**4.2**	16.0	**4.9**
TF-IDF_Bin_Louvain_P	**1.2**	**3.4**	**2.4**	**2.4**	10.0	**1.0**	11.2	**4.0**	**4.5**
TF-IDF_Bin_Louvain_C	**1.0**	**3.0**	**1.8**	**2.4**	**5.2**	**1.0**	**1.2**	11.5	**3.4**
TF-IDF_Bin_Louvain_J	9.4	8.2	9.8	5.4	5.6	**1.0**	12.0	6.5	7.2

Outline

- Motivation and challenges for scRNA-Seq data analysis
- Background: TF-IDF transformation
- Methods: Existing and TF-IDF based methods
- Experimental setup
- Results and Discussion
- Conclusions

Conclusion & Ongoing Work

- The range of single-cell applications continues to expand, fueled by advances in technology
- New algorithms for scRNA-Seq clustering still needed
  - Preliminary results using TF-IDF transformation promising
  - Scalable to millions of cells in conjunction with graph-based clustering
- Ongoing work
  - Modified TF-IDF definition
  - Study effect of cell cycle analysis/removal on clustering
  - Imputation effect on dropout events and clustering accuracy.
  - Clustering based on chromosomal copy number variations (CNVs) as first tier for tumor/normal data.

>

Modified TF-IDF Transformation

- Term Frequency x Inverse Document Frequency for scRNA-Seq data:

$$f' = \log(f + 1)$$

  - For gene i in cell j with count f:

$$TF_{ij} = f'_{ij} / \max_k f'_{kj}$$

  - If gene i is detected with $f_i \geq \mathrm{t}$ in n_i out of N cells:

$$IDF_i = \log_2(N / n_i)$$

  Possible choice for $t = mean\ TF$

  - TF-IDF score:

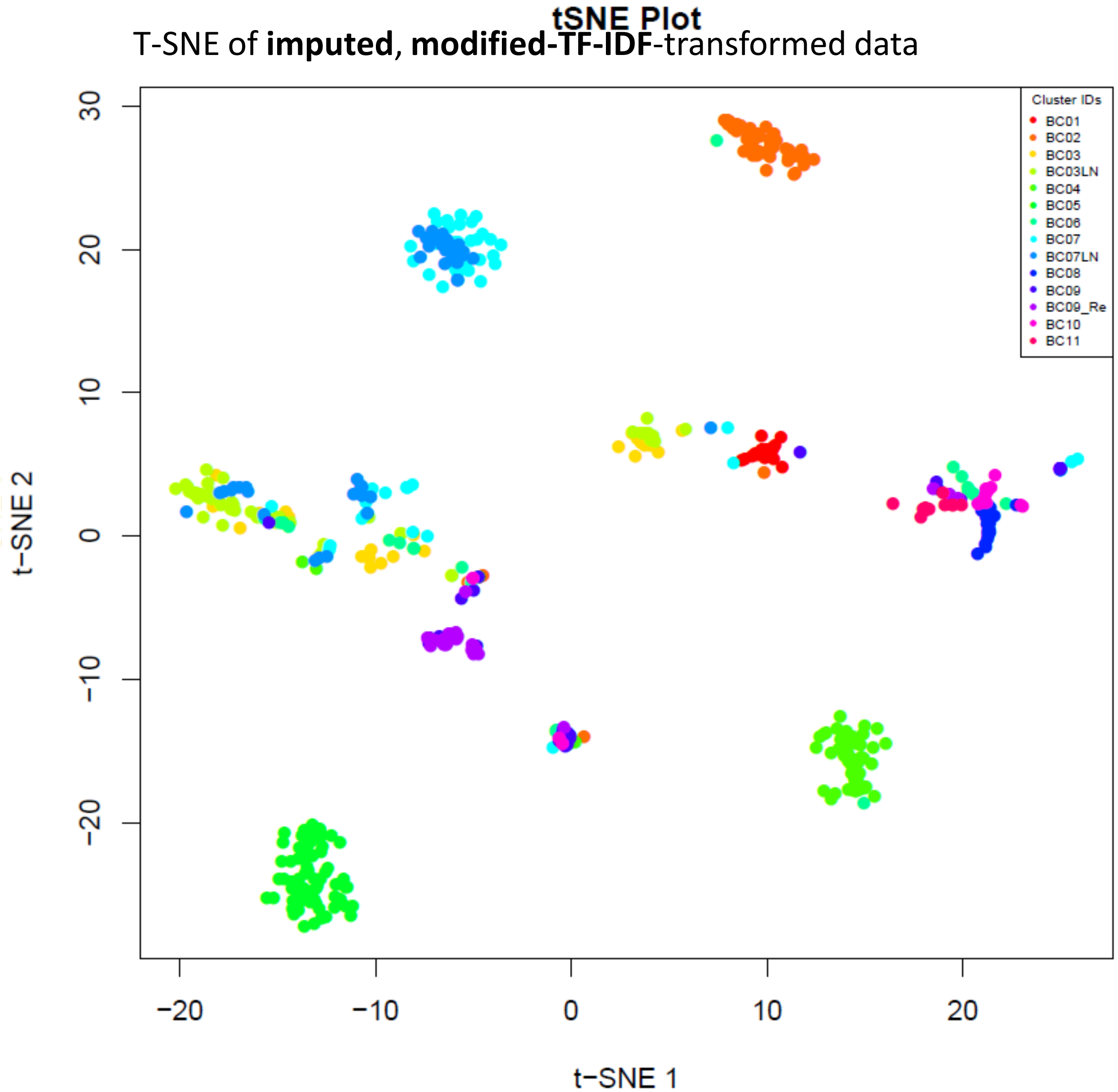
$$TF_{ij} \ * IDF_i$$

Breast Cancer data [Chung et al., 2017]

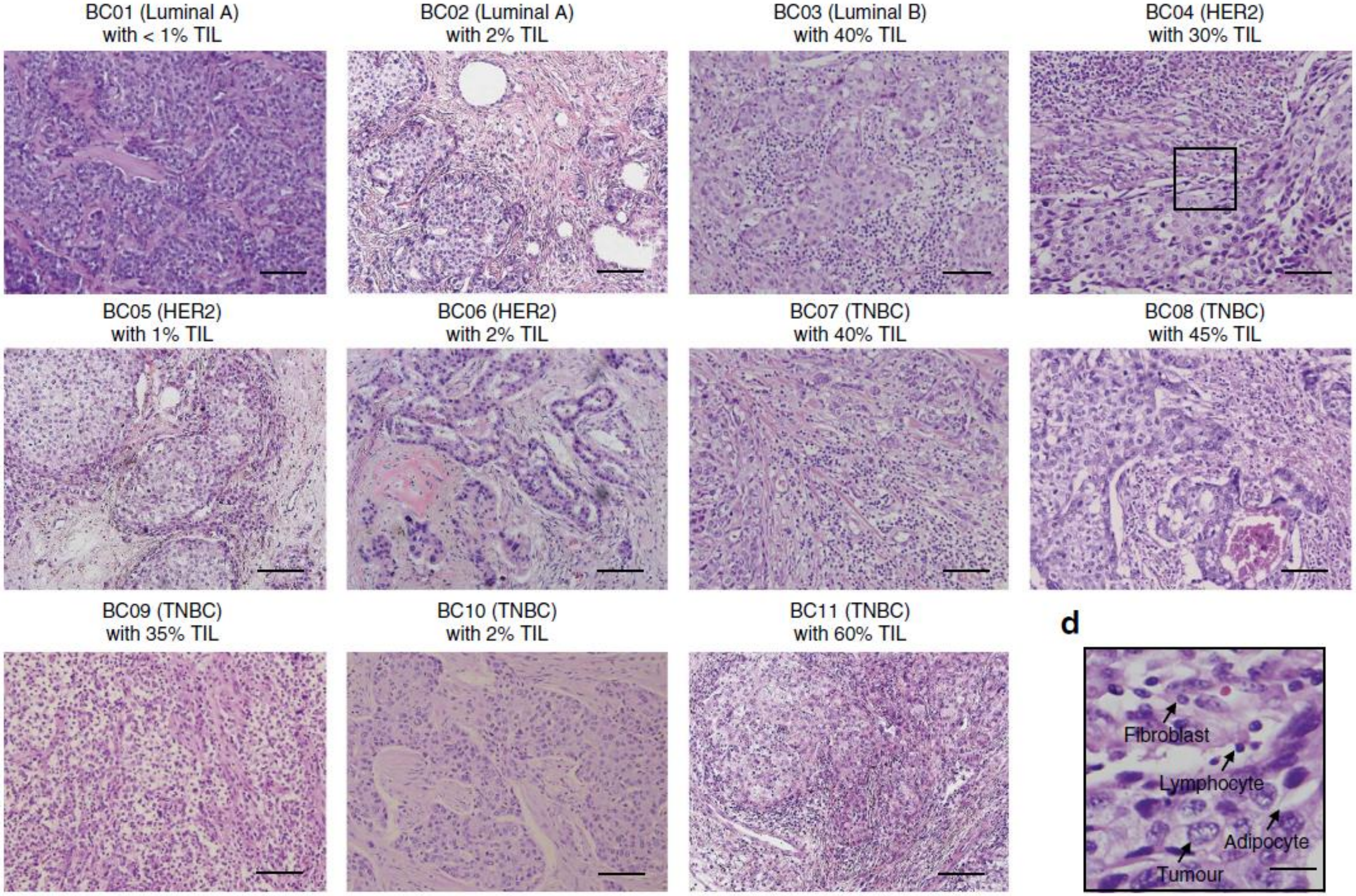
11 patients representing the four subtypes of BC: luminal A; luminal B; HER2; and triple negative breast cancer (TNBC).

Markers:

- ER-positive (BC01 and BC02; luminal A),
- ER/HER2-positive (BC03; luminal B),
- HER2-positive (BC04, BC05 and BC06; HER2)
- and triple negative (BC07–BC11; TNBC) invasive ductal carcinoma.
- Regional metastatic lymph nodes were collected from the luminal B (BC03LN) sample
- and a triple negative breast cancer (BC07LN) sample.



T-SNE of **imputed**, **modified-TF-IDF**-transformed data



BC01 (Luminal A) with < 1% TIL

BC02 (Luminal A) with 2% TIL

BC03 (Luminal B) with 40% TIL

BC04 (HER2) with 30% TIL

BC05 (HER2) with 1% TIL

BC06 (HER2) with 2% TIL

BC07 (TNBC) with 40% TIL

BC08 (TNBC) with 45% TIL

BC09 (TNBC) with 35% TIL

BC10 (TNBC) with 2% TIL

BC11 (TNBC) with 60% TIL

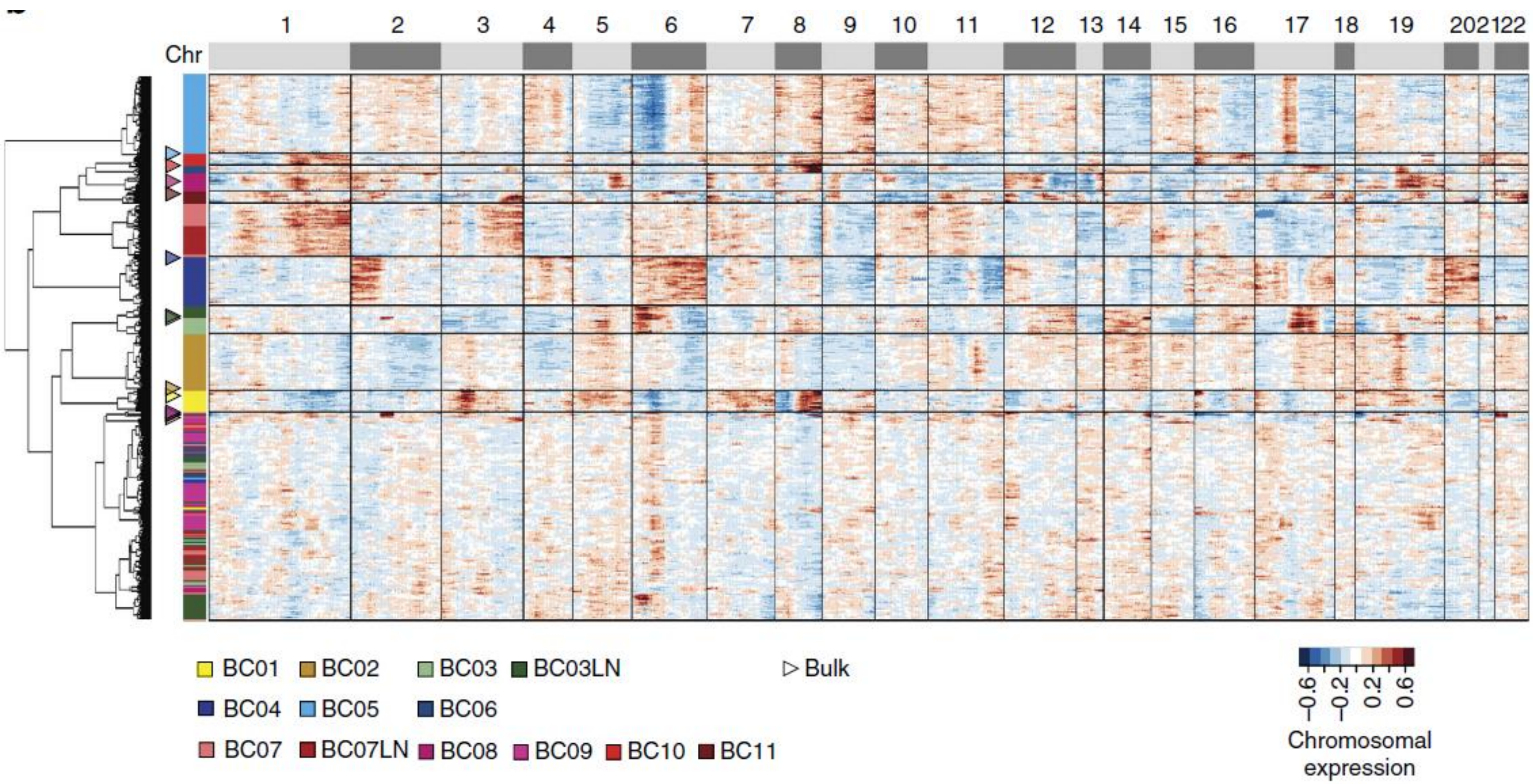
Microscopic findings indicated carcinoma and non-carcinoma cells, including tumor-infiltrating lymphocytes9 (TIL, 1–60%). Most of the TNBC tumors except BC10 were heavily infiltrated with lymphocytes, whereas luminal A tumors showed enrichment with carcinoma cells.

Chromosomal copy number variations based clustering

- sorted genes by their genomic locations (chromosome number, then gene start position)
- moving average of 100 analyzed genes
- estimate of chromosomal CNVs in each cell and at each analyzed gene:

$$CNV_k(i) = \frac{\sum_{j=i-50}^{i+50} E_k\left(o_j\right)}{101}$$

  - $CNV(i)$ is the estimated relative copy number of cell *k* at the *i*’th gene in the genomically-ordered list of genes,
  - o_j is the *j*’th gene in the genomically-ordered list of genes,
  - and $E_k\left(o_j\right)$is the relative normalized expression of that gene in cell *k*

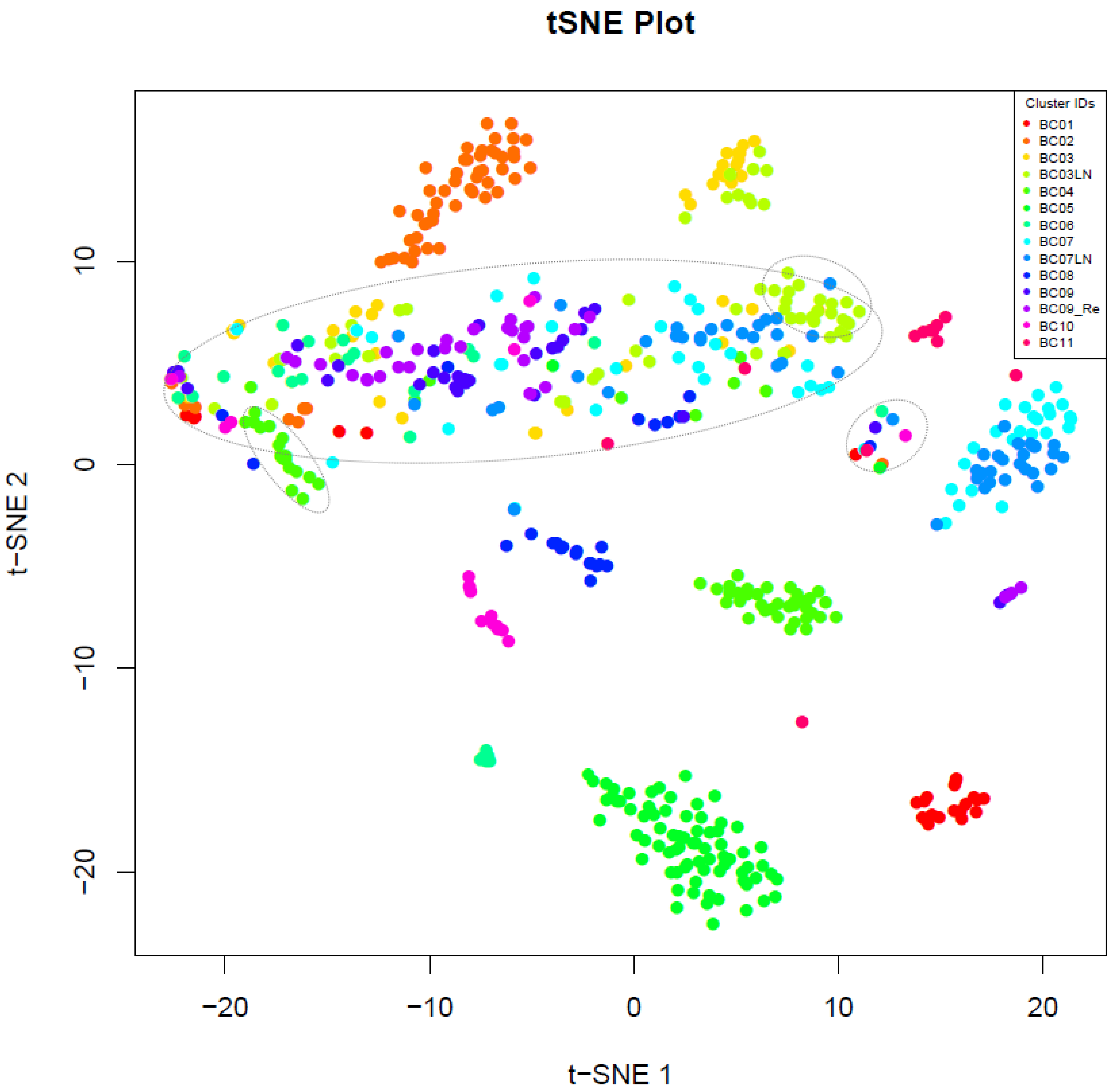


Hierarchical clustering of the chromosomal gene expression pattern separating the patient-specific carcinoma cell groups from the non-carcinoma cell cluster. For each chromosome, the chromosomal gene expression pattern was estimated from the moving average of 150 genes. These patterns implicate chromosomal amplification and deletion.[Chung, 2017]

t-SNE of CNV matrix

- ER-positive (BC01 and BC02; luminal A) *< 2% TIL*
- ER/HER2-positive (BC03; luminal B) *~ 30% TIL*
- HER2-positive (BC04 *~ 30% TIL*, BC05 and BC06 *~ 2% TIL*; HER2)
- and triple negative (BC07 *~40% –* BC11 *~ 70% TIL*; TNBC) invasive ductal carcinoma.

≤



References

- Tang, F., et al. "mRNA-Seq whole-transcriptome analysis of a single cell." *Nature methods* 6.5 (2009): 377-382.
- Macosko, E. Z., et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." Cell 161.5 (2015): 1202-1214.
- Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature biotechnology 33(5), 495{502 (2015)
- Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells.
- Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(2), 411{423 (2001).
- Moussa, M., and Mandoiu, I.: Clustering scRNA-Seq Data using TF-IDF. Bioinformatics Research and Applications. 13th International Symposium, ISBRA 2017, Honolulu, HI, USA. Lecture Notes in Computer Science book series (LNCS, volume 10330).
- Moussa, M., and Mandoiu, I.: Single Cell RNA-seq Data Clustering using TF-IDF based Methods. (BMC Special Issues, 2017).
- Chung, W., et al. "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer." Nature Communications 8 (2017).

Thank You.

Questions?