KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(오프라인)

Best practice for the single-cell data analysis: from basics to advanced topics

박종은 _ KAIST

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 (2.6 월)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	개회사/공지사항전달			
09:30-10:50 (80)	**Best practice for single-cell data analysis**	**박종은 교수**	**Introduction to ML & DNN (이론)**	**이상근 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Practice1: Scanpy basic workflow**	**김우석 김성룡 조교**	**CNN (이론)**	**이상근 교수**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**Public data, batch correction, cell annotation**	**박종은 교수**	**RNN, GAN, XAI (이론)**	**이상근 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**Practice2: Advanced single-cell analysis**	**김우석 김성룡 조교**	**AI 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)**	**이정현 한성민 조교**

DAY2 (2.7 화)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**Introduction to protein structure prediction - Homology modeling - Coevolution-guided modeling Early AI-based approaches**	**백민경 교수**	**Pre-trained Models for Transfer Learning (이론)**	**전민지 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**단백질 구조 예측 실습 - MSA generation, template search - homology modeling contact prediction & modeling**	**백민경 교수**	**Pre-trained Models for Transfer Learning (실습)**	**정민수 조교**
12:10-13:40 (90)	점심			
13:40-15:10 (90)	**AI-based protein structure prediction - AlphaFold/RoseTTAFold Applications to PPI prediction & protein design**	**백민경 교수**	**Deep learning in Bioinformatics**	**노미나 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**단백질 구조 예측 실습 II AlphaFold, RoseTTAFold 실습 및 응용**	**백민경 교수**	**Deep learning model을 이용한 실습**	**곽호진 박예솔 조교**

DAY3 (2.8 수)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**화학정보학 기초(Cheminformatics) 약물특성 및 약물다움(druglikeness) Molecular Notations & Descriptors AI 신약개발을 위한 Databases AI 신약개발을 위한 Programming 기초**	**김동섭 교수**	**마이크로바이옴 기본 이론**	**이선재 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Google Colab에 RDKit 설치 화합물 정보 읽기 실습 Bioactivity database 검색 및 정보 읽기 실습 Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습**	**문채영 나민주 조교**	**16S rRNA amplicon seq. - DADA2**	**서영창 조준우 조교**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**AI 신약개발을 위한 기계학습법 기초 QSAR 모델링 기초 AI 신약개발을 위한 딥러닝 모델 Virtual screening (ligand-based, structure-based) 및 de novo design**	**김동섭 교수**	**최신 메타지놈 분석 기법의 현황**	**이선재 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**QSAR modeling 전체 과정 실습 화합물의 Bioactivity 예측 모델 개발 Virtual screening 과정을 통한 신약후보물질 발굴 실습**	**문채영 나민주 조교**	**Shotgun metagenome 분석 (Linux)**	**서영창 조준우 조교**

강의개요

Best practice for the single-cell data analysis: from basics to advanced topics

우리 몸을 세포 수준에서 이해하고자 하는 노력은 single-cell genomics 라는 새로운 기술의 발달로 이어졌으며, 최근 쏟아지고 있는 single-cell 데이터는 생명정보학의 새로운 중요한 재료가 되고 있다. 본 강의에서는 single-cell 데이터 분석을 위한 best practice를 정의해 보고자 한다. 초심자를 위해 single-cell 데이터의 특성과 기본 분석법, 자주 발생하는 오류들과 이를 피하기 위한 방법들을 설명하고, 공공 데이터의 활용법, 머신 러닝을 활용한 손쉬운 세포 타입 annotation, 딥러닝 기반의 batch correction 방법 등도 간단한 실습을 통해 소개한다. Python과 google colab 을 활용한 실습 진행을 포함한다.

- Single-cell data structure (multi-dimension data analysis, data sparsity)
- Basic analysis pipeline
- Common erros in single-cell data analysis
- Batch correction and assessing the integration
- Public data analysis
- Automatic cell type annotation

* 참고 웹사이트:
  https://scanpy.readthedocs.io/en/stable/index.html

* 교육생준비물: 노트북

* 강의 난이도: 초급-중급

* 강의: 박종은 교수 (한국과학기술원 의과학대학원)

Curriculum Vitae

Speaker Name: Jong-Eun Park, Ph.D.

▶ **Personal Info**

Name	Jong-Eun Park
Title	Assistant Professor
Affiliation	KAIST, GSMSE

▶ **Contact Information**

Address	Graduate School of Medical Science and Engineering, KAIST, Daejeon, 34141
Email	jp24@kaist.ac.kr
Phone Number	010-4528-8702

Research Interest

Single-cell genomics, Immunology, Cancer

Educational Experience

2009	B.S. in Seoul National University, Biological Science, South Korea
2015	Ph.D. in Seoul National University, Biological Science, South Korea

Professional Experience

2015-2017	Post-doc research fellow, IBS center for RNA biology, Seoul National University
2017-2020	Post-doc research fellow, Wellcome Sanger Institute, United Kingdom
2020-	Assistant Professor, KAIST

Selected Publications (5 maximum)

1. **Park, J.-E.***, … , Taghon, T., Haniffa, M., Teichmann, S.A., **2020**, A cell atlas of human thymic development defines T cell repertoire formation. **Science** 367, eaay3224
2. Polański, K.*, Young, M.D.*, Miao, Z., Meyer, K.B., Teichmann, S.A. and **Park, J.-E.**, 2019. BBKNN: Fast Batch Alignment of Single Cell Transcriptomes. **Bioinformatics.** 36, 964-965
3. Son, A.*, **Park, J.-E.***, Kim, V.N., **2018**. PARN and TOE1 Constitute a 3′ End Maturation Module for Nuclear Non-coding RNAs. **Cell Reports.** 23, 888–898.
4. **Park, J.-E.***, Yi, H.*, Kim, Y.*, Chang, H., Kim, V.N., **2016.** Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. **Molecular Cell.** 62, 462–471.

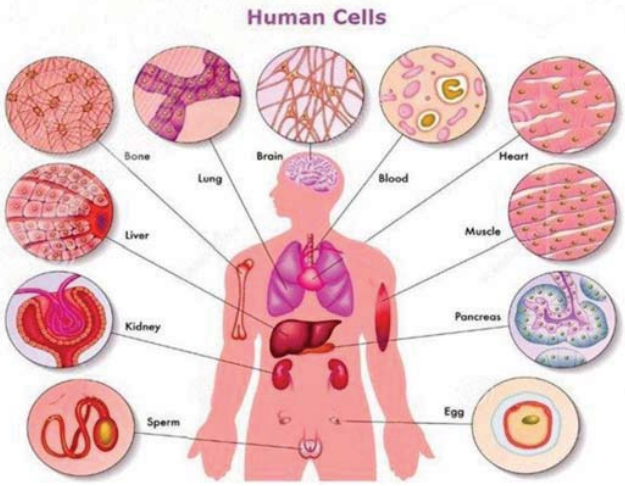
KSBi-BIML 2023

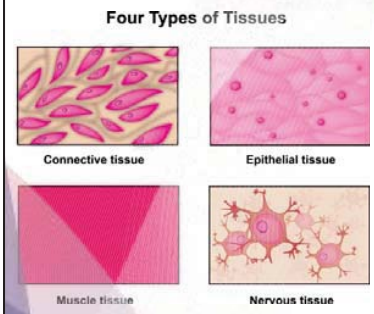
Best practice in single-cell analysis
(python version)

Understanding the complexity of human tissue

“Cellular heterogeneity of human tissue” created by DALL-E

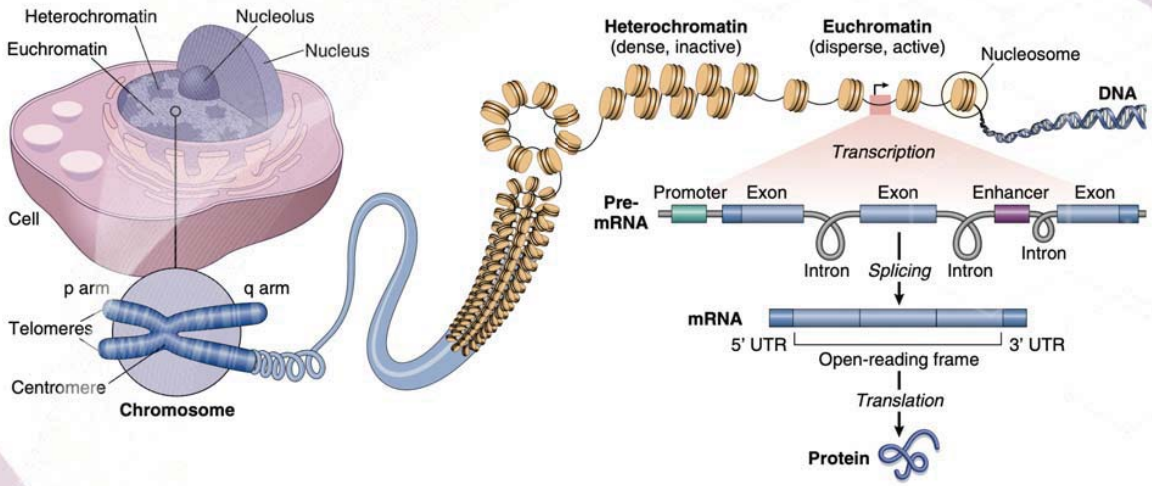
Diversity of the cells in our body





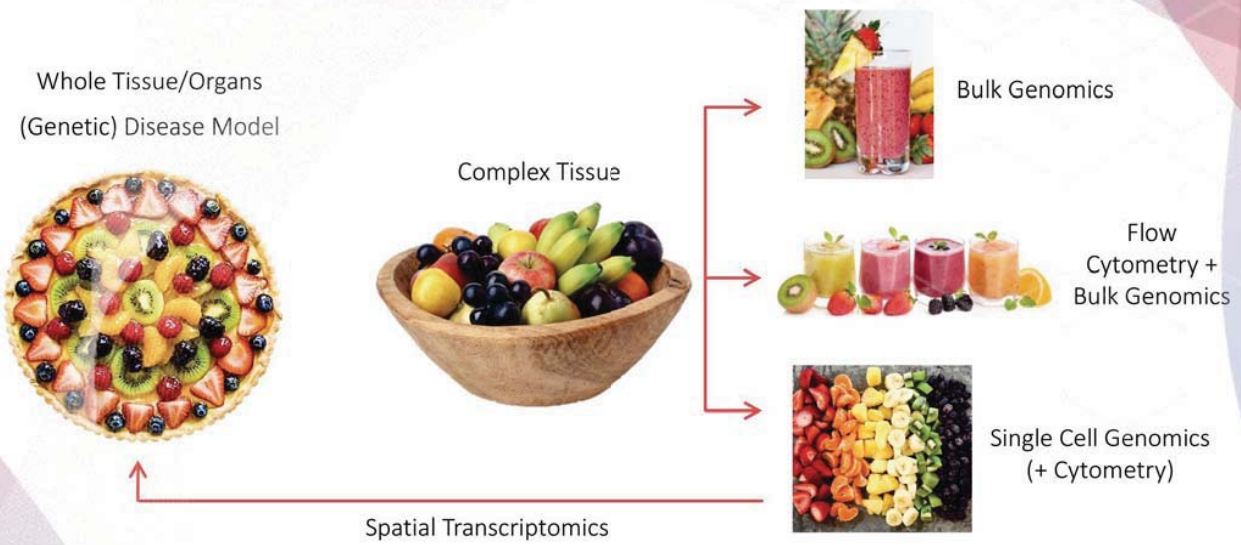
4 tissues, 18 tissue types, 78 organs, > 200 cell types, 37 trillion cells

Data stored in our genome



>20,000 protein coding genes, ~20,000 non-coding genes (more variants)
?? Number of regulatory elements
3 billion DNA letters (nucleotides) of reference, 6 billion personal genome

Single-cell analysis toolkits to understand CellxGene



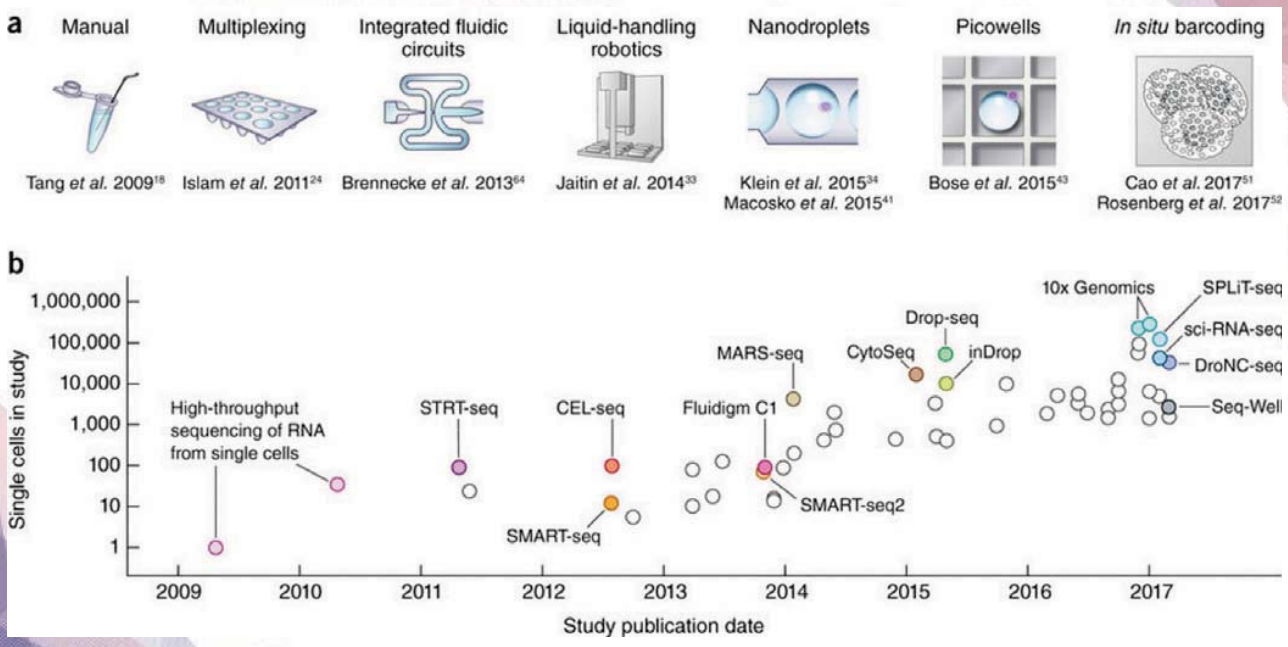
강의 개요

1. Single-cell data 의 특성 및 분석의 개요
2. Best practice in single-cell data analysis
3. Public databases & data integration
4. Single-cell multi-omics data analysis

6

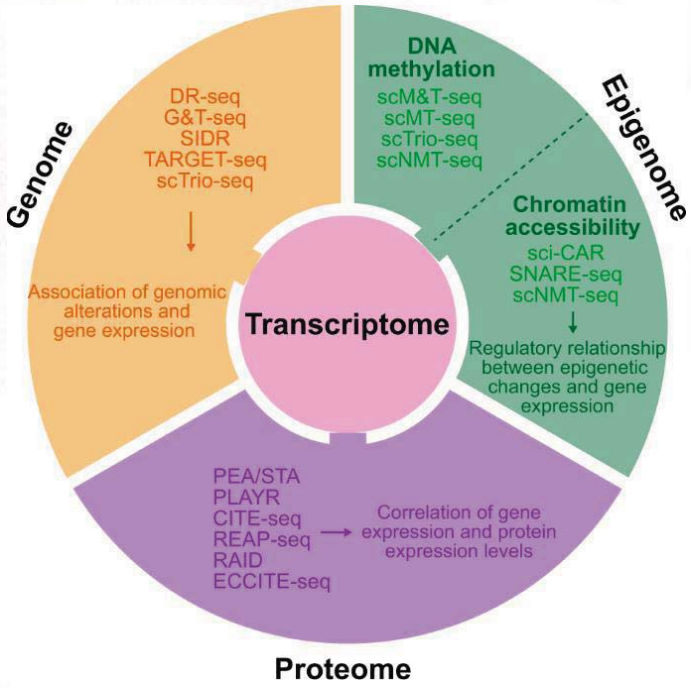
1. Single-cell data 의 특성과 분석 개요

Advancement in single-cell technologies



Svensson & Vento-Tormo, 2018, Nature biotechnology

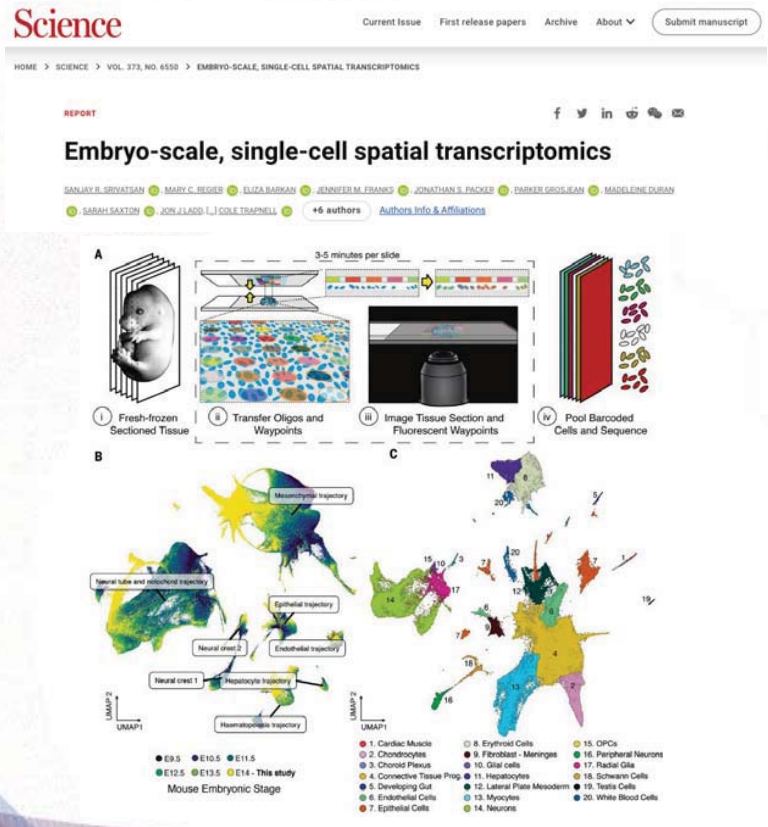
Multi-omics at single-cell resolution



Lee, J., Hyeon, D.Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* **52**, 1428–1442 (2020).

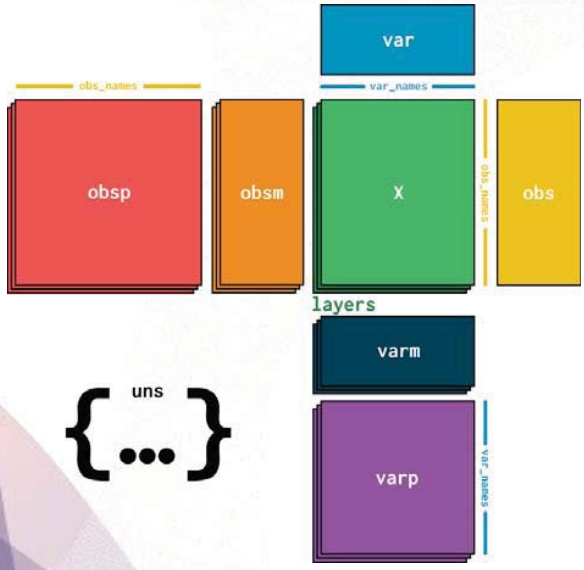
9

Multi-omics at single-cell resolution



10

Single-cell data structure: Anndata format



- X: Gene expression values
- Obs: observations, cells
- Var: variables, genes
- Obsm: PCA dimensions, …
- Obsp: Graph connectivities, …
- Varm: PCA loadings, …
- Varp: Gene-wise correlations, …
- Uns: anything

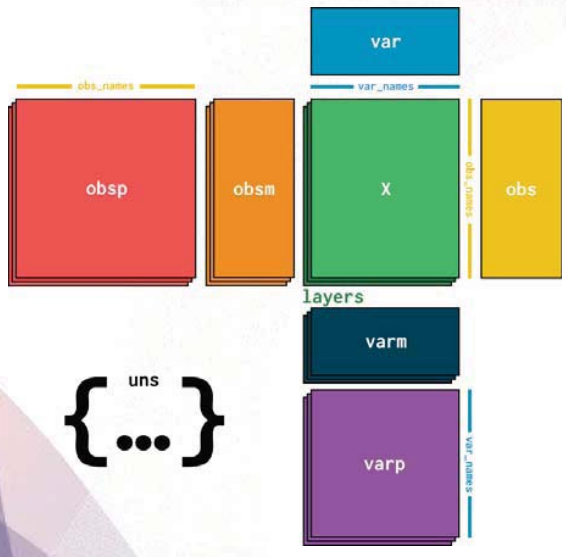
- Based on hdf5 format
- Efficient memory usage and storage
- Easy plug-in with machine-learning packages

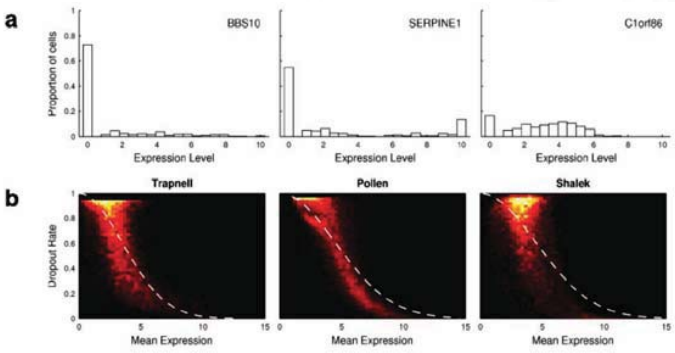
Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf
bioRxiv 2021.12.16.473007; doi: https://doi.org/10.1101/2021.12.16.473007

https://anndata.readthedocs.io/en/latest/index.html

11

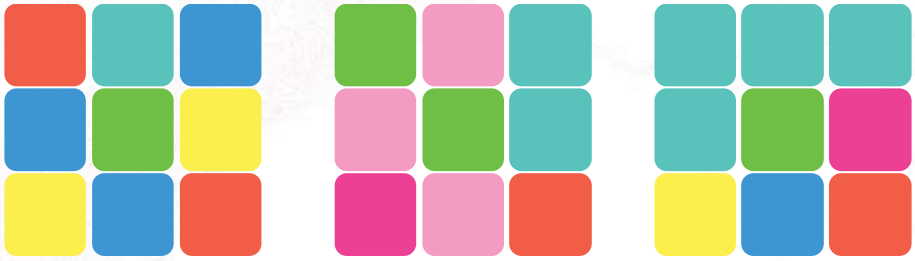
Sparsity, dropouts in single-cell data



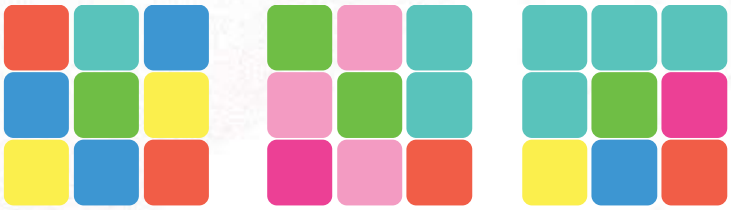


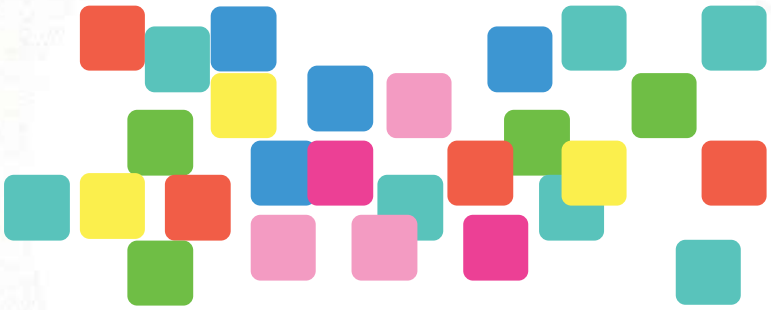
12

Example: 3 different cell types



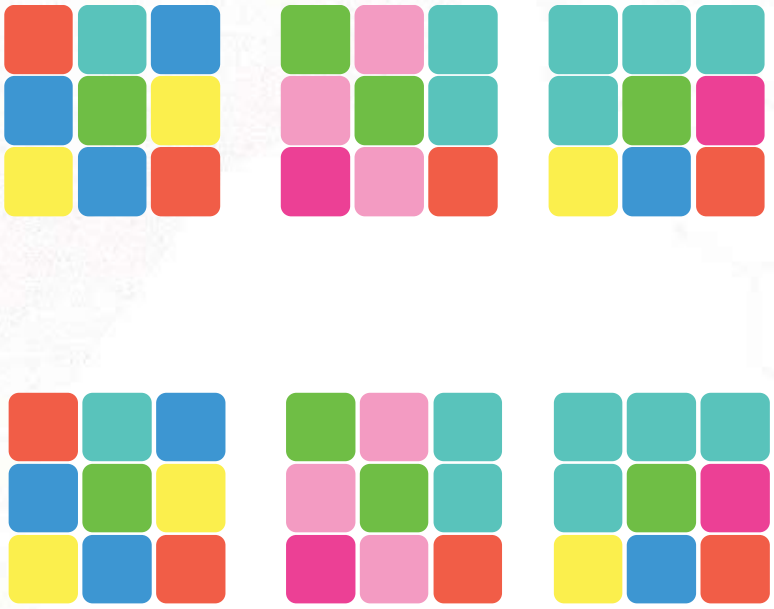
Example: 3 different cell types





bulk RNA-seq

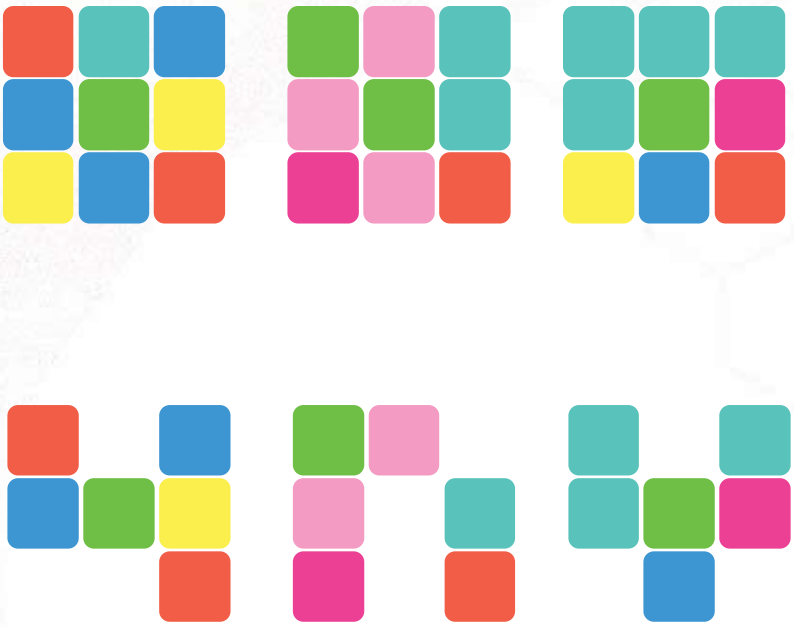
Example: 3 different cell types



ideal world single cell data

15

Dropouts in single-cell data

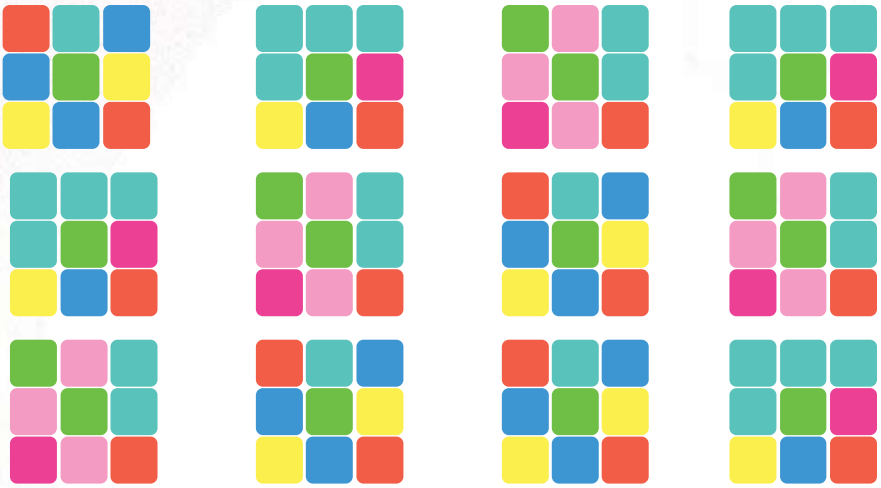


real world single cell data

16

Profiling same cells for multiple times

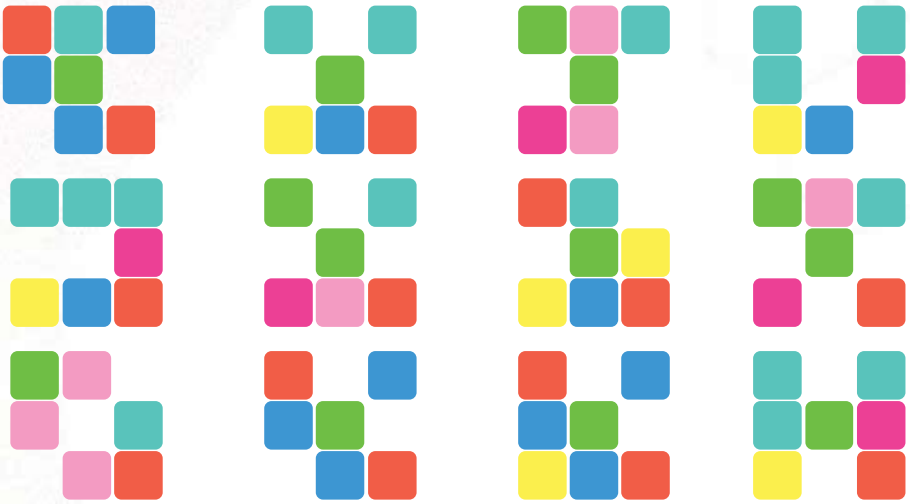
multiple cells sequenced



17

Simulating random dropouts

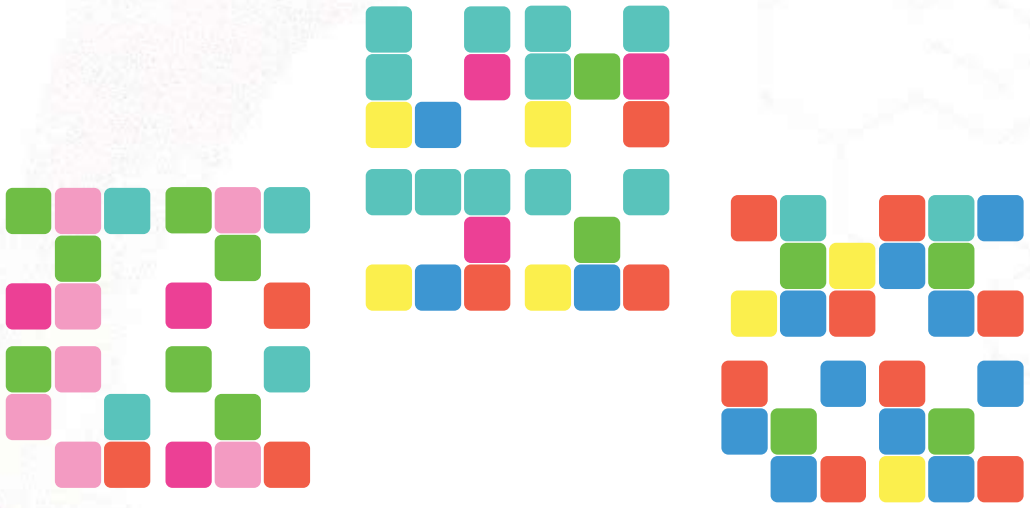
assuming low mRNA capture rate



18

Clustering keeping the resolution

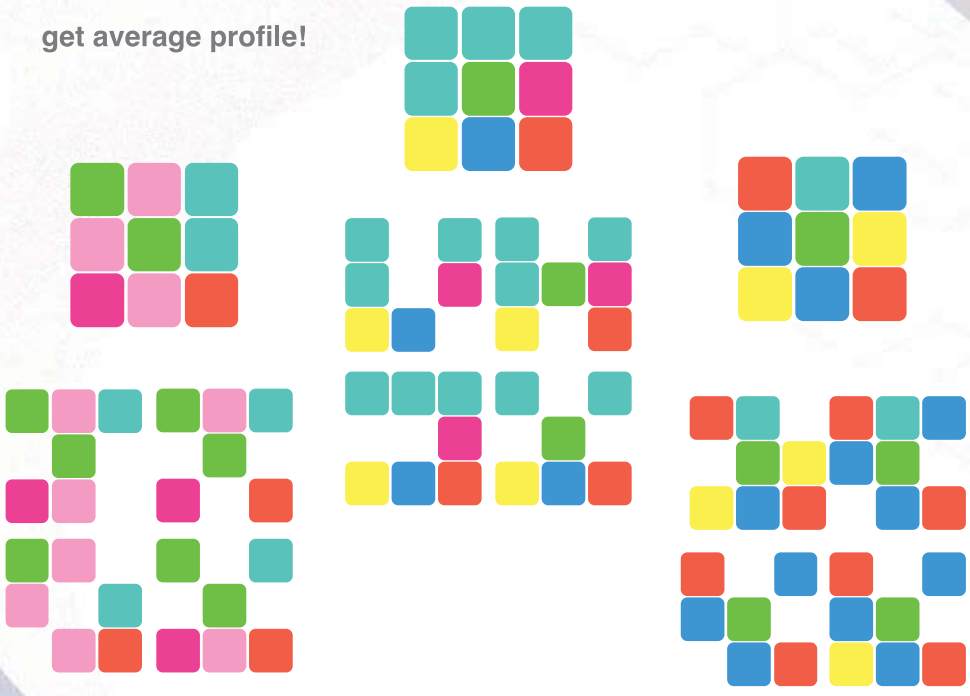
compare to each other & create clusters



19

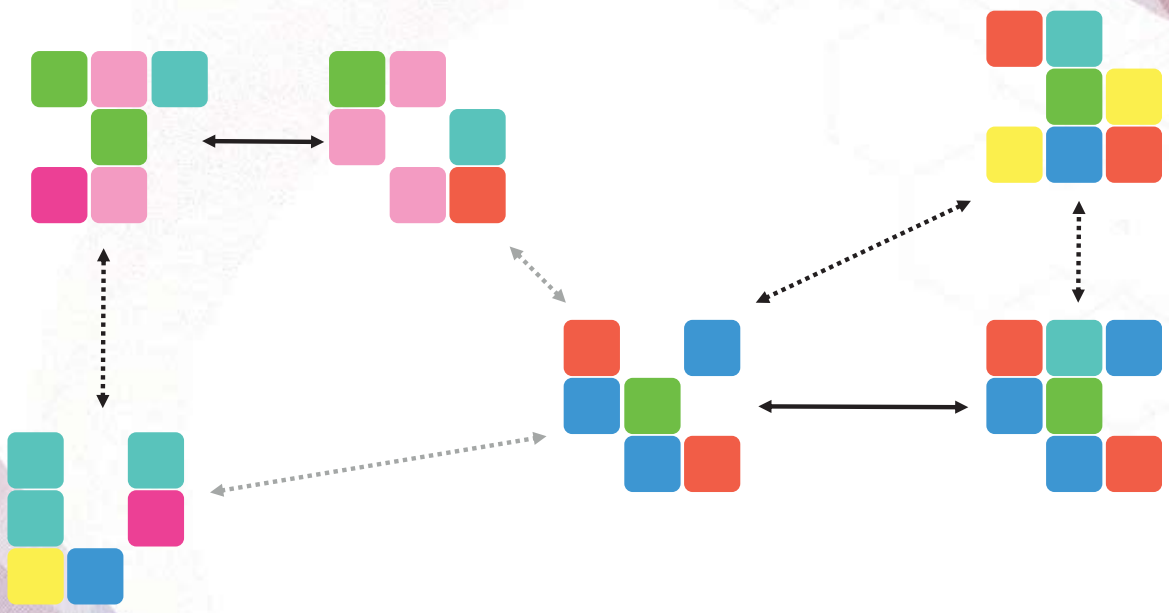
Power of taking average!

get average profile!



20

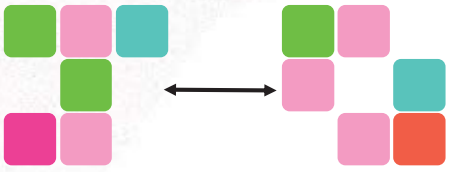
It's all about finding neighbors

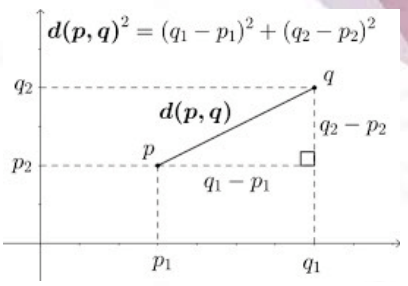


21

How to find neighbors?

How to measure distance between cells?



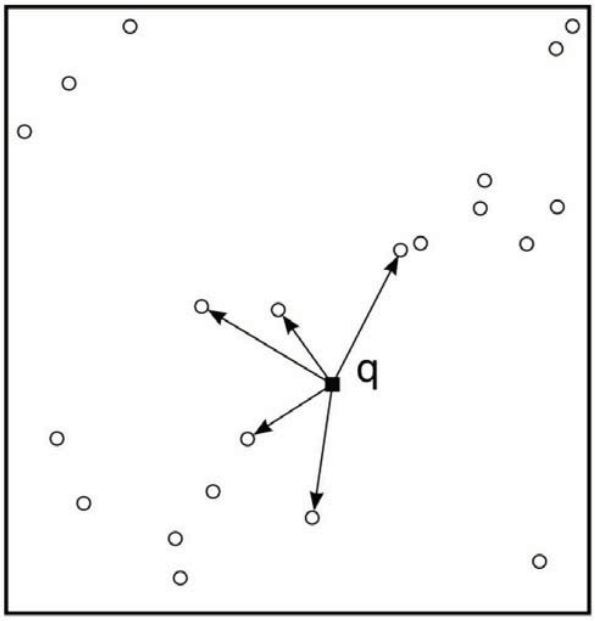


2	1	1
1	1	0
2	3	1
0	1	1
1	0	1

22

Finding k-nearest neighbors

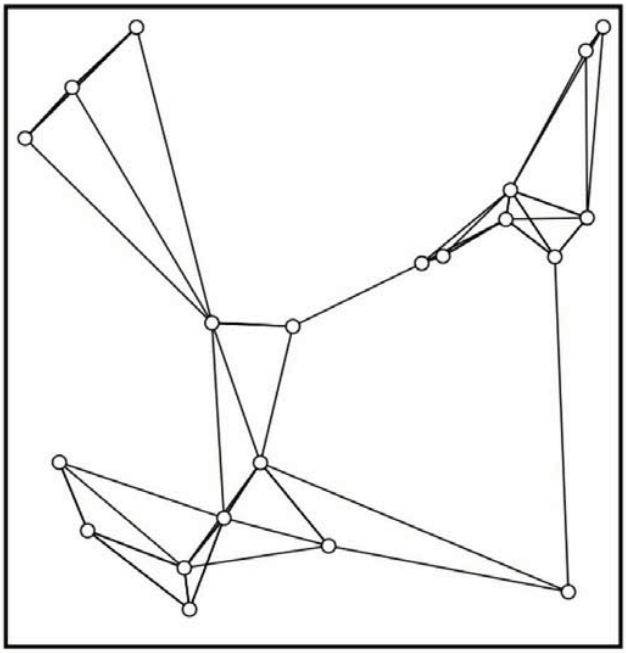
k-nearest neighbors, $k = 5$



23

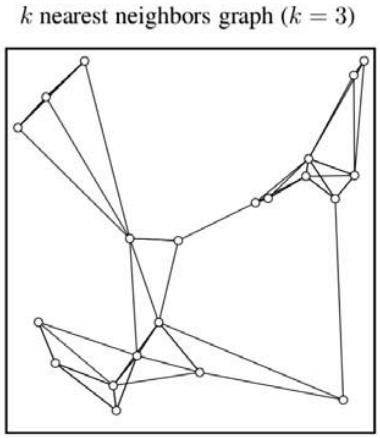
Finding k-nearest neighbors

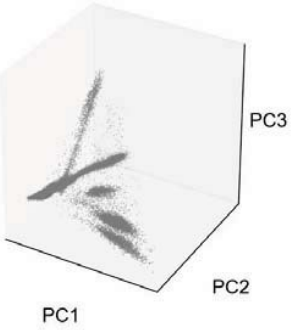
k nearest neighbors graph ($k = 3$)

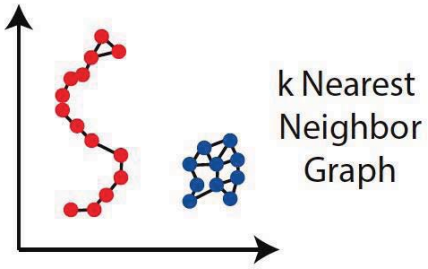


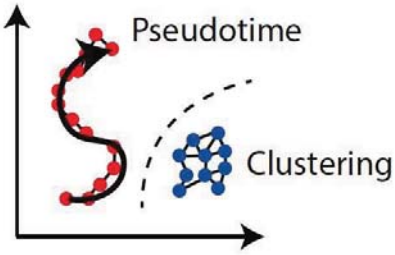
24

Power of neighbor graph in single-cell analysis









25

Overview of single-cell analysis

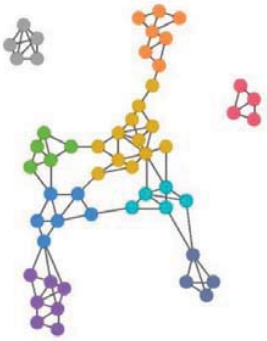
Gene expression

↓

Dimension reduction (e.g. PCA)

↓

KNN-graph



→ UMAP

→ Force-directed graph

→ Diffusion map

Graph-embedding

↓

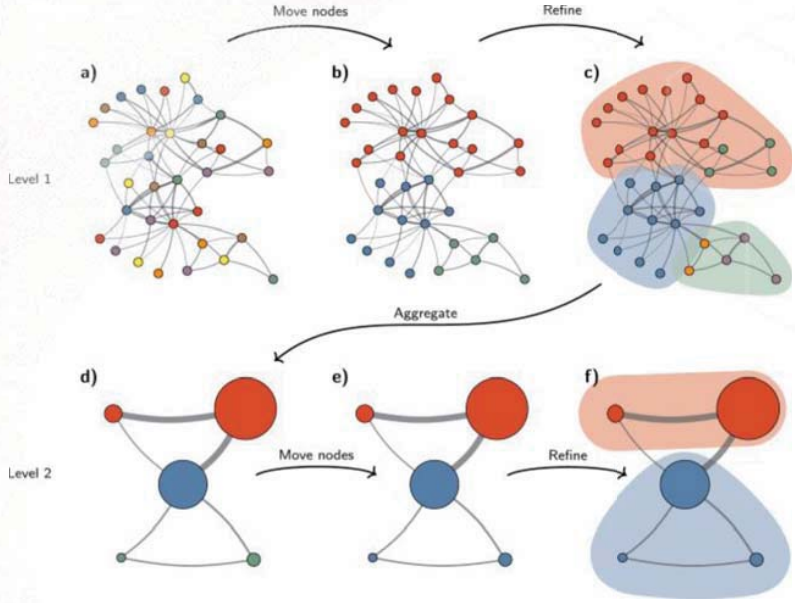
Clustering, annotation, pseudotime….

Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746.

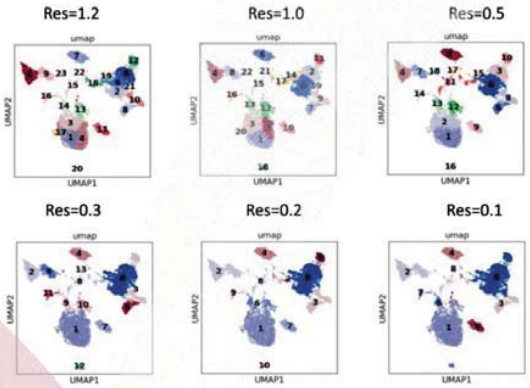
박종은(2018). 단일 세포 RNA 시퀀싱(Single-cell RNA sequencing) 기술 동향. BRIC View 2018-T28.

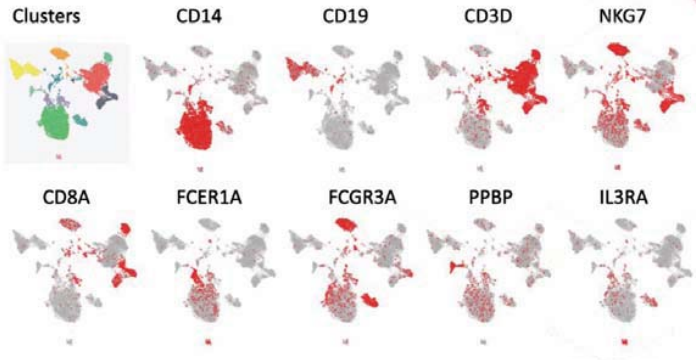
26

Graph based clustering



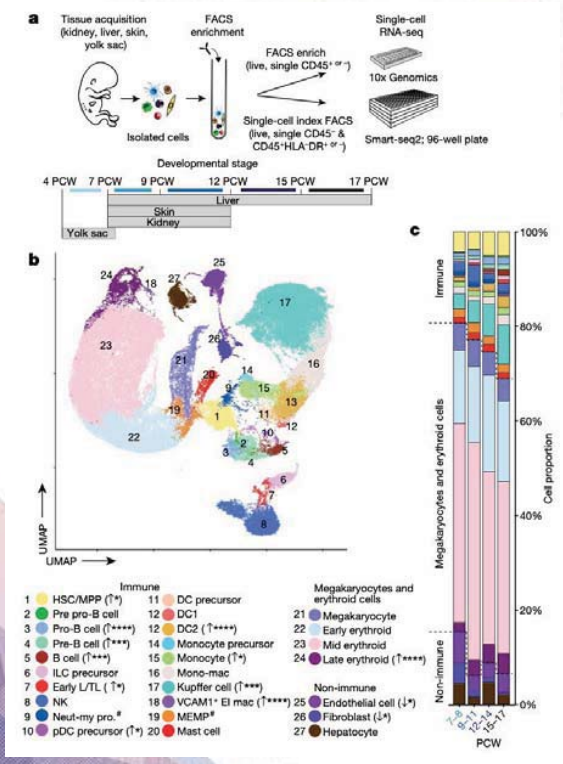
Cell type annotation





Annotating cell types

Overview of single-cell analysis



Popescu DM, Botting RA, Stephenson E, et al. Decoding human fetal liver haematopoiesis. *Nature*. 2019 29

2. Best practice in single-cell data analysis

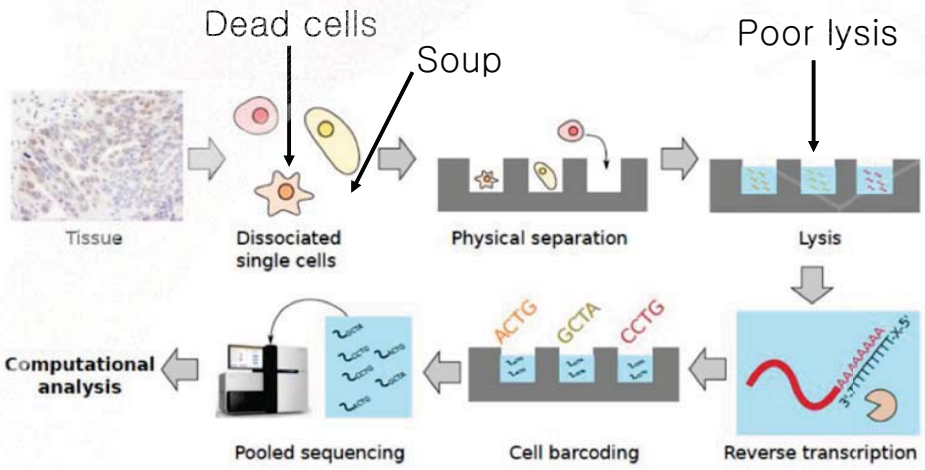
30

2-1. Sample QC

31

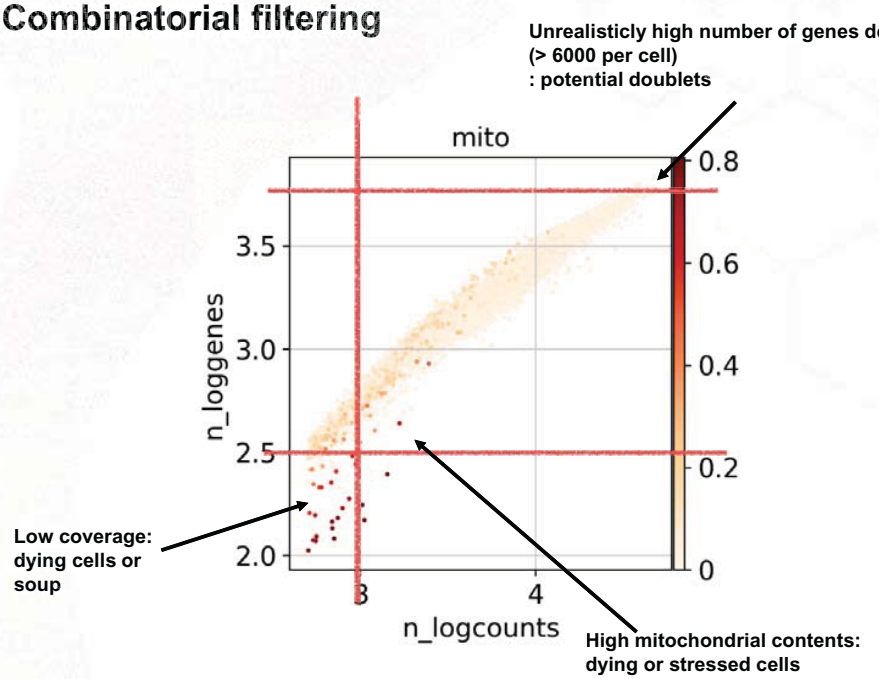
How can we define cells?

Sample QC: why?



Loss during library prep

Combinatorial filtering



Best practice

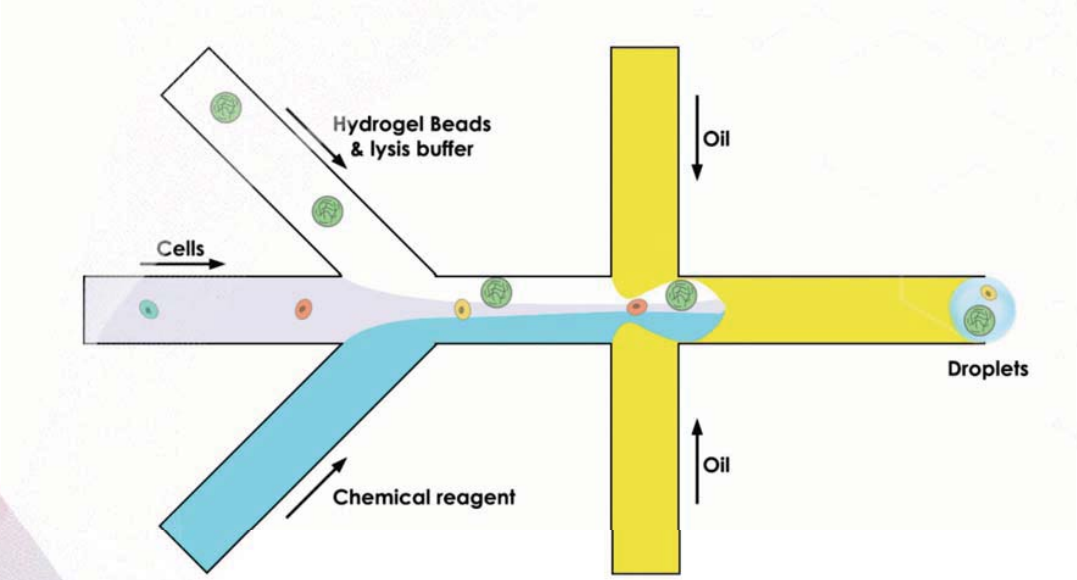
- Perform QC for individual sample
- Draw UMAP for individual sample

: color by well-known markers, mito-genes, n_genes, Immunoglobulins, hemoglobins, etc…

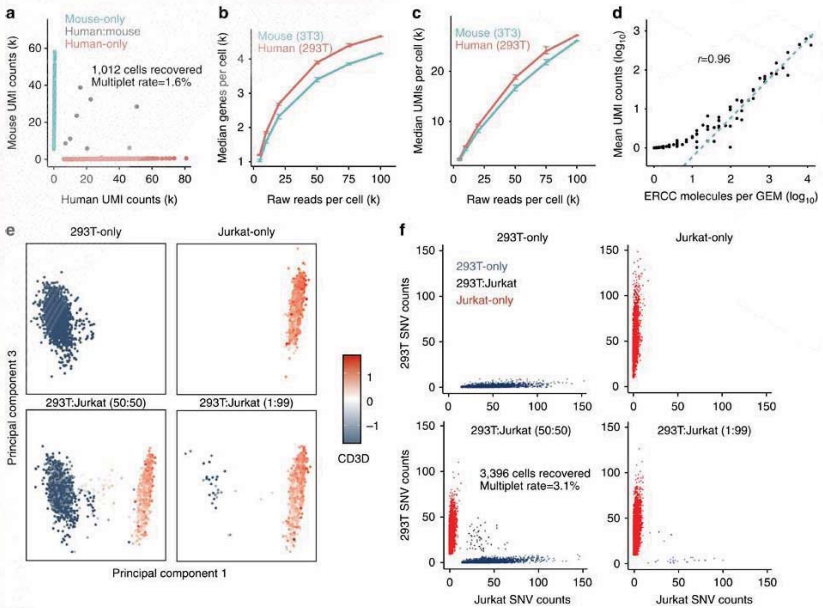
- Try to find best universal cutoff
- You might need to adjust cutoff for some low-quality samples (or simply discard them)

Problem 1: Doublets

Doublets expected from droplet based methods



Doublets expected from droplet based methods

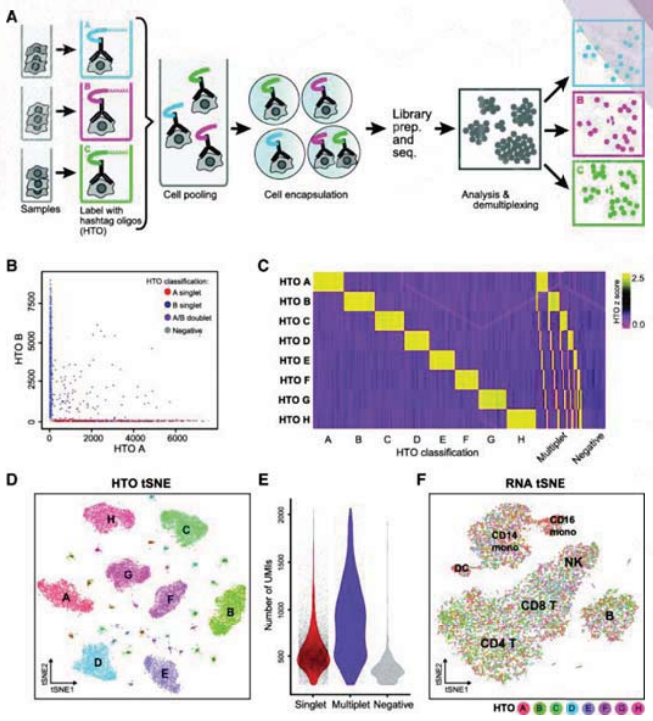


https://www.nature.com/articles/ncomms14049/figur

Doublet rates from 10X Genomics platform

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~800	~500
~0.8%	~1,600	~1,000
~1.6%	~3,200	~2,000
~2.3%	~4,800	~3,000
~3.1%	~6,400	~4,000
~3.9%	~8,000	~5,000
~4.6%	~9,600	~6,000
~5.4%	~11,200	~7,000
~6.1%	~12,800	~8,000
~6.9%	~14,400	~9,000
~7.6%	~16,000	~10,000

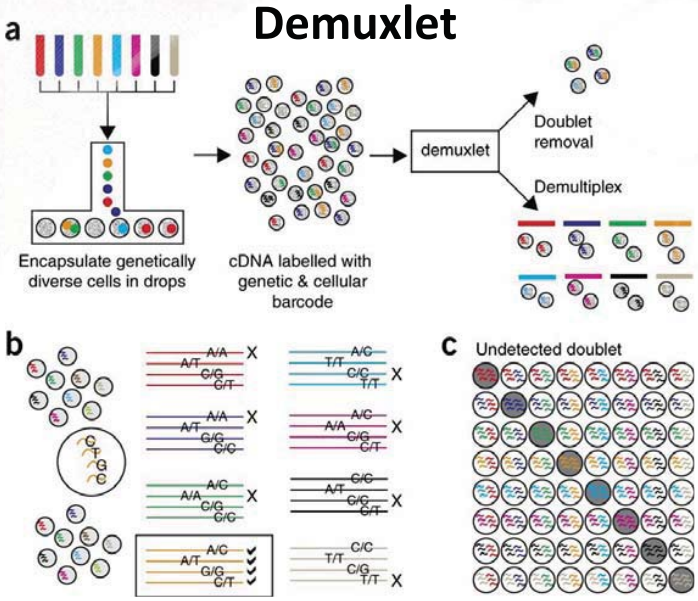
Cell Hashing allows detection of doublets



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1603-1

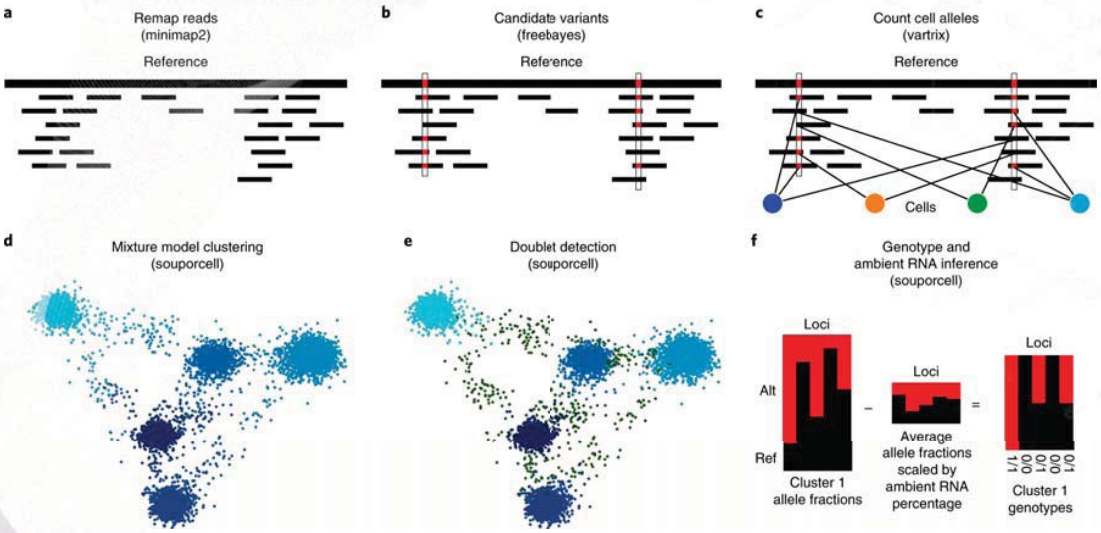
Genotype mixing allows detection of doublets

Demuxlet



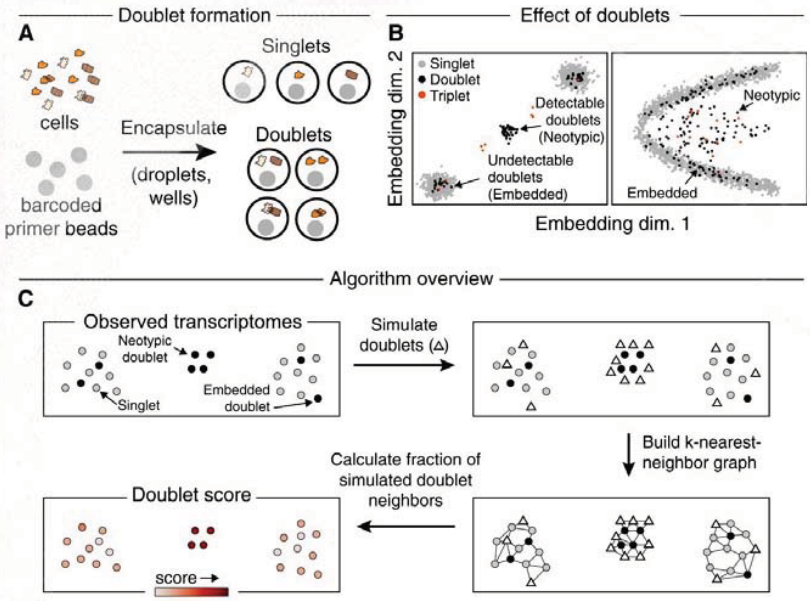
https://www.nature.com/articles/nbt.4042

Souporcell



https://www.nature.com/articles/s41592-020-0820-1

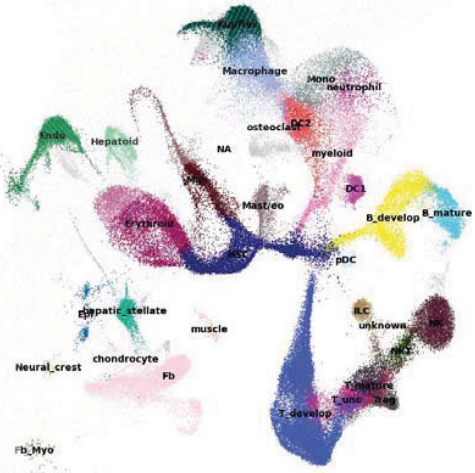
Computationally predicting doublets

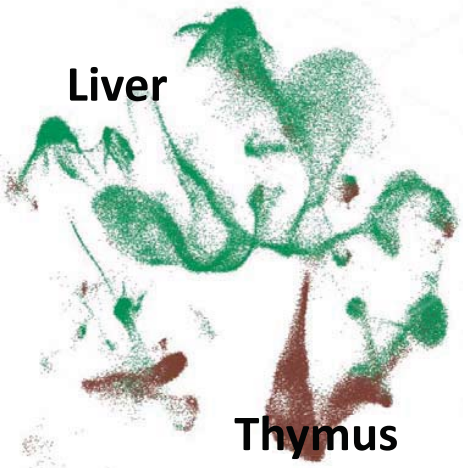


https://www.sciencedirect.com/science/article/pii/S2405471218304745

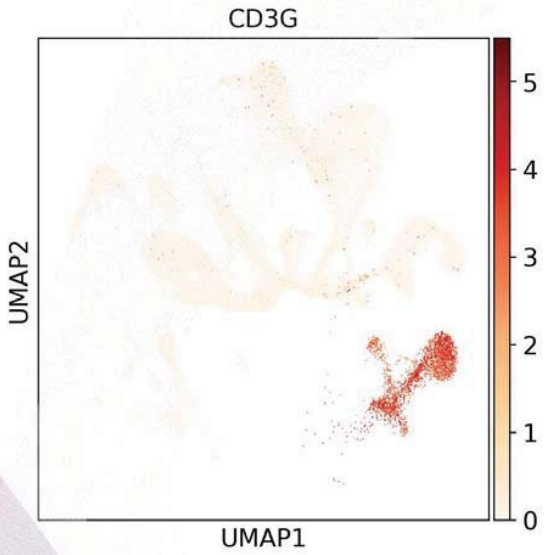
Problem 2: Contaminating reads from ambient RNAs

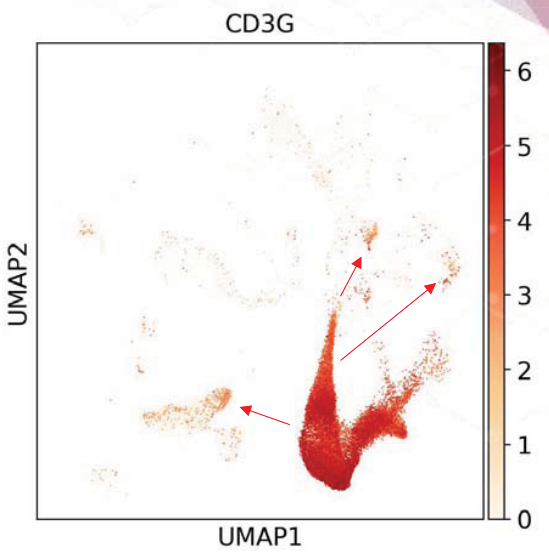
Example for the ambient RNA contamination



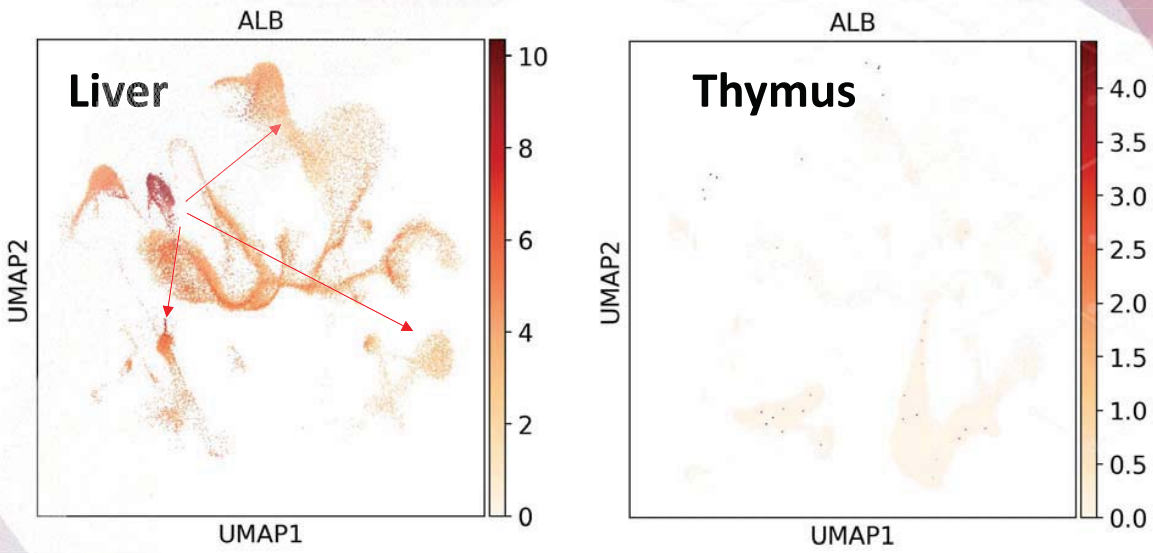


Example for the ambient RNA contamination

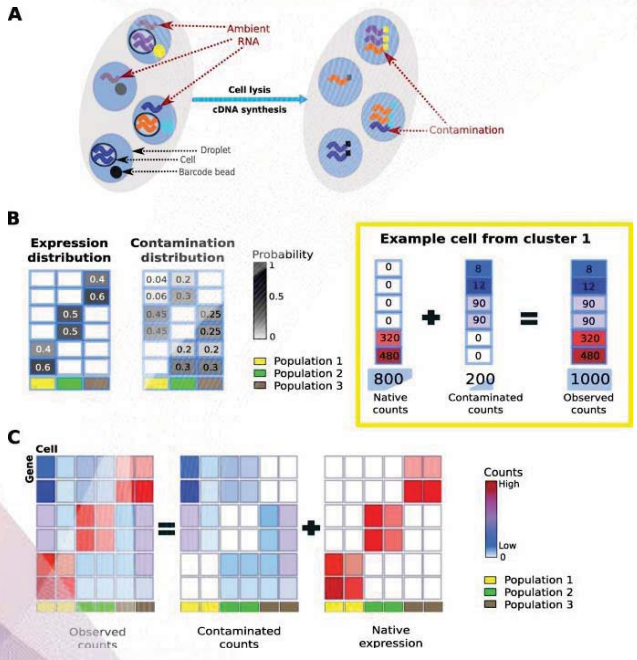




Example for the ambient RNA contamination



DecontX

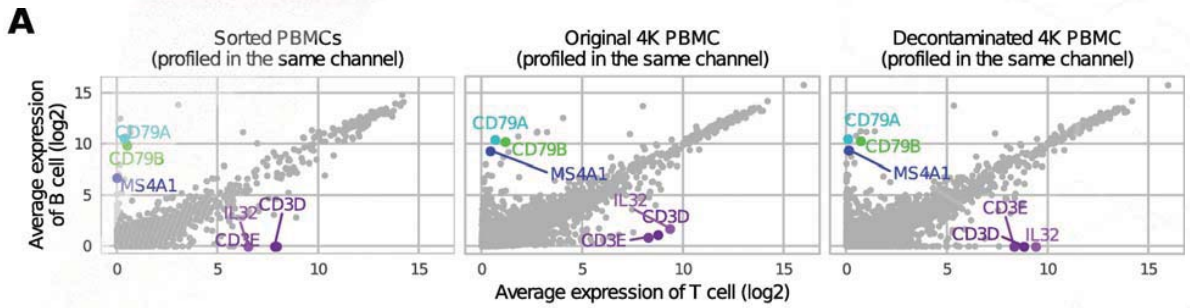


Decontamination of ambient RNA in single-cell RNA-seq with DecontX

Shivi Yang, Sean E. Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima & Joshua D. Campbell

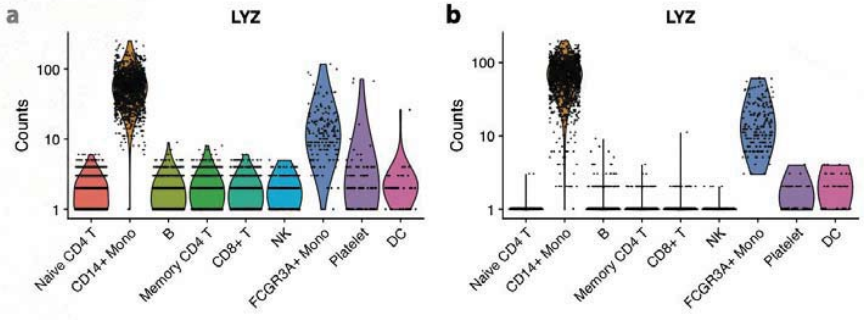
Genome Biology **21**, Article number: 57 (2020) | Cite this article

Cell Hashing allows detection of doublets



CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets

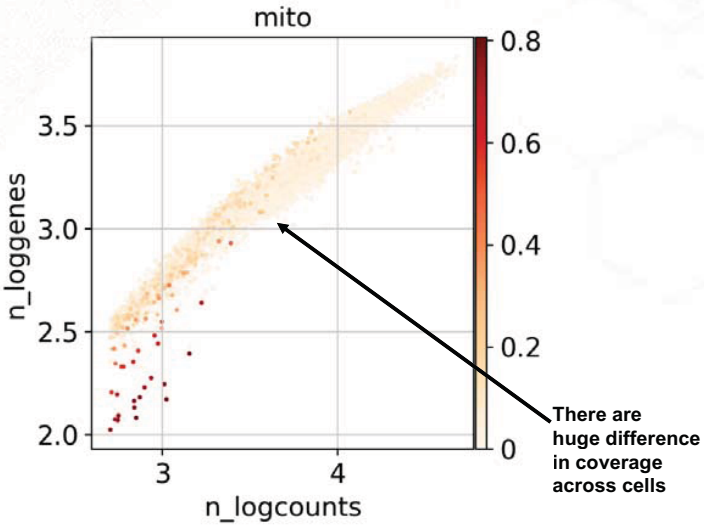
Stephen J. Fleming[1,2], John C. Marioni[3,4], and Mehrtash Babadi[1,2]



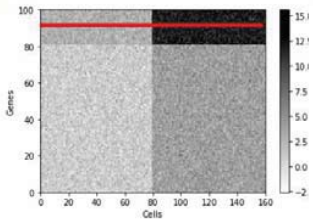
2-2. Preprocessing (normalization, scaling)

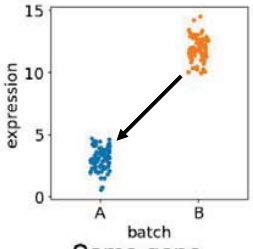
51

Normalisation



Why do we need normalisation?





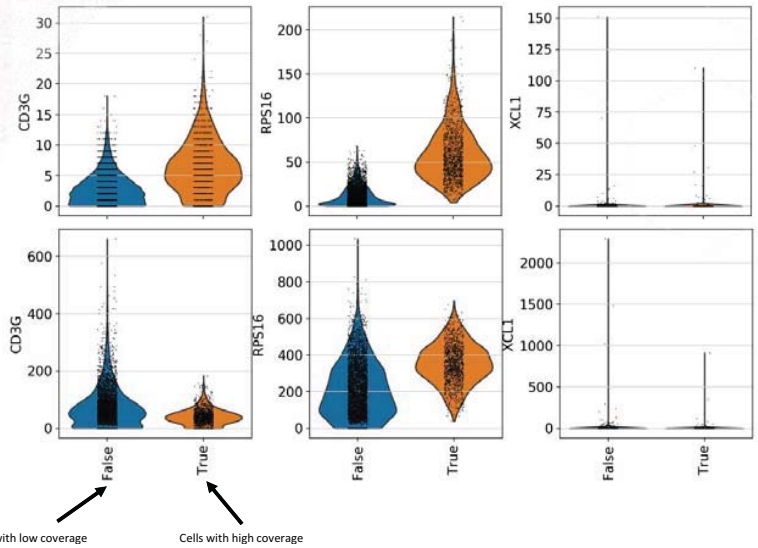
Same gene

Normalisation method

Xnorm = (X / sum(X)) * 10000

before

after



Cells with low coverage

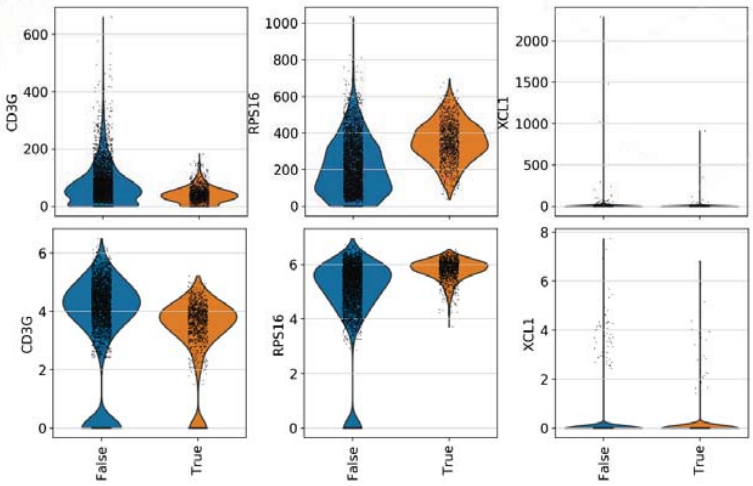
Cells with high coverage

Log-transformation

Xlog = log(Xnorm+1)

before

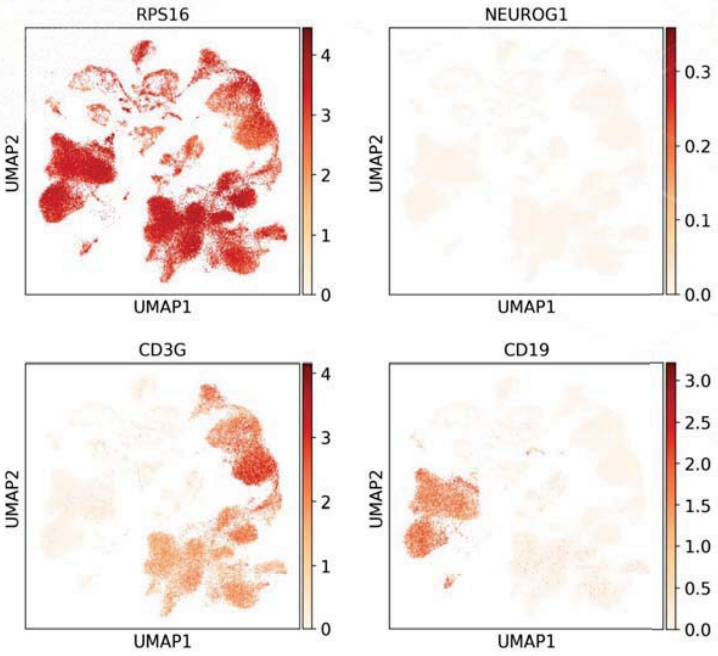
after



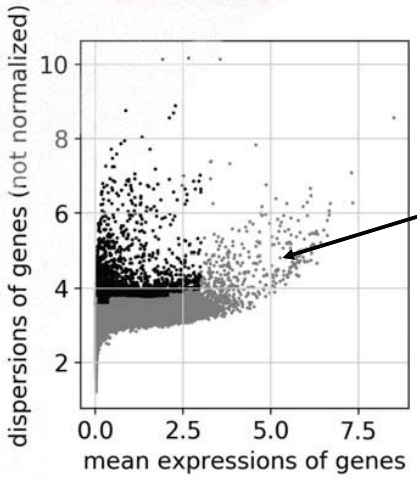
Selecting highly variable genes

Not informative

Highly informative

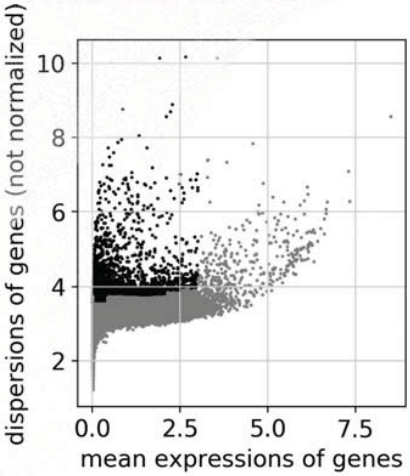


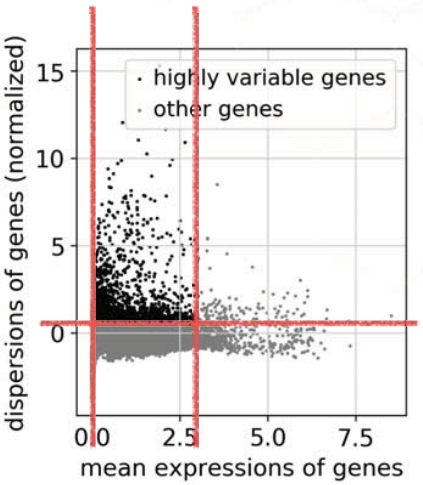
Selecting highly variable genes



Higher the expression level, also higher the dispersion

Selecting highly variable genes





After normalisation, we can apply flat cutoff!

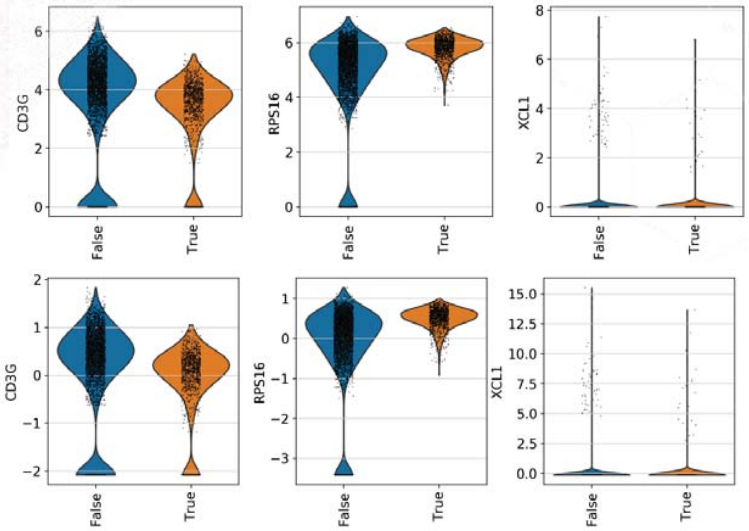
Scaling

centralise gene expression with zero mean and unit variance

"Making all genes equally important"

before

after



Raw count

Integer between 0-10000

Normalized

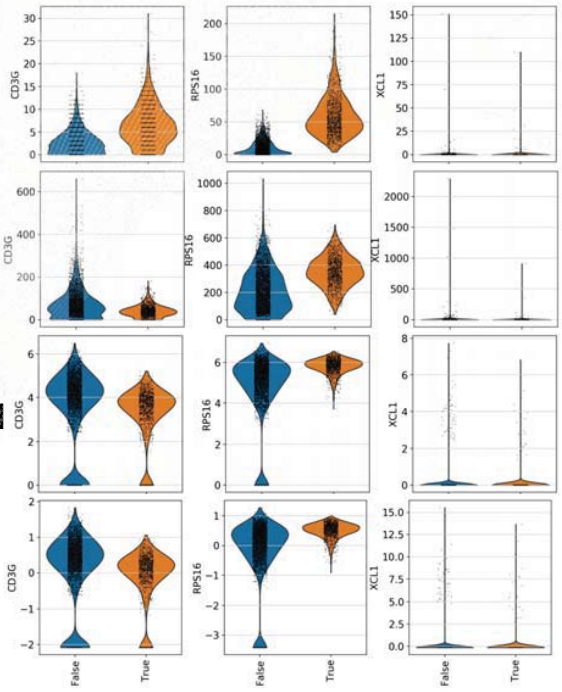
float between 0-10000

Log-transformed

float between 0-20

scaled

float between –x.xx ~ + x.xx



Hvg selected: < 5000 gene number

2-3. Dimension reduction / neighborhood graphs

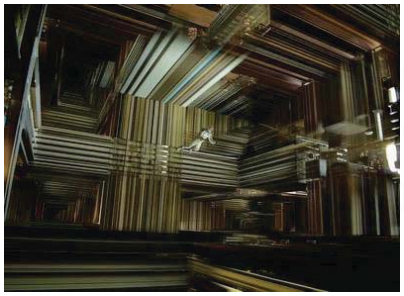
Single-cell data is high dimensional!

Cells in high-dimensional space (> 30000 genes)

Genes (30,000)

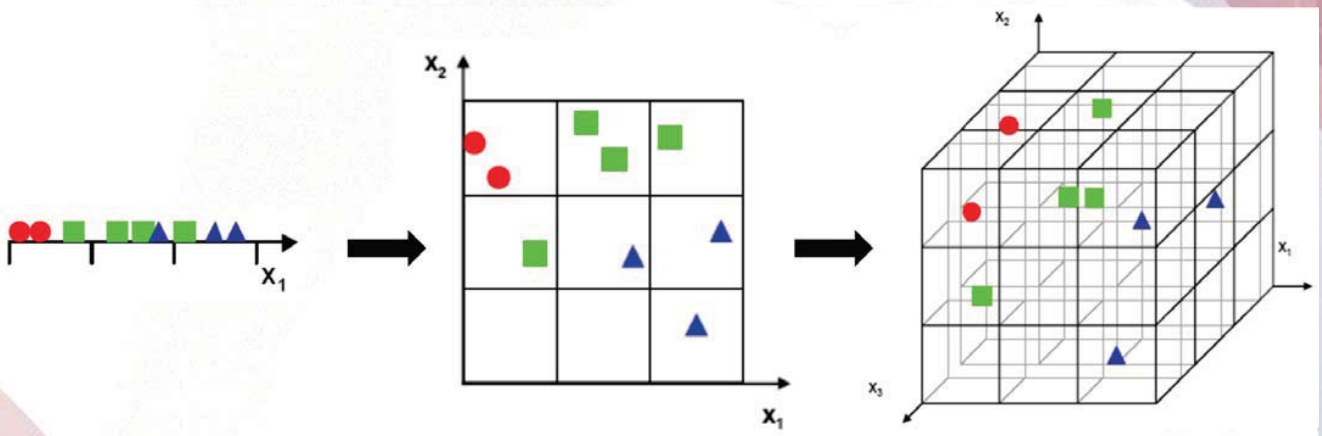
Cells (up to 1,000,000)

0	4	0	3	…	9	0	5	2
0	6	1	0	…	7	3	0	4
2	5	0	5	…	6	0	3	3
:	:	:	:		:	:	:	:
0	3	0	0	…	8	2	1	1
1	9	2	2	…	5	1	4	5
0	0	1	0	…	0	2	1	0



Interstellar, 2014

Curse of dimensionality



low dimension
dense
easy to compare

high dimension
sparse
difficult to compare

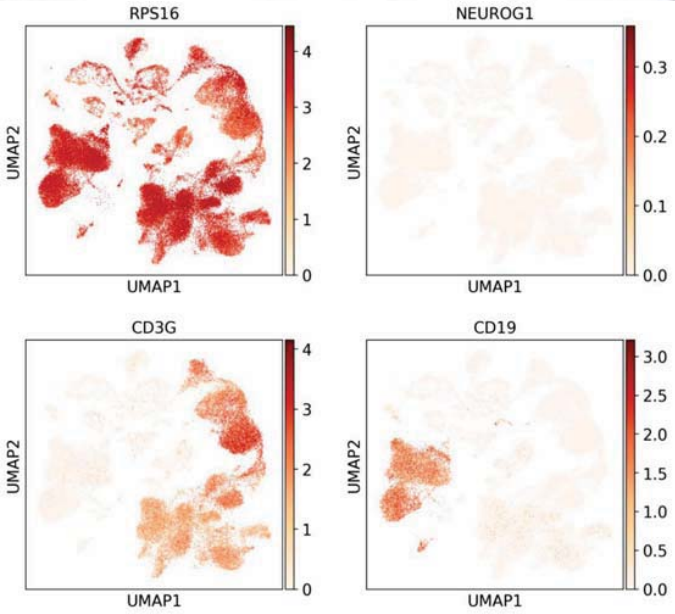
63

Feature selection

Genes (30,000)

Cells (up to 1,000,000)

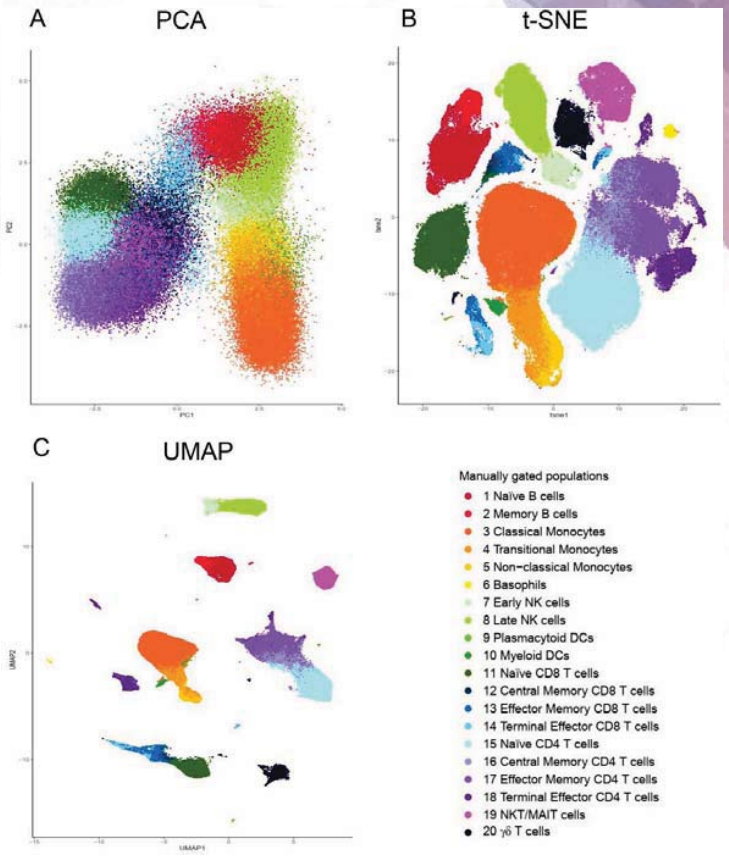
0	4	0	3	…	9	0	5	2
0	6	1	0	…	7	3	0	4
2	5	0	5	…	6	0	3	3
:	:	:	:		:	:	:	:
0	3	0	0	…	8	2	1	1
1	9	2	2	…	5	1	4	5
0	0	1	0	…	0	2	1	0



64

Dimension reduction

Finding latent space embedding
Finding key axis
PCA, CCA, NMF, t-SNE, UMAP, FDG



Liu, Peng, et al. "Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data." *Frontiers in cell and developmental biology* 8 (2020): 234.

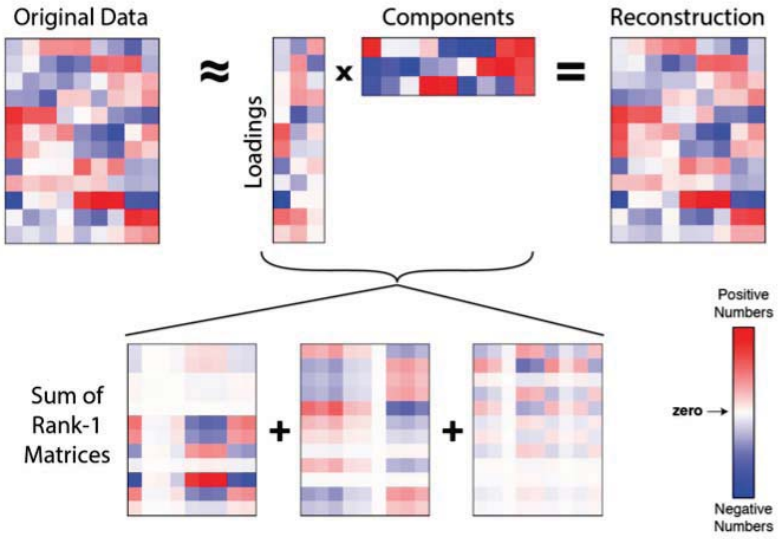
65

Dimension reduction

Gene expression

↓

Principal components



Matrix decomposition

Other examples:
Non-Negative Matrix Factorization
Canonical Correlation Analysis
Bayesian modelling

Importance of neighborhood graph in single-cell data analysis

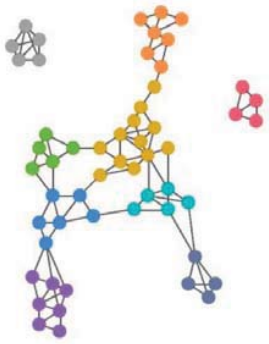
Gene expression

↓

Dimension reduction (e.g. PCA)

↓

KNN-graph



→ UMAP

→ Force-directed graph

→ Diffusion map

Graph-embedding

↓

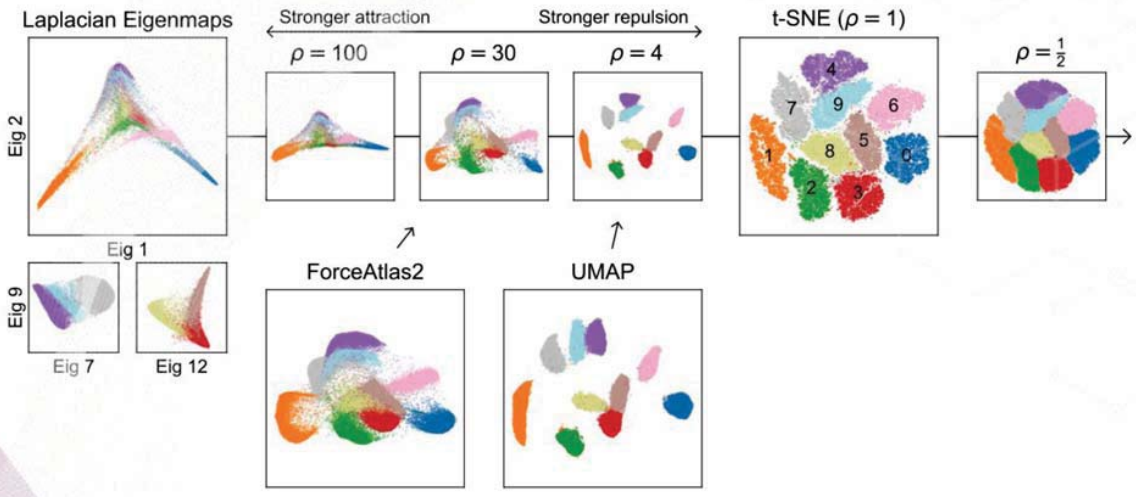
Clustering, annotation, pseudotime….

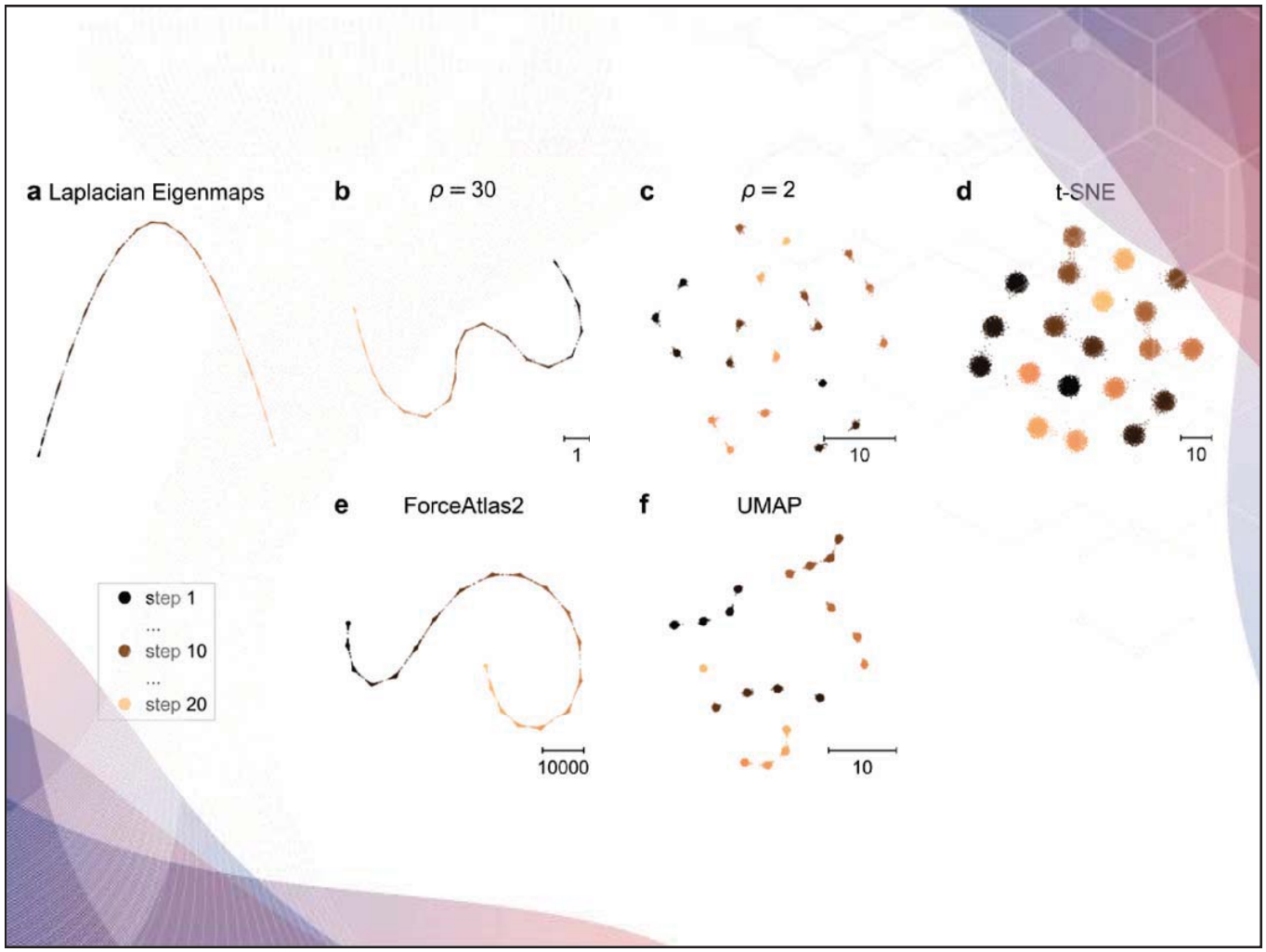
Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746.

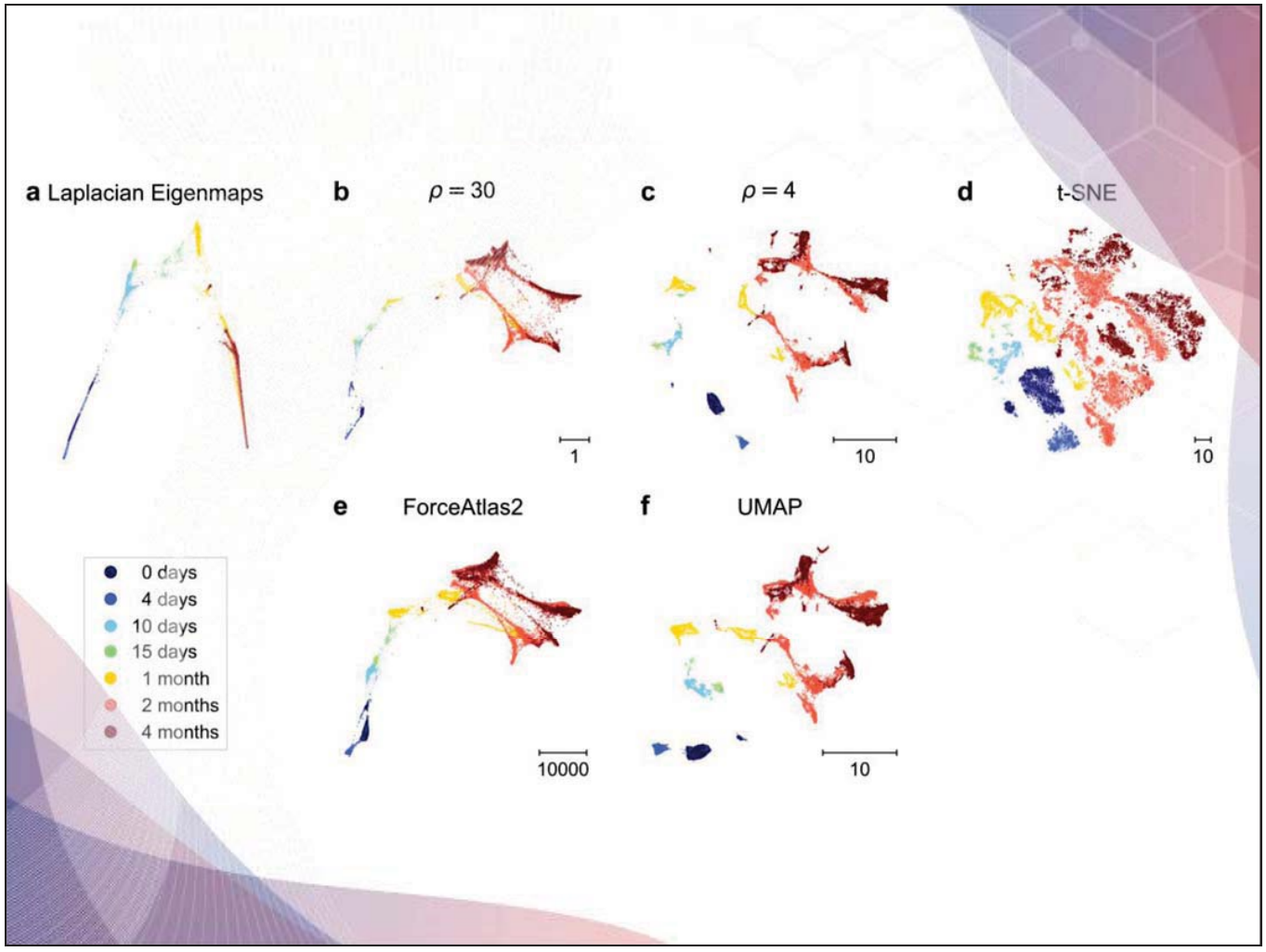
박종은(2018). 단일 세포 RNA 시퀀싱(Single-cell RNA sequencing) 기술 동향. BRIC View 2018-T28.

67

Many methods for graph embedding



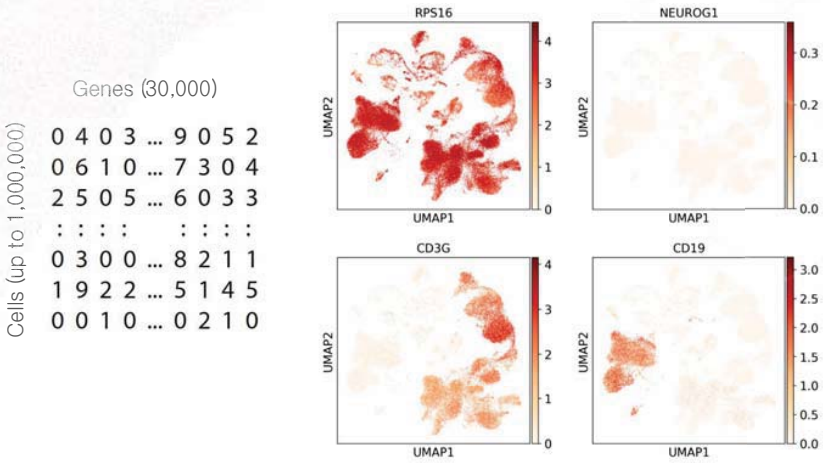

a Laplacian Eigenmaps
b ρ = 30
c ρ = 2
d t-SNE
e ForceAtlas2
f UMAP
step 1
...
step 10
...
step 20
1
10
10
10000
10


a Laplacian Eigenmaps
b ρ = 30
c ρ = 4
d t-SNE
e ForceAtlas2
f UMAP
0 days
4 days
10 days
15 days
1 month
2 months
4 months
1
10
10
10000
10

How to annotate cells?
How to define cluster resolution?

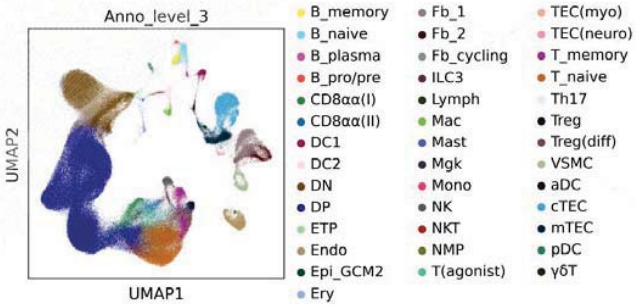
2-4. Clustering, finding marker genes and cell type annotation

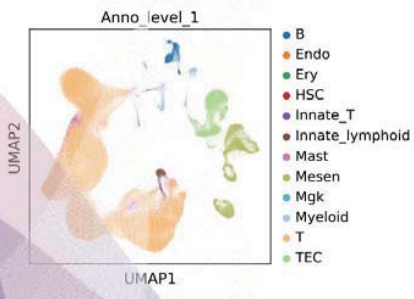
Choice of highly variable genes / PCA projection / batch correction methods… all affect clustering outcome

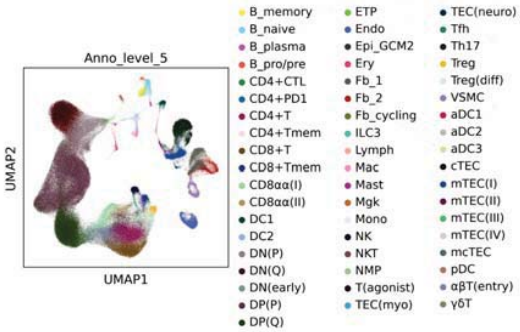


highly variable gene selection

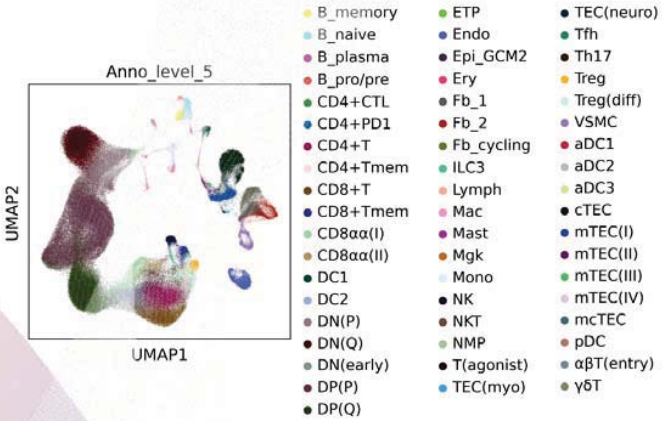
Hierarchy in cell annotation

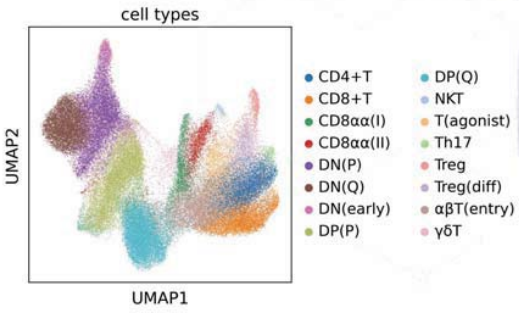






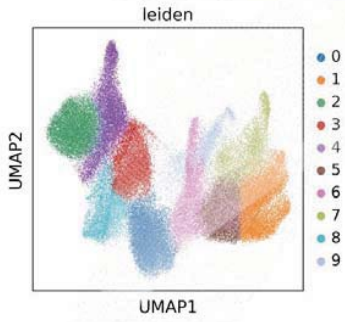
Zooming into subset to get better resolution



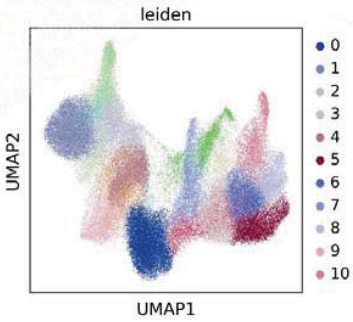


Subset, balancing, different genes, changed PCs

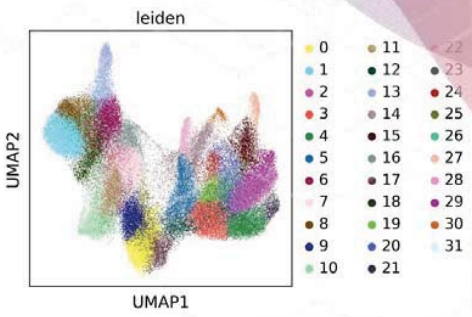
Resolution = 1

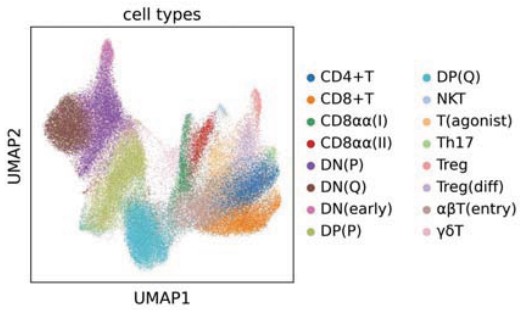


Resolution = 2

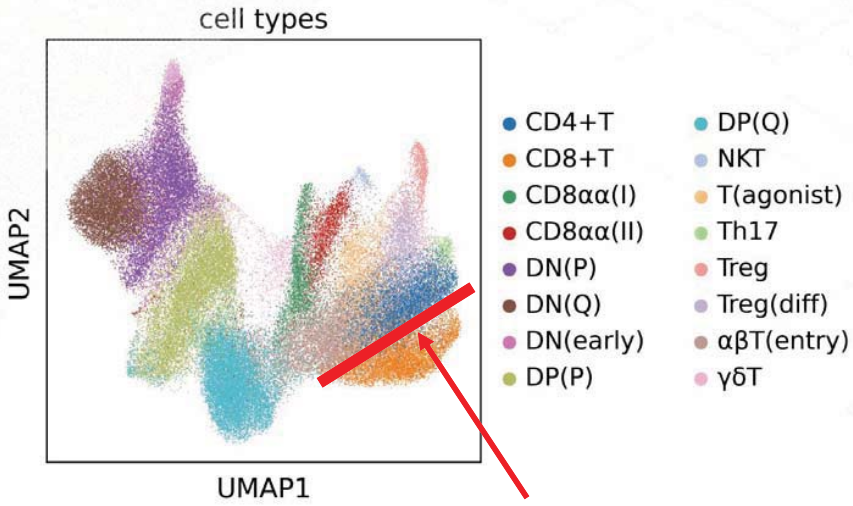


Resolution = 3





Defining markers



Focusing on the borders
Any binary classifier?

DEG analysis?

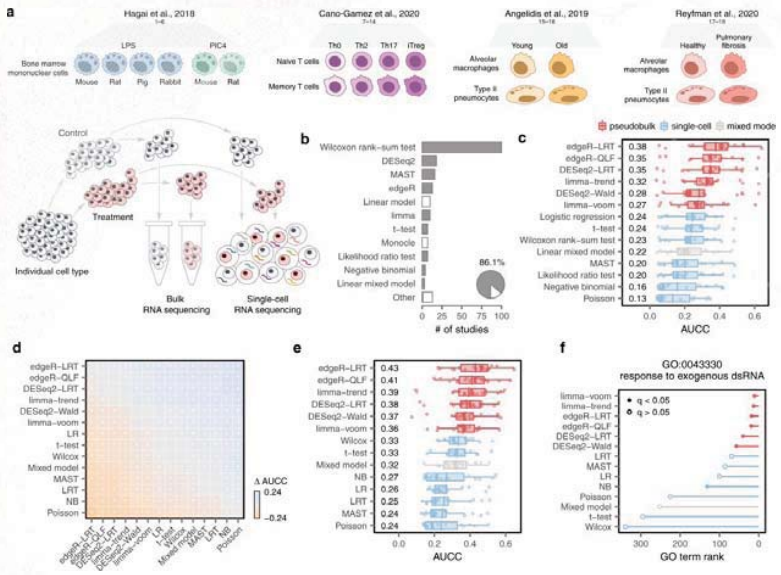
Article | Open Access | Published: 28 September 2021

Confronting false discoveries in single-cell differential expression

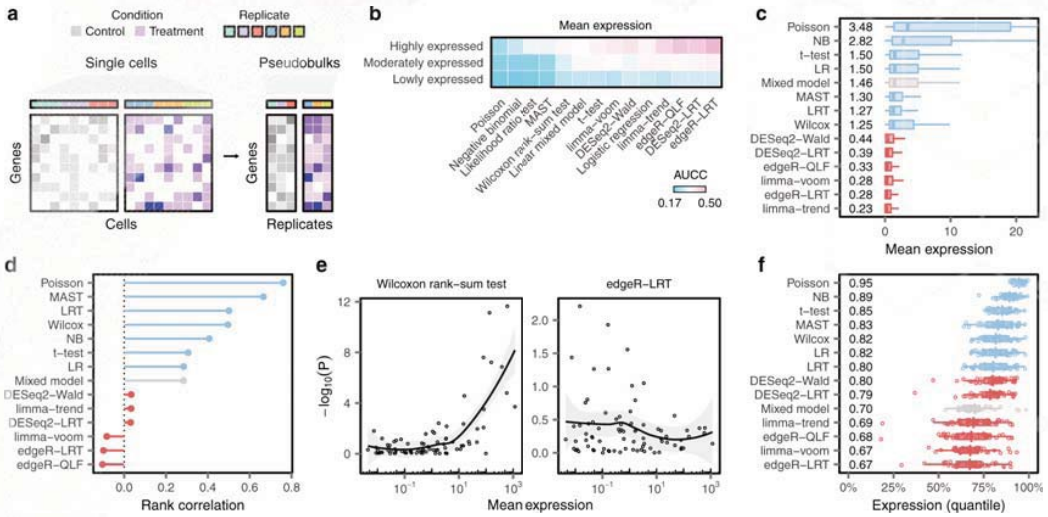
Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider ✉ & Grégoire Courtine ✉

Nature Communications **12**, Article number: 5692 (2021) | Cite this article

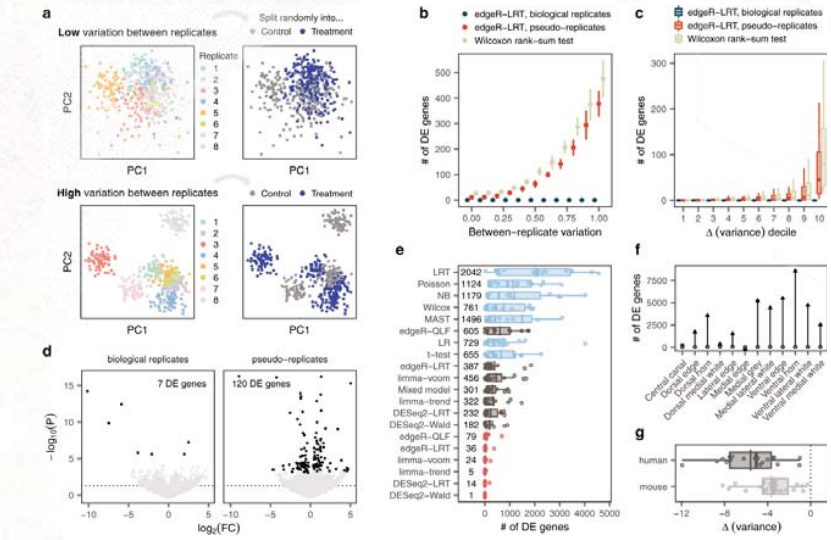
Pseudo-bulk methods outperform generic and specialized single-cell DE methods



Single−cell DE methods are biased towards highly expressed genes.

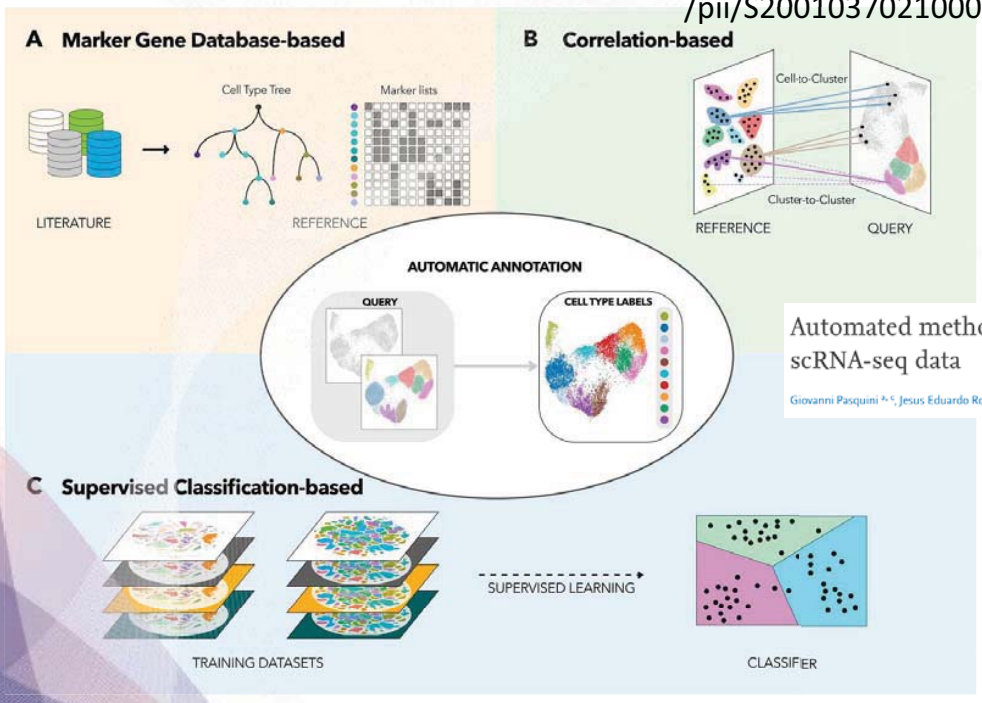


Single−cell DE methods often leads to false discoveries of DEGs



Automatic cell type annotation

https://www.sciencedirect.com/science/article/pii/S2001037021000192#b0120

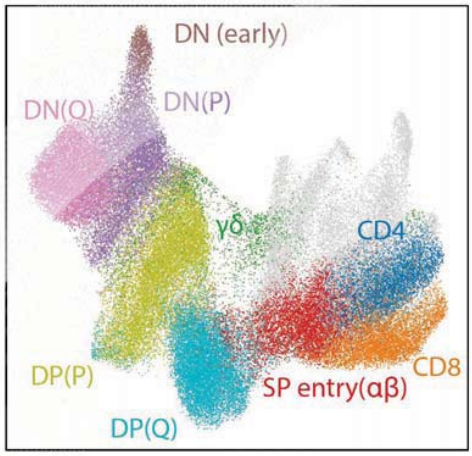


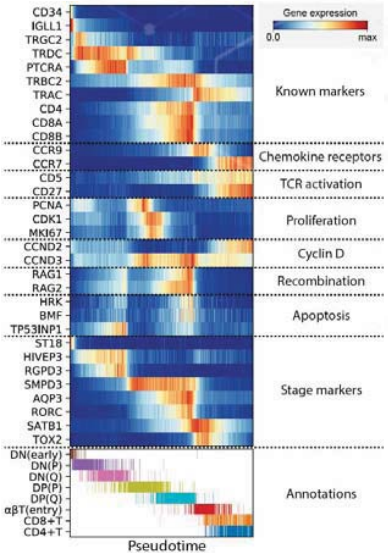
Automated methods for cell type annotation on scRNA-seq data

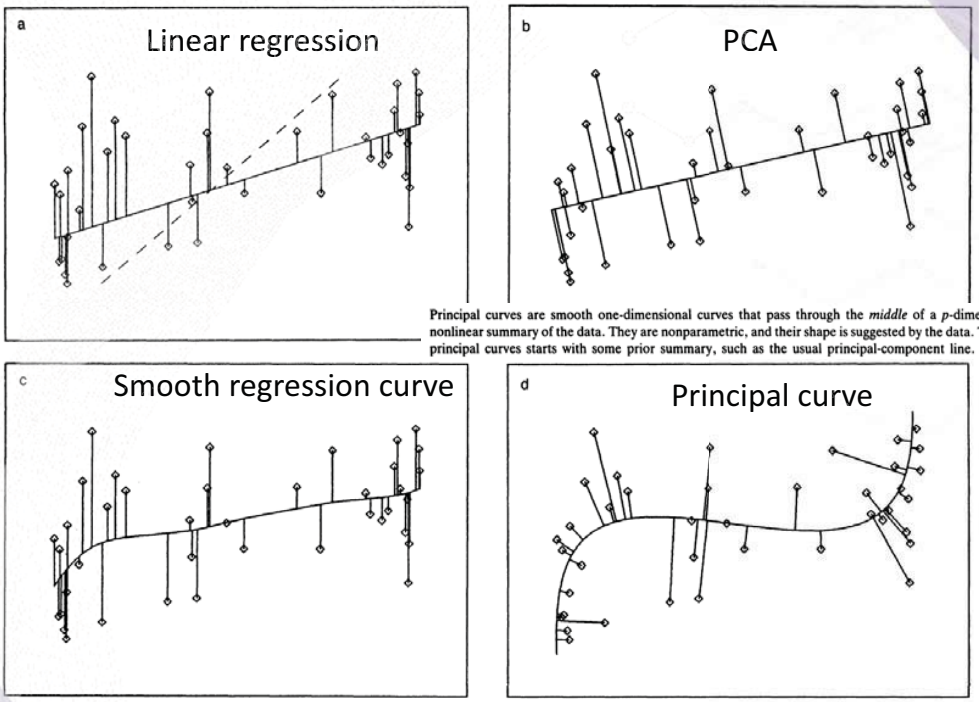
Giovanni Pasquini [a, c], Jesus Eduardo Rojo Arias [b], Patrick Schäfer [a], Volker Busskamp [a, c]

2-5. Trajectory inference

Modelling the cell differentiation trajectory



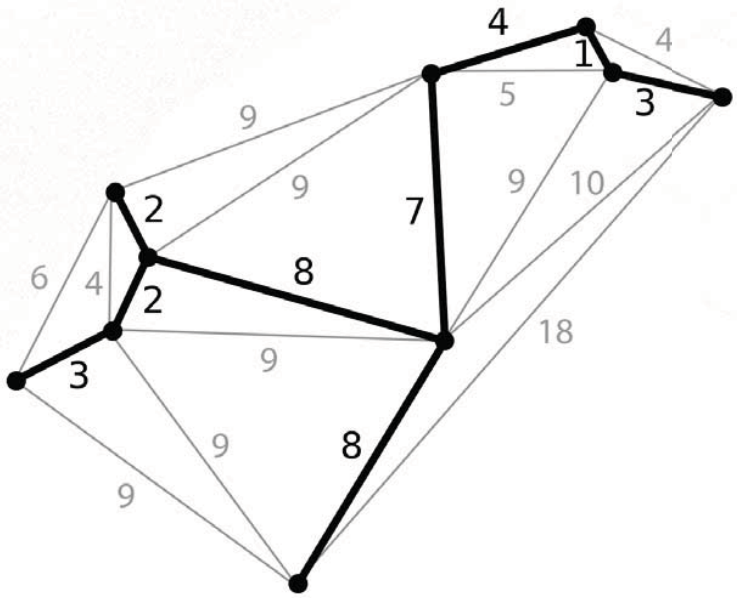




Principal curves are smooth one-dimensional curves that pass through the *middle* of a *p*-dimensional data set, providing a nonlinear summary of the data. They are nonparametric, and their shape is suggested by the data. The algorithm for constructing principal curves starts with some prior summary, such as the usual principal-component line. The curve in each successive

Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.

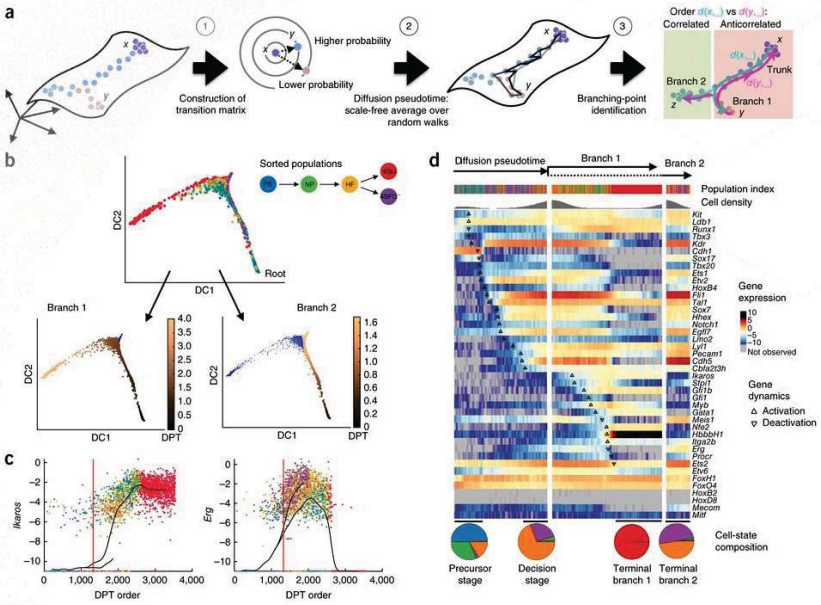
Minimum Spanning Tree



Published: 29 August 2016

Diffusion pseudotime robustly reconstructs lineage branching

Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner & Fabian J Theis



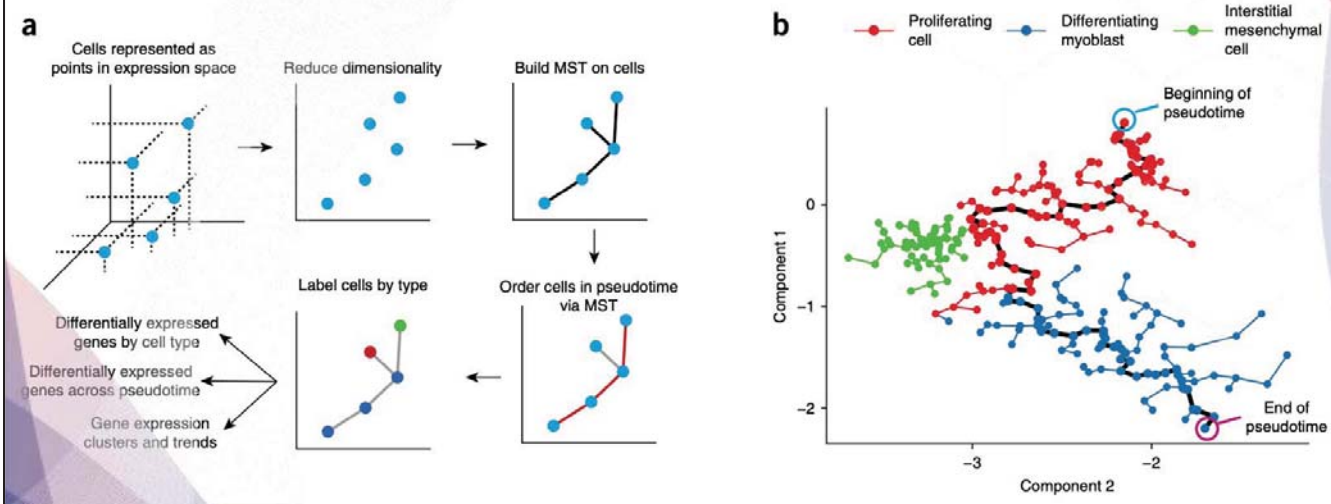
Monocle

nature biotechnology

LETTERS

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

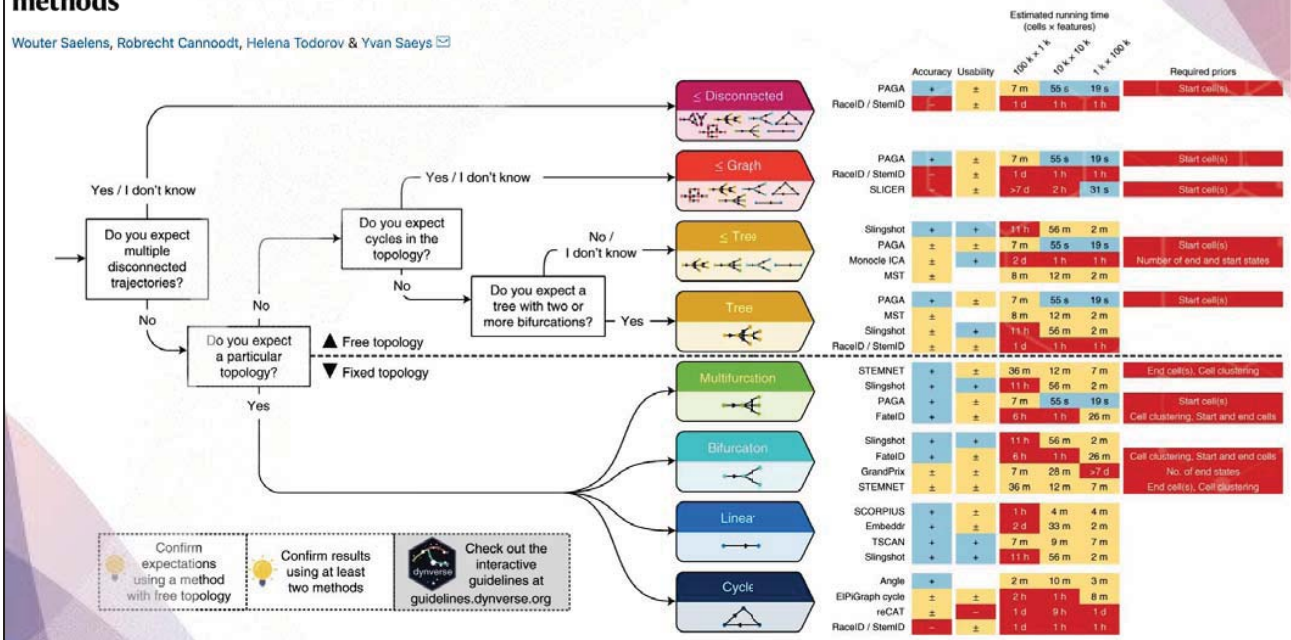
Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[2], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]



Article | Published: 01 April 2019

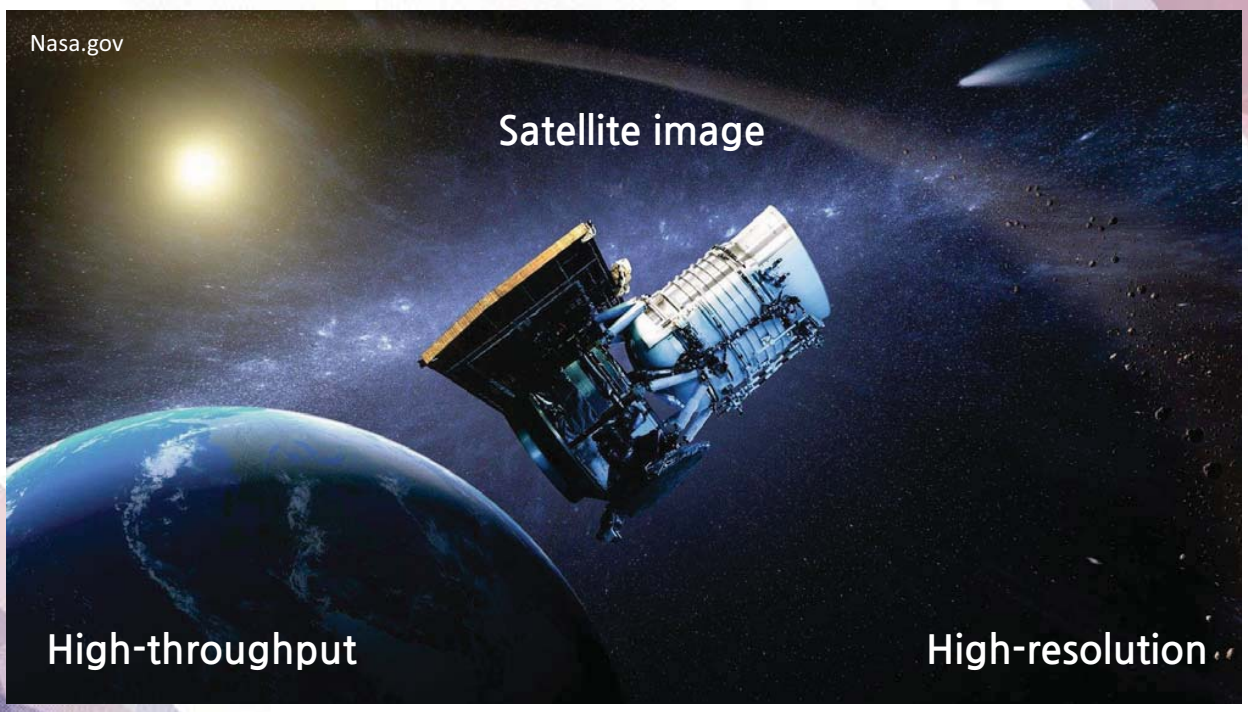
A comparison of single-cell trajectory inference methods

Wouter Saelens, Robrecht Cannoodt, Helena Todorov & Yvan Saeys



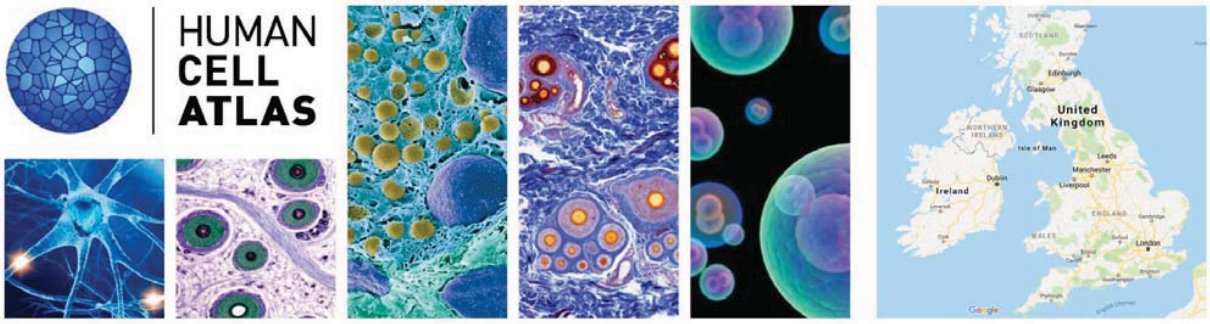
3. Public databases & Data integration

Analogy between singel-cell omics vs satellite images



Human cell atlas

Human cell atlas : “Google map” of human body



Cells & Genes

City & buildings

93

Human cell atlas: timeline and scope

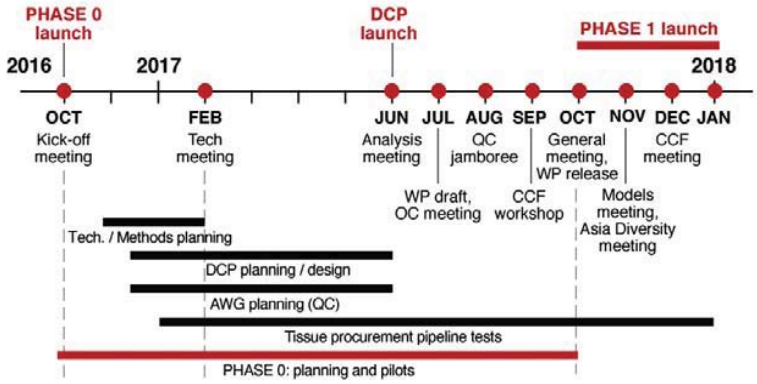
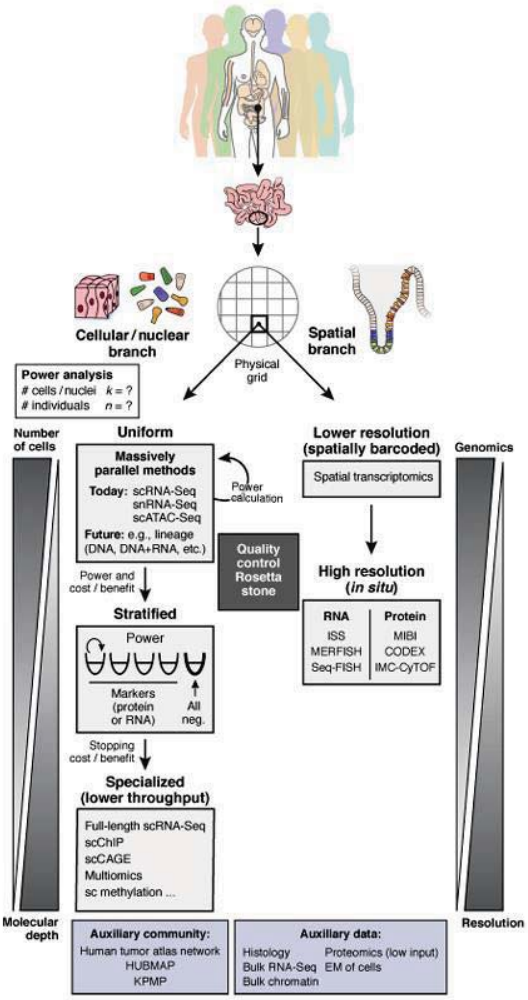
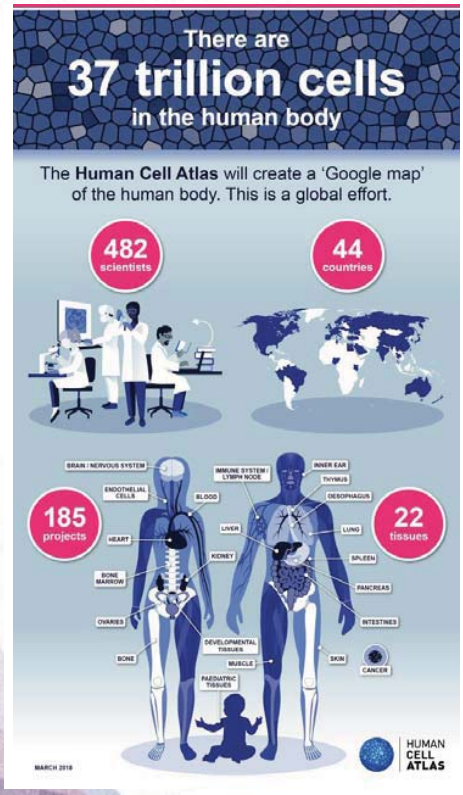
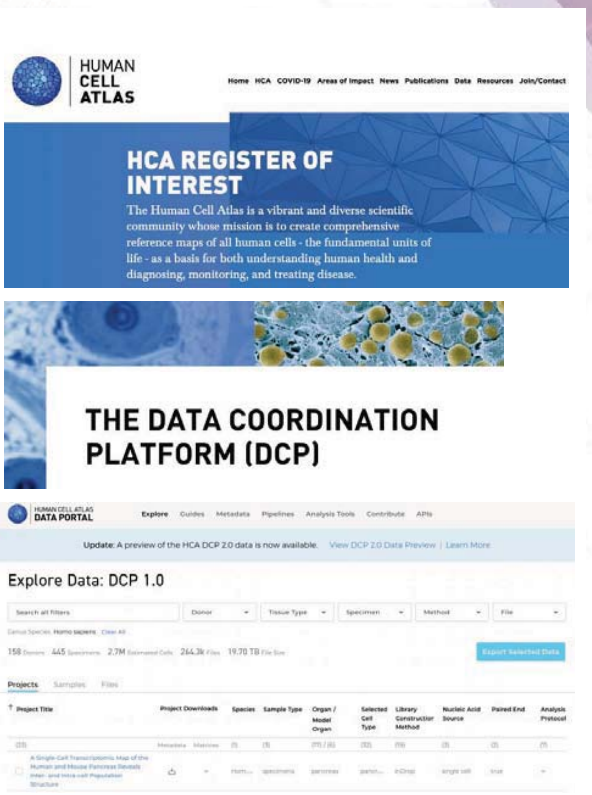


Figure 2. Timeline of HCA activities, October 2016 through January 2018.

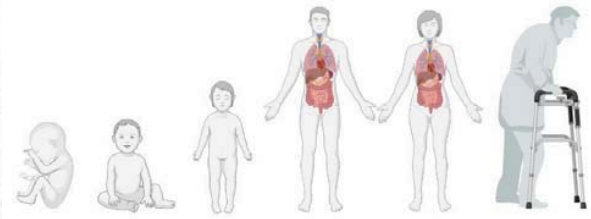


Human cell atlas: data coordination platform





Divide and conquer strategy



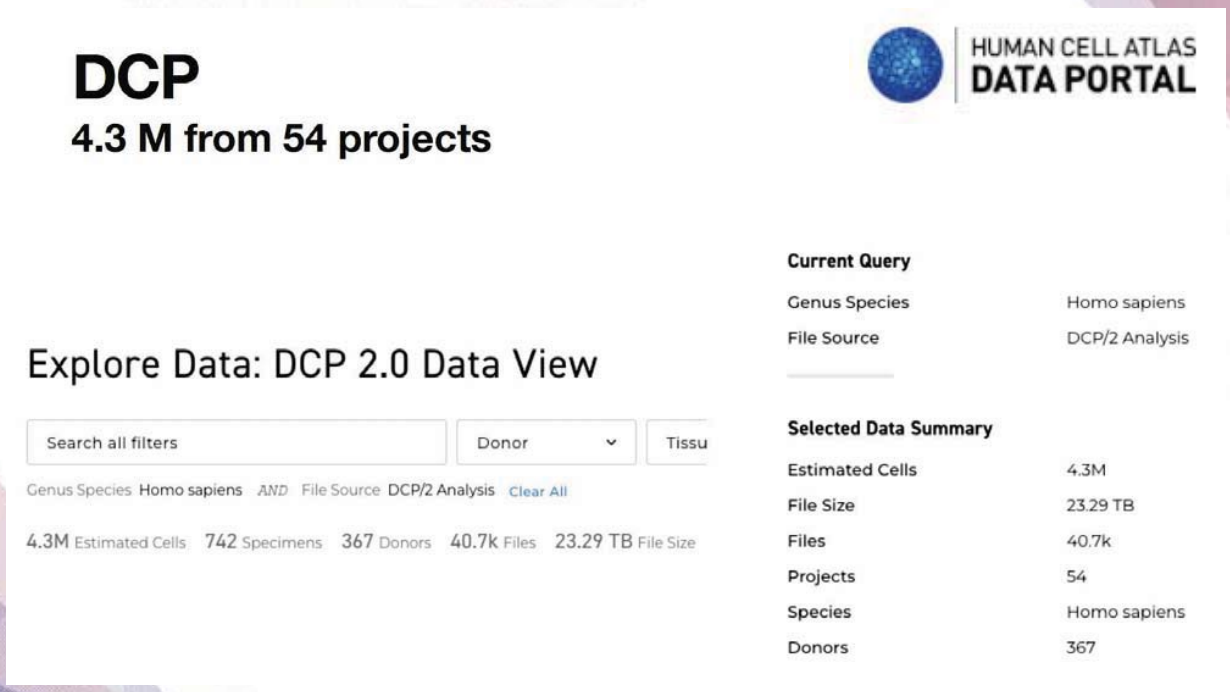
164 datasets in DCP

Organs | Developmental stages | Disease status

Data integration

HUMAN CELL ATLAS

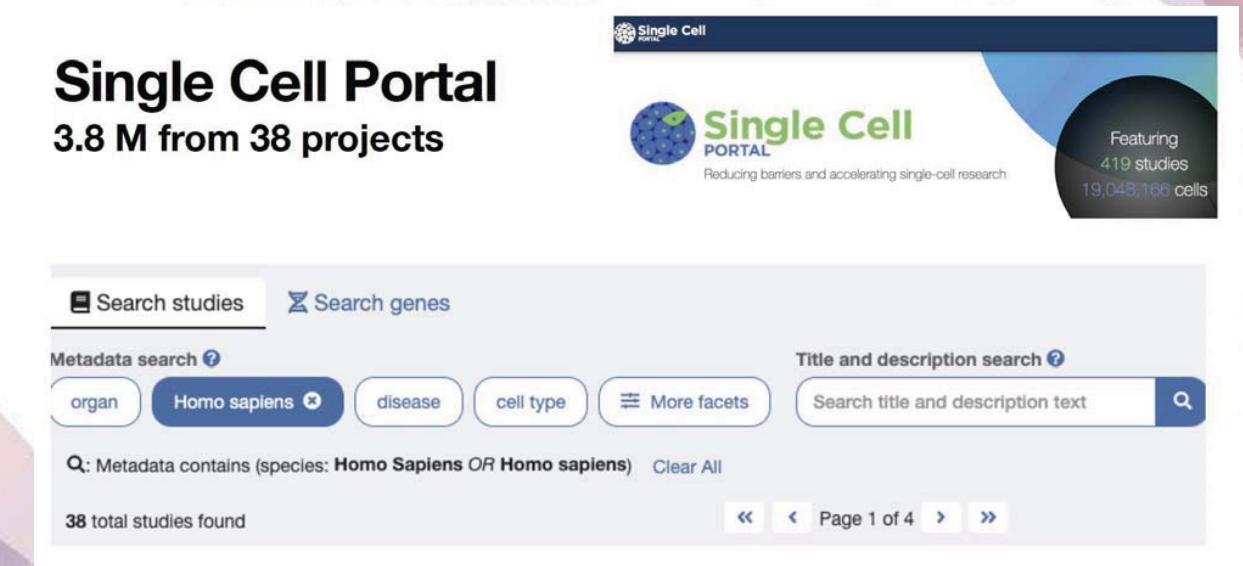
Example of single-cell atlas database



https://data.humancellatlas.org/explore/projects

97

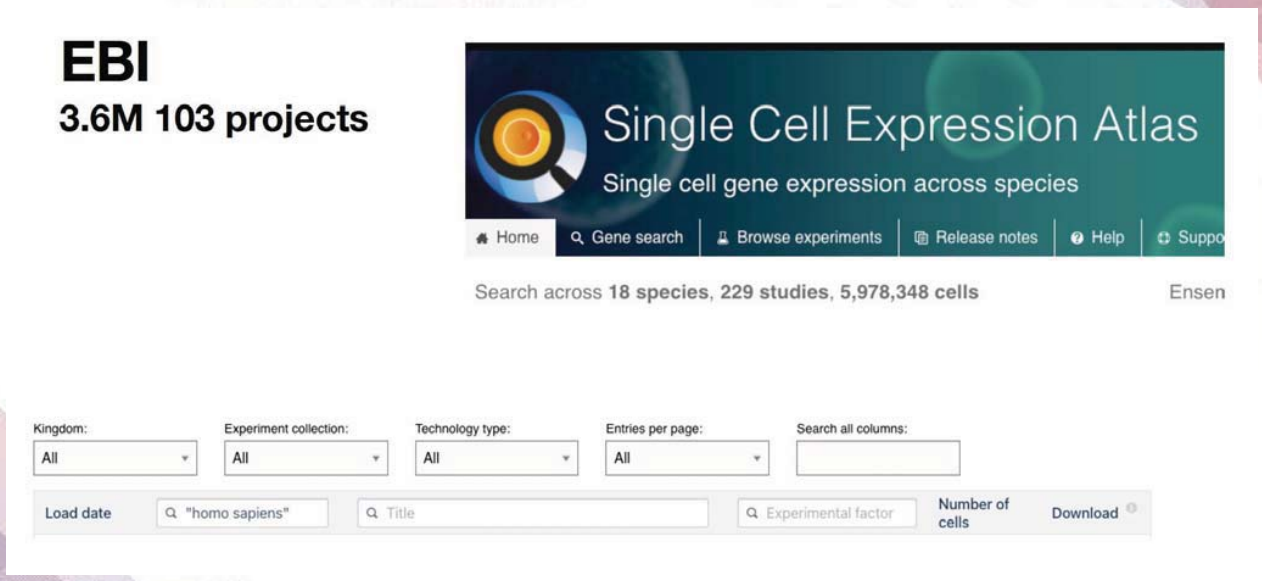
Example of single-cell atlas database



https://singlecell.broadinstitute.org/single_cell

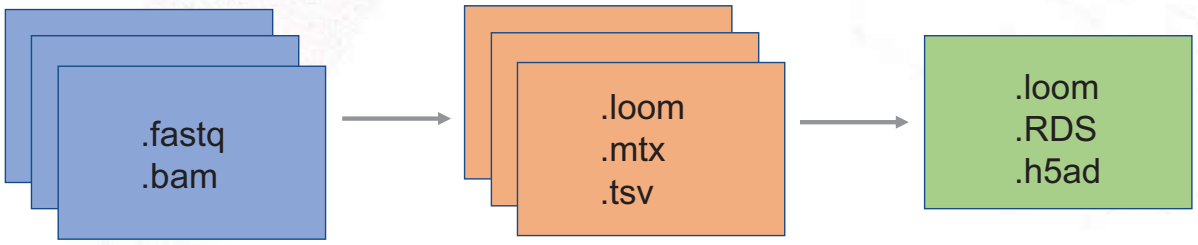
98

Example of single-cell atlas database



https://www.ebi.ac.uk/gxa/sc/home

Structure of single-cell data files

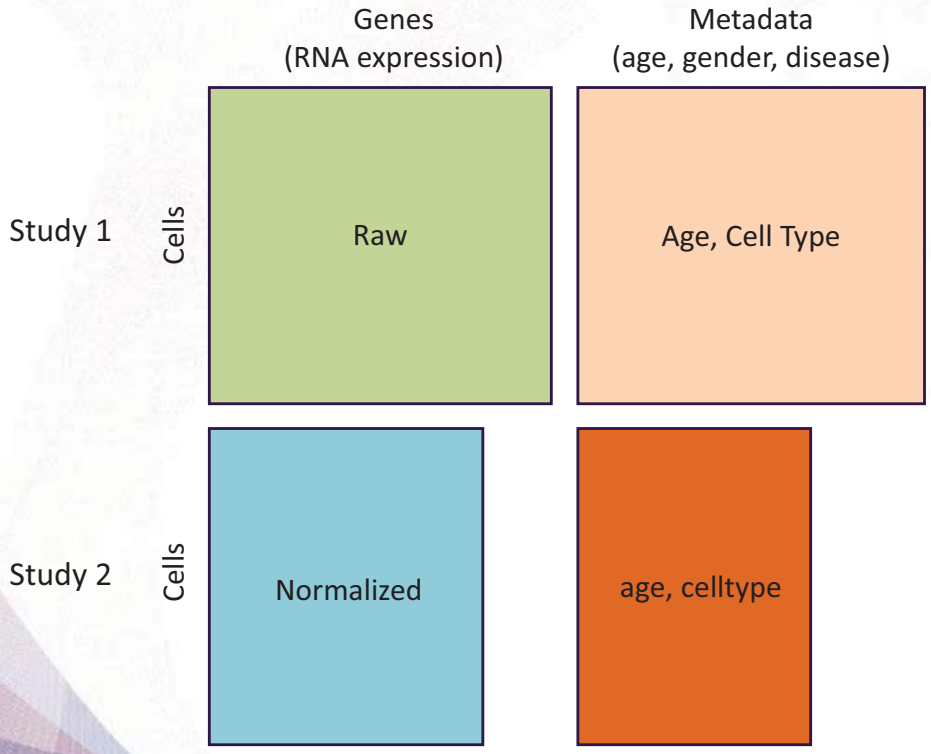


Sequence files | Count matrix | Annotated matrix

Takes long time to download/process
Difficult to match with metadata

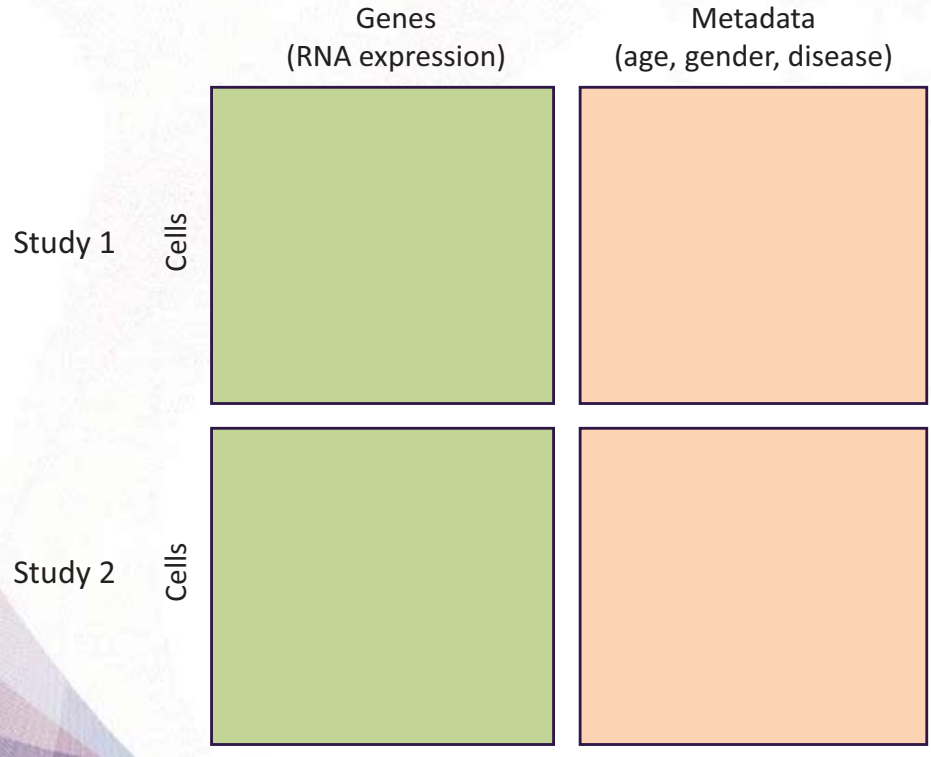
Easy to download
Matched metadata

Potential problems in utilizing annotated matrix



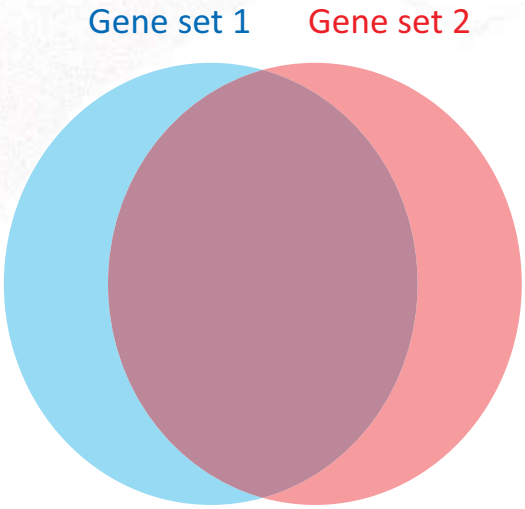
101

Desired outcome for utilization of annotated matrix files

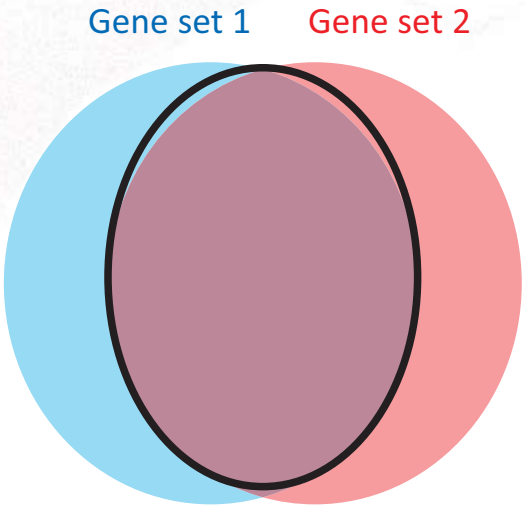


102

Harmonizing gene sets



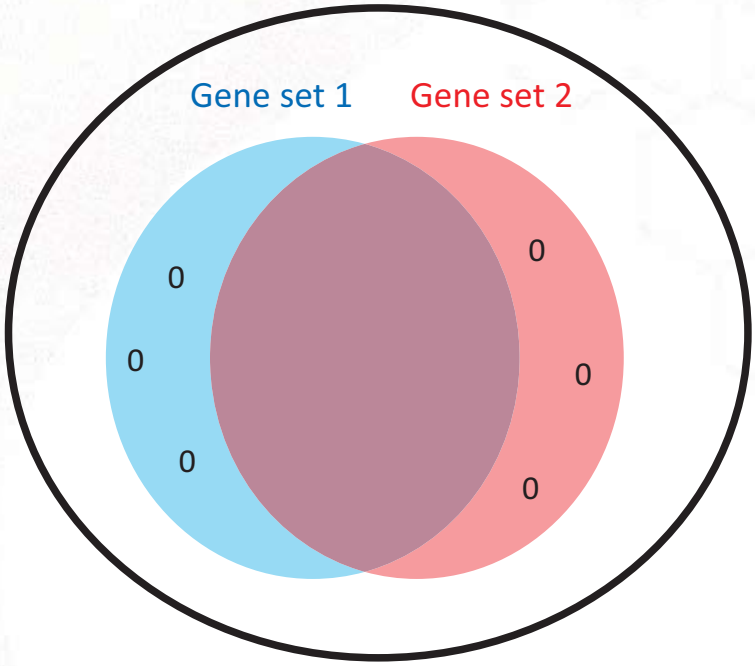
Harmonizing gene sets



Approach 1. Intersection

Efficient/accurate batch integration
Downside: Loosing sample specific marker genes

Harmonizing gene sets

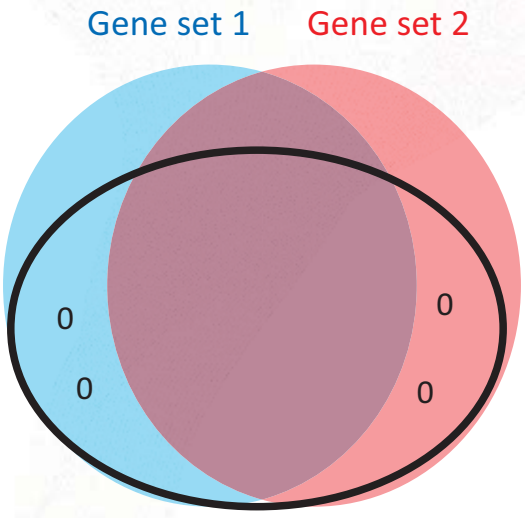


Approach 2. Union

Downside: Increased ”batch effect”

105

Harmonizing gene sets



Approach 3. Filter

Define universal gene annotation
Apply gene mapper (converting)
Assess the quality of mapping
Impute non-existing genes as 0

106

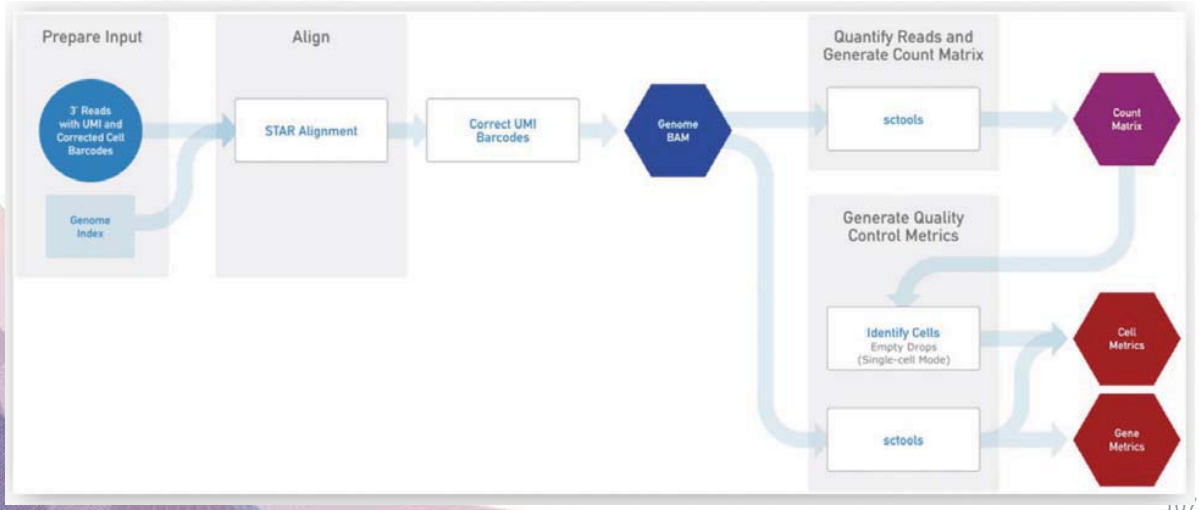
Remapping (sequence alignment using uniform pipeline)

Introduction to the Optimus Workflow

The long-term goal of the Optimus workflow is to support any 3 prime single-cell or single- nucleus transcriptomics assay selected by the HCA project. Using the correct modularity, we hope to grow a generic pipeline that has specific modules to address differences in assays, while leveraging common code where steps of the assays are the same. We offer this as a community resource for community development and improvement.

HUMAN CELL ATLAS
DATA PORTAL

The workflow supports the 10x v2 and v3 gene expression assay and has been validated for analyzing single-cell and single-nucleus from both human and mouse data sets.

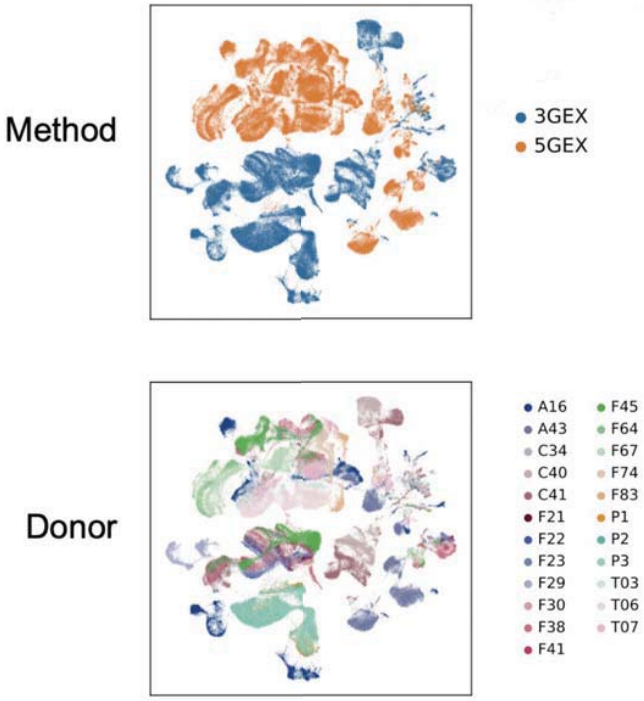


107

3-2. Single-cell RNA-seq data integration (batch correction)

108

Problem of batch effect in single-cell data

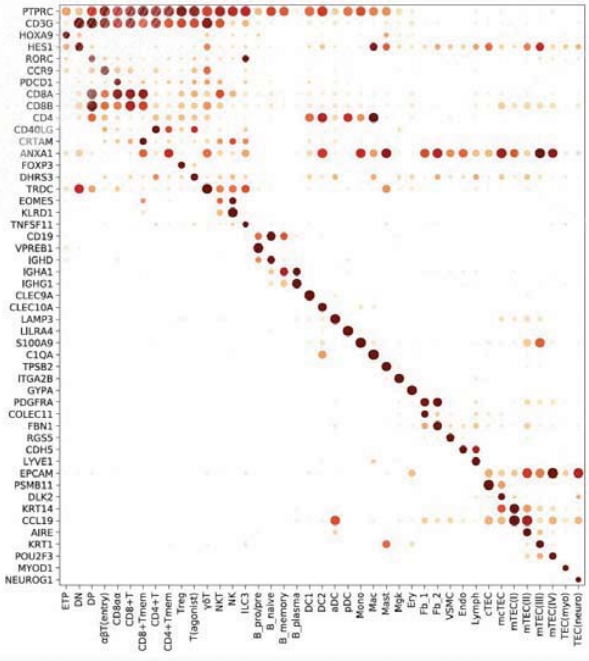


Park et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science*.

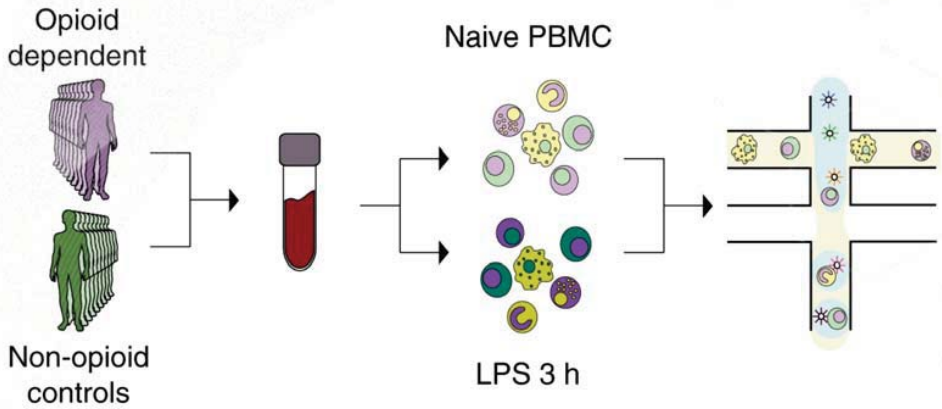
109

Source of variation for gene expression?

Source of variation (1) Biology – Cell types

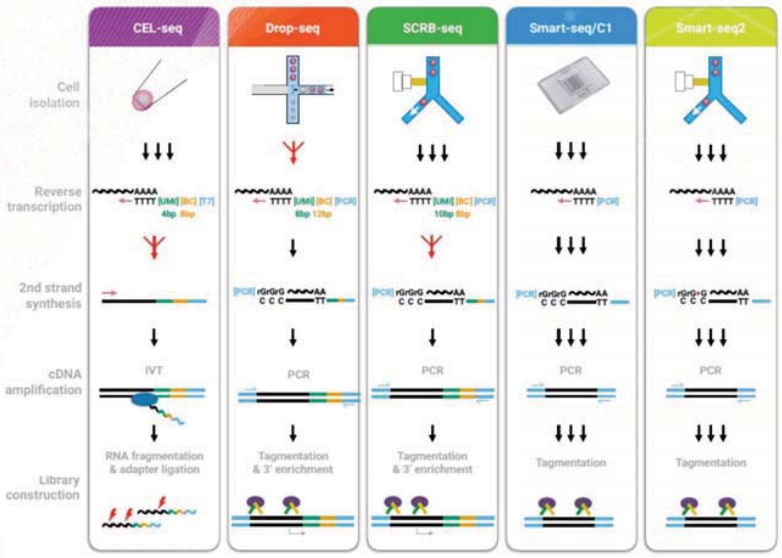


Source of variation (2) Experimental conditions



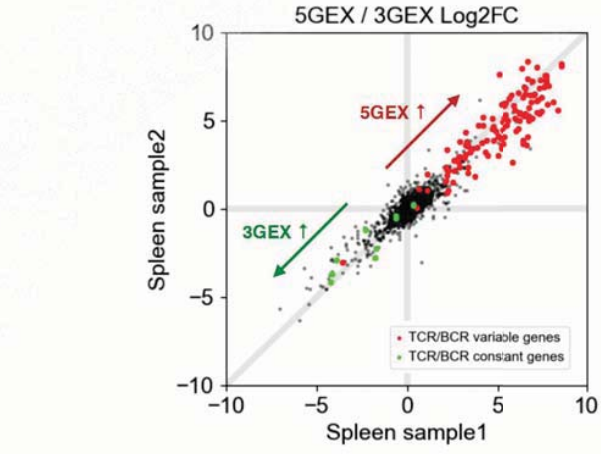
Source of variation (3) Replicates

Source of variation (3) Technology

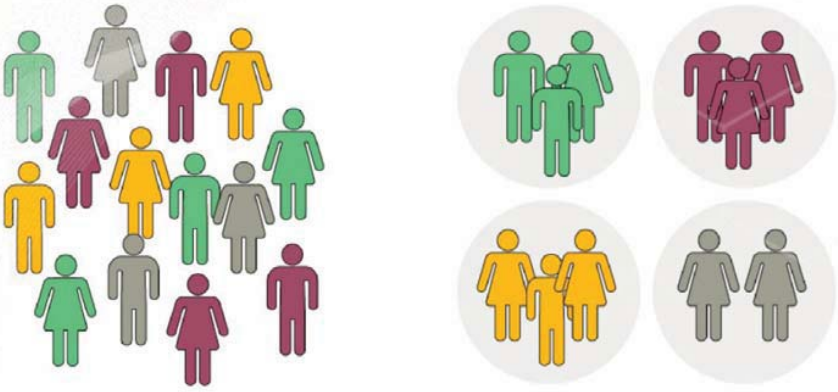


https://www.biorxiv.org/content/10.1101/035758v3.full#disqus_thread

Source of variation (3) Technology



Source of variation (4) Individuals (genotypes)



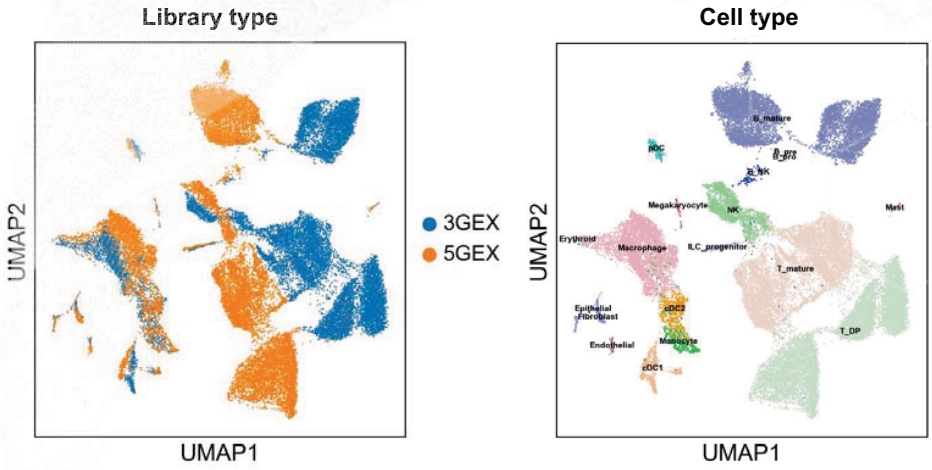
Genders (XIST) or HLA genes…

Expression of gene X

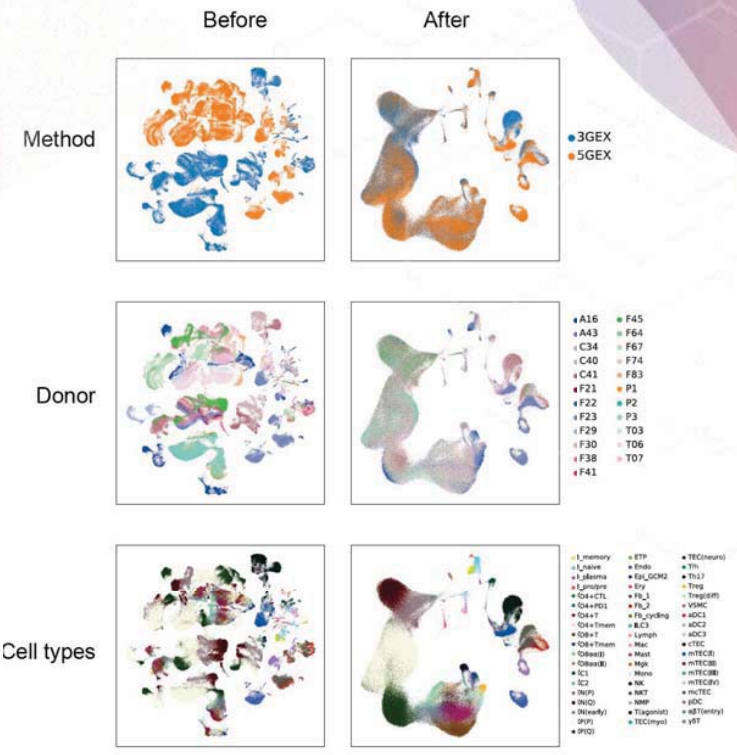
Cell type + Experimental conditions +	Keep!
Replicate + Technology + Genotype + Bioinformatic pipeline	Remove!

Batch effect!

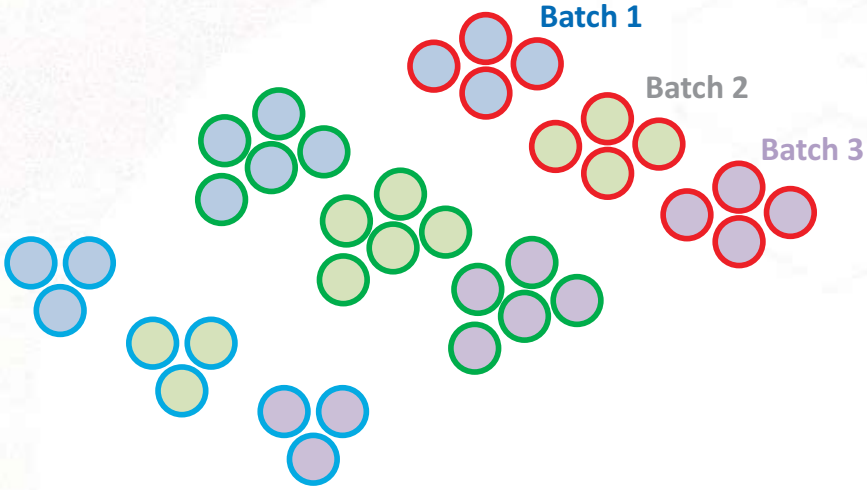
Example of batch effect



Correcting for the batch effect



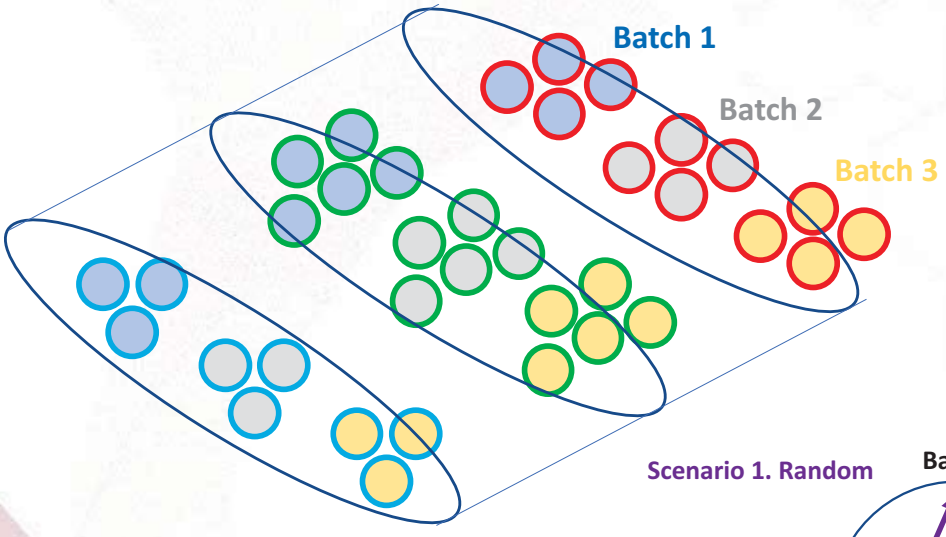
Visualizing batch effect in single-cell data



gene expression space…

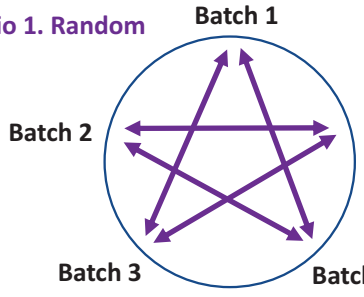
119

Structure of batch effect



gene expression space…

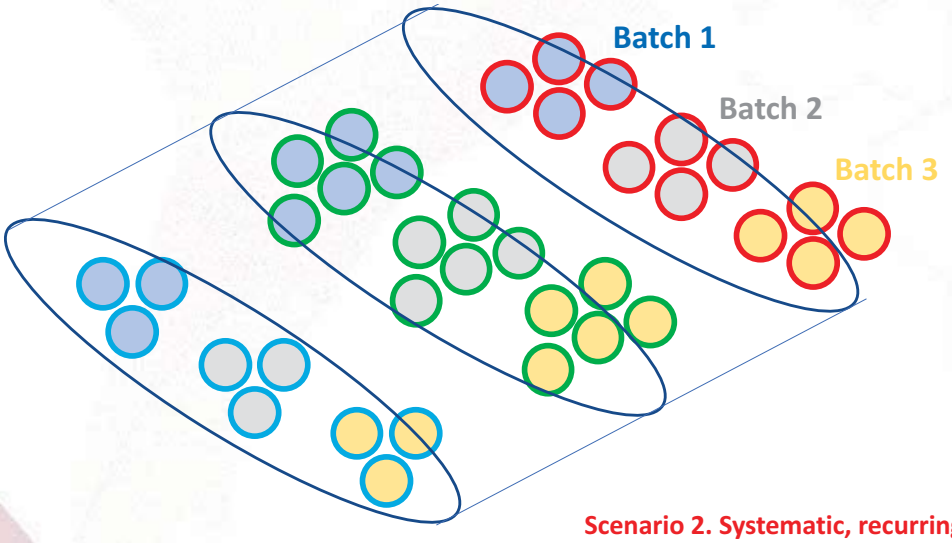
Scenario 1. Random



Genotype,
Exp. Batch…

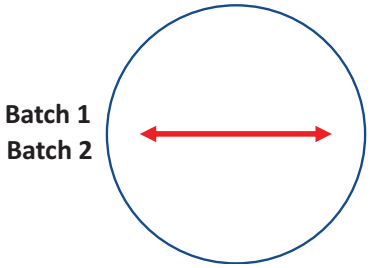
120

Structure of batch effect



gene expression space…

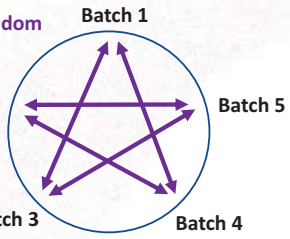
Scenario 2. Systematic, recurring



10X 3' vs 5'
Development
Disease

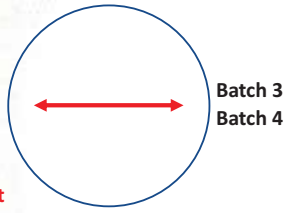
Structure of batch effect

Scenario 1. Random

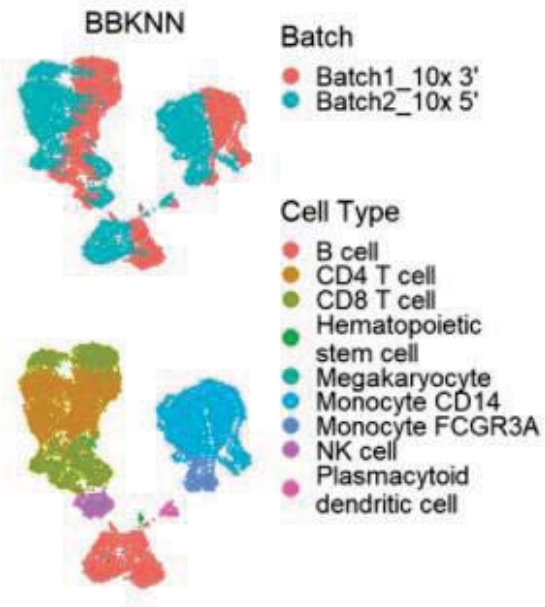


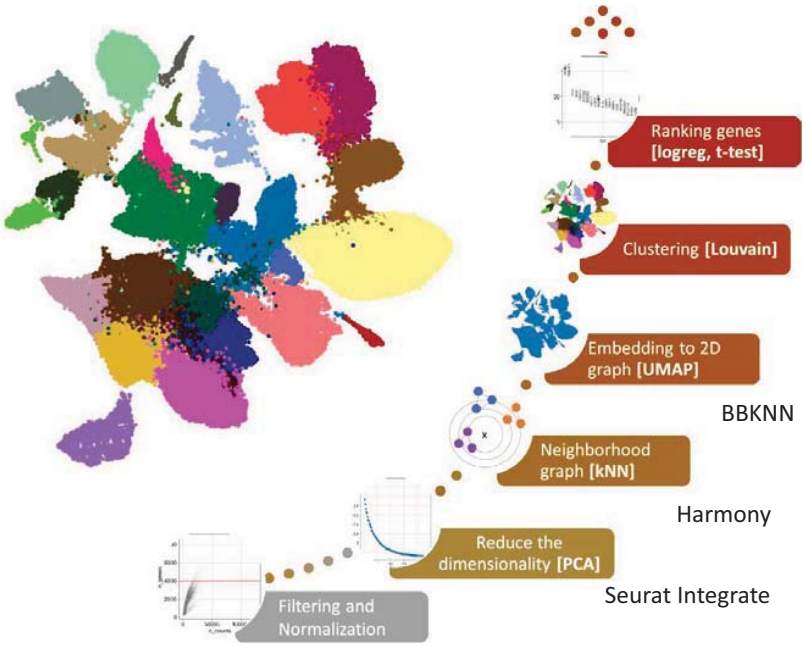
Genotype,
Exp. Batch…

Scenario 2. Systematic, recurring

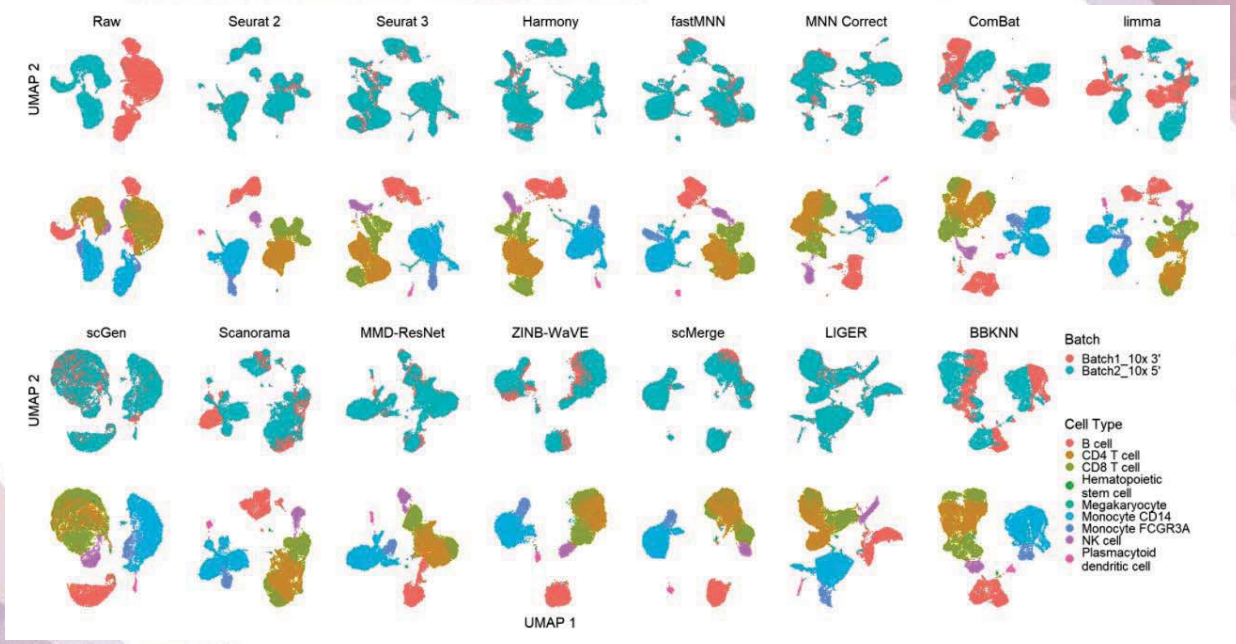


10X 3' vs 5'
Development
Disease





Different methods, different results

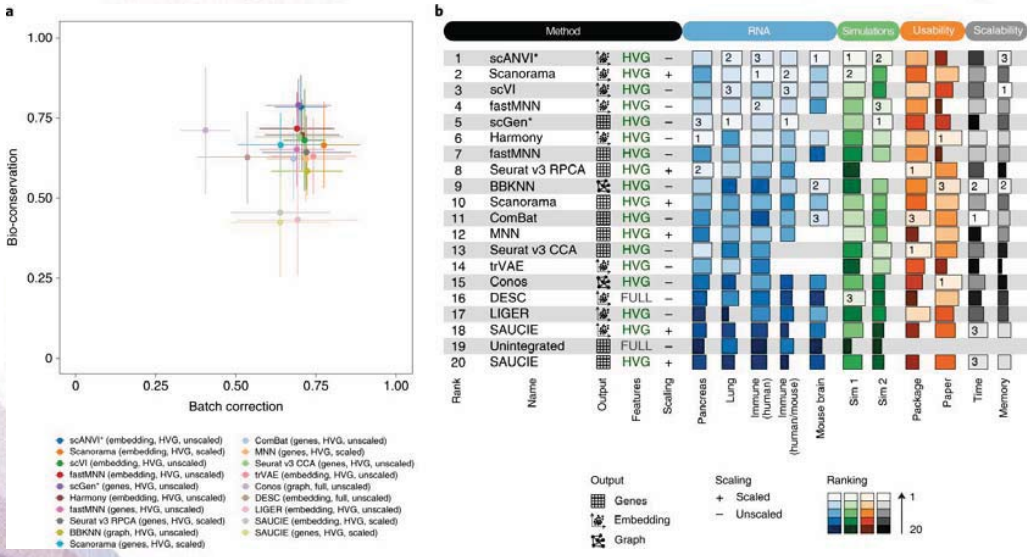


Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21, 12

124

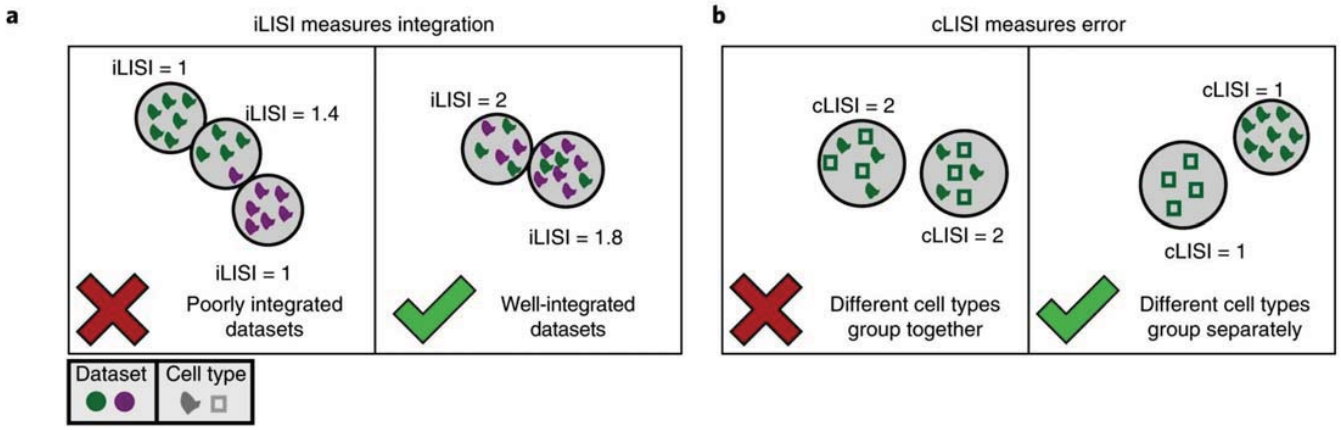
Goals for ideal batch correction

- **Batch removal -> Good harmonization across batches**
- **Bio conservation -> Maintaining biological integrity (no distortion or over-correction)**
- **Scalability -> Can deal with large scale datasets**



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat*

Metrics to assess the quality of batch correction

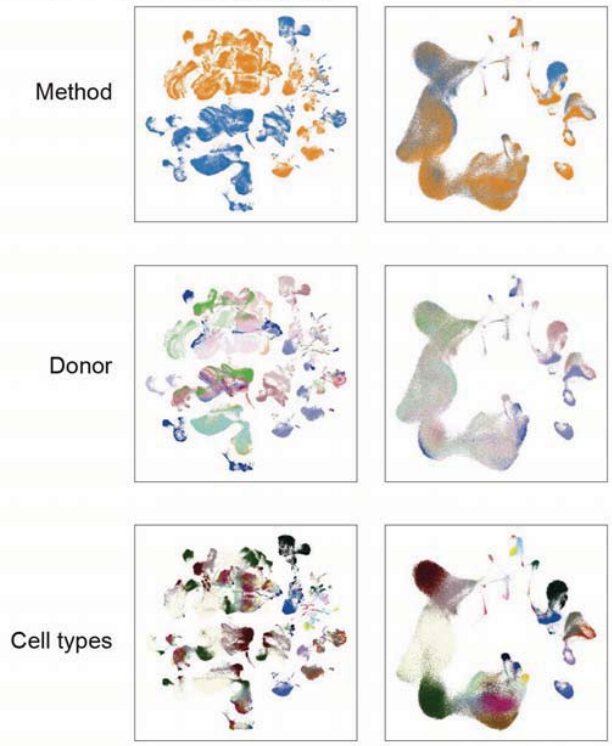


Measuring integration

Measuring cell type preservation

Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat*

Ideal batch correction example



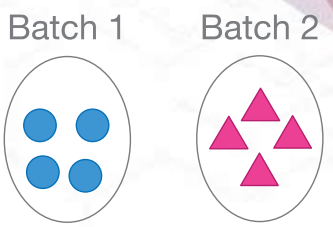
Park et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science*.

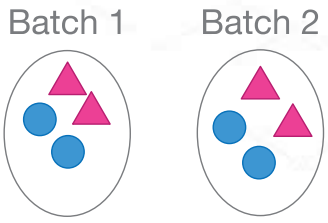
127

Linear regression

Y ~ Tech + Donor + Gender + residual

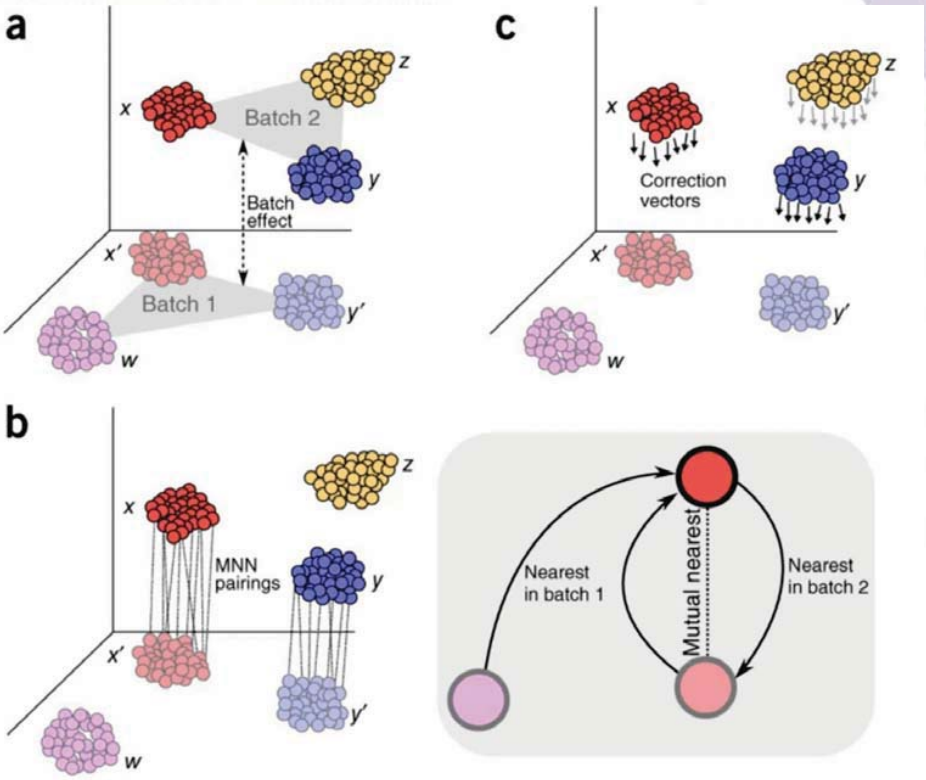
- Regress out unwanted variations
  - Limma, ComBat
- Assumption: each batch contains similar cell composition
  - Risk of over-correction





128

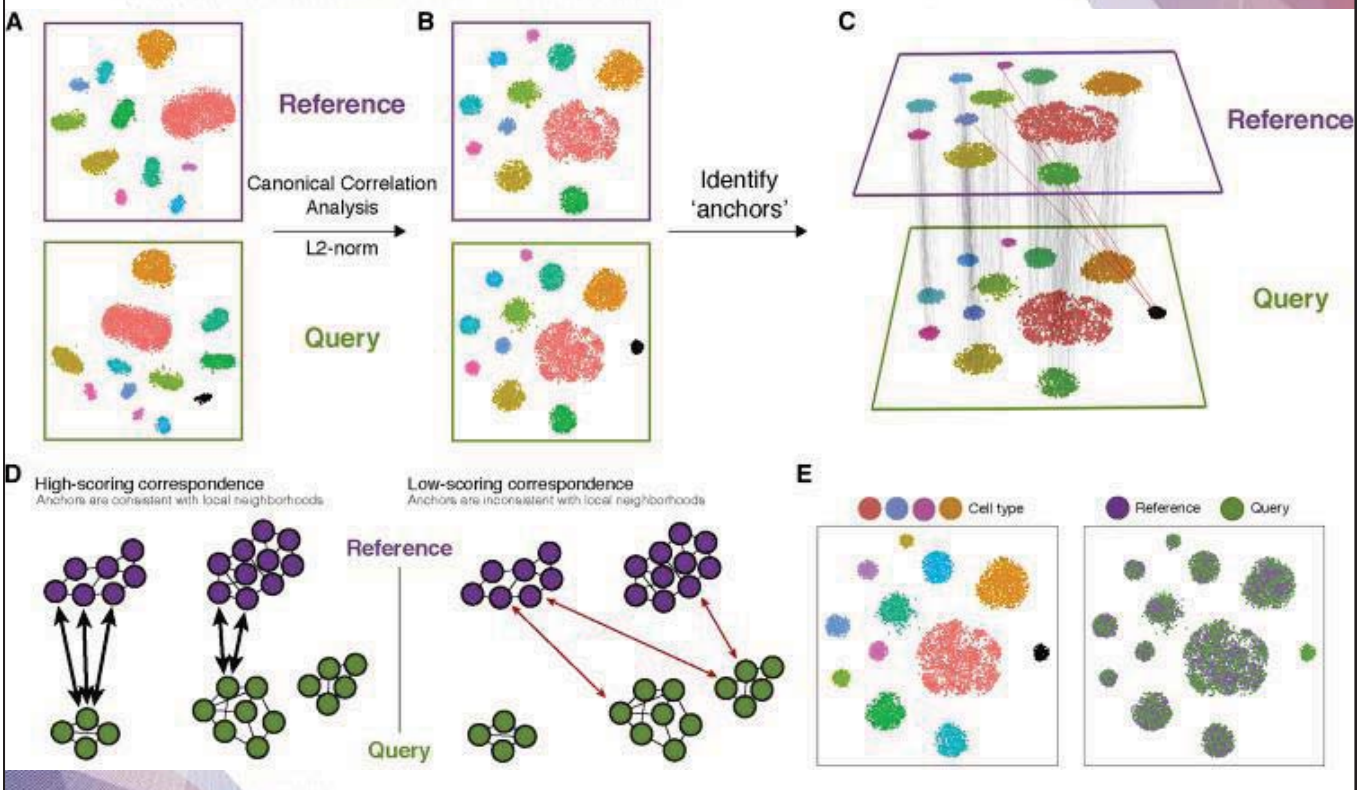
Mutual nearest neighbors



Haghverdi, L., Lun, A., Morgan, M. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat*

129

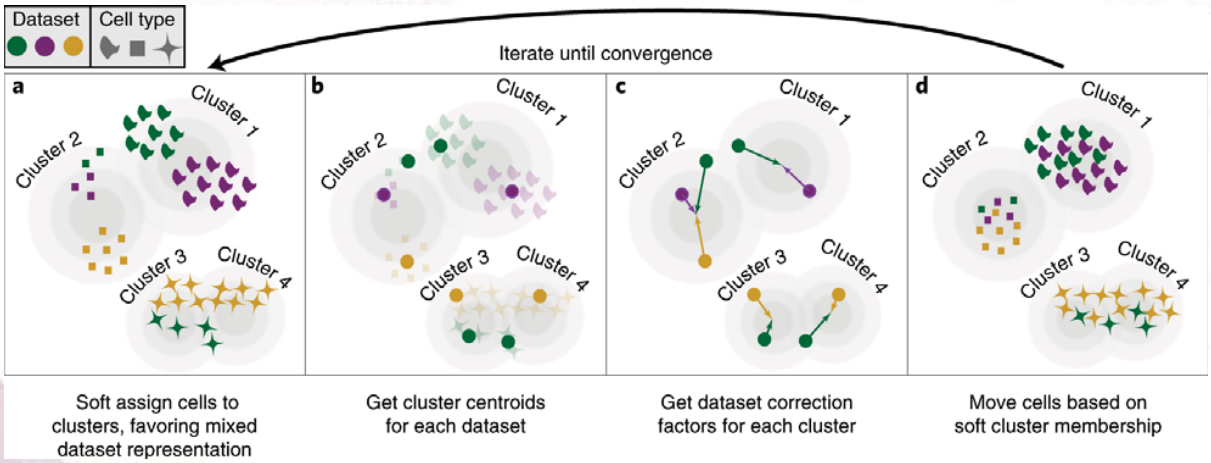
Finding anchors (with CCA dimension reduction)



Stuart, Tim, et al. "Comprehensive integration of single-cell data." *Cell* 177.7 (2019): 1888-1902.

130

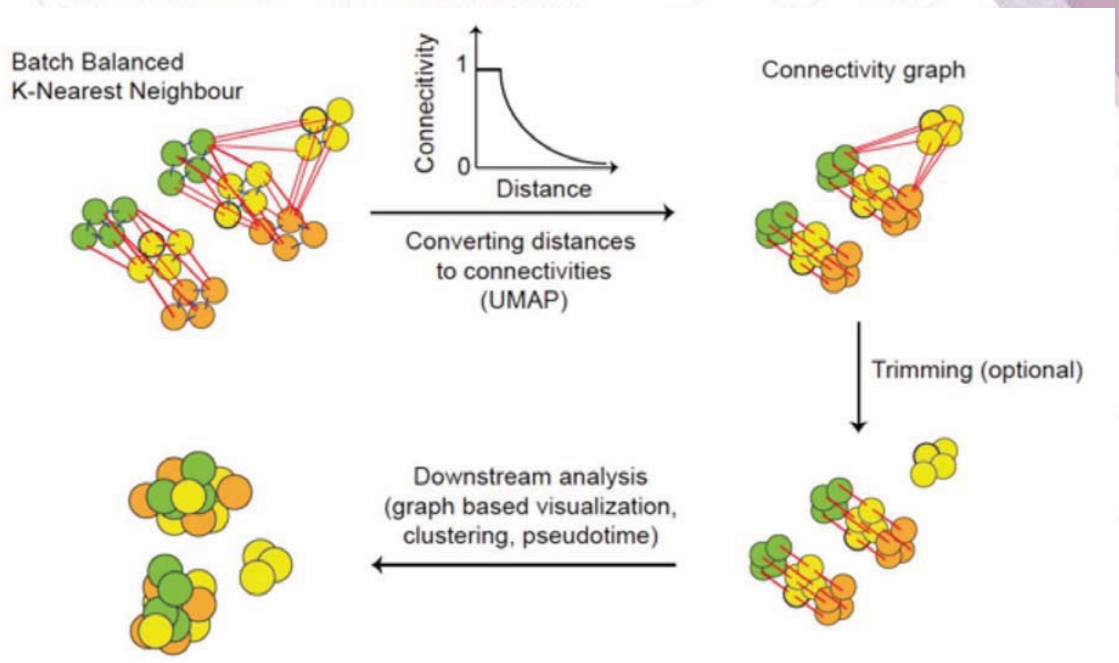
Harmony: batch correction at cluster level



Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16, 1289–

131

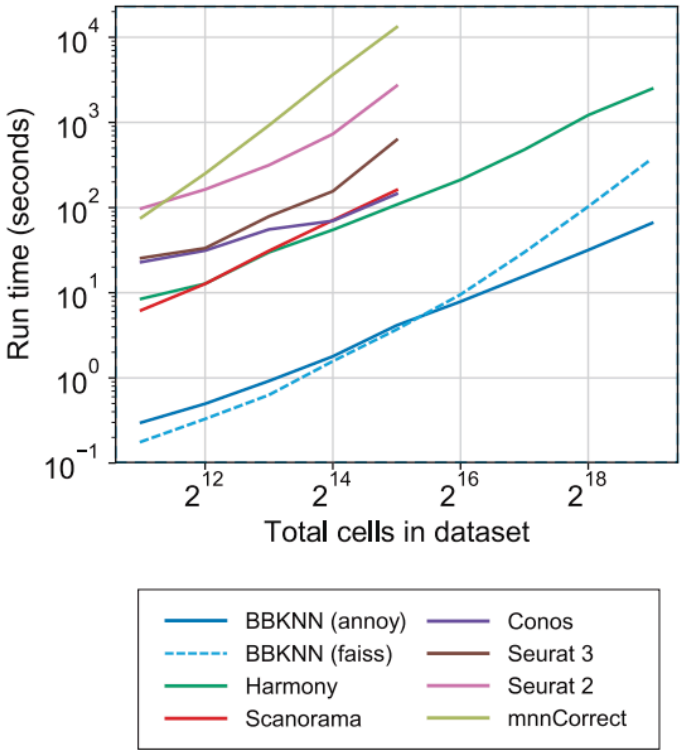
BBKNN: Biology correction at graph level



Polański, Krzysztof, et al. "BBKNN: fast batch alignment of single cell transcriptomes." *Bioinformatics* 36.3 (2020): 964-965.

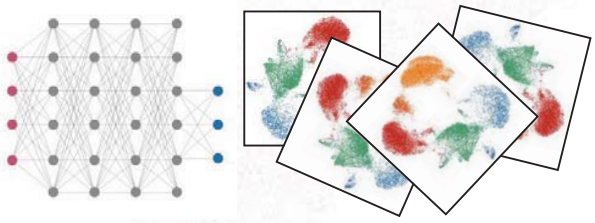
132

Importance of efficiency and speed for large data integration



Polański, Krzysztof, et al. "BBKNN: fast batch alignment of single cell transcriptomes." *Bioinformatics* 36.3 (2020): 964-965.

Tandem batch correction



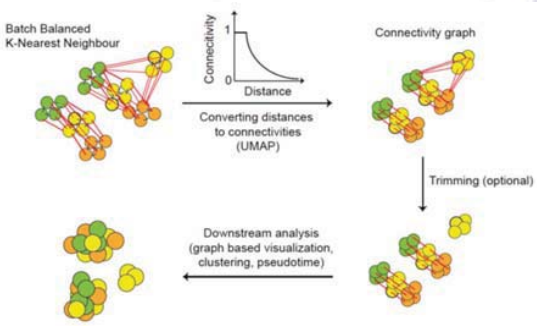
Model based cell type prediction

Correct systematic batch

Keeping biology

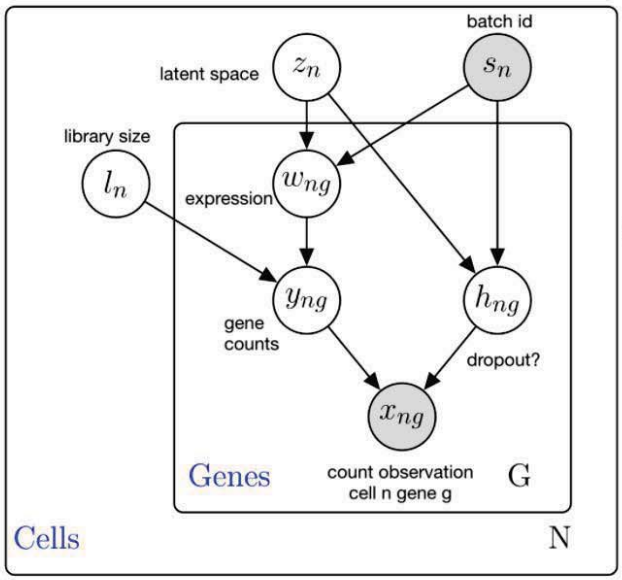
Gene Expression ~ Batch + Cell Type

Regularized linear model with batch + cell type design



BBKNN

Deep learning & scRNA-seq data integration



$$z_n \sim \mathcal{N}(0, I) \qquad s_n \in \{1, \cdots, K\}$$

$$w_{ng} \sim \text{Gamma}(f_w^g(z_n, s_n), \theta_g)$$

$$l_n \sim \text{LogNormal}(l_\mu^{s_n}, l_\nu^{s_n})$$

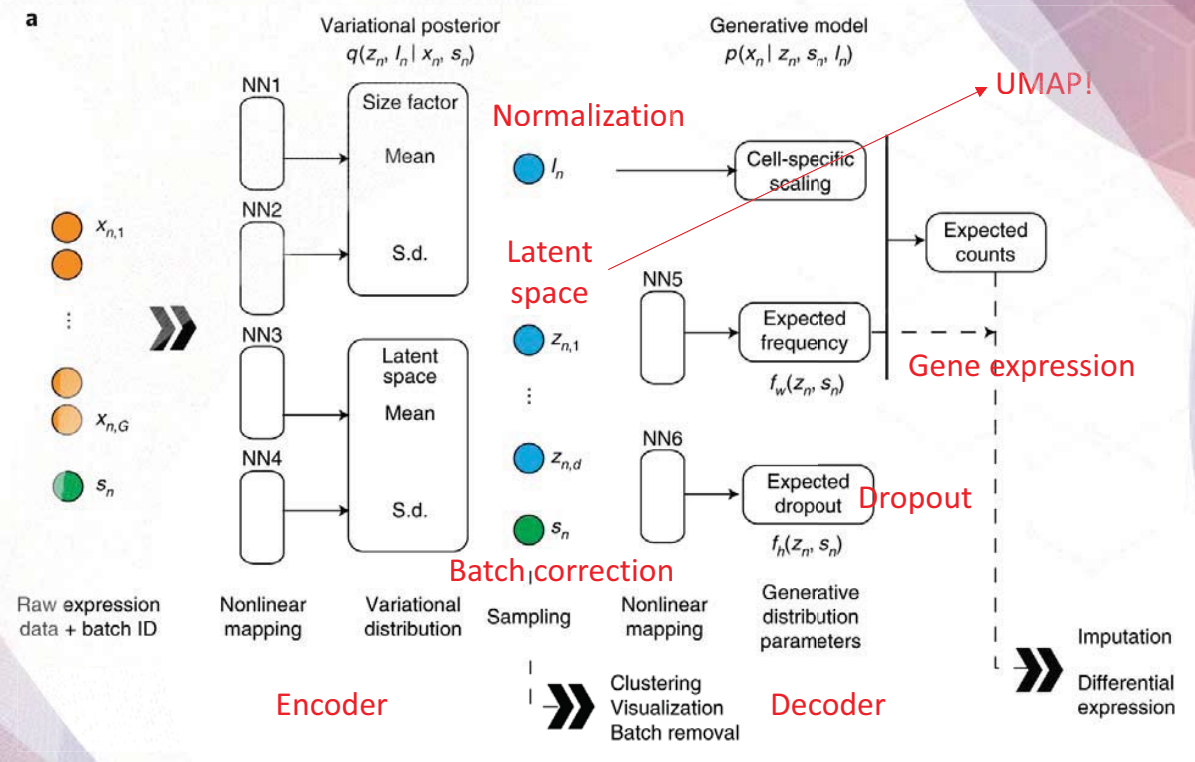
$$y_{ng} \sim \text{Poisson}(l_n w_{ng}) \qquad h_{ng} \sim \text{Bernoulli}(f_h^g(z_n, s_n))$$

$$x_{ng} = \begin{cases} y_{ng} & \text{if } h_{ng} = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Lopez, R., Regier, J., Cole, M.B. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat*

135

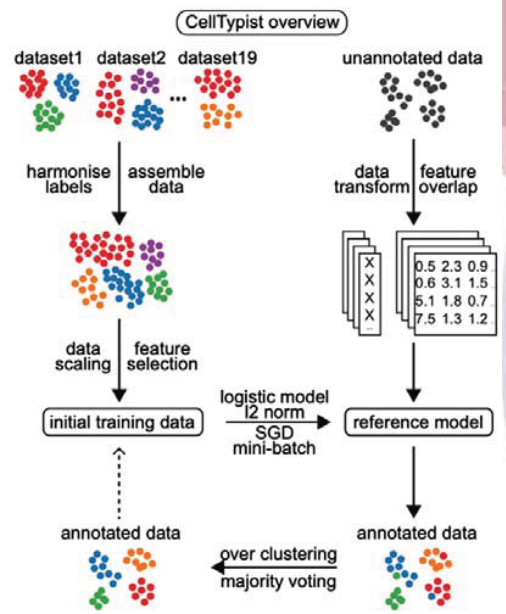
Batch effect & Embedding -> SCVI



Variational autoencoder

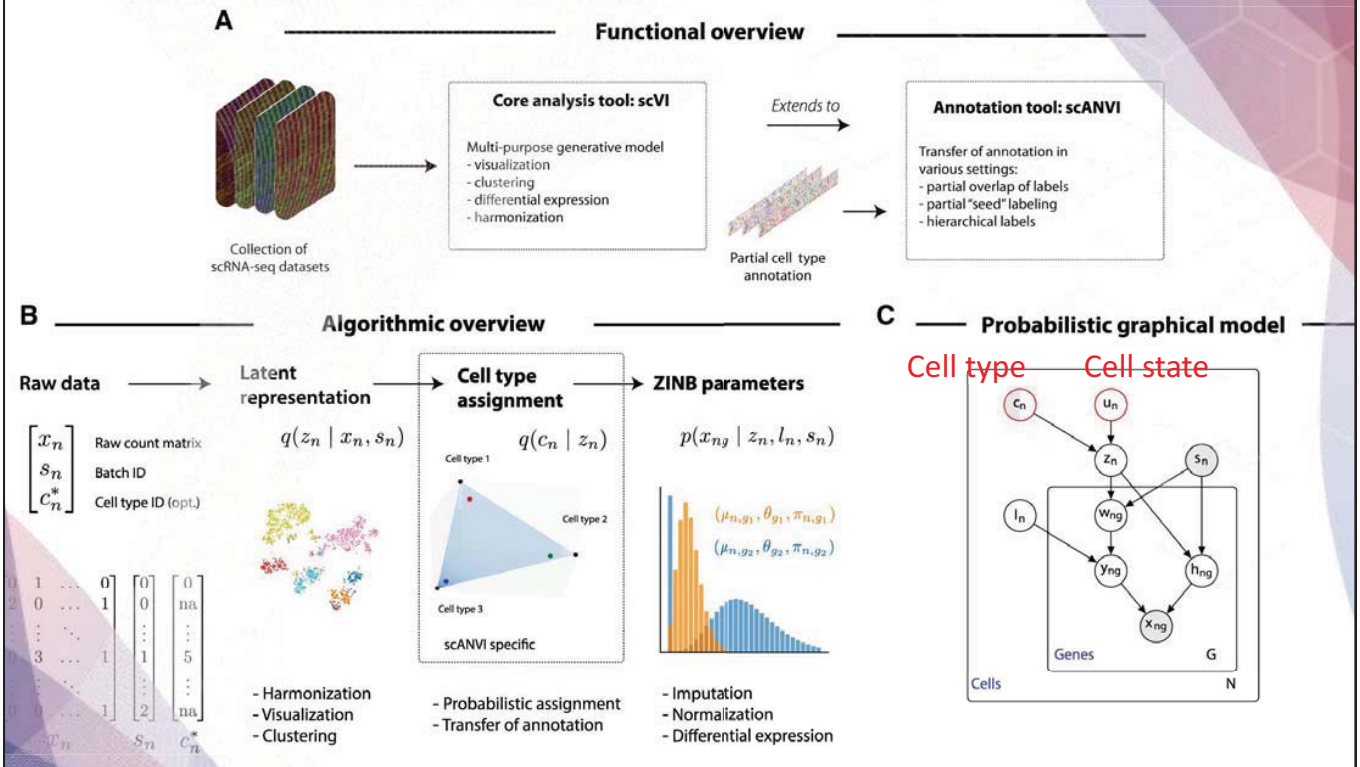
136

Machine learning based general cell annotation

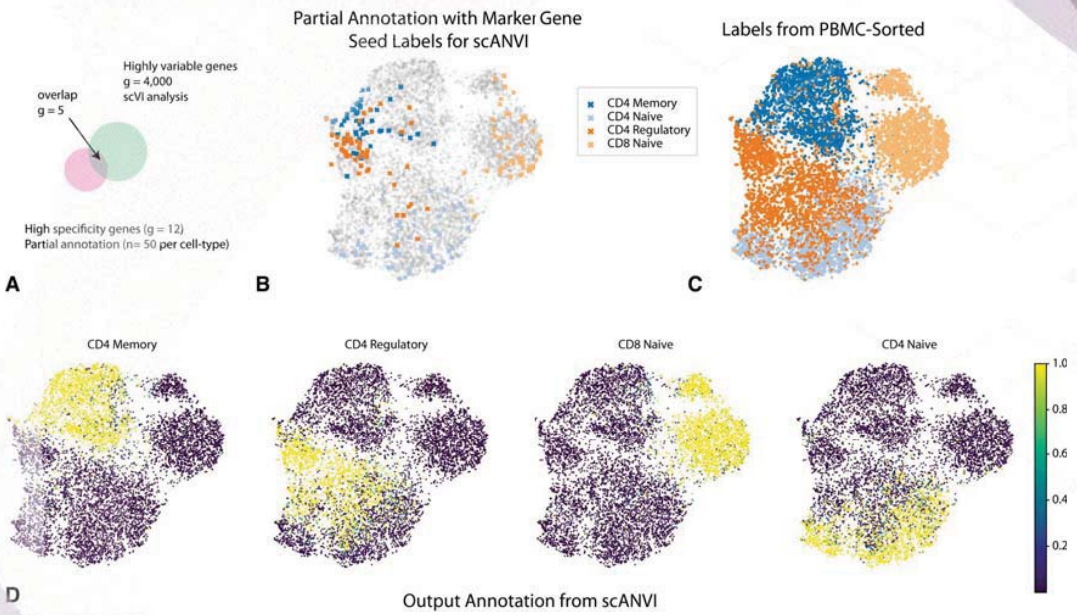


Dom'nguez Conde, C., et al. "Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture in humans." (2021).

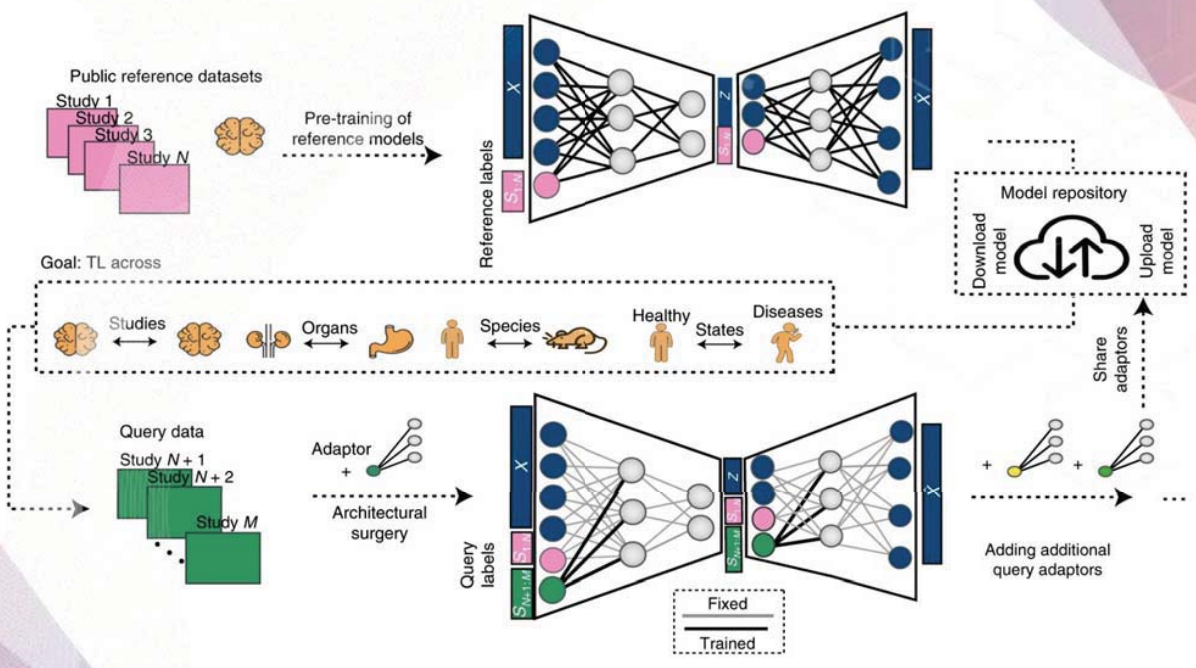
Annotation transfer -> SCANVI



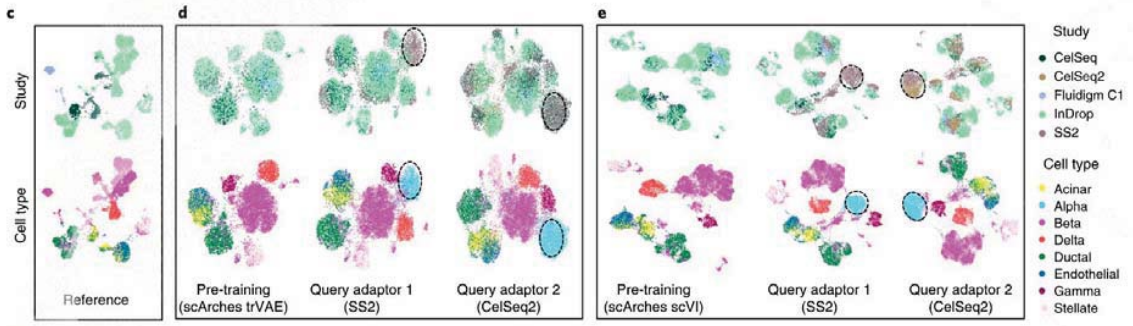
Annotation transfer -> SCANVI



Annotation transfer -> scArches (Transfer learning)



Annotation transfer -> scArches (Transfer learning)



Biology at single-cell resolution

https://docs.scvi-tools.org/

scvi-tools

Installation **Tutorials** User guide API Release notes References Contributing Discussion

Search the docs ...

- Introduction to scvi-tools
- Data loading and preparation
- Using Python in R with `reticulate`
- **Atlas-level integration of lung data**
- Integrating datasets with scVI in R
- Integration and label transfer with Tabula Muris
- Reference mapping with scvi-tools
- Seed labeling with scANVI
- Linearly decoded VAE
- Identification of zero-inflated genes
- Annotation with CellAssign
- Topic Modeling with Amortized LDA
- PeakVI: Analyzing scATACseq data
- ATAC-seq analysis in R
- Multi-resolution deconvolution of spatial transcriptomics
- Multi-resolution deconvolution of spatial transcriptomics in R

Note

This page was generated from harmonization.ipynb. Interactive online version: Open in Colab.

Atlas-level integration of lung data

An important task of single-cell analysis is the integration of several samples, which we can perform with scVI. For integration, scVI treats the data as unlabelled. When our dataset is fully labelled (perhaps in independent studies, or independent analysis pipelines), we can obtain an integration that better preserves biology using scANVI, which incorporates cell type annotation information. Here we demonstrate this functionality with an integrated analysis of cells from the lung atlas integration task from the scIB manuscript. The same pipeline would generally be used to analyze any collection of scRNA-seq datasets.

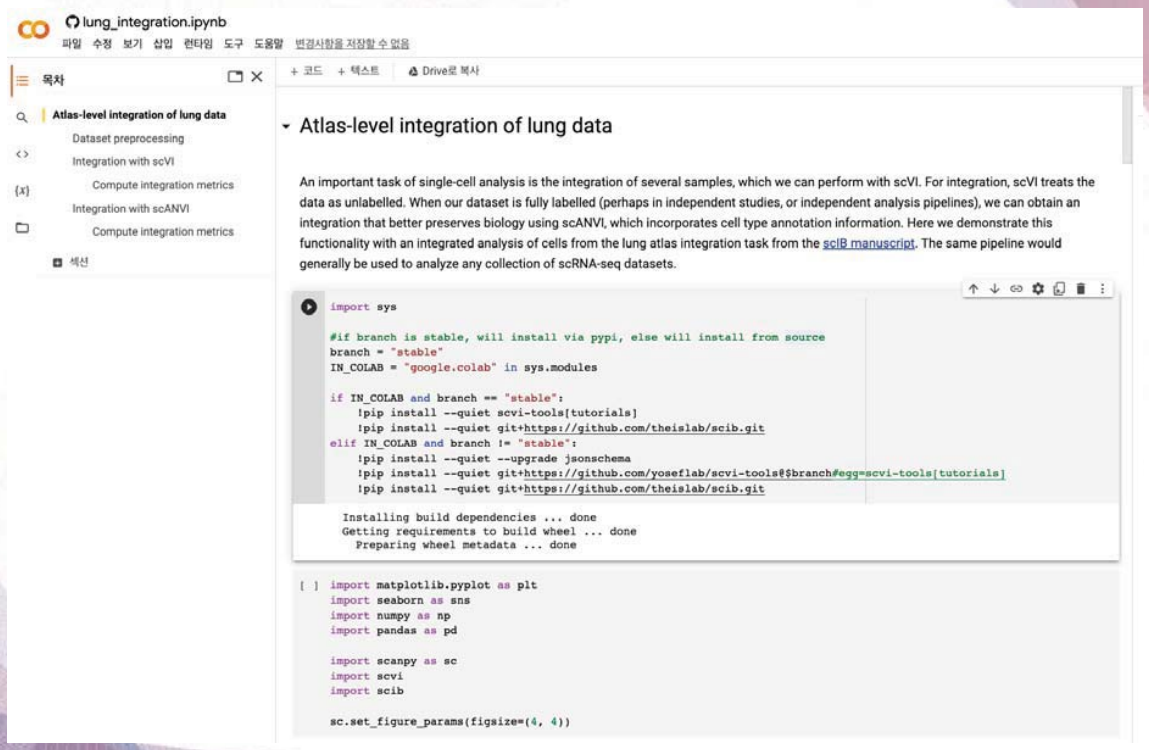
```
import sys

#if branch is stable, will install via pypi, else will install from source
branch = "stable"
IN_COLAB = "google.colab" in sys.modules

if IN_COLAB and branch == "stable":
    !pip install --quiet scvi-tools[tutorials]
    !pip install --quiet git+https://github.com/theislab/scib.git
elif IN_COLAB and branch != "stable":
    !pip install --quiet --upgrade jsonschema
    !pip install --quiet git+https://github.com/yoseflab/scvi-tools@$branch#egg=scvi-tools[tut
    !pip install --quiet git+https://github.com/theislab/scib.git
```

Try google colab platform

https://docs.scvi-tools.org/



Tutorial - SCVI

Atlas-level integration of lung data

An important task of single-cell analysis is the integration of several samples, which we can perform with scVI. For integration, scVI treats the data as unlabelled. When our dataset is fully labelled (perhaps in independent studies, or independent analysis pipelines), we can obtain an integration that better preserves biology using scANVI, which incorporates cell type annotation information. Here we demonstrate this functionality with an integrated analysis of cells from the lung atlas integration task from the scIB manuscript. The same pipeline would generally be used to analyze any collection of scRNA-seq datasets.

```
!pip install --quiet scvi-colab
!pip install --quiet git+https://github.com/theislab/scib.git
from scvi_colab import install
install()
```

Integration with scVI

As a first step, we assume that the data is completely unlabelled and we wish to find common axes of variation between the two datasets. There are many methods available in scanpy for this purpose (BBKNN, Scanorama, etc.). In this notebook we present scVI. To run scVI, we simply need to:

- Register the AnnData object with the correct key to identify the sample and the layer key with the count data.
- Create an SCVI model object.

```
scvi.model.SCVI.setup_anndata(adata, layer="counts", batch_key="batch")
```

We note that these parameters are non-default; however, they have been verified to generally work well in the integration task.

```
vae = scvi.model.SCVI(adata, n_layers=2, n_latent=30, gene_likelihood="nb")
```

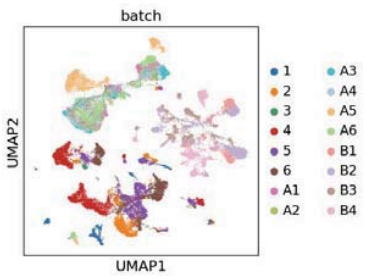
Now we train scVI. This should take a couple of minutes on a Colab session

```
vae.train()
```

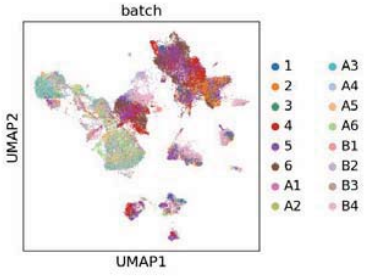
```
Epoch 246/246: 100%|| 246/246 [09:19<00:00,  2.27s/it, loss=553, v_num=1]
```

Tutorial - SCVI

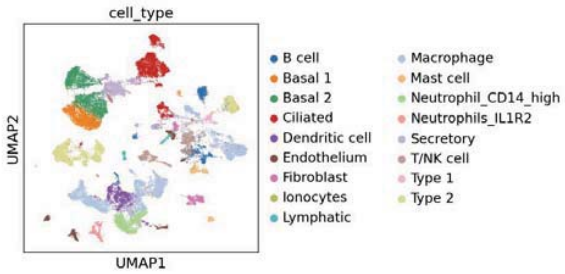
```
[ ] sc.pl.umap(adata,color='batch')
```



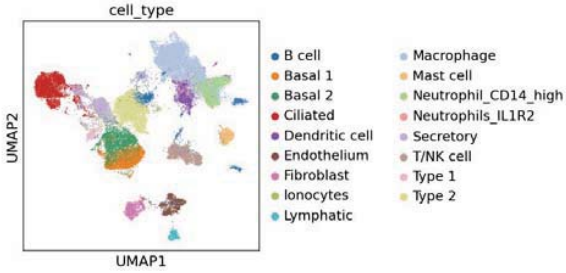

```
sc.pl.umap(adata,color='batch')
```




```
sc.pl.umap(adata,color='cell_type')
```




```
[ ] sc.pl.umap(adata,color='cell_type')
```

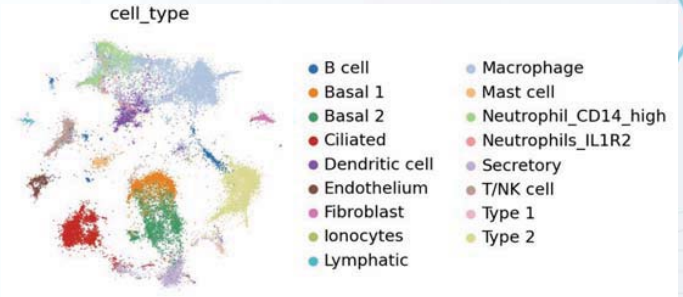


Tutorial - SCANVI

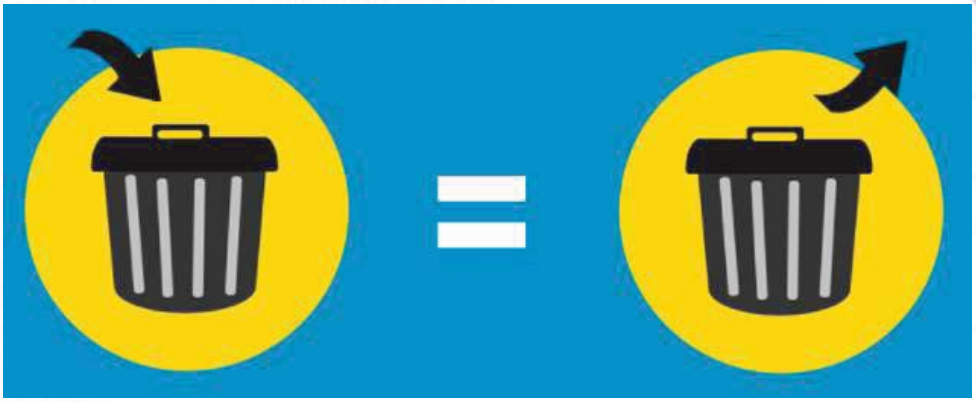
```
lvae = scvi.model.SCANVI.from_scvi_model(
    vae,
    adata=adata,
    labels_key="cell_type",
    unlabeled_category="Unknown",
)
```

```
lvae.train(max_epochs=20, n_samples_per_label=100)
```

```
INFO     Training for 20 epochs.
Epoch 20/20: 100%|| 20/20 [01:39<00:00,  4.96s/it, loss=628, v_num=1]
```



More things to consider…



- Removing bad quality data – “Garbage in garbage out”
- Removing doublets
- Considering ‘soup effect’

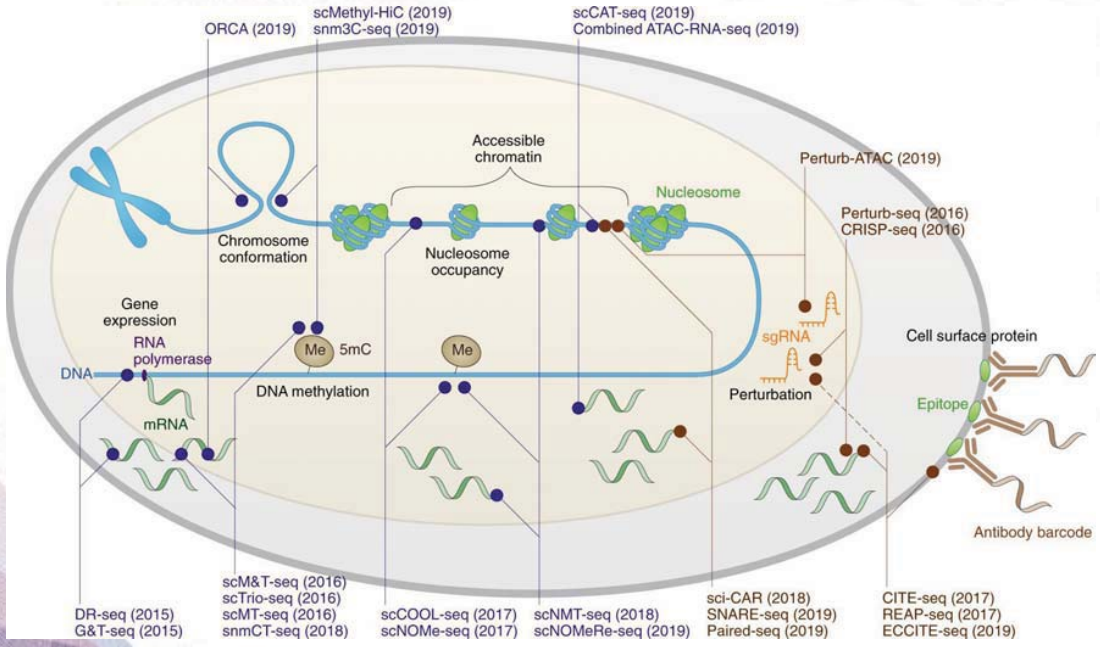
147

4. Single-cell multi-omics analysis

148

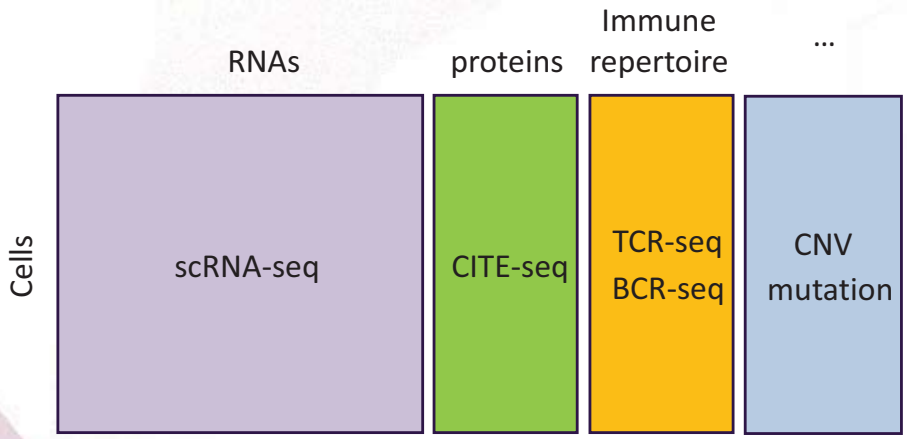
What about other single-cell omics data?

Beyond RNA: Single-cell multiomics



149

Beyond gene expression matrix



Structure of ATAC-seq data

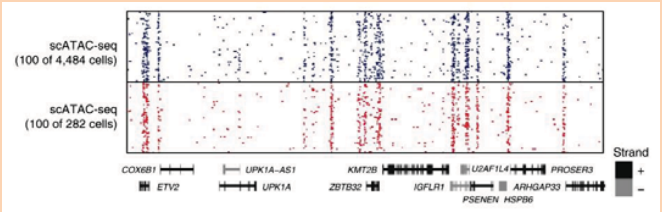
genes

Cells (set A)

scRNA-seq

genomic coordinates

Cells (set B)



Peak calling for ATAC-seq

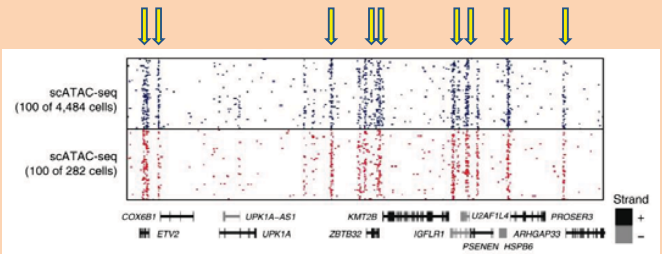
genes

Cells (set A)

scRNA-seq

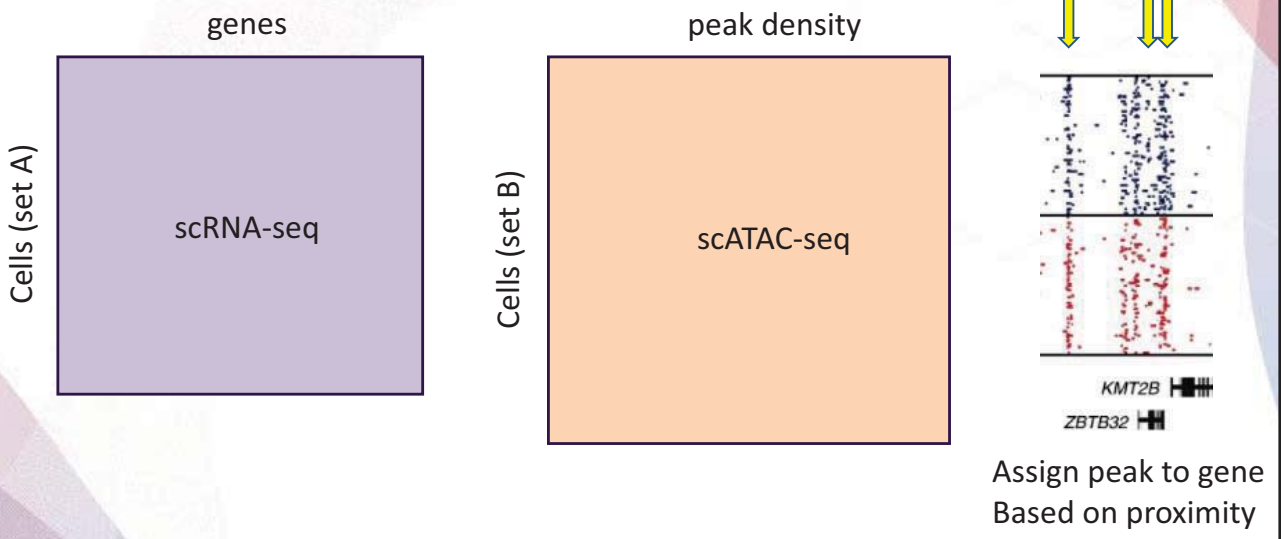
genomic coordinates

Cells (set B)



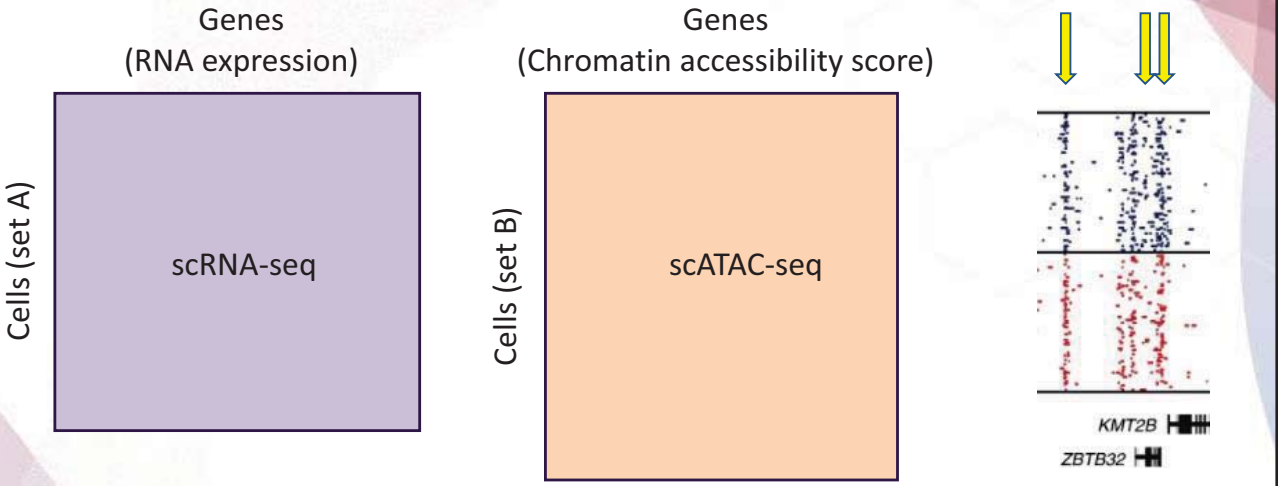
Peak calling (de novo)
Prior knowledge (bulk ATAC-seq, ChIP-seq)

Linking peaks to genes



Assign peak to gene
Based on proximity

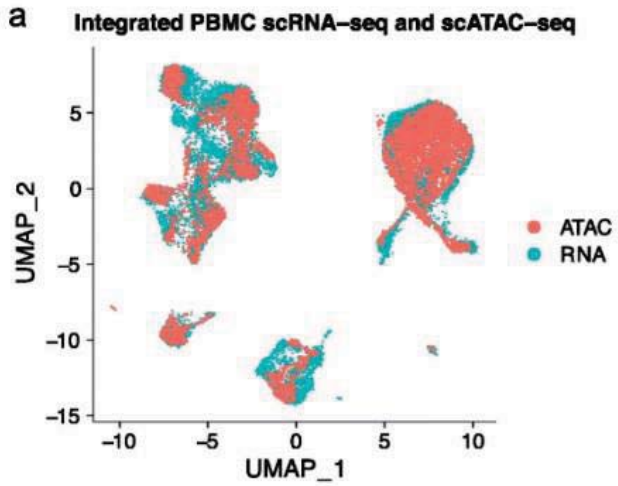
Linking peaks to genes



Batch effect correction problem

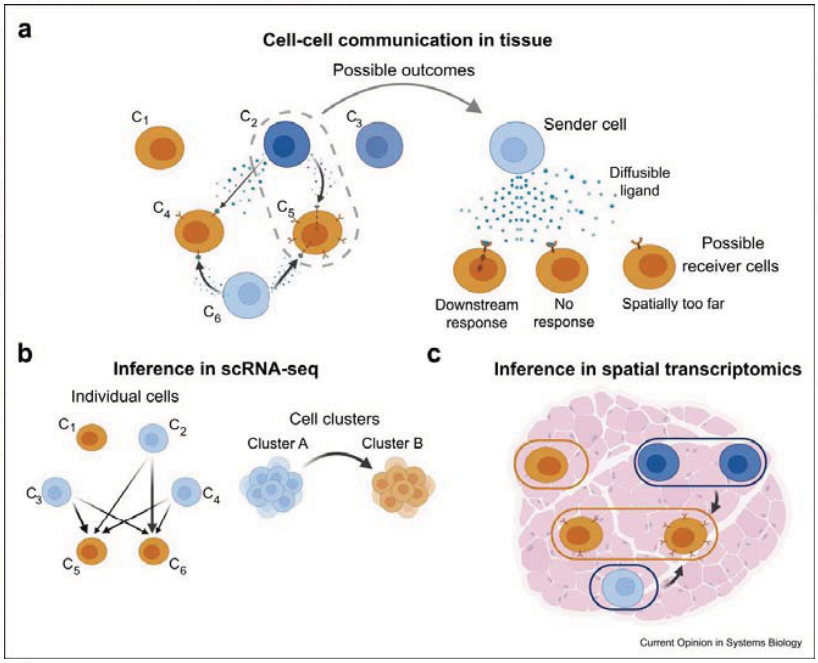
Assign peak to gene
Based on proximity

Combining RNA with chromatin information



Link epigenetics to RNA expression

Importance of spatial information

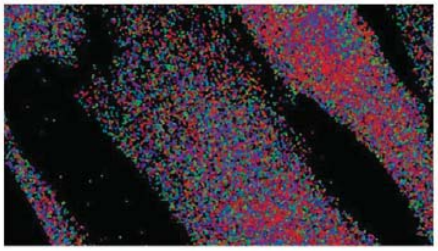


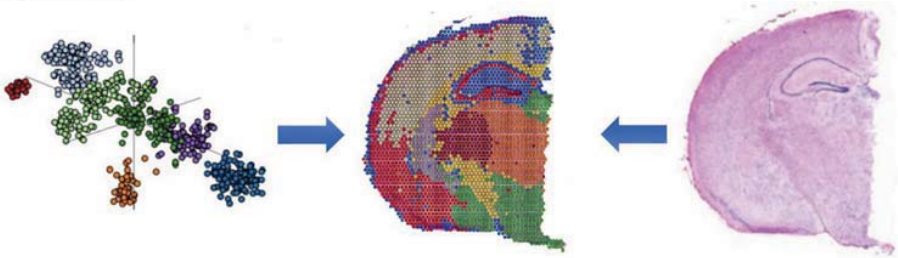
Single cell transcriptomics with spatial resolution

FOCUS | 06 JANUARY 2021

Method of the Year 2020: spatially resolved transcriptomics

Spatially resolved transcriptomics is our Method of the Year 2020, for its ability to provide valuable insights into the biology of cells and tissues while retaining information about spatial context.

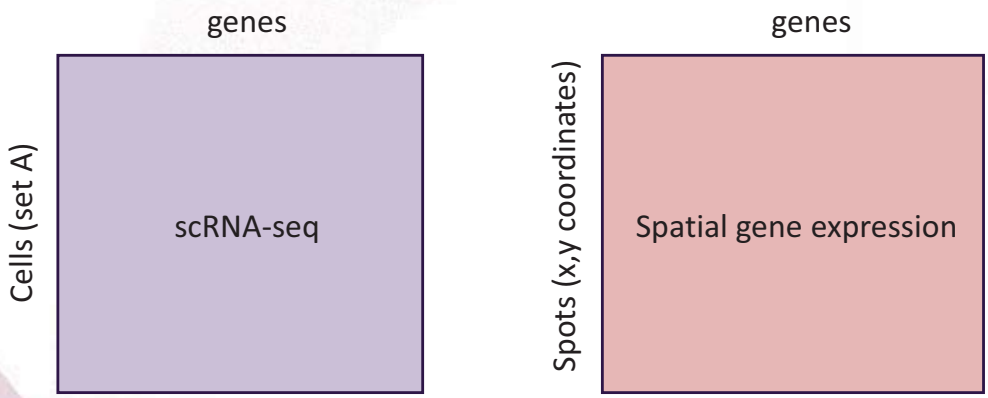




Single Cell Gene Expression | Spatially Resolved Gene Expression | Tissue Section

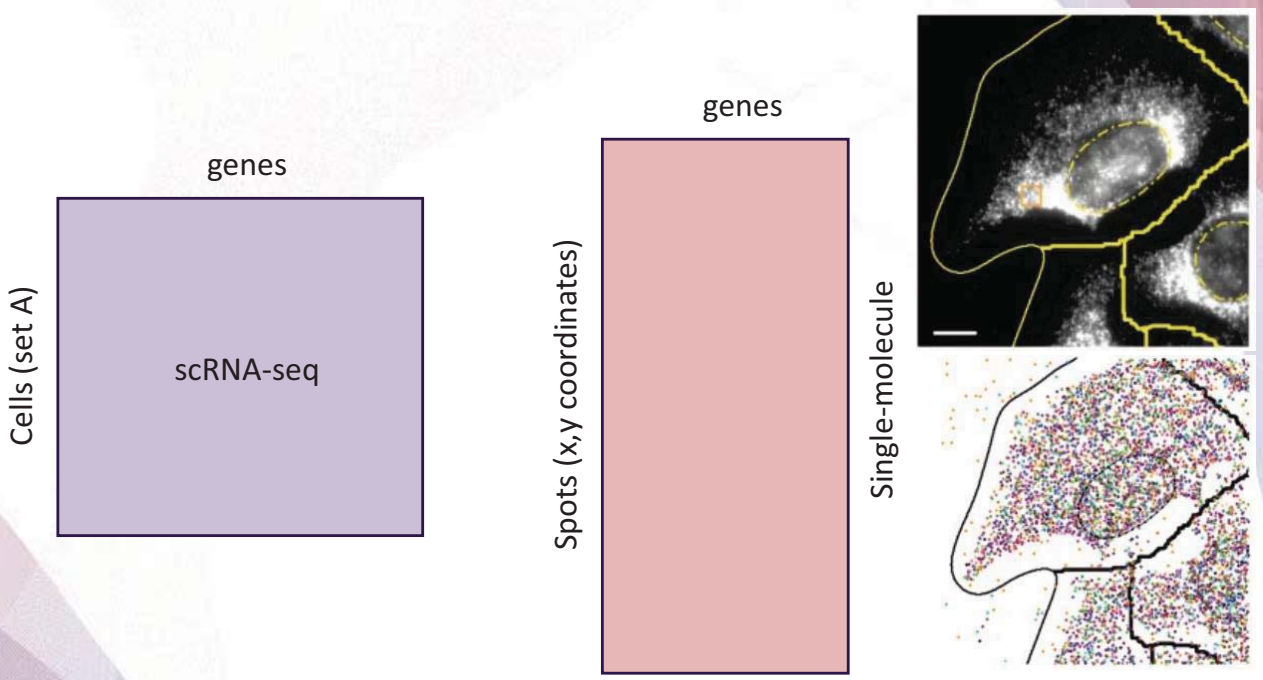
Adapted from 10x Genomics

Structure of spatial gene expression data

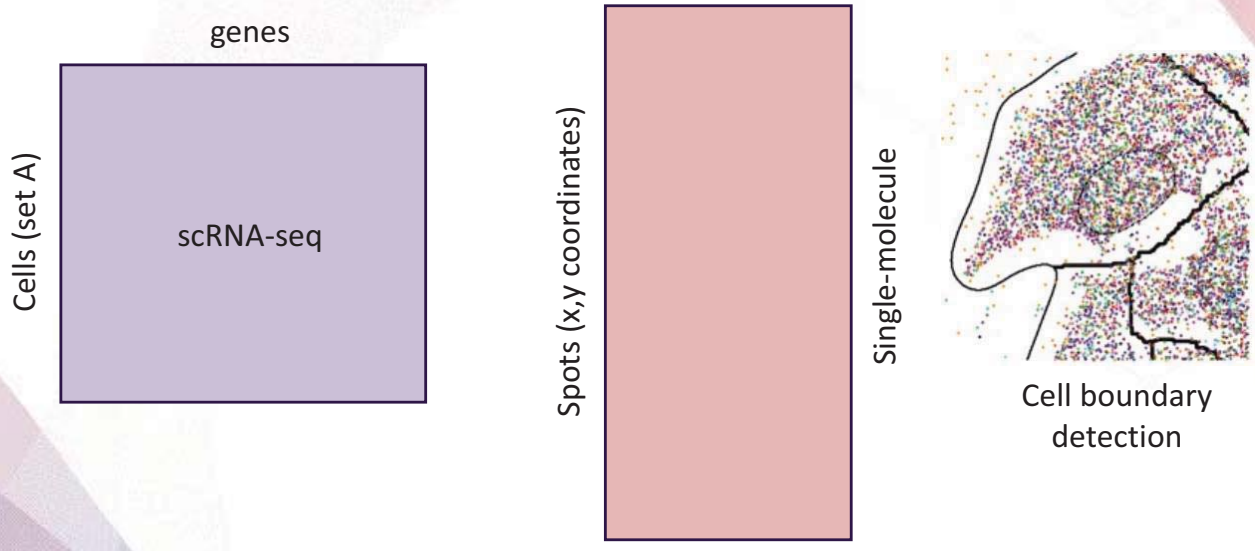


Structure of spatial gene expression data (Visium)

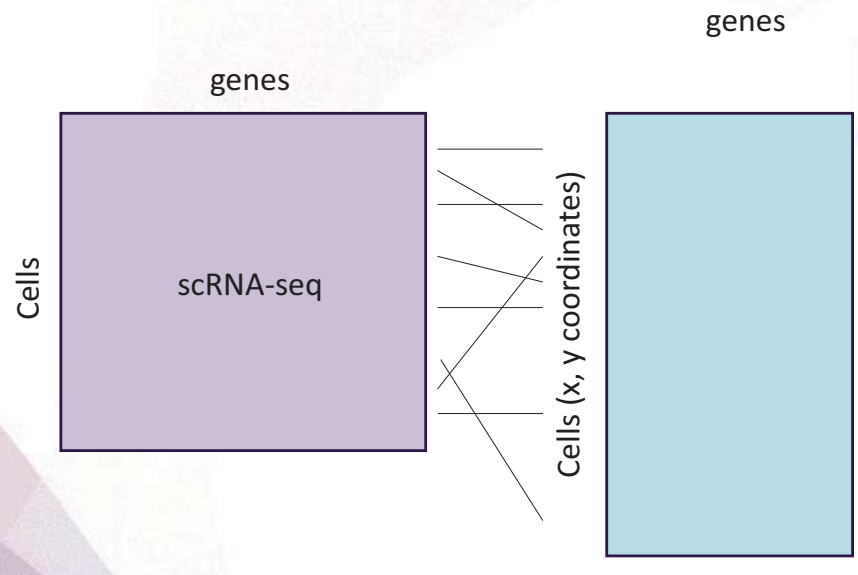
Structure of spatial gene expression data (smFISH)



Structure of spatial gene expression data (smFISH)

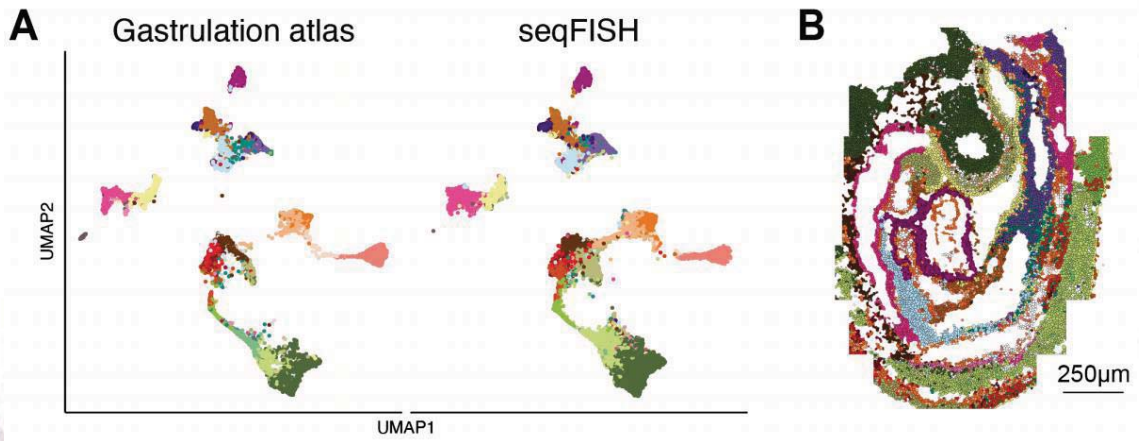


Mapping cells to space

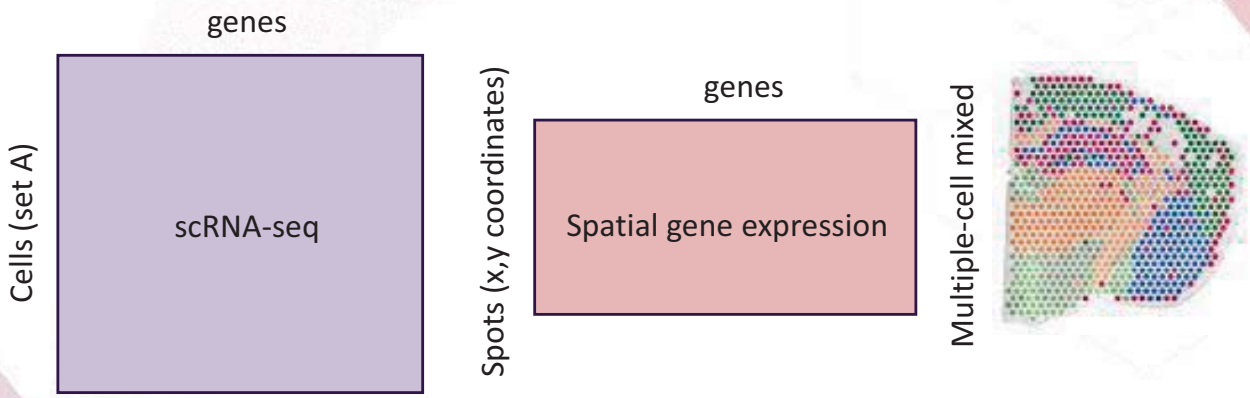


Batch effect correction problem

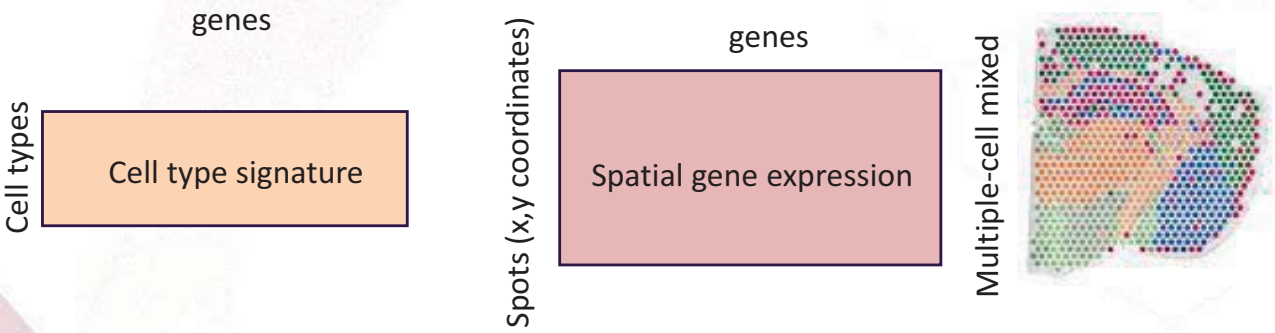
Spatial data alignment with single-cell data



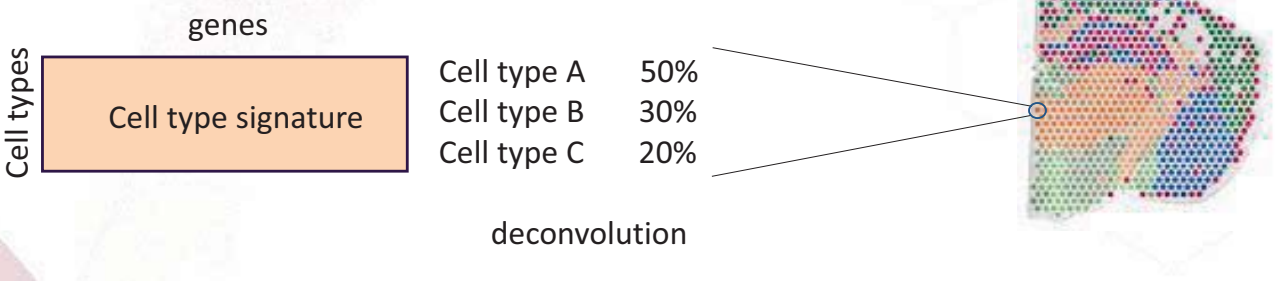
Problem of spatial data deconvolution (e.g. Visium)



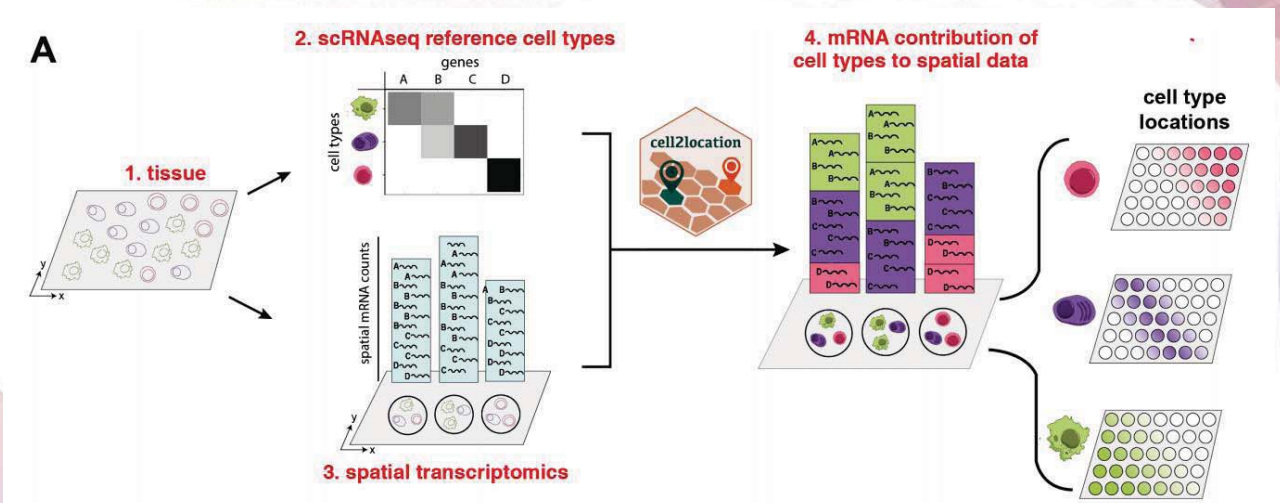
Problem of spatial data deconvolution (e.g. Visium)



Problem of spatial data deconvolution (e.g. Visium)

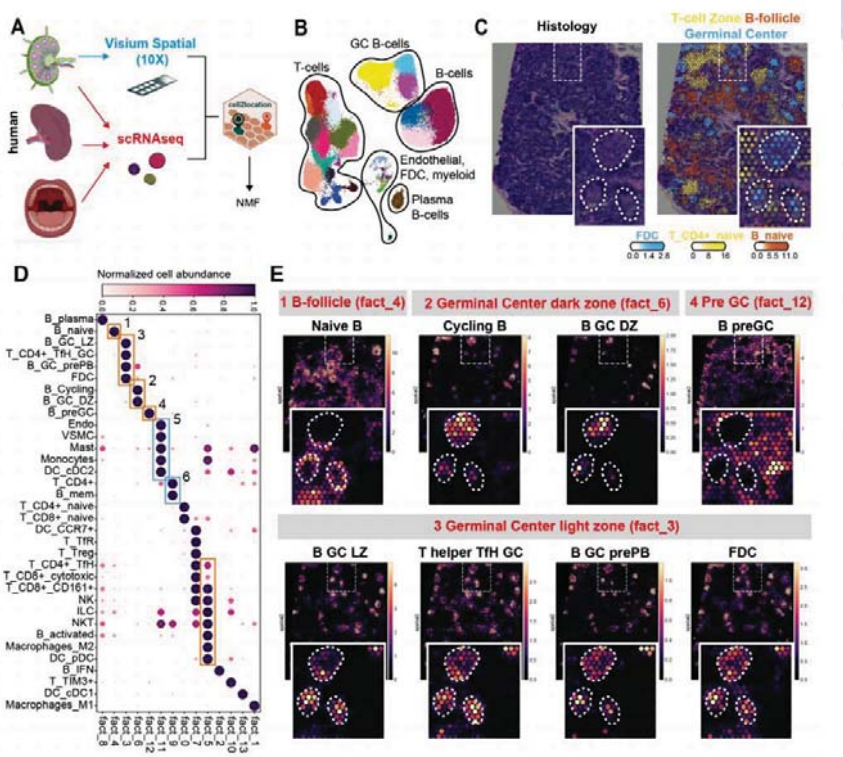


Problem of spatial data deconvolution (e.g. cell2location)



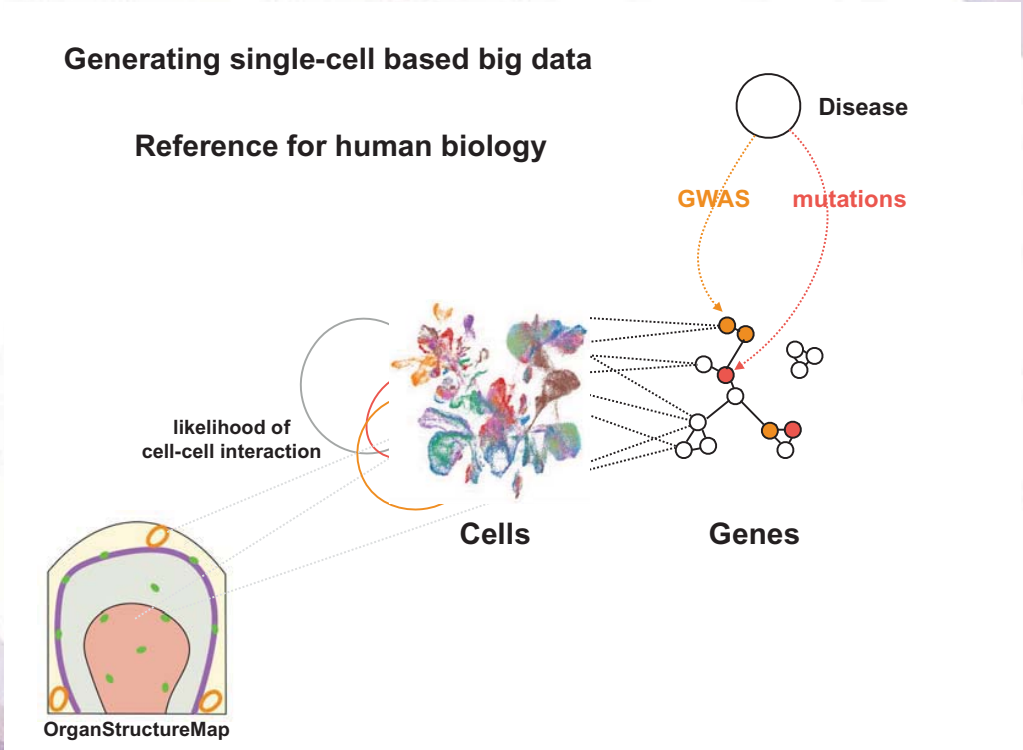
Kleshchevnikov, V., Shmatko, A., Dann, E. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* (2022).

Problem of spatial data deconvolution (cell2location)



Kleshchevnikov, V., Shmatko, A., Dann, E. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* (2022).

Conclusion



Link to the practice

https://drive.google.com/drive/folders/1GYq-gM3X9JlV2608UGw9AgElvu8_q-5M?usp=sharing

170

감사합니다

171