Preliminaries

- Mechanism to exchange code, links, etc.:
  - https://yopad.eu/p/freiburg-scrnaseq-workshop
- Code and slides are available online:
  - https://github.com/markrobinsonuzh/workshop-single-cell-transcriptomics-freiburg-25-Nov-2022
- Some data you'll need is available here:
  - https://www.dropbox.com/s/hwqx49qh6msa6ul/workshop-single-cell-transcriptomics-freiburg-25-Nov-2022-data.zip?dl=0

Strategy

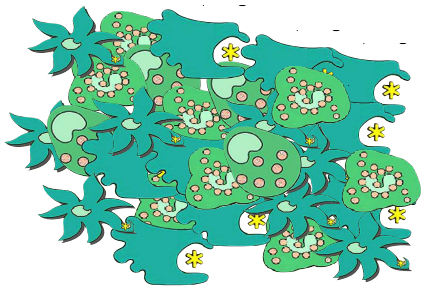
- 4 sessions:
  - 1. Preliminaries (setup, environment, data containers)
  - 2. scRNA-seq QC / normalisation / viz
  - 3. scRNA-seq integration / DA / DS
  - 4. scRNA-seq “advanced”: trajectories, knn differential discovery
- mix of slides, demos, hands-on
- balance of ‘wrappers’ and ‘make it work’
- live coding
- questions at any time

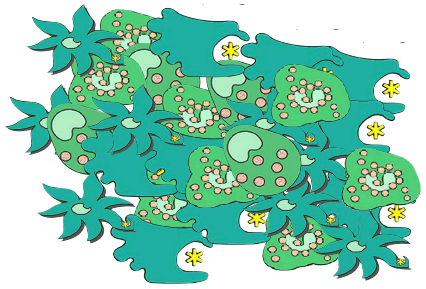
Bulk vs single-cell RNA-sequencing

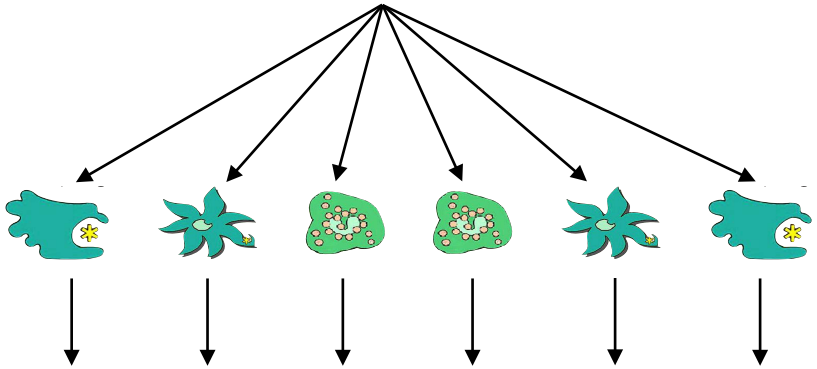
Cell sorting, tissue dissociation

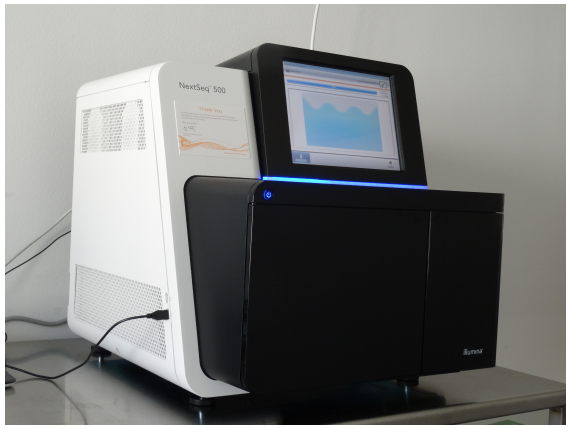
RNA extraction, preparation of cDNA, cell barcoding, UMIs (scRNA-seq only)

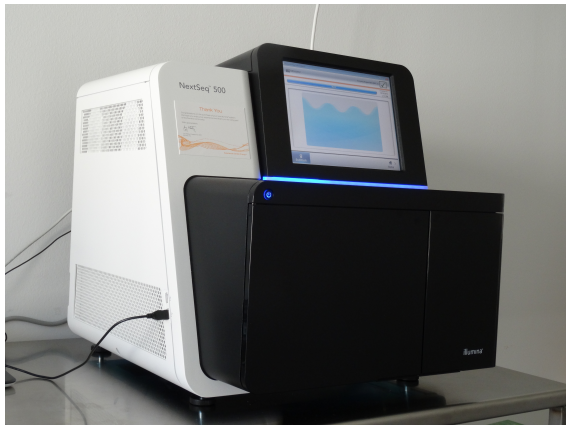
sequencing











Images modified from https://www.flickr.com/photos/konradfoerstner/21264667663 and https://commons.wikimedia.org/wiki/File:Innate_Immune_cells.svg

Introduction to Single-Cell RNA Sequencing

Thale Kristin Olsen[1] and Ninib Baryawno[1]

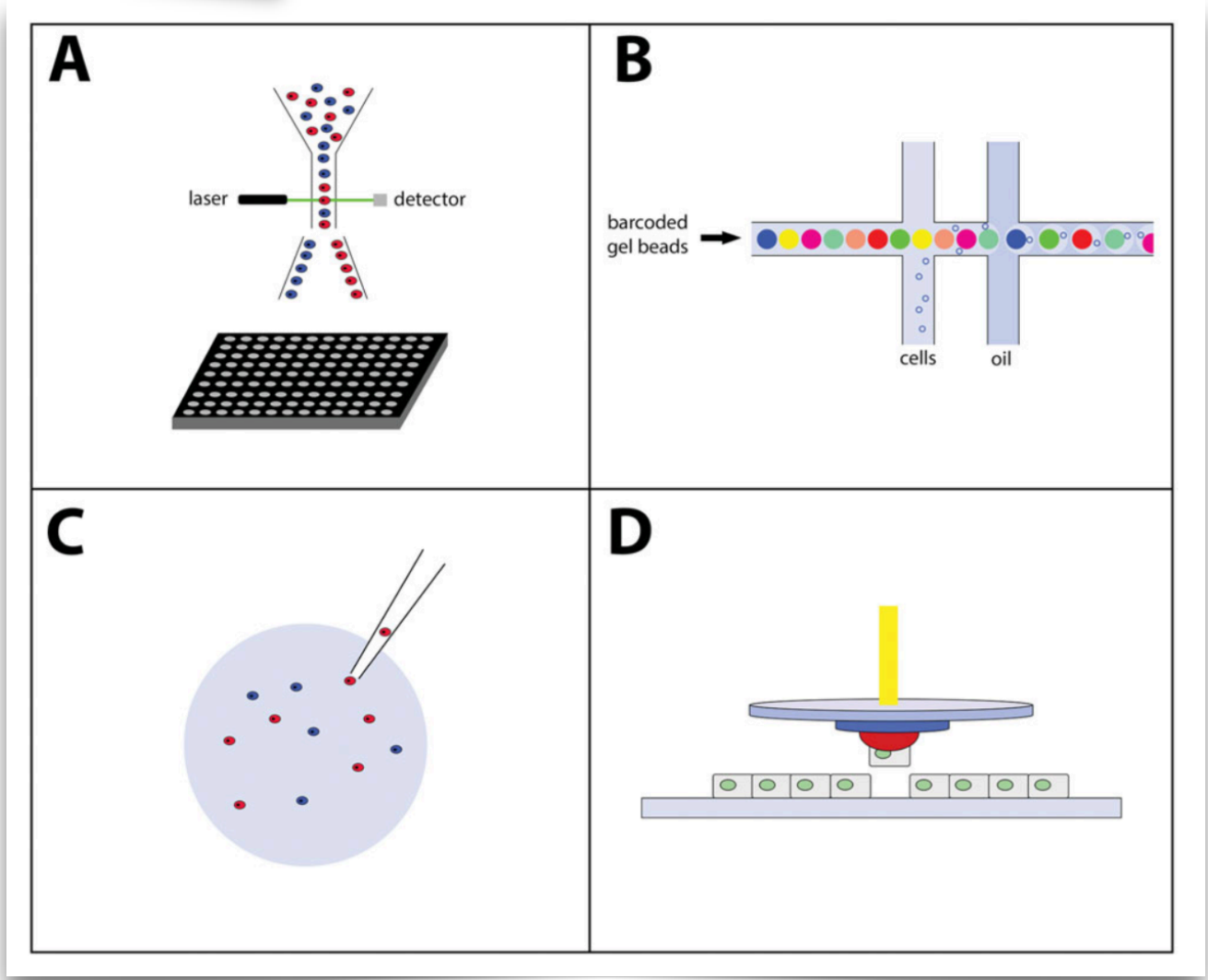
[1]Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

Single cell isolation

FACS

Tissue dissociation

Micro-pipetting



10X GENOMICS

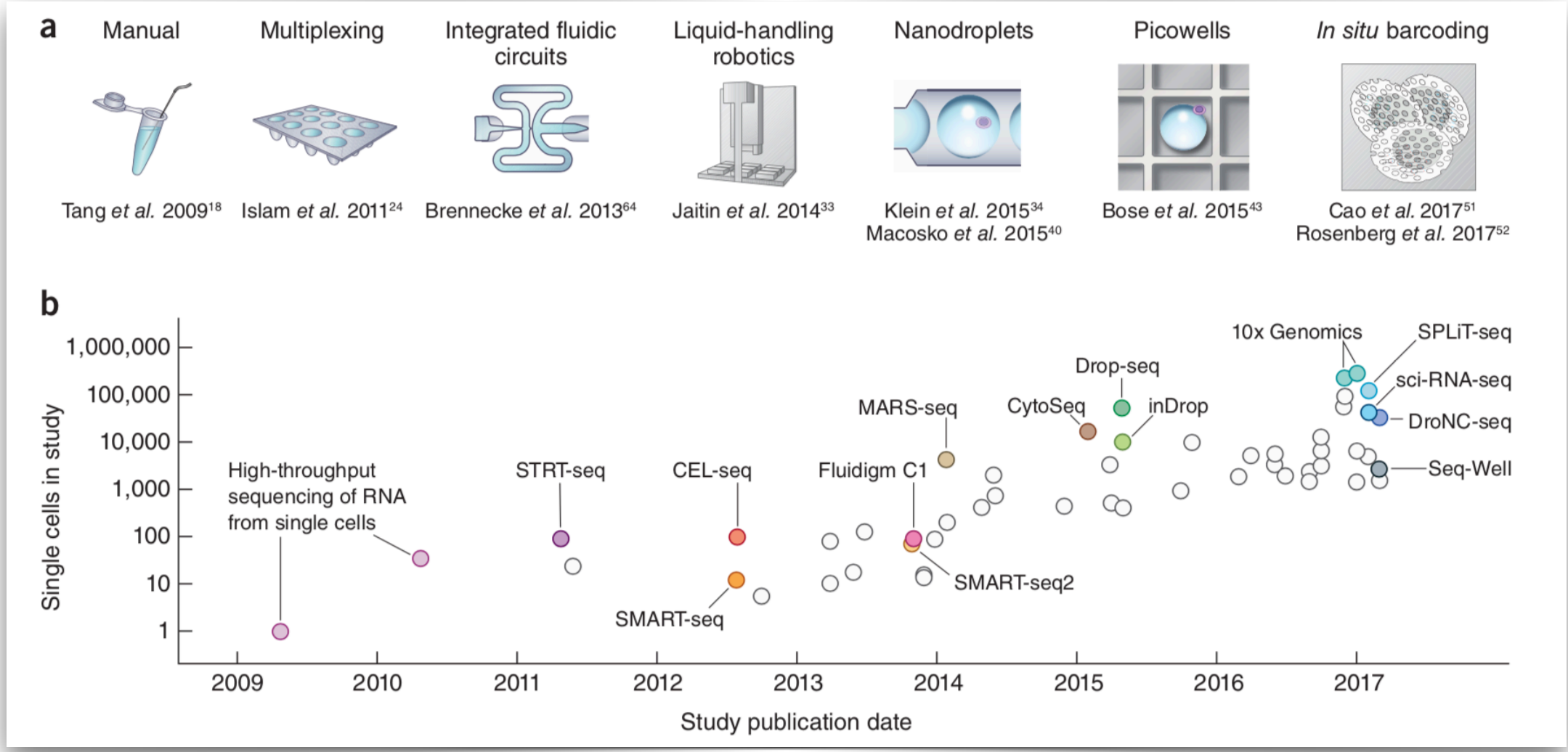
Single-Cell Analysis System

Brand BD Rhapsody™

droplet / microwell

Laser Capture Microdissection

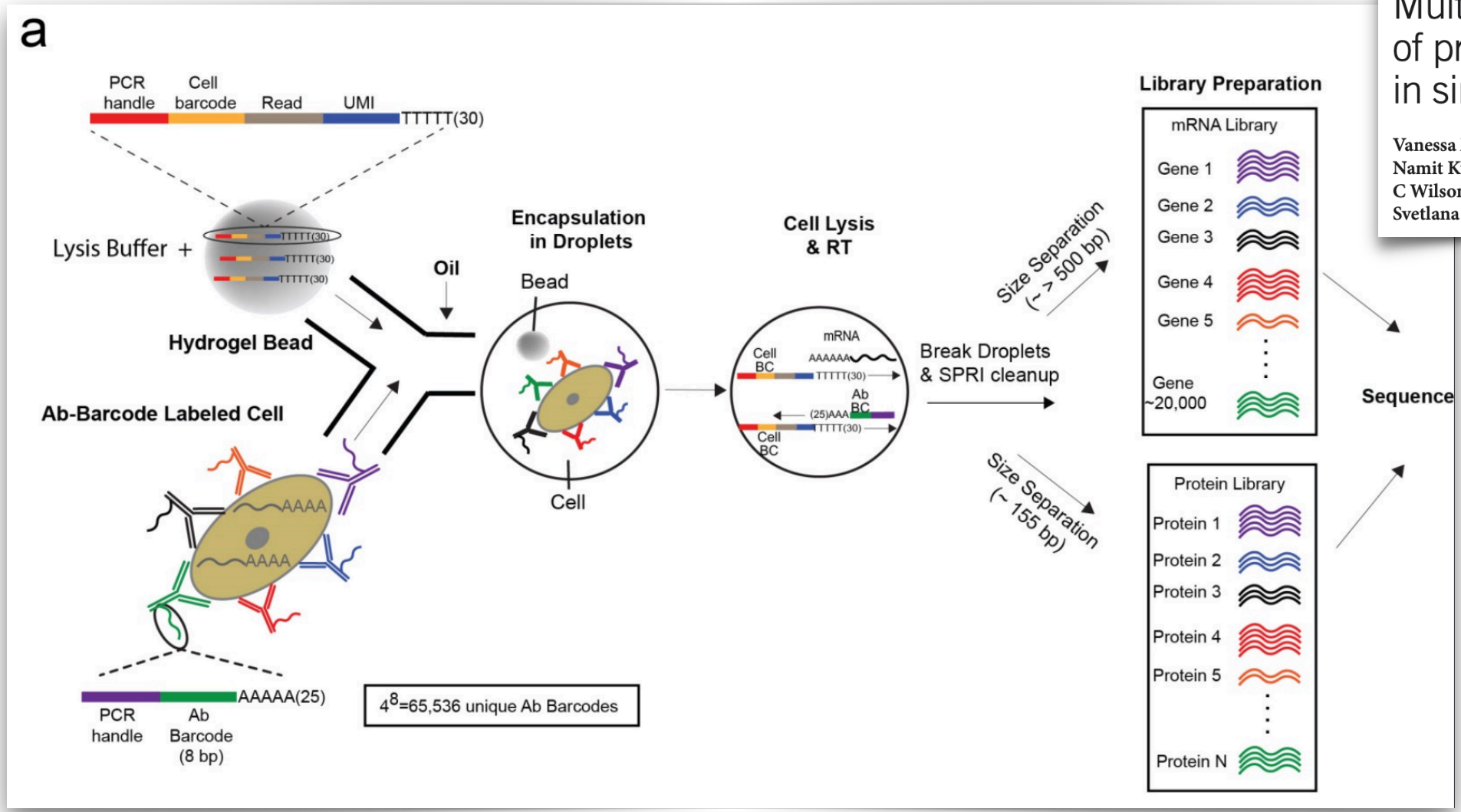
Platforms



Note: tradeoff between number of cells and **depth per cell**

Svensson et al. 2018

Related technologies 1: expression + protein (antibody)



Multiplexed quantification of proteins and transcripts in single cells

Vanessa M Peterson[1,5], Kelvin Xi Zhang[2,5], Namit Kumar[1], Jerelyn Wong[3], Lixia Li[1], Douglas C Wilson[3], Renee Moore[4], Terrill K McClanahan[3], Svetlana Sadekova[3] & Joel A Klappenbach[1]

Related technologies 2: expression + (epi)genome/ proteome

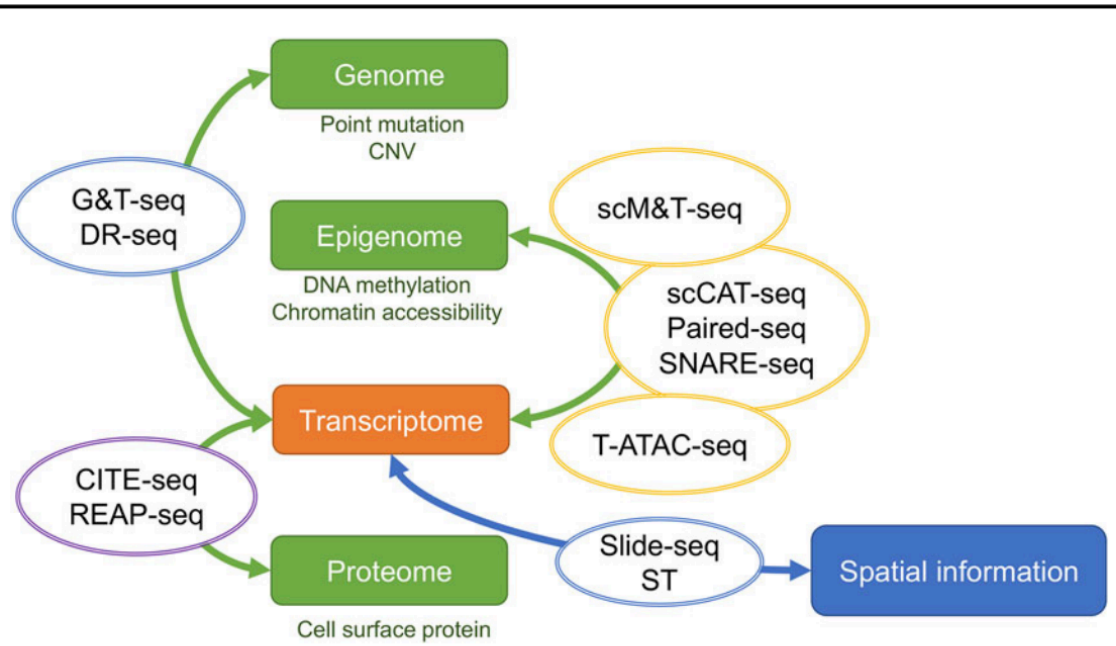


Fig. 3 Multilayered single-cell sequencing. Representative single-cell multimodal sequencing methods. Genomic, epigenomic, and proteomic information can be simultaneously profiled with the transcriptome. Spatial information for a tissue section can also be obtained with gene expression data at the level of one to tens of cells. ST spatial transcriptomics (Visium).

REVIEW ARTICLE Open Access

Single-cell sequencing techniques from individual to multiomics analyses

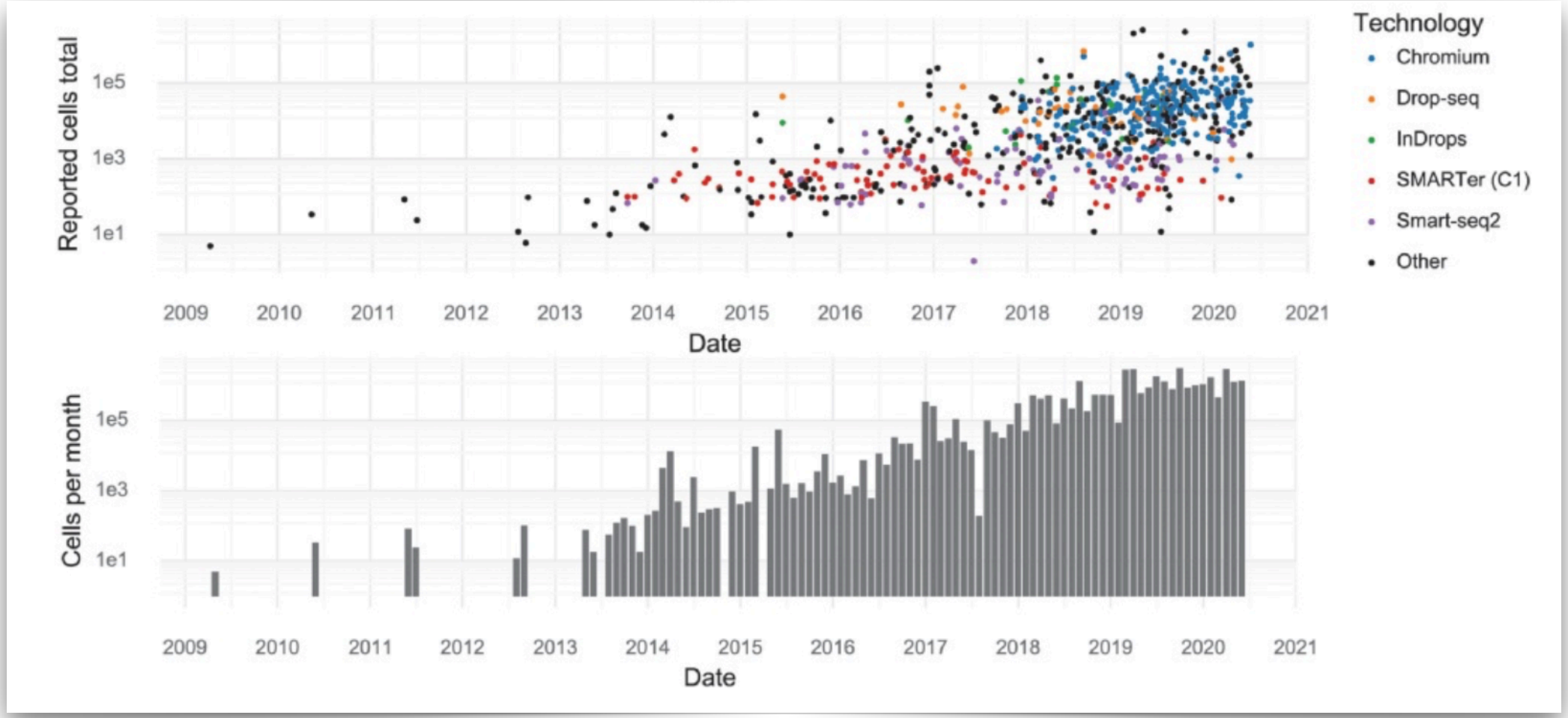
Yukie Kashima[1,2], Yoshitaka Sakamoto[1], Keiya Kaneko[1], Masahide Seki[1], Yutaka Suzuki[1] and Ayako Suzuki[1]

A curated database reveals trends in single-cell transcriptomics

Valentine Svensson*, Eduardo da Veiga Beltrame and Lior Pachter

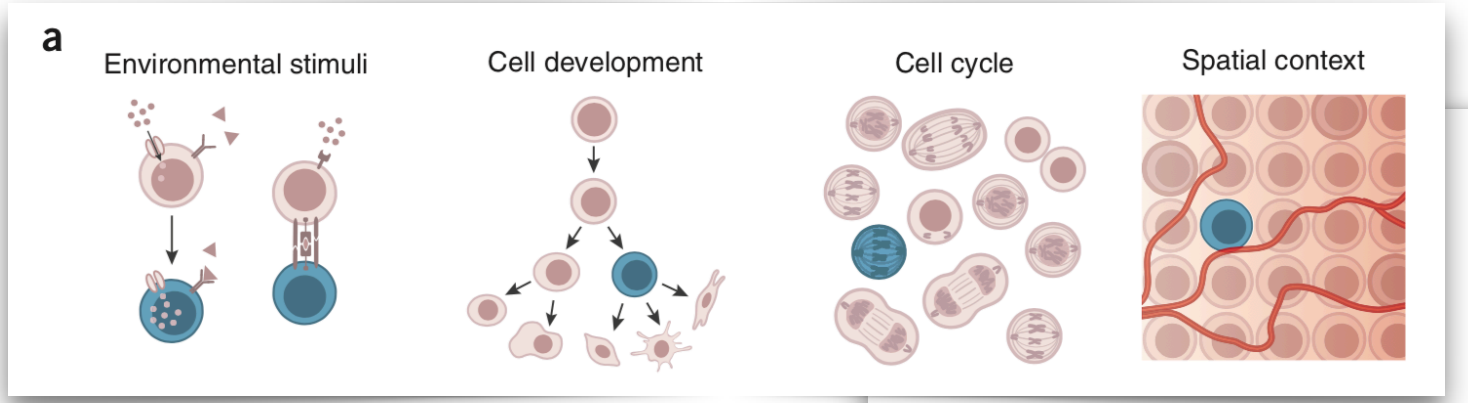
Division of Biology and Biological Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA, 91125, USA

(scRNA-seq) data velocity



Note: tradeoff between number of cells and **depth per cell**

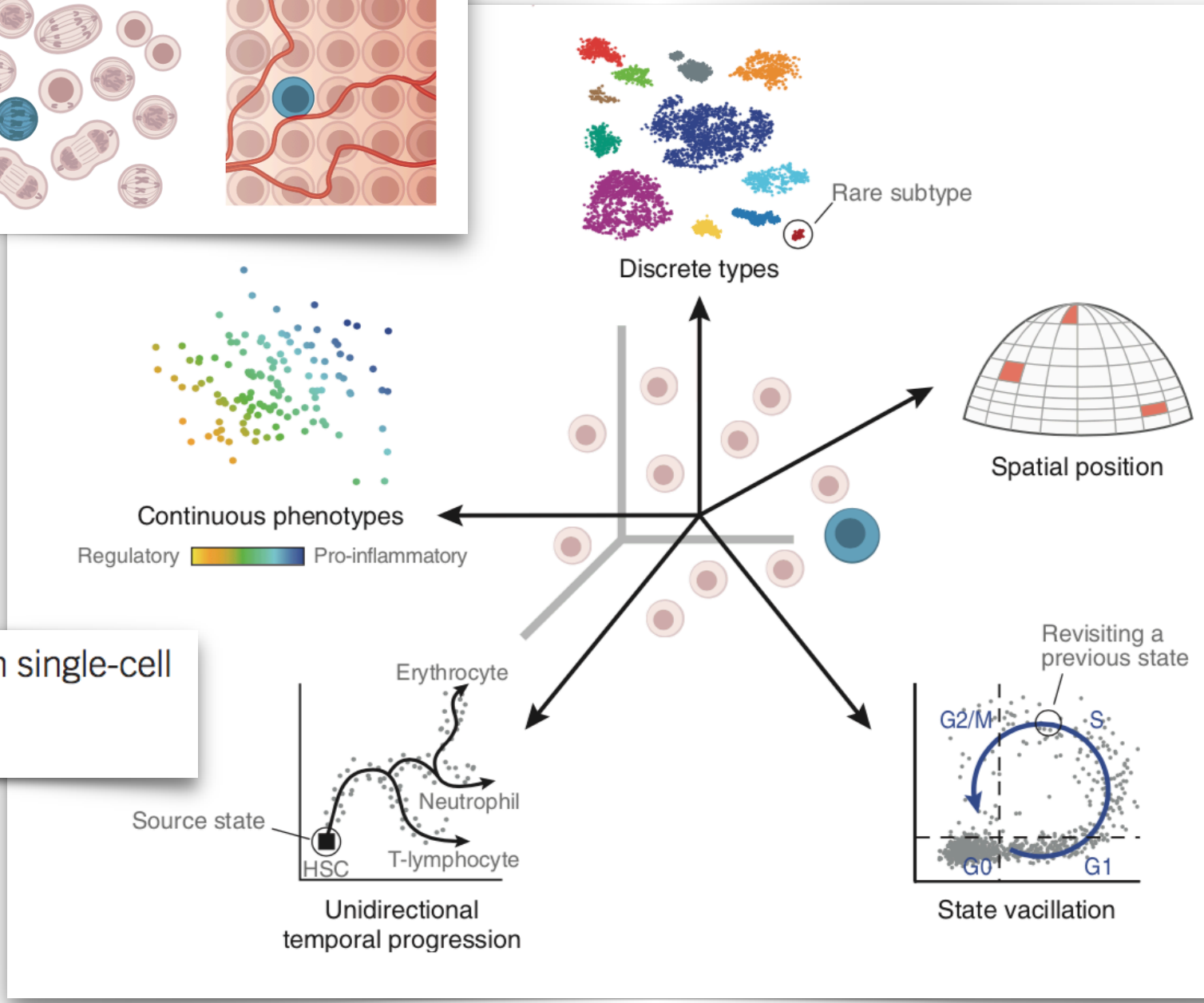
Svensson et al. 2020



Applications

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]



SINGLE-CELL OMICS

Computational and analytical challenges in single-cell transcriptomics

Oliver Stegle[1], Sarah A. Teichmann[1,2] and John C. Marioni[1,2]

Why single cell?

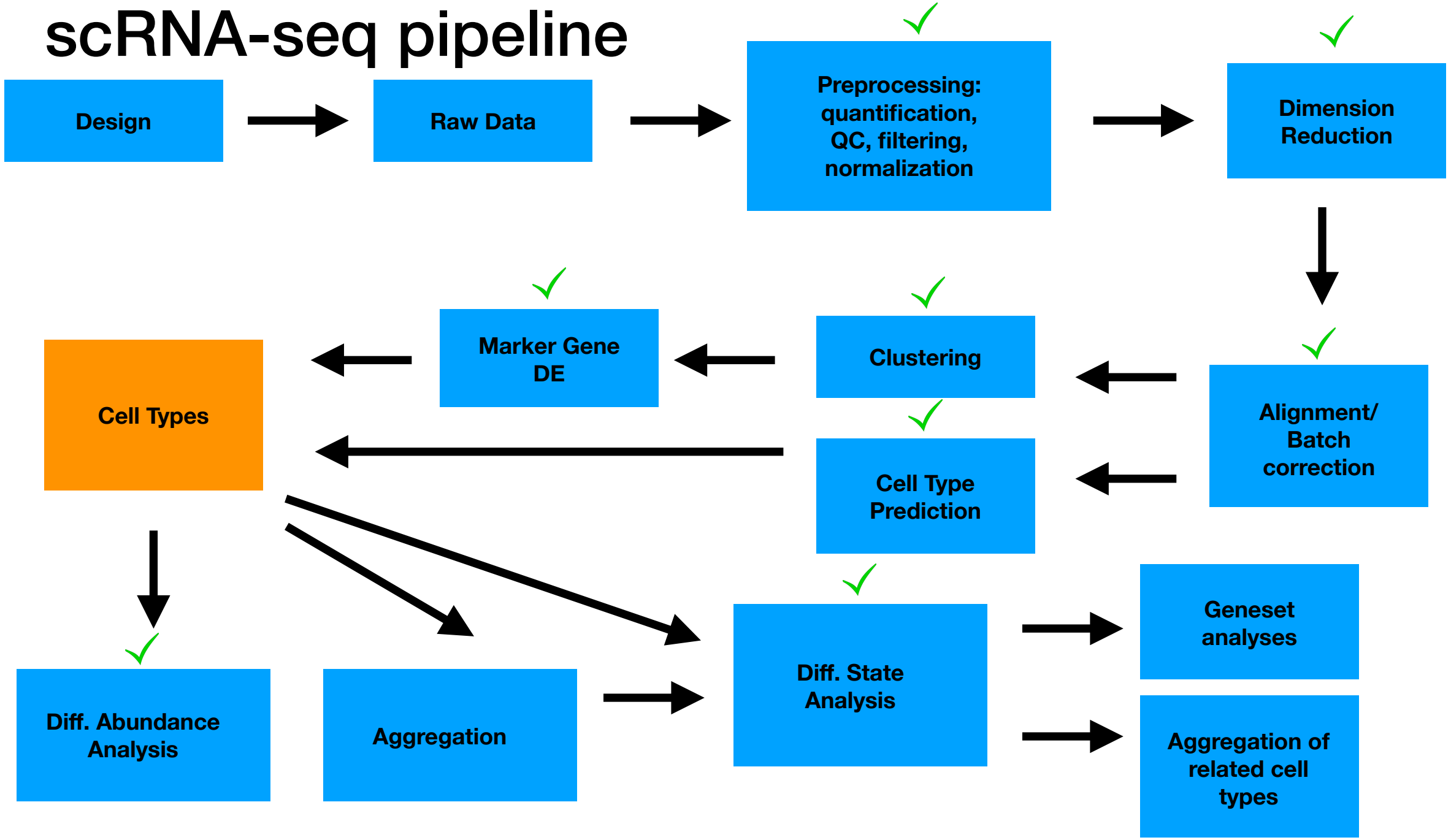
"Bulk" versus single-cell

Discover and quantify abundance of (new) cell types

Study heterogeneity of gene expression

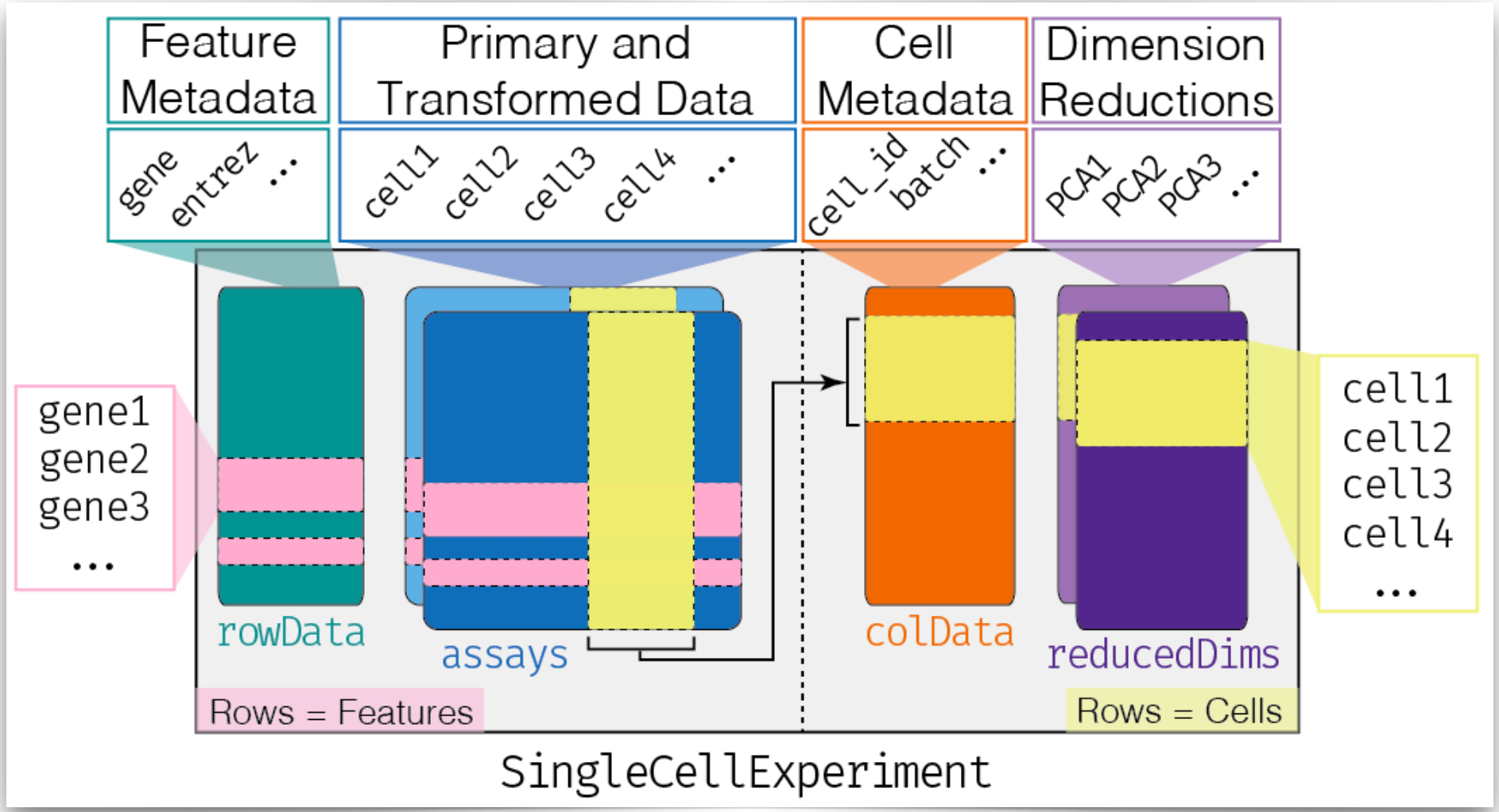
However, there are also important biological questions for which bulk measures of gene expression are insufficient[14]. For instance, during early development, there are only a small number of cells, each of which can have a distinct function and role[15–17]. Moreover, complex tissues, such as brain tissues, are composed of many distinct cell types that are typically difficult to dissect experimentally[18]. Consequently, bulk-based approaches may not provide insight into whether differences in expression between samples are driven by changes in cellular composition (that is, the abundance of different cell types) or by changes in the underlying phenotype. Finally, ensemble measures do not provide insights into the stochastic nature of gene expression[19,20].

scRNA-seq pipeline

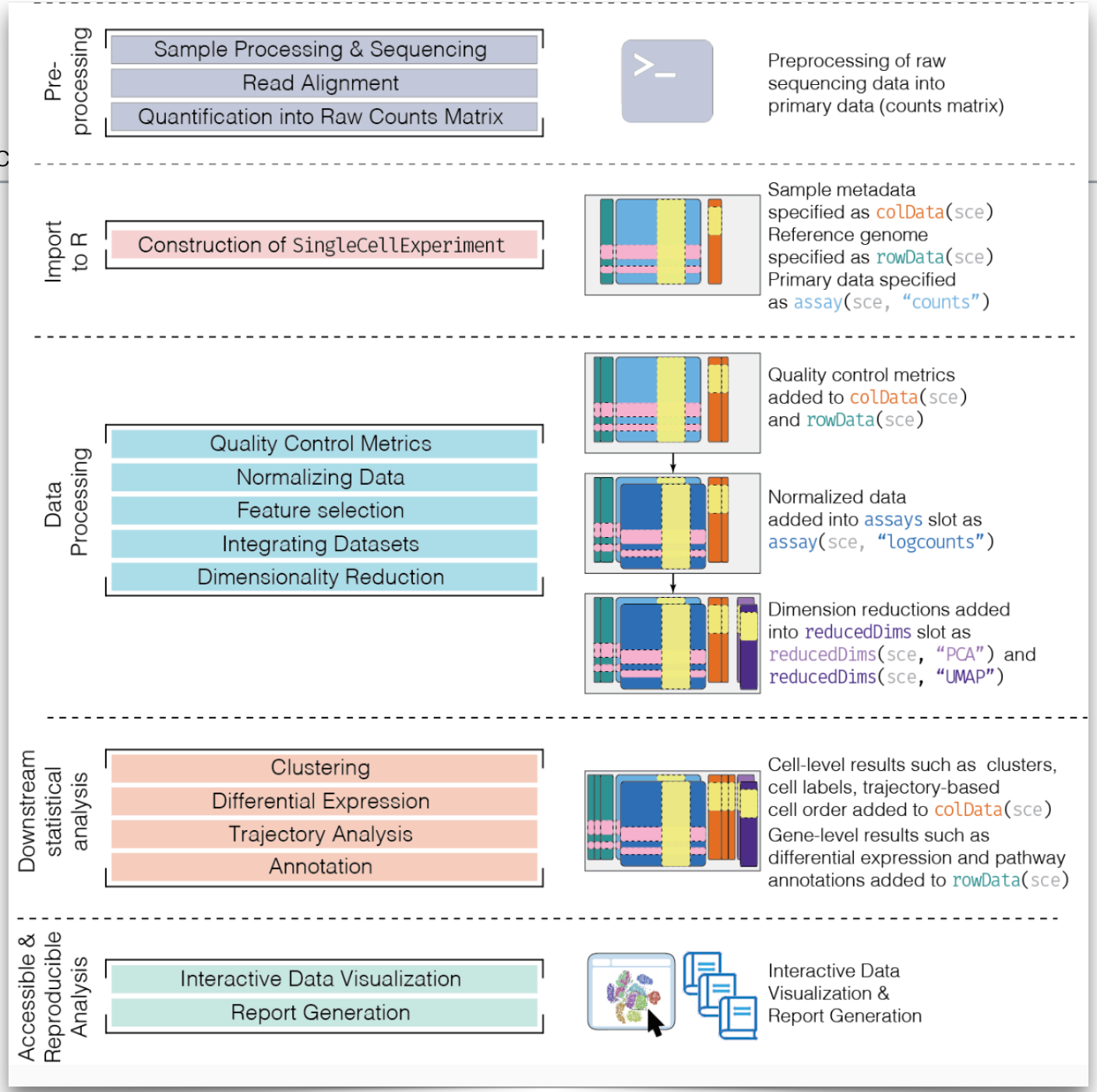


https://bioconductor.org/books/3.12/OSCA/data-infrastructure.html
https://bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html

SingleCellExperiment



SingleCellExperiment



scRNAseq package

```
out <- listDatasets()
```

Reference	Taxonomy	Part	Number	Call
Aztekin et al. (2019)	Xenopus	tail	13199	`AztekinTailData()`
Bach et al. (2017)	Mouse	mammary gland	25806	`BachMammaryData()`
Bacher et al. (2020)	Human	T cells	104417	`BacherTCellData()`
Baron et al. (2016)	Human	pancreas	8569	`BaronPancreasData('human')`
Baron et al. (2016)	Mouse	pancreas	1886	`BaronPancreasData('mouse')`
Bhaduri et al. (2020)	Human	cortical organoids	242349	`BhaduriOrganoidData()`
Buettner et al. (2015)	Mouse	embryonic stem cells	288	`BuettnerESCData()`
(**???**)	Human	haematopoietic stem and progenitor	5183	`BunisHSPCData()`
Campbell et al. (2017)	Mouse	brain	21086	`CampbellBrainData()`
Chen et al. (2017)	Mouse	brain	14437	`ChenBrainData()`

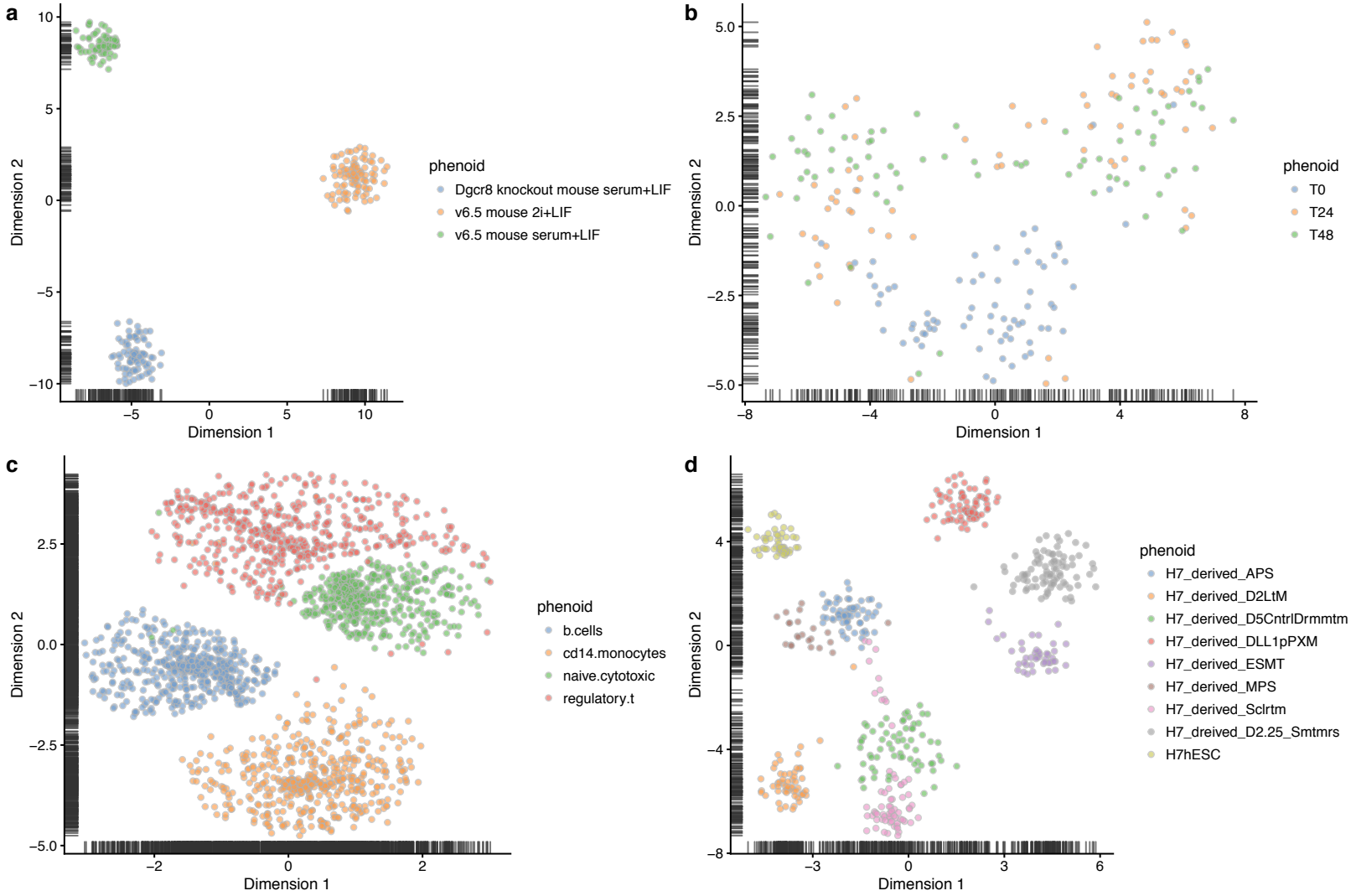
Quick start

```r
library(scRNAseq)
sce <- MacoskoRetinaData()

# Quality control (using mitochondrial genes).
library(scater)
is.mito <- grepl("^MT-", rownames(sce))
qcstats <- perCellQCMetrics(sce, subsets=list(Mito=is.mito))
filtered <- quickPerCellQC(qcstats, percent_subsets="subsets_Mito_percent")
sce <- sce[, !filtered$discard]

# Normalization.
sce <- logNormCounts(sce)

# Feature selection.
library(scran)
dec <- modelGeneVar(sce)
hvg <- getTopHVGs(dec, prop=0.1)

# PCA.
library(scater)
set.seed(1234)
sce <- runPCA(sce, ncomponents=25, subset_row=hvg)

# Clustering.
library(bluster)
colLabels(sce) <- clusterCells(sce, use.dimred='PCA',
    BLUSPARAM=NNGraphParam(cluster.fun="louvain"))

# Visualization.
sce <- runUMAP(sce, dimred = 'PCA')
plotUMAP(sce, colour_by="label")
```

http://bioconductor.org/books/3.16/OSCA.intro/analysis-overview.html#quick-start-simple

After clustering, how to find cell type markers ?

- here, **predefined groups**: range of difficulty



Differential expression: zero inflation / model dropout, mixture models, etc.

MAST

mixture model

hurdle model

SCDE

Single-cell RNA-seq hurdle model

We model the $\log_2(\text{TPM} + 1)$ expression matrix as a two-part generalized regression model. The gene expression rate was modeled using logistic regression and, conditioning on a cell expressing the gene, the expression level was modeled as Gaussian.

Given normalized, possibly thresholded (see Additional file 1), scRNA-seq expression $Y = [y_{ig}]$, the rate of expression and the level of expression for the expressed cells are modeled conditionally independent for each gene g. Define the indicator $Z = [z_{ig}]$, indicating whether gene g is expressed in cell i (i.e., $z_{ig} = 0$ if $y_{ig} = 0$ and $z_{ig} = 1$ if $y_{ig} > 0$). We fit logistic regression models for the discrete variable Z and a Gaussian linear model for the continuous variable $(Y \mid Z = 1)$ independently, as follows:

$$logit(Pr(Z_{ig} = 1)) = \mathrm{X_i}\,\beta_g^D$$

$$\Pr(Y_{ig} = y | Z_{ig} = 1) = \mathrm{N}\left(\mathrm{X_i}\beta_g^C,\ \sigma_g^2\right)$$

The regression coefficients of the discrete component are regularized using a Bayesian approach as implemented in the *bayesglm* function of the *arm* R package, which uses weakly informative priors [30] to provide sensible estimates under linear separation (See Additional file 1 for details). We also perform regularization of the continuous model variance parameter, as described below, which helps to increase the robustness of gene-level differential expression analysis when a gene is only expressed in a few cells.

Differential expression analysis. With a Bayesian approach, the posterior probability of a gene being expressed at an average level x in a subpopulation of cells S was determined as an expected value (E) according to

$$p_S(x) = E\left[\prod_{c \in B} p(x \mid r_c, \Omega_c)\right]$$

where B is a bootstrap sample of S, and $p(x|r_c,\Omega_c)$ is the posterior probability for a given cell c, according to

$$p(x \mid r_c, \Omega_c) = p_d(x)p_{Poisson}(x) + (1 - p_d(x))p_{NB}(x \mid r_c)$$

where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S, $p_{Poisson}(x)$ and $p_{NB}(x|r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) of a gene expressed at level x in cell c, with the parameters of the distributions determined by the Ω_c fit. For the differential expression analysis, the posterior probability that the gene shows a fold expression difference of f between subpopulations S and G was evaluated as

$$p(f) = \sum_{x \in X} p_S(x)p_G(fx)$$

where x is the valid range of expression levels. The posterior distributions were renormalized to unity, and an empirical P value was determined to test for significance of expression difference.

Differential distributions

Equivalent **E**xpression

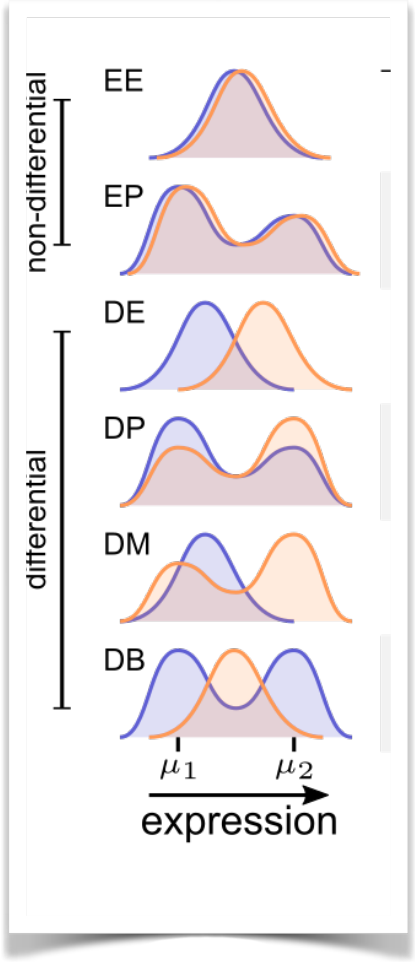
Equivalent **P**roportions

Differential **E**xpression

Differential **P**roportions

Differential **M**odality

Both, **D**ifferential modality & component means



(shift in means)

(identical means)

Korthauer *et al. Genome Biology* (2016) 17:222
DOI 10.1186/s13059-016-1077-y

Genome Biology

METHOD Open Access

A statistical approach for identifying differential distributions in single-cell RNA-seq experiments

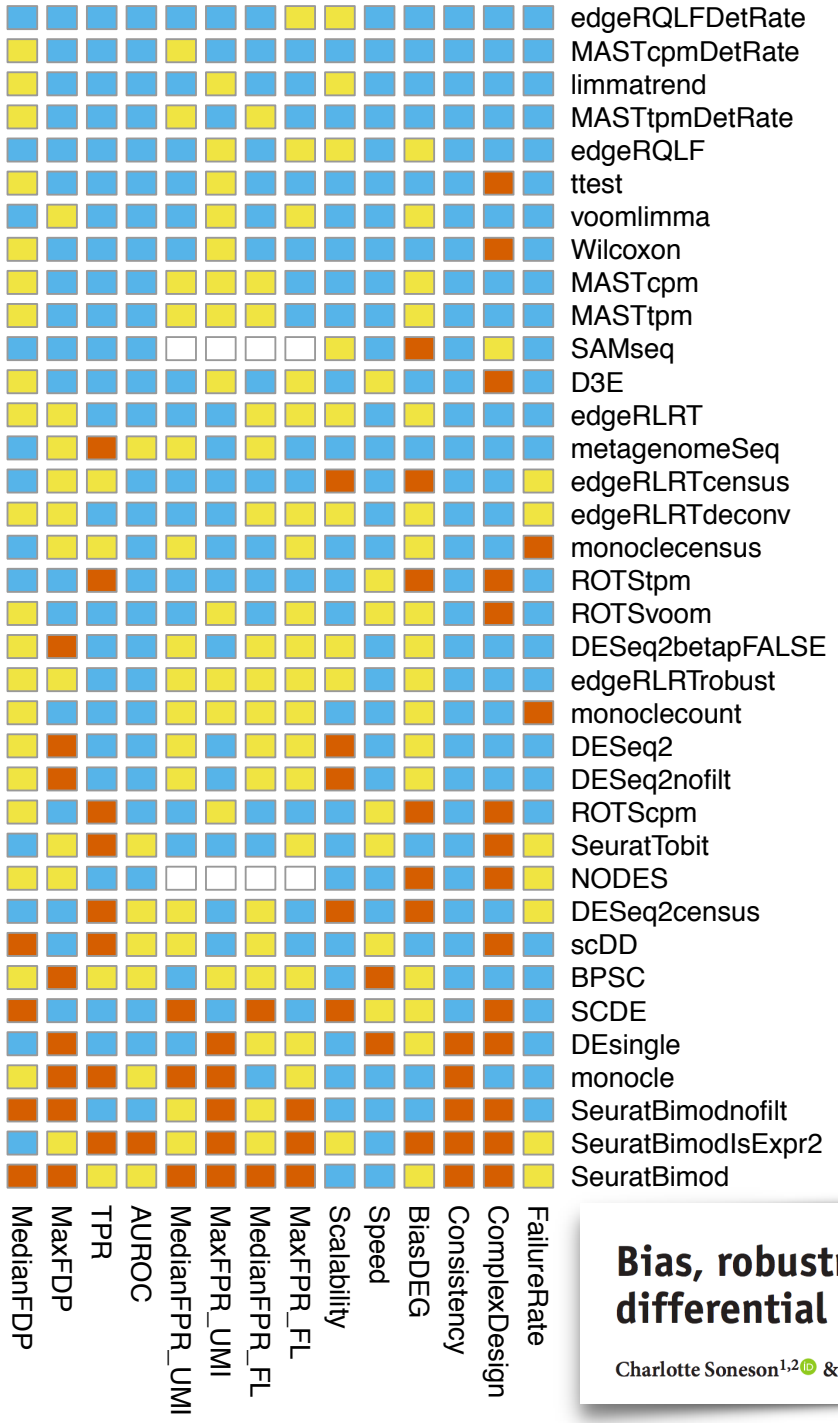
Keegan D. Korthauer[1,2], Li-Fang Chu[3], Michael A. Newton[4,5], Yuan Li[5], James Thomson[3,6,7], Ron Stewart[3] and Christina Kendziorski[4,5*]

Punchline

Several methods work well, including a mix of single-cell-specific and bulk methods

t-test and Wilcoxon perform surprisingly well

“we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq”



Methods

Criteria

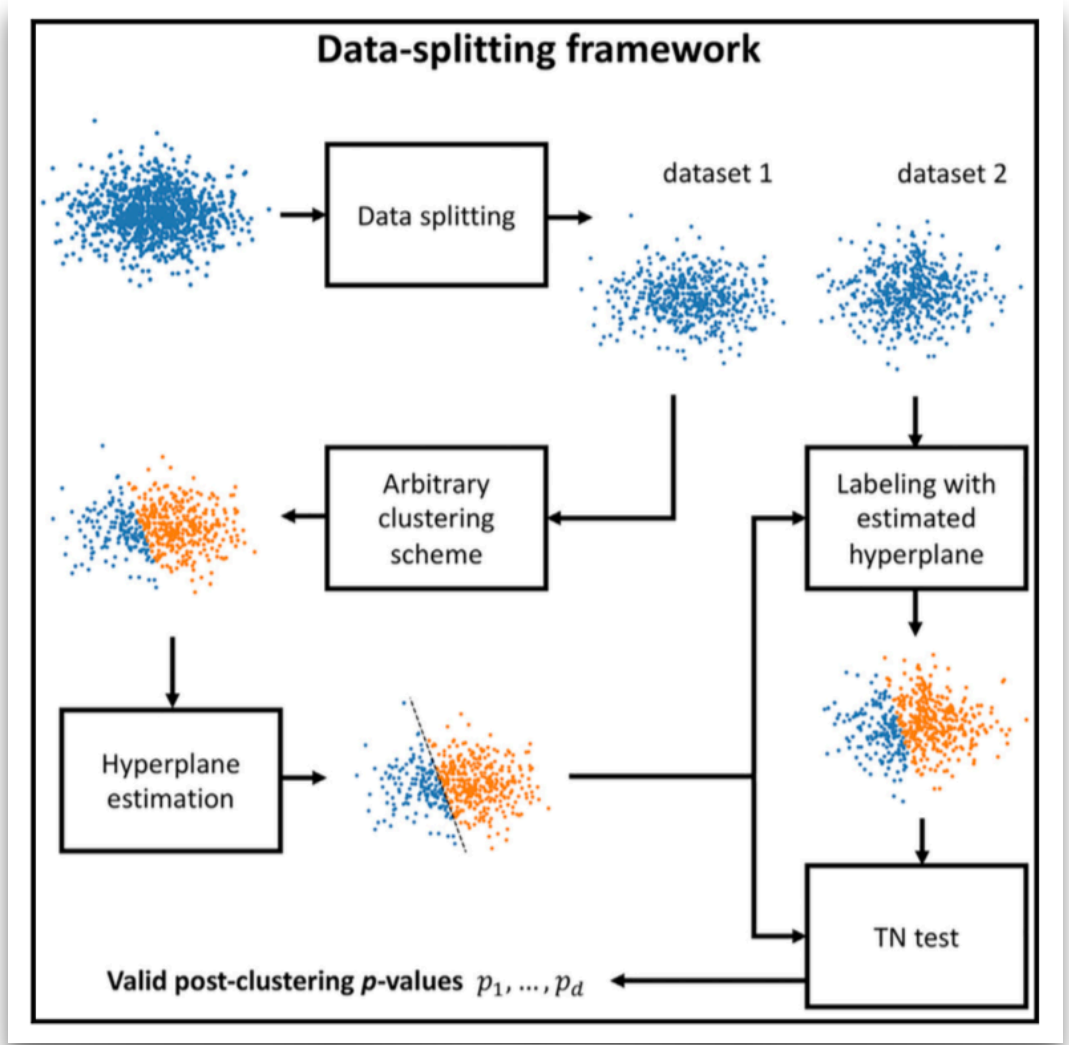
Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson[1,2] & Mark D Robinson[1,2]

Statistical issue lurking here: clustering and then testing differences between clusters leads to invalid P-values

Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq

Graphical Abstract



Authors

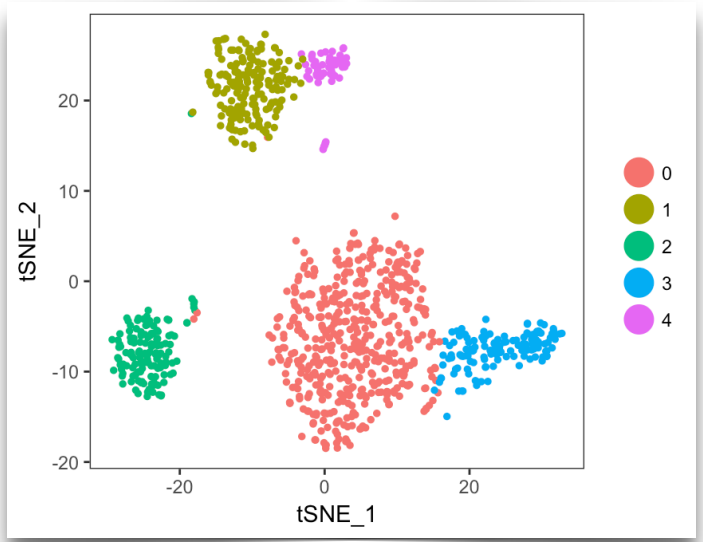
Jesse M. Zhang, Govinda M. Kamath, David N. Tse

Dimension reduction: general introduction

- Single cell data comes as a matrix of N cells x G features
- Each cell is a point in G-dimensional space
- Goal: represent the data in 2-3 dimensions, but preserve **structure** as best as possible (i.e., points that are **close** in G dimensions should be close in 2/3 dimensions)

Dimension reduction is versatile

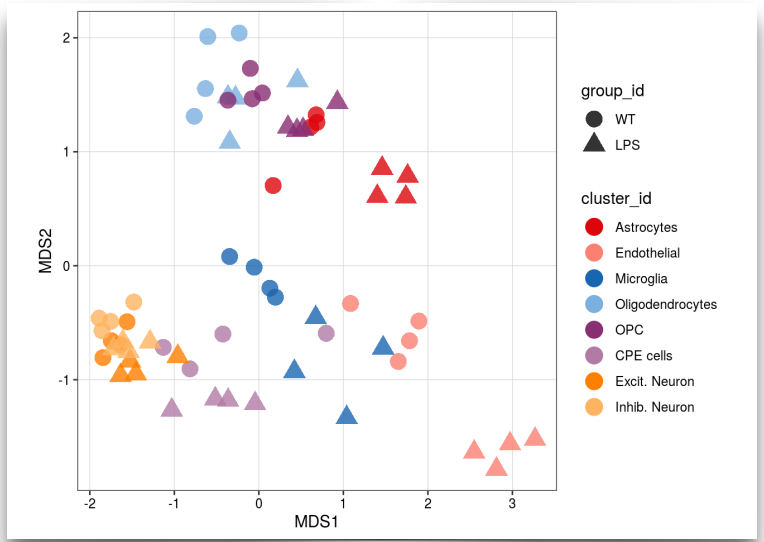
K features x N cells —> 2 dimensions x N cells



Each point = **single cell**

(10x PBMC)

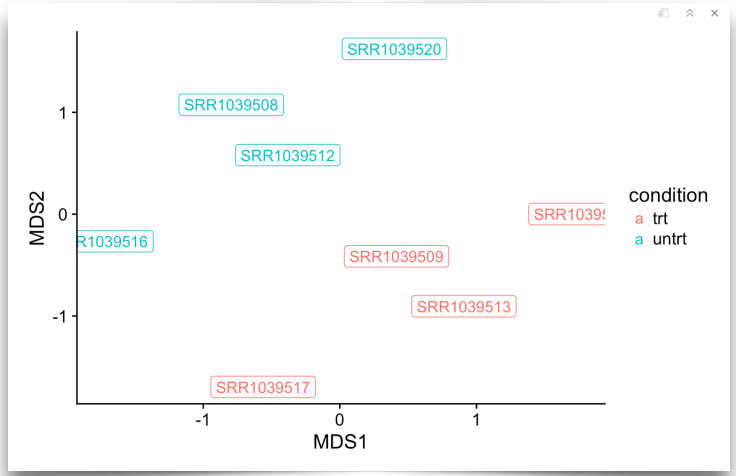
N cells x K features —> N cell subpopulations x 2 dimensions



Each point = **subpopulation from a single sample**

(LPS mouse cortex)

P samples x K features —> P samples x 2 dimensions

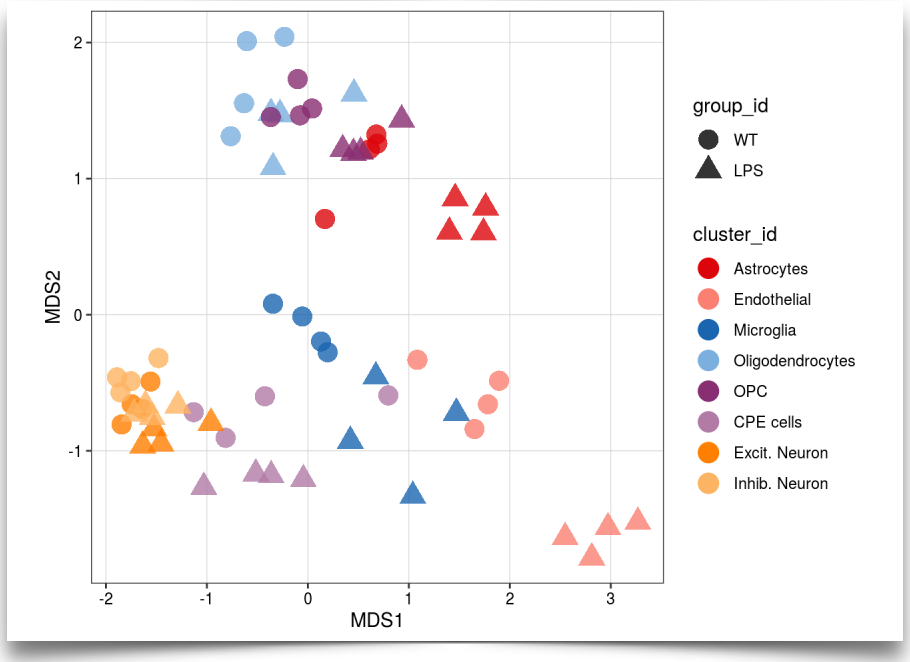


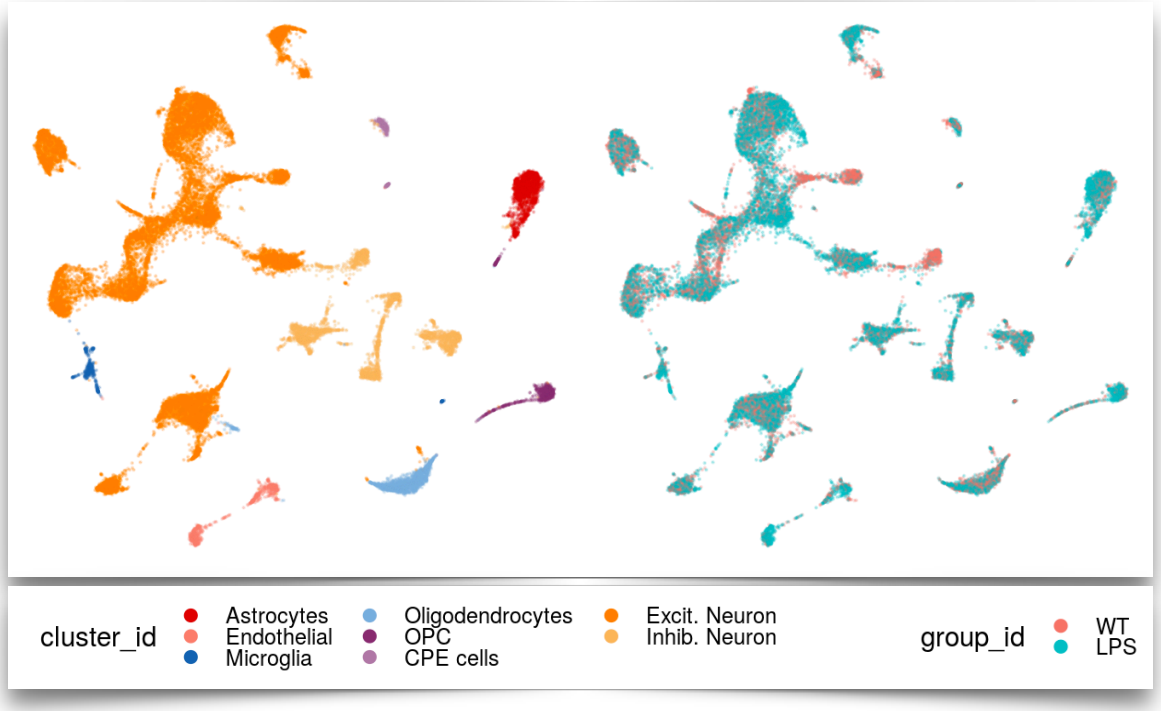
Each point = **sample**

(airway)

Dimension reduction: interpretation (single cell)

- Distances (in low-dimensional space) represent transcriptional changes: the larger the distance, the larger the transcriptional differences
- Hides: what are the transcriptional differences (e.g., few feature with large differences or many features with subtle differences)
- Local structure is typically better preserved than global structure





Cell type and cell state

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

Perspective

Defining cell types and states with single-cell genomics

Cole Trapnell

Department of Genome Sciences, University of Washington, Seattle, Washington 98105, USA

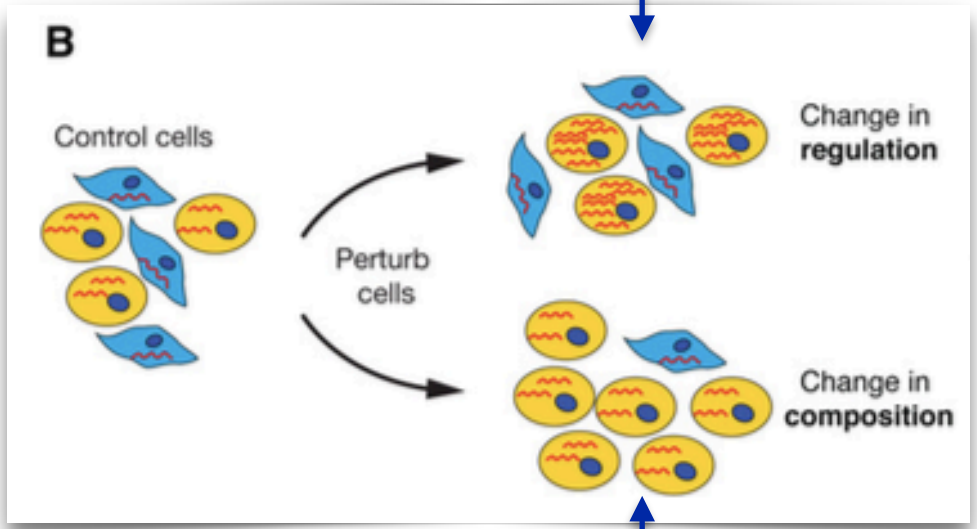
Box 1 The many facets of a cell's identity

We define a cell's *identity* as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its *type* (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its *state*. Cell types are often organized in a hierarchical

Type: more permanent

State: more transient

Differential state analysis



Differential abundance analysis

HYPOTHESIS

A periodic table of cell types

Bo Xia[1] and Itai Yanai[1,2,*]

„We view a cell state as a secondary module operating in addition to the general cell type regulatory program.“

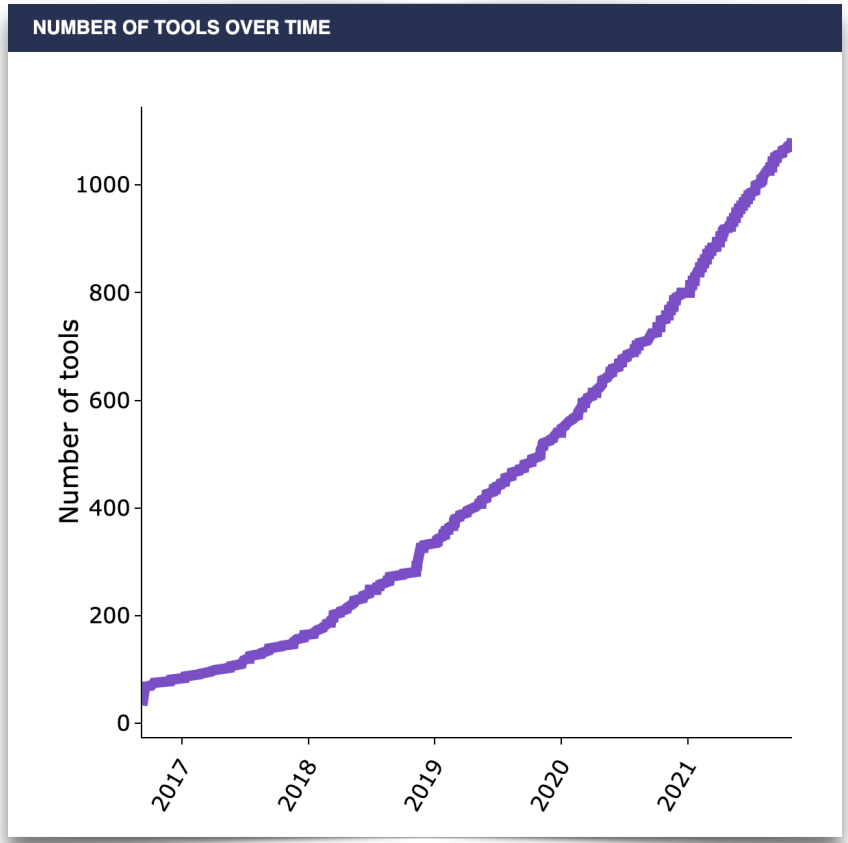
SPOTLIGHT

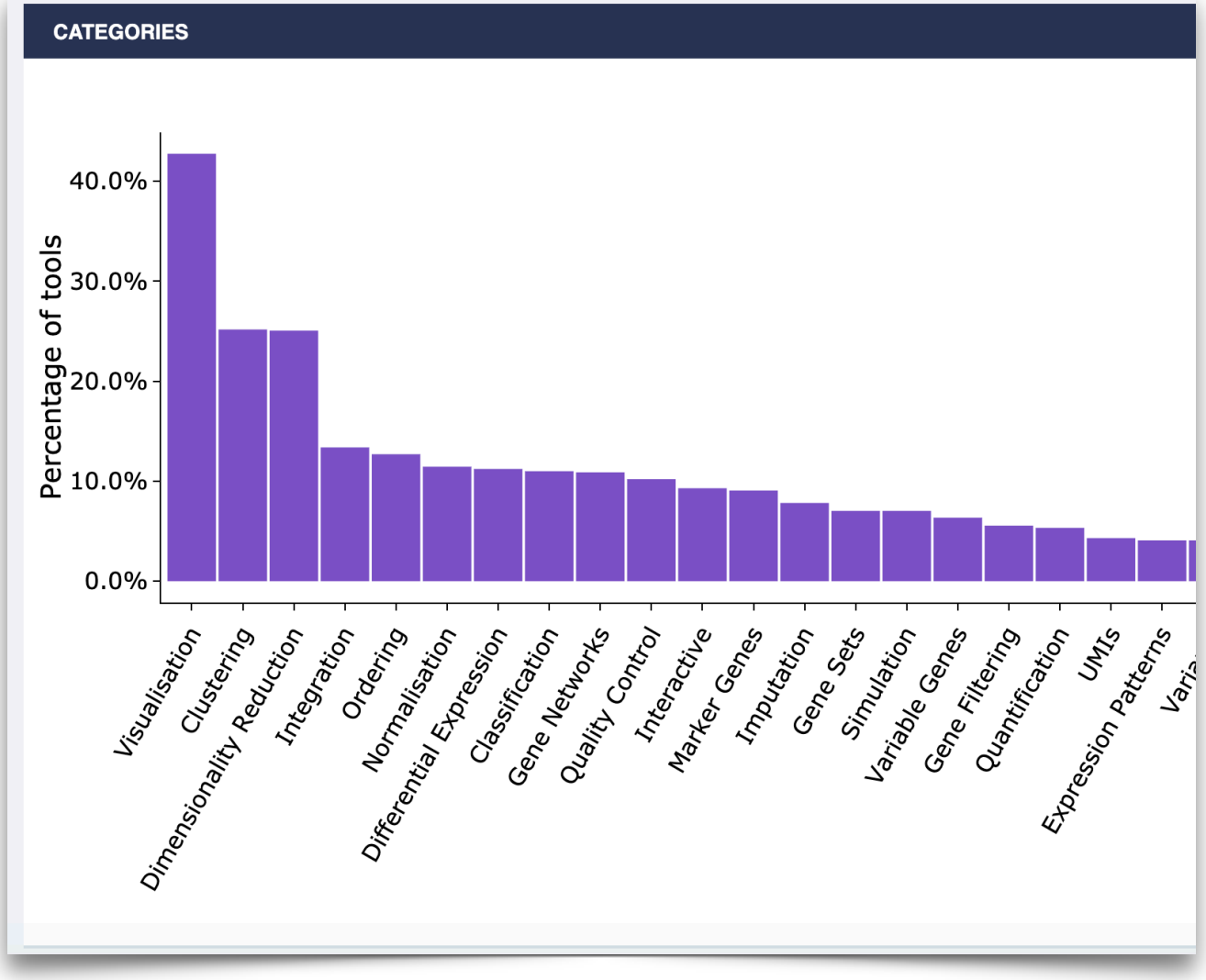
The evolving concept of cell identity in the single cell era

Samantha A. Morris[1,2,3,*]

„how can we be confident that a novel transcriptional signature represents a new cell type rather than a known cell type in an unrecognized state?“

Method velocity





REVIEW Open Access

Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape

Luke Zappia[1,2] and Fabian J. Theis[1,2,3*]

RESEARCH ARTICLE

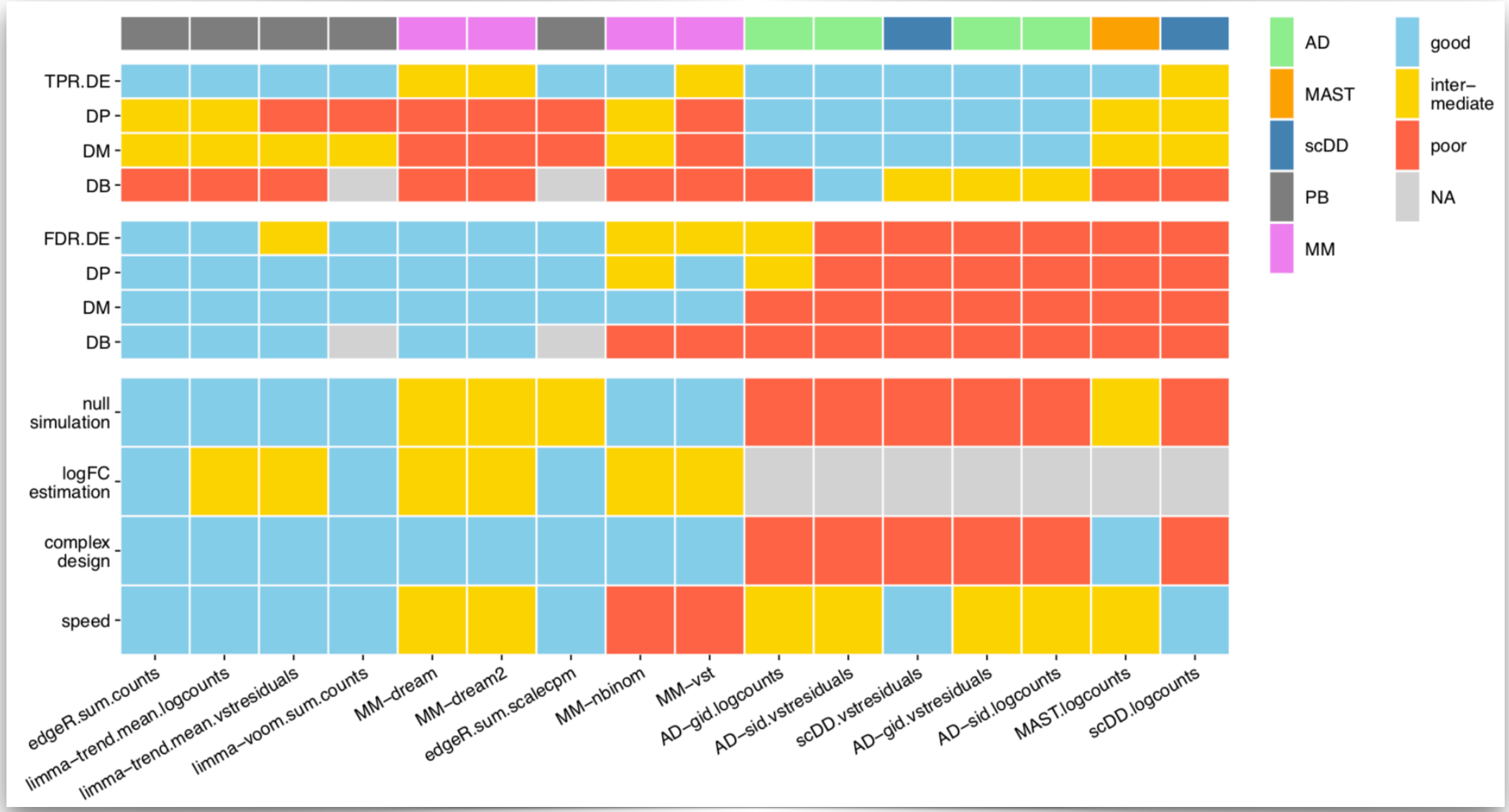
Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database

Luke Zappia[1,2], Belinda Phipson[1], Alicia Oshlack[1,2*]

1 Bioinformatics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia, **2** School of Biosciences, Faculty of Science, University of Melbourne, Melbourne, Victoria, Australia

Current rating: differential state analysis

PB = pseudobulk
AD = Anderson-Darling
MM = mixed models



Helena

Pseudobulks FTW!

.. pseudobulk approaches provide an excellent trade-off between speed and accuracy for single-cell DE analysis.

Here, we demonstrate that the central principle underlying valid DE analysis [for cross-sample single cell RNA-Seq] is the ability of statistical methods to account for the intrinsic variability of biological replicates.

there is an urgent need for a paradigm shift in the statistical methods that are used for DE analysis of single-cell data. The need is underscored by our observation that most studies published in the past two years have used inappropriate statistical methods for DE analysis

ARTICLE

Check for updates

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

Confronting false discoveries in single-cell differential expression

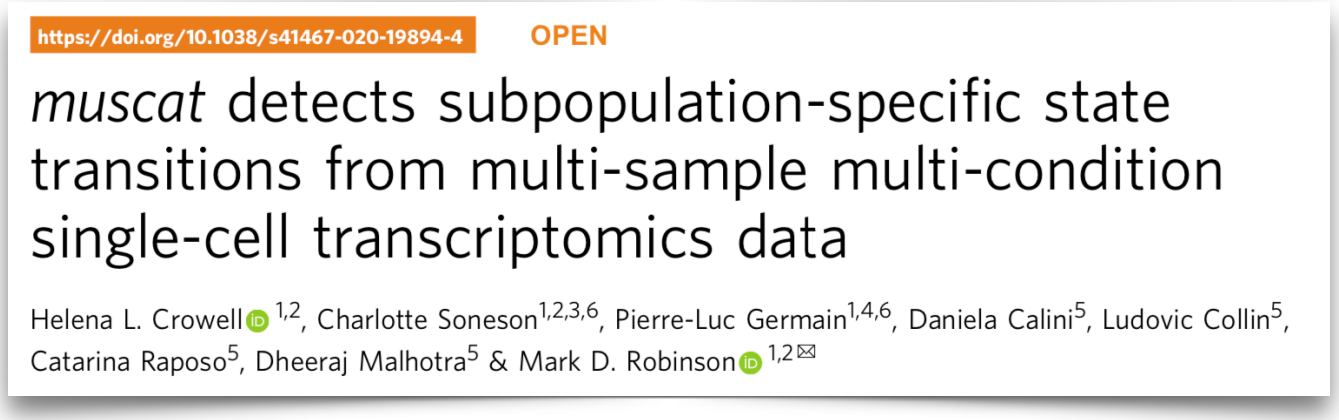
Jordan W. Squair[1,2,3], Matthieu Gautier[1,2], Claudia Kathe[1,2], Mark A. Anderson[1,2], Nicholas D. James[1,2], Thomas H. Hutson[1,2], Rémi Hudelle[1,2], Taha Qaiser[3], Kaya J. E. Matson[4], Quentin Barraud[1,2], Ariel J. Levine[4], Gioele La Manno[1], Michael A. Skinnider[1,2,5,6]✉ & Grégoire Courtine[1,2,6]✉

Compositional data

- "*The most important characteristic of compositional data is that they carry only relative information*" [1]
- "*Standard statistical methods are not appropriate for analyzing compositional data*" [2]
- "*DESeq2 and EdgeR implicitly assume that the absolute abundances do not change due to the treatment. This is equivalent to using the Centered Log-Ratio (CLR) transformation from the CoDA methods*" [3]

[1] http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf
[2] https://www.nature.com/articles/s41467-020-17041-7
[3] https://towardsdatascience.com/relative-vs-absolute-how-to-do-compositional-data-analyses-part-2-f554eb9b26e

What would make DS more flexible?


https://doi.org/10.1038/s41467-020-19894-4 OPEN

muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data

Helena L. Crowell[1,2], Charlotte Soneson[1,2,3,6], Pierre-Luc Germain[1,4,6], Daniela Calini[5], Ludovic Collin[5], Catarina Raposo[5], Dheeraj Malhotra[5] & Mark D. Robinson[1,2]✉

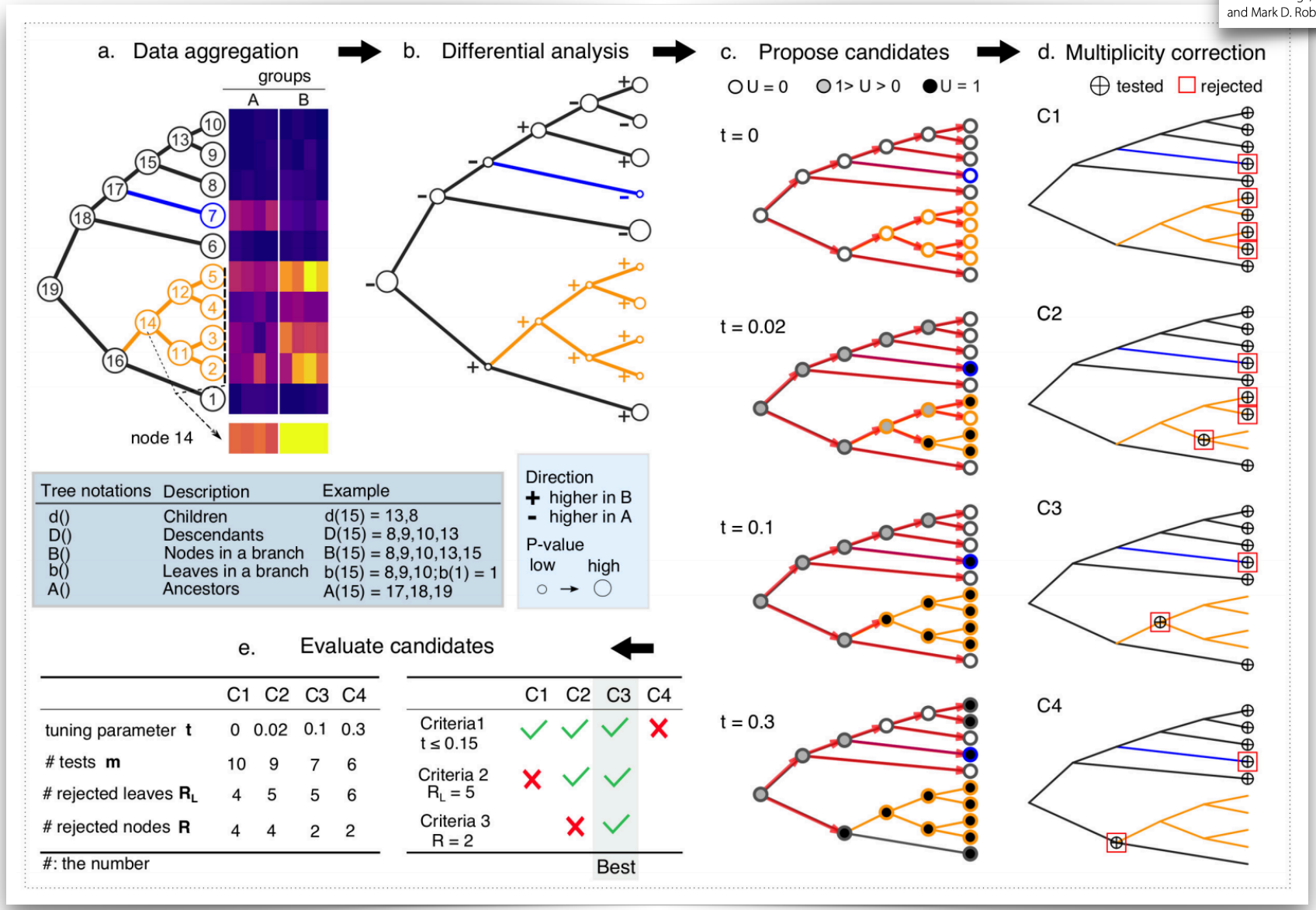
- What (clustering/annotation) resolution should you use? —> options include: i) use a tree of subpopulation relations to guide the inference of differential expression; ii) operate at neighbourhood level
- Can we do better by looking at full distributions instead of aggregating?

Motivation: can we use the tree information and perform differential inferences across resolutions?

METHOD Open Access

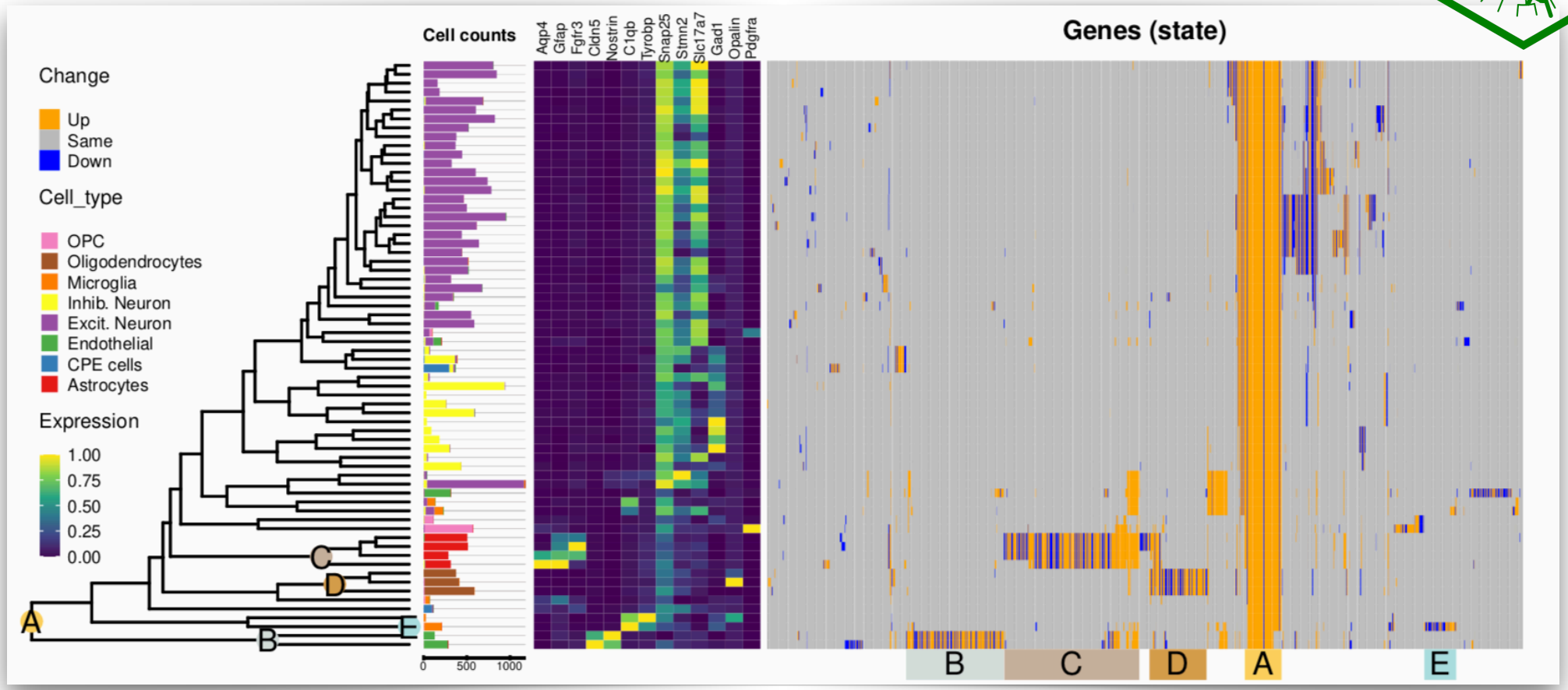
treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses

Ruizhu Huang[1], Charlotte Soneson[1,2], Pierre-Luc Germain[1,3], Thomas S.B. Schmidt[1,4], Christian Von Mering[1] and Mark D. Robinson[1*]



Fiona

Analyses across resolutions

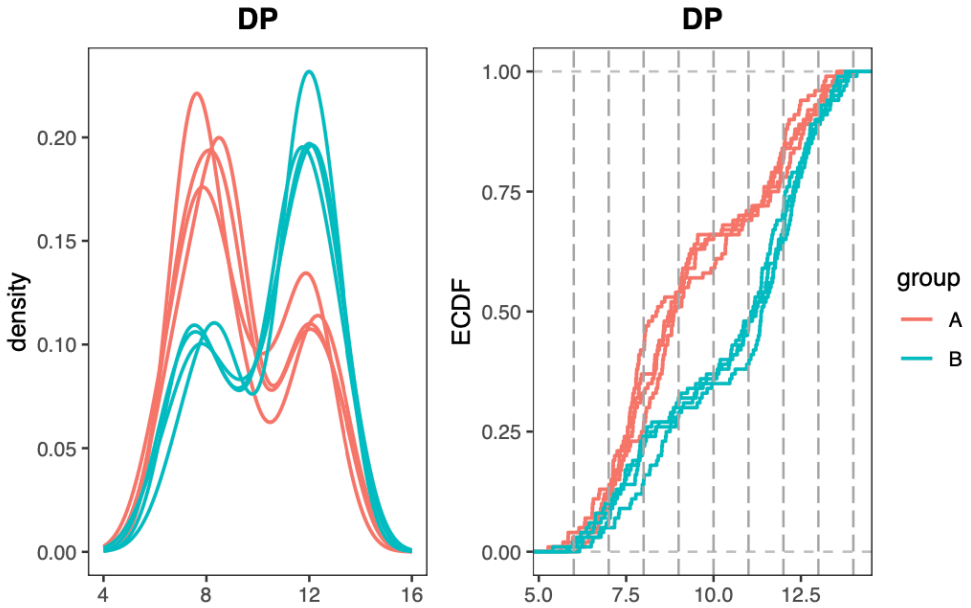


(here: scRNA-seq LPS dataset **over-clustered** to subdivide certain subpopulations)

Fiona

Differential state from full distributions

Simone Tiberi



- We consider several cut-offs at which we compute the mean difference between the ECDFs in the two groups.
- Our test statistic is the summation of these differences across all cut-offs:

$$s^{obs} = \sum_{c=1}^{N_{cutoffs}} \left| \overline{ECDF}_c^{(A)} - \overline{ECDF}_c^{(B)} \right|. \quad (1)$$

Bioconductor package ‘distinct’

null distribution by shuffling cells across samples.

Related literature to explore ..

scCODA: A Bayesian model for compositional single-cell data analysis

Büttner M.[1+], Ostner J.[1,2+], Müller CL.[1,2,3*], Theis FJ.[1,4,5†], Schubert B.[1,4†*]

Differential abundance testing on single-cell data using *k*-nearest neighbor graphs

Emma Dann[1], Neil C. Henderson[2,3], Sarah A. Teichmann[1,4], Michael D. Morgan[5,6] ✉ and John C. Marioni[1,5,6] ✉

Differential abundance

Cell Type Composition Analysis: Comparison of statistical methods

Sean Simmons
Stanley Center, Broad Institute

Differential abundance + Differential state

Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign

Sisi Chen[a,b,1], Paul Rivaud[a,b], Jong H. Park[a,b], Tiffany Tsou[a,b], Emeric Charles[c], John R. Haliburton[d], Flavia Pichiorri[e], and Matt Thomson[a,b,1]

Case-control analysis of single-cell RNA-seq studies

Viktor Petukhov[1,2,*], Anna Igolkina[3,*], Rasmus Rydbirk[1], Shenglin Mei[2], Lars Christoffersen[1], Konstantin Khodosevich[1,§], Peter V. Kharchenko[2,4,§]

Quantifying the effect of experimental perturbations at single-cell resolution

Daniel B. Burkhardt[1,8], Jay S. Stanley III[2,8], Alexander Tong[3], Ana Luisa Perdigoto[4], Scott A. Gigante[2], Kevan C. Herold[4], Guy Wolf[6,7,9], Antonio J. Giraldez[1,9], David van Dijk[5,9] ✉ and Smita Krishnaswamy[1,3,9] ✉

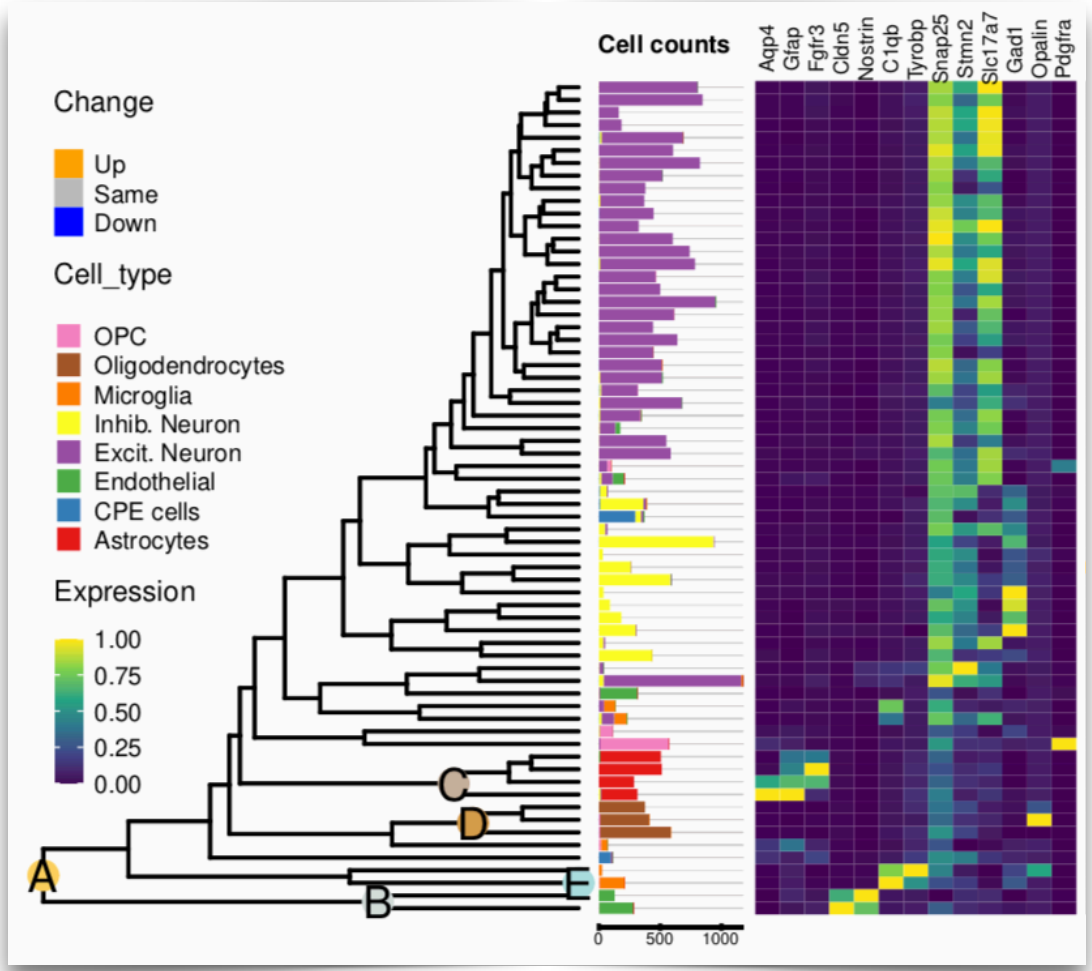
Differential state

■ ■

METHOD Open Access

treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses

Ruizhu Huang[1], Charlotte Soneson[1,2], Pierre-Luc Germain[1,3], Thomas S.B. Schmidt[1,4], Christian Von Mering[1] and Mark D. Robinson[1*]



Differential abundance testing on single-cell data using *k*-nearest neighbor graphs

Emma Dann[1], Neil C. Henderson[2,3], Sarah A. Teichmann[1,4], Michael D. Morgan[5,6] ✉ and John C. Marioni[1,5,6] ✉