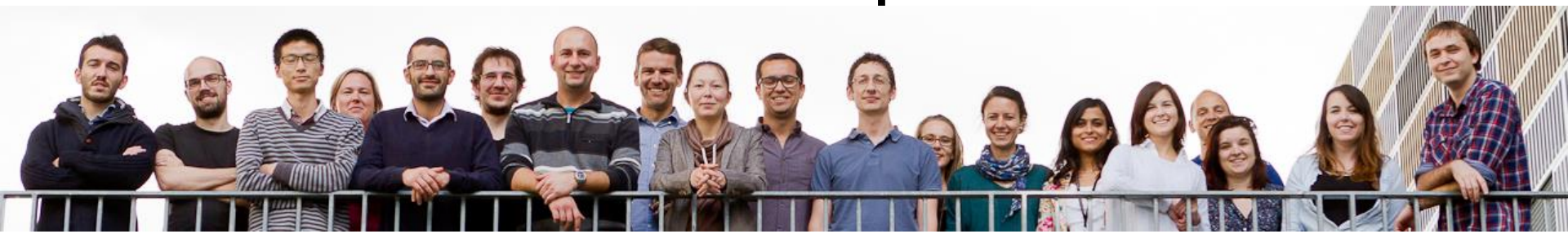
Single-cell protocols

From raw reads to cell clusters in a few easy steps

PART II

VINCENT GARDEUX
MARJAN BIOCANIN

Few words about Deplancke's lab



- Located at EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland
- 1 Bart, 5 Postdocs, 10 PhD students, 3 Master students, 1 lab tech
- Working in Fly genomics, Aging, Obesity, TF binding, and Microfluidics tech

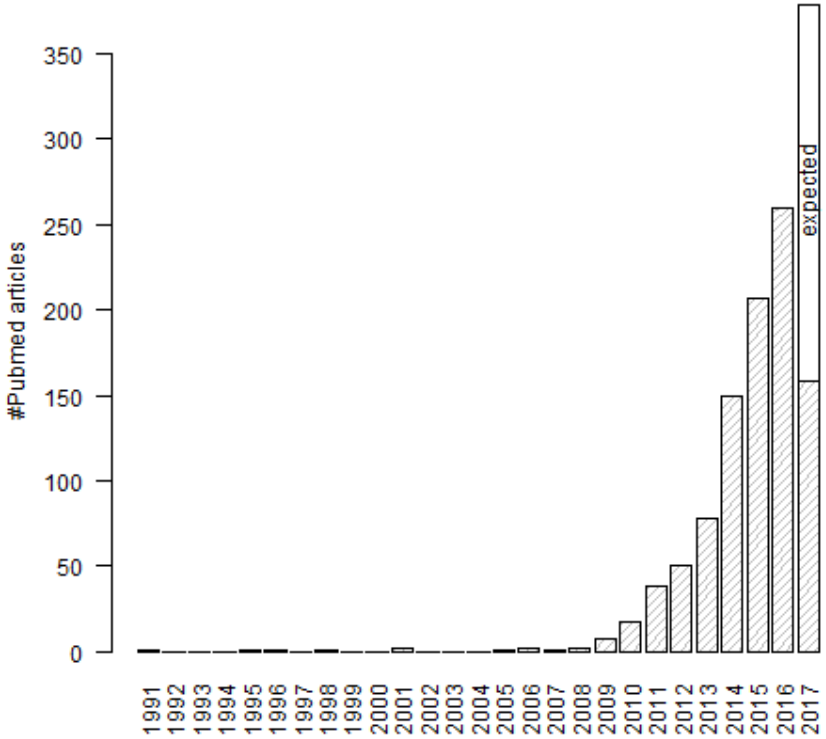
⇒RNAseq, ATACseq, ChIPseq, WES, Proteomics, …

- Also, more and more datasets generated from single-cell technologies: Dropseq, Smartseq2, 10x Genomics

⇒And not enough bioinformaticians!

Single-cell RNA-seq undergoes tremendous expansion

PubMed articles for "Single-Cell RNA-seq"

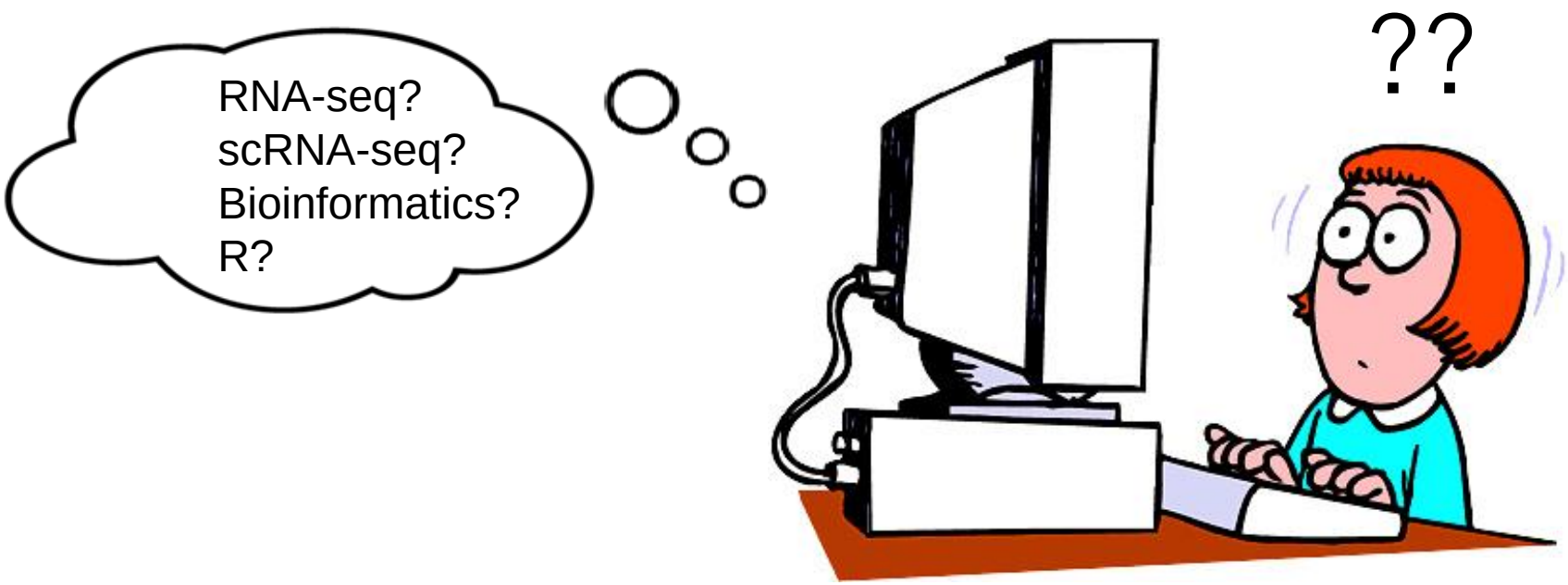


Google Trends

Community needs: « I want to do scRNA-seq ! »

I finally got good single-cell data!!

Community needs: « But how do I analyze scRNA-seq data? »



⇒ This is a typical bottleneck, and was routinely experienced in our own lab

ASAP is designed to handle the downstream analysis

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Read/transcript matrix
(raw counts/UMI or normalized data)

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

Study-specific downstream analyses
to generate new biological insight

Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

PREPROCESSING

DOWNSTREAM ANALYSIS

ASAP is designed to handle the downstream analysis

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Read/transcript matrix
(raw counts/UMI or normalized data)

dropSeqPipe

PREPROCESSING

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

Study-specific downstream analyses
to generate new biological insight

Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

ASAP is designed to handle the analysis workflow

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Read/transcript matrix
(raw counts/UMI or normalized data)

ASAP

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

Study-specific downstream analyses
to generate new biological insight

Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

DOWNSTREAM ANALYSIS

scRNA-seq Computational Workflow

Quality control & filtering (I)

↓

Mapping to reference genome

↓

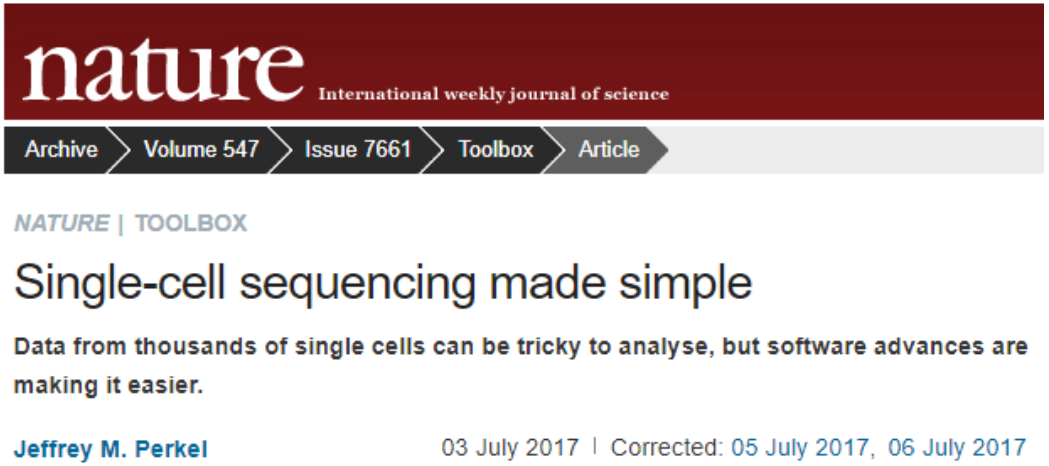
Read/transcript matrix
(raw counts/UMI or normalized data)

ASAP

"Black box" fixed pipeline?

e.g. 10x cell Ranger pipeline

scRNA-seq pipeline may not be applicable to all datasets


nature International weekly journal of science

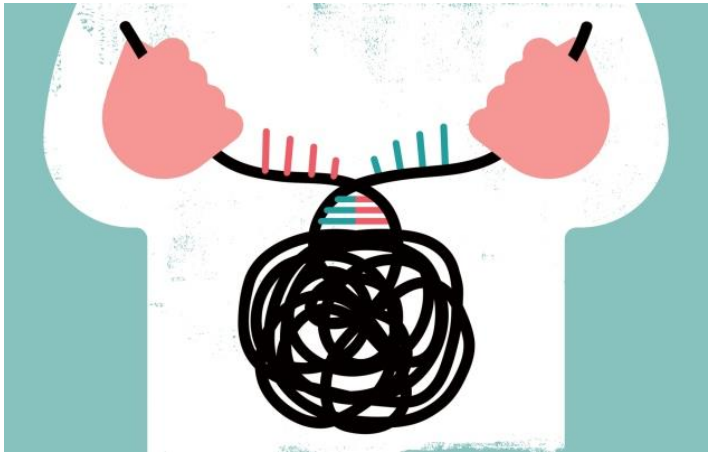
Archive > Volume 547 > Issue 7661 > Toolbox > Article

NATURE | TOOLBOX

Single-cell sequencing made simple

Data from thousands of single cells can be tricky to analyse, but software advances are making it easier.

Jeffrey M. Perkel 03 July 2017 | Corrected: 05 July 2017, 06 July 2017



"The tools aren't perfect for every situation"

⇒ "A pipeline that excels at identifying cell types, for instance, might stumble with pseudo-time analysis"

"Appropriate methods are 'very data-set dependent'", says Sandrine Dudoit, (biostatistician at the University of California, Berkeley).

⇒ "The methods and tuning parameters may need to be adjusted to account for variables such as sequencing length"

scRNA-seq Computational Workflow

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

Study-specific downstream analyses
to generate new biological insight

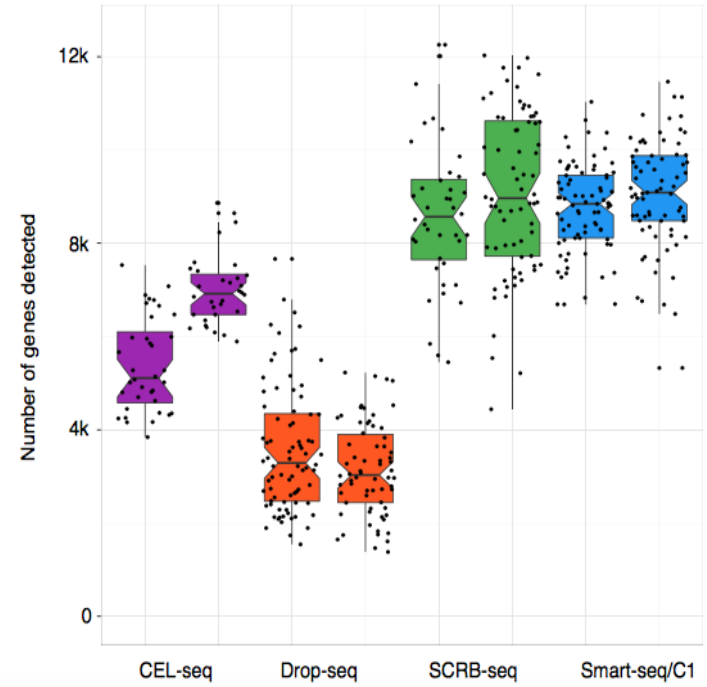
Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

Quality control (post-alignment)

Number of expressed genes/cell

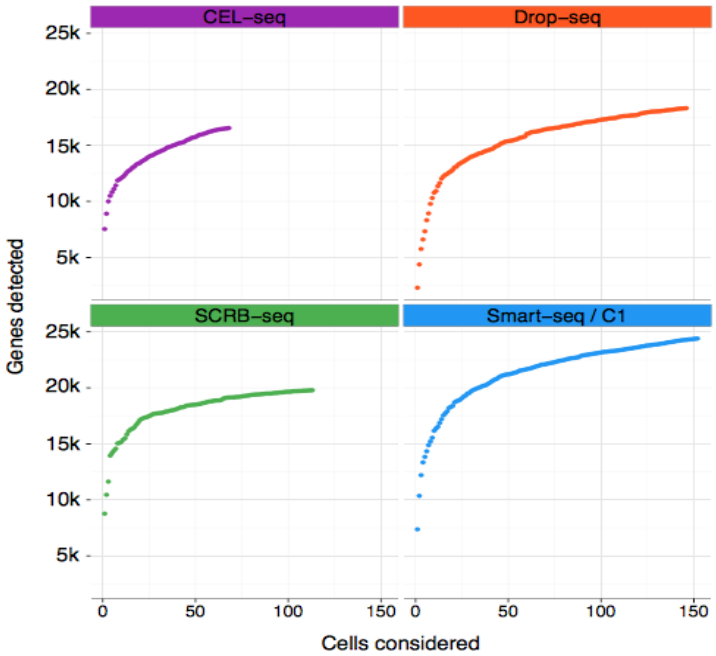
Low number of expressed genes/cell can indicate low quality cells (e.g. degraded RNA) => remove

Genes detected/cell



Different experimental protocols result in different nrs. of expressed genes/cell

Genes detected/experiment



The ability to analyze a higher number of cells => similar gene detection rate per experiment despite lower per cell nrs.

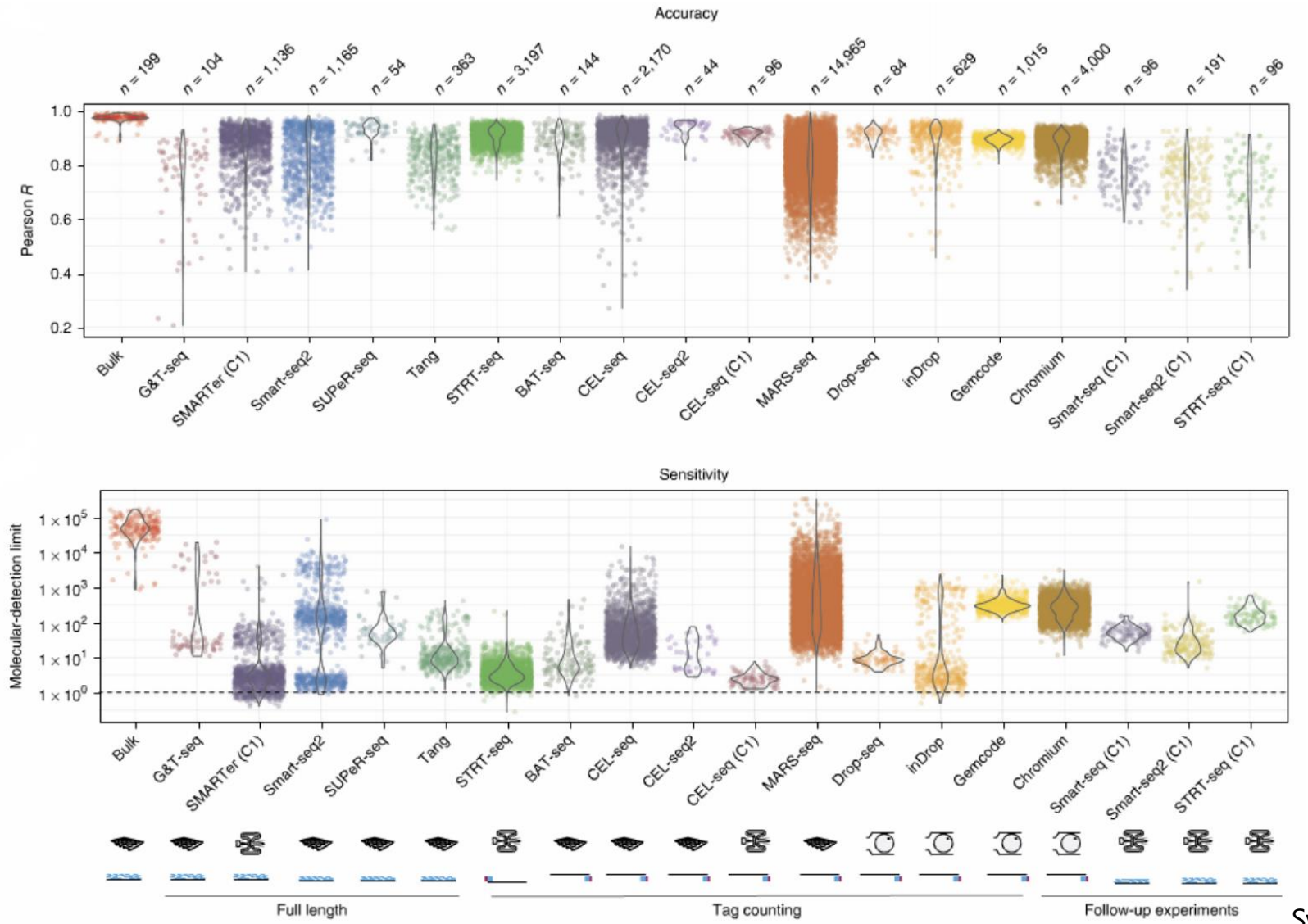
Ziegenhain & Enard (2016) BiorXiv

Quality control (post-alignment)

Accuracy = cell vs cell correlation

Sensitivity = number of UMI per cell

Complexity = number of genes detected per cell



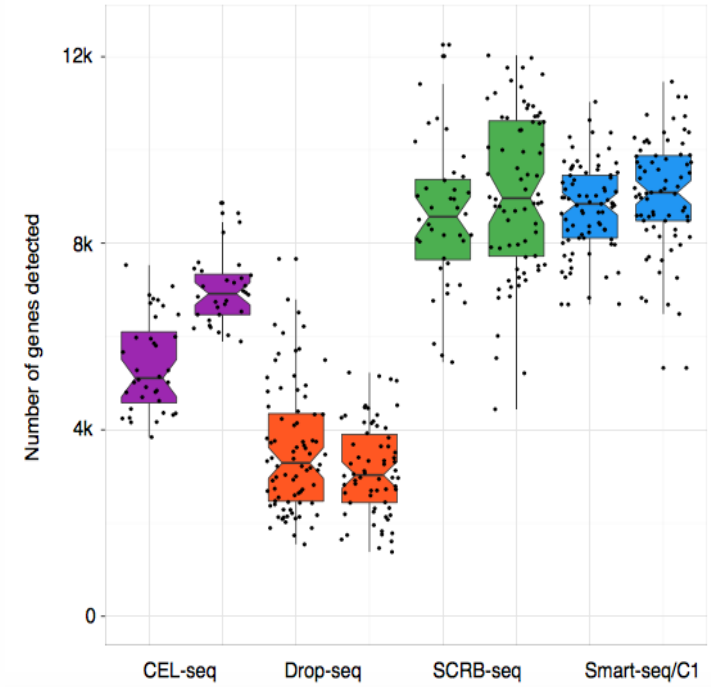
Svensson *et al.* 2017

Quality control (post-alignment)

Number of expressed genes/cell

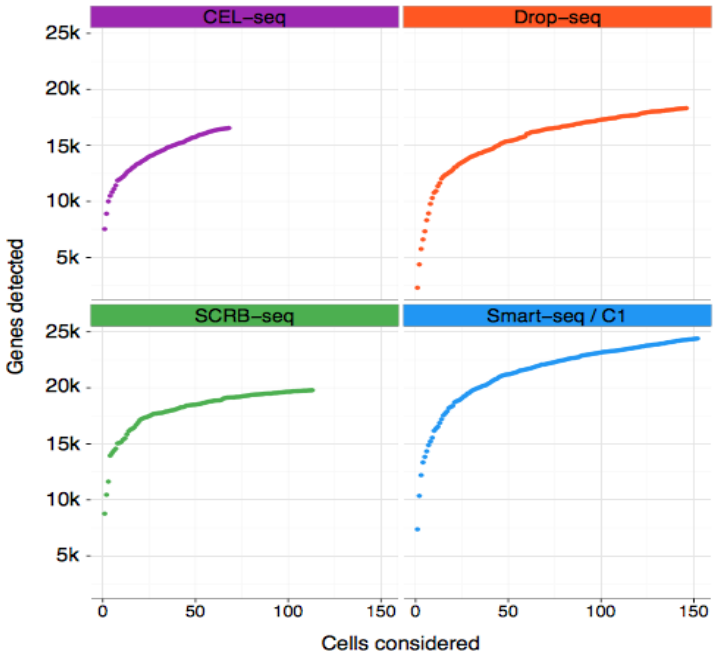
Low number of expressed genes/cell can indicate low quality cells (e.g. degraded RNA) => remove

Genes detected/cell



Different experimental protocols result in different nrs. of expressed genes/cell

Genes detected/experiment



The ability to analyze a higher number of cells => similar gene detection rate per experiment despite lower per cell nrs.

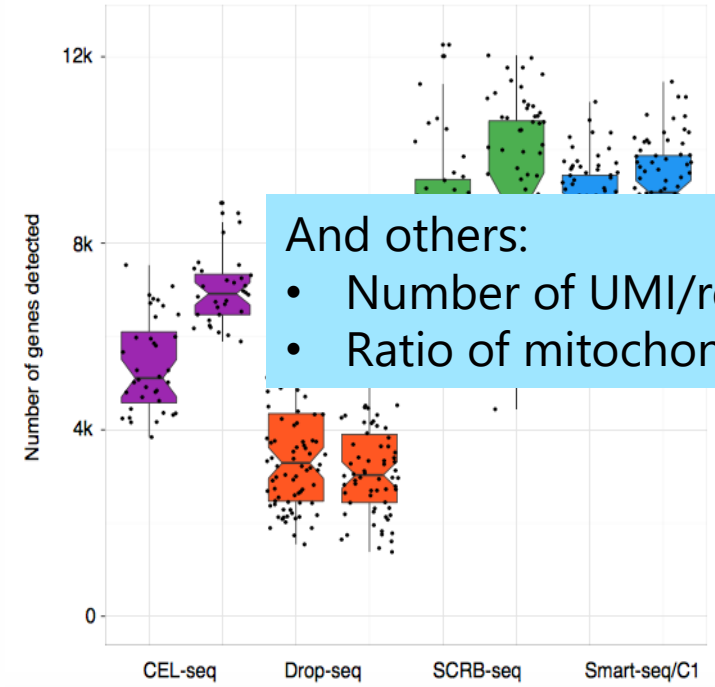
Ziegenhain & Enard (2016) BiorXiv

Quality control (post-alignment)

Number of expressed genes/cell

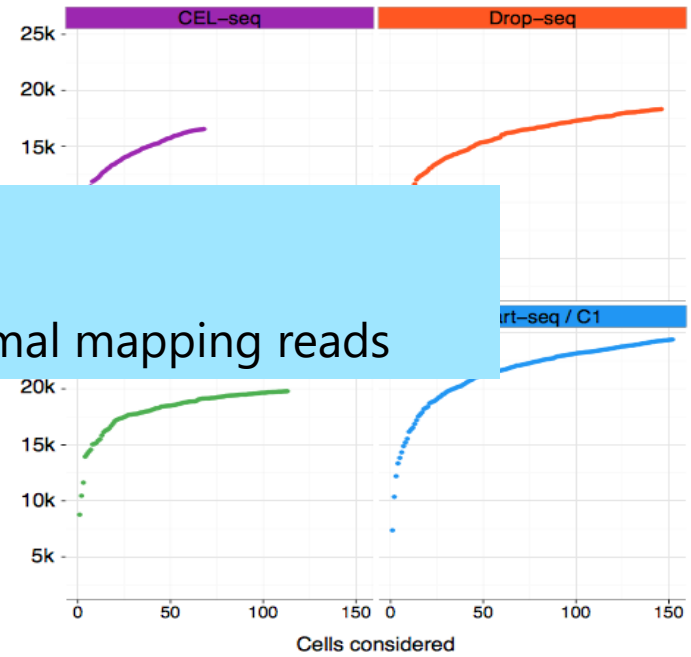
Low number of expressed genes/cell can indicate low quality cells (e.g. degraded RNA) => remove

Genes detected/cell



Different experimental protocols result in different nrs. of expressed genes/cell

Genes detected/experiment



The ability to analyze a higher number of cells => similar gene detection rate per experiment despite lower per cell nrs.

And others:

- Number of UMI/reads per cell
- Ratio of mitochondrial/ribosomal mapping reads

Ziegenhain & Enard (2016) BiorXiv

scRNA-seq Computational Workflow

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

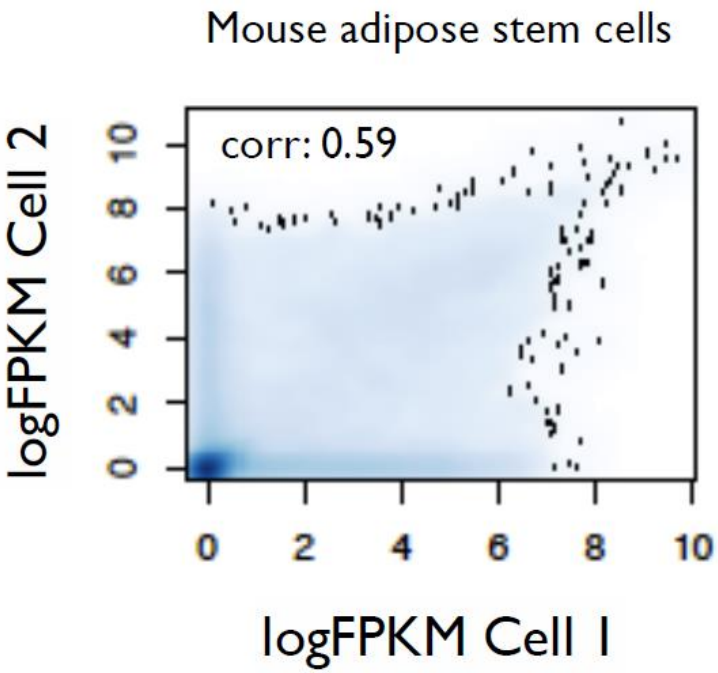
Study-specific downstream analyses
to generate new biological insight

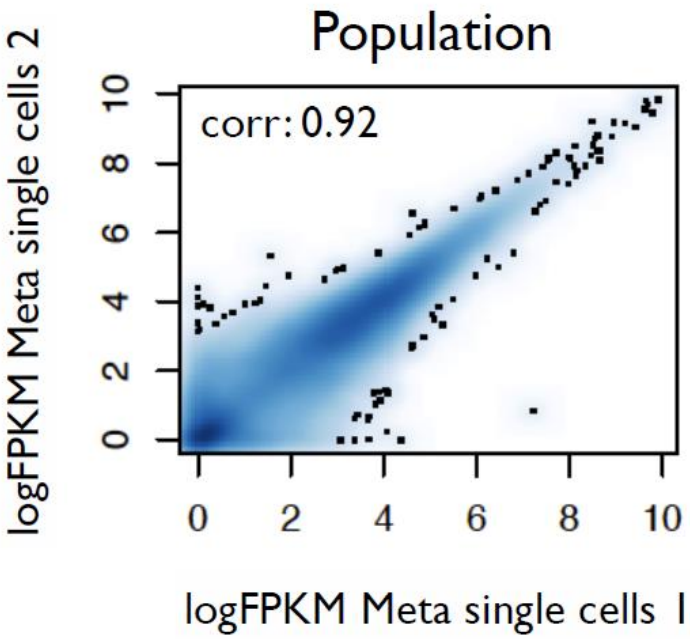
Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

scRNA-seq normalisation: noise

Cell-to-cell variability?

Single-cell gene expression is noisy



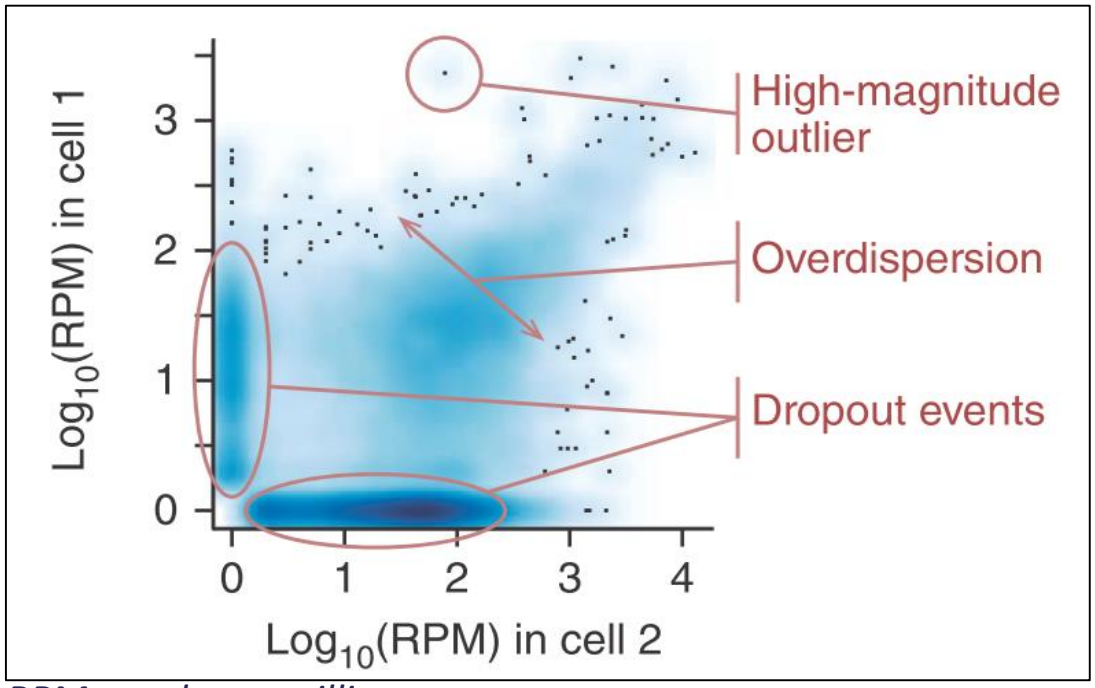


Single-cell RNA-seq challenges

Non linear behavior of PCR amplification : overdispersion

Dropouts events

Biological & Technical variations are difficult to detangle



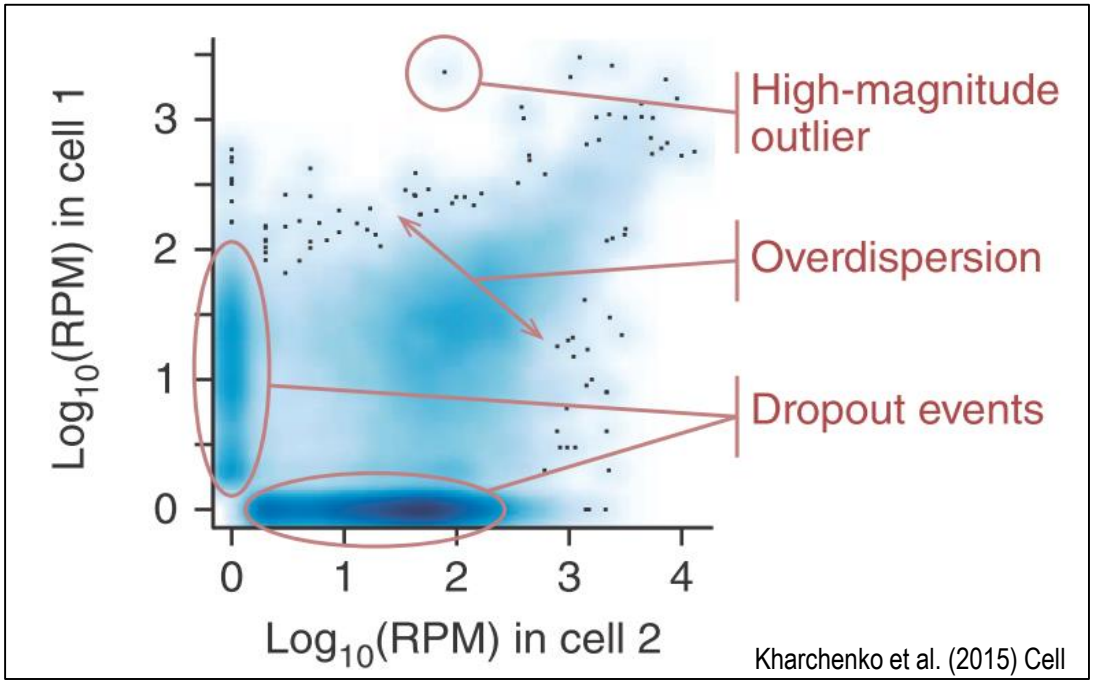
RPM: reads per million

Kharchenko et al. (2015) Cell

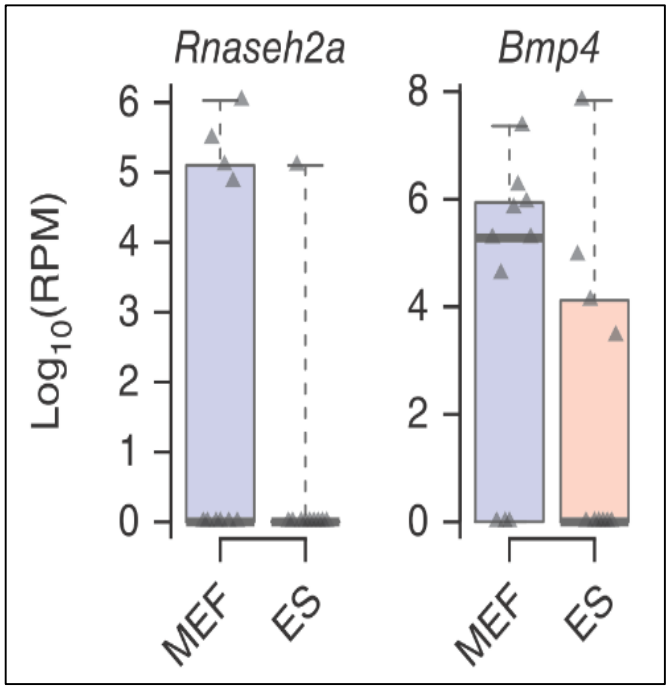
Gene expression estimated from two cells of the same type, illustrating the types of cell-to-cell variability observed.

Dropout events

Low starting amount of mRNA increases the probability of missing the transcript in the reverse transcription step. The expression of gene is observed in all cells excepted in one.



Kharchenko et al. (2015) Cell



Expression of Rnaseh2a and Bmp4: two top differentially expressed genes in MEF and ES samples.

=> The variance is high and caused by frequent dropout events.

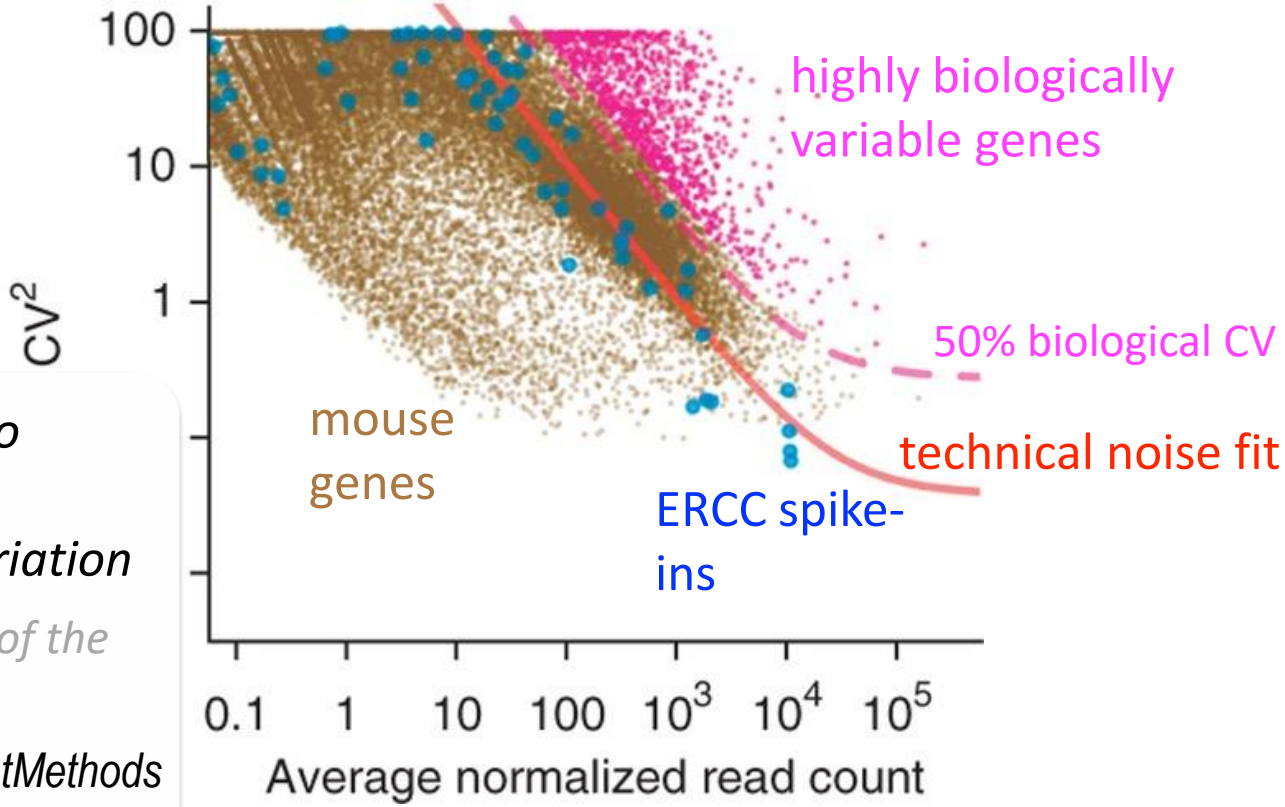
Solution: modeling of drop-out events (e.g. SCDE method or M3Drop)

scRNA-seq: estimation of technical variability

How much of the cell-cell variability is purely technical?
Use of ERCC spike-ins to distinguish technical from biological noise

CV: coefficient of variation

measure of dispersion of values of samples/populations around the mean



Statistical model to select only genes with high biological variation

generalized linear model of the gamma family

Brennecke & Heisler (2013) NatMethods
Method implemented in scLVM

Principle: estimate technical variation from spike-ins: for each mean - expect certain variation. Above this expectation => biological variation

scRNA-seq Computational Workflow

Quality control & filtering (I)

↓

Mapping to reference genome

↓

Quality control & filtering (II)

↓

Normalization, noise removal

↓

Study-specific downstream analyses
to generate new biological insight

Cell type identification
Cell type characterization
Gene network analysis
Kinetics of transcription

Study-specific downstream analyses

Read counts

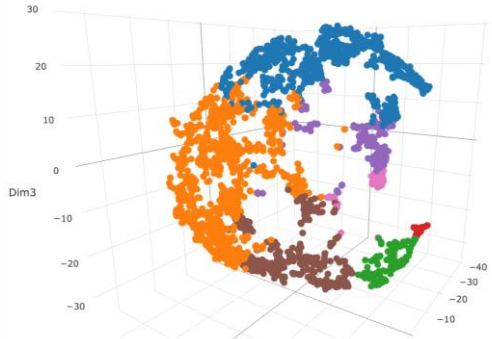
	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			
Spike 1	103	180	
Spike 2	13	19	
...			

4.

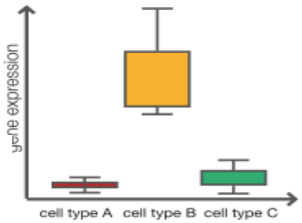
Filtering + Normalization

Cell type identification & characterisation

1. Cell grouping



2. Finding markers of cell type



3. Functional enrichment analysis

the Gene Ontology

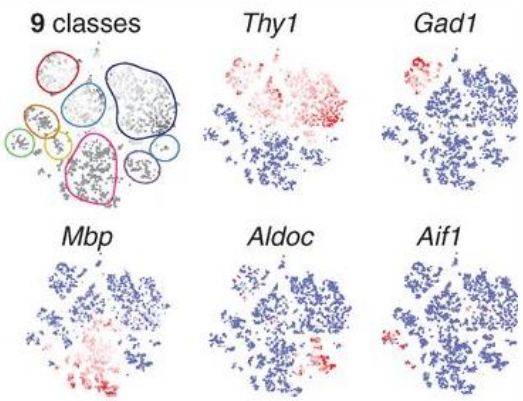
4. Trajectory analysis

adapted from Kolodziejczykl ...Teichman Cell 2015

T-distributed stochastic neighbor embedding (tSNE)

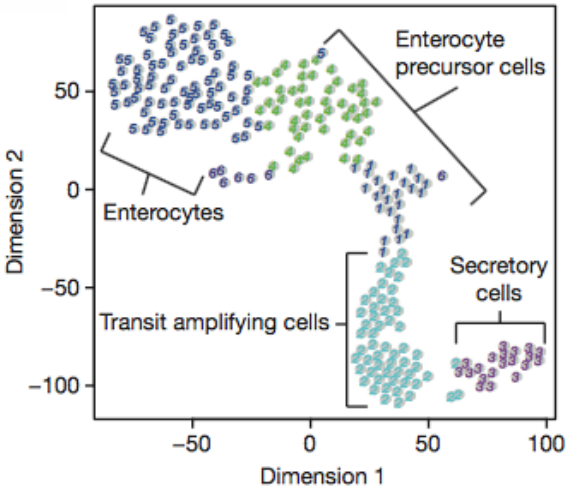
tSNE has been broadly applied to many scRNA-seq datasets in the past 1-2 years

Cell types in the mouse brain



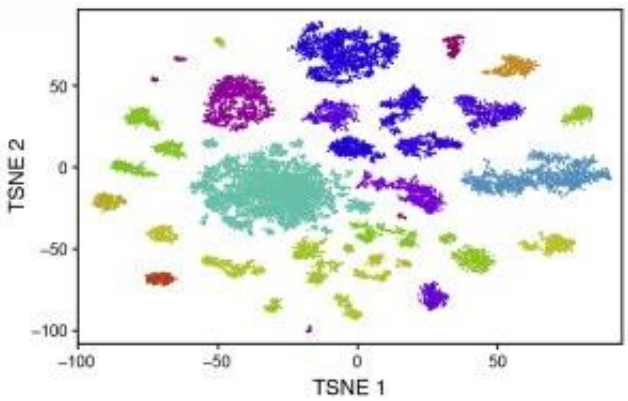
ZeiselLinnarsonScience2015

Cell types in the mouse gut



GrunOudenaardenCell2015

Cell types in the mouse retina



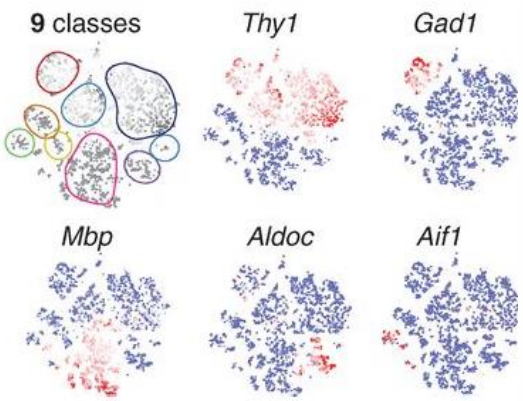
MacoskoMcCarrollCell2015

Did you use it before?
Can you cite at least one issue with t-SNE?

T-distributed stochastic neighbor embedding (tSNE)

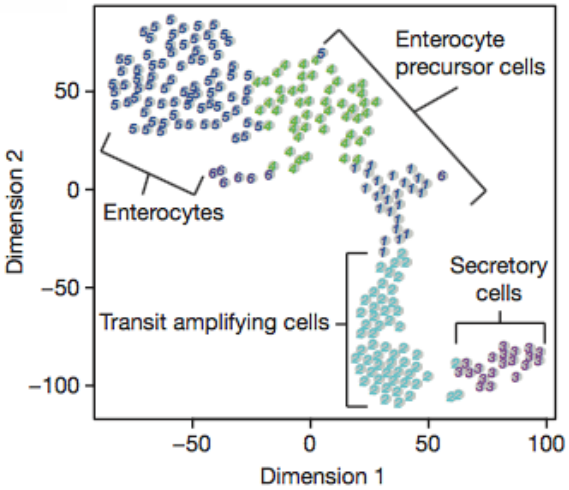
tSNE has been broadly applied to many scRNA-seq datasets in the past 1-2 years

Cell types in the mouse brain



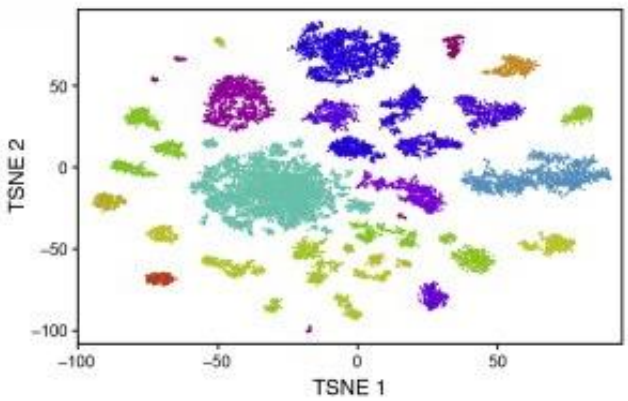
ZeiselLinnarsonScience2015

Cell types in the mouse gut



GrunOudenaardenCell2015

Cell types in the mouse retina



MacoskoMcCarrollCell2015

Did you use it before?
Can you cite at least one issue with t-SNE?

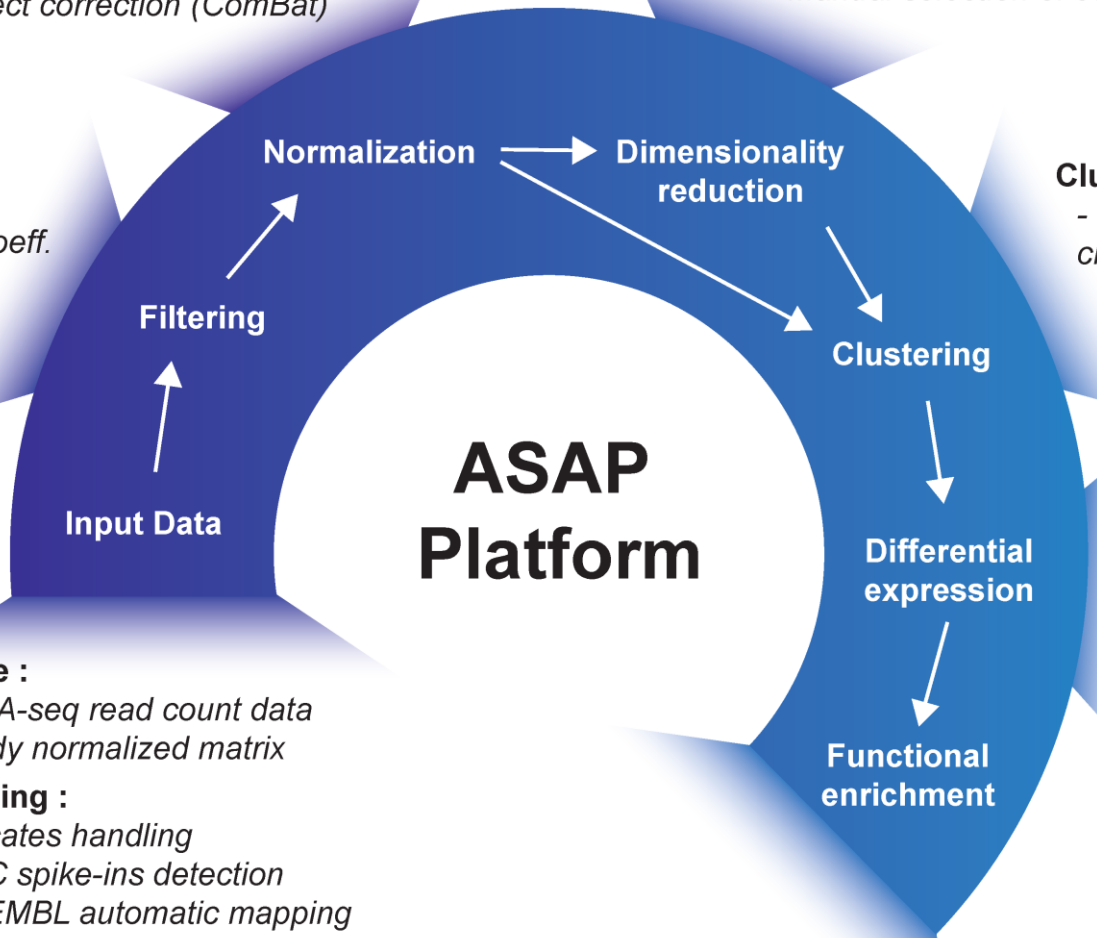
- It's stochastic, i.e. running it several times will generate different results
- Distances between cells is arbitrary and cannot be related to any metric distance (as compared to PCA)
- Perplexity parameter is somewhat very vague and complex to tune
- Running t-SNE with different number of components changes the results

ASAP: Automated Single-cell Analysis Pipeline

Normalization :
- *Scaling, Log2, RPKM*
- *Voom, TMM, DESeq2*
- *scLVM (can use spike ins)*
- *Batch effect correction (ComBat)*

2D and 3D interactive visualization for :
- *PCA, tSNE, MDS, ZIFA*
- *Cell colouring according to expression or clustering*
- *Manual selection of cells from the plots*

Filtering :
- *Expression, Coeff. of Var., CPM*
- *PAGODA*
- *SCAN/UPC*



Clustering algorithms :
- *K-Means, PAM, Hierarchical Clustering, SC3*

DE algorithms :
- *Marker genes or 2 groups comparison*
- *limma-Voom, edgeR, DESeq2, SCDE*

Input file :
- *scRNA-seq read count data*
- *already normalized matrix*

Processing :
- *duplicates handling*
- *ERCC spike-ins detection*
- *ENSEMBL automatic mapping*

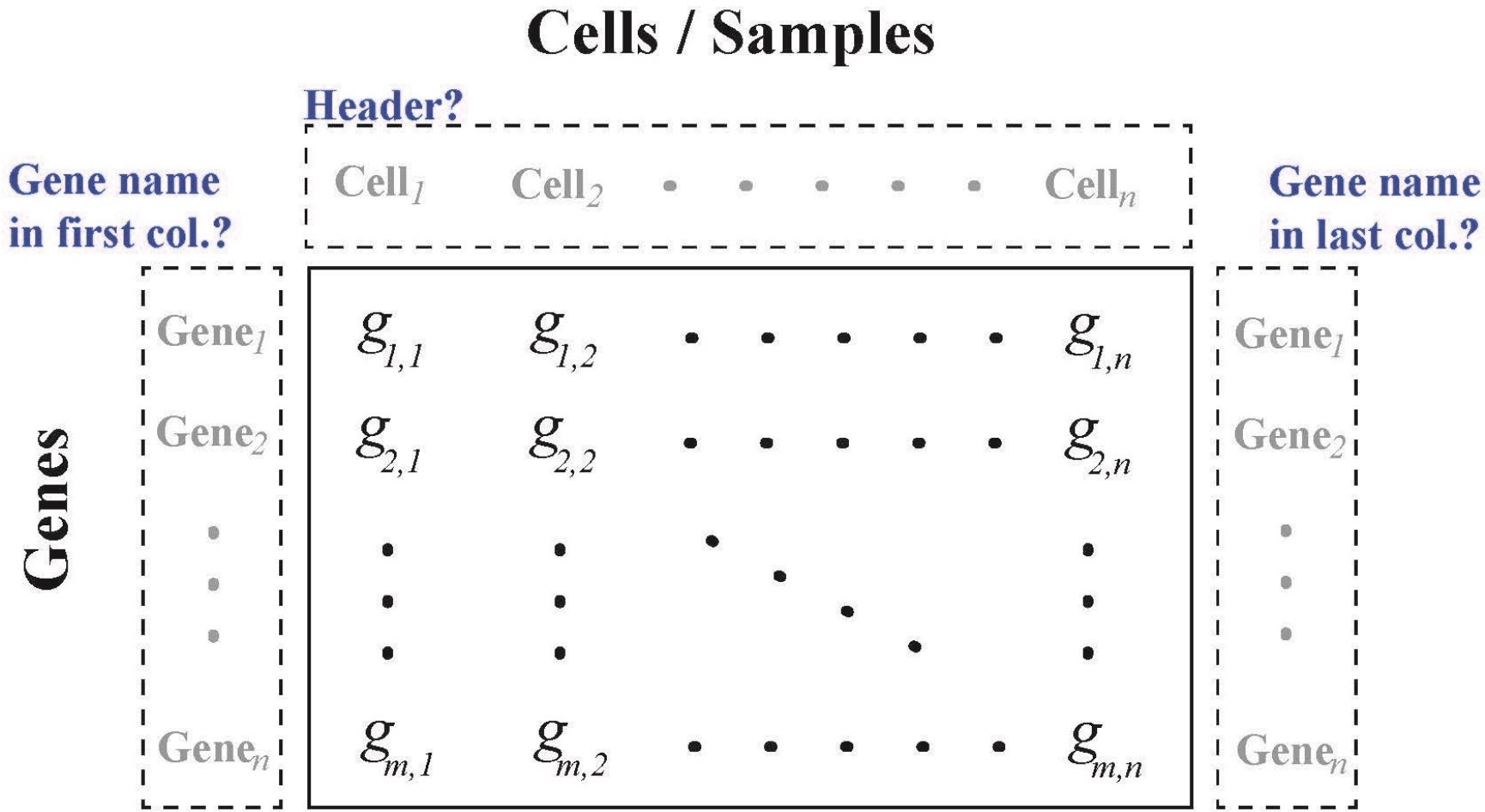
Functional annotation databases :
- *Gene atlas, Gene ontology (GO), KEGG Pathways*

And more...
- *Trajectory analyses*
- *Imputation*
- *scMAP*
- *Velocity*

Published in Bioinformatics: https://doi.org/10.1093/bioinformatics/btx337

What kind of data is output from the preprocessing step?

ASAP input file: count matrix or already normalized gene expression matrix
=> Soon, possibility of sending HDF5 (.h5) 10x files and .loom files



Other software analysis pipelines vs ASAP…

Existing software:
MAST, PAGODA, SCell, Seurat, FastProject, FastGenomics

Disadvantages:

- Restricted set of algorithms and methods
- **Lack of interactivity** and visualization
- Required knowledge of **programming** and/or statistics
- Require **installation of libraries and dependencies**
- Depends on personal computer processing power

ASAP

Combines state-of-the art single-cell specific algorithms written in R, Python and Java

- **Interactive** and user-friendly web interface with 2D/3D visualization
- Nothing to install for the end user (**web-based, no command-line**)
- **Sharing projects** and publishing analysis for complete reproducibility
- **Light for the user** => everything run on the server

Granatum: Zhu et al. (Dec 2017) *Genome Medicine*
MAST: Finak et al. (2015) *Genome biology*
PAGODA: Kharchenko et al. (2014) *Nature Methods*
SCell: Diaz et al. (2016) *Bioinformatics*
FastProject: DeTomaso et al. (2016) *BMC bioinformatics*
FastGenomics: wp.fastgenomics.org

Technically, what is ASAP?

- **Centralized computational resources**: Ruby-on-rails server currently hosted at the EPFL

- Implementation of the "*delayed-jobs*" gem that allows job queuing management

- Single-cell analysis scripts are written in R (mostly), Python (dimension reduction) and Java (parsing, functional enrichment)

⇒Generates JSON files that are interpreted by the browser

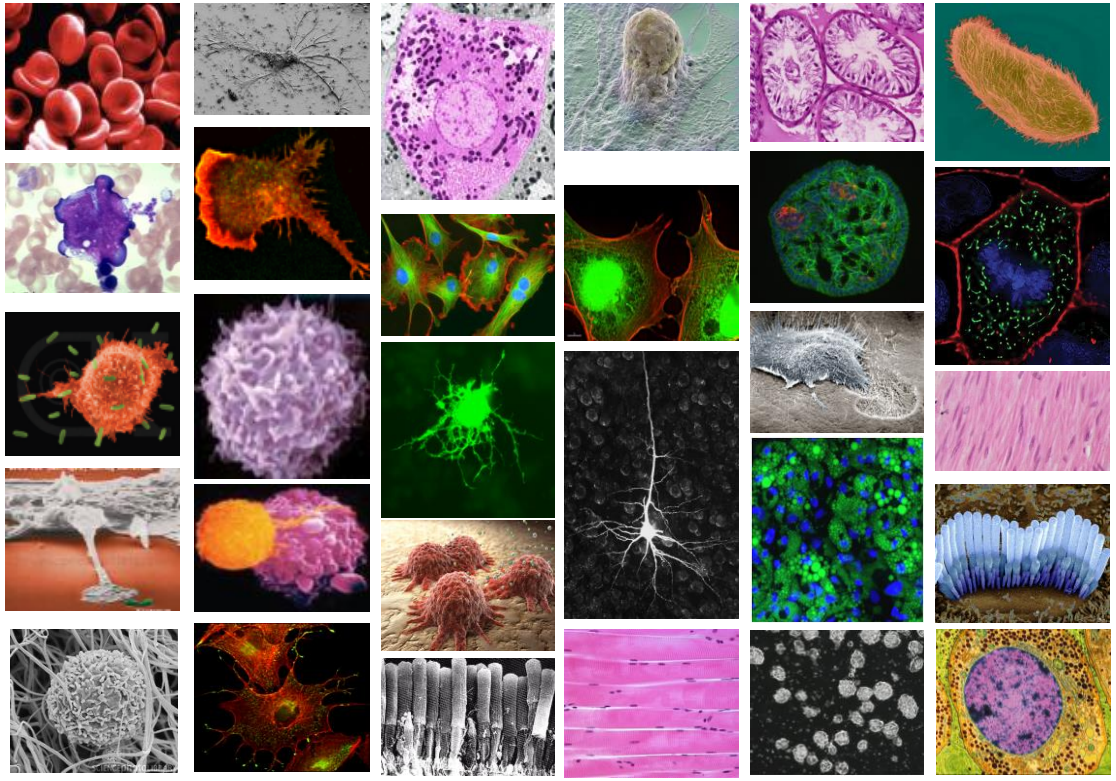
- **Interactive** and user-friendly web interface with 2D/3D visualization (using plotly JS) [currently moving to plotly webGL]

External needs for such a platform?

ASAP is supported by the Chan Zuckerberg Initiative (CZI)

- **Cell atlas initiative**
- Infectious disease initiative

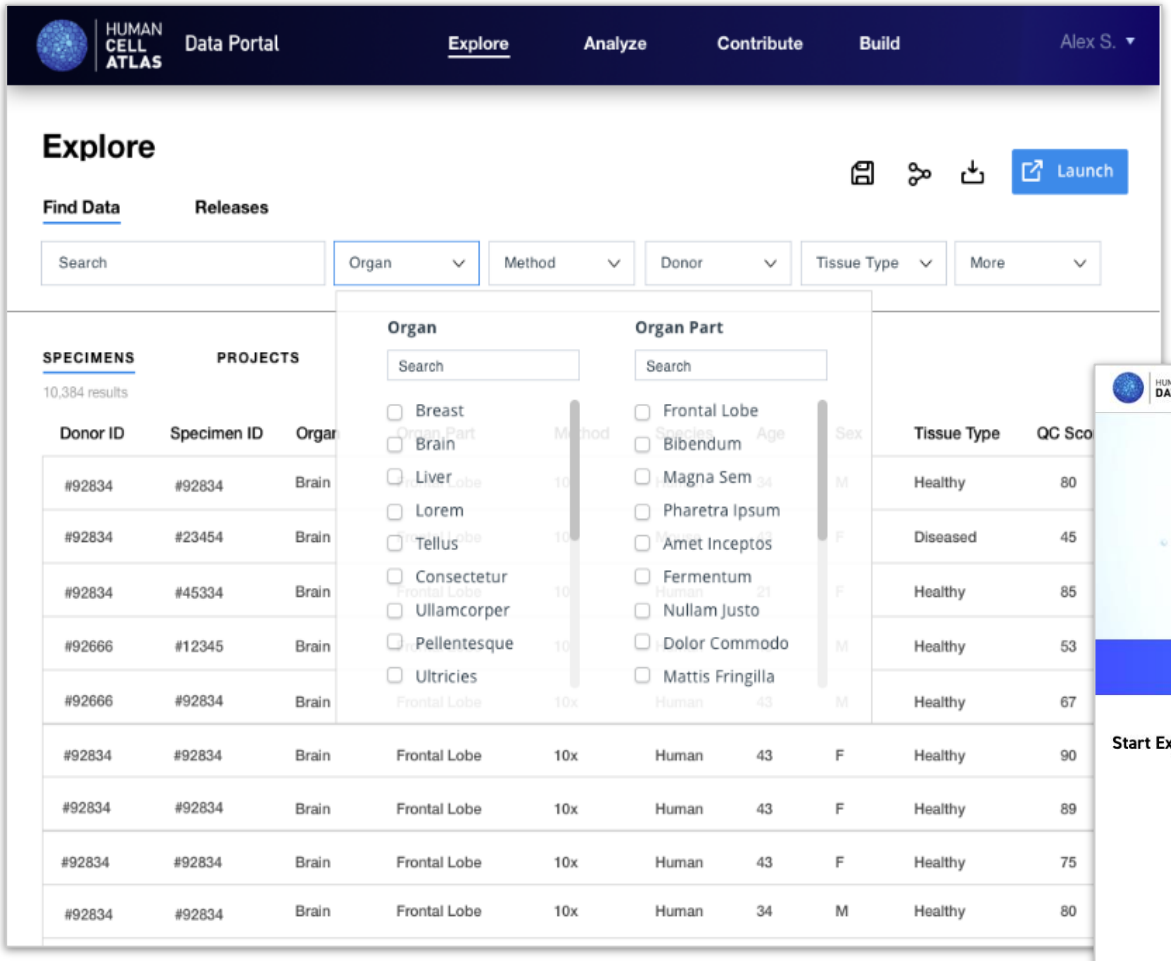
HUMAN CELL ATLAS

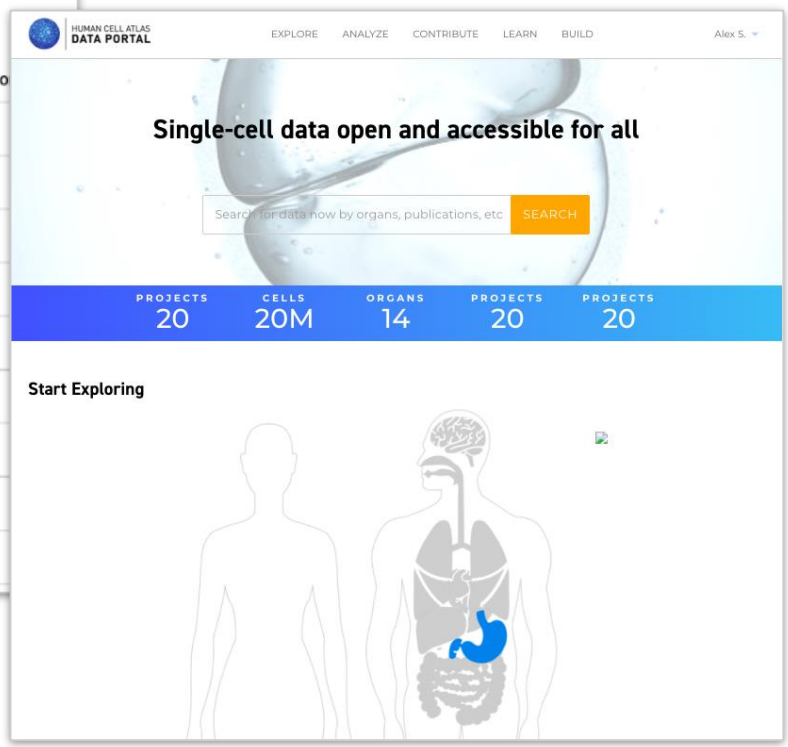


A free, open reference map of all cells in the healthy human body

www.humancellatlas.org

HCA web-based data browser

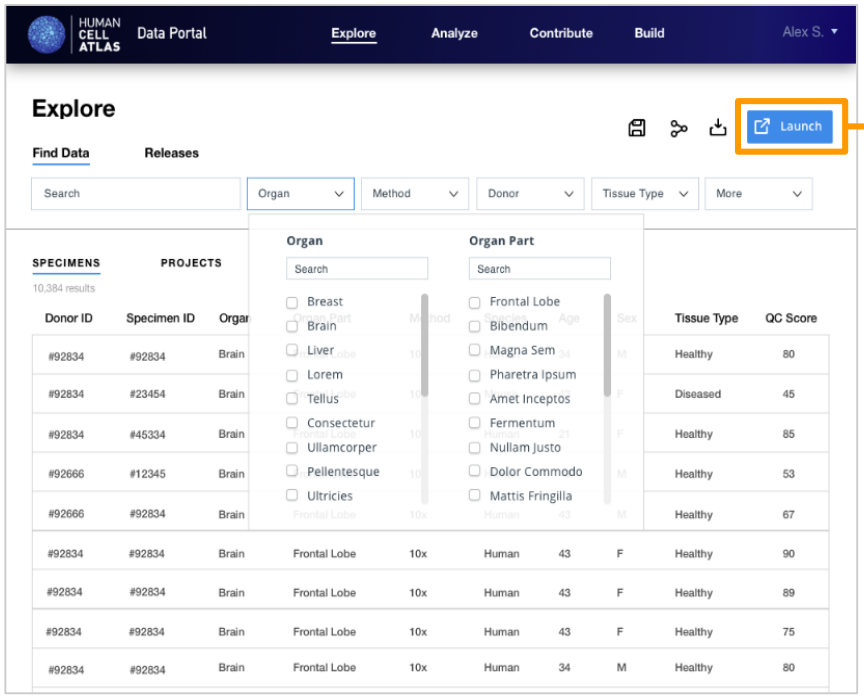


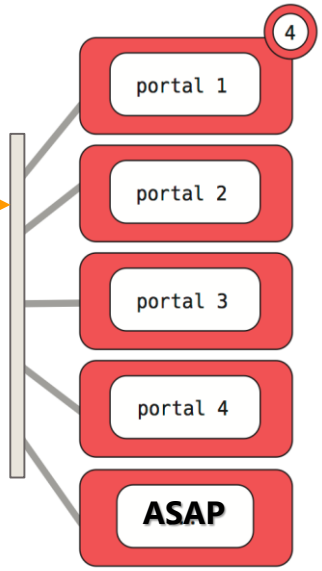


preview.data.humancellatlas.org

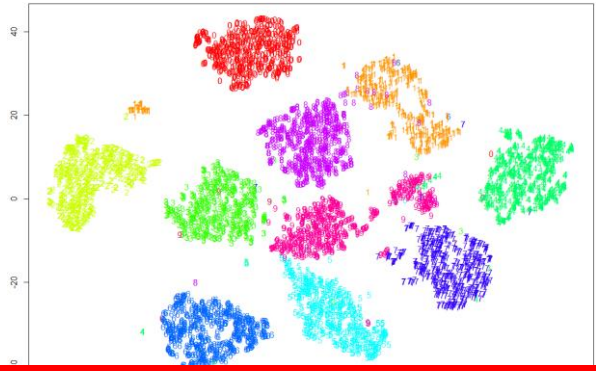
Goal: connect to community tools/portals

- "Handoff" search results to Community Tools for analysis/visualization





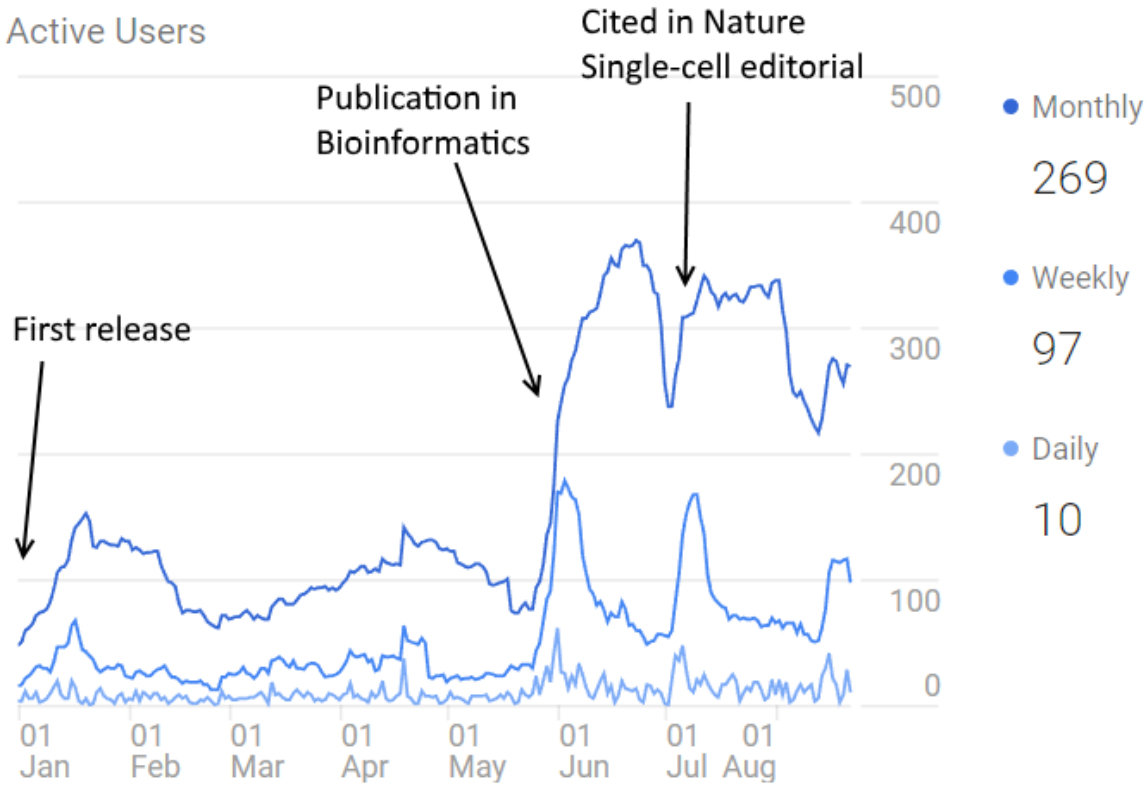
multiple portals
and tertiary
analysis tools



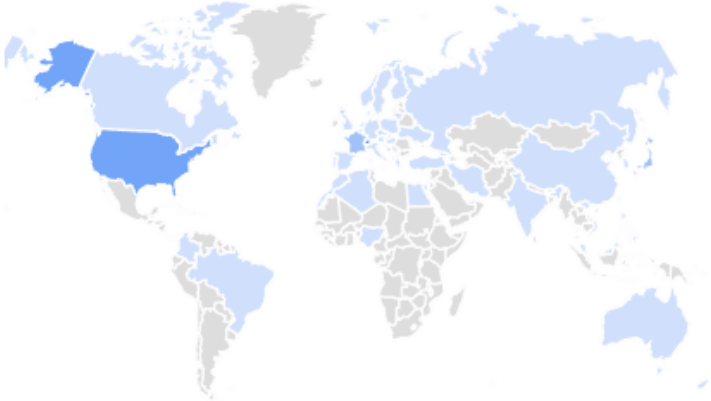
ASAP usage since first release

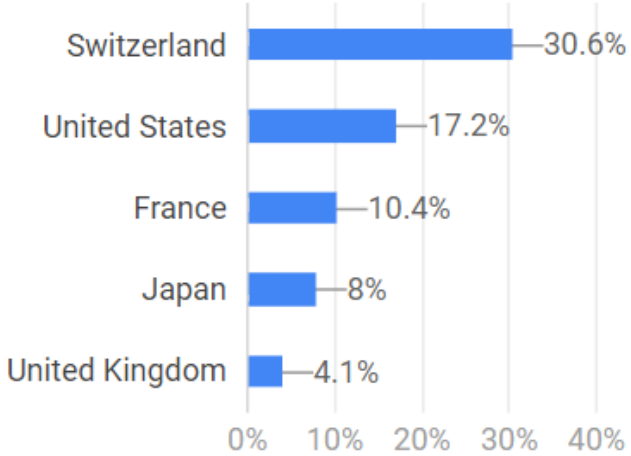
Users	Sessions	Bounce Rate	Session Duration
1.2K	3.7K	37.31%	9m 0s
↑2,941.5%	↑3,870.7%	↓1.9%	↓9.9%

Active Users



Sessions by country





Example – Hands on

https://asap.epfl.ch

Public 10x dataset already uploaded on ASAP:

http://cf.10xgenomics.com/samples/cell-exp/2.1.0/t_3k/t_3k_web_summary.html

Can be viewed/cloned as a public project by anyone

Current challenges in scRNA-seq

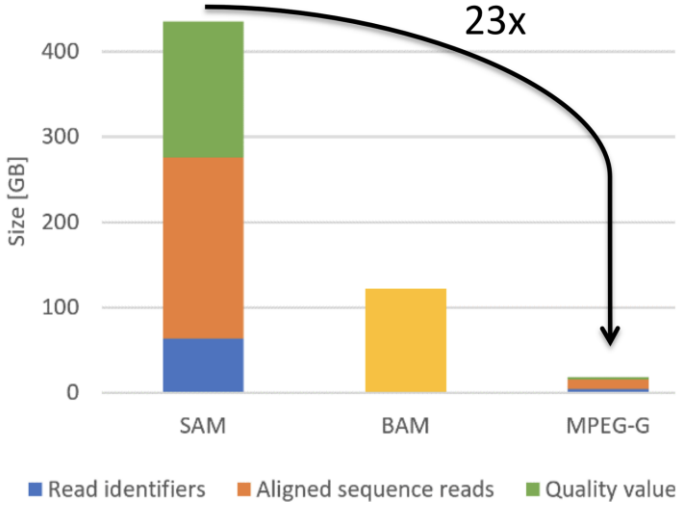
- **Manifold alignment:** Define novel methods for integration of multiomics/multiplatform datasets (e.g. batch effect)
- **Scaling:** HCA plans to generate datasets of > 10 billions cells. How to t-SNE that??

⇒Cloud computing, scalable methods (scanpy, Seurat?), out-of-RAM computation

- **Compression**

⇒Standardized data format for .fastq/BAM files (MPEG-G)

⇒Data formats (HDF5, loom?)



Most of this involves benchmarking

Future developments for ASAP

- Add the ability to share projects and work simultaneously on the same project
- Add new tools / algorithms as they are now published (M3DROP, MAGIC, scanpy, scMap, …)
- Add other databases for functional enrichment (pharmgkb, oncogenes, …)

Scalability

- Being able to display ultra huge datasets (10 billion cells?)
- Implement HDF5/Loom for storing files and faster access/out-of-RAM computations

Future developments for ASAP

- ~~Add the ability to share projects and work simultaneously on the same project~~ **Released on Monday**
- Add new tools / algorithms as they are now published (M3DROP, MAGIC, scanpy, scMap, UMAP…)
- Add other databases for functional enrichment (pharmgkb, oncogenes, …)

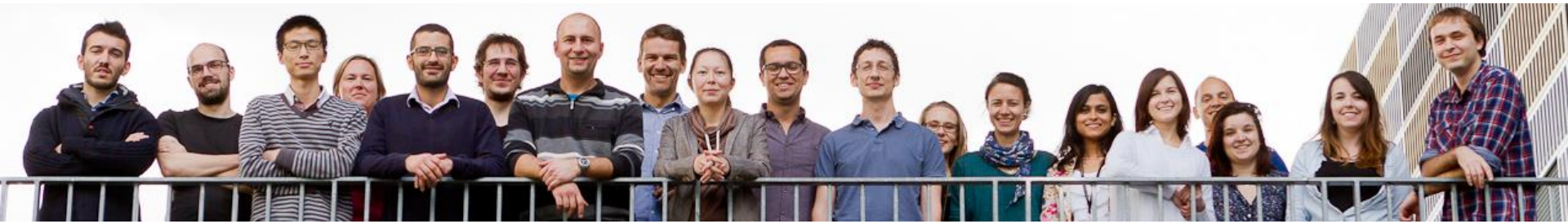
Scalability

- Being able to display ultra huge datasets (10 billion cells?)
- Implement HDF5/Loom for storing files and faster access/out-of-RAM computations

Currently under implementation

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Laboratory of Systems Biology and Genetics



@Deplancke's lab

Thanks

Vincent Gardeux