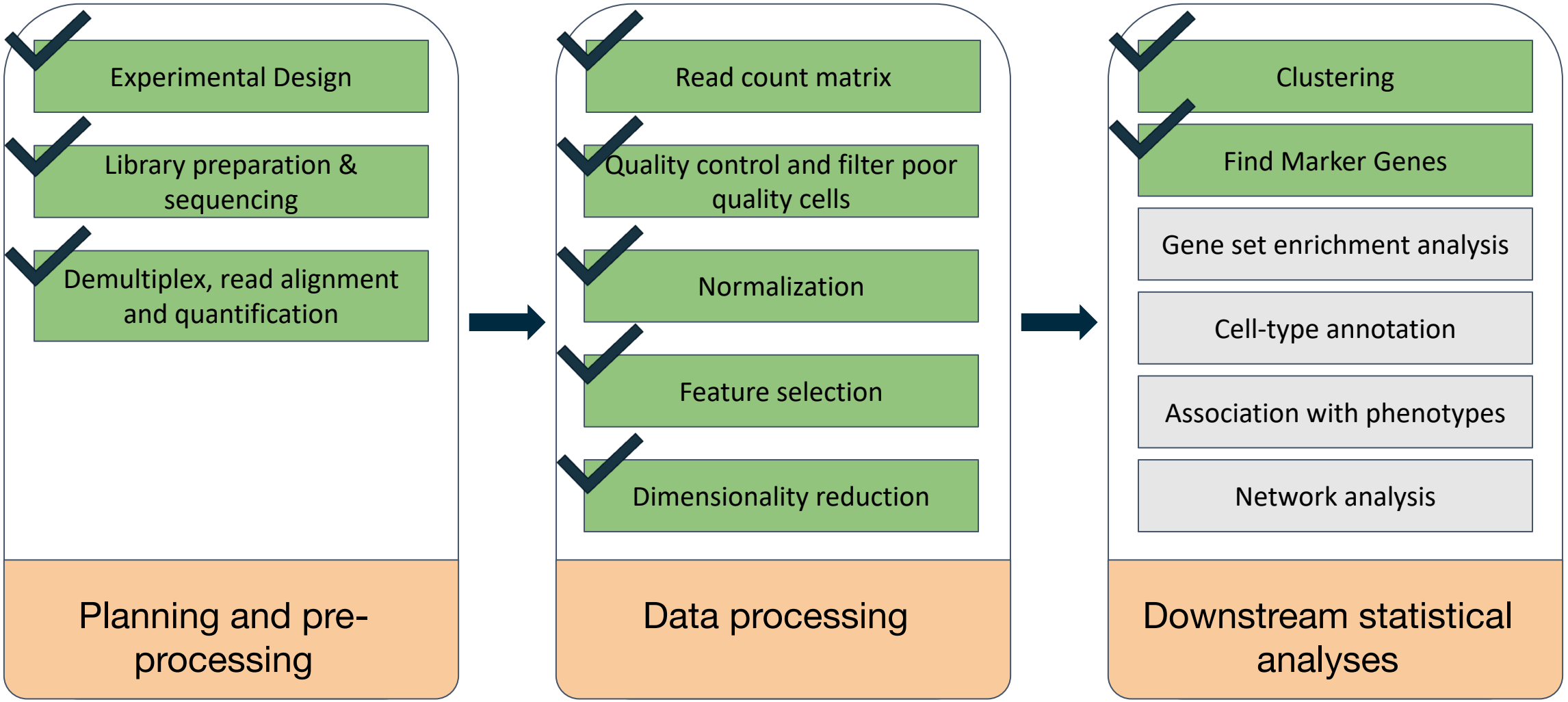
Session 3: Single-cell RNA-seq
Normalization, Differential Expression and Batch Correction

Reuben Thomas
Gladstone Bioinformatics Core
Jan 27th, 2023

scRNA-seq workflow



Typical scientific questions

- **Identify the cell types** in your sample of interest
- **Identify which cell types change** under condition of interest
- **Identify genes whose expression** is modulated in relevant cell types

Outline for this session

- **Main messages**
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- Normalization (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

Main points I want to convey

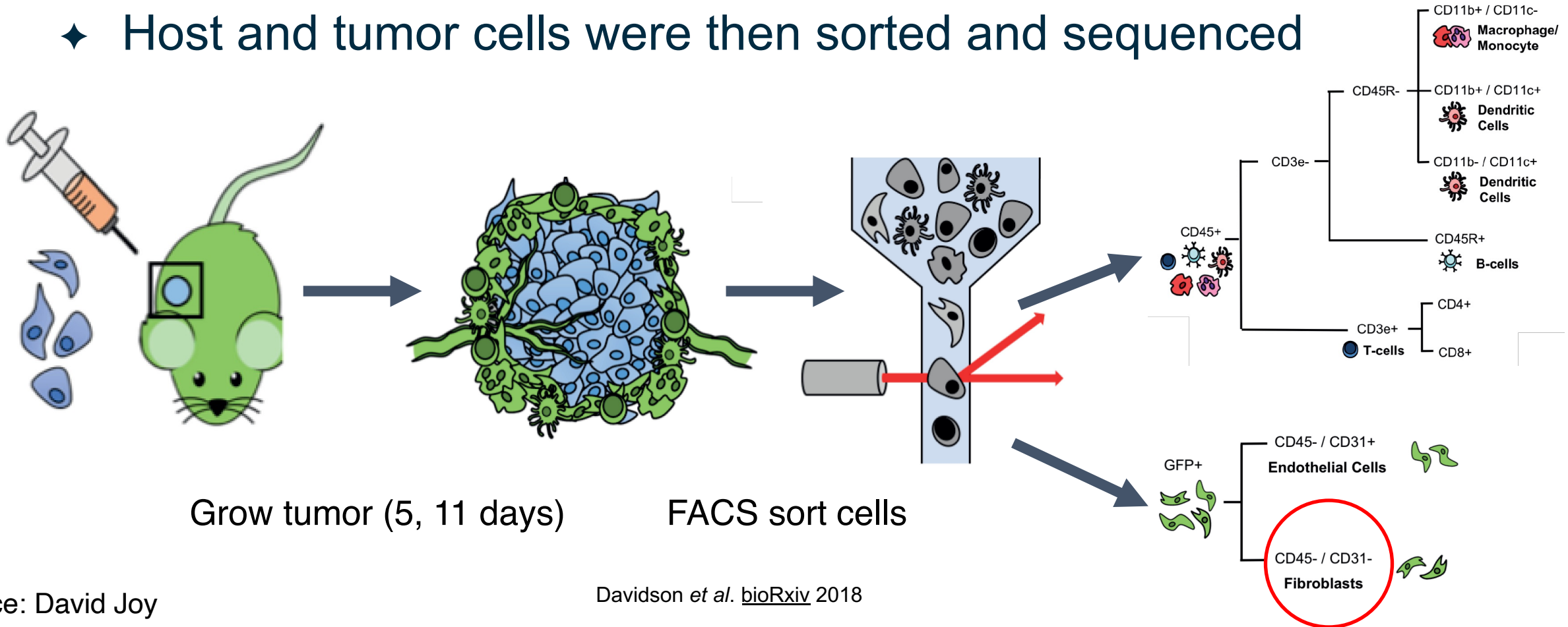
- State of art of the **design and statistical analyses of scRNAseq data is still in flux**
  - Benchmarks in different areas are becoming increasingly available
- There are **better designs**
  - **Most important:** Please include replicates drawn from the population you want to make a claim about
- There are **better statistics**/ways to get “more” reproducible results.
  - Normalization
  - Identification of marker genes
  - Multi-sample multi-condition comparison
  - Batch correction

Outline for this session

- Main messages
- **Introduce the data and the research questions (5 min)**
- Primer on statistics (5 min)
- Normalization (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

Dissecting the Tumor Microenvironment

- GFP+ Mice were injected with a melanoma cell line
- Host and tumor cells were then sorted and sequenced

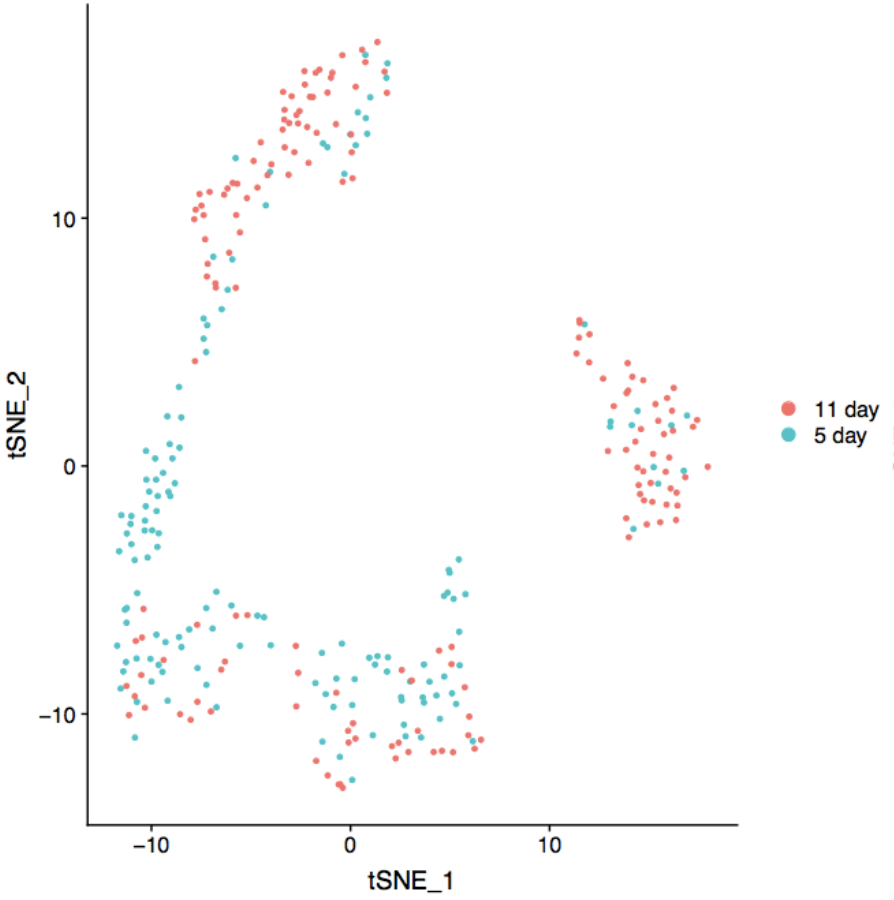


Source: David Joy

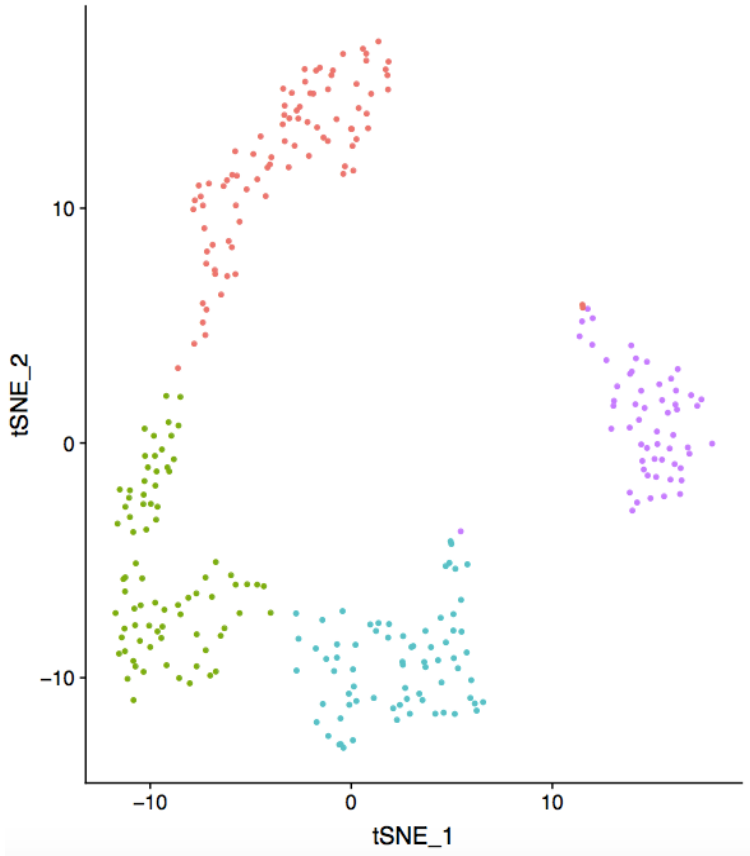
Davidson *et al*. bioRxiv 2018

Fibroblast cells over **Time, Clusters and Individuals**

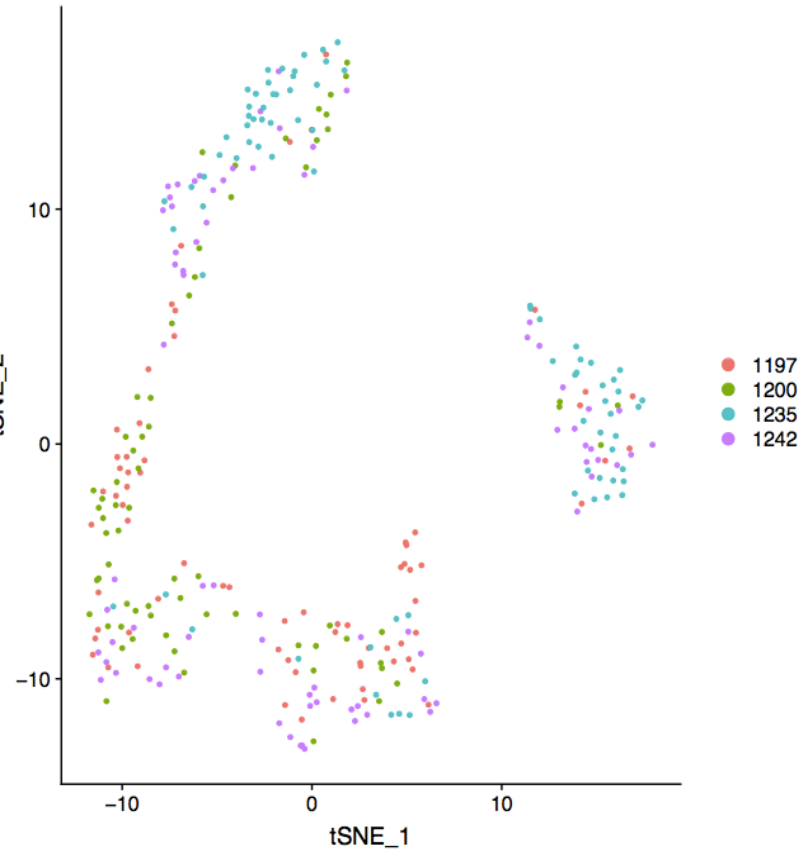
Time



Cluster



Individual



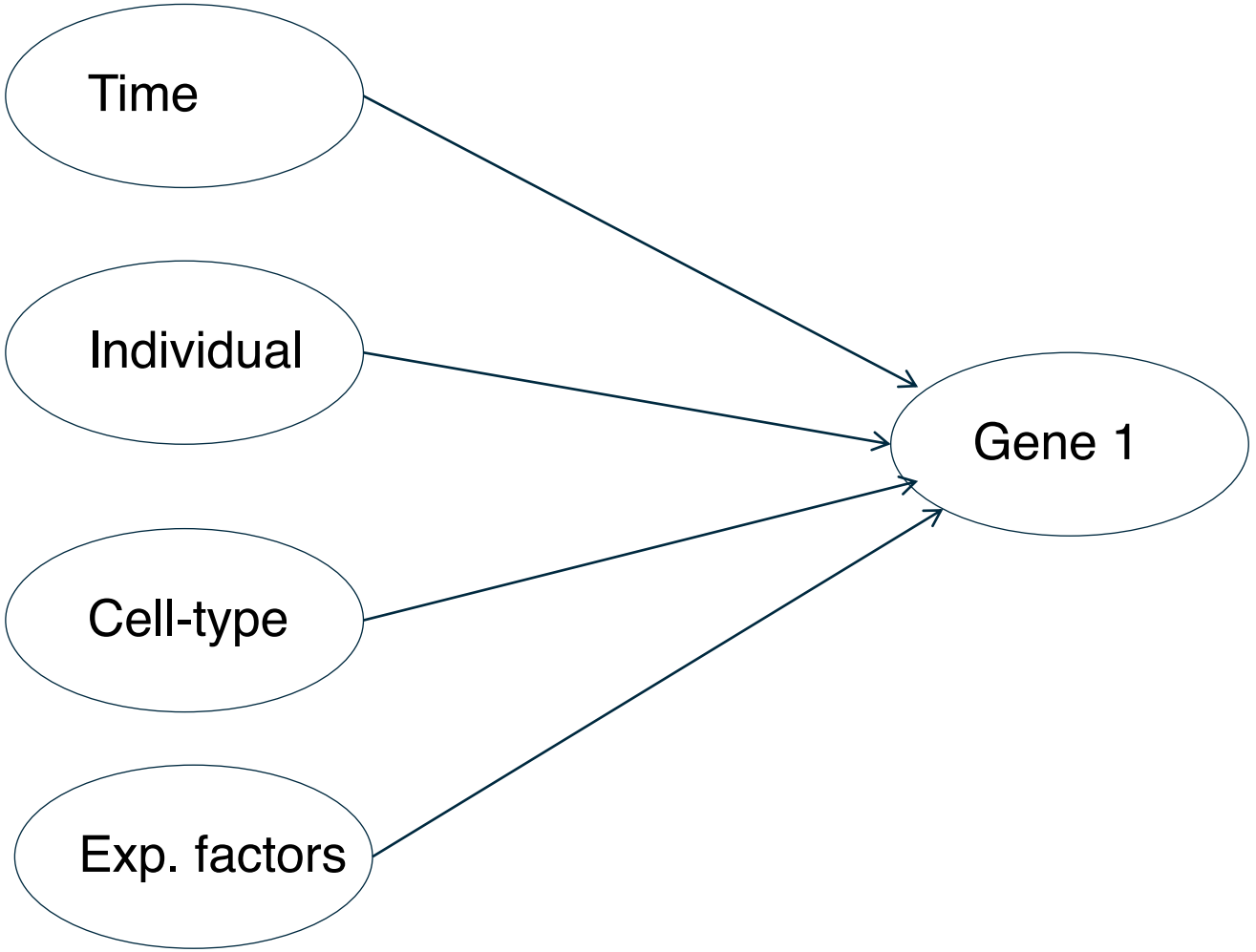
Research Questions

- **Genes associated with time:** which genes change their expression in the host fibroblast cells from the 5 day time-point to the 11 day time-point?
- Would like to take into account inter-individual variability – *make generalizable claims*
- Would make claims specific to particular clusters or implied cell-types to this population of cells – *after all we have scRNA-seq data, ☺.*
- Would like to avoid the multiple testing problem and make statistically rigorous claims

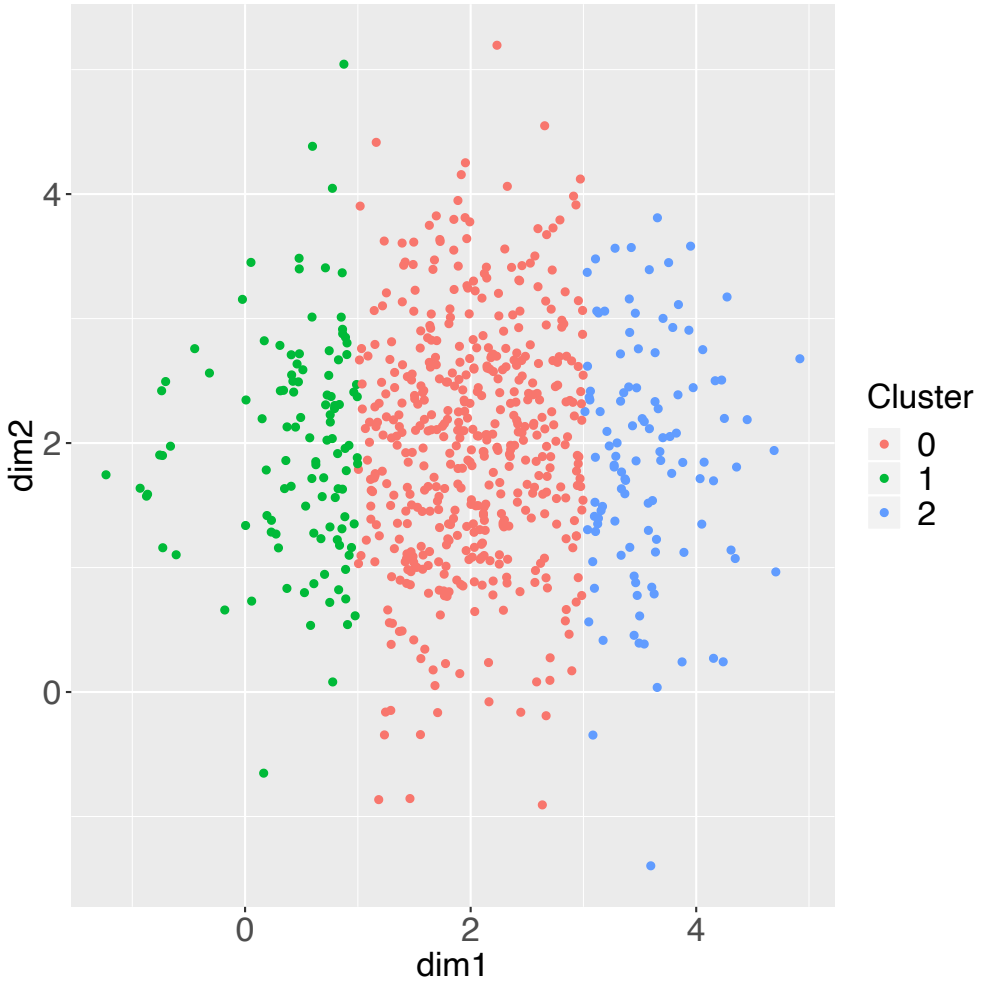
Outline for this session

- Main messages
- Introduce the data and the research questions (5 min)
- **Primer on statistics (5 min)**
- Normalization (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

Predictors of gene expression in a given cell



We could use cluster as surrogate for cell-type

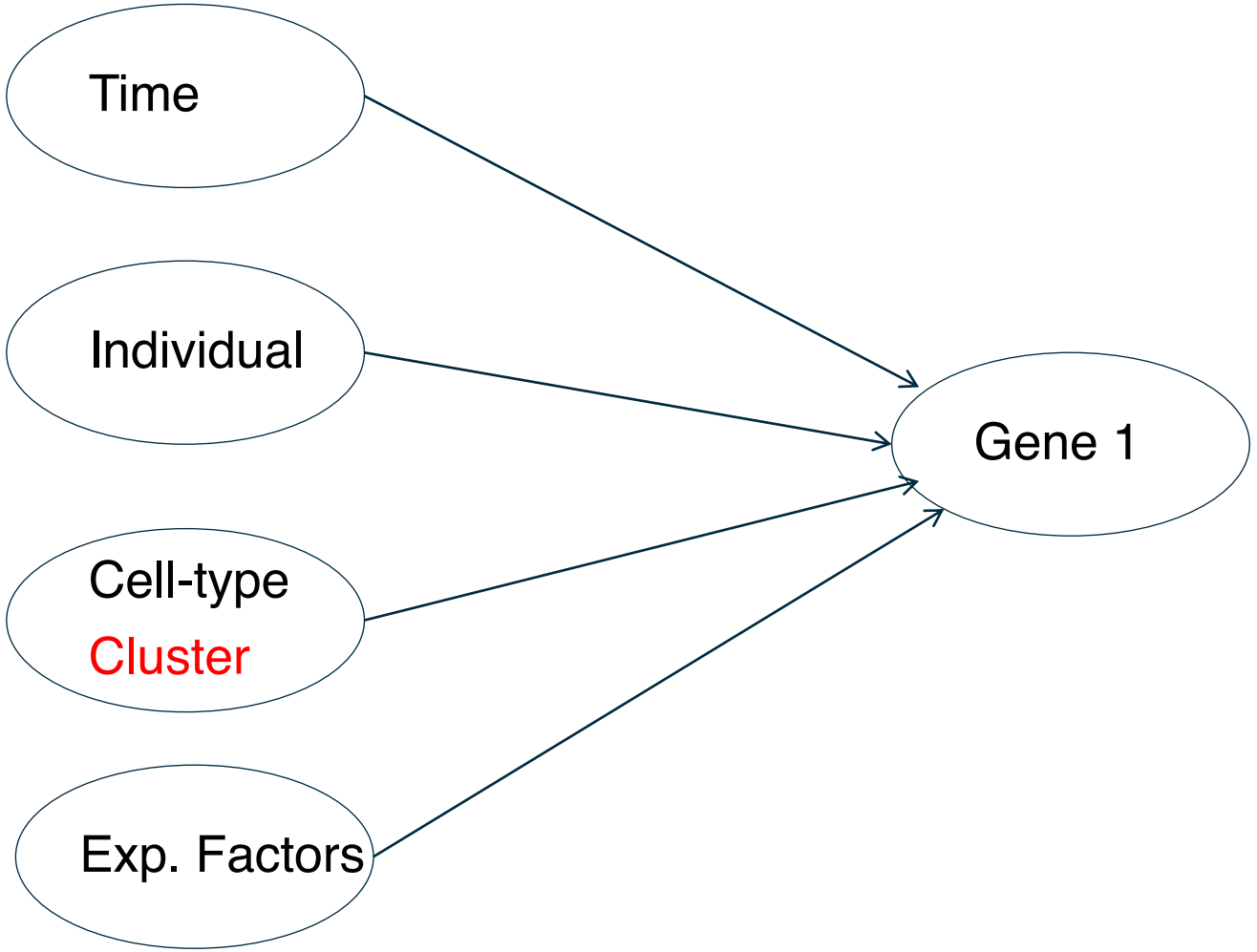


Cell-type ~ cluster

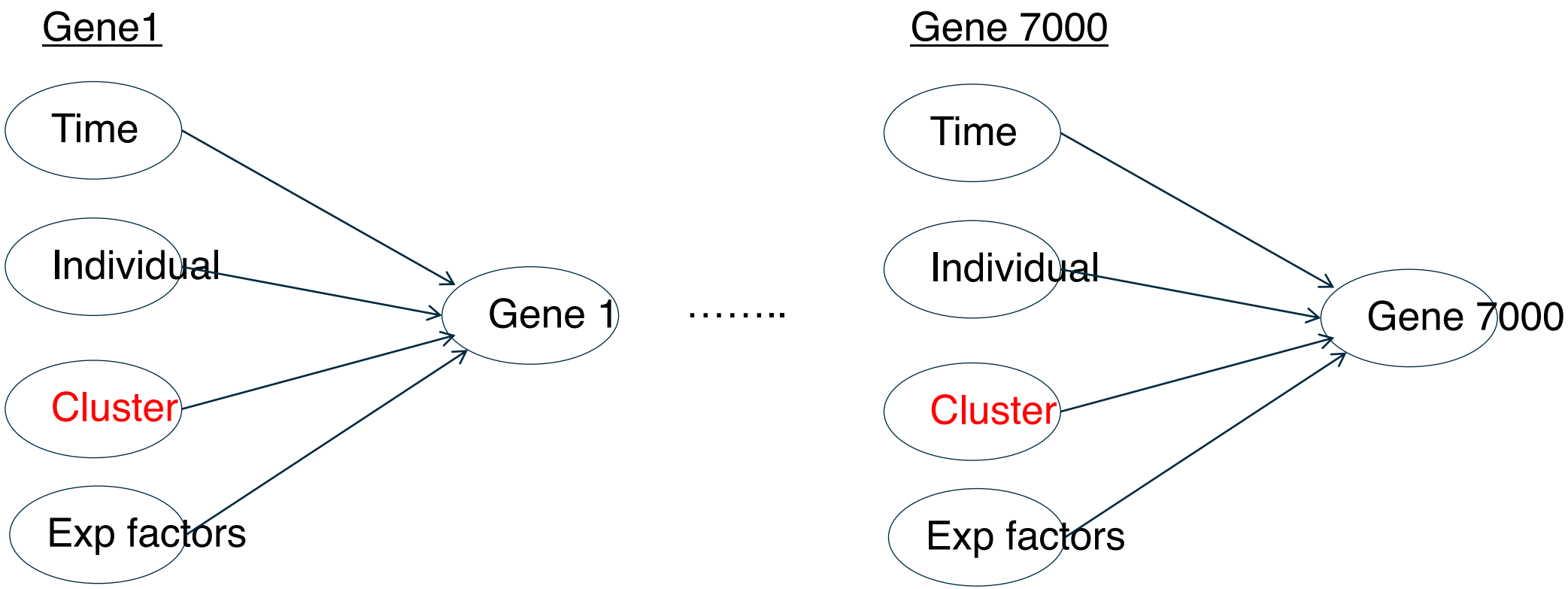
WARNING:

Data dredging: *use the same data to define clusters of cells and differences in gene expression between these clusters*

Predictors of gene expression

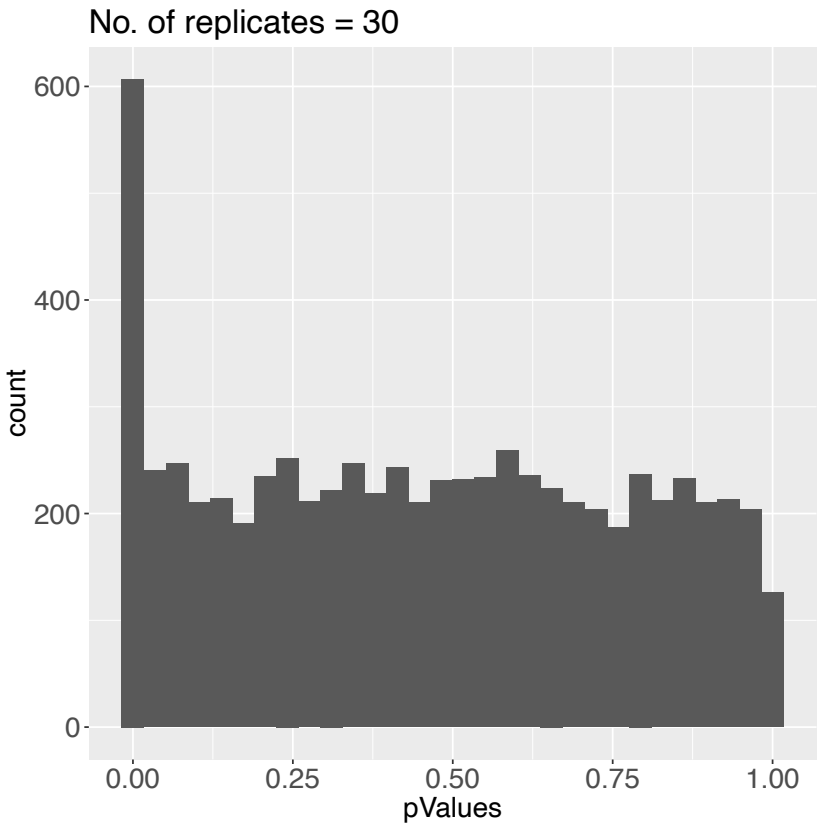


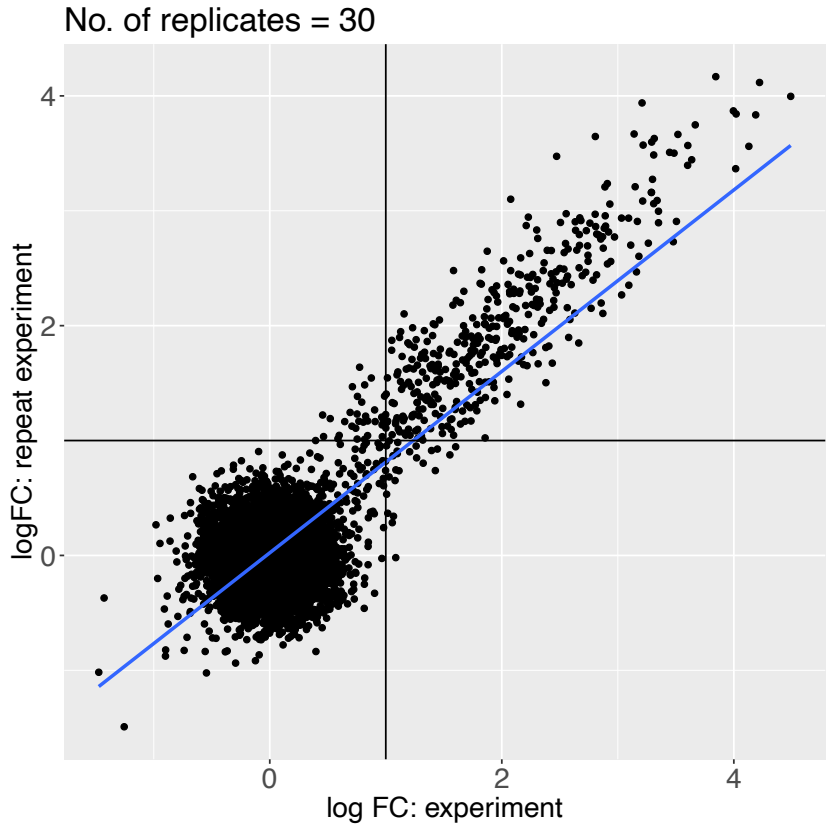
Predictors of gene expression



Distribution of p-values across all genes and reproducibility***(simulated data)***

We would like for our claims to reproducible across repeated experiments!





Log odds ratio of the same gene having a logFC > 1 in two independent experiments = 8

Outline for this session

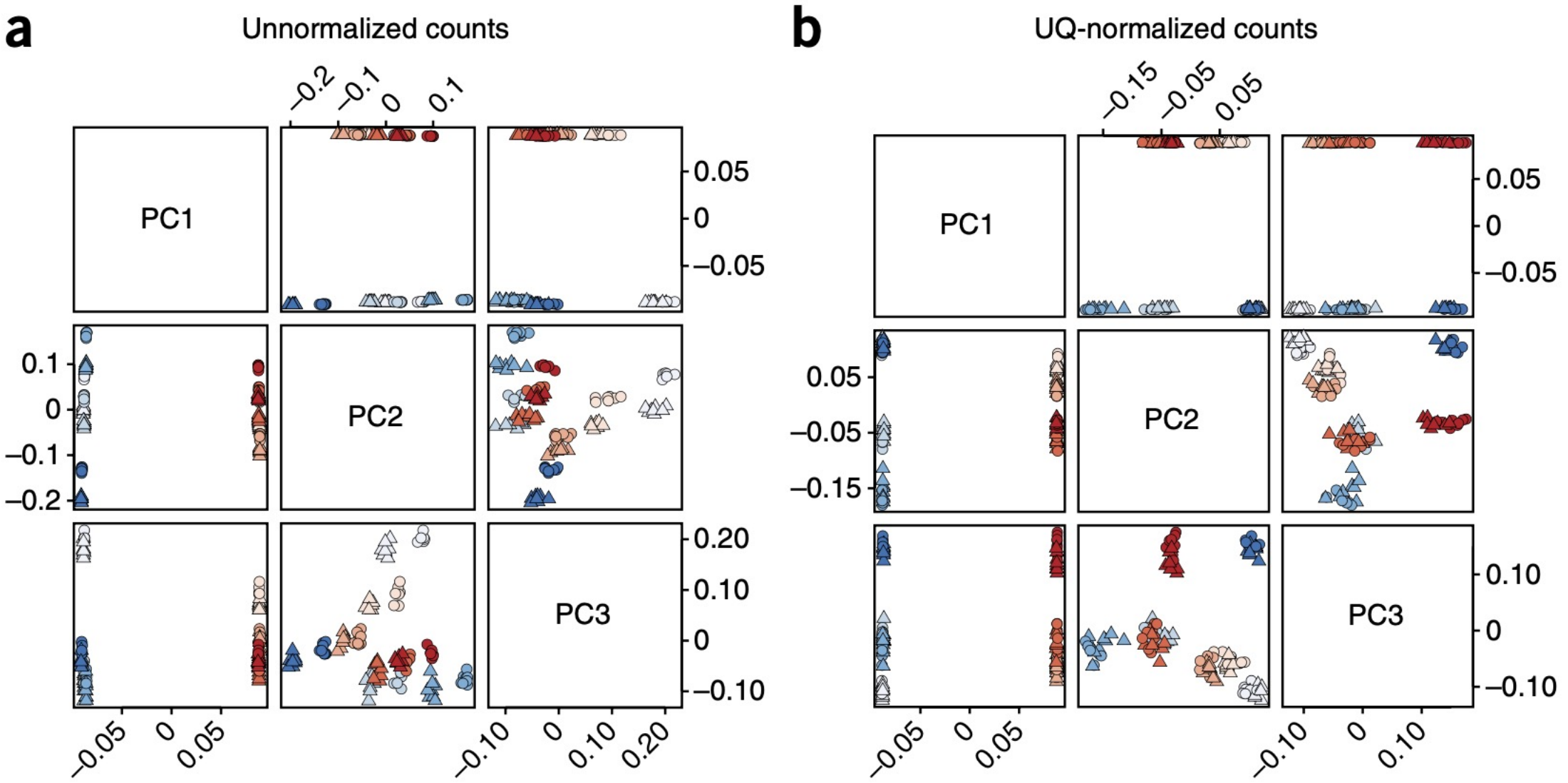
- Main messages
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- **Normalization** (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

Why all this fuss about normalization and batch correction?

- We would like to minimize variation/bias in gene expression due to technical and unwanted biological sources
- Significance of conclusions will be questioned if one does perform these steps
- RNA-seq counts are representative of RELATIVE and NOT ABSOLUTE levels of gene expression
  - Covid-19 sample1 in USA has body temperature 102 while Covid-19 sample2 in UK has body temperature 39.
  - Gata4 gene has 103 read counts in replicate 1 at E9.5, has 576 read counts in replicate 3 at E11.5

Normalization vs batch effects

- **Normalization** aims to remove/reduce technical sources of variation
- Technical sources: *sequencing depth*, *cDNA capture* or *PCR amplification efficiency* across cells
- Technical sources tend to affect all genes
- **Batch effects** aims to remove both technical and biological sources of variations arising due to processing in different batches
- Batch effects may not affect all genes equally



Red/Blue: different biological conditions

Shade of color: library prep date

Circles and Triangles: two different flow cells

Why normalize for scRNA-seq

- Effects performance in all essential questions asked of data
  - Cell-type/cluster identification
  - Low-dimensional representation
  - Differential expression analyses
  - Trajectory analysis

Four main classes of normalization

1. **Library-size normalization**: *LogNormalize* in Seurat
2. **Deconvolution normalization**: *scran* Bioconductor package
3. **Spike-in normalization**
4. **Variance stabilizing transform-based normalization**: *sctransform* in Seurat

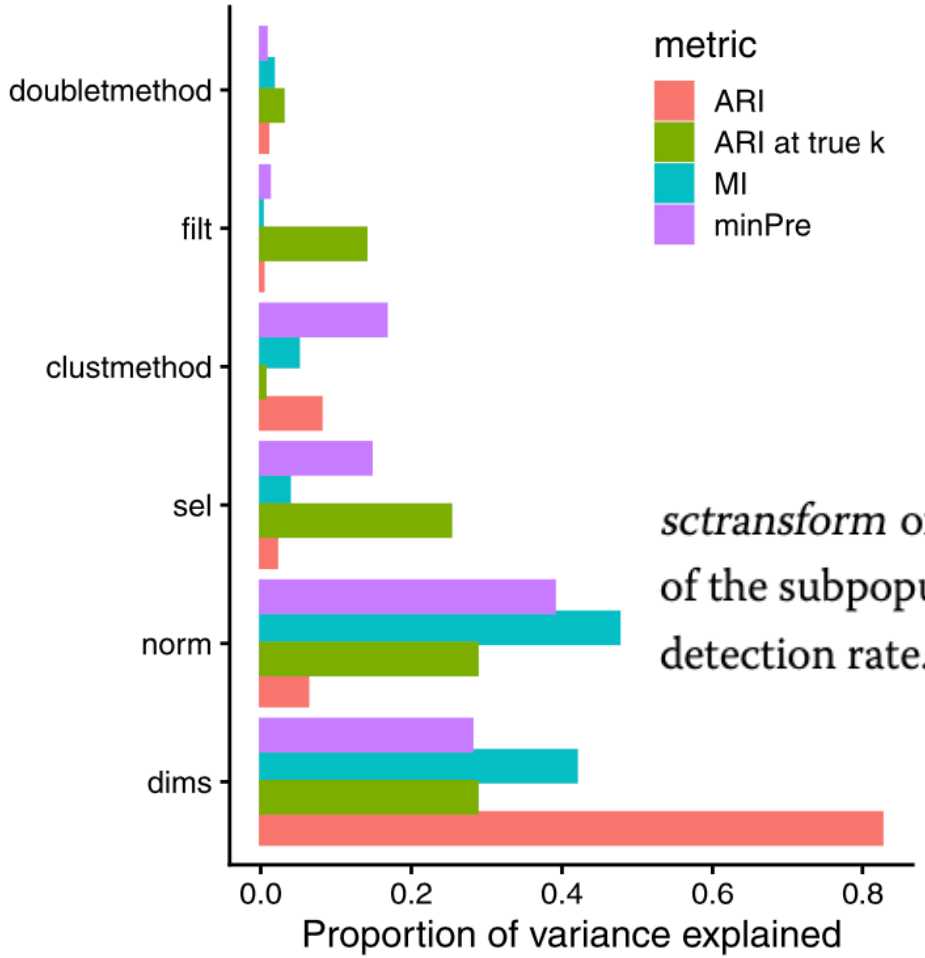
Germain *et al. Genome Biology* (2020) 21:227
https://doi.org/10.1186/s13059-020-02136-7

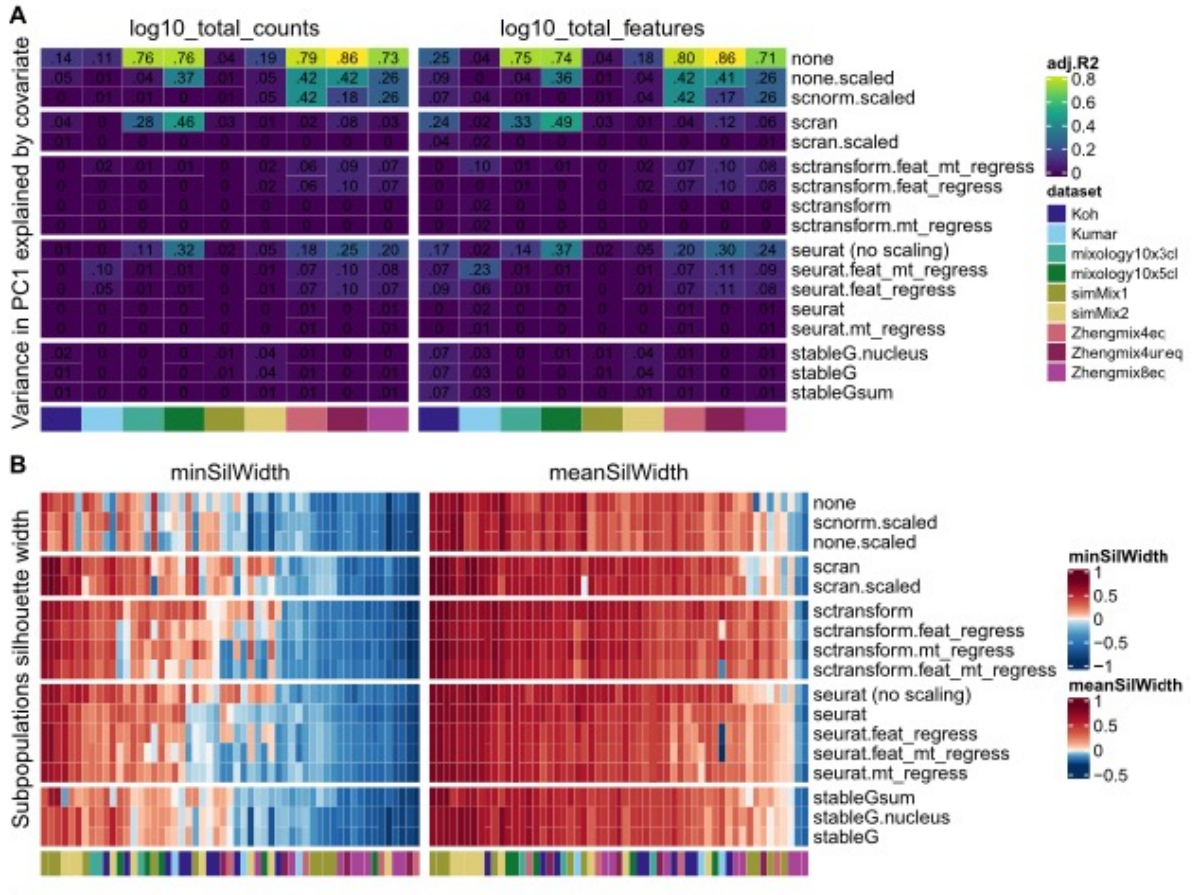
Genome Biology

METHOD Open Access

pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools

Pierre-Luc Germain[1,2,3*], Anthony Sonrel[1,2] and Mark D. Robinson[1,2*]





sctransform offered the best overall performance in terms of the separability of the subpopulations, as well as removing the effect of library size and detection rate.

- Deviance [40] offered the best ranking of genes for feature selection.
- Increasing the number of features included tended to lead to better classifications, plateauing from 4000 features in our datasets.

Outline for this session

- Main messages
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- Normalization (10 min)
- **Finding marker genes (5 min)**
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

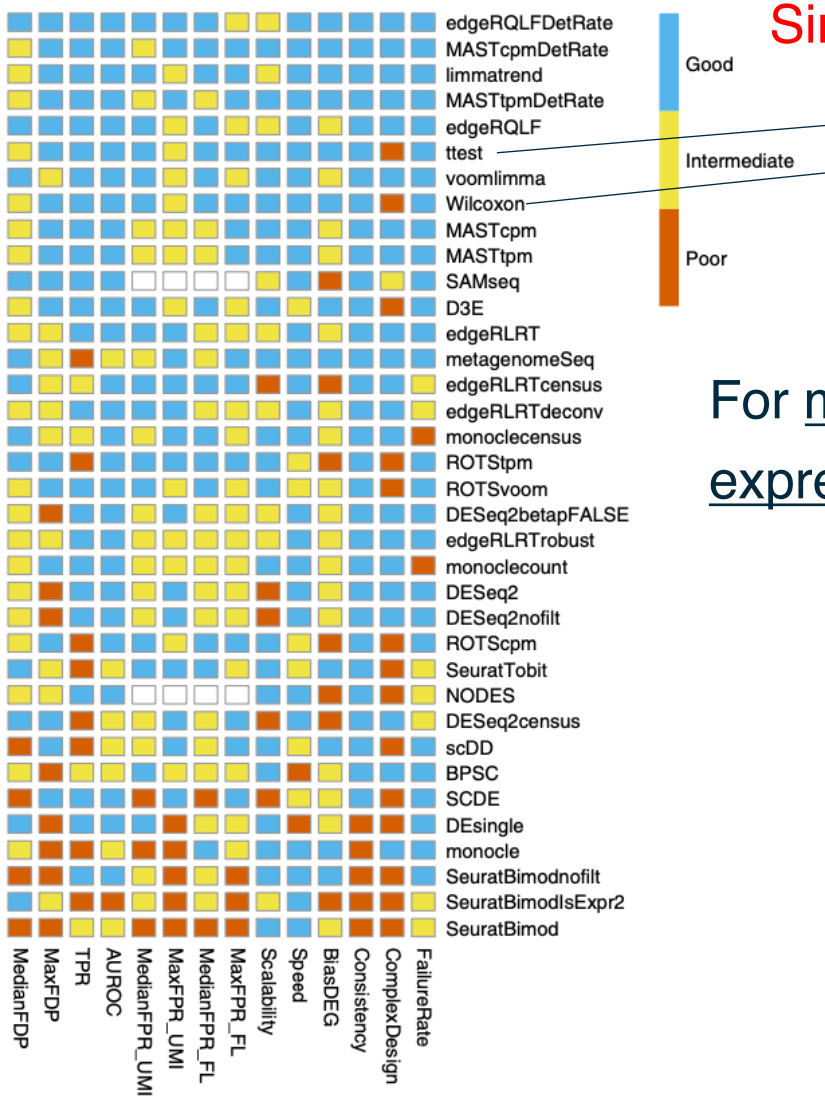
Cluster identification: marker genes

- Identify genes whose expressions define particular clusters
- Compare the distribution of normalized expression of each gene among cells in given cluster versus cells from all other clusters
- Two sample t-test, Wilcoxon tests?
- More sophistical methods?

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

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

NATURE METHODS | VOL.15 NO.4 | APRIL 2018 | 255



Simple methods like the two-sample t-test and Wilcoxon tests do well!!

For more sophisticated methods like edgeR, MAST – filtering of low expressed genes very important for good performance

Outline for this session

- Main messages
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- Normalization (10 min)
- Finding marker genes (5 min)
- **Batch correction (15 min)**
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- Reiterate the main messages

Define batch effects

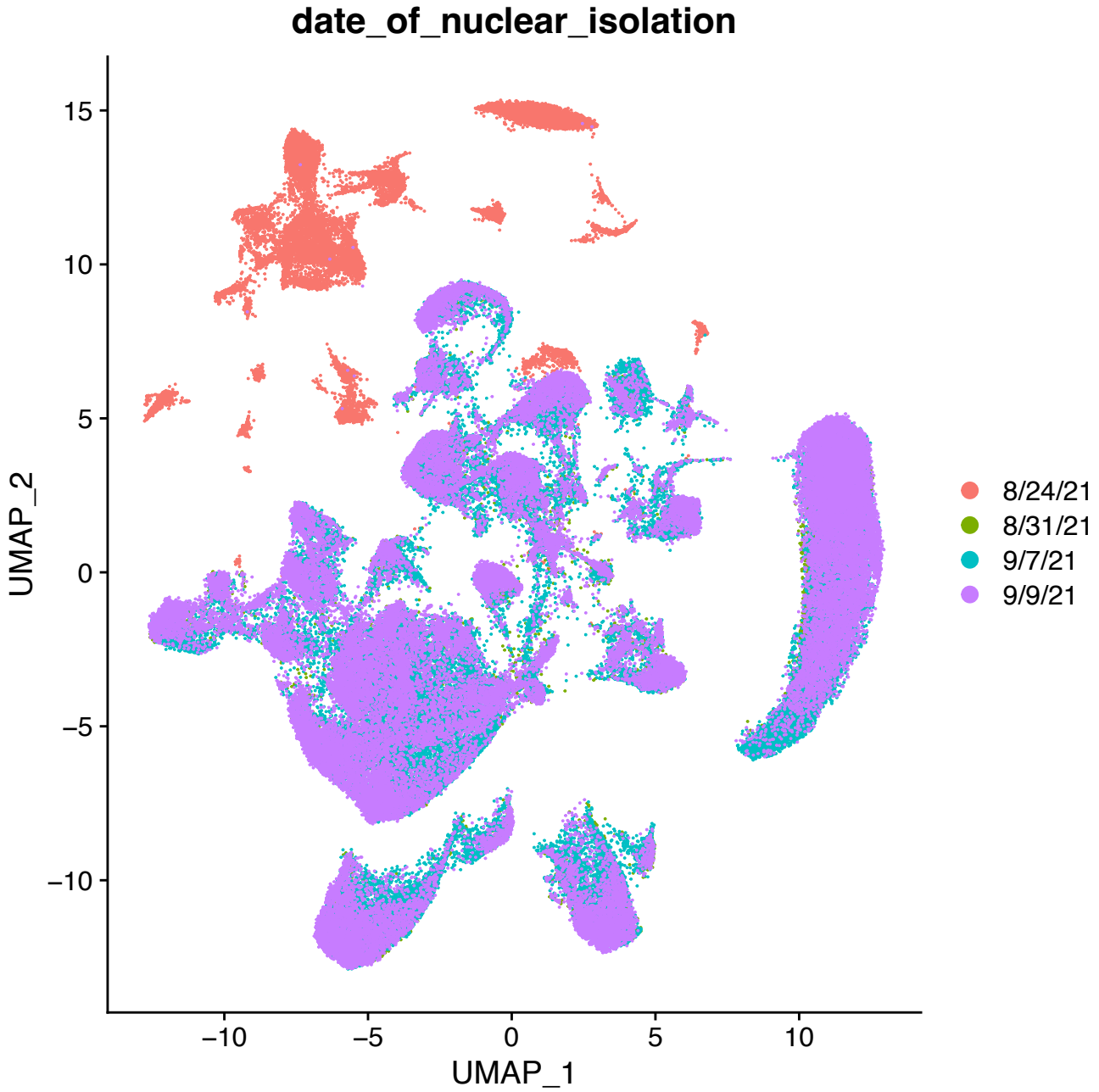
- “Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study”: *Leek et al 2010*

Batch effects could arise

- In your experiments where subset of samples are processed are differently – different times, people, library prep
- When you want to compare your data to publicly accessible data

Factors to be controlled for in your experimental design

- Date of sample perfusion
- Date of nuclear isolation
- Litter
- Date of sequencing
- Age of mouse
- Person processing the sample
- The reagents used
- Person doing the preps
- Animal models
- Drug treatments



Choice of samples within a batch

- **Idea:** Samples within batches are balanced in terms of biological end-points of interest
- Ideal: to have samples from each of the end-points being compared
- Don't: have all samples from the same end-point
- Ok: to have samples from a smaller subset of end-points

Caution: Avoid sample-wise batch correction!

Functional, metabolic and transcriptional maturation of human pancreatic islets derived from stem cells

Diego Balboa[1,2,3,11], Tom Barsby[1,11], Väinö Lithovius[1,11], Jonna Saarimäki-Vire[1], Muhmmad Omar-Hmeadi[4], Oleg Dyachok[4], Hossam Montaser[1], Per-Eric Lund[4], Mingyu Yang[4], Hazem Ibrahim[1], Anna Näätänen[1], Vikash Chandra[1], Helena Vihinen[5], Eija Jokitalo[5], Jouni Kvist[1], Jarkko Ustinov[1], Anni I. Nieminen[6], Emilia Kuuluvainen[7], Ville Hietakangas[7,8], Pekka Katajisto[7,9], Joey Lau[4], Per-Ola Carlsson[4], Sebastian Barg[4], Anders Tengholm[4] and Timo Otonkoski[1,10] ✉

and read counting. Also, the extremely high level of ambient RNA contamination (>70%) precludes using SoupX[79]. Instead we used the read counts (UMI) and metadata provided in the GEO submission (GSE114412) as a starting point for the comparison. Since the analysis by Veres et al. used a different genome annotation, we had to exclude genes that were not included in both datasets (retaining 19,170 shared genes). We combined the datasets with those of Seurat[77] and normalized the expression with default settings. The variable genes (top 1,000 per sample) were identified separately for each sample, the data was scaled, and the top 50 PCs were identified with default settings. The resulting datasets were harmonized with Harmony[78] using sample ID as grouping variable, theta set to 2, using 50 clusters and maximum iterations per cluster set to 40 and maximum iterations for harmony set to 10. The harmonized PCA values were used as input for UMAP, and the UMAP values were used to identify neighboring cells with default settings. Clustering was carried out at resolution set to 2. Clusters that were highly correlated with previously identified cell types were identified using clustifyr[88]. Some clusters were annotated manually as we had excluded the nonendocrine cells for our endocrine cell dataset.

Sample-wise batch correction is often applied

Article

Single-cell delineation of lineage and genetic identity in the mouse brain

https://doi.org/10.1038/s41586-021-04237-0

Received: 15 January 2021

Accepted: 12 November 2021

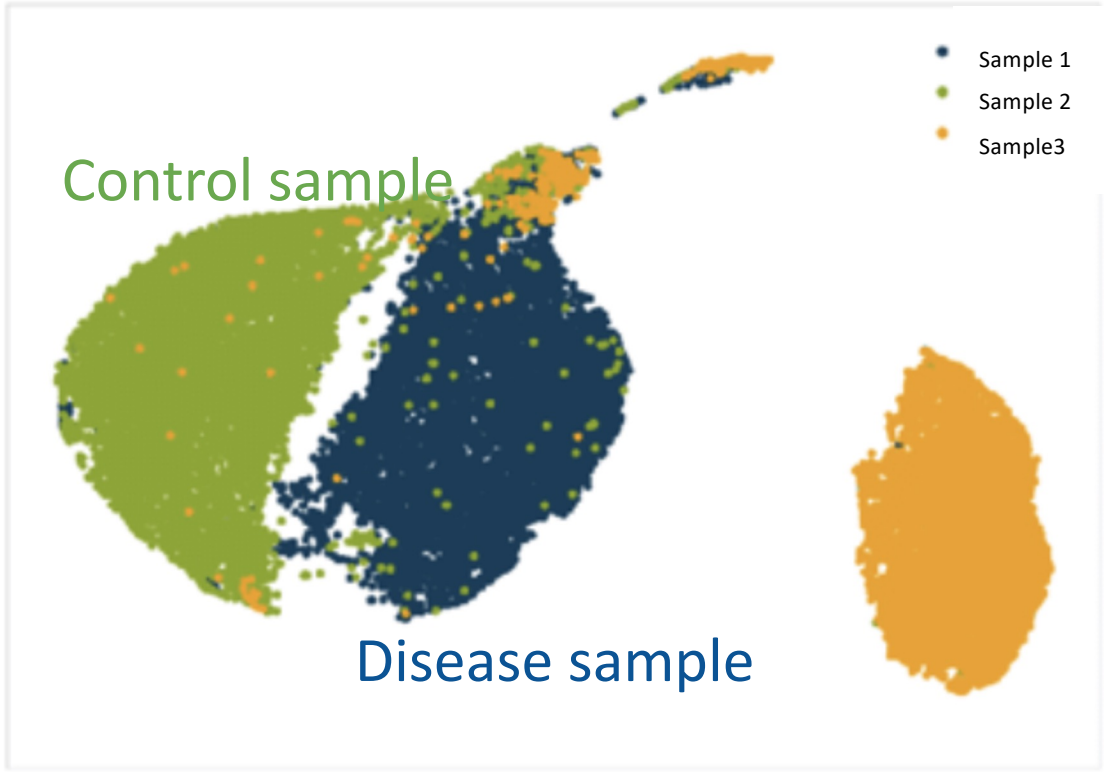
Published online: 15 December 2021

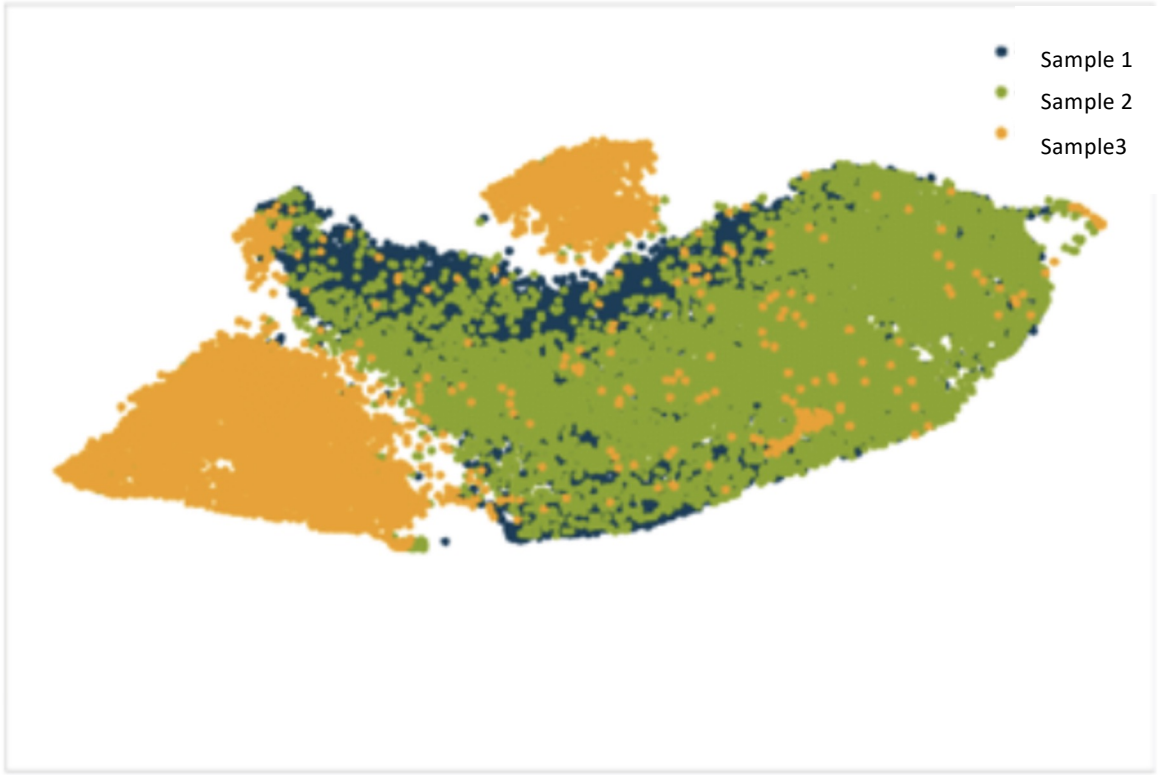
Rachel C. Bandler[1,2,3,11], Ilaria Vitali[1,11], Ryan N. Delgado[4,5,6,11], May C. Ho[1], Elena Dvoretskova[1], Josue S. Ibarra Molinas[1], Paul W. Frazel[2], Maesoumeh Mohammadkhani[2], Robert Machold[2], Sophia Maedler[7], Shane A. Liddelow[2,8,9], Tomasz J. Nowakowski[4,5,6], Gord Fishell[3,10] & Christian Mayer[1]✉

Capture of gene expression and lineages

To determine lineage relationships of diverse cell types in the mouse forebrain, we first implemented a lentiviral lineage barcoding method called STICR (scRNA-seq-compatible tracer for identifying clonal relationships; Fig. 1a, Extended Data Fig. 1a, see companion paper[26]), which enables massively parallel tagging of single cells using a high-diversity lentiviral library that encodes synthetic oligonucleotide sequences (lineage barcodes). The STICR tag library was introduced via in utero injections into the lateral ventricles of mouse embryos at embryonic day 10.5 (E10.5; STICR^{E10}), E12.5 (STICR^{E12}), E13.5 (STICR^{E13}) and E14.5 (STICR^{E14}), stages that encompass the peak of neurogenesis. This resulted in labelling of mitotic progenitors along the ventricles and their daughter cells that migrated throughout the forebrain, including the cortex, basal ganglia, hippocampus and olfactory bulb (OB) (Fig. 1b, Extended Data Fig. 1b). We waited until postnatal stages when labelled cells differentiated into mature cell types, then dissociated forebrain tissue, FACS-enriched the virally infected cells by selecting for enhanced GFP (eGFP) expression, and performed scRNA-seq with the 10x Chromium System (Fig. 1a, Extended Data Fig. 1c). We analysed transcriptomes from 65,700 high-quality cells that passed filtering (see Methods). To group cells on the basis of patterns of gene expression, we performed a principal components analysis[27] and batch normalized the different replicates using Harmony[28], followed by a UMAP visualization and clustering analysis (Extended Data Fig. 1d, Supplementary Data 1), and tracked the position of clonally related cells in the transcriptomic cell-state landscape (Extended Data Fig. 1e–h). The average and maximum size of multicellular clones was larger when the lentiviral library was introduced at E10.5 than at E12.5 or E14.5, when mitotic progenitors presumably undergo fewer divisions (Fig. 1c).

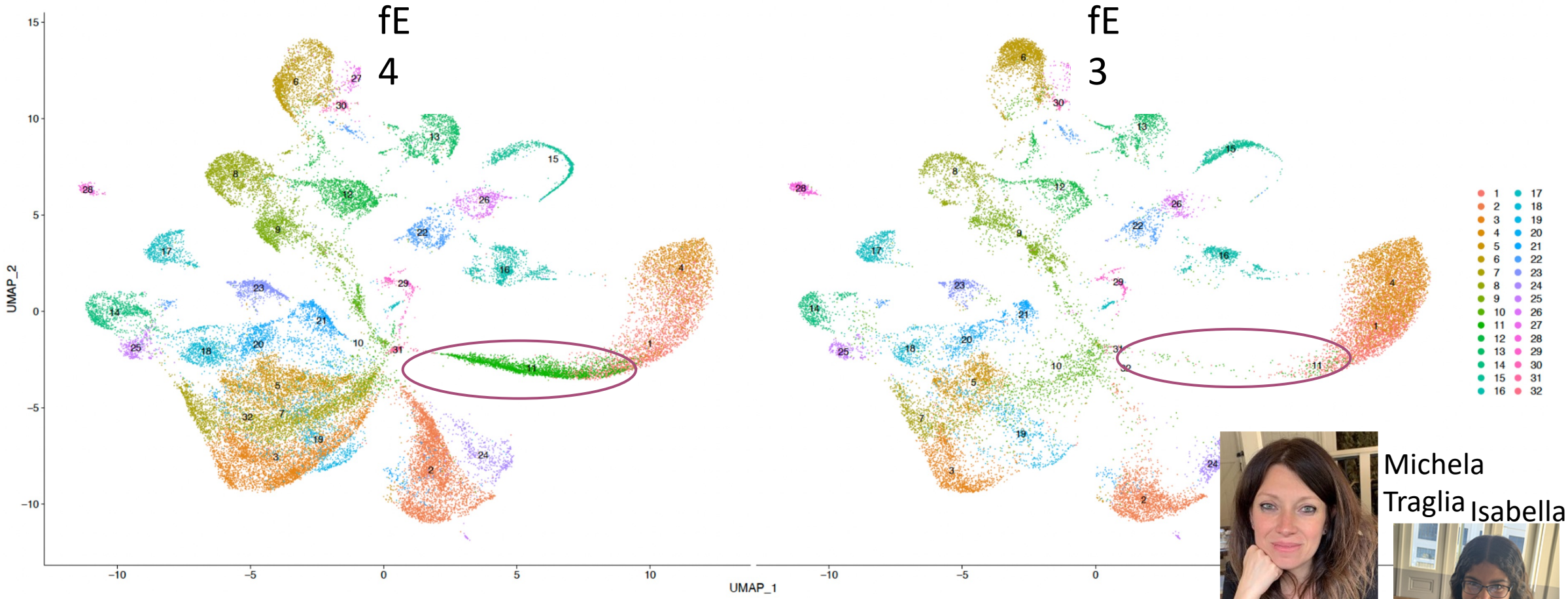
Potential loss of biological variation between conditions/genotypes





Adapted from Xu et al., 2019

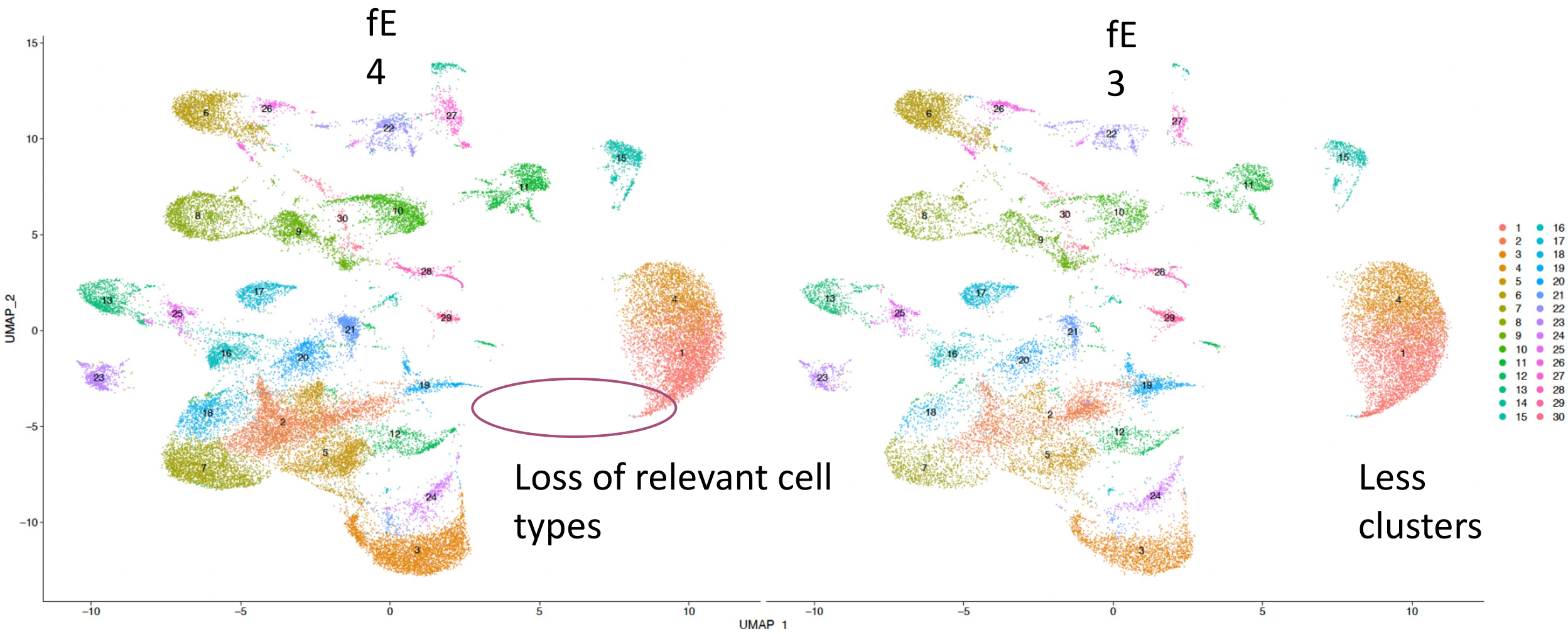
Cells clustering without sample-wise batch correction suggests gain of cluster 11-like cells in fE4 mice



Michela Traglia

Isabella DeLoa

Cells clustering after sample-wise batch correction showed less diseases-relevant cell types



Important assumptions for batch correction methods

1. There is at least one cell-type that is common between the two data sets being integrated
2. Batch effects are orthogonal to the biological effects
3. Variation in batch effects across cells is much smaller than variation in biological effects

Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors

Laleh Haghverdi[1,2], Aaron T L Lun[3], Michael D Morgan[4] & John C Marioni[1,3,4]

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

Genome Biology

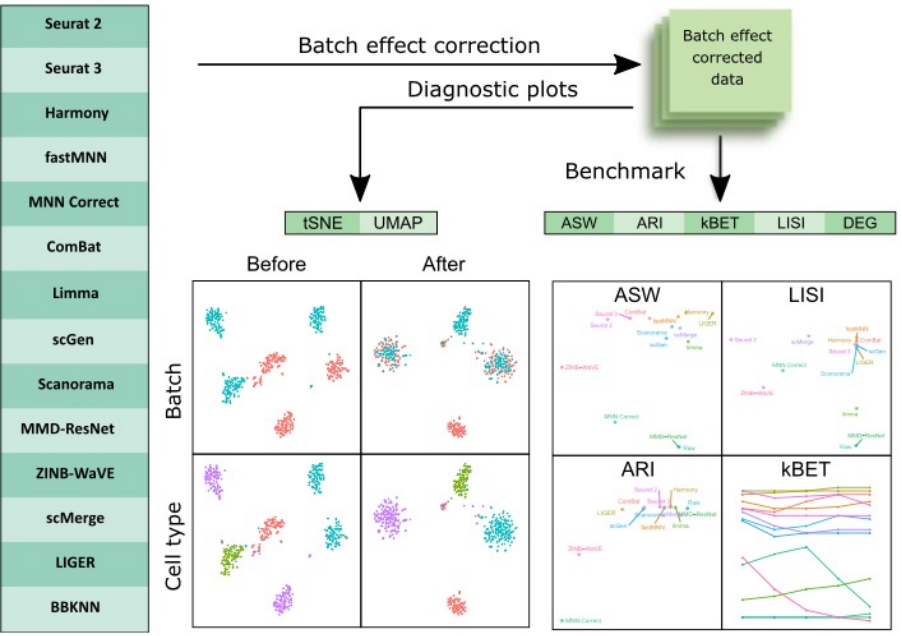
RESEARCH Open Access

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*

LIGER, Harmony and Seurat 3 top 3 performing methods!!

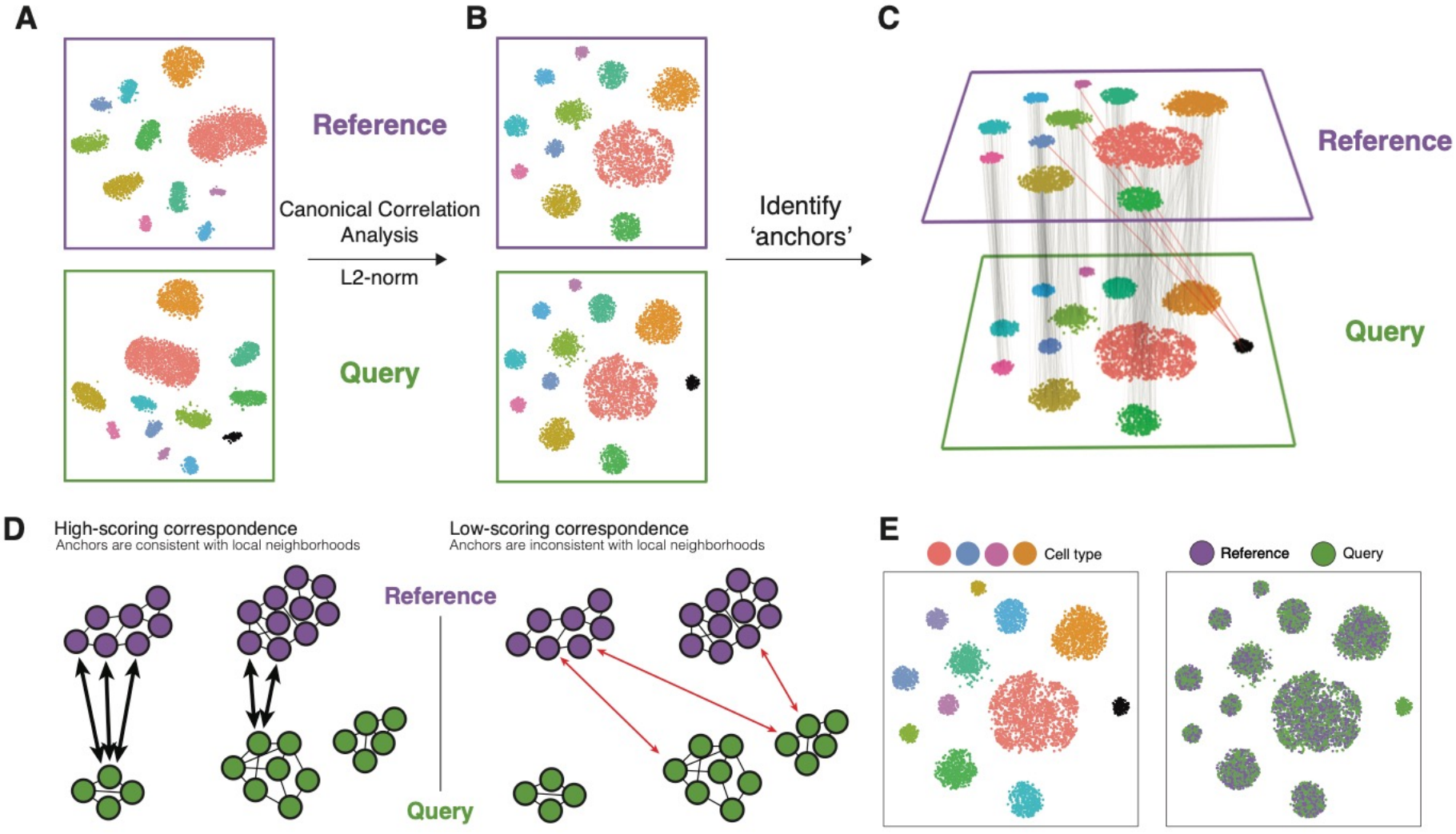
A



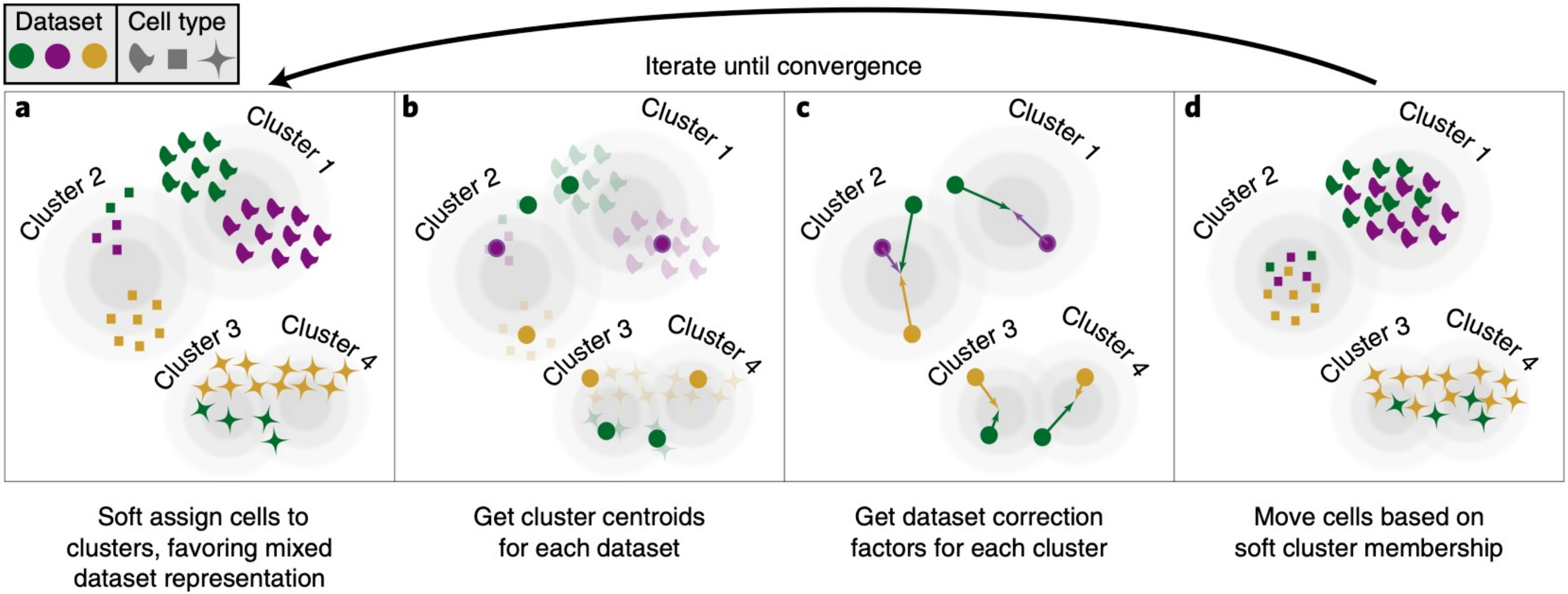
B

Dataset	Description	Number of batches	Total cell number	Technologies
1	Human Dendritic Cells	2	576	Smart-Seq2
2	Mouse Cell Atlas	2	6,954	Microwell-Seq
				Smart-Seq2
3	Simulation	Refer to Simulation table		
4	Human Pancreas	5	14,767	inDrop
				CEL-Seq2
				Smart-Seq2
				SMARTer
				SMARTer
5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3'
				10x 5'
6	Cell line	3	9,530	10x
7	Mouse Retina	2	71,638	Drop-seq
8	Mouse Brain	2	833,206	Drop-seq
				SPLiT-seq
9	Human Cell Atlas	2	621,466	10x
10	Mouse Haematopoietic Stem and Progenitor Cells	2	4,649	MARS-seq
				Smart-Seq2

Seurat 3: IntegrateData



Harmony



Korunsky et al. 2019

Hands-on: Normalization, FindMarkers and Batch correction

Outline for this session

- Main messages
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- Normalization (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- **Multi-sample multi-condition comparison (30 min)**
  - Hands-on
- Reiterate the main messages

Multi-sample-multi condition comparison

- Replicate mice, human patients to be compared across multiple conditions
- Order 1000s of cells per mouse, human patient
- Inference should be at the level of mouse/human-patient and not individual cells
- Would indicate reproducibility of detected associations
- To replicate the study – one would sample a new set of mice or human patients
- To replicate the study – there is really not an infinite pool of cells from which one can sample

Gene expression is more correlated among cells within subjects

ARTICLE

https://doi.org/10.1038/s41467-021-21038-1 OPEN

Check for updates

A practical solution to pseudoreplication bias in single-cell studies

Kip D. Zimmerman [1,2✉], Mark A. Espeland[1] & Carl D. Langefeld [1,2,3✉]

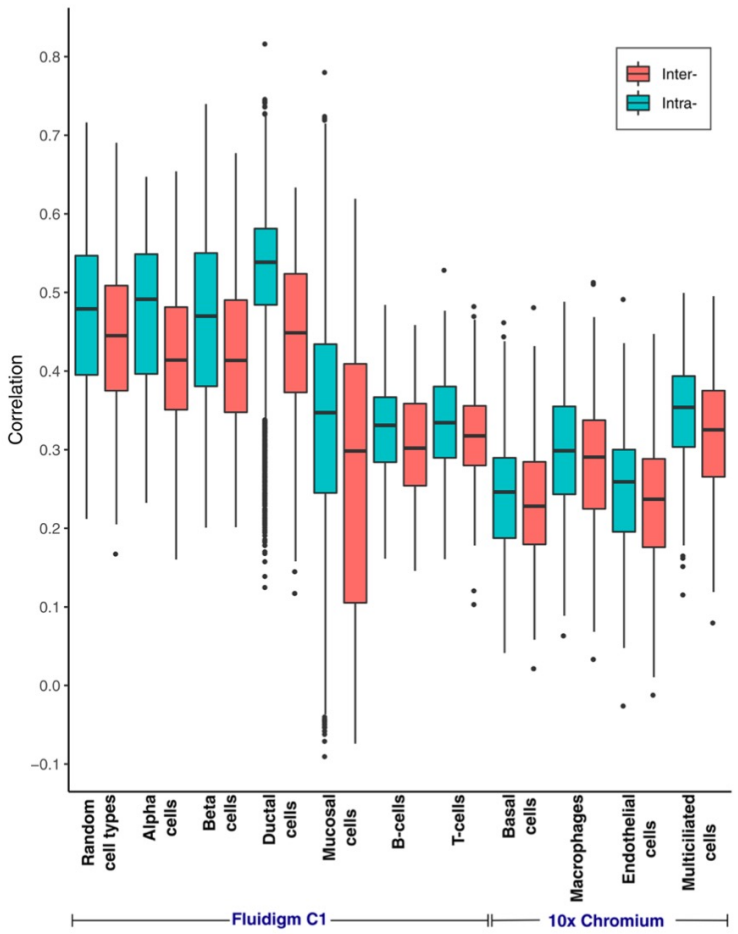


Fig. 1 Intra-individual correlation. Intra- and inter-individual Spearman's correlations for gene expression values across ten different pancreatic cell types and a random sample of different cell types. The respective

Multi-sample-multi condition comparison

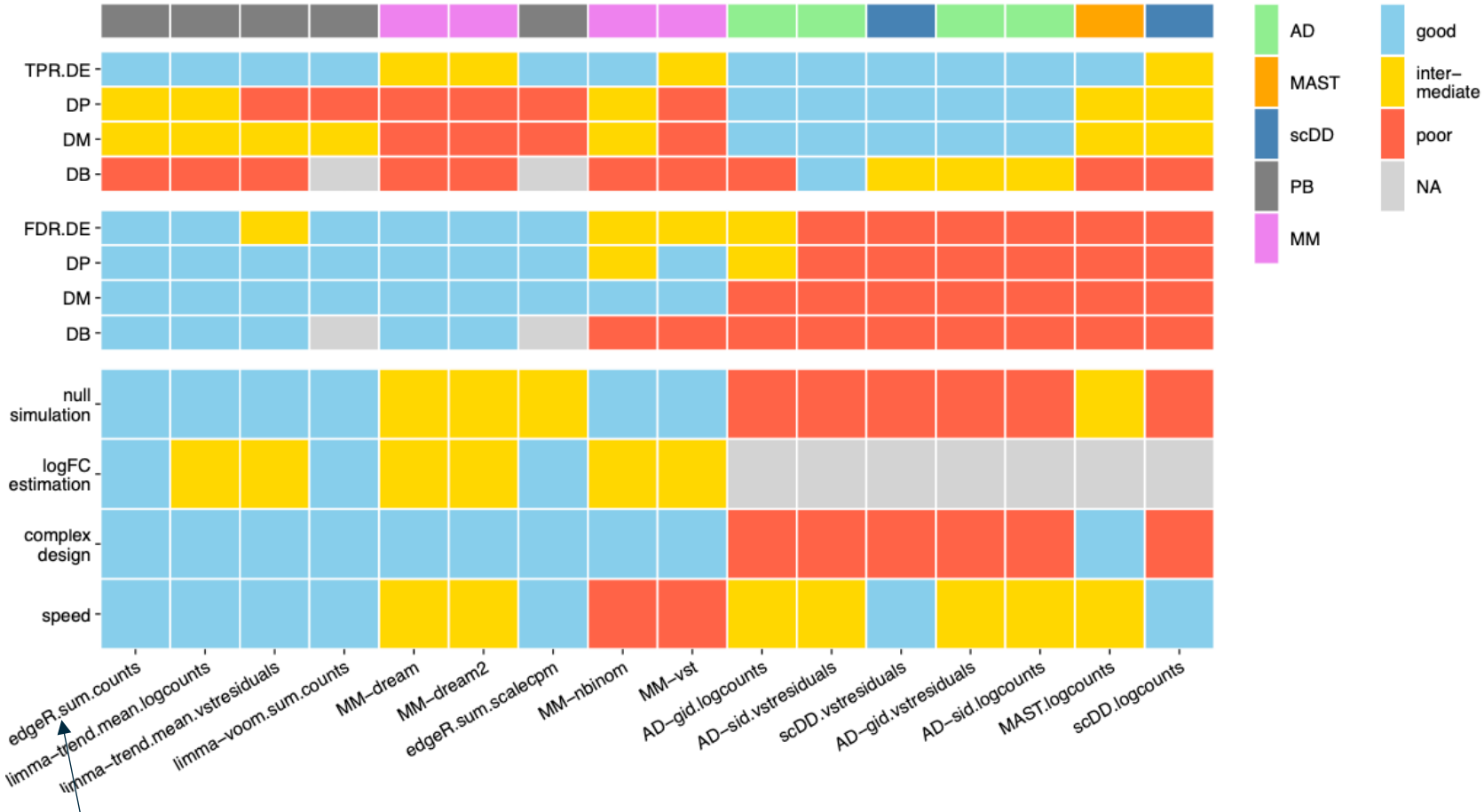
- **Question 1 (Within-cluster):** Find genes whose mean (across replicates) expression is different between condition within each cluster
- **Question 2 (Between-cluster):** Find clusters with disproportionate number of cells between the two conditions

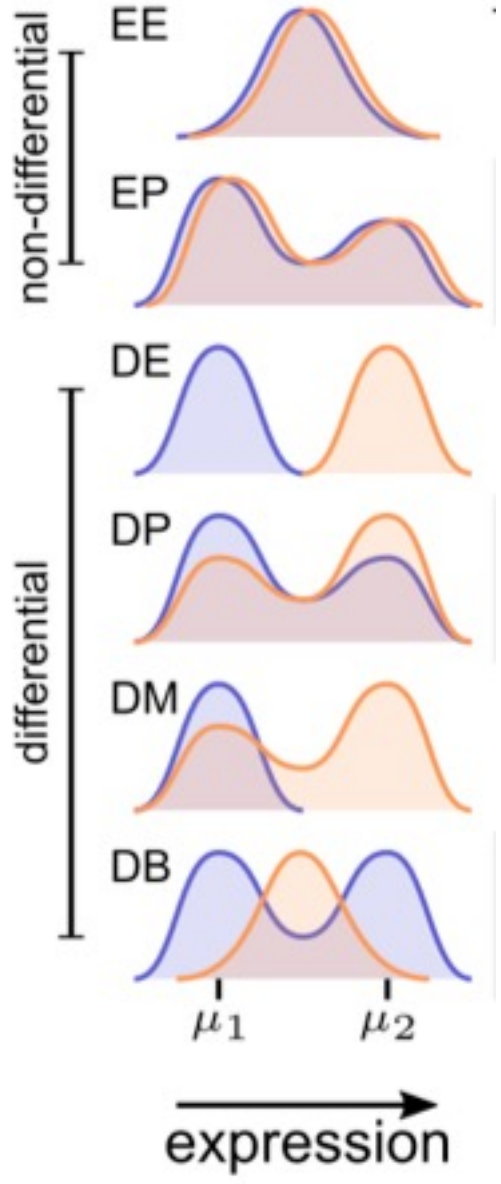
Within cluster: Multi-sample-multi condition comparison

- Within clusters, all cells are not independent of each other
- Cells are dependent on the animal of origin
- How does one model this dependence of reads across cells from the same animal?
- *Simple approach*: Aggregate counts
- *More complex approach*: Use linear mixed effects model

On the discovery of subpopulation-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data

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





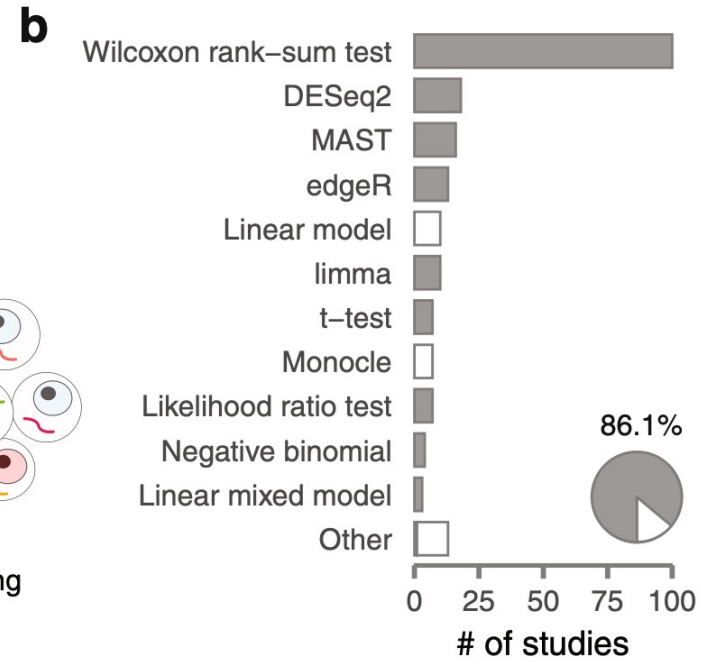
edgeR with aggregate counts (across cells from given individual) simple and most powerful!!

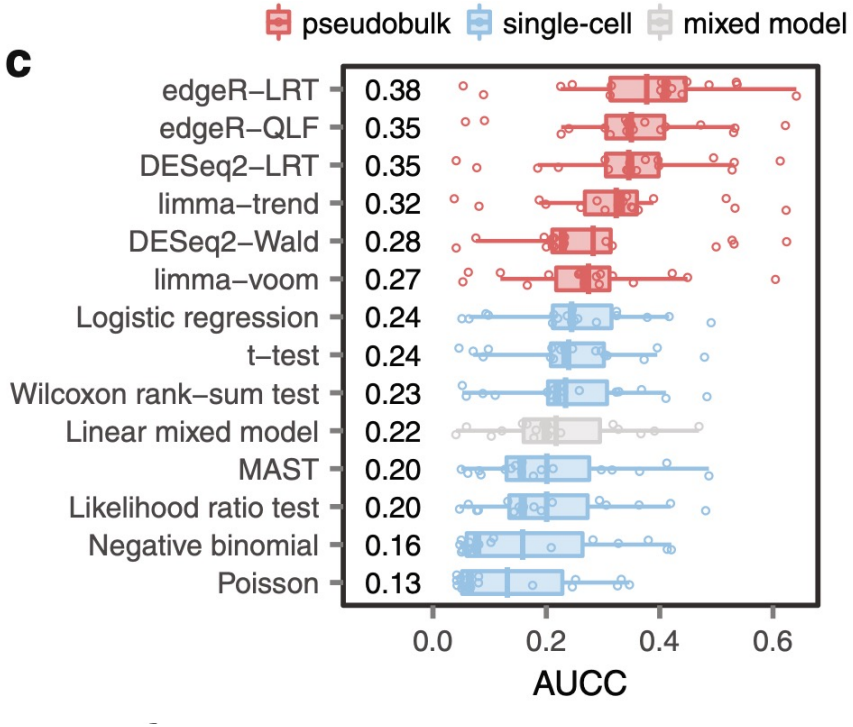
ARTICLE

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

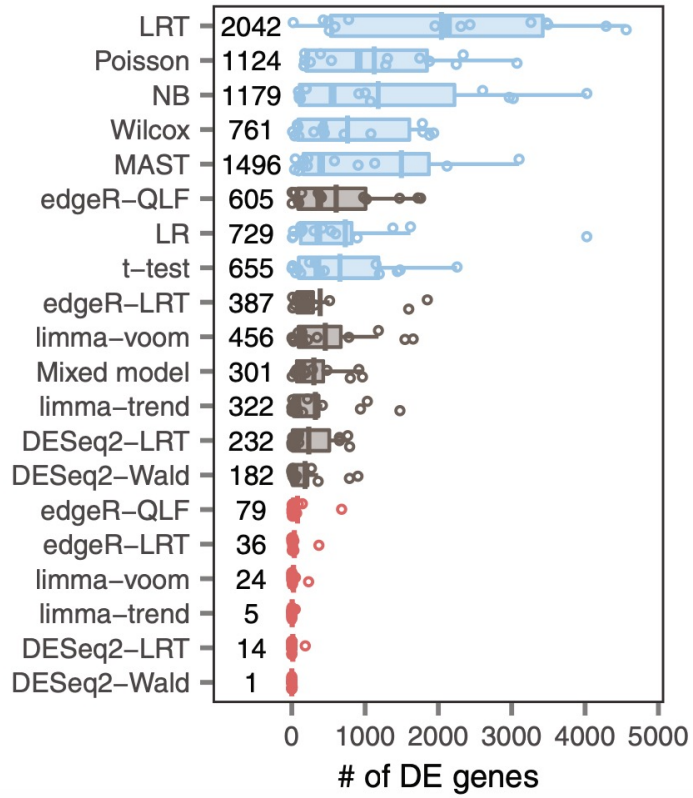
Confronting false discoveries in single-cell differential expression

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]✉





Assessment of True Positives



Assessment of False Positives

Between cluster: Multi-sample-multi condition comparison

- Cell-type (cluster) proportions from each subject may be associated with condition under study
- We can model the change in the odds of cluster membership of cells from a given subject with change in condition
- Use generalized linear mixed effects models

Hands on:
Multi-sample-multi condition comparison

Outline for this session

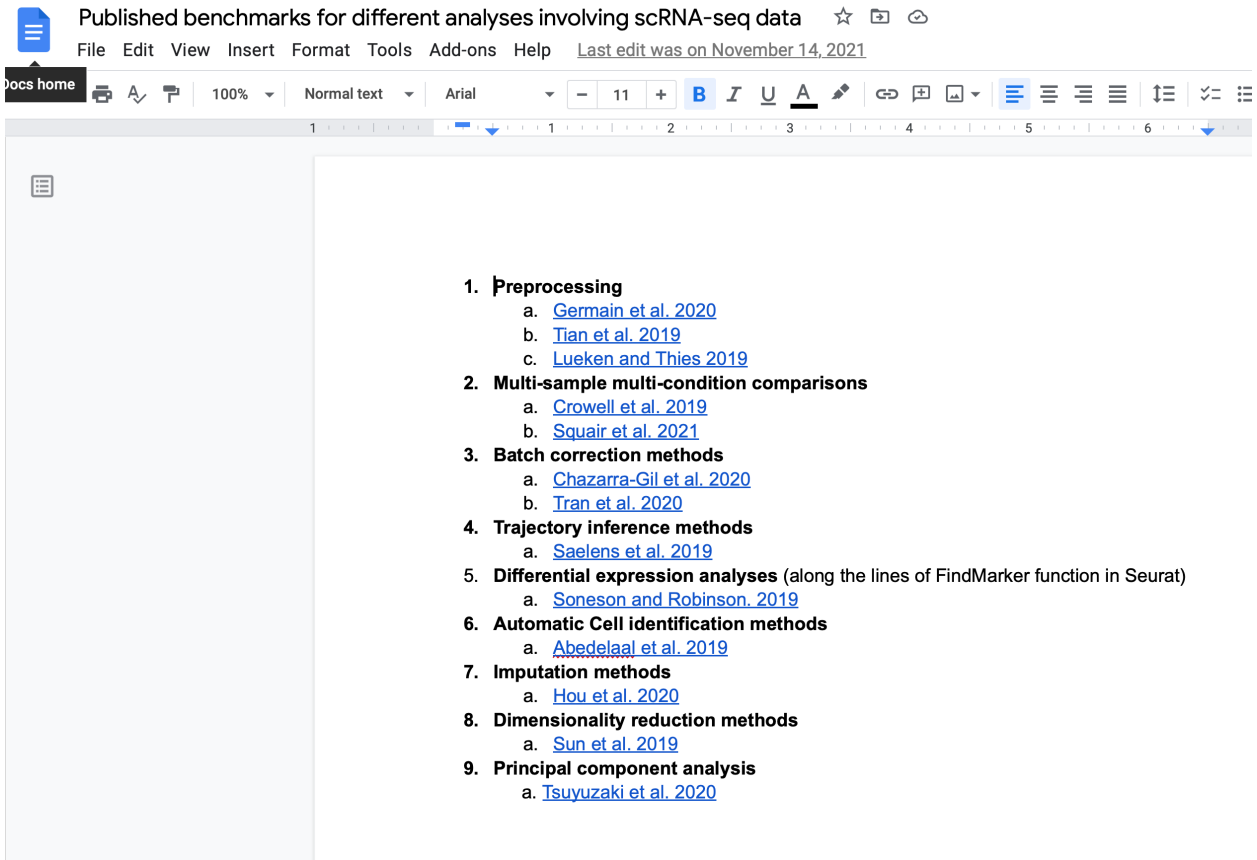
- Main messages
- Introduce the data and the research questions (5 min)
- Primer on statistics (5 min)
- Normalization (10 min)
- Finding marker genes (5 min)
- Batch correction (15 min)
  - Hands-on
- Multi-sample multi-condition comparison (30 min)
  - Hands-on
- **Reiterate the main messages**

Main points I want to convey

- State of art of the **design and statistical analyses of scRNAseq data is still in flux**
  - Benchmarks in different areas are being increasingly available
- There are **better designs**
  - **Most important:** Please include replicates drawn from the population you want to make a claim about
- There are **better statistics**/ways to get “more” reproducible results.
  - Normalization
  - Identification of marker genes
  - Multi-sample multi-condition comparison
  - Batch correction

Benchmarks will help you narrow down your choice of right tools

- Published benchmarks for different classes methods/analyses on single-cell RNA-seq data



Published benchmarks for different analyses involving scRNA-seq data

File Edit View Insert Format Tools Add-ons Help Last edit was on November 14, 2021

1. **Preprocessing**
   a. Germain et al. 2020
   b. Tian et al. 2019
   c. Lueken and Thies 2019
2. **Multi-sample multi-condition comparisons**
   a. Crowell et al. 2019
   b. Squair et al. 2021
3. **Batch correction methods**
   a. Chazarra-Gil et al. 2020
   b. Tran et al. 2020
4. **Trajectory inference methods**
   a. Saelens et al. 2019
5. **Differential expression analyses** (along the lines of FindMarker function in Seurat)
   a. Soneson and Robinson. 2019
6. **Automatic Cell identification methods**
   a. Abedelaal et al. 2019
7. **Imputation methods**
   a. Hou et al. 2020
8. **Dimensionality reduction methods**
   a. Sun et al. 2019
9. **Principal component analysis**
   a. Tsuyuzaki et al. 2020

Helpful resources

- Wynton Slack channel
  - ucsf-wynton.slack.com
- Gladstone Bioinformatics Core slack channel
  - https://gladstoneinstitutes.slack.com/archives/C0145F1L7QS
- Wynton tutorials
  - https://github.com/ucsf-wynton/tutorials/wiki

Your feedback is important to us!

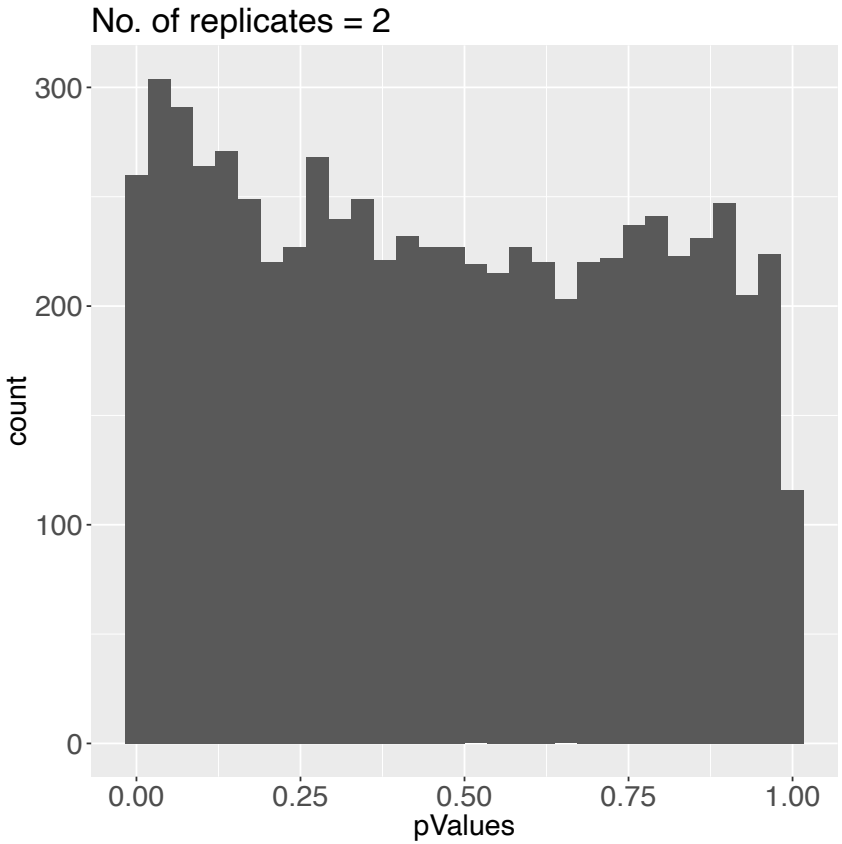
- https://www.surveymonkey.com/r/F75J6VZ
- ~3 min.

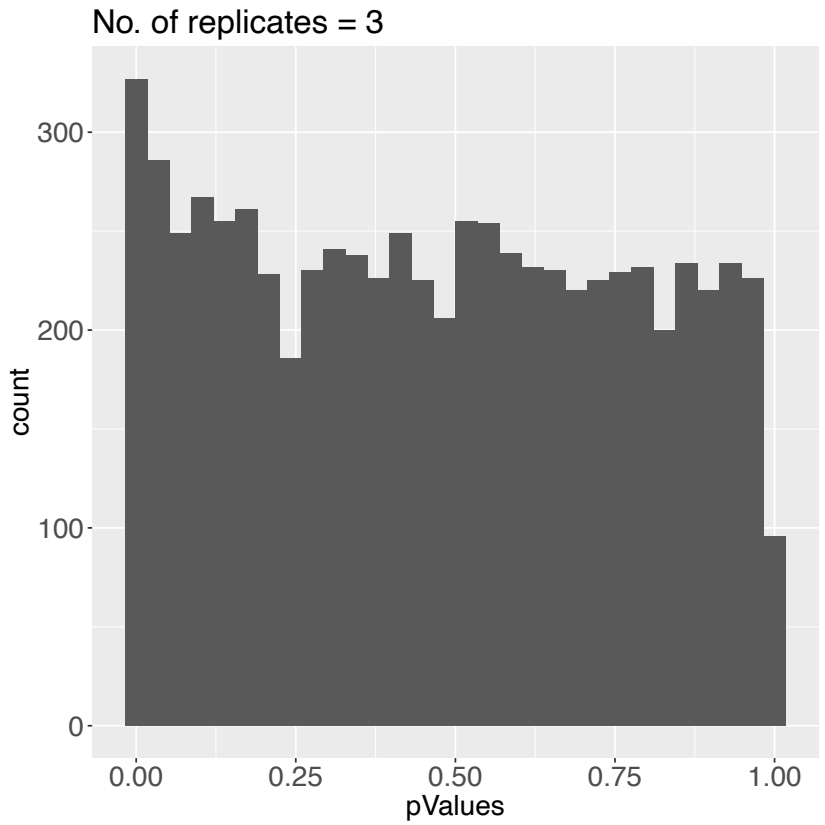
Thank you

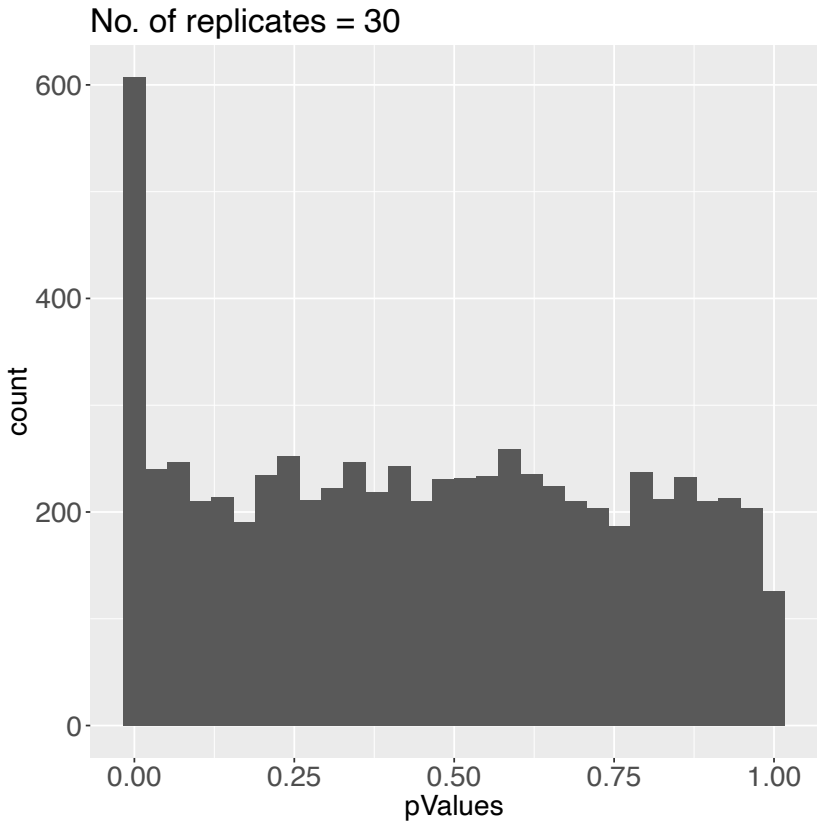
- **Krishna Chaudhary**, Bioinformatics Core
- Alex Pico, Bioinformatics Core
- Min-Gyoung Shin, Bioinformatics Core


GLADSTONE
INSTITUTES

Distribution of p-values across all genes *(simulated data)*

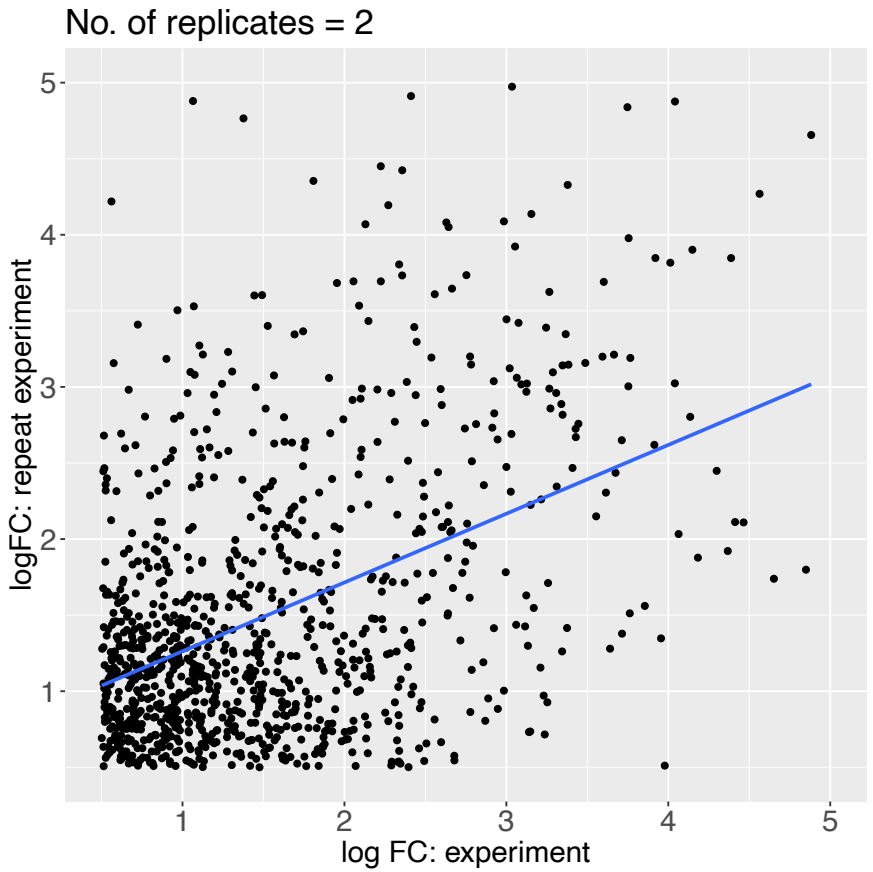


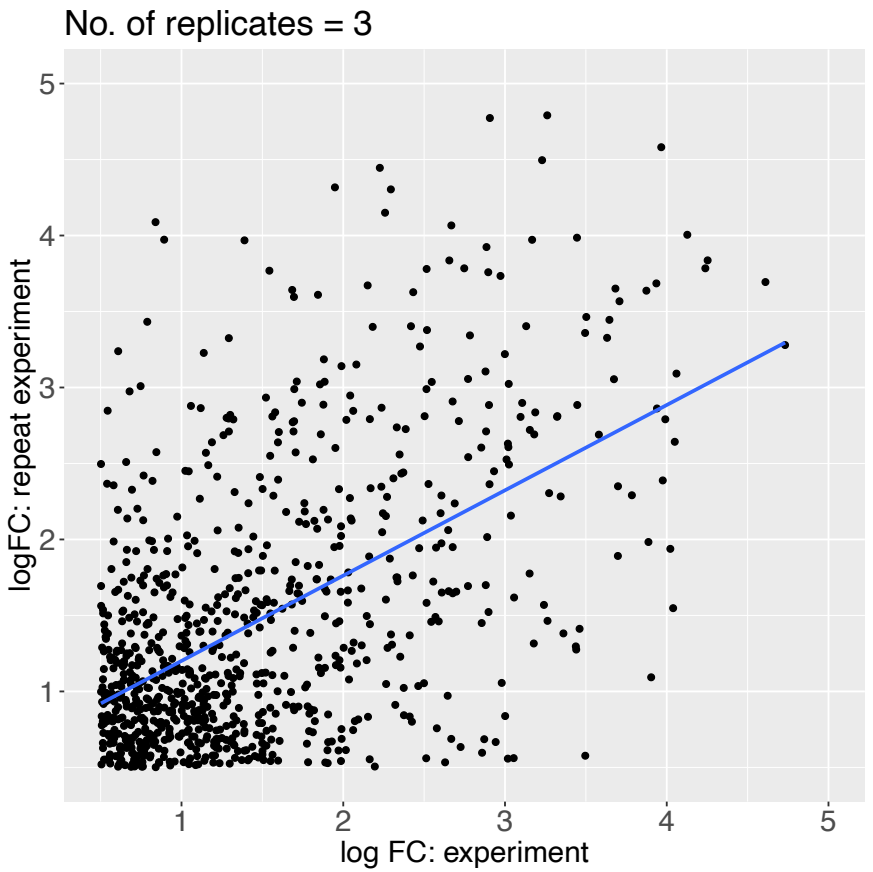


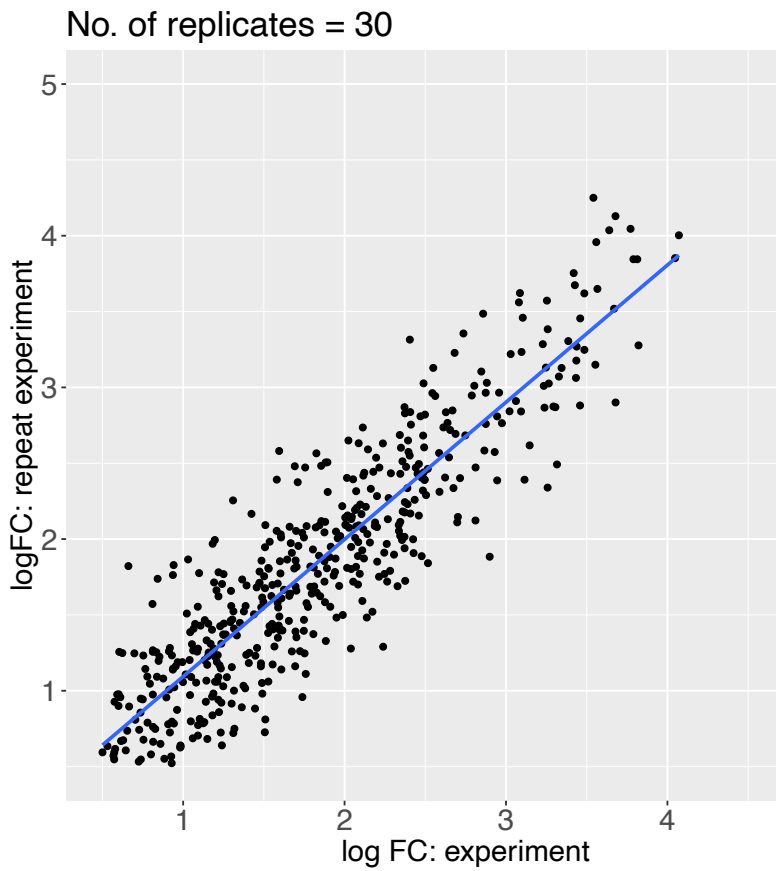


Scatter of fold-changes over repeated experiments *(simulated data)*

More left skewed p-value distribution, the more reproducible the results!







1. Library size normalization

- E.g.: *LogNormalize* in Seurat
- Assumption: No “imbalance” in the DE genes between any pair of cells
- Advantages: Simple to understand, quick to implement
- Disadvantages: Could introduce biases in clustering, low-dimensional reduction, differential expression
- Method: Normalize reads by sequencing depth per cell

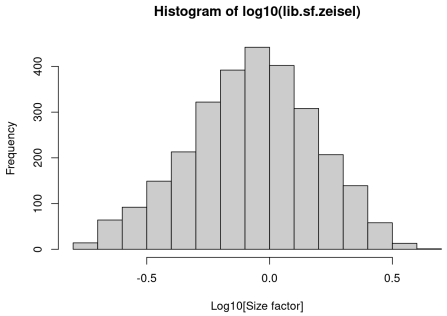
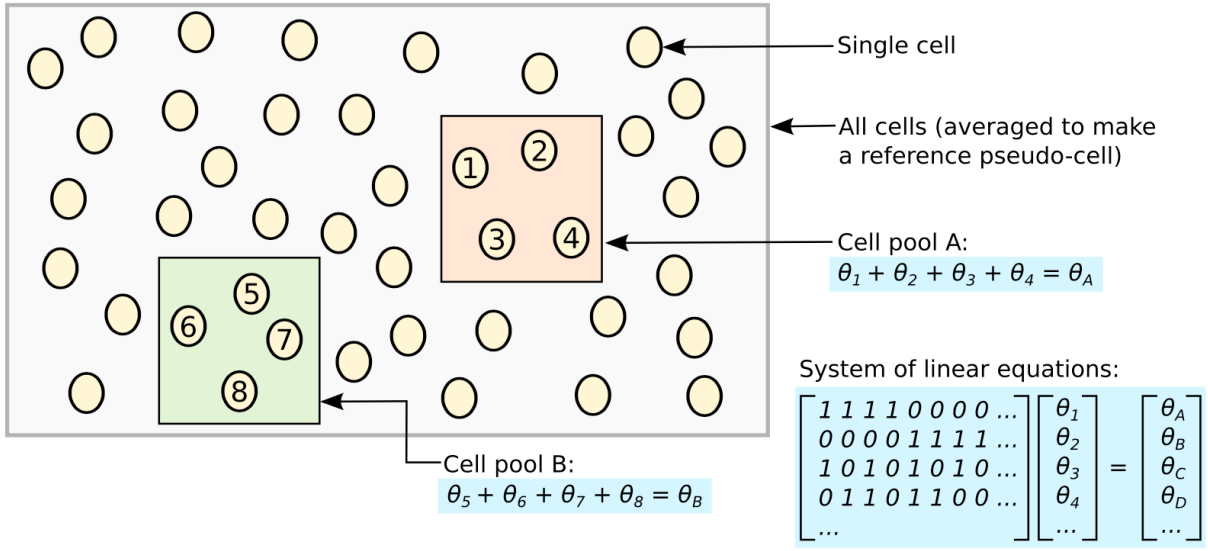


Figure 7.1: Distribution of size factors derived from the library size in the Zeisel brain dataset.

2. Deconvolution normalization

- E.g.: scran, edgeR, DESeq2 in Bioconductor;
- Assumption: input RNA quantity from non-diff expressed genes is the same across all cells. Most genes are not differentially expressed
- Advantage: Treats data as counts unlike *LogNormalize*
- Method: Compute size factors by pooling cells



3. Spike-in normalization

- Assumption: Same amount of spike-in RNA in each cell
- Cell differences in the coverage of the spike-in transcripts due to capture efficiency or sequencing depth
- Advantage: Does not require the assumption of same amount of starting RNA material per cell
- Advantage: Can treat data as count data
- Method: Use differences in coverage of spike-in transcripts to normalize reads assigned to genes

4. Variance-stabilizing transformation

- *SCTransfrom* in Seurat, *zinbwave* in Bioconductor
- Assumption: Most genes are not differentially expressed
- Advantage: Treats data as counts unlike *LogNormalize*
- Disadvantage: Can be computationally intensive
- Method: “Regularized” dependence of count distribution parameters on gene abundance

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m_i$$

Mean expression of gene i

Sequencing depth