2017 EDITION

BORDEAUX SUMMER SCHOOLS

An experience of excellence

PCA & t-SNE Visualize Single-Cell RNA-seq datasets

Statistical analysis of big data in systems immunology

Hadrien LORENZO, PhD student, SISTM team

Bordeaux school of public health ISPED Institut de Santé Publique d'Epidémiologie et de Développement

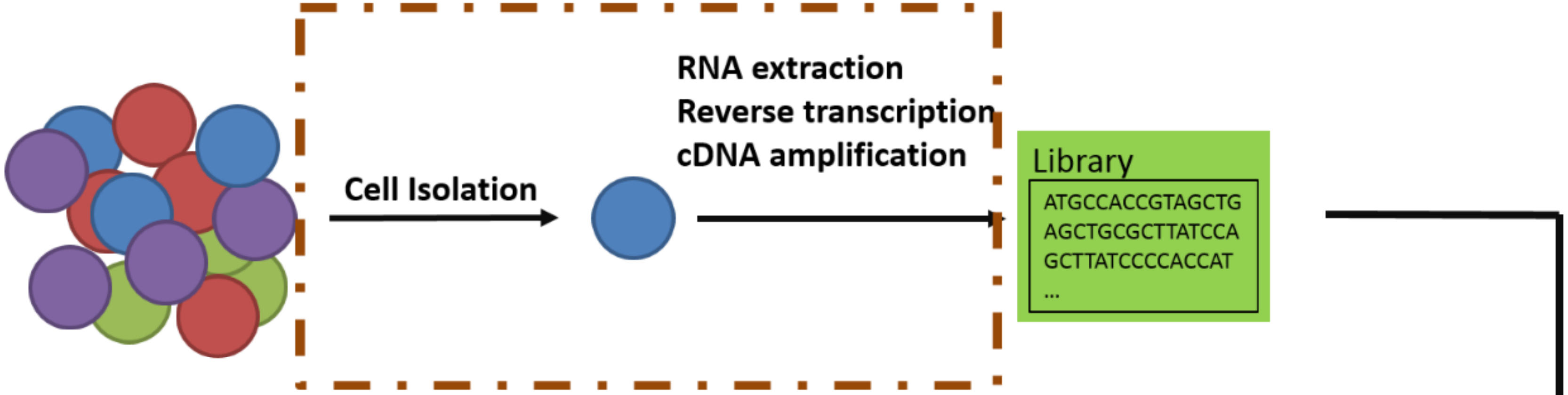
BORDEAUX POPULATION HEALTH Research Center - U1219

SISTM / Statistics in Systems biology and Translational Medicine

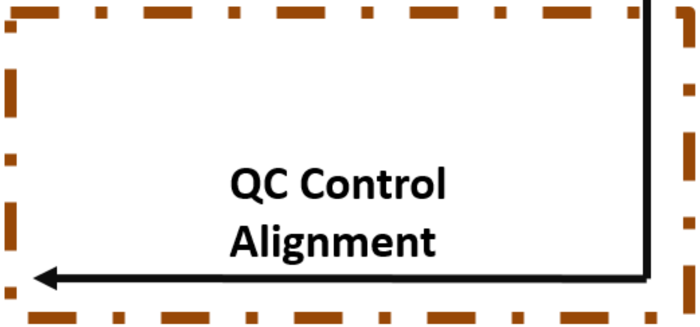
Inria INVENTEURS DU MONDE NUMÉRIQUE

Single-cell RNA-Seq Overview

Single-Cell RNA-Seq theory & methods



Single-Cell RNA-Seq Methods, see [Ziegenhain et al., 2017]



Gene 1	Gene 2	...	Gene g
0	5	...	20

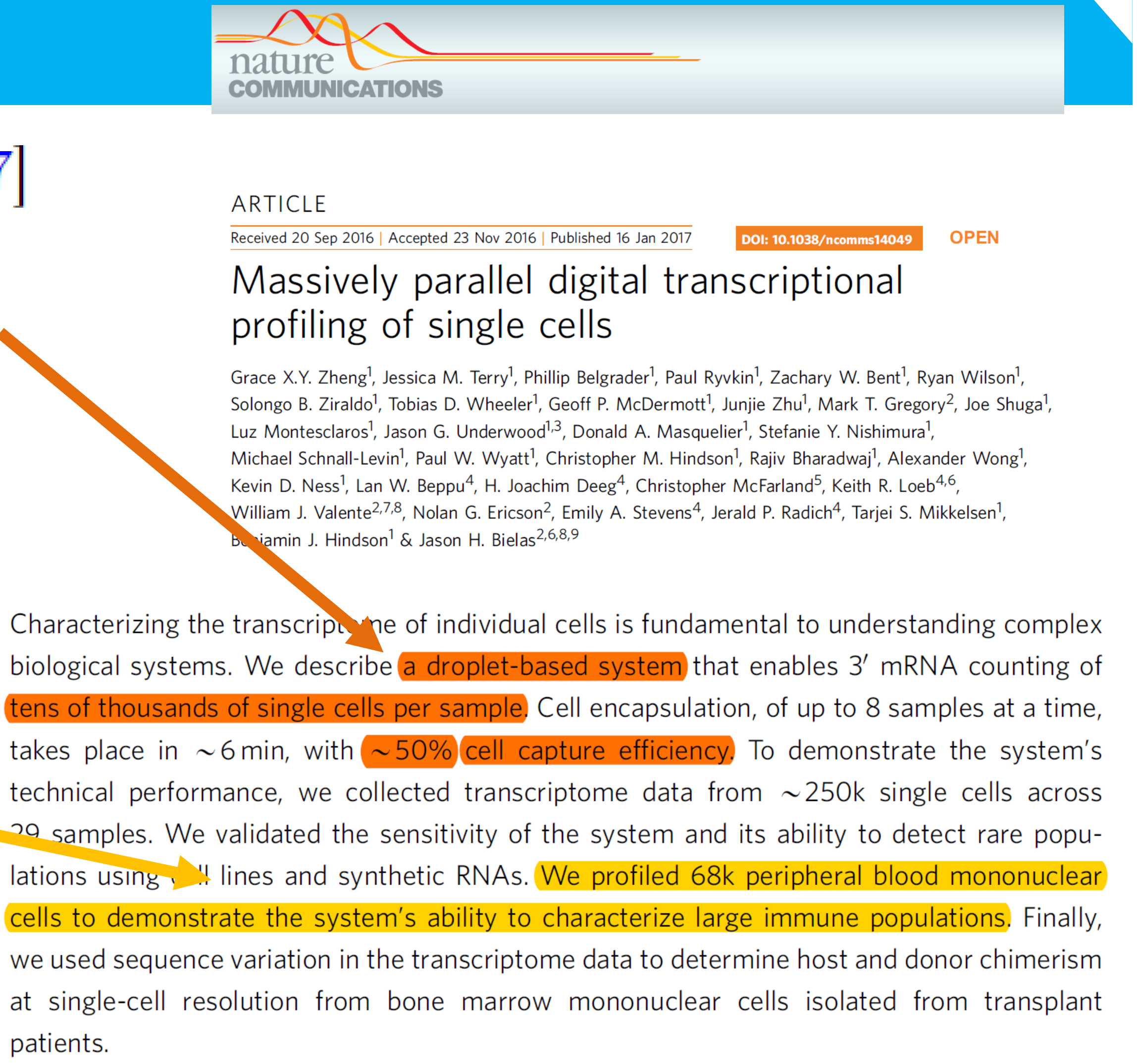
Most often : STAR, see [Dobin et al., 2013]
Fast and accurate

An handy dataset

see [Zheng et al., 2017]

Technical work, technology : **10x**

PBMC Single-Cell RNA sequences :



nature COMMUNICATIONS

ARTICLE

Received 20 Sep 2016 | Accepted 23 Nov 2016 | Published 16 Jan 2017

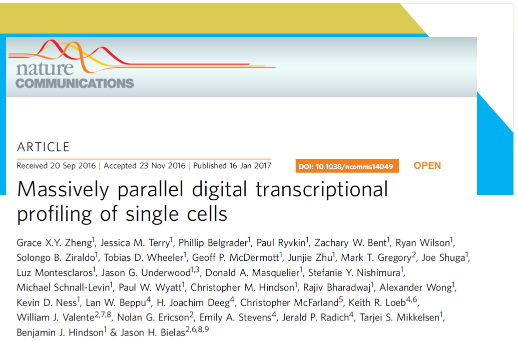
DOI: 10.1038/ncomms14049 OPEN

Massively parallel digital transcriptional profiling of single cells

Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1], Stefanie Y. Nishimura[1], Michael Schnall-Levin[1], Paul W. Wyatt[1], Christopher M. Hindson[1], Rajiv Bharadwaj[1], Alexander Wong[1], Kevin D. Ness[1], Lan W. Beppu[4], H. Joachim Deeg[4], Christopher McFarland[5], Keith R. Loeb[4,6], William J. Valente[2,7,8], Nolan G. Ericson[2], Emily A. Stevens[4], Jerald P. Radich[4], Tarjei S. Mikkelsen[1], Benjamin J. Hindson[1] & Jason H. Bielas[2,6,8,9]

Characterizing the transcriptome of individual cells is fundamental to understanding complex biological systems. We describe a droplet-based system that enables 3′ mRNA counting of tens of thousands of single cells per sample. Cell encapsulation, of up to 8 samples at a time, takes place in ∼6 min, with ∼50% cell capture efficiency. To demonstrate the system's technical performance, we collected transcriptome data from ∼250k single cells across 29 samples. We validated the sensitivity of the system and its ability to detect rare populations using cell lines and synthetic RNAs. We profiled 68k peripheral blood mononuclear cells to demonstrate the system's ability to characterize large immune populations. Finally, we used sequence variation in the transcriptome data to determine host and donor chimerism at single-cell resolution from bone marrow mononuclear cells isolated from transplant patients.

An handy dataset



nature COMMUNICATIONS

ARTICLE

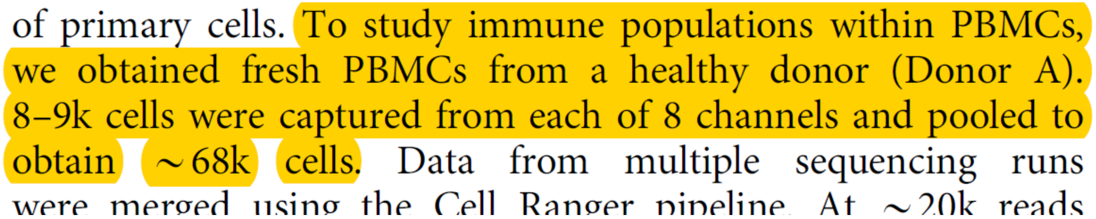
Received 20 Sep 2016 | Accepted 23 Nov 2016 | Published 16 Jan 2017 DOI: 10.1038/ncomms14049 OPEN

Massively parallel digital transcriptional profiling of single cells

Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1], Stefanie Y. Nishimura[1], Michael Schnall-Levin[1], Paul W. Wyatt[1], Christopher M. Hindson[1], Rajiv Bharadwaj[1], Alexander Wong[1], Kevin D. Ness[1], Lan W. Beppu[4], H. Joachim Deeg[4], Christopher McFarland[5], Keith R. Loeb[4,6], William J. Valente[2,7,8], Nolan G. Ericson[2], Emily A. Stevens[4], Jerald P. Radich[4], Tarjei S. Mikkelsen[1], Benjamin J. Hindson[1] & Jason H. Bielas[2,6,8,9]

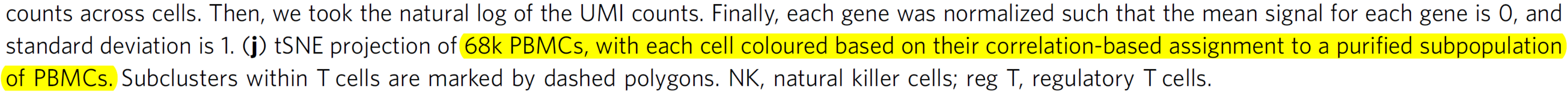
see [Zheng et al., 2017]

- Immune population from **1 donor** :

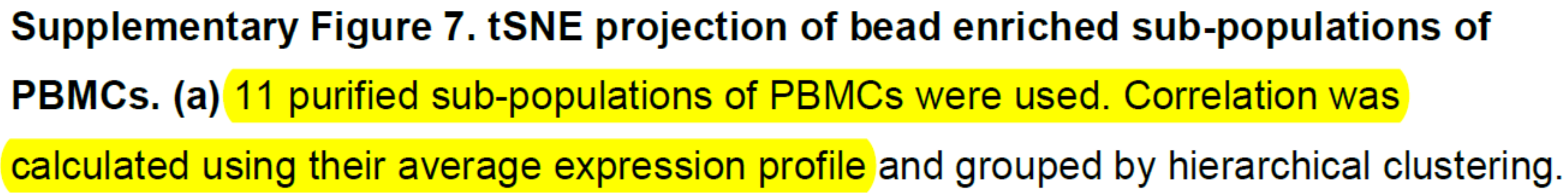


of primary cells. To study immune populations within PBMCs, we obtained fresh PBMCs from a healthy donor (Donor A). 8–9k cells were captured from each of 8 channels and pooled to obtain ~68k cells. Data from multiple sequencing runs were merged using the Cell Ranger pipeline. At ~20k reads

- … Cells labellised with **purified subpopulation of PBMCs** :



counts across cells. Then, we took the natural log of the UMI counts. Finally, each gene was normalized such that the mean signal for each gene is 0, and standard deviation is 1. (**j**) tSNE projection of 68k PBMCs, with each cell coloured based on their correlation-based assignment to a purified subpopulation of PBMCs. Subclusters within T cells are marked by dashed polygons. NK, natural killer cells; reg T, regulatory T cells.



Supplementary Figure 7. tSNE projection of bead enriched sub-populations of PBMCs. (a) 11 purified sub-populations of PBMCs were used. Correlation was calculated using their average expression profile and grouped by hierarchical clustering.

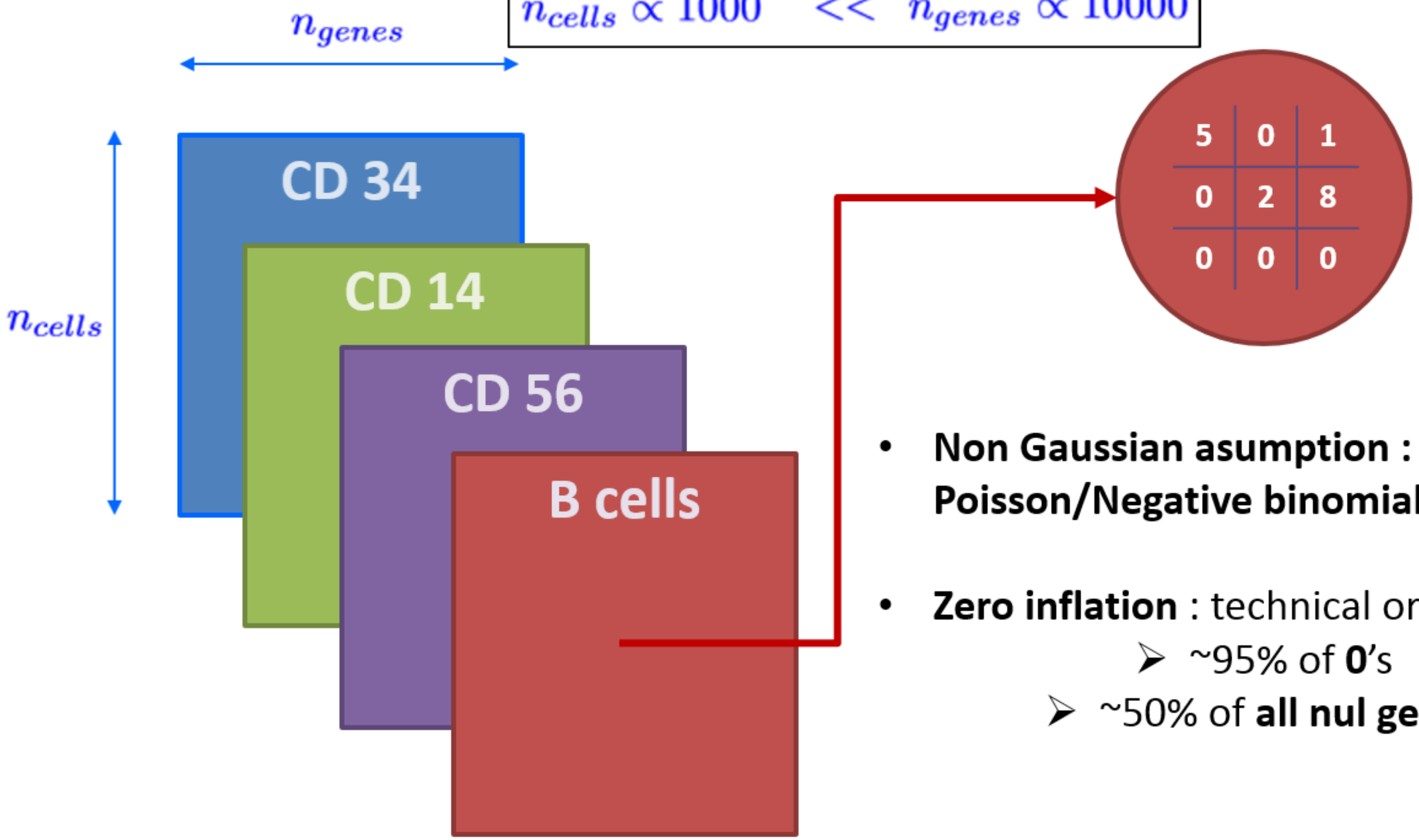
- In this course :
  - 4 populations
  - 300 cells per population

CD 34 CD 14 CD 56 B cells

Single-cell RNA-Seq Dataset Structure

Single-Cell RNA-Seq dataset structures

$n_{cells} \propto 1000 \quad << \quad n_{genes} \propto 10000$

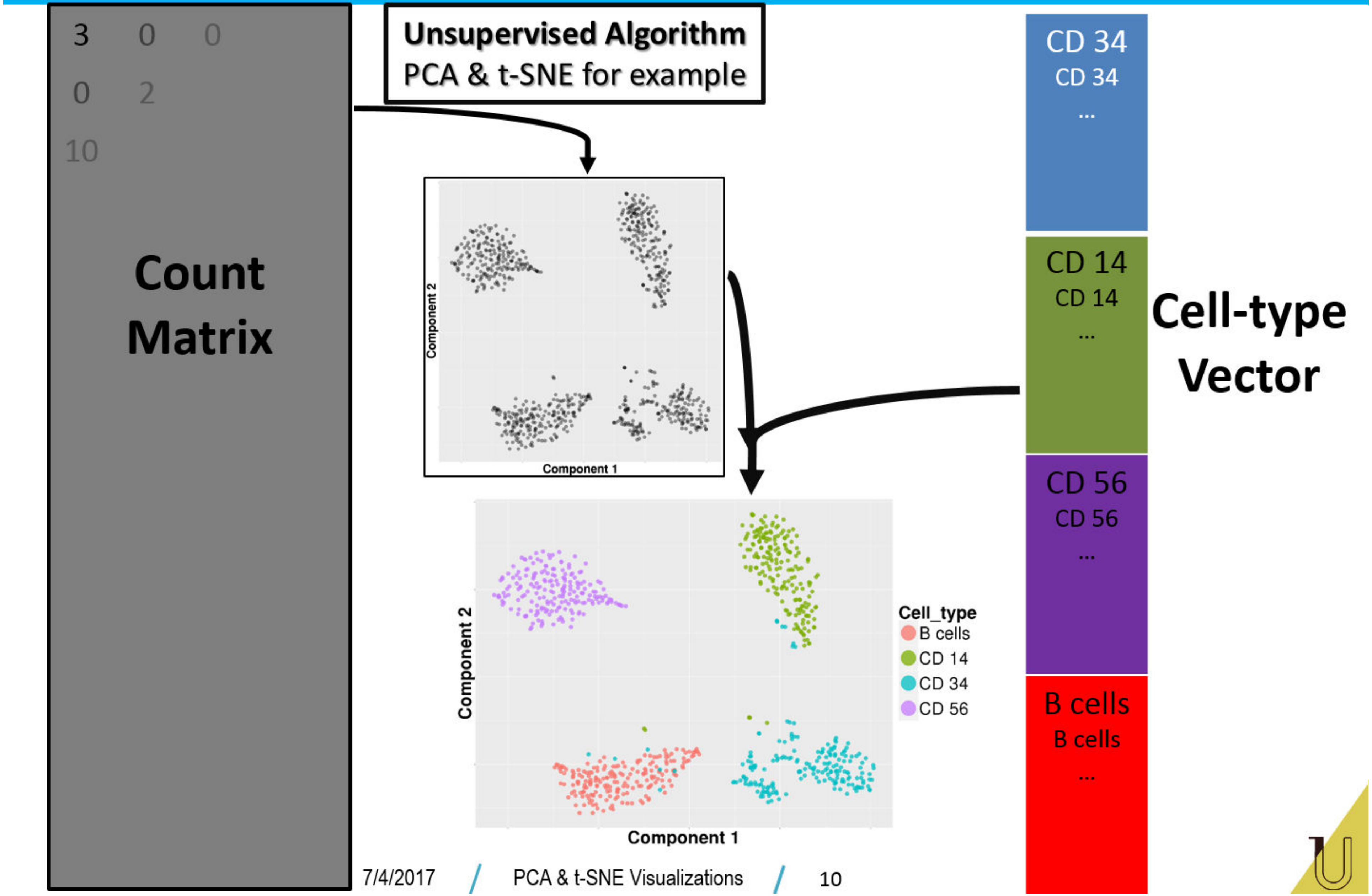


5	0	1
0	2	8
0	0	0

- **Non Gaussian asumption : Poisson/Negative binomial**

- **Zero inflation** : technical or biological ?
  - ➢ ~95% of **0**'s
  - ➢ ~50% of **all nul genes**

Single-Cell RNA-Seq dataset structures



3 0 0

0 2

10

Count Matrix

Unsupervised Algorithm

PCA & t-SNE for example

CD 34
CD 34
...

CD 14
CD 14
...

CD 56
CD 56
...

B cells
B cells
...

Cell-type Vector

Single-cell RNA-Seq : Prepare datasets

Prepare datasets

Not specific to single Cell RNA Seq data

Remove batch effects :

Based on clinical design (if any…)

Think PCA ;)

Parametric : **edgeR, DESeq2** see [Robinson et al., 2010][Love et al., 2014]

Non parametric : **Voom + Limma** see [Law et al., 2014]

Longitudinal & model Free: **tcgsaseq** see [Agniel and Hejblum, 2017]

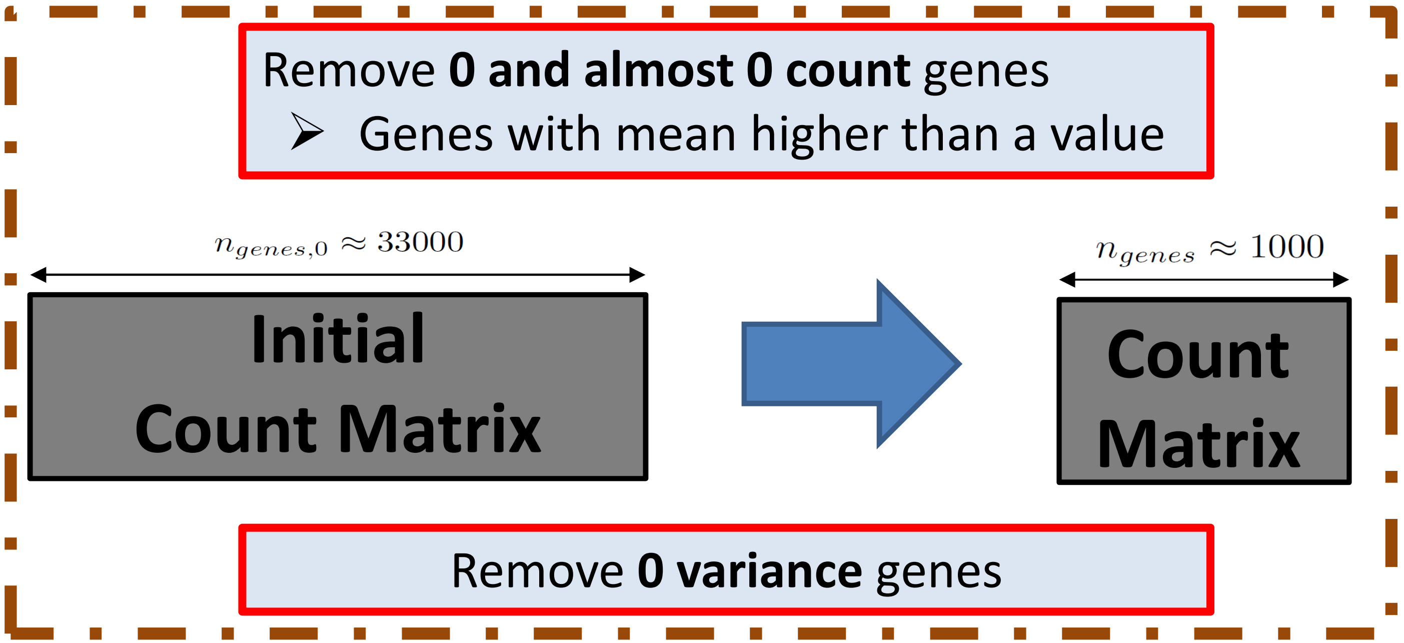
PCA & t-SNE

➢ **Gaussian models**

Final transformation

➢ **log-CPM** for example

Prepare datasets



Remove **0 and almost 0 count** genes

- ➢ Genes with mean higher than a value

$n_{genes,0} \approx 33000$

Initial Count Matrix

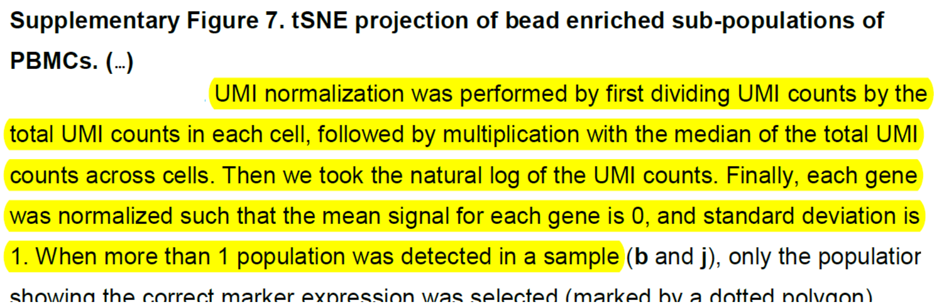
$n_{genes} \approx 1000$

Count Matrix

Remove **0 variance** genes

Specific to single Cell RNA Seq data

In the paper :



Supplementary Figure 7. tSNE projection of bead enriched sub-populations of PBMCs. (...)

UMI normalization was performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells. Then we took the natural log of the UMI counts. Finally, each gene was normalized such that the mean signal for each gene is 0, and standard deviation is 1. When more than 1 population was detected in a sample (**b** and **j**), only the populatior showing the correct marker expression was selected (marked by a dotted polygon).

$$n_{genes} = 200$$

$$n_{cells} = 1200$$

Visualize Single-Cell RNA-Seq

- **Objectives**
  - Appealing visualizations
  - Interpretable results
- **Biological challenges**
  - Low number of replicates : a few participants
  - Samples sensible to lab conditions : long chain of manipulations
- **Mathematical constraints**
  - Positive counts data with zero inflated values
  - High dimensionnal settings : thousands of genes
  - Unsupervised analysis : no cell labels

Ven diagramm as a first analysis tool

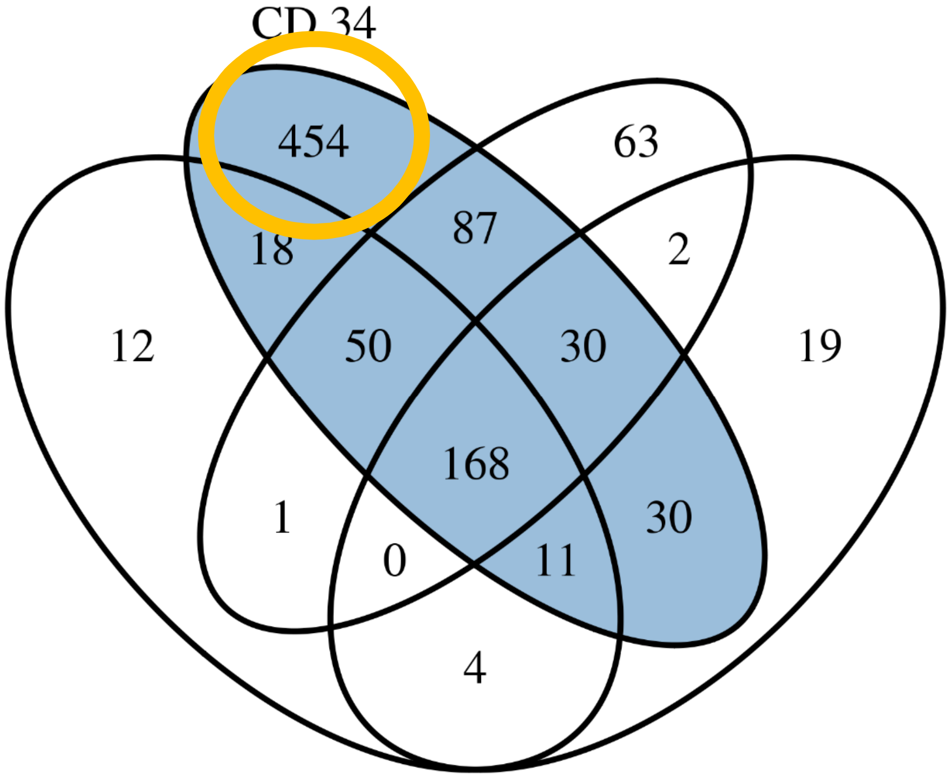
Explore dataset low counts

Per cell type :
Genes with count means higher than a value

Cell-Type	# Genes selected
B-cells	264
CD 14	264
CD 34	848
CD 56	401

$$n_{CD\ 34} >> n_{others}$$

? Normal ?
Check deeper : **Venn Diagram**



Huge count!
? Multi modalities ?

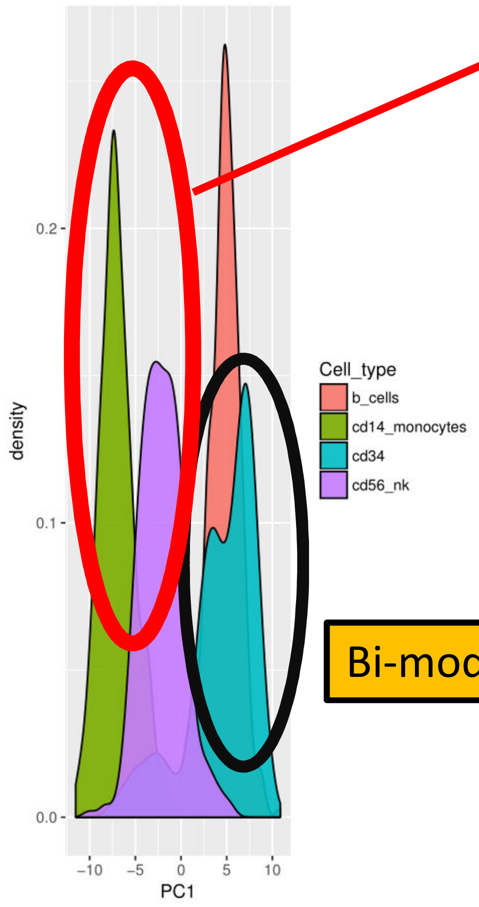
PCA for scRNA-Seq data visualization & interpretation

First Principal Component

Heterogeneity in CD 14 and CD 56

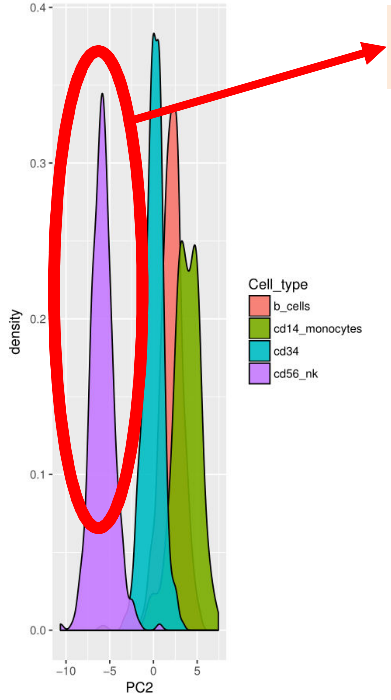
First principal component :

✓ **Heterogeneity in CD 14 and CD 56**



Bi-modality

Second Principal Component



Unimodality

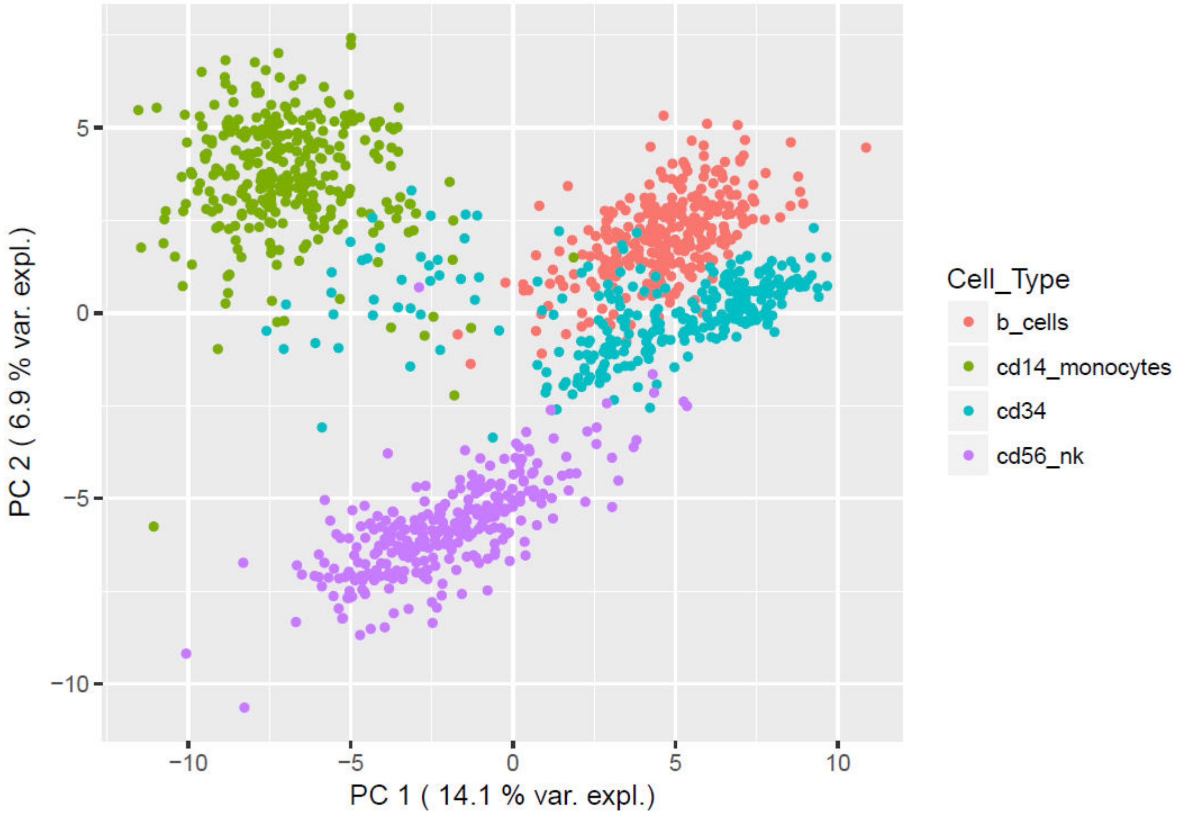
Second principal component :

✓ **Variability in CD 56**

… where to stop ?

PCA for scRNA-Seq data visualization & interpretation

Two first axes interpretable
Independantly
Biological meaning

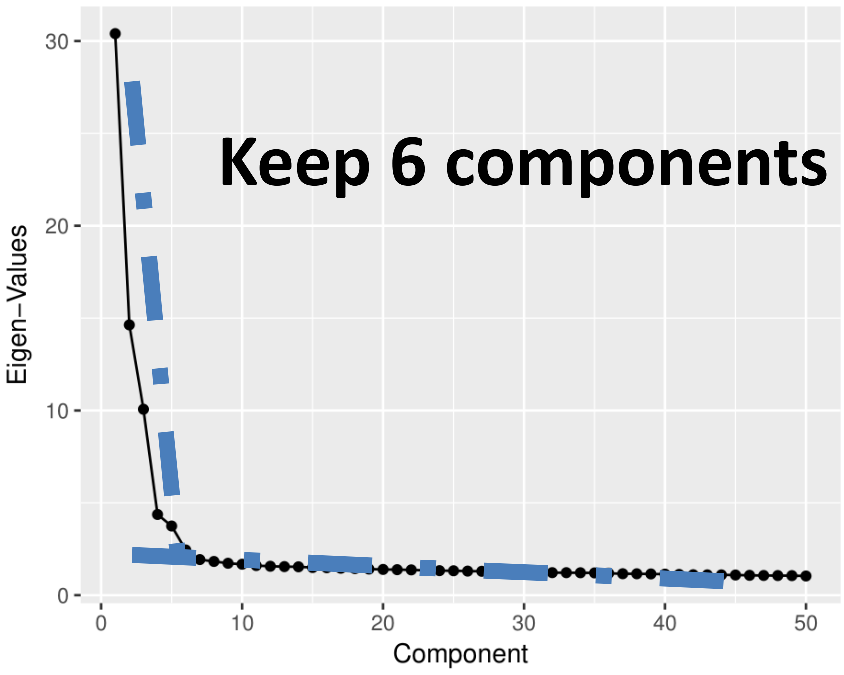


PCA for scRNA-Seq data : where to stop ?

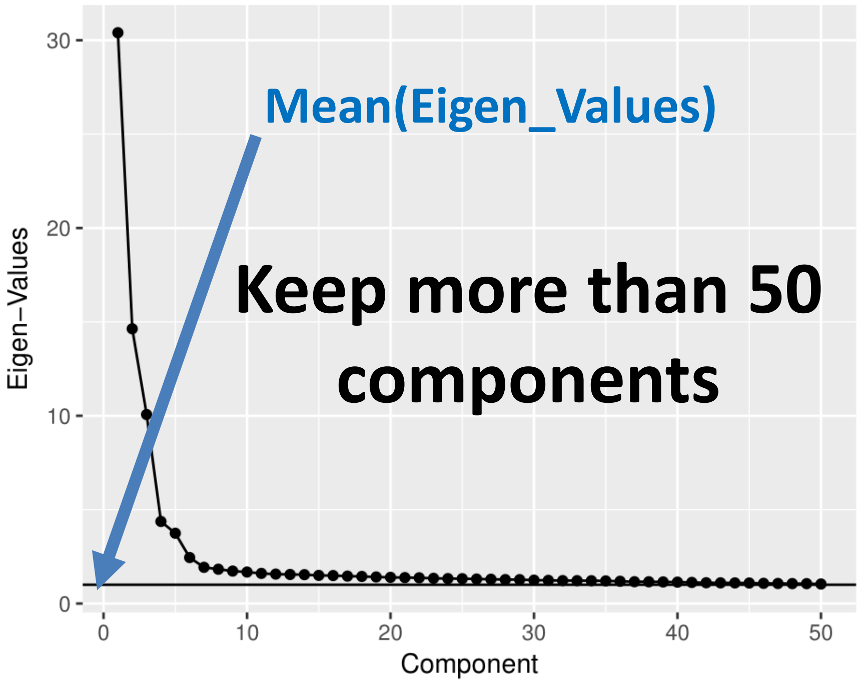
2 usual criterions

Elbow criterion

Kaiser criterion



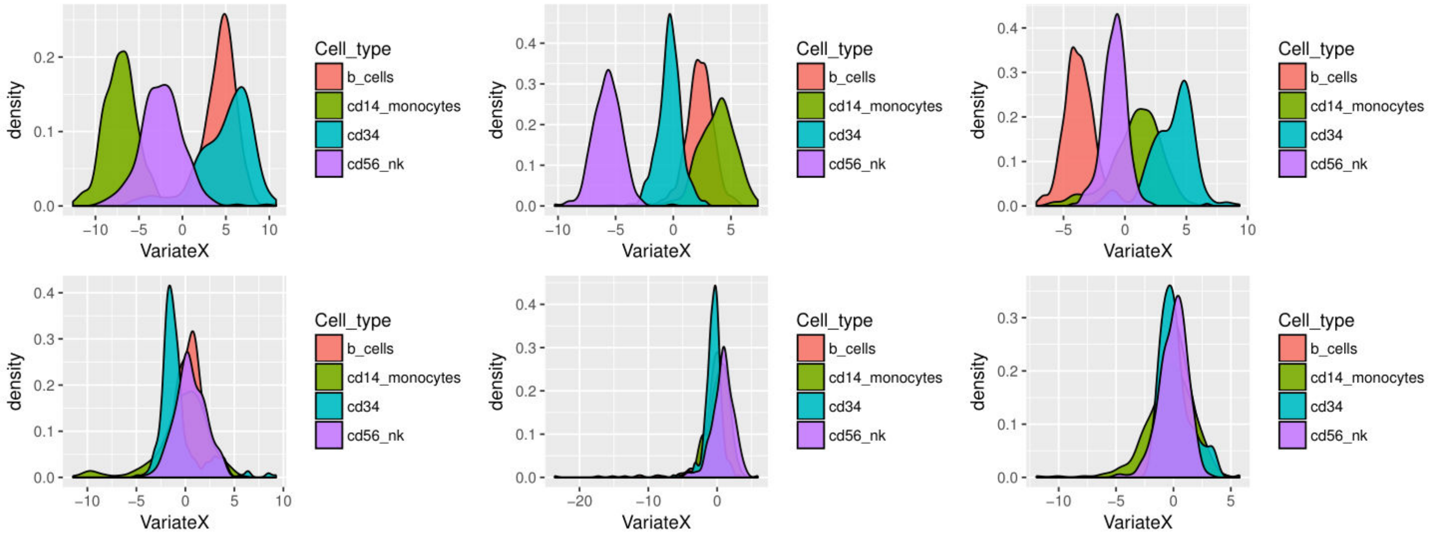
Keep 6 components



Keep more than 50 components

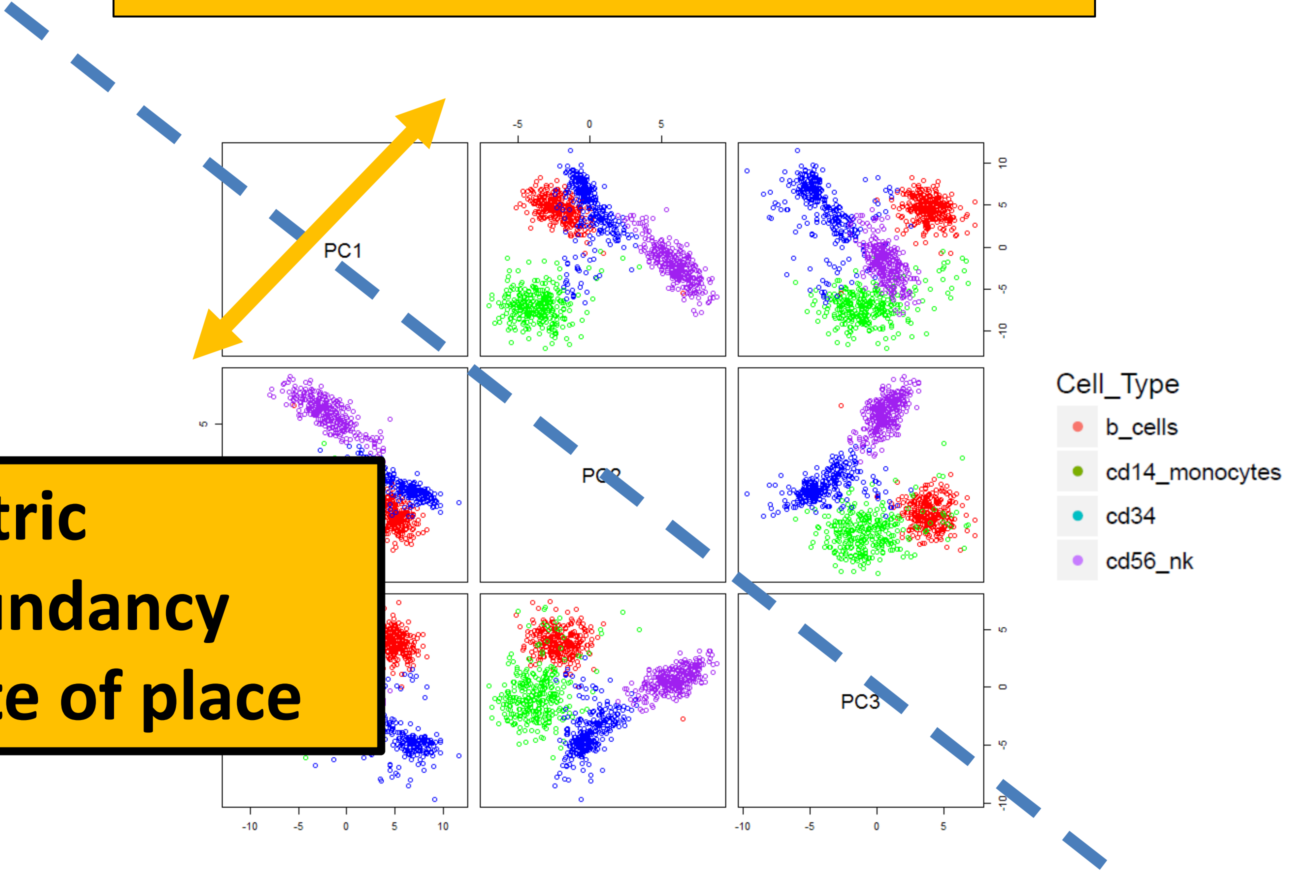
PCA for scRNA-Seq data : where to stop ?

How to visualize so many components ? (6)



PCA for scRNA-Seq data : where to stop ?

3 components
Still hard to communicate!



Symmetric
→ Redundancy
→ Waste of place

PCA for scRNA-Seq data : In a nutschell ?

PCA powers

Interpretability of each axis (independantly)

~~**Stopping criterion (mutual)**~~

Maybe not for communication

Other way
Fix the number of dimensions where to project the data and then build those

2/3d

t-SNE

see [Maaten and Hinton, 2008]

PCA for scRNA-Seq data : In a nutschell ?

PCA powers

Interpretability of each axis (independantly)

~~Stopping criterion (mutual)~~

Maybe not for communication

Other way
Fix the number of dimensions where to project the data and then build those

2/3d

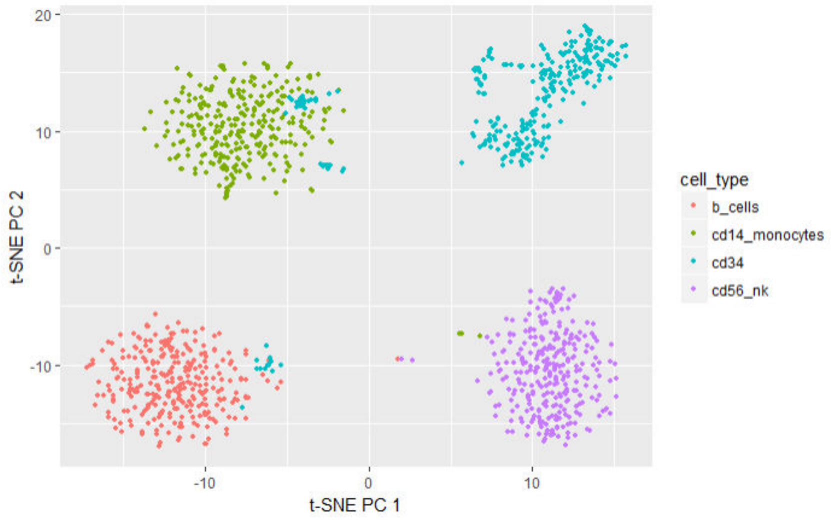
t-SNE

see [Maaten and Hinton, 2008]

t-SNE : A visualization object for communication

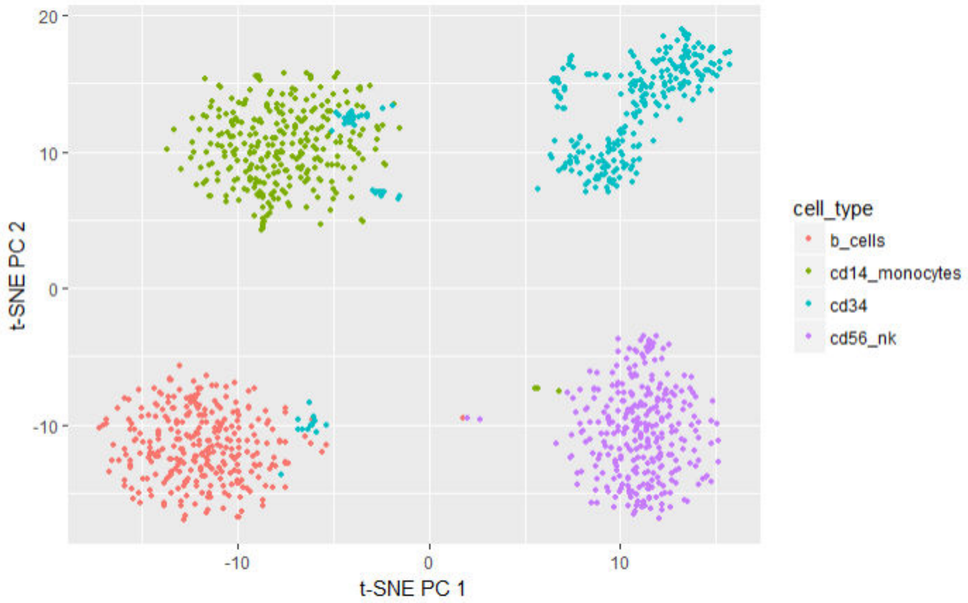
- Fix the number of dimensions : 2
- Random initialization
- Iterative process
- Tune a few parameters :
  - PCA Pre-process : t-SNE faster!
  - Perplexity : Balance between local and global aspects.

see [Wattenberg et al., 2016]



t-SNE : A visualization object for communication

t-SNE permits :

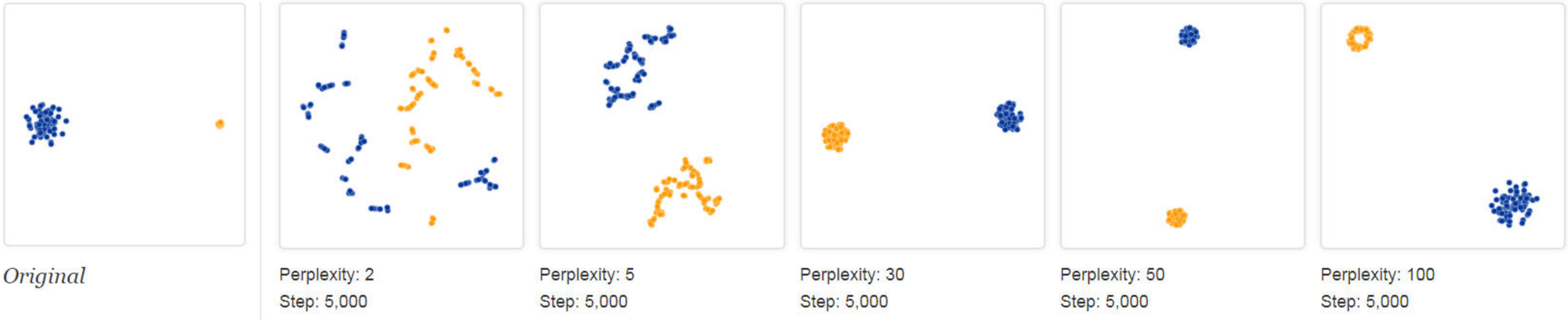


- ✓ Find clusters with non linear bounders
- ✓ Interpret some cells as badly classified
- ✓ Give an appealing 2-d visualization

t-SNE must not be interpreted too easily!

t-SNE : A visualization object for communication

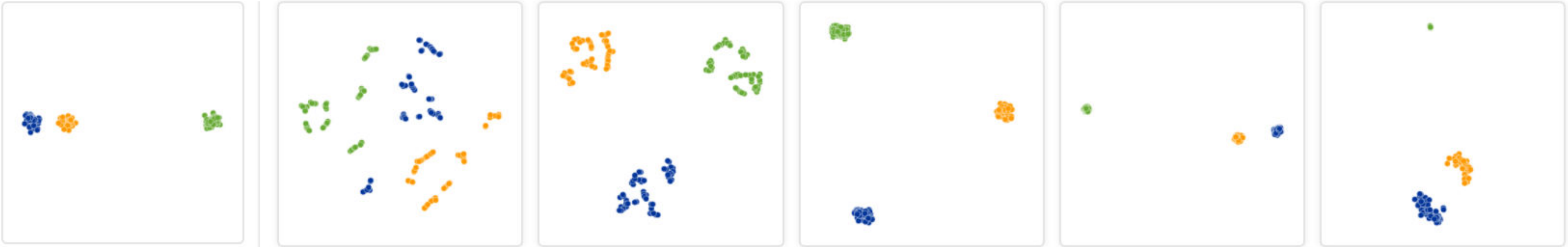
t-SNE cluster sizes mean nothing!



[Wattenberg et al., 2016]

t-SNE : A visualization object for communication

t-SNE between cluster distances mean nothing!



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

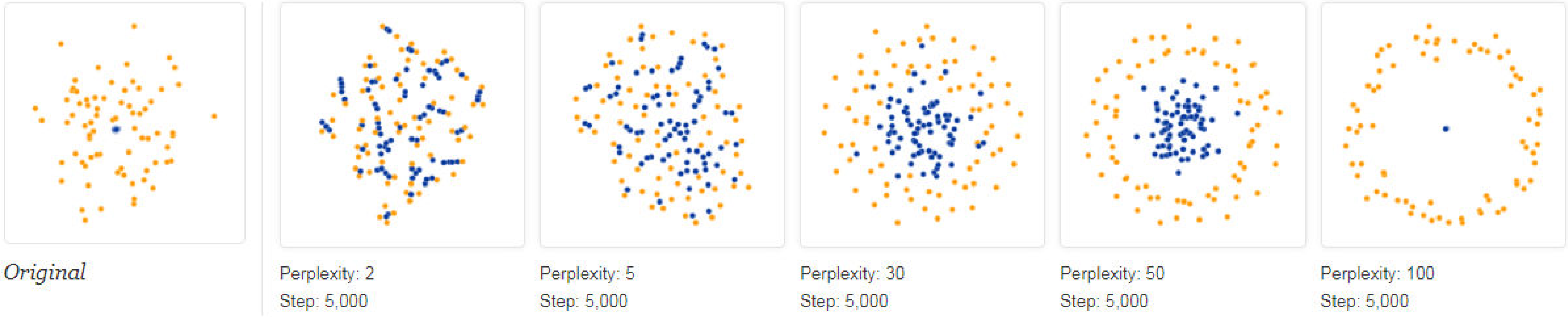
Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

[Wattenberg et al., 2016]

t-SNE : A visualization object for communication

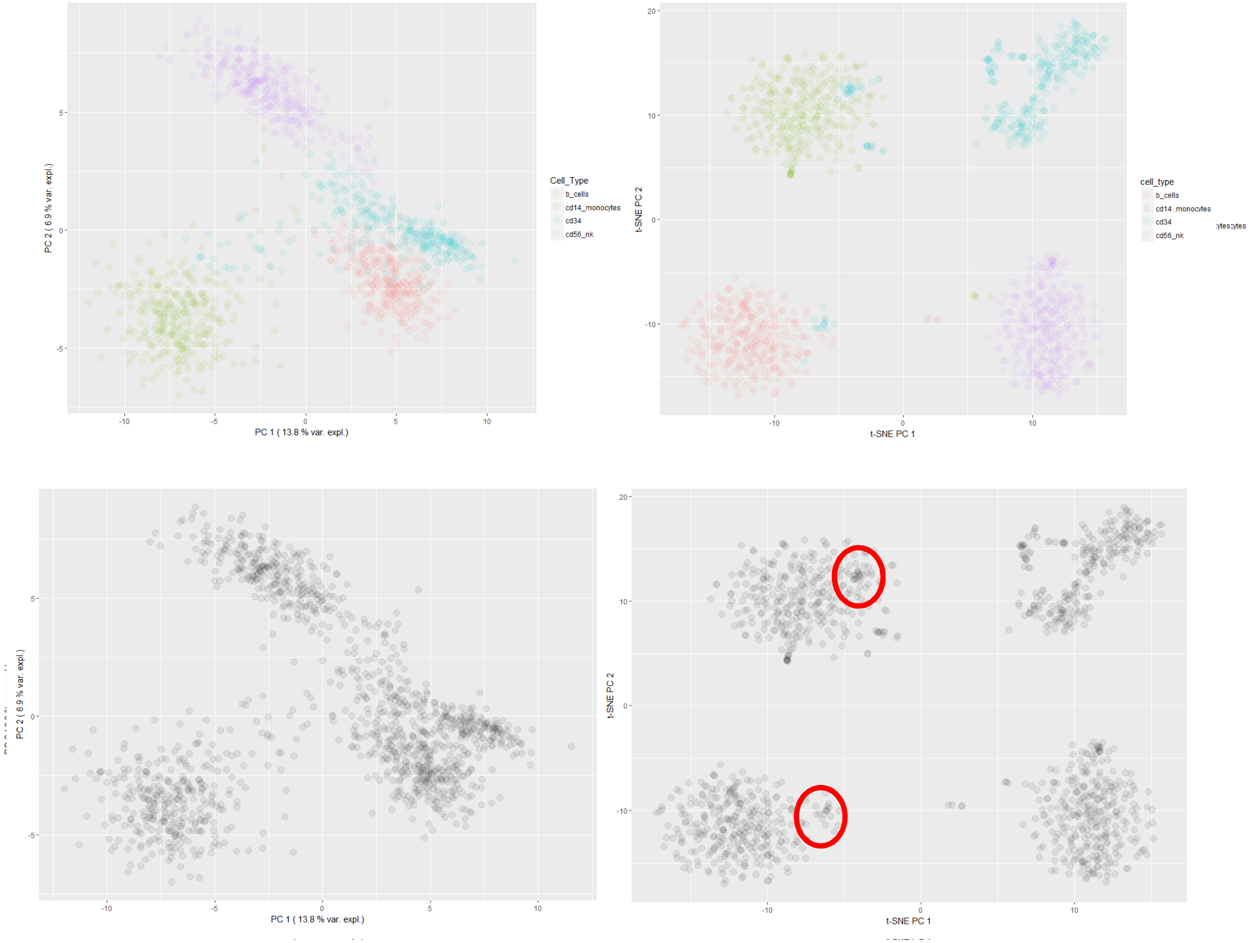
t-SNE shapes may be just fantasy!



[Wattenberg et al., 2016]

… mainly due to previous problems

t-SNE : Not conserving the distances : a matter of transparency!



Transparency parameter :

$$\alpha = 0.07$$

Crowding problem

Conclusion

université de BORDEAUX

Conclusion

Two methods of representation :

- PCA :
  - ➢ Interpretable from A to Z
  - ➢ Not strong enough for too complex datasets
  - ➢ Difficult for communication
- t-SNE :
  - ➢ Flexible : fill 2 dimensions
  - ➢ Strong to non linear relashionships
  - ➢ Relative distances/positions not interpretable
  - ➢ Crowding effect not solved

Watch out naive conclusions !

Thank you!

References

[Agniel and Hejblum, 2017] Agniel, D. and Hejblum, B. P. (2017). Variance component score test for time-course gene set analysis of longitudinal rna-seq data. *Biostatistics*, page kxx005.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[Law et al., 2014] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Wattenberg et al., 2016] Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*, 1(10):e2.

[Zheng et al., 2017] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049.

[Ziegenhain et al., 2017] Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643.