KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (오프라인)

Best practice for single-cell data analysis

박종은 _ KAIST

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 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 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 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 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	**의료빅데이터/인공지능 총론** 김헌성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	**의료영상 인공지능의 이해 및 의료영상 레이블링 실습** 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	**의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기** 김선근 대표(원닥 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	**EMR 데이터를 활용한 머신러닝 기반 예후예측:** **Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)** 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	**Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단:** **Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)** 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	**심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network +** **Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)** 고태훈 교수(가톨릭대학교)

DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**DNN (이론)** 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	**CNN (이론)** 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	**RNN, ChatGPT, XAI (이론)** 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	**CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)** 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Best practice for single-cell data analysis** 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	**Practice1: Scanpy basic workflow** 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	**Public database, data integration, reference mapping, multiomics** 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	**Practice2: Advanced single-cell analysis (siVI universe)** 정성민 조교, 고용준 조교

DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**AI-based protein structure prediction** **- Intro to protein structure prediction** **- Early AI-based approaches** **- AlphaFold and RoseTTAFold** 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	**단백질 구조 예측 실습** **- ColabFold를 활용한 단백질 구조 및 상호작용 예측** **- Tips &Tricks for better structure modeling** 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	**AI-based protein design** **- Intro to protein design** **- Protein backbone design using RFdiffusion** **- Protein sequence design using ProteinMPNN** 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	**단백질 디자인 실습** **- RFdiffusion 및 ProteinMPNN의 활용법 실습** 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Single-cell biology** 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	**i. Unsupervised Spatial transcriptome analysis** **ii. Tumor Boundary Determination in Spatial Transcriptomics** 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	**i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics** **ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics** 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	**i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset** **ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets** 천하림 조교, 이호진 조교

DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Transformers (이론)** 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	**Introduction to Transformers (실습)** 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	**Deep learning in Bioinformatics** 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	**Deep learning model을 이용한 실습** 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**마이크로바이옴 기본 이론** 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	**16S rRNA amplicon seq. - DADA2** 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	**최신 메타지놈 분석 기법의 현황** 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	**Shotgun metagenome 분석 (Linux)** 조준우 조교, 백재우 조교

DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)** **Molecular Notations &Descriptors / AI 신약개발을 위한 Databases** **AI 신약개발을 위한 Programming 기초** 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	**Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습** **Bioactivity database 검색 및 정보 읽기 실습** **Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습** 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	**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(KAIST)
15:10-15:20	휴식
15:20-16:50	**QSAR modeling 전체 과정 실습/ 화합물의 Bioactivity 예측 모델 개발** **Virtual screening 과정을 통한 신약후보물질 발굴 실습** 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Single cell multiomics 이론 / Gene regulatory network 이론** 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	**Seurat/Signac, ArchR, TENET+ 실습** 김현규 조교, 정회빈 조교
12:40-14:10	점심
14:10-15:40	**롱리드 시퀀싱 소개 및 유전체 조립 실습** 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	**변이 분석 및 시각화 실습** 김준 교수(충남대학교)

강의개요

Best practice for single-cell data analysis

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. Kwon, J.*, **Kang, J.***, ..., An H. J.#, Lee H.-O.#, **Park J.-E**#, Choi, J. K.# (2023). Single-cell mapping of combinatorial target antigens for CAR switches using logic gates. ***Nature Biotechnology***, 1-13.
2. **Kang, J.**,* Kim, M.*, Yoon, D. Y.*, **Kim, W. S.**, ..., **Park, M.**, Lee, J. S., **Park J.-E**.#, & Kim, S. M.# (2023). AXL+ SIGLEC6+ dendritic cells in cerebrospinal fluid and brain tissues of patients with autoimmune inflammatory demyelinating disease of CNS. ***Clinical Immunology***, 109686.
3. **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
4. **Park, J.-E.***, Jardine, L.*, Gottgens, B., Teichmann, S.A., Haniffa, M., (2020). Prenatal development of human immunity. ***Science*** 368, 600–603.
5. Polański, K.*, Young, M.D.*, Miao, Z., Meyer, K.B., Teichmann, S.A.# and **Park, J.-E.**#, (2020). BBKNN: Fast Batch Alignment of Single Cell Transcriptomes. ***Bioinformatics.*** 36, 964-965

KSBi-BIML 2024

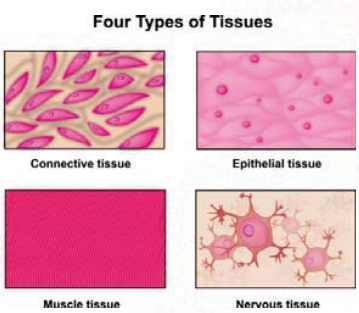
Best practice in single-cell analysis
(python version)

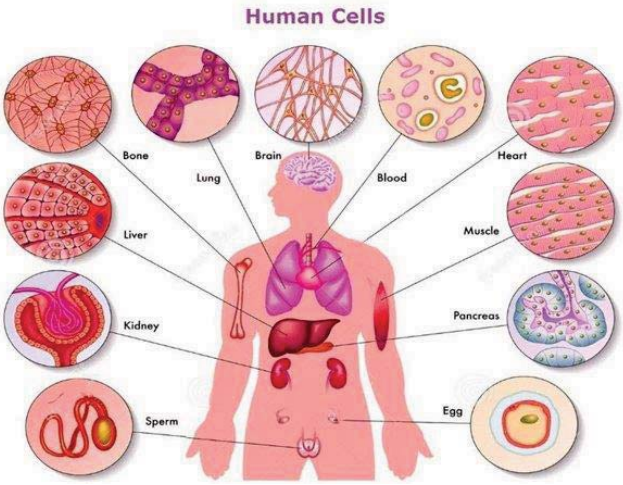
카이스트 의과학대학원
박종은
조교: 정성민/고용준

Understanding the complexity of human tissue

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

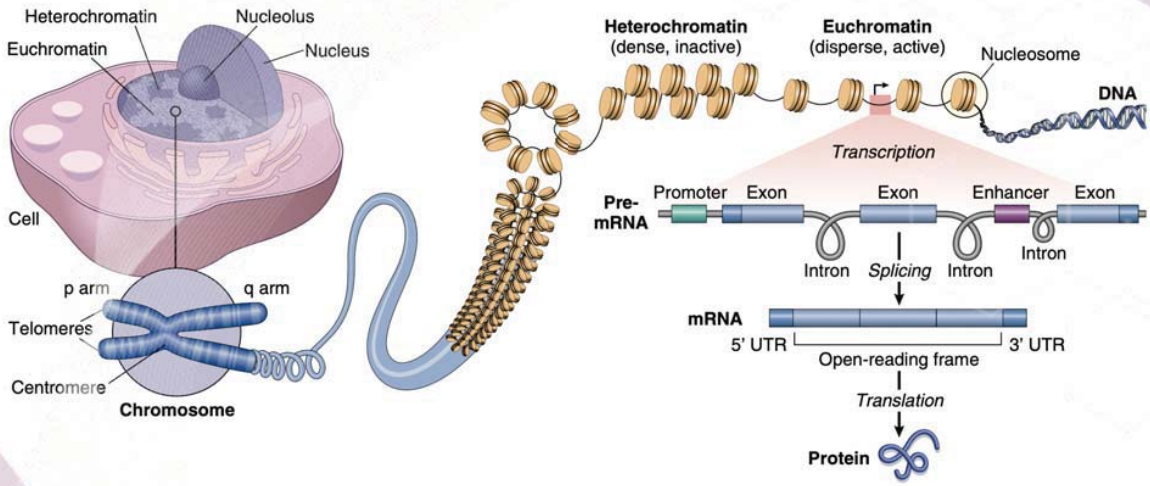
Diversity of the cells in our body





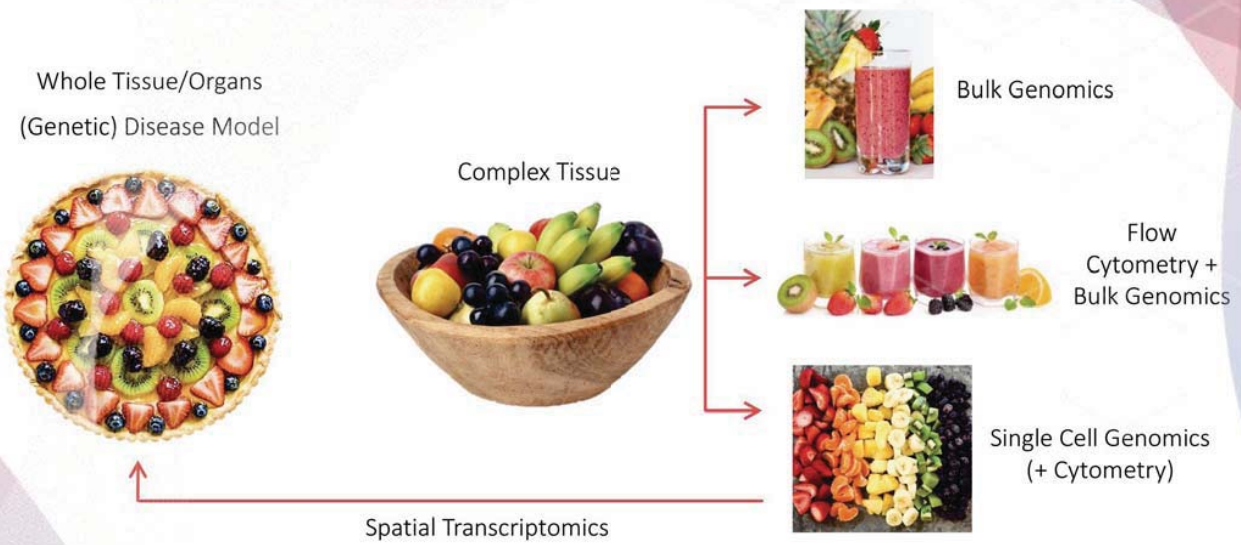
4 tissues, 18 tissue types, 78 organs, > 200 cell types, 37 trillion cells

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

From bulk to single cell and spatial approaches



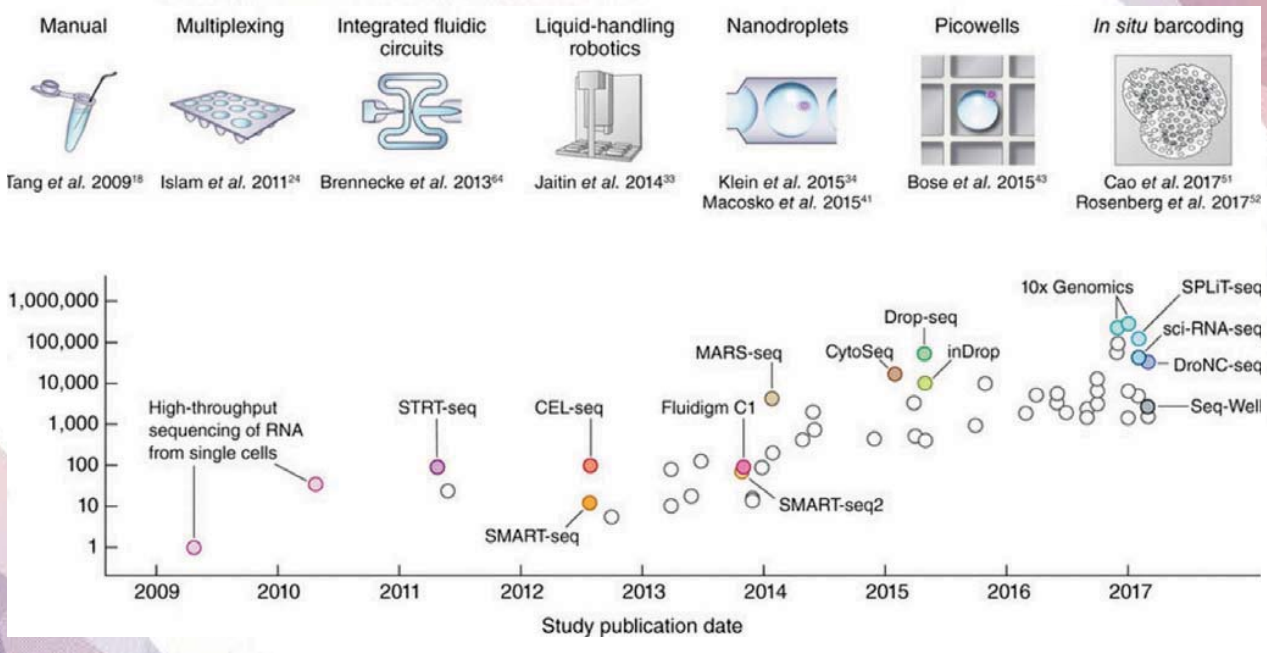
강의 개요

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

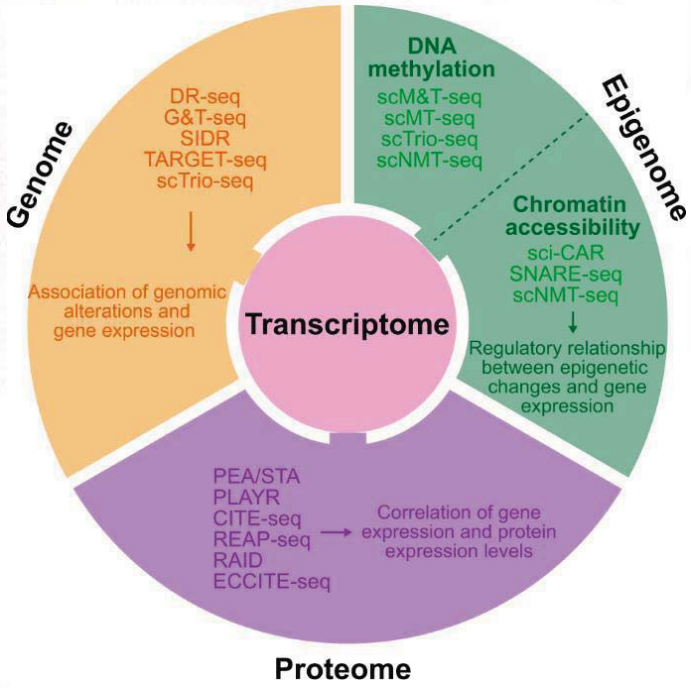
1. Single-cell data 의 특성과 분석 개요

Advancement in single-cell technologies



Svensson & Vento-Tormo, 2018, Nature biotechnology

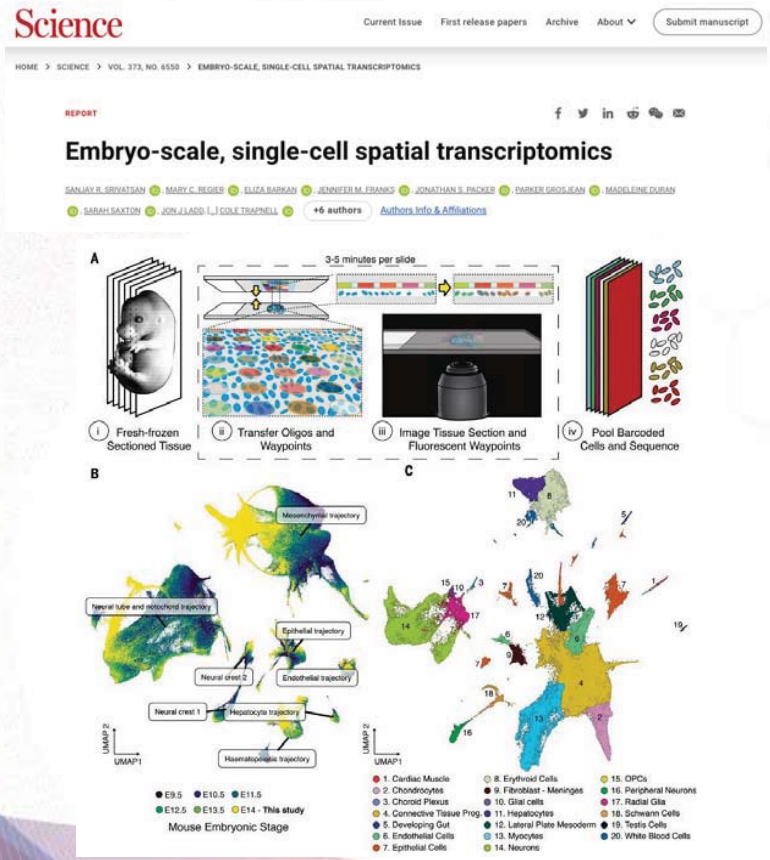
Multi-omics at single-cell resolution



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

9

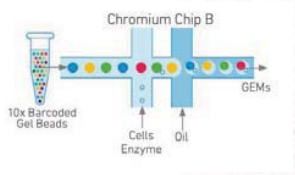
Spatial transcriptomics



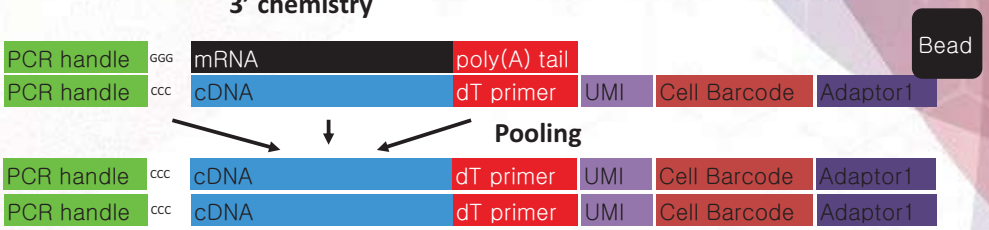
10

Single-cell library generation

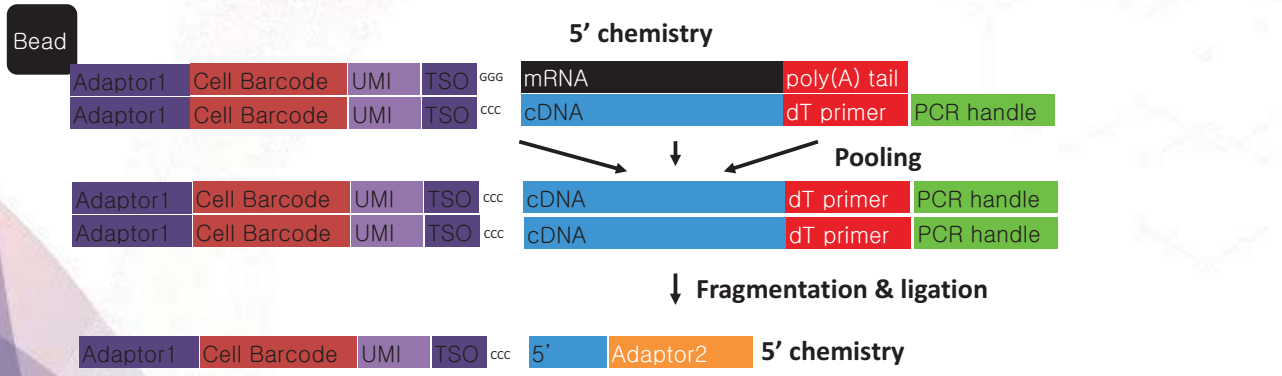
10X GENOMICS



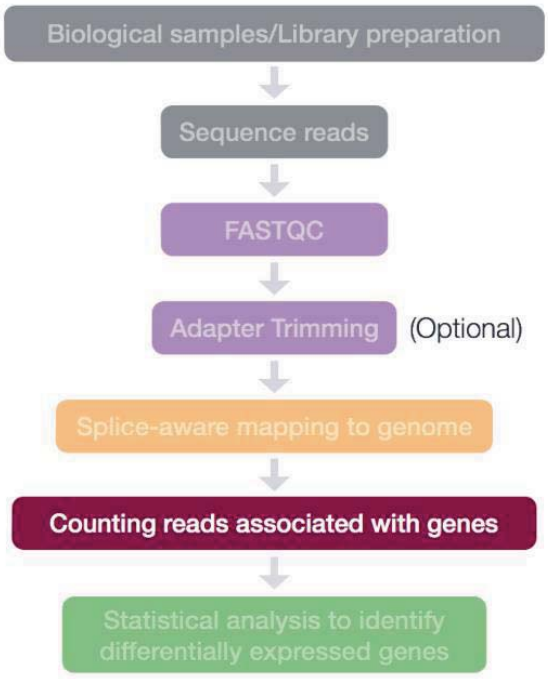
3' chemistry



5' chemistry



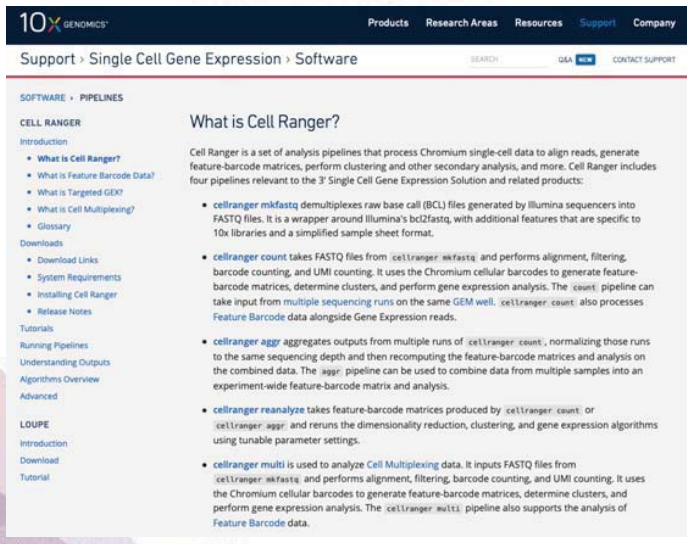
RNA-seq analysis

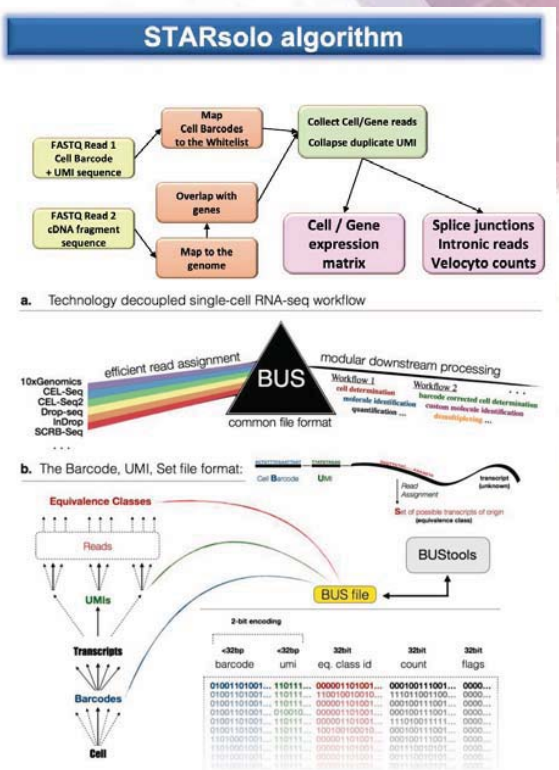


https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.html

Single-cell RNA-seq data alignment & counting

Packages for single-cell mapping





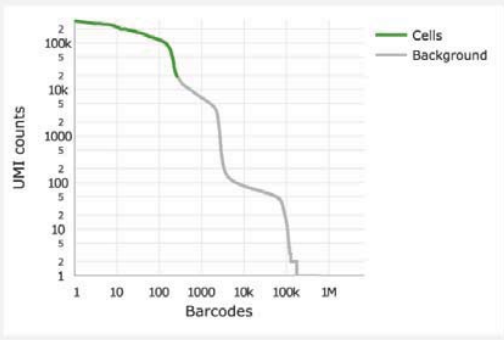
Cellranger output

Cellranger output and cell barcode filtering

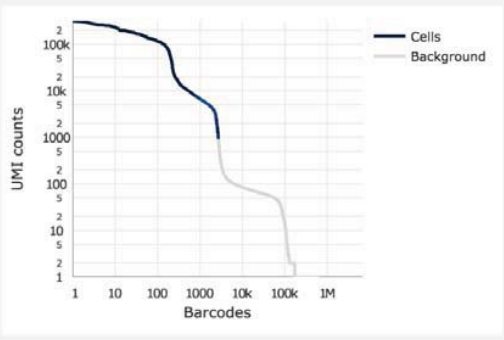
```
drwxr-xr-x 7 user1 genomics       4096  2월 15 21:07 analysis
-rwxr-xr-x 1 user1 genomics   57216204  2월 16 05:15 cloupe.cloupe
drwxr-xr-x 2 user1 genomics       4096  2월 16 06:45 filtered_feature_bc_matrix
-rwxr-xr-x 1 user1 genomics   15736383  2월 16 05:15 filtered_feature_bc_matrix.h5
drwxr-xr-x 3 user1 genomics       4096  2월 15 21:07 filtered_gene_bc_matrices
-rwxr-xr-x 1 user1 genomics        682  2월 16 05:15 metrics_summary.csv
-rwxr-xr-x 1 user1 genomics  182565559  2월 16 05:16 molecule_info.h5
-rwxr-xr-x 1 user1 genomics 24271610464  2월 16 06:45 possorted_genome_bam.bam
-rwxr-xr-x 1 user1 genomics    7117296  2월 16 06:45 possorted_genome_bam.bam.bai
drwxr-xr-x 2 user1 genomics       4096  2월 16 06:46 raw_feature_bc_matrix
-rwxr-xr-x 1 user1 genomics   33432810  2월 16 06:45 raw_feature_bc_matrix.h5
-rwxr-xr-x 1 user1 genomics    6509377  2월 16 06:45 web_summary.html
```

```
drwxr-xr-x 2 user1 genomics 4.0K  2월 16 06:45 .
drwxr-xr-x 6 user1 genomics 4.0K  2월 16 06:45 ..
-rwxr-xr-x 1 user1 genomics  30K  2월 16 06:45 barcodes.tsv.gz
-rwxr-xr-x 1 user1 genomics 298K  2월 16 06:45 features.tsv.gz
-rwxr-xr-x 1 user1 genomics  35M  2월 16 06:45 matrix.mtx.gz
```

```
drwxr-xr-x 2 user1 genomics 4.0K  2월 16 06:46 .
drwxr-xr-x 6 user1 genomics 4.0K  2월 16 06:45 ..
-rwxr-xr-x 1 user1 genomics 2.5M  2월 16 06:46 barcodes.tsv.gz
-rwxr-xr-x 1 user1 genomics 298K  2월 16 06:46 features.tsv.gz
-rwxr-xr-x 1 user1 genomics  45M  2월 16 06:46 matrix.mtx.gz
```

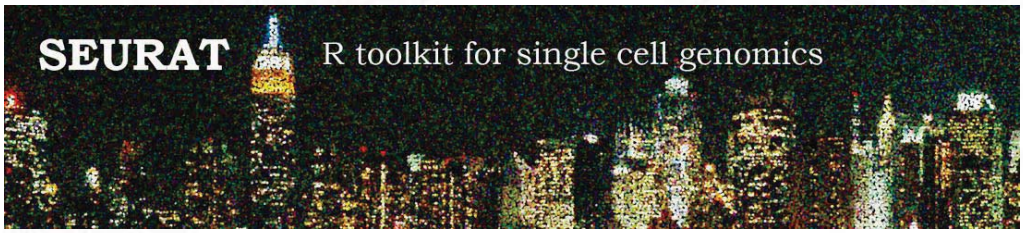


Cell Ranger 2.2

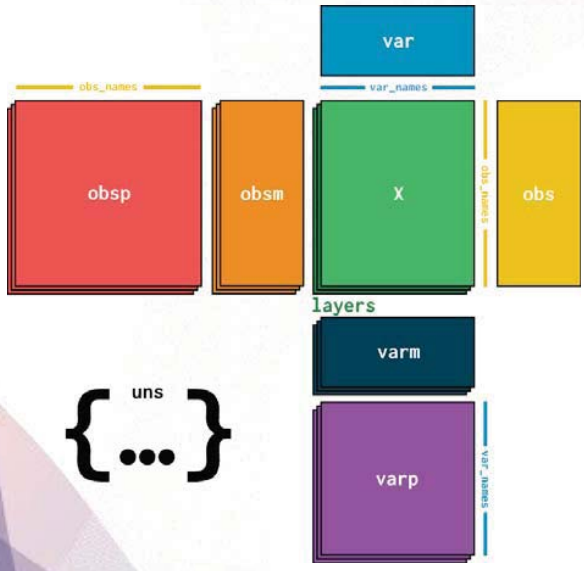


Cell Ranger 3.0

Data analysis package



Single-cell data structure: Anndata format



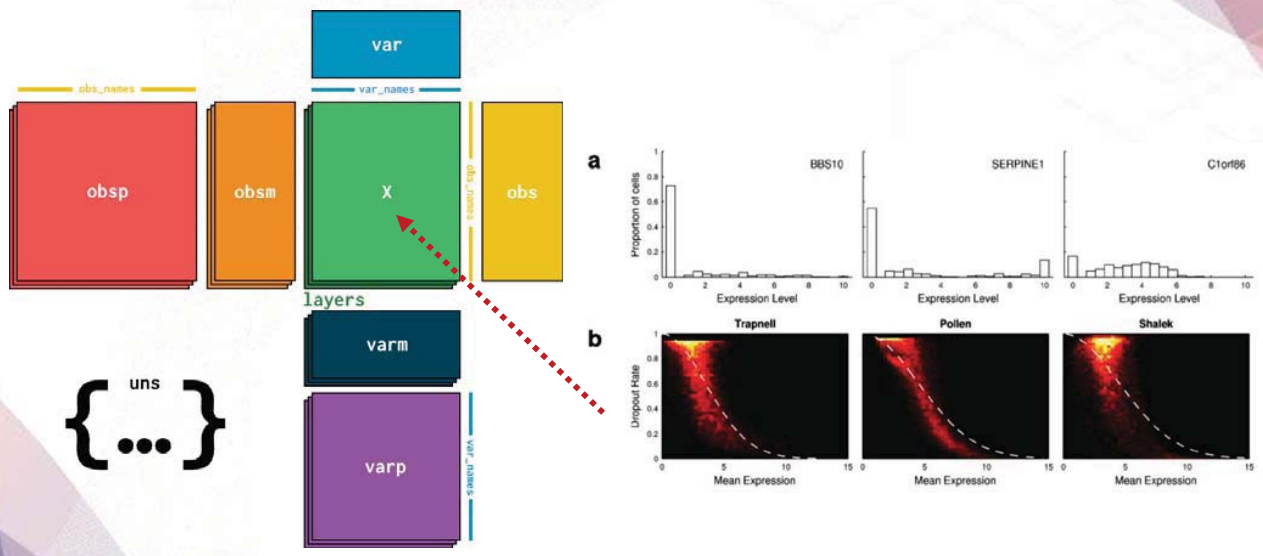
- X: Gene expression values
- Obs: observations, cells
- Var: variables, genes
- Obsm: PCA dimensions, …
- Obsp: Graph connectivities, …
- Varm: PCA loadings, …
- Varp: Gene-wise correlations, …
- Uns: anything

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

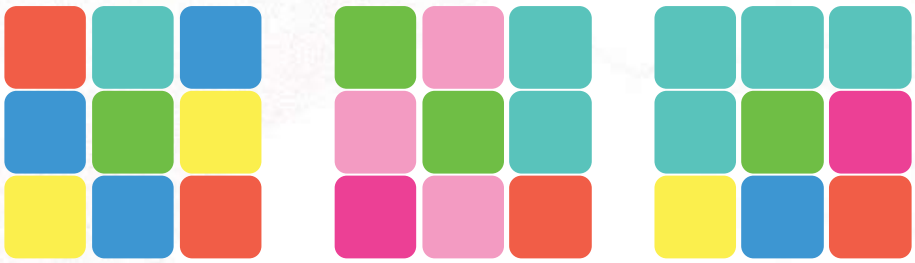
Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf
bioRxiv 2021.12.16.473007; doi: https://doi.org/10.1101/2021.12.16.473007
https://anndata.readthedocs.io/en/latest/index.html

16

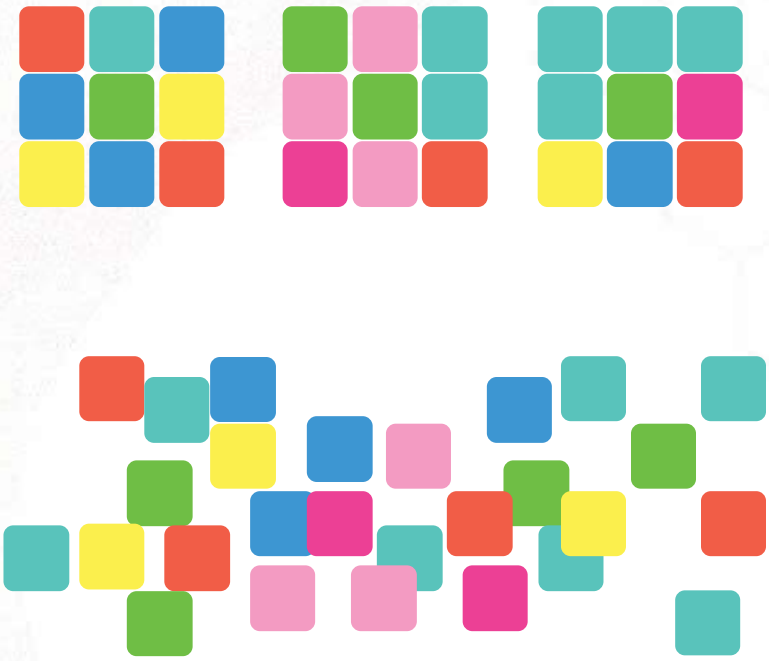
Sparsity, dropouts in single-cell data



Example: 3 different cell types



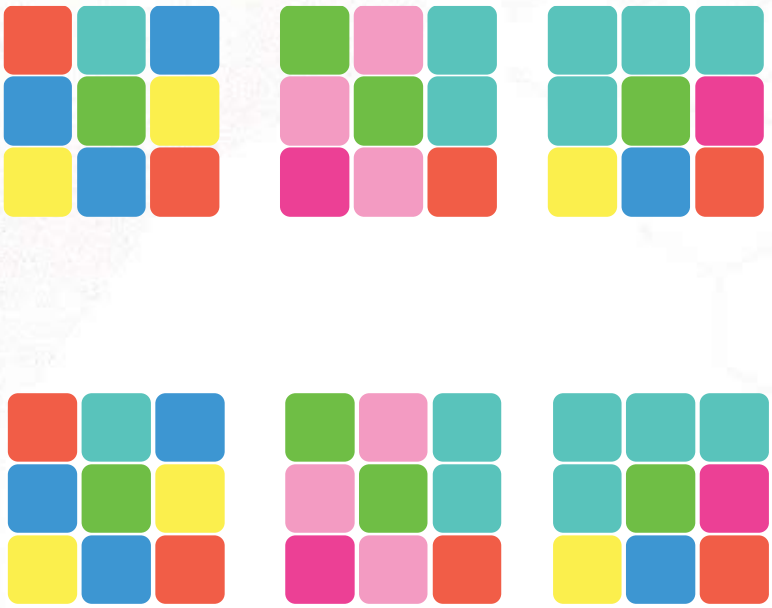
Example: 3 different cell types



bulk RNA-seq

19

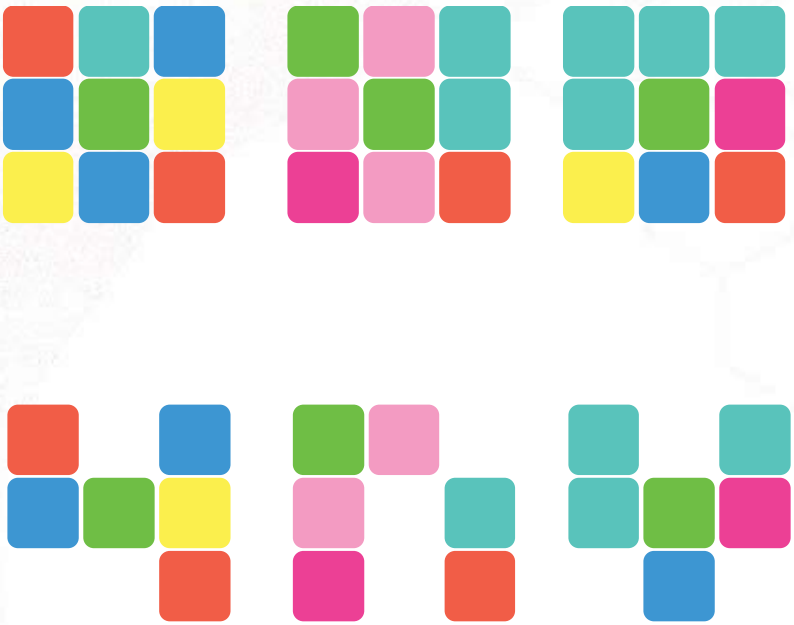
Example: 3 different cell types



ideal world single cell data

20

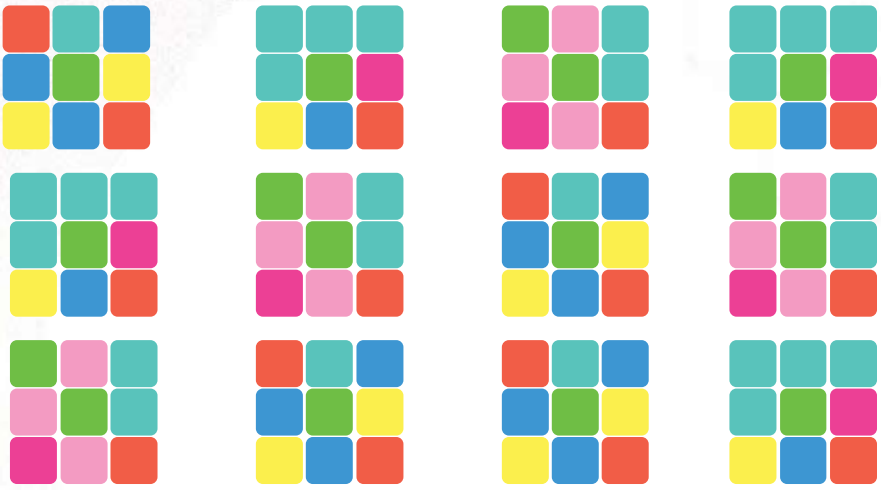
Dropouts in single-cell data



real world single cell data

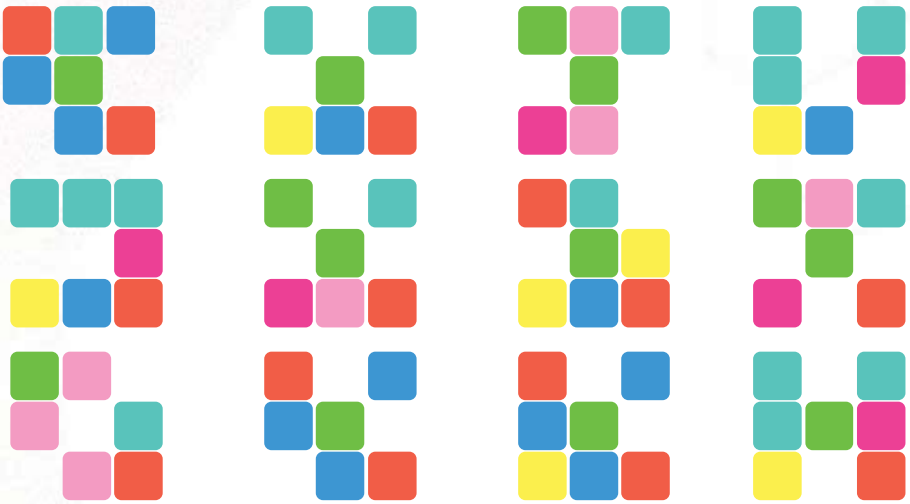
Profiling same cells for multiple times

multiple cells sequenced



Simulating random dropouts

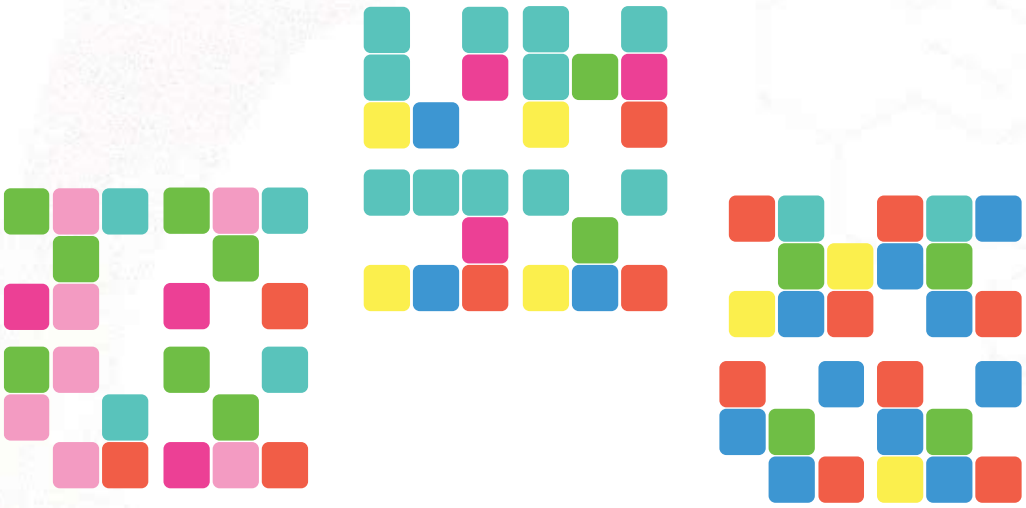
assuming low mRNA capture rate



23

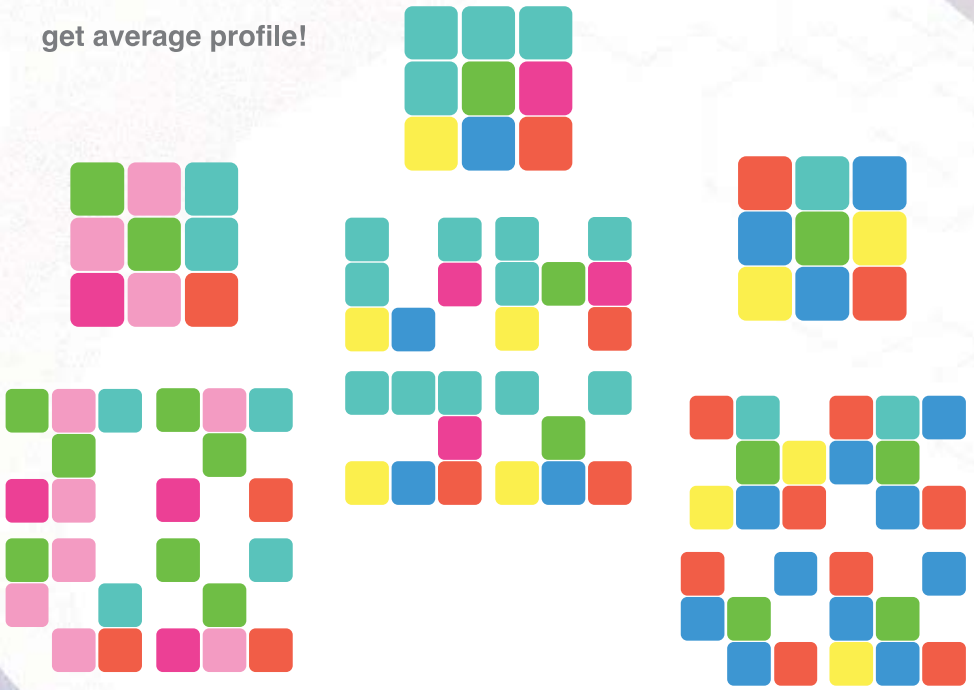
Clustering keeping the resolution

compare to each other & create clusters



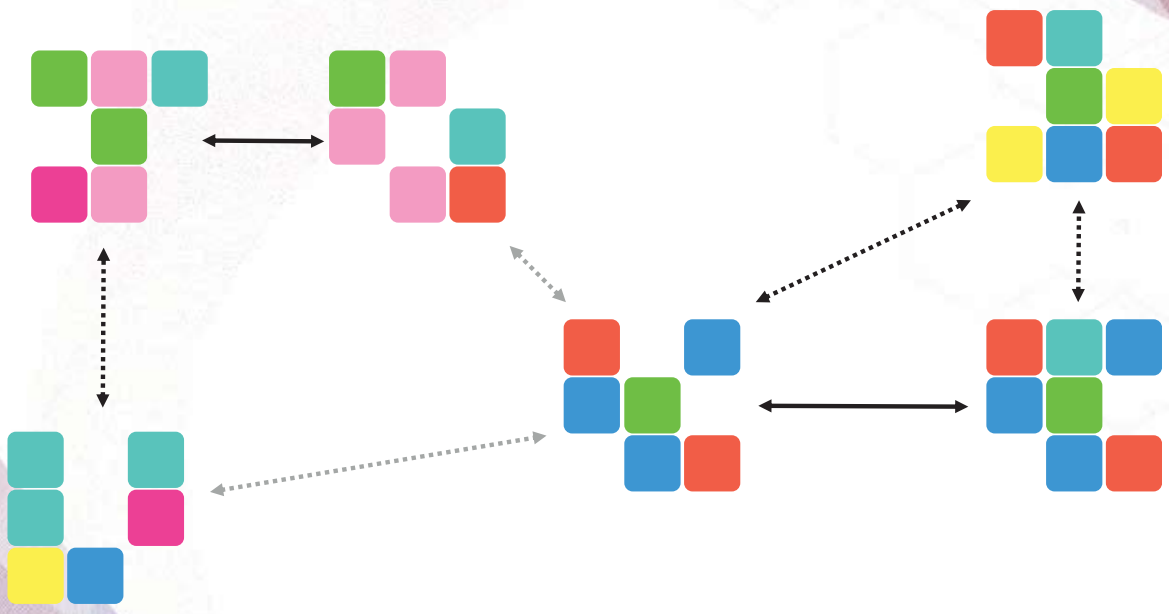
24

Power of taking average!



25

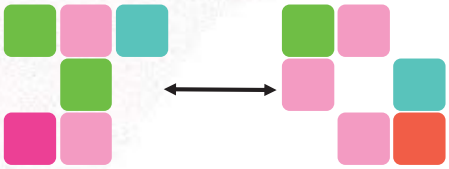
It's all about finding neighbors

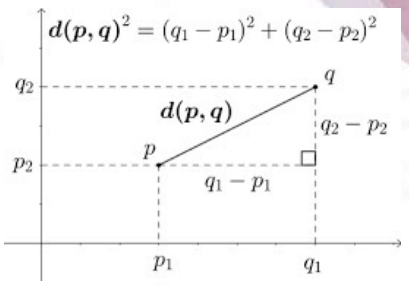


26

How to find neighbors?

How to measure distance between cells?



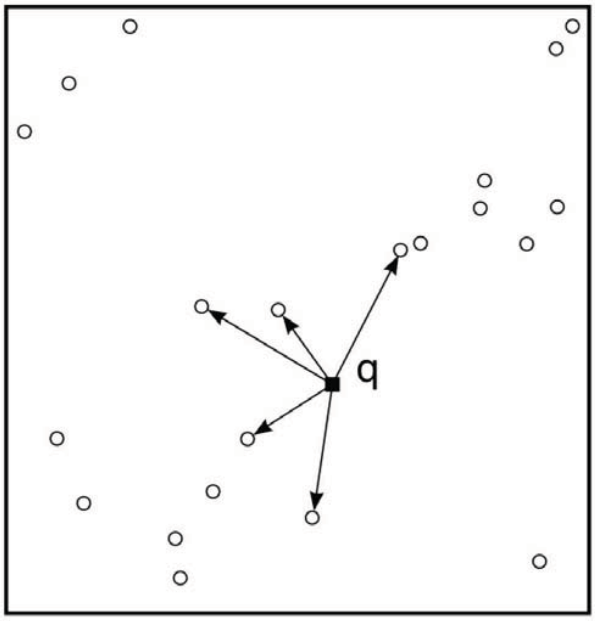


2	1	1
1	1	0
2	3	1
0	1	1
1	0	1

27

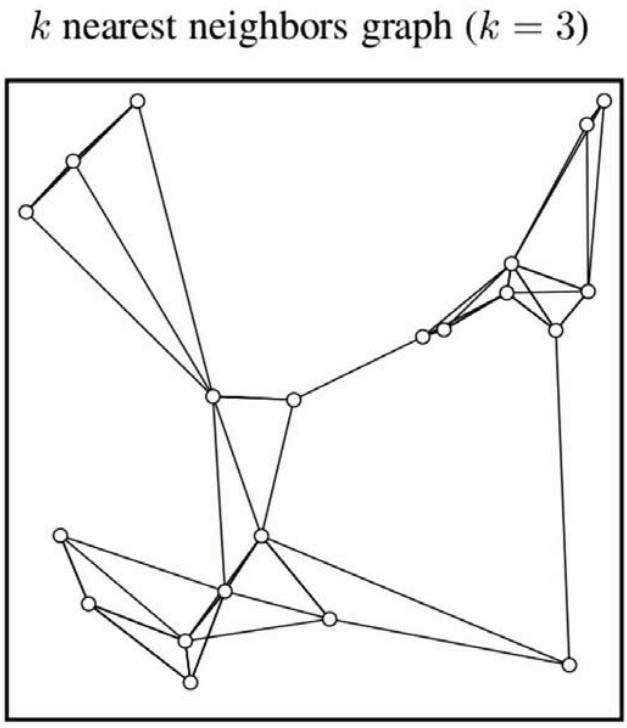
Finding k-nearest neighbors

k-nearest neighbors, $k = 5$



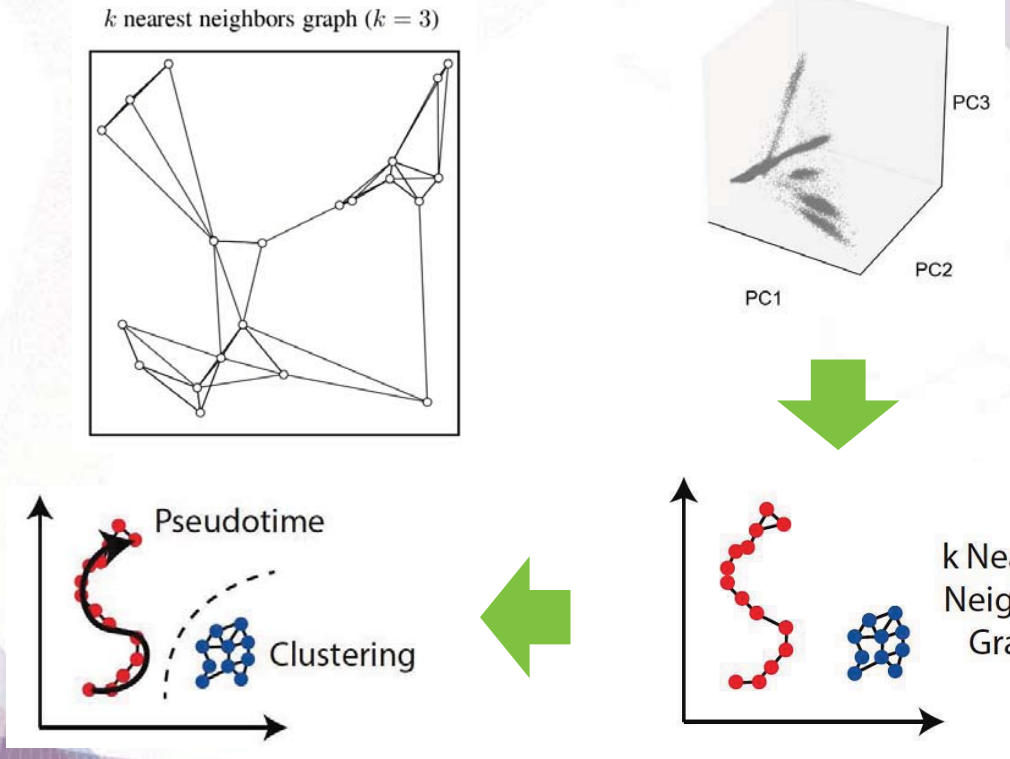
28

Finding k-nearest neighbors



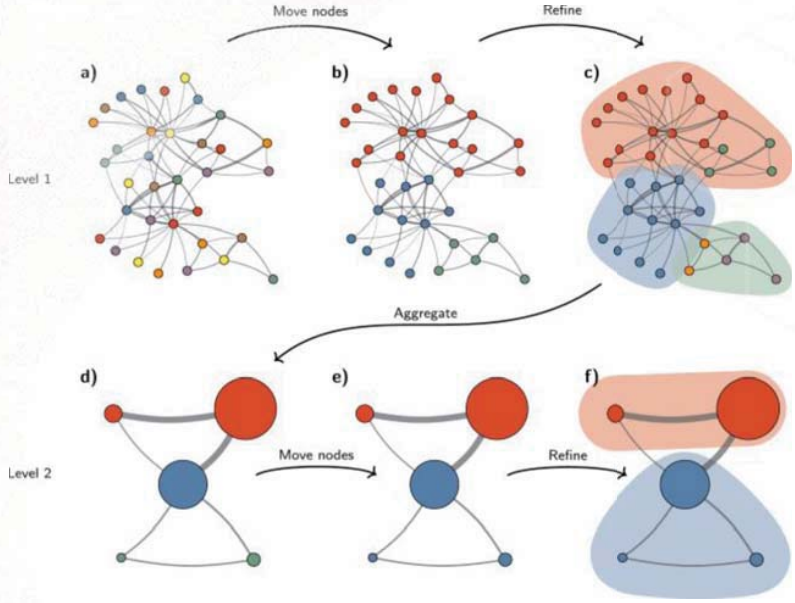
29

Power of neighbor graph in single-cell analysis

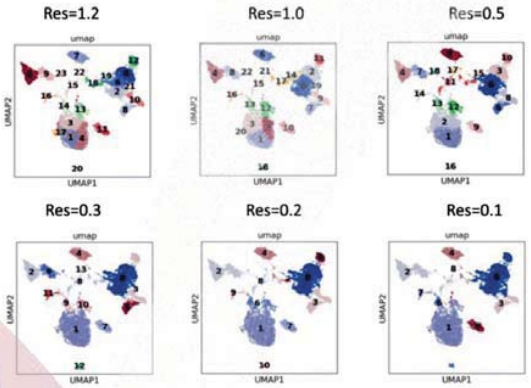


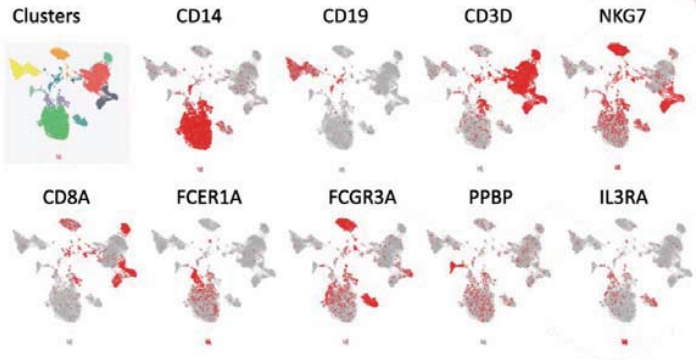
30

Graph based clustering



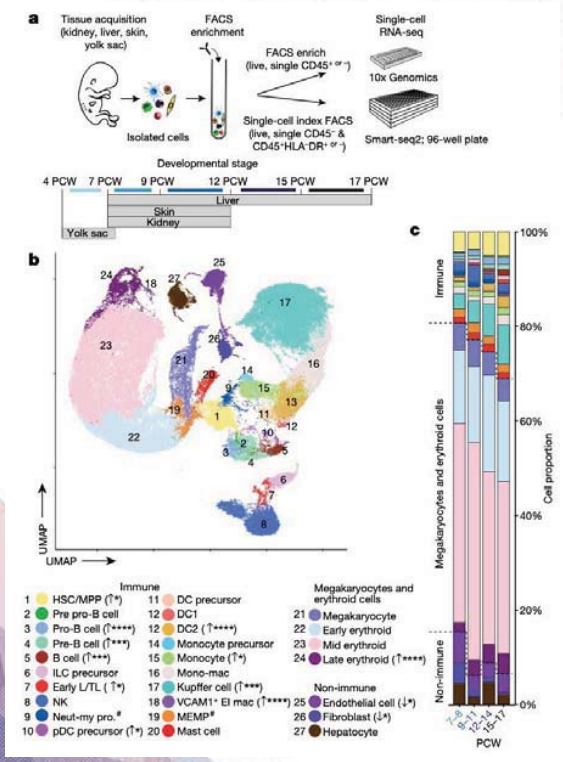
Cell type annotation





Annotating cell types

Overview of single-cell analysis



Popescu DM, Botting RA, Stephenson E, et al. Decoding human fetal liver haematopoiesis. *Nature*. 2019 33

Overview of single-cell analysis

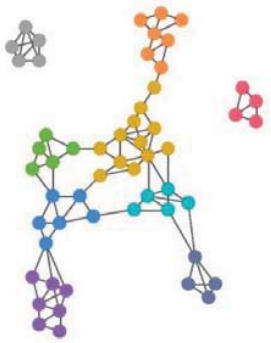
Gene expression

↓

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. 34

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

2. Best practice in single-cell data analysis

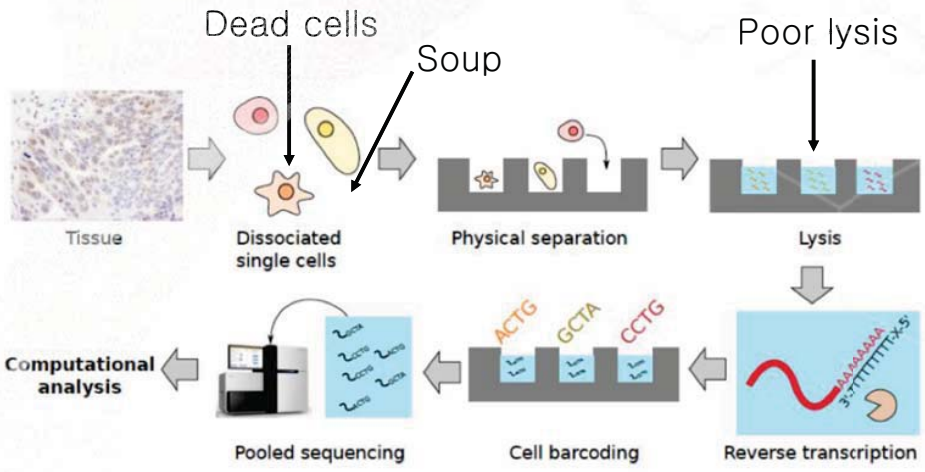
35

2-1. Sample QC

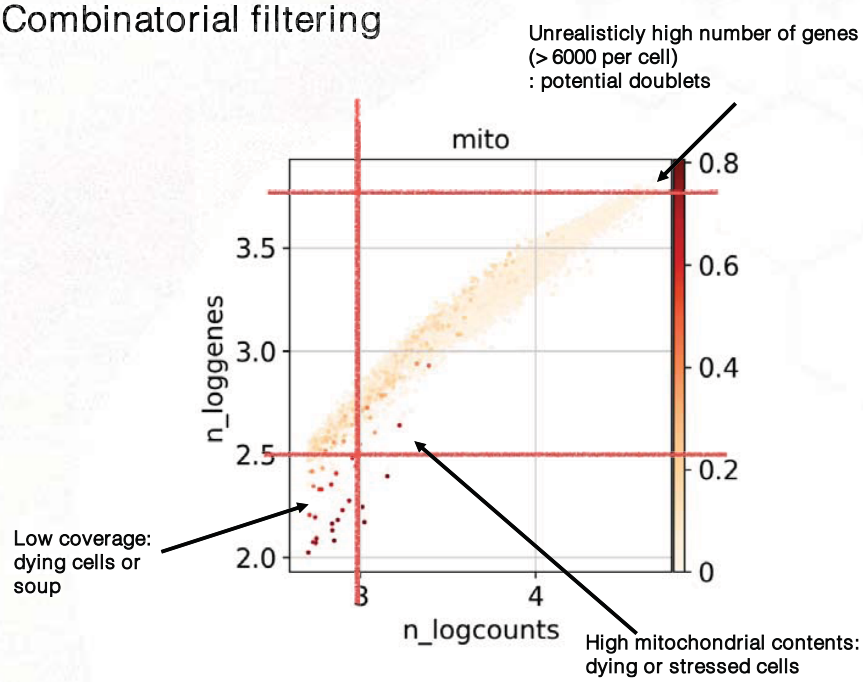
36

How can we define cells?

Sample QC: why?



Combinatorial filtering



Best practice

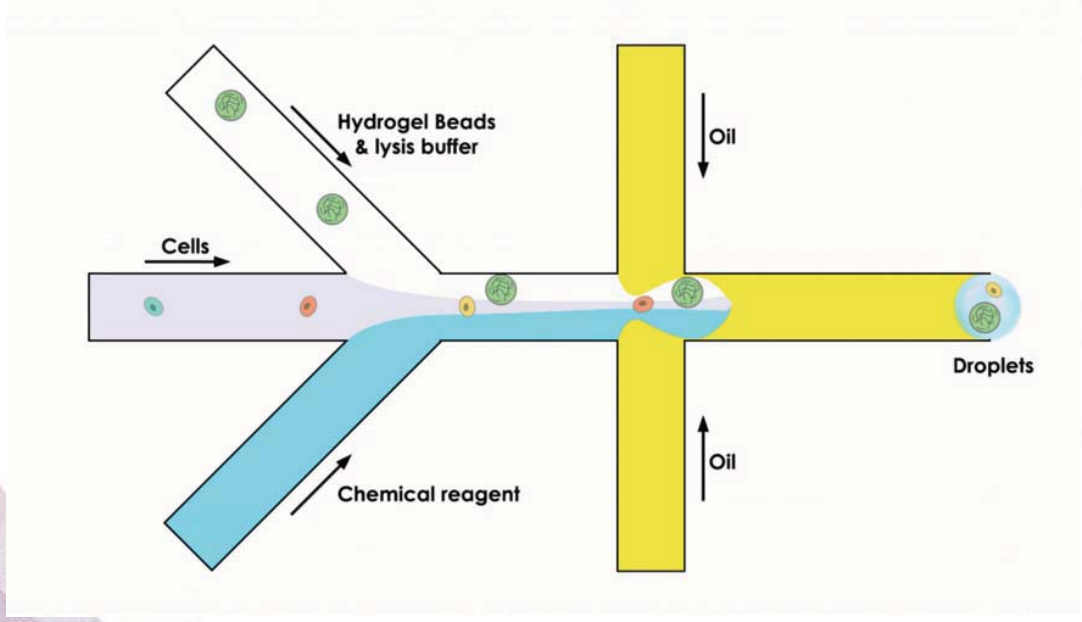
- Perform QC for individual sample
- Draw UMAP for individual sample

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

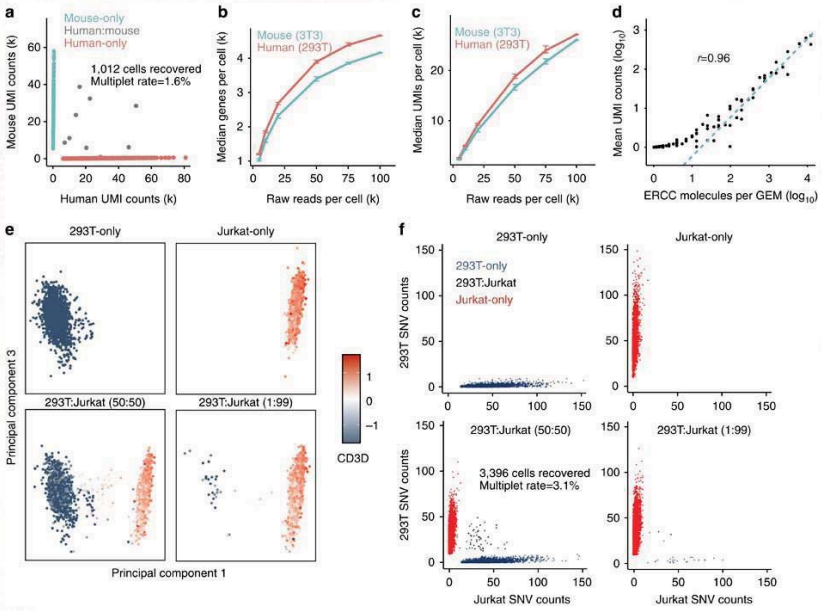
- Try to find best universal cutoff
- You might need to adjust cutoff for some low-quality samples (or simply discard them)

Problem 1: Doublets

Doublets expected from droplet based methods



Doublets expected from droplet based methods

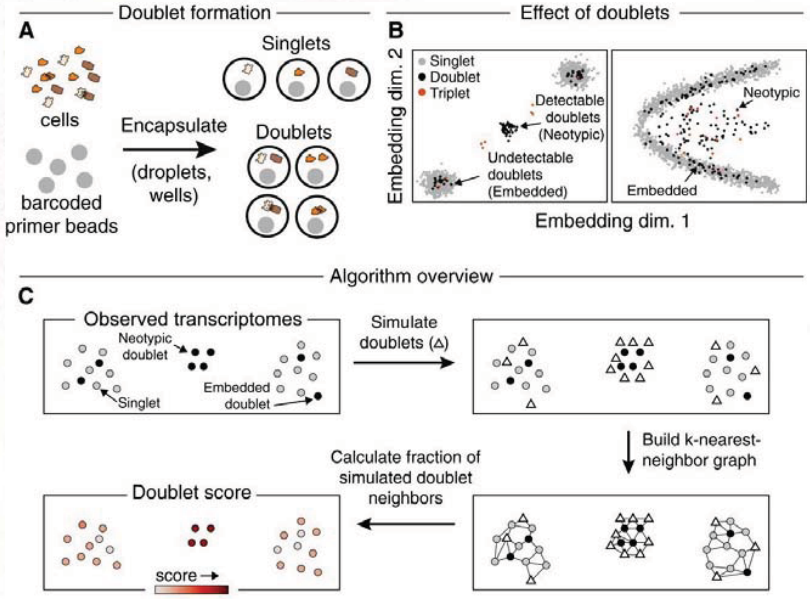


https://www.nature.com/articles/ncomms14049/figures/2

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

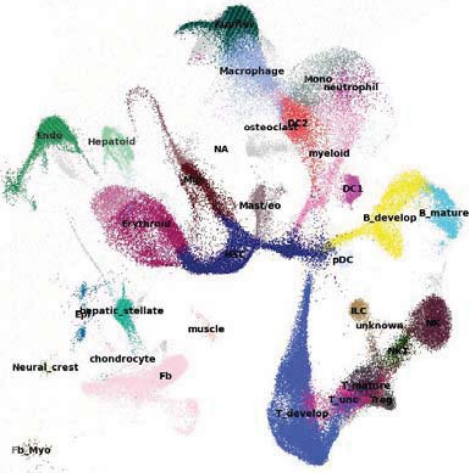
Computationally predicting doublets

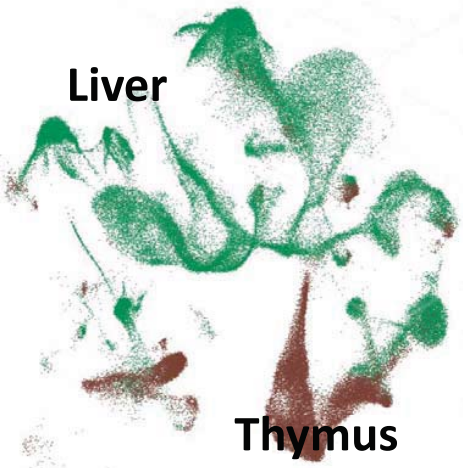


https://www.sciencedirect.com/science/article/pii/S2405471218304745

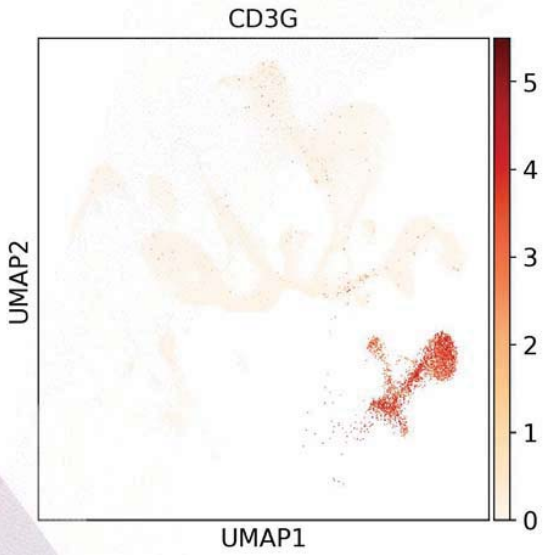
Problem 2: Contaminating reads from ambient RNAs

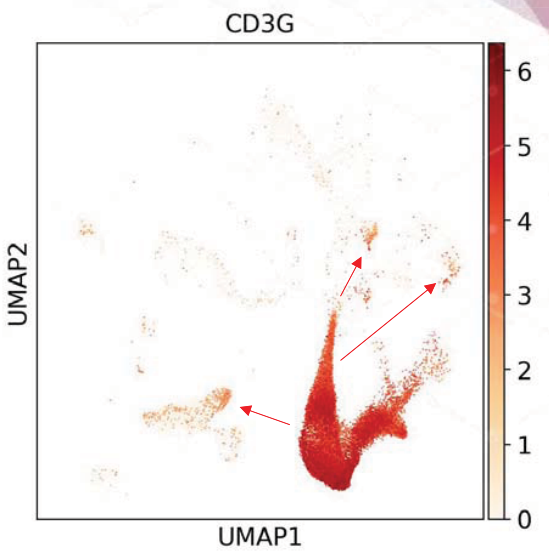
Example for the ambient RNA contamination



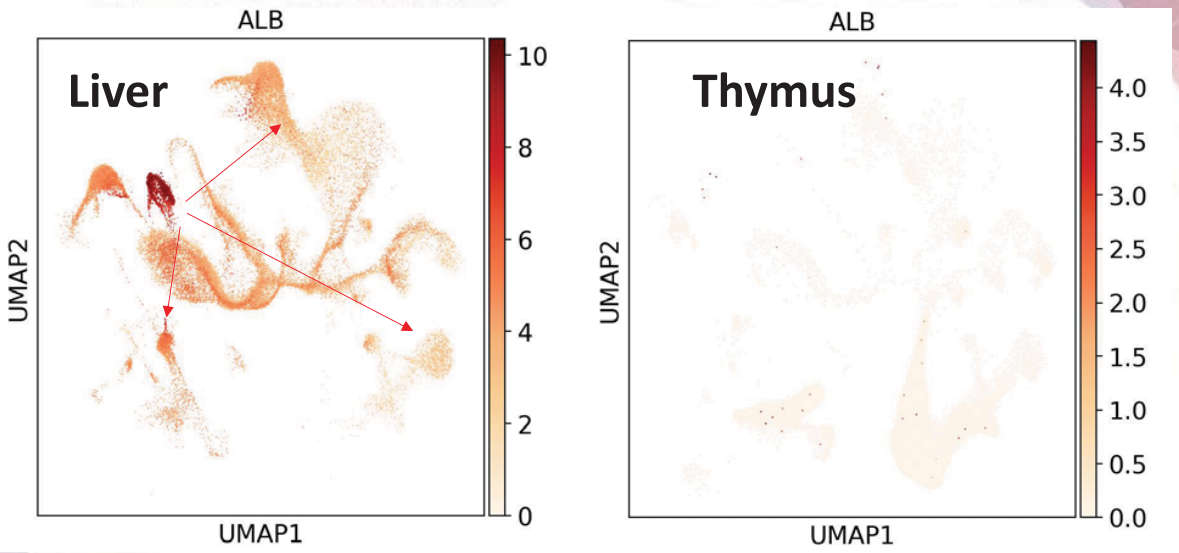


Example for the ambient RNA contamination

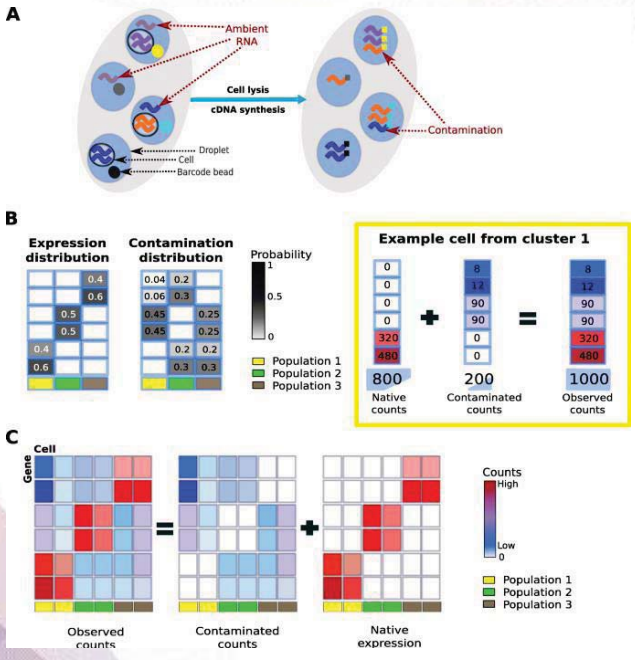




Example for the ambient RNA contamination



DecontX

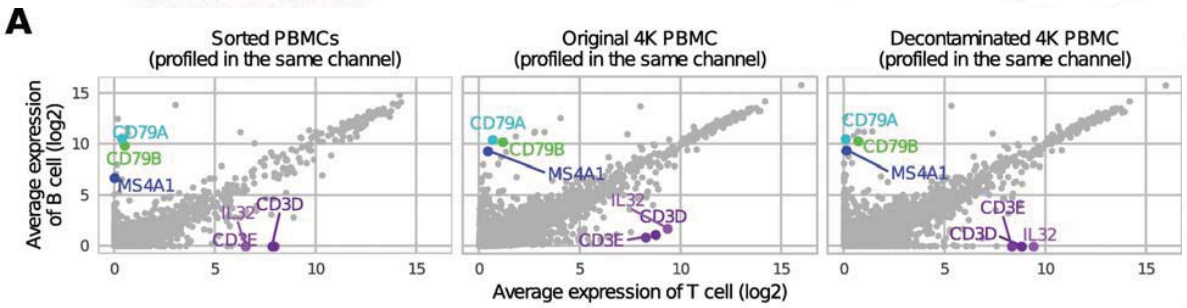


Decontamination of ambient RNA in single-cell RNA-seq with DecontX

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

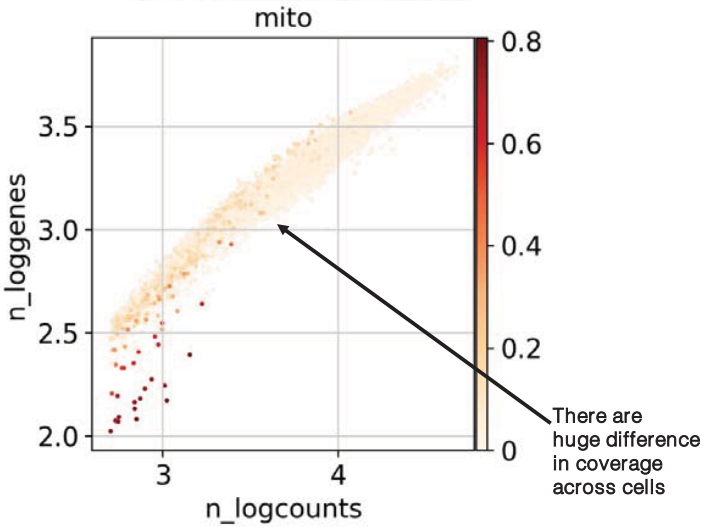
Genome Biology **21**, Article number: 57 (2020) | Cite this article

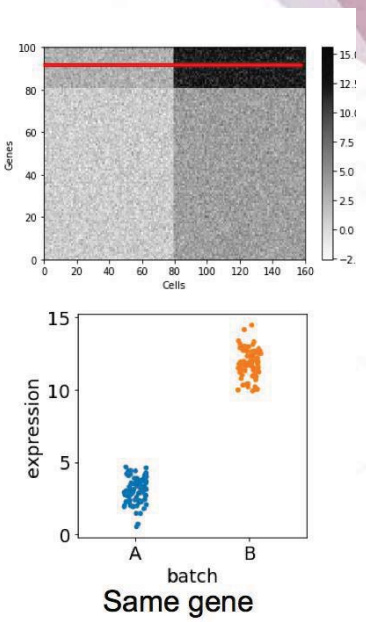
Cell Hashing allows detection of doublets



2-2. Preprocessing (normalization, scaling)

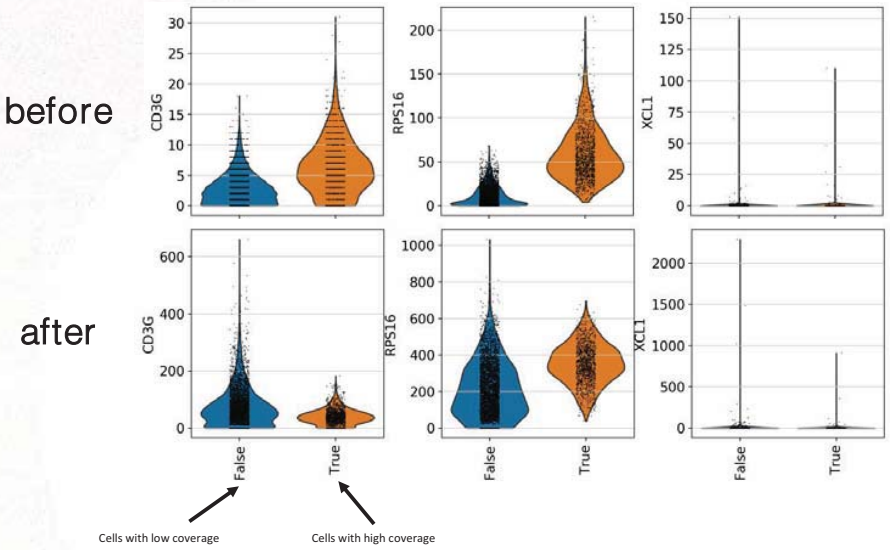
Why do we need normalisation?





Normalisation method

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

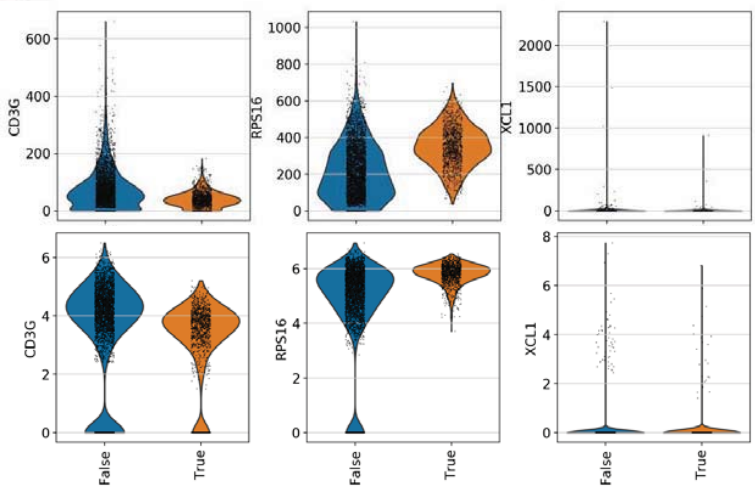


Log-transformation

$$Xlog = log(Xnorm+1)$$

before

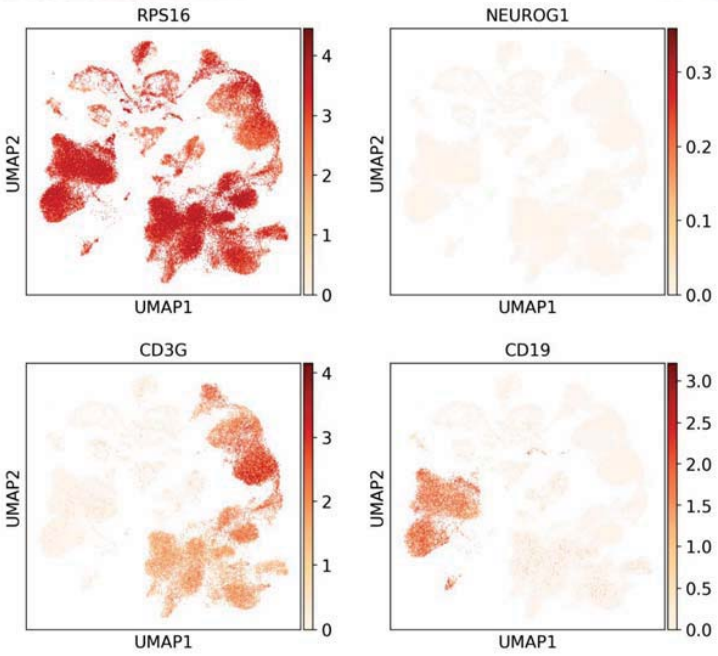
after



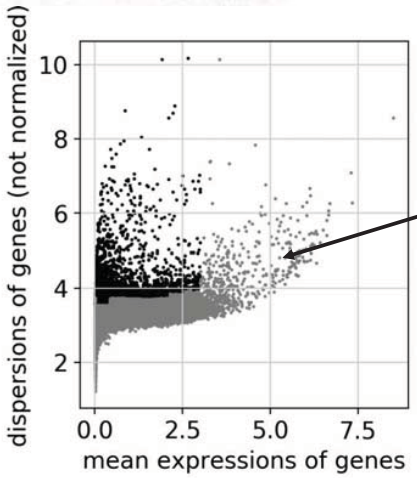
Selecting highly variable genes

Not informative

Highly informative

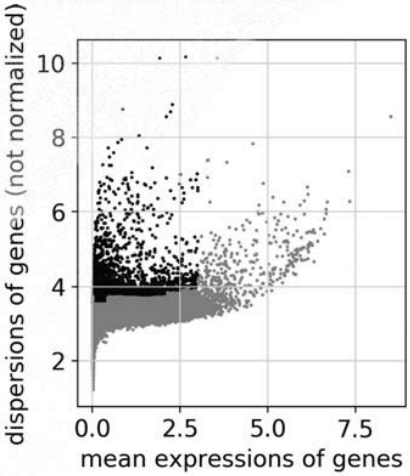


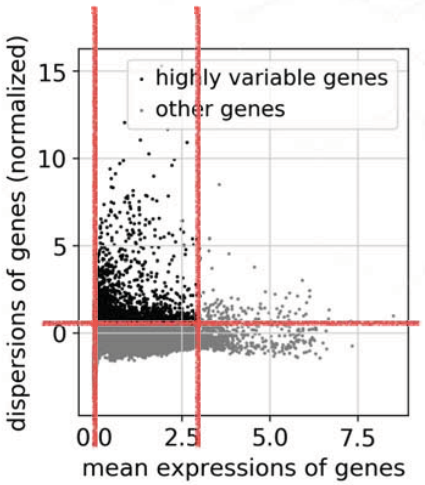
Selecting highly variable genes



Higher the expression level, also higher the dispersion

Selecting highly variable genes





After normalisation, we can apply flat cutoff!

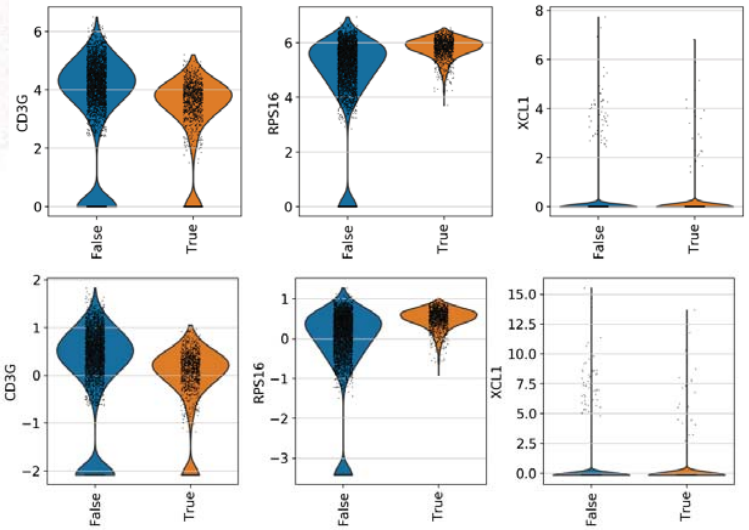
Scaling

centralise gene expression with zero mean and unit variance

“Making all genes equally important”

before

after

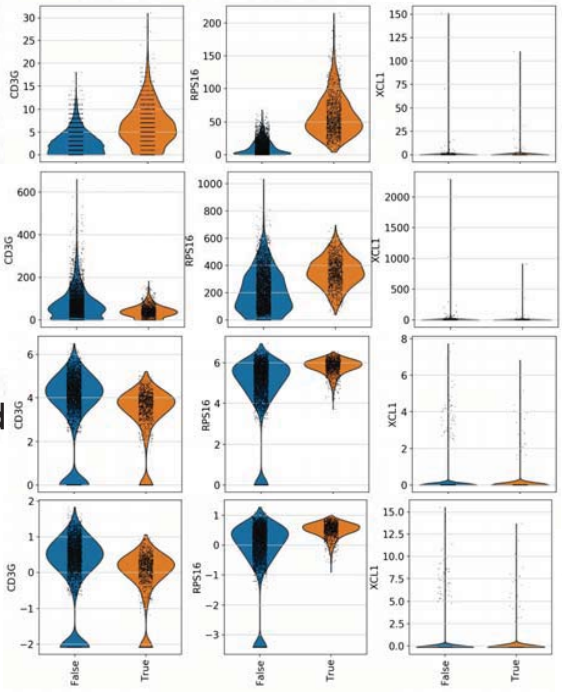


Raw count — Integer between 0-10000

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

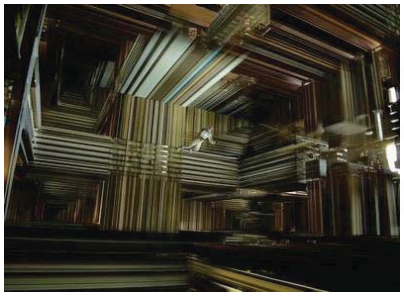
Single-cell data is high dimensional!

Cells in high-dimensional space (> 30000 genes)

Genes (30,000)

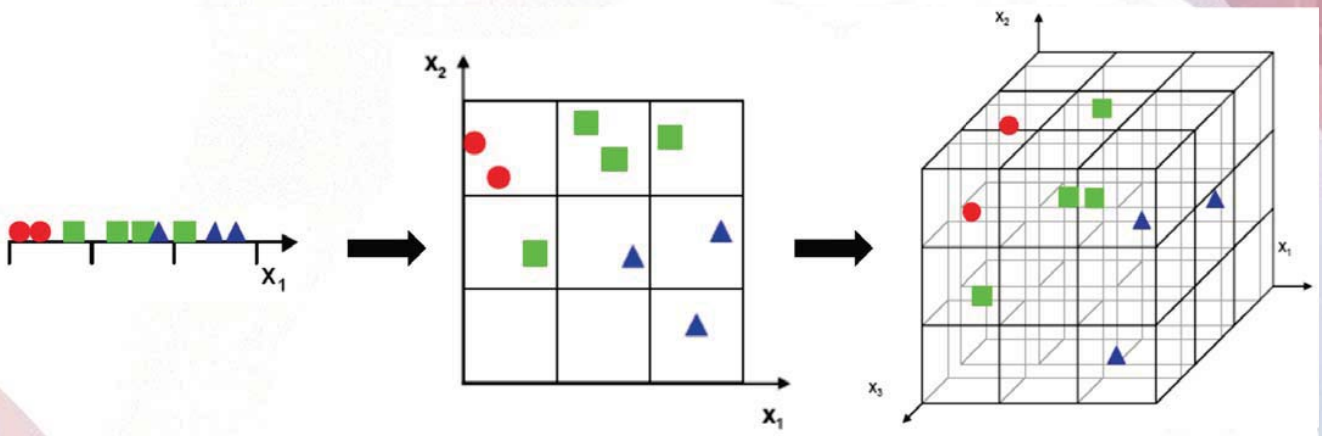
Cells (up to 1,000,000)

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

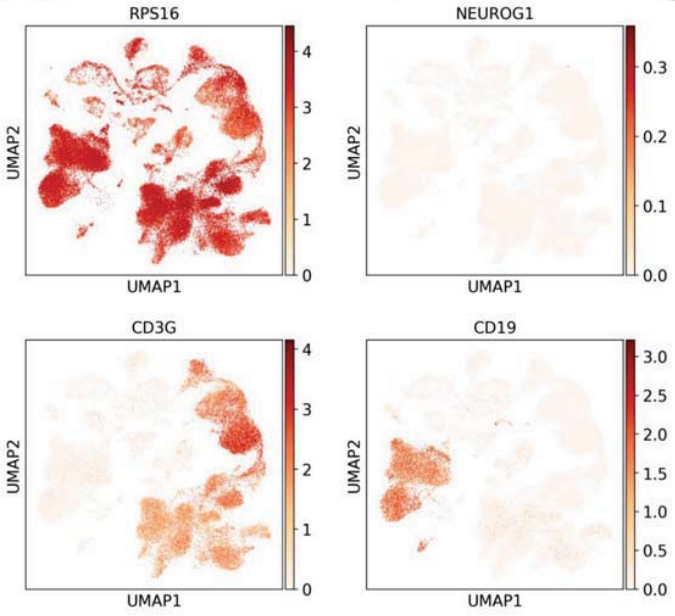
67

Feature selection

Genes (30,000)

Cells (up to 1,000,000)

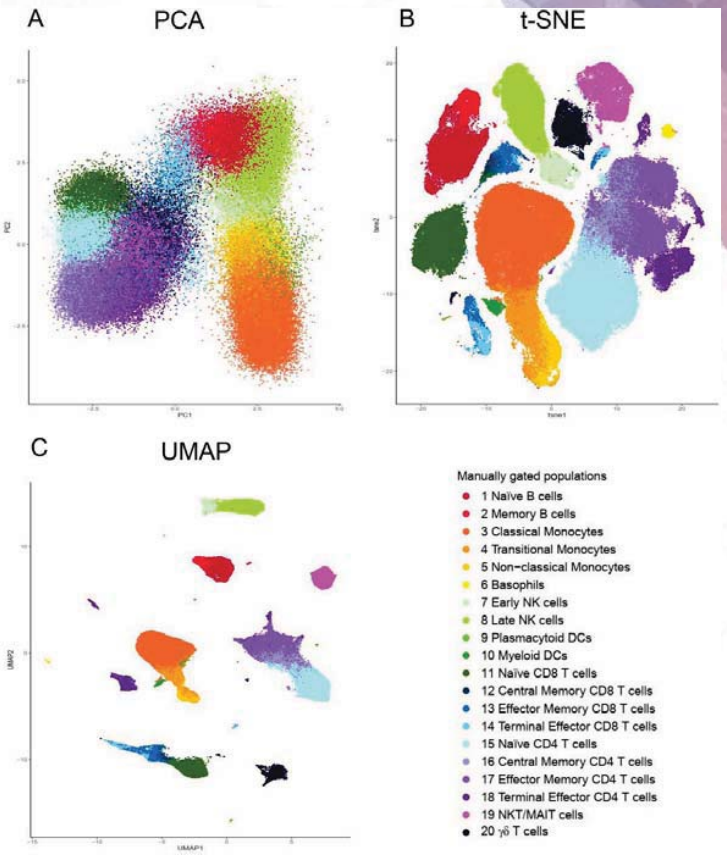
0	4	0	3	…	9	0	5	2
0	6	1	0	…	7	3	0	4
2	5	0	5	…	6	0	3	3
:	:	:	:		:	:	:	:
0	3	0	0	…	8	2	1	1
1	9	2	2	…	5	1	4	5
0	0	1	0	…	0	2	1	0



68

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.

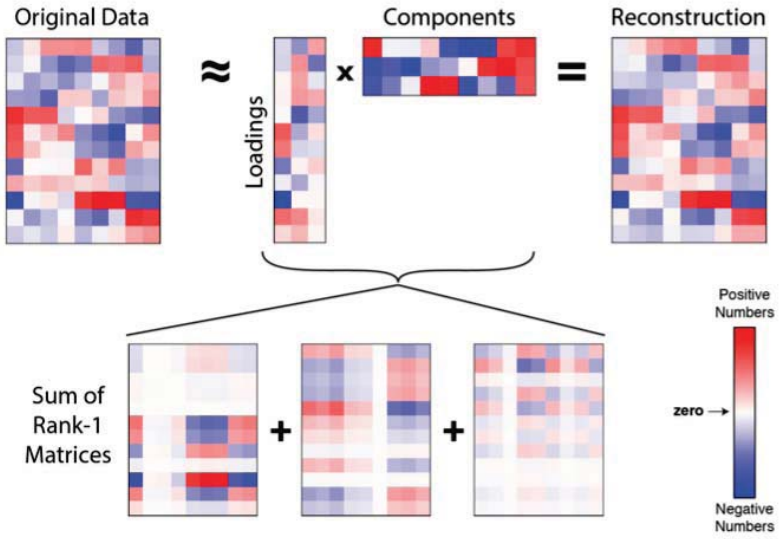
69

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

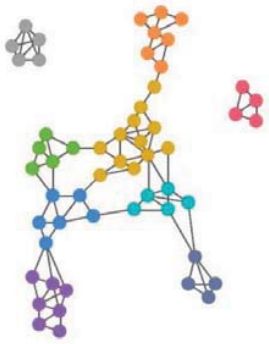
Gene expression

↓

Dimension reduction (e.g. PCA)

↓

KNN-graph



→ UMAP

→ Force-directed graph

→ Diffusion map

Graph-embedding

↓

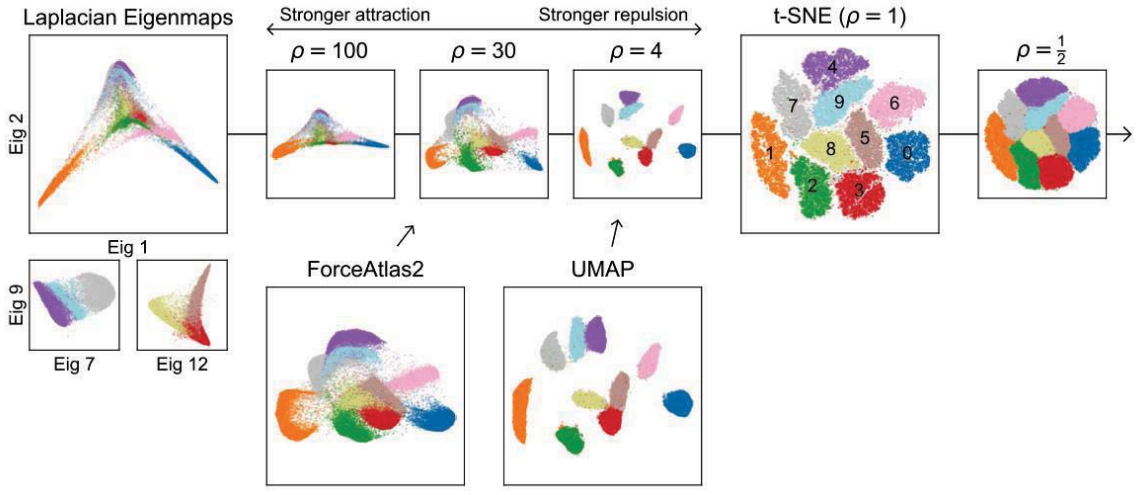
Clustering, annotation, pseudotime….

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.

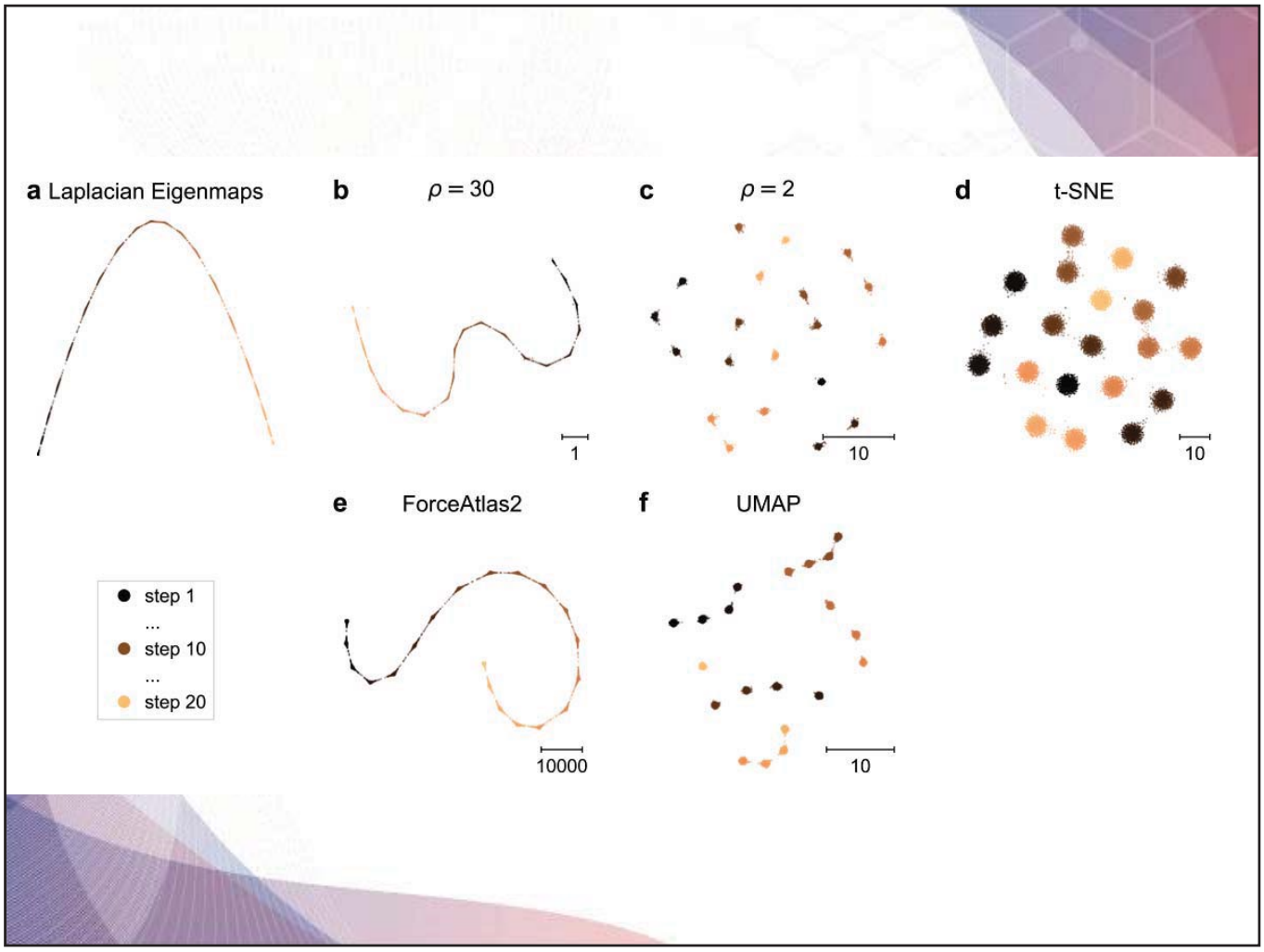
71

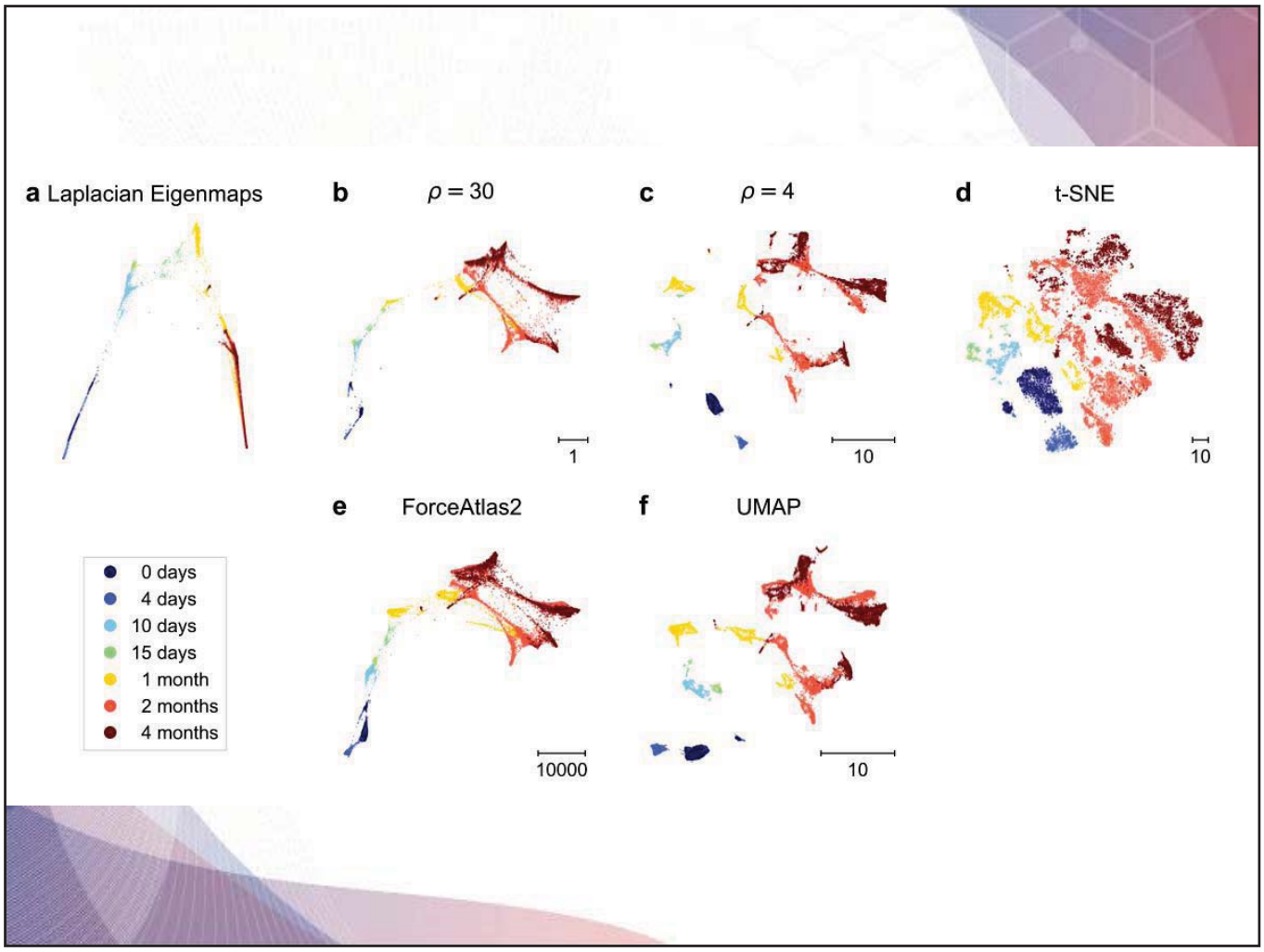
Many methods for graph embedding



Attraction & connectivity

Repulsion & separation


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

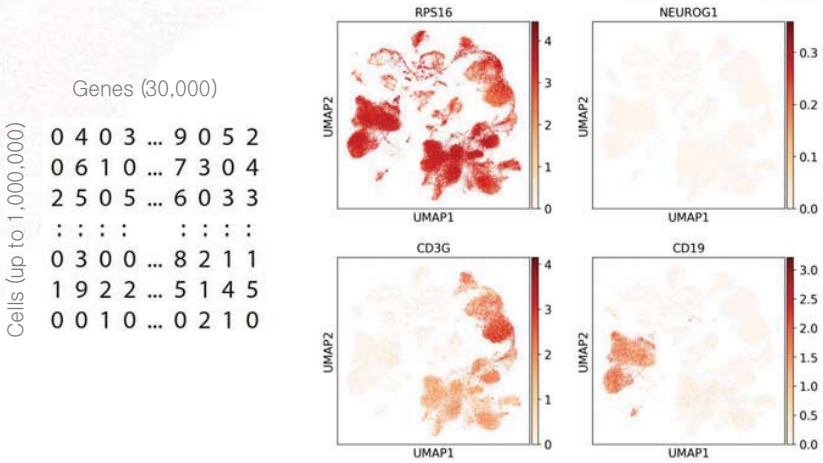

a Laplacian Eigenmaps
b ρ = 30
c ρ = 4
d t-SNE
e ForceAtlas2
f UMAP
0 days
4 days
10 days
15 days
1 month
2 months
4 months
1
10
10
10000
10

How to annotate cells?
How to define cluster resolution?

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

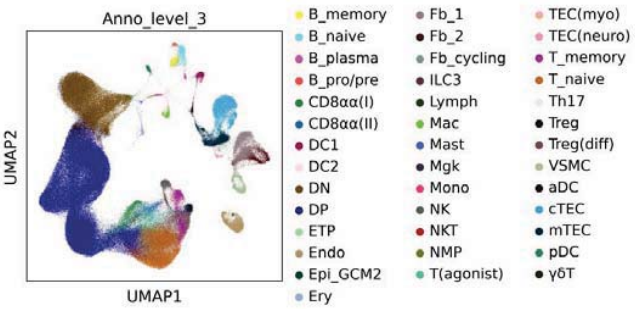
76

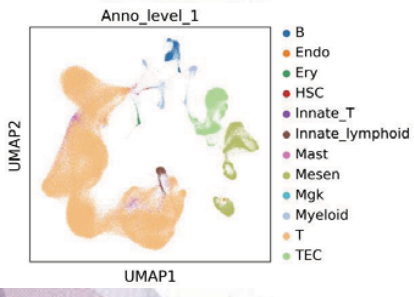
Choice of highly variable genes / PCA projection / batch correction methods… all affect clustering outcome

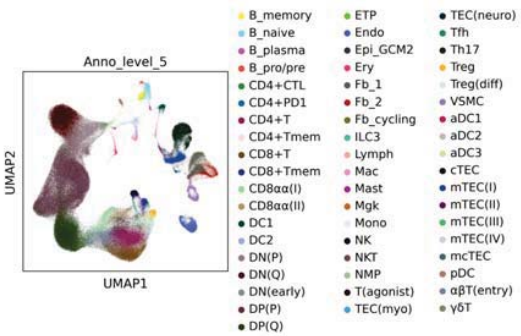


highly variable gene selection

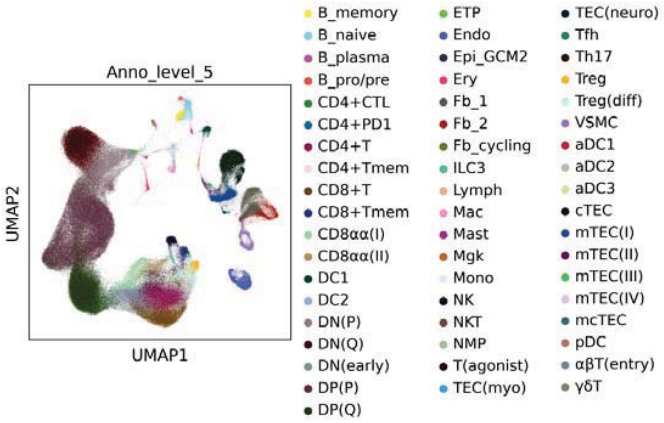
Hierarchy in cell annotation

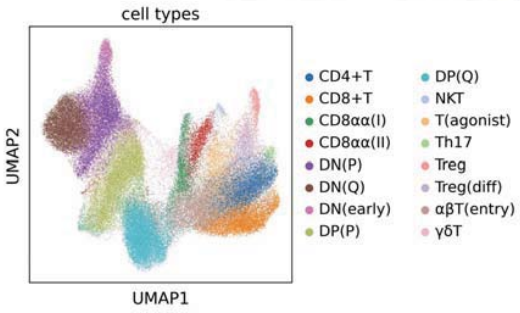






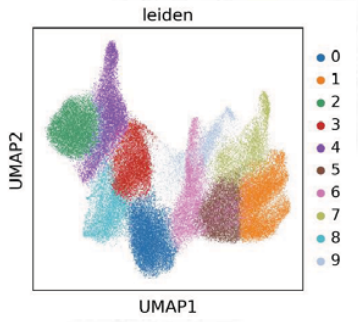
Zooming into subset to get better resolution



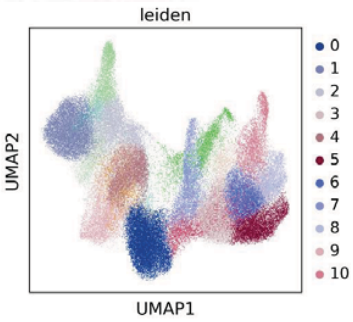


Subset, balancing, different genes, changed PCs

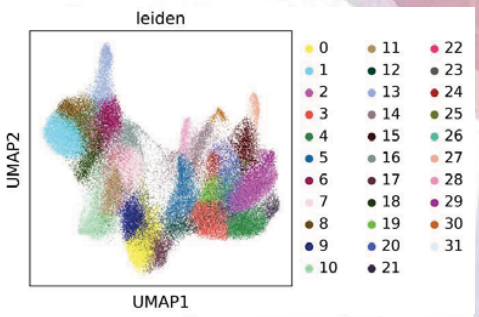
Resolution = 1

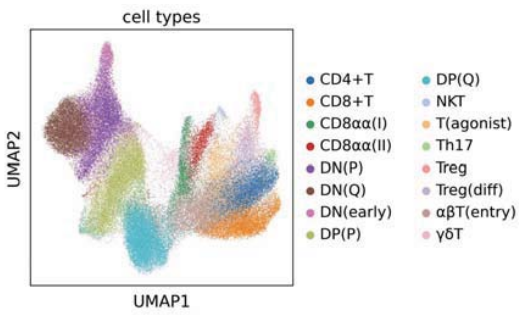


Resolution = 2

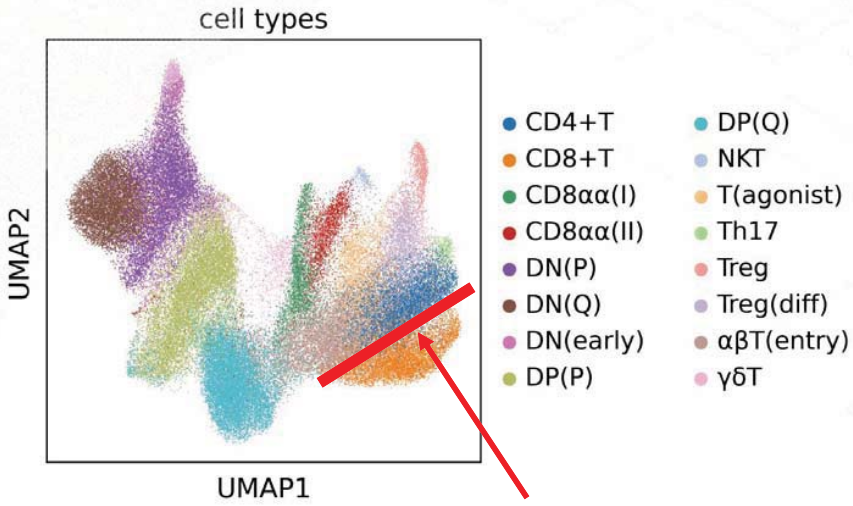


Resolution = 3





Defining markers



Focusing on the borders
Any binary classifier?

DEG analysis?

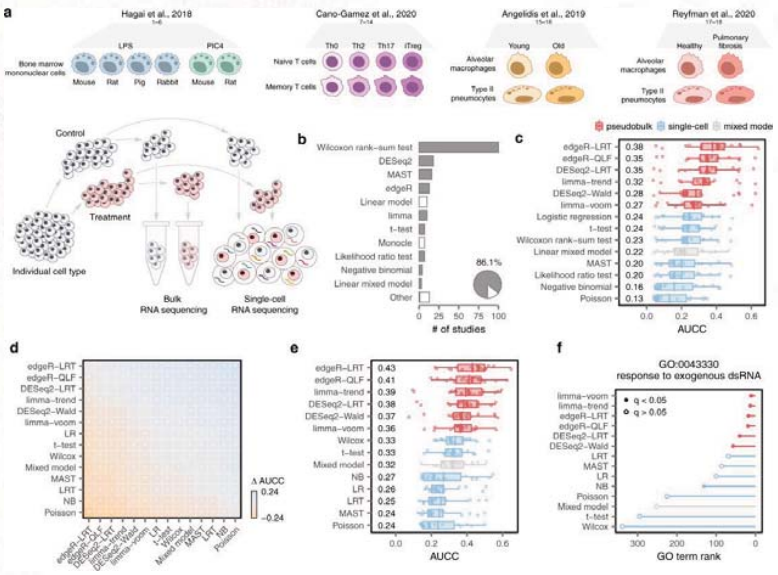
Article | Open Access | Published: 28 September 2021

Confronting false discoveries in single-cell differential expression

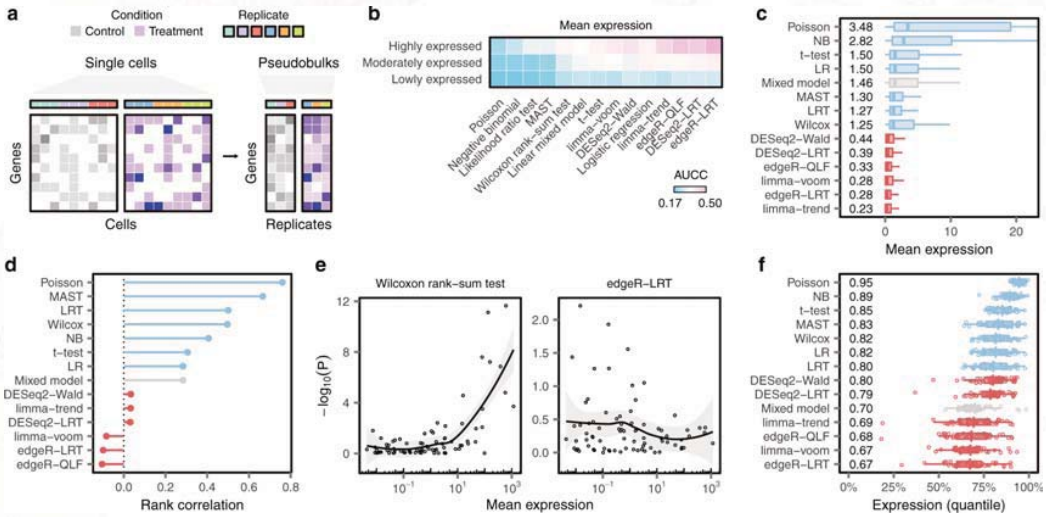
Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider ✉ & Grégoire Courtine ✉

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

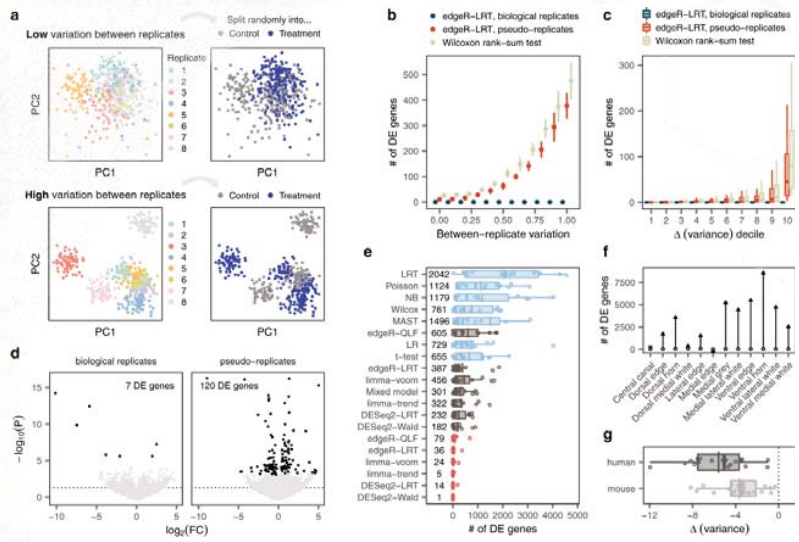
Pseudo-bulk methods outperform generic and specialized single-cell DE methods



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

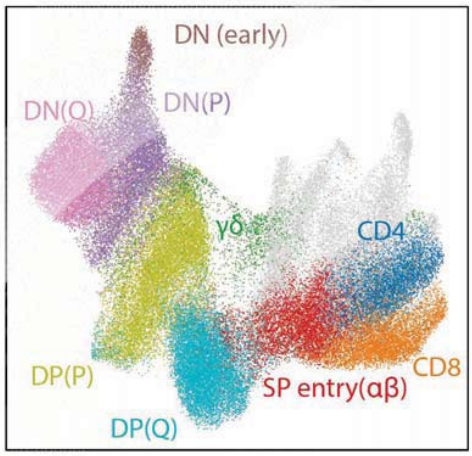


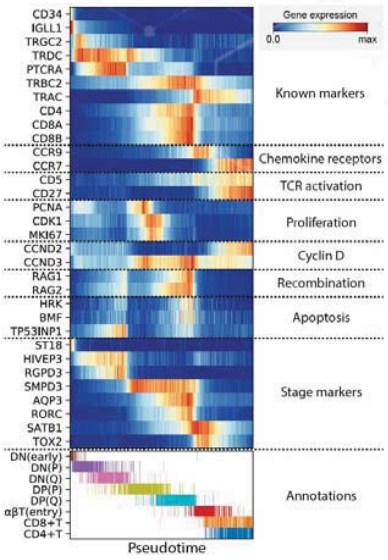
Single–cell DE methods often leads to false discoveries of DEGs

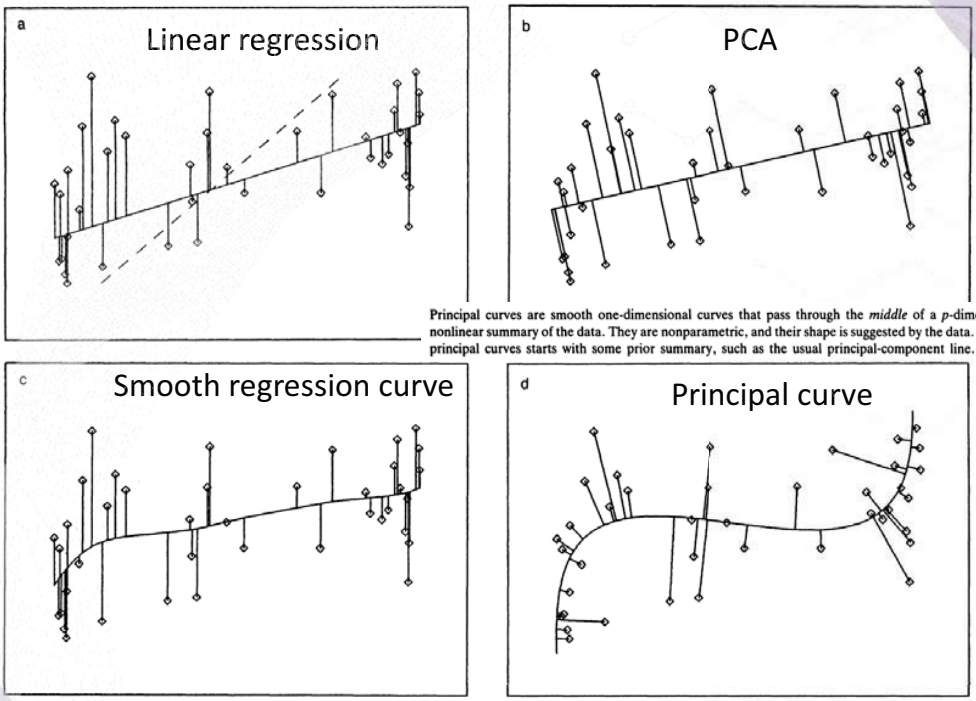


2-5. Trajectory inference

Modelling the cell differentiation trajectory



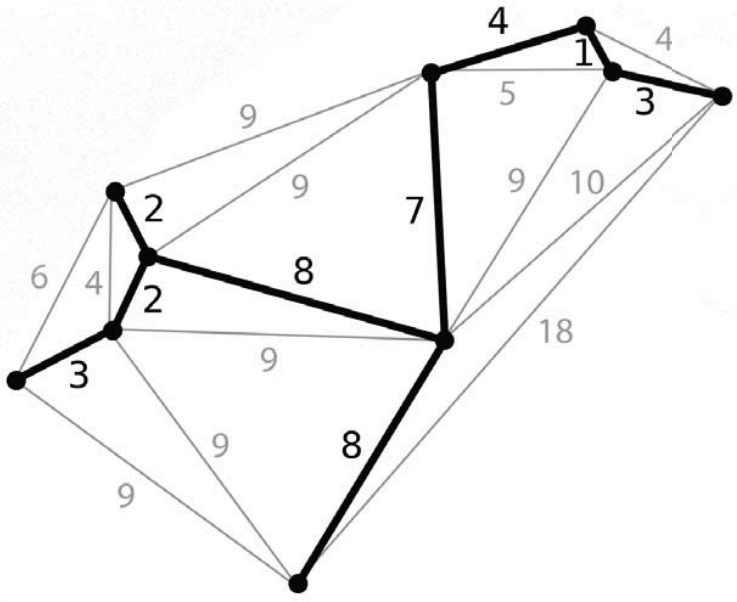




Principal curves are smooth one-dimensional curves that pass through the *middle* of a *p*-dimensional data set, providing a nonlinear summary of the data. They are nonparametric, and their shape is suggested by the data. The algorithm for constructing principal curves starts with some prior summary,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

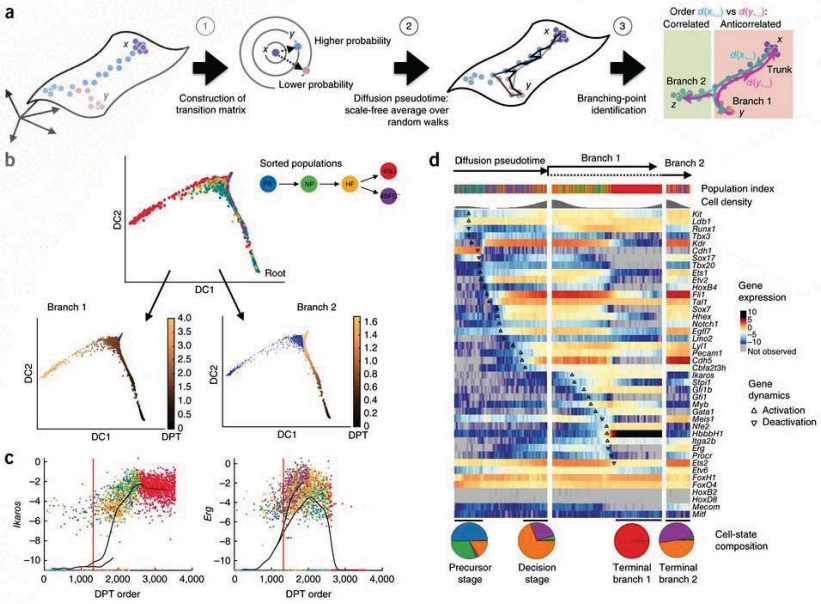
Minimum Spanning Tree



Published: 29 August 2016

Diffusion pseudotime robustly reconstructs lineage branching

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

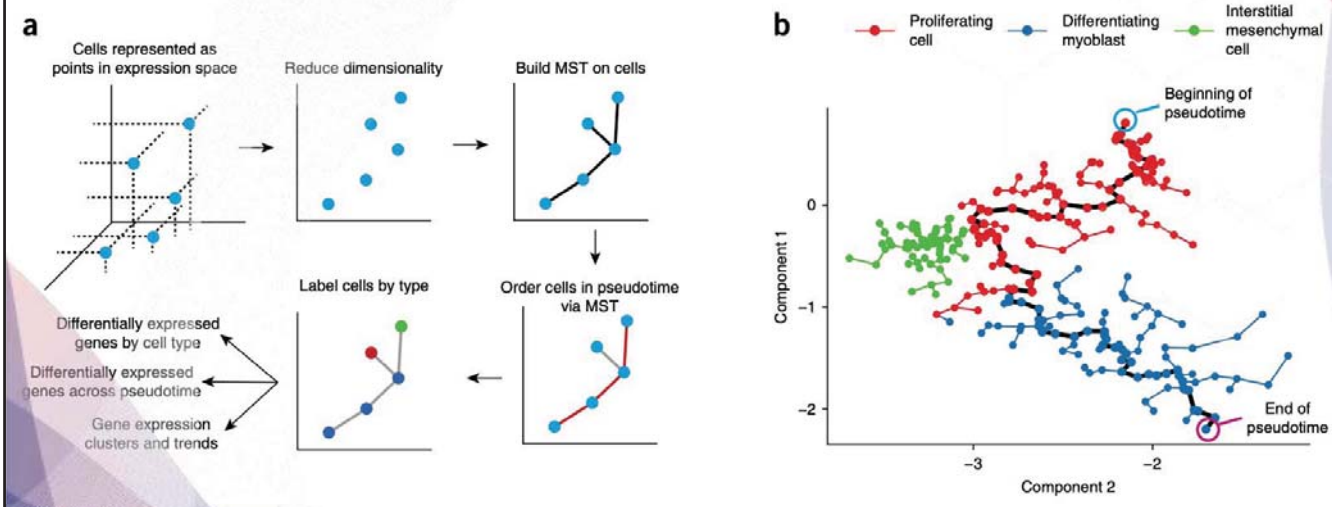


Monocle

nature biotechnology

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

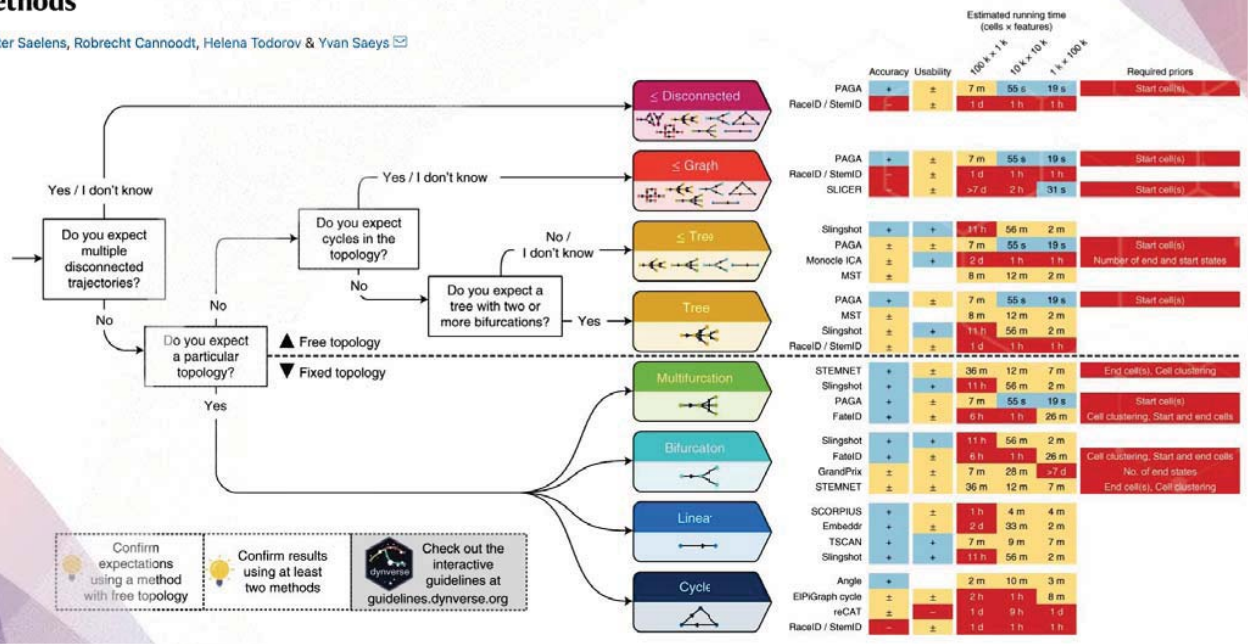
Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[5], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]



Article | Published: 01 April 2019

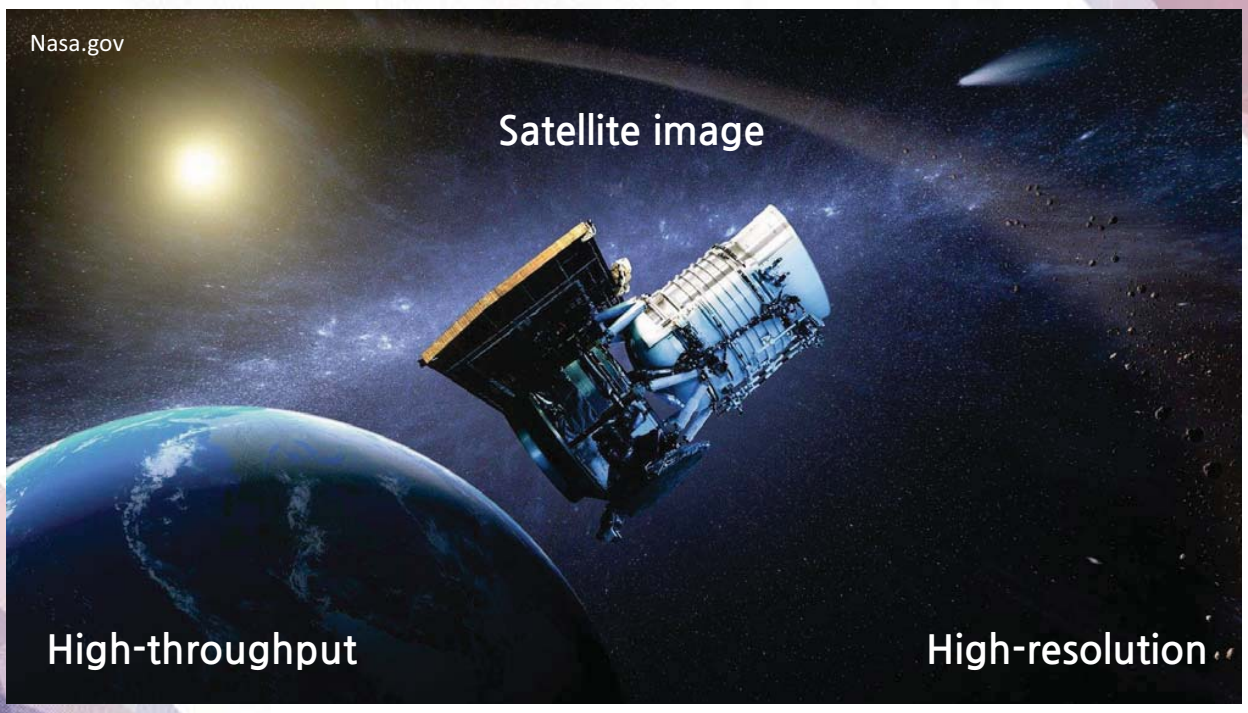
A comparison of single-cell trajectory inference methods

Wouter Saelens, Robrecht Cannoodt, Helena Todorov & Yvan Saeys



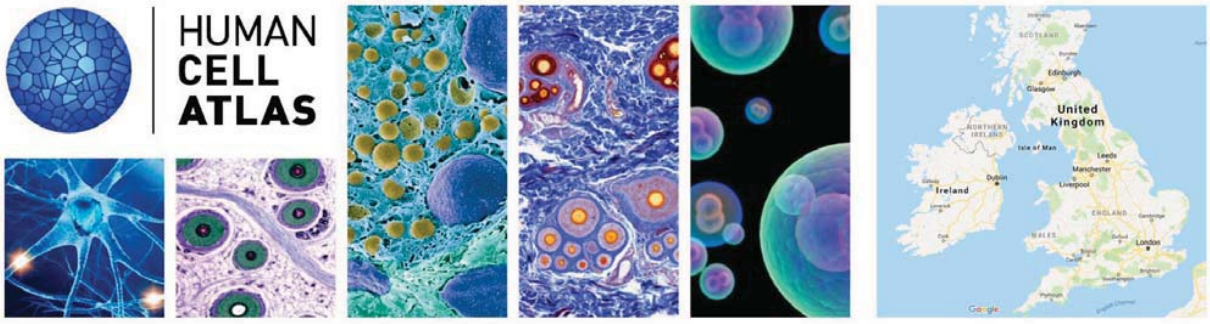
3. Public databases & Data integration

Analogy between singel-cell omics vs satellite images



Human cell atlas

Human cell atlas : “Google map” of human body



Cells & Genes

City & buildings

Human cell atlas: timeline and scope

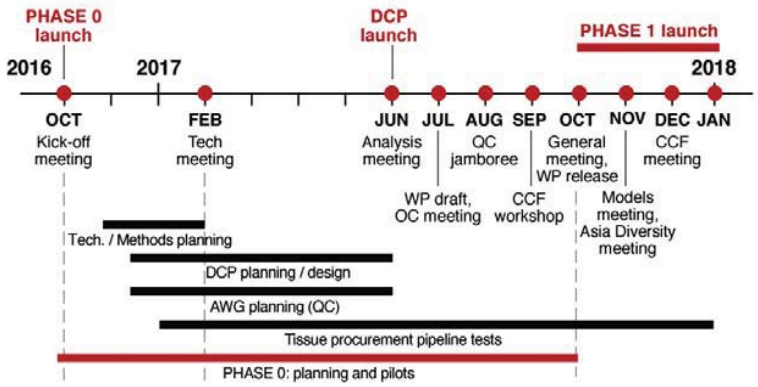
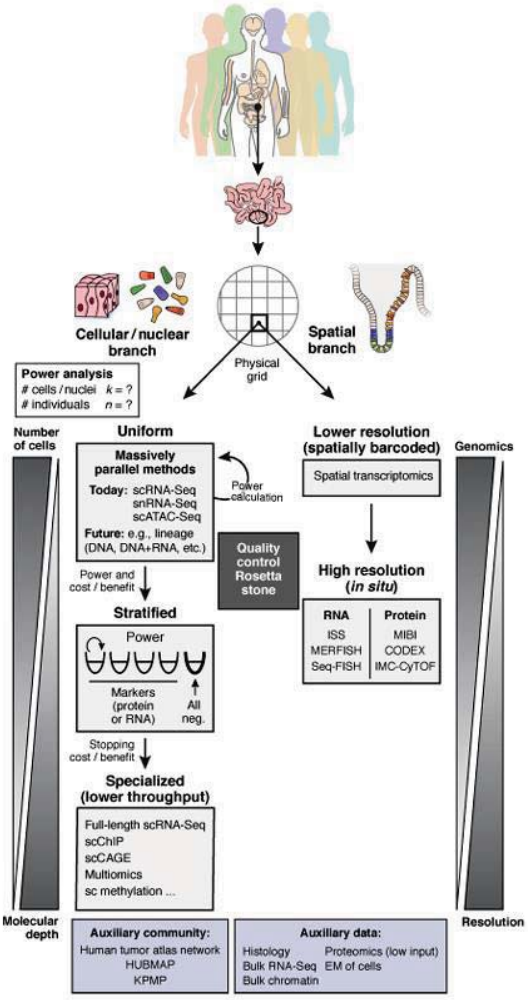
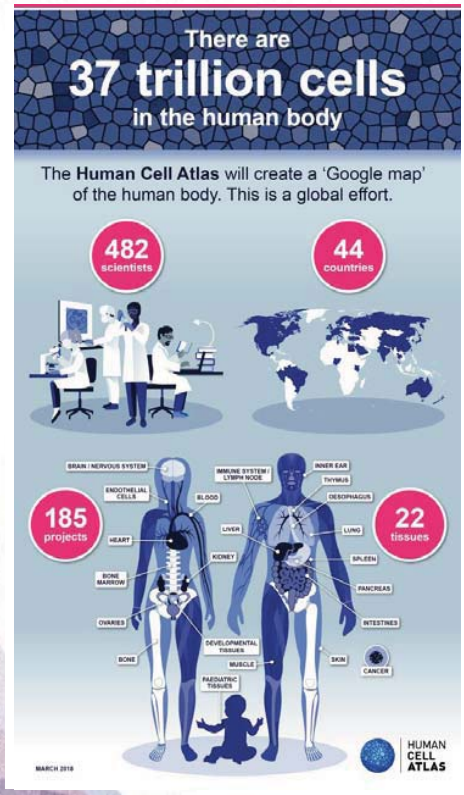
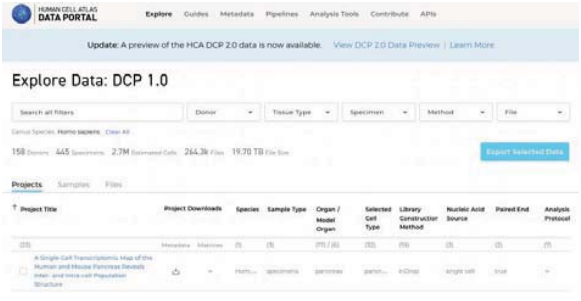


Figure 2. Timeline of HCA activities, October 2016 through January 2018.

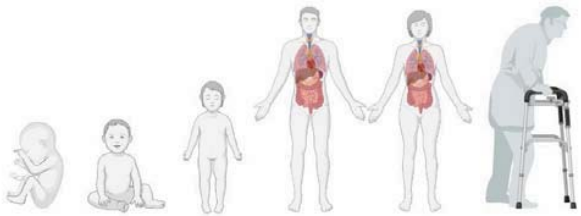


Human cell atlas: data coordination platform





Divide and conquer strategy



164 datasets in DCP

Organs

Developmental stages

Disease status

Data integration

HUMAN CELL ATLAS

99

Example of single-cell atlas database

DCP

4.3 M from 54 projects

HUMAN CELL ATLAS
DATA PORTAL

Explore Data: DCP 2.0 Data View

Search all filters | Donor | Tissu

Genus Species Homo sapiens *AND* File Source DCP/2 Analysis Clear All

4.3M Estimated Cells 742 Specimens 367 Donors 40.7k Files 23.29 TB File Size

Current Query

Genus Species	Homo sapiens
File Source	DCP/2 Analysis

Selected Data Summary

Estimated Cells	4.3M
File Size	23.29 TB
Files	40.7k
Projects	54
Species	Homo sapiens
Donors	367

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

100

Example of single-cell atlas database

Single Cell Portal

3.8 M from 38 projects

https://singlecell.broadinstitute.org/single_cell

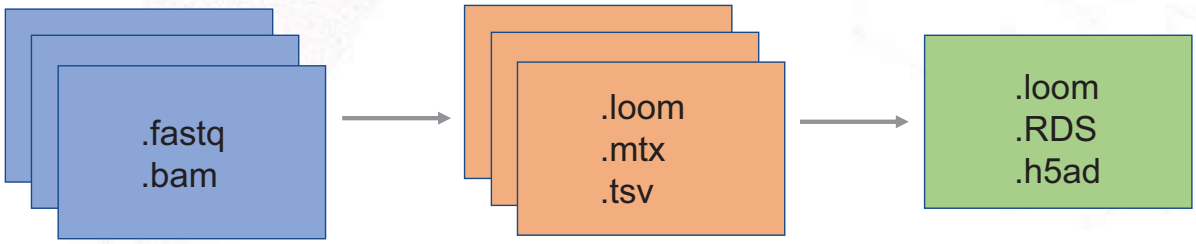
Example of single-cell atlas database

EBI

3.6M 103 projects

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

Structure of single-cell data files



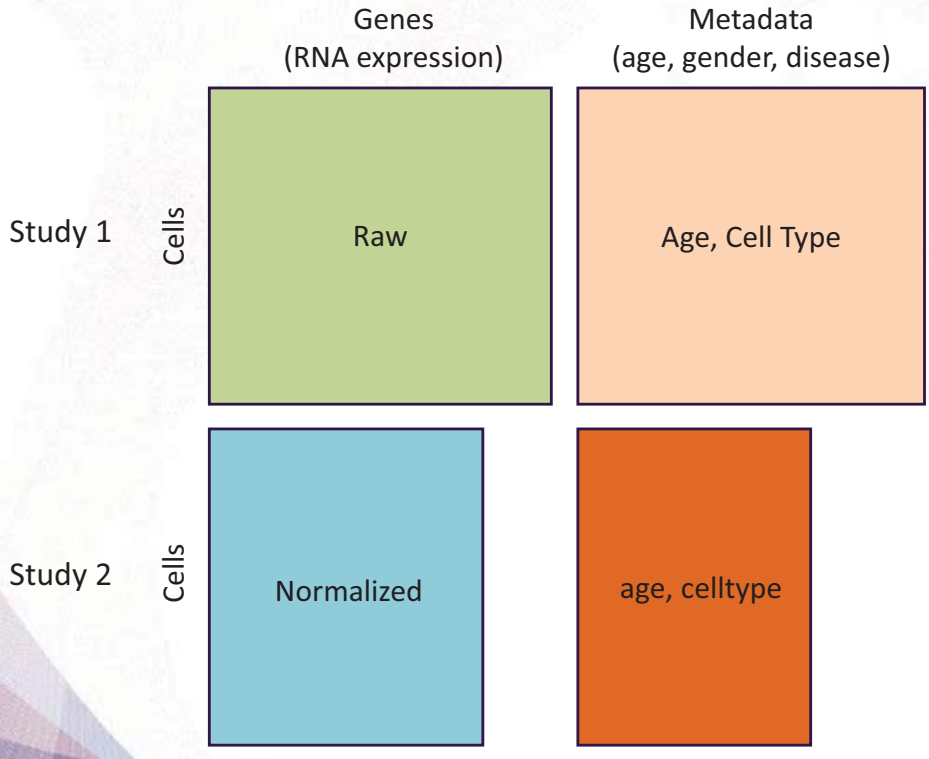
Sequence files | Count matrix | Annotated matrix

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

Easy to download
Matched metadata

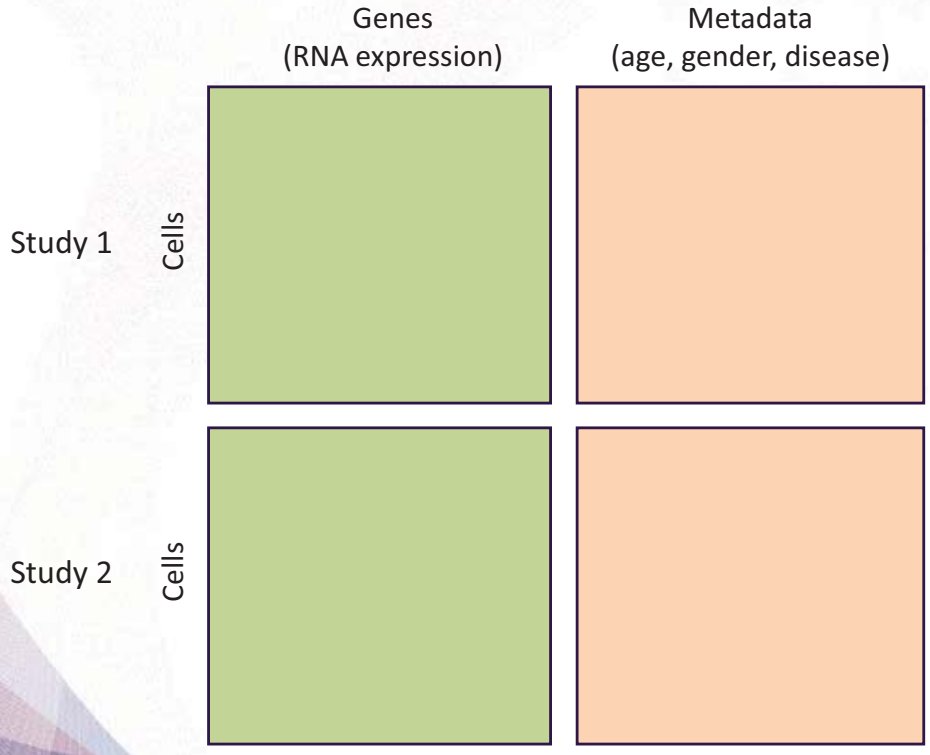
103

Potential problems in utilizing annotated matrix



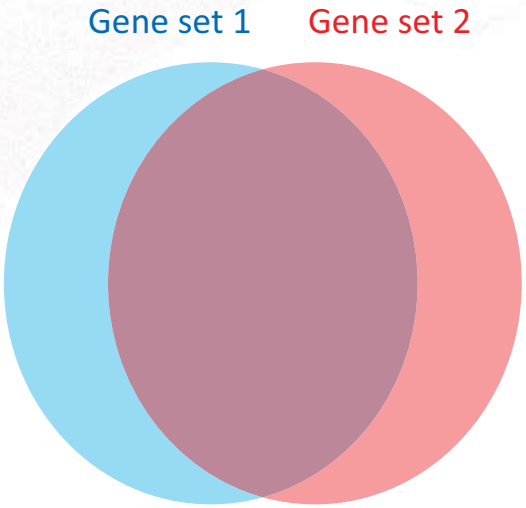
104

Desired outcome for utilization of annotated matrix files



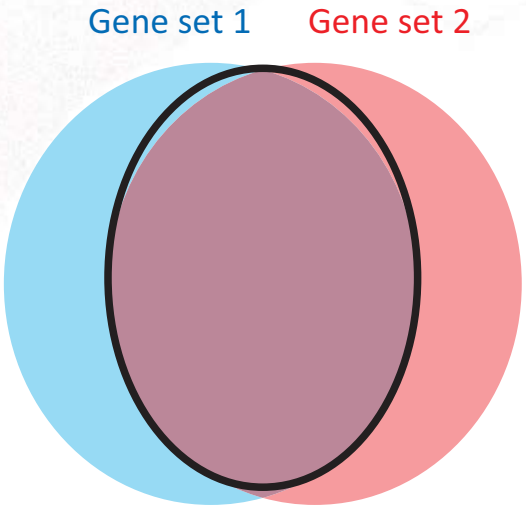
105

Harmonizing gene sets



106

Harmonizing gene sets

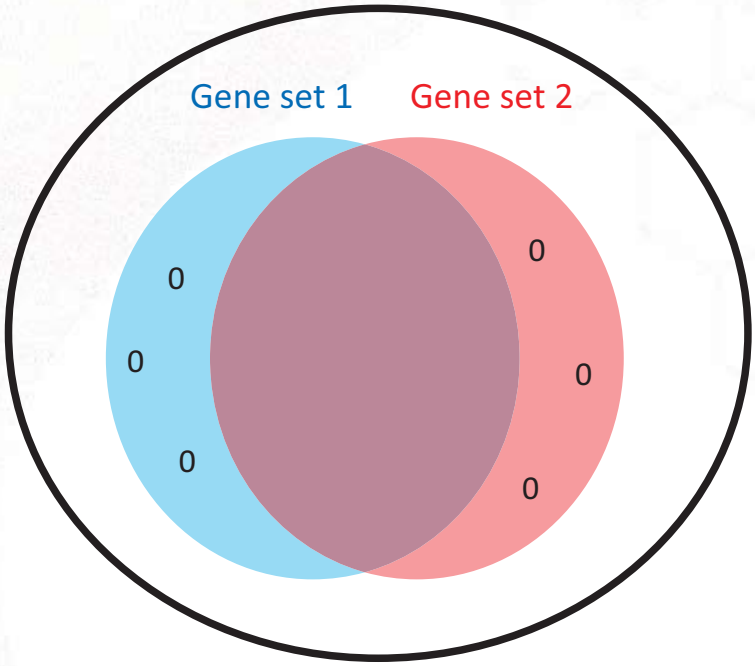


Approach 1. Intersection

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

107

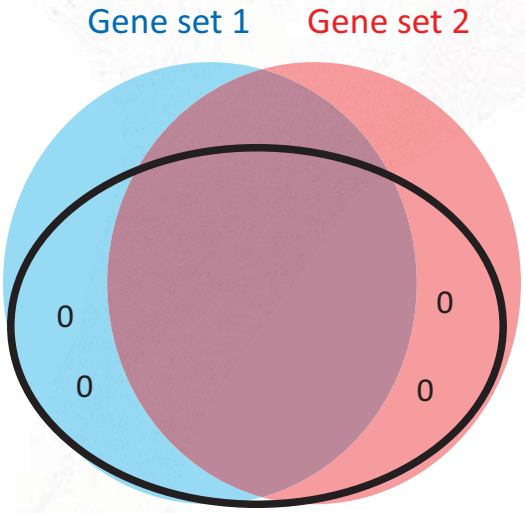
Harmonizing gene sets



Approach 2. Union
Downside: Increased ”batch effect”

108

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

109

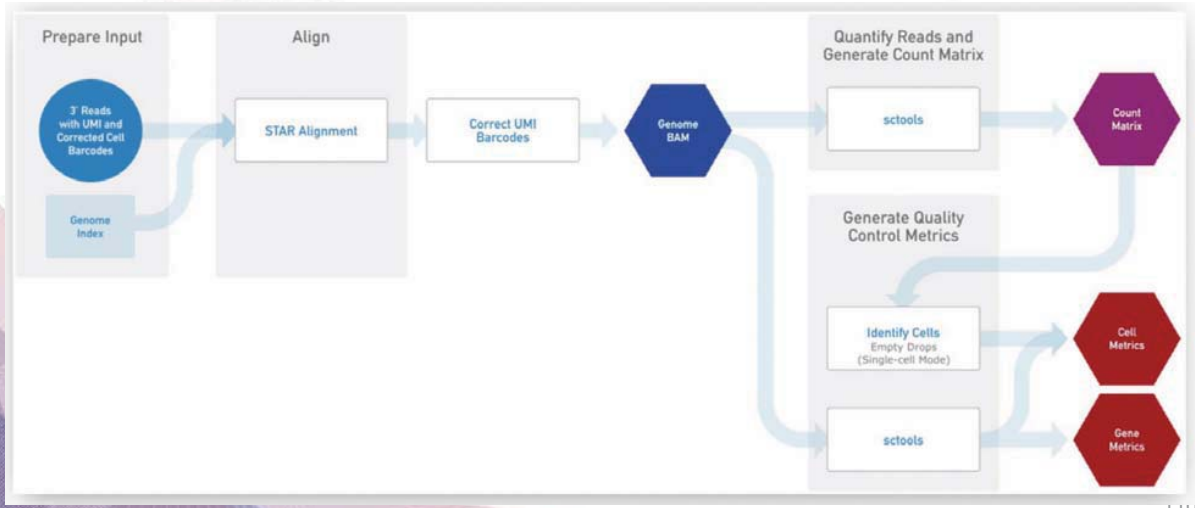
Remapping (sequence alignment using uniform pipeline)

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.

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.

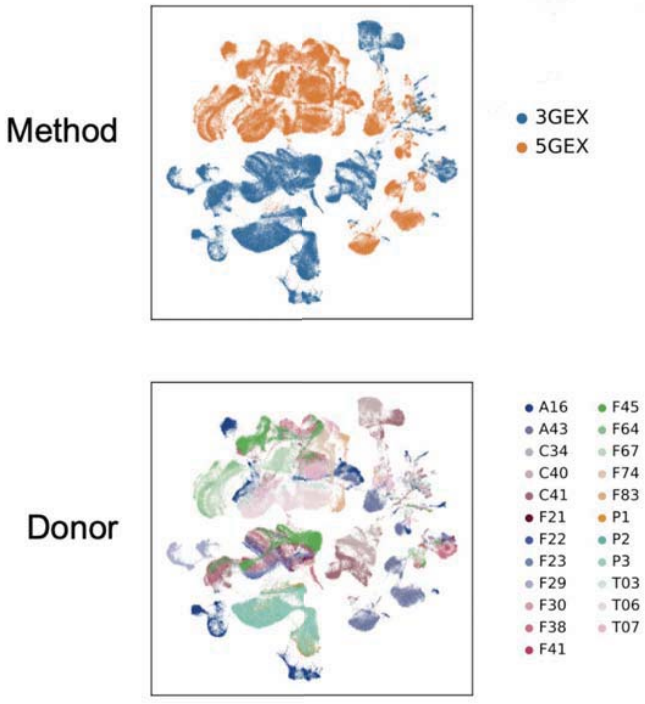
HUMAN CELL ATLAS
DATA PORTAL



110

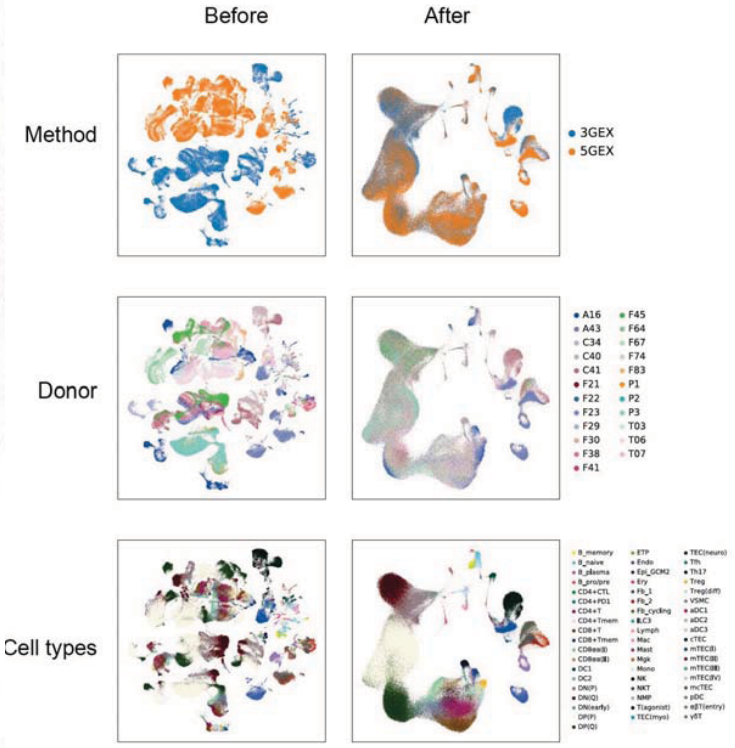
3-2. Single-cell RNA-seq data integration (batch correction)

Problem of batch effect in single-cell data



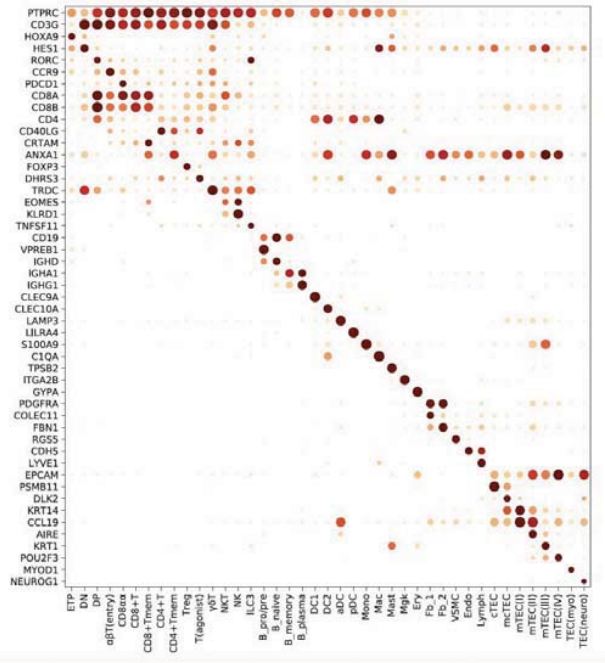
Park et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science*.

Correcting for the batch effect

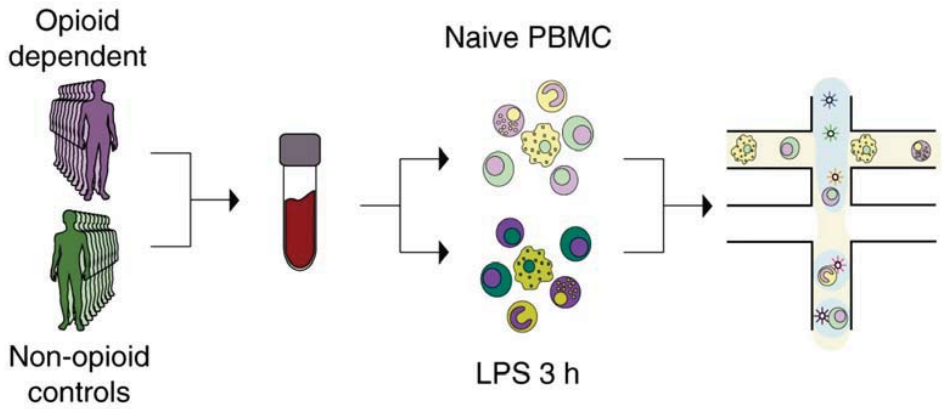


Source of variation for gene expression?

Source of variation (1) Biology – Cell types

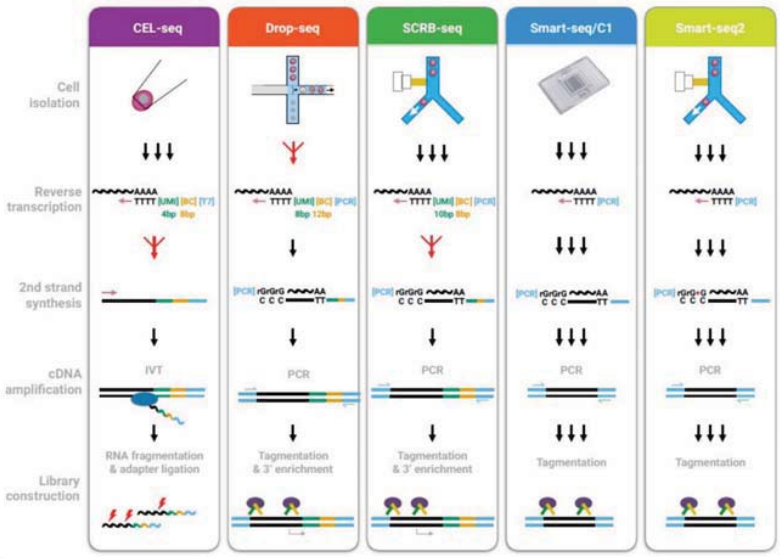


Source of variation (2) Experimental conditions



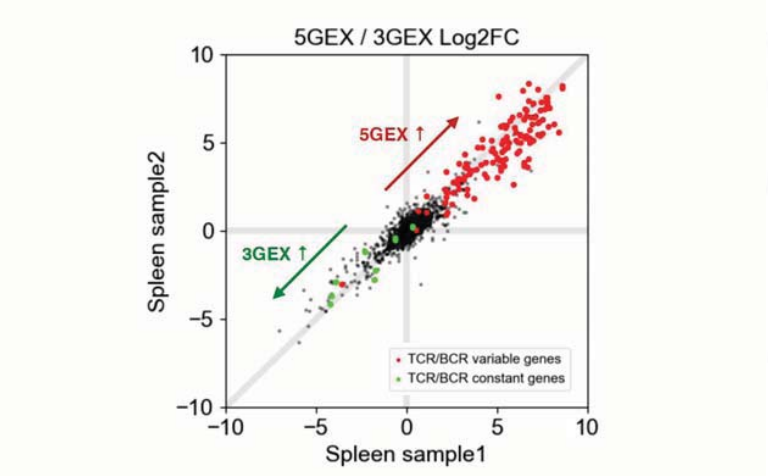
Source of variation (3) Replicates

Source of variation (3) Technology

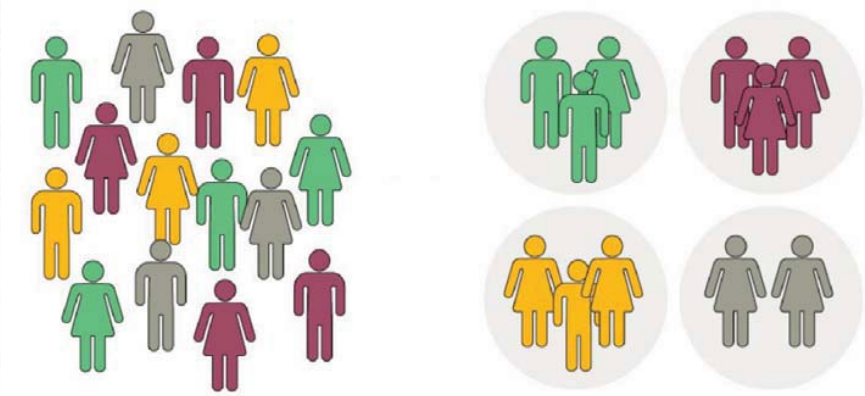


https://www.biorxiv.org/content/10.1101/035758v3.full#disqus_thread

Source of variation (3) Technology

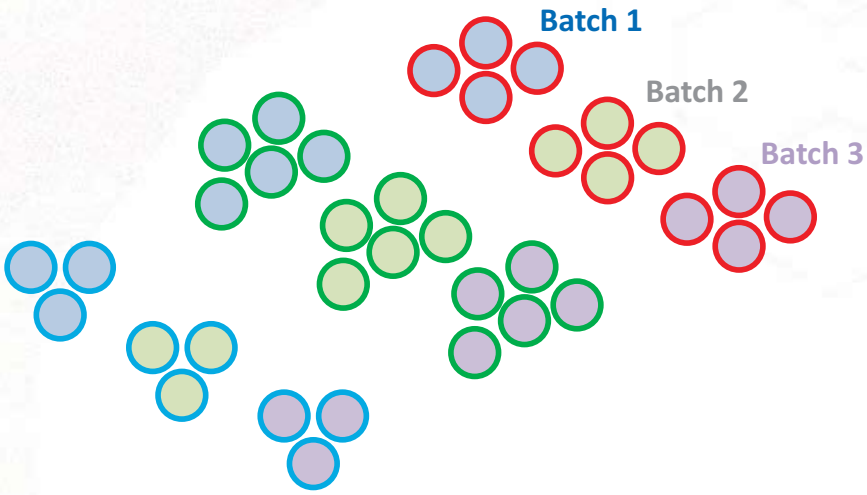


Source of variation (4) Individuals (genotypes)



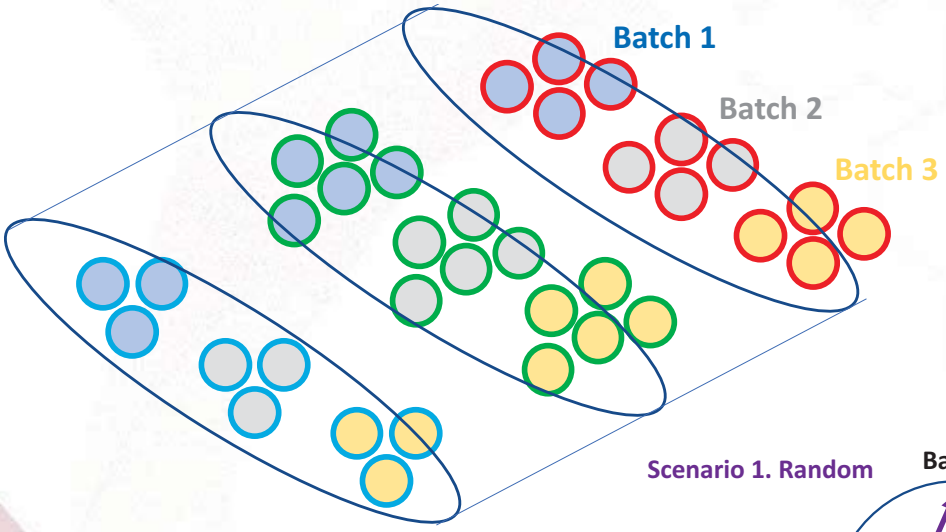
Genders (XIST) or HLA genes…

Visualizing batch effect in single-cell data



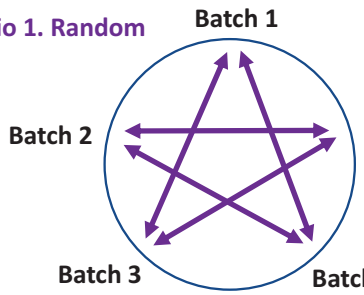
gene expression space…

Structure of batch effect



gene expression space…

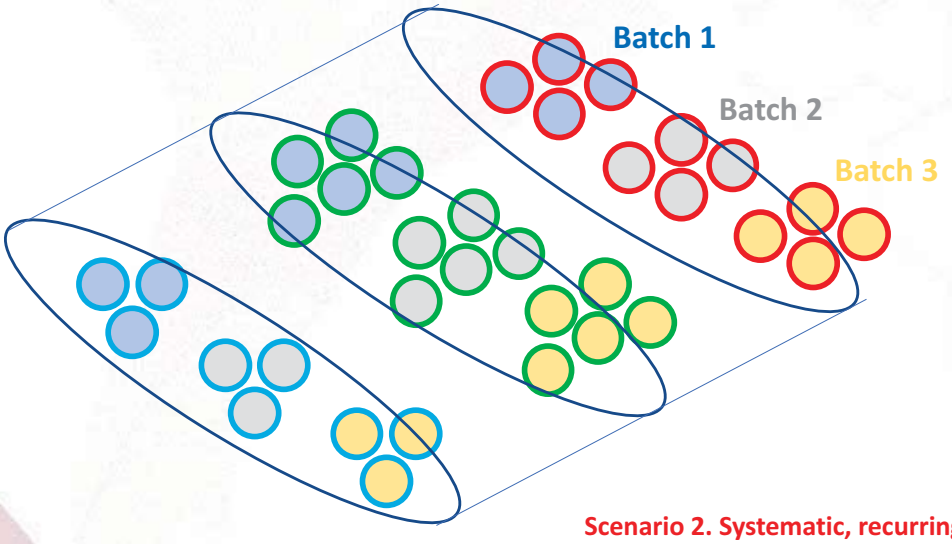
Scenario 1. Random



Genotype,
Exp. Batch…

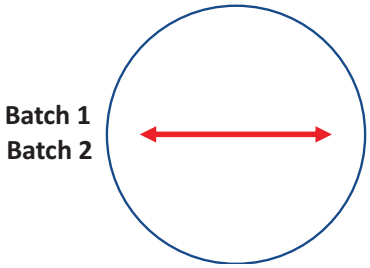
121

Structure of batch effect



gene expression space…

Scenario 2. Systematic, recurring



10X 3’ vs 5’
Development
Disease

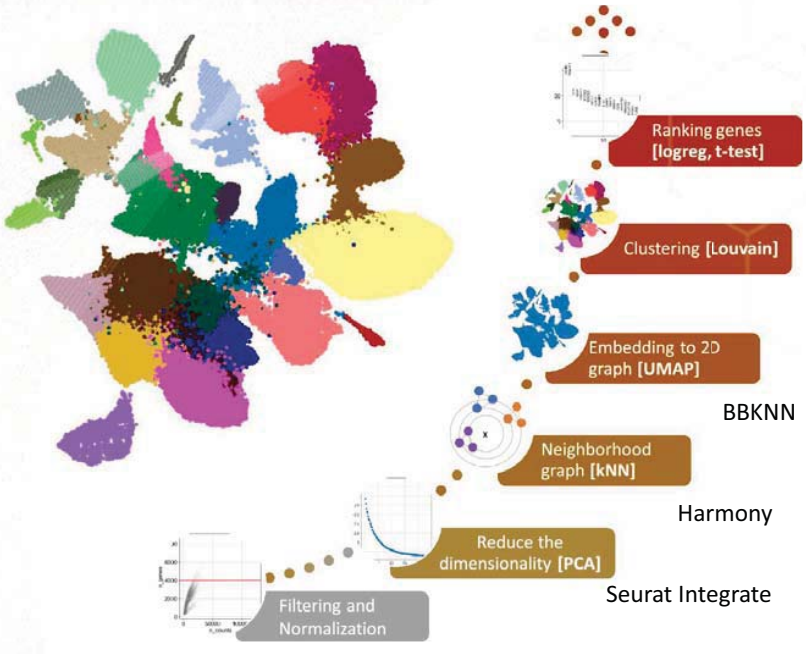
122

Expression of gene X

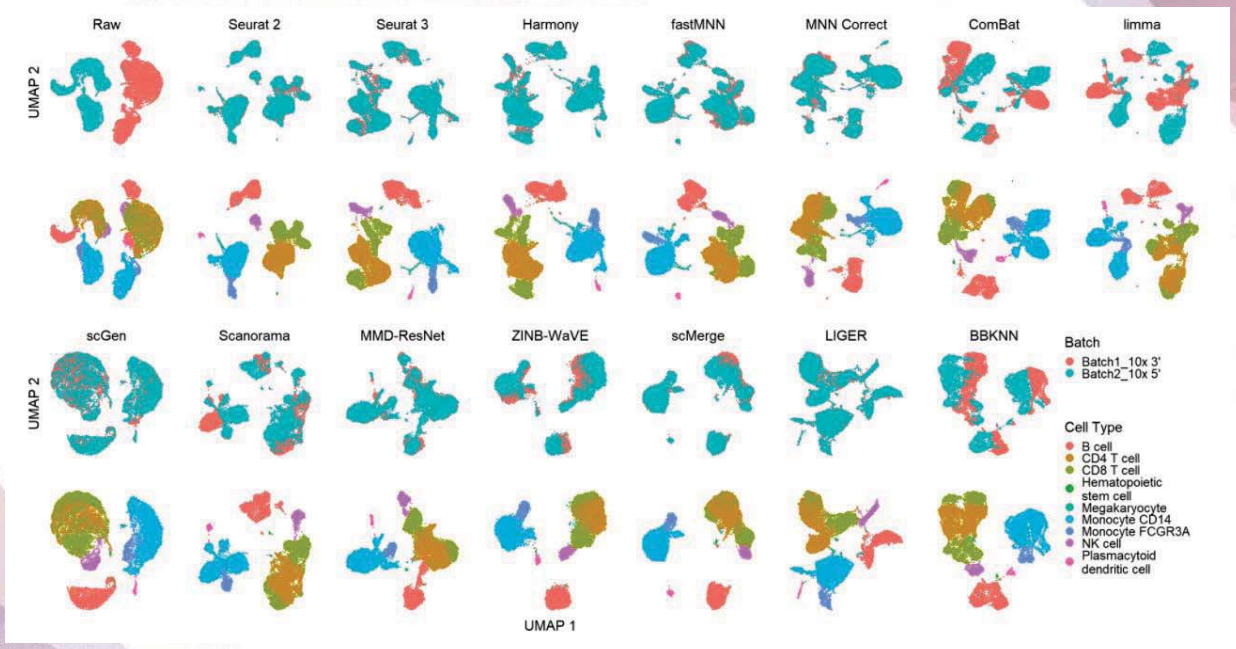
Cell type + Experimental conditions +	Keep!
Replicate + Technology + Genotype + Bioinformatic pipeline	Remove!

Batch effect!

Multiple batch correction strategy



Different methods, different results

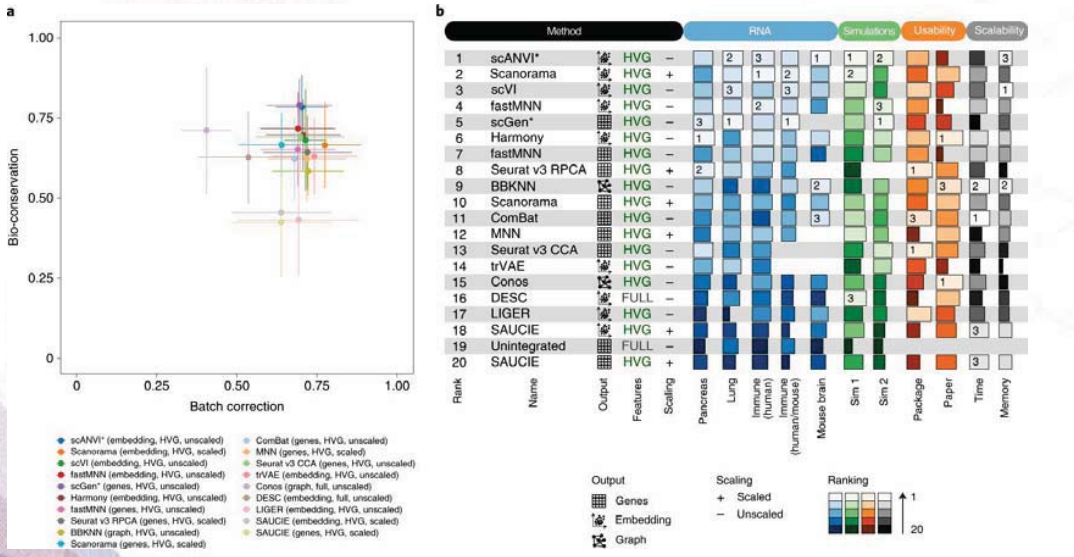


Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21, 12

125

Goals for ideal batch correction

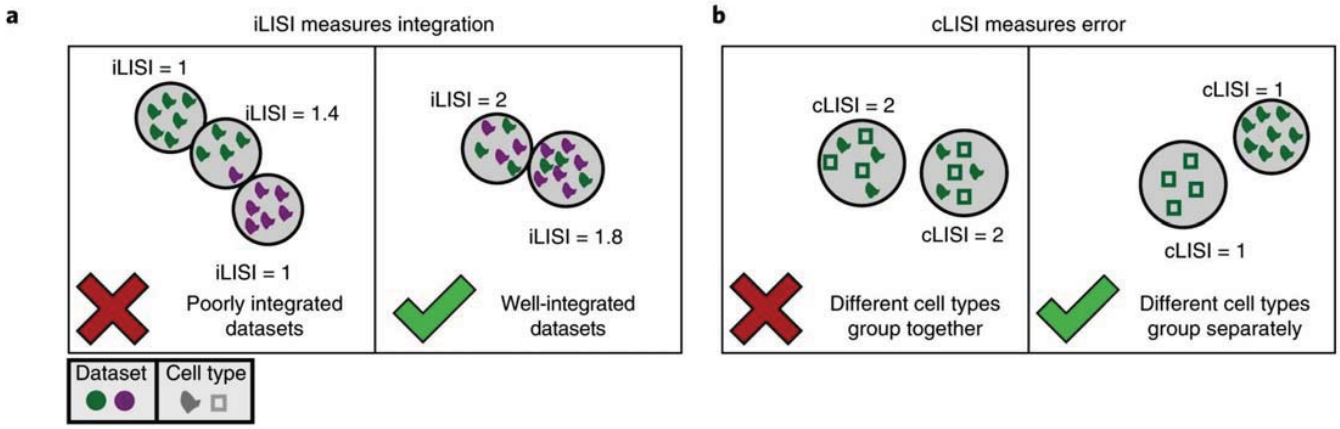
- **Batch removal -> Good harmonization across batches**
- **Bio conservation -> Maintaining biological integrity (no distortion or over-correction)**
- **Scalability -> Can deal with large scale datasets**



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

126

Metrics to assess the quality of batch correction



Measuring integration

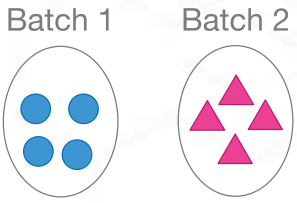
Measuring cell type preservation

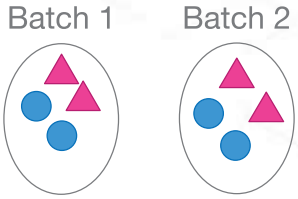
Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat*

Linear regression

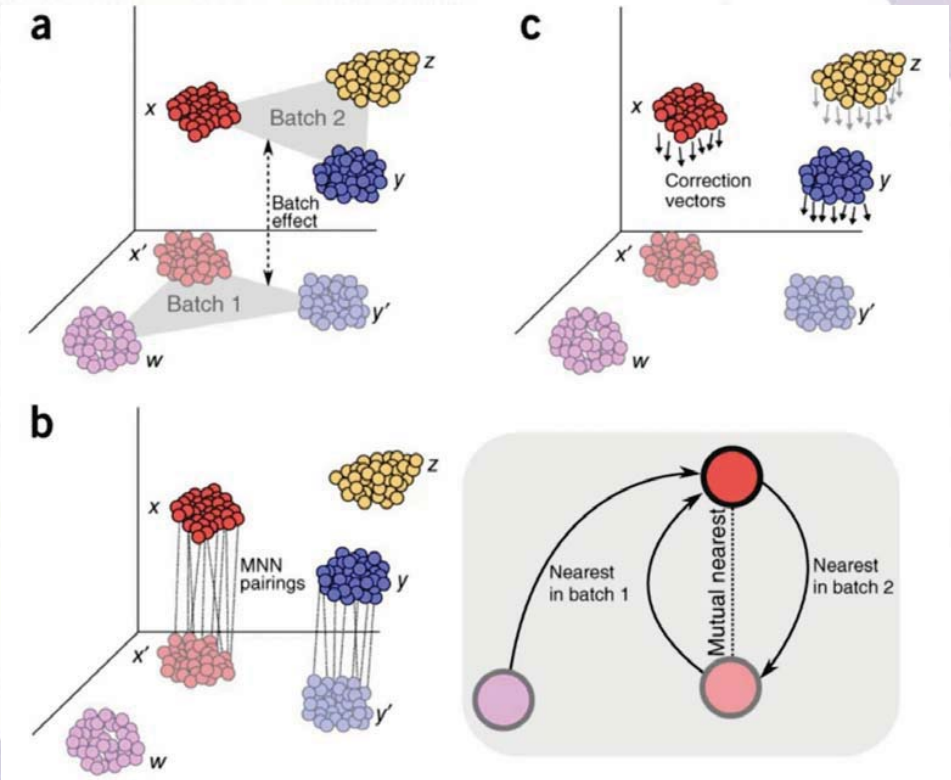
Y ~ Tech + Donor + Gender + residual

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





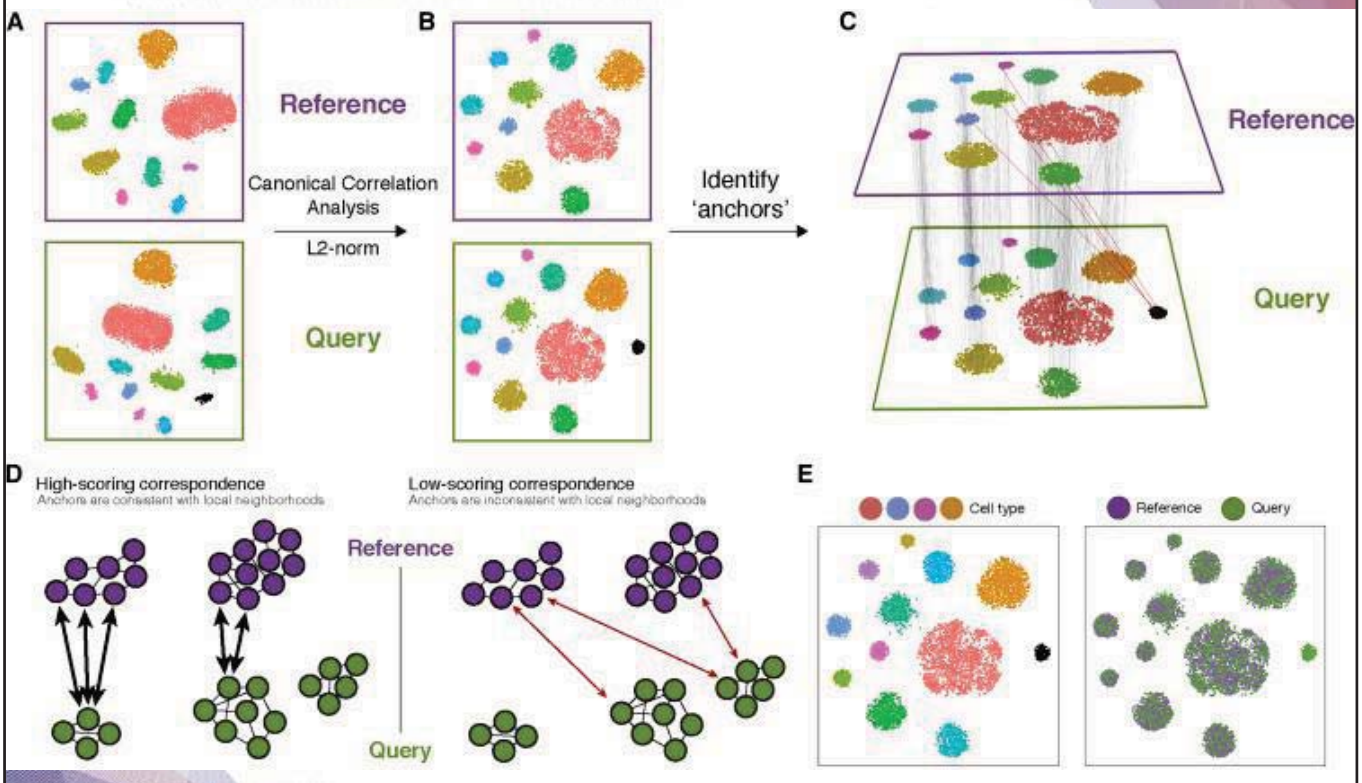
Mutual nearest neighbors



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

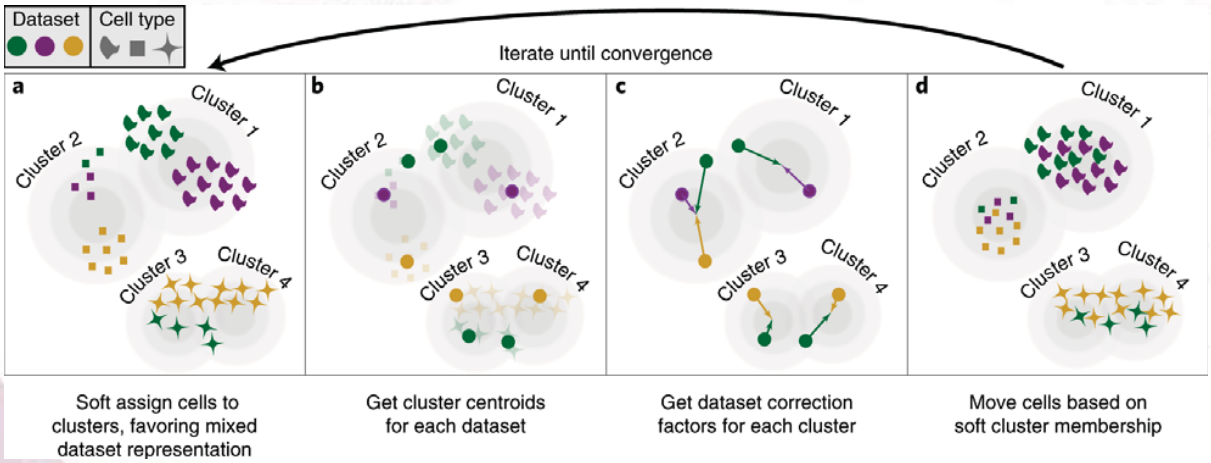
Finding anchors (with CCA dimension reduction)



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

130

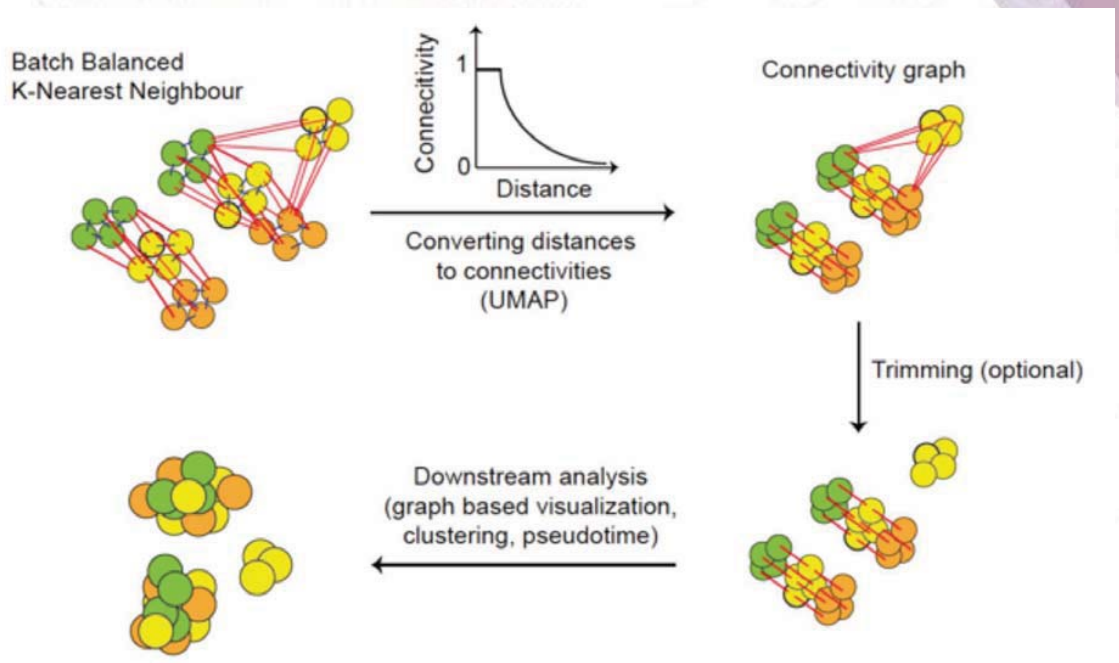
Harmony: batch correction at cluster level



Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16, 1289–1296 (2019).

131

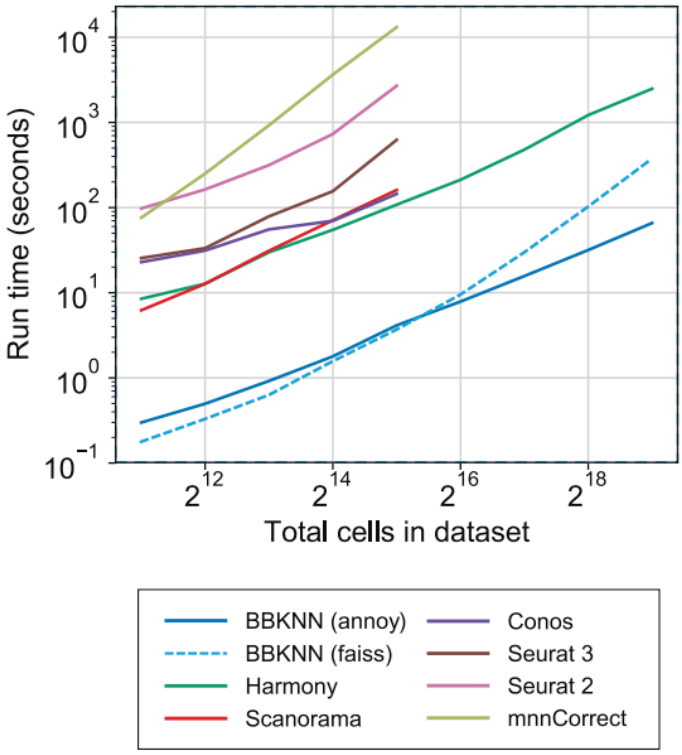
BBKNN: Biology correction at graph level



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

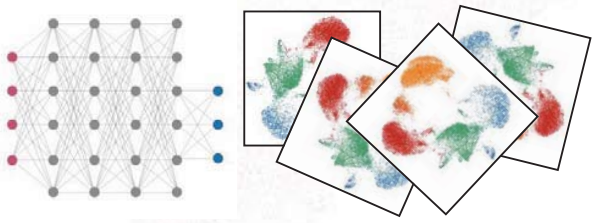
132

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



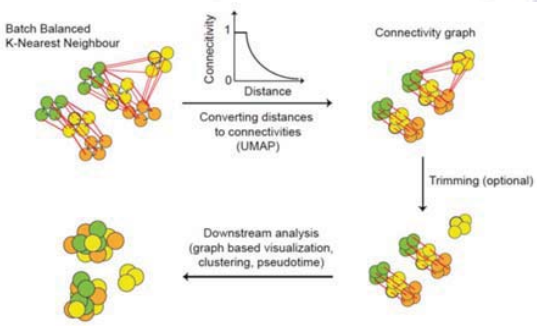
Model based cell type prediction

Correct systematic batch

Keeping biology

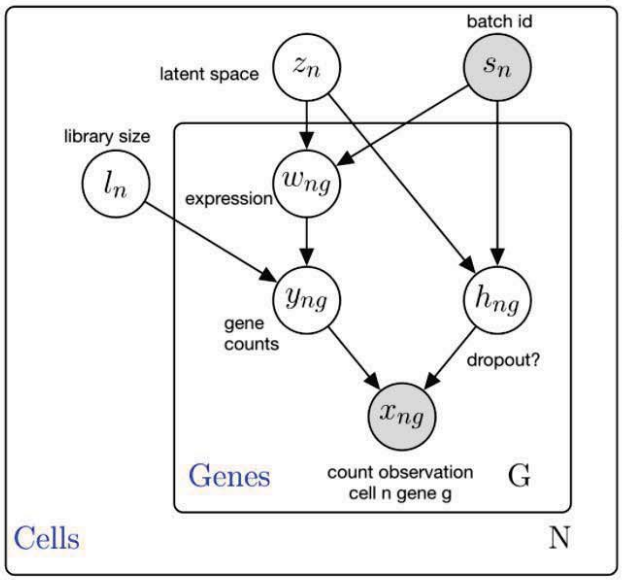
Gene Expression ~ Batch + Cell Type

Regularized linear model with batch + cell type design



BBKNN

Deep learning & scRNA-seq data integration



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

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

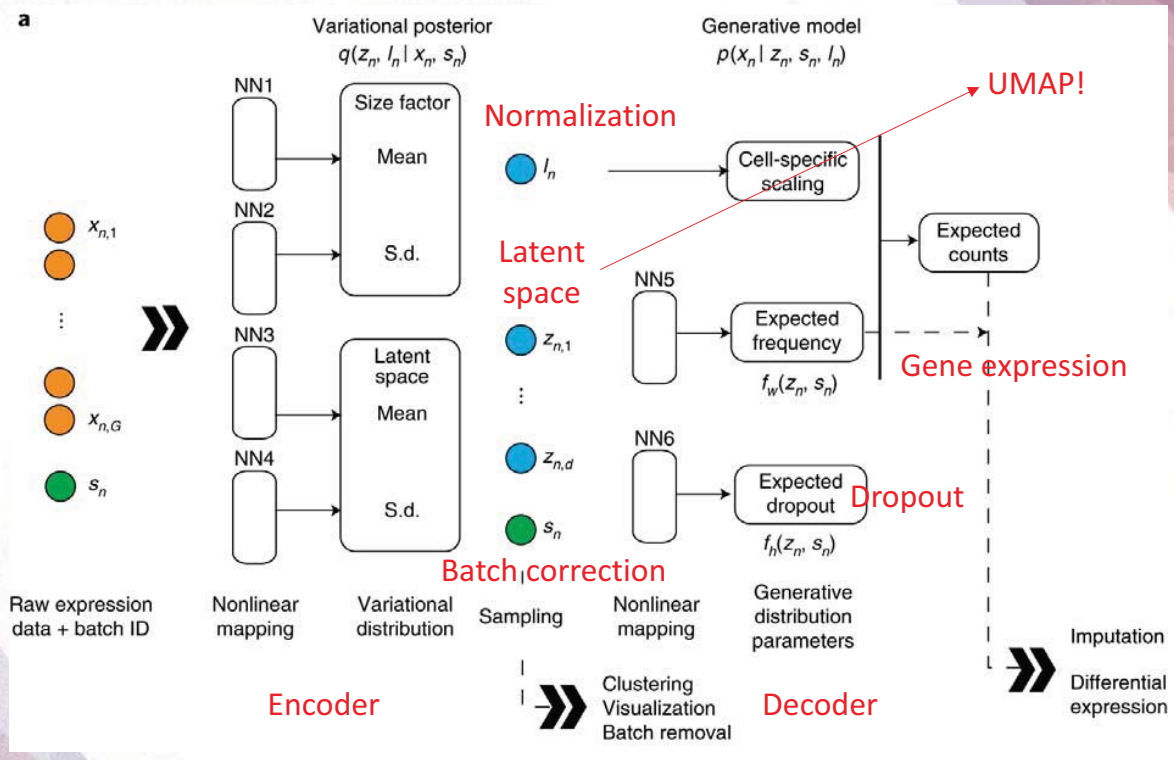
$$l_n \sim \text{LogNormal}(l_\mu^{s_n}, l_\nu^{s_n})$$

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

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

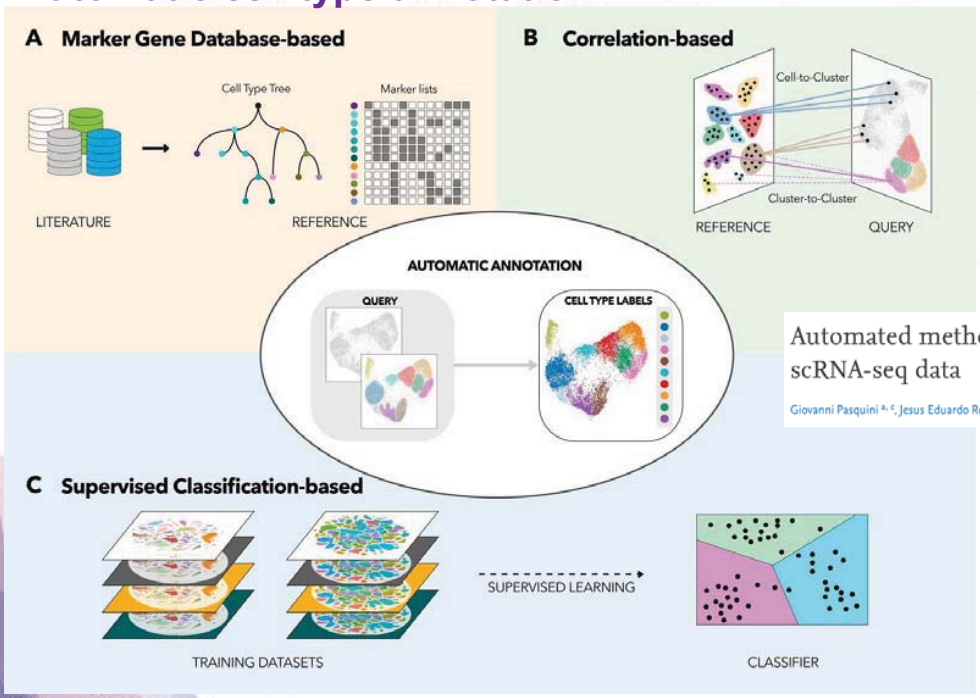
Lopez, R., Regier, J., Cole, M.B. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat*

Batch effect & Embedding -> SCVI



Variational autoencoder

Automatic cell type annotation

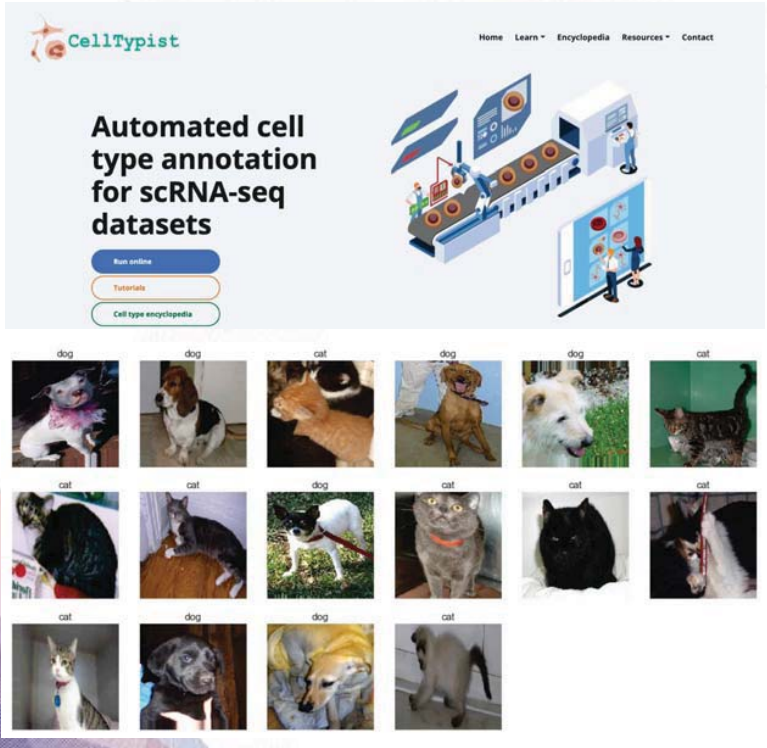


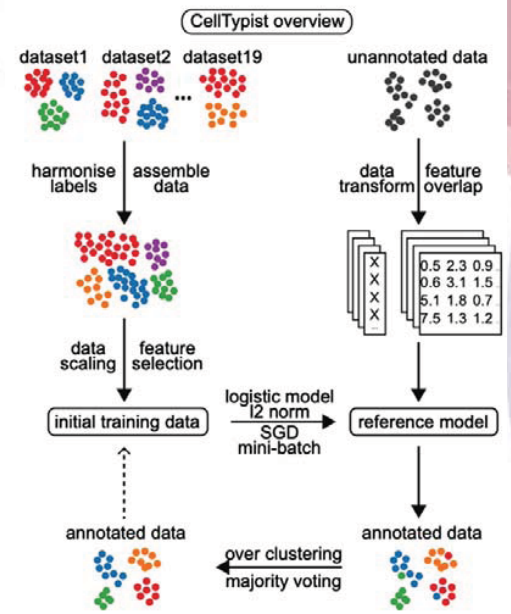
Automated methods for cell type annotation on scRNA-seq data

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

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

Machine learning based general cell annotation

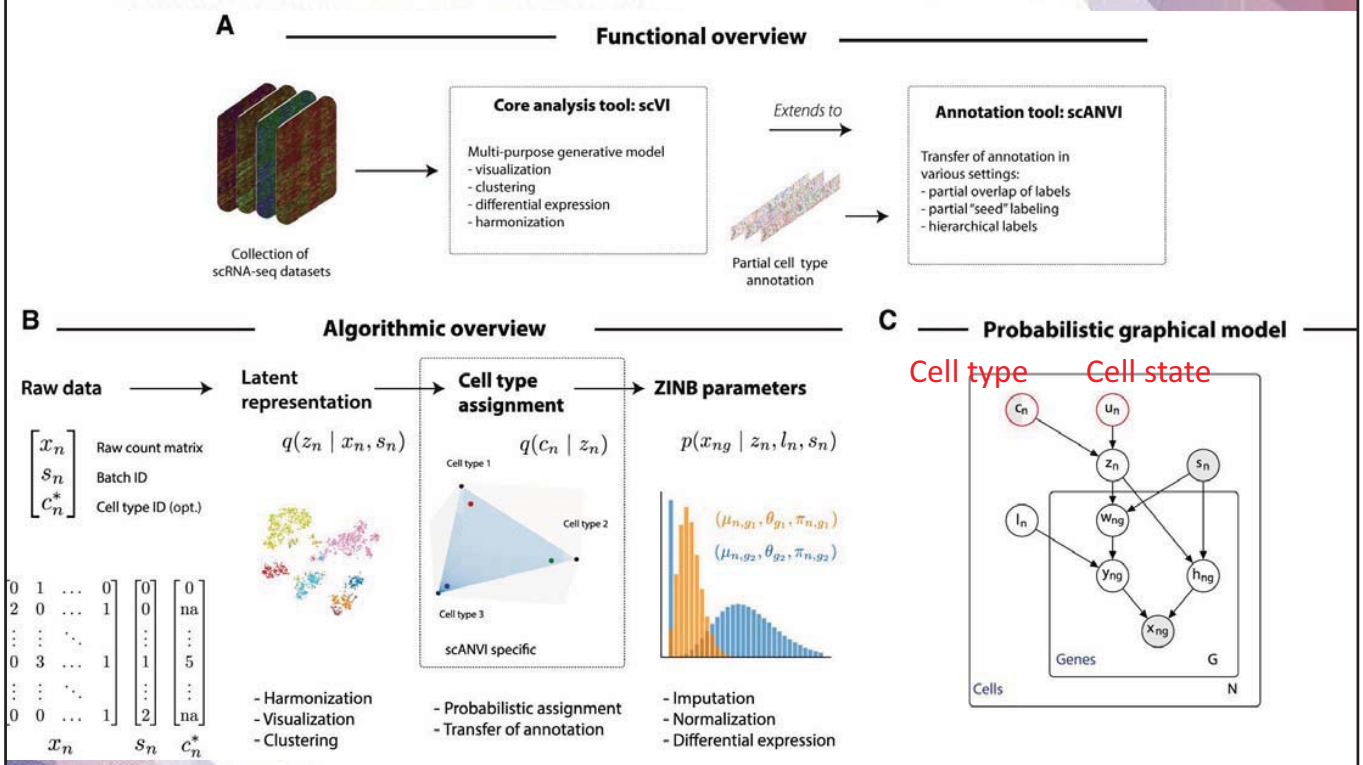




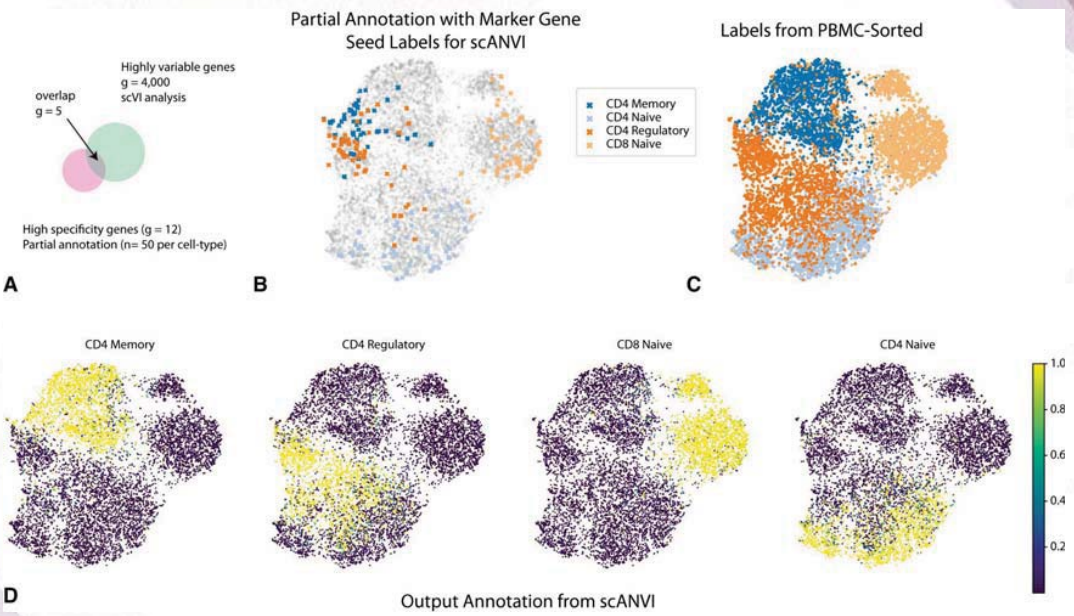
Domínguez Conde, C., et al. "Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture in humans." (2021).

138

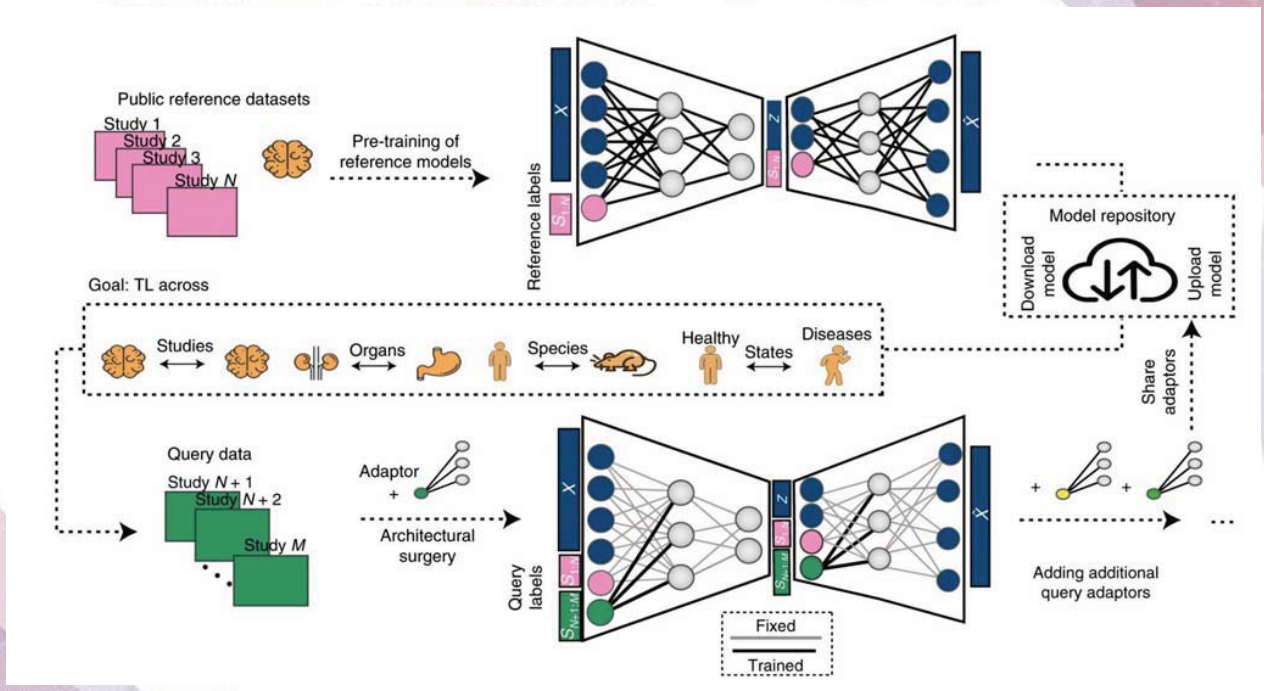
Annotation transfer -> SCANVI



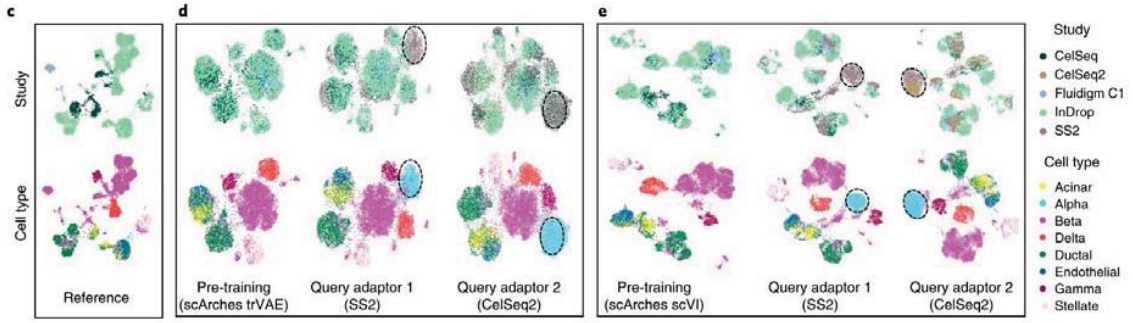
Annotation transfer -> SCANVI



Annotation transfer -> scArches (Transfer learning)

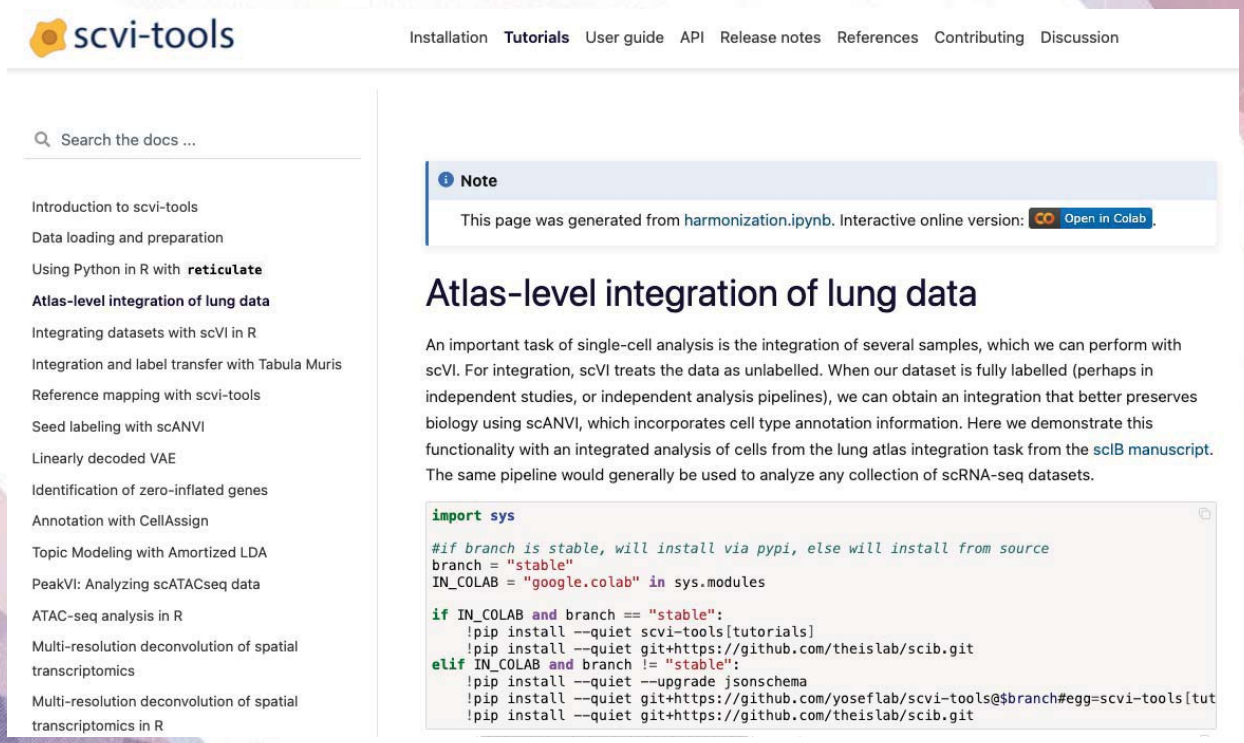


Annotation transfer -> scArches (Transfer learning)



Biology at single-cell resolution

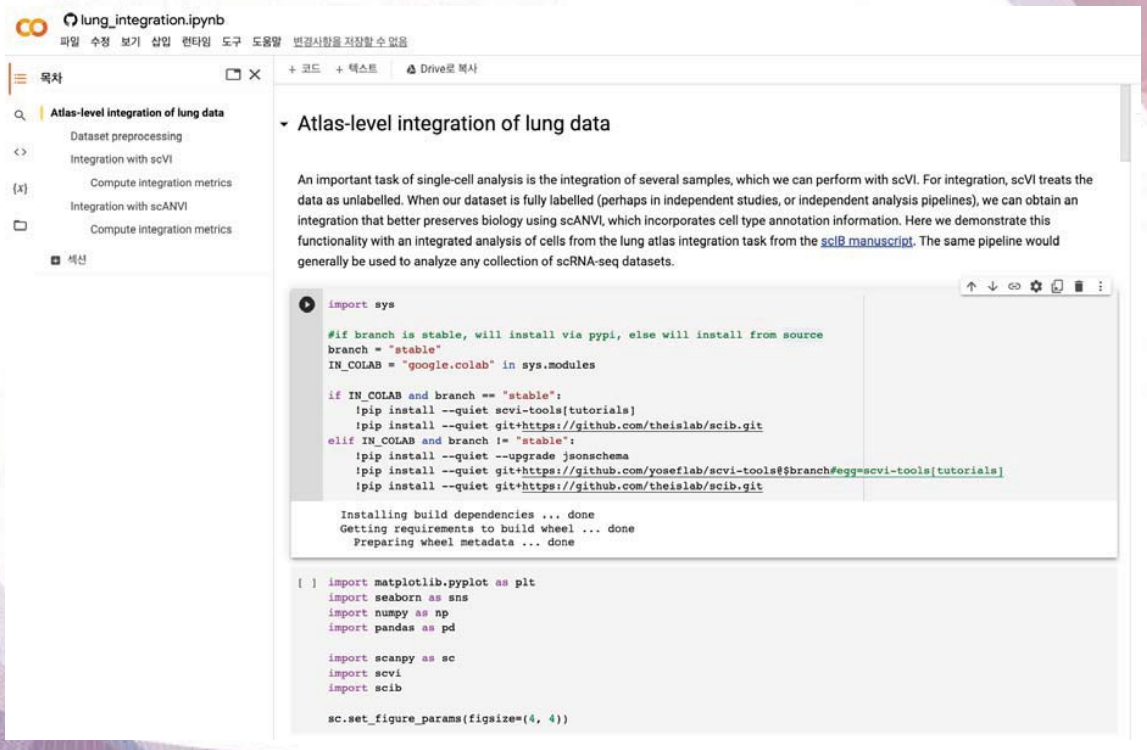
https://docs.scvi-tools.org/



143

Try google colab platform

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



144

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")
```

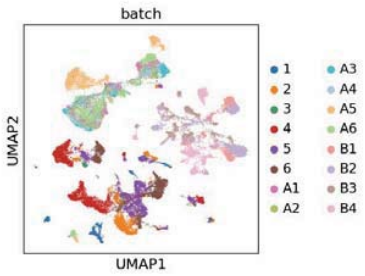
Now we train scVI. This should take a couple of minutes on a Colab session

```
vae.train()
```

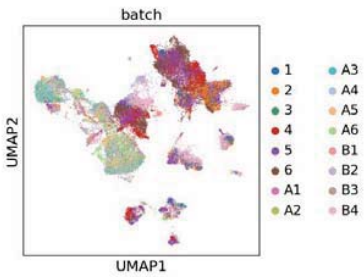
```
Epoch 246/246: 100%|| 246/246 [09:19<00:00,  2.27s/it, loss=553, v_num=1]
```


Tutorial - SCVI

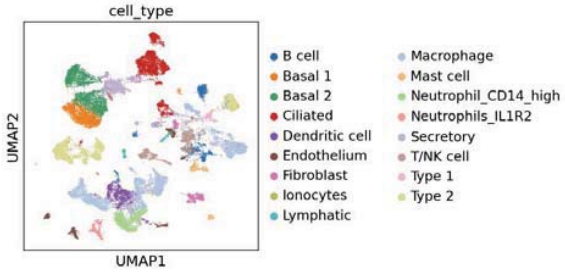
```
[ ] sc.pl.umap(adata,color='batch')
```



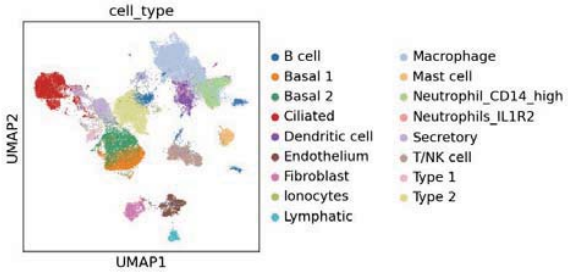

```
sc.pl.umap(adata,color='batch')
```




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




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

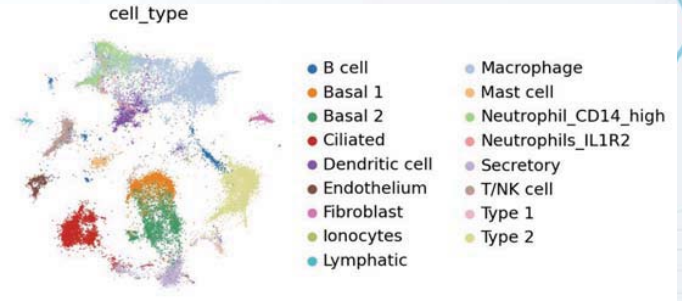


Tutorial - SCANVI

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

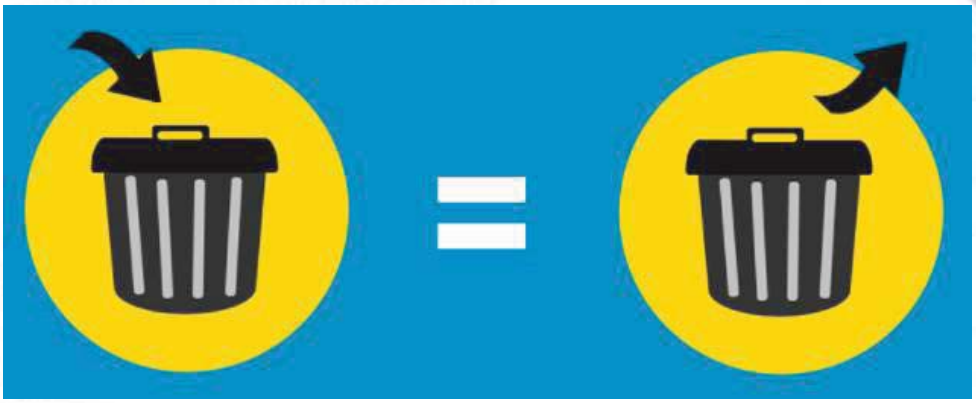
```
lvae.train(max_epochs=20, n_samples_per_label=100)
```

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



147

More things to consider…



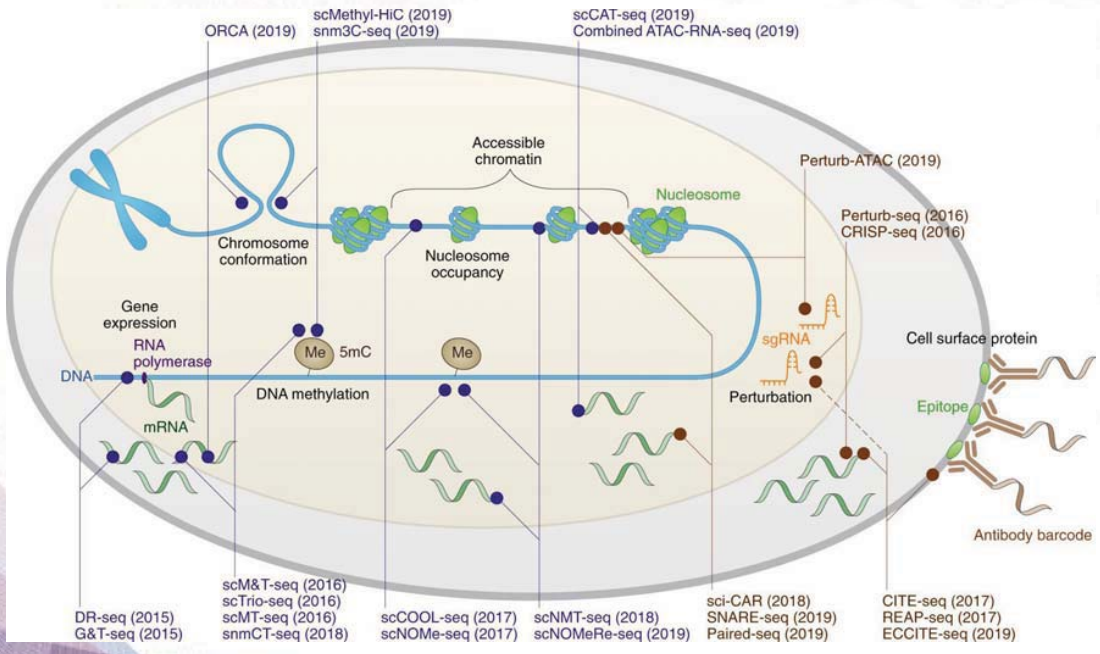
- Removing bad quality data – “Garbage in garbage out”
- Removing doublets
- Considering ‘soup effect’

148

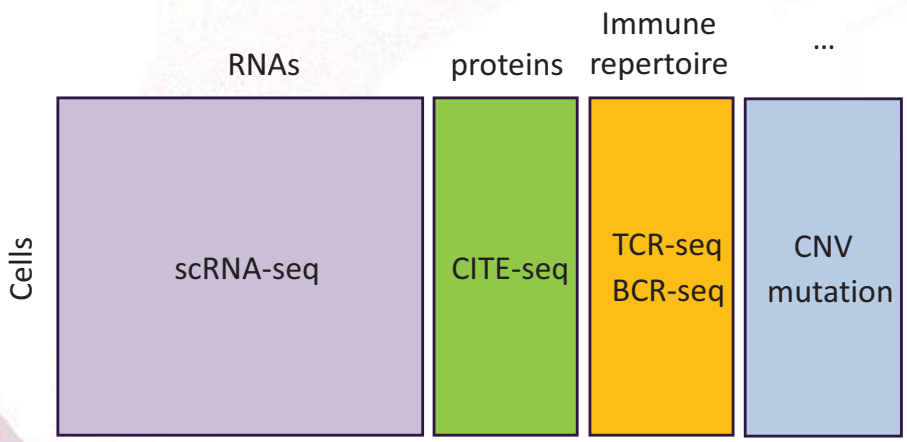
4. Single-cell multi-omics analysis

What about other single-cell omics data?

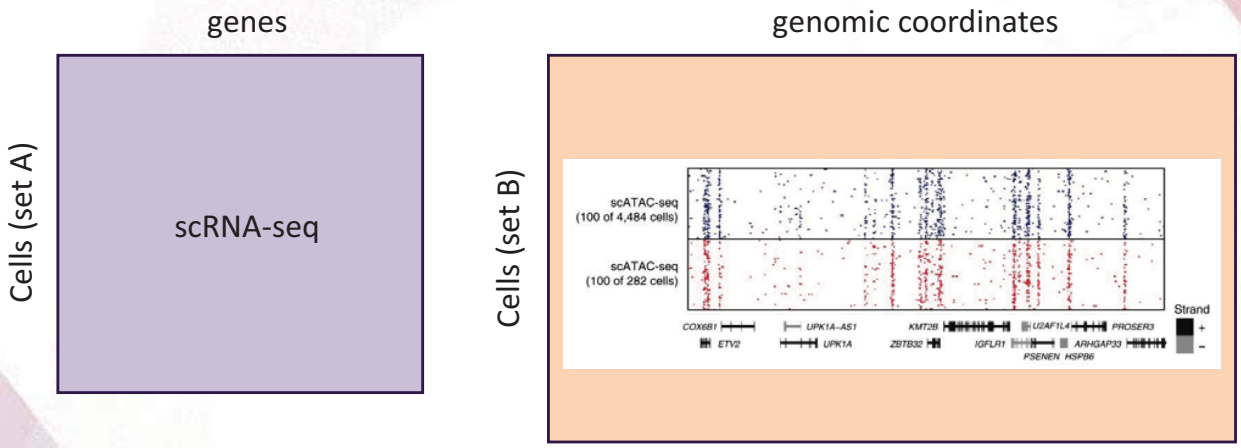
Beyond RNA: Single-cell multiomics



Beyond gene expression matrix



Structure of ATAC-seq data



Peak calling for ATAC-seq

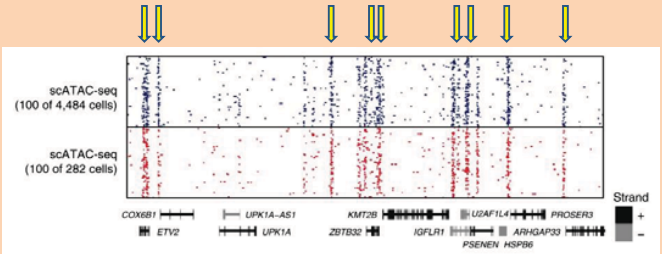
genes

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

genes

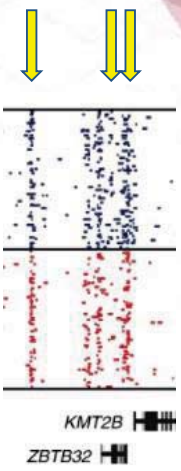
Cells (set A)

scRNA-seq

peak density

Cells (set B)

scATAC-seq



Assign peak to gene
Based on proximity

Linking peaks to genes

Genes
(RNA expression)

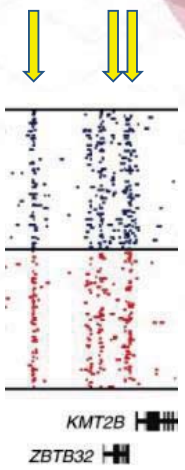
Cells (set A)

scRNA-seq

Genes
(Chromatin accessibility score)

Cells (set B)

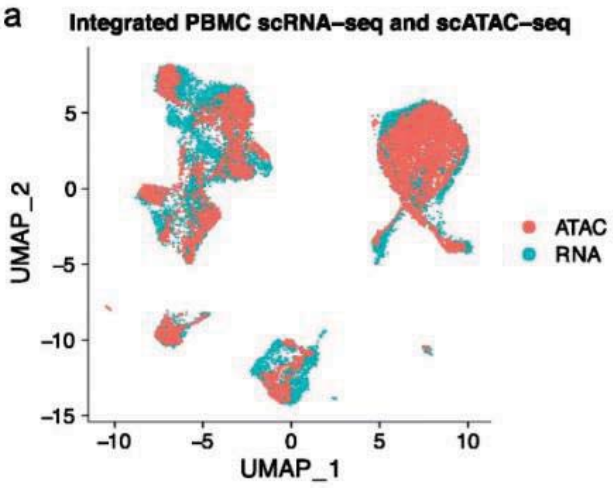
scATAC-seq



Assign peak to gene
Based on proximity

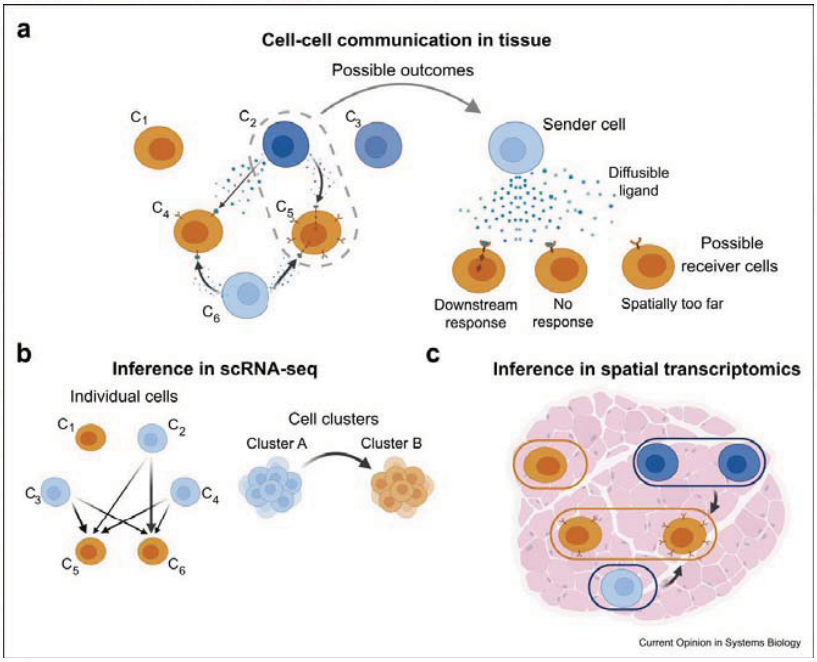
Batch effect correction problem

Combining RNA with chromatin information

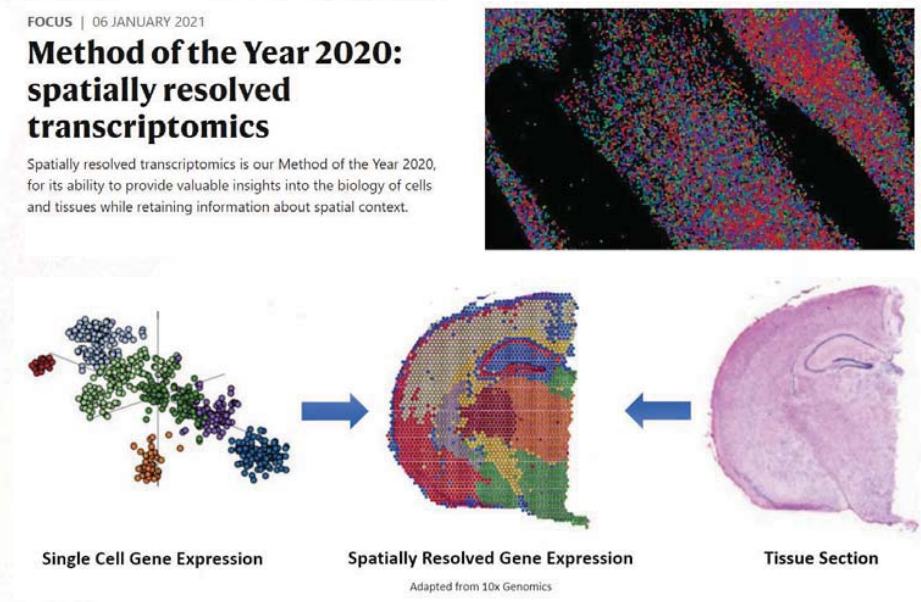


Link epigenetics to RNA expression

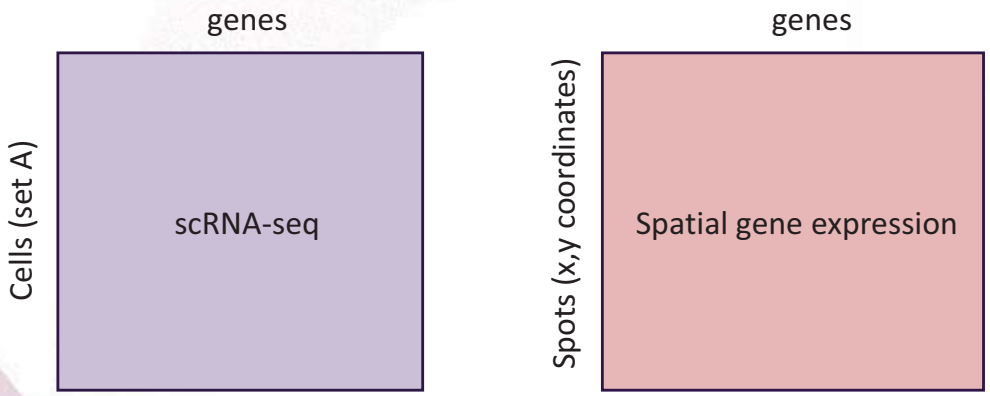
Importance of spatial information



Single cell transcriptomics with spatial resolution

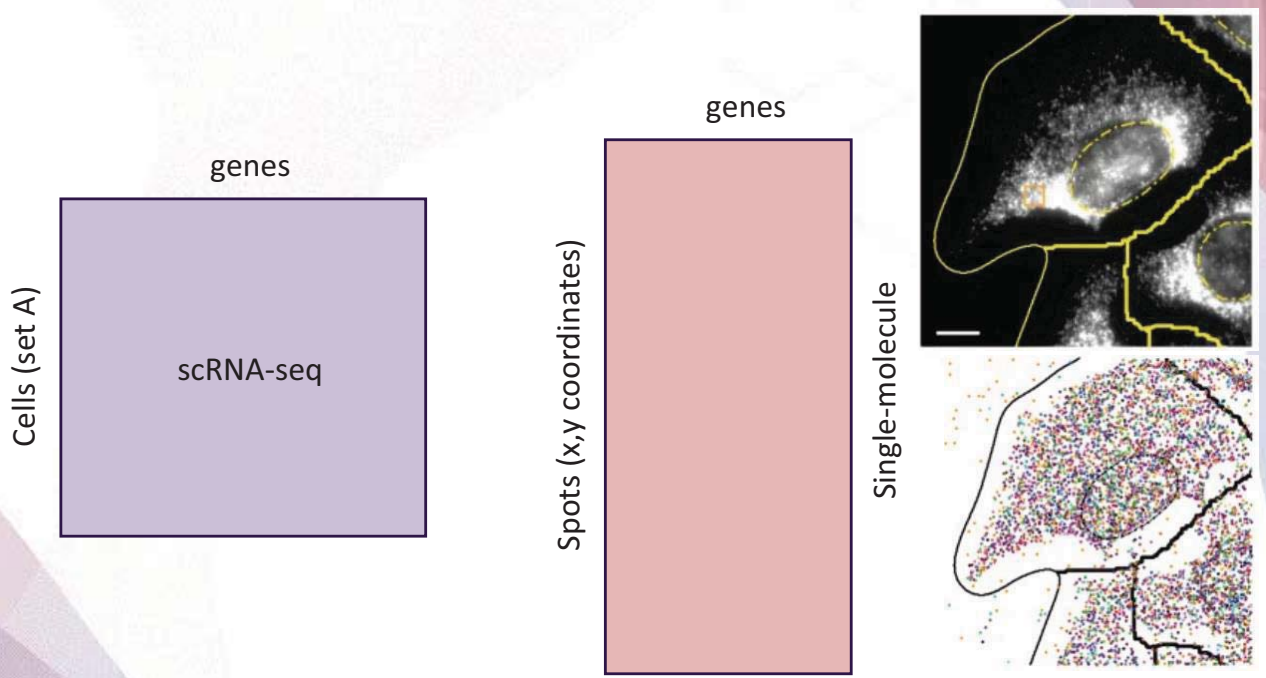


Structure of spatial gene expression data

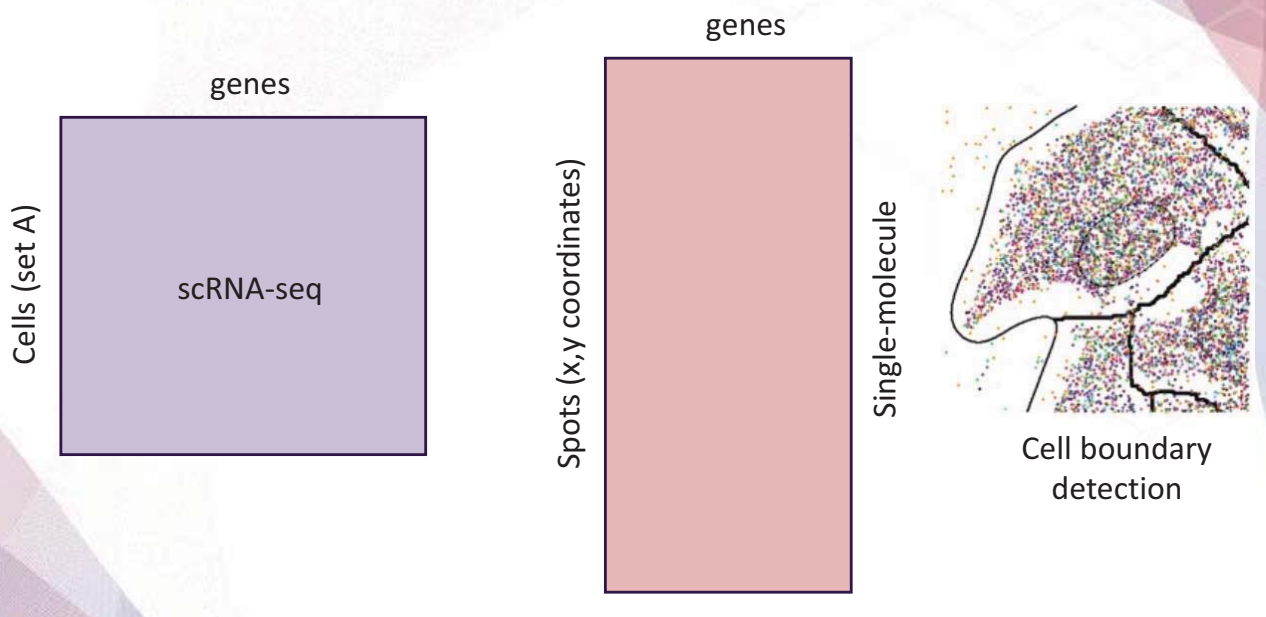


Structure of spatial gene expression data (Visium)

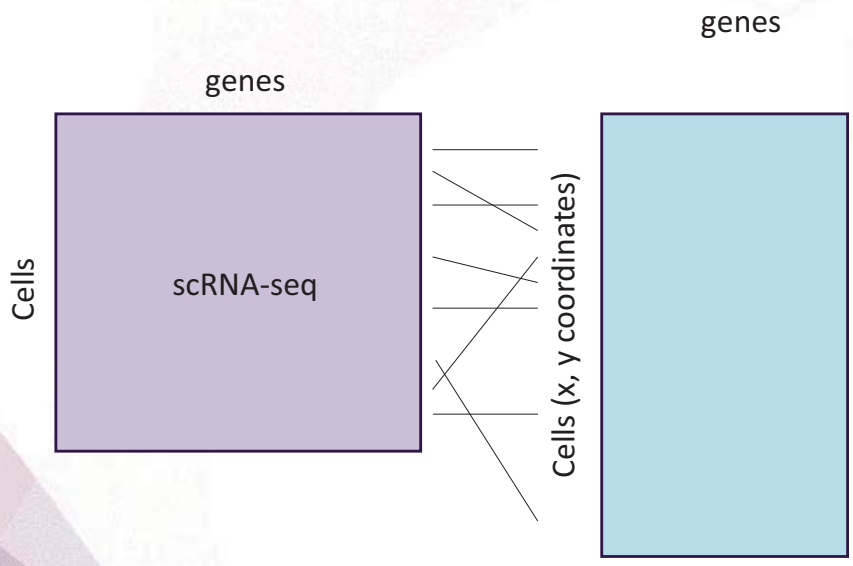
Structure of spatial gene expression data (smFISH)



Structure of spatial gene expression data (smFISH)

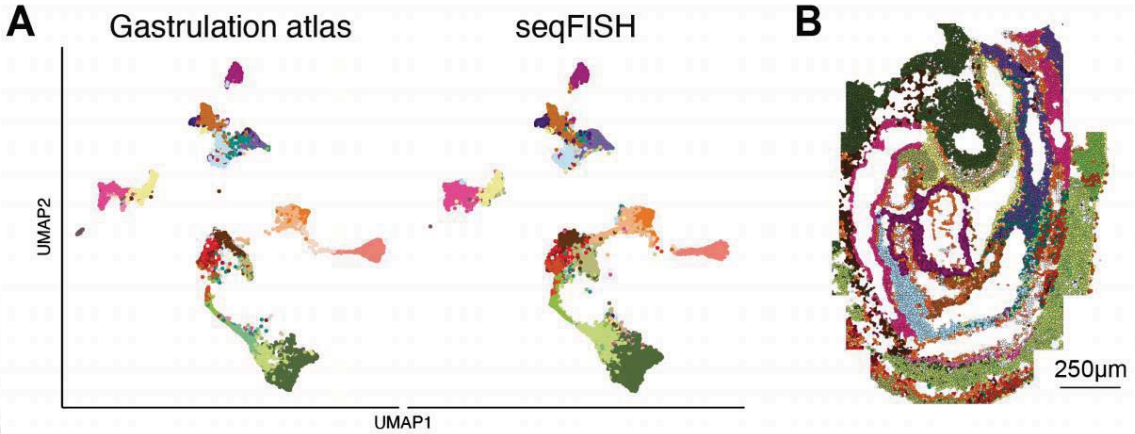


Mapping cells to space

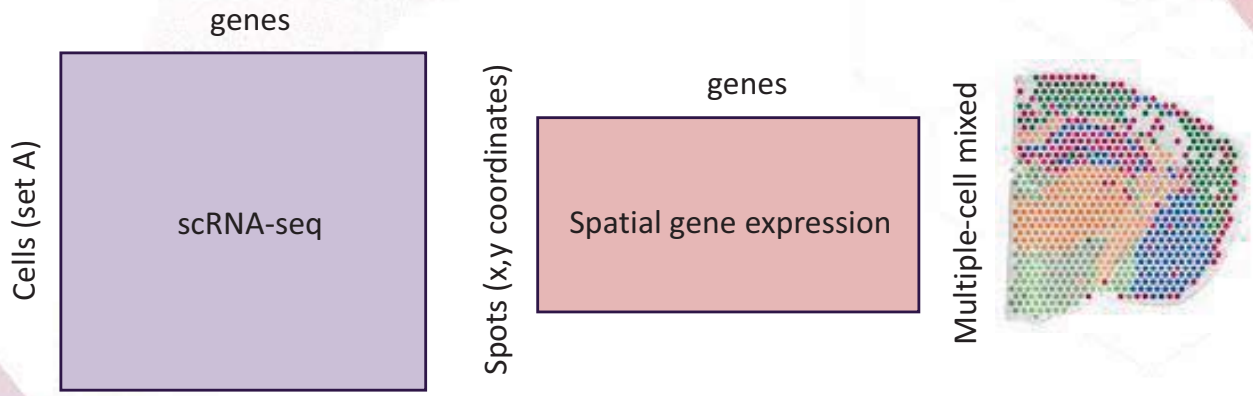


Batch effect correction problem

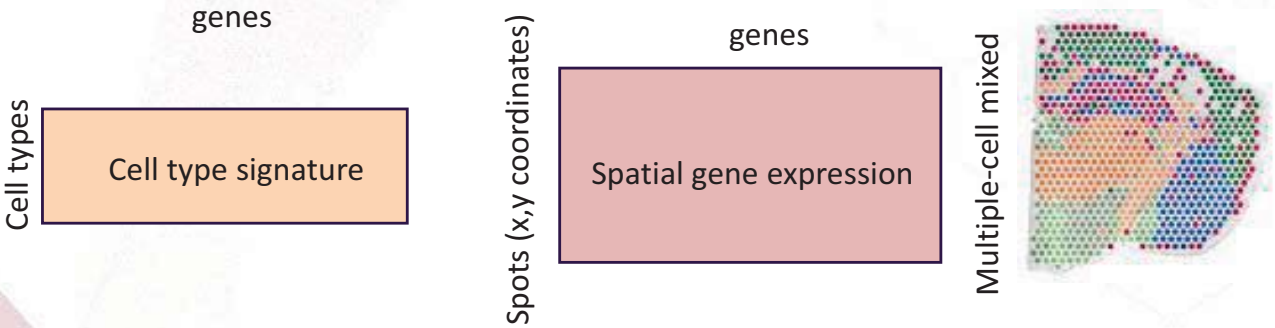
Spatial data alignment with single-cell data



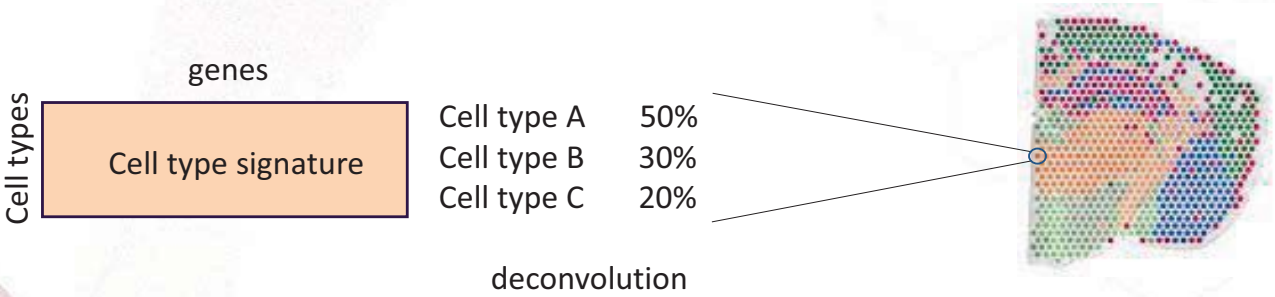
Problem of spatial data deconvolution (e.g. Visium)



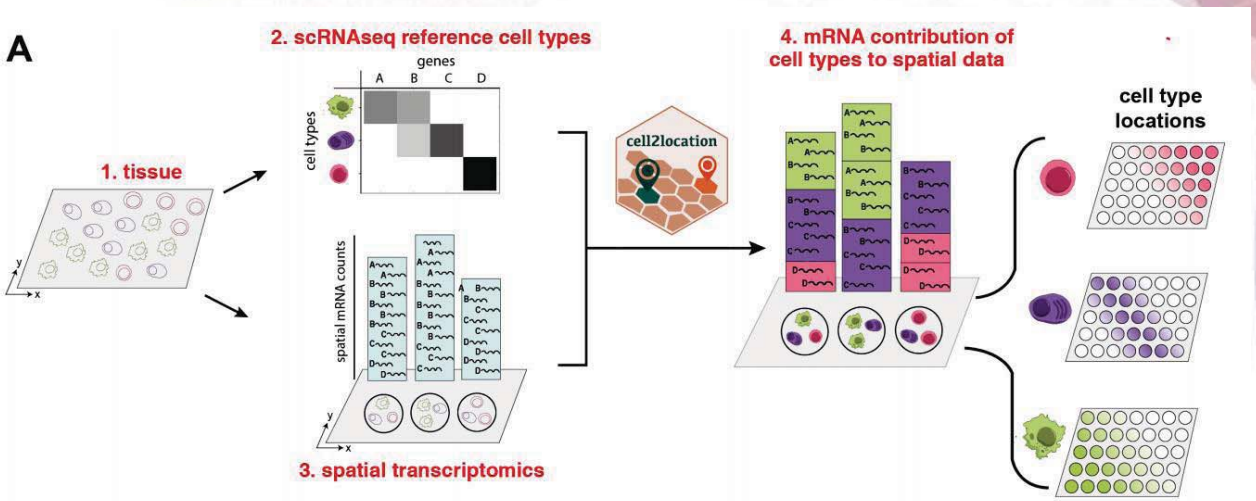
Problem of spatial data deconvolution (e.g. Visium)



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

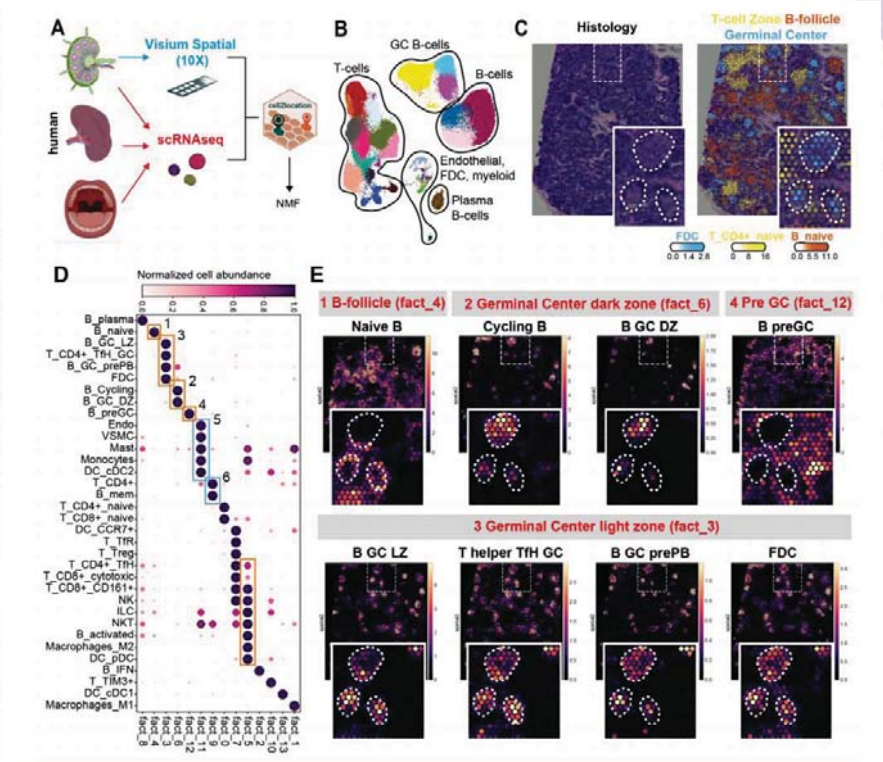


Problem of spatial data deconvolution (e.g. cell2location)



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)

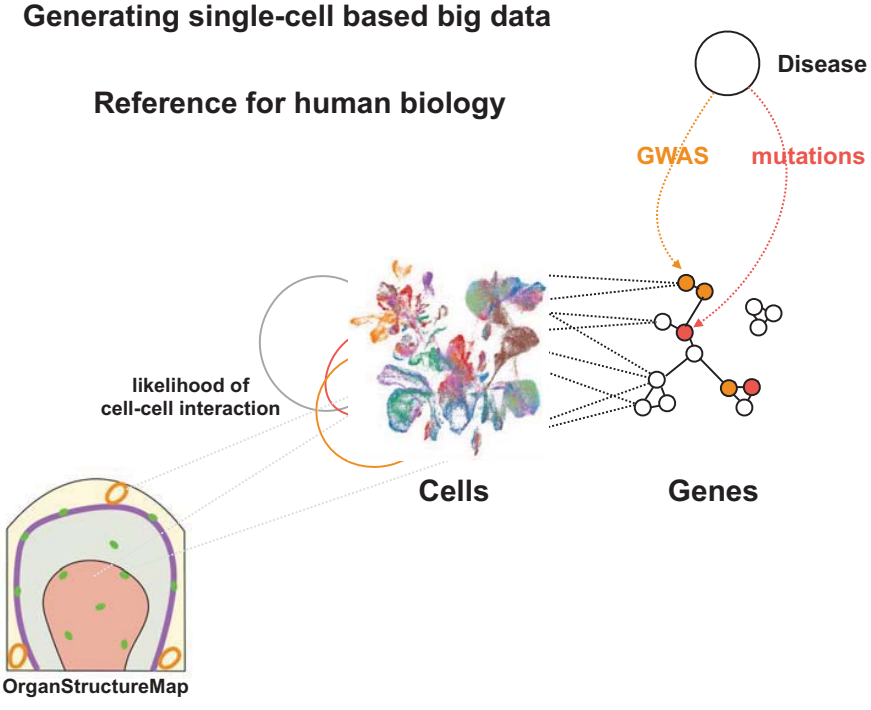


Kleshchevnikov, V., Shmatko, A., Dann, E. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* (2022).

Conclusion

Generating single-cell based big data

Reference for human biology



Link to the practice

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

171

감사합니다

172