Single-Cell RNA-Seq Module
Introductory Lecture

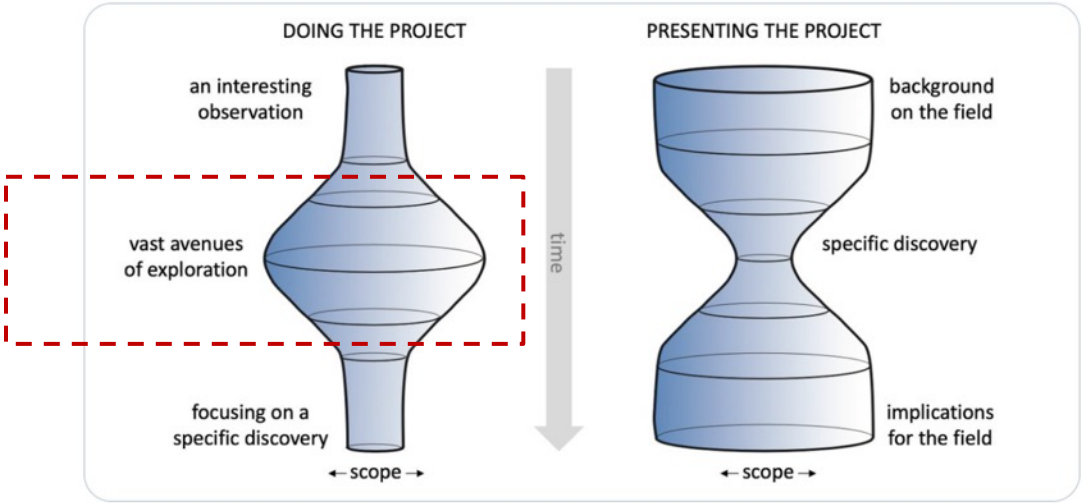
Igor Dolgalev

Bioinformatics (Fall 2021)

December 7, 2021

Itai Yanai
@ItaiYanai

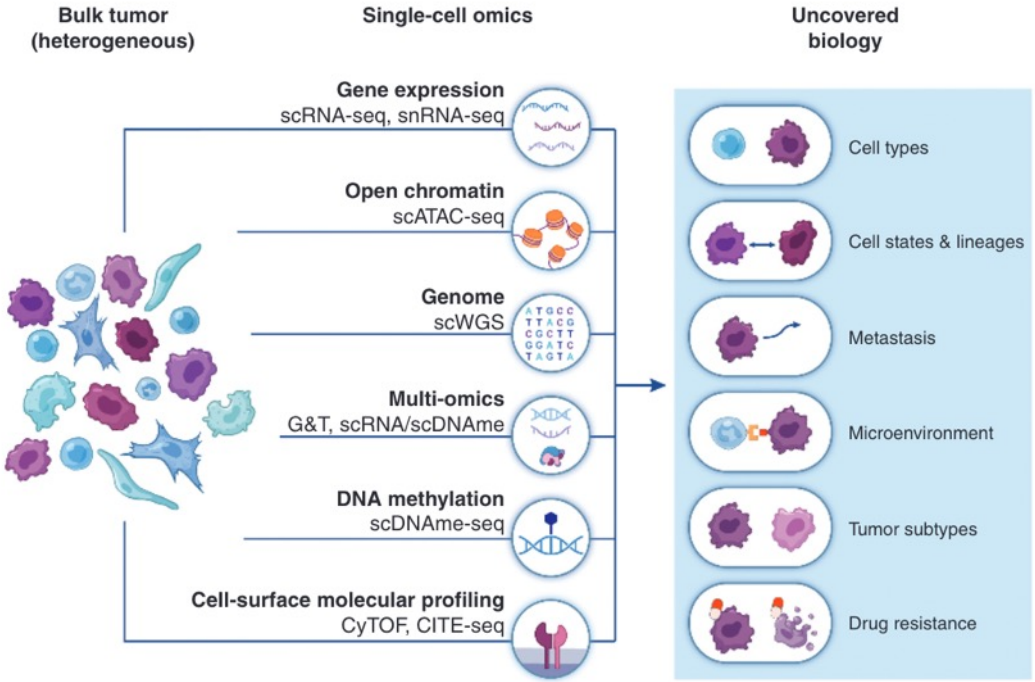
The difference between doing a project and presenting it. An observation can lead to many avenues of explorations before focus turns to a specific discovery. Presenting it, in a talk / paper, follows inversely, with broad perspectives coming before & after the specific discovery.



Background

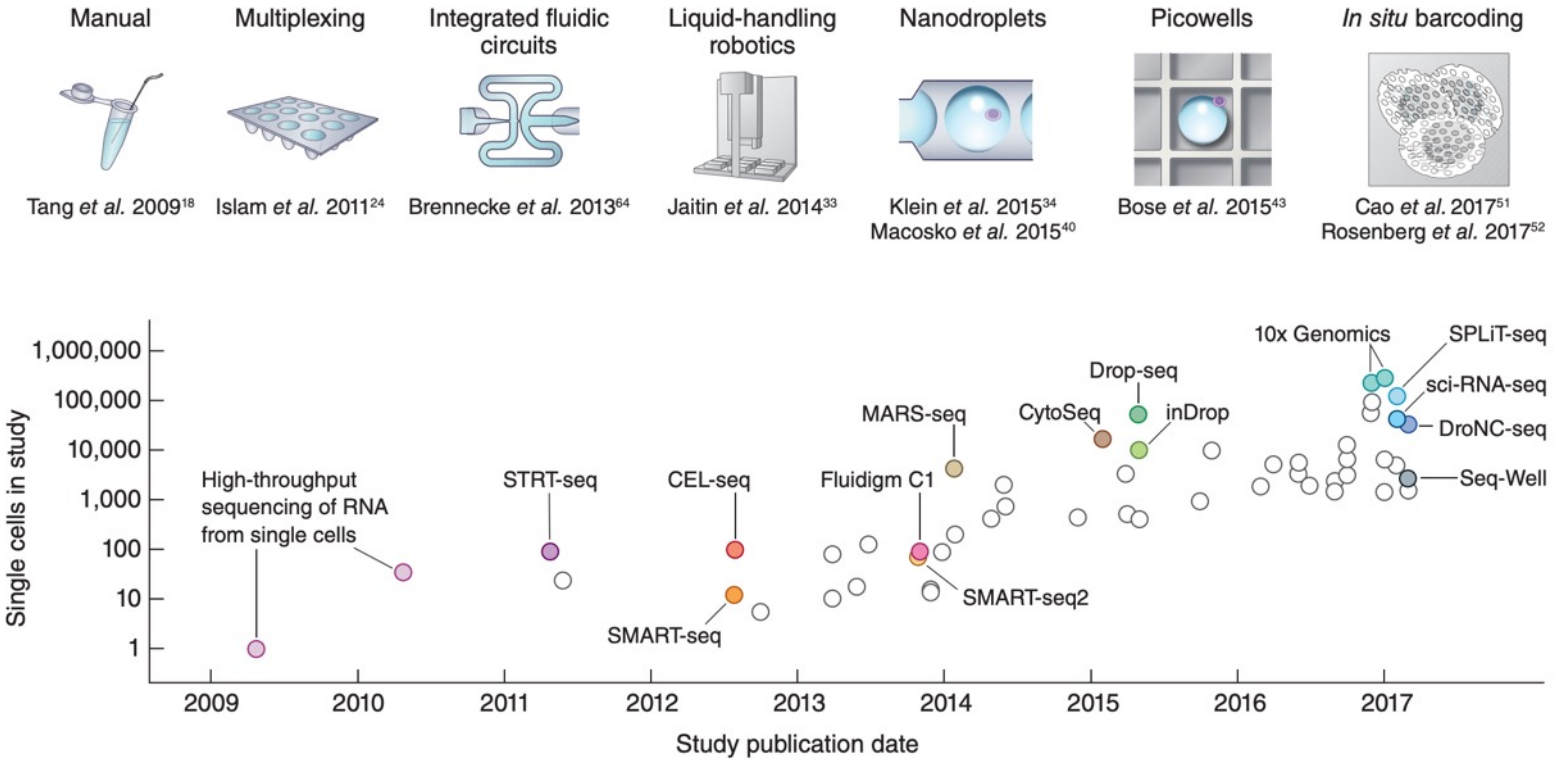
- Microscope improvements allowed Robert Hooke and Anton van Leeuwenhoek to describe the cell as the structural unit of life in the 17th century.
- The cell can be viewed as the structural and functional unit of life.
- A cell's function is based on its molecular composition, so it should be possible to define cell types on the basis of gene expression.
- Distinct gene expression programs can drive cellular activity and identity.
- Single-cell proteomic technologies (assaying the final functional product of gene expression) are limited to a pre-selected set of markers.
- Assay of the full transcriptome in single cells provides a means to classify and characterize them at the molecular level.

Biological information at single-cell resolution



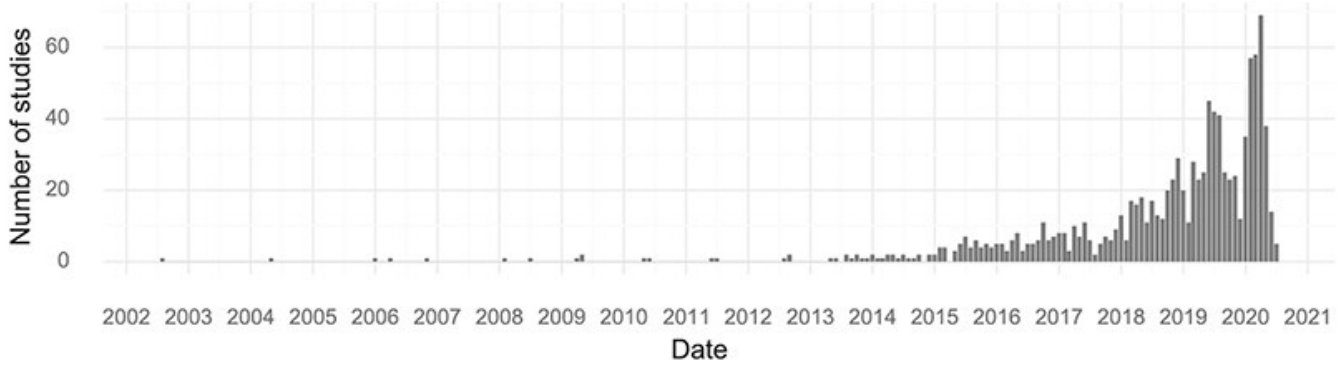
Castro et al, *Cancer Discovery* 2021

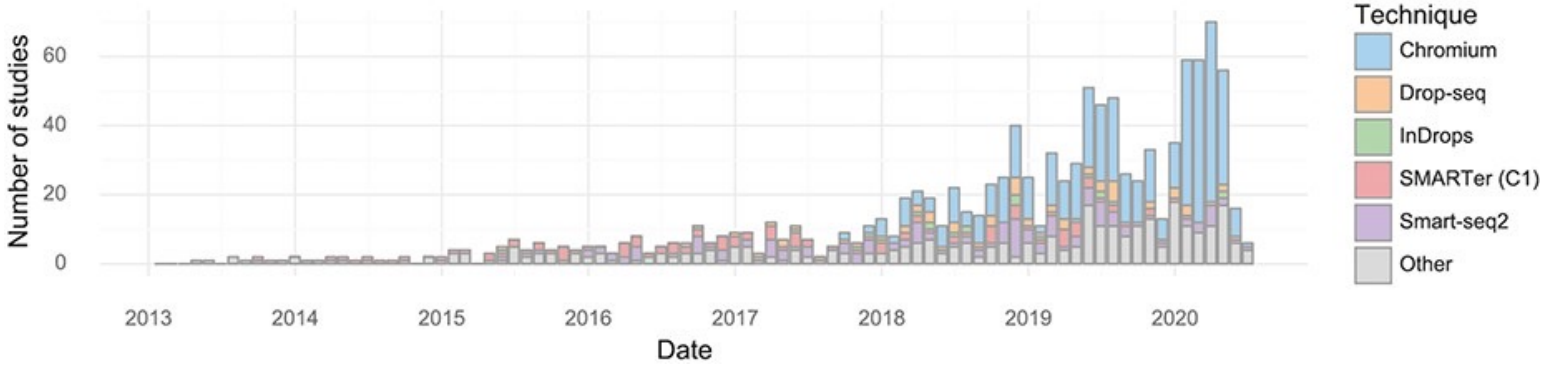
Exponential growth of scale in single-cell transcriptomic experiments



Svensson et al, *Nature Protocols* 2018

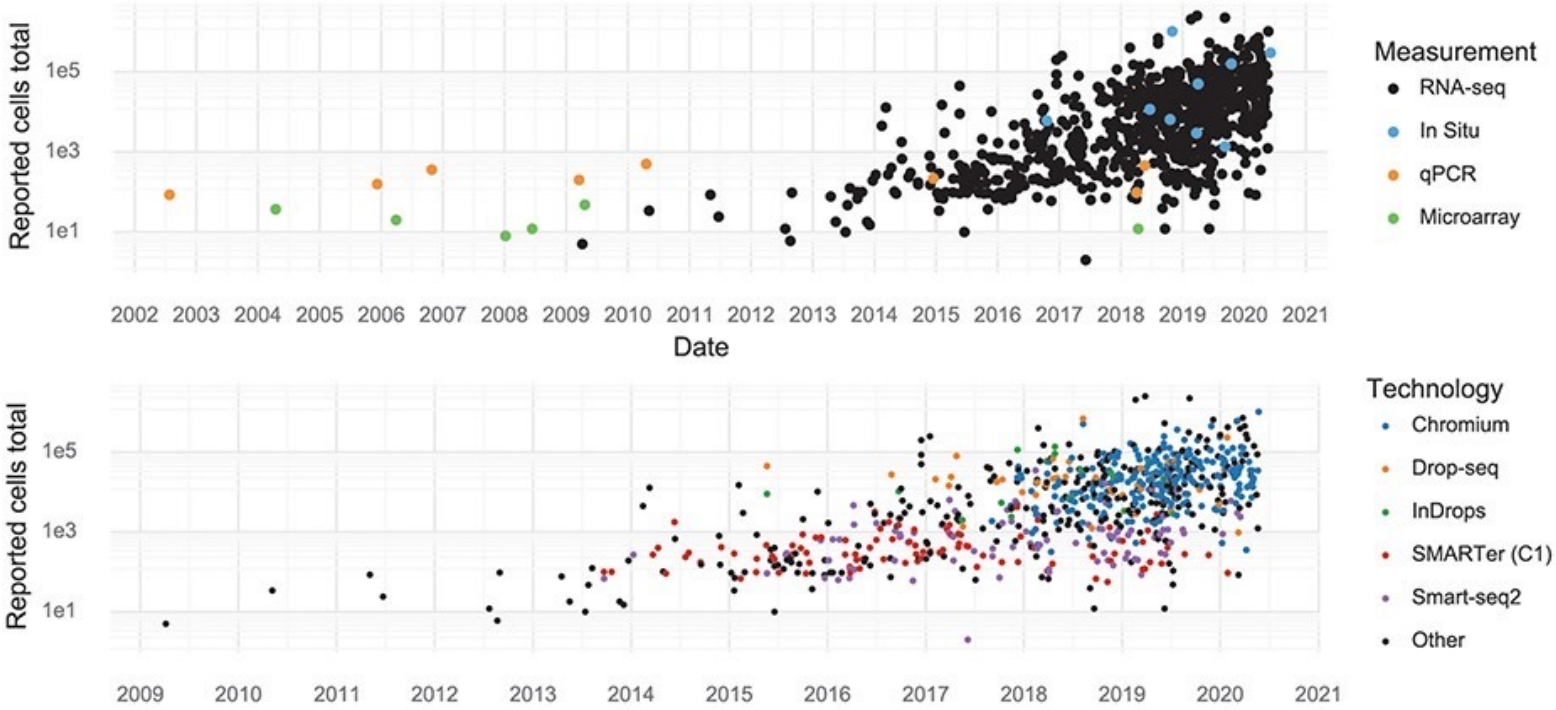
Single-cell study trends





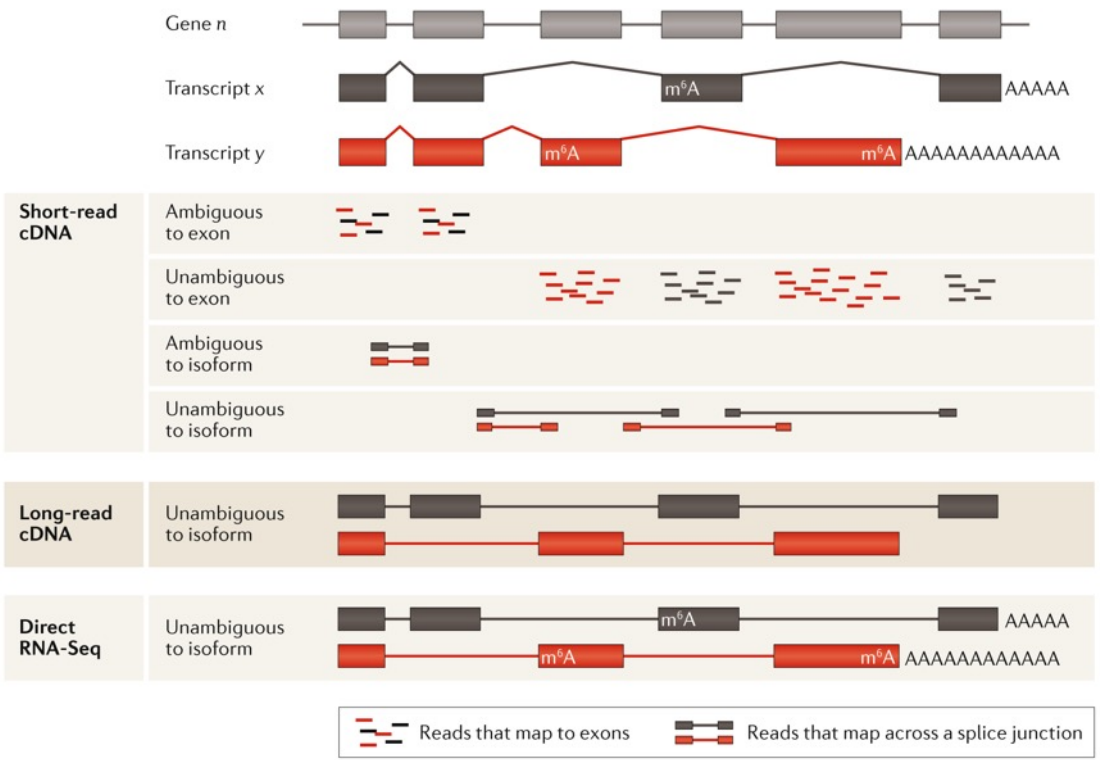
Svensson et al, *Database (Oxford)* 2020

Scale of experiments and data over time



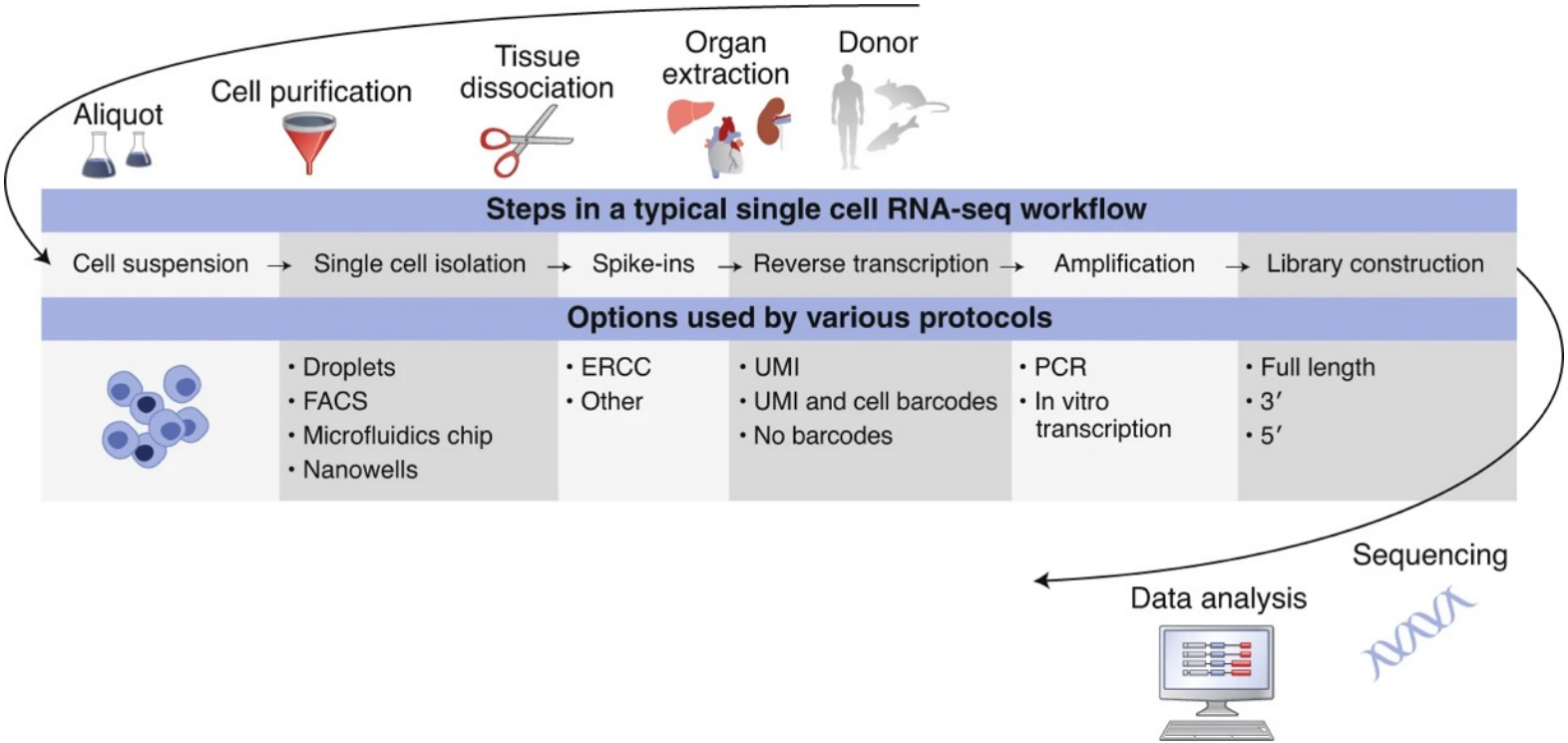
Svensson et al, *Database (Oxford)* 2020

Short-read, long-read, and direct (single-molecule) RNA-seq



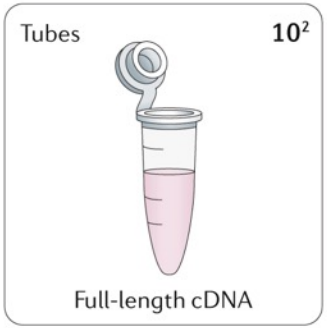
Stark et al, *Nature Reviews Genetics* 2019

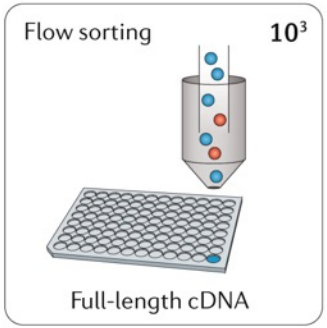
Typical steps of scRNA-seq experiments

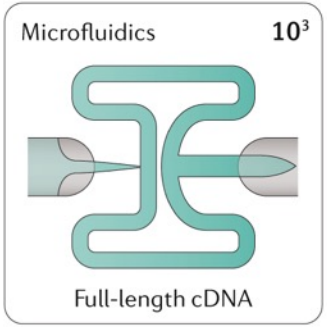


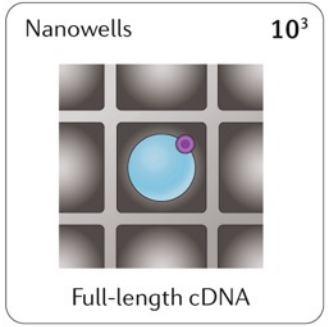
Fullgrabe, *Nature Biotechnology* 2020

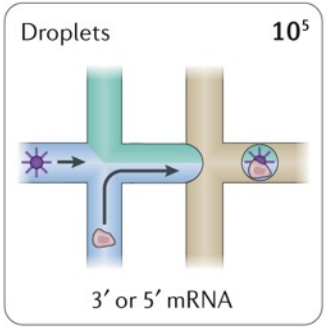
Single-cell capture methods

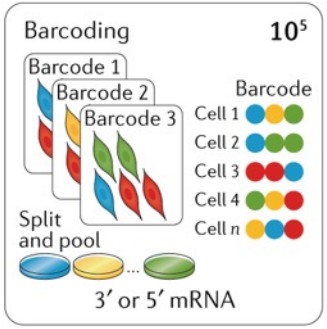






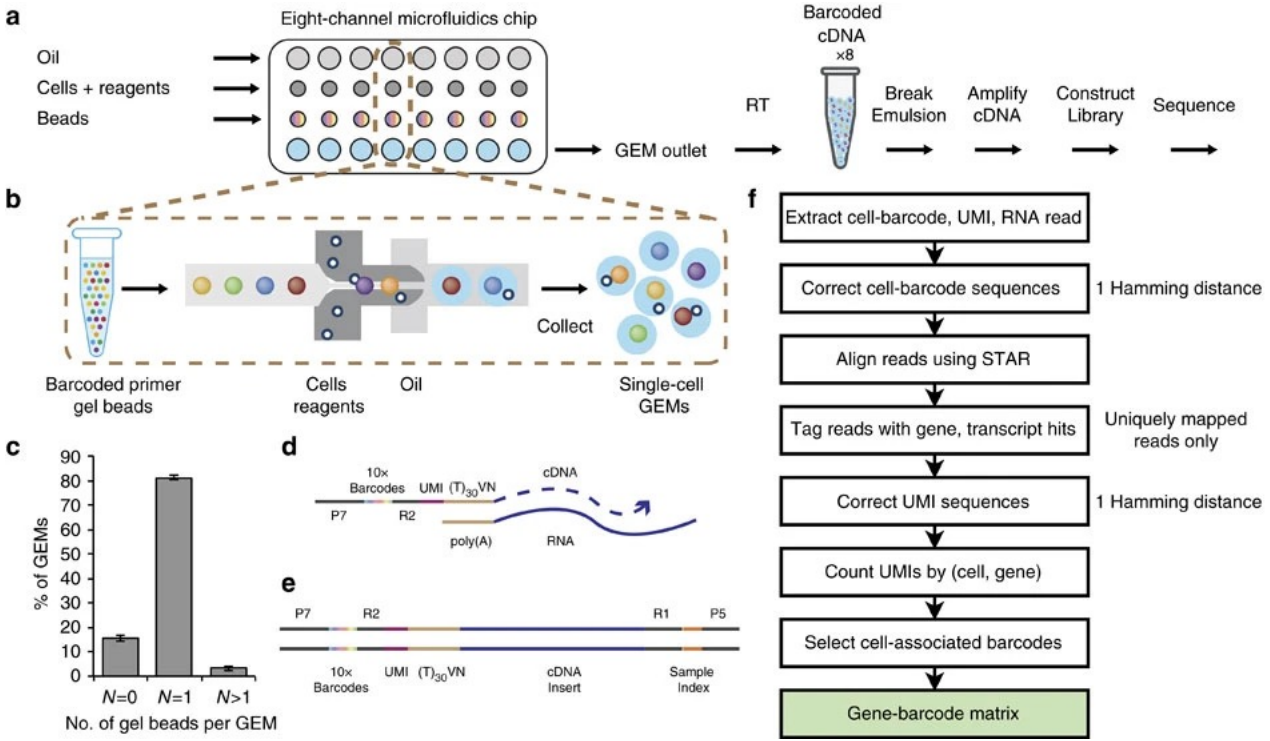






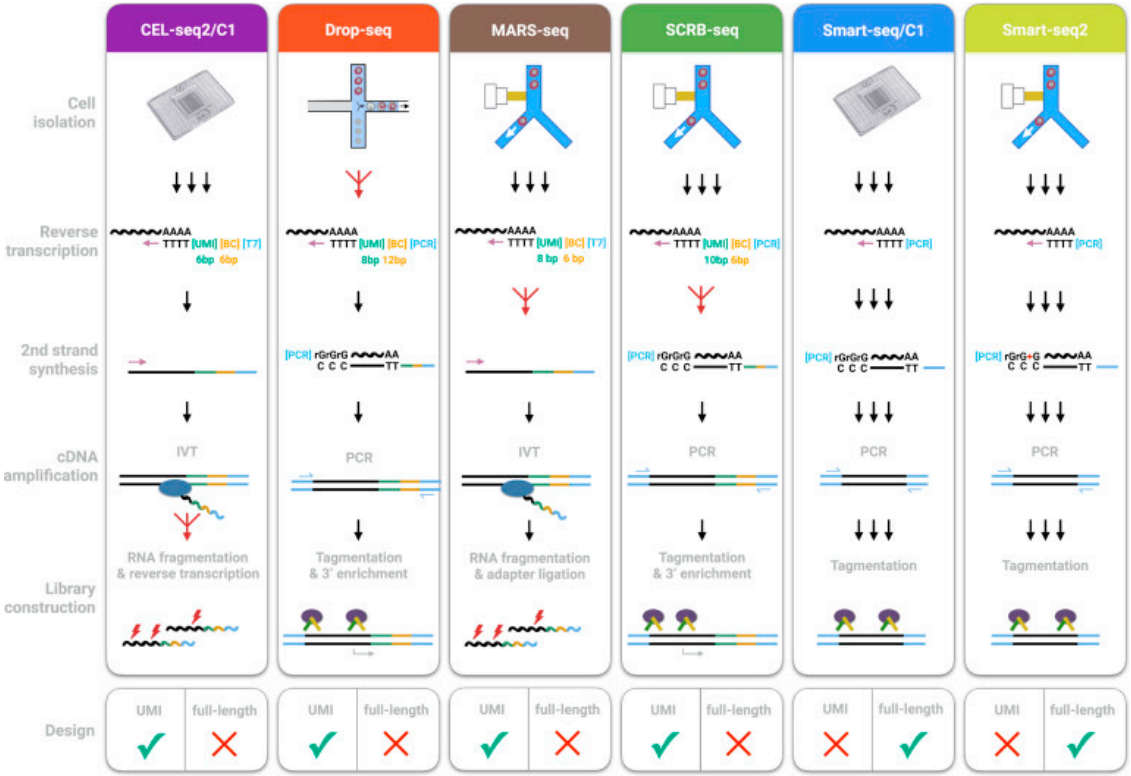
Stark et al, *Nature Reviews Genetics* 2019

10x Genomics GemCode technology



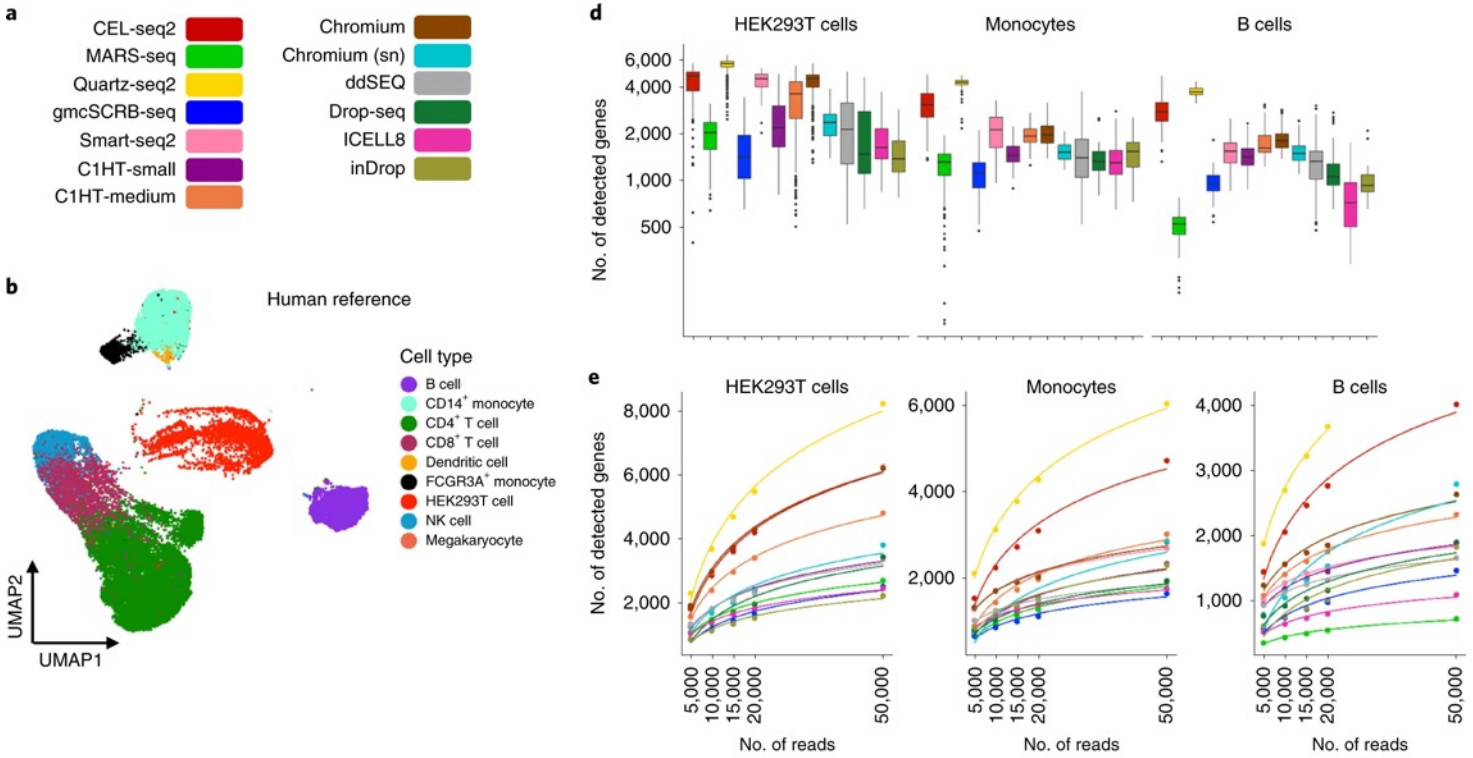
Zheng et al, *Nature Communications* 2017

Schematic overview of library preparation steps



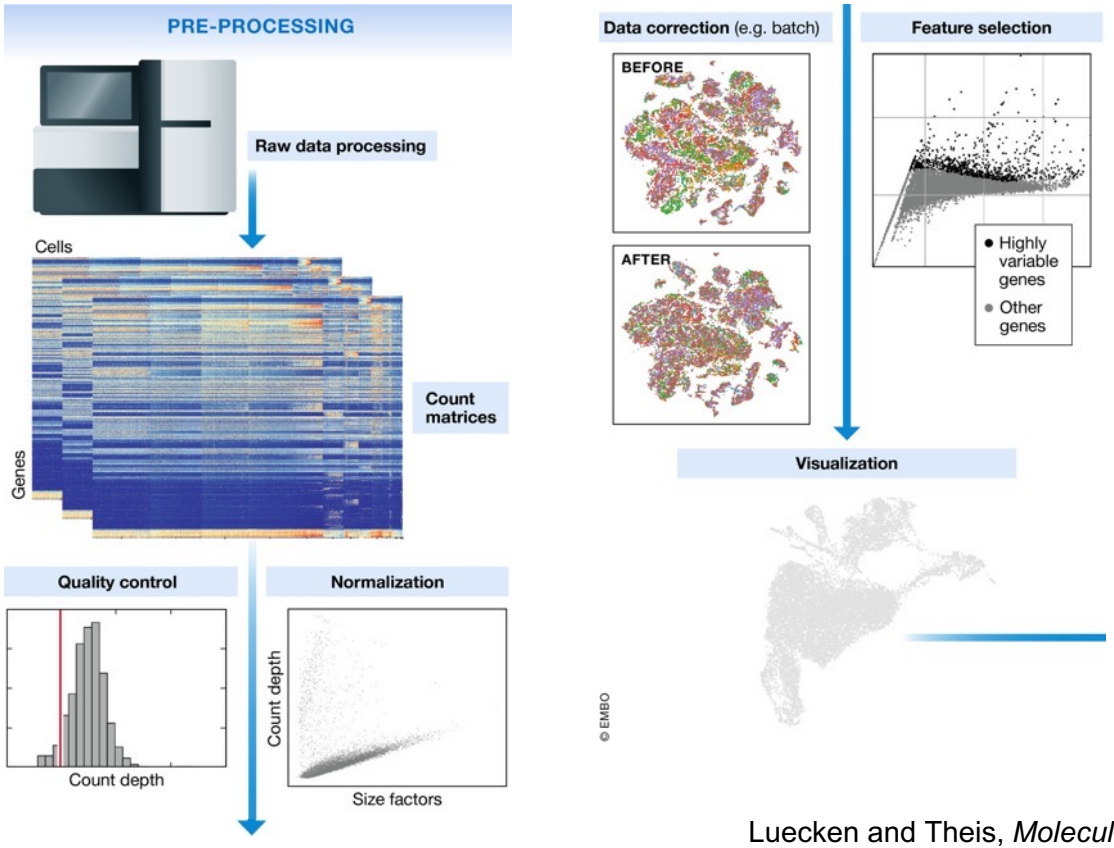
Ziegenhain et al, *Molecular Cell* 2017

Comparison of 13 sc/snRNA-seq methods



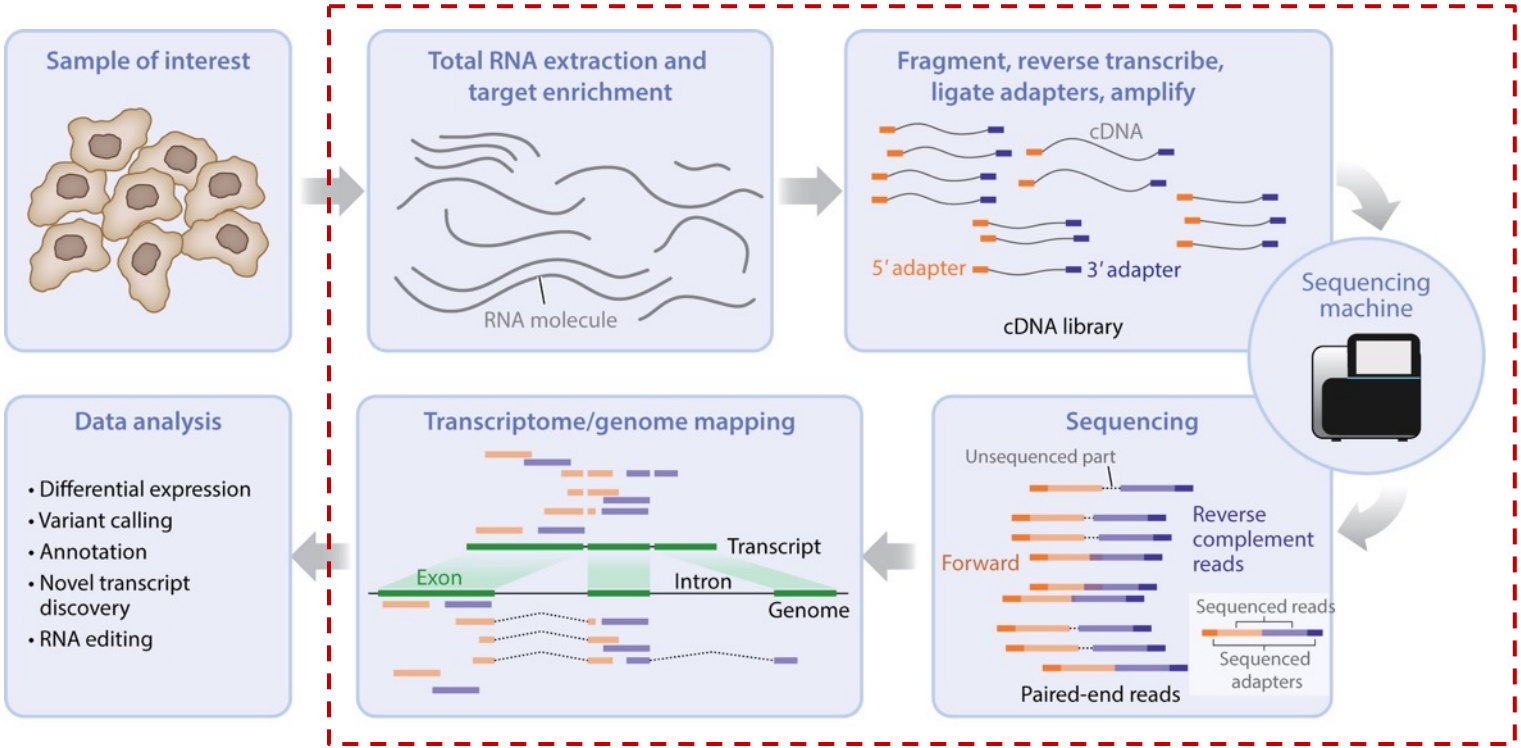
Mereu et al, *Nature Biotechnology* 2020

Data analysis first half: pre-processing and visualization



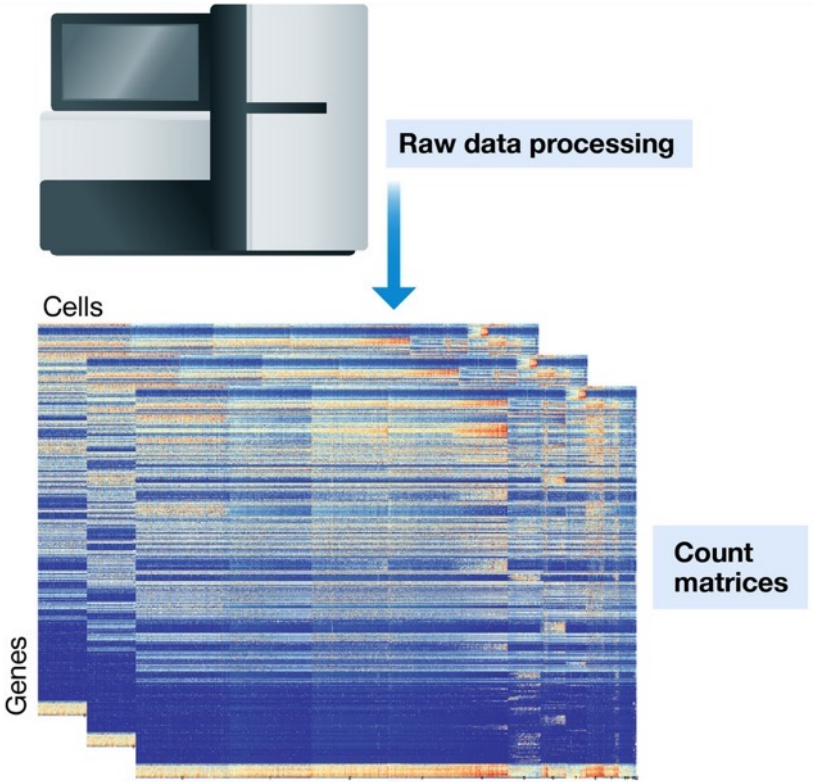
Luecken and Theis, *Molecular Systems Biology* 2019

A basic overview of the steps in a standard (bulk) RNA-seq experiment



Van den Berge K, et al. 2019.
Annu. Rev. Biomed. Data Sci. 2:139–73

Building a counts matrix



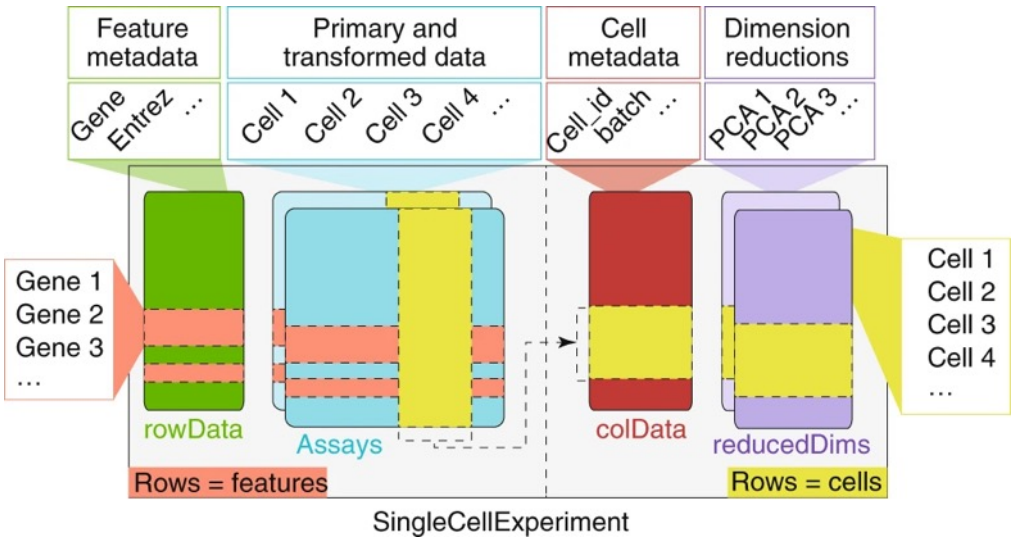
Cell Ranger, STARsolo, Salmon/Alevin, Kallisto, indrops, SEQC, zUMIs, etc.

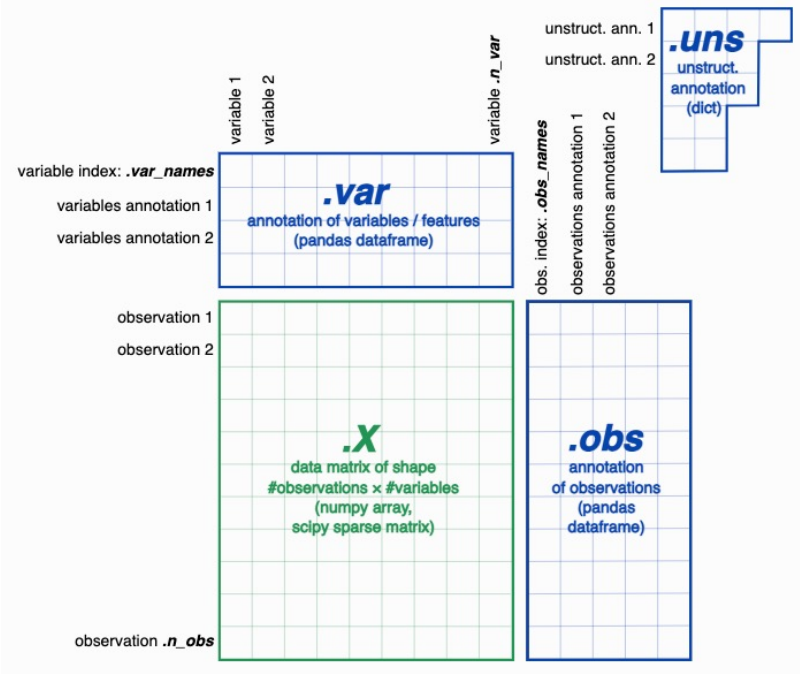
Analysis (QC, dimensionality reduction, visualization, clustering, differential expression) tools

- Analysis packages:
  - Seurat
  - scater (scran/scuttle/SingleCellExperiment)
  - scanpy (anndata)

- Workflows/tutorials:
  - Seurat vignettes
    - https://satijalab.org/seurat/articles/get_started.html
  - Orchestrating Single-Cell Analysis with Bioconductor (OSCA)
    - https://bioconductor.org/books/release/OSCA/

Objects for storing single-cell assay data and metadata

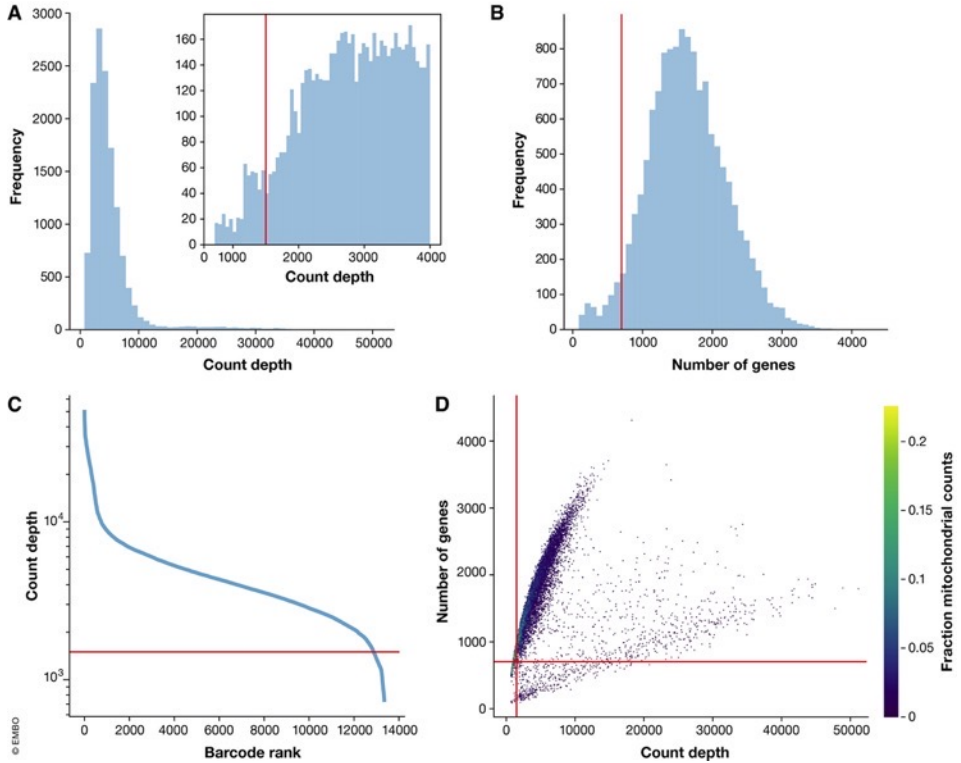




Amezquita et al, *Nature Methods* 2019
https://anndata.readthedocs.io/

Quality control and filtering

- Potential issues:
  - Dying cells
  - Doublets/multiplets
  - Empty droplets
- QC covariates:
  - number of counts per barcode (count depth)
  - number of genes per barcode
  - fraction of counts from mitochondrial genes per barcode



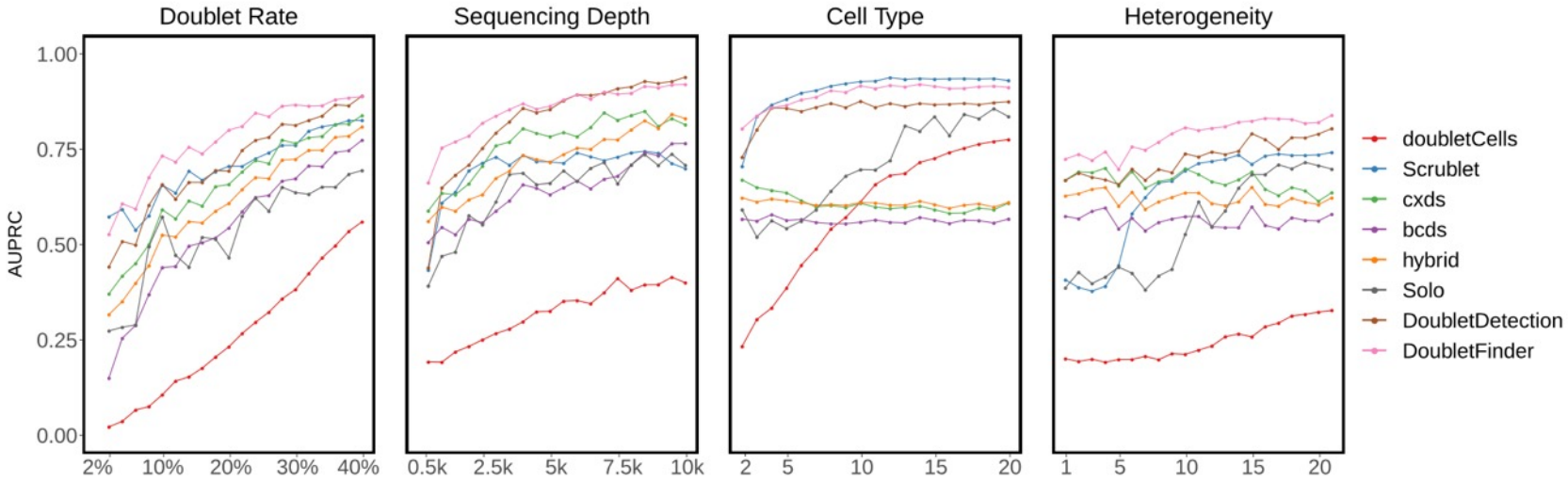
Luecken and Theis, *Molecular Systems Biology* 2019

Doublet rate
(10x Genomics Single Cell 3' Gene Expression v3.1 assay)

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~825	~500
~0.8%	~1,650	~1,000
~1.6%	~3,300	~2,000
~2.4%	~4,950	~3,000
~3.2%	~6,600	~4,000
~4.0%	~8,250	~5,000
~4.8%	~9,900	~6,000
~5.6%	~11,550	~7,000
~6.4%	~13,200	~8,000
~7.2%	~14,850	~9,000
~8.0%	~16,500	~10,000

https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled-

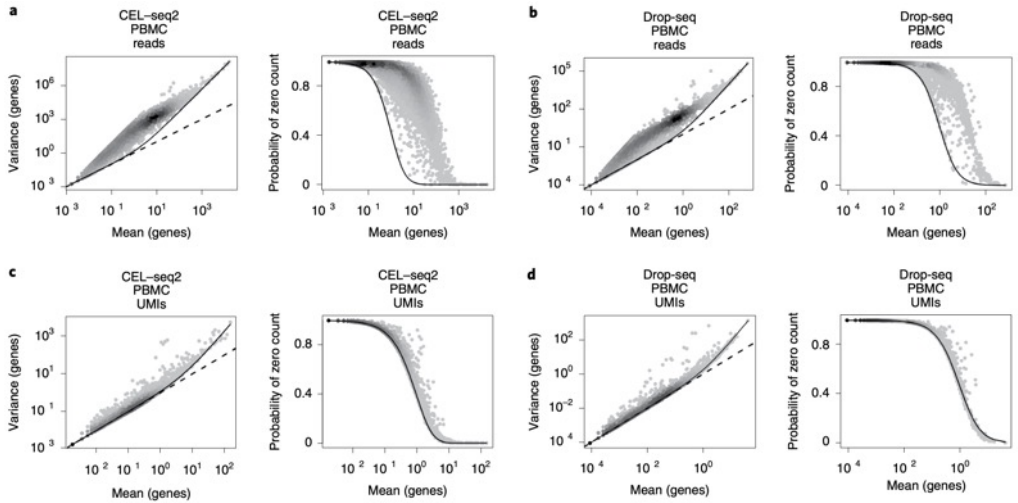
Doublet detection benchmark



Xi and Li, *Cell Systems* 2020

Dropouts and zero-inflation

- Large fraction of observed zeroes (“dropouts”)
  - Methodological noise (expressed but not detected)
  - Absence of expression
- Suppressed zero-inflation with UMIs, stronger zero-inflation with read counts



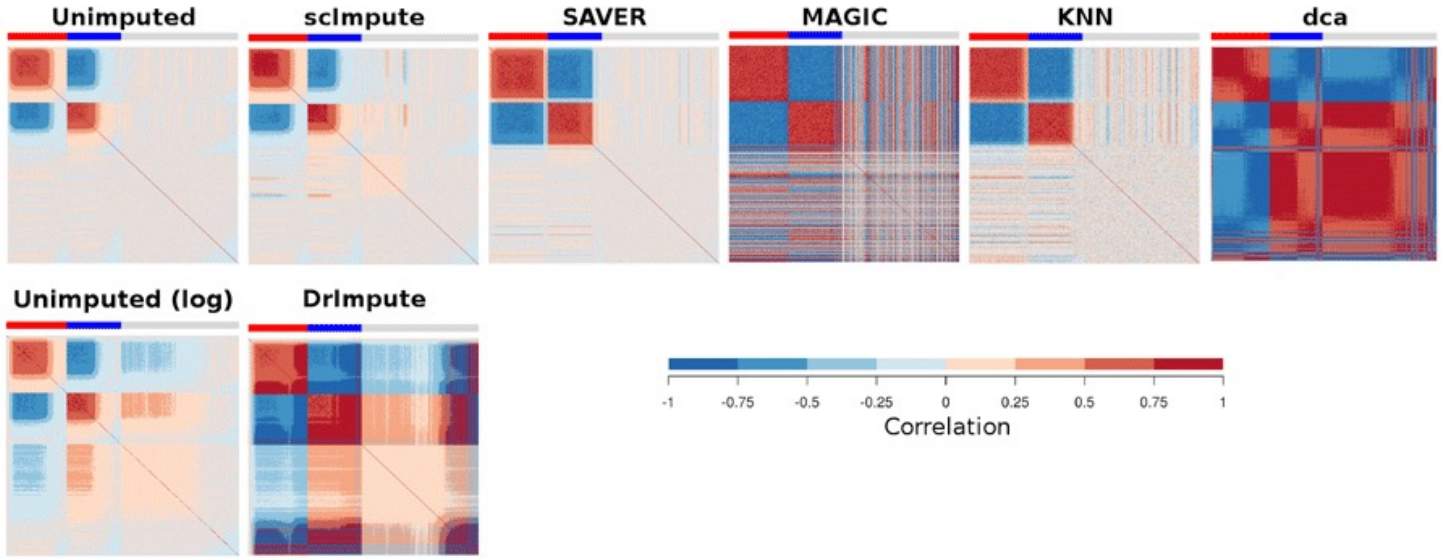
Cao et al, *Nature Biotechnology* 2020

Resolving sparsity with expression recovery or imputation

- Imputation approaches have been proposed to achieve adjusted data values that better represent the true expression values.

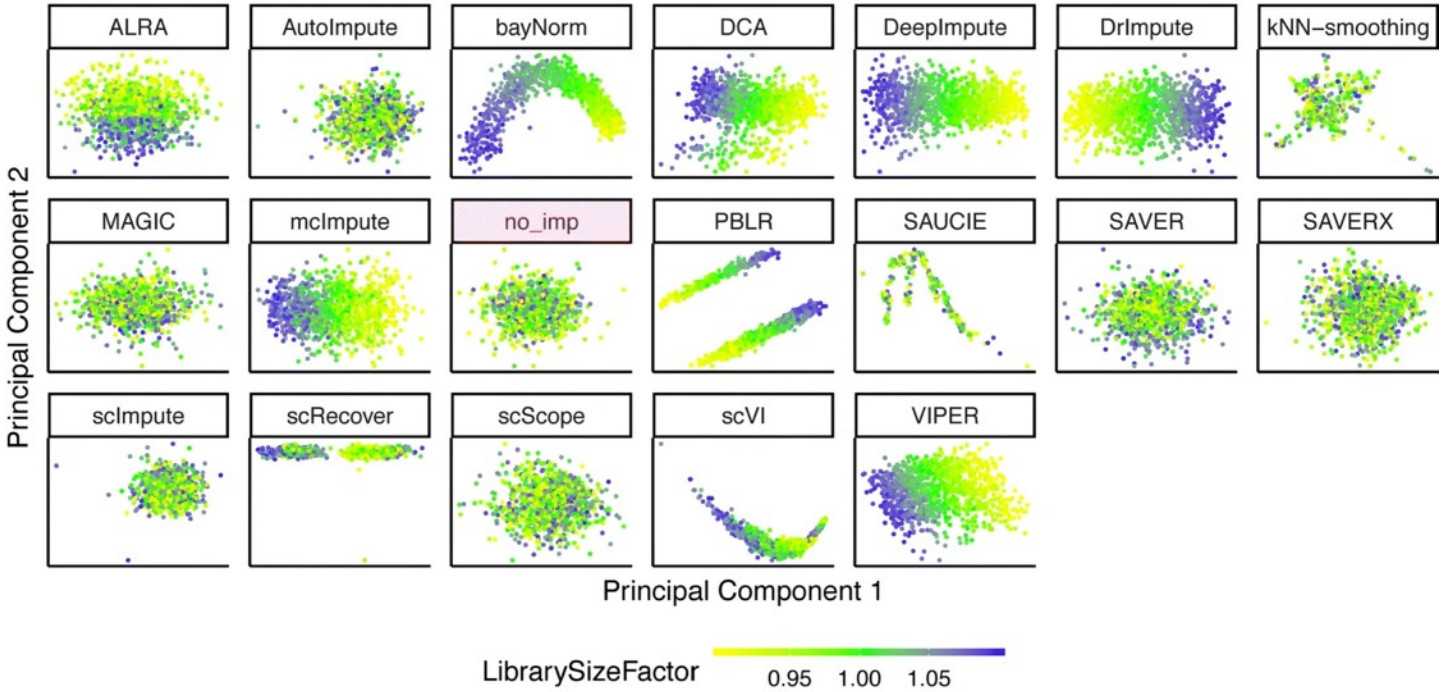
- Categories of methods for imputation:
  - Model-based: use probabilistic models to identify which observed zeros represent technical rather than biological zeros.
  - Data-smoothing: adjust expression values for each cell based on the expression values in similar/nearby cells.
  - Data-reconstruction: use of matrix factorization methods or autoencoders to reconstruct the observed data matrix from low-rank or simplified representations.
  - With an external dataset or reference, using it for transfer learning.

False gene-gene correlations induced by single-cell imputation methods



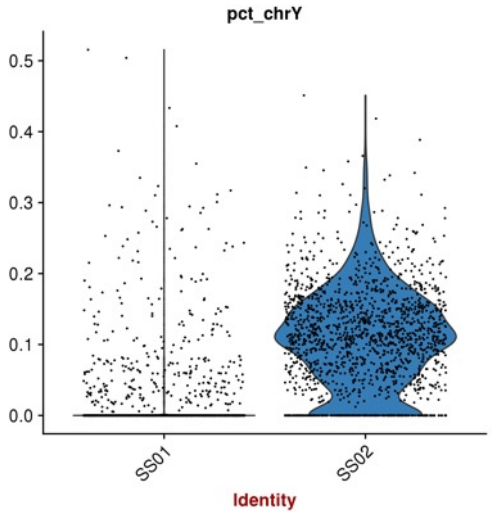
Andrews and Hemberg, *F1000Research* 2020

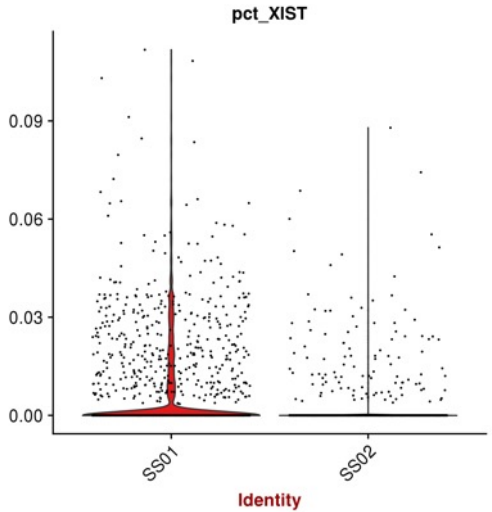
Spurious structural patterns in low dimensional representations of imputed data

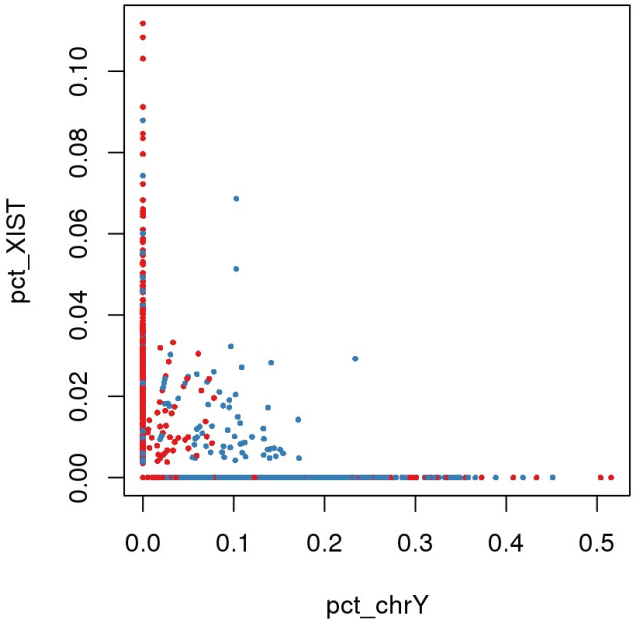


Hou et al, *Genome Biology* 2020

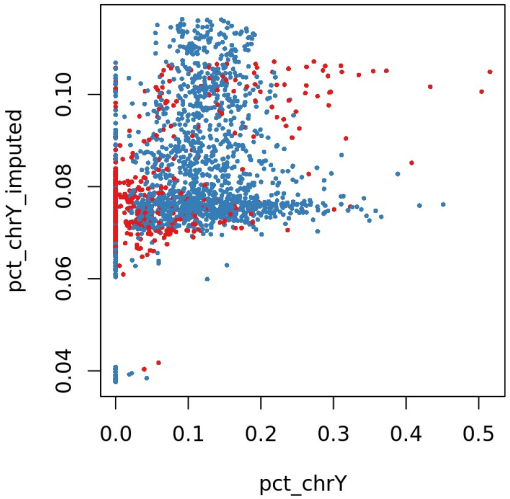
Attempt to demultiplex mixed male/female libraries based on chrY and XIST expression

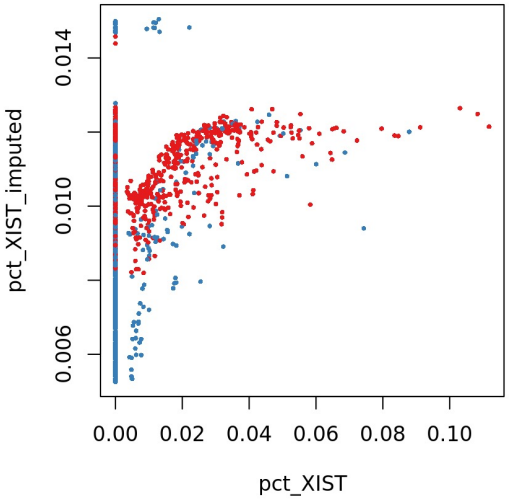


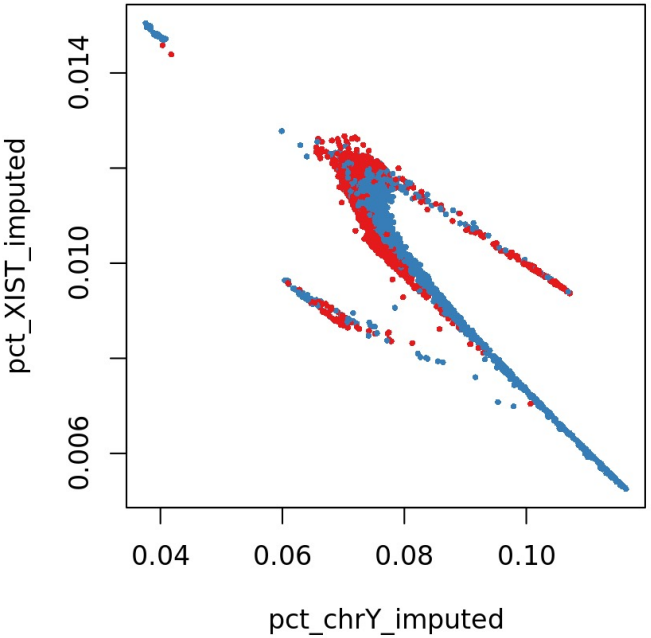




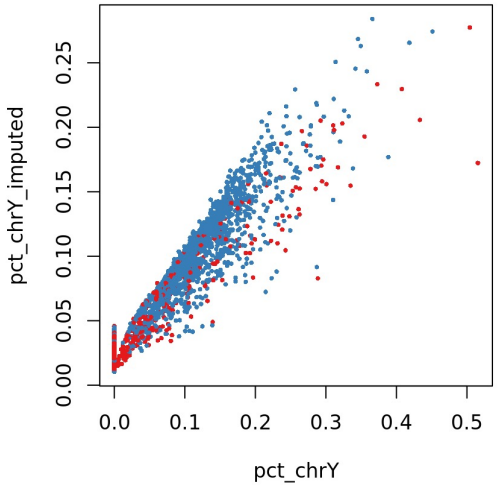
MAGIC: recovered all the dropouts and inferred negative correlation

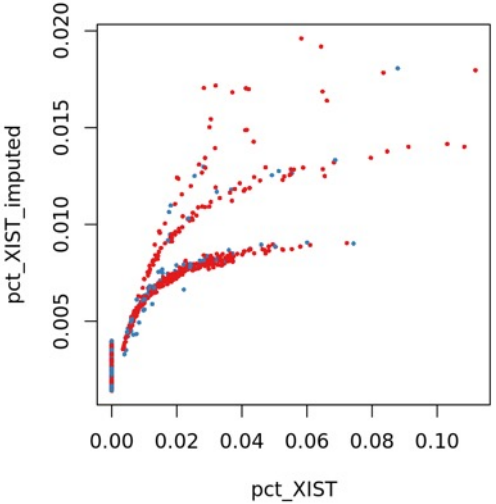


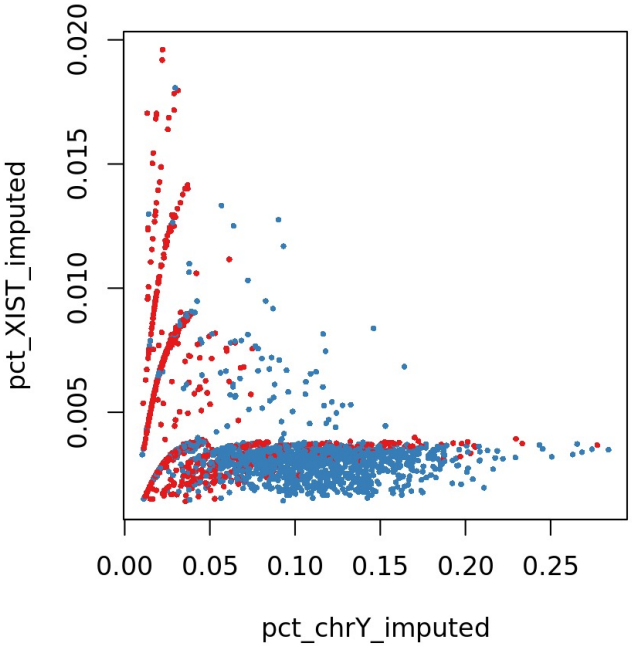




SAVER: more conservative, but dropouts remain even for chrY genes







Normalization

- Each count represents the successful capture, reverse transcription, and sequencing of a molecule of mRNA.
- Any difference between counts may have arisen solely due to sampling effects.
- Normalization adjusts count data to obtain more accurate relative gene expression abundances between cells.

- Approaches:
  - Count depth scaling or CPM/TPM: assumes all cells contain an equal number of mRNA molecules, normalizes count data using a size factor proportional to the count depth per cell.
  - Pooling-based size factor estimation (scran): cells are pooled to avoid technical dropout effects and size factors are estimated based on a linear regression over genes.
  - sctransform: variance stabilizing transformation using the Pearson residuals from regularized negative binomial regression, removes the need for pseudocount addition or log-transformation.

Normalization

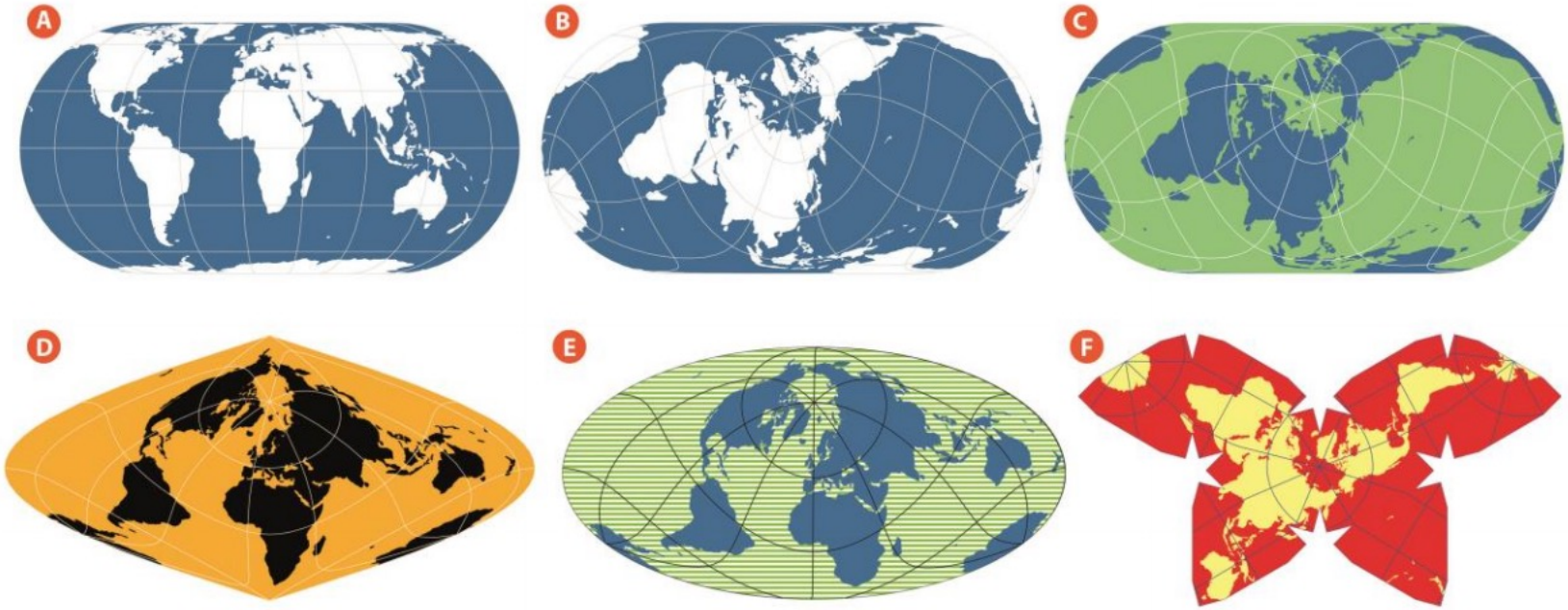
“In most cases, the log-transformation is probably satisfactory. It is simple to compute, preserves sparsity and allows interpretation of differences as log-fold changes (for large counts, at least). It upweights genes with large relative changes in expression that are most likely to be biologically relevant. And if we put aside theoretical arguments, the widespread use of the log-transformation “in the wild” reflects its adequacy and reliability for most analysts.”

Aaron Lun (https://ltla.github.io/SingleCellThoughts/general/transformation.html)

Dimensionality reduction and visualization

- Biological manifold on which cellular expression profile lies can be well described by fewer dimensions than the number of genes.
- Dimensionality reduction algorithms embed the expression matrix into a low-dimensional space, designed to capture the underlying structure in the data.
- Reduced dimensions can be used as coordinates on a scatter plot to obtain a visual representation of the data.
- Summarization techniques reduce the data to its essential components for downstream analysis.

World map projections (2D representation of 3D data)

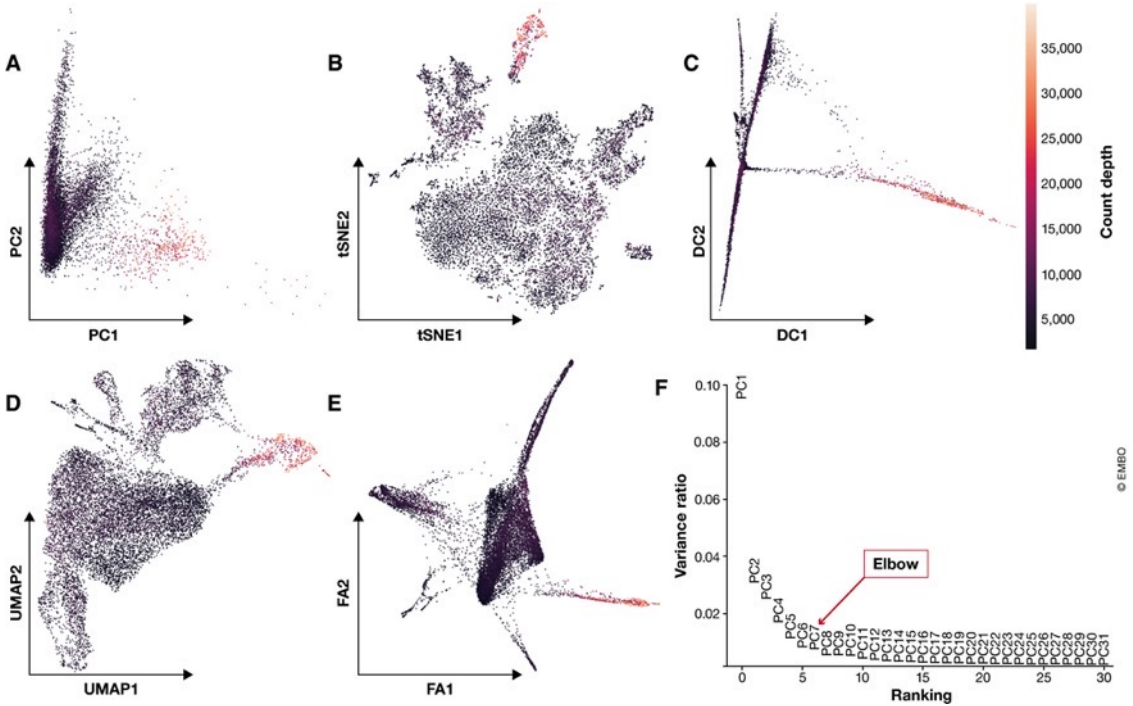


Stirnemann, *Blucher Design Proceedings* 2016

Dimensionality reduction techniques

- PCA: A linear deterministic approach. Basis of clustering and trajectory inference.
- Diffusion maps: Distances between all points are calculated and local relationships are preserved. Diffusion components emphasize transitions in the data and can be used for continuous processes.
- t-SNE: Focus on capturing local similarity at the expense of global structure.
- UMAP: Attempts to preserve both local and most of the global structure in the data.

Common visualization methods

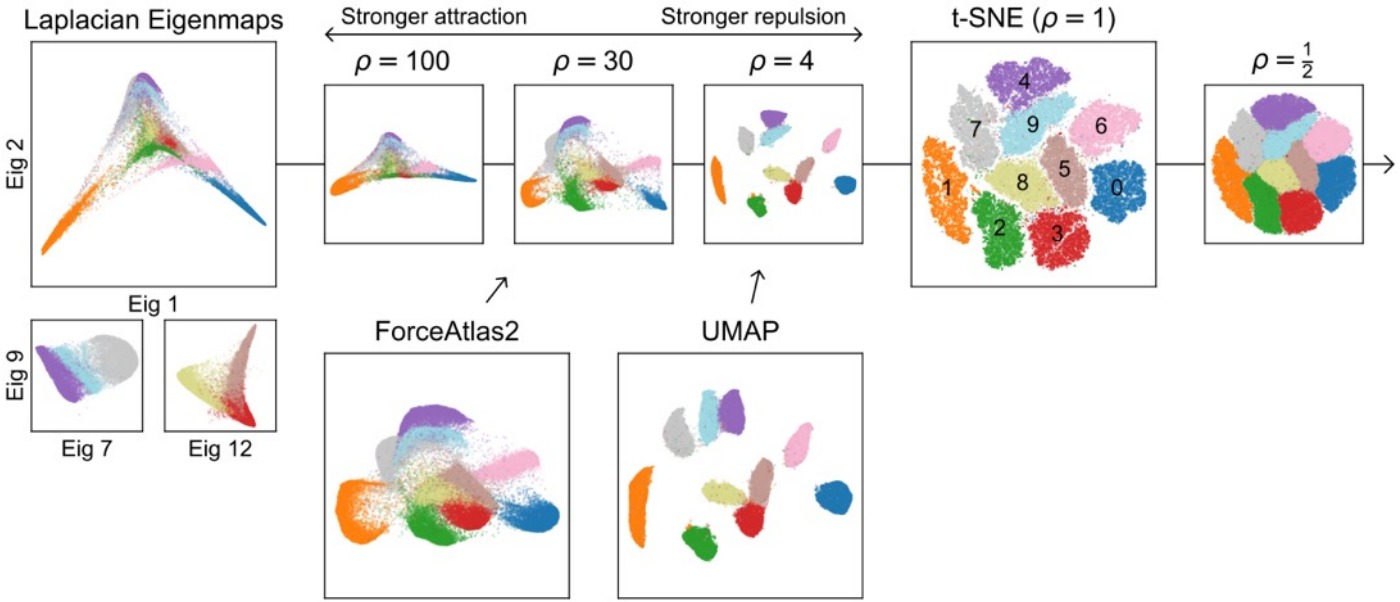


Luecken and Theis, *Molecular Systems Biology* 2019

t-SNE hyperparameters

https://distill.pub/2016/misread-tsne/

Spectrum of t-SNE embeddings depending on the attractive forces between kNN graph neighbors (MNIST data)



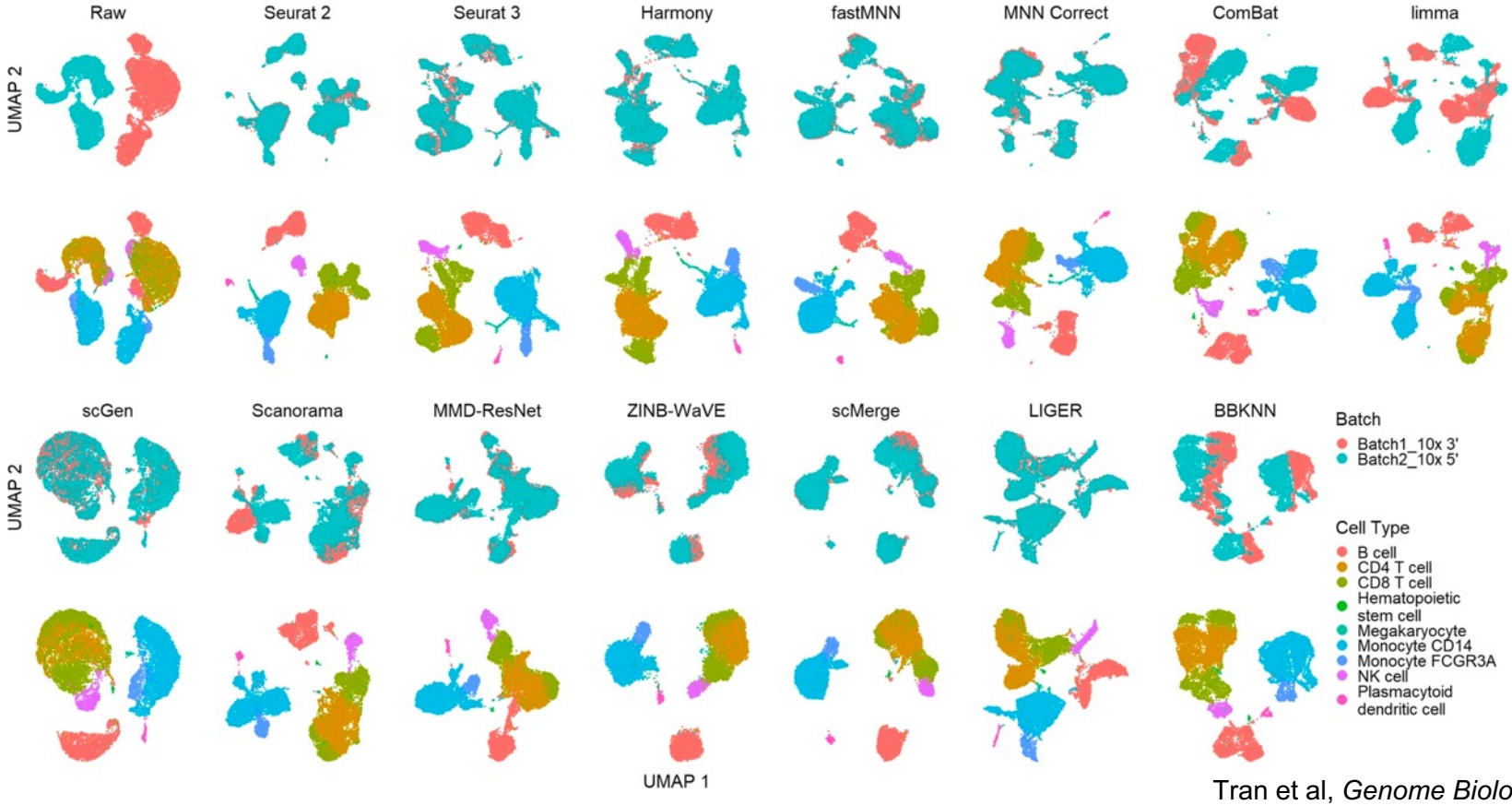
Bohm et al, *arXiv* 2020

t-SNE and UMAP with random and non-random initialization

The t-SNE implementations by default used random initialization, while the UMAP implementation used Laplacian eigenmaps (LE) to initialize the embedding

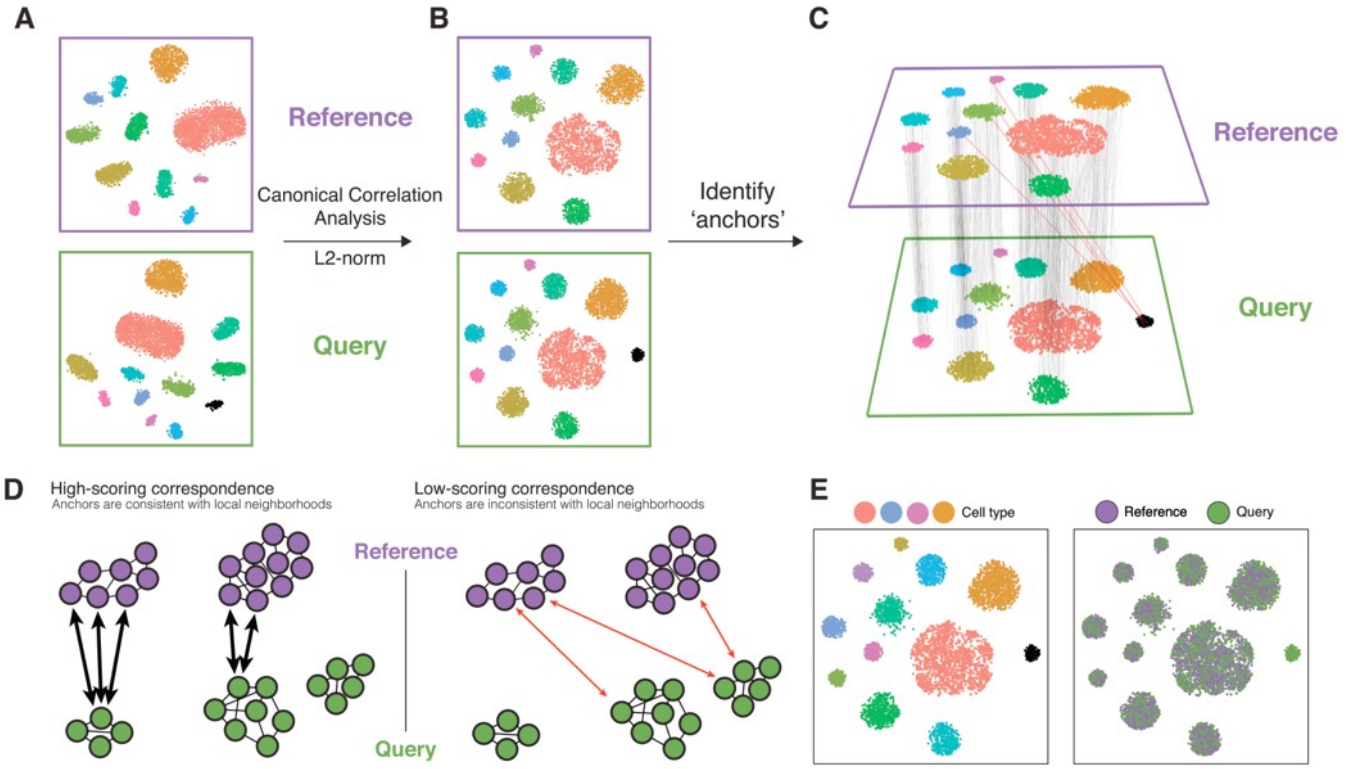
Kobak and Linderman, *Nature Biotechnology* 2021

Multi-dataset integration (batch correction)



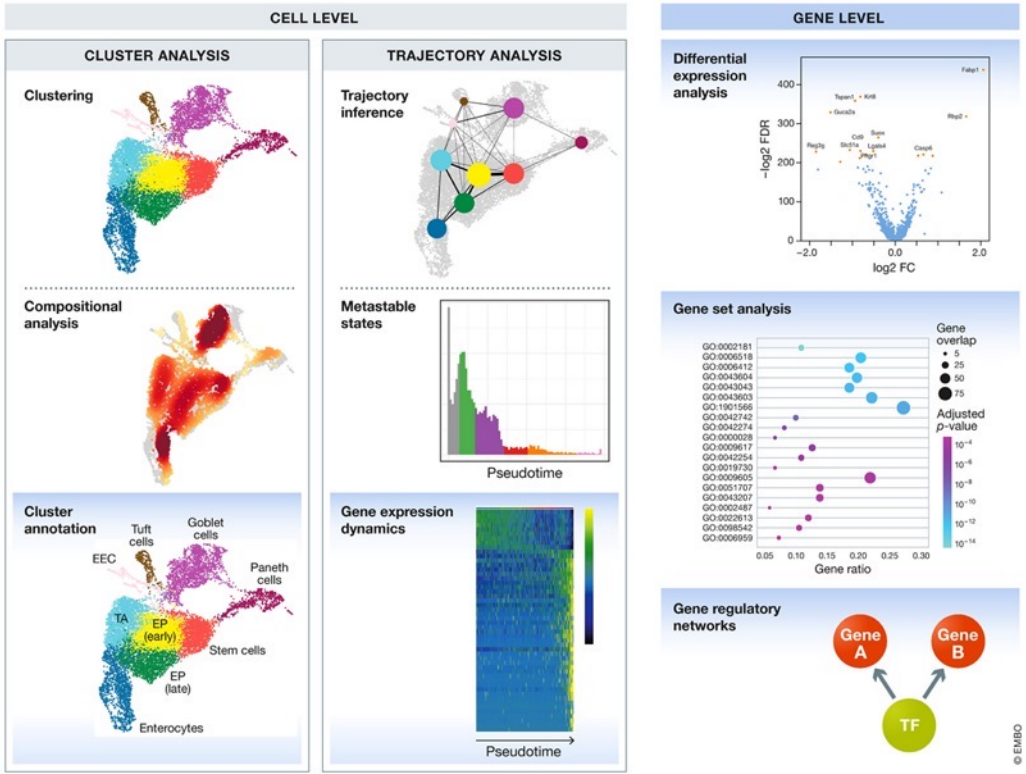
Tran et al, *Genome Biology* 2020

Seurat reference “assembly” integration



Stuart et al, *Cell* 2019

Data analysis second half: downstream analysis



Luecken and Theis, *Molecular Systems Biology* 2019

Clustering

- Unsupervised learning procedure to define groups of transcriptionally similar cells.
- Describes heterogeneity of the high-dimensional manifold on which the cells reside in terms of discrete labels.
- Classical unsupervised machine learning problem, based on a distance matrix:
  - Distances: Euclidean, cosine similarity, correlation-based distance metrics, SIMLR (distance metric based on Gaussian kernels)
  - K-means clustering algorithm divides cells into k clusters by determining cluster centroids and assigning cells to the nearest cluster centroid. Centroid positions are iteratively optimized.
  - Graph-partitioning algorithms that rely on a graph representation of data.
  - Louvain algorithm implements multi-resolution modularity optimization and detects communities as groups of cells that have more links between them than expected.

The correct number of clusters

“it is helpful to realize that clustering, like a microscope, is simply a tool to explore the data. We can zoom in and out by changing the resolution of the clustering parameters, and we can experiment with different clustering algorithms to obtain alternative perspectives of the data. … As such, questions about the “correctness” of the clusters or the “true” number of clusters are usually meaningless. We can define as many clusters as we like, with whatever algorithm we like - each clustering will represent its own partitioning of the high-dimensional expression space, and is as “real” as any other clustering. A more relevant question is “how well do the clusters approximate the cell types or states of interest?””

https://bioconductor.org/books/3.15/OSCA.basic/clustering.html

Cell type identification

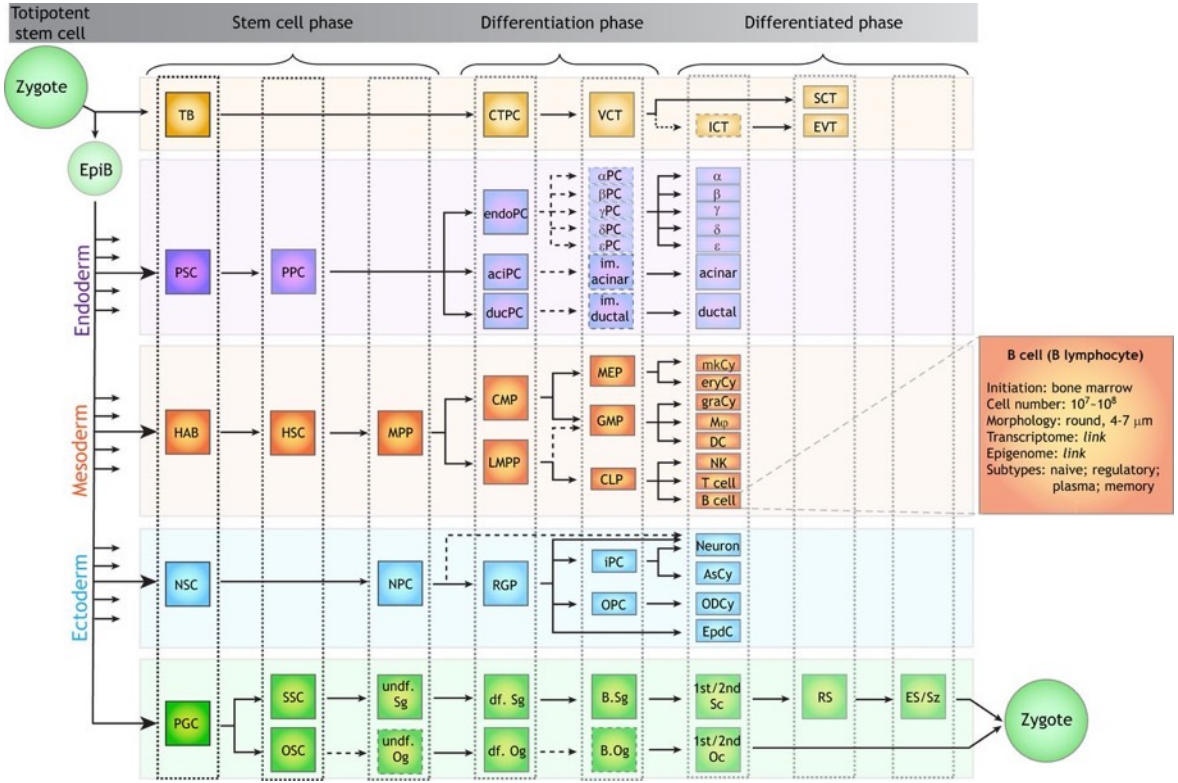
- Identification of cell subpopulations from complex scRNA-seq mixtures includes automated steps from normalization to cell clustering.

- Assigning cell type labels to cell clusters is often conducted manually, a process that is time-consuming, poorly documented, not reproducible, prone to bias and error.

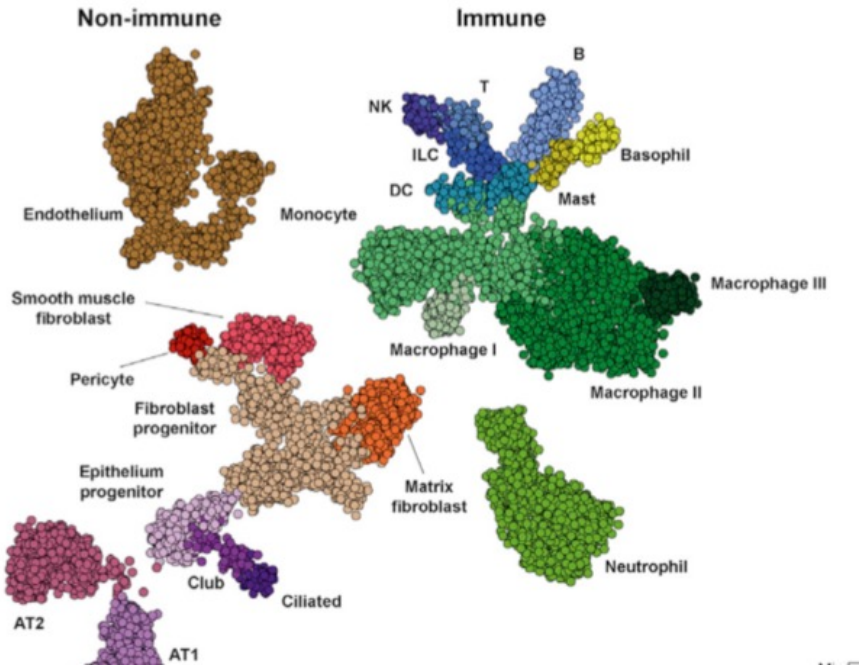
Annotation challenges

- Stability of the supporting signals.
- Relative scarcity of reference populations or cell type signatures.
- Markers may not be specific enough to differentiate the cell subpopulations in the same dataset, or may not be generic enough to be applied from one study to another.
- There is not a canonical set of markers.
- Intermediate cell states and developmental trajectories.
- Cell type hierarchies.

Periodic table of cell types

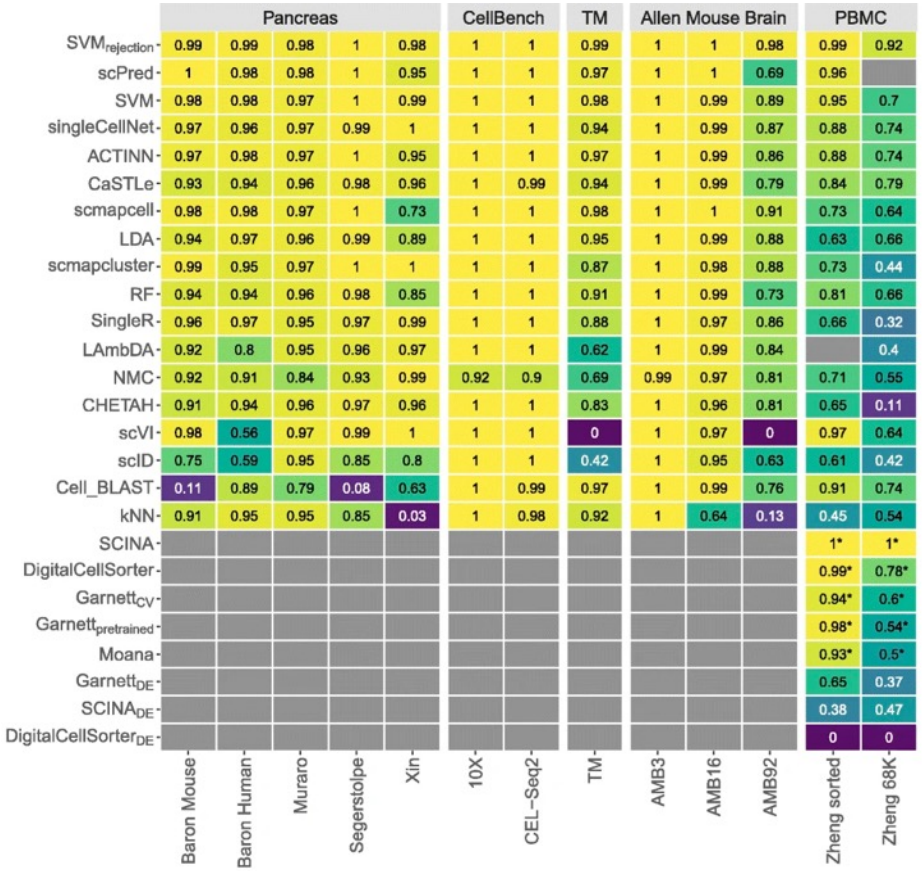


Xia and Yanai, Development 2019



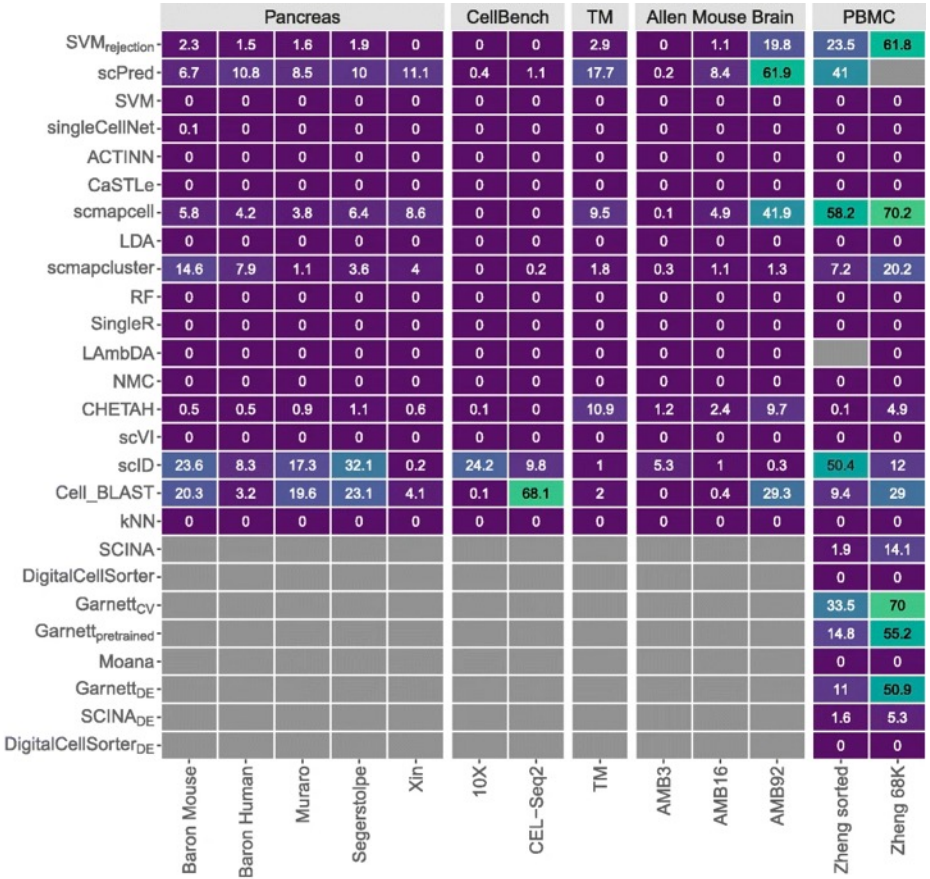
group	gene	priority	fold change
Epithel	Epcam	1	2
AT1	Clic5	3	5
AT2	Sftpc	3	40
Endothel	Cdh5	4	4
Fibro	Col1a2	1	2
Pericytes	Gucy1a3	3	5
Club	Scgb3a2	3	2
Matrix	Mfap4	3	10
Smooth	Tgfbi	2	8
Ciliated	Ccdc19	3	2
Ciliated	Foxj1	3	2
B	Cd79b	1	2
Baso	Mcpt8	5	2
DC	Flt3	4	2
MacI	Cx3cr1	4	6
MacII	Ear2	3	2
MacIII	Ccl6	5	20
MacIII	Cd9	5	7
Mast	Mcpt4	4	2
Mast	Gata2	3	3
Mon	Ccr2	2	2
Mon	F13a1	3	4
Mon	Fcgr4	5	3.5
Mon	Csf1r	3	4
Neut	S100a8	1	20
Neut	Csf3r	4	5
NK	Gzma	3	5
T	Trbc2	2	2
ILC	Rora	4	2

Comparison of supervised classifiers for cell identification: median F1-scores



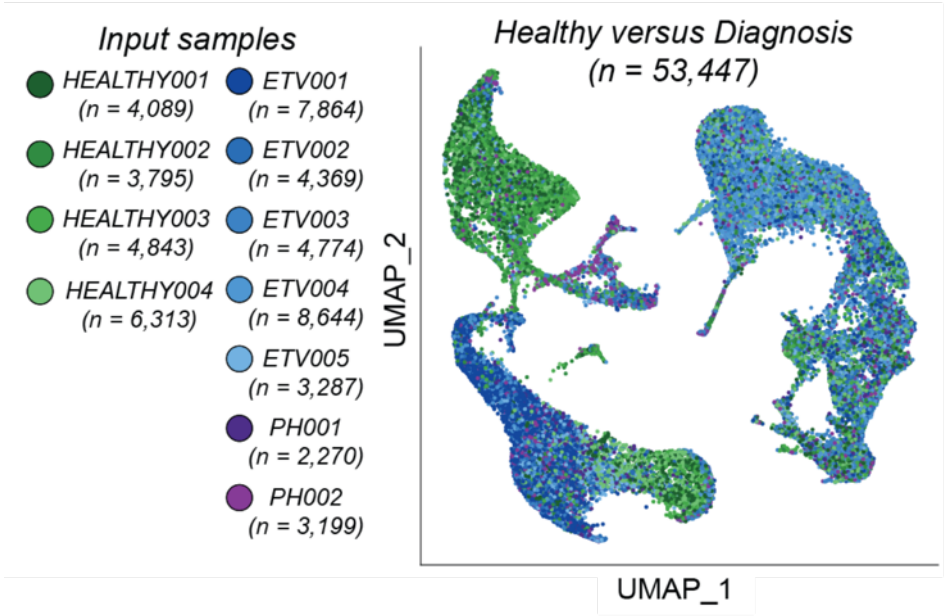
Abdelaal et al, Genome Biology 2019

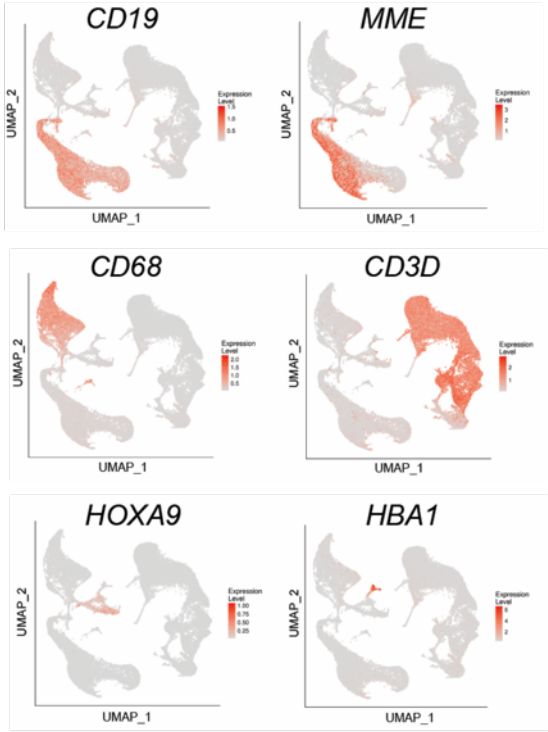
Comparison ofsupervised classifiers for cell identification: unlabeled cells



Abdelaal et al, *Genome Biology* 2019

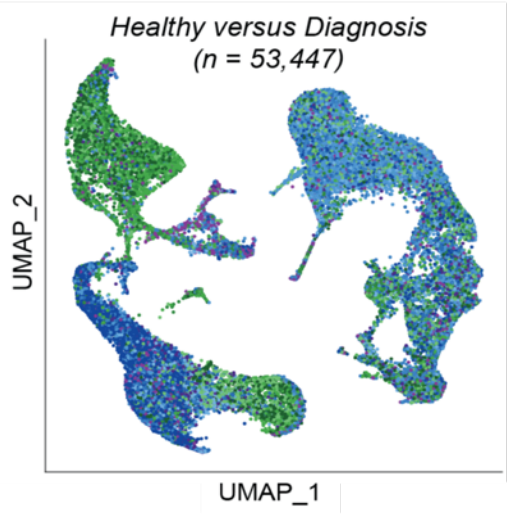
Single cell landscape of the B-ALL bone marrow immune microenvironment

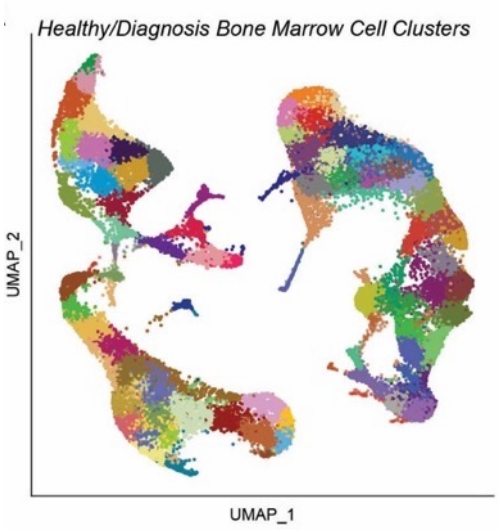


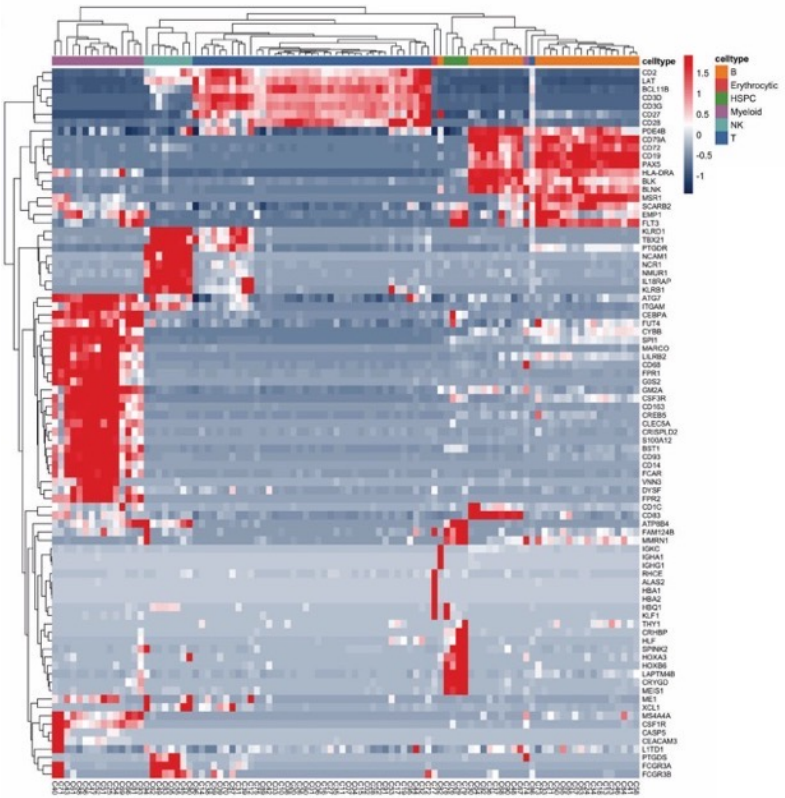


Witkowski & Dolgalev et al, *Cancer Cell* 2020

Signature-based identification of major immune cell types







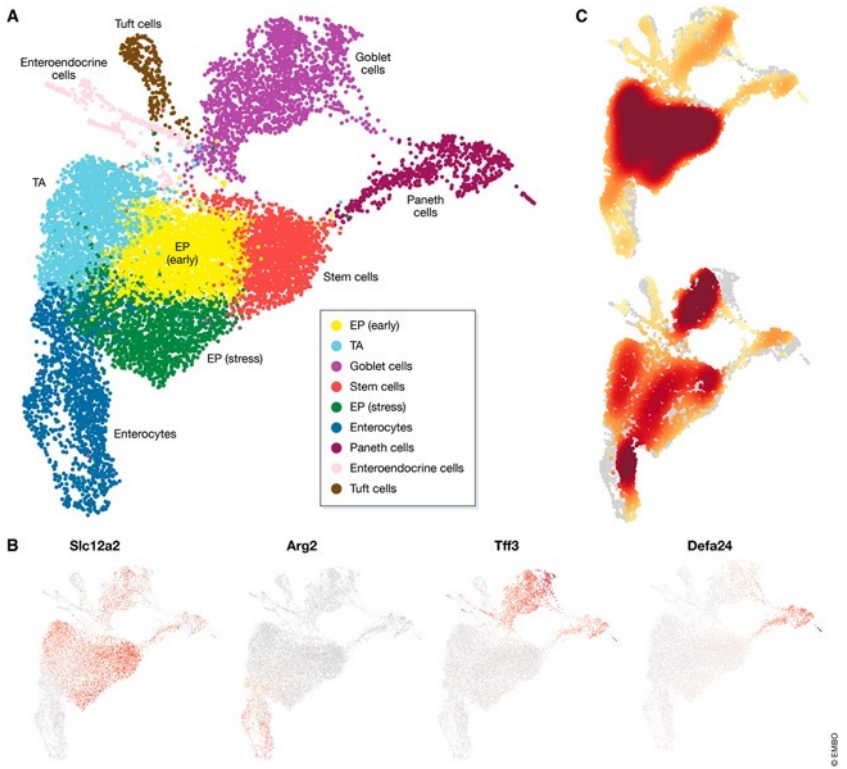
Exploratory cell type “digging” with clustermole

Unbiased enrichment of >2,500 cell types/states sourced from a variety of databases

interactive tutorial: http://bit.ly/nyucells2020

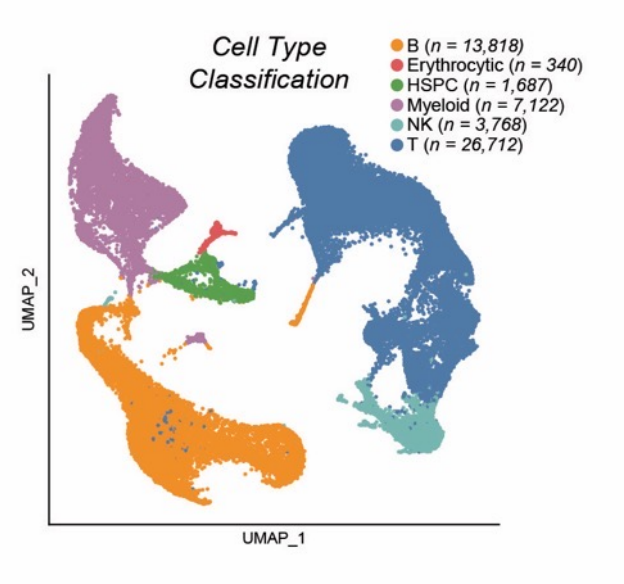
https://github.com/igordot/clustermole
https://cran.r-project.org/package=clustermole

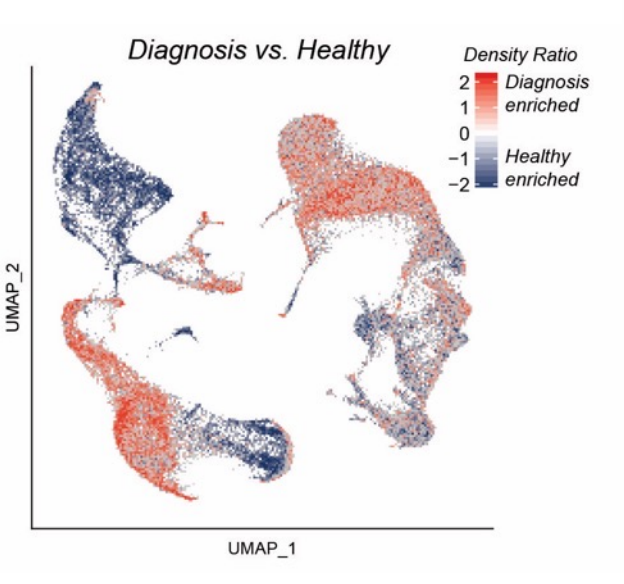
Compositional analysis

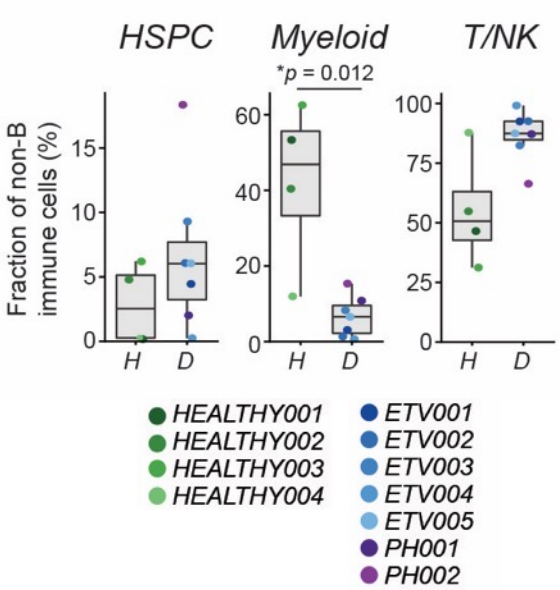


Luecken and Theis, *Molecular Systems Biology* 2019

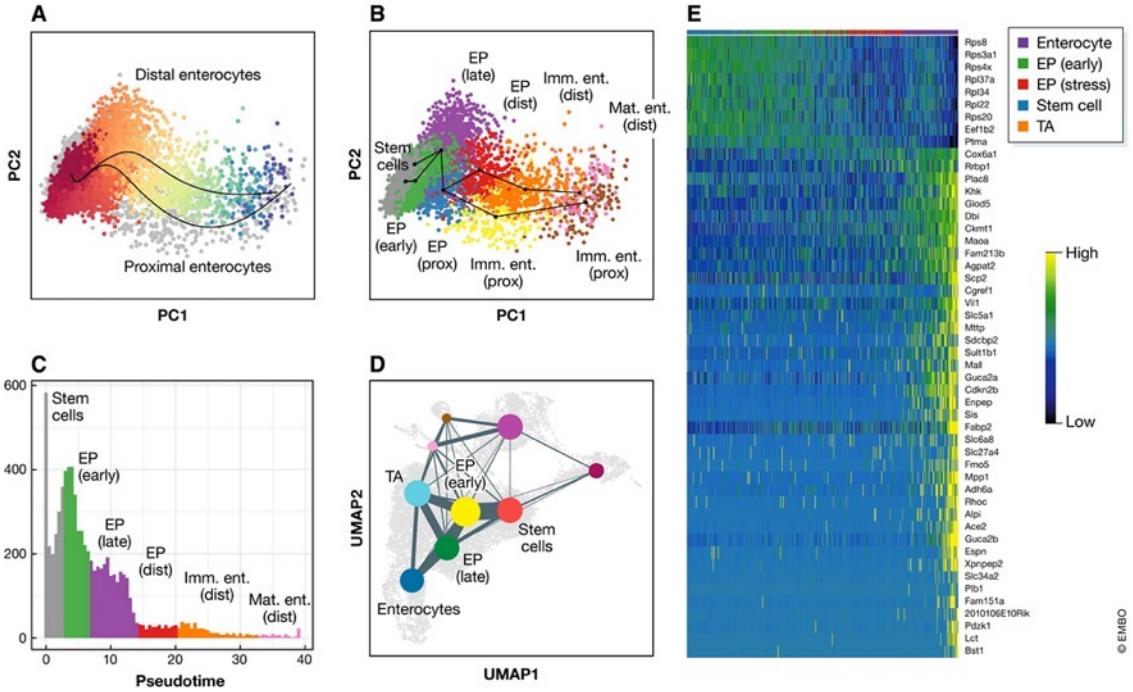
Myeloid compartment diminished due to B-ALL emergence





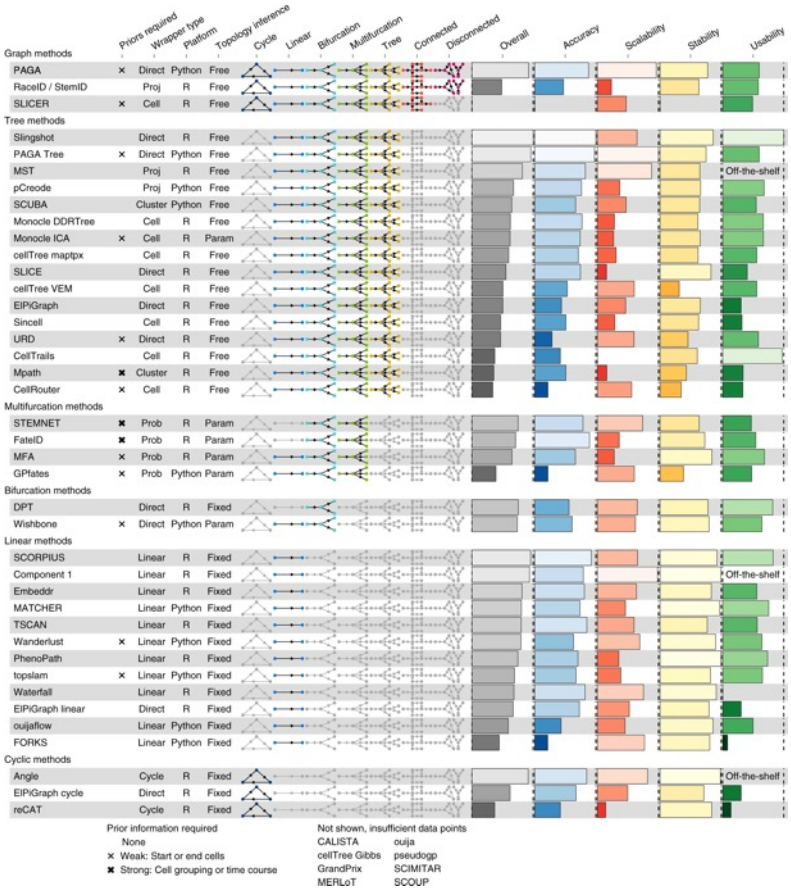


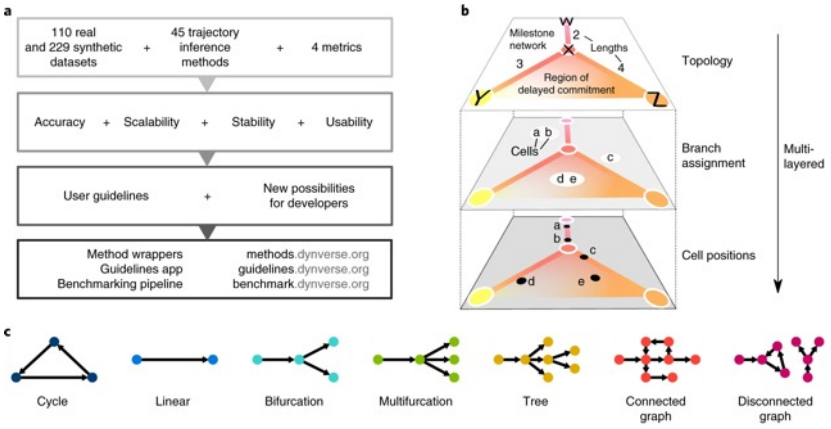
Trajectory analysis



Luecken and Theis, *Molecular Systems Biology* 2019

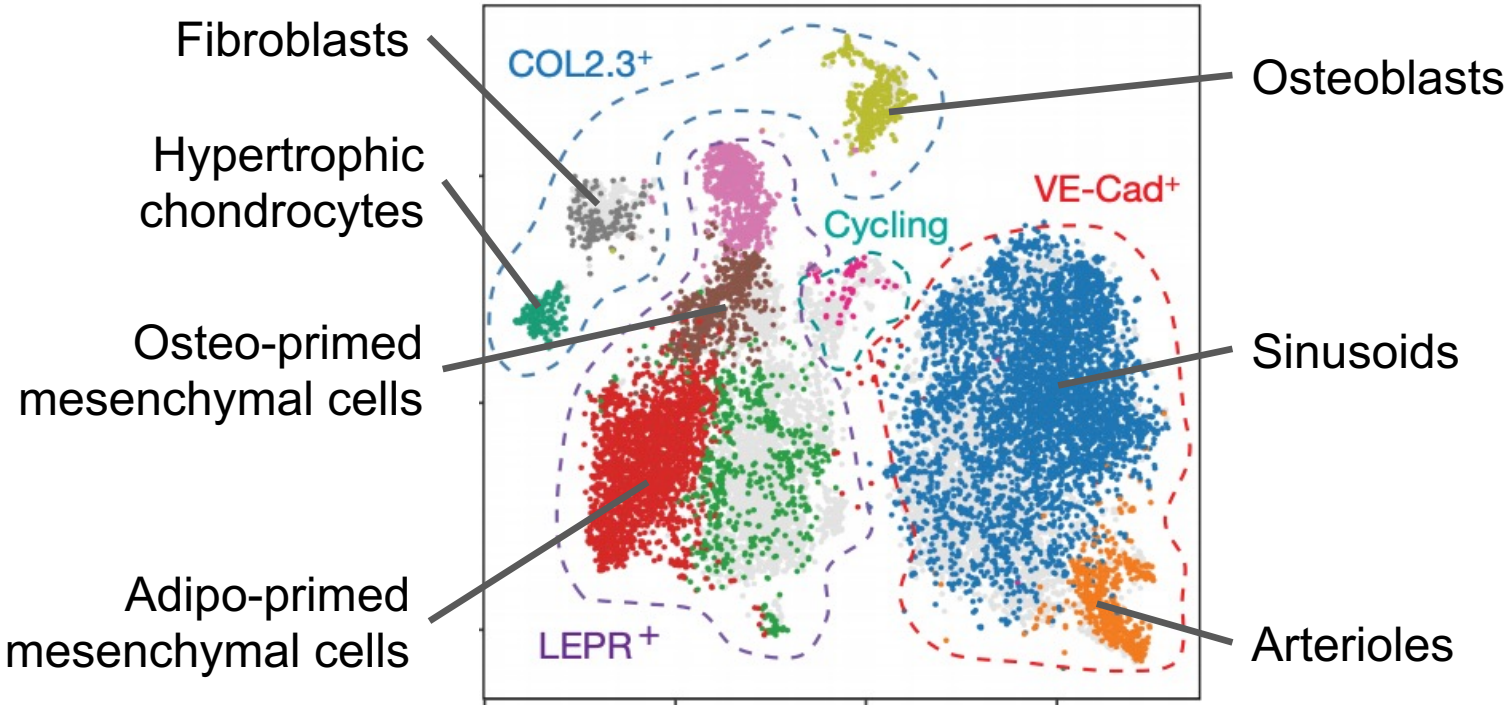
A characterization of 45 trajectory inference methods





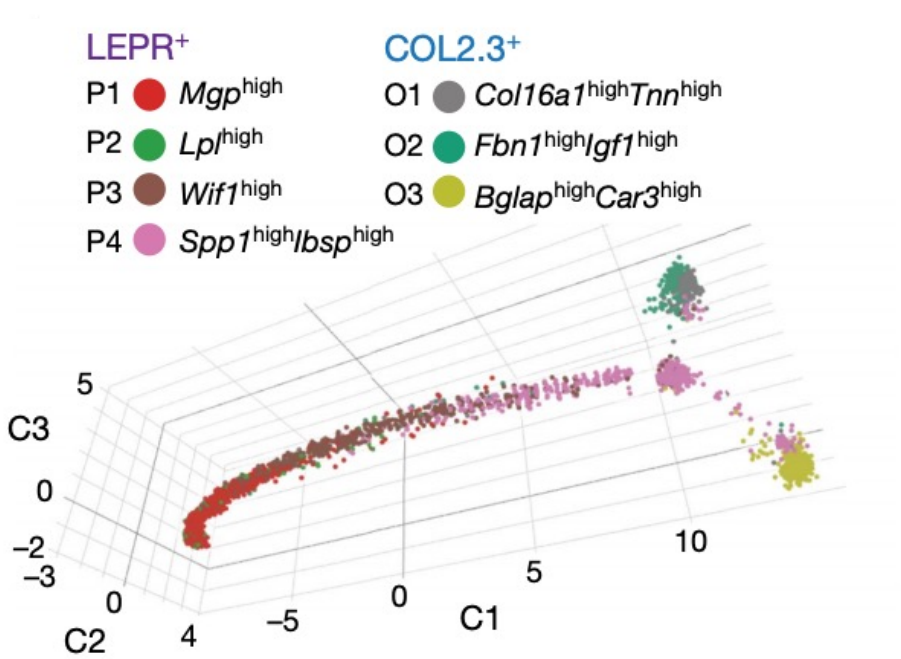
Saelens et al, *Nature Biotechnology* 2019

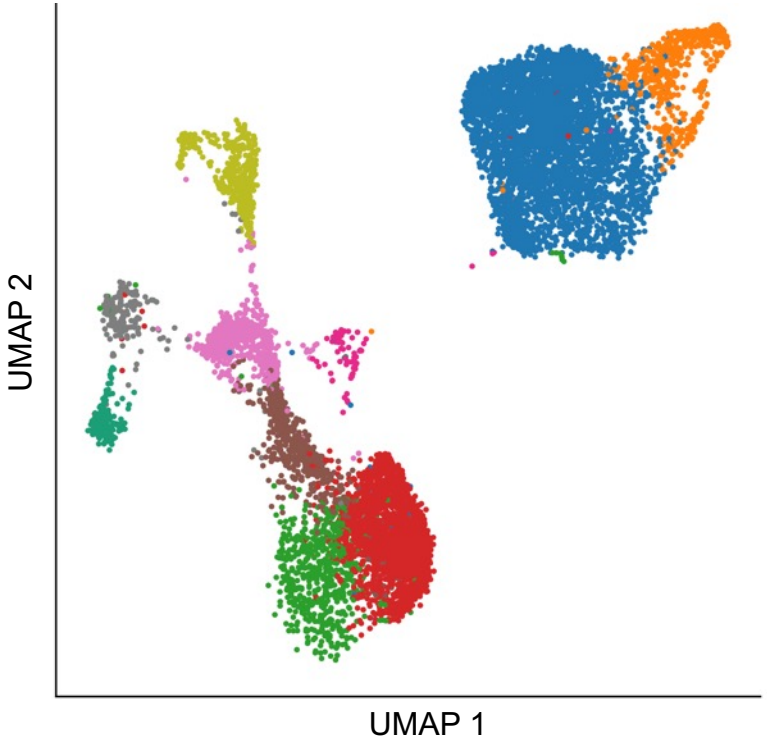
Transcriptional profiling identified specific niche subpopulations



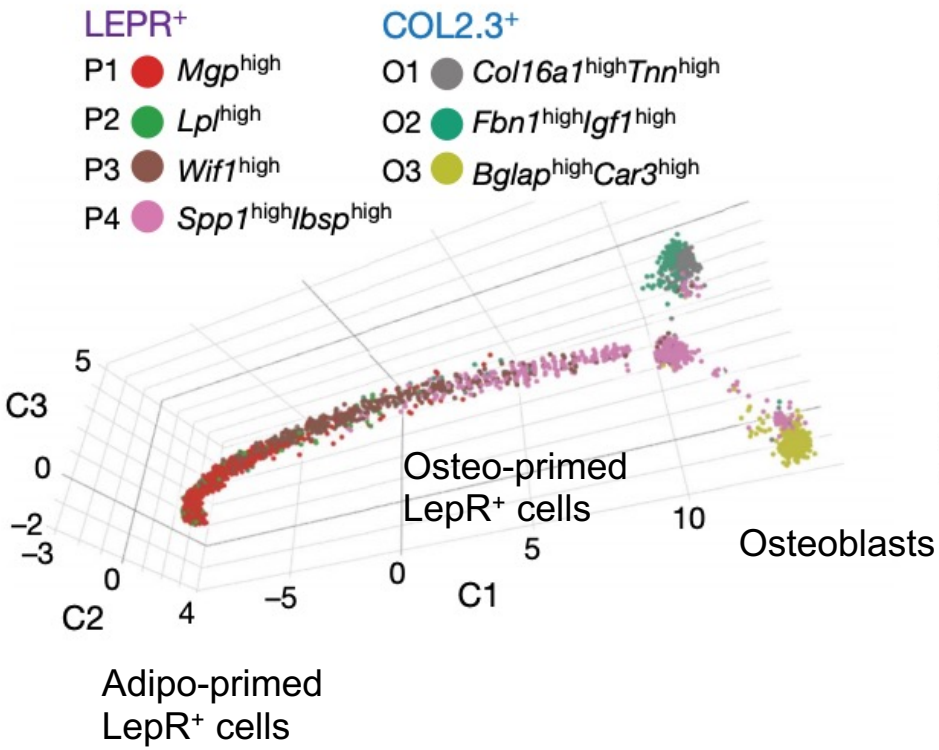
Tikhonova & Dolgalev et al, *Nature* 2019

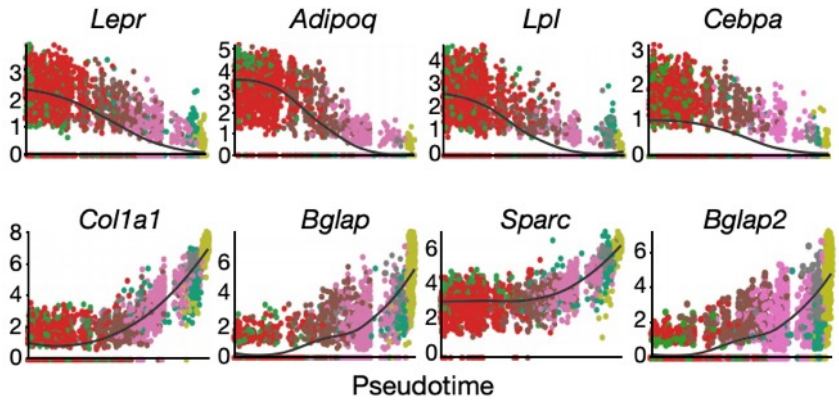
Reconstructed cell differentiation trajectory of LEPR+ and COL2.3+ populations





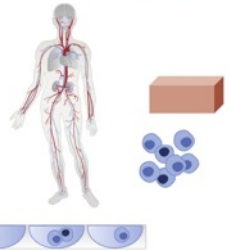
Pseudotime ordering revealed a transcriptional continuum of LepR+ cellular states, with known adipogenic and osteogenic markers rising towards the opposite ends of this range



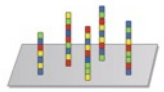


Metadata categories that can be used to describe a single-cell assay

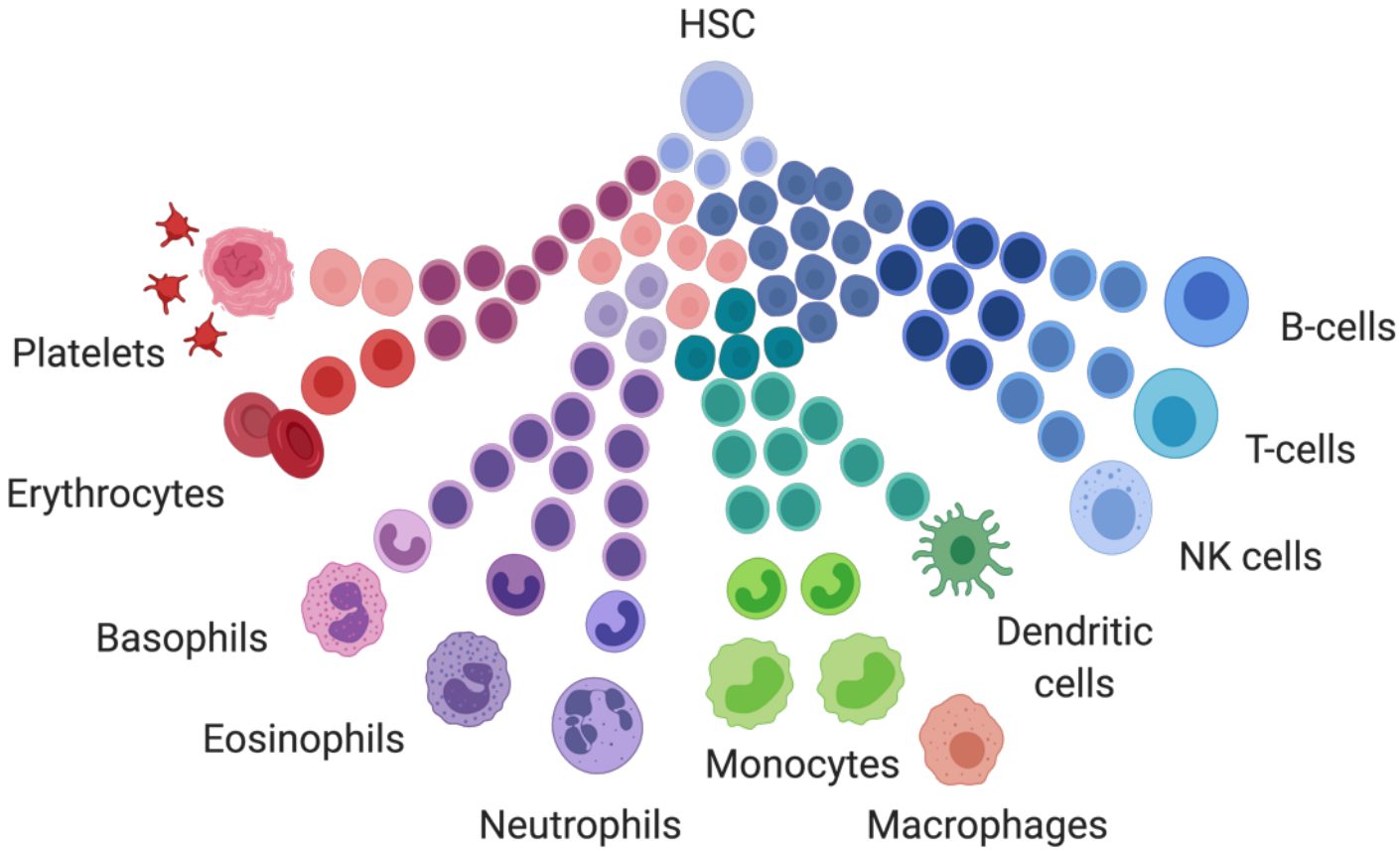
Component	Subtype	Attributes
Biosource	Donor Specimen from organism Organoid Cell suspension Cell line	Species/taxon ID *Other attributes (depending on experiment)
Single cell isolation		Single cell identifier Single cell entity Tissue dissociation Cell enrichment Single cell isolation Single cell quality Cell number
Library construction		Library construction End bias Library strand Input molecule + primer Amplification method Spike-in + dilution
Sequencing		Instrument Library layout/paired-end Technical replicate group/library reference



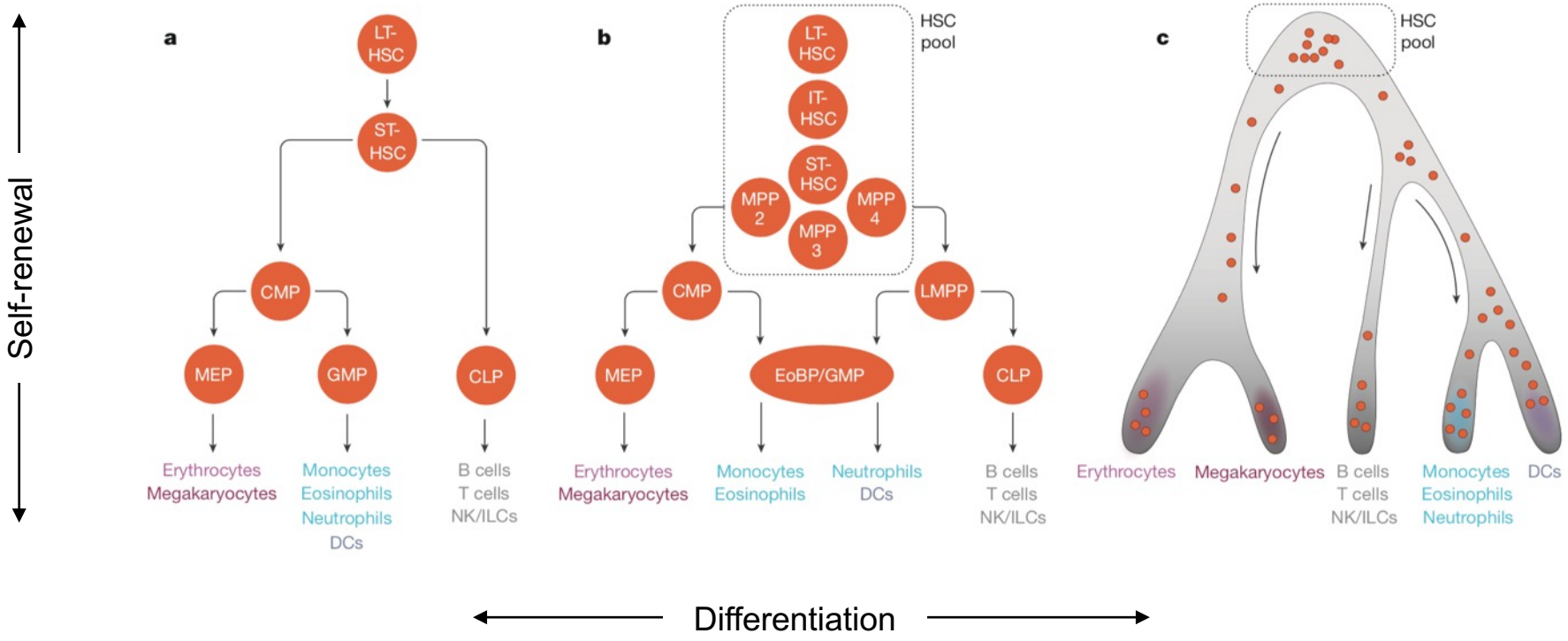
Component	Subtype	Attributes
Raw data		Read type (read1/read2/index) UMI barcode read UMI barcode offset UMI barcode size Cell barcode read Cell barcode offset Cell barcode size cDNA read cDNA read size cDNA read offset Filename Checksum
Processed data		Inferred cell type Postanalysis cell quality *Other derived attributes
Protocols	Cell isolation/enrichment Library construction Sequencing Data analysis	Description



Fullgrabe, *Nature Biotechnology* 2020


HSC
Platelets
Erythrocytes
Basophils
Eosinophils
Neutrophils
Monocytes
Macrophages
Dendritic cells
NK cells
T-cells
B-cells

Evolution of the hierarchical models of haematopoiesis and a continuum of differentiation

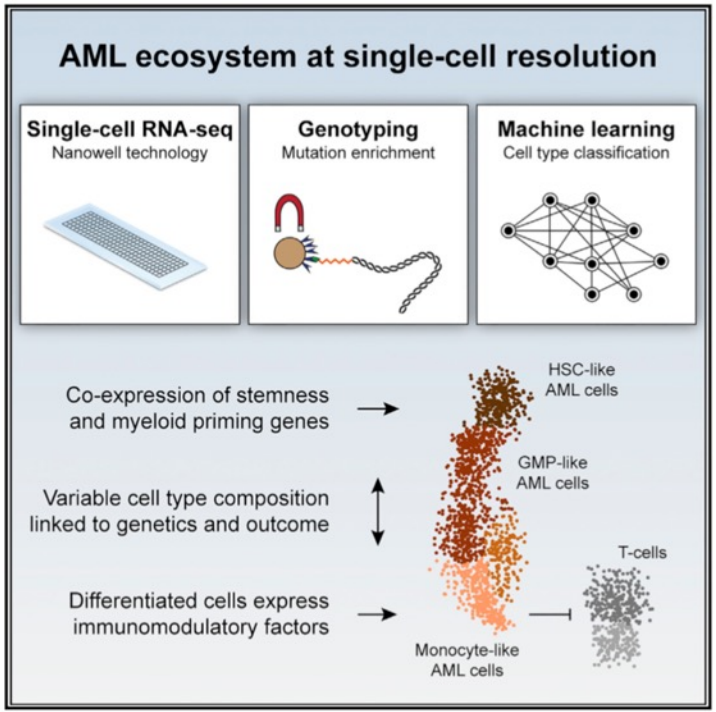


Laurenti & Göttgens, Nature 2018

Cell

Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity

Graphical Abstract



Authors

Peter van Galen, Volker Hovestadt, Marc H. Wadsworth II, ..., Jon C. Aster, Andrew A. Lane, Bradley E. Bernstein

Correspondence

bernstein.bradley@mgh.harvard.edu

In Brief

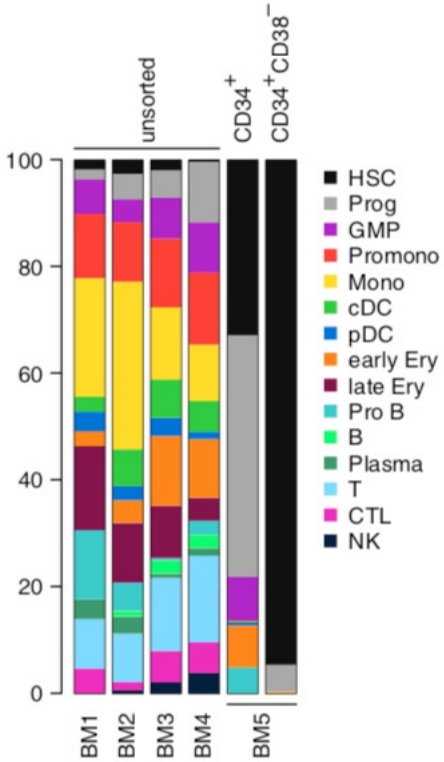
A combination of transcriptomics and mutational analyses in single cells from acute myeloid leukemia patients reveals the existence of distinct functional subsets and their associated drivers.

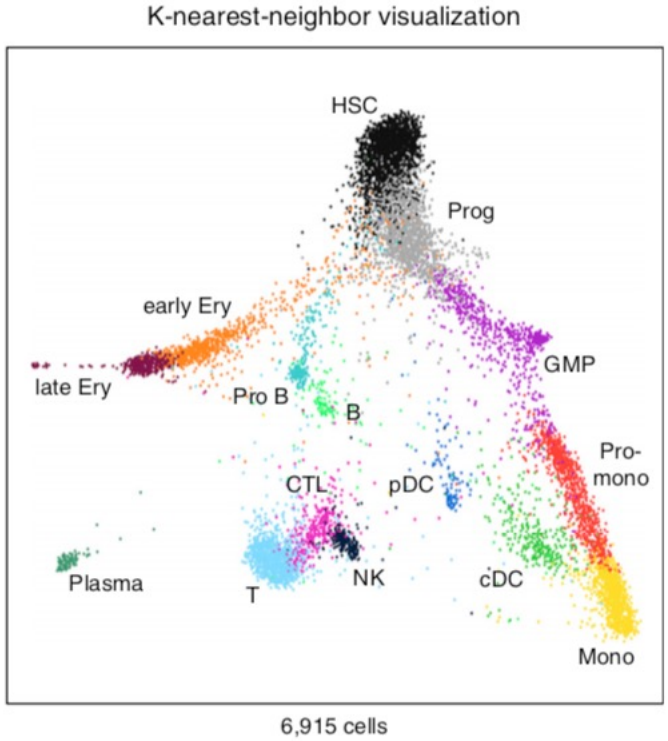

BROAD
INSTITUTE

MGH
1811
MASSACHUSETTS
GENERAL HOSPITAL

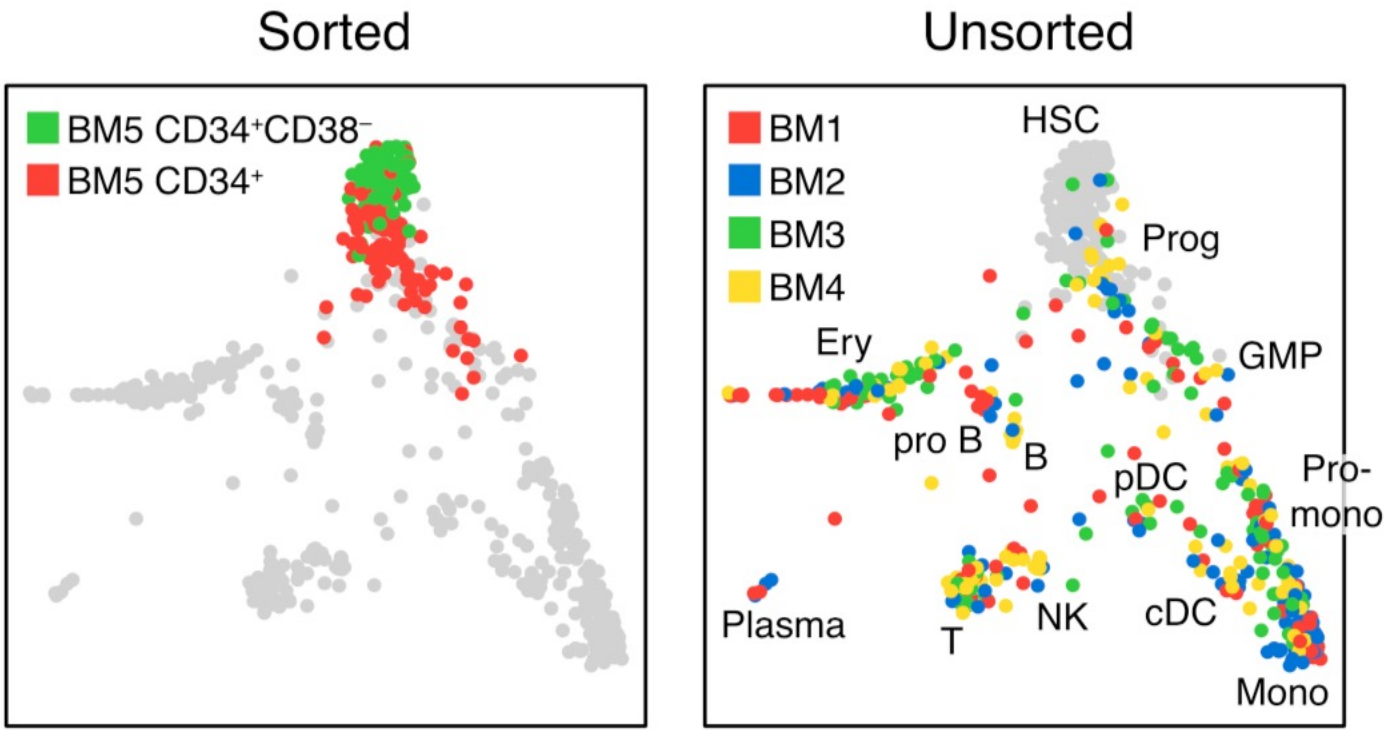
CellPress

Identification of Cell Populations in Healthy BM Samples



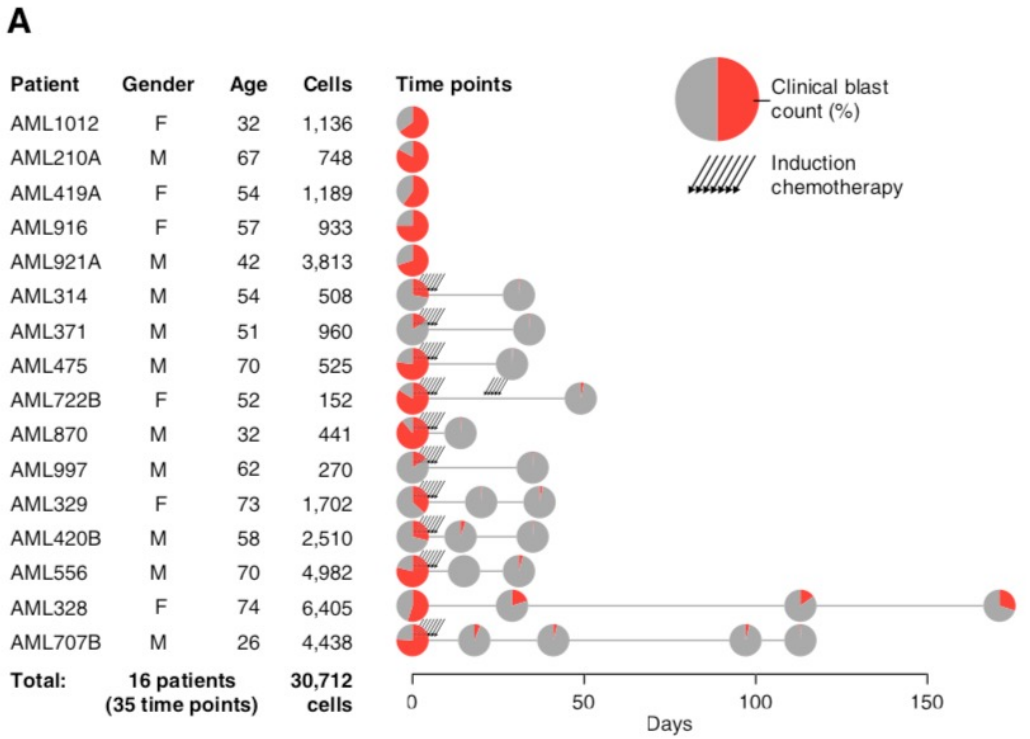


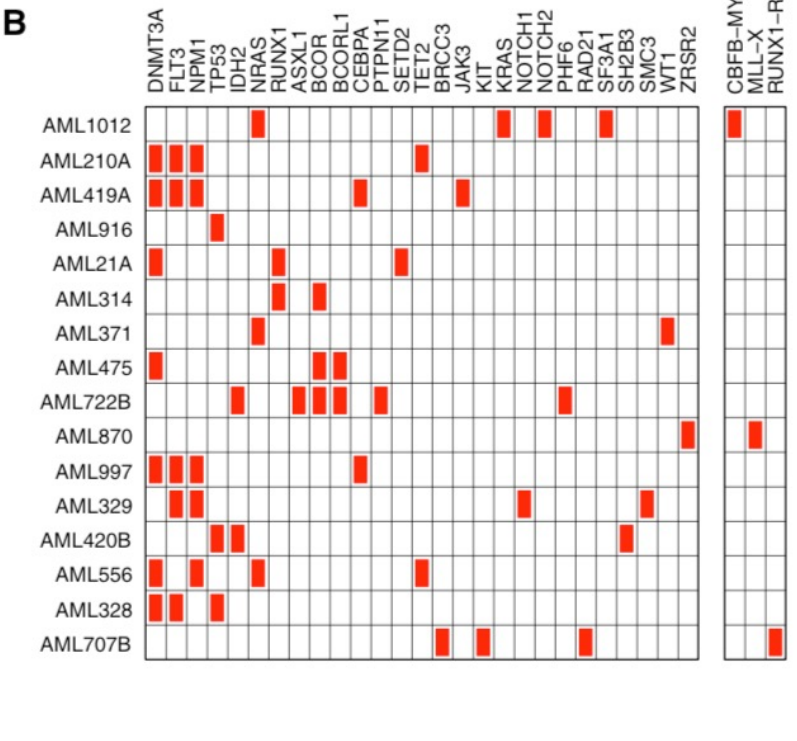
“we did not observe batch dependent clustering”



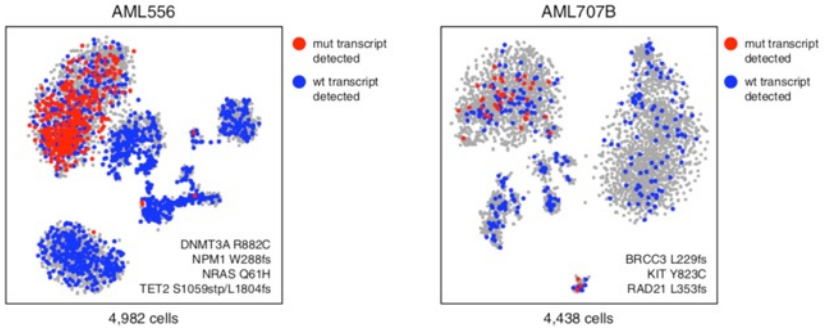
Each sample was downsampled to 100 cells for visualization

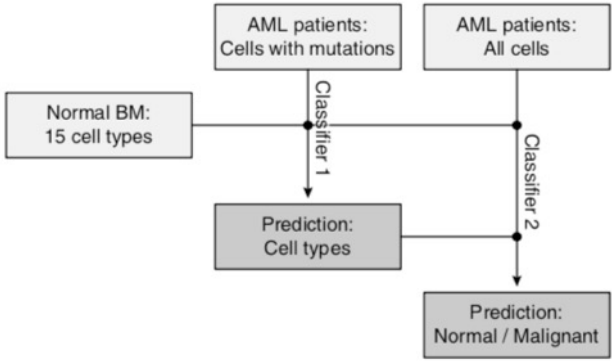
Single-Cell Profiling of AML Tumor Ecosystems

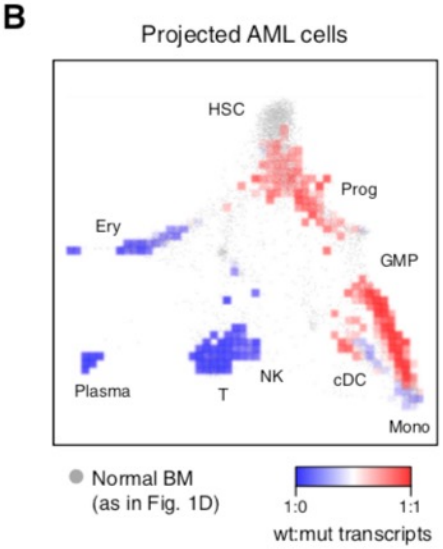


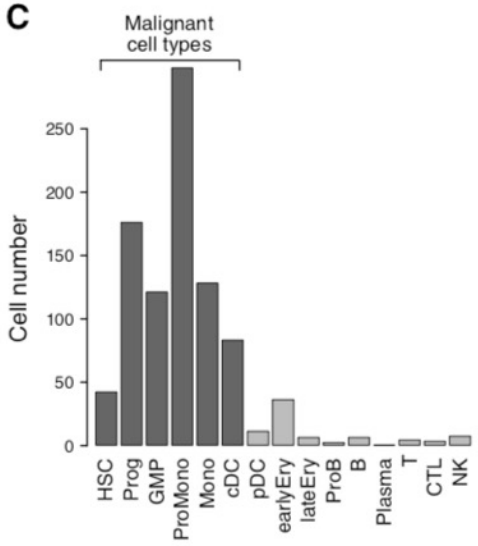


Classifier Distinguishes Cell Types in the AML Ecosystem

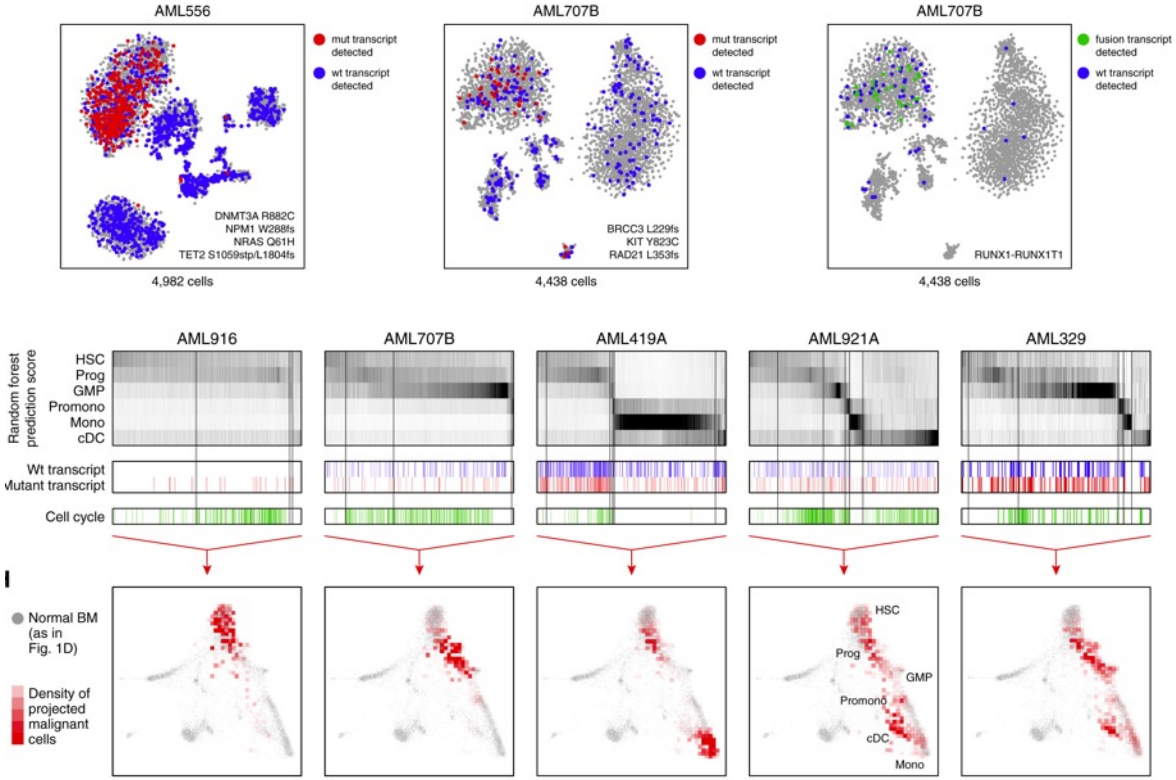








Subtle Differences:
Distinguishing Cell Types in the AML Ecosystem



van Galen et al, *Cell* 2019

Thank You

 igor.dolgalev@nyulangone.org