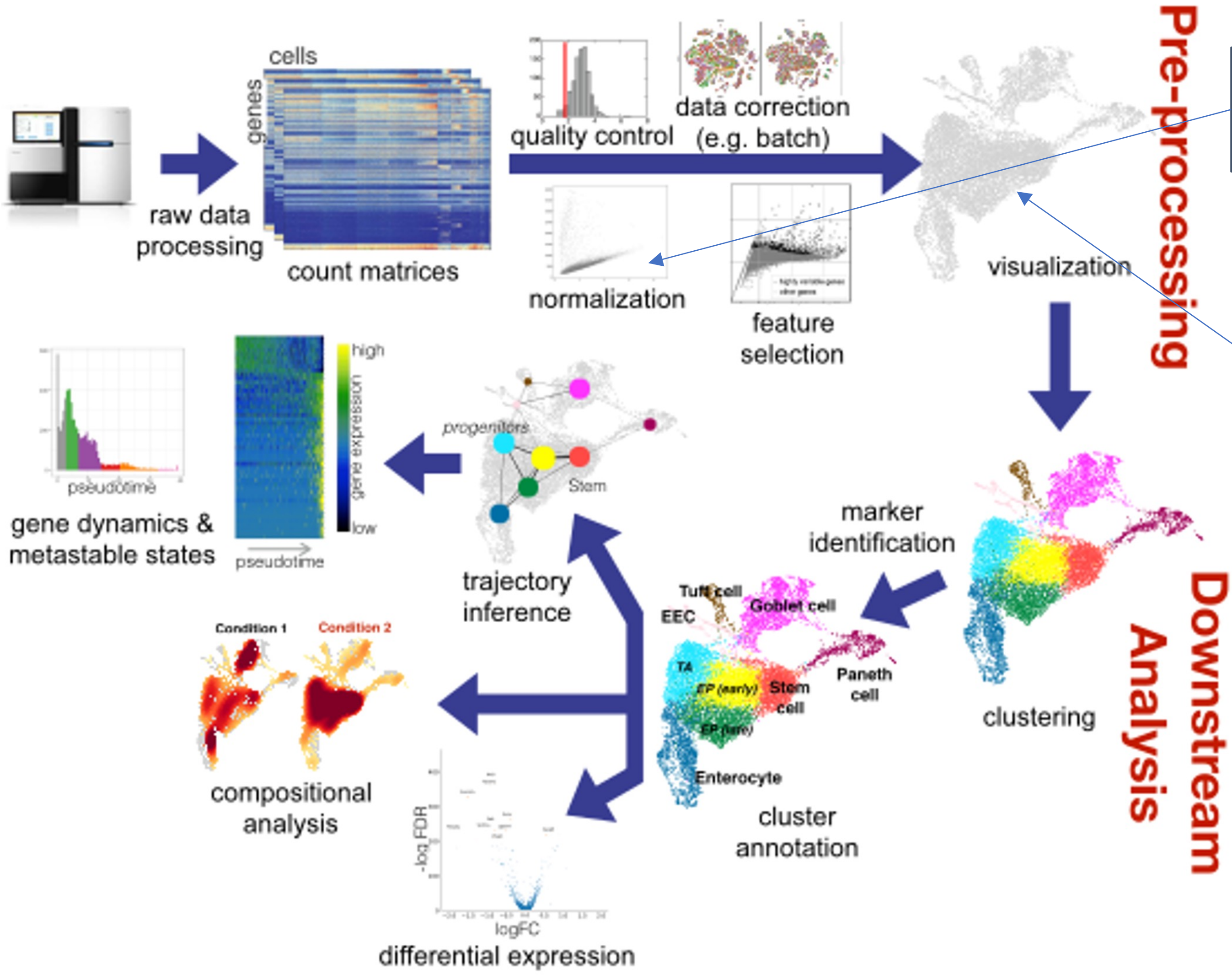
Single-cell analysis: best practices and challenges

Ming 'Tommy' Tang
Director of Bioinformatics at AstraZeneca
X: @tangming2005
https://divingintogeneticsandgenomics.com/
YouTube: chatomics
09/25/2024

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



Lib size etc.
SCTransform in Seurat

Dimension Reduction:
PCA
TSNE
UMAP

Credit to Peter Hickey

[Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).]

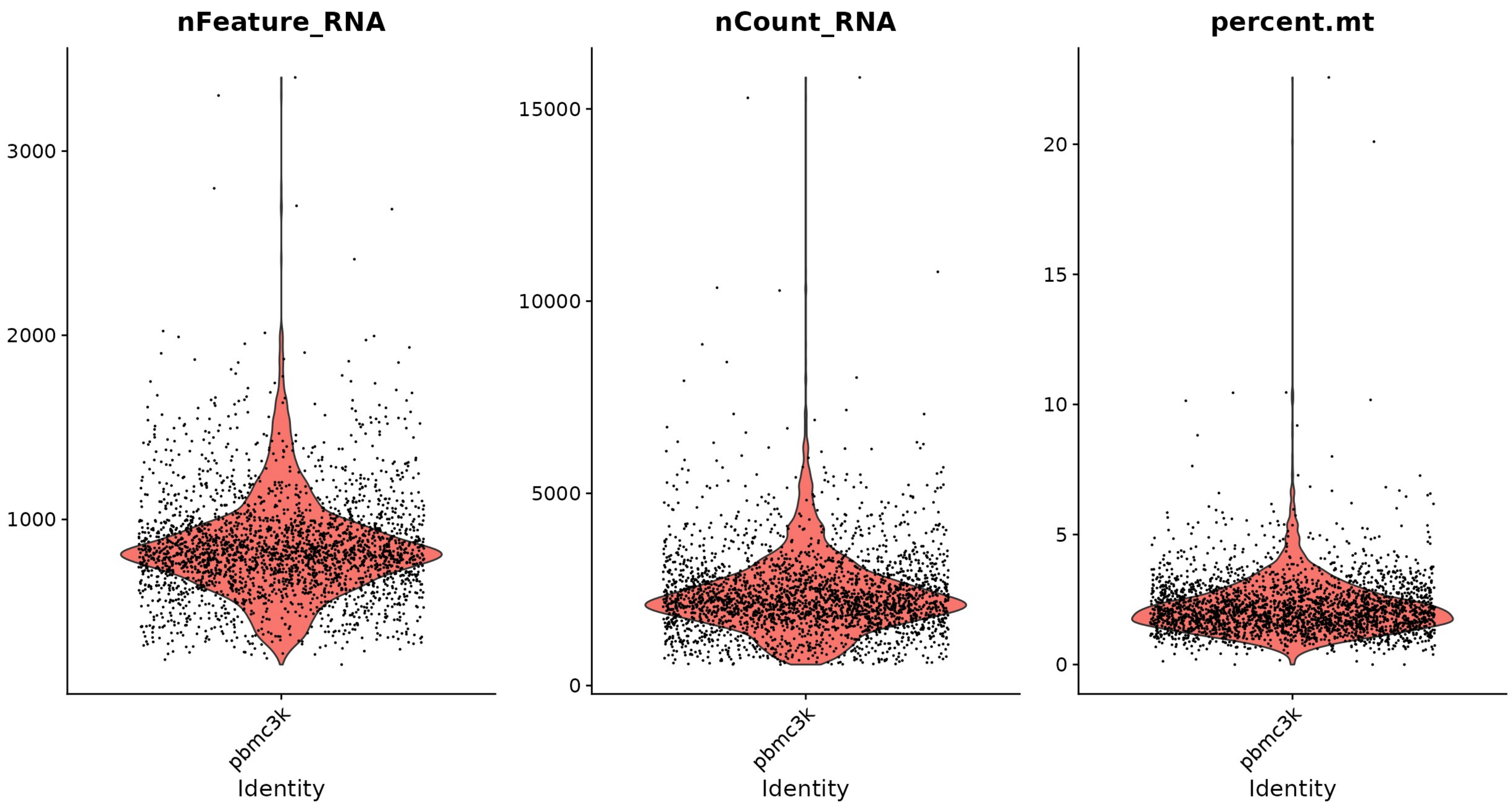
Sparse matrix

	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

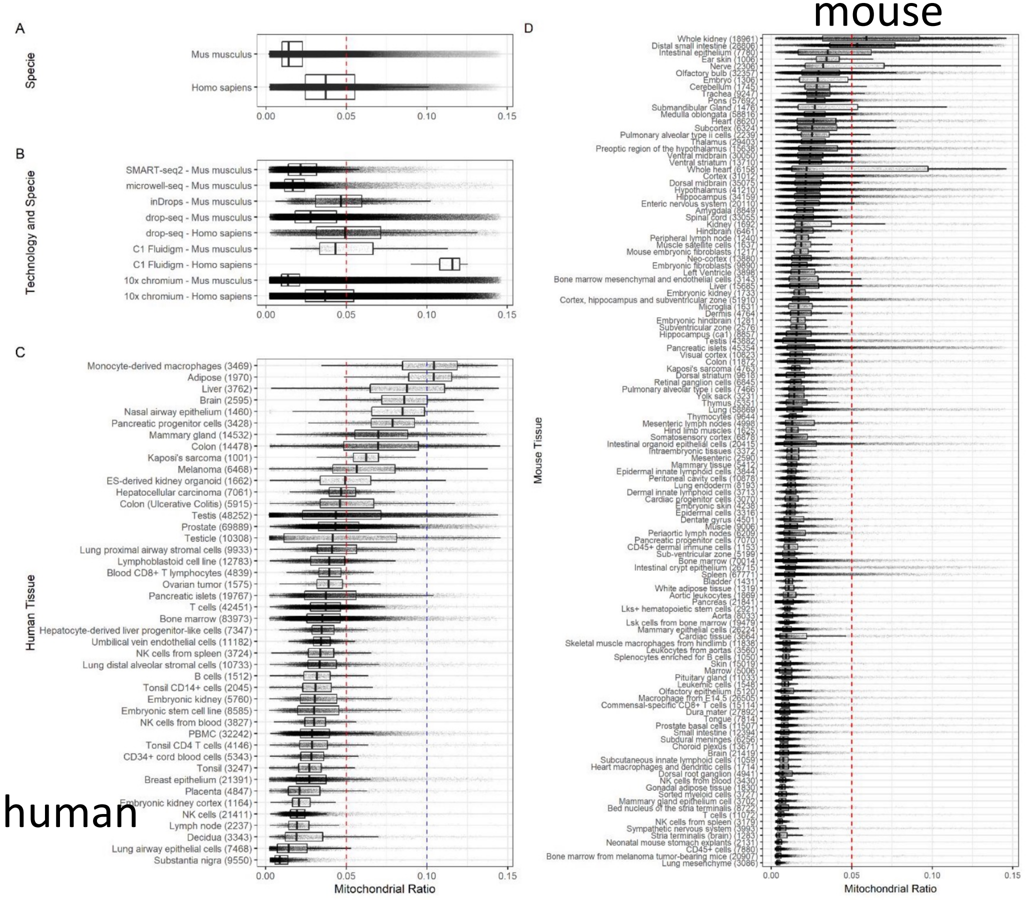
https://www.nature.com/articles/s41596-018-0073-y

Spend time for Quality control



Mitochondrial gene content cutoff

mouse



human

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.

Osorio et al 2020 Bioinformatics

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.

Doublet detection and ambient RNA

- DoubletFinder https://github.com/chris-mcginnis-ucsf/
- Scrublet - https://github.com/AllonKleinLab/scrublet
- DoubletCell in Scran::DoubletCell
- https://github.com/broadinstitute/CellBender
- https://github.com/constantAmateur/SoupX

Normalization and scaling

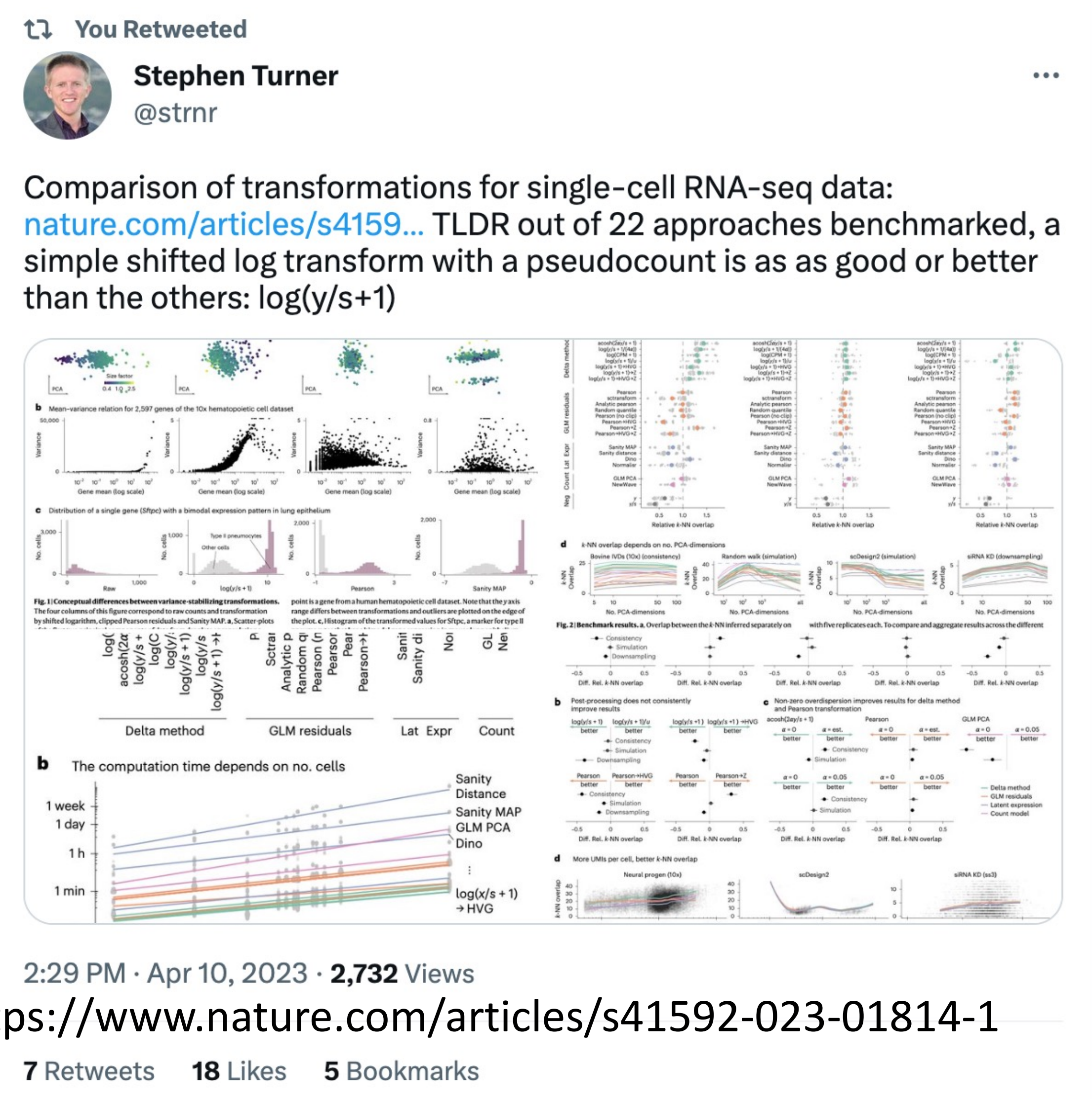
- Bulk-RNAseq
  - Reads per kilobase of exon per million reads mapped (RPKM)
  - Transcript per million (TPM)

- Single-cell RNAseq
  - LogNormalize: log(n/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

Normalization cont't

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

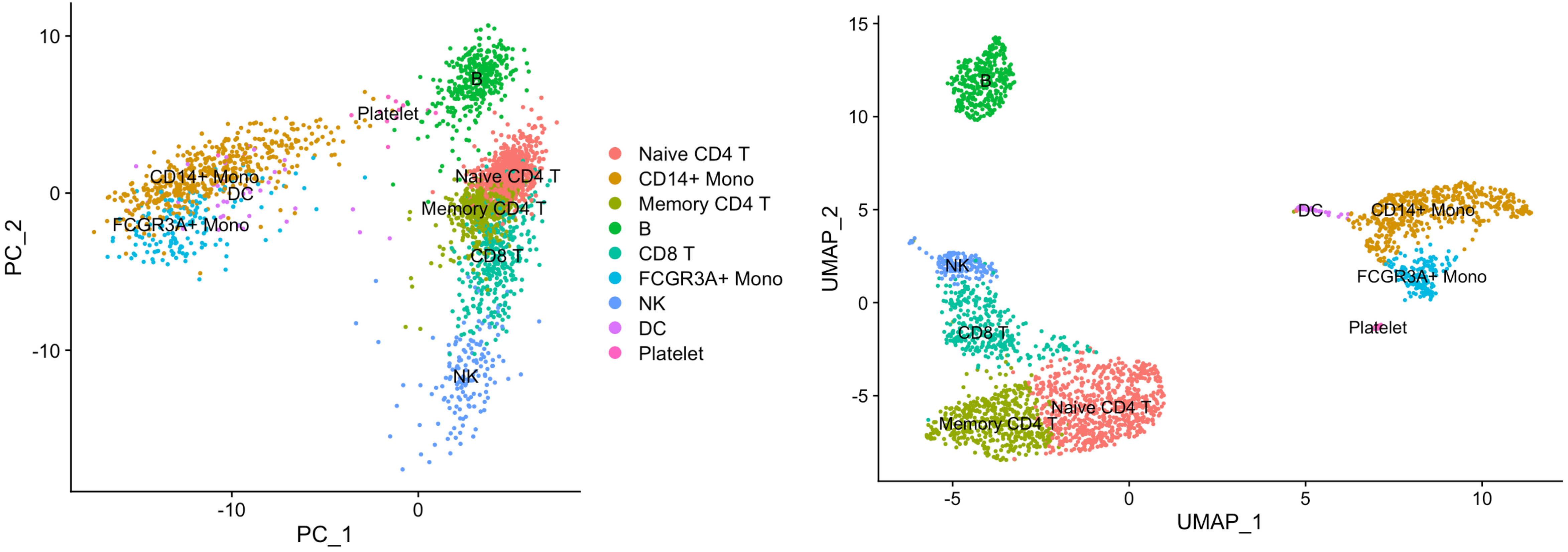


https://www.nature.com/articles/s41592-023-01814-1

Normalize to library size and log transform

More sophisticated methods: SCTransform in Seurat

Dimension reduction (PCA vs UMAP)



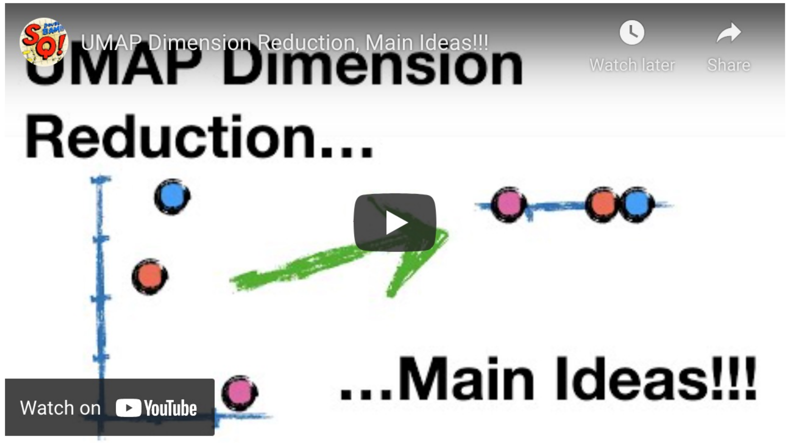
https://divingintogeneticsandgenomics.rbind.io/post/pca-in-action/

UMAP and TSNE

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

UMAP Dimension Reduction: Part 1 – Main Ideas

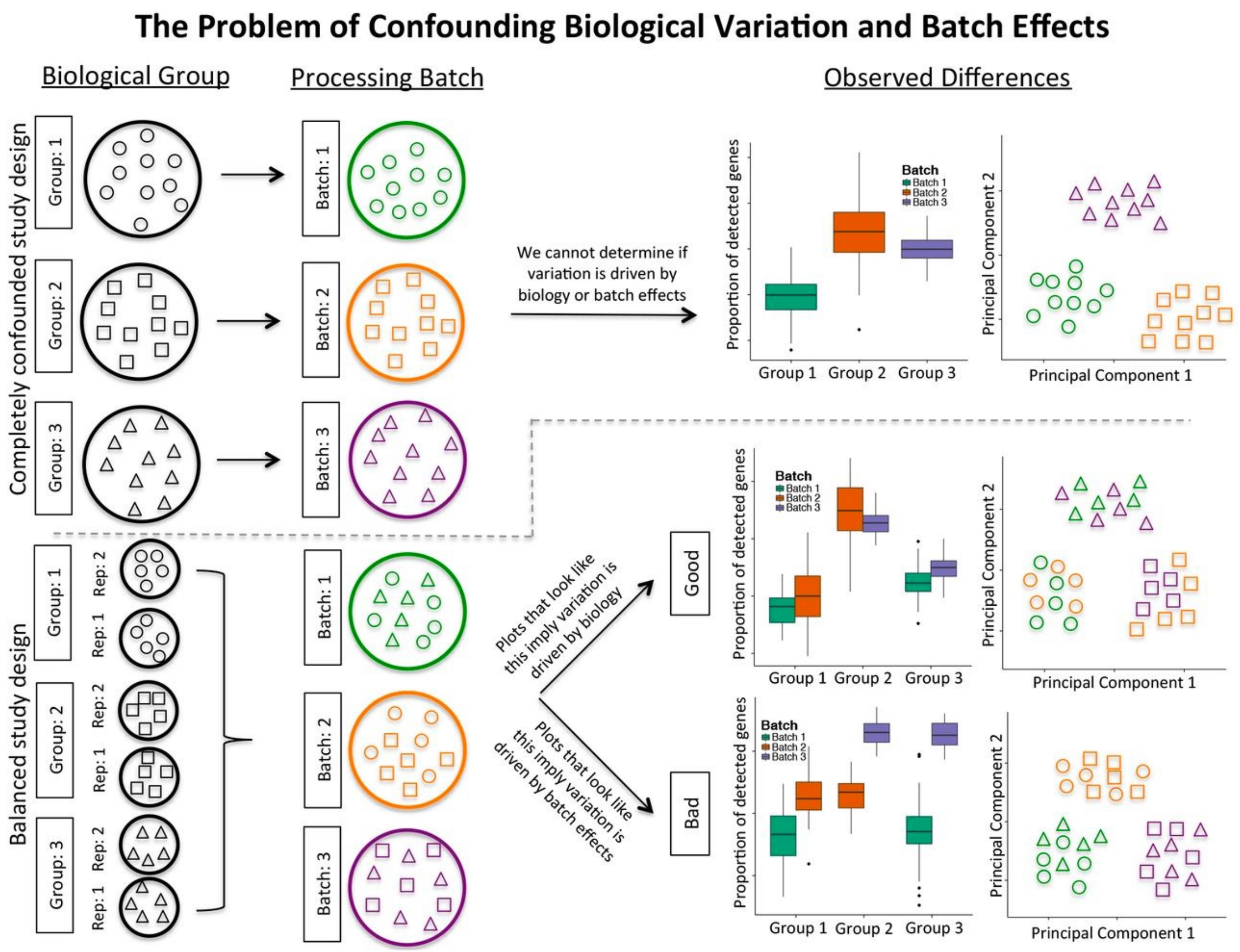
March 7, 2022



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

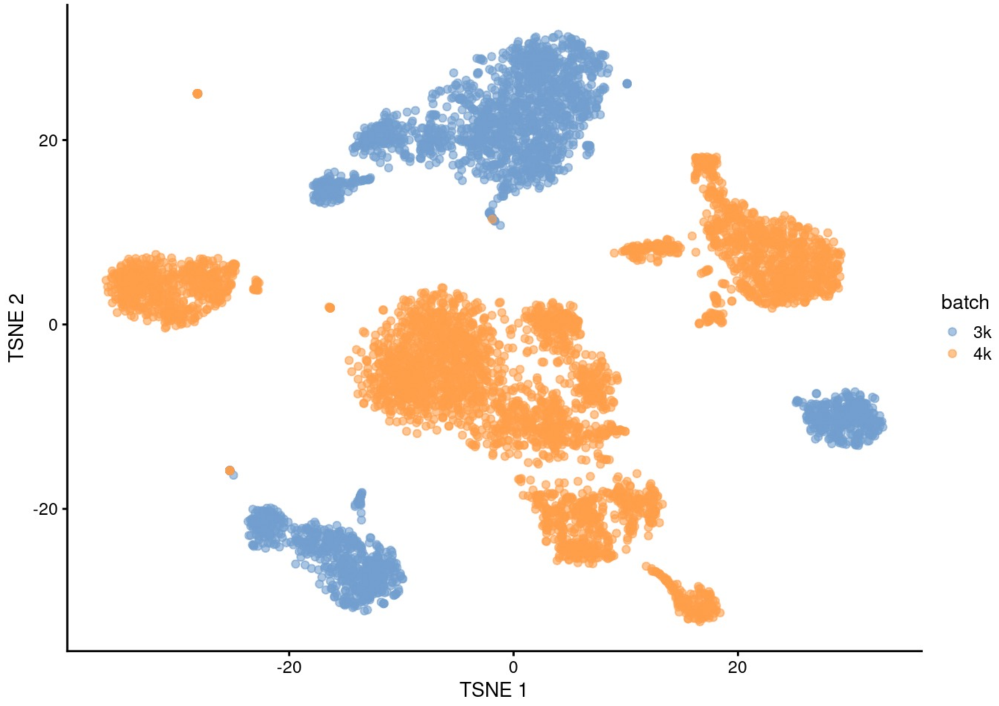
https://statquest.org/

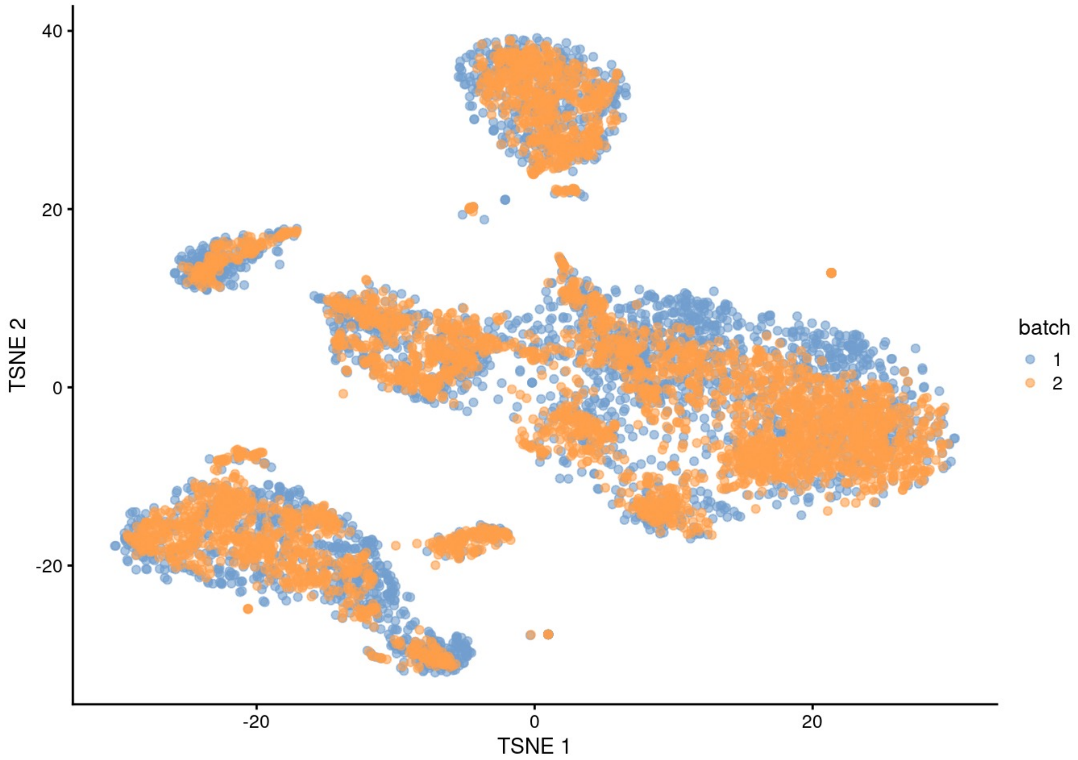
Avoid batch and confounding effects: experimental design



Hicks *et al.*, *Biostatistics*. 2018

Data integration/batch correction

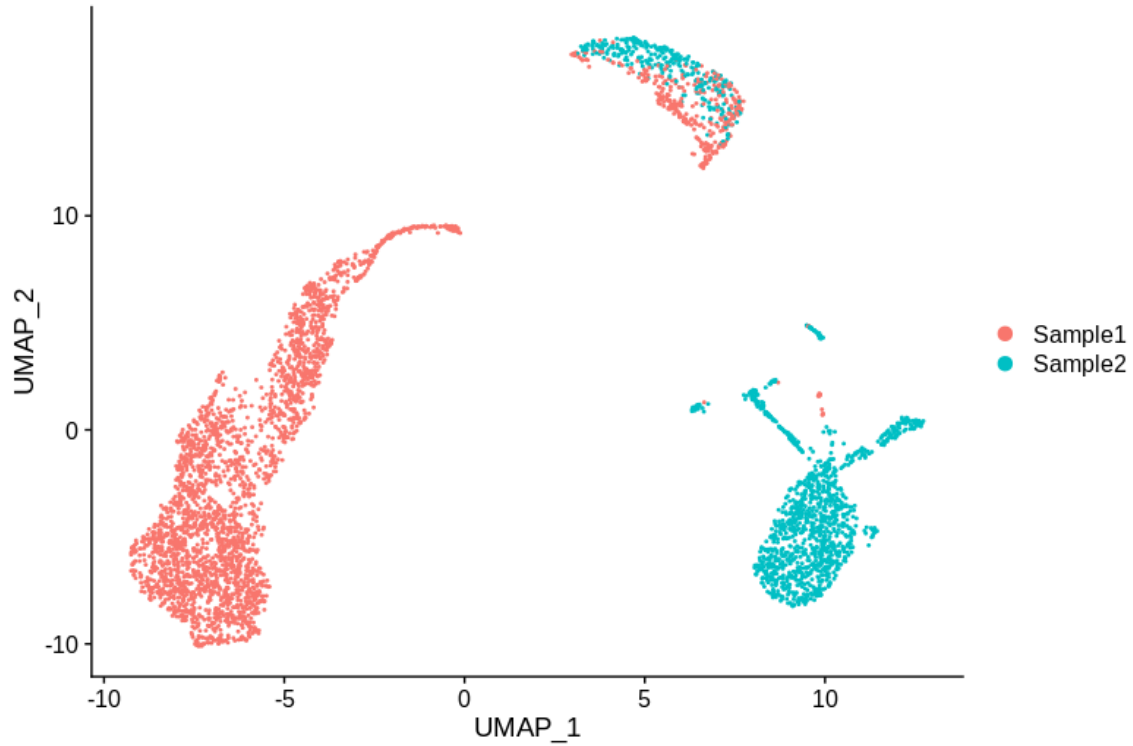


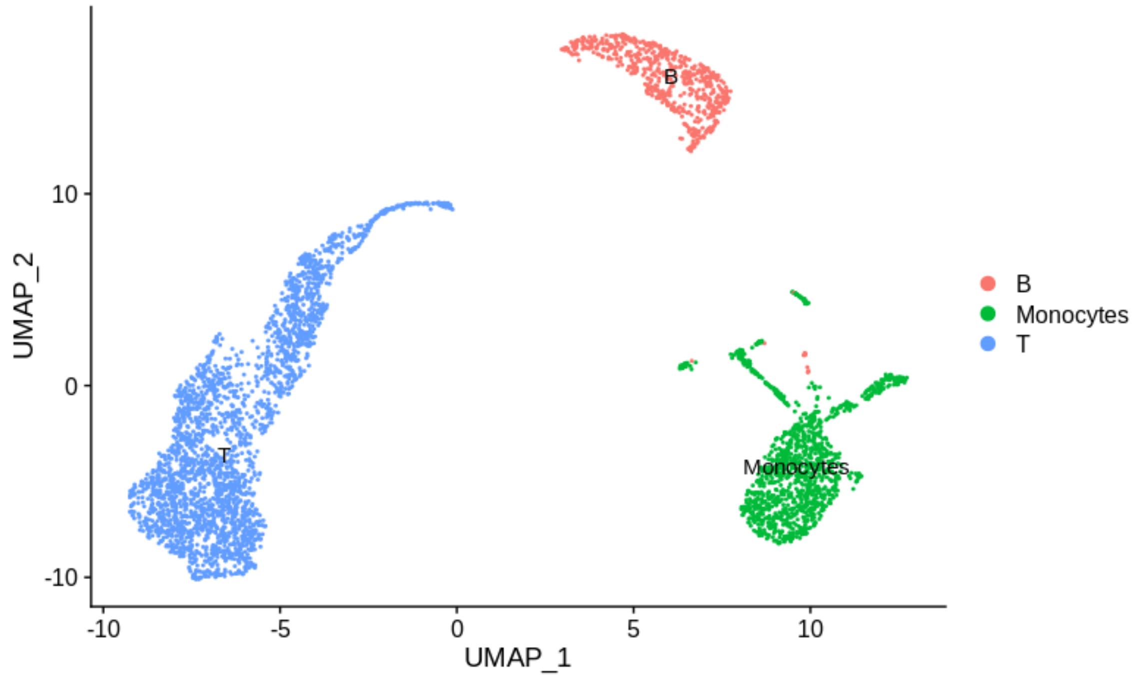


http://bioconductor.org/books/3.14/OSCA.multisample/integrating-datasets.html#motivation

Data integration

- Batch effect or not? Correct or not





https://constantamateur.github.io/2020-06-09-scBatch1/

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.

New Results

PMD Uncovers Widespread Cell-State Erasure by scRNAseq Batch Correction Methods

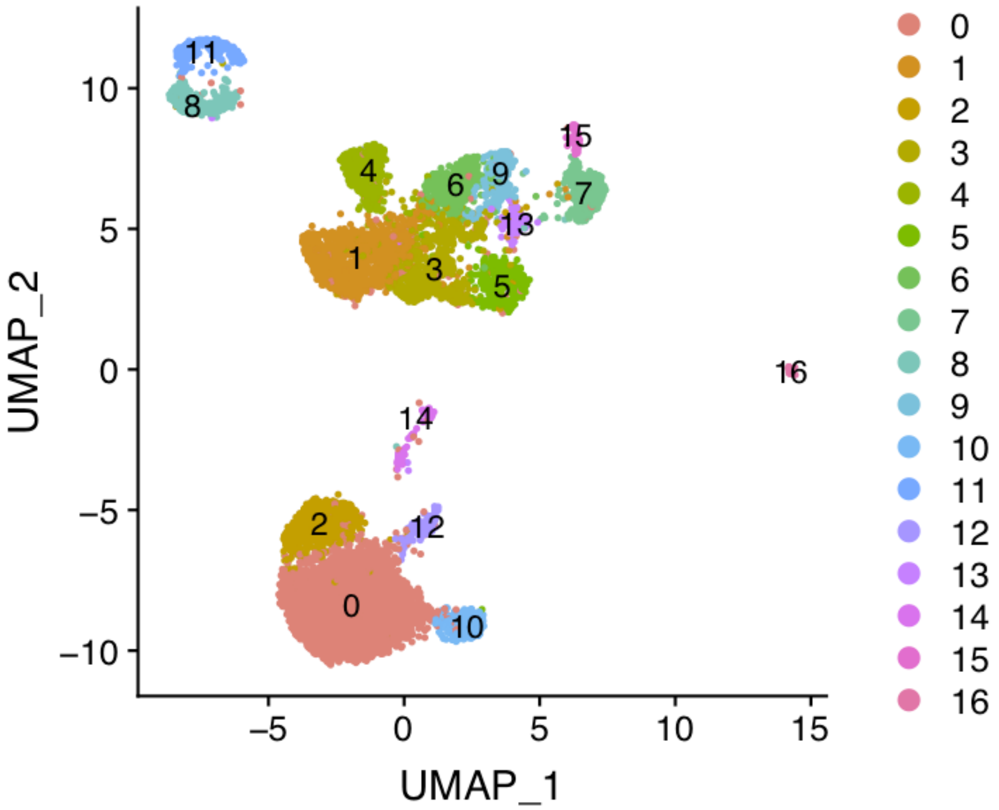
Scott R Tyler, Supinda Bunyavanich, Eric E Schadt

doi: https://doi.org/10.1101/2021.11.15.468733

http://bioconductor.org/books/3.14/OSCA.multisample/differential-abundance.html#sacrificing-differences

Clustering

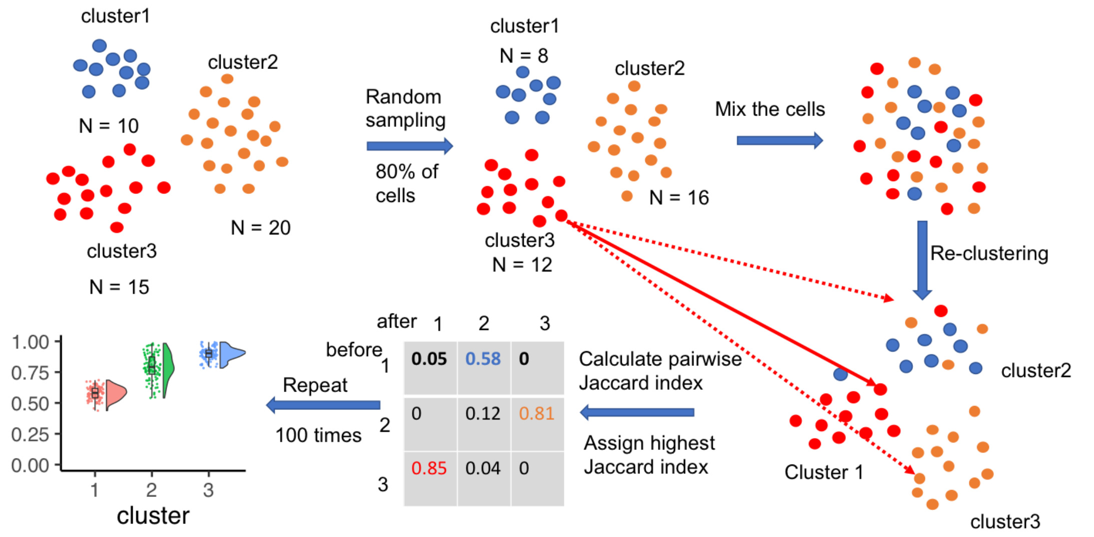
- Dimension reduction (PCA)
- k-means, hierarchical clustering etc
- Cluster cells (on the reduced dimensions) using graph-based method in Seurat v3 (Stuart et al, Cell 2019). KNN graph + community detection algorithm
- Can visualize using t-SNE / UMAP



Evaluating cluster stability

5.4 Evaluating cluster stability

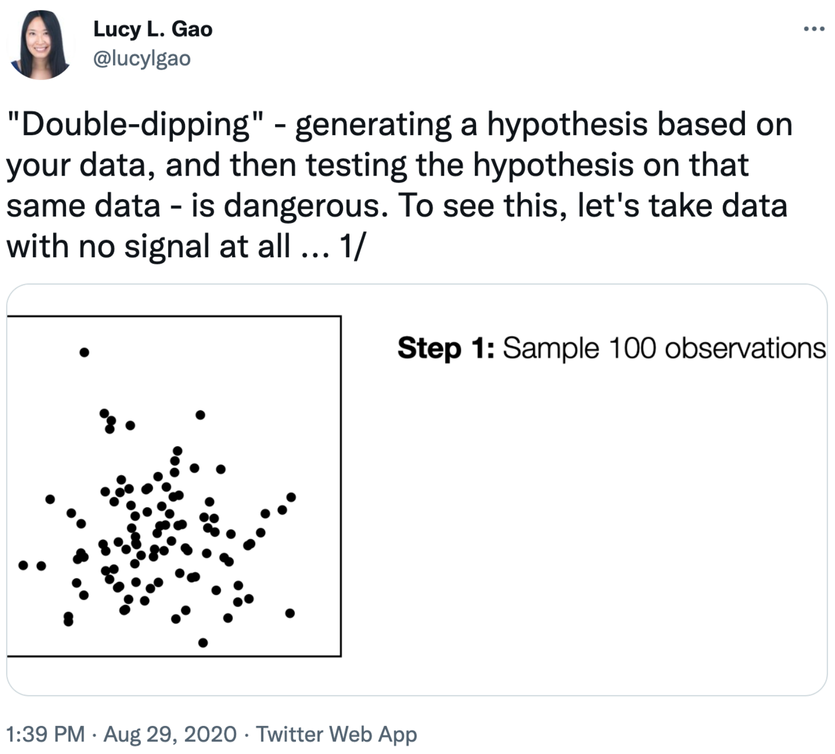
A desirable property of a given clustering is that it is stable to perturbations to the input data (Von Luxburg 2010). Stable clusters are logistically convenient as small changes to upstream processing will not change the conclusions; greater stability also increases the likelihood that those conclusions can be reproduced in an independent replicate study. *scran* 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.

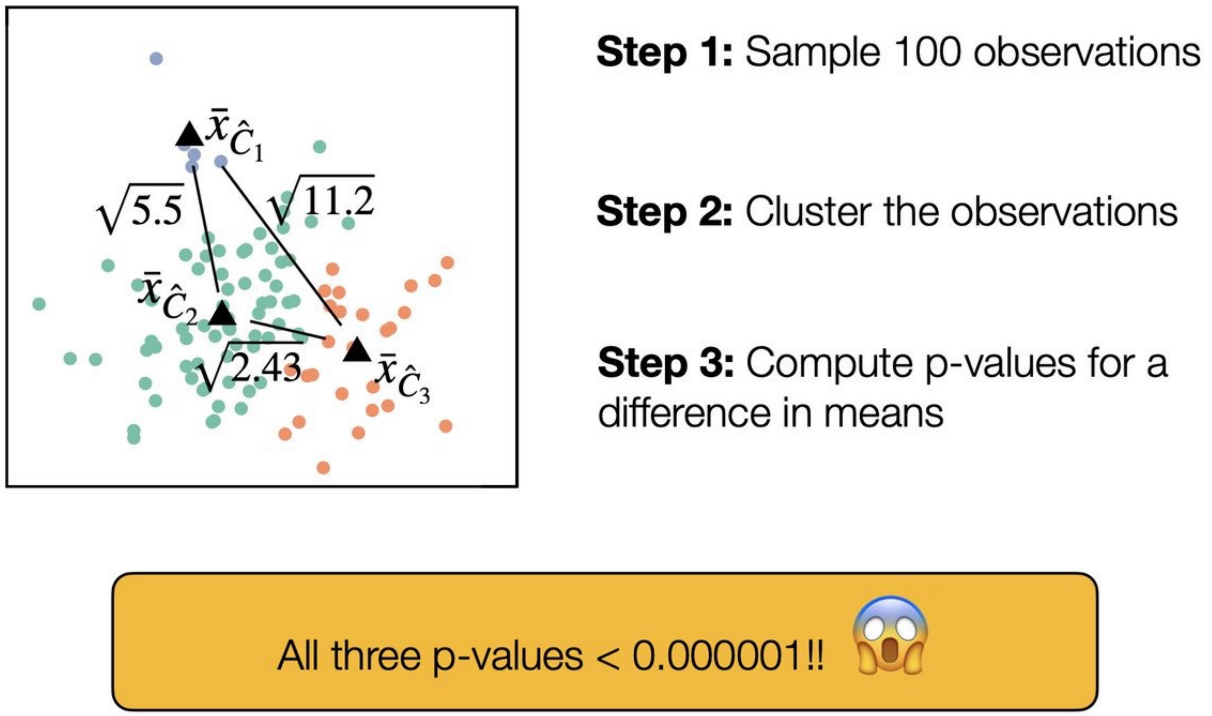


Tang et al 2021 Bioinformatics

http://bioconductor.org/books/3.14/OSCA.advanced/clustering-redux.html#cluster-bootstrapping
https://github.com/crazyhottommy/scclusteval
https://divingintogeneticsandgenomics.com/post/scrnaseq-clustering-significant-test-an-unsolvable-problem/. scSHC
https://divingintogeneticsandgenomics.com/post/fine-tune-the-best-clustering-resolution-for-scrnaseq-data-trying-out-callback/

Marker gene p-value is inflated





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

https://www.youtube.com/watch?v=voseWZIaFm4

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

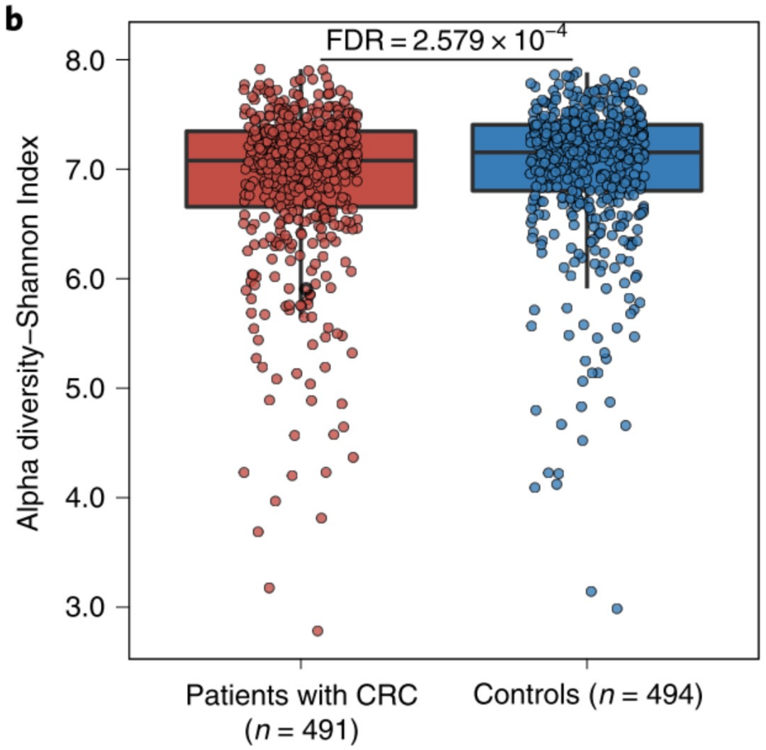
Large number of data points will make p-value tiny

Ming "Tommy" Tang
@tangming2005

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

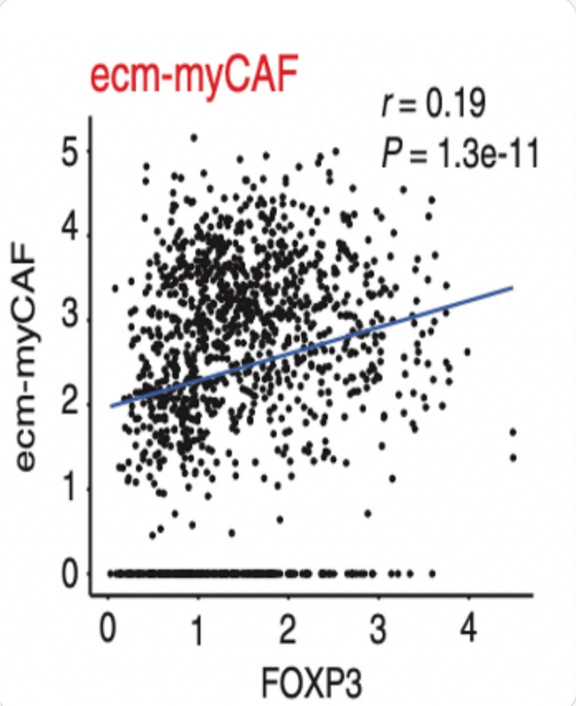
Daniel Martínez @dan_martimarti · Feb 4

This effect size...



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

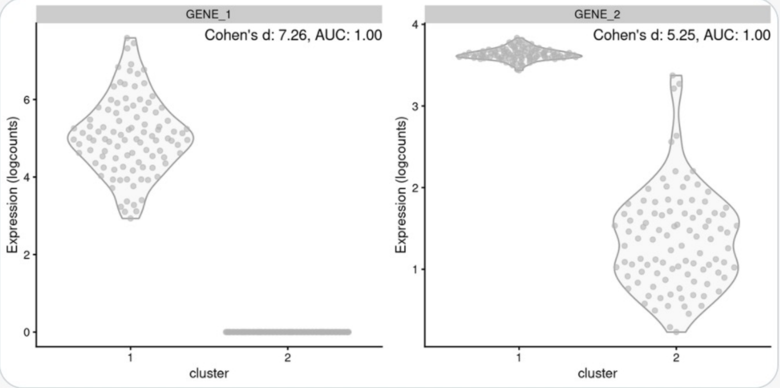
Question: if you have tens of thousands of data points with a **correlation** of 0.2 and a p-value 10^-11. Is it meaningful to show that? you always get a tiny p-value when you have a lot of data points.



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.

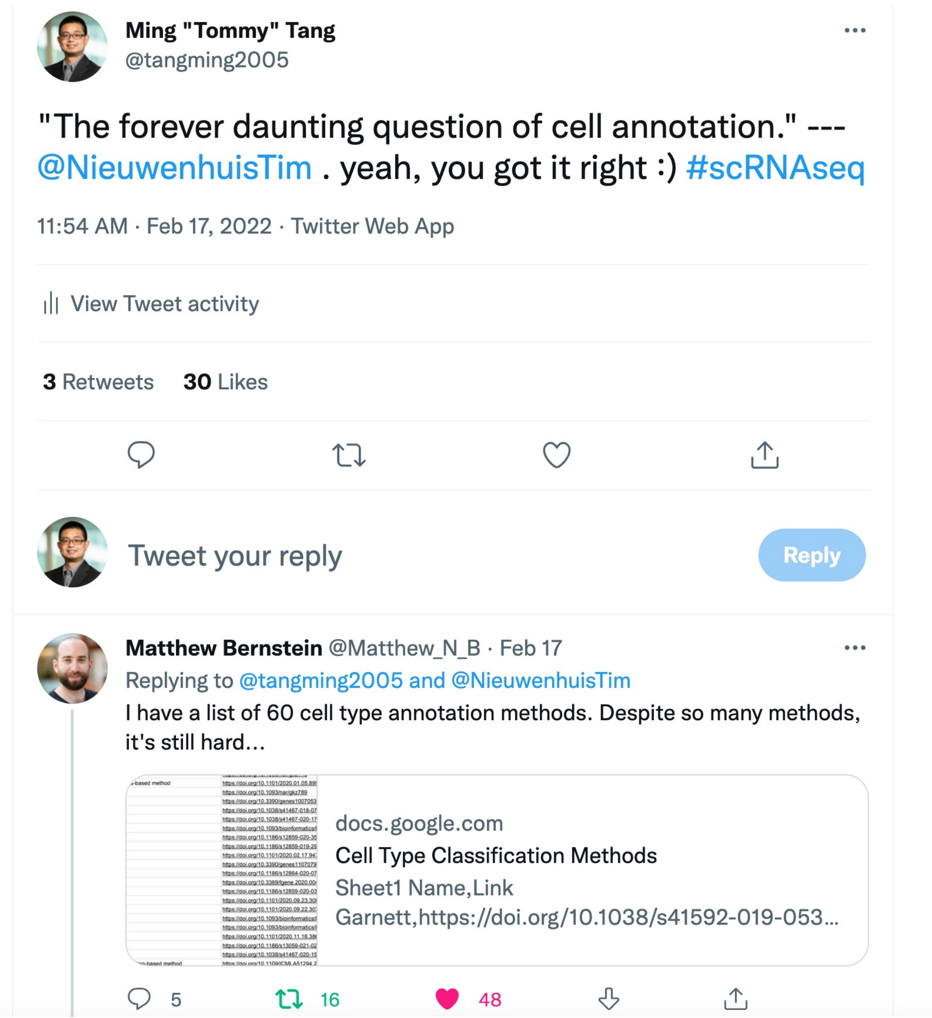
(Image source: tinyurl.com/y8fqlkkx)



1 4

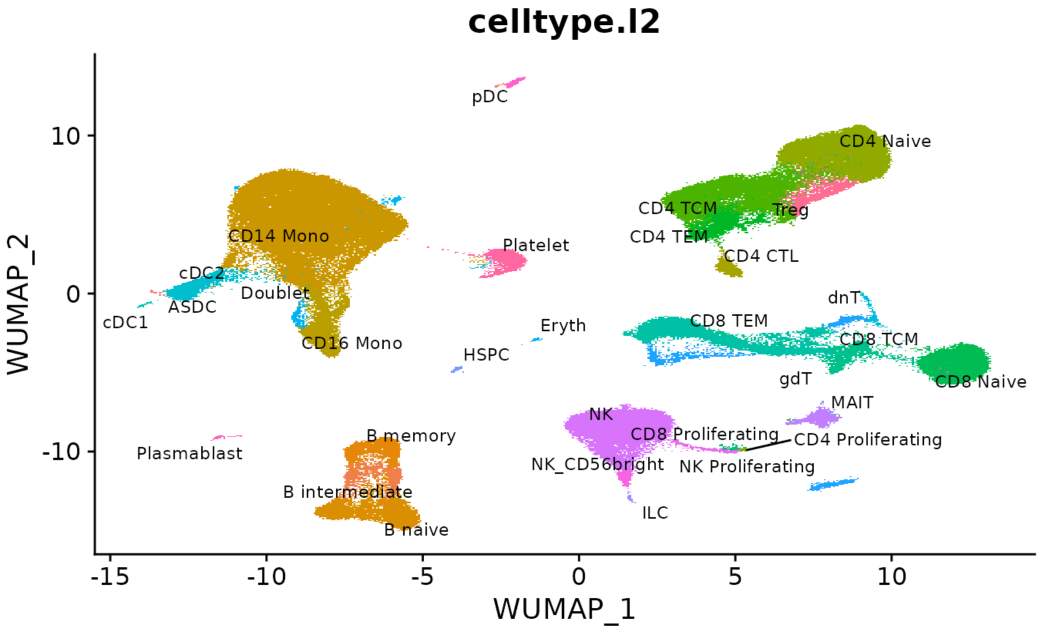
https://twitter.com/tangming2005/status/1489964367336648707
https://mobile.twitter.com/mikhaeldito313/status/1505204061506715649

Cell annotation



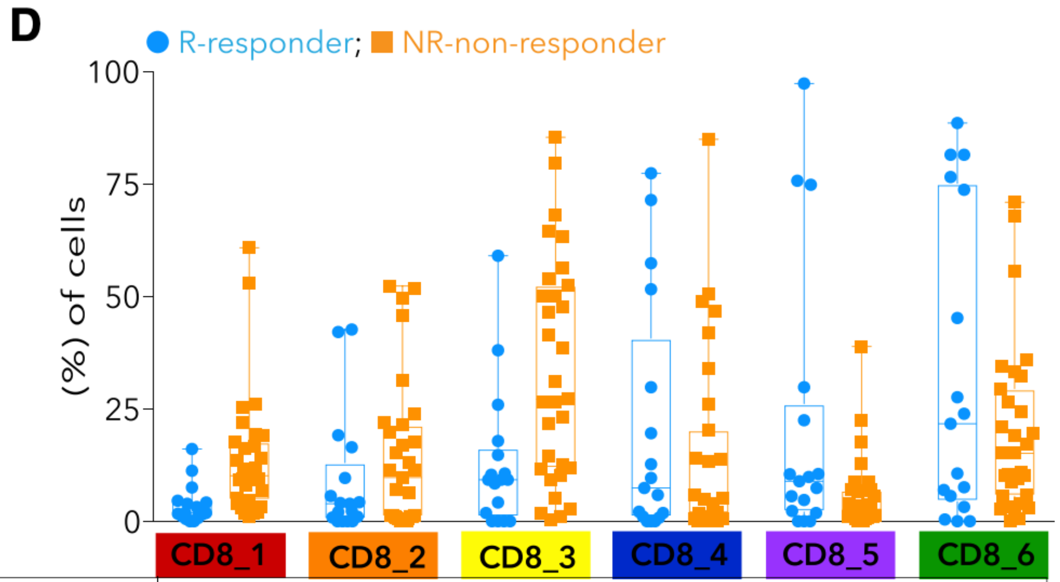
SingleR

Seurat V4 reference based mapping



https://twitter.com/tangming2005/status/1494354585602998274 https://bioconductor.org/books/3.14/SingleRBook/

Differential cell abundance analysis



	CD8_1	CD8_2	CD8_3	CD8_4	CD8_5	CD8_6
All samples	6.7×10^{-5}	n.s	0.001	n.s	n.s	n.s
Ag presentation	1×10^{-5}	n.s	6.6×10^{-5}	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
```

http://bioconductor.org/books/3.14/OSCA.multisample/differential-abundance.html#overview

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

nature COMMUNICATIONS

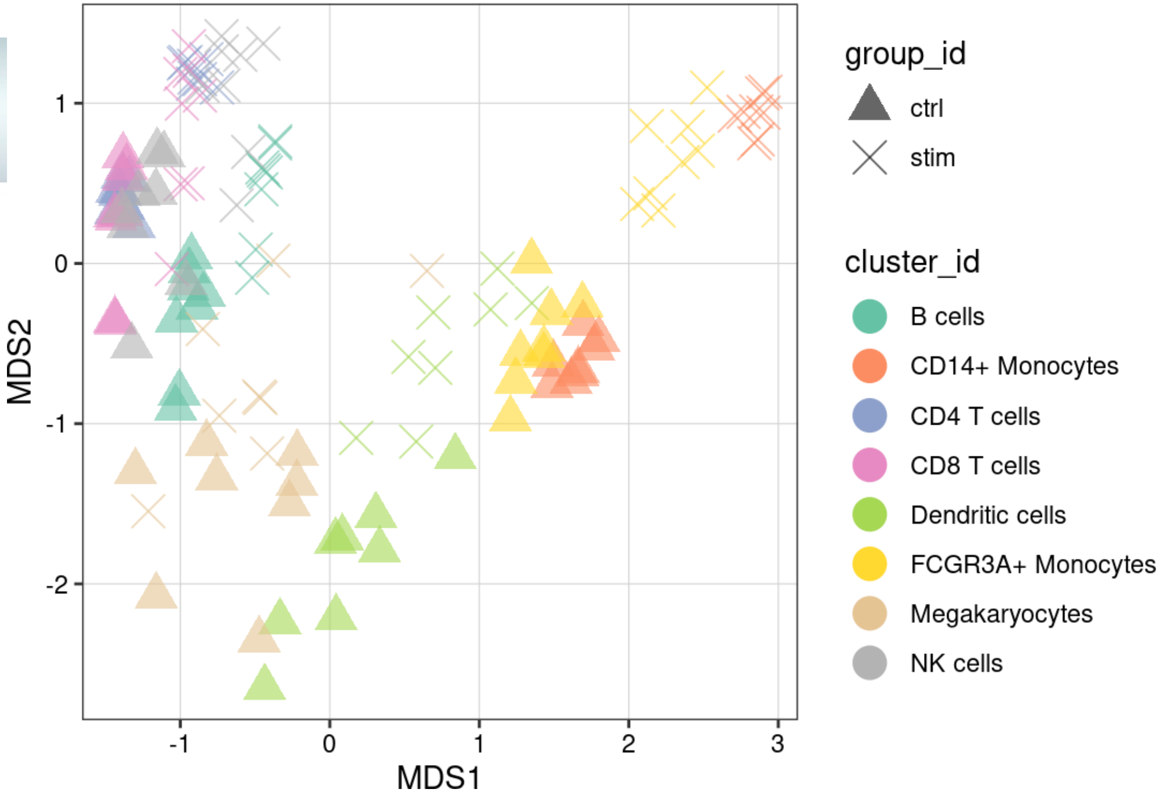
ARTICLE

Check for updates

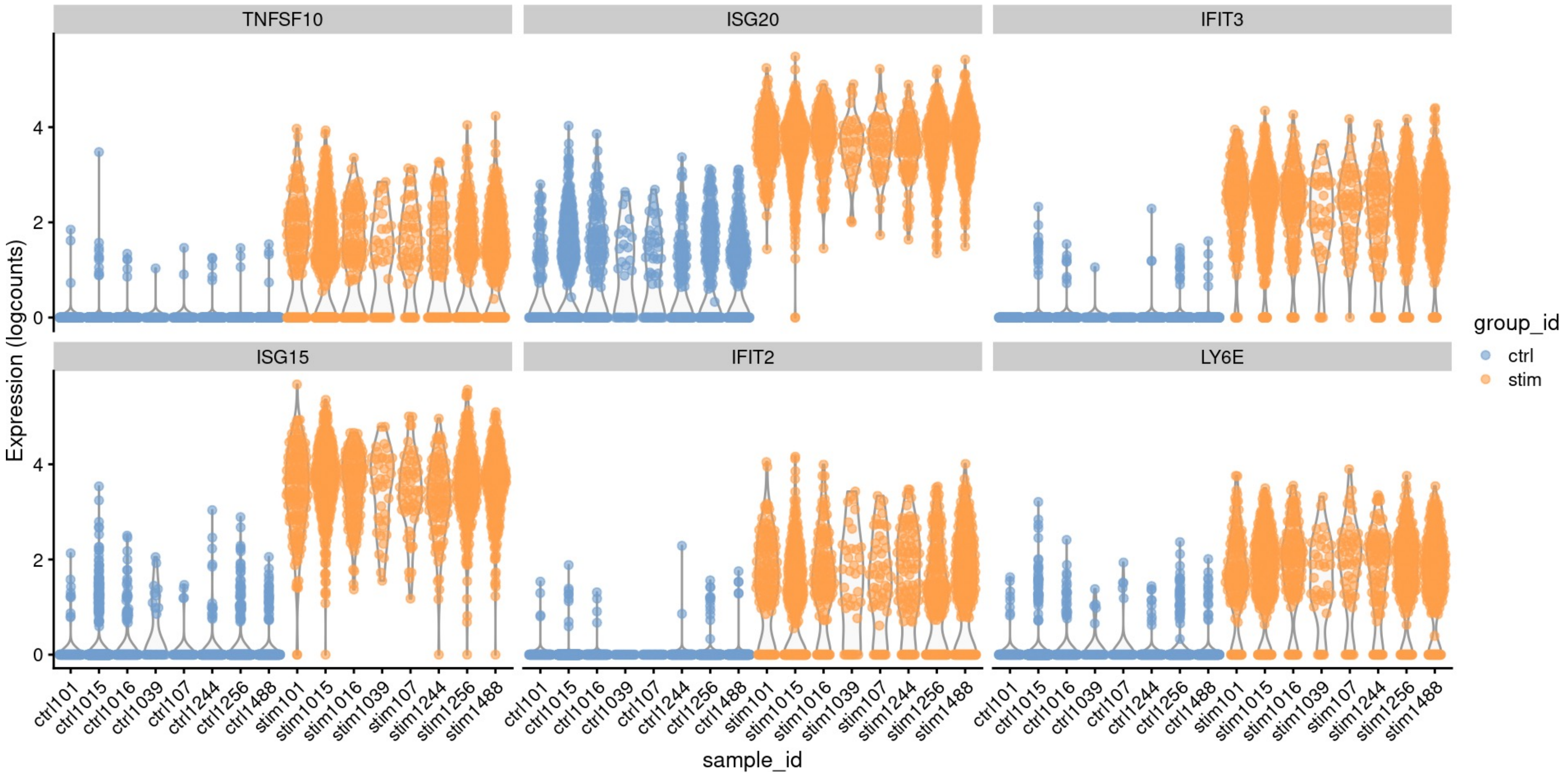
https://doi.org/10.1038/s41467-021-25960-2 OPEN

Confronting false discoveries in single-cell differential expression

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



Muscat::pbDS() or Scran::pseudoBulkDEG



https://www.nature.com/articles/s41467-020-19894-4

Differential expression (DE) vs Differential abundance (DA)

14.6.1 DE or DA? Two sides of the same coin

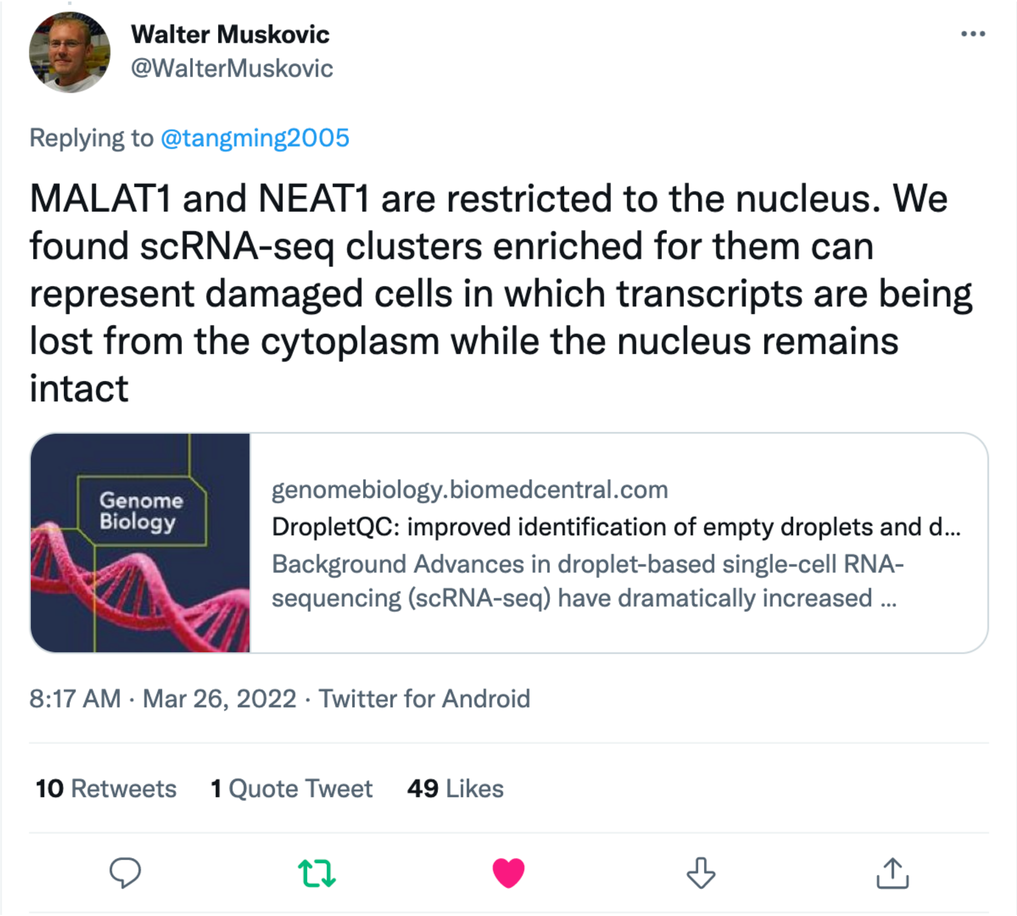
While useful, the distinction between DA and DE analyses is inherently artificial for scRNA-seq data. This is because the labels used in the former are defined based on the genes to be tested in the latter. To illustrate, consider a scRNA-seq experiment involving two biological conditions with several 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.

https://www.biorxiv.org/content/10.1101/2022.03.15.484475v1

Be aware of technical artifacts



Walter Muskovic
@WalterMuskovic

Replying to @tangming2005

MALAT1 and NEAT1 are restricted to the nucleus. We found scRNA-seq clusters enriched for them can represent damaged cells in which transcripts are being lost from the cytoplasm while the nucleus remains intact

genomebiology.biomedcentral.com
DropletQC: improved identification of empty droplets and d...
Background Advances in droplet-based single-cell RNA-sequencing (scRNA-seq) have dramatically increased ...

8:17 AM · Mar 26, 2022 · Twitter for Android

10 Retweets **1** Quote Tweet **49** Likes

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-

Dissociation methods can induce artificial gene signatures

nature neuroscience

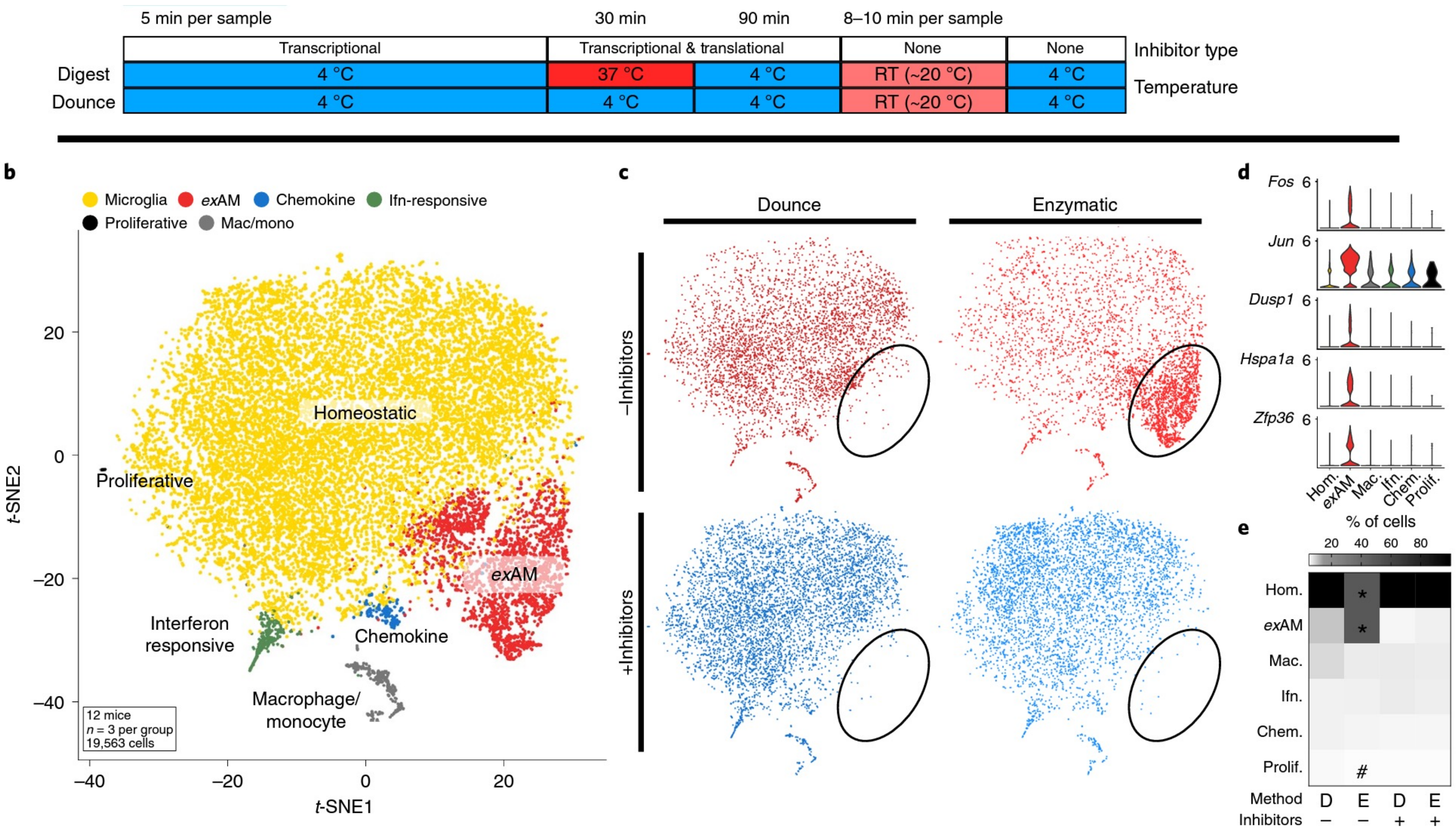
Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › nature neuroscience › articles › article

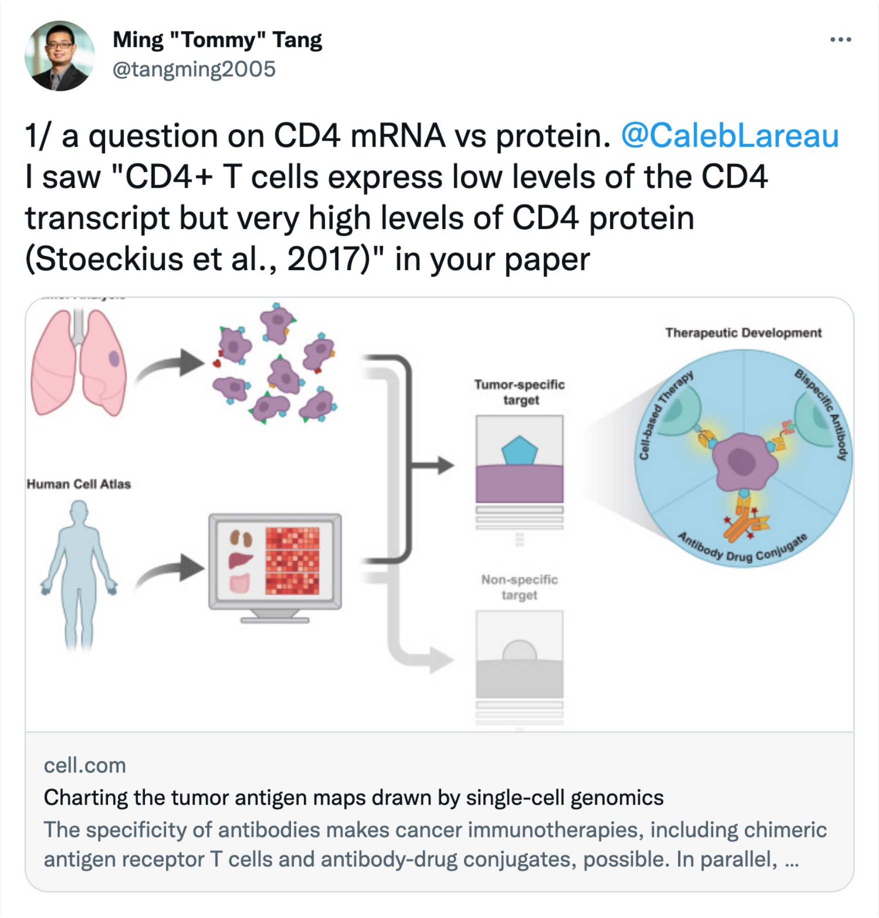
Article | Published: 08 March 2022

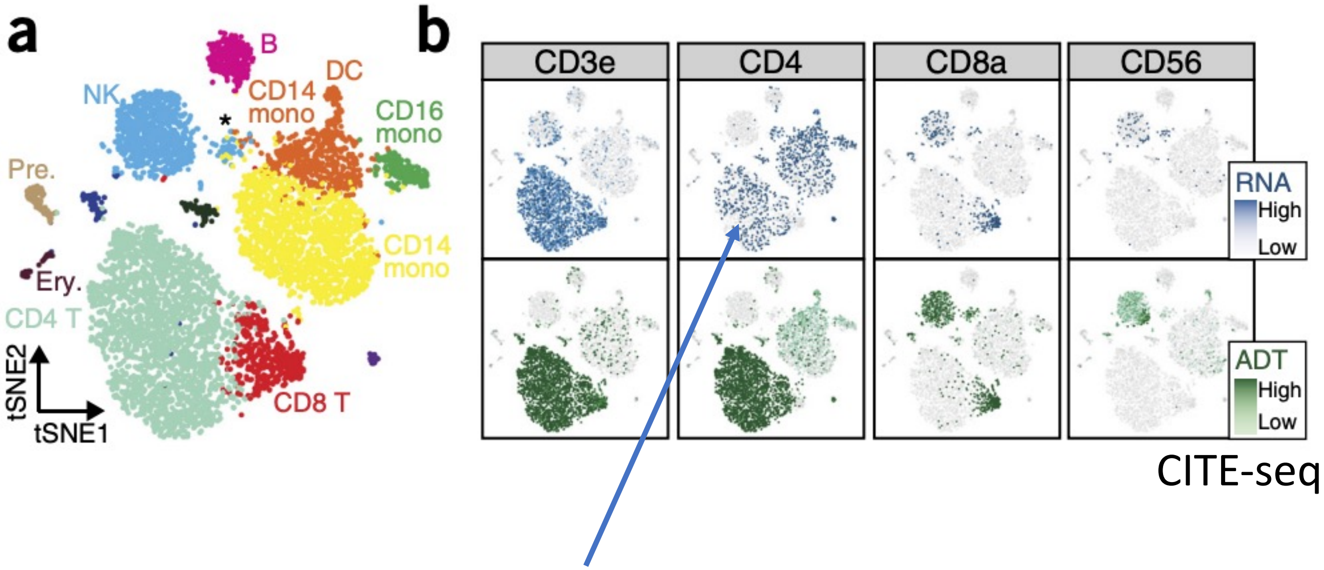
Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain

Samuel E. Marsh, Alec J. Walker, Tushar Kamath, Lasse Dissing-Olesen, Timothy R. Hammond, T.



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





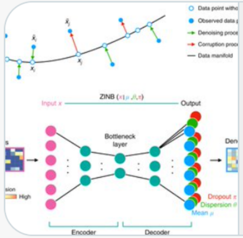
CITE-seq

https://twitter.com/tangming2005/status/1501766040686108678
https://www.nature.com/articles/nmeth.4380

https://rpubs.com/MattPM/cd4

CD56/NCAM1 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

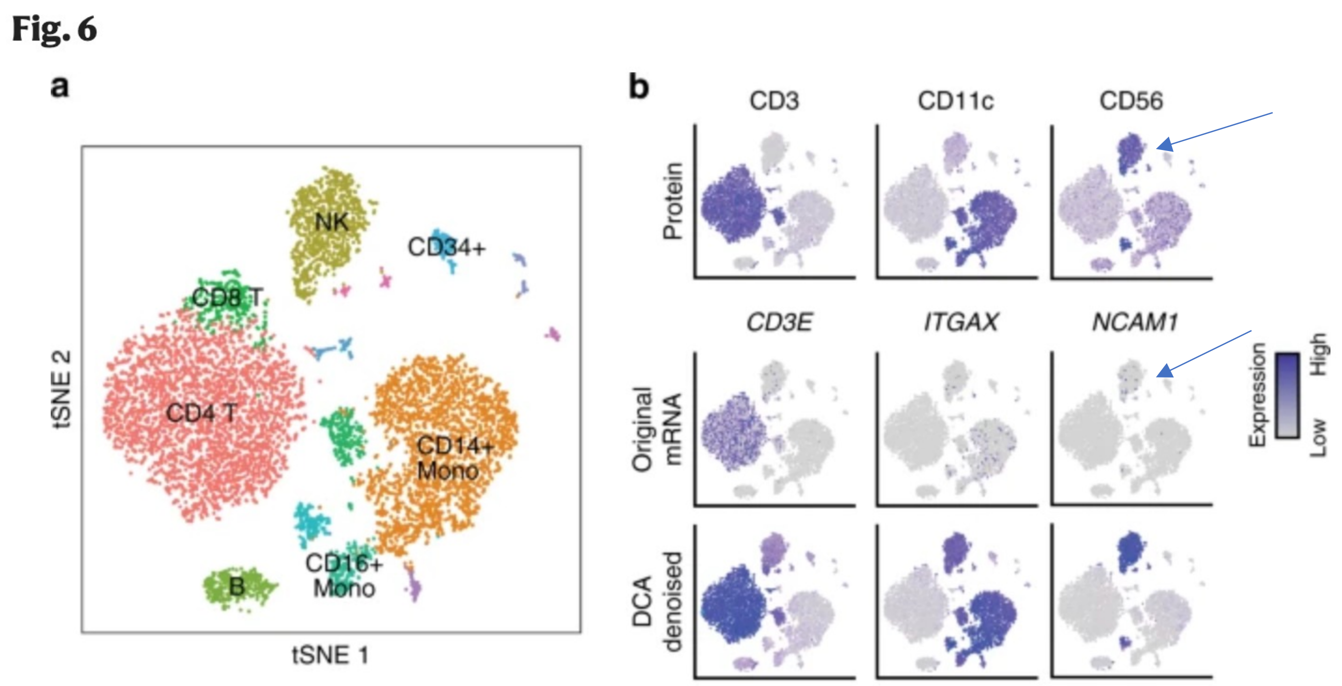


nature.com
Single-cell RNA-seq denoising using a deep count ...
Nature Communications - Single-cell RNA sequencing is a powerful method to study gene ...

1 1 6

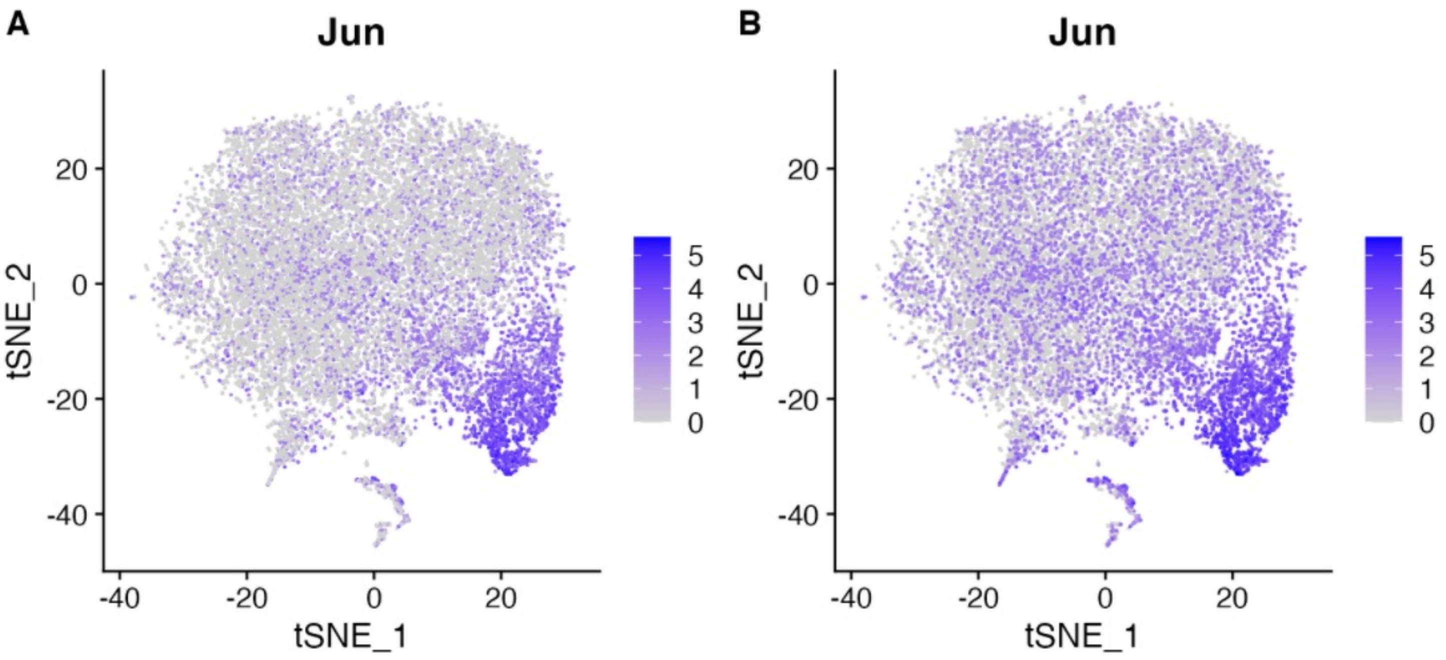
Ming "Tommy" Tang @tangming2005 · Mar 10
Yes! Had the same experience with CD56 myself.

1



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

Gazillions of point, data can be misleading

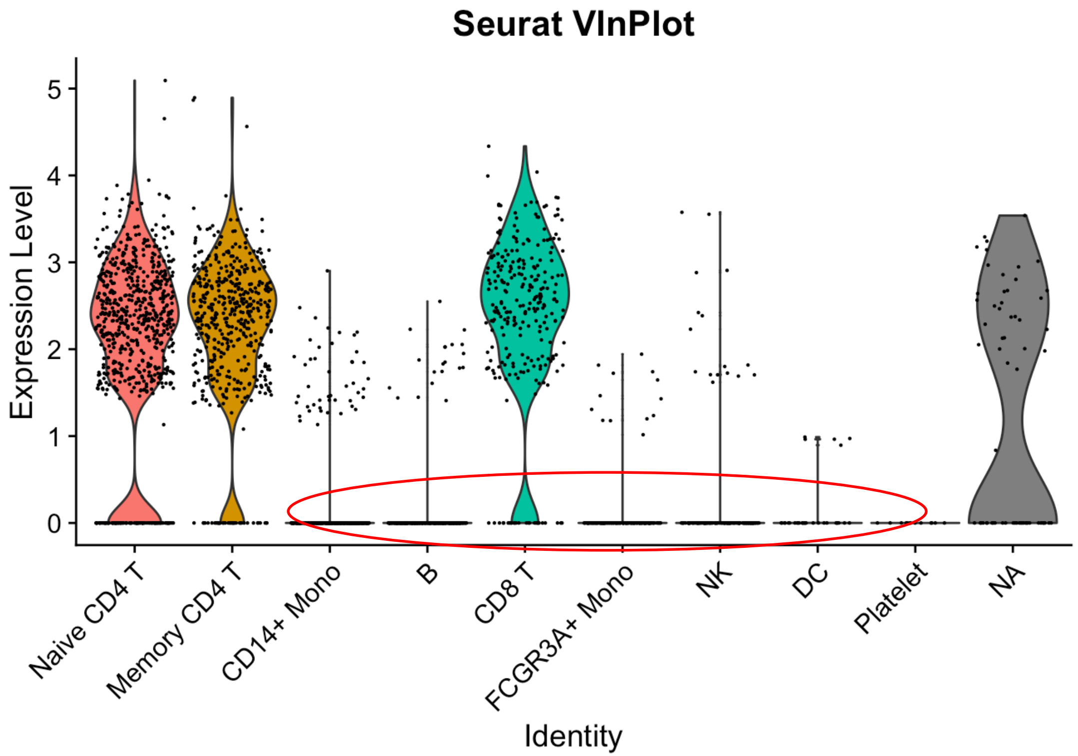


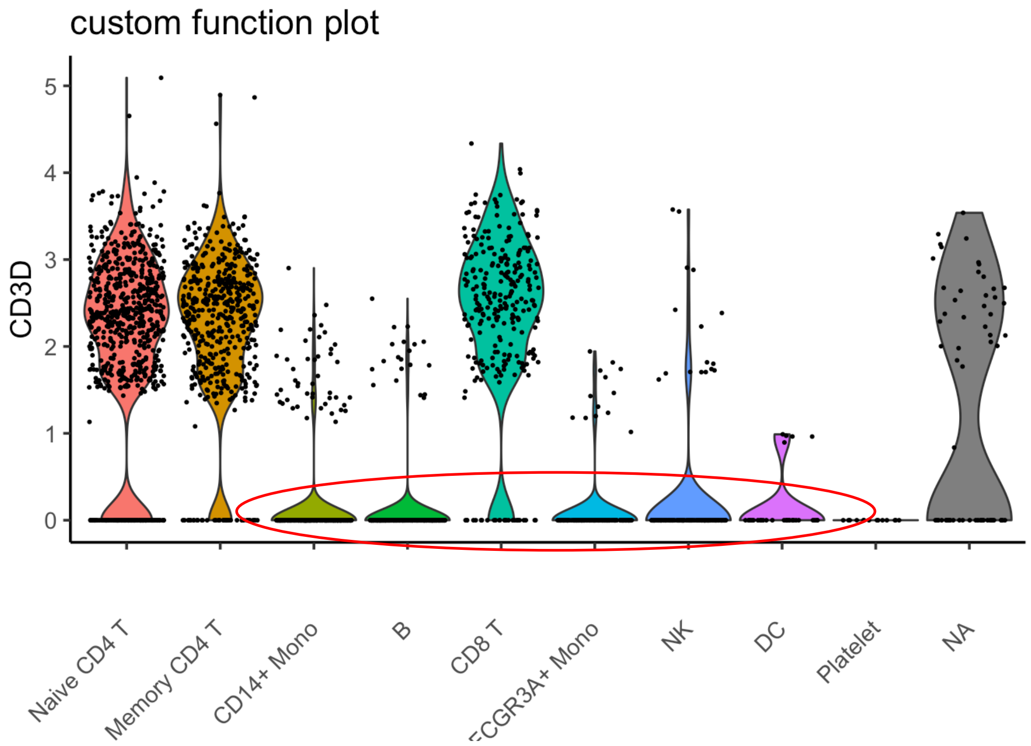
Order =FALSE

Order =TRUE

https://github.com/exaexa/scattermore

Understanding the details of methods





yuhanH commented on Jul 31, 2020 — Collaborator

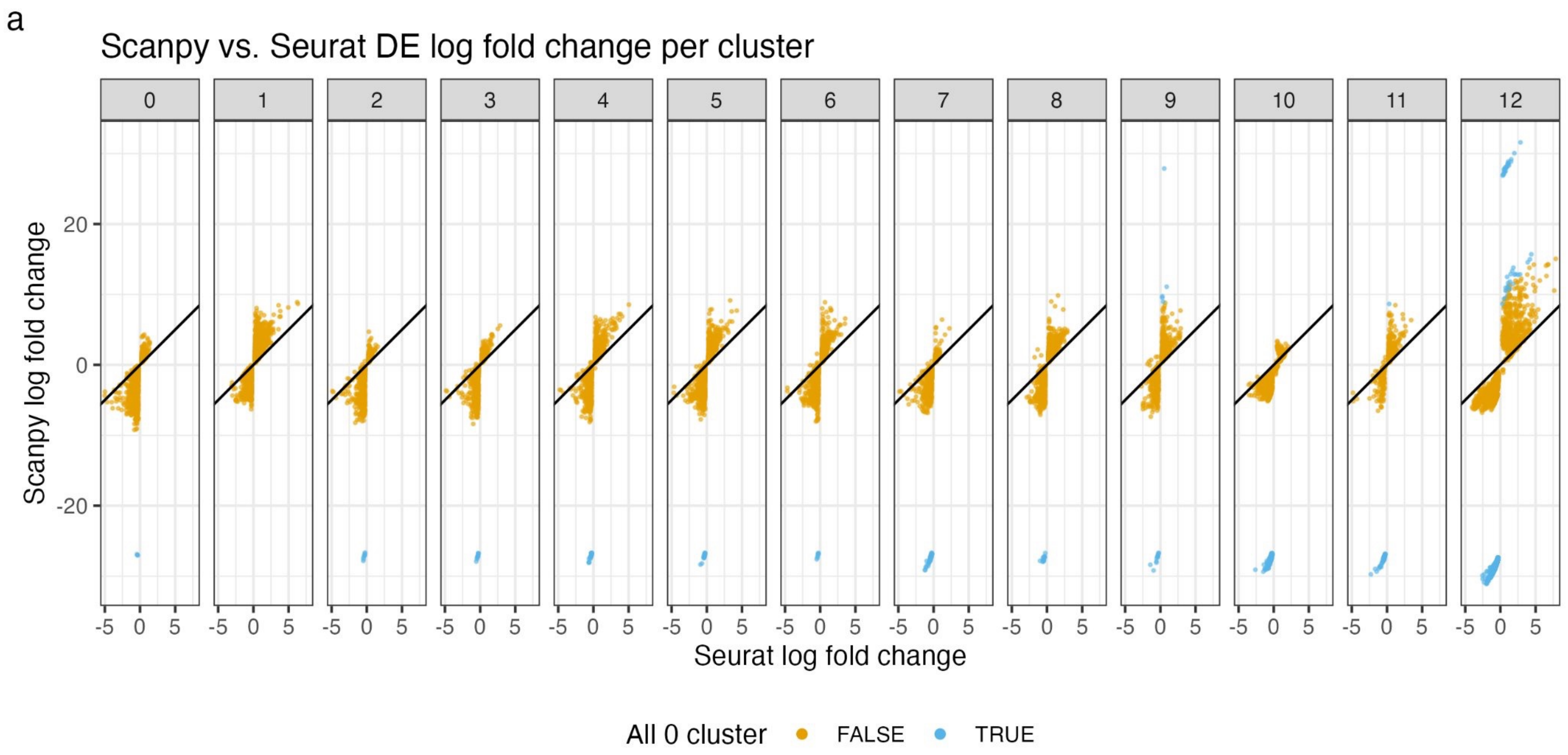
Actually, we add a small noise into data before VlnPlot.

seurat/R/visualization.R
Lines 6733 to 6738 in 72a0c7b

```
6733        noise <- rnorm(n = length(x = data[, feature])) / 100000
6734      }
6735      if (all(data[, feature] == data[, feature][1])) {
6736        warning(paste0("All cells have the same value of ", feature, "."))
6737      } else{
6738        data[, feature] <- data[, feature] + noise
```

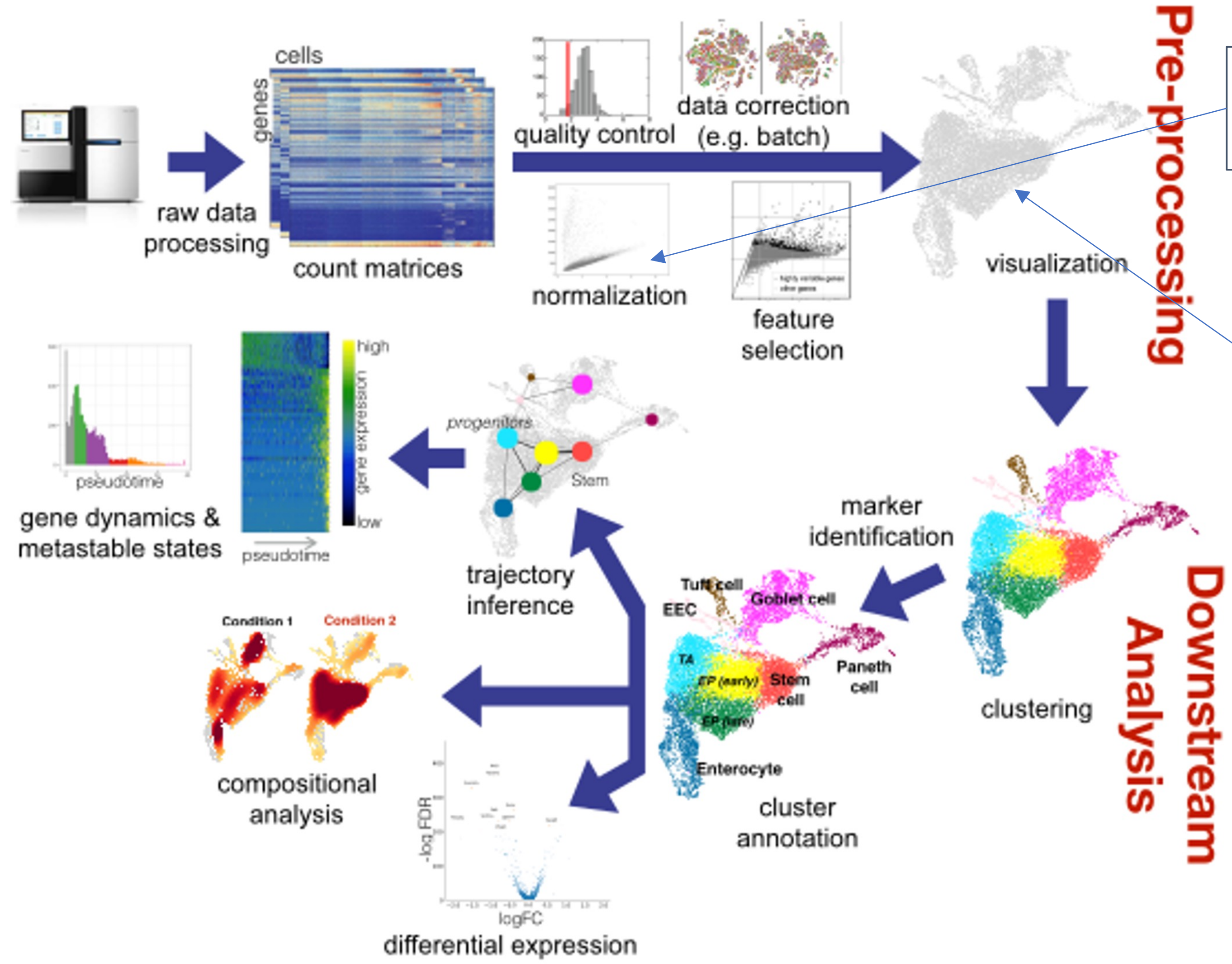
https://github.com/satijalab/seurat/issues/3322

Discrepancy of log2Fold change for marker genes between scanpy and Seurat



https://divingintogeneticsandgenomics.com/post/do-you-really-understand-log2fold-change-in-single-cell-rnaseq-data/

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



Lib size etc.
SCTransform in Seurat

Dimension Reduction:
PCA
TSNE
UMAP

Credit to Peter Hickey

Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

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.

Bioconductor

What you will learn

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/crazyhottommy/scRNAseq-analysis-notes

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

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

Acknowledgements

DFCI:
Shirley Liu
Margaret Shipp

Harvard FAS informatics:
Tim Sackton

Jackson Lab:
Roel Verhaak
Samir Amin

Almighty Tweeps

Follow me on twitter @tangming2005

Divingintogeneticsandgenomics.com

What questions do you have?