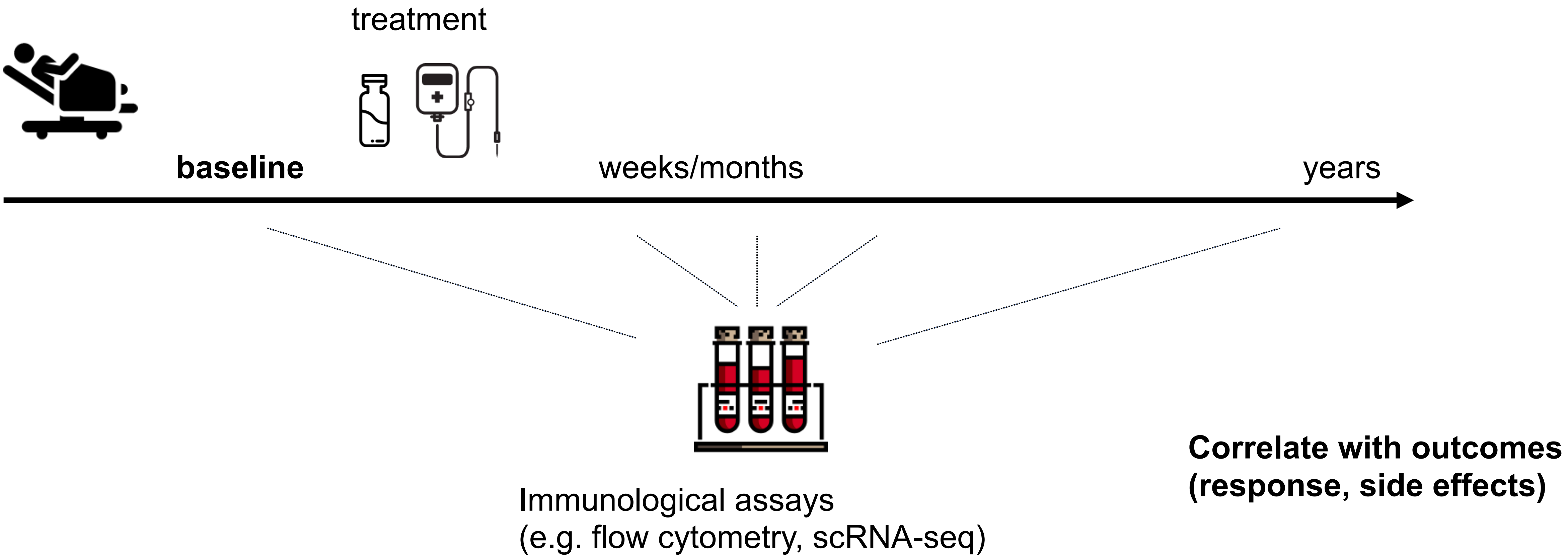
Computational challenges in single-cell data analysis

Raphael Gottardo, PhD

Professor of Biomedical Data Sciences

University of Lausanne and University Hospital of Lausanne

Basics of immune correlate studies



treatment

baseline

weeks/months

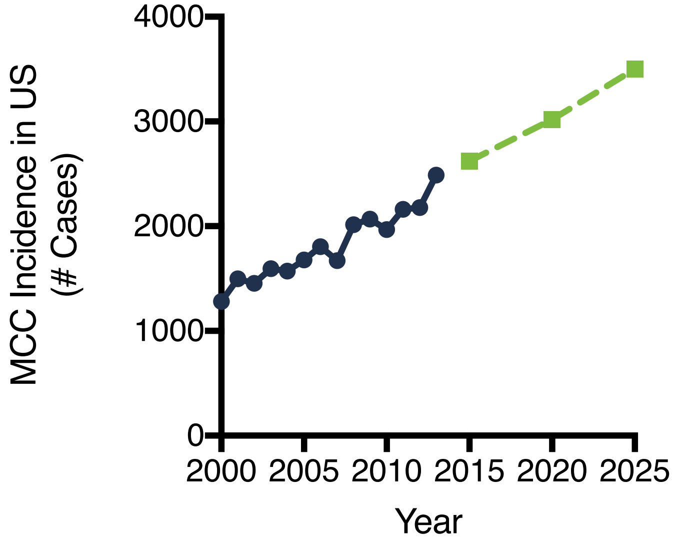
years

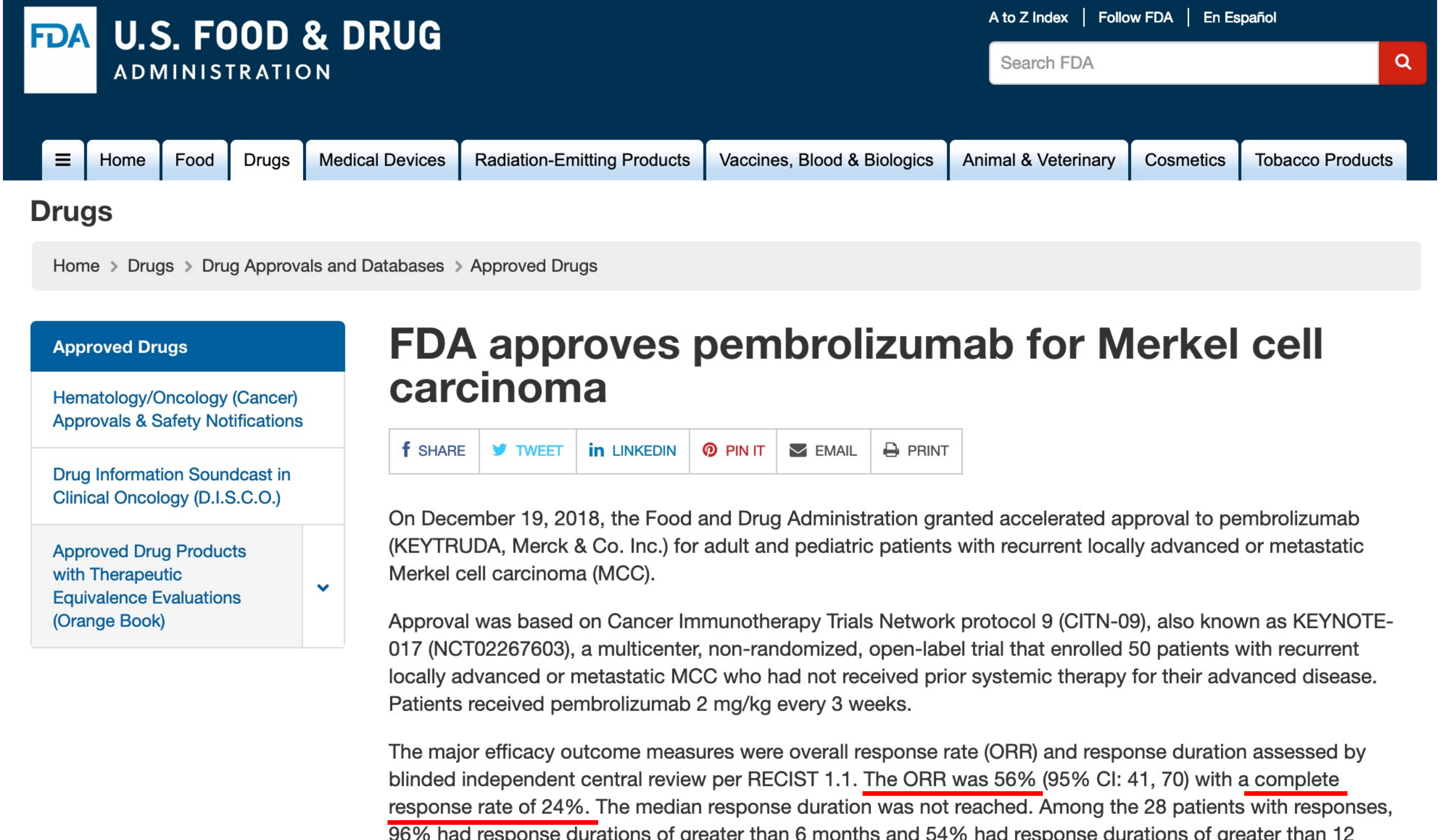
Immunological assays
(e.g. flow cytometry, scRNA-seq)

Correlate with outcomes
(response, side effects)

Motivation: The Case of Merkel Cell Carcinoma

- Increasingly common skin cancer
- Very aggressive: twice as likely to metastasize as melanoma
- 80%+ of cases are caused by a virus (Merkel cell polyomavirus)
- Very difficult to treat until recently




FDA U.S. FOOD & DRUG ADMINISTRATION
A to Z Index
Follow FDA
En Español
Search FDA
Home
Food
Drugs
Medical Devices
Radiation-Emitting Products
Vaccines, Blood & Biologics
Animal & Veterinary
Cosmetics
Tobacco Products
Drugs
Home > Drugs > Drug Approvals and Databases > Approved Drugs
Approved Drugs
Hematology/Oncology (Cancer) Approvals & Safety Notifications
Drug Information Soundcast in Clinical Oncology (D.I.S.C.O.)
Approved Drug Products with Therapeutic Equivalence Evaluations (Orange Book)
SHARE
TWEET
LINKEDIN
PIN IT
EMAIL
PRINT

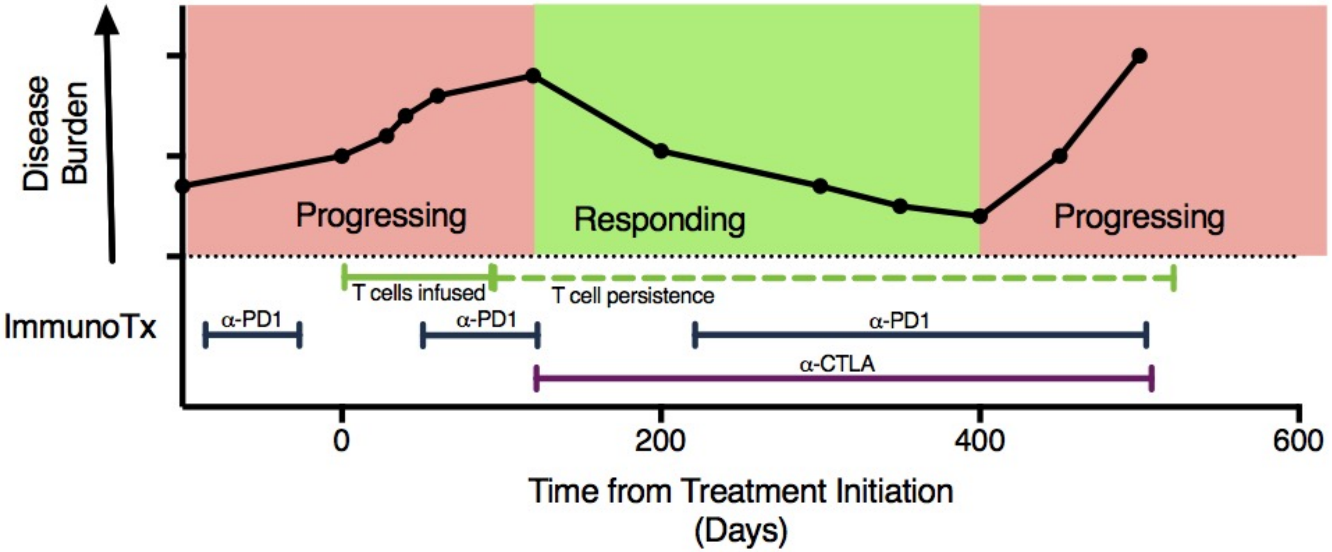
FDA approves pembrolizumab for Merkel cell carcinoma

On December 19, 2018, the Food and Drug Administration granted accelerated approval to pembrolizumab (KEYTRUDA, Merck & Co. Inc.) for adult and pediatric patients with recurrent locally advanced or metastatic Merkel cell carcinoma (MCC).

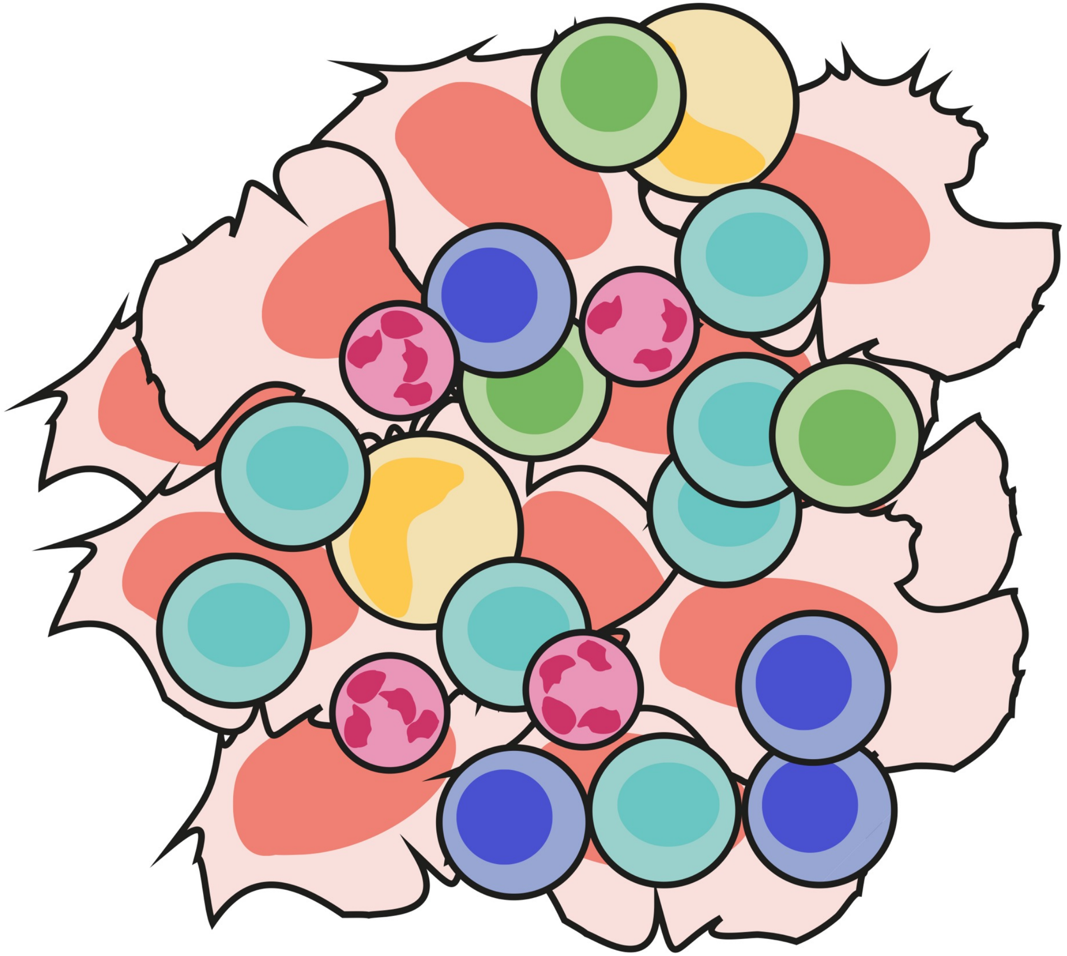
Approval was based on Cancer Immunotherapy Trials Network protocol 9 (CITN-09), also known as KEYNOTE-017 (NCT02267603), a multicenter, non-randomized, open-label trial that enrolled 50 patients with recurrent locally advanced or metastatic MCC who had not received prior systemic therapy for their advanced disease. Patients received pembrolizumab 2 mg/kg every 3 weeks.

The major efficacy outcome measures were overall response rate (ORR) and response duration assessed by blinded independent central review per RECIST 1.1. The ORR was 56% (95% CI: 41, 70) with a complete response rate of 24%. The median response duration was not reached. Among the 28 patients with responses, 96% had response durations of greater than 6 months and 54% had response durations of greater than 12

A case of immune escape



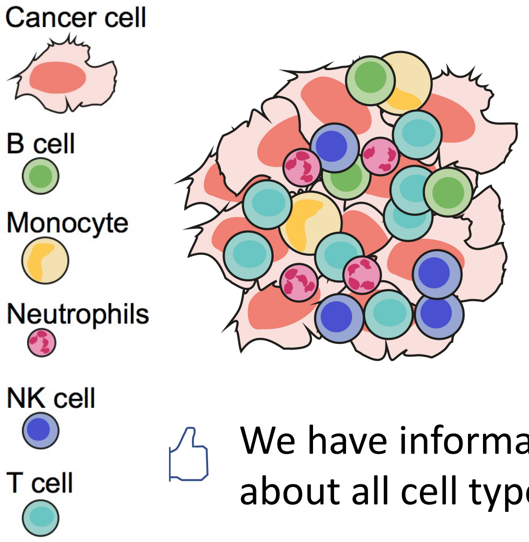
- 60 year old man with metastatic Merkel cell carcinoma, virus positive
- Has HLA-B*3502 restricted T cells recognizing an epitope of the Merkel cell polyomavirus
- Receives T cell therapy and 2 immune checkpoint inhibitors
- Has an initial impressive partial response of >1 year, but then progression.
- At time of progression, has antigen (MCPyV positive tumor) and persistent T cells recognizing MCPyV



“Bulk” (e.g. RNA-seq) profiling cannot reveal *within* tissue heterogeneity

RNA-seq limitations

Bulk sample gene expression analysis

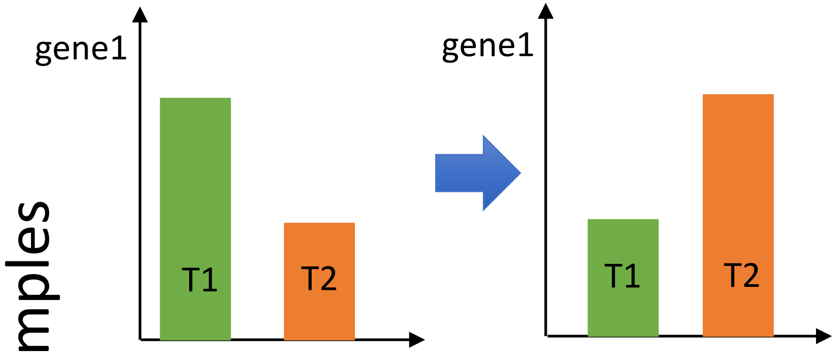


👍 We have information about all cell types

👎 The signal is mixed with other cell types

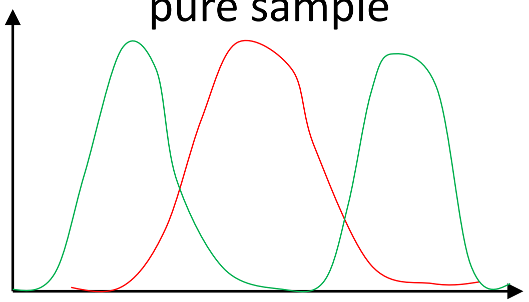
Examples

Different expression in different cell types in mixed sample



Same average

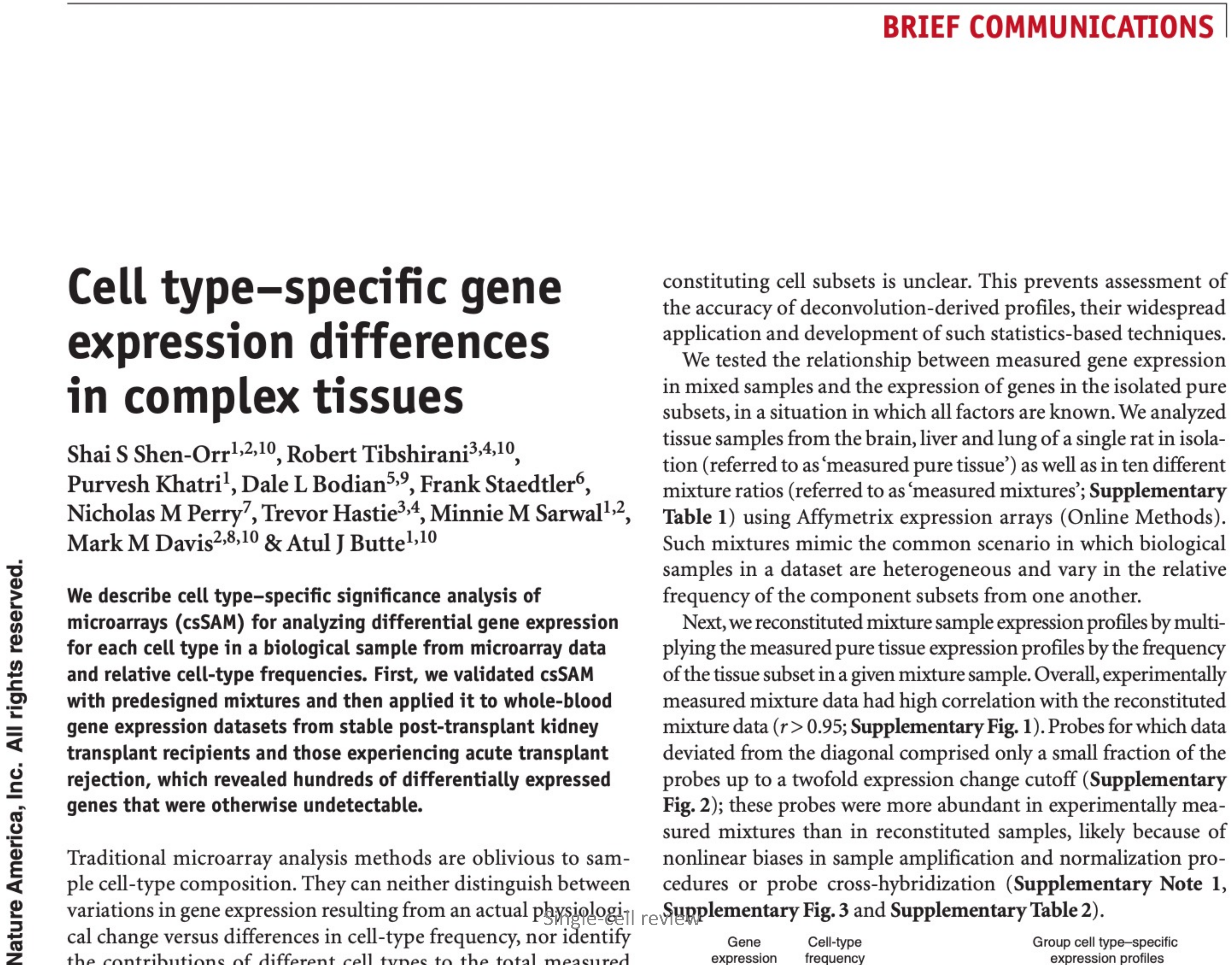
Bimodal expression in a pure sample



Same average

Possible in some cases to infer the fraction of the different cell types, but impossible to infer their actual gene expression profile

Credit: David Gfeller



Cell type–specific gene expression differences in complex tissues

Shai S Shen-Orr[1,2,10], Robert Tibshirani[3,4,10], Purvesh Khatri[1], Dale L Bodian[5,9], Frank Staedtler[6], Nicholas M Perry[7], Trevor Hastie[3,4], Minnie M Sarwal[1,2], Mark M Davis[2,8,10] & Atul J Butte[1,10]

We describe cell type–specific significance analysis of microarrays (csSAM) for analyzing differential gene expression for each cell type in a biological sample from microarray data and relative cell-type frequencies. First, we validated csSAM with predesigned mixtures and then applied it to whole-blood gene expression datasets from stable post-transplant kidney transplant recipients and those experiencing acute transplant rejection, which revealed hundreds of differentially expressed genes that were otherwise undetectable.

Traditional microarray analysis methods are oblivious to sample cell-type composition. They can neither distinguish between variations in gene expression resulting from an actual physiological change versus differences in cell-type frequency, nor identify the contributions of different cell types to the total measured

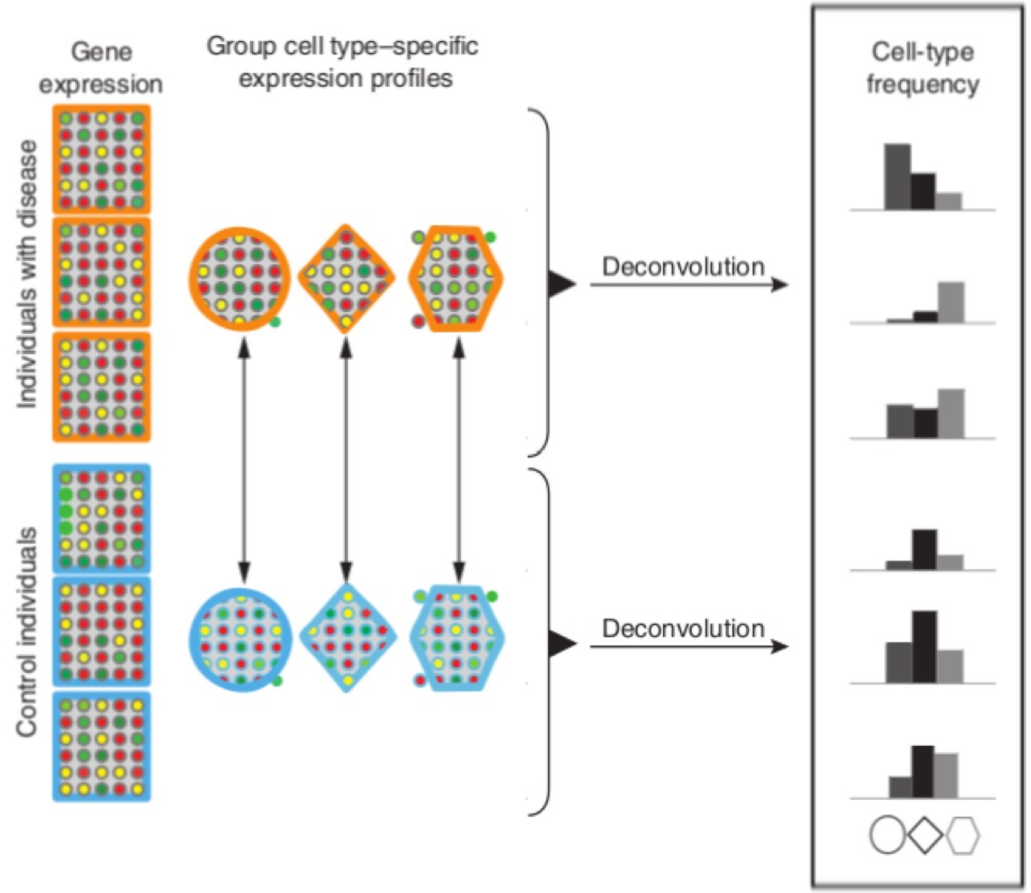
constituting cell subsets is unclear. This prevents assessment of the accuracy of deconvolution-derived profiles, their widespread application and development of such statistics-based techniques.

We tested the relationship between measured gene expression in mixed samples and the expression of genes in the isolated pure subsets, in a situation in which all factors are known. We analyzed tissue samples from the brain, liver and lung of a single rat in isolation (referred to as 'measured pure tissue') as well as in ten different mixture ratios (referred to as 'measured mixtures'; **Supplementary Table 1**) using Affymetrix expression arrays (Online Methods). Such mixtures mimic the common scenario in which biological samples in a dataset are heterogeneous and vary in the relative frequency of the component subsets from one another.

Next, we reconstituted mixture sample expression profiles by multiplying the measured pure tissue expression profiles by the frequency of the tissue subset in a given mixture sample. Overall, experimentally measured mixture data had high correlation with the reconstituted mixture data ($r > 0.95$; **Supplementary Fig. 1**). Probes for which data deviated from the diagonal comprised only a small fraction of the probes up to a twofold expression change cutoff (**Supplementary Fig. 2**); these probes were more abundant in experimentally measured mixtures than in reconstituted samples, likely because of nonlinear biases in sample amplification and normalization procedures or probe cross-hybridization (**Supplementary Note 1, Supplementary Fig. 3** and **Supplementary Table 2**).

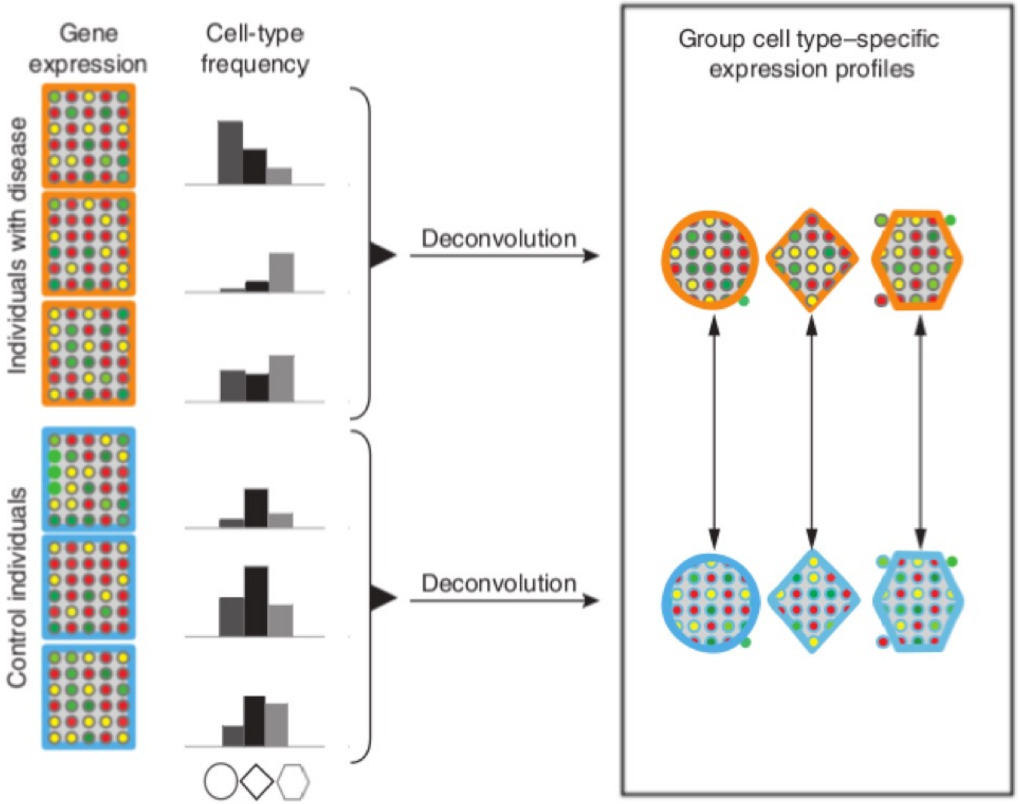
Computational deconvolution

(a) *Partial deconvolution using available signatures*



G=**F**xC + ϵ (infer F)

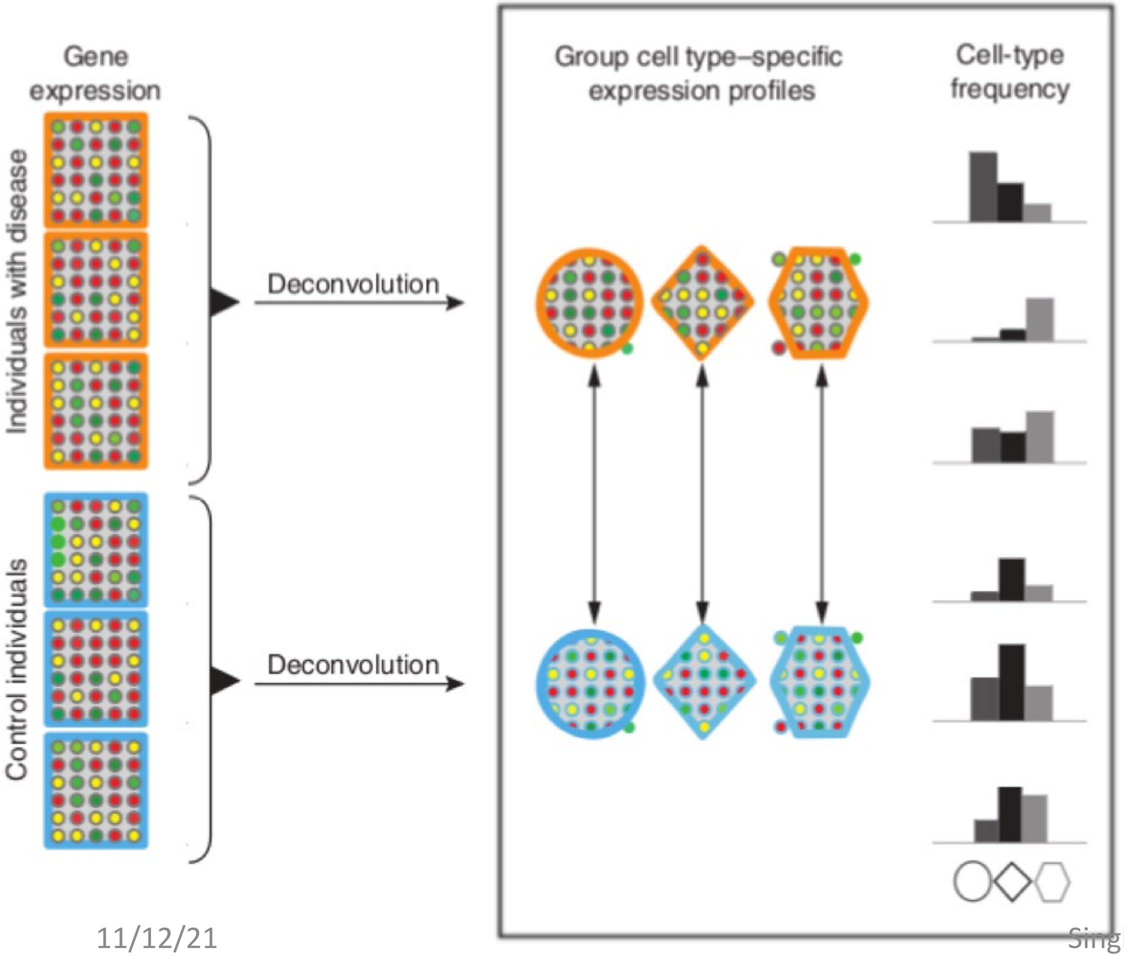
(b) *Partial deconvolution using available proportions*



G=Fx**C** + ϵ (infer C)

Computational deconvolution

(c) *Complete deconvolution from global expression*



$G=\mathbf{F}x\mathbf{C} + \epsilon$ (infer F and C). Need additional constraints.
Can be done, e.g., via non-negative matrix factorization or
Bayesian modeling, to name a few methods.

A large number of available tools

- csSAM: General method for cell-type specific differential expression. (Shen-Orr et al. *Nat. Methods* 2010)

- EPIC: Estimating the Proportion of Immune and Cancer cells from bulk tumor gene expression data (Racle et al.,*eLife* 2017).

- CIBERSORT: Many types of immune cell reference profiles (Newman et al. *Nat Meth*. 2015).

- TIMER: Main immune cell types (Li et al. *Genome Biology* 2016).

- MCPcounter: Immune + stromal cells, good performance, but results not comparable across different cell types (Becht E et al. *Genome Biology* 2016).

- CellMix package in R is a good resource (Gaujoux and Seoighe, *Bioinformatics* 2013)

Many of these algorithms have now improved given the large number of gene expression signatures in the public domain (including single-cell level data).

Bulk vs. single-cell

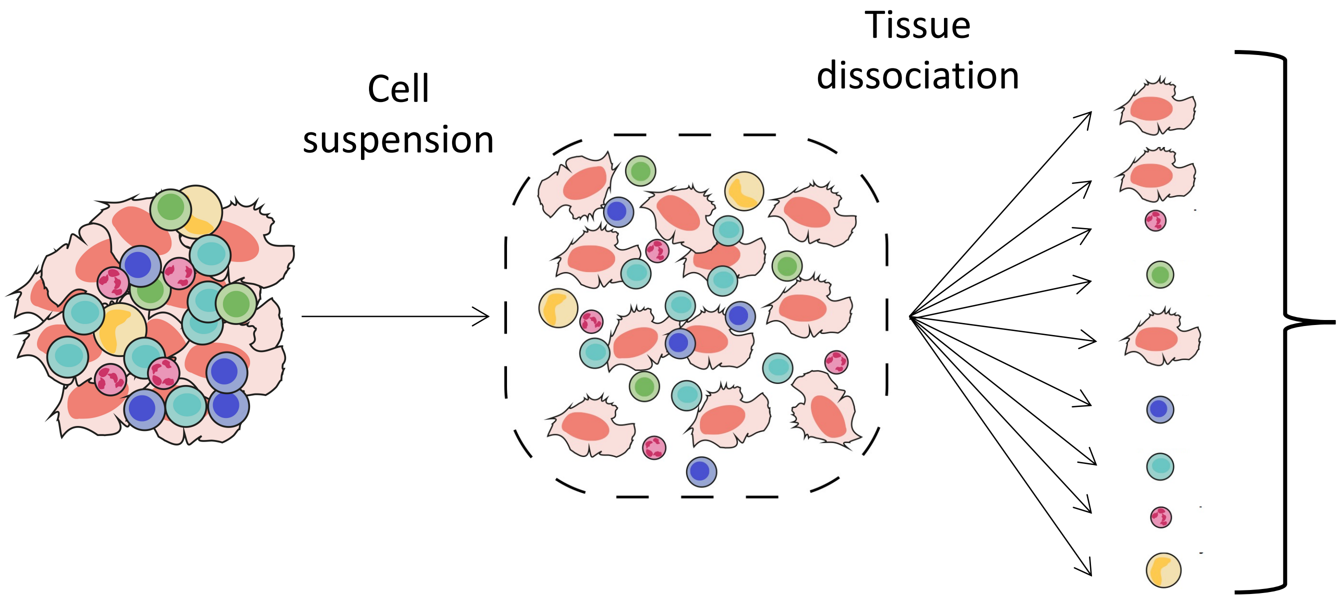
Bulk RNA-seq
(~2008)

Single-cell RNA-seq
(~2014)

While deconvolution can be done, it fails when the cell-type proportions are small and/or cell types not well defined.

Hard to taste 1-2 blueberries in a smoothie or tell that there are raspberries (if you've never tasted a raspberry).

Single-cell analysis



Cell suspension

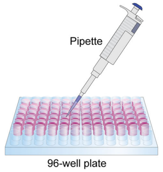
Tissue dissociation

Quantify RNA and/or protein in each single-cell

The progression of single-cell technologies over time (much simplified!)

RNA

Thousands of genes, 1 cell



2009

Thousands of genes, thousands of cells/sample

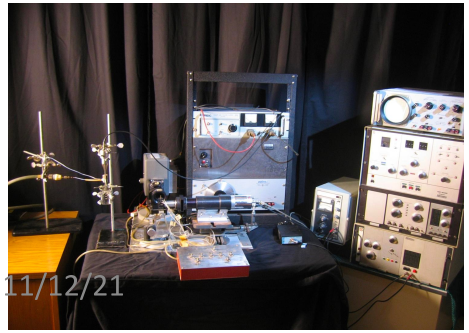
sciRNA-seq
Drop-seq
Well-seq
iCell 8

Today

Protein

1970

10,000 of cells, 1 protein

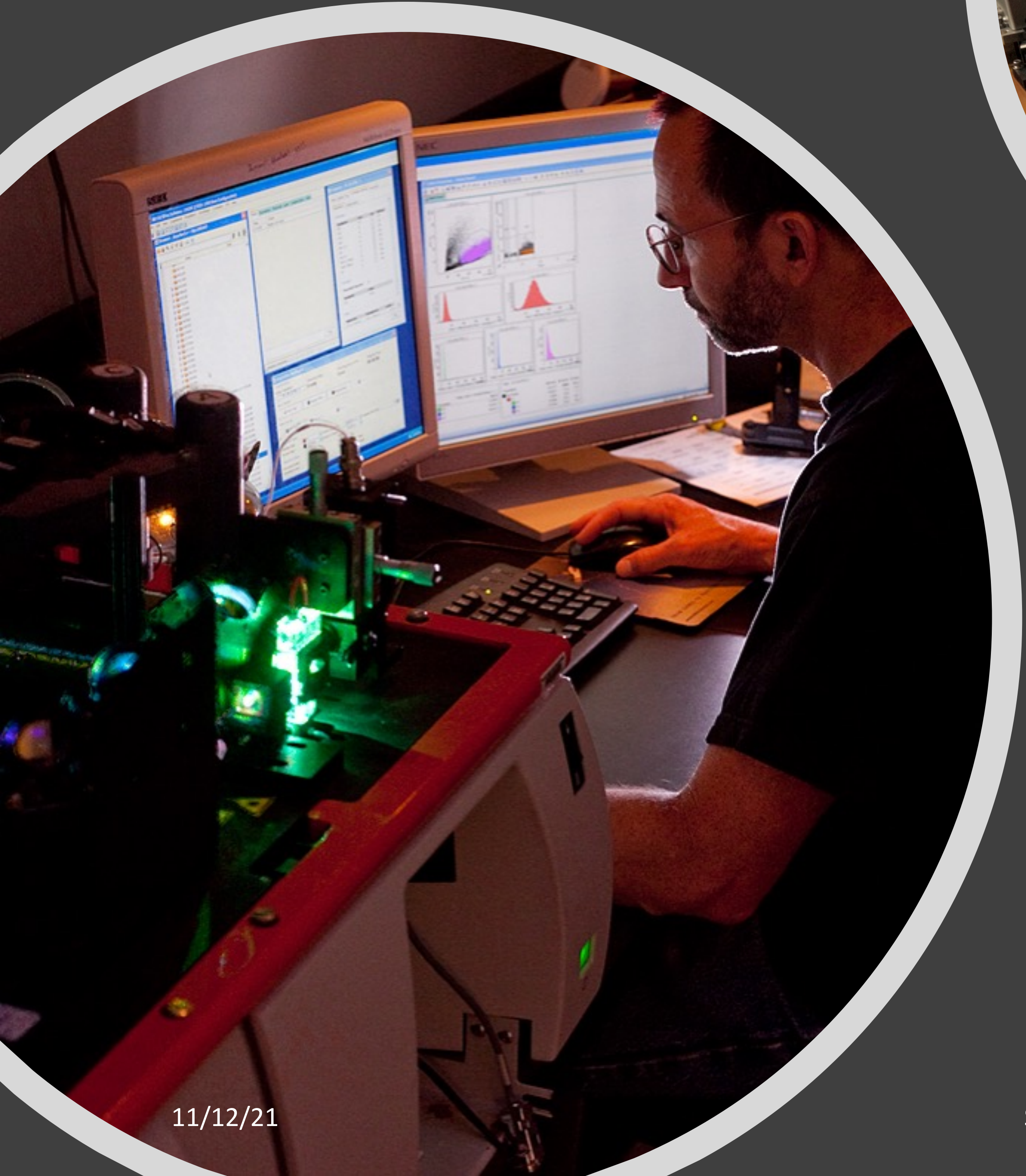


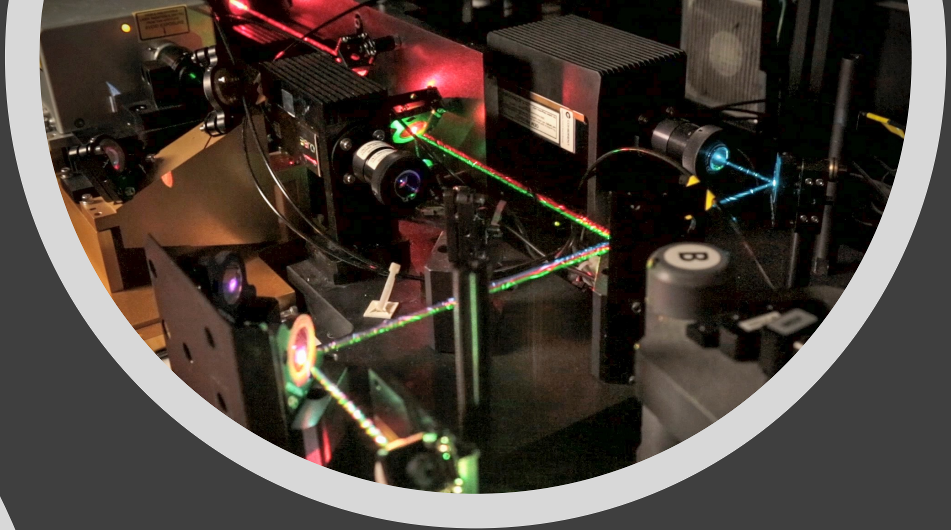
50-100 proteins, 1M-100M cells/sample

Single-cell analysis: Two main technologies

	Throughput	**Cost**	**Dimensionality**	**Pros**	**Cons**
Flow cytometry 	Very high (100,000 to millions)	Low (<.1$/cell)	15-50 proteins	Standardized, Targeted, Sorting	Limited number of proteins
scRNAseq (and variants) **10x Genomics** split-seq sciRNA-seq, etc 	High (10'000s cells)	High (~1$/cell)	High (1000s of genes)	High-dimensional, no need to select genes. Can now be combined with protein/epigenomics	Still expensive, much slower, more noisy

The two technologies are complementary!

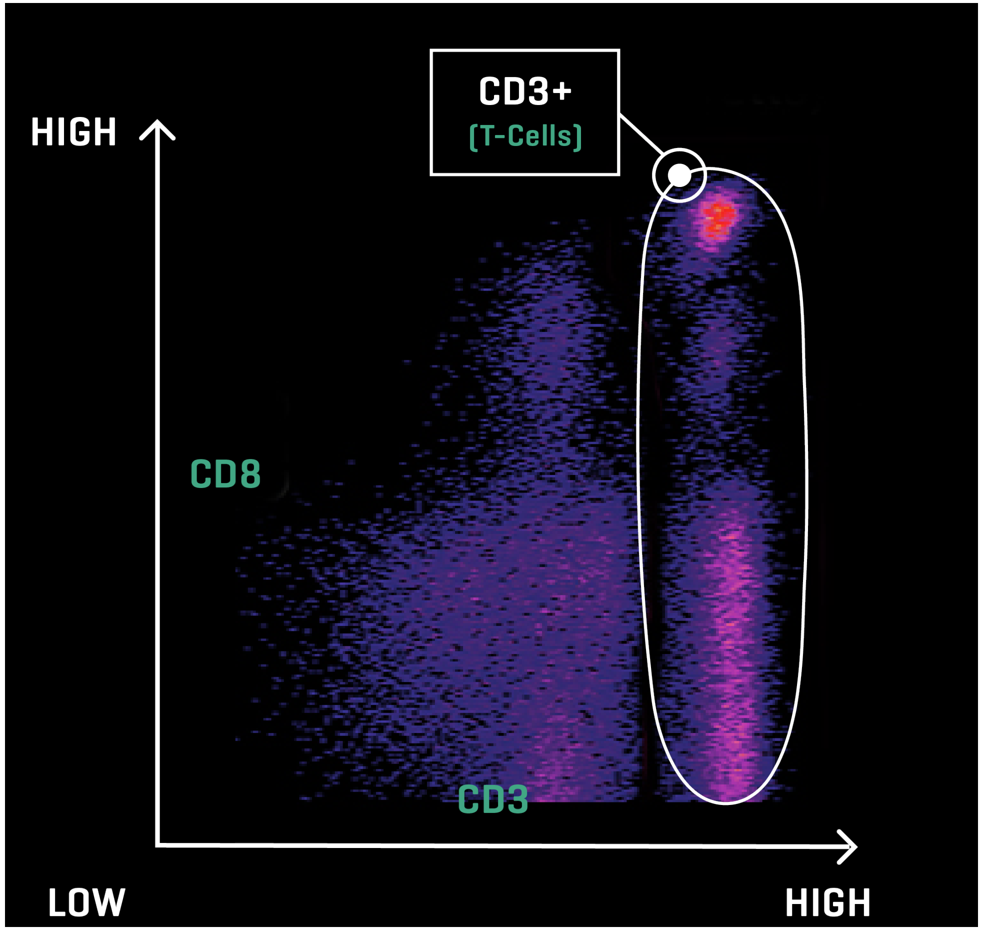


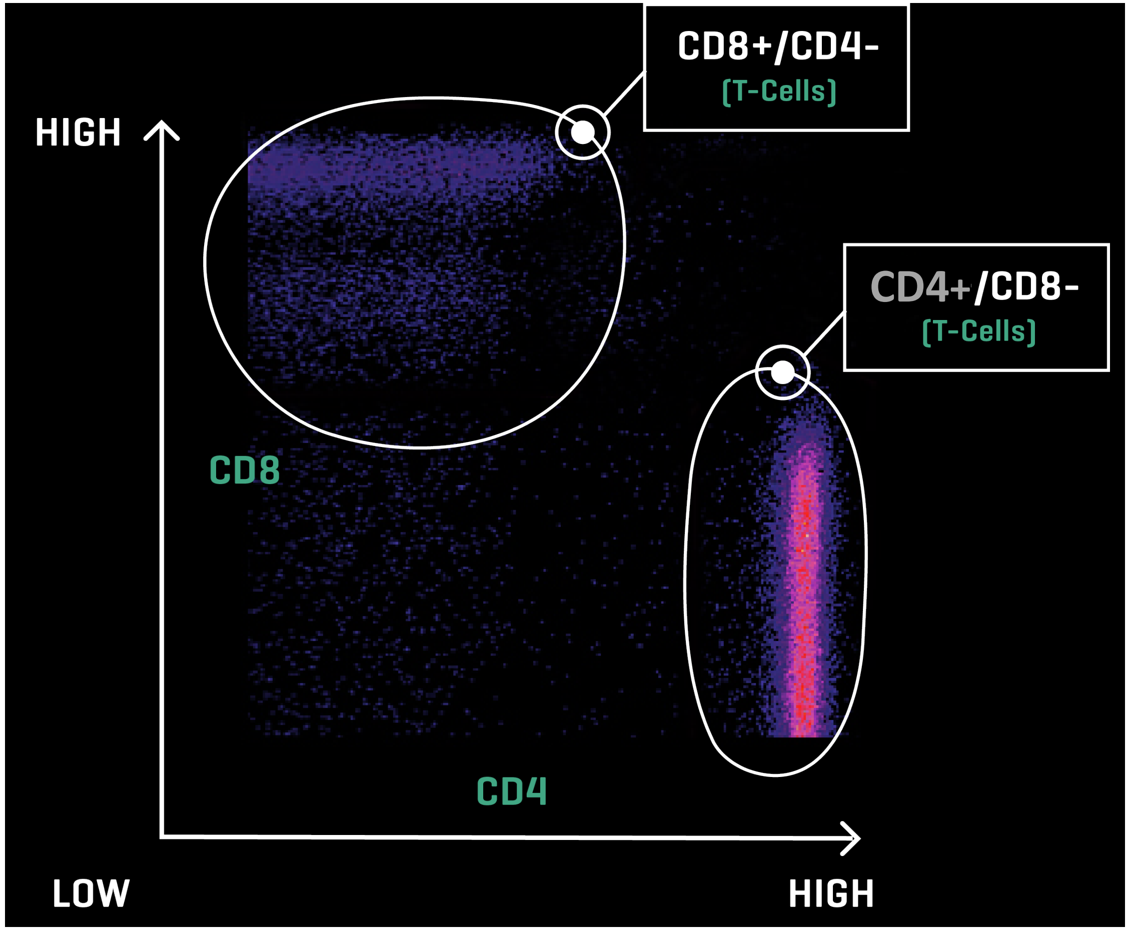


Deep immune profiling via single-cell cytometry

Credit: Fred Hutch

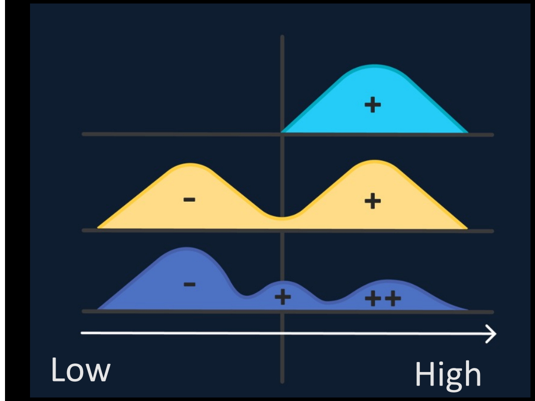
Manual analysis is still the gold standard



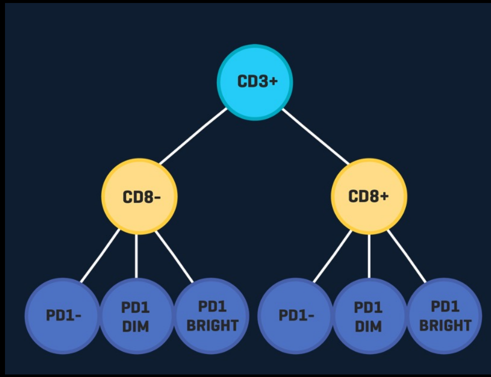


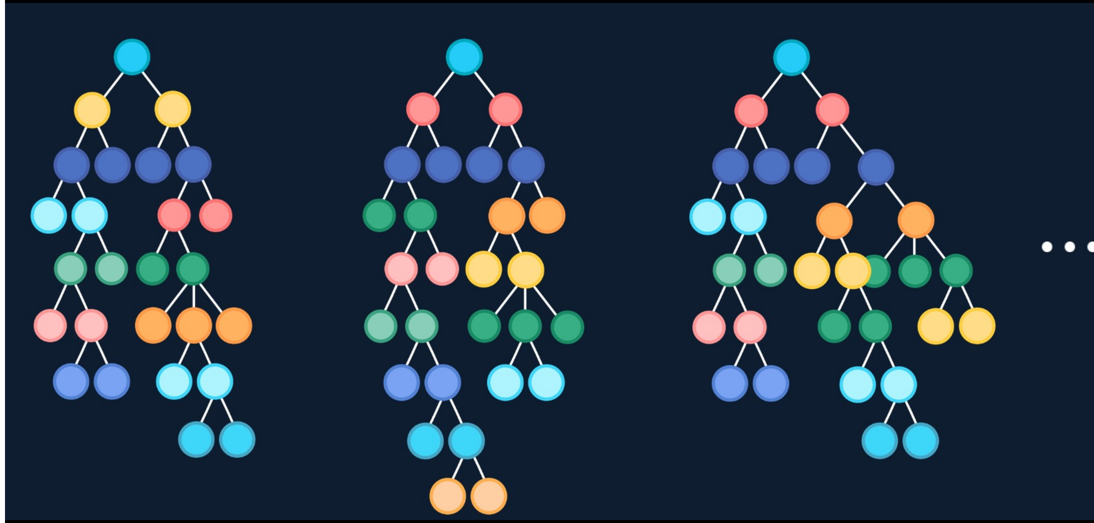
The FAUST algorithm

- Interpretable machine learning approach
- Unambiguously finds all cell populations in a data-driven manner
- Complete phenotypic annotations and cell counts for biomarker screening, e.g. CD3+/CD4-/CD8+/PD1 Dim
- Robust to biological and technological heterogeneity
- Compares favorably to many other competing approaches (e.g. Phenograph, flowSOM)
- (**Greene et al. *Patterns* 2021**)

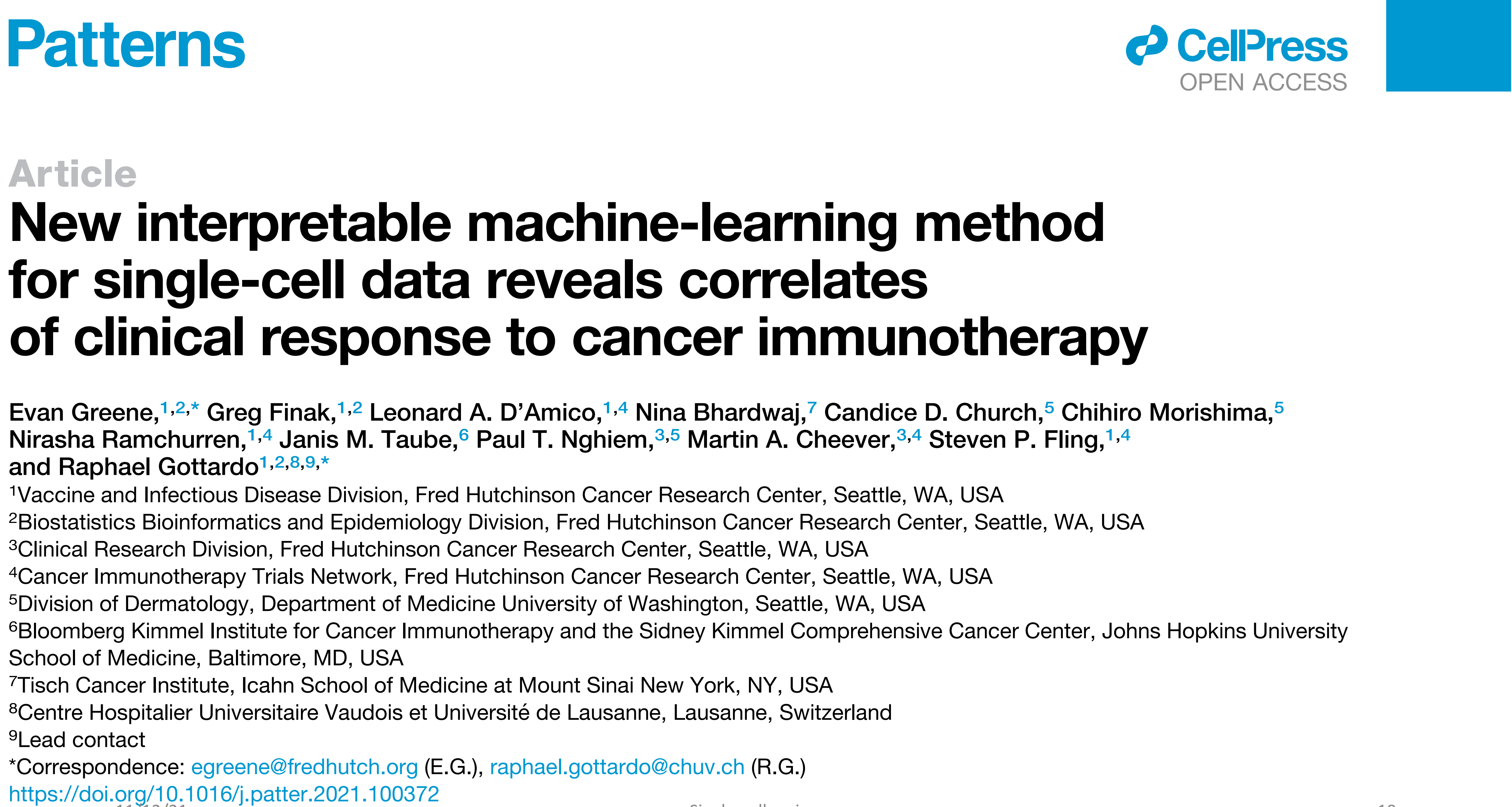


Protein expression





Evan Greene



Patterns

CellPress OPEN ACCESS

Article

New interpretable machine-learning method for single-cell data reveals correlates of clinical response to cancer immunotherapy

Evan Greene,[1,2,*] Greg Finak,[1,2] Leonard A. D'Amico,[1,4] Nina Bhardwaj,[7] Candice D. Church,[5] Chihiro Morishima,[5] Nirasha Ramchurren,[1,4] Janis M. Taube,[6] Paul T. Nghiem,[3,5] Martin A. Cheever,[3,4] Steven P. Fling,[1,4] and Raphael Gottardo[1,2,8,9,*]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[2]Biostatistics Bioinformatics and Epidemiology Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[4]Cancer Immunotherapy Trials Network, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[5]Division of Dermatology, Department of Medicine University of Washington, Seattle, WA, USA
[6]Bloomberg Kimmel Institute for Cancer Immunotherapy and the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[7]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai New York, NY, USA
[8]Centre Hospitalier Universitaire Vaudois et Université de Lausanne, Lausanne, Switzerland
[9]Lead contact
*Correspondence: egreene@fredhutch.org (E.G.), raphael.gottardo@chuv.ch (R.G.)

https://doi.org/10.1016/j.patter.2021.100372

Analysis tools for cytometry analysis and standardization from our lab

LARGE INFRASTRUCTURE TO MANIPULATE AND PREPROCESS FLOW CYTOMETRY DATA IN R AND BIOCONDUCTOR

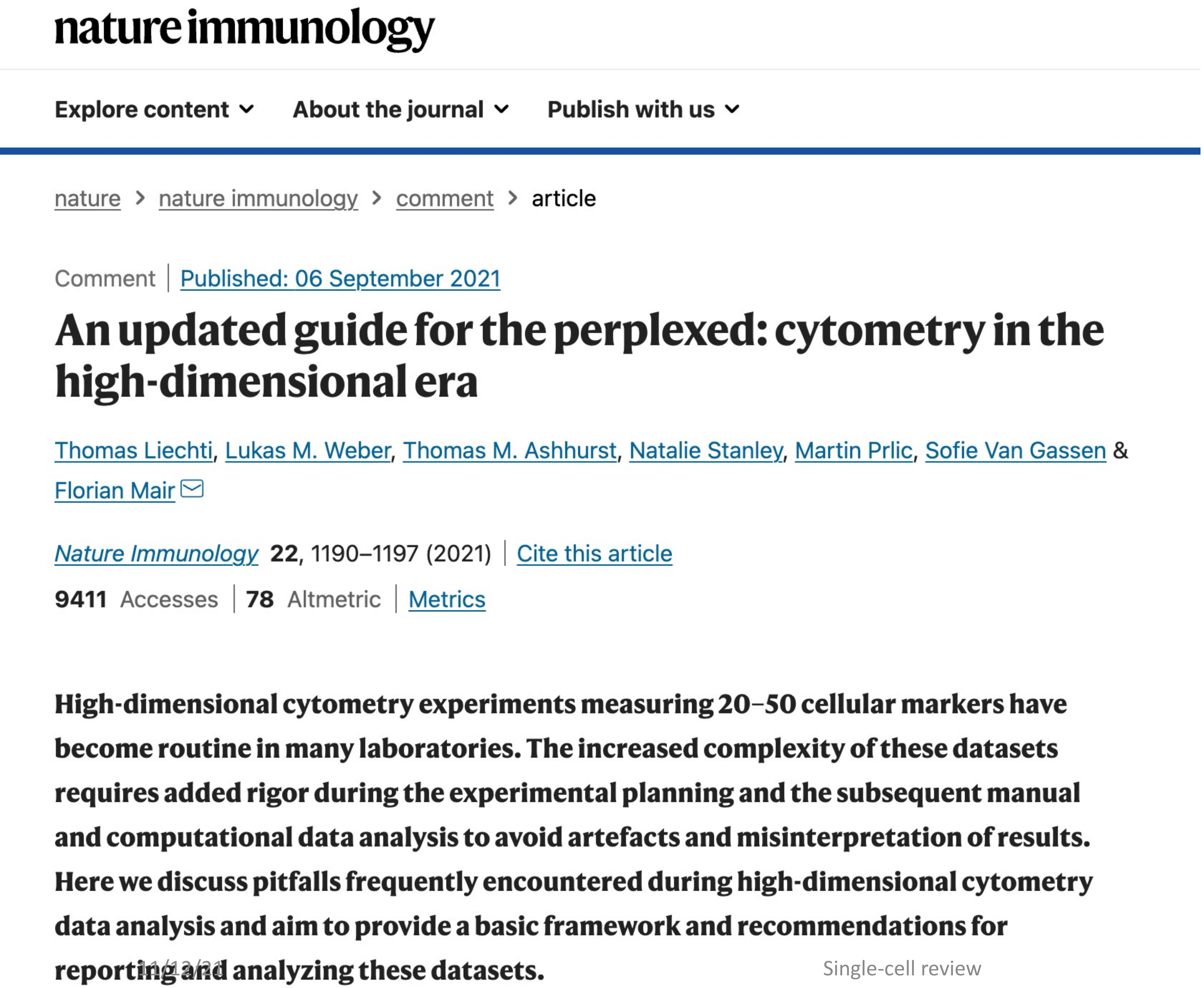
OPENCYTO SUITE (HDF5, FLOWJO INTEROPERABILITY, VISUALIZATION, ETC)

AUTOMATED GATING

STATISTICAL MODELING OF CELL COUNTS

FUNDED BY NIH R01 METHODS GRANT

1. Finak, G., Jiang, W., Gottardo, R., 2018. CytoML for cross-platform cytometry data sharing. *Cytometry A* 93, 1189–1196.
2. Van, P., Jiang, W., Gottardo, R., Finak, G., 2018. ggCyto: Next Generation Open-Source Visualization Software for Cytometry. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty441
3. Finak et al. Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. *Scientific Reports* (2016).
4. Lin et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nat. Biotech.* (2015)
5. Lin et al. Identification and Visualization of Multidimensional Antigen-Specific T-Cell Populations in Polychromatic Cytometry Data. *Cytometry A* (2015).
6. Finak, G. *et al.* OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. *PLoS CB* (2014).
7. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods (*2013*).*



nature immunology

Comment | Published: 06 September 2021

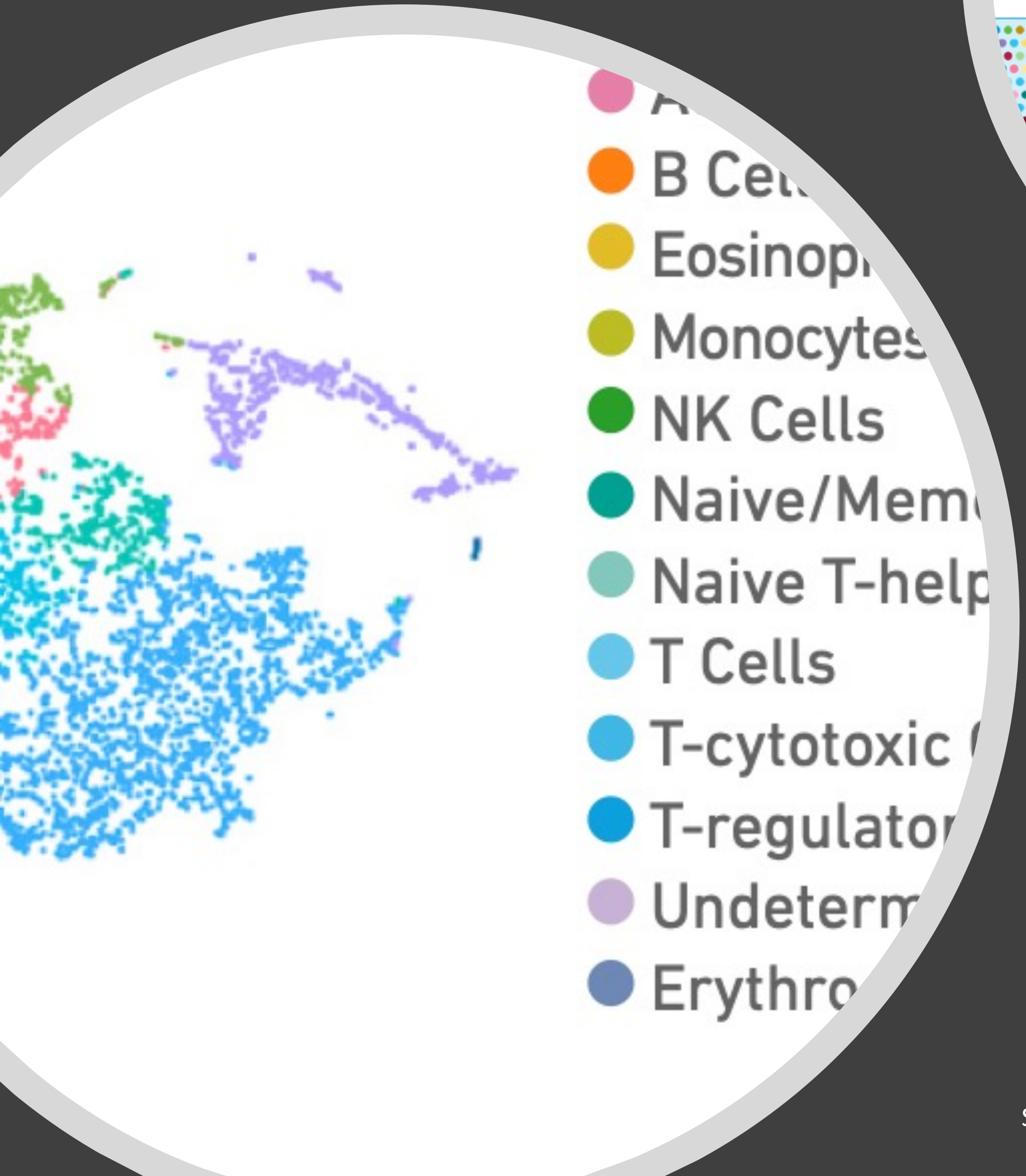
An updated guide for the perplexed: cytometry in the high-dimensional era

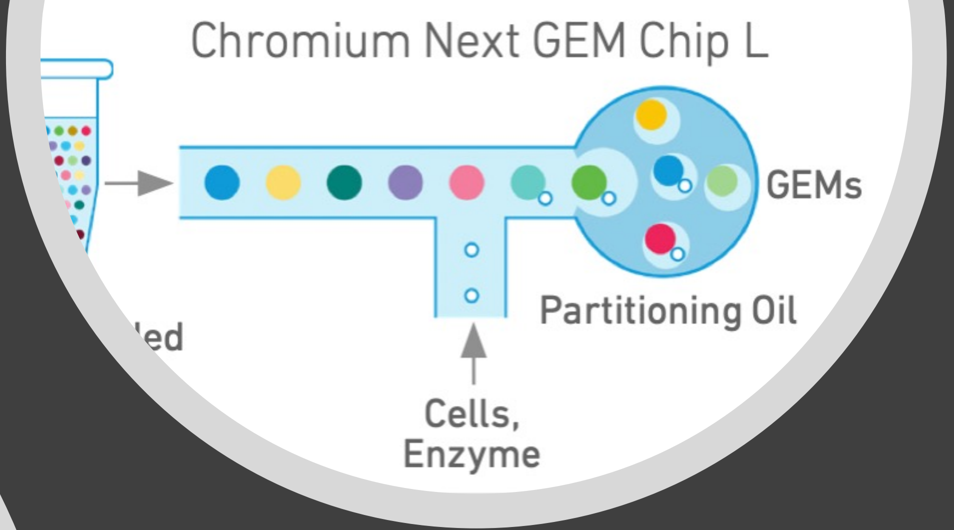
Thomas Liechti, Lukas M. Weber, Thomas M. Ashhurst, Natalie Stanley, Martin Prlic, Sofie Van Gassen & Florian Mair

Nature Immunology **22**, 1190–1197 (2021) | Cite this article

9411 Accesses | **78** Altmetric | Metrics

High-dimensional cytometry experiments measuring 20–50 cellular markers have become routine in many laboratories. The increased complexity of these datasets requires added rigor during the experimental planning and the subsequent manual and computational data analysis to avoid artefacts and misinterpretation of results. Here we discuss pitfalls frequently encountered during high-dimensional cytometry data analysis and aim to provide a basic framework and recommendations for reporting and analyzing these datasets.

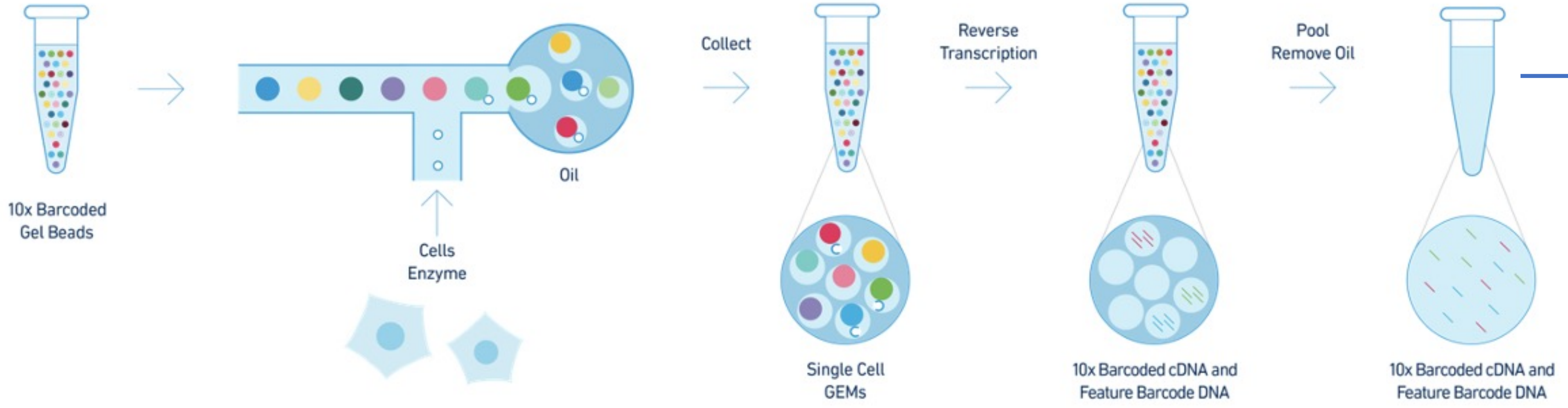




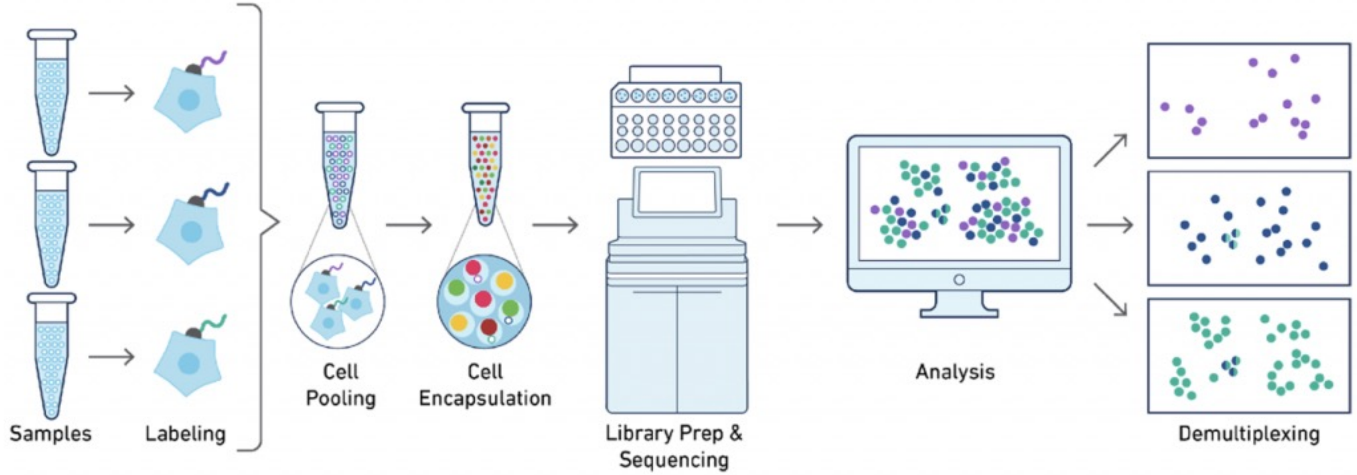
Deep immune profiling via scRNAseq

Credit: 10X Genomics

10x Genomics

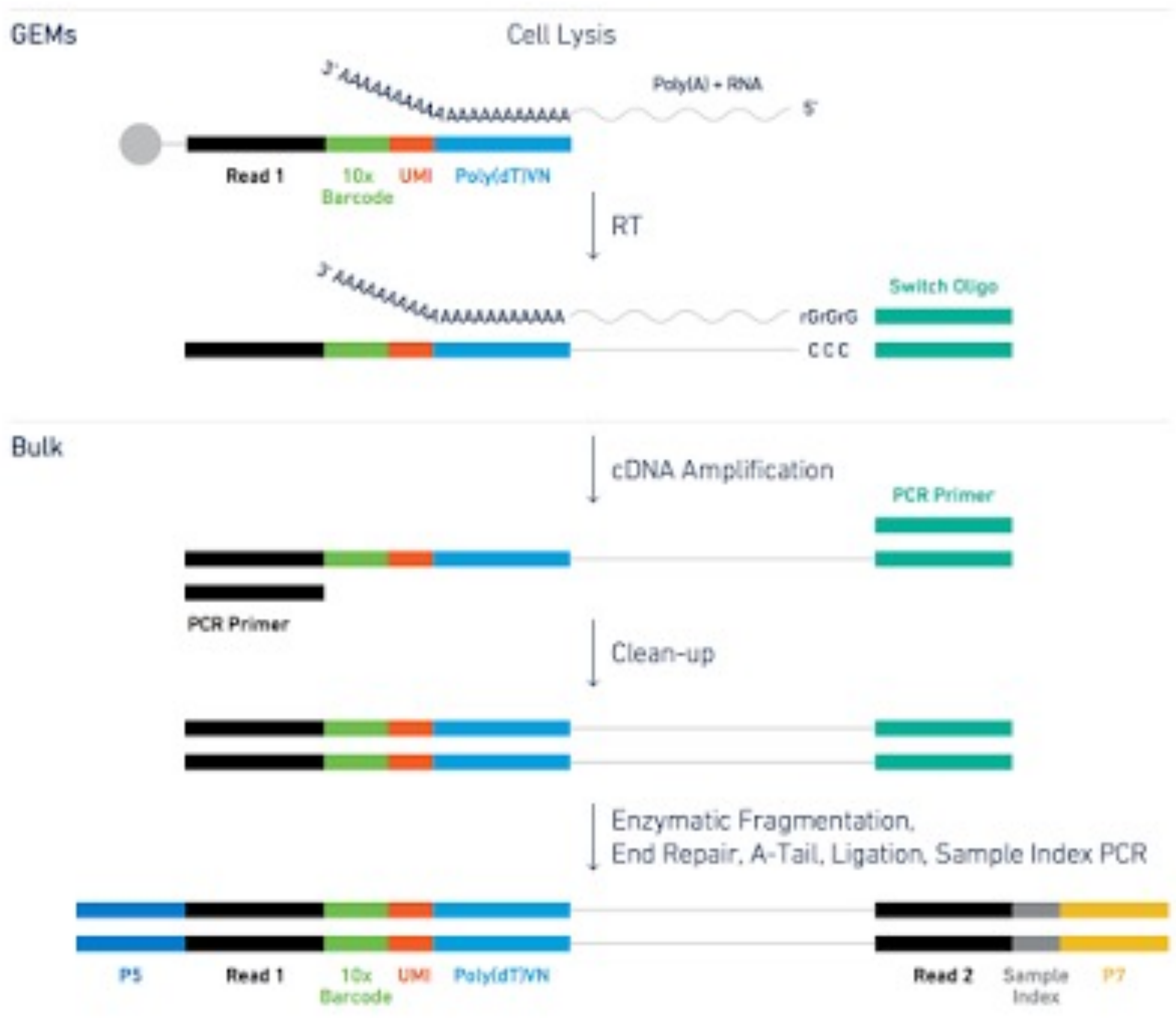


Library prep and sequencing



Credit: https://brcf.medicine.umich.edu/cores/advanced-genomics/technologies/single-cell/10x-genomics/

Single cell 3’ mRNA-seq



- UMI: Unique molecular identifier
  - Enable counting of unique molecules
  - Normalize for PCR biases

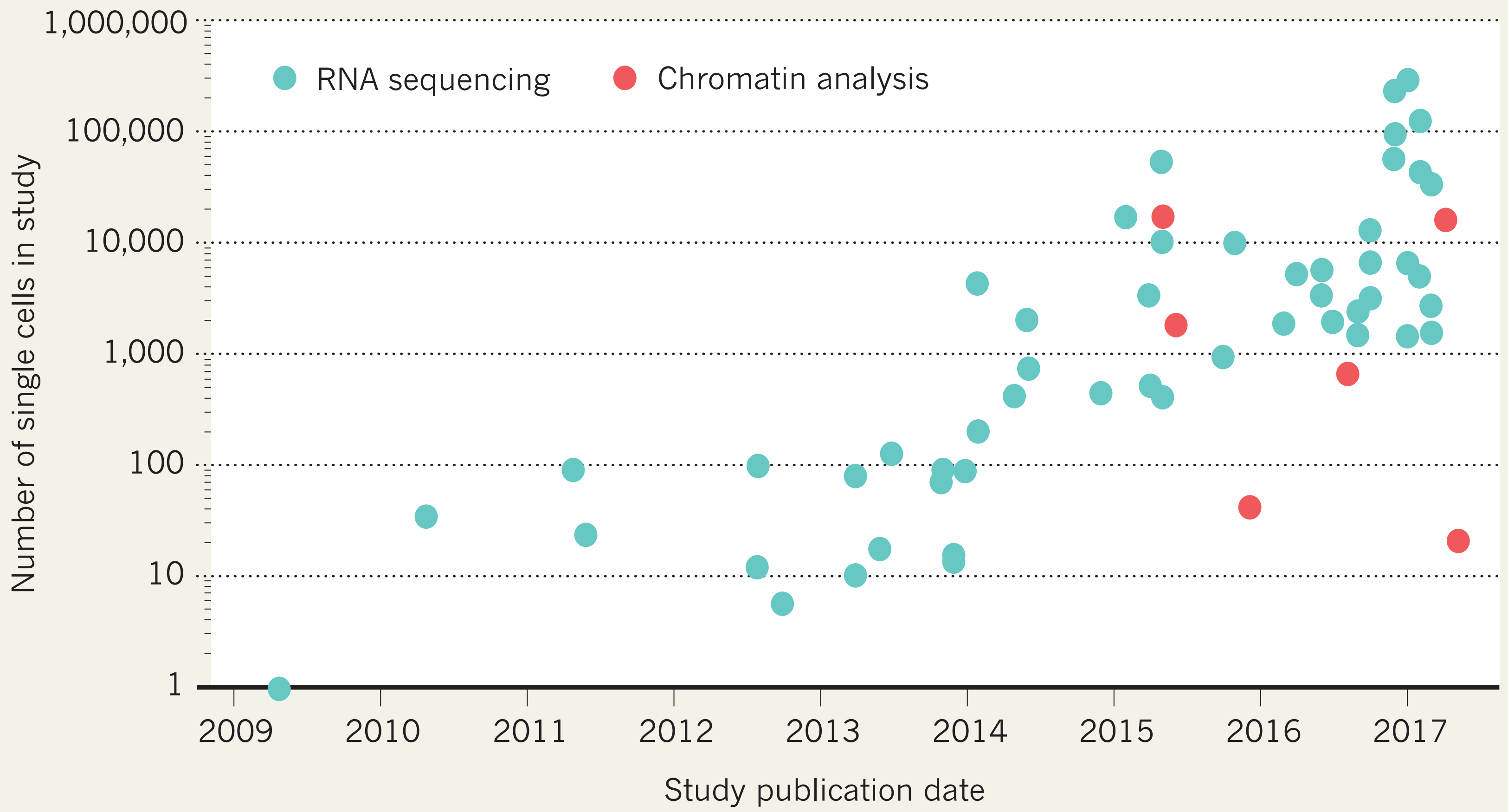
- 3’ sequencing: Can only quantify gene expression, no alternative splicing

- 5’ sequencing can also be used to infer T/B cell receptors

From a single-cell to one million cells!

FROM ONE TO MILLIONS

Biologists can now analyse RNA transcripts or chromatin accessibility in thousands or even millions of individual cells in parallel.



Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., Teichmann, S.A., 2017. The Human Cell Atlas: from vision to reality. Nature 550, 451–453.

Challenges in single-cell genomics

- **Technical issues**
  - Unwanted cell-to-cell variability
  - Assay failure (e.g. due to cell capture, RNA extraction), empty droplets, etc
  - Batch effects (Experimental design)

- **Bi-modality**
  - A gene can be off or on at the single-cell level. Standard statistical models might not be appropriate
  - Data are sparse (lots of zeros)

- **Large datasets**
  - Possibly thousands of genes in thousands of cells with complex designs. Pay attention to computational implementations.

Common computational problems

- Alignment and gene expression quantification
- Normalization and batch effect correction
- Dimension reduction
- Clustering and cell annotation
- Differential gene expression
- Trajectory analysis

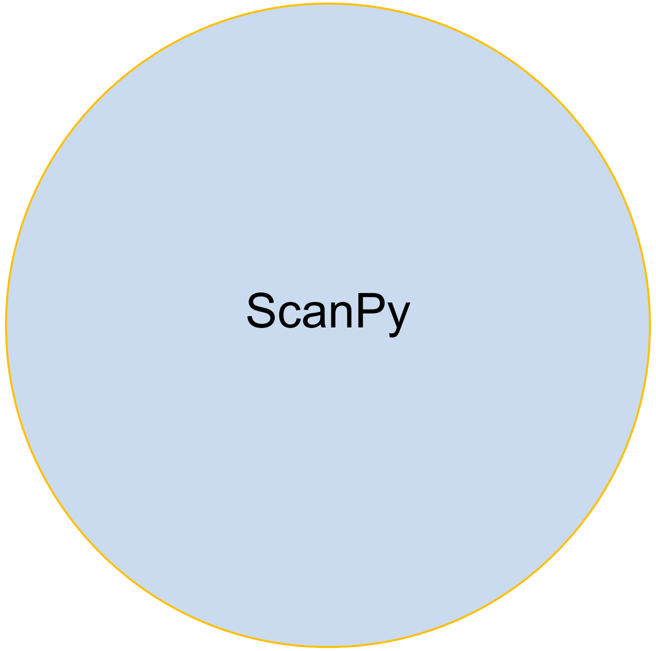
Methods/tools for single-cell genomics from our lab

- Optimization of primers Filtering criteria for cells/probes
- Statistical model for zero-inflated distributions (bimodality)
- Gene-set enrichment analysis
- Support for multiple single-cell platforms (Fluidigm, NanoString, 10X)
- Tools development within Bioconductor

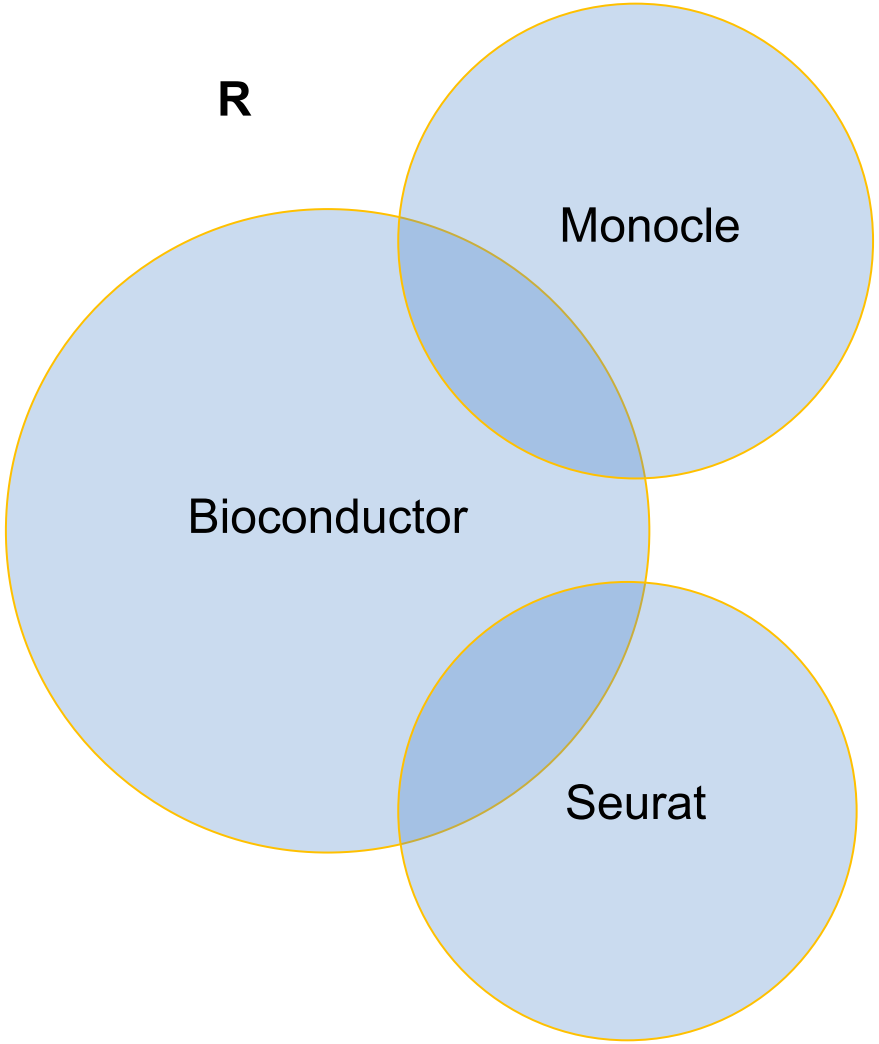
- Hao, Y et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29.
- Amezquita, …, Gottardo, R., Hicks, S.C., 2020. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* 17, 137–145.
- Mair, F., Erickson, J.R., Voillet, V., Simoni, Y., Bi, T., Tyznik, A.J., Martin, J., Gottardo, R., Newell, E.W., Prlic, M., 2020. A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. *Cell Re*p. 31, 107499.
- McDavid, Finak, and Gottardo The Contribution of Cell Cycle to Heterogeneity in Single-Cell RNA-Seq Data. *Nature Biotechnology* (2016).
- Finak, G. *et al.* MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA-Seq Data. *Genome Biology (2015)*.
- McDavid, A. *et al.* Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS CB* (2015).
- McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 29, 461–467 (2013).

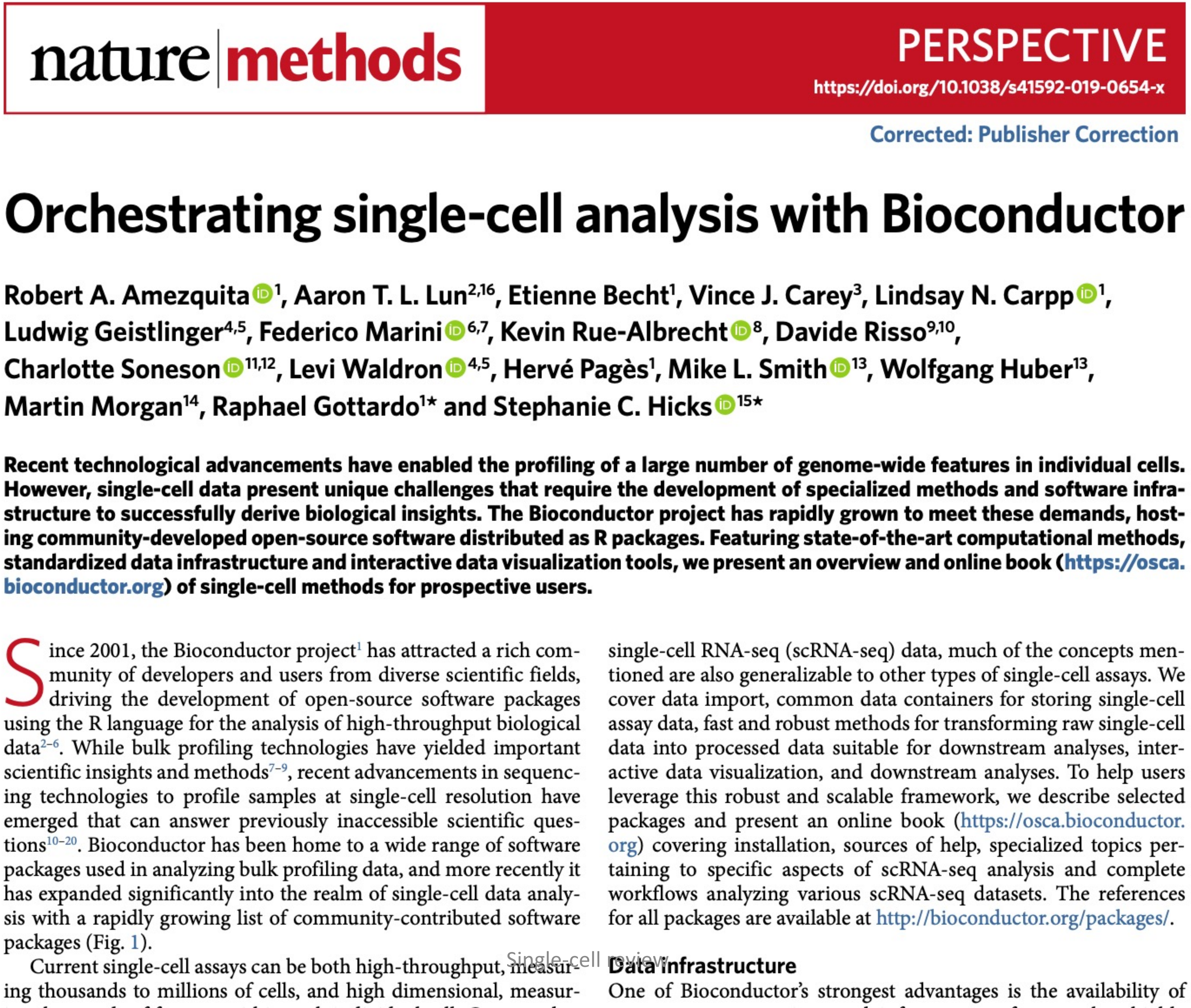
Environments for data analysis

Python



R





nature methods

PERSPECTIVE

https://doi.org/10.1038/s41592-019-0654-x

Corrected: Publisher Correction

Orchestrating single-cell analysis with Bioconductor

Robert A. Amezquita[1], Aaron T. L. Lun[2,16], Etienne Becht[1], Vince J. Carey[3], Lindsay N. Carpp[1], Ludwig Geistlinger[4,5], Federico Marini[6,7], Kevin Rue-Albrecht[8], Davide Risso[9,10], Charlotte Soneson[11,12], Levi Waldron[4,5], Hervé Pagès[1], Mike L. Smith[13], Wolfgang Huber[13], Martin Morgan[14], Raphael Gottardo[1]* and Stephanie C. Hicks[15]*

Recent technological advancements have enabled the profiling of a large number of genome-wide features in individual cells. However, single-cell data present unique challenges that require the development of specialized methods and software infrastructure to successfully derive biological insights. The Bioconductor project has rapidly grown to meet these demands, hosting community-developed open-source software distributed as R packages. Featuring state-of-the-art computational methods, standardized data infrastructure and interactive data visualization tools, we present an overview and online book (https://osca.bioconductor.org) of single-cell methods for prospective users.

Since 2001, the Bioconductor project[1] has attracted a rich community of developers and users from diverse scientific fields, driving the development of open-source software packages using the R language for the analysis of high-throughput biological data[2–6]. While bulk profiling technologies have yielded important scientific insights and methods[7–9], recent advancements in sequencing technologies to profile samples at single-cell resolution have emerged that can answer previously inaccessible scientific questions[10–20]. Bioconductor has been home to a wide range of software packages used in analyzing bulk profiling data, and more recently it has expanded significantly into the realm of single-cell data analysis with a rapidly growing list of community-contributed software packages (Fig. 1).

Current single-cell assays can be both high-throughput, measuring thousands to millions of cells, and high dimensional, measur-

single-cell RNA-seq (scRNA-seq) data, much of the concepts mentioned are also generalizable to other types of single-cell assays. We cover data import, common data containers for storing single-cell assay data, fast and robust methods for transforming raw single-cell data into processed data suitable for downstream analyses, interactive data visualization, and downstream analyses. To help users leverage this robust and scalable framework, we describe selected packages and present an online book (https://osca.bioconductor.org) covering installation, sources of help, specialized topics pertaining to specific aspects of scRNA-seq analysis and complete workflows analyzing various scRNA-seq datasets. The references for all packages are available at http://bioconductor.org/packages/.

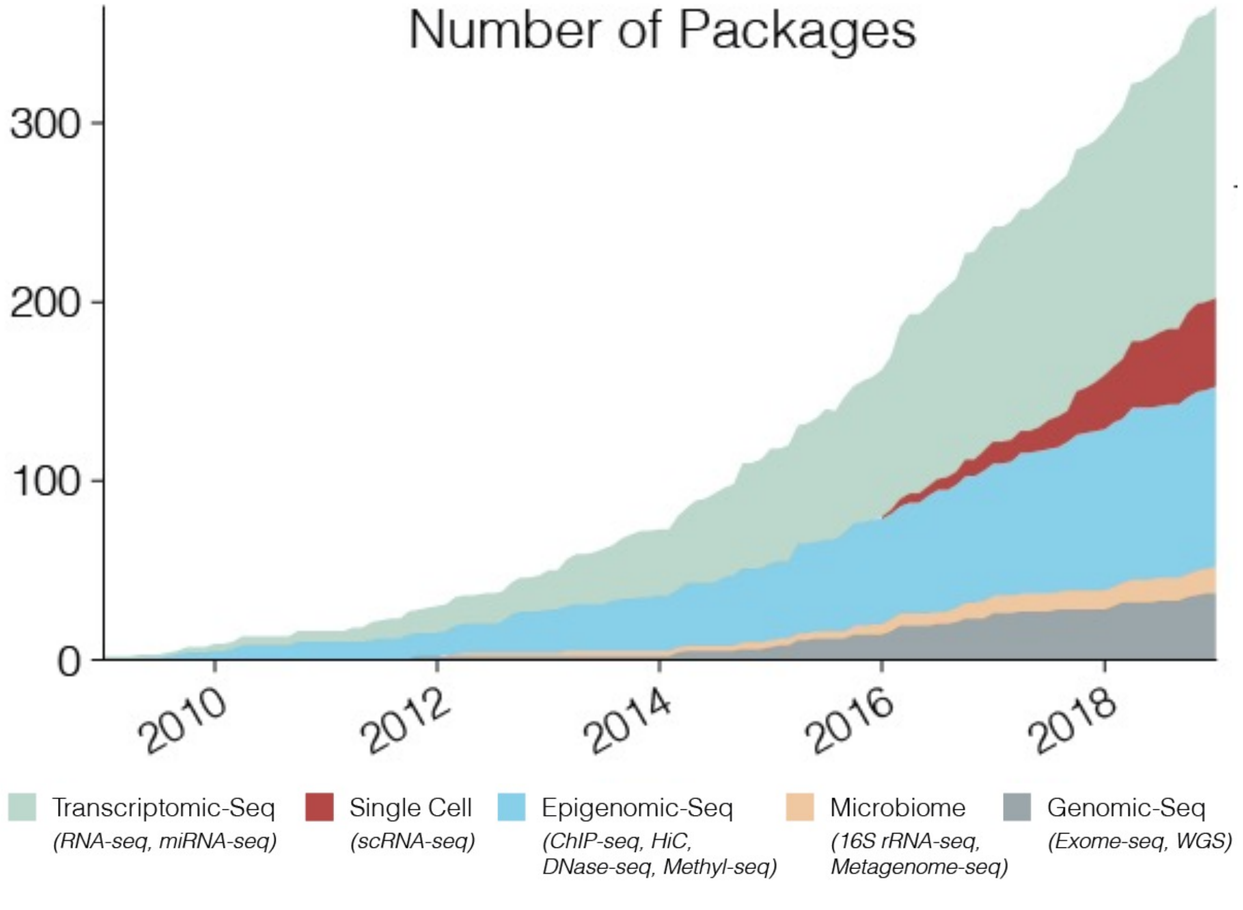
Data infrastructure

One of Bioconductor's strongest advantages is the availability of

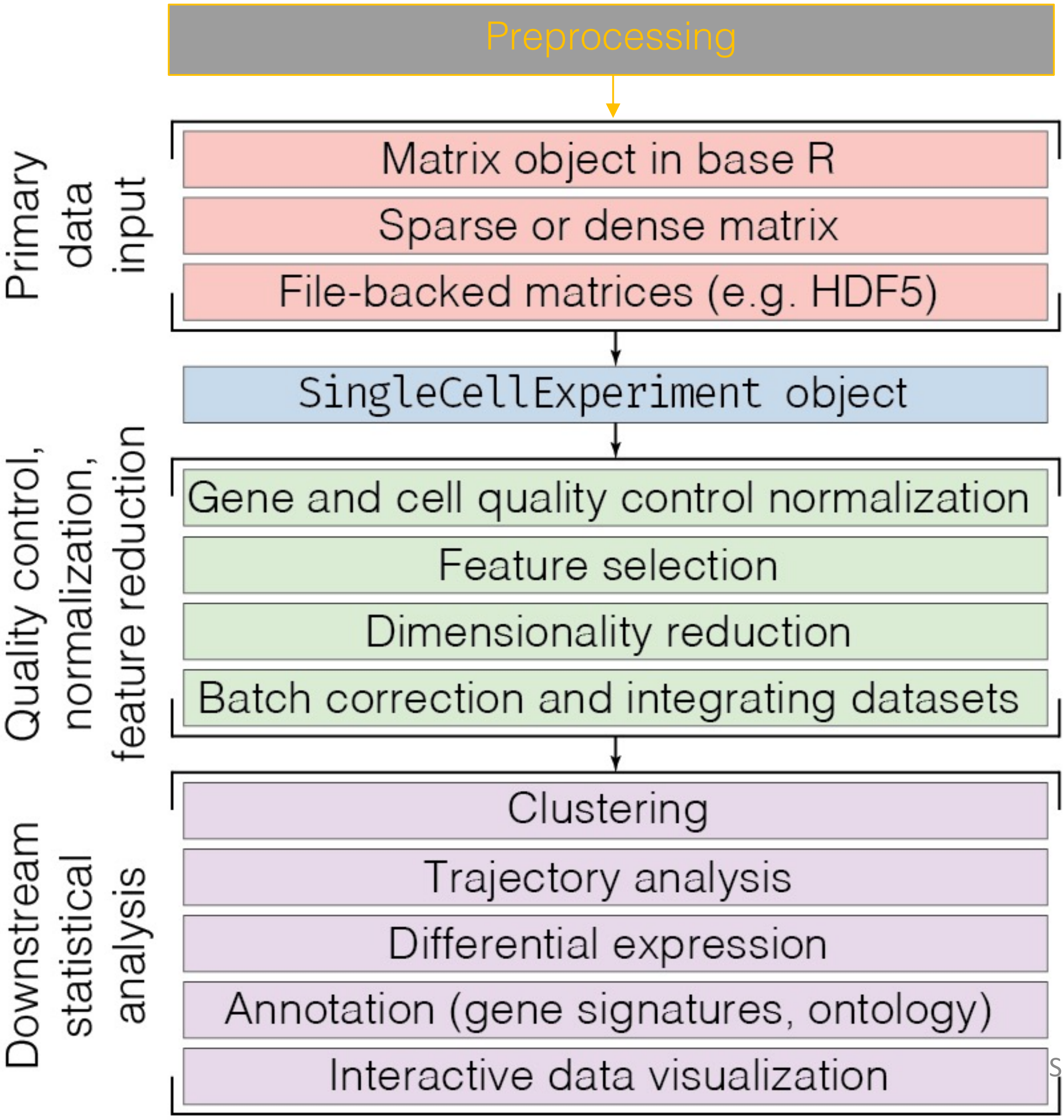
Bioconductor – Open Source Software for Bioinformatics in R

- Built in and for R

- Focused on the analysis of genomic data, with a rich history of software and methods development that has spanned the era of sequencing



Overall Framework for Analysis of scRNA-seq



Preprocessing

Primary data input

Matrix object in base R

Sparse or dense matrix

File-backed matrices (e.g. HDF5)

SingleCellExperiment object

Quality control, normalization, feature reduction

Gene and cell quality control normalization

Feature selection

Dimensionality reduction

Batch correction and integrating datasets

Downstream statistical analysis

Clustering

Trajectory analysis

Differential expression

Annotation (gene signatures, ontology)

Interactive data visualization

Typically done outside of R, e.g. *CellRanger, kallisto*

We can read 10X data into a *SingleCellExperiment* using the *DropletUtils* package

Common data model used by most packages

e.g. scater, zinbwave

e.g. scran, scFeatureFilter

e.g. scater, destiny

e.g. scMerge, batchelor

e.g. BiocNeighbors, SC3

e.g. slingshot, TSCAN

e.g. MAST, SCDE, muscat

e.g. MAST, singleR

e.g. iSEE

Gene expression quantification

Alignment and quantification – Overview

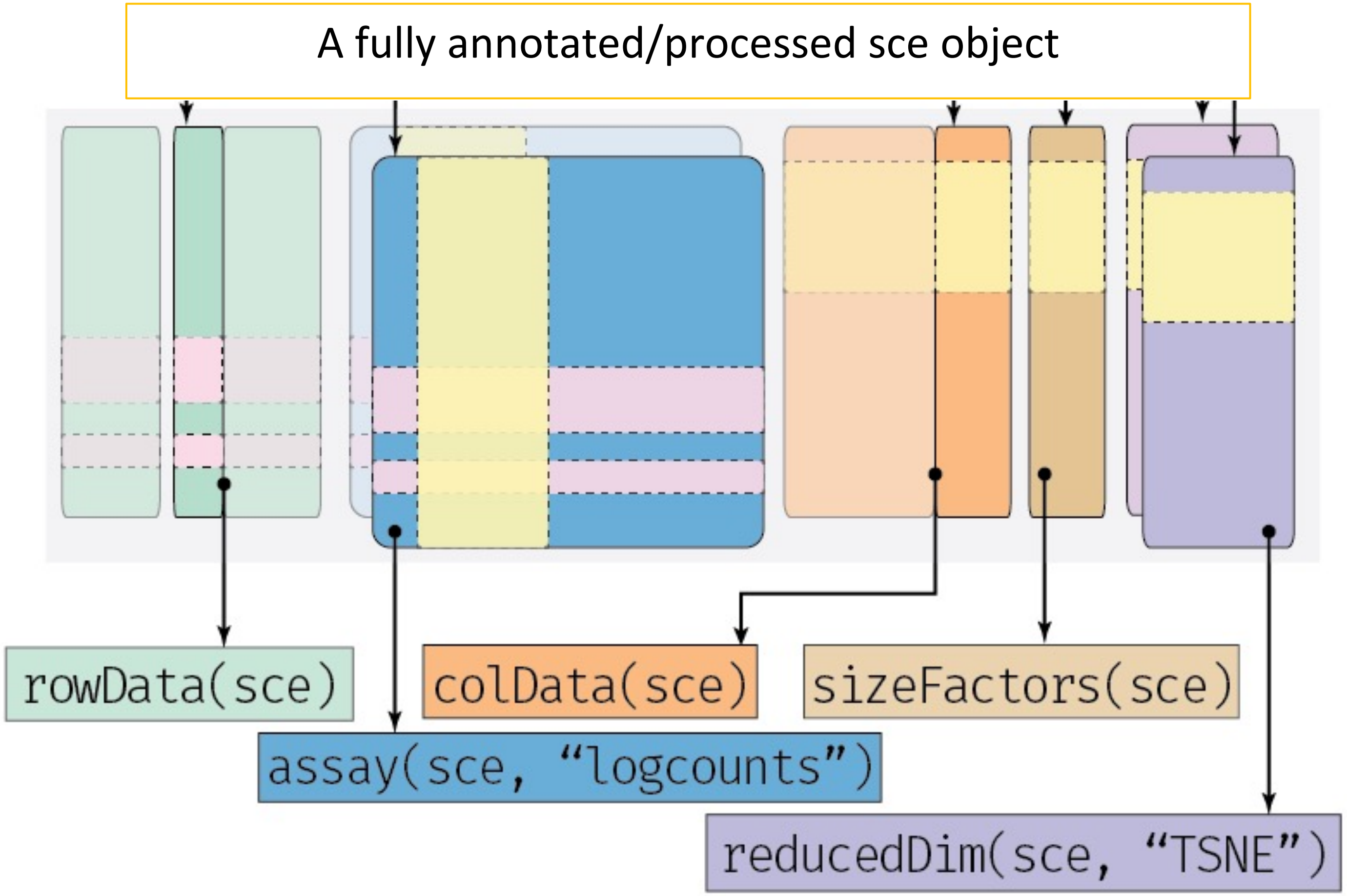
- First step of any analysis is to align reads to a reference genome to quantify transcript/gene expression for each gene within each cell
- The reference genome will typically be specific for your study (e.g. human/mouse, etc) and might include additional sequence/annotations for quantifying non-host expression (e.g. viral genes, CAR T gene, spike-in controls)
- Large number of available tools for sequence alignment and quantification (e.g. CellRanger, kalisto (Bray et al. *Nat. Biotech.* 2016))

Introducing the SingleCellExperiment class

An S4 class object used to store the primary data, data transformations, and metadata associated with a single-cell experiment or collection thereof

Arranges the data into specialized sub-containers called *slots*, which can be accessed via standardized *accessor* functions that allow for *creation of, modification, and supplementation* of (meta)data

```
library(SingleCellExperiment)
sce <- SingleCellExperiment(…)
```



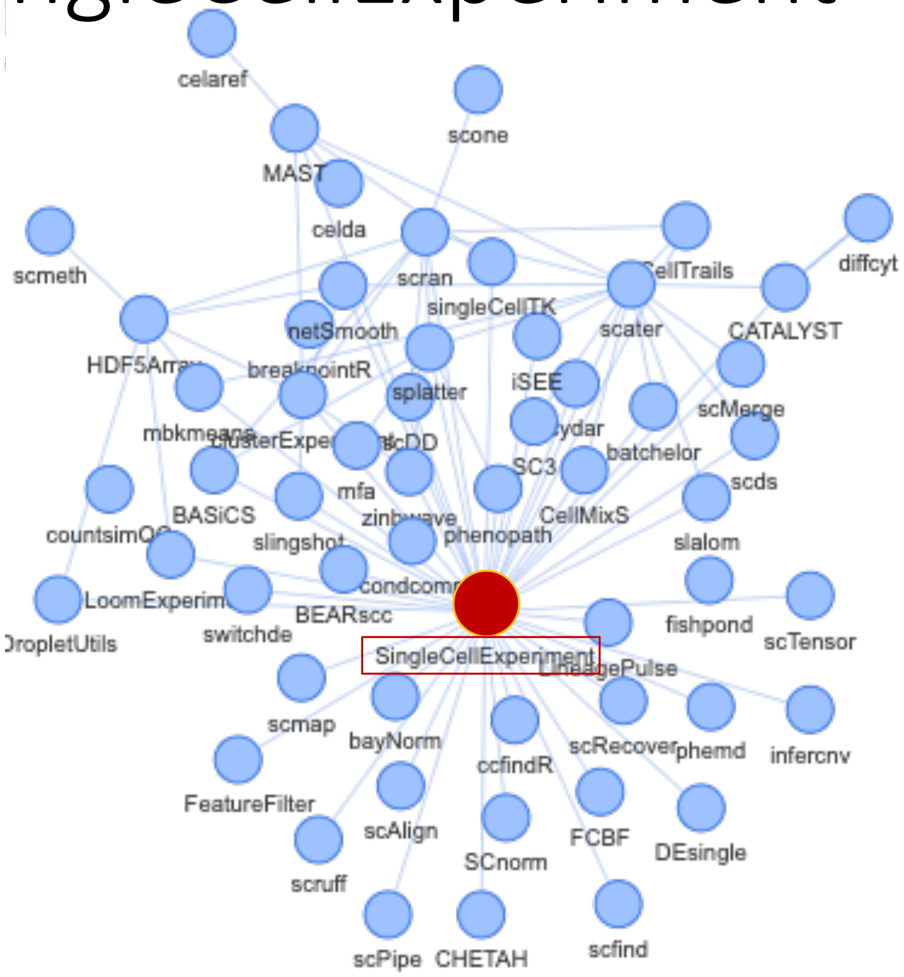
Infrastructure Matters: Why SingleCellExperiment is Awesome

Ability to analyze across a universe of Bioconductor packages

The consistent, extensible interface makes it possible for plethora of packages with various objectives to work with one another in a workflow, from reading in raw data to normalization to visualization

This is why a singular workflow is possible that for each step can leverage various Bioconductor tools (where each tool has been individually tested, peer-reviewed, and published)

Conversion to other types of containers is possible as well to Monocle's *CellDataSet* and Seurat's *SeuratObject*



Packages which have a dependency on the *SingleCellExperiment* class

https://seandavi.github.io/2019/05/single-cell-packages-and-dependencies-in-bioconductor-using-biocpkgtools/

Dimensionality reduction

Dimensionality reduction – Overview

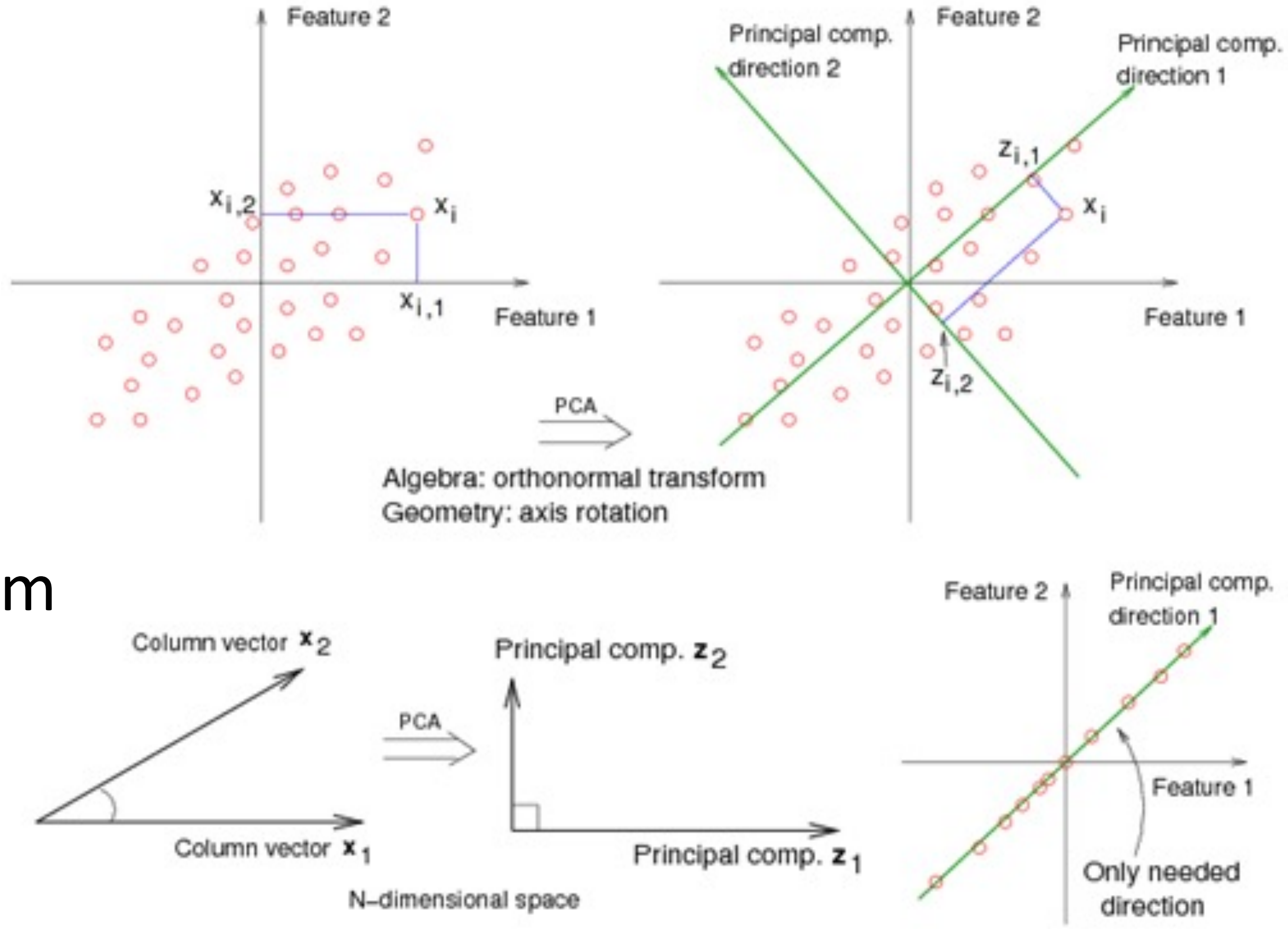
- Basic idea: Reduce the number of variables while retaining most of the (biological) information in the original dataset

- **Feature down-selection:** Basic filtering of non-informative genes, e.g. zero expression values or low variability
- **Data transformation and compression:** Define a minimal set of new variables that maximize information content. Transformation can be linear and non-linear (e.g. PCA, tSNE)

Principal component analysis

- Capture the intrinsic variability in the data.
- Reduce the dimensionality of a data set, either to ease interpretation or as a way to avoid overfitting and to prepare for subsequent analysis.
- Linear, so principal components (PC) can be more interpretable (compared to non-linear approaches t-SNE)
- Not efficient for capturing non-linear structures
- More than 2 PCs might be needed for efficient compression

Principal component analysis – Toy example


Feature 2
Feature 1
$x_{i,2}$
x_i
$x_{i,1}$
PCA
Algebra: orthonormal transform
Geometry: axis rotation
Principal comp. direction 2
Feature 2
Principal comp. direction 1
$z_{i,1}$
x_i
Feature 1
$z_{i,2}$
Column vector $\mathbf{x}_2$
Column vector $\mathbf{x}_1$
PCA
Principal comp. $\mathbf{z}_2$
Principal comp. $\mathbf{z}_1$
N–dimensional space
Feature 2
Principal comp. direction 1
Feature 1
Only needed direction

Linear transformation

Simple rotation of the coordinate system

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

t-Distributed Stochastic Neighboring Embedding

- Non-linear dimensionality reduction
- Originally developed by van der Maaten and Hinton
- Became popular for single-cell analysis when applied to CyTOF data (viSNE, Amir et al. 2013)
- Efficient data compression even with 2 components
- Sensitive to initial values and tuning parameters. Different runs can output very different plots
- t SNE dimensions and distances are not interpretable
- Generally speaking, the resulting structure is fairly robust
- https://distill.pub/2016/misread-tsne/

tSNE intuition

X observed (high-dimensional) data
Y lower dim representation

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)} \xrightarrow{\text{Approximate}} q_{ij} = \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l\neq k}(1+\|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

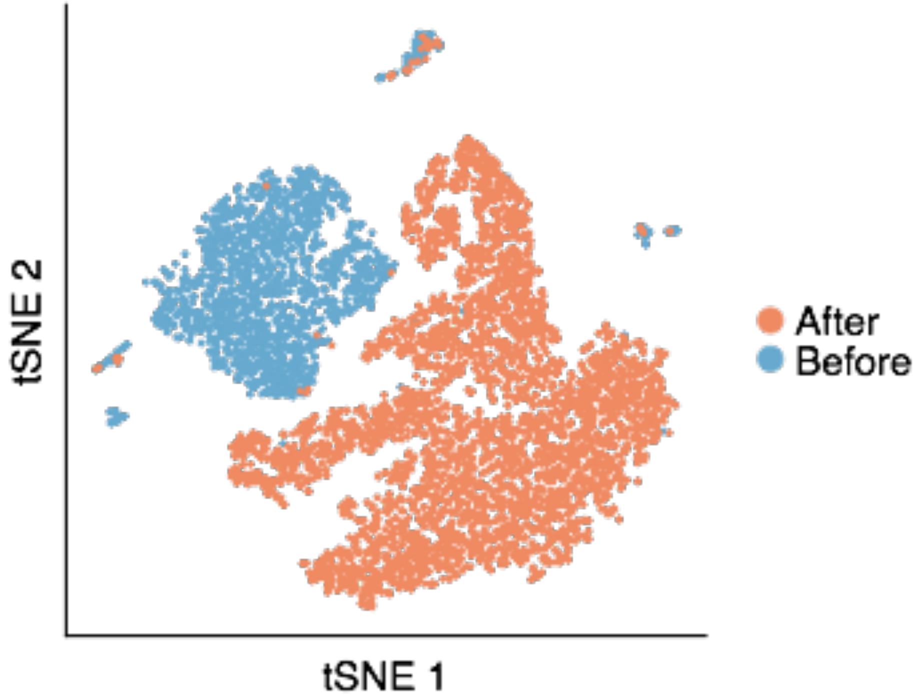
By minimizing the KL divergence

$$\mathrm{KL}\left(P \parallel Q\right) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

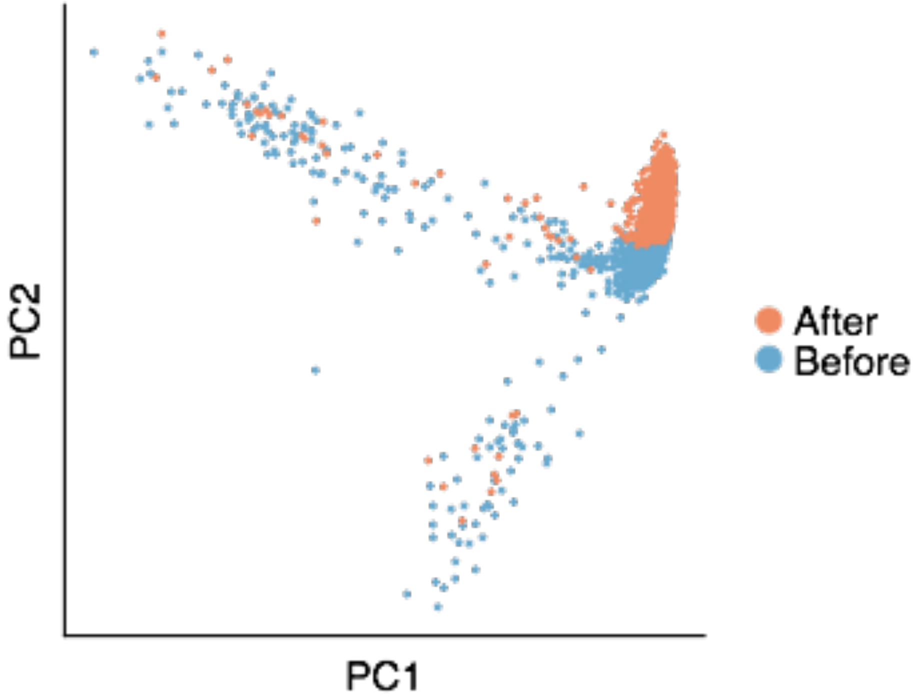
tSNE vs. PCA – Example

MCC project

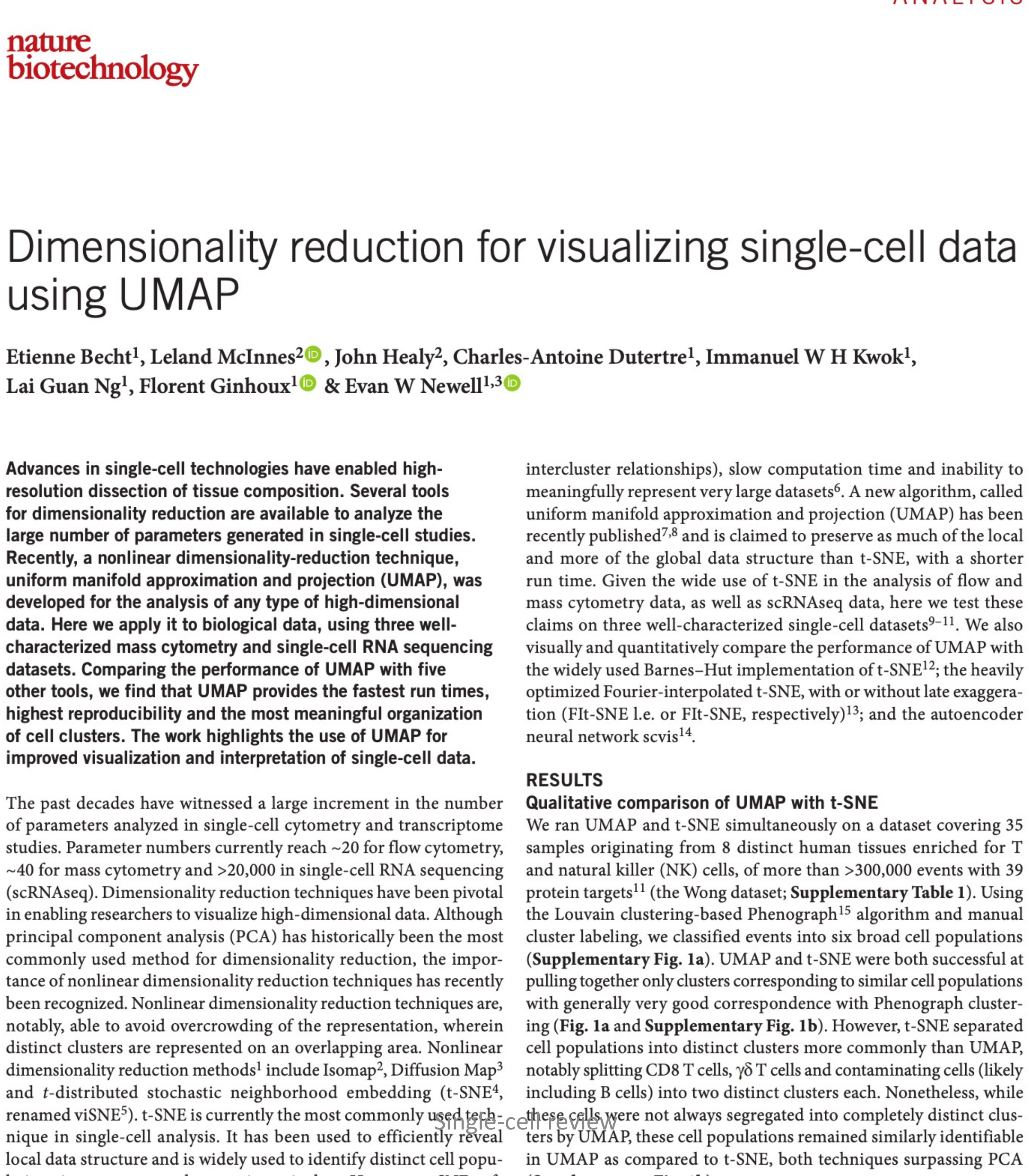
Data normalized and unwanted sources of variation (nUMI) are removed - Seurat Analysis



tSNE



PCA



nature biotechnology

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht[1], Leland McInnes[2], John Healy[2], Charles-Antoine Dutertre[1], Immanuel W H Kwok[1], Lai Guan Ng[1], Florent Ginhoux[1] & Evan W Newell[1,3]

Advances in single-cell technologies have enabled high-resolution dissection of tissue composition. Several tools for dimensionality reduction are available to analyze the large number of parameters generated in single-cell studies. Recently, a nonlinear dimensionality-reduction technique, uniform manifold approximation and projection (UMAP), was developed for the analysis of any type of high-dimensional data. Here we apply it to biological data, using three well-characterized mass cytometry and single-cell RNA sequencing datasets. Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters. The work highlights the use of UMAP for improved visualization and interpretation of single-cell data.

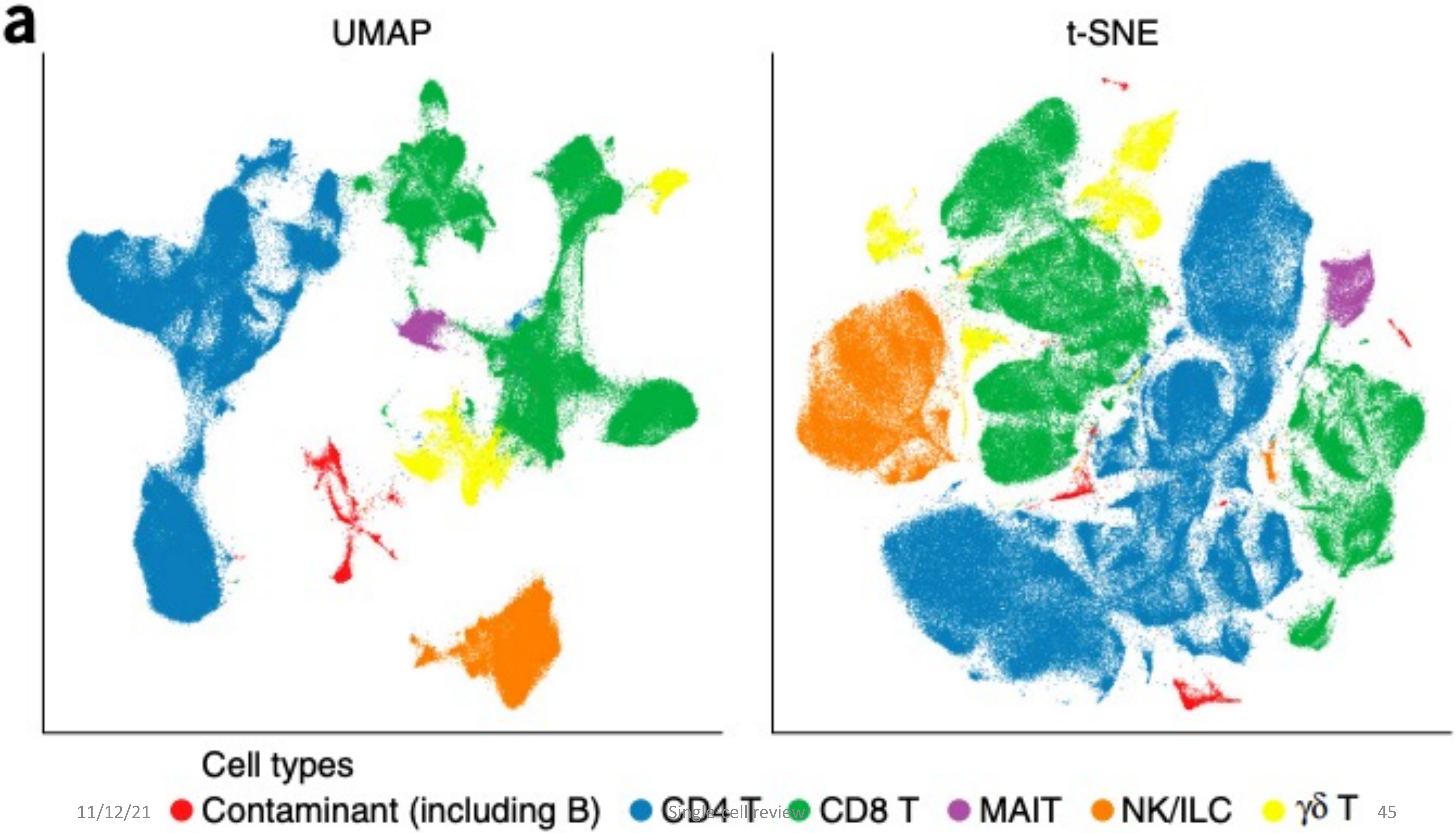
The past decades have witnessed a large increment in the number of parameters analyzed in single-cell cytometry and transcriptome studies. Parameter numbers currently reach ~20 for flow cytometry, ~40 for mass cytometry and >20,000 in single-cell RNA sequencing (scRNAseq). Dimensionality reduction techniques have been pivotal in enabling researchers to visualize high-dimensional data. Although principal component analysis (PCA) has historically been the most commonly used method for dimensionality reduction, the importance of nonlinear dimensionality reduction techniques has recently been recognized. Nonlinear dimensionality reduction techniques are, notably, able to avoid overcrowding of the representation, wherein distinct clusters are represented on an overlapping area. Nonlinear dimensionality reduction methods[1] include Isomap[2], Diffusion Map[3] and *t*-distributed stochastic neighborhood embedding (t-SNE[4], renamed viSNE[5]). t-SNE is currently the most commonly used technique in single-cell analysis. It has been used to efficiently reveal local data structure and is widely used to identify distinct cell popu-

intercluster relationships), slow computation time and inability to meaningfully represent very large datasets[6]. A new algorithm, called uniform manifold approximation and projection (UMAP) has been recently published[7,8] and is claimed to preserve as much of the local and more of the global data structure than t-SNE, with a shorter run time. Given the wide use of t-SNE in the analysis of flow and mass cytometry data, as well as scRNAseq data, here we test these claims on three well-characterized single-cell datasets[9–11]. We also visually and quantitatively compare the performance of UMAP with the widely used Barnes–Hut implementation of t-SNE[12]; the heavily optimized Fourier-interpolated t-SNE, with or without late exaggeration (FIt-SNE l.e. or FIt-SNE, respectively)[13]; and the autoencoder neural network scvis[14].

RESULTS

Qualitative comparison of UMAP with t-SNE

We ran UMAP and t-SNE simultaneously on a dataset covering 35 samples originating from 8 distinct human tissues enriched for T and natural killer (NK) cells, of more than >300,000 events with 39 protein targets[11] (the Wong dataset; **Supplementary Table 1**). Using the Louvain clustering-based Phenograph[15] algorithm and manual cluster labeling, we classified events into six broad cell populations (**Supplementary Fig. 1a**). UMAP and t-SNE were both successful at pulling together only clusters corresponding to similar cell populations with generally very good correspondence with Phenograph clustering (**Fig. 1a** and **Supplementary Fig. 1b**). However, t-SNE separated cell populations into distinct clusters more commonly than UMAP, notably splitting CD8 T cells, γδ T cells and contaminating cells (likely including B cells) into two distinct clusters each. Nonetheless, while these cells were not always segregated into completely distinct clusters by UMAP, these cell populations remained similarly identifiable in UMAP as compared to t-SNE, both techniques surpassing PCA


a
UMAP
t-SNE
Cell types
● Contaminant (including B)
● CD4 T
● CD8 T
● MAIT
● NK/ILC
● γδ T

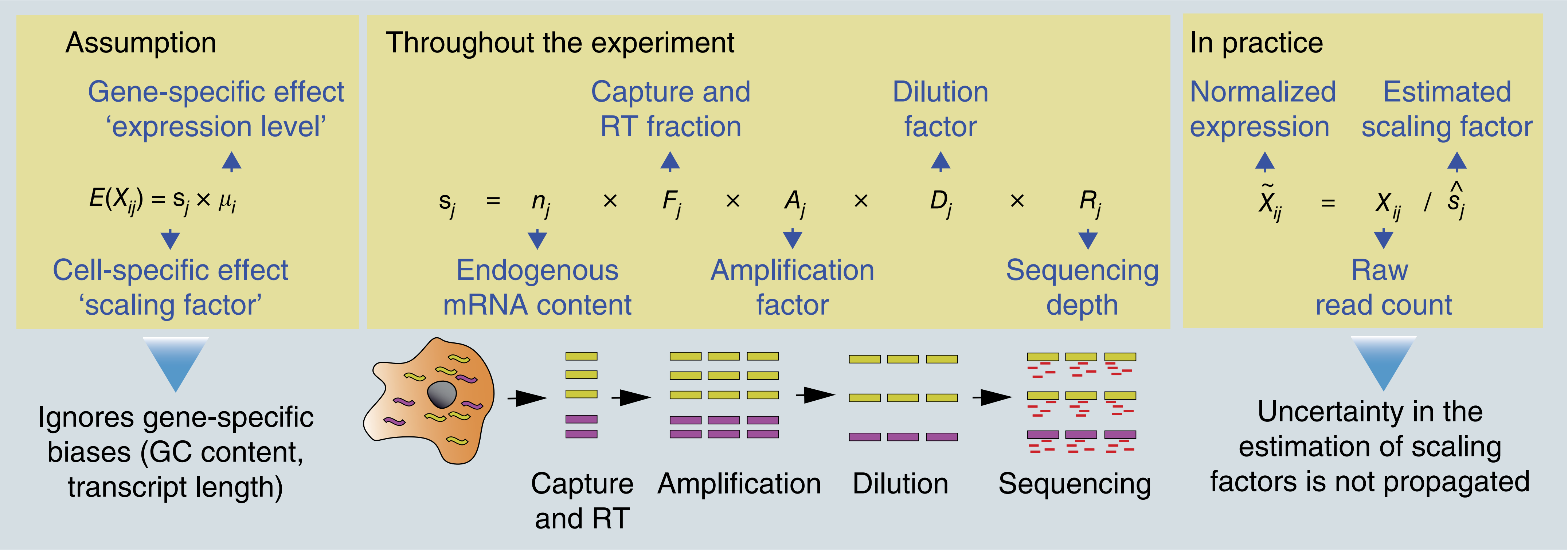
Normalization

Overall goals of normalization

- Normalization is needed to ensure that observed differences in counts (across samples, biological conditions) are indeed biological and not due to some technical artifact (e.g. array batch, technician, sequencing depth, etc)
- Normalization strategies must capture biases that are specific to the technology of interest
- Many techniques exist for bulk RNA-seq
- <u>Problem here:</u> What is a technical replicate? How can we differentiate technical vs biological variation?

Global-scaling normalization

- From molecules to reads



Assumption

Gene-specific effect ‘expression level’

$E(X_{ij}) = s_j \times \mu_i$

Cell-specific effect ‘scaling factor’

Ignores gene-specific biases (GC content, transcript length)

Throughout the experiment

Capture and RT fraction

Dilution factor

$s_j = n_j \times F_j \times A_j \times D_j \times R_j$

Endogenous mRNA content

Amplification factor

Sequencing depth

Capture and RT

Amplification

Dilution

Sequencing

In practice

Normalized expression

Estimated scaling factor

$\tilde{X}_{ij} = X_{ij} / \hat{s}_j$

Raw read count

Uncertainty in the estimation of scaling factors is not propagated

Global-scaling normalization

- Developed for bulk RNA-seq experiments (mixture of 1,000-1,000,000 cells) where gene expression is more homogeneous
- Reads per million (RPM), (or related RPKM and TPM) Standardizes the total number of reads between samples (library-size normalization)
- These estimates can be dominated by a handful of highly expressed genes, and this can bias results
- Alternative approaches are the trimmed mean of M values (TMM) and DEseq. Use a reference samples and exclude extreme values.
- No much data on how TMM and DEseq perform on scRNA-seq

Spike-in sequences and normalization

- Scaling factor normalization cannot distinguish between technical biases and true biological variation
- Extrinsic control genes. Spike-in sequences are added to each cell lysate at (theoretically) constant and fixed amounts.
- External RNA Control Consortium (ERCC) molecules
- <u>Critical assumption:</u> technical effects equally affect the intrinsic and extrinsic genes
- <u>Major technical challenge:</u> Calibrating the added number of spike-in molecules
- In practice spike-ins based normalization is not very effective

Tailored normalization strategies for scRNA-seq

- BASiCS (Vallejos et al. *PLoS CB* 2015)
- SCNorm (Bacher et al. *Nat. Methods* 2017)
- scTransform (Hafemeister & Satija, *Genome Biology* 2019) implemented in Seurat. scTransform performs (~log) transformation and normalization at the same time.
- MAST (Finak et al. *Genome Biology* 2015) include a normalization factor in our differential gene expression analysis (More on this later)

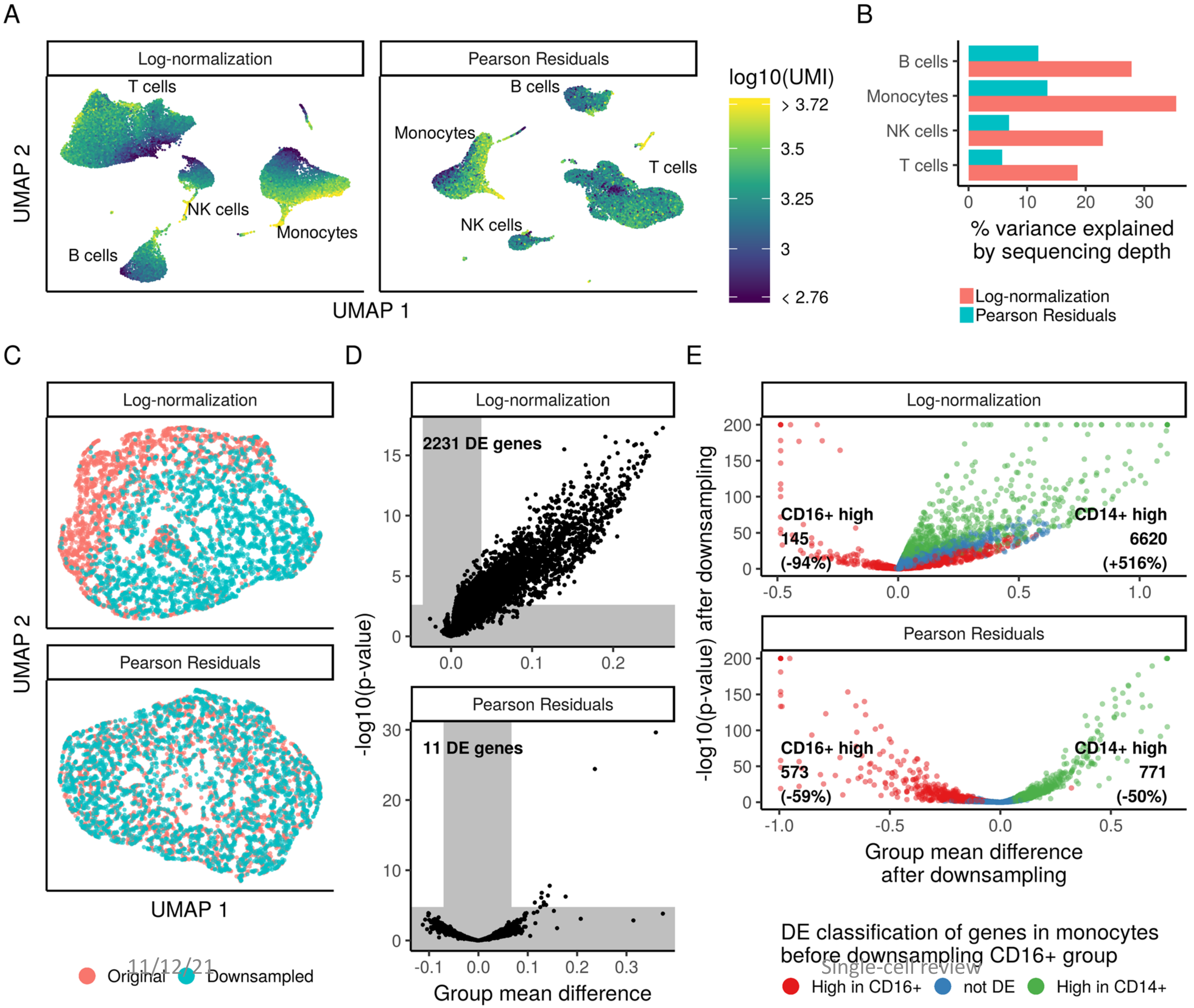
Basics of scTranform

- Inspired by methods from bulk RNA-seq (e.g. DEseq, edgeR)
- Parameters estimated by regularized Negative-Binomial regression

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m, \qquad \xrightarrow{\text{Transformation}} \qquad \begin{aligned} z_{ij} &= \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}}, \\ \mu_{ij} &= \exp\left(\beta_{0_i} + \beta_{1_i} \log_{10} m_j\right), \\ \sigma_{ij} &= \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}}, \end{aligned}$$

Gene i (count)

Total count (over all genes) for a cell


A
Log-normalization
Pearson Residuals
T cells
NK cells
Monocytes
B cells
B cells
Monocytes
T cells
NK cells
UMAP 2
UMAP 1
log10(UMI)
> 3.72
3.5
3.25
3
< 2.76
B
B cells
Monocytes
NK cells
T cells
0
10
20
30
% variance explained
by sequencing depth
🟥 Log-normalization
🟦 Pearson Residuals
C
Log-normalization
Pearson Residuals
UMAP 2
UMAP 1
● Original ● Downsampled
D
Log-normalization
2231 DE genes
15
10
5
0
0.0
0.1
0.2
Pearson Residuals
11 DE genes
30
20
10
0
-0.1
0.0
0.1
0.2
0.3
-log10(p-value)
Group mean difference
E
Log-normalization
CD16+ high
145
(-94%)
CD14+ high
6620
(+516%)
200
150
100
50
0
-0.5
0.0
0.5
1.0
Pearson Residuals
CD16+ high
573
(-59%)
CD14+ high
771
(-50%)
200
150
100
50
0
-1.0
-0.5
0.0
0.5
-log10(p-value) after downsampling
Group mean difference
after downsampling
DE classification of genes in monocytes
before downsampling CD16+ group
● High in CD16+ ● not DE ● High in CD14+

Batch effect correction: regression methods

Motivation

- Batch effects are technical sources of variation that have been added to the samples during handling. Example of batch variables: lot number, technician, instrument settings, plate, etc.
- More common with high throughput technologies
- If not adjusted for, these batch variables can have a substantial effects on downstream analysis.
- Normalization will not always correct for batch effects. Technical variation due to batch effects might only affect a subset of the genes.

Adjusting for batch effects

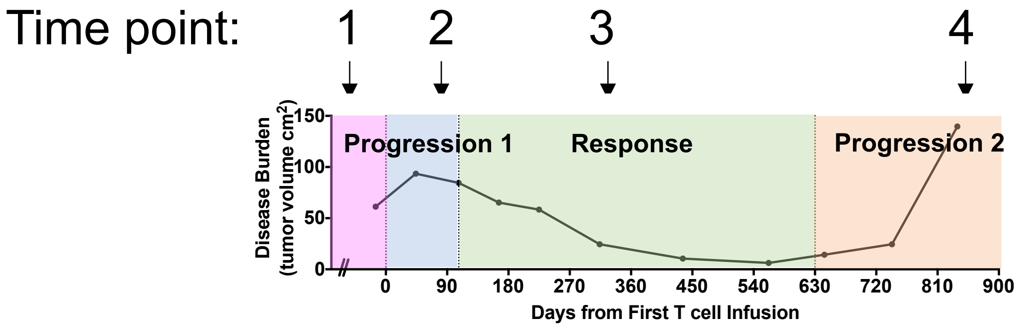
- **Two scenarios:**

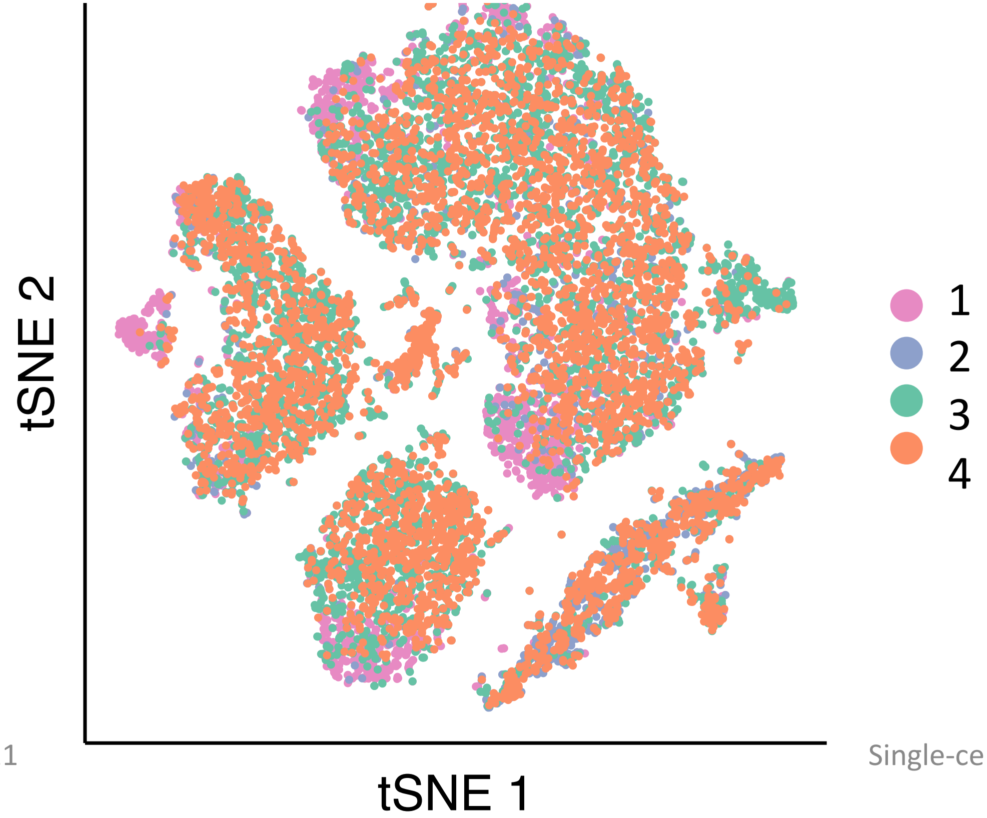
1. **You have information about the batch variable**
   Use your batch effect as a covariate in your analysis (e.g. MAST, Harmony)
2. **You suspect a batch effect, but you don't know where it is coming from**
   The batch effect needs to be estimated first and then corrected for, by adding the estimated variables as co-variates

Quality control and back variability

- **Look at your data!**
- Use a dimension reduction technique (e.g. PCA/tSNE) and plot individual samples (i.e. cells) coloring each cell by various metadata variable (e.g. processing time, plate, etc)
- Do you see variation that can be explain by your experimental factors?
  - Yes: Include them in your model
  - No: Try to estimate potential batch effects

Single-sell RNA-seq: 4 PBMC samples





- Cells arrange themselves by gene expression
- Different time points overlay (processing effects minimal)

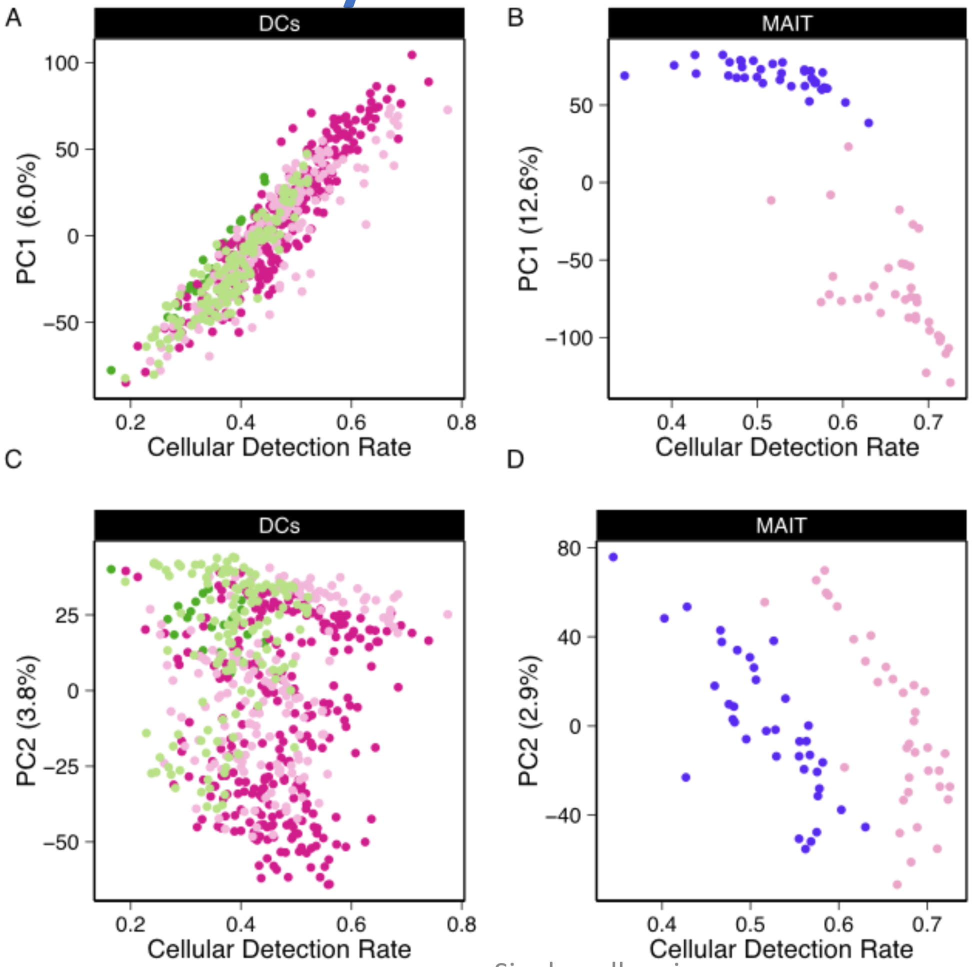
Surrogate variable analysis

- Why not use principal component analysis (PCA) to detect potential batch effects?
- Problem: Biological variation or technical variation?
- Solution: Regress out all known biological/technical variation (e.g. using a linear regression) and use PCA on residuals
- Use the first *k* principle components as estimated batch effects (surrogate variables)
- How many surrogate variables should one use (k=1, 2, etc)?

CDR effects in single-cell RNAseq

- **Principal component analysis leads to cellular detection rate (CDR)**

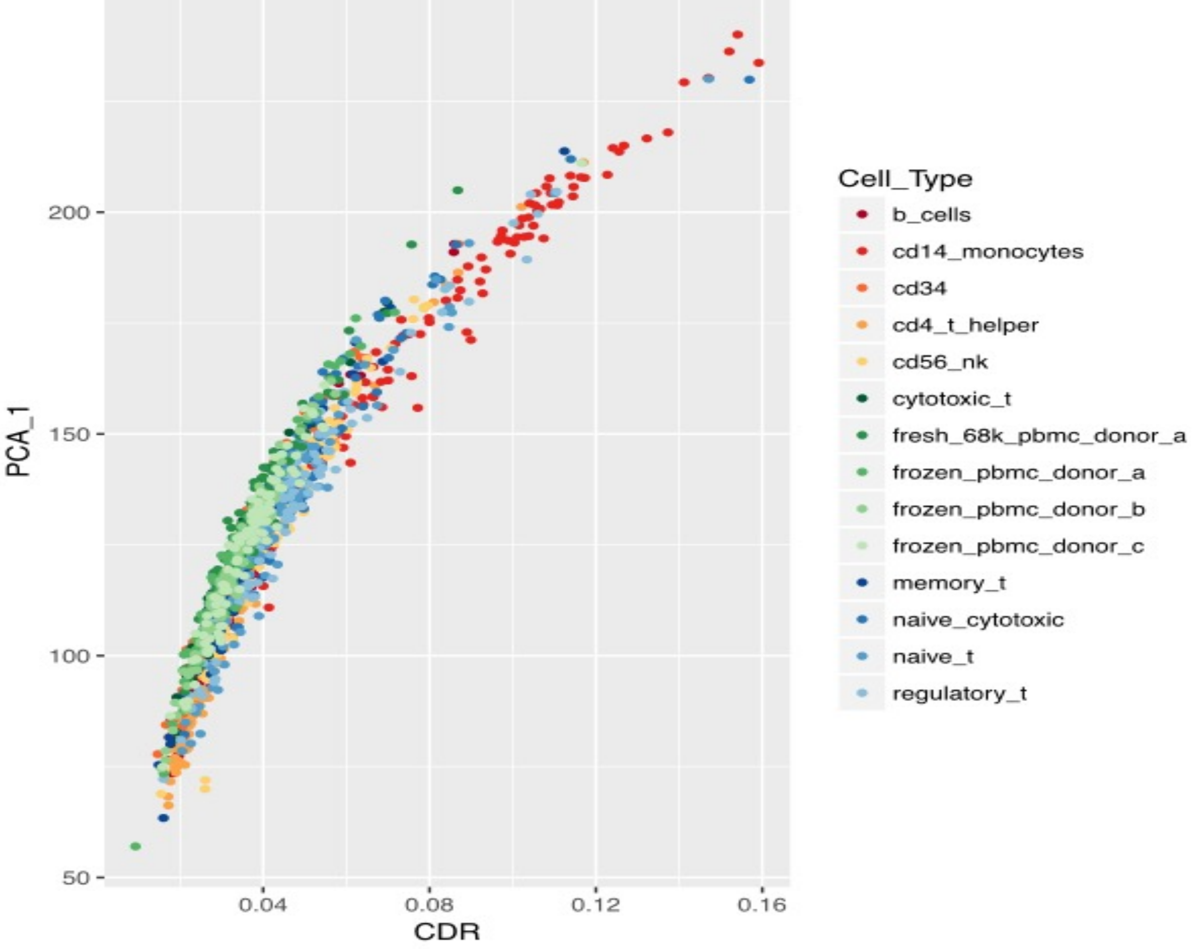
DCs:
Shalek et al. Nature (2014)
Mouse dendritic cells
(color coded by time of stimulation)



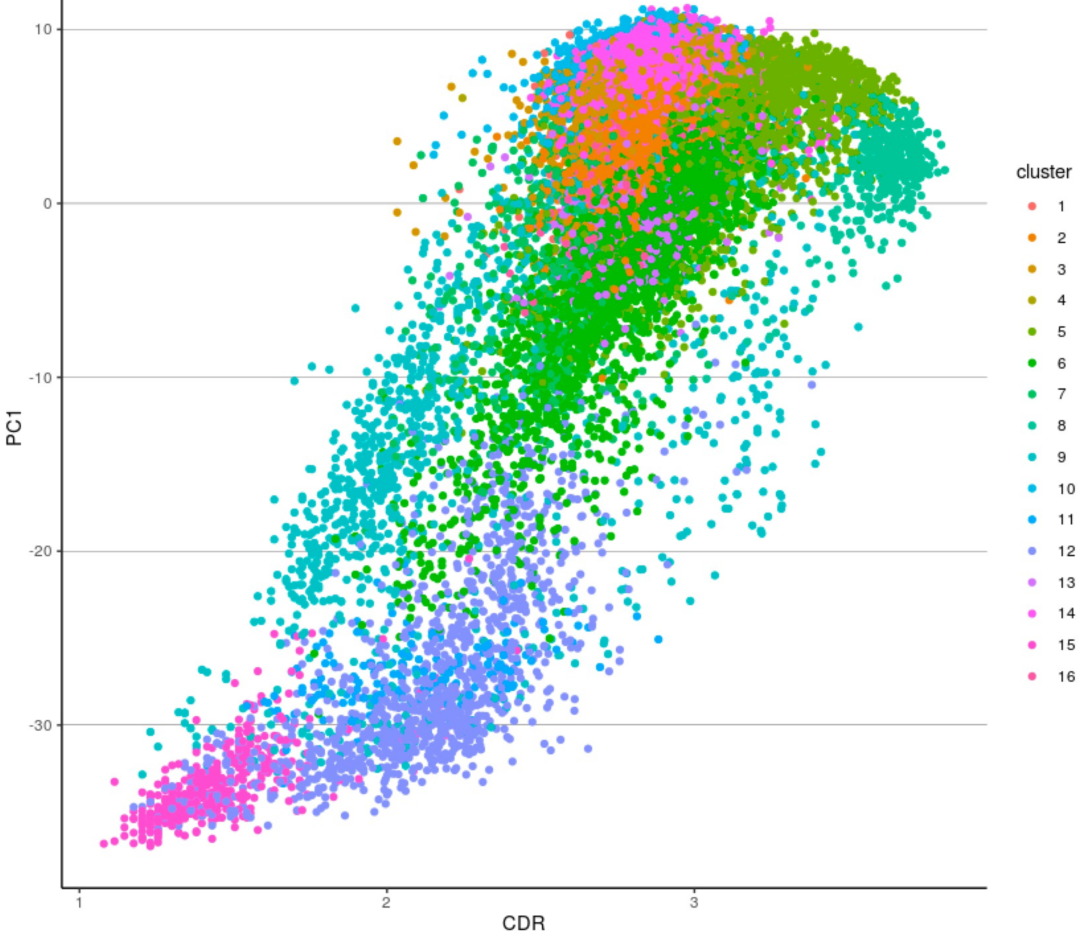
MAITs:
Finak et al. (2014)
Mucosal Associated Invariant T-cells
(color coded by stimulation)

unstimulated
stimulated with IL18, IL15, IL12

CDR effects in 10X Genomics



Zheng et al. (2017). “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.



Human Cell Atlas (Bone marrow) (2019)

The cellular detection rate (CDR): Unwanted variability

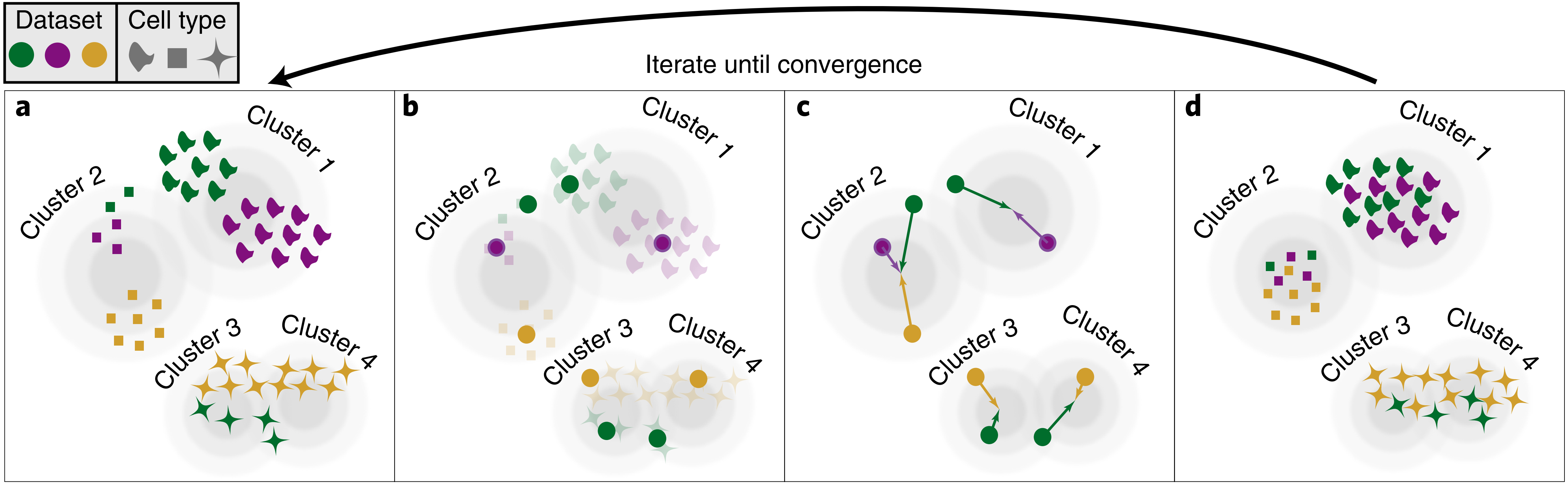
- CDR = proportion of expressed gene in a cell (Finak et al. 2015)
  - Later described in Hicks et al. 2017
- Huge source of variability, possibly confounded with treatment effects
  - Mixed of technical and biological (e.g. size, see Padovan-Merhar et al. 2015)
- Empirical estimates (e.g. mean expression) and associated effects (i.e. differentially expressed genes) can be inflated if not adjusted for this
- CDR is highly correlated with UMI numbers
- We always adjust for CDR in differential expression analysis (more on this latter)

Batch effect correction: alignment methods

Batch correction through data alignment

- Many algorithms available:
  - Seurat v3 (Butler et al. 2018), MNN (Haghverdy et al. 2018), Harmony (Korsunsky et al. 2019), etc.
- These algorithms basically transform the original data in a way to minimize between batch variation and retain within sample variation (i.e. cell type variation)
- The result is a corrected expression matrix

Example: Harmony

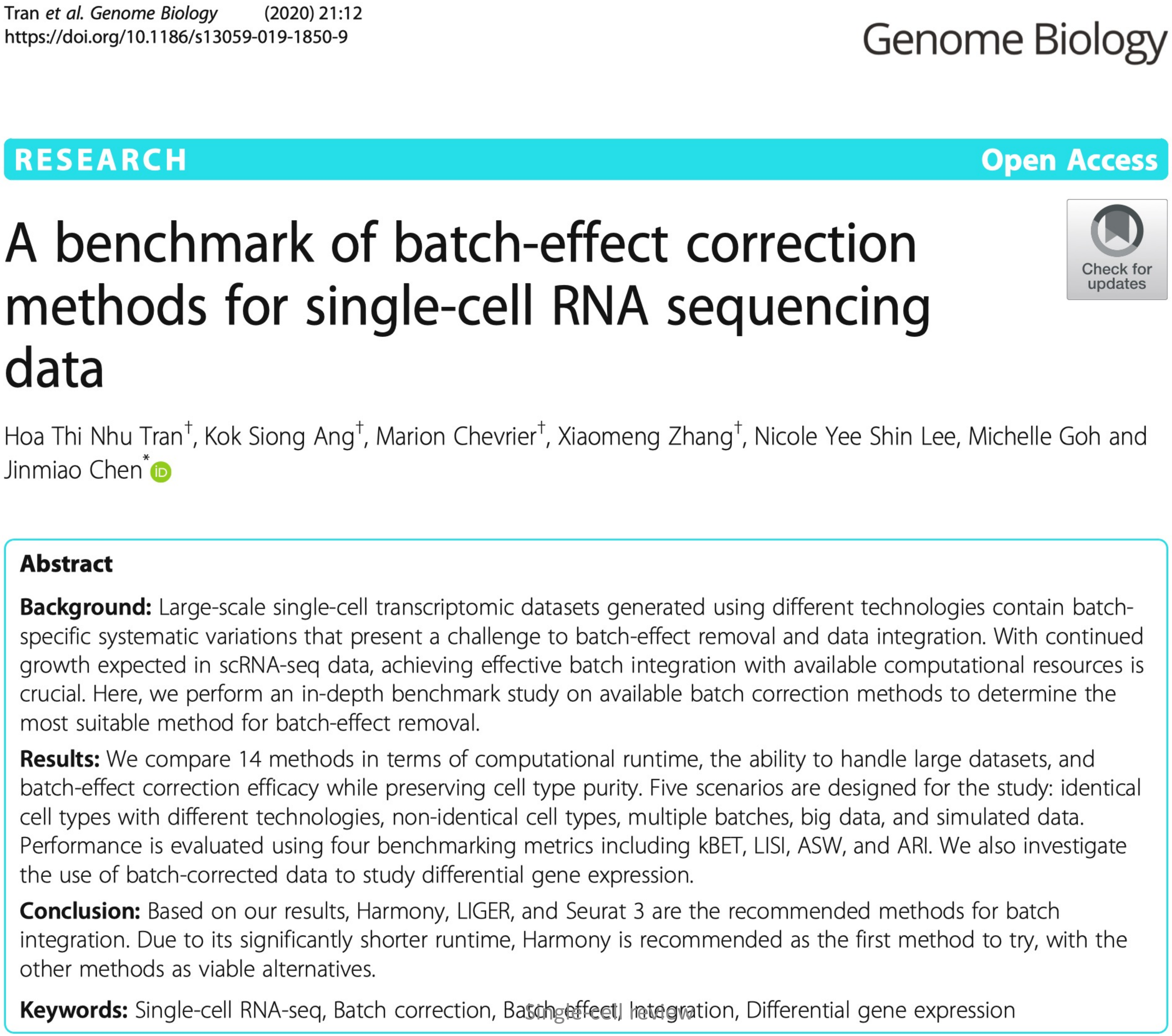


Soft assign cells to clusters, favoring mixed dataset representation

Get cluster centroids for each dataset

Get dataset correction factors for each cluster

Move cells based on soft cluster membership



Tran *et al. Genome Biology* (2020) 21:12
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

RESEARCH — Open Access

A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Hoa Thi Nhu Tran†, Kok Siong Ang†, Marion Chevrier†, Xiaomeng Zhang†, Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen*

Abstract

Background: Large-scale single-cell transcriptomic datasets generated using different technologies contain batch-specific systematic variations that present a challenge to batch-effect removal and data integration. With continued growth expected in scRNA-seq data, achieving effective batch integration with available computational resources is crucial. Here, we perform an in-depth benchmark study on available batch correction methods to determine the most suitable method for batch-effect removal.

Results: We compare 14 methods in terms of computational runtime, the ability to handle large datasets, and batch-effect correction efficacy while preserving cell type purity. Five scenarios are designed for the study: identical cell types with different technologies, non-identical cell types, multiple batches, big data, and simulated data. Performance is evaluated using four benchmarking metrics including kBET, LISI, ASW, and ARI. We also investigate the use of batch-corrected data to study differential gene expression.

Conclusion: Based on our results, Harmony, LIGER, and Seurat 3 are the recommended methods for batch integration. Due to its significantly shorter runtime, Harmony is recommended as the first method to try, with the other methods as viable alternatives.

Keywords: Single-cell RNA-seq, Batch correction, Batch effect, Integration, Differential gene expression

Alignment methods and statistical inference

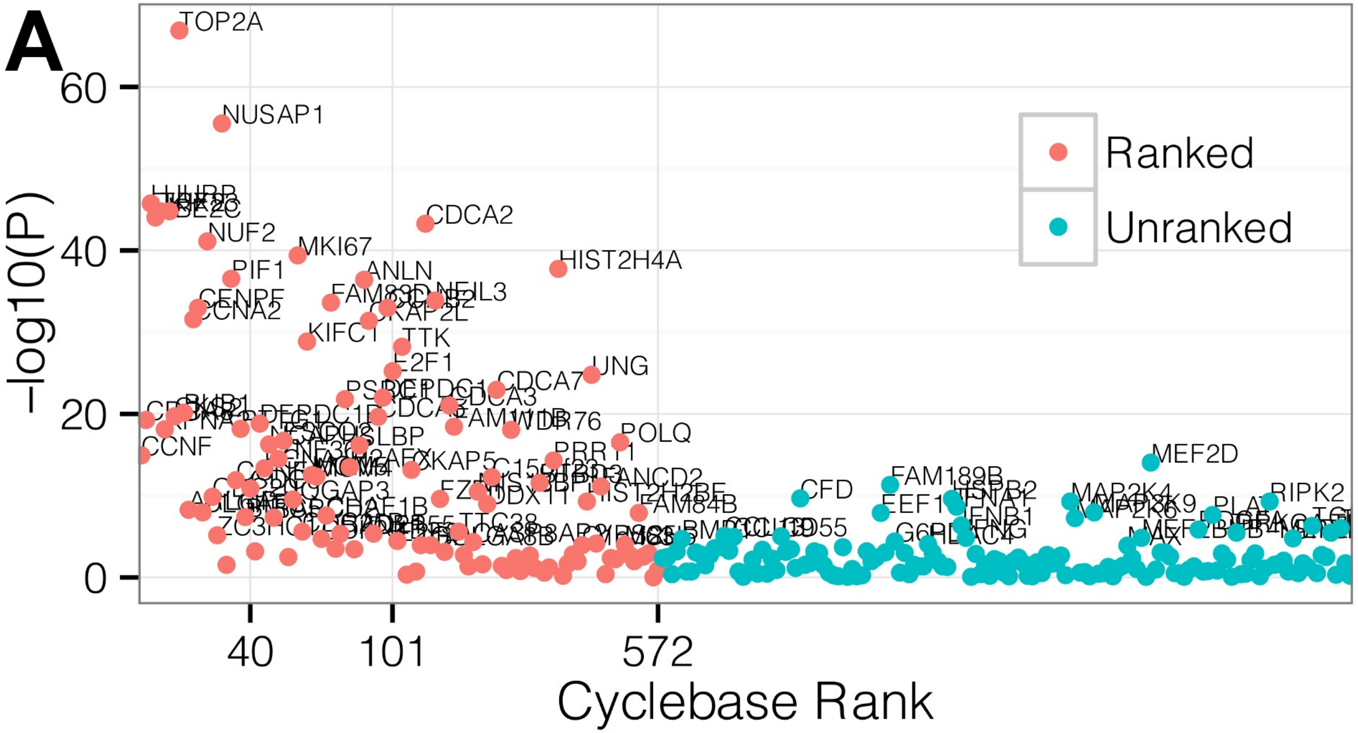
- Correction/alignment methods should only be used for visualization and/or clustering or cell type assignment (more on this later)
- It is not recommended to perform statistical inference on corrected values. This could lead to poorly calibrated results due to possible confounding of batch and biological variables
- The preferred approach is to model the uncorrected data and account for batch variables (e.g. add these as covariates in your statistical model) possibly conditional on cell type assignments (or clustering results)

Experimental design and batch effects

- Is there a way to avoid batch effects altogether?
- <u>Maybe:</u>
- Optimize/standardize your assay to reduce variability as much as possible
- Group samples of interest (i.e. samples/conditions you wish to compare) within the same batch (if possible)
- Use controls and/or technical replicates (if feasible) that can be used to estimate batch variability and correct for it

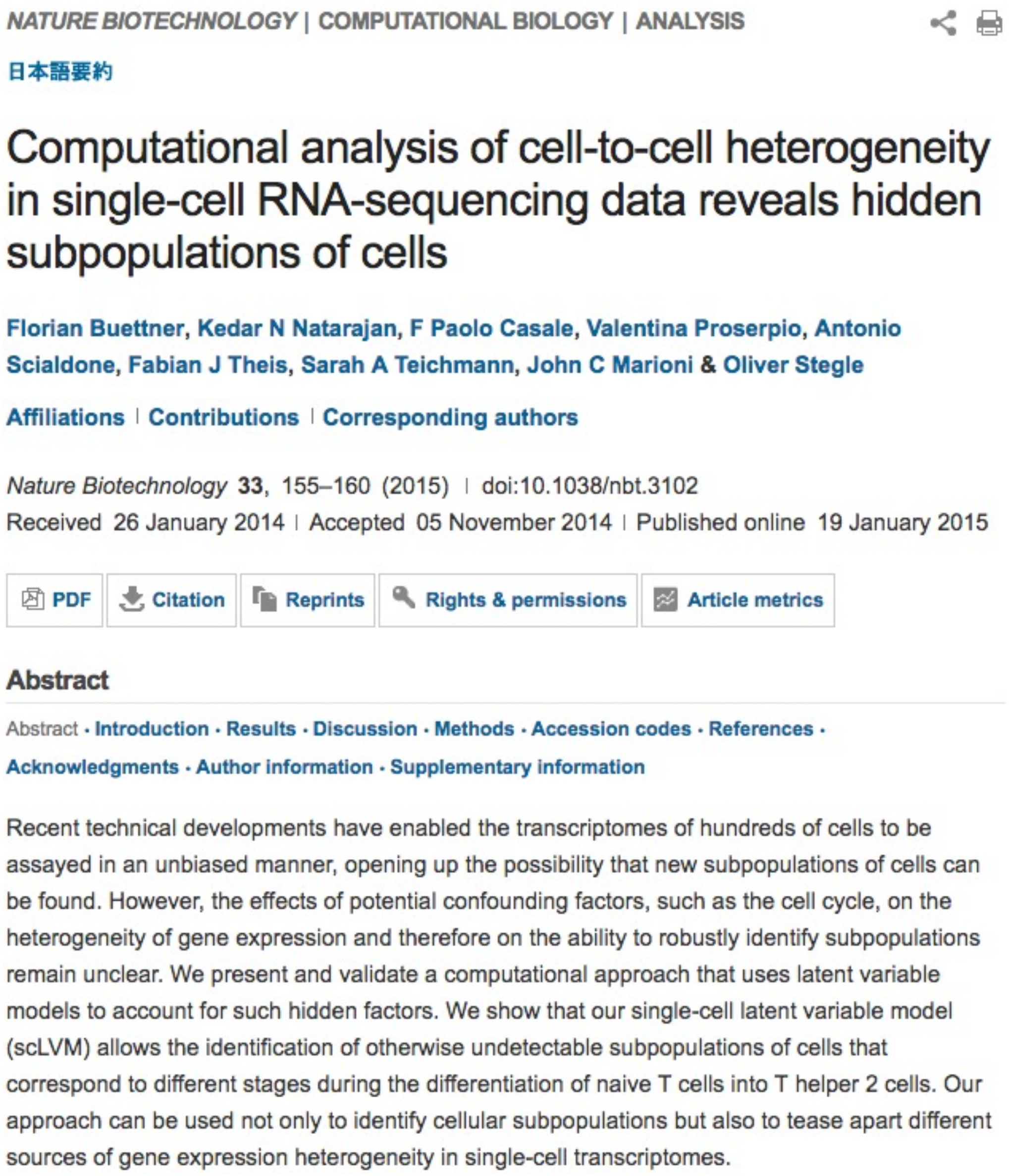
Why experimental design matters!

McDavid, A. *et al.* Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS CB* (2015).



69

Why experimental design matters!



NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

日本語要約

Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni & Oliver Stegle

Affiliations | Contributions | Corresponding authors

Nature Biotechnology **33**, 155–160 (2015) | doi:10.1038/nbt.3102
Received 26 January 2014 | Accepted 05 November 2014 | Published online 19 January 2015

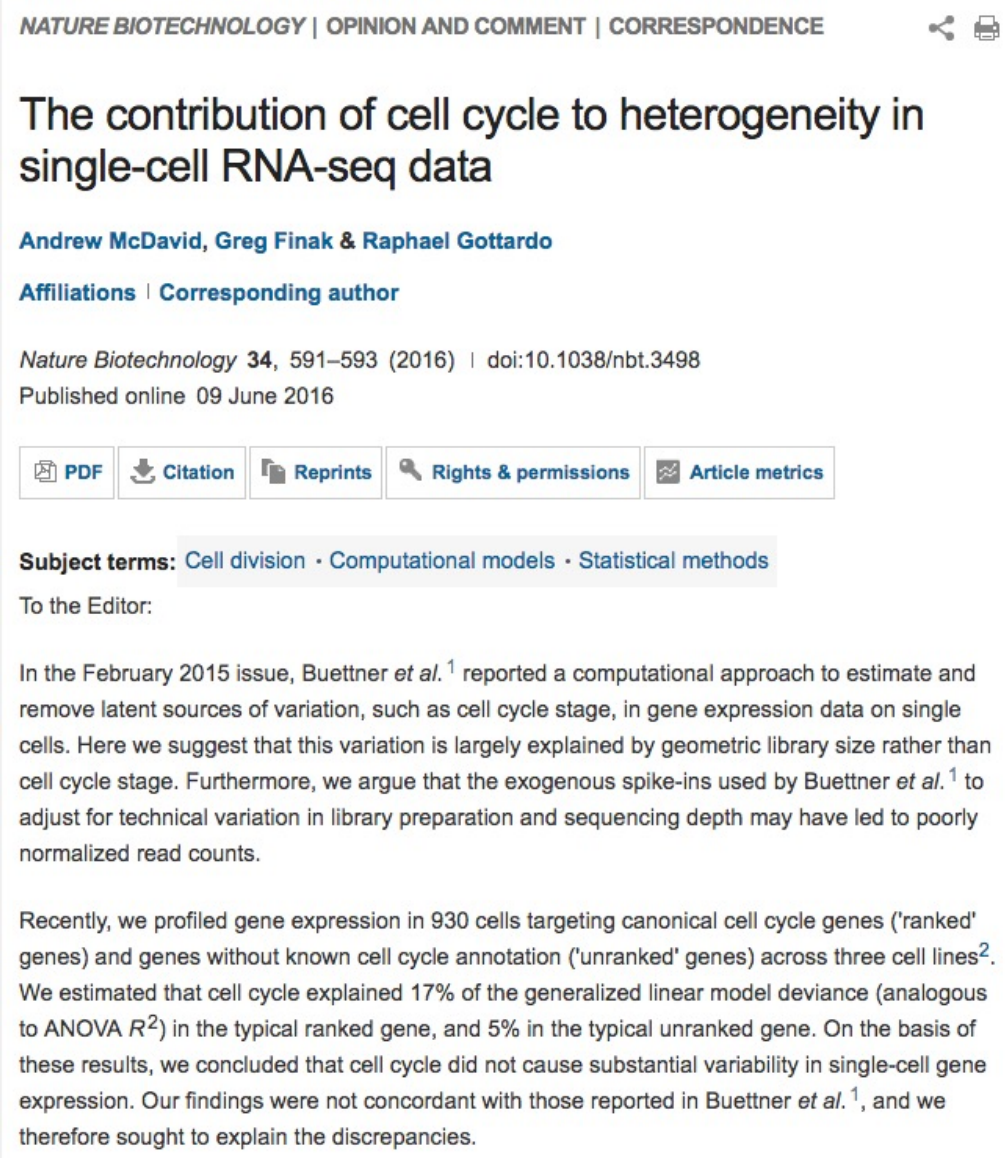
Abstract

Abstract · Introduction · Results · Discussion · Methods · Accession codes · References · Acknowledgments · Author information · Supplementary information

Recent technical developments have enabled the transcriptomes of hundreds of cells to be assayed in an unbiased manner, opening up the possibility that new subpopulations of cells can be found. However, the effects of potential confounding factors, such as the cell cycle, on the heterogeneity of gene expression and therefore on the ability to robustly identify subpopulations remain unclear. We present and validate a computational approach that uses latent variable models to account for such hidden factors. We show that our single-cell latent variable model (scLVM) allows the identification of otherwise undetectable subpopulations of cells that correspond to different stages during the differentiation of naive T cells into T helper 2 cells. Our approach can be used not only to identify cellular subpopulations but also to tease apart different sources of gene expression heterogeneity in single-cell transcriptomes.

“Cell cycle variation affects global gene expression”

Cell-cycle variation or technical variation?



NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

The contribution of cell cycle to heterogeneity in single-cell RNA-seq data

Andrew McDavid, Greg Finak & Raphael Gottardo

Affiliations | Corresponding author

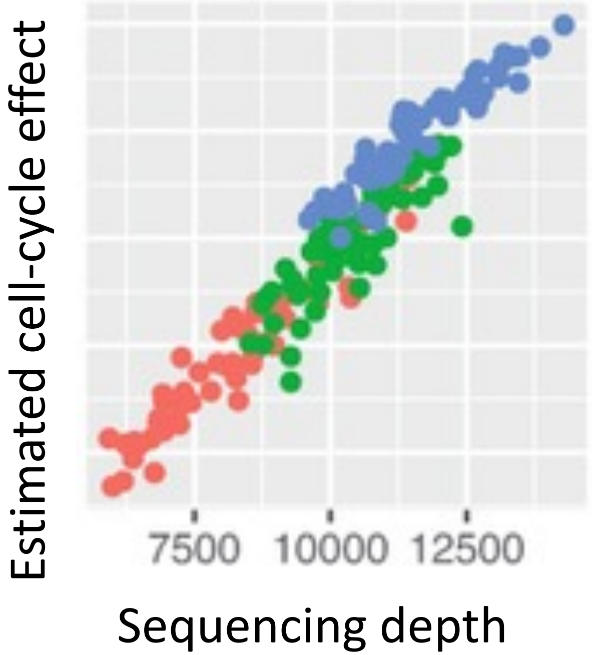
Nature Biotechnology **34**, 591–593 (2016) | doi:10.1038/nbt.3498
Published online 09 June 2016

Subject terms: Cell division · Computational models · Statistical methods

To the Editor:

In the February 2015 issue, Buettner *et al.*[1] reported a computational approach to estimate and remove latent sources of variation, such as cell cycle stage, in gene expression data on single cells. Here we suggest that this variation is largely explained by geometric library size rather than cell cycle stage. Furthermore, we argue that the exogenous spike-ins used by Buettner *et al.*[1] to adjust for technical variation in library preparation and sequencing depth may have led to poorly normalized read counts.

Recently, we profiled gene expression in 930 cells targeting canonical cell cycle genes ('ranked' genes) and genes without known cell cycle annotation ('unranked' genes) across three cell lines[2]. We estimated that cell cycle explained 17% of the generalized linear model deviance (analogous to ANOVA R^2) in the typical ranked gene, and 5% in the typical unranked gene. On the basis of these results, we concluded that cell cycle did not cause substantial variability in single-cell gene expression. Our findings were not concordant with those reported in Buettner *et al.*[1], and we therefore sought to explain the discrepancies.



- Improper normalization
- Poor experimental design
  - Cell cycle phase confounded with C1 chip
  - No replication

Clustering

Clustering overview

Goal: Group data points (cells) that “look similar”

- Most methods are distance based, groups cells that are close together → What distance metrics?

- How many cell subsets? → Trade-off between overfitting and lack of fit

- What are my cells? Most algorithms do not return a phenotype

- Getting results out of a clustering algorithm is easy making sense of it harder

- Usually helpful to use t-SNE plots and gene expression to visualize and label cell-populations

Clustering scRNA-seq data

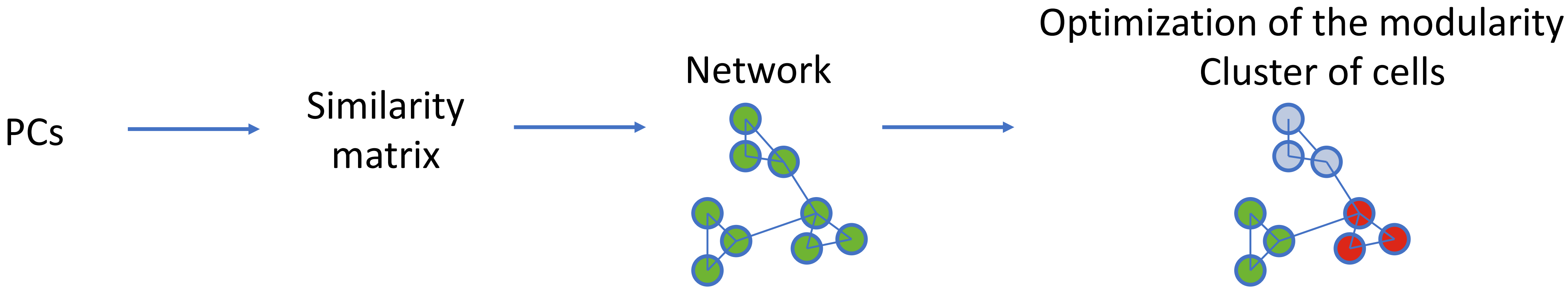
- Clustering is not a new problem!
- For example, lots of prior work in flow cytometry
- scRNA-seq data present new challenges: High dimensionality, zero-inflation, large datasets
- Lack of ground truth and benchmark datasets

Clustering tools for scRNA-seq data

- Several tools have been developed (non-exhaustive list)
  - Phenograph (Levine et al. *Cell* 2015) originally developed for CyTOF data
  - FlowSOM (Van Gassen et al. *Cytometry A* 2015) originally developed for CyTOF data
  - Louvain and Leiden graph clustering (Traag et al. *Sci. Rep.* 2019), also implemented in Seurat and the igraph R package
  - SC3 (Kiselev et al. *Nat. Meth.* 2017) an ensemble clustering approach
  - SNN-cliq (Xu & Su, *Bioinformatics* 2015)

Graph-based clustering

- Many of the algorithms listed on the previous slide are based on graph-based clustering
- These methods embed cells in a graph structure (ex: a K-nearest neighbor (KNN) network), with edges drawn between cells with similar gene expression patterns, and then attempts to partition this graph into highly interconnected communities.



PCs

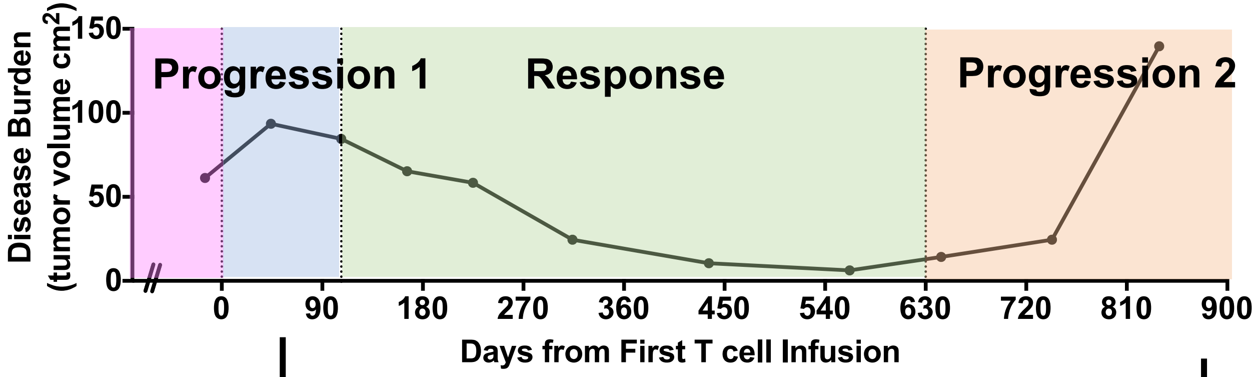
Similarity matrix

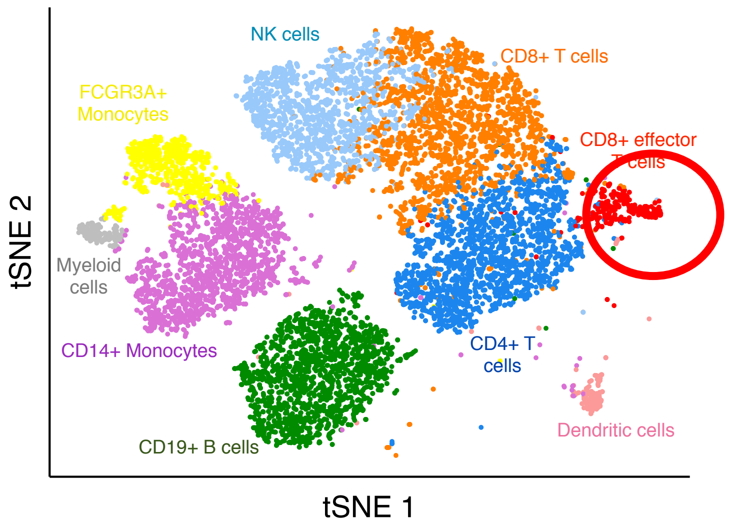
Network

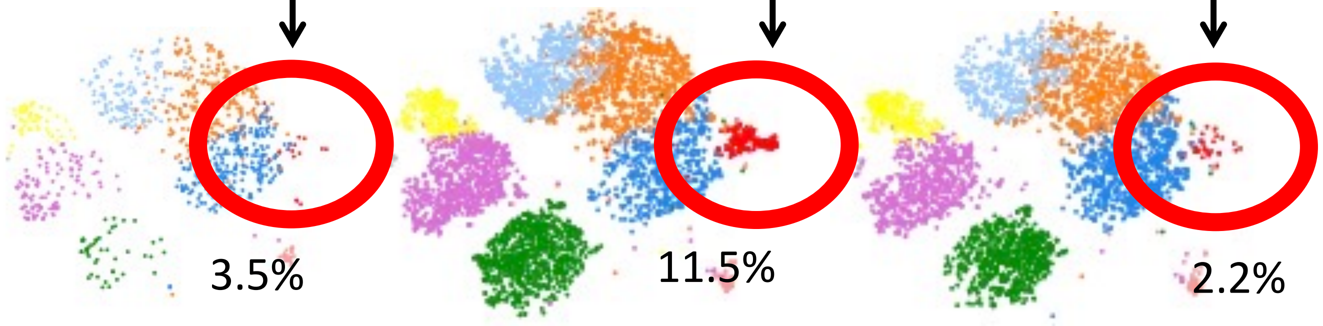
Optimization of the modularity

Cluster of cells

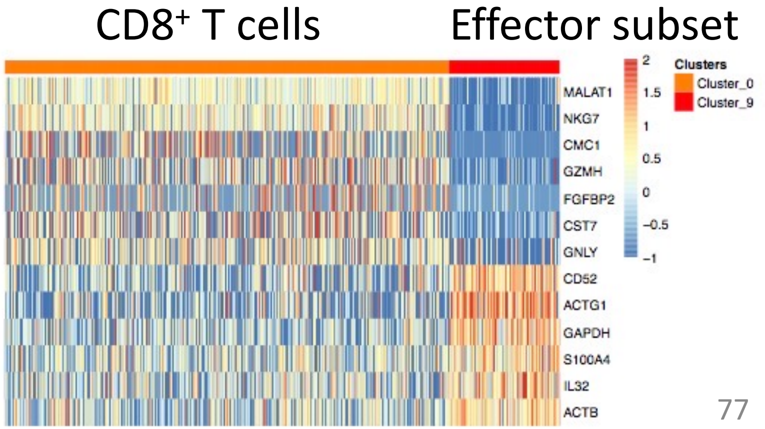
CD8+ effector population enriched in blood at time of treatment response





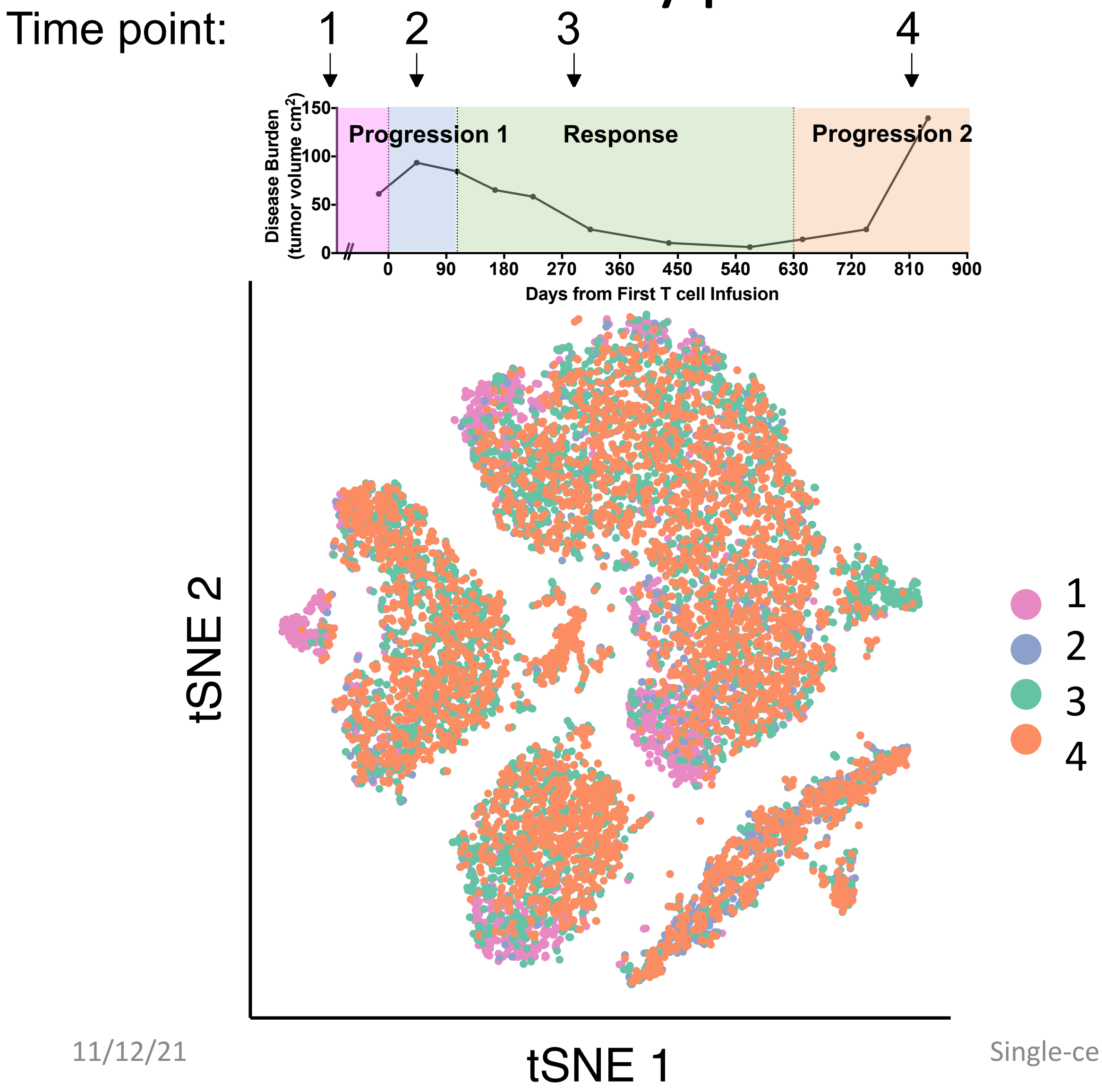


Overexpressing genes suggestive of activation, metabolism and cell division →



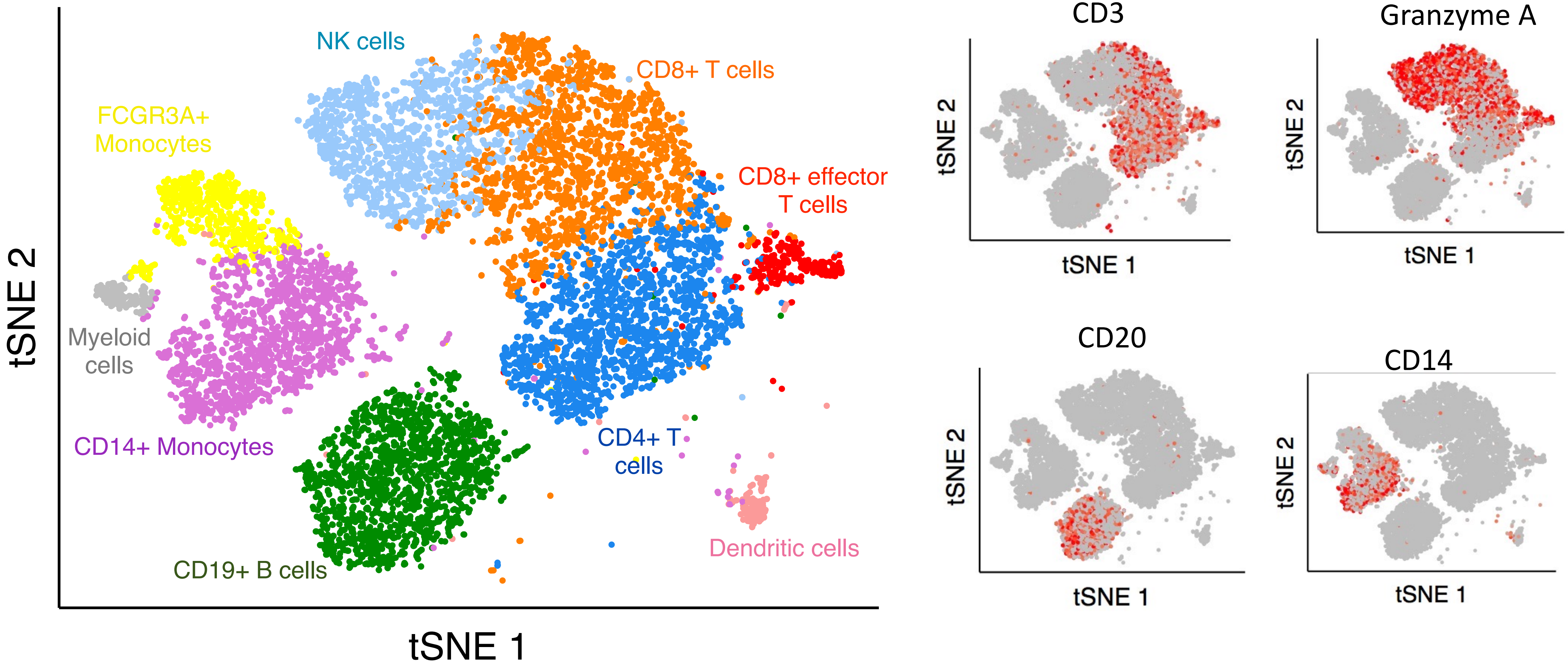
Cell type annotation

Manual cell type annotation

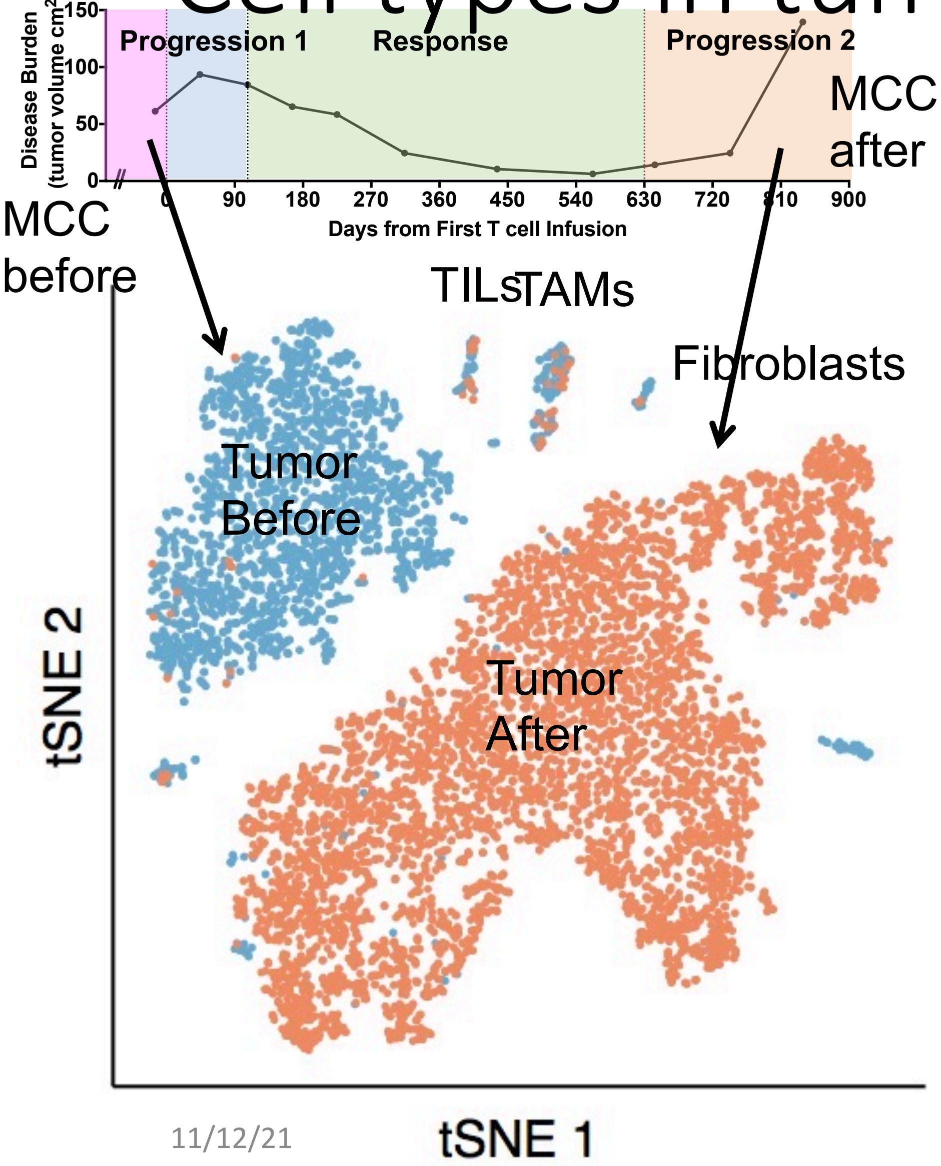


- **Cells arrange themselves by gene expression**
- **Different time points overlay (processing effects minimal)**

Manual cell type annotation

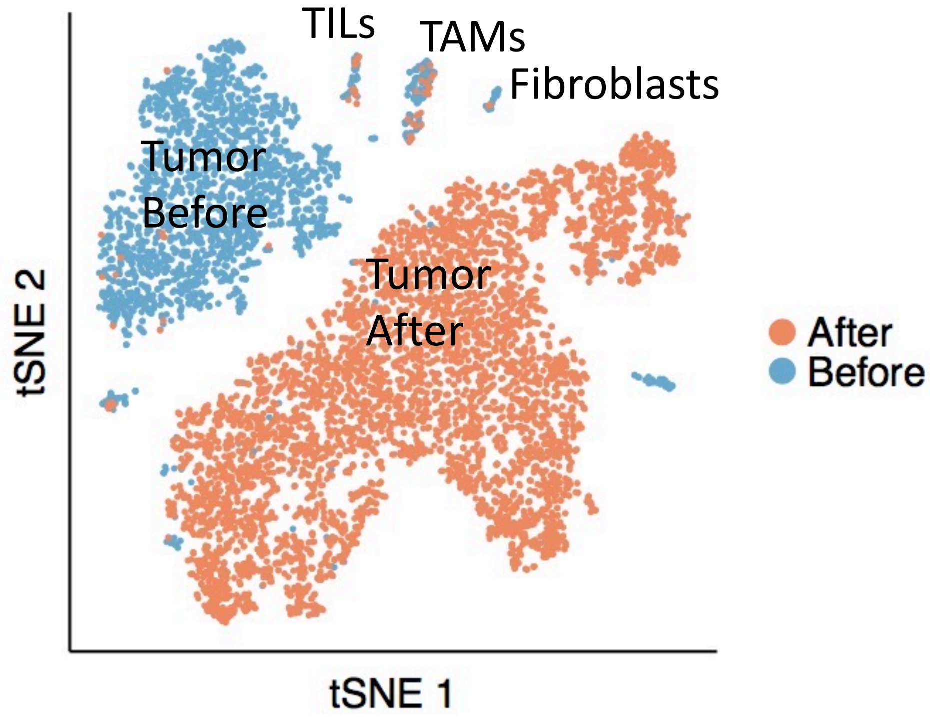


Cell types in tumor biopsies

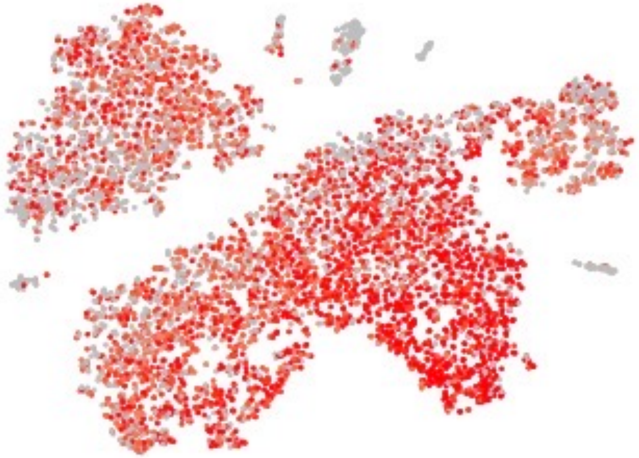


- TILs (tumor infiltrating lymphocytes), TAMs (tumor associated macrophages) and fibroblasts cluster together
- Tumor cells pre- and post-treatment cluster separately
- All tumor cells changed their gene-expression profile, suggestive of intense selective pressure

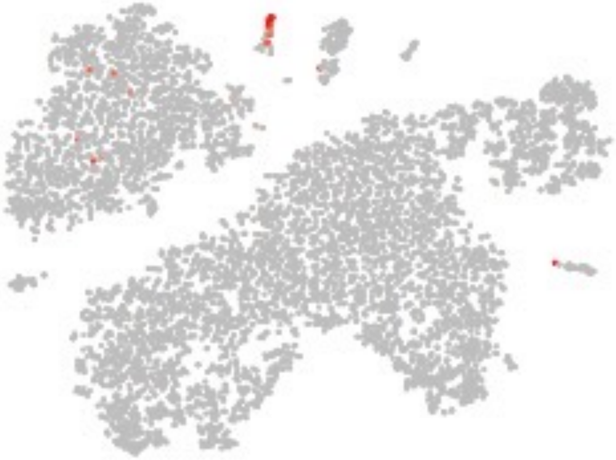
Cell types in tumor biopsies



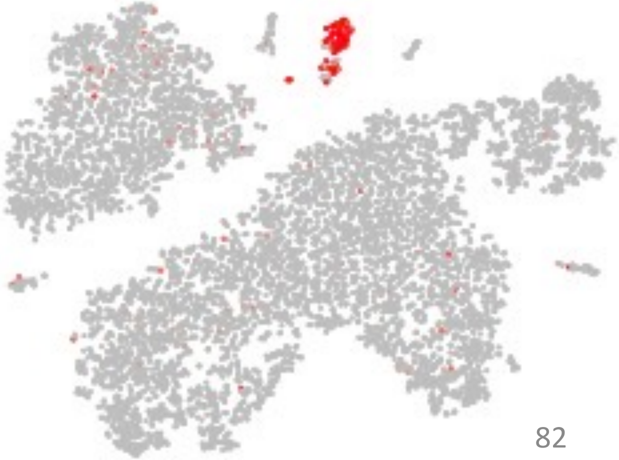
Chromogranin
(tumor marker)



CD3 (T cells)



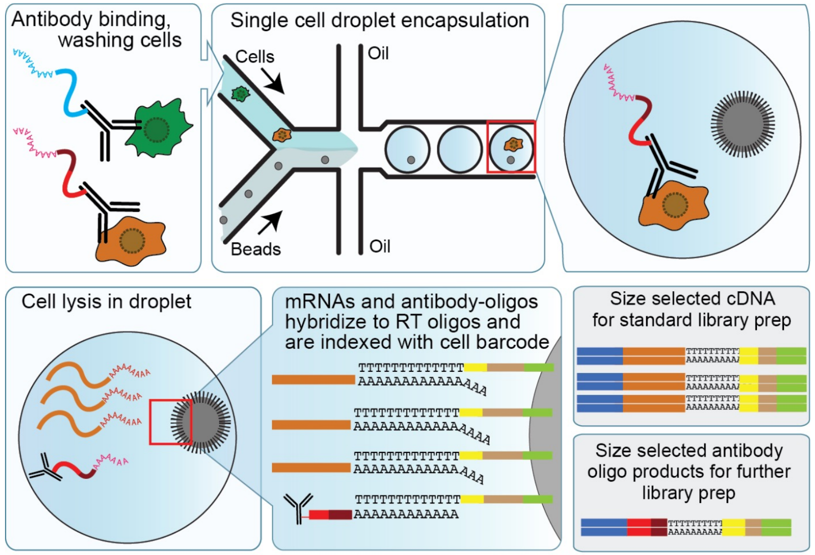
CD68 (macrophages)



Supervised cell-type annotation

- Most cell-type annotation are supervised (or semi-supervised) in the sense that they used either known marker genes or cell-type signatures (derived from sorted bulk or single-cell data) to predict cell type labels
  - Can be done at the single-cell level or cluster level
- Many methods available:
  - SingleR (Aran et al. *Nat. Immunol.* 2019)
  - scPre (Alquicira-Hernandez et al. *Genome Biology* 2019)
  - Azimuth (Hao et al. *Cell* 2021)

CITE-seq: RNA + protein in single-cells



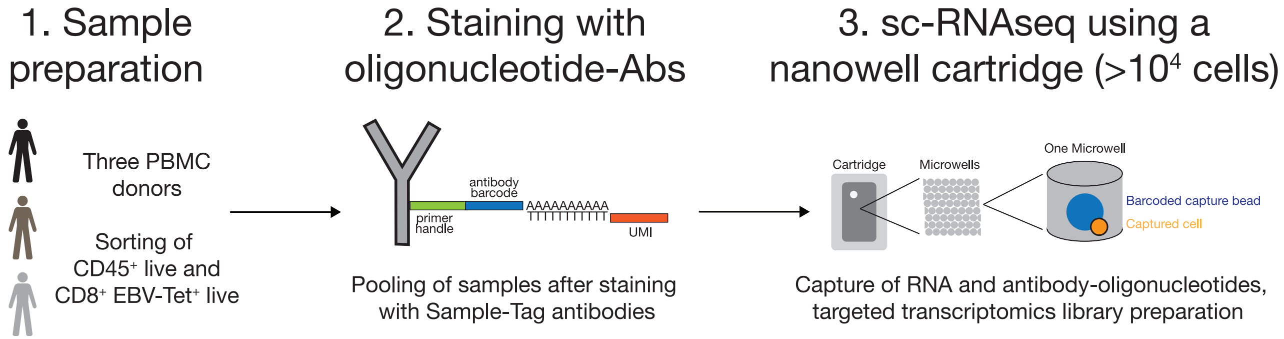
Antibody binding, washing cells
Single cell droplet encapsulation
Cells
Oil
Beads
Oil
Cell lysis in droplet
mRNAs and antibody-oligos hybridize to RT oligos and are indexed with cell barcode
Size selected cDNA for standard library prep
Size selected antibody oligo products for further library prep

CITE-seq: Stoeckius, et al. *Nat. Methods* 2017.

- Can interrogate 100s of protein in single-cells along with targeted or unbiased gene expression

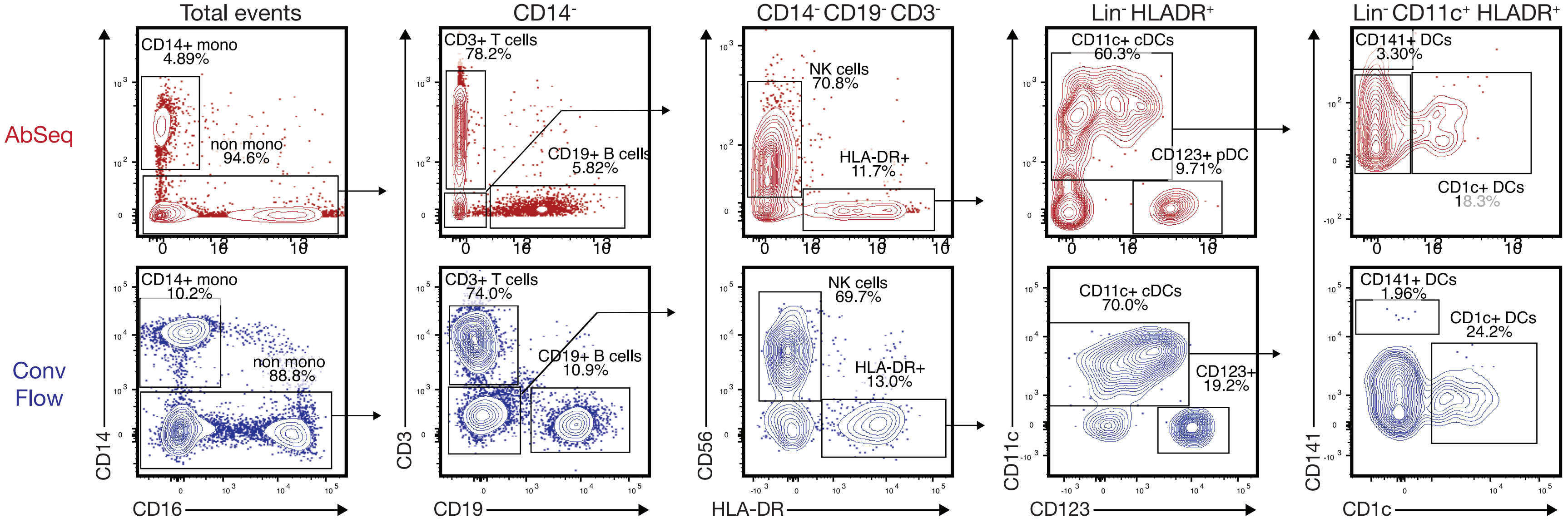
- The best of both worlds:
- **cytometry + RNA-seq in the same cell**

Collaboration with Prlic, Newell Labs
Abseq: Mair, et al. *Cell Reports* 2020.



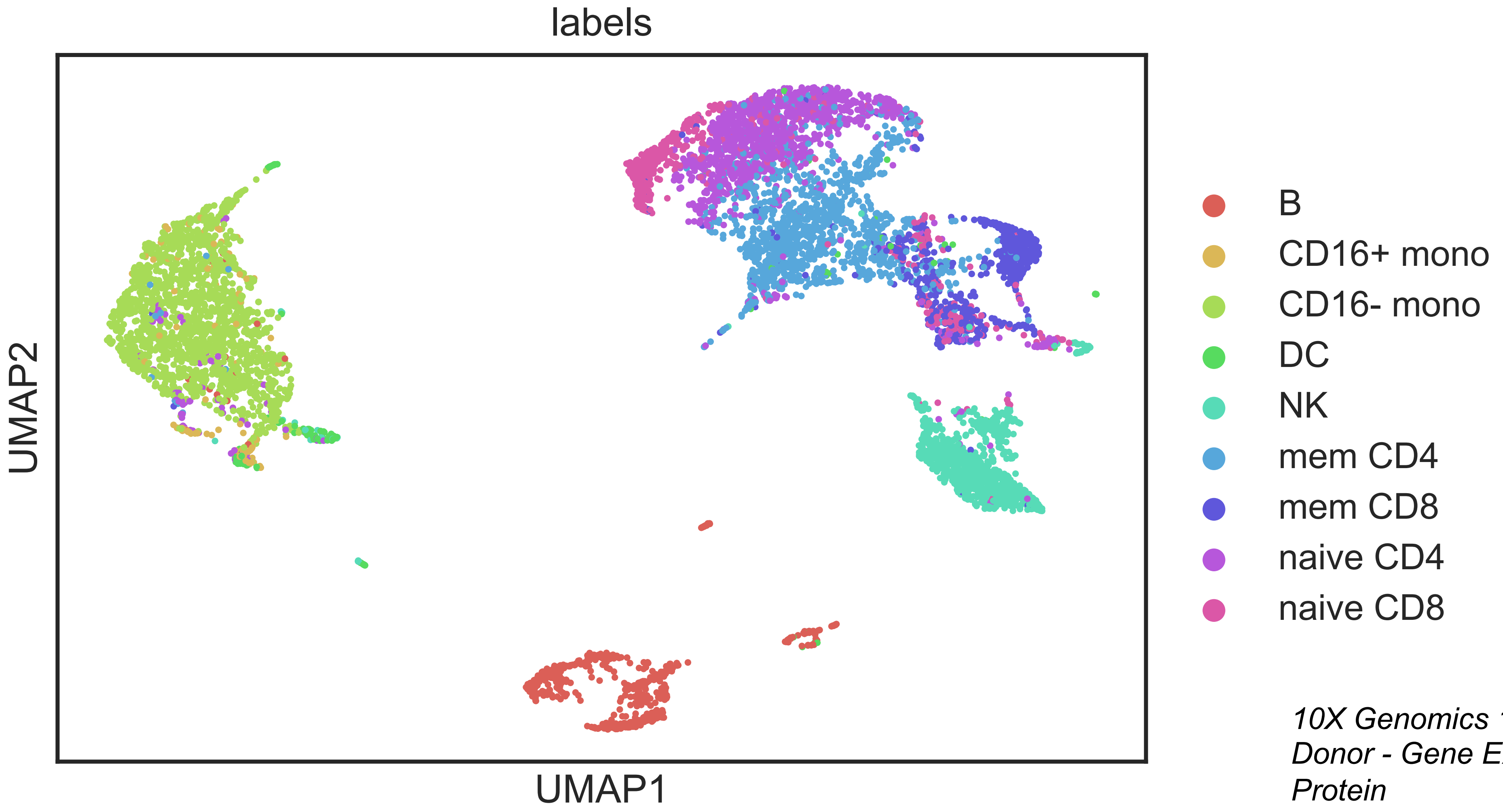
1. Sample preparation
2. Staining with oligonucleotide-Abs
3. sc-RNAseq using a nanowell cartridge (>10⁴ cells)
Three PBMC donors
Sorting of CD45⁺ live and CD8⁺ EBV-Tet⁺ live
antibody barcode
primer handle
AAAAAAAAAA
UMI
Pooling of samples after staining with Sample-Tag antibodies
Cartridge
Microwells
One Microwell
Barcoded capture bead
Captured cell
Capture of RNA and antibody-oligonucleotides, targeted transcriptomics library preparation

Identifying major cells types with proteins



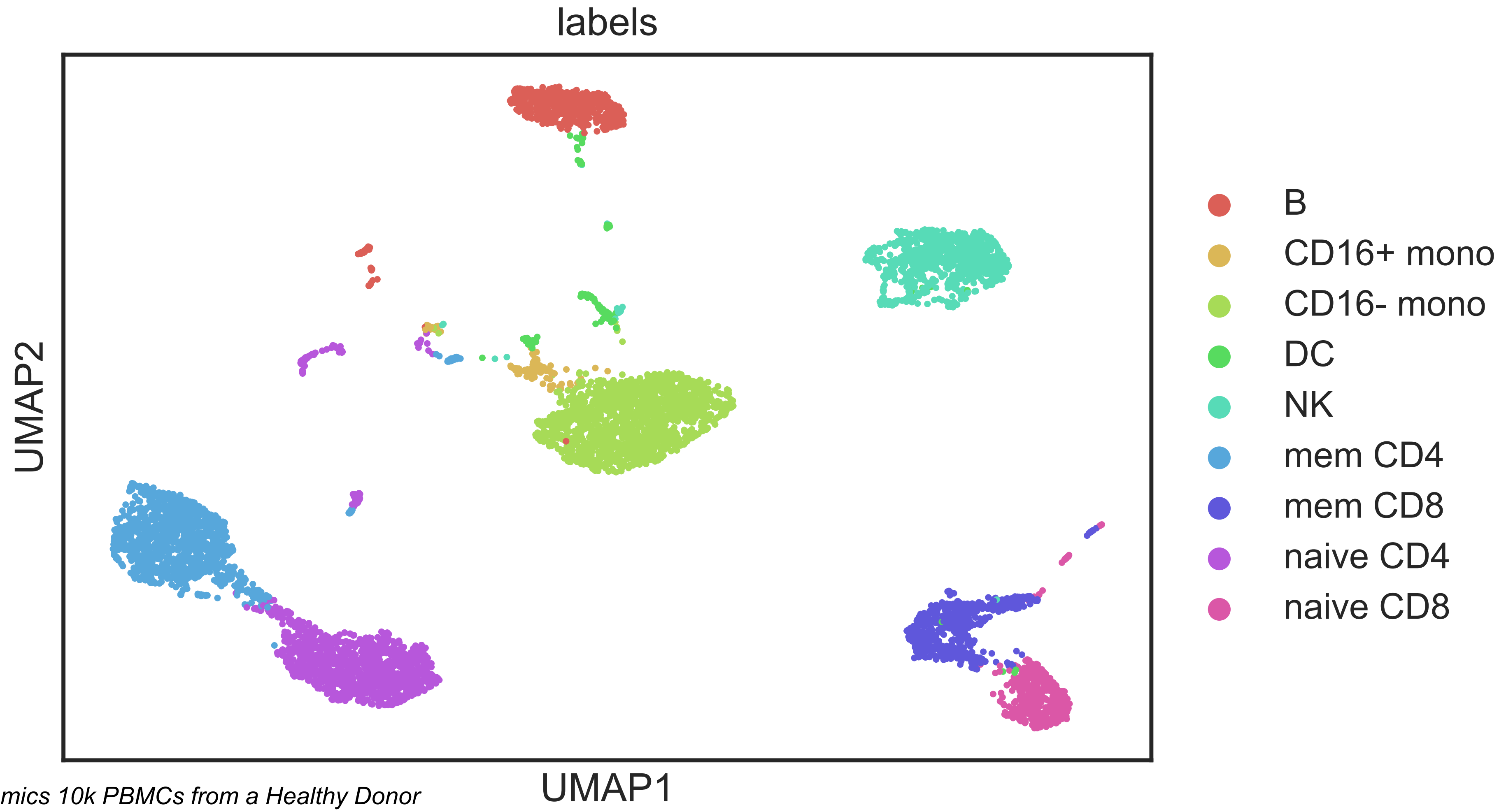
- Abseq and CITE-seq can be analyzed just like traditional cytometry data using manual (or automated) gating
- Define cell phenotypes using protein expression data, and then quantify RNA expression changes for each phenotype

Visualizing phenotypes in RNA space



10X Genomics 10k PBMCs from a Healthy Donor - Gene Expression and Cell Surface Protein

Visualizing phenotypes in protein space



10X Genomics 10k PBMCs from a Healthy Donor
- Gene Expression and Cell Surface Protein

Weighted-nearest neighbor: combining

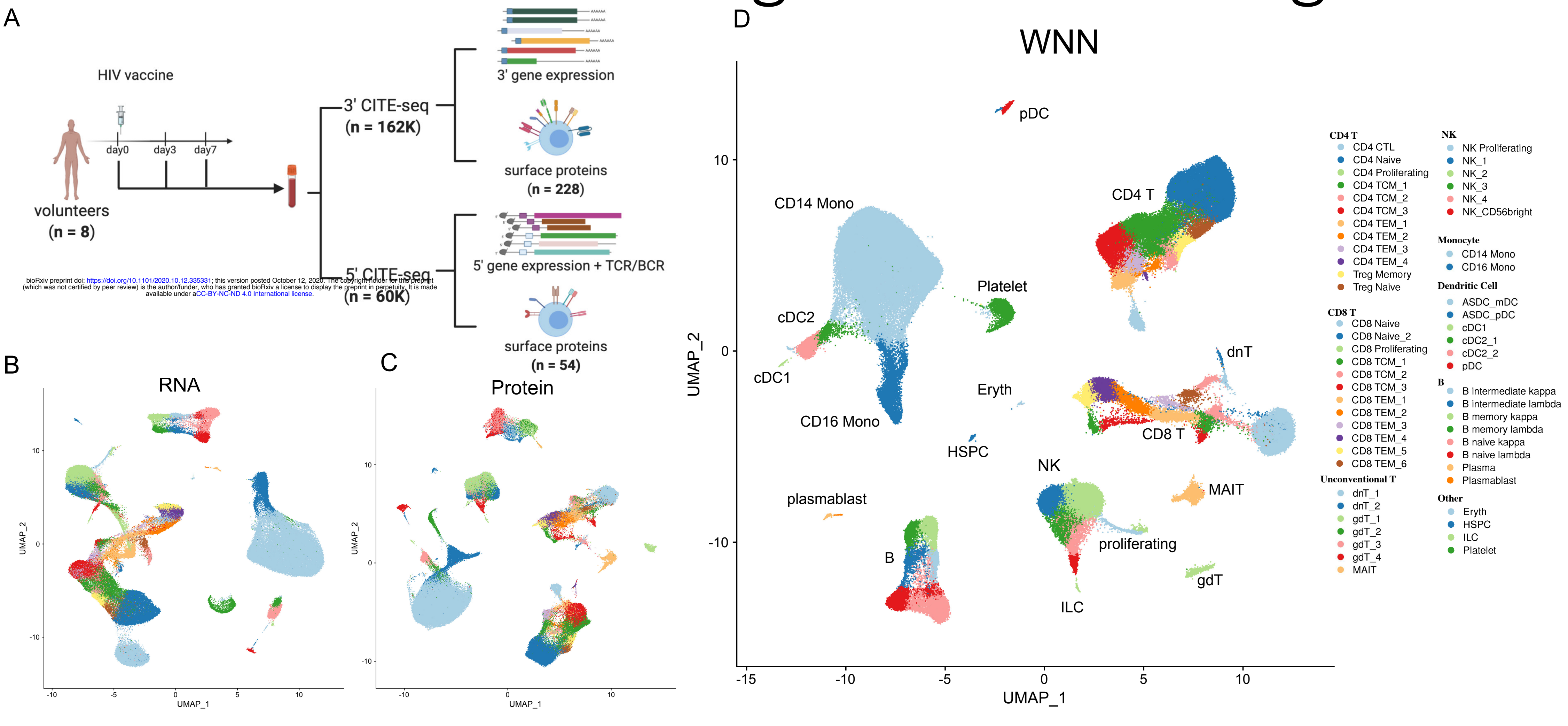
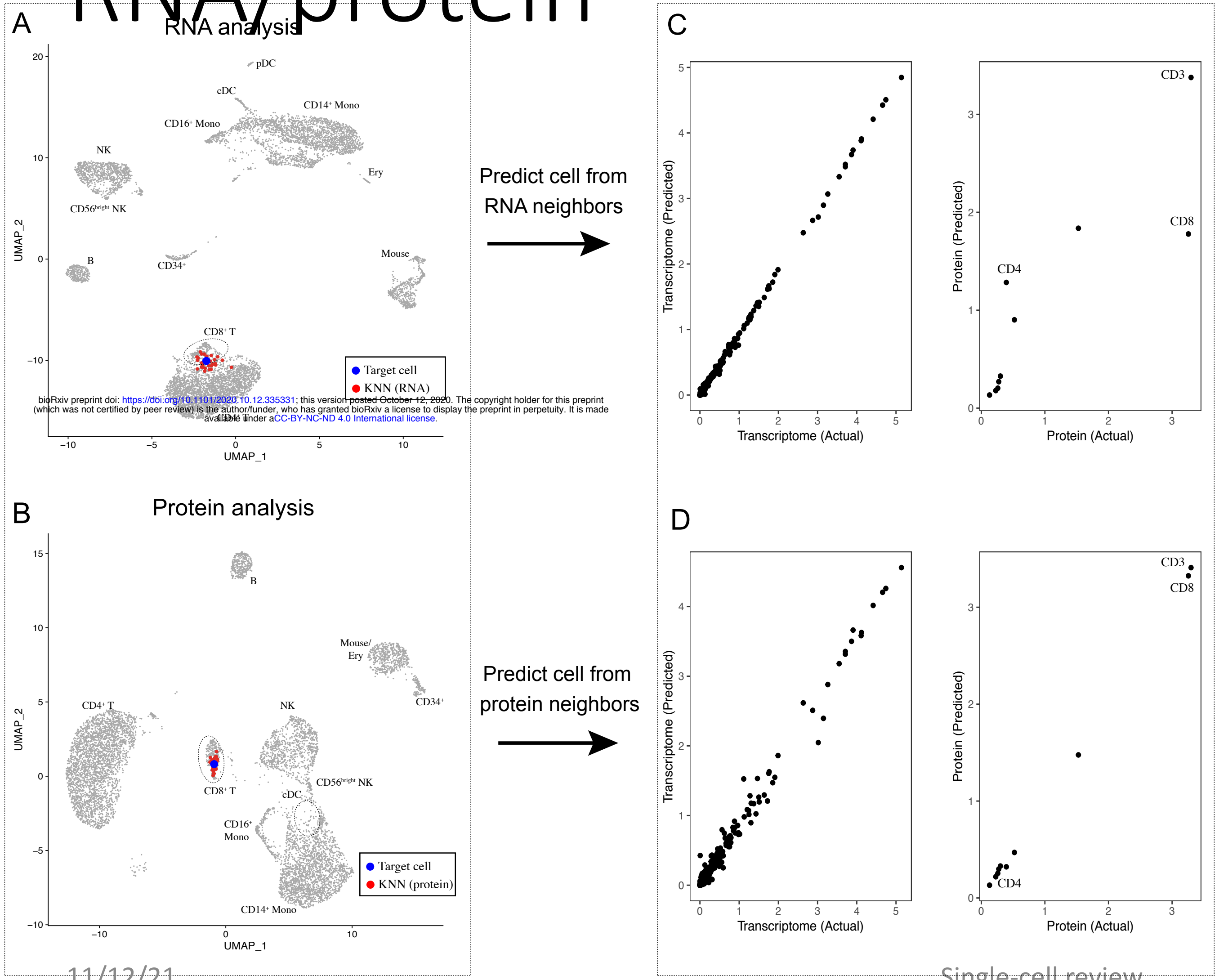
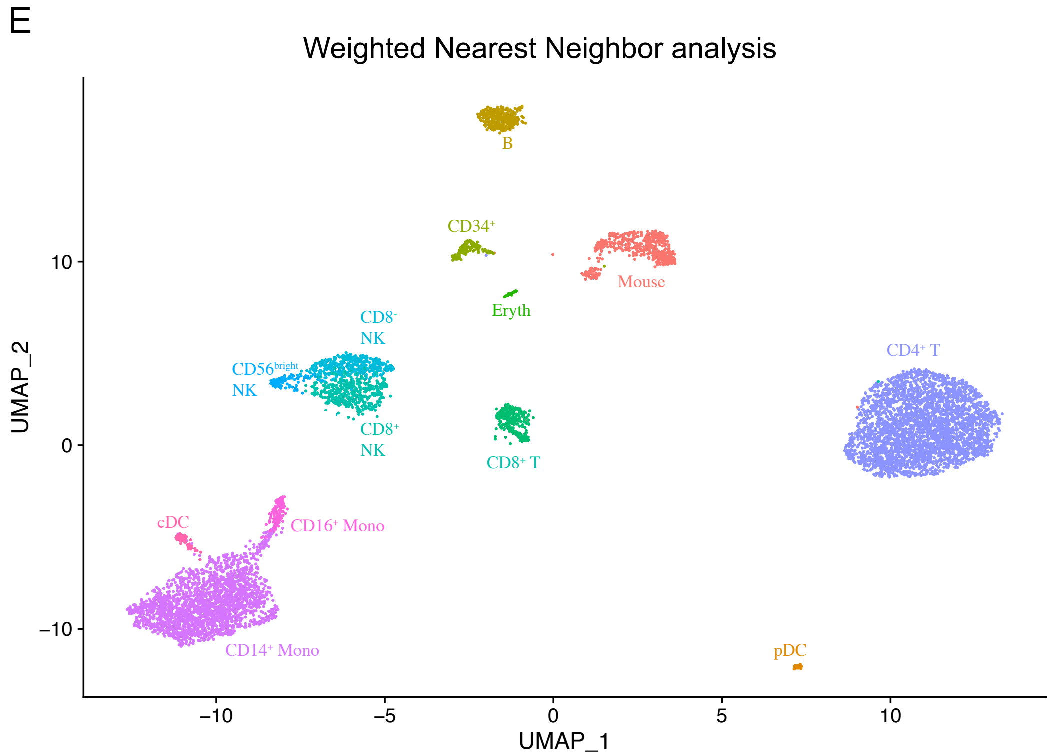


Figure 3: A multimodal atlas of human PBMC

Weighted-nearest neighbor: combining RNA/protein

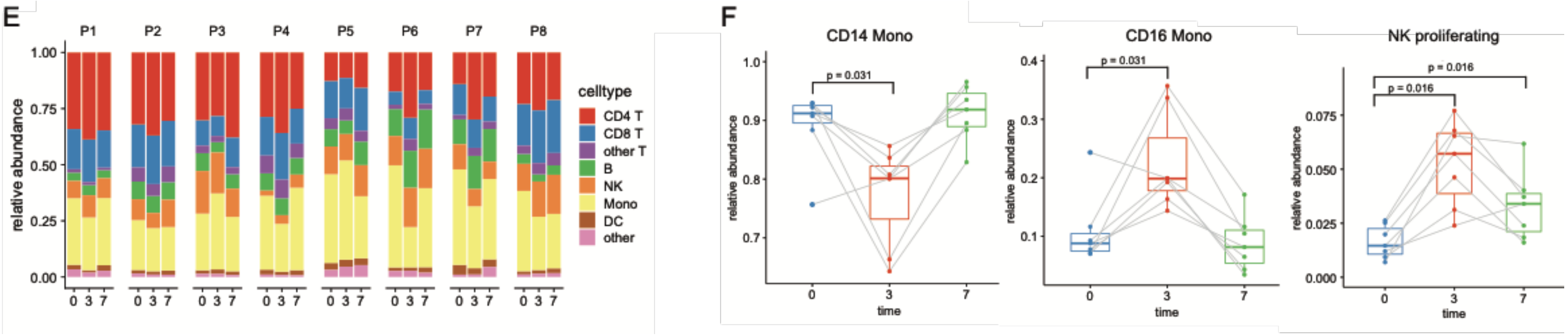


- Weights can then be used to compute a weighted nearest neighbor (WNN) graph
- WNN can be used for visualization and clustering

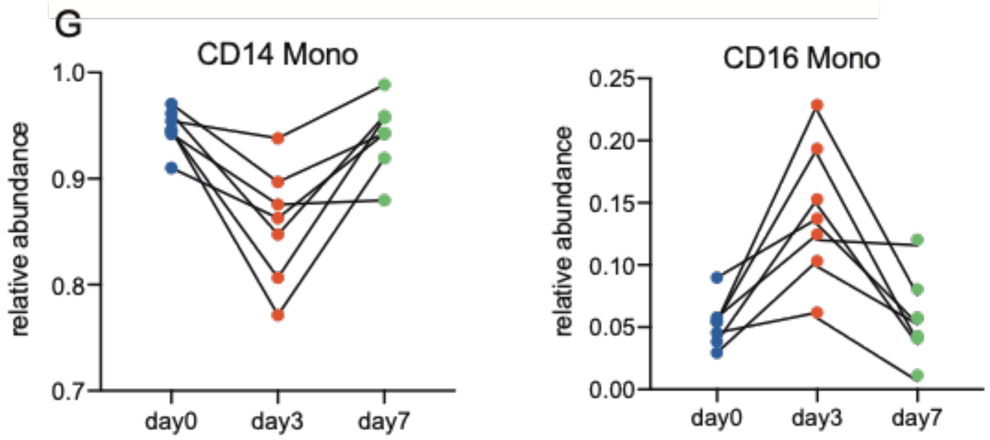


Detecting vaccine induced changes over time

CITE-Seq



Flow cytometry

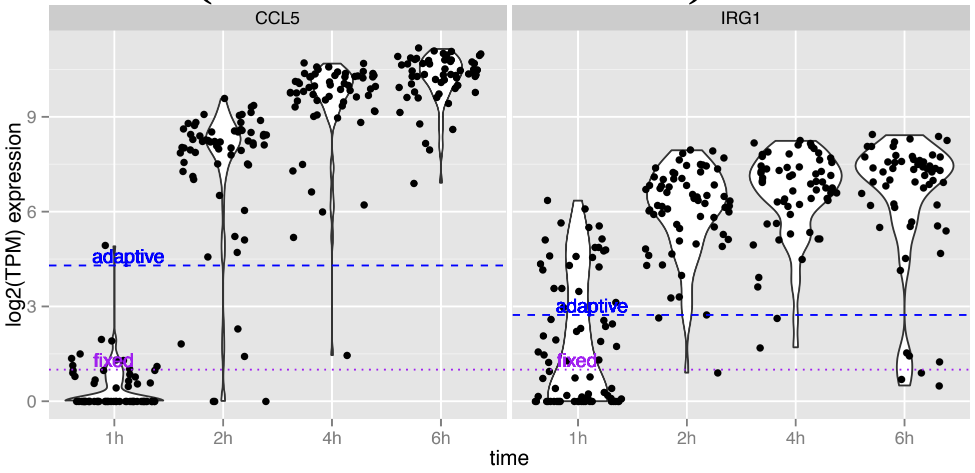


Differential gene expression analysis

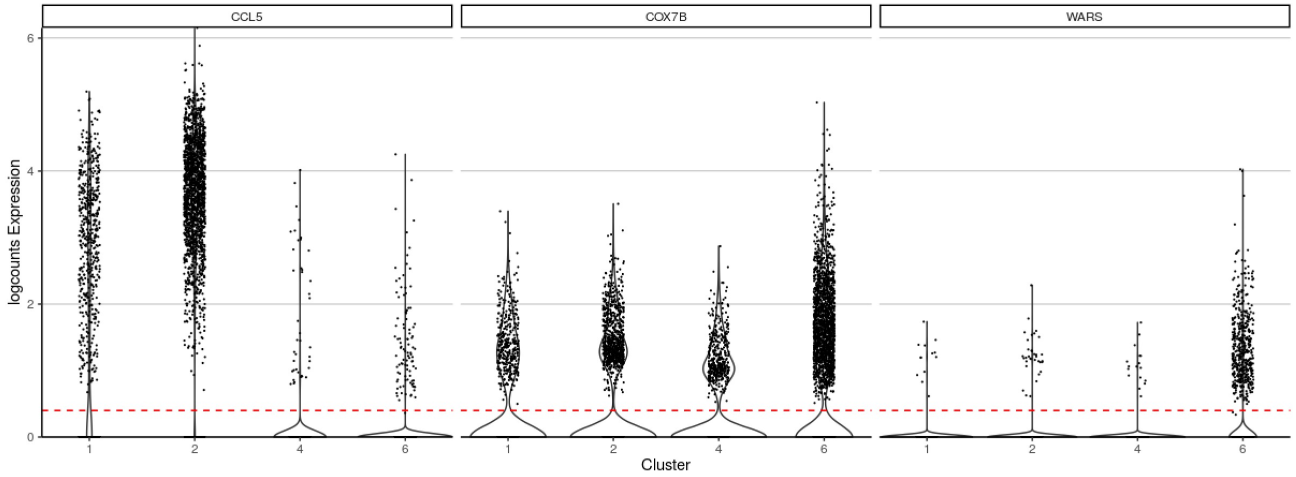
Bimodality (zero-inflation) in scRNA-seq

- Clear bimodality of expression
- Threshold expression values and set the low expression to zero
- **Adaptive:** Gene specific thresholds derived through density estimation
- **Fixed:** log2(TPM+1)>1

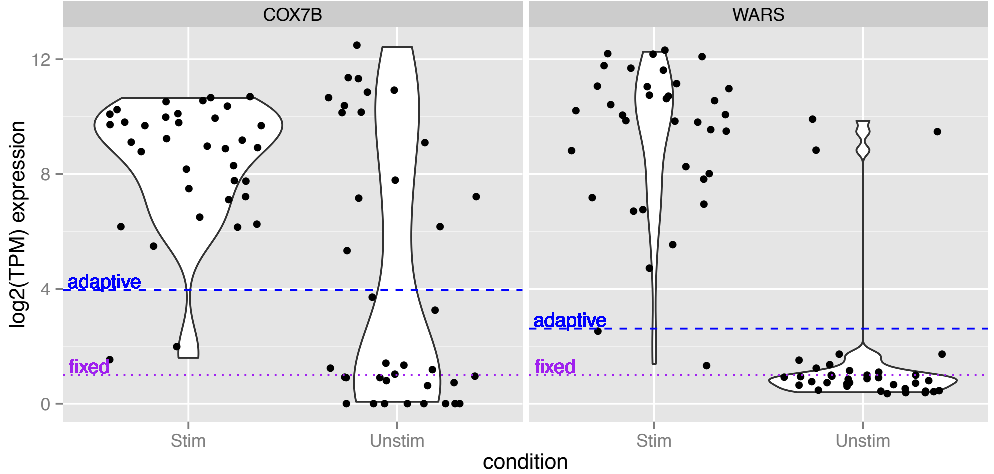
DCs (Shalek et al. 2014)



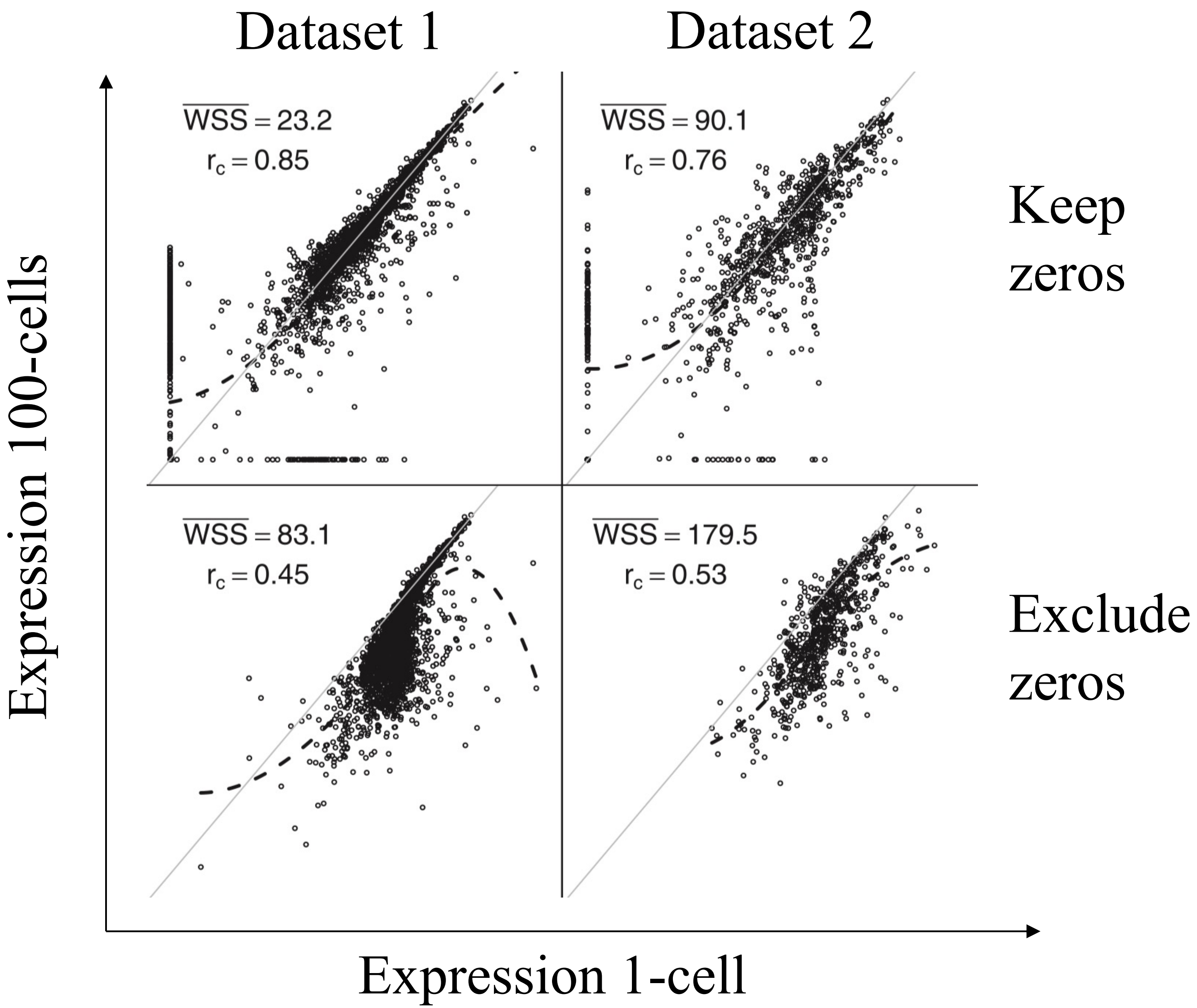
Human Cell Atlas (Bone Marrow, 2019)



MAITs (Finak et al. 2016)

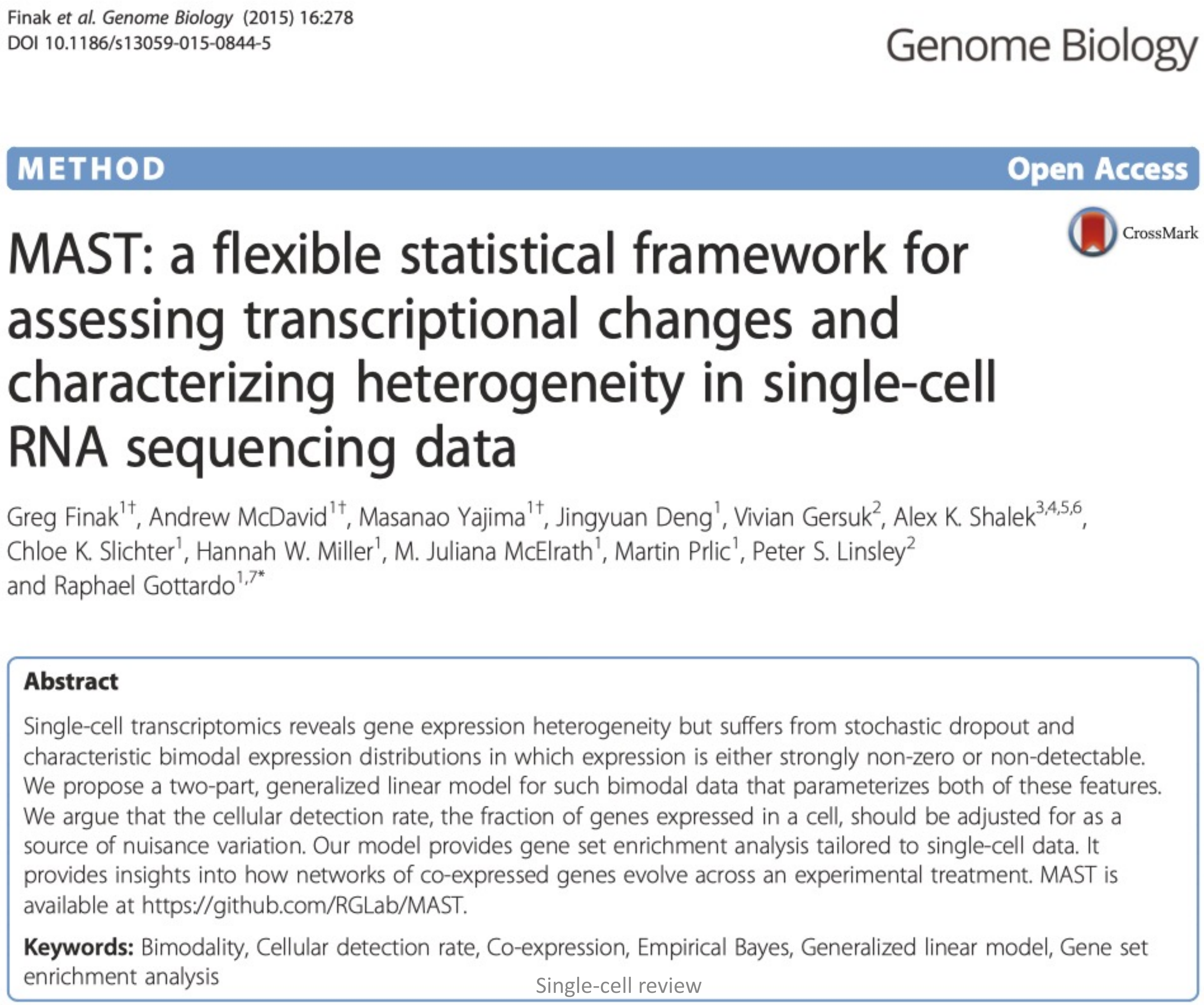


Zeros: real or noise?



- Fluidigm Biomark profiling HIV-specific T-cells
- Dataset from McDavid et al. (2013)
- Zeros reflect true expression values, removing them leads to poor concordance
- Need to model zeros explicitly

Andrew McDavid



Finak *et al. Genome Biology* (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

Abstract

Single-cell transcriptomics reveals gene expression heterogeneity but suffers from stochastic dropout and characteristic bimodal expression distributions in which expression is either strongly non-zero or non-detectable. We propose a two-part, generalized linear model for such bimodal data that parameterizes both of these features. We argue that the cellular detection rate, the fraction of genes expressed in a cell, should be adjusted for as a source of nuisance variation. Our model provides gene set enrichment analysis tailored to single-cell data. It provides insights into how networks of co-expressed genes evolve across an experimental treatment. MAST is available at https://github.com/RGLab/MAST.

Keywords: Bimodality, Cellular detection rate, Co-expression, Empirical Bayes, Generalized linear model, Gene set enrichment analysis

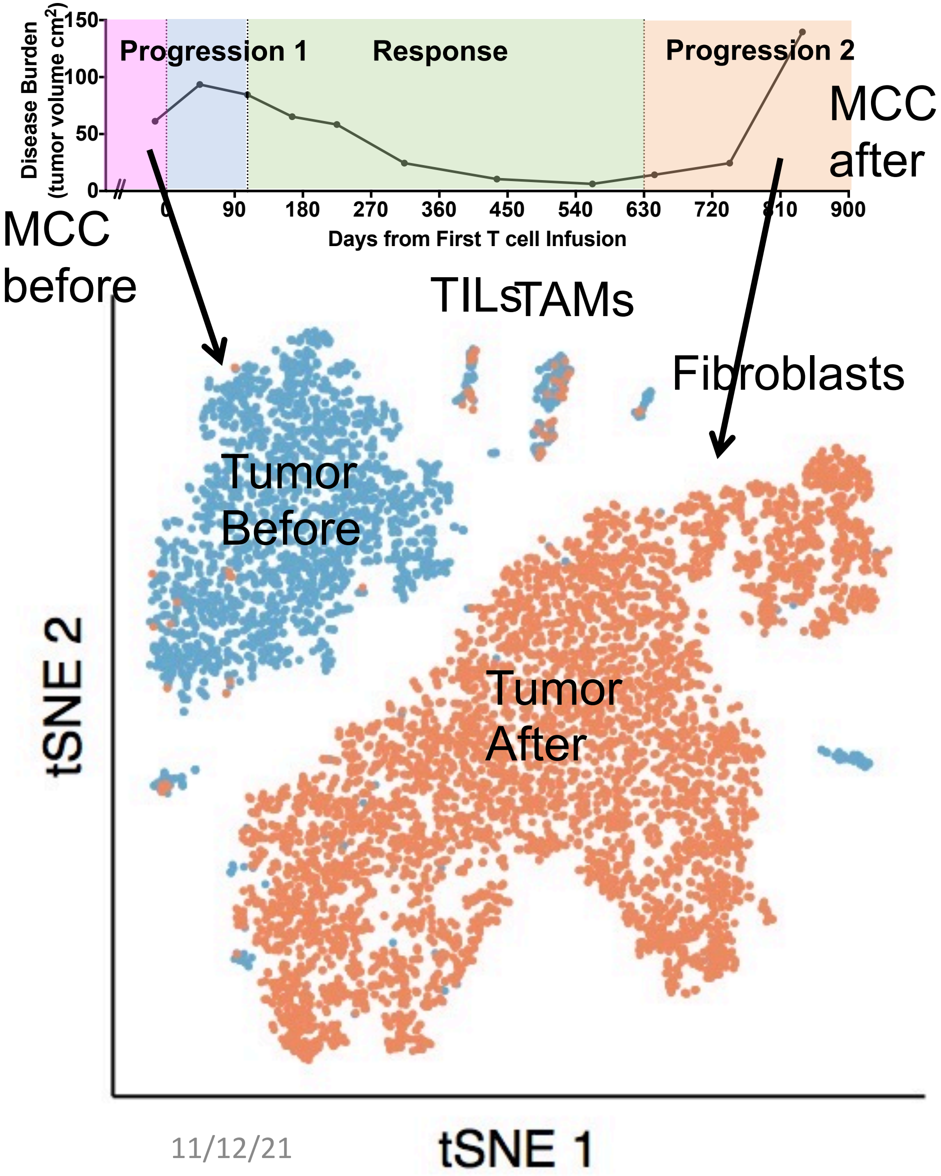
MAST: A unified computational framework

- Model-based Analysis for Single-cell Transcriptomics
  - Support for multiplexed-qPCR, NanoString, and scRNA-seq
  - Thresholding and filtering methodology
  - Semi-continuous model for estimation and inference
  - Flexible framework for modeling effects and covariates (e.g. **CDR**, subjects)
- Gene set enrichment analysis
- Finak et al. *Genome Biology* (2015).
- R package: https://github.com/RGLab/MAST
- Builds on McDavid et al. (2013) and McDavid et al. (2014)

MAST performance

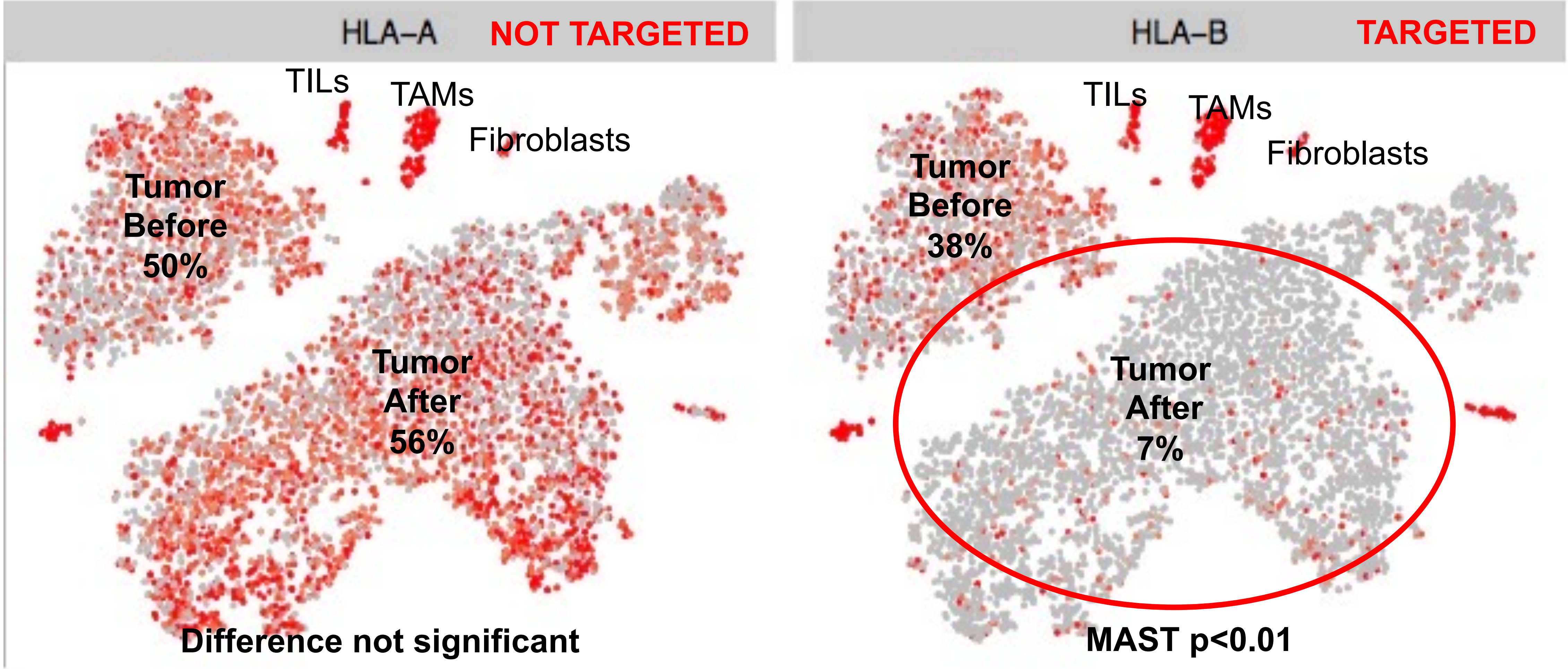
- Among the best tools for differential expression in single-cell RNA-seq
  - Dal Molin, A., Baruzzo, G., Di Camillo, B., 2017. Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. *Front. Genet.* 8, 62.
  - Jaakkola, M.K., Seyednasrollah, F., Mehmood, A., Elo, L.L., 2016. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.*
  - Soneson, C., Robinson, M.D., 2017. Bias, Robustness And Scalability In Differential Expression Analysis Of Single-Cell RNA-Seq Data. *bioRxiv*.
- Good false discovery rate control, good power, fast and applicable to different data types

RNA expression of MCC tumor before treatment and at the time of progression



- TILs (tumor infiltrating lymphocytes), TAMs (tumor associated macrophages) and fibroblasts cluster together
- Tumor cells pre- and post-treatment cluster separately
- All tumor cells changed their gene-expression profile, suggestive of intense selective pressure

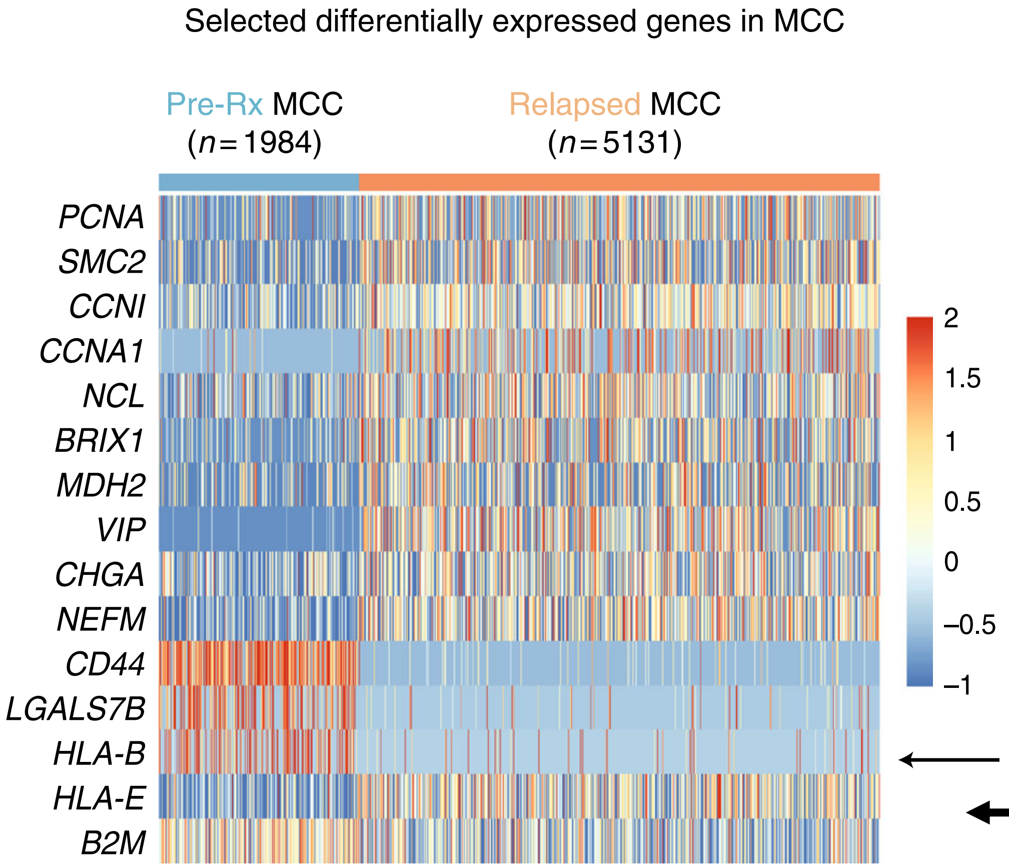
Targeted approach reveals selective immune escape post-treatment with HLA-B35 Targeted T cells



MAST: Finak et al., Genome Biology 2015

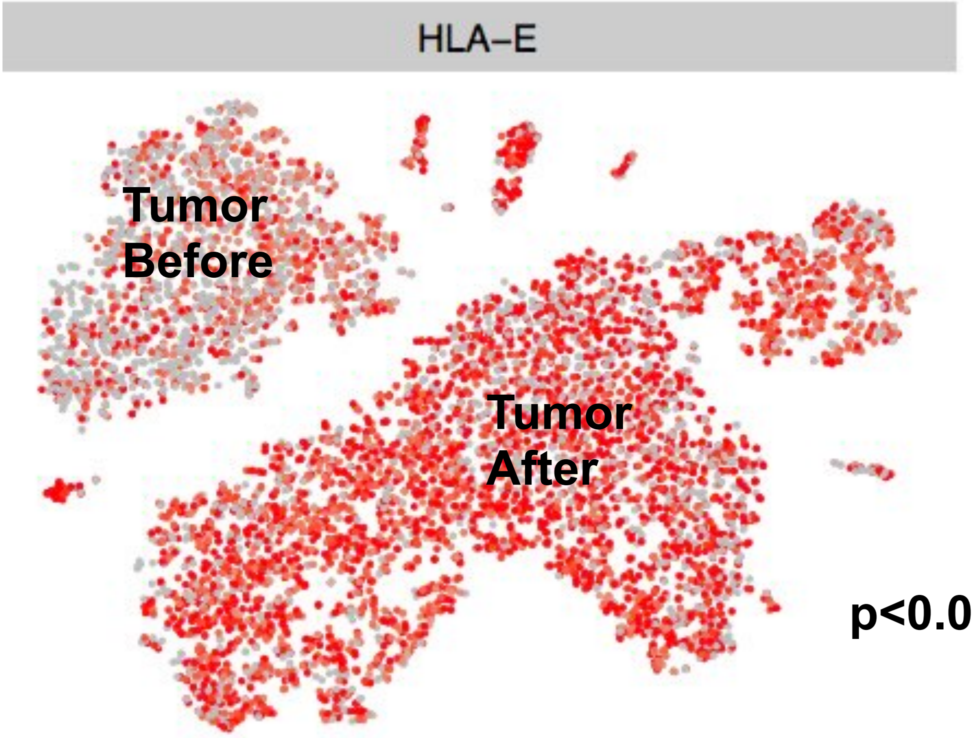
Unbiased approach: significant increase in HLA-E expression

Selected differentially expressed genes



HLA-B

HLA-E

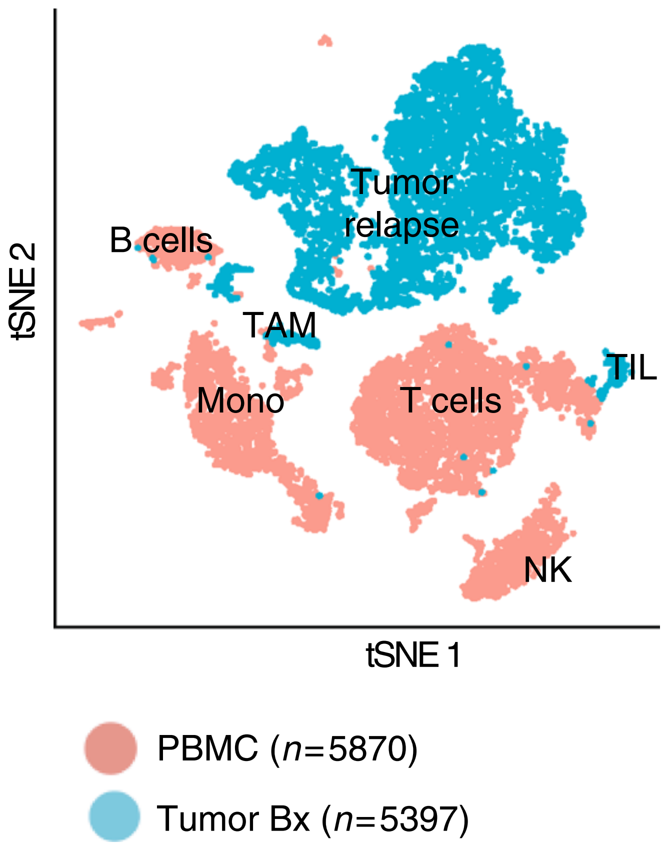


p<0.001

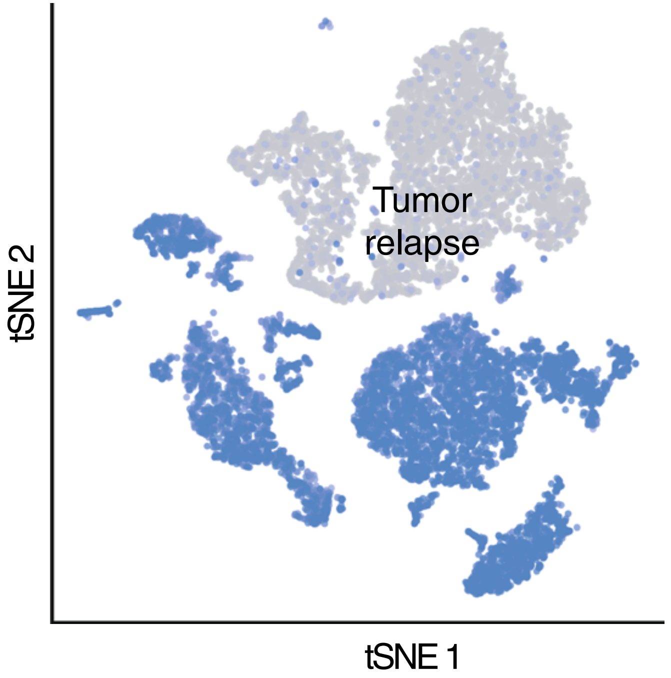
- **HLA-E functions as an inhibitory ligand for NK cells**
- **Patient was subsequently non-responsive to NK boosting strategies**

Validation patient (targeting HLA-A)

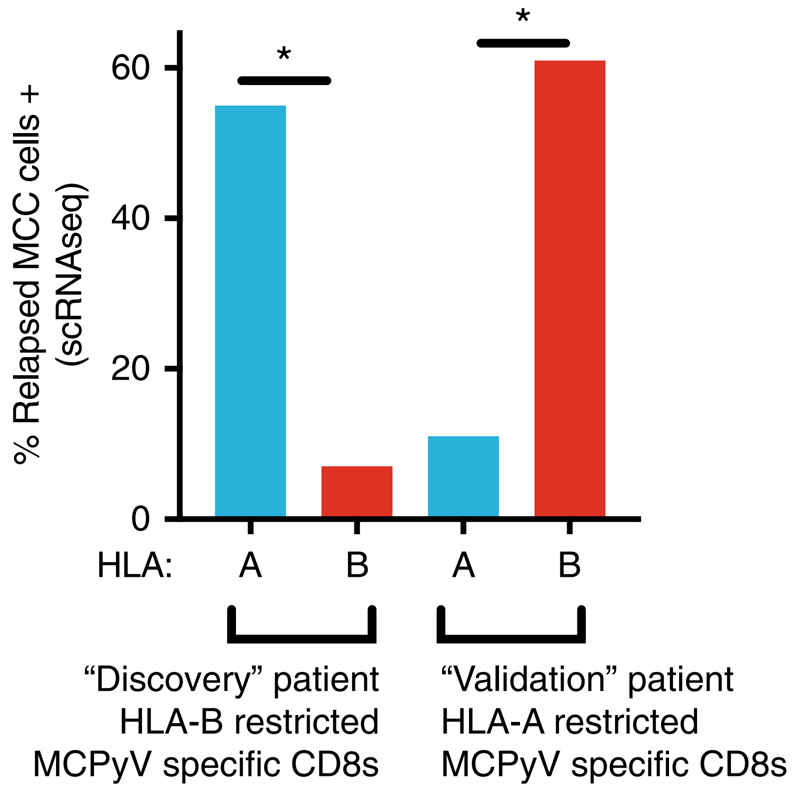
scRNAseq validation patient



Tumor-specific *HLA-A* loss



Differential *HLA-A, -B*



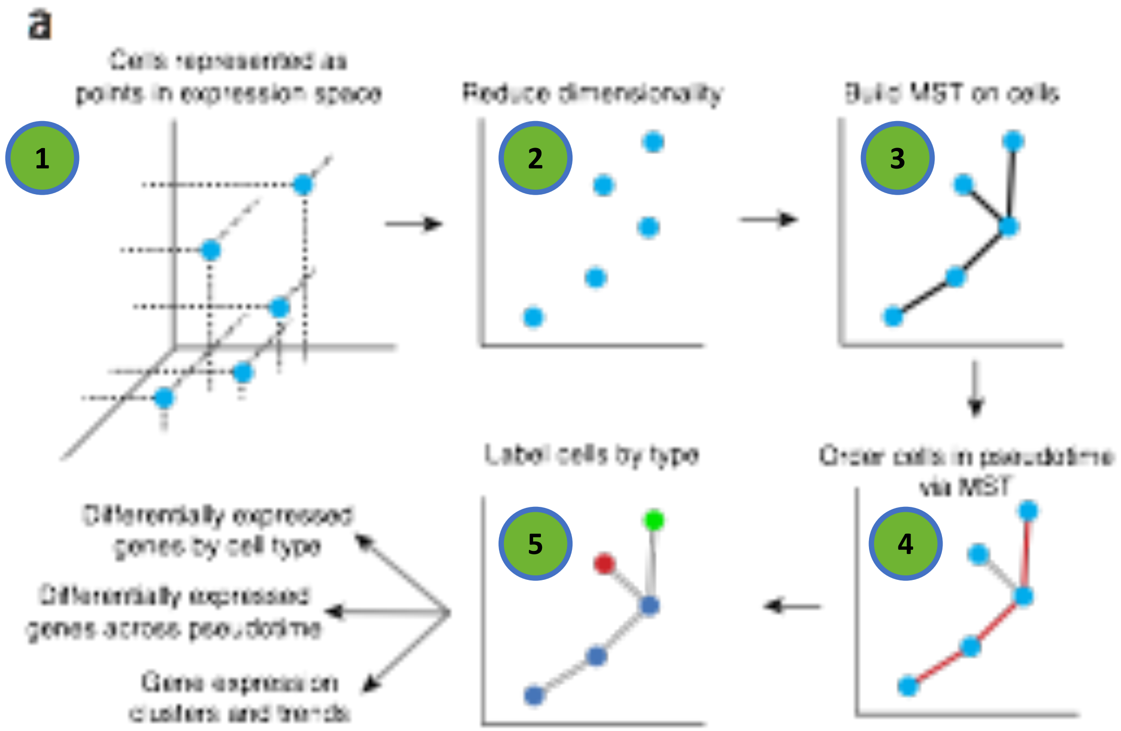
Trajectory analysis

Problem overview

- Pseudotemporal ordering and inference: elucidate the regulatory mechanisms that drive and control changes in gene expression states.
- Challenge of learning temporal dynamics (e.g. state of differentiation) from static measurements.
- Reorder the cells according to their position along a differentiation path.
- Several tools have been recently developed (Review in Babtie et al., 2017)
  - Monocle 2 (Qiu et al., 2017)
  - TSCAN (Ji et al., 2016)
  - RNA velocity (Manno et al. *Nature* 2018).

Trajectory analysis with Monocle

- Monocle - Trapnell et al., 2014 - The original Monocle paper, which introduced the **concept of pseudotime ordering for single-cell analysis**


a
Cells represented as points in expression space
1
Reduce dimensionality
2
Build MST on cells
3
Order cells in pseudotime via MST
4
Label cells by type
5
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends

Pseudotime is a quantitative measure of biological progression through a process such as cell differentiation.

1 Expression profile of each cell as a point in a high-dimensional Euclidean space - one dimension for each gene.

2 Dimensionality reduction - transforms the cell data from a high-dimensional space into a low-dimensional one that preserves essential relationships between cell populations but is much easier to visualize and interpret.

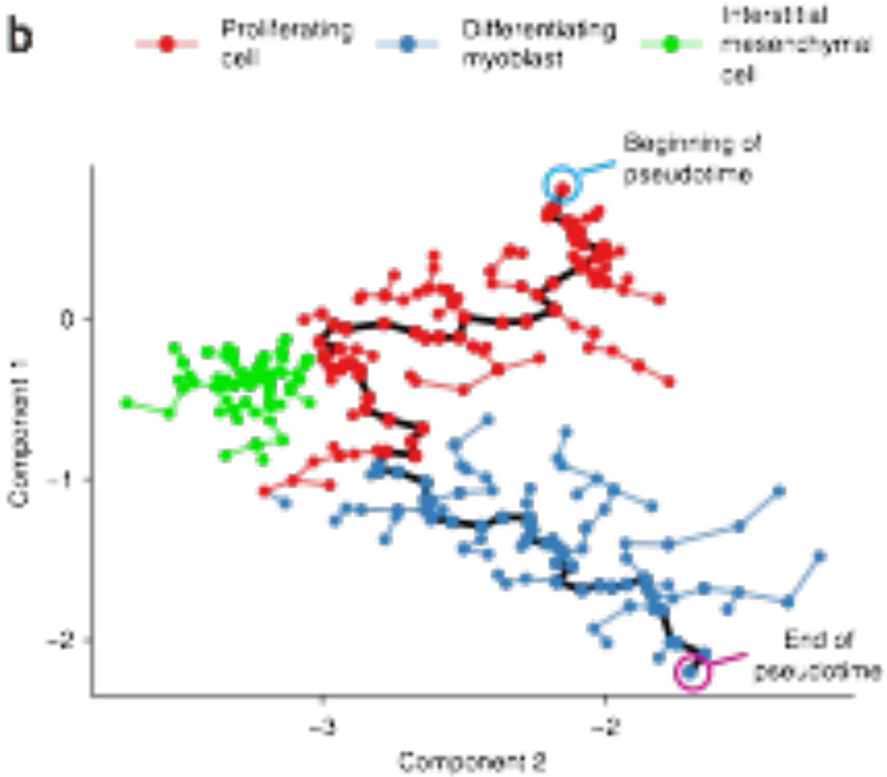
3 Construction of a Minimum Spanning Tree (MST) on the cells (connects all the vertices (cells) together, without any cycles and with the minimum possible total edge weight) - already used w/ flow or mass cytometry.

4 Find the longest path through the MST, corresponding to the longest sequence of transcriptionally similar cells.

5 Use this sequence to produce a trajectory.

Trajectory analysis with Monocle

- Monocle - Trapnell et al., 2014 - The original Monocle paper, which introduced the **concept of pseudotime ordering for single-cell analysis**

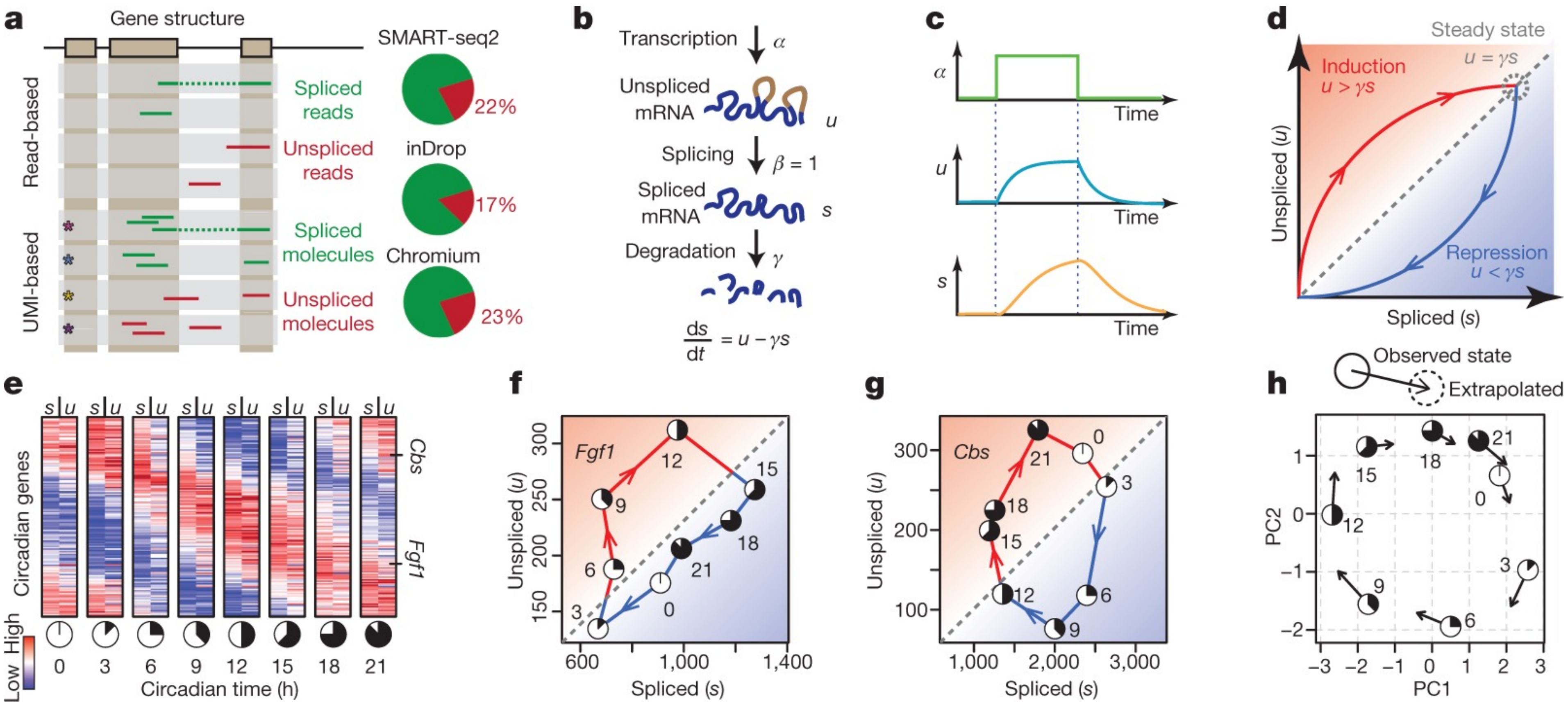


Primary human myoblast cultured in high-serum medium. After a switch to low-serum medium, cells were dissociated and individually captured at 24h intervals (0h, 24h, 48h and 72h). mRNA sequencing: C1 Fluidigm.

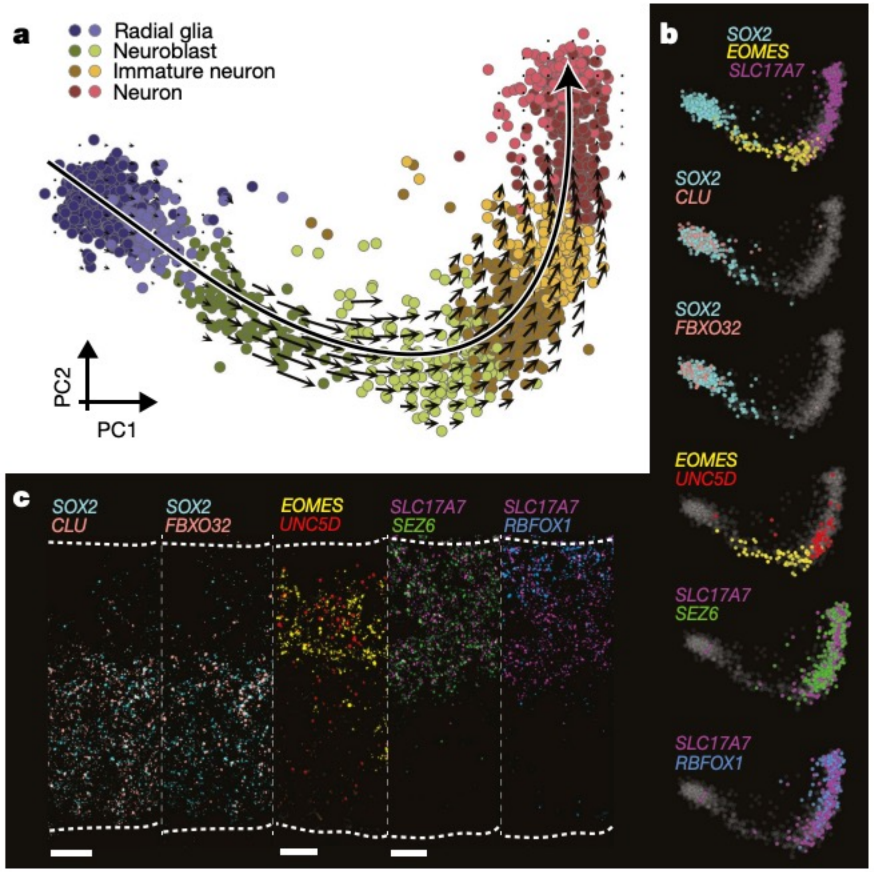
Monocle decomposed myoblast differentiation into a **two-phase trajectory** and **isolated a branch of non-differentiating cells**.
The first phase of the trajectory was primarily composed of cells collected under high-mitogen conditions (**proliferating cells**). Cells in the second phase were positive for markers of muscle differentiation (**differentiating myoblasts**). A tightly grouped third population of cells branched from the trajectory near the transition between phases. These cells lacked myogenic markers, but expressed PDGFRA and SPHK1, suggesting that they are contaminating **interstitial mesenchymal cells** and did not arise from the myoblasts.

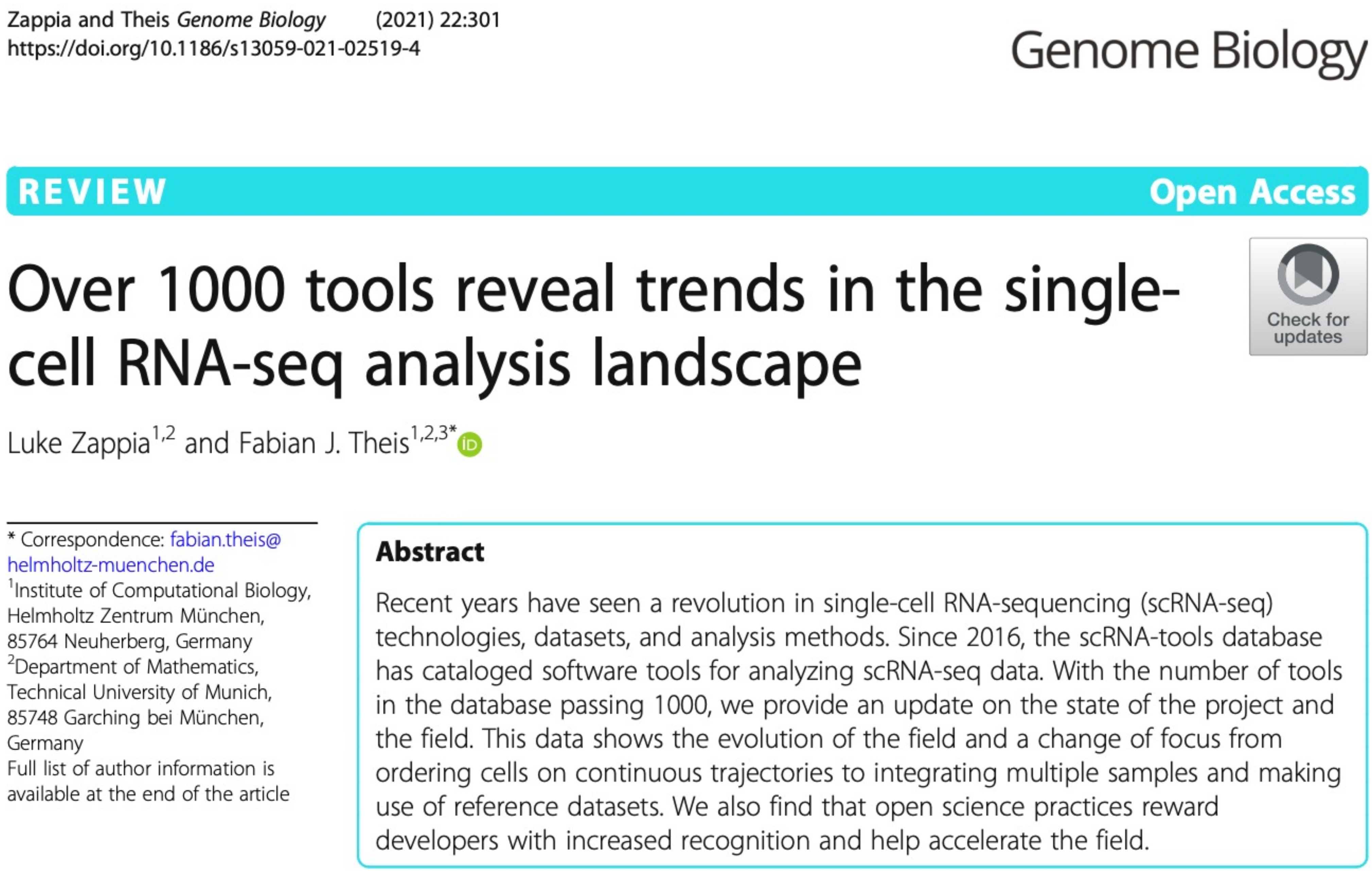
New version of Monocle - Monocle 2 - applies reversed graph embedding (RGE), a recently developed machine learning strategy, to accurately reconstruct complex single-cell trajectories.
Works on every data type (UMIs, TPM, etc.)

Trajectory analysis with RNA velocity



Kinetics of transcription during human embryonic glutamatergic neurogenesis





Zappia and Theis *Genome Biology* (2021) 22:301
https://doi.org/10.1186/s13059-021-02519-4

Genome Biology

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*]

* Correspondence: fabian.theis@helmholtz-muenchen.de
[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Department of Mathematics, Technical University of Munich, 85748 Garching bei München, Germany
Full list of author information is available at the end of the article

Abstract

Recent years have seen a revolution in single-cell RNA-sequencing (scRNA-seq) technologies, datasets, and analysis methods. Since 2016, the scRNA-tools database has cataloged software tools for analyzing scRNA-seq data. With the number of tools in the database passing 1000, we provide an update on the state of the project and the field. This data shows the evolution of the field and a change of focus from ordering cells on continuous trajectories to integrating multiple samples and making use of reference datasets. We also find that open science practices reward developers with increased recognition and help accelerate the field.

Spatial gene expression: The next frontier

Bulk RNA-seq
(~2008)

Single-cell RNA-seq
(~2014)

Spatial transcriptomics
(2016)