Ecole thématique sincellTE 2019

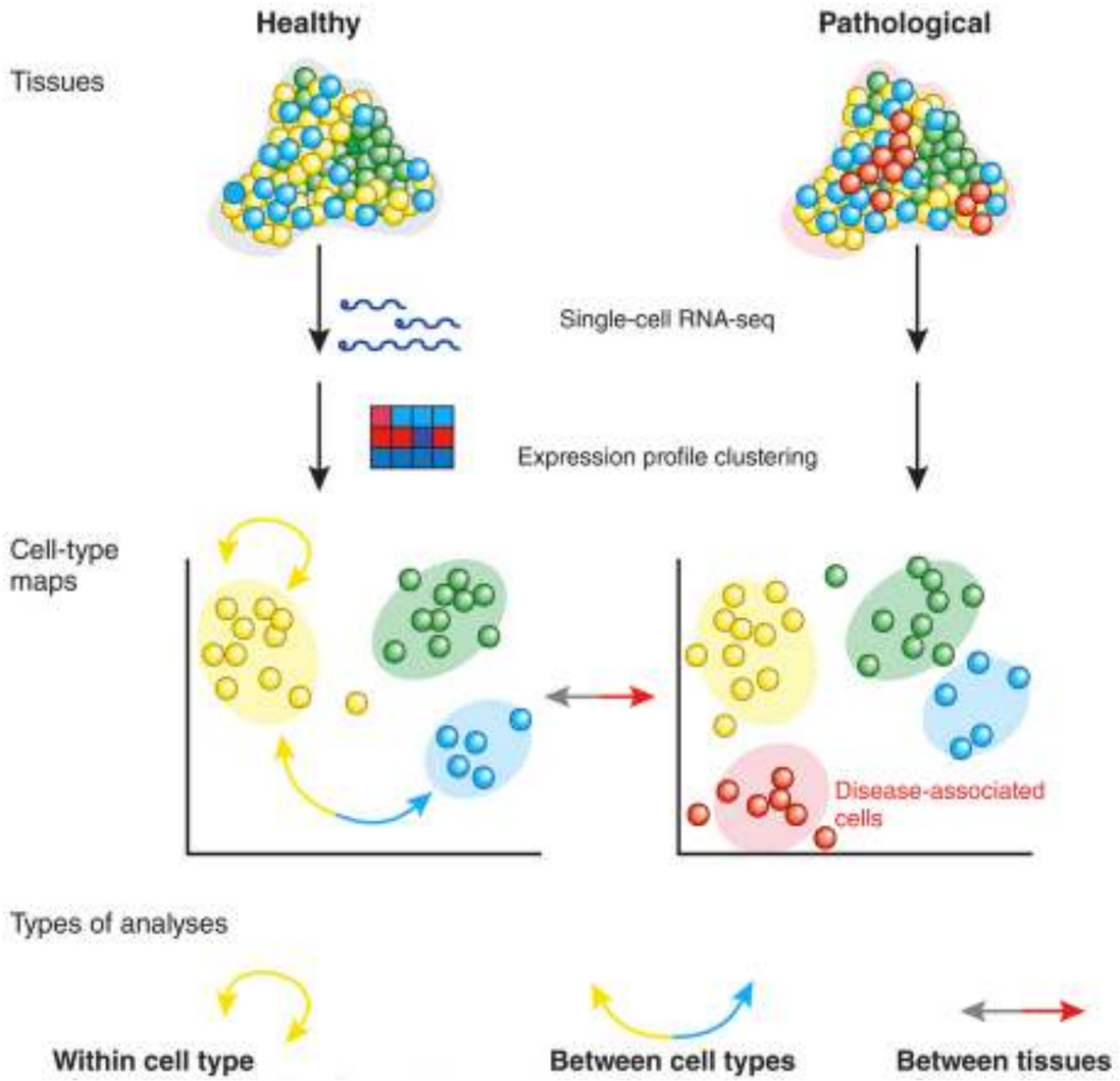
Statistical models and analysis

Antonio Rausell, Ph.D.

Roscoff, February 5th 2019

imagine
INSTITUT DES MALADIES GÉNÉTIQUES

The context: Single-cell RNA-seq to uncover cell heterogeneity associated to distinct cellular phenotypes



Sandberg R. Nature Methods 11, 22–24 (2014) doi:10.1038/nmeth.2764

Heterogeneity between samples arising from:

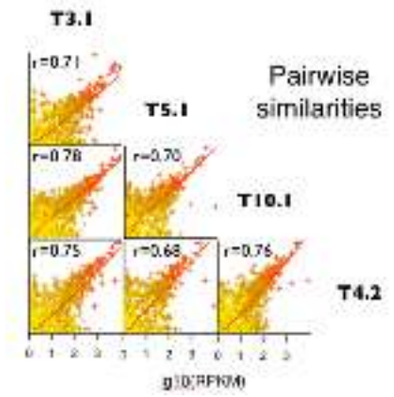
- Genetic factors (Donor)
- Environmental factors
- Treatments / Times of activation
- History of cells (e.g. clonal selection/expansion)
- Natural aging

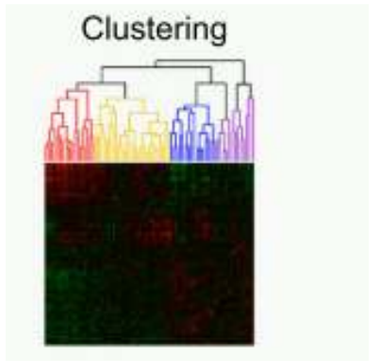
Heterogeneity within samples arising from:

- **Cell fate (permanent):**
  - Different lineages of differentiation
  - Different compositions of cell types
- **Cell state (transient):**
  - Stochasticity of gene expression
  - Pulsation / Circadian-like
  - Associated to cell cycle
  - **Different stages of activation**
- **Technical noise**

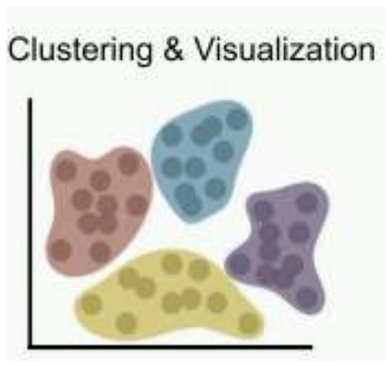
The questions

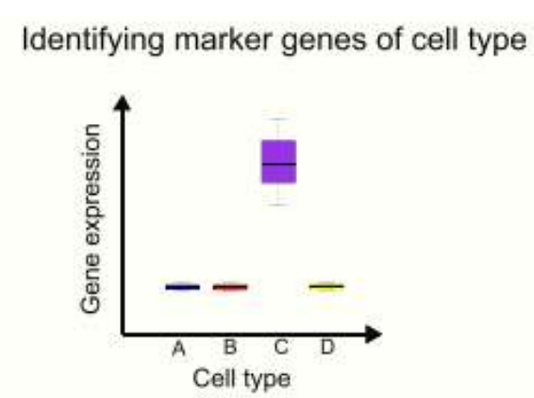
1- Is there **functionally relevant** cell heterogeneity in my data?



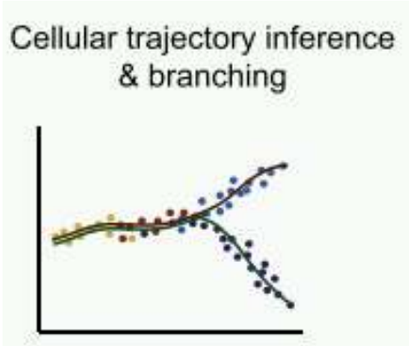


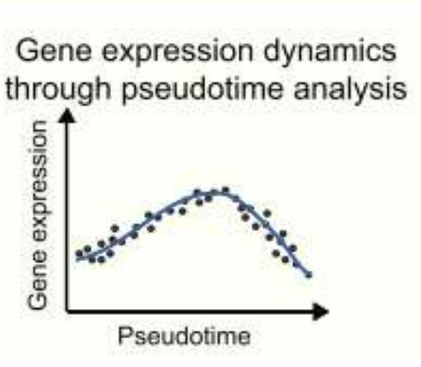
2- Are there distinct subpopulations of cells?





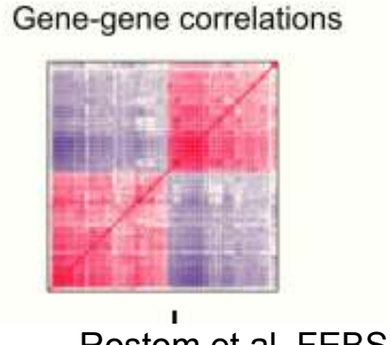
3- Are there continuums of differentiation / activation cell states?

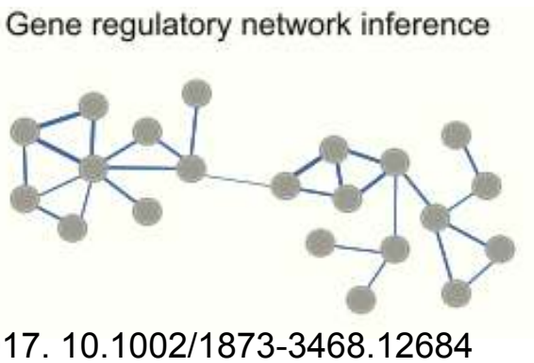




4- Which are the genes driving such heterogeneity?

5- May we learn something about the cellular / molecular mechanisms involved?: e.g. cell differentiation, biological process, pathways, regulatory modules, etc?





Rostom et al, FEBS 2017. 10.1002/1873-3468.12684

The bioinformatics / biostatistics mindset evolution

what to do

learn **how**

understand **why**

critical thinking on it

creative thinking: alternatives?

from following the crowd

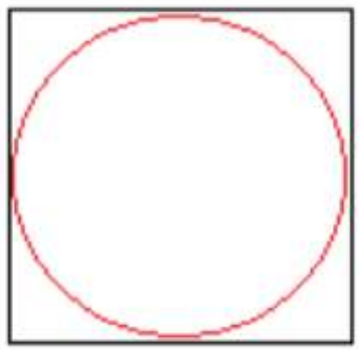
to proposing solutions

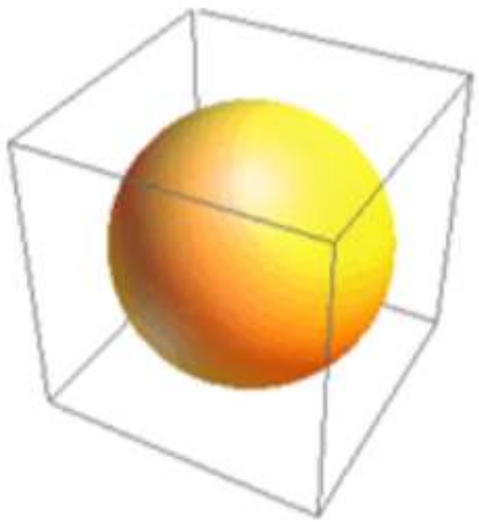
The computational challenges

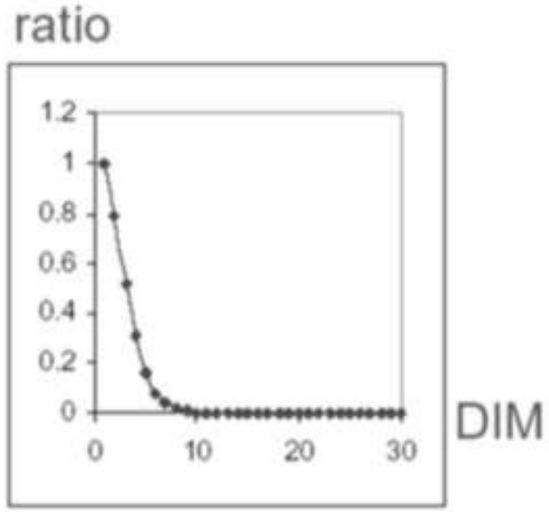
1. Lower coverage/depth than bulk RNA-seq
2. Technical & biological noise
3. High dimensionality
4. High variability
5. Dropouts => Zero-inflated data
6. Multimodality

The challenges - High dimensionality

The curse of dimensionality: When dimensionality increases, data becomes increasingly sparse in the space that it occupies



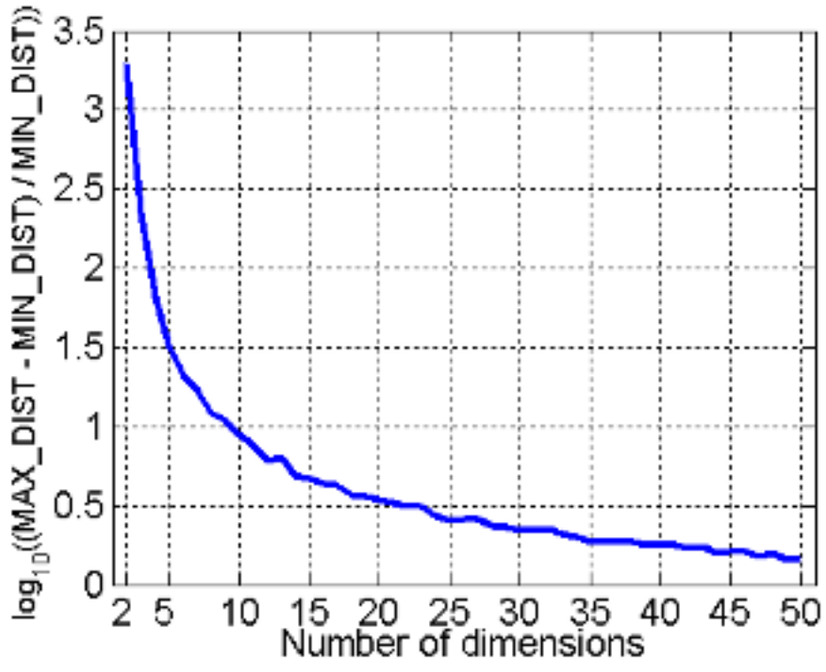




Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

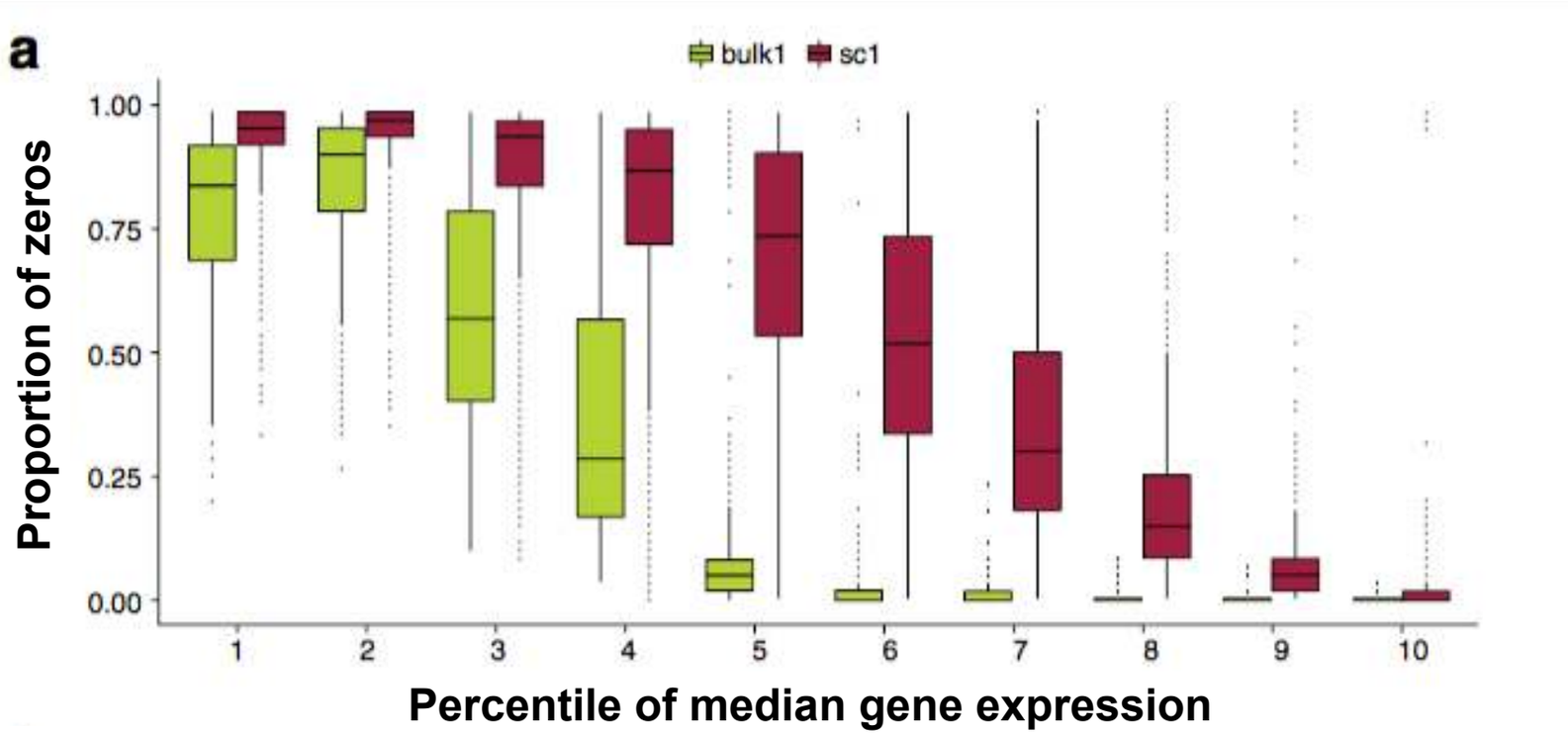
Randomly generate 500 points

Compute difference between max and min distance between any pair of points



Taken from Tan, Steinbach & Kumar, Introduction to Data Mining course
http://slideplayer.com/slide/6194466

The computational challenges - Zero Inflated data

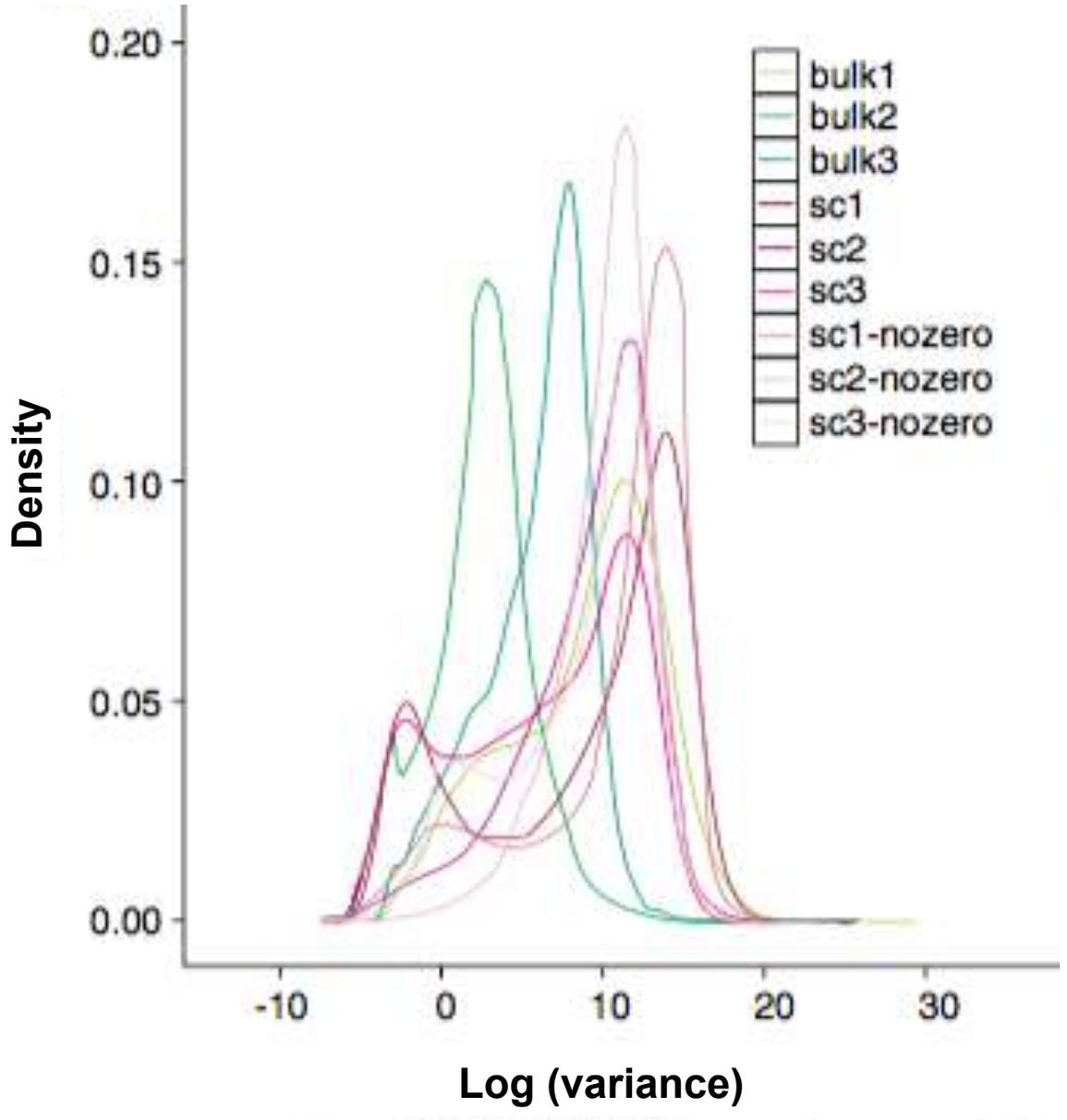


Bulk and SC sets with comparable depths

Bulk 1: 60 female bulk RNA-seq samples of individual Drosophila flies
SC1: 60 individual Mus musculus embryonic cells at various developmental time points

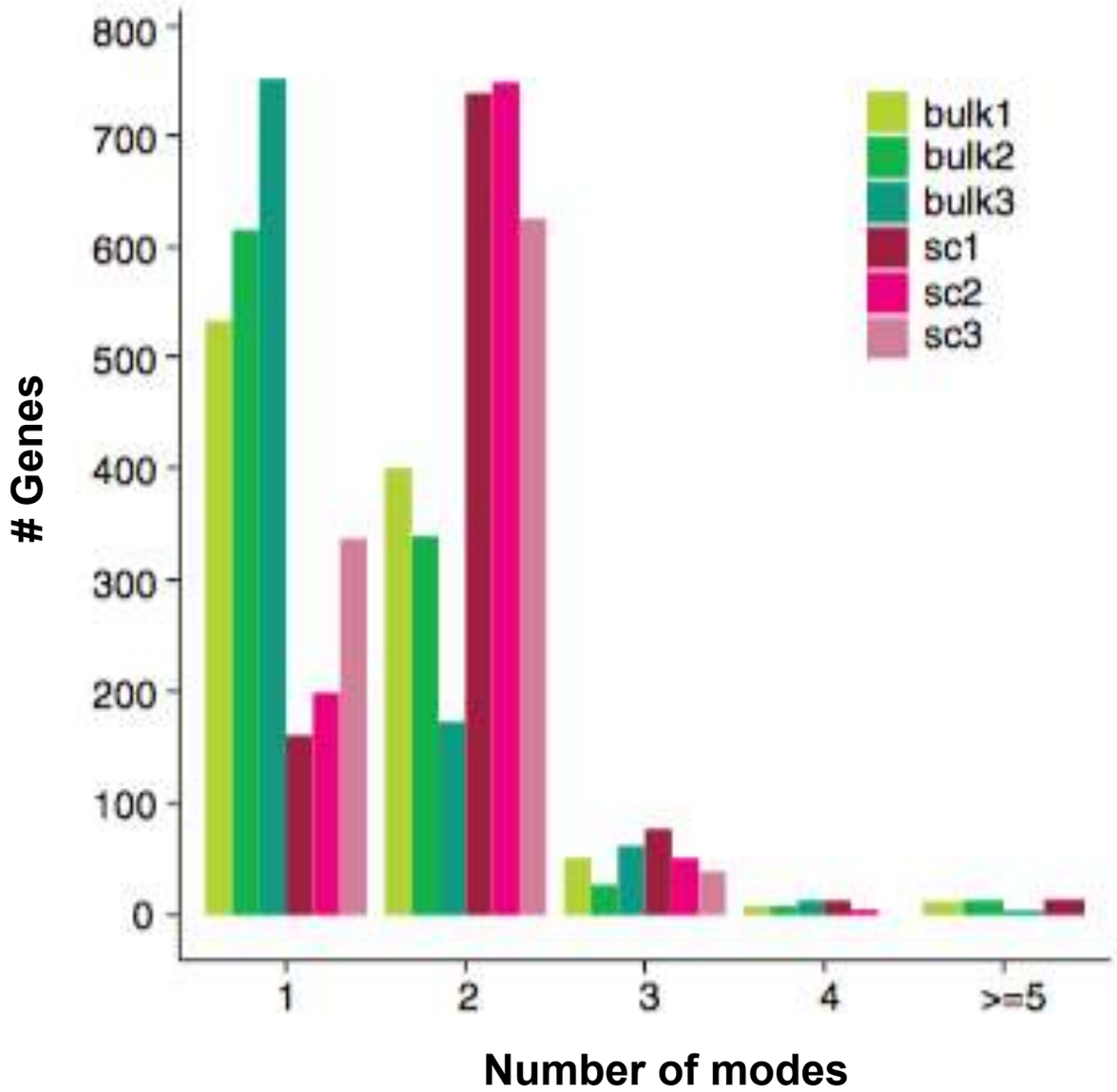
Bacher and Kendziorski Genome Biology (2016) 17:63

The computational challenges - High variability (overdispersion)



Bacher and Kendziorski Genome Biology (2016) 17:63

The computational challenges - Multimodality



Bacher and Kendziorski Genome Biology (2016) 17:63

The bioinformatics pipeline: main “modular” components

This afternoon

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies (L. Albergante)

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

Tomorrow

8 - Batch Effect correction and data integration

9 - Single-cell matching across datasets

The bioinformatics pipeline: Example 1

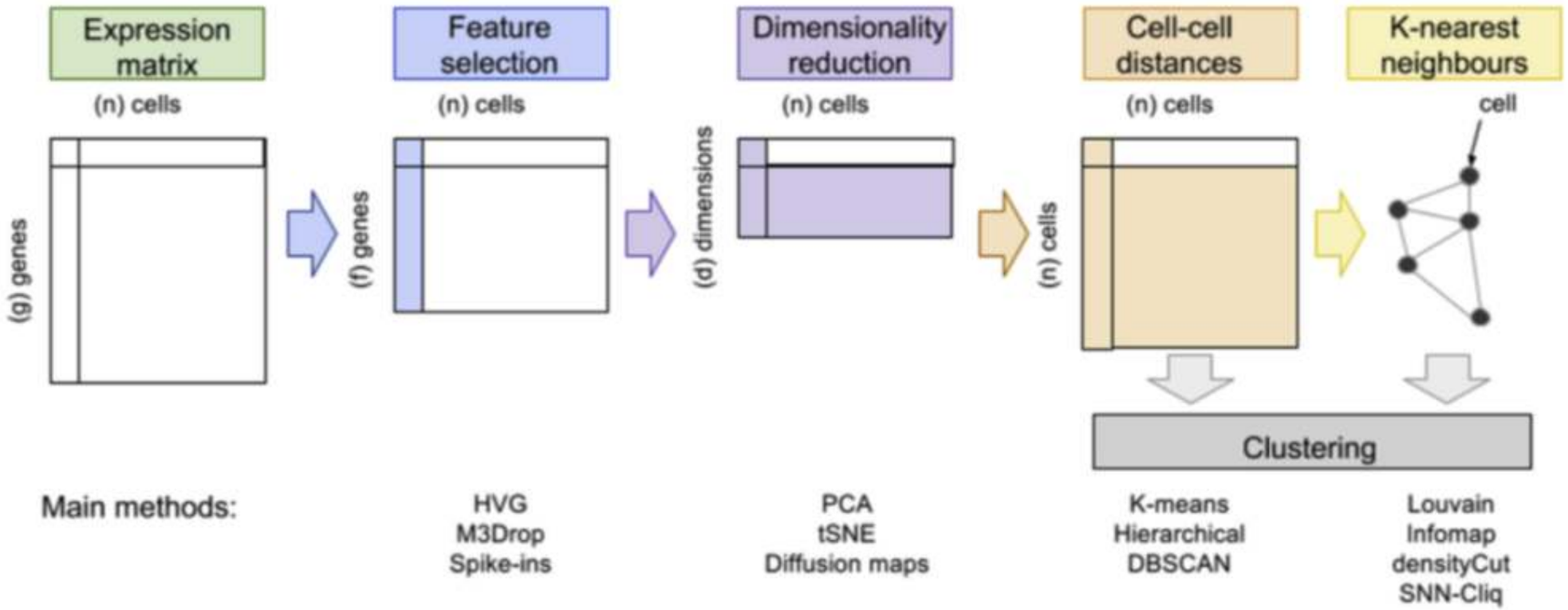
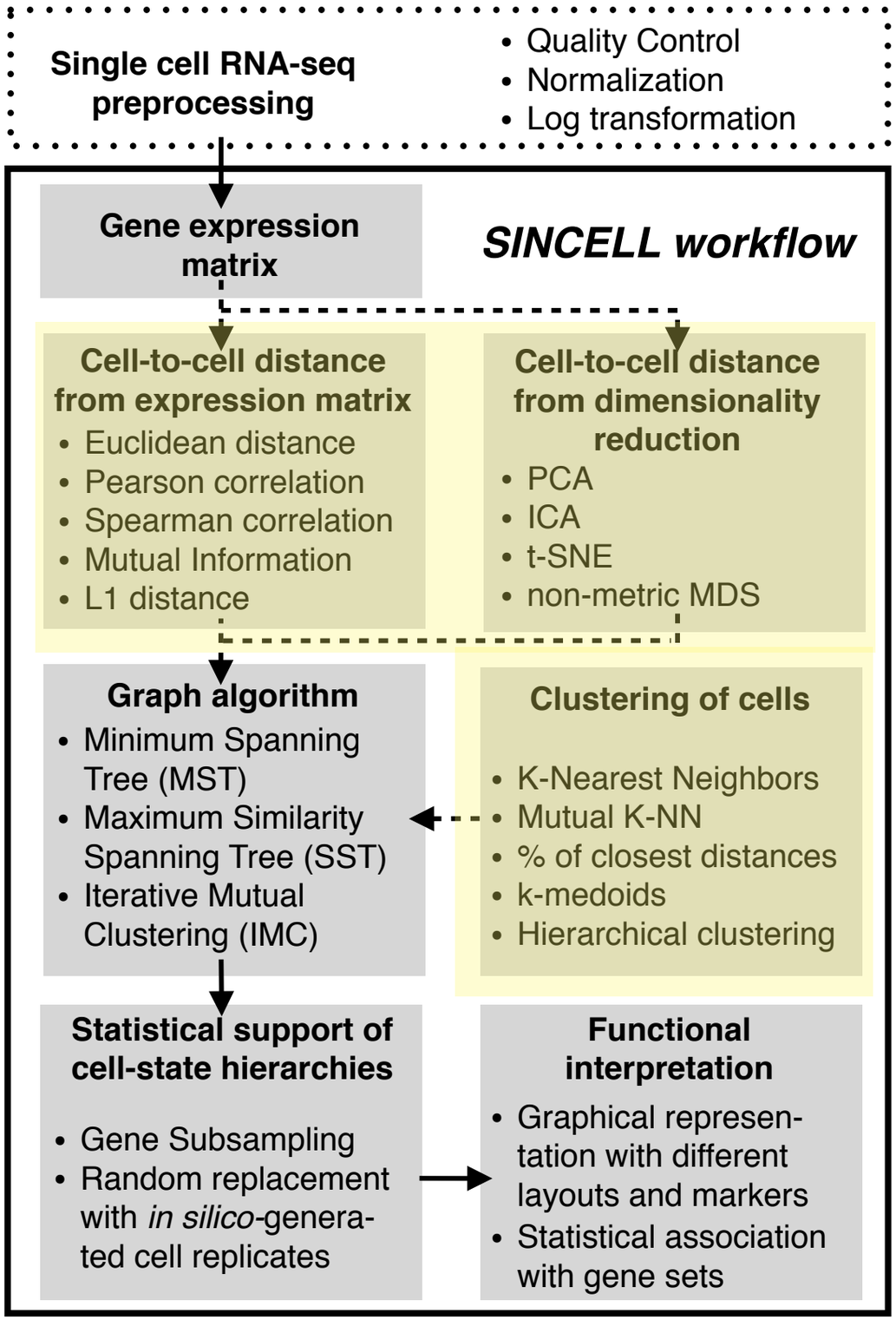


Fig. 1. Overview of methods covered in this review. Colour indicates which parts of the expression matrix are adjusted after each step, for instance feature selection only removes rows from the expression matrix, whereas dimensionality reduction calculates a new matrix composed of meta-features. Preprocessing steps not covered in detail in this review include quality control and normalization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Design and computational analysis of single-cell RNA-sequencing experiments. Bacher R & Kendziorski C. Genome Biology (2016) 17:63. https://doi.org/10.1186/s13059-016-0927-y

The bioinformatics pipeline: Example 2

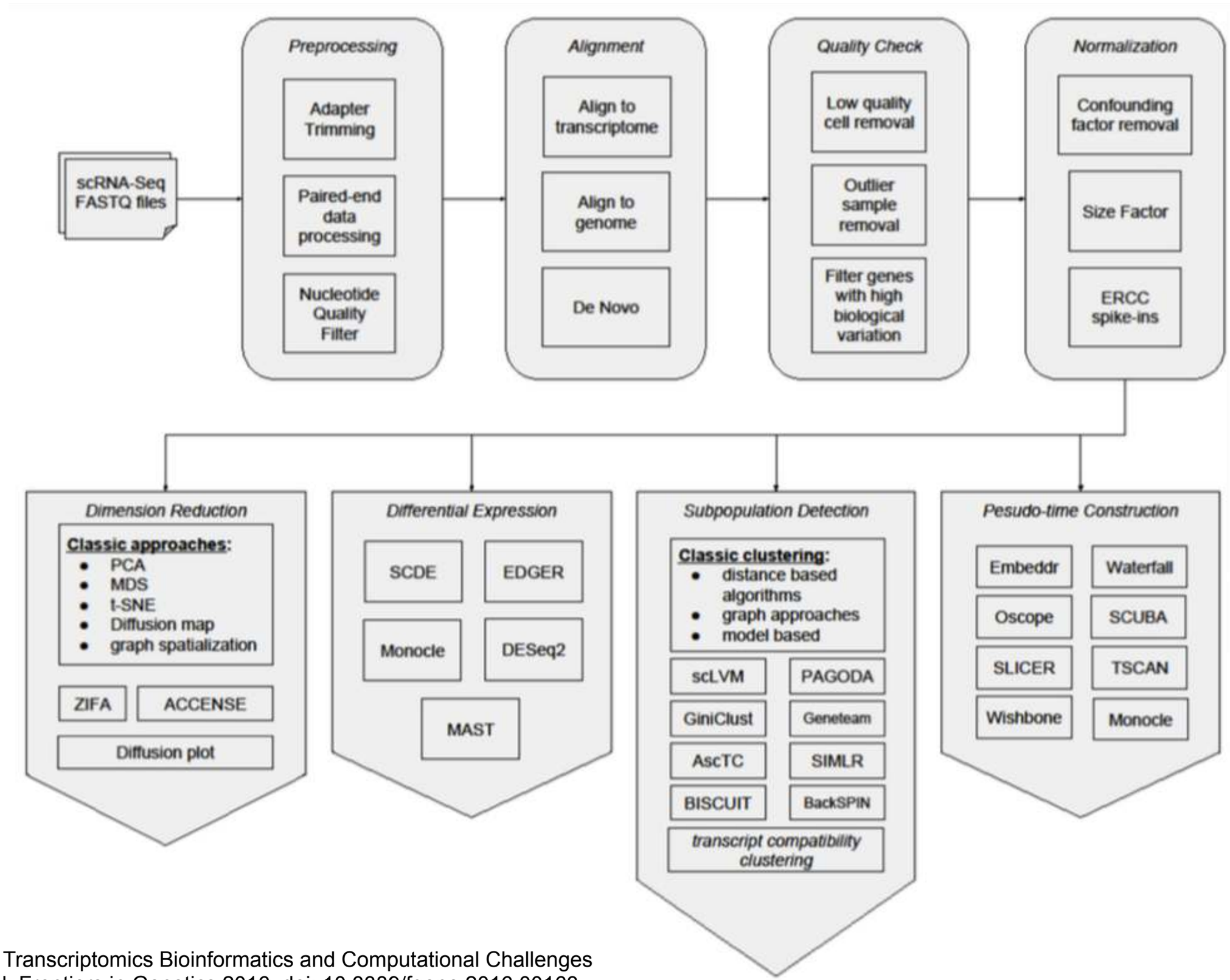


sincell software

http://bioconductor.org/packages/sincell

Molina, Telenti, Rausell*. Bioinformatics 2015

The bioinformatics pipeline: Example 3



Single-Cell Transcriptomics Bioinformatics and Computational Challenges
Poirion et al. Frontiers in Genetics 2016. doi: 10.3389/fgene.2016.00163

The bioinformatics pipeline: Example 4

F1000Research F1000Research 2016, 5:2122 Last updated: 20 APR 2018

Check for updates

SOFTWARE TOOL ARTICLE

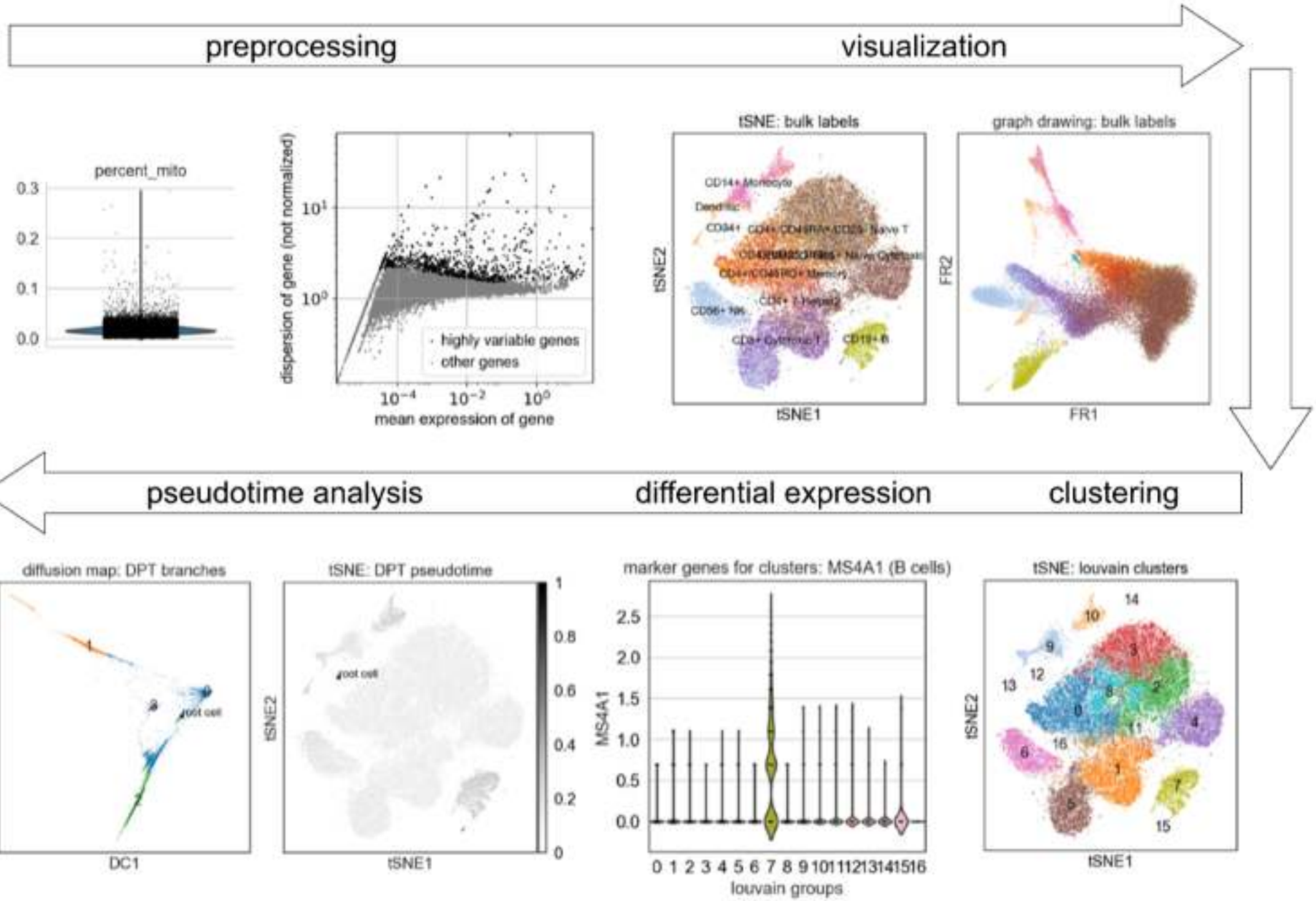
REVISED A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; referees: 3 approved, 2 approved with reservations]

Aaron T.L. Lun [1], Davis J. McCarthy[2,3], John C. Marioni[1,2,4]

See Rausell A. Referee Report For: **A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor** [version 2; referees: 3 approved, 2 approved with reservations]. *F1000Research* 2016, **5**:2122 (doi: 10.5256/f1000research.10712.r17328)

Examples of *all-in-one* environments: SCANPY

SCANPY: large-scale single-cell gene expression data analysis.
Wolf et al. Genome Biology 2018, 19:15 https://doi.org/10.1186/s13059-017-1382-0



Examples of *all-in-one* environments: SEURAT

Expression QC

Normalization

Highly variable genes

Dealing with confounders

Dimensional Reduction

Visualization

Marker genes

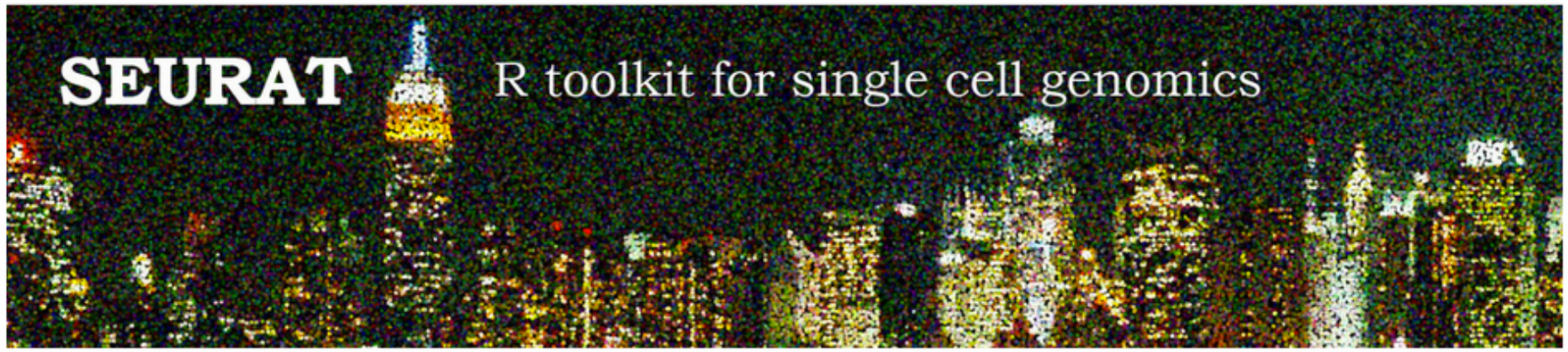
Cell Cycle Regression

Clustering cells

Differential expression

Multimodal Analysis

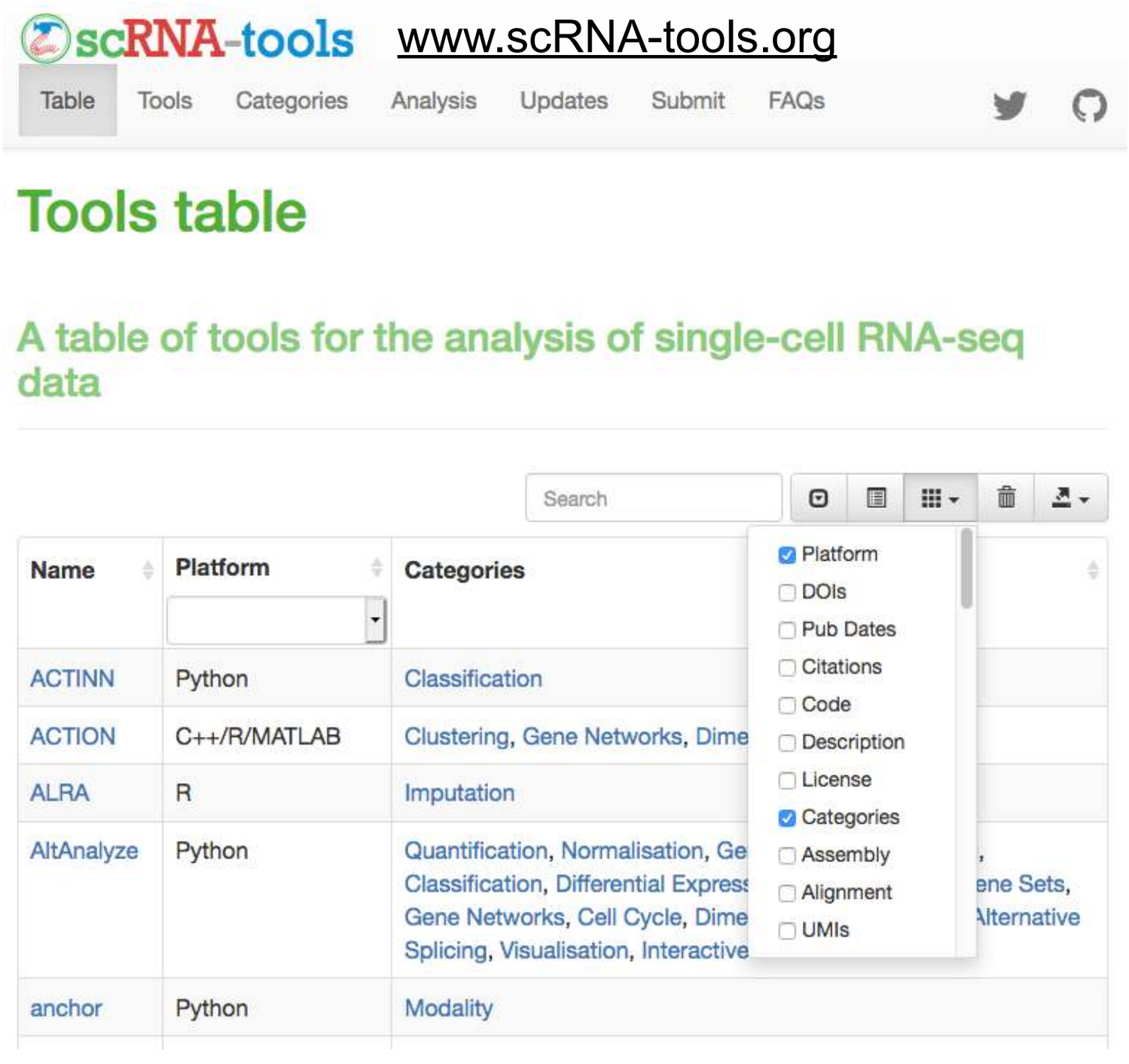
https://satijalab.org/seurat/



See tutorial at:
https://hemberg-lab.github.io/scRNA.seq.course/seurat-chapter.html

Online catalogue: scRNA-tools database

Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. Zappia, et al. Plos Comp Biol 2018. https://doi.org/10.1371/journal.pcbi.1006245



Outline

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Trajectories (Wouter Saelens)

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

Feature extraction (I)

A. Simple filtering criteria, eg:

- Filtering of lowly expressed genes (see e.g Lun et al F1000Res 2016; Sonesson & Robinson Nature Methods 2018):
  - genes expressed in < x% of cells
  - genes with a mean average of expression < threshold

- Restrict to protein coding genes

B. M3Drop: Dropout-based feature selection for scRNASeq:

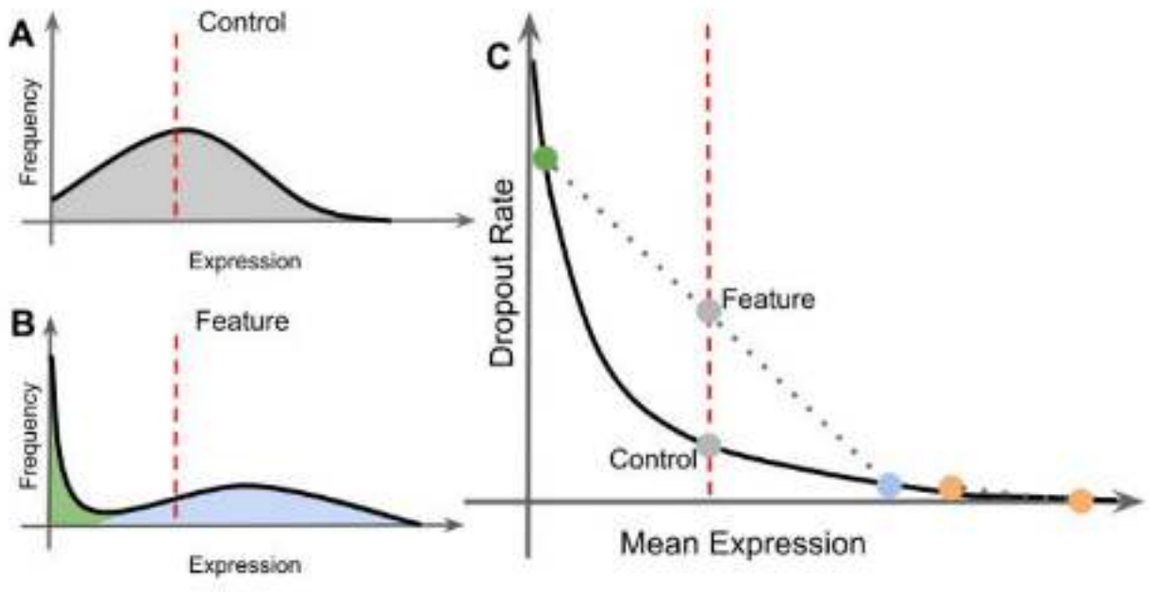


Figure 1: Differentially expressed genes exhibit bimodal expression which increases the dropout rate relative to the mean expression. (**A & B**) Genes with the same mean expression (dashed red line), but (A) is expressed evenly across cells, whereas (B) is highly expressed in some cells (blue) and lowly expressed in others (green). (**C**)This leads to a surplus of dropouts since mean and dropout rate average linearly (dotted line) whereas the expectation (black line) is non-linear. Orange points indicate a gene with very high expression where differential expression leads to only a small increase in dropout-rate.

Michaelis-Menten function to the relationship between mean expression (S) and dropout-rate (M3Drop).

$$P_{dropout} = 1 - \frac{S}{K_M + S}$$

Since the Michaelis-Menten function has a single parameter (Km), we can test the hypothesis that the gene-specific Ki is equal to the Km that was fit for the whole transcriptome. This can be done by propagating errors on both observed dropout rate and observed mean expression to estimate the error of each Ki. The significance can then be evaluated using a t-test

Andrews and Hemberg. Bioinformatics (2018)
https://doi.org/10.1093/bioinformatics/bty1044

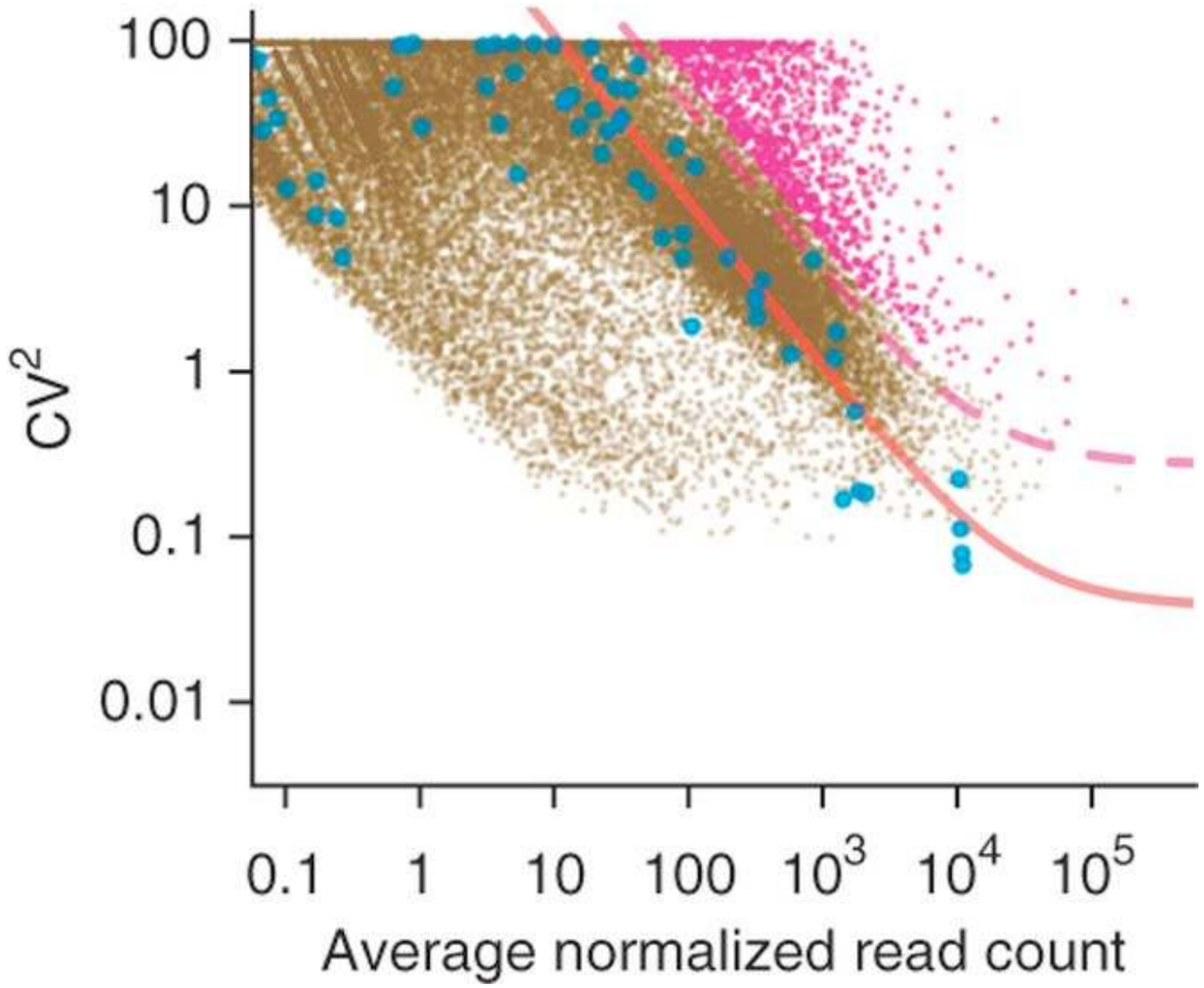
Feature extraction (II) - Highly Variable Genes

B. Selection of Highly Variable Genes

Identifying highly variable genes

Brennecke et al. [48]	A gamma generalized linear model fit to the mean-variance relationship quantified by the square of the coefficient of variation (CV^2) of the spike-ins estimates technical noise parameters. These parameters are then used to estimate technical variability for endogenous genes and to test whether each gene exceeds a variability threshold	Spike-ins and endogenous genes are normalized separately using the median normalization method. Gene specific *P* values are provided to identify highly variable genes
Kim et al. [63]	Uses spike-ins to estimate parameters related to technical variance, allowing for differences in variability across cells. Estimates gene-specific biological variability by subtracting technical variability from total variance	Normalization factors are estimated using the median normalization method. A simulation based framework to test for highly variable genes is provided
BASiCS [54]	Jointly models spike-ins and endogenous genes as two Poisson-Gamma hierarchicalmodels with shared parameters	Estimates normalization parameters jointly across all genes. Gene-specific posterior probabilities are provided to identify both lowly and highly variable genes

Feature extraction (II) - Highly Variable Genes (Brennecke et al)

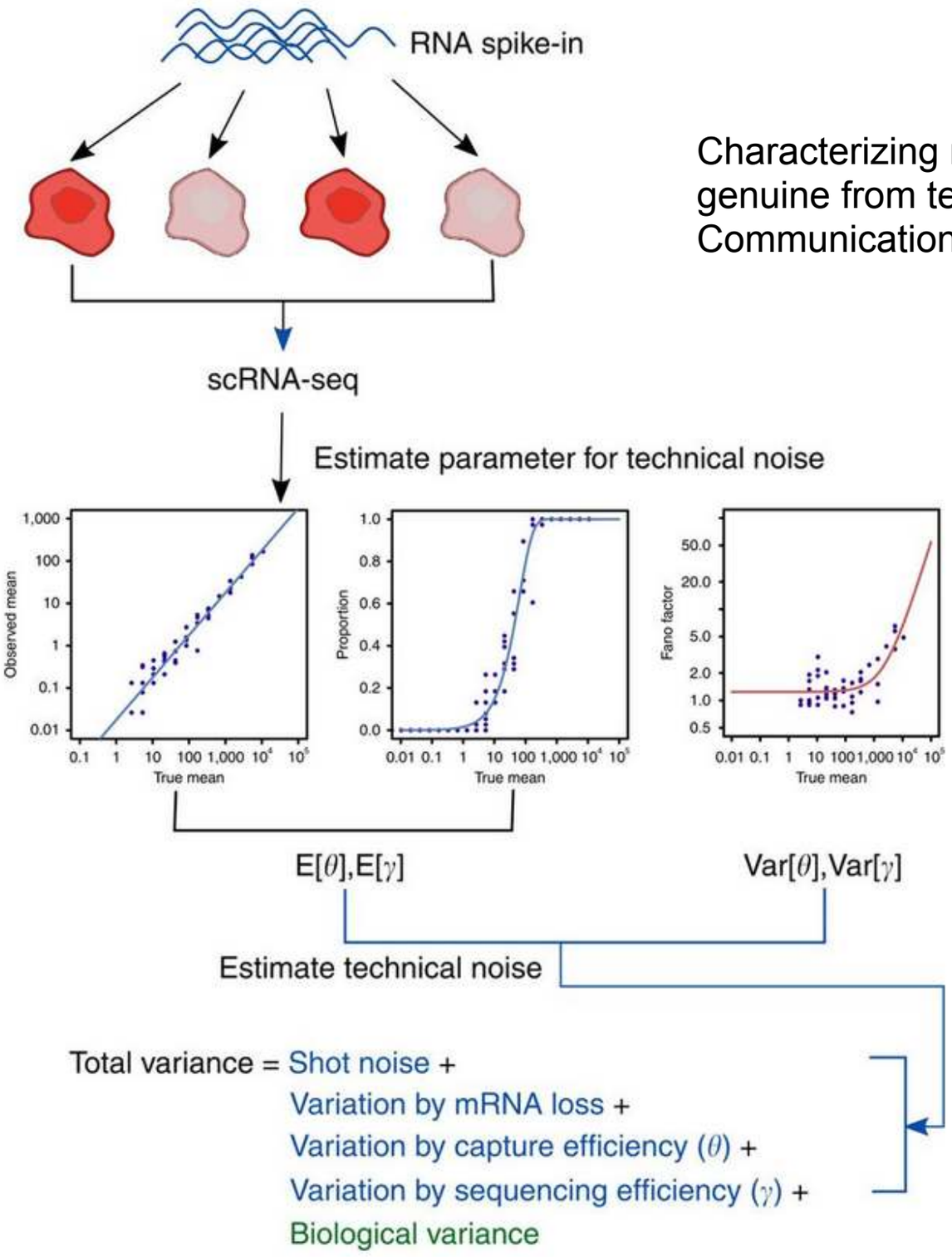


Accounting for technical noise in single-cell RNA-seq experiments.
Brennecke et al. Nature Methods (2013) 10:1093–1095

The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean

$$c_v = \frac{\sigma}{\mu}.$$

Feature extraction (II) - Highly Variable Genes (Kim et al)



Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Kim et al. Nature Communications, 2015.

With the help of external RNA spike-in molecules, added at the same quantity to each cell's lysate, we first estimate four parameters capturing technical variability, which are the expectation and variance of capture (θ) and sequencing (γ) efficiency. Then, by the general variance decomposition formula, the total observed variance of read counts can be decomposed into technical (blue) and biological (green) variance terms. The estimate of biological variance can be obtained by subtracting technical variance terms from the total observed variance. Shot noise (or Poisson noise) is cell-to-cell variability that can be modelled by a Poisson process.

Feature extraction (II) - Highly Variable Genes (BASiCS, Vallejos et al)

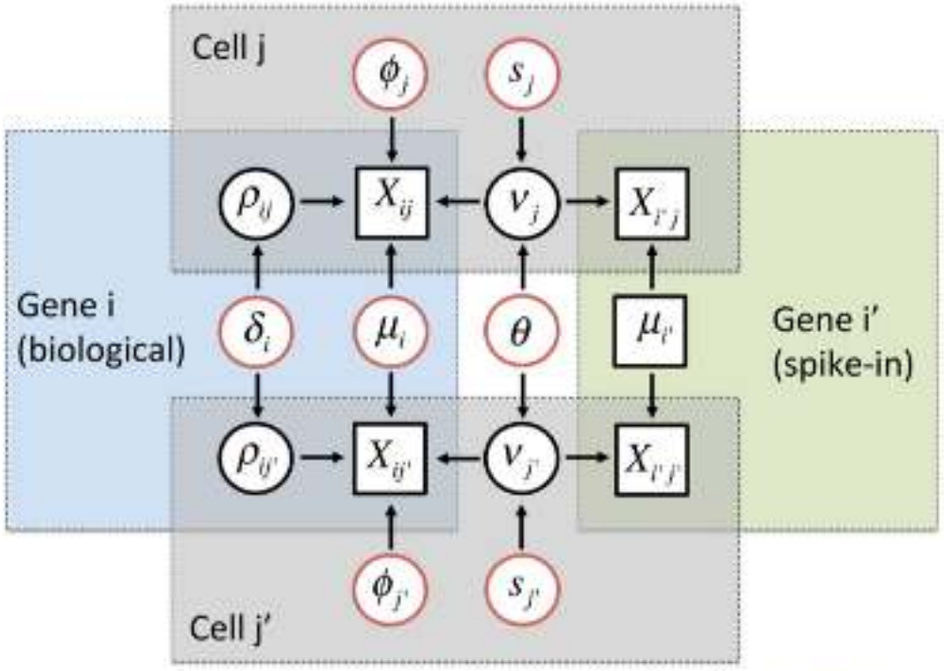


Fig 2. Graphical representation of the hierarchical model implemented in BASiCS. Diagram based on

BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. Vallejos, Marioni & Richardson. Plos Comp Biol 2015. doi:10.1371/journal.pcbi.1004333

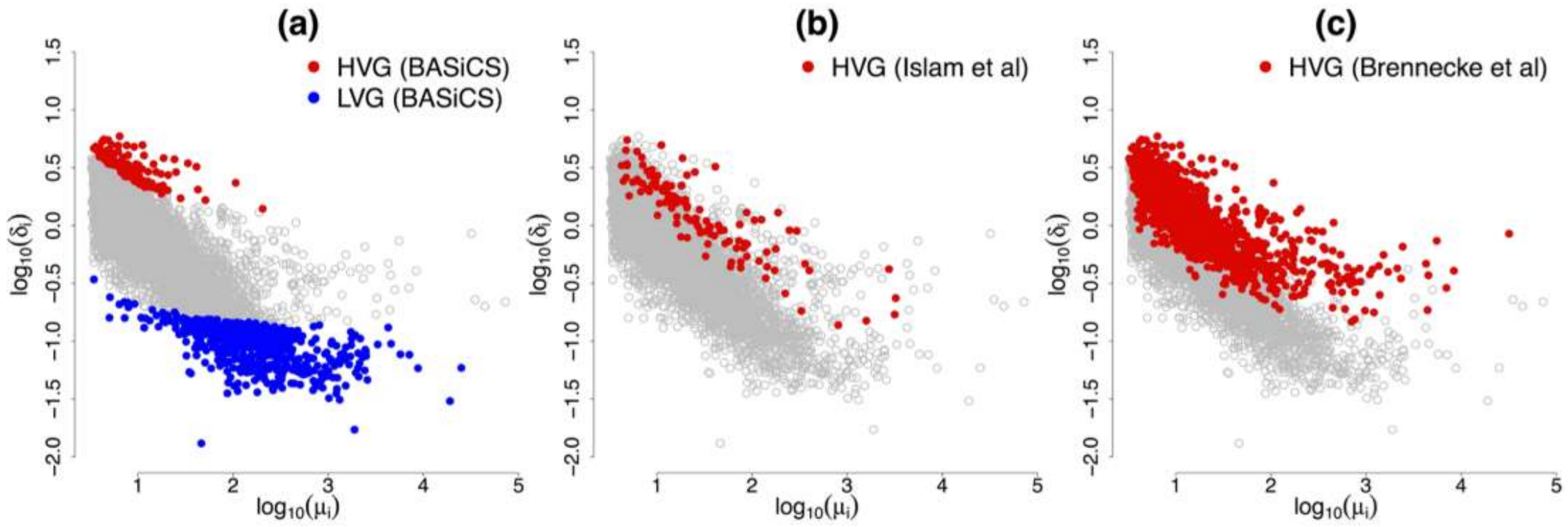


Fig 8. Comparison of HVG detection among different methods. For each of the 7,895 biological genes, posterior medians of biological cell-to-cell heterogeneity term δ_i (log scale) against posterior medians of expression level μ_i (log scale). While the methods described in [16] and [5] only provide a characterisation of HVG, BASiCS is able to detect those genes whose expression rates are stable among cells.

doi:10.1371/journal.pcbi.1004333.g008

Benchmark of HVG-detection methods: low overlap among methods

Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data.
Yip et al. Briefings in Bioinformatics, 2018. doi: 10.1093/bib/bby011

Table 2. Percent overlap between a method's detected genes with the same number of highest expressing genes

Method	Dataset			Average (%)
	Deng (%)	Islam (%)	Yan (%)	
Brennecke	23.22	25.49	15.78	21.50
scLVM_Log	82.84	63.68	87.75	78.09
scLVM_LogVar	68.28	50.09	63.44	60.60
scran	22.65	29.29	26.04	25.99
scVEGs	NA	1.10	NA	1.10
Seurat	10.03	22.17	5.27	12.49

Note: Results from the Deng, Islam and Yan data sets are reported. BASiCS is not tested because of its low stability.

- BASiCS' results ranked top with simulated data sets but ranked last in real data set. This shows that noises in real scRNA-seq can cause lower performances in methods that use spike-ins.
- Reproducibilities are low, among different tools and among different samples analyzed by the same tool. A higher number of cells can improve rediscovery rates.
- BASiCS, Brennecke, scVEGs and Seurat are shown to have high type I/II error rates.

Feature extraction (III) - Highly correlated genes (Lun et al. 2016)

- Identify the genes highly correlated with one another.
- Correlations between genes are quantified by computing Spearman's rho
- Gene pairs with significantly large positive or negative values of rho are retained
- This distinguishes between HVGs caused by random noise and those involved in driving systematic differences between subpopulations.

Other examples of the use of highly correlated genes (see Andrews & Hemberg, Molecular Aspects of Medicine 59 (2018), p117, 118):

- Andrews and Hemberg, 2016
- Macosko et al., 2015
- Pollen et al., 2014
- Usoskin et al., 2015
- Fan et al., 2016

NOTE

"We only apply this function to the set of HVGs, because these genes have large biological components and are more likely to exhibit strong correlations driven by biology. In contrast, calculating correlations for all possible gene pairs would require too much computational time and increase the severity of the multiple testing correction. It may also prioritize uninteresting genes that have strong correlations but low variance, e.g., tightly co-regulated house-keeping genes"

Outline

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies (L. Albergante)

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

Dimensionality reduction - Why?

1. Need of an orthogonal space

2. Minimize curse of dimensionality

3. Filter out noise

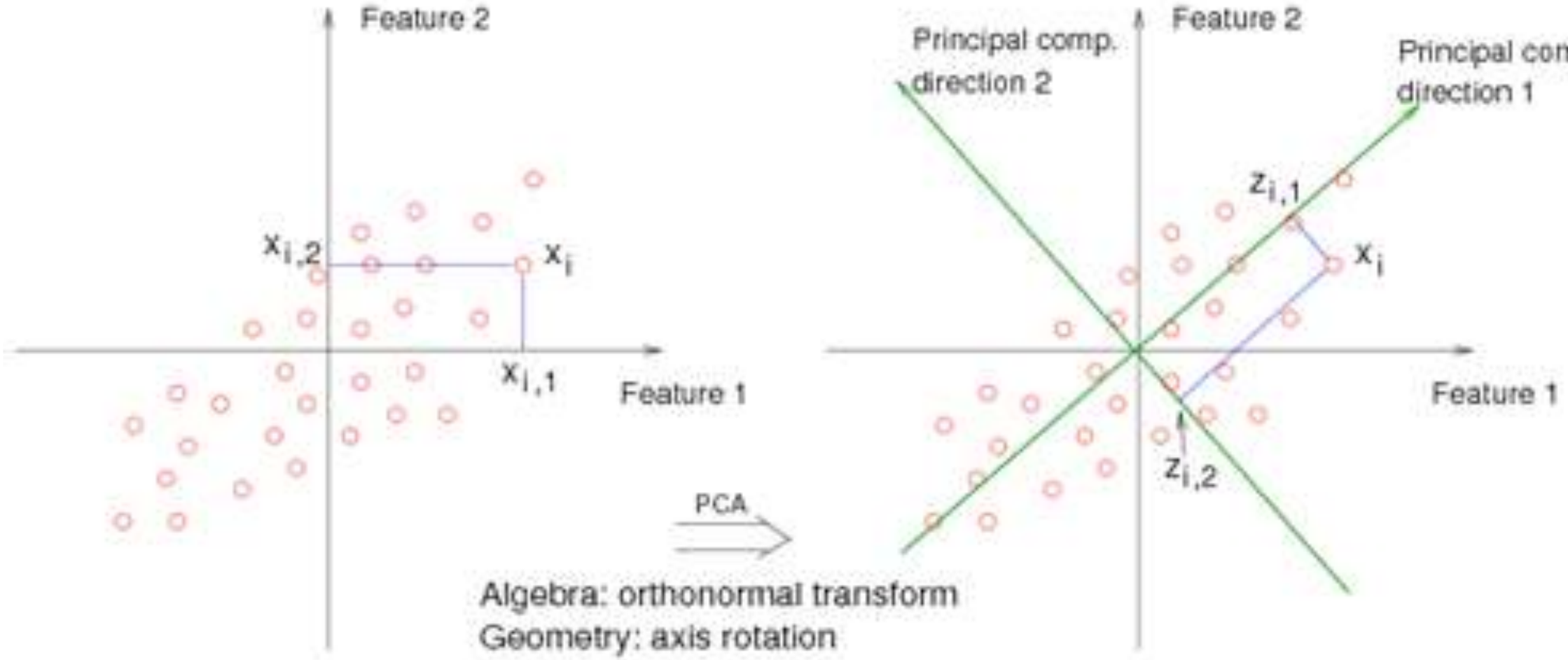
4. Allow visualization

5. Reduce computational load

Popular methods used for single-cell data analysis:

1. PCA
2. ICA
3. tSNE
4. UMAP
5. Others : Diffusion map, Isomap

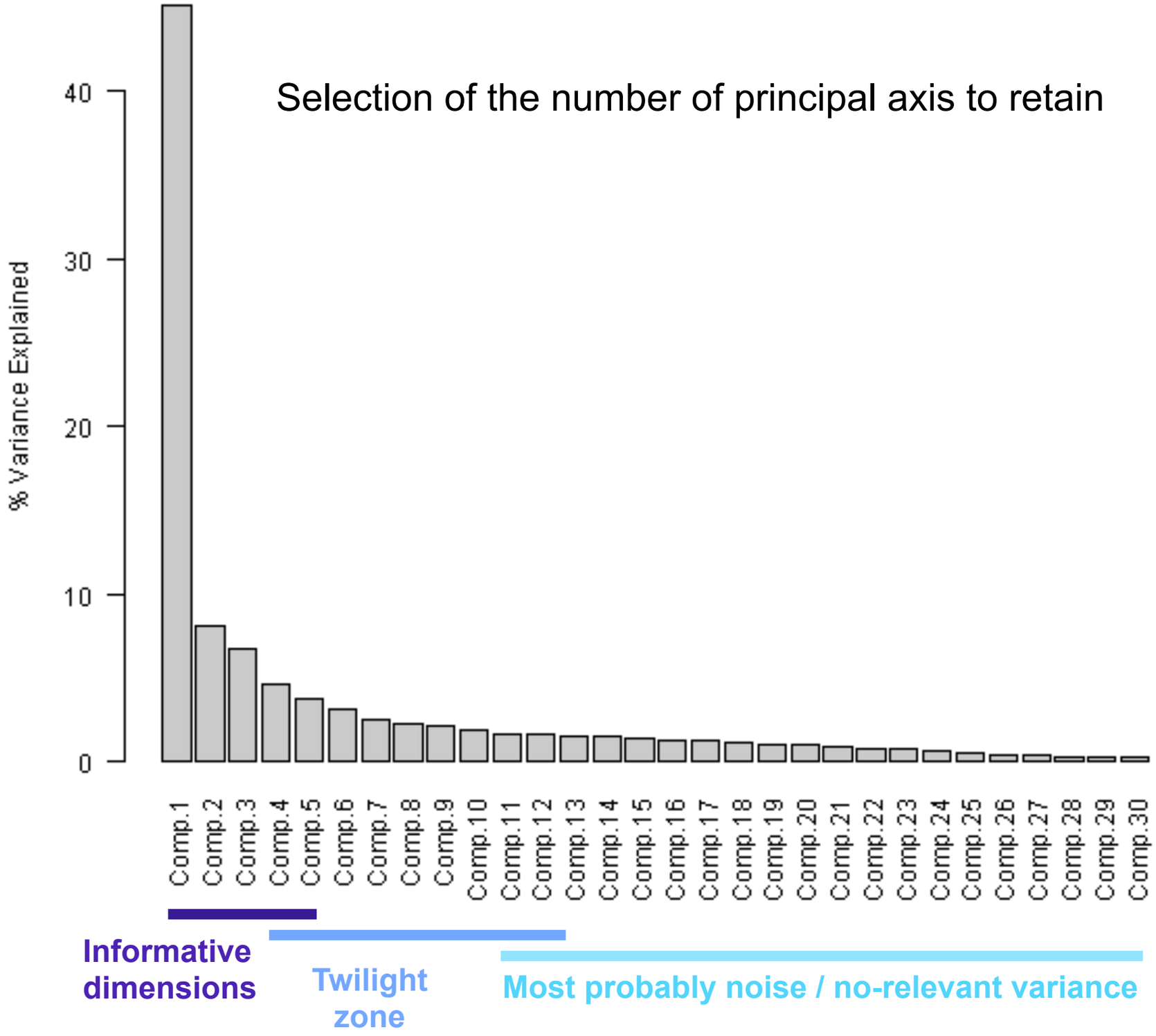
Dimensionality reduction (I) - Principal Component Analysis (PCA) (I)



Source URL: https://onlinecourses.science.psu.edu/stat857/node/35

Dimensionality reduction (I) - Principal Component Analysis (PCA) (II)

Selection of the number of principal axis to retain



https://www.r-bloggers.com/visualizing-principal-components/

Further reading: https://hemberg-lab.github.io/scRNA.seq.course/seurat-chapter.html#significant-pcs

Dimensionality reduction (I) - Principal Component Analysis (PCA) (III)

PCA Advantages and limitations

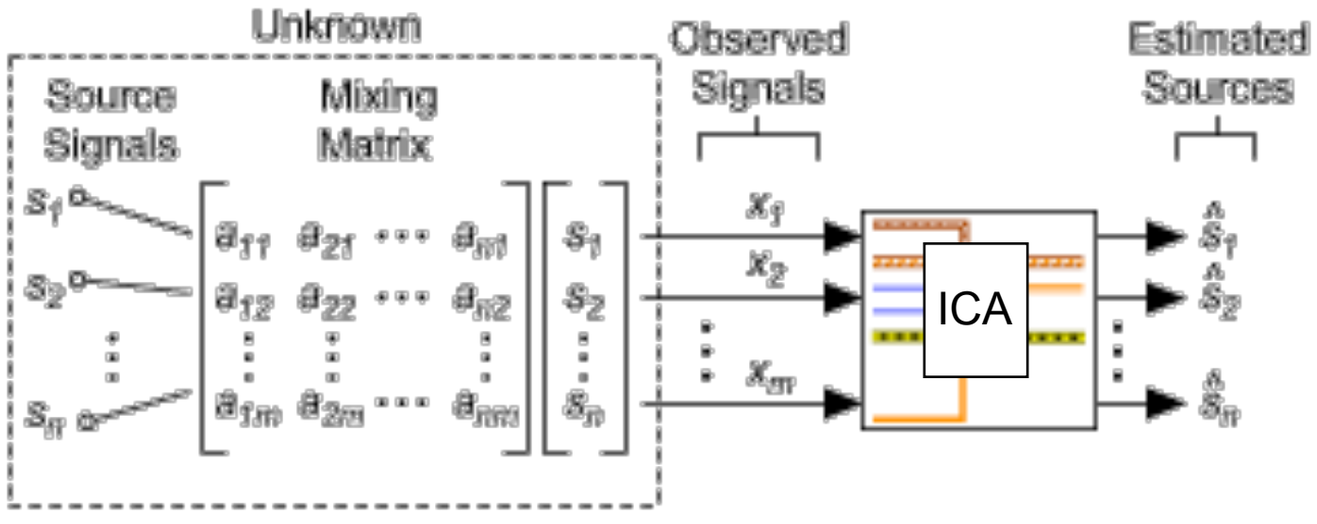
- Based on linear transformations
- Captures the dimensions with higher variance
- Objective control on the amount of retained dimensions
- Fast & scalable
- Preserves both long-range and short-range relationships

Extensions of the PCA approach

- A variation of PCA which explicitly deals with the large number of zero-values in scRNASeq data has been developed (ZIFA, Pierson and Yau, 2015) but the zero-inflation model employed may not fit all datasets (Andrews and Hemberg, 2016).

- Recently Risso et al. (2017) proposed a method similar to PCA based on a zero- inflated negative binomial model instead of a Gaussian model.

Dimensionality reduction (II) - Independent Component Analysis (ICA)

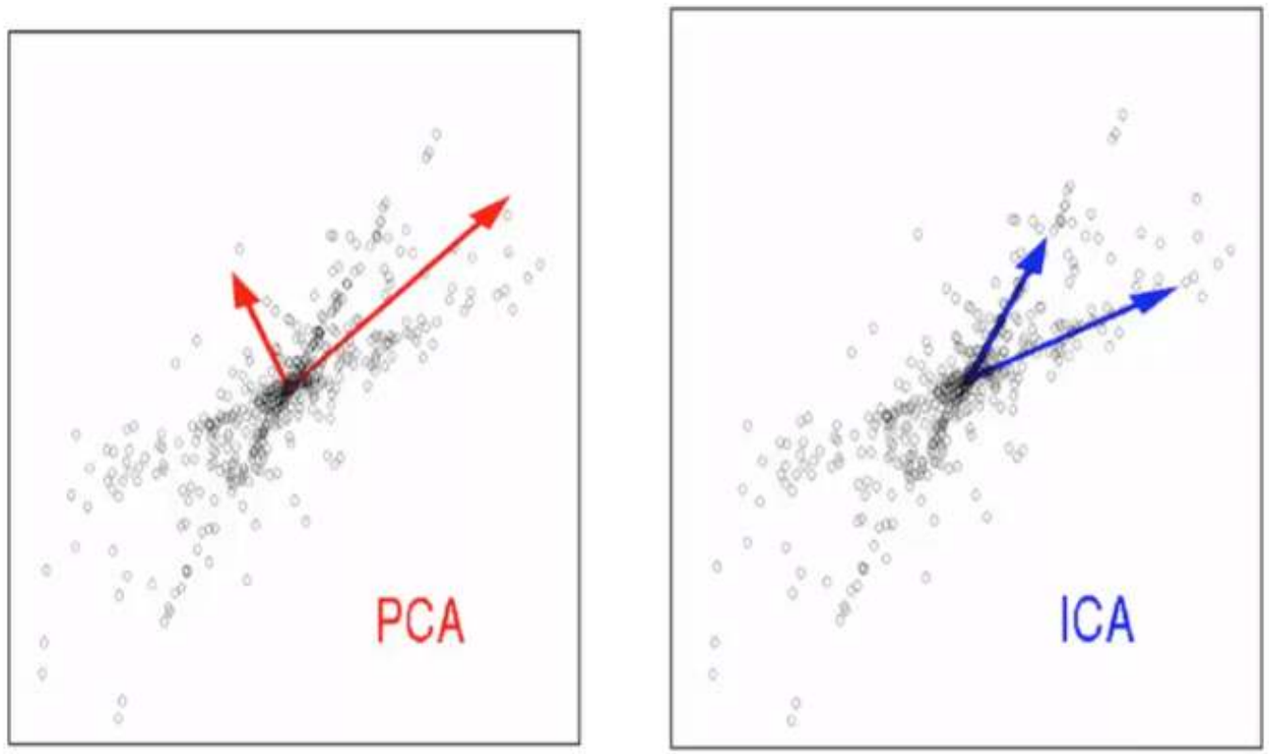


URL source: https://zone.ni.com/reference/en-XX/help/371419D-01/lvasptconcepts/tsa_multivariate_stat_analysis/

Independent component analysis (ICA):

1. Decompose a multivariate signal into statistically independent components

2. Estimated components optimize a proxy for "independence":
   Minimization of mutual information
   Maximization of non-Gaussianity

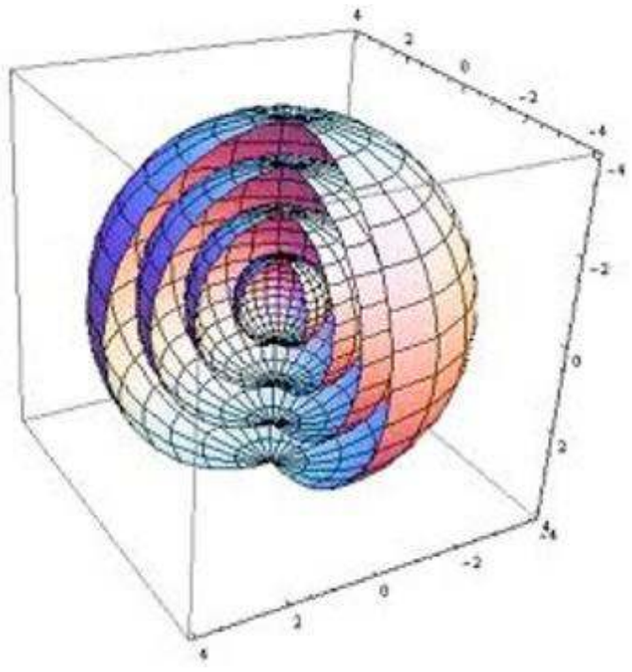


URL source: https://mr-why.com/machinelearning/tom-mitchell-machine-learning-09-learning-representations

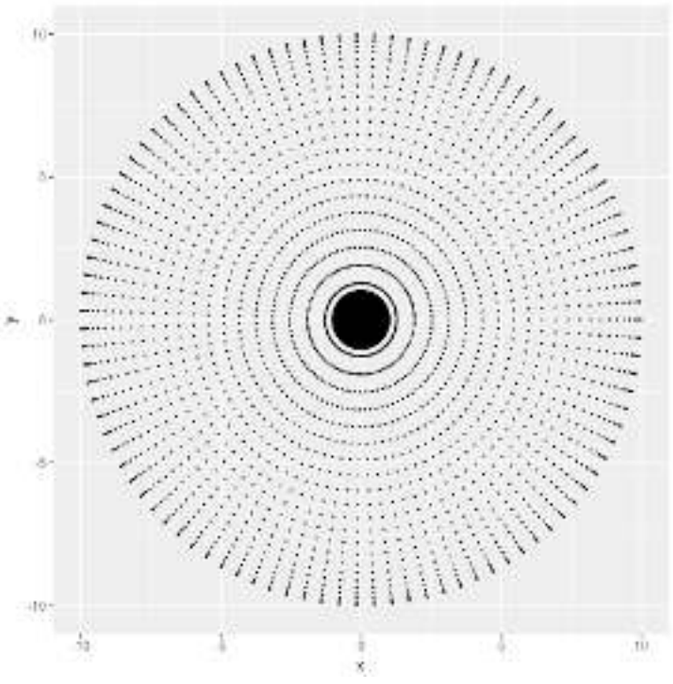
Dimensionality reduction (III) - tSNE
t-distributed stochastic neighbor embedding

tSNE is a non-linear dimension reduction technique able to show structures in the data that cannot be found simply by changing the direction in which you look

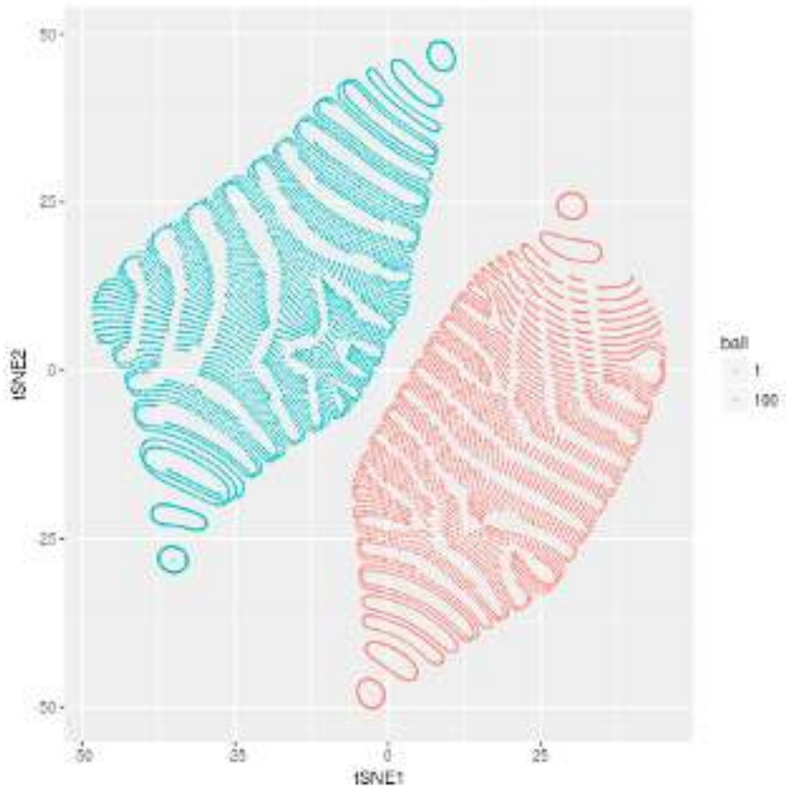
Original



PCA



tSNE



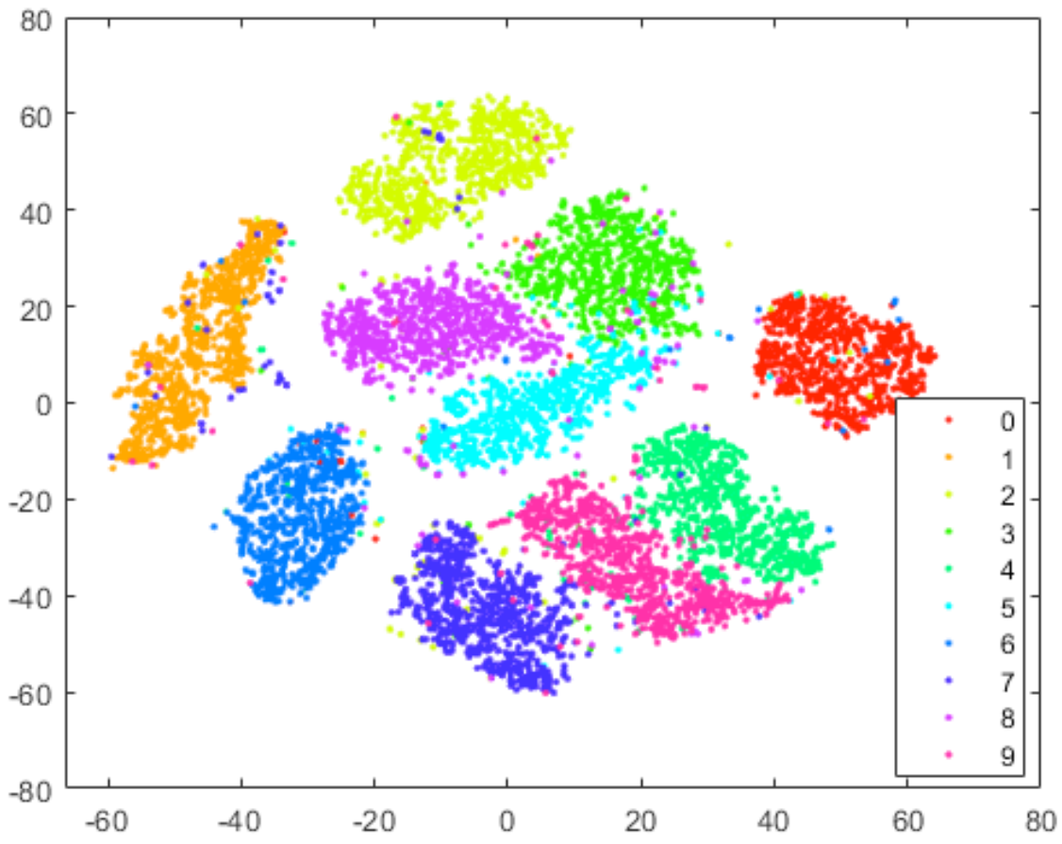
tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

Dimensionality reduction (III) - tSNE

tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

How to Use t-SNE Effectively, Wattenberg, et al. Distill, 2016
https://distill.pub/2016/misread-tsne/

t-SNE in wikipedia
https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding



Dimensionality reduction (III) - t-SNE

This is sort of what tSNE is trying to do, but rather than operate on the distances it operates on transformed distances. tSNE defines two quantities,

$$p_{ij} \sim \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)$$

and

$$q_{ij} \sim (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$$

where $\mathbf{x}$ are positions in the original space, $\mathbf{y}$ are positions in the 2D tSNE map and i and j indicate different points. tSNE aims to pick values of $\mathbf{y}$ such that $p_{ij} \approx q_{ij}$.

You might also think that tSNE should be required to place points that are distant in the original space far apart in the 2D map, but this is not the case. This is because tSNE makes q_{ij} match p_{ij} by minimising the Kullback-Leibler divergence, defined as,

$$\sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

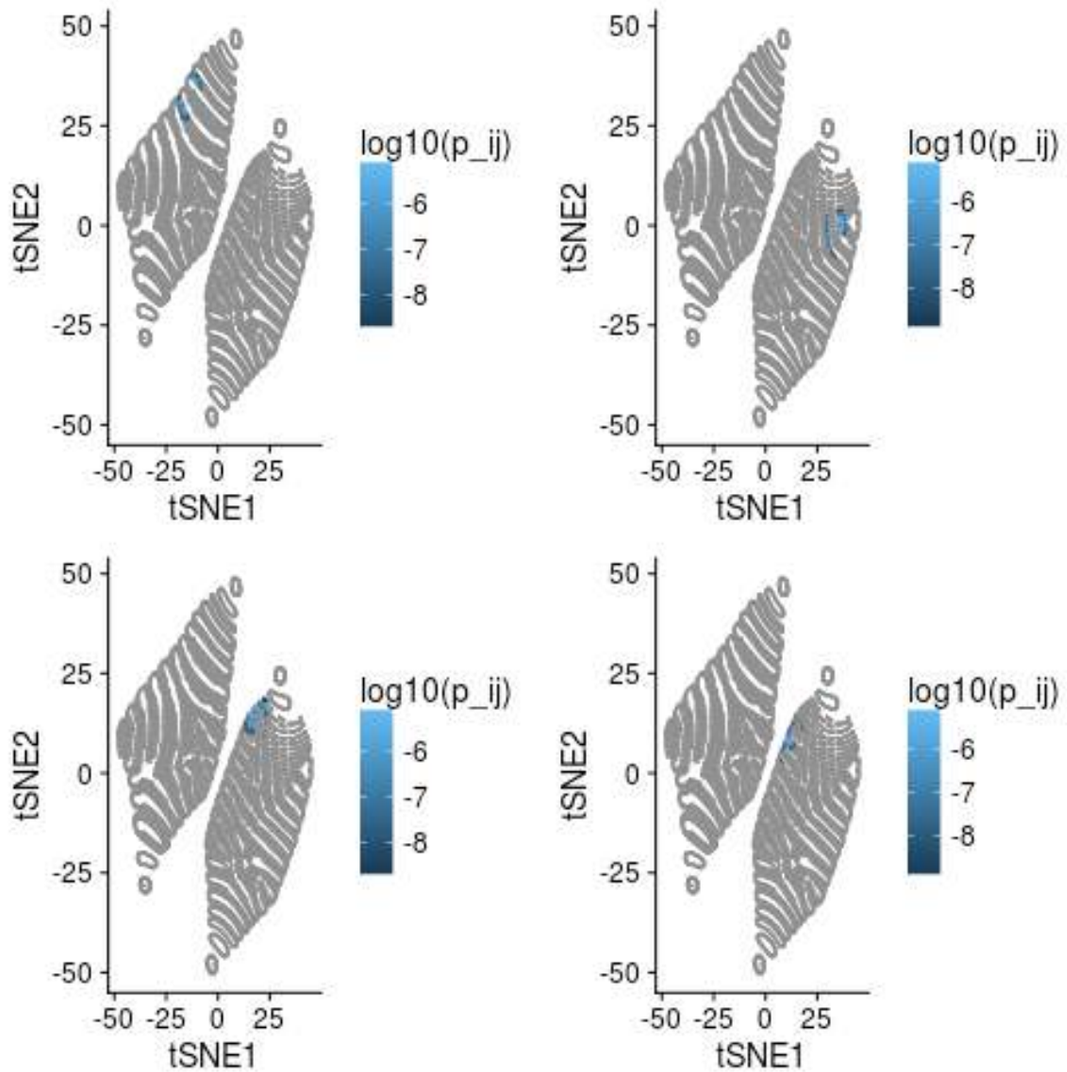
which is equal to 0 for any q_{ij} when $p_{ij} = 0$.

tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

Dimensionality reduction (III) - t-SNE

the algorithm only cares about placing the nearest neighbours of each point correctly -> Long distances are meaningless

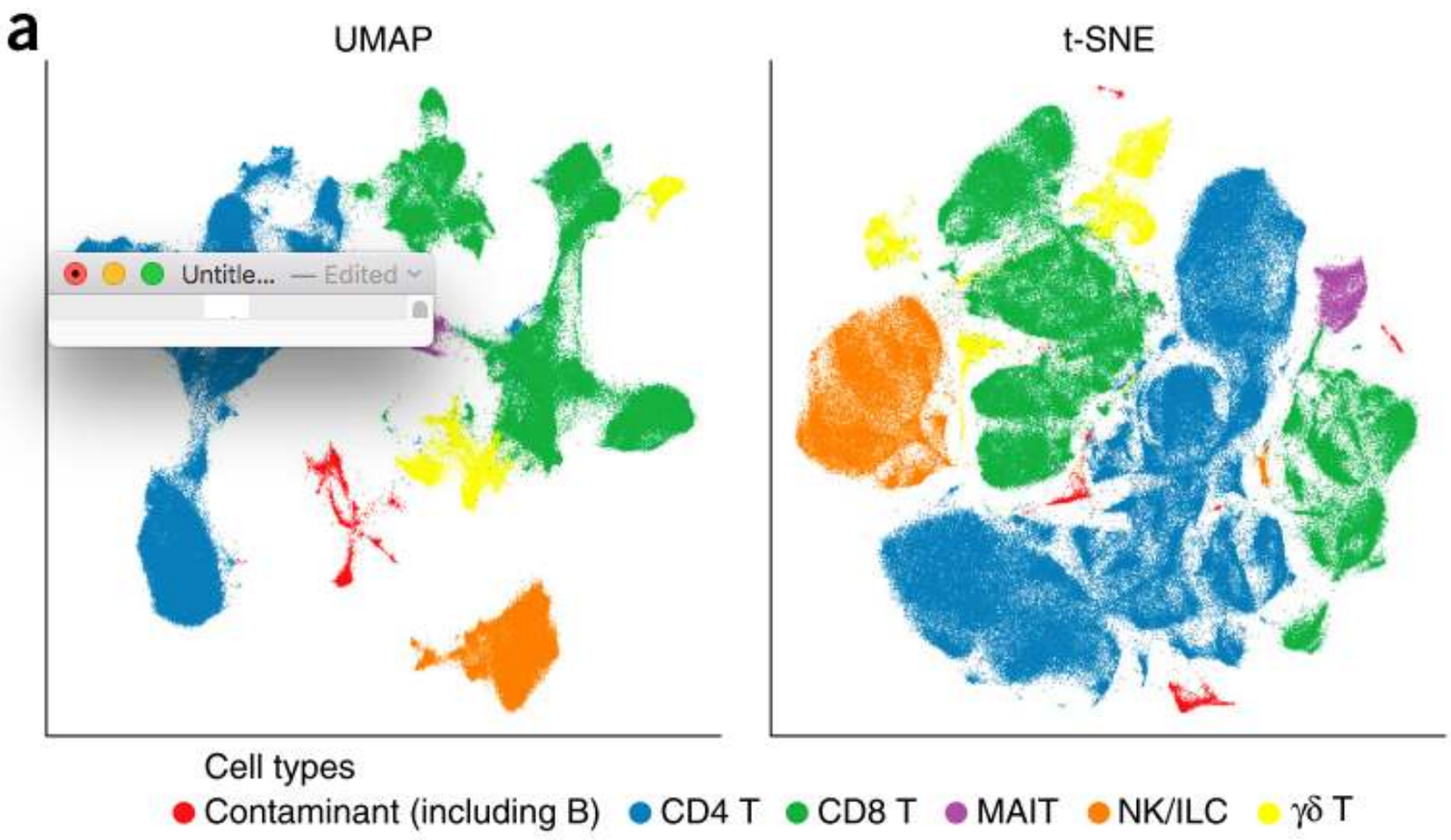
WARNING: Strong influence of hyperparameters + Non-deterministic



tSNE: What the hell is it?, by Matthew Young
https://constantamateur.github.io/2018-01-02-tSNE/

Dimensionality reduction (IV) - Uniform Manifold Approximation and Projection (UMAP)

Uniform manifold approximation and projection (UMAP):
- Claimed to preserve as much of the local & more of the global data structure than t-SNE
- Shorter run time.



Dimensionality reduction for visualizing single-cell data using UMAP
Becht et al. Nature Biotechnology 2018. http://www.nature.com/doifinder/10.1038/nbt.4314

Dimensionality reduction (IV) - Uniform Manifold Approximation and Projection (UMAP)

UMAP : class of k-neighbour based graph learning algorithms (similar to Isomap and t-SNE)

1st: Construct a **weighted k-neighbour graph**: **nearest neighbor descent algorithm** (Dong et al. 2011)

$$\bar{G} = (V, E, w)$$

2nd: Compute a **low dimensional layout** of such graph

UMAP uses a **force directed graph layout** algorithm in low dimensional space.
Set of **attractive forces** applied along edges
Set of **repulsive forces** applied among vertices.

The **attractive force** between two vertices i and j at coordinates yi and yj respectively, is determined by:

$$\frac{-2ab\|\mathbf{y_i} - \mathbf{y_j}\|_2^{2(b-1)}}{1 + \|\mathbf{y_i} - \mathbf{y_j}\|_2^2} w((x_i, x_j))\,(\mathbf{y_i} - \mathbf{y_j})$$

a, b: hyperparameters
w: weight function

The **repulsive force** is given by:

$$\frac{b}{(\epsilon + \|\mathbf{y_i} - \mathbf{y_j}\|_2^2)\,(1 + \|\mathbf{y_i} - \mathbf{y_j}\|_2^2)} (1 - w((x_i, x_j)))\,(\mathbf{y_i} - \mathbf{y_j})$$

, ϵ = 0.001

Untitled — Edited

The algorithm proceeds by iteratively applying attractive and repulsive forces at each edge or vertex. Convergence is achived by slowly decreasing the attractive and repulsive forces, similarly to simulated annealing.

UMAP: uniform manifold approximation and projection for dimension reduction.
McInnes & Healy (2018) https://arxiv.org/abs/1802.03426

Outline

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies (L. Albergante)

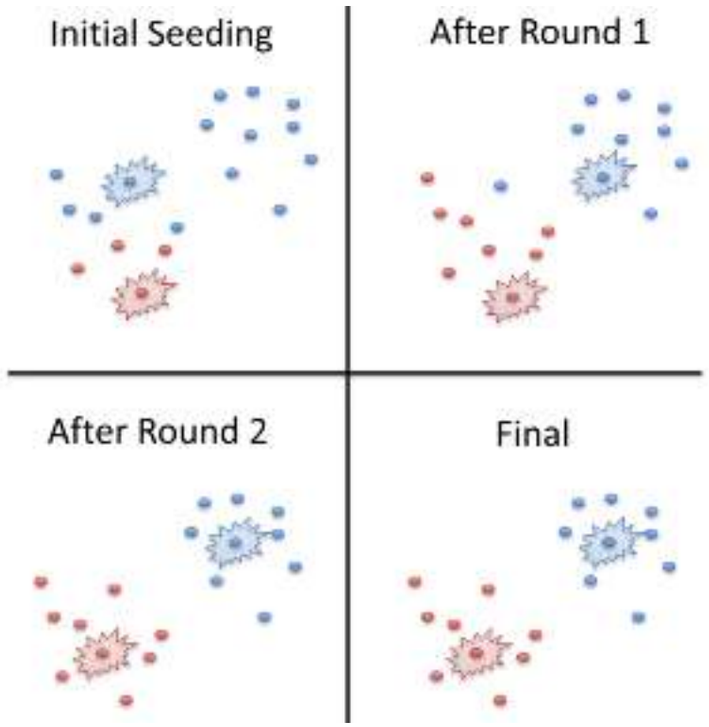
5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

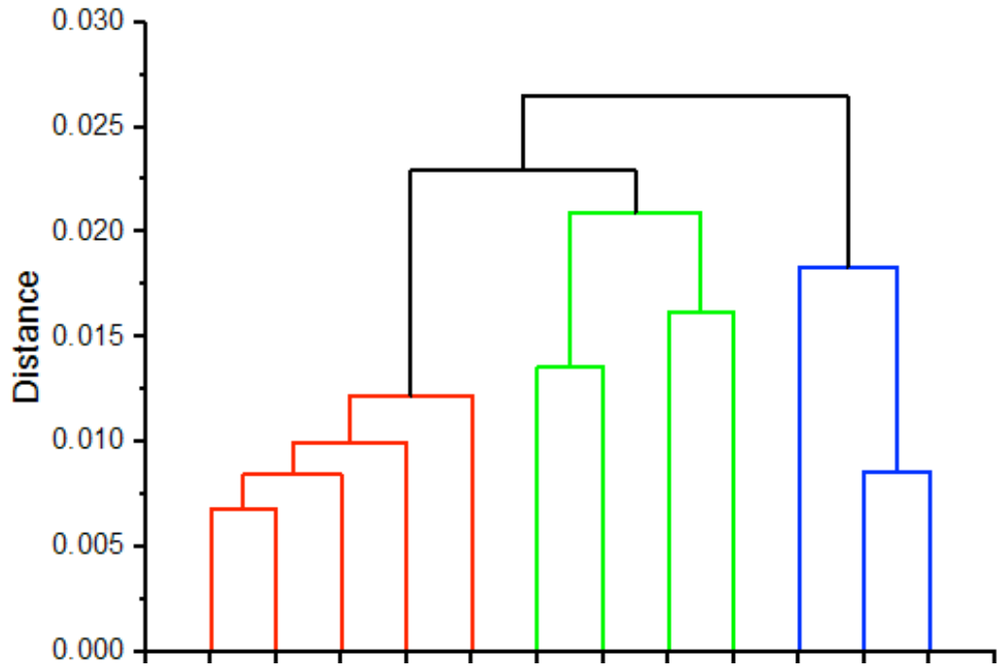
Unsupervised clustering: broad method catergories borrowed for scRNA-seq data analysis

1) K-means based

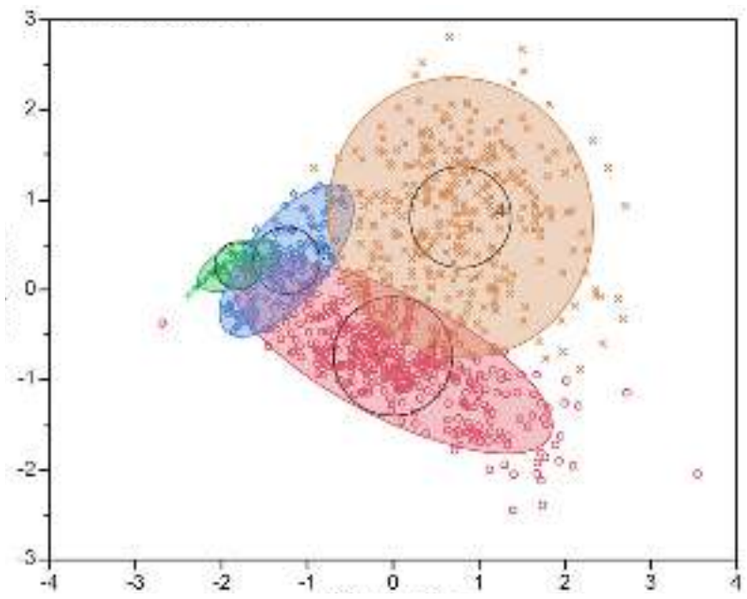


Page et al. BMC Research Notes 2014, 7:829
http://www.biomedcentral.com/1756-0500/7/829

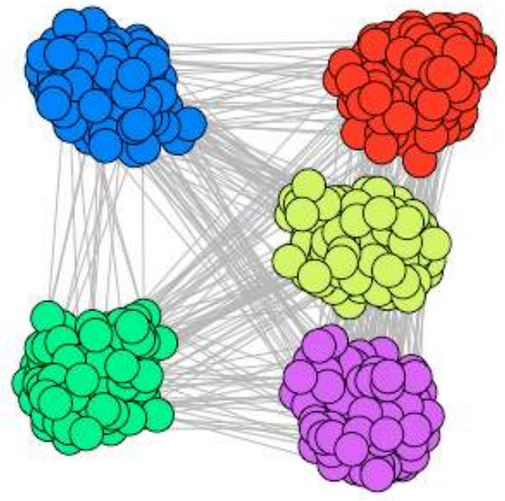
2) Hierarchical clustering



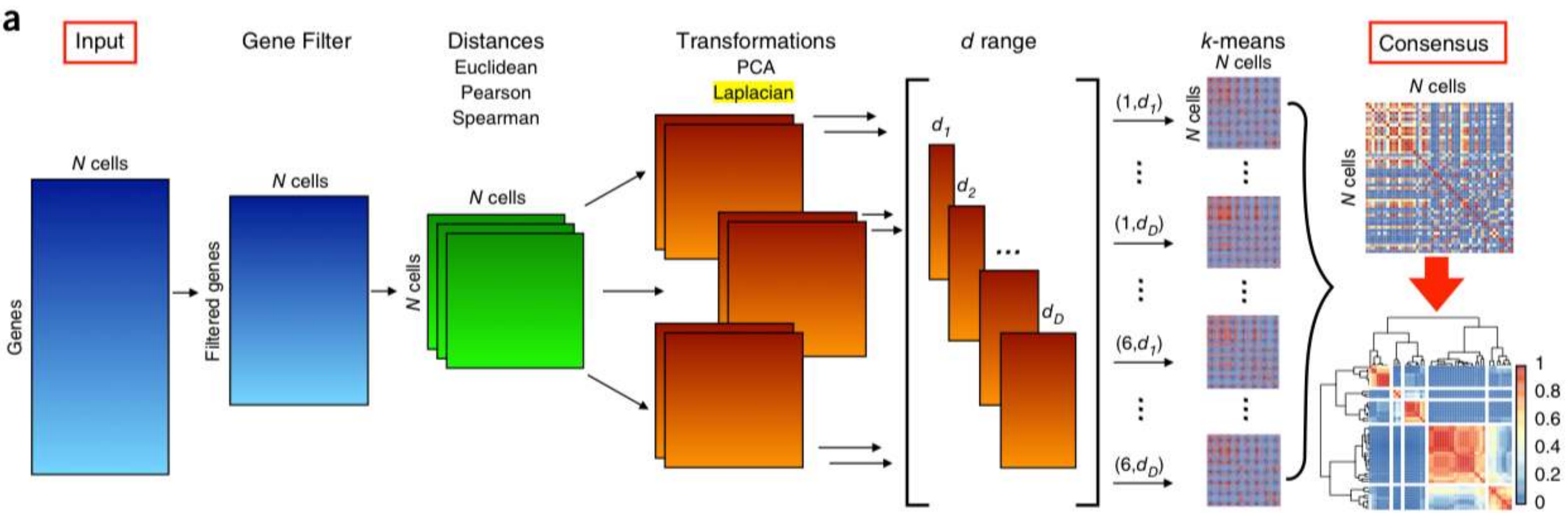
3) Model-based clustering (Mclust)



4) Graph-based clustering (iGraph)

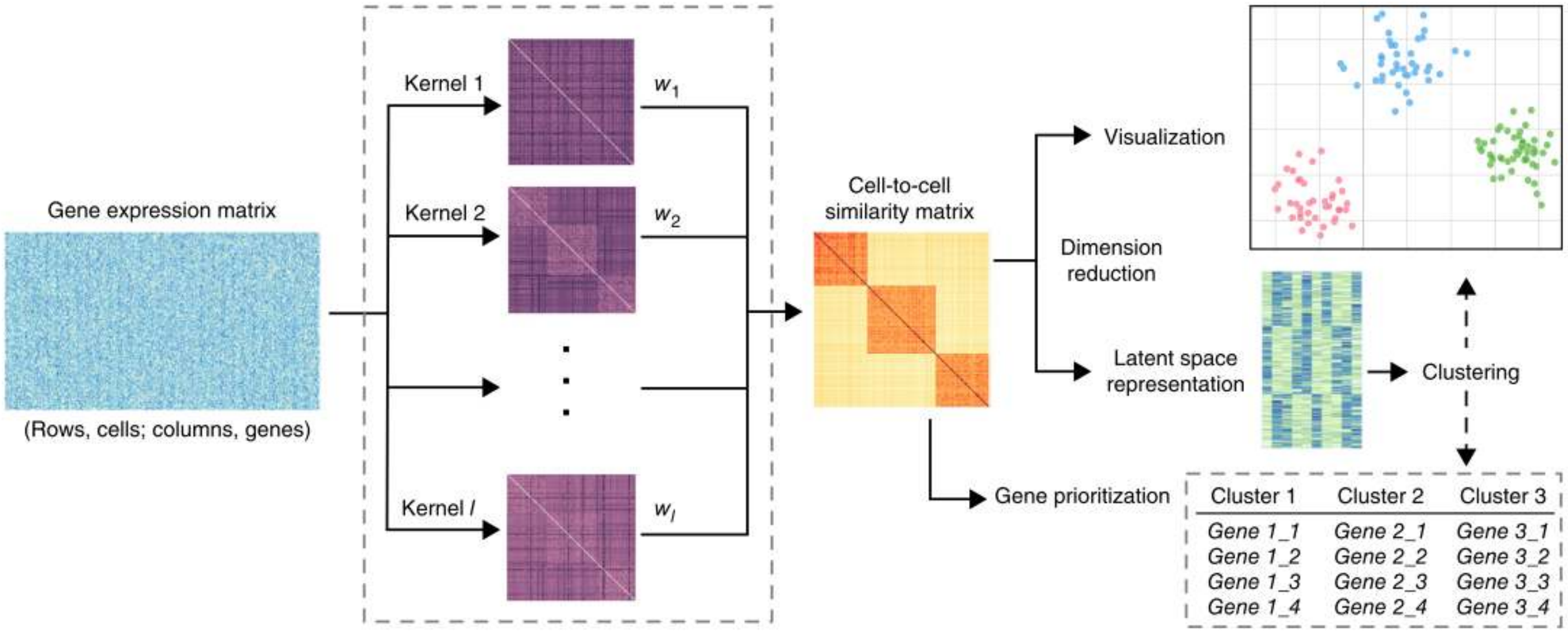


Unsupervised clustering. Examples of dedicated methods for scRNA-seq (I): SC3



SC3: consensus clustering of single-cell RNA-seq data. Kiselev et al. Nature Methods 2017, 14:483–486

Unsupervised clustering. Examples of dedicated methods for scRNA-seq (II): SIMLR



Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Wang et al. Nature Methods (2017) 14,414–416

Kernel construction for SIMLR. In the default implementation of SIMLR, we use Gaussian kernels with various hyperparameters. Gaussian kernels, which generate better empirical performance than do other types of kernels[9,13], take the form

$$K\left(c_i, c_j\right) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} \exp\left(-\frac{\| c_i - c_j \|_2^2}{2\epsilon_{ij}^2}\right) \quad (3)$$

where $||c_i - c_i||_2$ is the Euclidean distance between cell i and cell j. The variance, ϵ_{ij}, can be calculated with different scales:

$$\mu_i = \frac{\sum_{l \in \mathrm{KNN}(c_i)} \|c_i - c_j\|_2}{k}, \varepsilon_{ij} = \frac{\sigma\left(\mu_i + \mu_j\right)}{2} \quad (4)$$

where $\mathrm{KNN}(c_i)$ represents cells that are top k neighbors of the cell i. Hence, each kernel is decided by a pair of parameters (σ, k). We set k = 10,12,14,…,30 and σ = 1.0,1.25,1.5,1.75,2, resulting in 55 different kernels in total. However, we empirically show that our method is insensitive to the number of kernels and choices of parameters (**Supplementary Fig. 23**).

Unsupervised clustering. Other dedicated methods for scRNA-seq (III)

Name	Description	Requirements/deliverables
SNN-Cliq [71]	Clusters cells by identifying and merging sub-graphs (quasi-cliques) in a shared nearest neighbor (SNN) graph; the number of clusters is chosen automatically	Requires a reduced set of genes. Xu and Su [71] recommend using genes with average RPKM >20 and using a log transformation to reduce the effect of outliers. Relies on a valid choice of graph parameters
RaceID [59]	Uses k-means applied to a similarity matrix of Pearson's correlation coefficients for all pairs of cells; the number of clusters is chosen using the gap statistic. Outlier cells are those that cannot be explained by a background model that accounts for technical and biological noise. In a second step, rare subpopulations can be identified and outlier cells may be merged to an outlier cluster; new cluster centers are then computed and each cell is assigned to the most highly correlated cluster center	Requires a reduced set of genes. Grün et al. [59] consider genes with a minimum of five transcripts in at least one cell
SCUBA [73]	Uses k-means to cluster data along a binary tree detailing bifurcation events for time-course data. Models expression regulation along the tree using bifurcation theory	Requires a reduced set of genes. Marco et al. [73] recommend using the 1000 most variable genes that are expressed in at least 30 % of cells
BackSPIN [60]	Iteratively splits a two-way sorted (by both genes and cells) expression matrix into two clusters containing independent cells and genes, for a maximum number of splits. The algorithm has a stopping condition to avoid splitting data that are very homogeneous	Requires a reduced set of genes and the maximum number of splits allowed. Zeisel et al. [60] recommend selecting the top 5000 genes that have the largest residuals after fitting a simple noise model
PAGODA [68]	Allows for both detection and interpretation of the transcriptional heterogeneity within a cell population. A weighted principal component analysis (PCA) is conducted for each gene set; those sets for which the variance explained by the first principal component significantly exceeds genome-wide background expectation are identified. To provide a non-redundant view of heterogeneity structure, principal components from different gene sets showing high similarity are combined to form a single component of heterogeneity	Requires un-normalized gene expression counts (performs internal correction as in SCDE). Uses gene ontology (GO) annotated or user-defined gene sets

Bacher and Kendziorski Genome Biology (2016) 17:63

Benchmark of (some) clustering methods (I)

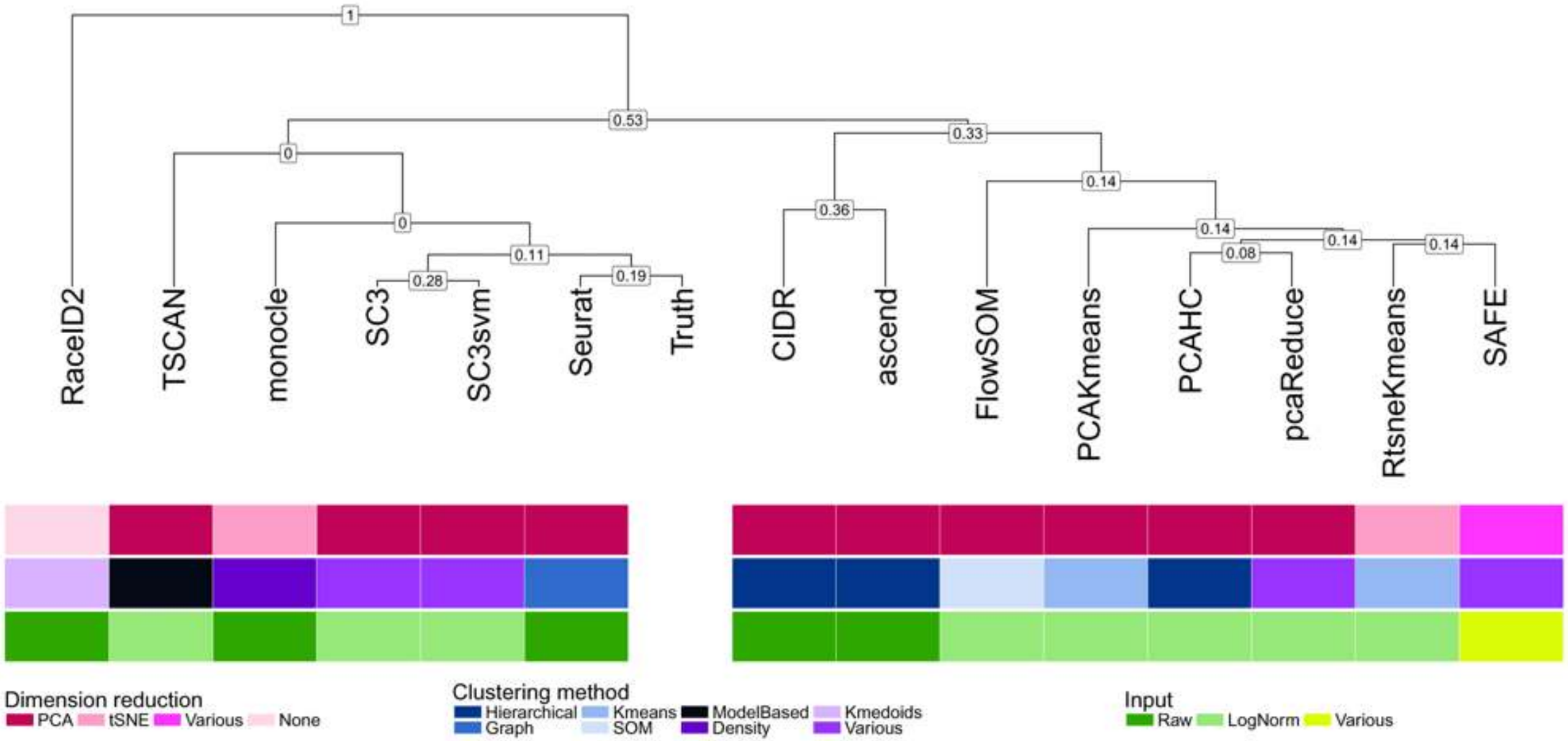


Figure 4. Clustering of the methods based on the average similarity of their partitions across data sets

Duò A, Robinson MD and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2]. F1000Research 2018, 7:1141 (doi: 10.12688/f1000research.15666.2)

Benchmark of (some) clustering methods (II)

... when forcing the methods to cluster with the right number of groups as truth...

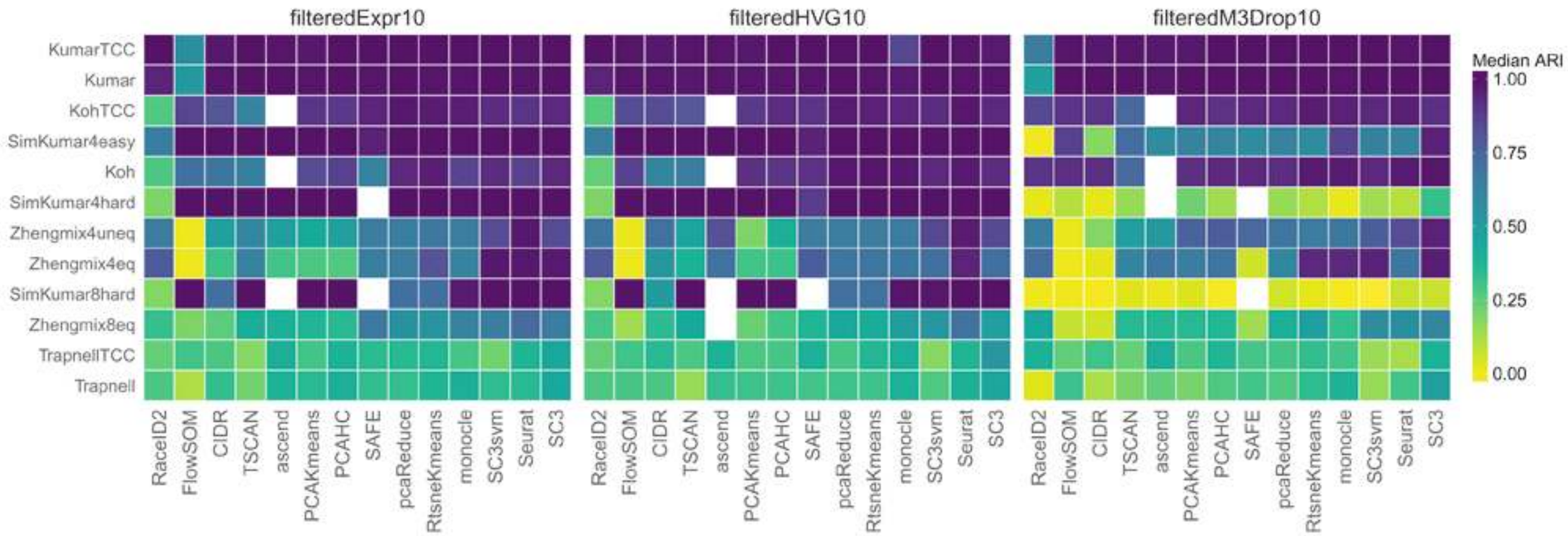


Figure 1. Median ARI scores, representing the agreement between the true partition and the one obtained by each method, when the number of clusters is fixed to the true number.

Each row corresponds to a different data set, each panel to a different gene filtering method, and each column to a different clustering method. The methods and the data sets are ordered by their mean ARI across the filterings and data sets. Some methods failed to return a clustering with the correct number of clusters for certain data sets (indicated by white squares).

Duò A, Robinson MD and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2]. F1000Research 2018, 7:1141 (doi: 10.12688/f1000research.15666.2)

Outline

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

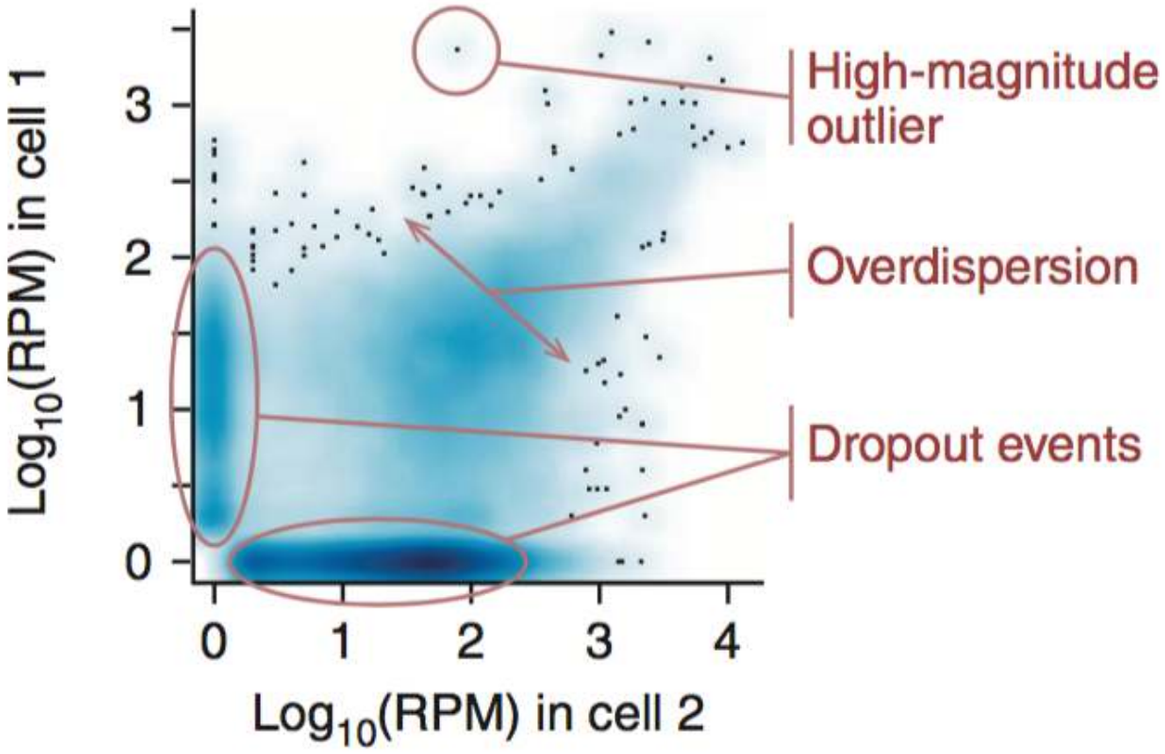
4- **Clustering** / Hierarchies (L. Albergante)

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

Differential expression analysis of single-cell data: The statistical concepts



Kharchenko et al. Nature Methods (2014)

1st. Modeling the measurement of cells as a mixture of two probabilistic processes:

i. The transcript is amplified & detected at a level correlating with its abundance (count data)

Negative binomial distribution

ii. The transcript fails to amplify or is not detected for other reasons (to account for abundance of dropout events)

Poisson distribution

Zero-inflated negative binomial

2nd. Empirical Bayesian framework to regularize model parameters

- helps to improve inference for genes with sparse expression
- based on measurements of individual cells in order to estimate both the likelihood of a gene being expressed at any given average level in each subpopulation and the likelihood of expression fold change between them

SCDE Kharchenko et al. Nature Methods (2014) 11:740

3rd. Extend to Generalized linear modeling (GLM) in order to:

- Accommodate complex experimental designs
- Controlling for covariates (including technical factors) in both the discrete and continuous parts of the model.

MAST. Finak et al. Genome Biology 2015, 16:278

Differential expression analysis: The methods (I)

Identification of differentially expressed genes			
Method	Description	Input	Availability
Designed specifically for single cell RNA-seq data			
SCDE	Bayesian method to compare two groups of single cells, taking into account variability in scRNAseq data due to dropout and amplification biases.	Raw gene expression counts	http://hms-dbmi.github.io/scde/ [68]
MAST	Uses two-part generalized linear model that is adjusted for cellular detection rate.	Normalized gene expression values	https://github.com/RGLab/MAST [66]
M3Drop	Applies Michaelis-Menten modelling of dropouts to identify differential expression.	Raw gene expression counts	https://github.com/tallulandrews/M3Drop [67]
scDD	A Bayesian modelling framework to identify genes that are differentially expressed and/or show a differential number of modes or differential proportion of cells within modes.	Normalized and log-scaled gene expression values	https://github.com/kdkorthauer/scDD [70]
SINCERA	Identifies DE genes based on simple statistical tests such as Wilcoxon rank sum and *t*-tests.	Raw gene expression values	https://research.cchmc.org/pbge/sincera.html [69]
Designed originally for bulk RNA-seq data			
DESeq2	Fits a GLM for each gene, uses shrinkage estimation for dispersions and fold changes, applies a Wald or LR test for significance testing.	Raw gene expression counts	https://bioconductor.org/packages/release/bioc/html/DESeq2.html [64]
EdgeR	Fits a negative binomial distribution for each gene, estimates dispersions by conditional maximum likelihood, identifies differential expression using an exact test adapted for overdispersed data. Supports arbitrary linear models.	Raw gene expression counts	http://bioconductor.org/packages/release/bioc/html/edgeR.html [65]

Computational approaches for interpreting scRNA-seq data.
Rostom et al. FEBS Letters 591 (2017) 2213–2225. doi: 10.1002/1873-3468.12684

Differential expression analysis: The methods (II)

Bias, robustness and scalability in single-cell differential expression analysis.
Soneson & Robinson. Nature Methods 2018. doi:10.1038/nmeth.4612

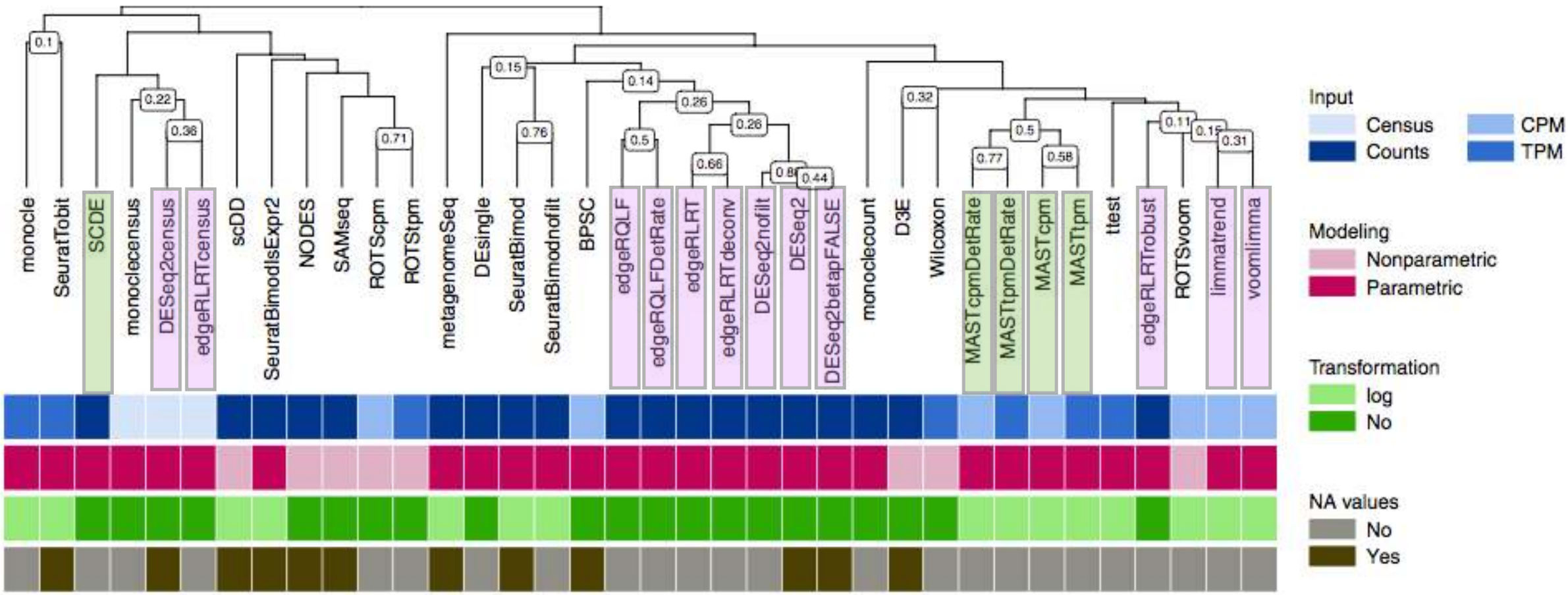
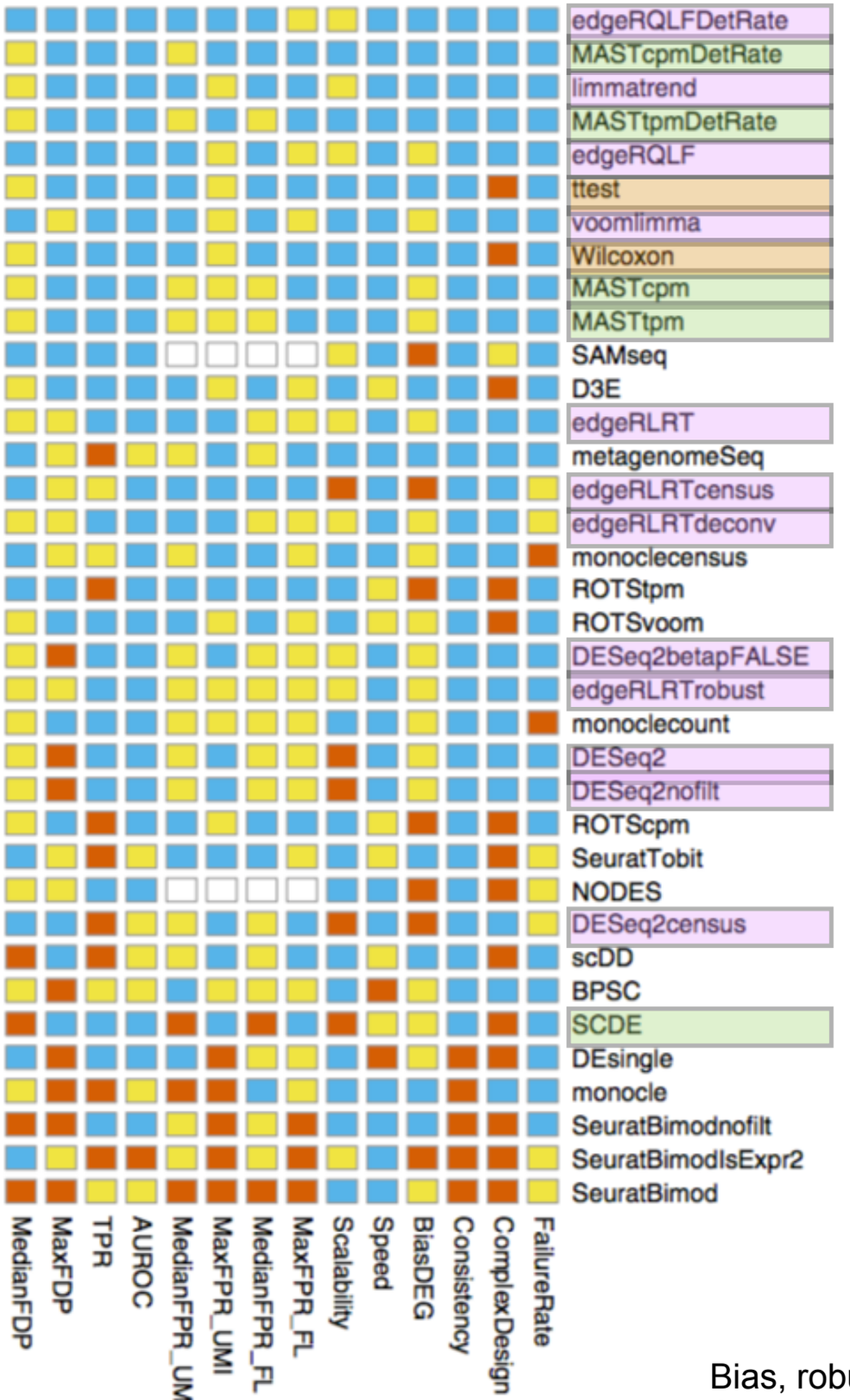


Figure 3 | Average similarities between gene rankings obtained by the evaluated DE methods. The dendrogram was obtained by complete-linkage hierarchical clustering based on the matrix of average AUCC values across all data sets. The labels of the internal nodes represent their stability across data sets (fraction of instances where they are observed). Only nodes with stability scores of at least 0.1 are labeled. Colored boxes represent method characteristics.

Differential expression analysis: The benchmark



Bayesian 3-component model adapted to single-cell data

Methods borrowed from bulk-RNA-seq

Naïve approaches

Bias, robustness and scalability in single-cell differential expression analysis. Soneson & Robinson. Nature Methods 2018. doi:10.1038/nmeth.4612

Outline

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies (L. Albergante)

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

A note on statistical robustness

Always check the robustness of your results upon random perturbation of the input data

Alternative strategies

a) Random subsampling of genes

b) Random subsampling of cells (when the number of cells is high)

c) Random substitution with in-silico pseudoreplicates (when the number of cells is low):
***Sincell package**, Julià, Telenti & Rausell, Bioinformatics 2015:*

Generation of in silico cell pseudoreplicates by perturbing the expression of all genes according to a model of stochastic variability

Model 1

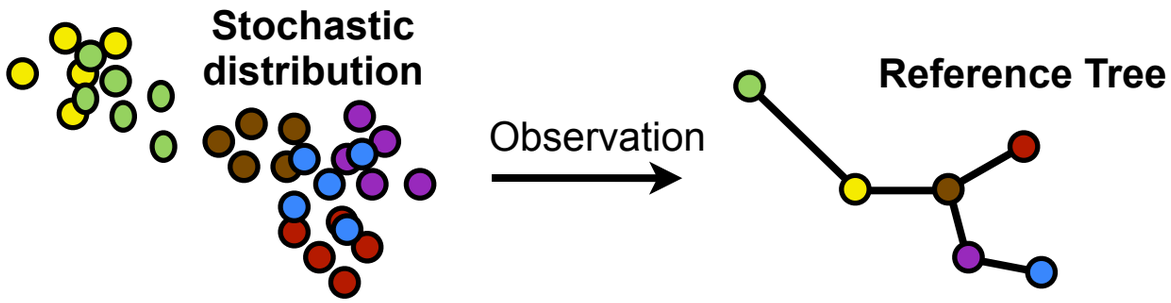
- Most genes follow a lognormal distribution log(x)~N(m,v) of mean m and variance v, with a third parameter alpha describing the proportion of cells where transcript expression was detected above a given threshold level (Shalek et al., 2014)

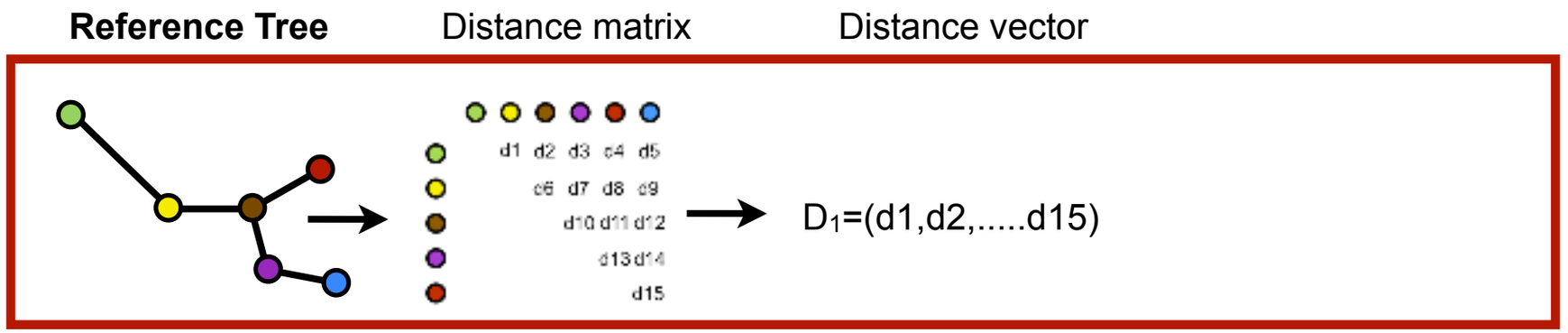
Model 2

- Mean-variance relationship (Anders and Huber, 2010; Brennecke et al., 2013)
- Genes are assigned to classes according to the deciles of mean they belong to.
- For a given gene g, a variance v is randomly chosen from the set of variances within the class of the gene.
- A random value drawn from a uniform distribution U(0,v) of mean zero and variance v is added to the expression value of a gene g in a cell c

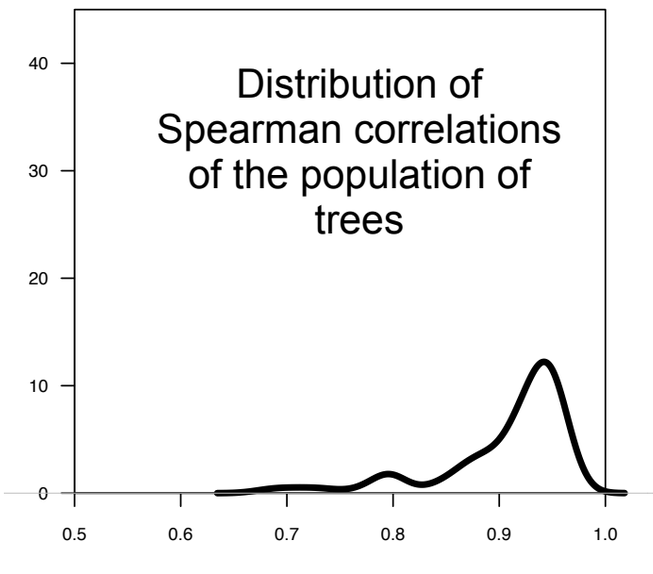
antonio.rausell@institutimagine.org

Sincell's statistical support and functional association tests





Population of trees drawn by either a, b, or c:
a) Random subsampling of genes
b) Random substitution with in-silico replicates: Negative Binomial, log-Normal
c) Restricting to the expression levels of a gene list: GO, Reactome, MSigDB etc



antonio.rausell@institutimagine.org

Further reading

Benchmarks evaluating each of the analytical steps:

Assessment of Single Cell RNA-Seq **Normalization** Methods.
http://www.g3journal.org/content/7/7/2039.long

Evaluation of tools for **highly variable gene discovery** from single-cell RNA-seq data
https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby011/4898116

A systematic performance evaluation of **clustering** methods for single-cell RNA-seq data
https://f1000research.com/articles/7-1141/v2

Comparison of **clustering** tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data
https://f1000research.com/articles/7-1297/v1

Bias, robustness and scalability in single-cell **differential expression** analysis.
https://www.nature.com/articles/nmeth.4612

A comparison of single-cell **trajectory inference** methods: towards more accurate and robust tools
https://www.biorxiv.org/content/10.1101/276907v1

A test metric for assessing single-cell RNA-seq **batch correction** (KBet)
https://www.nature.com/articles/s41592-018-0254-1

scRNA-seq mixology: towards better benchmarking of single cell RNA-seq **protocols** and analysis methods
https://www.biorxiv.org/content/10.1101/433102v2

Recommended online courses

Complete course on Single-cell RNA-seq data analysis from U. Cambridge
http://hemberg-lab.github.io/scRNA.seq.course/index.html

Bioinformatics Training channel on YouTube
https://www.youtube.com/channel/UCsc6r6UKxb2qRcDQPix2L5A

A step-by-step workflow for low-level analysis of single-cell RNA-seq data
https://f1000research.com/articles/5-2122/v1

Single-Cell Workshop 2014: RNA-seq, Harvard
http://pklab.med.harvard.edu/scw2014/

Recommended recent reviews on bioinformatics methods for scRNA-seq

Identifying cell populations with scRNASeq. Andrews & Hemberg. Molecular Aspects of Medicine 59 (2018) 114-122. dx.doi.org/10.1016/j.mam.2017.07.002

Computational approaches for interpreting scRNA-seq data. Rostom et al. FEBS Letters 591 (2017) 2213–2225. doi: 10.1002/1873-3468.12684

Design and computational analysis of single-cell RNA-sequencing experiments. Bacher and Kendziorski. Genome Biology (2016) 17:63. doi: 10.1186/s13059-016-0927-y

Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Poirion et al. Frontiers in Genetics 2016. 7: 163. doi: 10.3389/fgene.2016.00163