

# Single-cell analysis 101: challenges and opportunities

Ming 'Tommy' Tang

Director of computational Biology at Immunitas

Twitter: tangming2005

<https://divingintogeneticsandgenomics.rbind.io/>

03/29/2022 ABRF2022



# Who am I ?



**Ming Tang**

crazyhottommy

Director of Computational Biology at Immunitas working on single-cell RNAseq. Care about reproducible research and open science

Edit profile

🔗 1.6k followers · 39 following

🏢 Immunitas

📍 Waltham, MA

✉ tangming2005@gmail.com

🌐 <http://divingintogeneticsandgenomics.r...>

Achievements

📖 Overview

📁 Repositories 139

📁 Projects

📦 Packages

★ Stars 526

crazyhottommy / README.md

Hi there 🙌

- I am a computational biologist working on (single-cell) genomics, epigenomics and transcriptomics.
- I use R primary for data wrangling and visualization in the tidyverse ecosystem;
- I use python for writing Snakemake workflows and reformatting data;
- I am a unix geek learning shell tricks almost every month; I care about reproducible research and open science.

Learn more about me at my [blog](#)

Pinned

Customize your pins

📁 **ChIP-seq-analysis** Public

ChIP-seq analysis notes from Ming Tang

🐍 Python ☆ 559 📄 266

📁 **RNA-seq-analysis** Public

RNAseq analysis notes from Ming Tang

🐍 Python ☆ 654 📄 253

📁 **getting-started-with-genomics-tools-and-resources**

Public

Unix, R and python tools for genomics and data science

🐍 Shell ☆ 722 📄 239

📁 **pyflow-ChIPseq** Public

a snakemake pipeline to process ChIP-seq files from GEO or in-house

🐍 Python ☆ 88 📄 37

📁 **scRNAseq-analysis-notes** Public

scRNAseq analysis notes from Ming Tang

☆ 328 📄 98

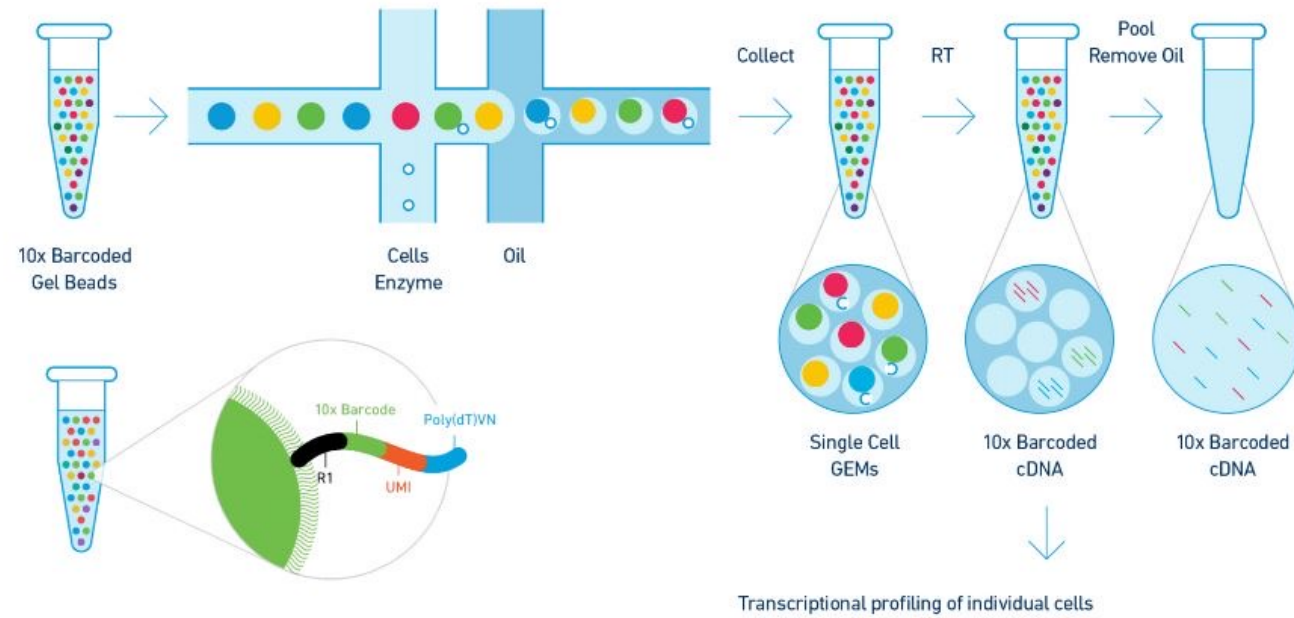
📁 **bioinformatics-one-liners** Public

Bioinformatics one liners from Ming Tang

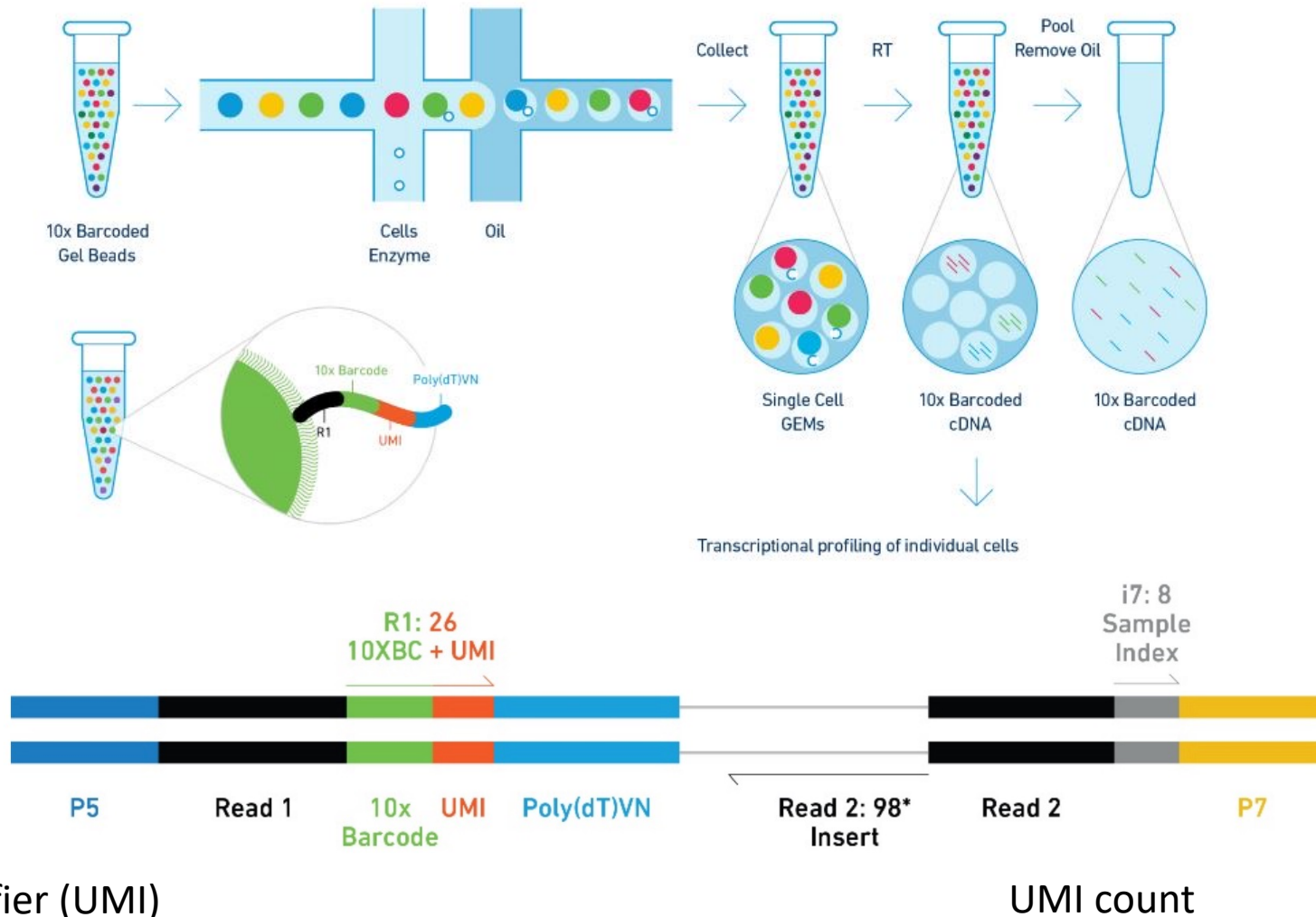
☆ 358 📄 96

<https://github.com/crazyhottommy>

# 10x single-cell gene expression solution



# 10x single-cell gene expression solution



16 bp cell barcode  
10bp (V2) vs 12bp (V3)  
Unique Molecular identifier (UMI)

UMI count



# Different scRNA-seq Techniques

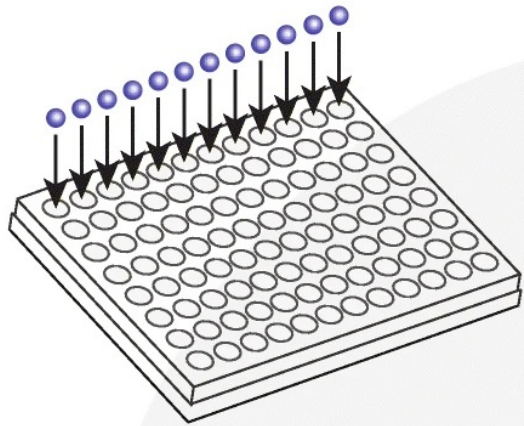
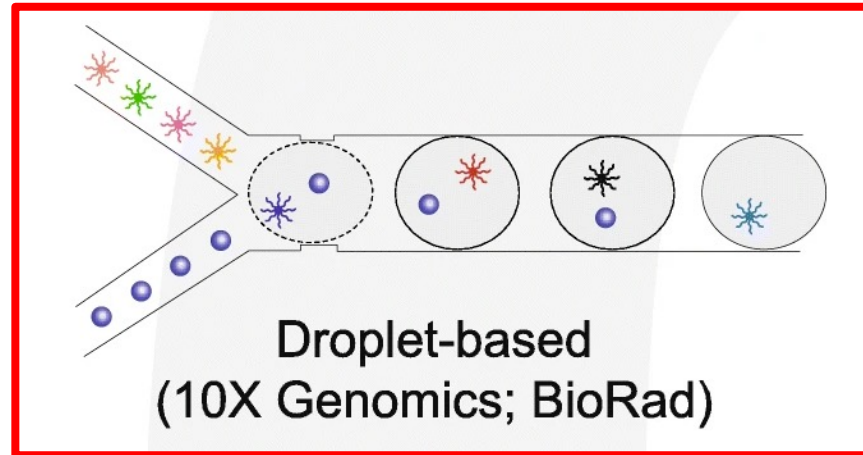
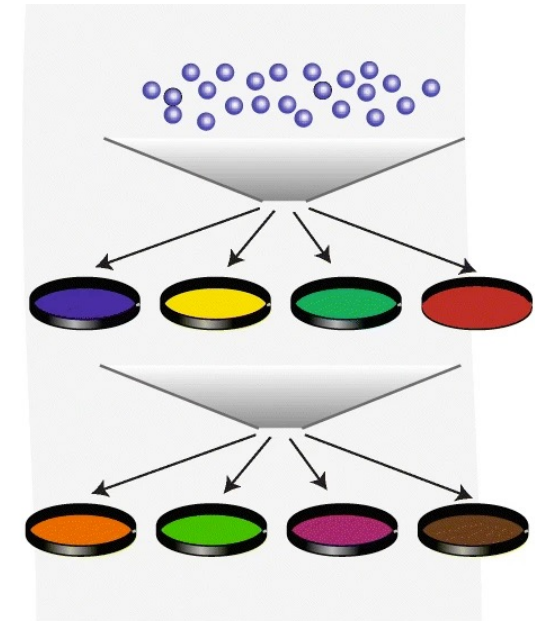


Plate or array  
(ICELL8, Fluidigm C1)

Takara Bio SMART-seq



Droplet-based  
(10X Genomics; BioRad)

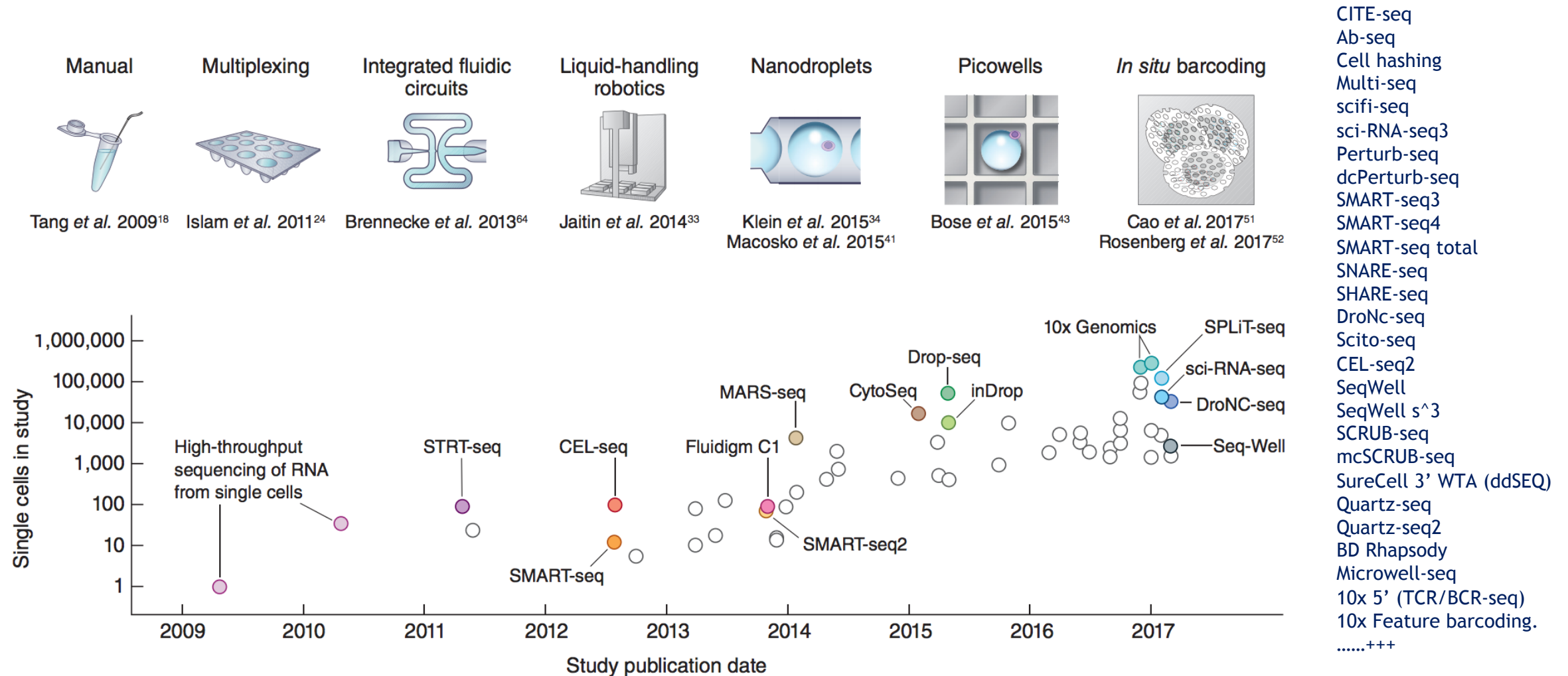


split-pool  
(sciATAC-seq)

SPLiT-seq

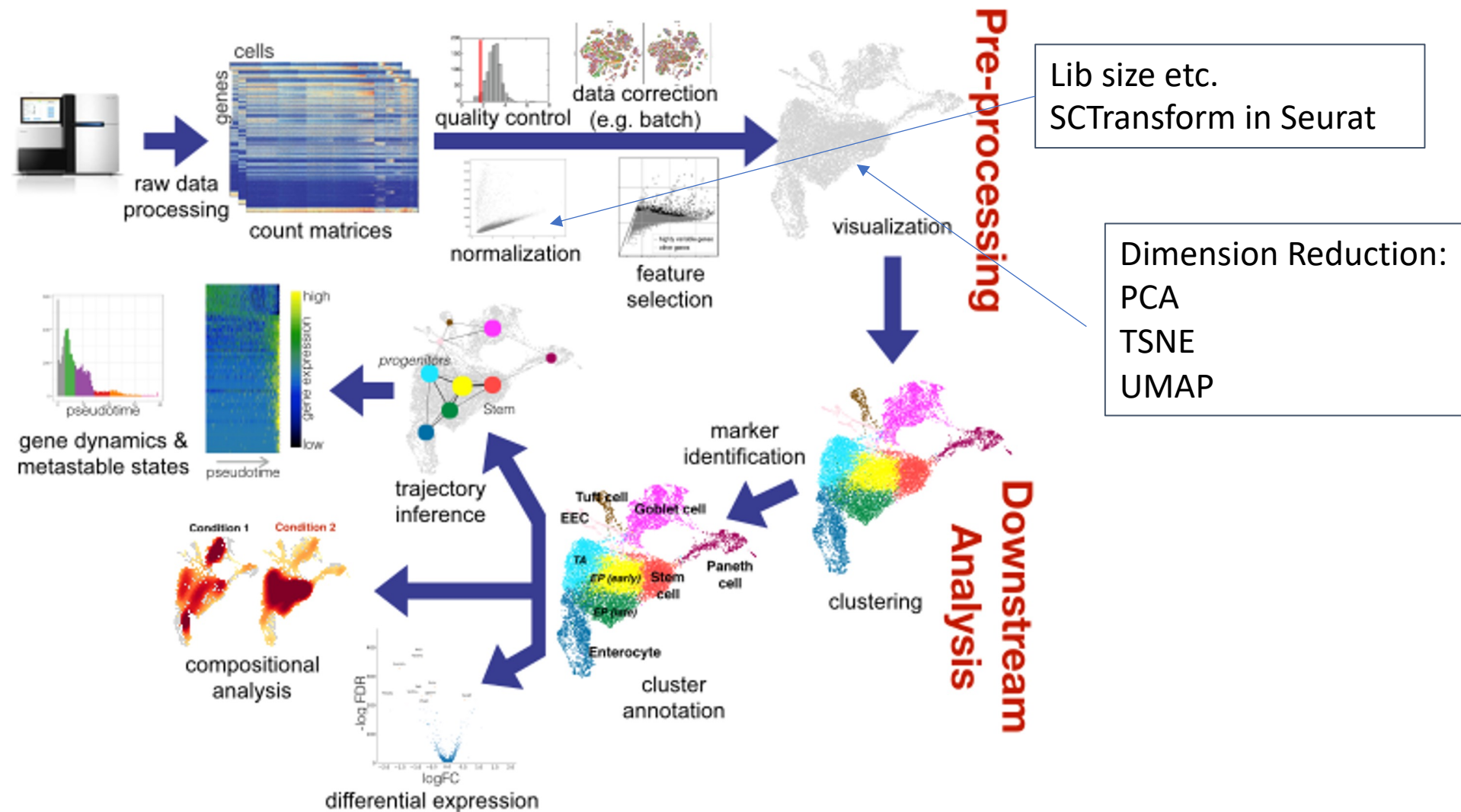
<https://www.youtube.com/watch?v=WqaeZe7mKUc>

# Exponential increase in throughput and new smart tech



Svensson *et al.*, *Nature Protocols* (2018)

# Let's ~~walk~~ sprint through a typical\* scRNA-seq analysis



Credit to Peter Hickey

# Read Alignment and gene quantification

- Cell Ranger (10X Genomics) solution: cellranger count
- RNA-seq: STAR
  - STARsolo (Blibaum et al, F1000 2019): 10X faster than CellRanger
- Alignment free:
  - Salmon Alevin
  - Kallisto bustool
- Resolve cell barcode and correct barcode sequencing errors

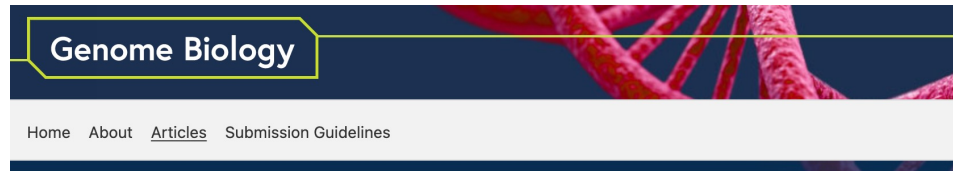
# Sparse matrix

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Sparse: many 0s in the matrix



# 0s: biological or not biological



Review | [Open Access](#) | Published: 21 January 2022

## Statistics or biology: the zero-inflation controversy about scRNA-seq data

[Ruochen Jiang](#), [Tianyi Sun](#), [Dongyuan Song](#) & [Jingyi Jessica Li](#)

[Genome Biology](#) 23, Article number: 31 (2022) | [Cite this article](#)

4654 Accesses | 80 Altmetric | [Metrics](#)

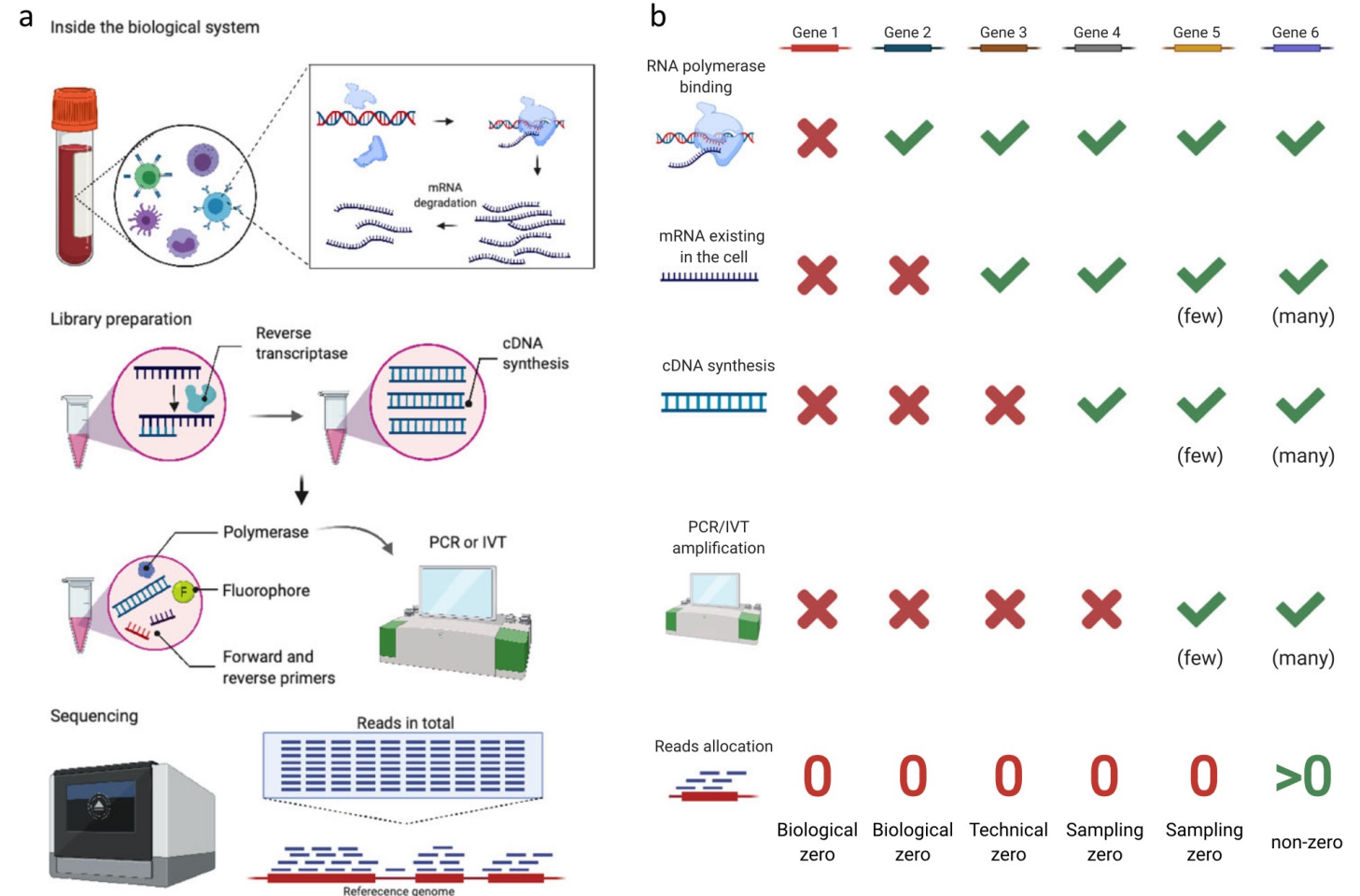
**Table 2** Clarification of zero-related terminology

In the current scRNA-seq literature, much ambiguity exists in the use of terms including “dropouts”, “excess zeros”, and “zero inflation” to describe the prevalence of zeros in scRNA-seq data [94]. We clarify the three terms by summarizing their various uses in the scRNA-seq field to facilitate our discussion.

**Dropout** or **dropouts** are widely used regarding the prevalence of zeros in scRNA-seq data. It was first introduced in the SCDE method paper: “dropout describes zero gene expression for the genes that show moderate or high expressions in only a proportion of cells [38]”. Hence, dropouts, as a data-driven concept, are not equivalent to either biological or non-biological zeros. Nevertheless, the use of “dropouts” in later papers became inconsistent and confusing: most papers meant non-biological zeros [20, 36, 40, 52, 55, 95, 96]; some meant non-biological zeros and low expression measurements [45, 97]; some meant all zeros [46, 47, 98]. In addition, “dropout” was often used as an adjective to mean the existence of many zeros [99]. Such inconsistent uses of “dropouts” are emphasized in a recent work [94]. To avoid possible confusion, we will not use “dropout” or “dropouts” in the following text.

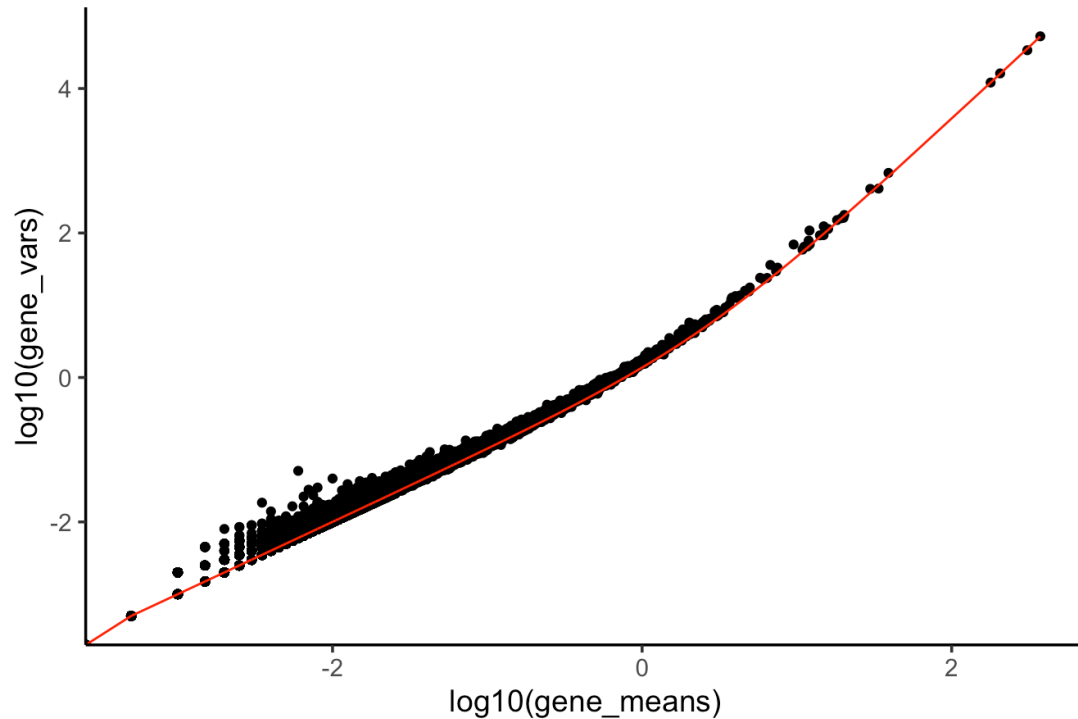
**Excess zeros** are used in various ways: some papers referred to the larger proportion of zeros in scRNA-seq data than in bulk RNA-seq data [40]; some meant non-biological zeros [45, 96]; some meant the additional zeros that cannot be explained by the negative binomial (NB) model [97]. To avoid confusion, we will not use “excess zeros” in the following text.

**Zero inflation**, unlike the first two terms, is a statistical concept that depends on a specified model, i.e., a count distribution such as the Poisson distribution and the NB distribution [95]. It means the proportion of zeros that exceeds what is expected under the specified model [40]. We will use “zero inflation” in the following discussion because its definition has no ambiguity.



# Droplet scRNA-seq is not zero-inflated

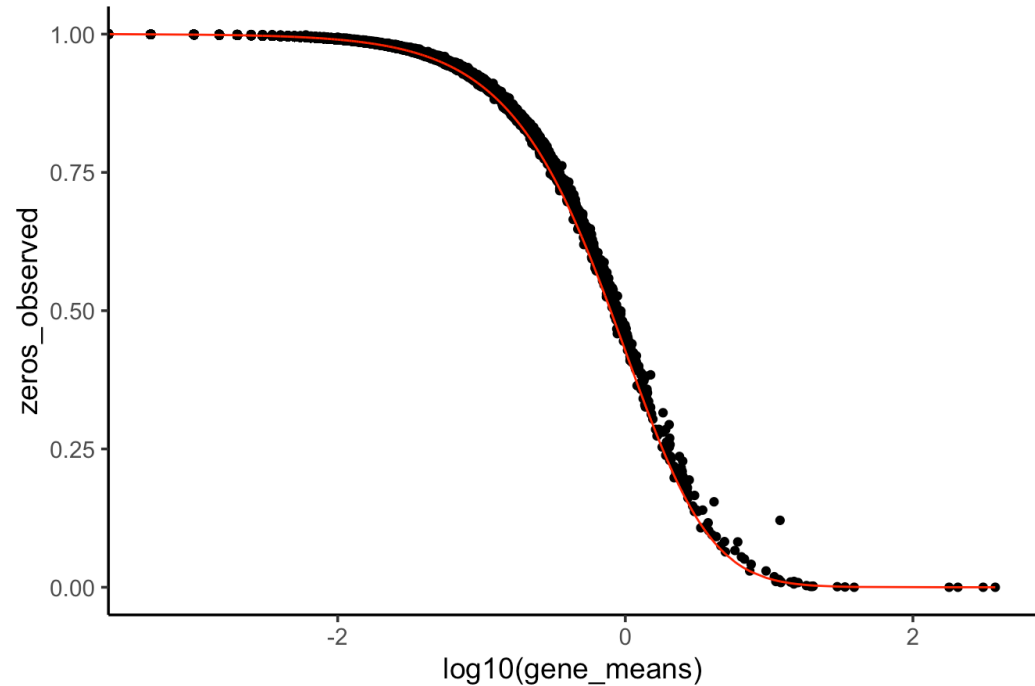
svensson et al



$$f(x; n, p) \equiv \Pr(X = x) = \binom{n + x - 1}{n - 1} (1 - p)^x p^n$$

This represents the number of failures which occur in a sequence of Bernoulli trials before a target number of successes (n) is reached. The mean is  $\mu = n(1-p)/p$  and variance  $n(1-p)/p^2$ .

Svensson et al 2

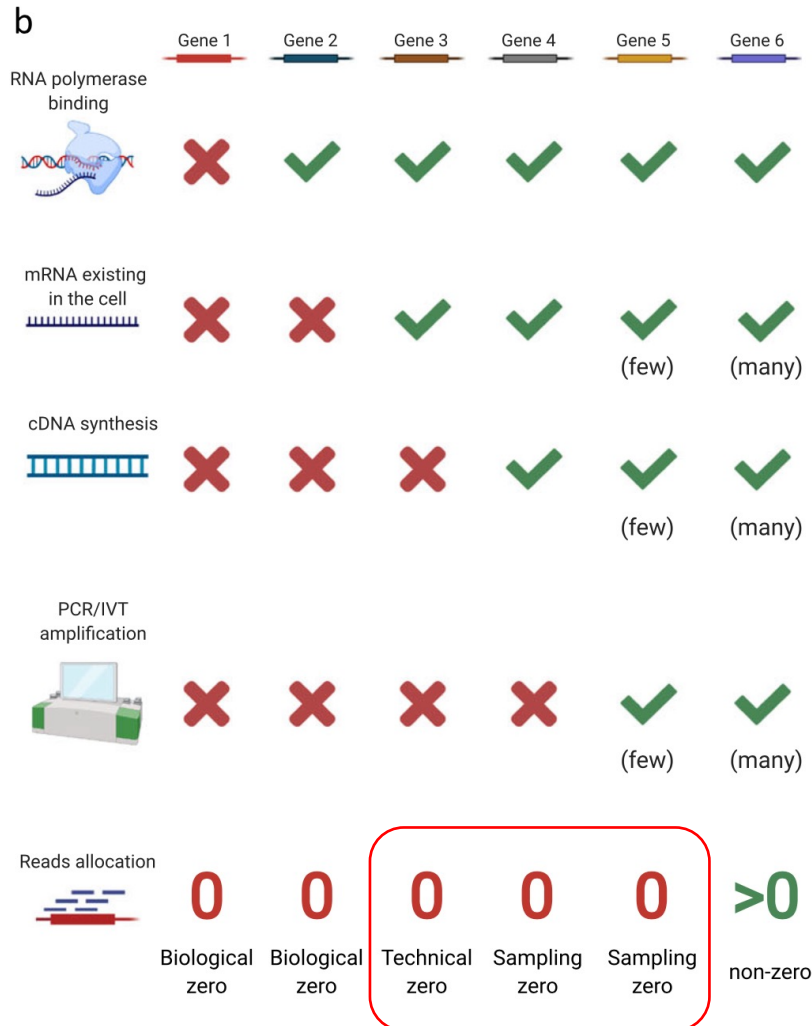


Svensson et al 2020

<https://divingintogeneticsandgenomics.rbind.io/post/negative-binomial-distribution-in-single-cell-rnaseq/>



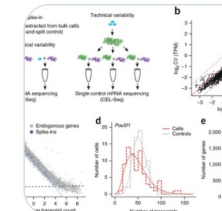
# Denoising vs imputation



**Florian Wagner** @flo\_compbio · Nov 2, 2020

Replying to @tangming2005 @fooliu and 2 others

The term "imputation" typically implies that there are "holes" in the matrix, i.e. missing data (zeros?). That's not what is happening. Rather, \*all\* measurements are associated with (different levels of) technical noise. Happy to elaborate, favorite ref:



nature.com

Validation of noise models for single-cell transcript...

Nature Methods - Noise models based on the identification of major sources of technical ...

1

4

15



**Florian Wagner** @flo\_compbio · Nov 2, 2020

From a technical perspective, measurements of 0 aren't any more special than measurements of 1 or 2 UMIs. Of course, if a gene is truly not expressed, then we should always measure 0. But if a gene is expressed, we could still fail to detect any of its transcripts in a given cell.

1

1

4



**Florian Wagner** @flo\_compbio · Nov 2, 2020

I think imputing (I prefer the term "denoising") makes a lot of sense if you want to visualize the data as a heatmap, but since it's often not totally clear how accurate denoising methods are, I would always double-check by plotting the raw data as well.

1

4



# To impute or not

Research | [Open Access](#) | [Published: 27 August 2020](#)

## A systematic evaluation of single-cell RNA-sequencing imputation methods

[Wenpin Hou](#), [Zhicheng Ji](#), [Hongkai Ji](#)  & [Stephanie C. Hicks](#) 

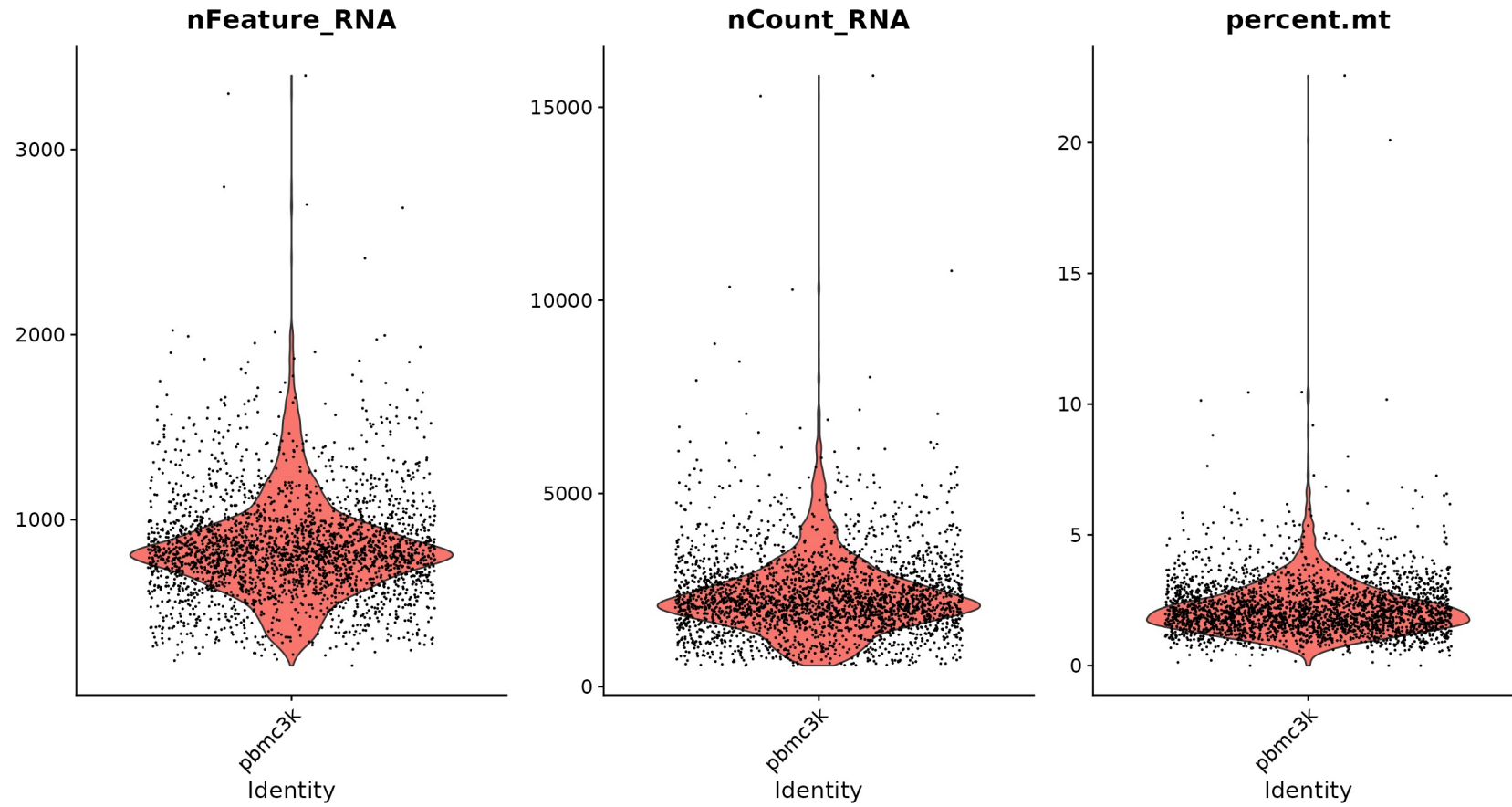
[Genome Biology](#) **21**, Article number: 218 (2020) | [Cite this article](#)

**15k** Accesses | **40** Citations | **100** Altmetric | [Metrics](#)

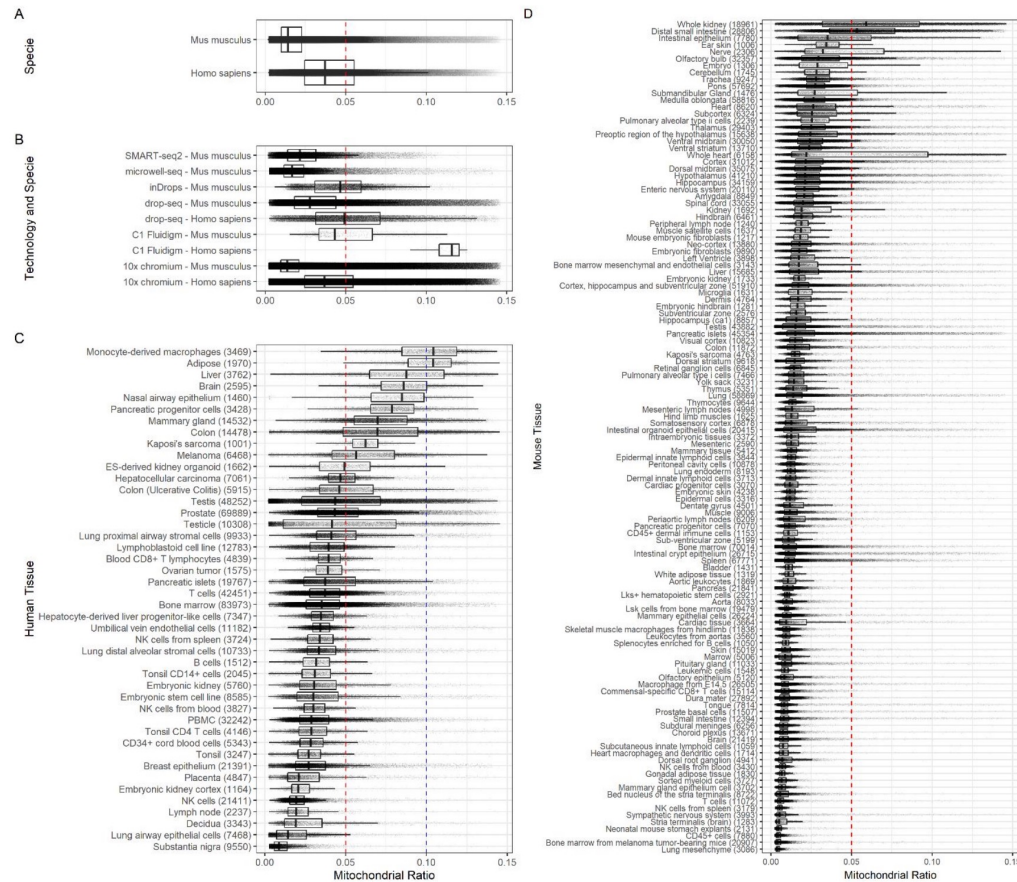
### Conclusions

We found that the majority of scRNA-seq imputation methods outperformed no imputation in recovering gene expression observed in bulk RNA-seq. However, the majority of the methods did not improve performance in downstream analyses compared to no imputation, in particular for clustering and trajectory analysis, and thus should be used with caution. In addition, we found substantial variability in the performance of the methods within each evaluation aspect. Overall, MAGIC, kNN-smoothing, and SAVER were found to outperform the other methods most consistently.

# Quality control



# Mitochondrial gene content cutoff



**Fig 1. Boxplots showing the differences in mtDNA% across species, technologies and tissues.** Each dot represents a cell; the red line is the early established 5% threshold, and the blue line is the 10% threshold for human cells proposed here. In parenthesis (panel C and D), the number of cells in the stated tissue. **(A)** The difference in mtDNA% between human and mice cells. **(B)** The differences in mtDNA% between human and mice cells by the technology used to generate the data. **(C)** Boxplots of mtDNA% across 44 human tissues. **(D)** Boxplots of mtDNA% across 121 mouse tissues.

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

# miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data

## Abstract

Single-cell RNA-sequencing (scRNA-seq) has made it possible to profile gene expression in tissues at high resolution. An important preprocessing step prior to performing downstream analyses is to identify and remove cells with poor or degraded sample quality using quality control (QC) metrics. Two widely used QC metrics to identify a ‘low-quality’ cell are (i) if the cell includes a high proportion of reads that map to mitochondrial DNA (mtDNA) encoded genes and (ii) if a small number of genes are detected. Current best practices use these QC metrics independently with either arbitrary, uniform thresholds (e.g. 5%) or biological context-dependent (e.g. species) thresholds, and fail to jointly model these metrics in a data-driven manner. Current practices are often overly stringent and especially untenable on certain types of tissues, such as archived tumor tissues, or tissues associated with mitochondrial function, such as kidney tissue [1]. We propose a data-driven QC metric (miQC) that jointly models both the proportion of reads mapping to mtDNA genes and the number of detected genes with mixture models in a probabilistic framework to predict the low-quality cells in a given dataset. We demonstrate how our QC metric easily adapts to different types of single-cell datasets to remove low-quality cells while preserving high-quality cells that can be used for downstream analyses. Our software package is available at <https://bioconductor.org/packages/miQC>.

# Normalization and scaling

- Bulk-RNAseq
  - Reads per kilobase of exon per million reads mapped (RPKM)
  - Transcript per million (TPM)
- Single-cell RNAseq
  - LogNormalize:  $\log(n/\text{library\_size} * 10^6)$
  - scTransform
- Scaling:
  - Shifts the expression of each gene, so that the mean expression across cells is 0
  - Scales the expression of each gene, so that the variance across cells is 1
  - This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate



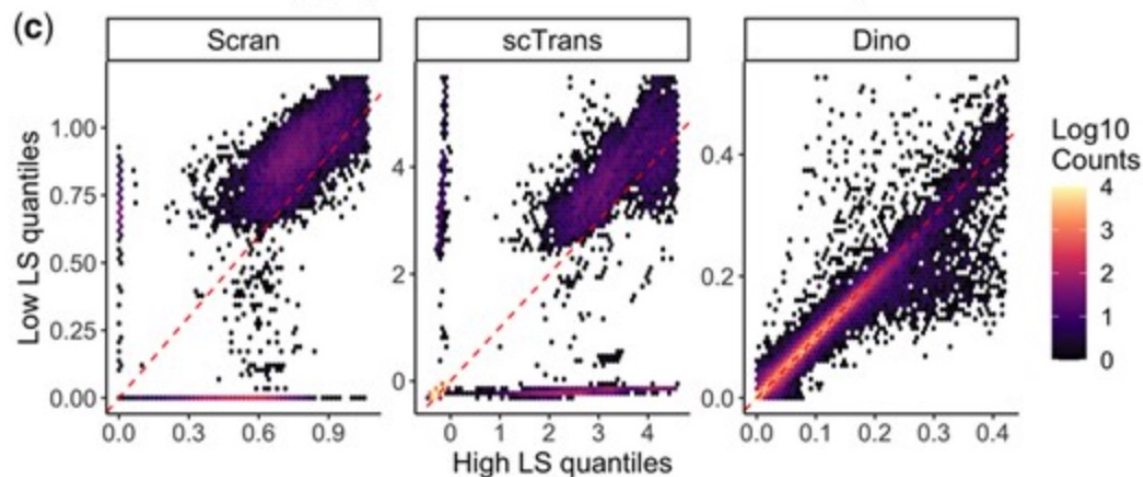
# How to calculate gene-gene correlation



Ming "Tommy" Tang  
@tangming2005

1/ single-cell RNAseq data matrix is sparse. dominant 0s makes gene-gene correlation calculation hard. Tools that I know to tackle this problem [#scRNAseq](#) : [bioconductor.org/packages/relea...](https://bioconductor.org/packages/release...)

9:58 PM · Mar 15, 2022 · Twitter Web App



Analysis | [Published: 08 April 2019](#)

## Evaluating measures of association for single-cell transcriptomics

[Michael A. Skinnider](#) ✉, [Jordan W. Squair](#) & [Leonard J. Foster](#) ✉

[Nature Methods](#) **16**, 381–386 (2019) | [Cite this article](#)

11k Accesses | 40 Citations | 57 Altmetric | [Metrics](#)

### Abstract

Single-cell transcriptomics provides an opportunity to characterize cell-type-specific transcriptional networks, intercellular signaling pathways and cellular diversity with unprecedented resolution by profiling thousands of cells in a single experiment. However, owing to the unique statistical properties of scRNA-seq data, the optimal measures of association for identifying gene–gene and cell–cell relationships from single-cell transcriptomics remain unclear. Here, we conducted a large-scale evaluation of 17 measures of association for their ability to reconstruct cellular networks, cluster cells of the same type and link cell-type-specific transcriptional programs to disease. Measures of proportionality were consistently among the best-performing methods across datasets and tasks. Our analysis provides data-driven guidance for gene and cell network analysis in single-cell transcriptomics.

Research | [Open Access](#) | [Published: 14 December 2021](#)

# Benchmarking UMI-based single-cell RNA-seq preprocessing workflows

[Yue You](#) , [Luyi Tian](#), [Shian Su](#), [Xueyi Dong](#), [Jafar S. Jabbari](#), [Peter F. Hickey](#)  & [Matthew E. Ritchie](#) 

[Genome Biology](#) **22**, Article number: 339 (2021) | [Cite this article](#)

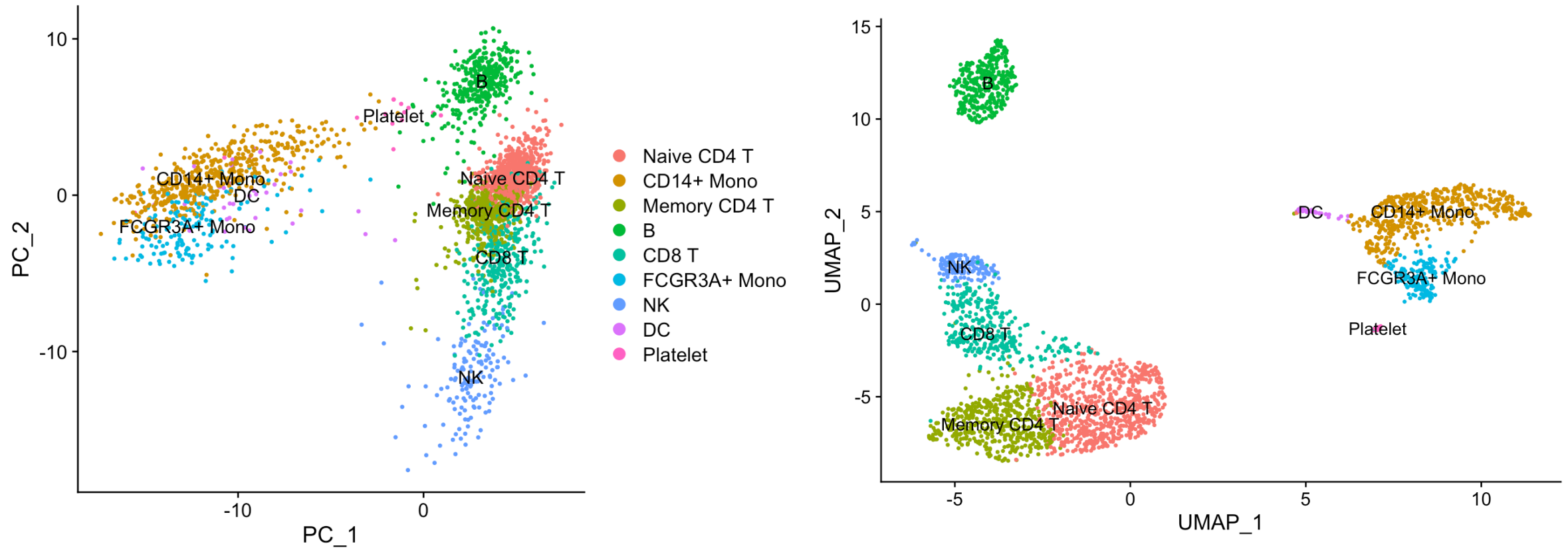
**4160** Accesses | **2** Citations | **56** Altmetric | [Metrics](#)

## Conclusions

In summary, the choice of preprocessing method was found to be less important than other steps in the scRNA-seq analysis process. Our study comprehensively compares common scRNA-seq preprocessing workflows and summarizes their characteristics to guide workflow users.



# Dimension reduction (PCA vs UMAP)

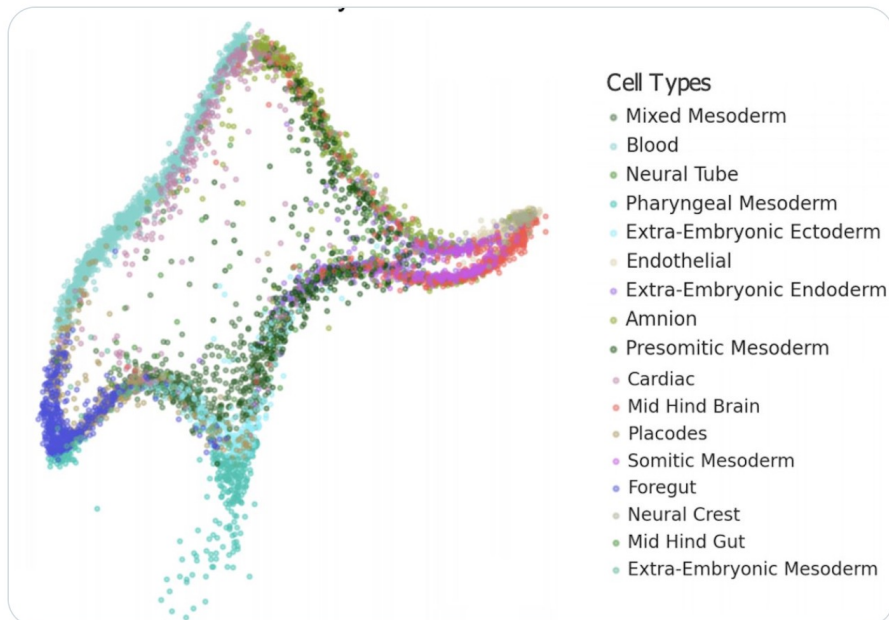


# UMAP and TSNE



**Lior Pachter** ✓ @lpachter · Aug 27, 2021

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary. [biorxiv.org/content/10.110...](https://doi.org/10.1101/2021.08.27.458111)



94

1,538

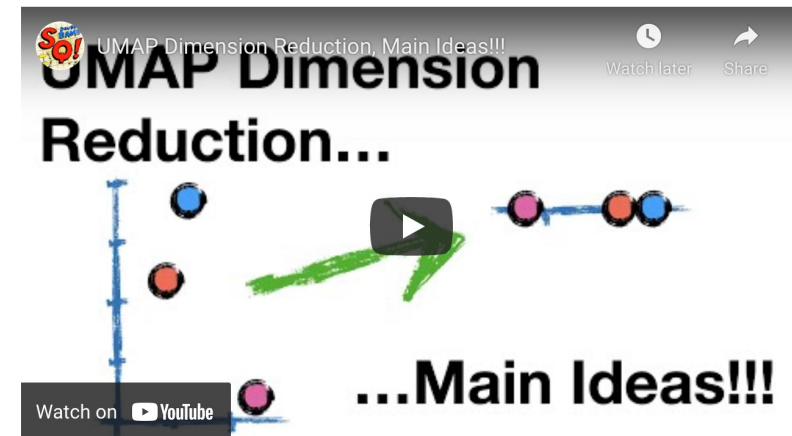
4,105



I personally think TSNE/UMAP is still useful  
To have a global view of your data.

## UMAP Dimension Reduction: Part 1 – Main Ideas

March 7, 2022

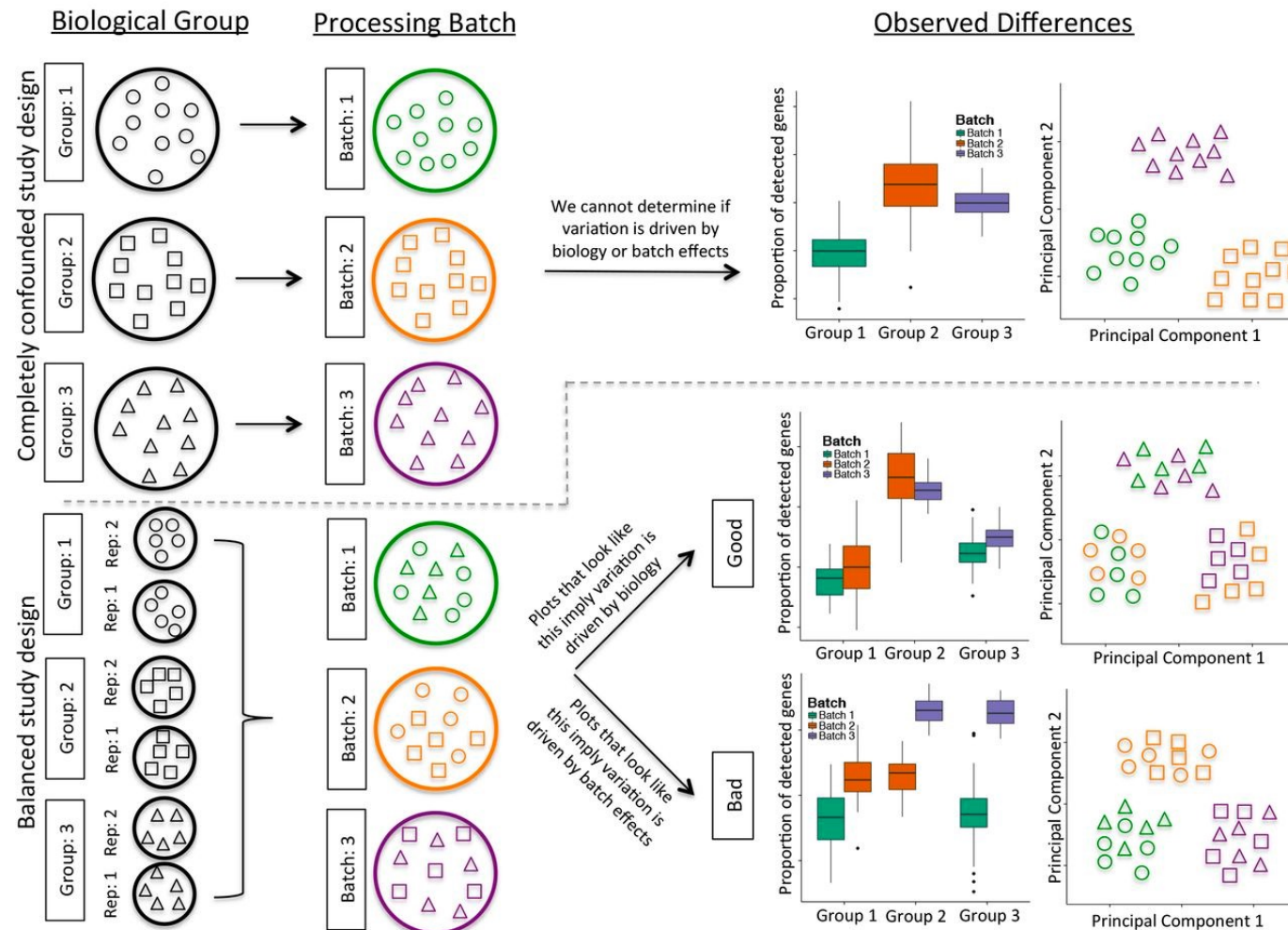


<https://twitter.com/lpachter/status/1431326048168202247>

<https://statquest.org/>

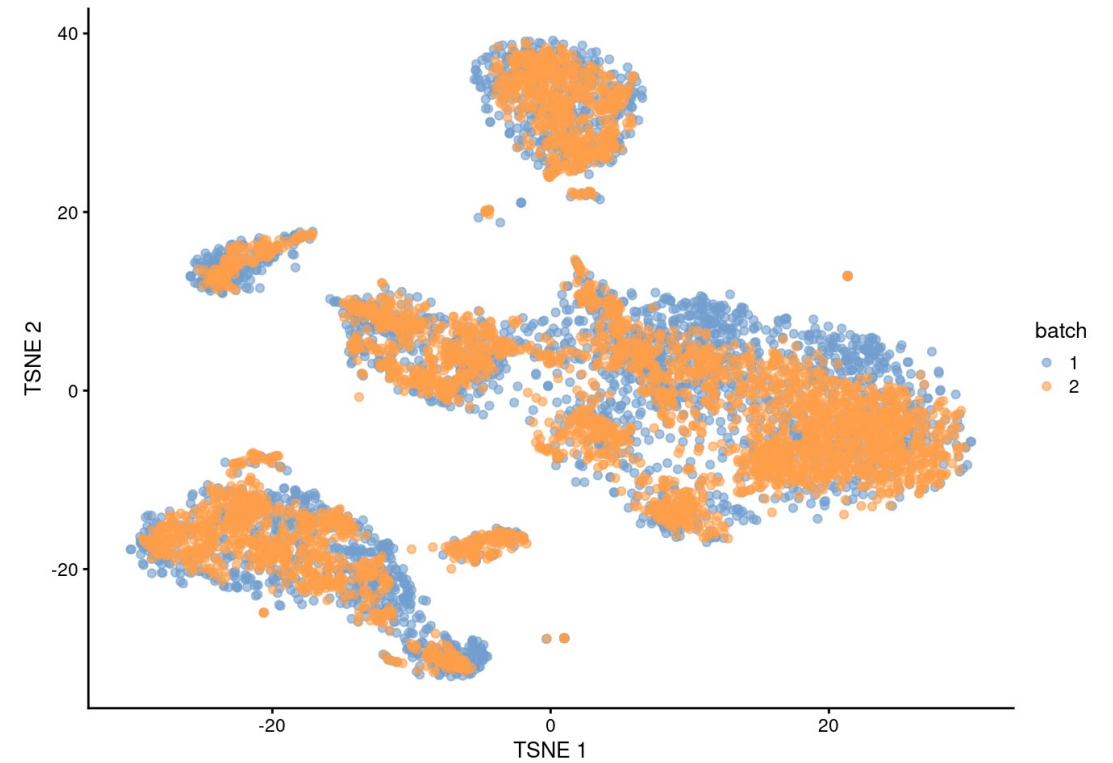
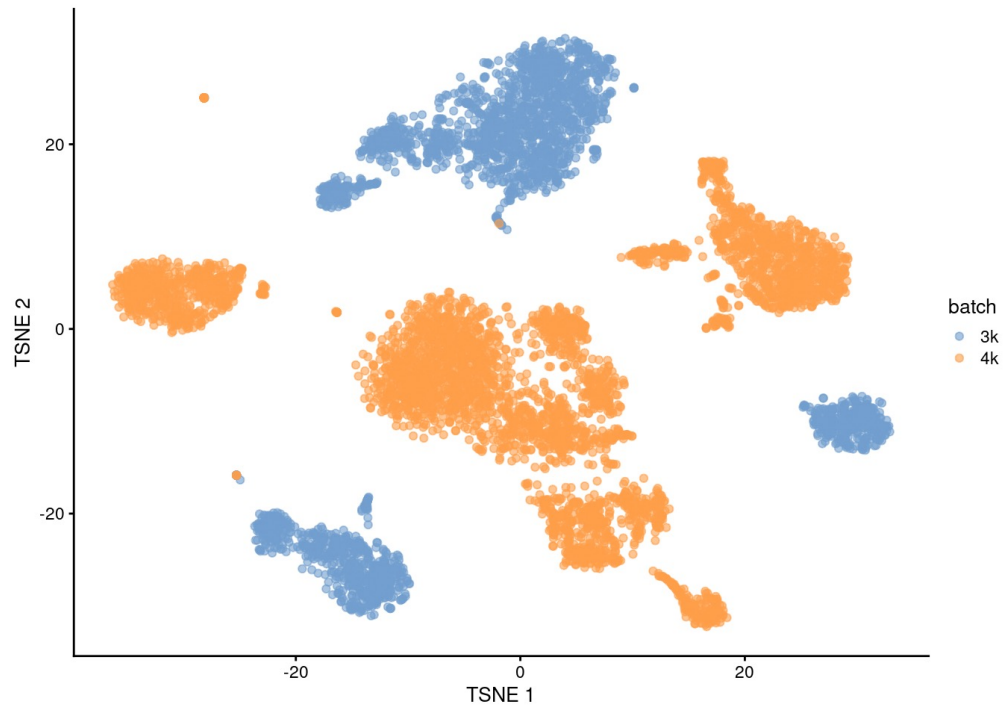
# Avoid batch and confounding effects: experimental design

## The Problem of Confounding Biological Variation and Batch Effects



Hicks et al., *Biostatistics*. 2018

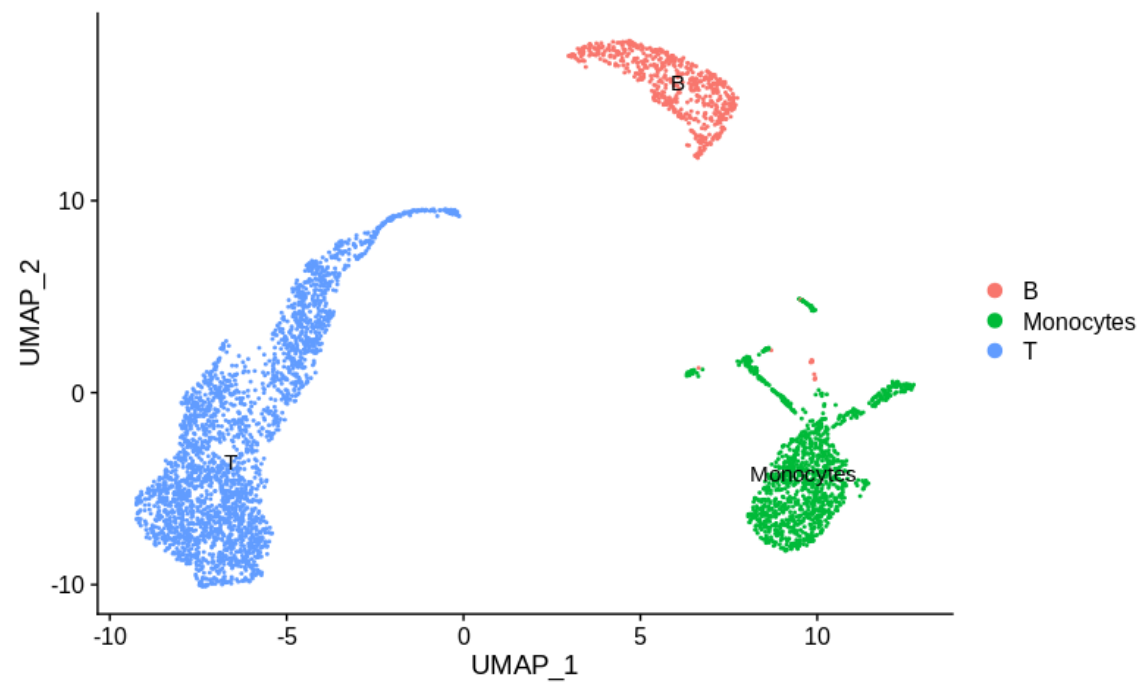
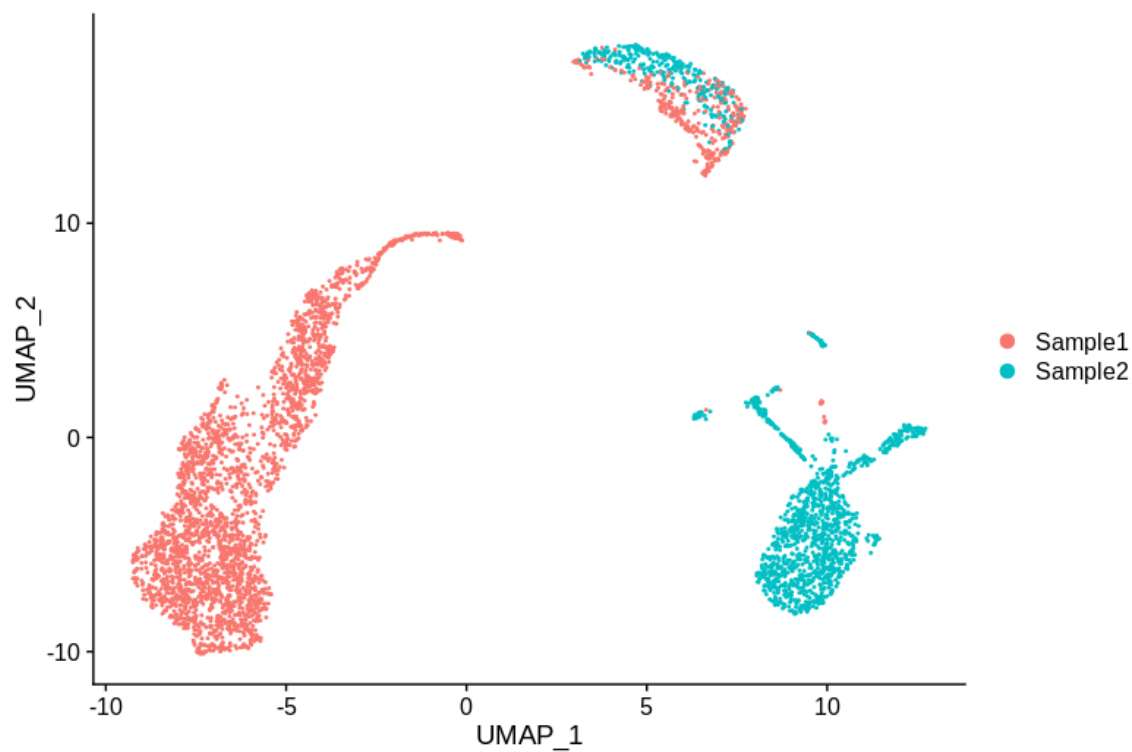
# Data integration/batch correction



<http://bioconductor.org/books/3.14/OSCA.multisample/integrating-datasets.html#motivation>

# Data integration

- Batch effect or not? Correct or not





# Sacrificing biology by integration

## 6.4.2 Sacrificing biology by integration

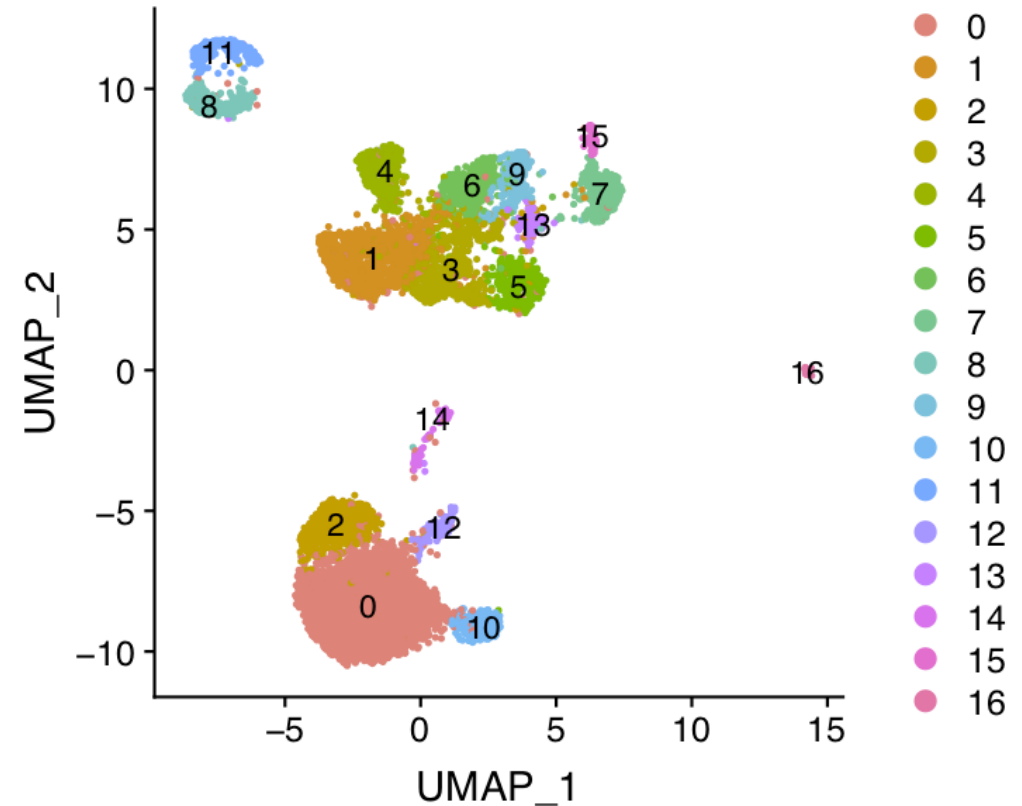
Earlier in this chapter, we defined clusters from corrected values after applying `fastMNN()` to cells from all samples in the chimera dataset. Alert readers may realize that this would result in the removal of biological differences between our conditions. Any systematic difference in expression caused by injection would be treated as a batch effect and lost when cells from different samples are aligned to the same coordinate space. Now, one may not consider injection to be an interesting biological effect, but the same reasoning applies for other conditions, e.g., integration of wild-type and knock-out samples (Section 5) would result in the loss of any knock-out effect in the corrected values.

This loss is both expected and desirable. As we mentioned in Section 3, the main motivation for performing batch correction is to enable us to characterize population heterogeneity in a consistent manner across samples. This remains true in situations with multiple conditions where we would like one set of clusters and annotations that can be used as common labels for the DE or DA analyses described above. The alternative would be to cluster each condition separately and to attempt to identify matching clusters across conditions - not straightforward for poorly separated clusters in contexts like differentiation.

It may seem distressing to some that a (potentially very interesting) biological difference between conditions is lost during correction. However, this concern is largely misplaced as the correction is only ever used for defining common clusters and annotations. The DE analysis itself is performed on pseudo-bulk samples created from the uncorrected counts, preserving the biological difference and ensuring that it manifests in the list of DE genes for affected cell types. Of course, if the DE is strong enough, it may result in a new condition-specific cluster that would be captured by a DA analysis as discussed in Section 6.4.1.

# Clustering

- Dimension reduction (PCA)
- k-means, hierarchical clustering etc
- Cluster cells (on the reduced dimensions) using graph-based method in Seurat v3 (Stuart et al, Cell 2019). KNN graph + community detection algorithm
- Can visualize using t-SNE / UMAP

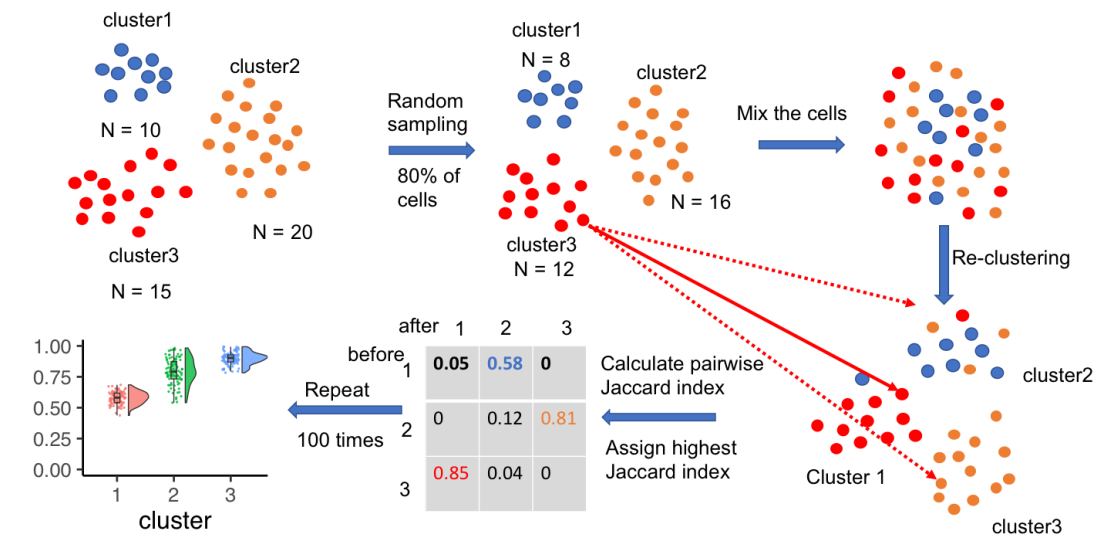




# Evaluating cluster stability

## 5.4 Evaluating cluster stability

A desirable property of a given clustering is that it is stable to perturbations to the input data (Von Luxburg 2010). Stable clusters are logistically convenient as small changes to upstream processing will not change the conclusions; greater stability also increases the likelihood that those conclusions can be reproduced in an independent replicate study. *scrn* uses bootstrapping to evaluate the stability of a clustering algorithm on a given dataset - that is, cells are sampled with replacement to create a “bootstrap replicate” dataset, and clustering is repeated on this replicate to see if the same clusters can be reproduced. We demonstrate below for graph-based clustering on the PCs of the PBMC dataset.



Tang et al 2021 Bioinformatics

<http://bioconductor.org/books/3.14/OSCA.advanced/clustering-redux.html#cluster-bootstrapping>

<https://github.com/crazyhottommy/scclusteval>

# Marker gene p-value is inflated



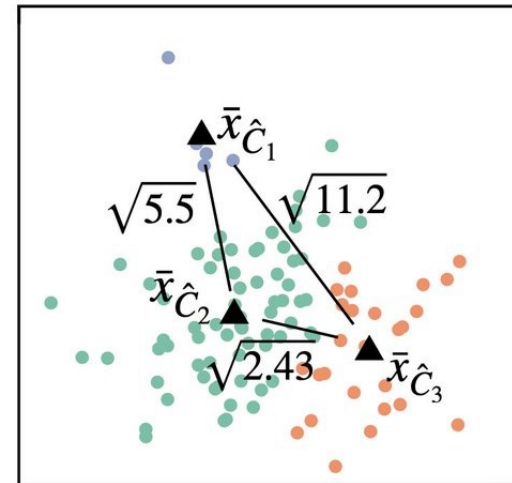
Lucy L. Gao  
@lucylgao

"Double-dipping" - generating a hypothesis based on your data, and then testing the hypothesis on that same data - is dangerous. To see this, let's take data with no signal at all ... 1/



Step 1: Sample 100 observations

1:39 PM · Aug 29, 2020 · Twitter Web App



**Step 1:** Sample 100 observations

**Step 2:** Cluster the observations

**Step 3:** Compute p-values for a difference in means

All three p-values < 0.000001!! 🤯

<https://www.lucylgao.com/clusterpval/>

<https://www.youtube.com/watch?v=voseWZlaFm4>

<https://www.sciencedirect.com/science/article/pii/S2405471219302698>

# Large number of data points will make p-value tiny



Ming "Tommy" Tang  
@tangming2005

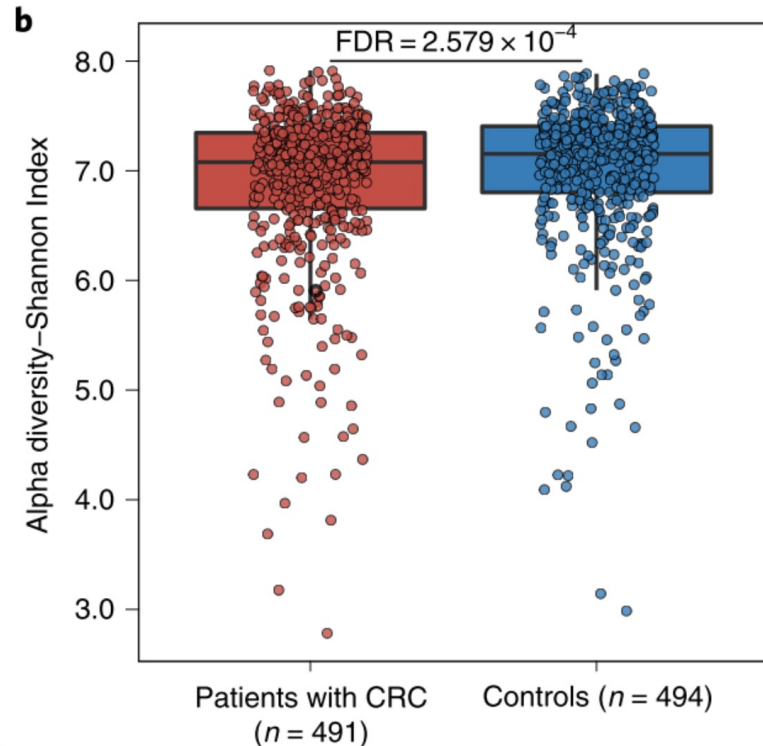
...

Reminder: You will get small p-values when your the number of data points is large



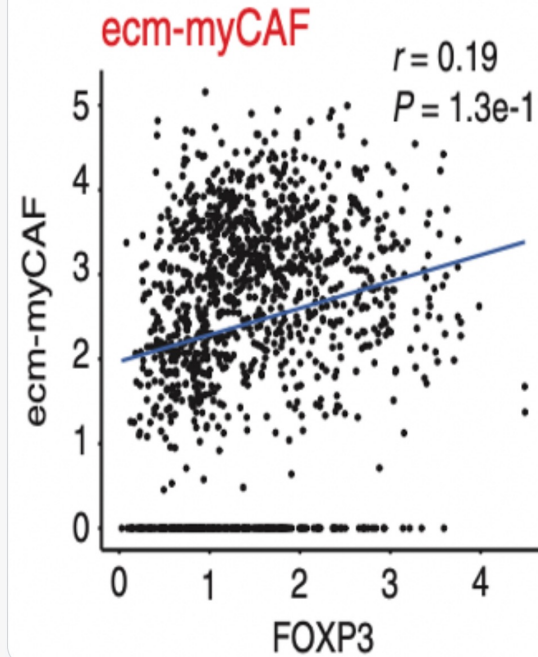
Daniel Martínez @dan\_martimarti · Feb 4

This effect size...



Ming "Tommy" Tang @tangming2005 · Sep 28, 2020

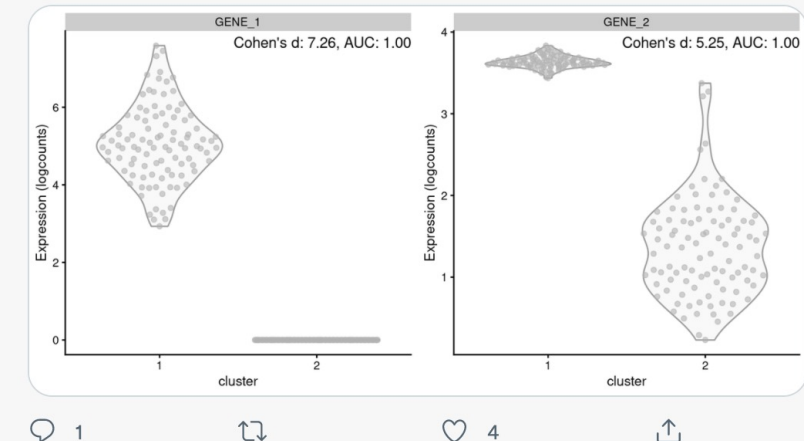
Question: if you have tens of thousands of data points with a **correlation** of 0.2 and a p-value  $10^{-11}$ . Is it meaningful to show that? you always get a tiny p-value when you have a lot of data points.



Mikhael Dito Manurung @mikhaeldito313 · Mar 19

2/T-test. It can give you Cohen's D, which is the number of standard deviations that separate the means of two groups. This accounts for the magnitude of difference in expression, which gives additional information over Wilcoxon's AUC.

(Image source: [tinyurl.com/y8fqkkx](https://tinyurl.com/y8fqkkx))



<https://twitter.com/tangming2005/status/1489964367336648707>

<https://mobile.twitter.com/mikhaeldito313/status/1505204061506715649>

# Cell annotation

SingleR

Seurat V4 reference based mapping



Ming "Tommy" Tang  
@tangming2005

"The forever daunting question of cell annotation." ---  
[@NieuwenhuisTim](#) . yeah, you got it right :) [#scRNAseq](#)

11:54 AM · Feb 17, 2022 · Twitter Web App

||| View Tweet activity

3 Retweets 30 Likes



Tweet your reply

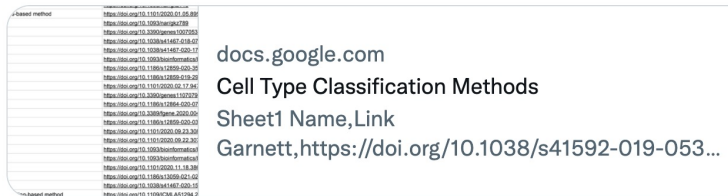
Reply



Matthew Bernstein @Matthew\_N\_B · Feb 17

Replying to [@tangming2005](#) and [@NieuwenhuisTim](#)

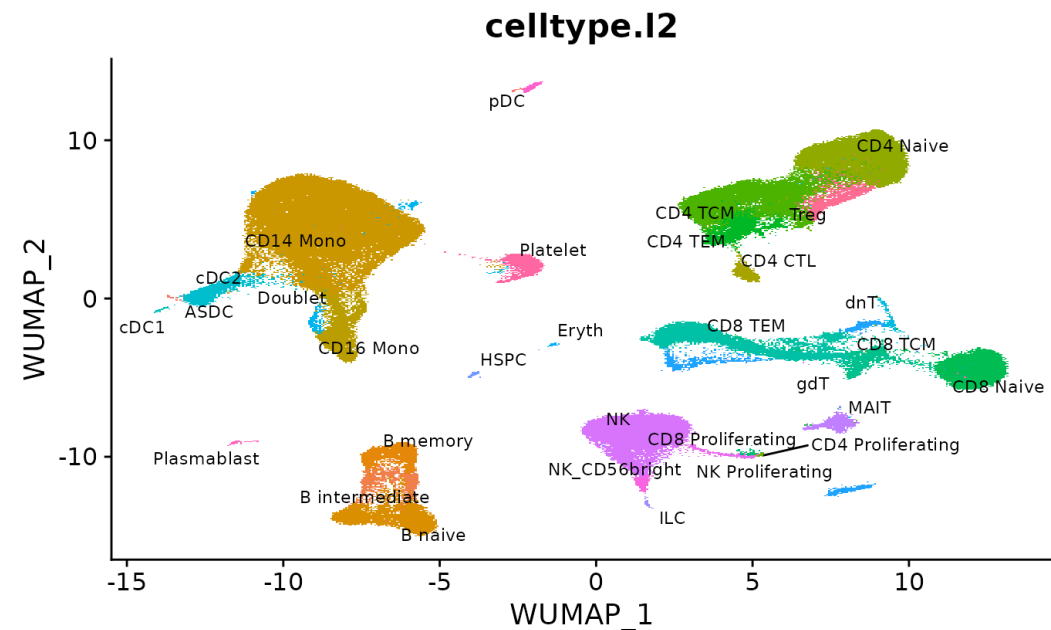
I have a list of 60 cell type annotation methods. Despite so many methods, it's still hard...



5

16

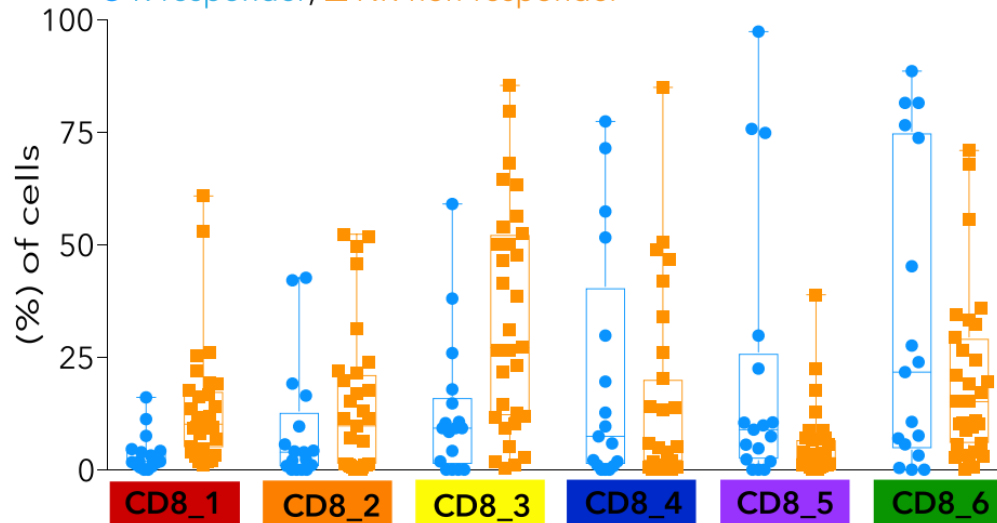
48



# Differential cell abundance analysis

**D**

● R-responder; ■ NR-non-responder



All samples	6.7x10 <sup>-5</sup>	n.s	0.001	n.s	n.s	n.s
Ag presentation	1x10 <sup>-5</sup>	n.s	6.6x10 <sup>-5</sup>	0.01	0.02	n.s

```
##
##           5  6  7  8  9 10
## Allantois    97 15 139 127 318 259
## Blood progenitors 1  6  3  16  6  8  17
## Blood progenitors 2 31  8  28  21  43 114
## Cardiomyocytes    85 21  79  31 174 211
## Caudal Mesoderm    10 10  9  3  10  29
## Caudal epiblast    2  2  0  0  22  45
```

## 6.2 Performing the DA analysis

Our DA analysis will again be performed with the *edgeR* package. This allows us to take advantage of the NB GLM methods to model overdispersed count data in the presence of limited replication - except that the counts are not of reads per gene, but of cells per label (Lun, Richard, and Marioni 2017). The aim is to share information across labels to improve our estimates of the biological variability in cell abundance between replicates.

```
library(edgeR)
# Attaching some column metadata.
extra.info <- colData(merged)[match(colnames(abundances), merged$sample),]
y.ab <- DGEList(abundances, samples=extra.info)
y.ab
```

# Multi-sample Differential expression: pseudo-bulk for the win



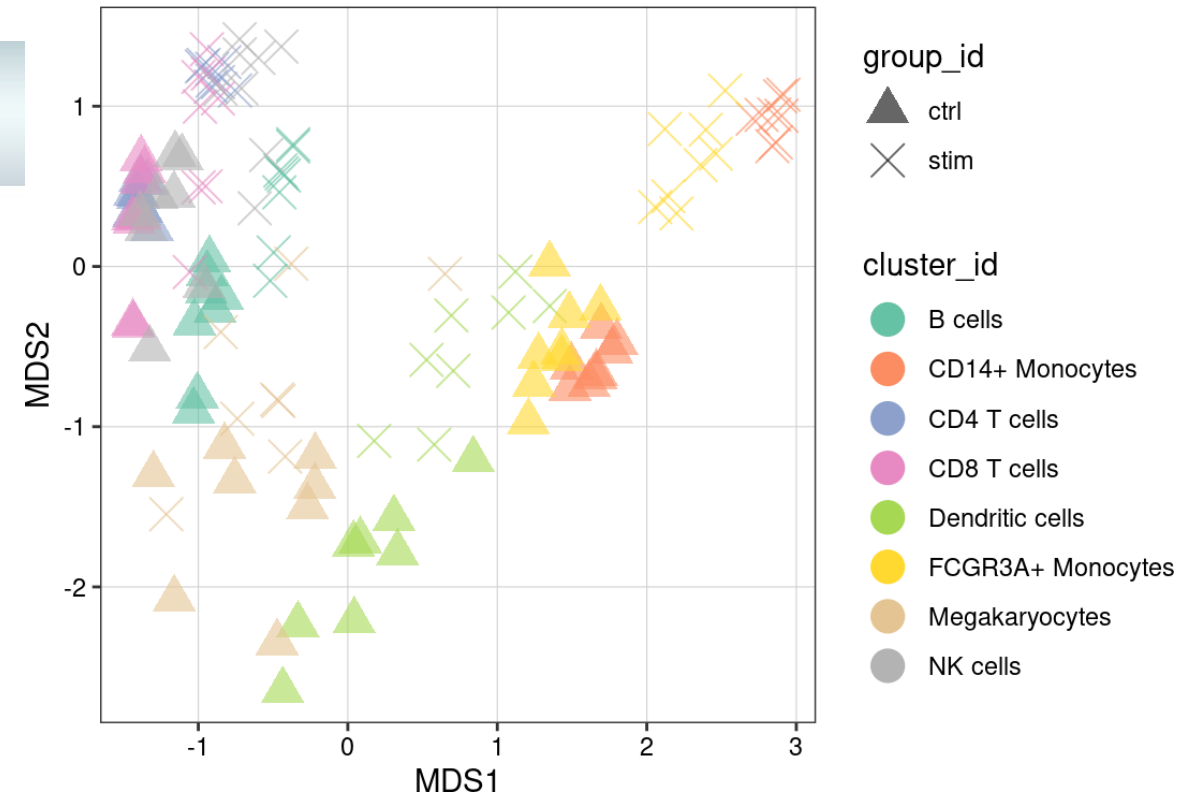
ARTICLE

<https://doi.org/10.1038/s41467-021-25960-2>

OPEN

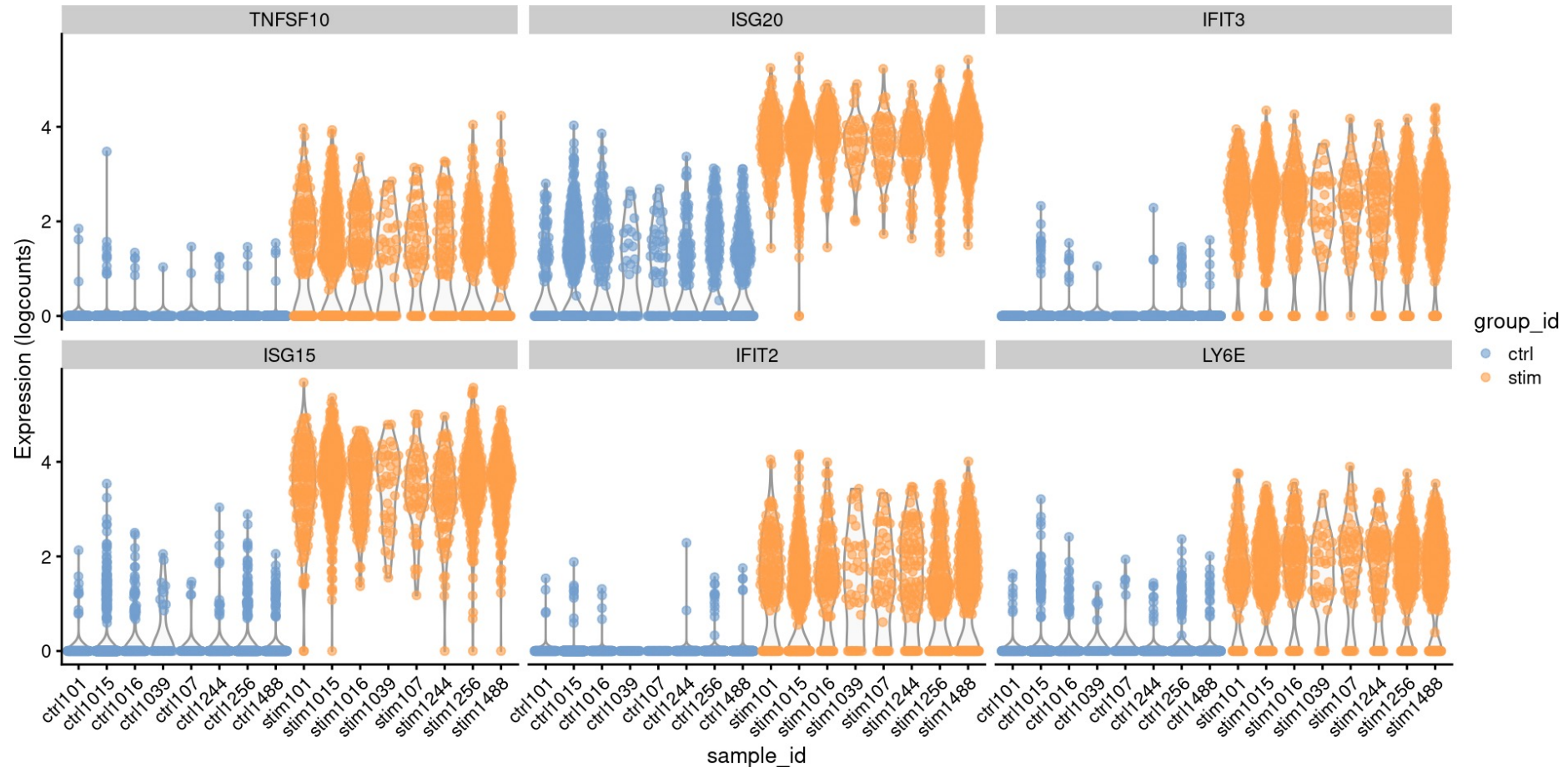
## Confronting false discoveries in single-cell differential expression

Jordan W. Squair<sup>1,2,3</sup>, Matthieu Gautier<sup>1,2</sup>, Claudia Kathe<sup>1,2</sup>, Mark A. Anderson<sup>1,2</sup>, Nicholas D. James<sup>1,2</sup>, Thomas H. Hutson<sup>1,2</sup>, Rémi Hudelle<sup>1,2</sup>, Taha Qaiser<sup>1,3</sup>, Kaya J. E. Matson<sup>4</sup>, Quentin Barraud<sup>1,2</sup>, Ariel J. Levine<sup>1,4</sup>, Gioele La Manno<sup>1</sup>, Michael A. Skinnider<sup>1,2,5,6</sup> & Grégoire Courtine<sup>1,2,6</sup>

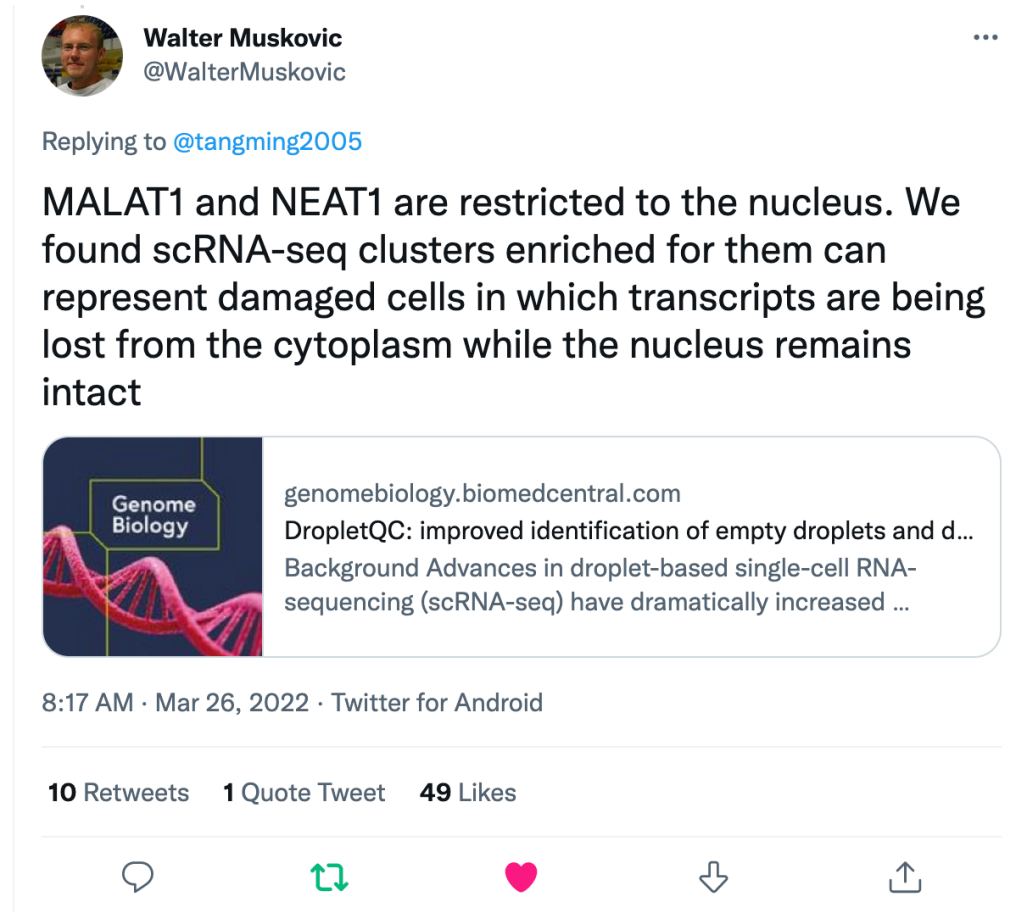




# Muscat::pbDS() or Scran::pseudoBulkDEG



# Be aware of technical artifacts



<https://twitter.com/tangming2005/status/1507520784792793090>

<https://kb.10xgenomics.com/hc/en-us/articles/360004729092-Why-do-I-see-high-levels-of-Malat1-in-my-gene-expression-data->

# Differential expression (DE) vs Differential abundance (DA)

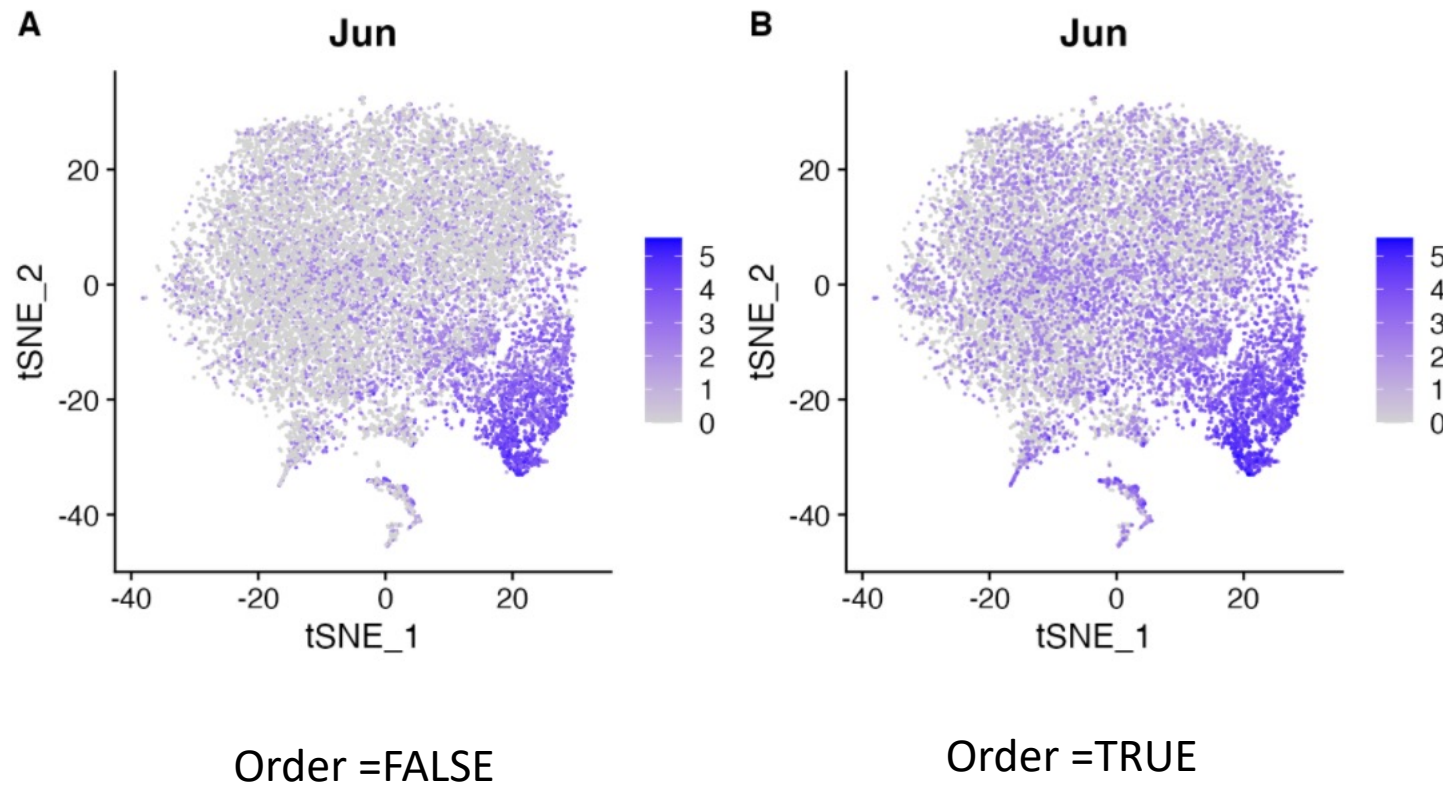
## 14.6.1 DE or DA? Two sides of the same coin

While useful, the distinction between DA and DE analyses is inherently artificial for scRNA-seq data. This is because the labels used in the former are defined based on the genes to be tested in the latter. To illustrate, consider a scRNA-seq experiment involving two biological conditions with several shared cell types. We focus on a cell type  $X$  that is present in both conditions but contains some DEGs between conditions. This leads to two possible outcomes:

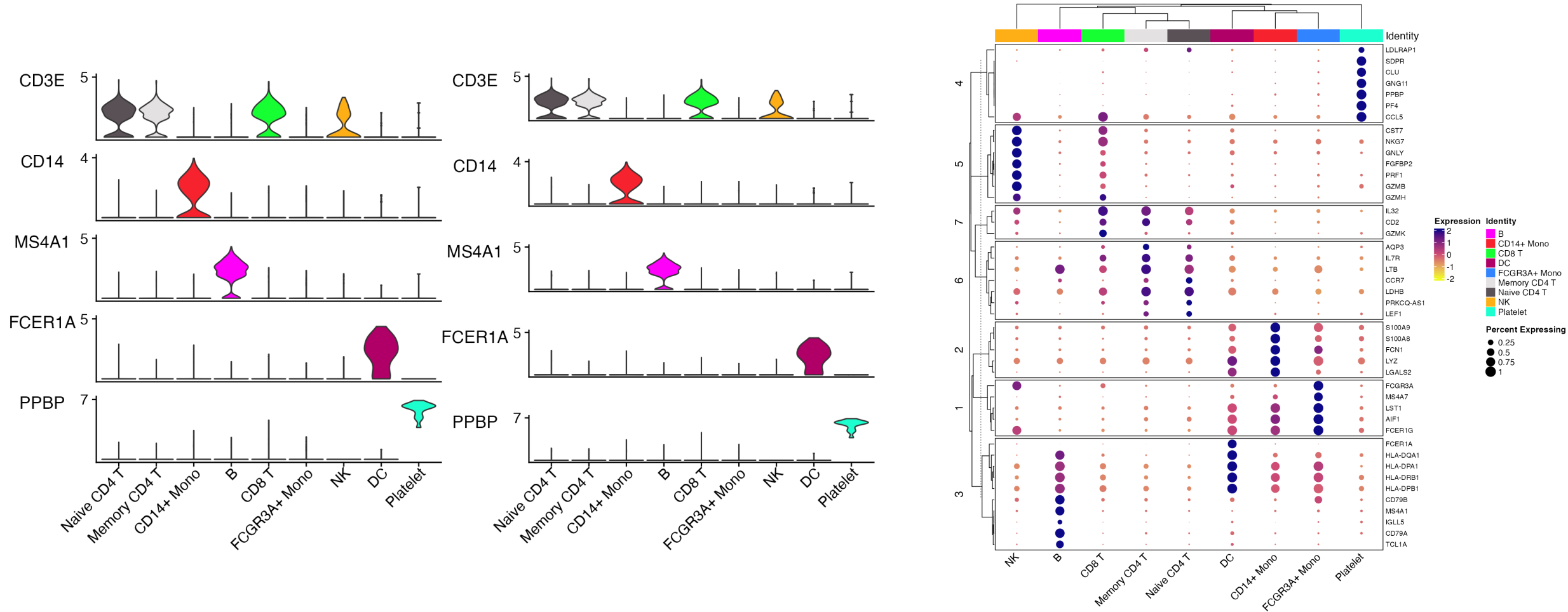
1. The DE between conditions causes  $X$  to form two separate clusters (say,  $X_1$  and  $X_2$ ) in expression space. This manifests as DA where  $X_1$  is enriched in one condition and  $X_2$  is enriched in the other condition.
2. The DE between conditions is not sufficient to split  $X$  into two separate clusters, e.g., because the data integration procedure identifies them as corresponding cell types and merges them together. This means that the differences between conditions manifest as DE within the single cluster corresponding to  $X$ .

We have described the example above in terms of clustering, but the same arguments apply for any labelling strategy based on the expression profiles, e.g., automated cell type assignment (Chapter 12). Moreover, the choice between outcomes 1 and 2 is made implicitly by the combined effect of the data merging, clustering and label assignment procedures. For example, differences between conditions are more likely to manifest as DE for coarser clusters and as DA for finer clusters, but this is difficult to predict reliably.

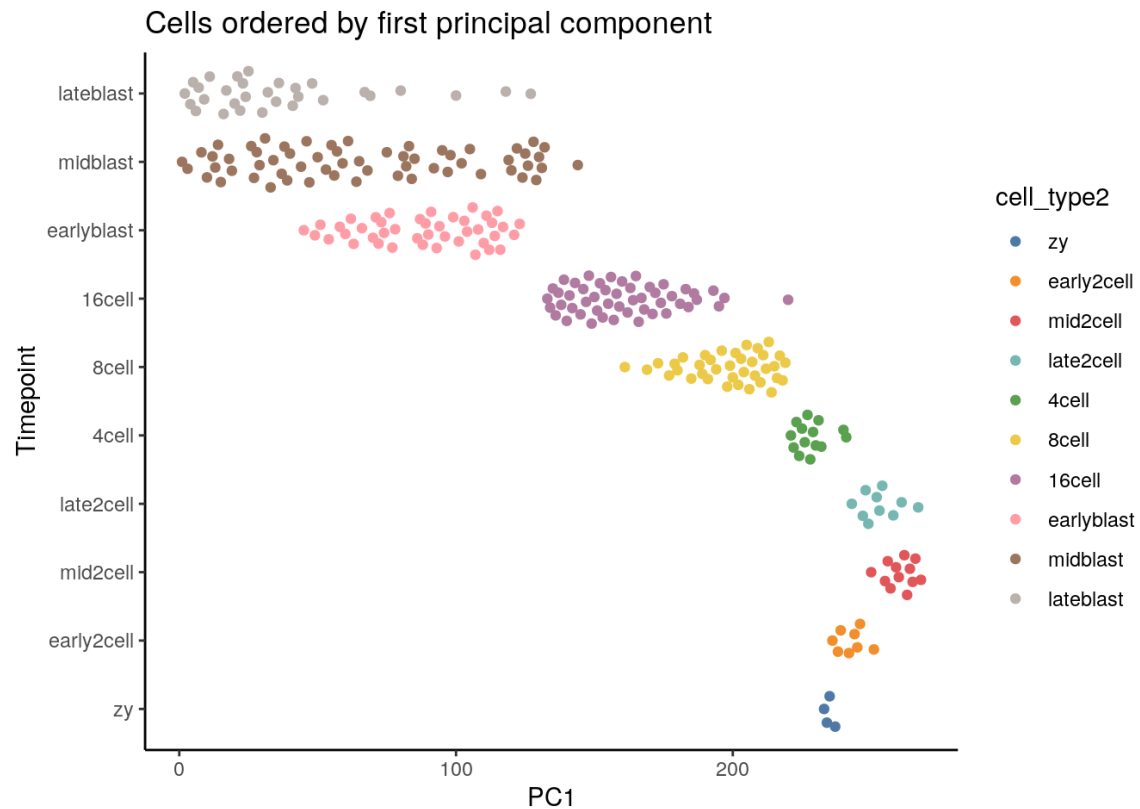
# Gazillions of point, data can be misleading



# Stacked violin plot and clustered dotplot

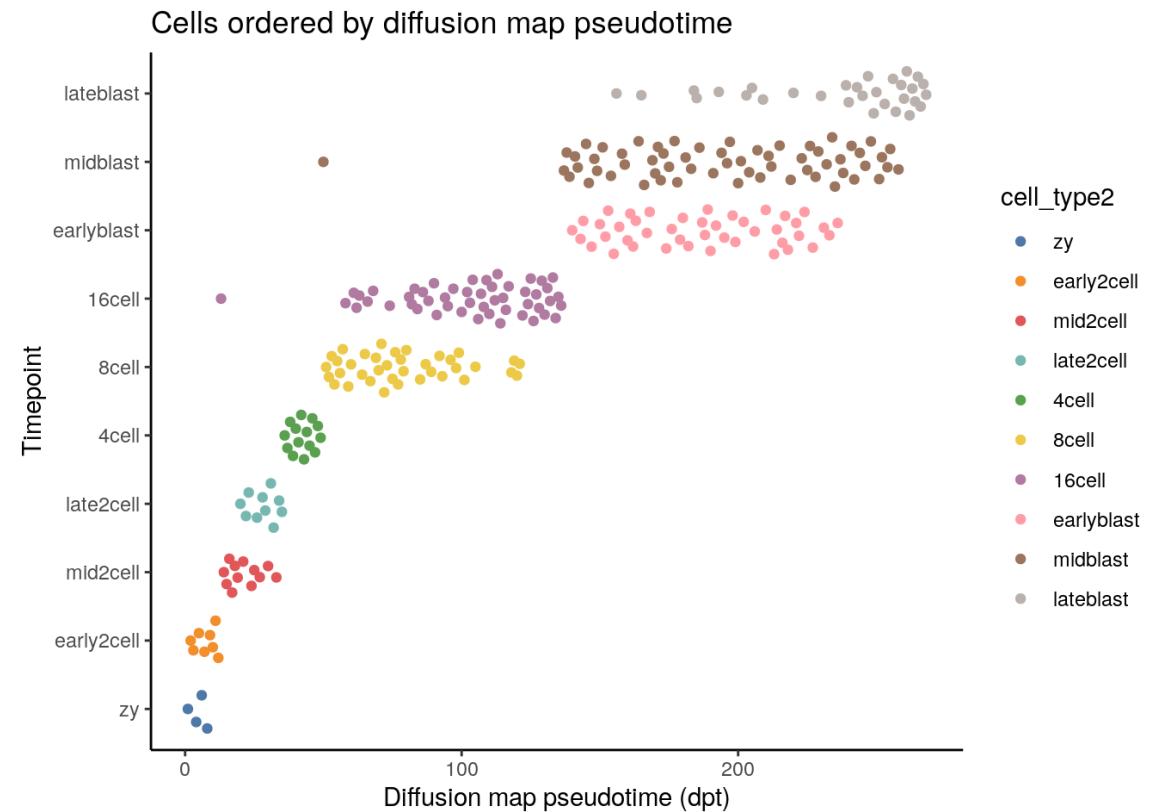
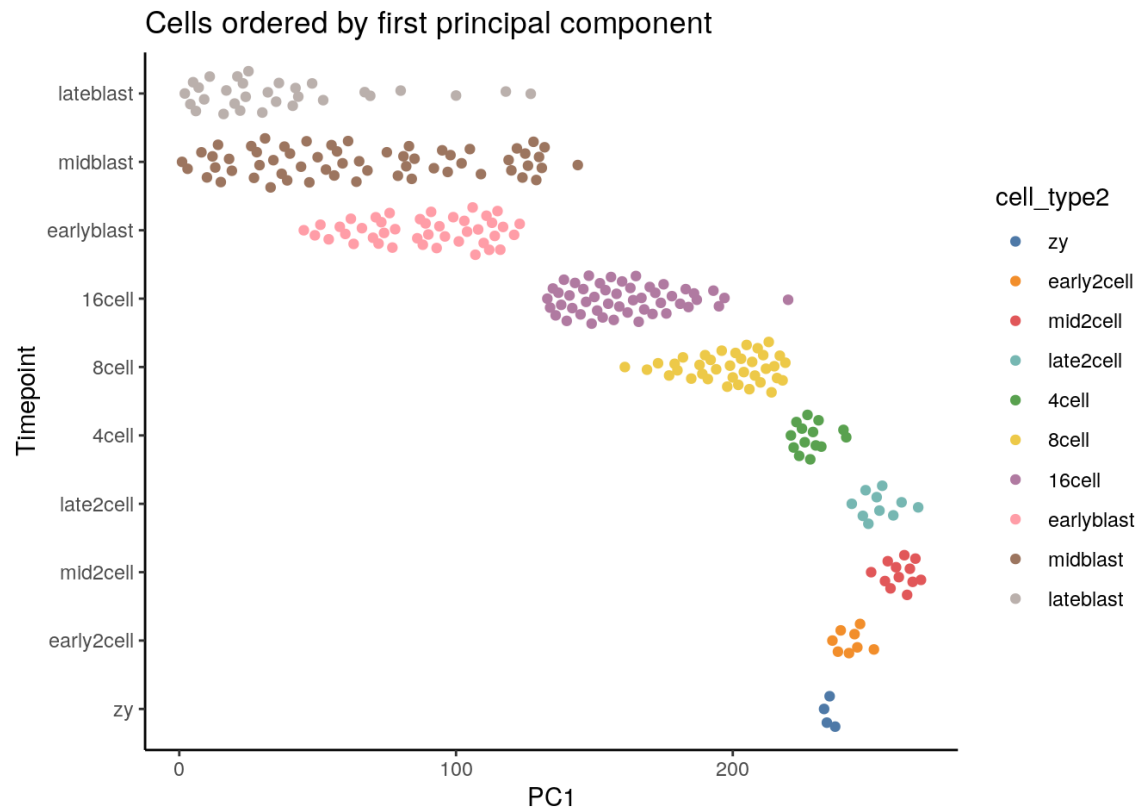


# Trajectory/pseudotime analysis





# Trajectory/pseudotime analysis



# RNA velocity

[nature](#) > [letters](#) > [article](#)

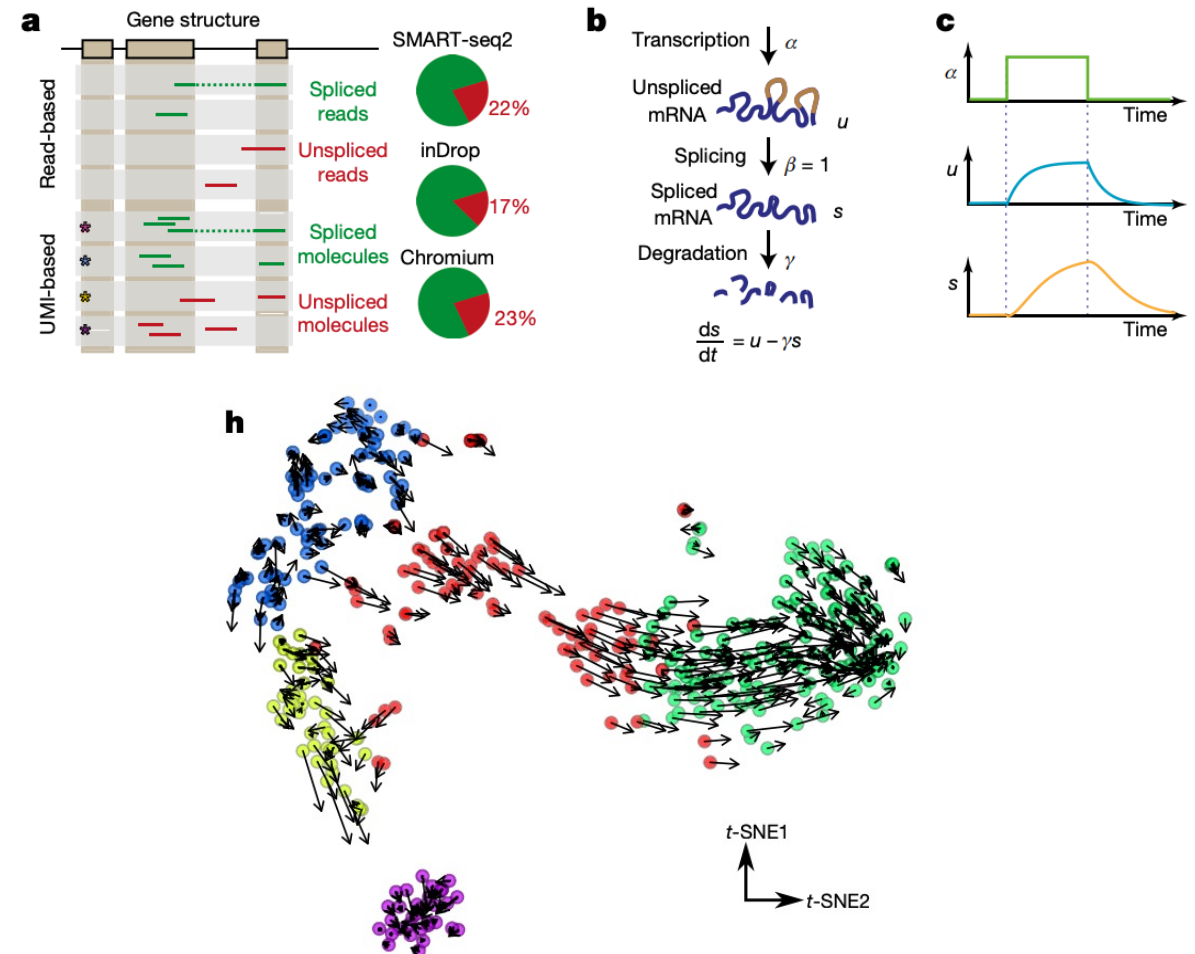
Letter | [Published: 08 August 2018](#)

## RNA velocity of single cells

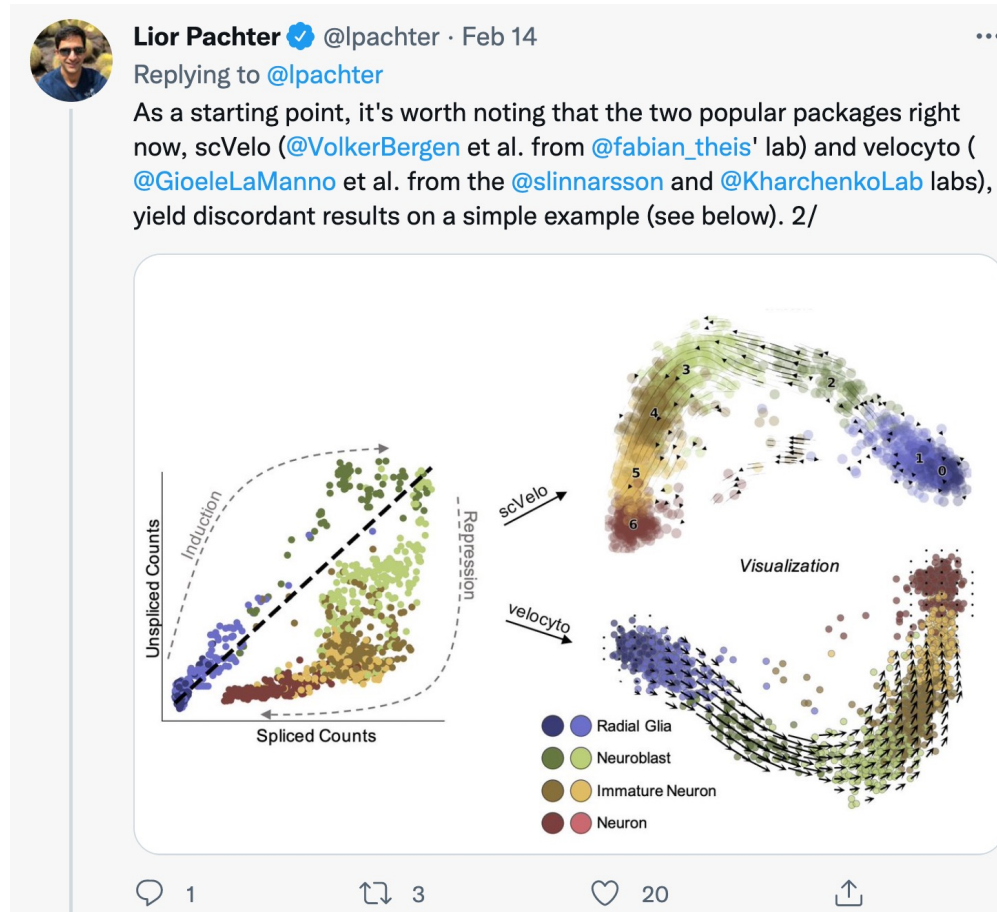
[Gioele La Manno](#), [Ruslan Soldatov](#), [Amit Zeisel](#), [Emelie Braun](#), [Hannah Hochgerner](#), [Viktor Petukhov](#),  
[Katja Lidschreiber](#), [Maria E. Kastriji](#), [Peter Lönnerberg](#), [Alessandro Furlan](#), [Jean Fan](#), [Lars E. Borm](#),  
[Zehua Liu](#), [David van Bruggen](#), [Jimin Guo](#), [Xiaoling He](#), [Roger Barker](#), [Erik Sundström](#), [Gonçalo Castelo-Branco](#),  
[Patrick Cramer](#), [Igor Adameyko](#), [Sten Linnarsson](#) ✉ & [Peter V. Kharchenko](#) ✉

[Nature](#) **560**, 494–498 (2018) | [Cite this article](#)

**156k** Accesses | **873** Citations | **670** Altmetric | [Metrics](#)



# Take a second thought on your velocity results

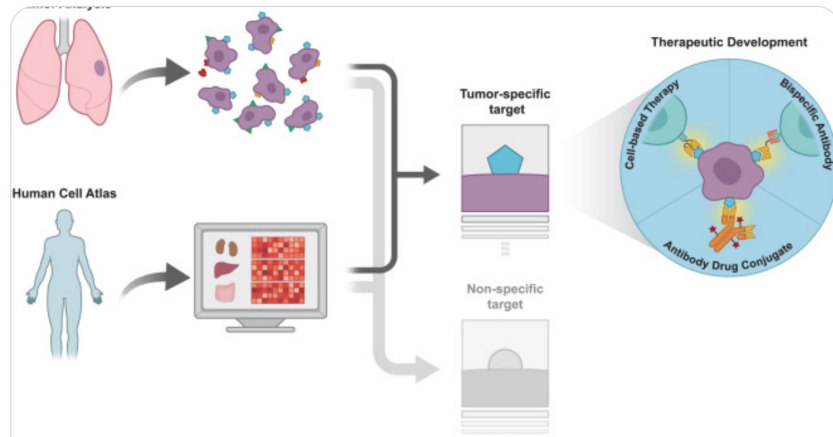


# CD4 is not expressed at high mRNA level in CD4+ cells



Ming "Tommy" Tang  
@tangming2005

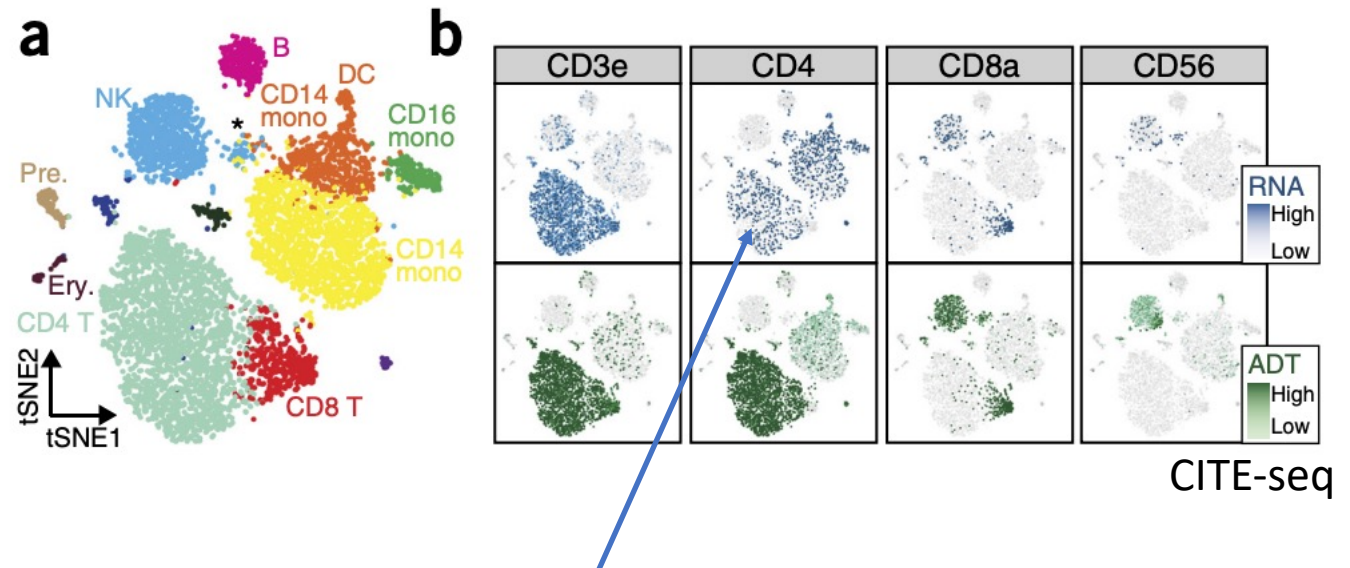
1/ a question on CD4 mRNA vs protein. @CalebLareau  
I saw "CD4+ T cells express low levels of the CD4 transcript but very high levels of CD4 protein (Stoeckius et al., 2017)" in your paper



cell.com

Charting the tumor antigen maps drawn by single-cell genomics

The specificity of antibodies makes cancer immunotherapies, including chimeric antigen receptor T cells and antibody-drug conjugates, possible. In parallel, ...



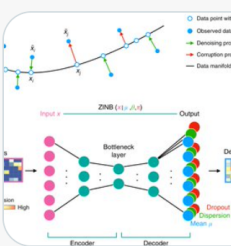
<https://twitter.com/tangming2005/status/1501766040686108678>

<https://www.nature.com/articles/nmeth.4380>

<https://rpubs.com/MattPM/cd4>

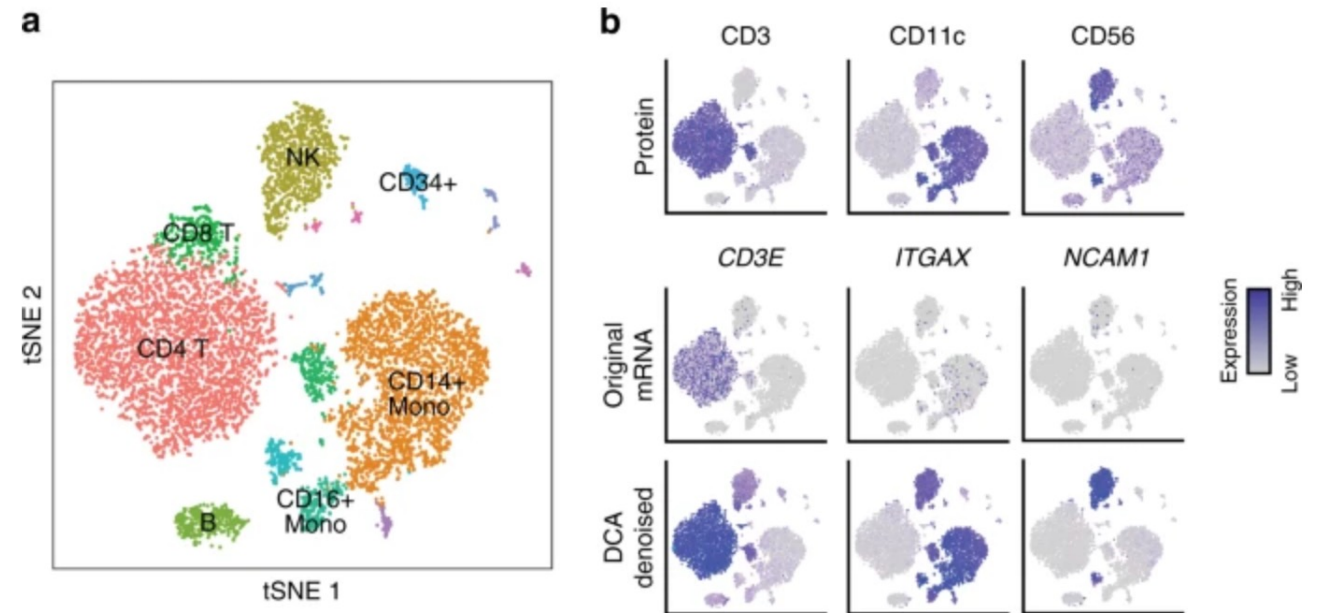
# CD56 is not expressed at high mRNA level in NK cells

**Ergün Tiryaki** @ErgnTiryaki · Mar 10  
Replying to @tangming2005  
@tomsgoms I think the same situation also applies to NCAM1 (CD56) mRNA in NK cells. Although Smart-seq2 captures more NCAM1 than 10X, it is still very low and zero for most of the NK cells. Fig 6B shows the NCAM1 mRNA and CD56 in NK cluster

nature.com  
Single-cell RNA-seq denoising using a deep count ...  
Nature Communications - Single-cell RNA sequencing is a powerful method to study gene ...

**Ming "Tommy" Tang** @tangming2005 · Mar 10  
Yes! Had the same experience with CD56 myself.

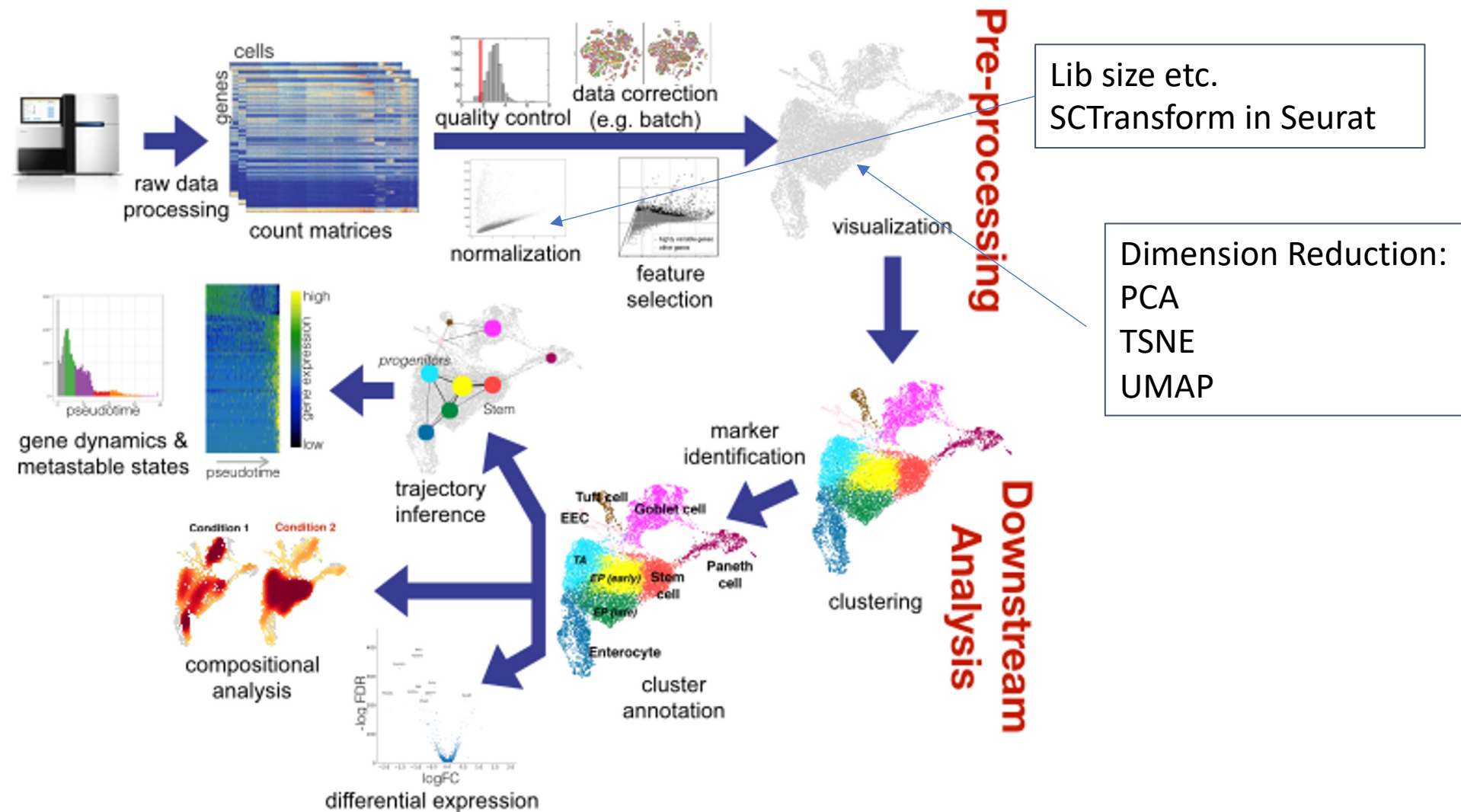
**Fig. 6**



<https://www.nature.com/articles/s41467-018-07931-2>



# Let's ~~walk~~ sprint through a typical\* scRNA-seq analysis



Credit to Peter Hickey



# Other resources

## Orchestrating Single-Cell Analysis with Bioconductor

**Authors:** Robert Amezquita [aut], Aaron Lun [aut, cre], Stephanie Hicks [aut], Raphael Gottardo [aut]

**Version:** 1.4.1

**Modified:** 2022-01-06

**Compiled:** 2022-01-07

**Environment:** R version 4.1.2 (2021-11-01), Bioconductor 3.14

**License:** CC BY 4.0

**Copyright:** Bioconductor, 2020

**Source:** <https://github.com/LTLA/OSCA>

## Welcome

This is the landing page for the “**Orchestrating Single-Cell Analysis with Bioconductor**” book, which teaches users some common workflows for the analysis of single-cell RNA-seq data (scRNA-seq). This book will show you how to make use of cutting-edge Bioconductor tools to process, analyze, visualize, and explore scRNA-seq data. Additionally, it serves as an online companion for the [paper of the same name](#).



## What you will learn

<https://bioconductor.org/books/release/OSCA/>

## nature methods

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾ [Subscribe](#)

[nature](#) > [nature methods](#) > [review articles](#) > [article](#)

Review Article | [Published: 21 June 2021](#)

## The triumphs and limitations of computational methods for scRNA-seq

[Peter V. Kharchenko](#) 

[Nature Methods](#) **18**, 723–732 (2021) | [Cite this article](#)

**18k** Accesses | **4** Citations | **240** Altmetric | [Metrics](#)

<https://github.com/seandavi/awesome-single-cell>

[https://github.com/mdozmorov/scRNA-seq\\_notes](https://github.com/mdozmorov/scRNA-seq_notes)

<https://github.com/crazyhottommy/scRNAseq-analysis-notes>

<https://liulab-dfci.github.io/bioinfo-combio/scatac.html>

# Acknowledgements

DFCI:  
Shirley Liu  
Margaret Shipp

Almighty Tweeps

Harvard FAS informatics:  
Tim Sackton

Jackson Lab:  
Roel Verhaak  
Samir Amin

What questions do you have?