

KSBi-BIML 2024

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

생명정보학 & 머신러닝 워크샵 (온라인)



Introduction to single cell transcriptomics analysis

김준일 _ 송실대학교



KSBI
KOREAN SOCIETY FOR
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 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 워크숍인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크숍은 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 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의를 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 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 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

Introduction to single cell transcriptomics analysis

2009년 단일세포에서 얻은 RNA sequencing 데이터가 처음 발표된 이후로 이제는 single cell RNA sequencing (scRNAseq)은 유전자발현 연구에 필수적인 도구로 자리잡고 있다. 이에 따라 데이터분석의 기본적인 파이프라인도 표준화되고 있지만 여전히 처음 접하는 연구자에게는 접근하기에 많은 어려움이 있다.

본 강의에서는 단일세포기술의 역사와 기본분석 파이프라인에 대해서 공부하고 가장 널리 쓰이는 Seurat과 Scanpy 사용법에 대해 설명한다. 이론강의에서 분석파이프라인이 만들어진 배경에 대해서 학습하고 그 원리를 탐구한 후에 예제 데이터를 통하여 Seurat과 Scanpy의 구성과 활용법에 대해서 익히고 연구자들이 생산한 데이터에 적용하여 분석하고 해석할 수 있는 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- 단일세포 기술의 역사
- scRNAseq 데이터분석 기본 파이프라인 (품질관리, 정규화, 차원축소, 시각화, 클러스터링)
- Seurat를 활용한 scRNAseq 데이터 분석
- Scanpy를 활용한 scRNAseq 데이터 분석

* 참고강의교재:

Current best practices in single-cell RNA-seq analysis: a tutorial / Malte Luecken, Fabian Theis

* 교육생준비물:

Seurat (R package), Scanpy (Python package)가 설치된 노트북 (메모리 16GB 이상)

* 강의 난이도: 중급

* 강의: 김준일교수 (송실대학교 컴퓨터공학부)

Curriculum Vitae

Speaker Name: Junil Kim, Ph.D.



► Personal Info

Name Junil Kim
Title Assistant Professor
Affiliation Soongsil University

► Contact Information

Address 369, Sangdo-Ro, Dongjak-Gu, Seoul, 06978
Email junilkim@ssu.ac.kr
Phone Number 010-3140-6567

Research Interest

Single Cell Genomics, Systems Biology, Network Biology

Educational Experience

2005 B.S. in Bioinformatics, Soongsil University, Republic of Korea
2008 M.S. in Bioinformatics, Seoul National University, Republic of Korea
2014 Ph.D. in Bio and Brain Engineering, KAIST, Republic of Korea

Professional Experience

2014-2016 Postdoctoral Researcher, CHA Cancer Institute, CHA University, Republic of Korea
2016-2018 Postdoctoral Researcher, Perelman School of Medicine, University of Pennsylvania, USA
2018-2021 Postdoctoral Researcher, BRIC, University of Copenhagen, Denmark
2021- Assistant Professor, School of Systems Biomedical Science, Soongsil University, Republic of Korea

Selected Publications (5 maximum)

- Guangzheng Weng, **Junil Kim***, and Kyoung Jae Won*, "VeTra: a tool for trajectory inference based on RNA velocity", *Bioinformatics* (IF: **6.937**), btab364, May 2021. (*Co-corresponding authors)
- Junil Kim**, Simon T. Jakobsen, Kedar N. Natarajan, Kyoung Jae Won, "TENET: gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data", *Nucleic Acids Research* (IF: **16.971**), Vol. 49, No. 1, e1-e1, Jan. 2021.
- Shibiao Wan, **Junil Kim**, Kyoung Jae Won, "SHARP: hyper-fast and accurate processing of single-cell RNA-seq data via ensemble random projection", *Genome Research* (IF: **9.043**), Vol. 30, Issue 2, 205-213, Jan. 2020. (Google Scholar Citations: **11** / Web of Science Citations: **4**)
- Junil Kim**, Diana E. Stanescu, and Kyoung Jae Won, "CellBIC: Bimodality-based top-down clustering of single-cell RNA sequencing data reveals hierarchical structure of the cell type", *Nucleic Acids Research* (IF: **16.971**), Vol. 46, Issue 21, e124, Aug. 2018. (Google Scholar Citations: **8** / Web of Science Citations: **4**)
- Junil Kim**, Sang-Min Park, and Kwang-Hyun Cho, "Discovery of a kernel for controlling biomolecular regulatory networks", *Scientific Reports*, Vol. 3, 2223, July 2013. (Google Scholar Citations: **87** / Web of Science Citations: **44**)

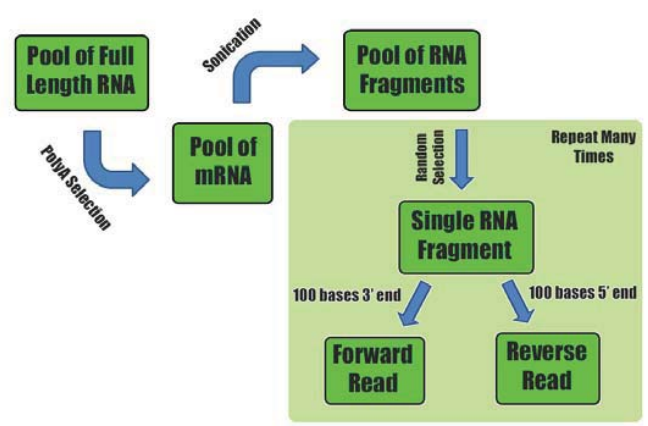
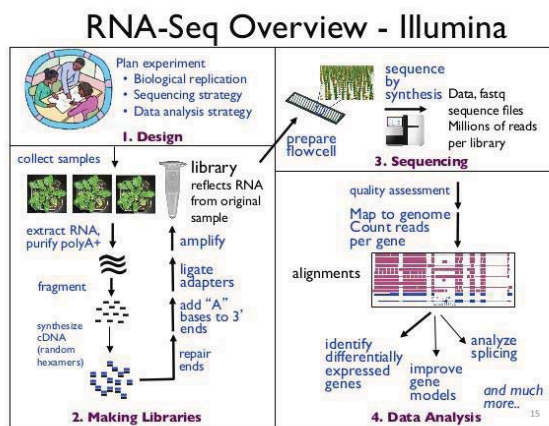
KSBi-BIML

Introduction to single cell transcriptomics analysis

Contents

1. Review of Bulk RNA sequencing
2. Why single cell?
3. Single cell technology
4. What can we do using single cell transcriptomics?
 - Dimension reduction and clustering (cell type identification)
 - Dynamics (Pseudotime, RNA velocity, GRN analysis)
 - Beyond single cell (Cell-cell interaction, Spatial transcriptomics)
5. Single cell transcriptomics analysis pipeline
 - Quality control and normalization
 - Feature selection
 - Dimension reduction and visualization
 - Clustering and cell type annotation

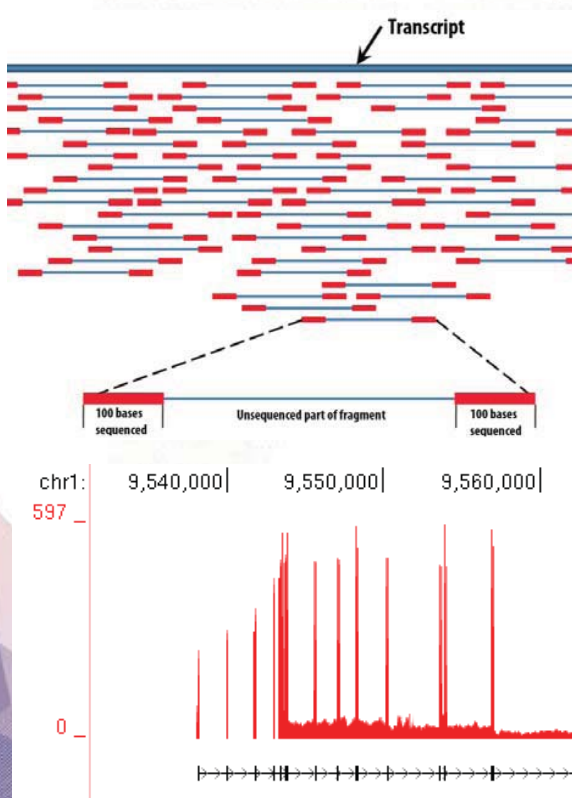
1. Review of Bulk RNA sequencing: mRNA sequencing exactly measures the quantity of mRNA molecule



- Full length mRNA cannot yet be sequenced routinely (Illumina).
 - Only short fragments can be sequenced accurately and cheaply.
- RNA are fragmented into small pieces, typically 200 - 500 bases.
- Approximately **100 bases** are sequenced from one, or both, ends of the fragments.

3

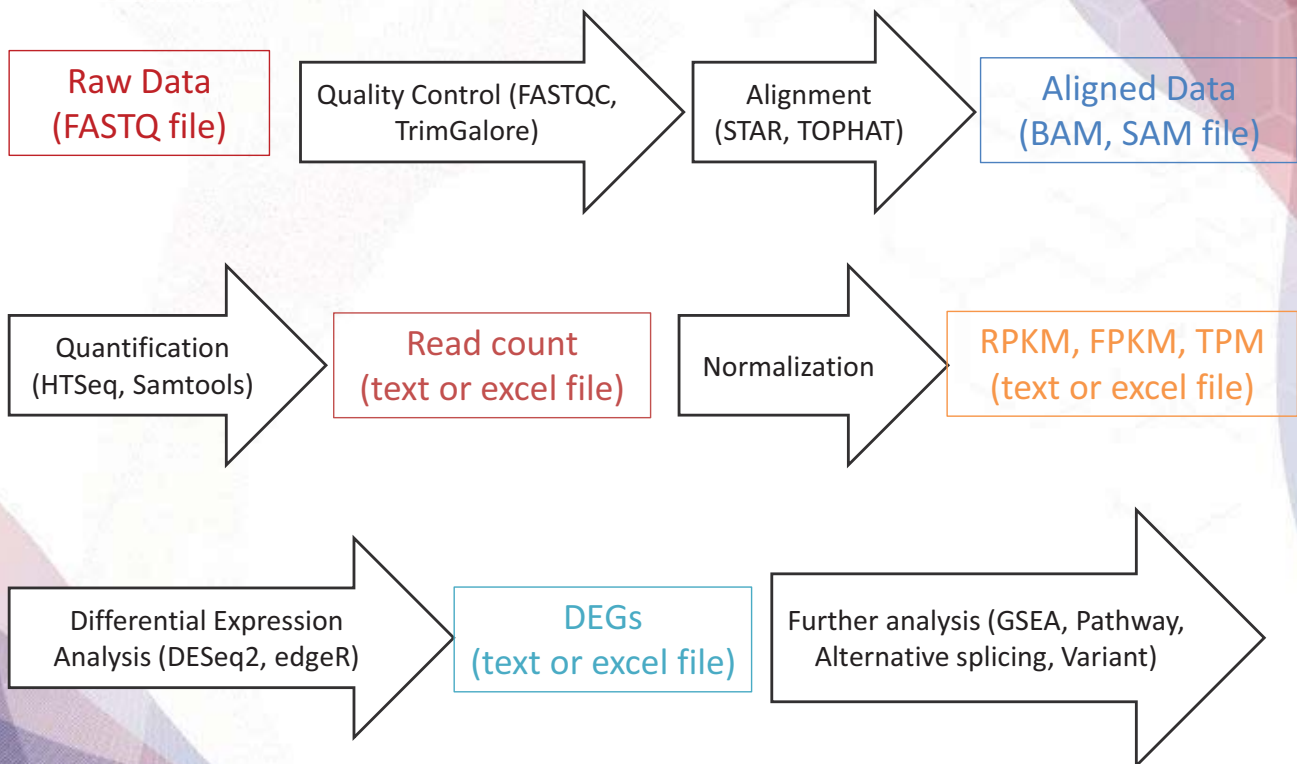
1. Review of Bulk RNA sequencing



- Reads are aligned to the genome.
- Data are represented as “depth of coverage” plots.
 - The height of the bar over a nucleotide is the number of reads which align across that location.
- The higher a gene is expressed, the more reads we find for that gene.
- The higher the peak, the higher the gene is expressed.

4

1. Review of Bulk RNA sequencing: RNA-seq analysis pipeline



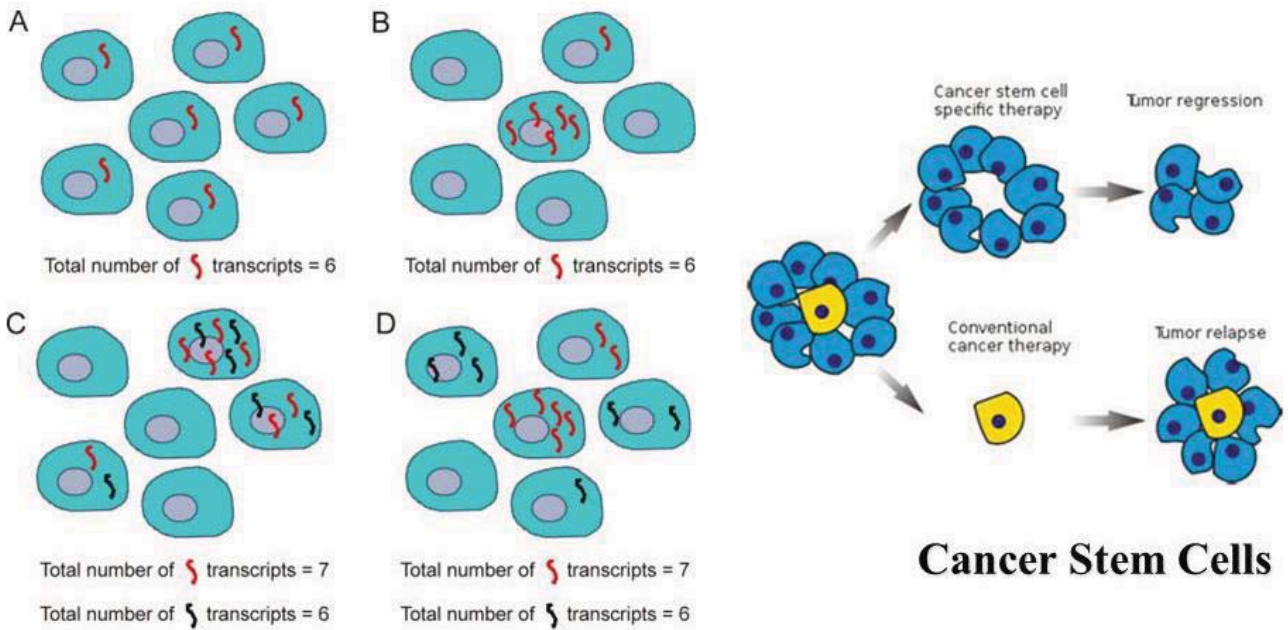
5

2. Why single cell?

- Cell identity and function can be characterized at the molecular level by unique transcriptome signatures (Stegle et al. Nature Reviews Genetics 2015)
- Ensemble-based approaches (bulk RNA-seq) only provide an average of each gene's expression across a large population of cells.
- What we cannot see using bulk measures?
 1. Heterogeneity – early embryo, complex tissues such as brain tissues
 2. Cellular composition – Differential gene expression between samples may be driven by cellular composition
 3. Stochasticity – gene expression bursts probabilistically

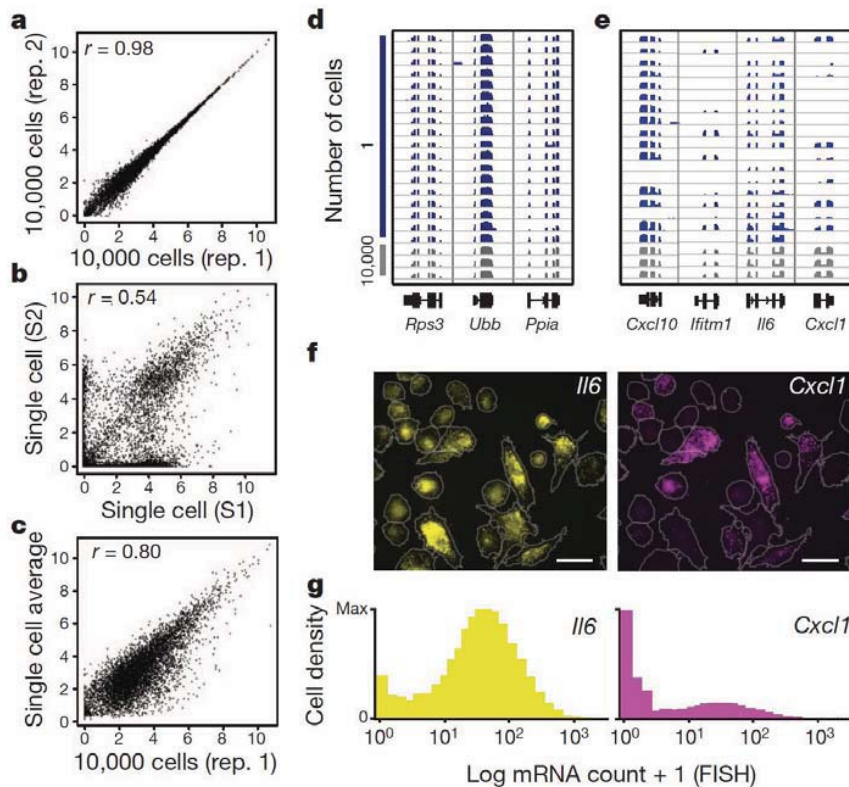
6

2. Why single cell? Bulk RNA-seq cannot be used for cellular heterogeneity



7

2. Why single cell? scRNA-seq uncover bulk cell expression patterns as well as single-cell-level heterogeneity

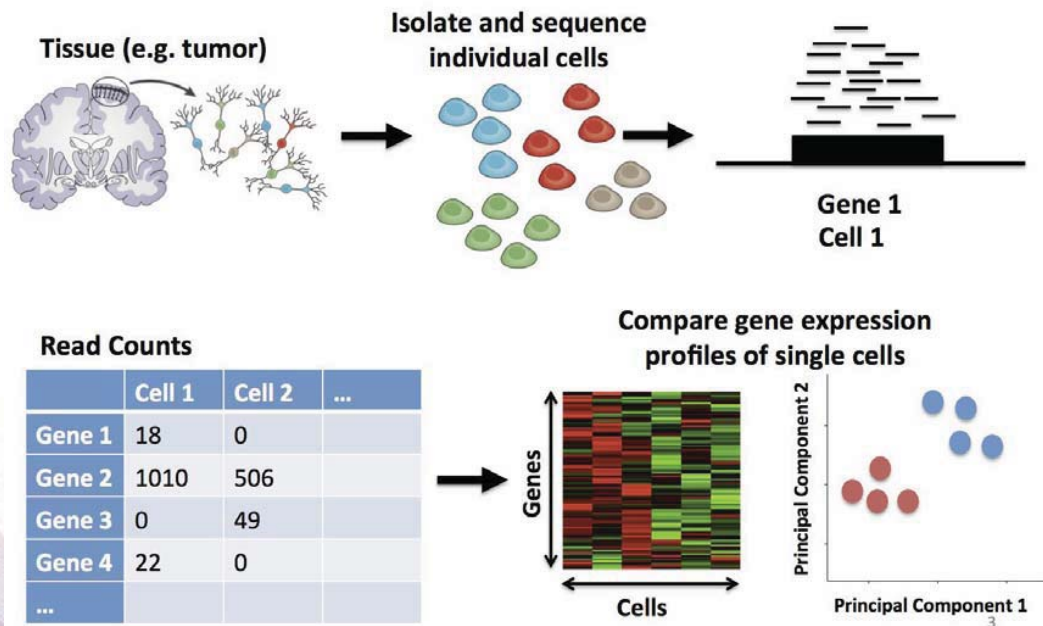


Shalek et al. (2013) Nature 498, 236-240

8

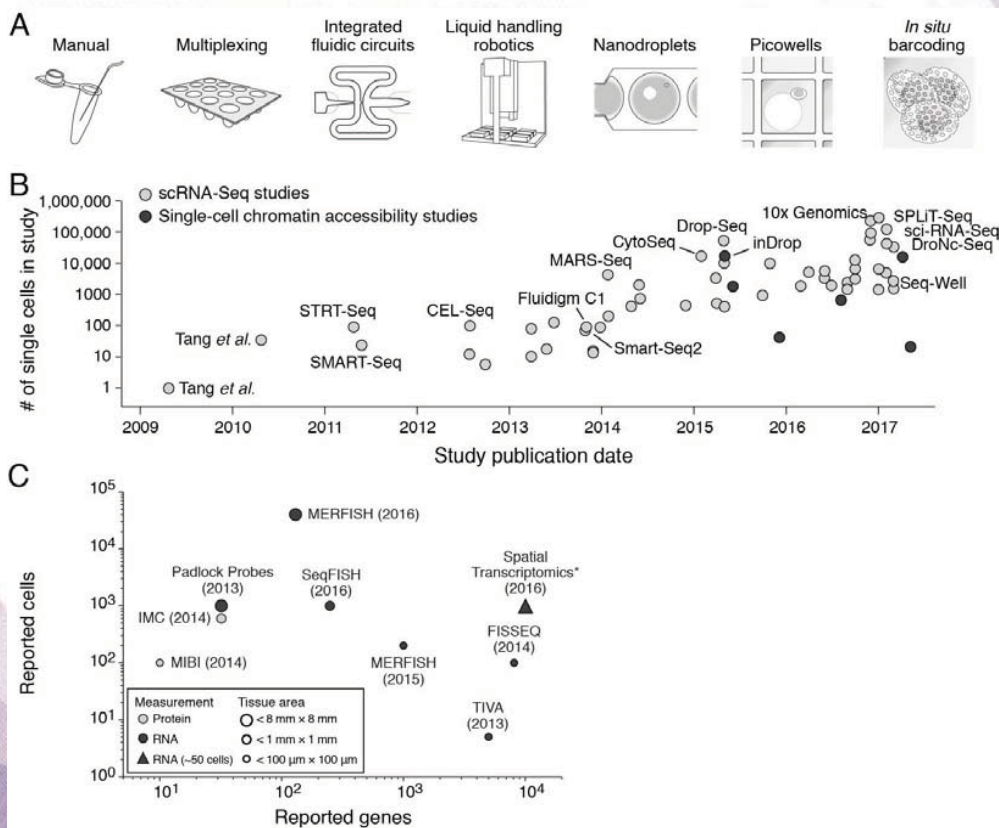
3. Single cell technology

Single-cell RNA-Seq (scRNA-Seq)



9

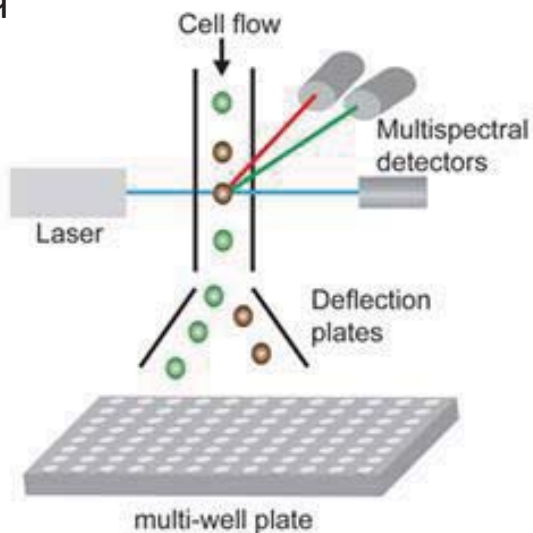
3. Single cell technology



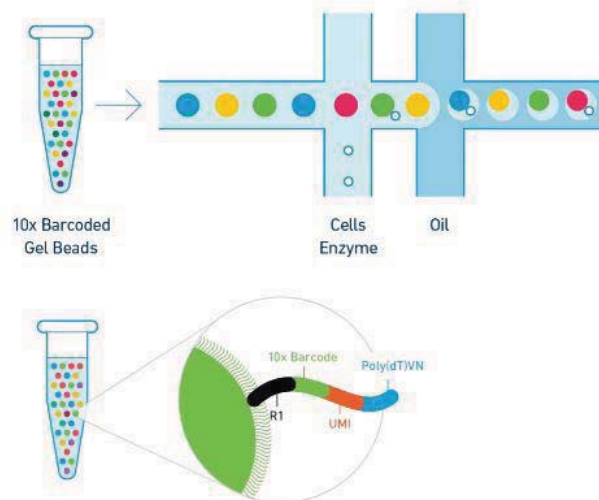
10

3. Single cell technology

Smart-seq: Well-based scRNA-seq



10x Genomics: Microfluidic droplet-based scRNA-seq



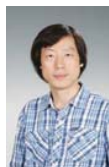
11

3. Single cell technology: First single cell RNA-seq paper (2009)

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang^{1,3}, Catalin Barbacioru^{2,4}, Yangzhou Wang², Ellen Nordman², Clarence Lee², Nanlan Xu², Xiaohui Wang², John Bodeau², Brian B Tuch², Asim Siddiqui², Kaiqin Lao² & M Azim Surani¹

Next-generation sequencing technology is a powerful tool for transcriptome analysis. However, under certain conditions, only a small amount of material is available, which requires more sensitive techniques that can preferably be used at the single-cell level. Here we describe a single-cell digital gene expression profiling assay. Using our mRNA-Seq assay with only a single mouse blastomere, we detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads. Moreover, 8–19% of the genes with multiple known transcript isoforms expressed at least two isoforms in the same blastomere or oocyte, which unambiguously demonstrated the complexity of the transcript variants at whole-genome scale in individual cells. Finally, for *Dicer1*^{-/-} and *Ago2*^{-/-} (*Eif2c2*^{-/-}) oocytes, we found that 1,696 and 1,553 genes, respectively, were abnormally upregulated compared to wild-type controls, with 619 genes in common.



Fuchou Tang



Azim Surani

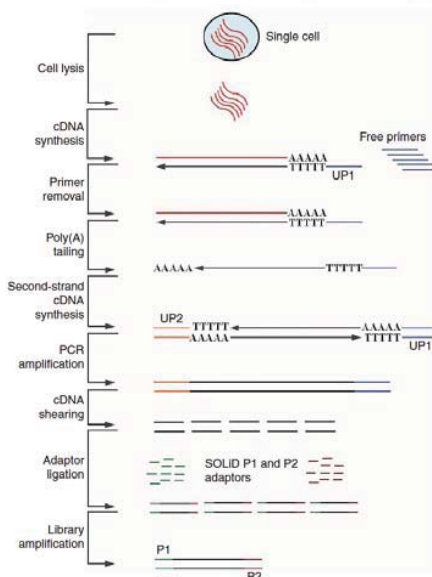


Figure 1 | Schematic of the single-cell whole-transcriptome analysis. A single cell is manually picked under a microscope and lysed. Then mRNAs are reverse-transcribed into cDNAs using a poly(T) primer with anchor sequence (UP1) and unused primers are digested. Poly(A) tails are added to the first-strand cDNAs at the 3' end, and second-strand cDNAs are synthesized using poly(T) primers with another anchor sequence (UP2). Then cDNAs are evenly amplified by PCR using UP1 and UP2 primers, fragmented, and P1 and P2 adaptors are ligated to the ends. Finally, emulsion PCR is performed by mixing libraries with 1 µm diameter beads with P1 primers covalently attached to their surfaces.

Question: How to capture the mRNA molecule in a small amount of material?

12

3. Single cell technology: STRT (single cell tagged reverse transcription) – Multiplexing (2011)

Method

Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq

Saiful Islam,^{1,4} Una Kjällquist,^{1,4} Annalena Moliner,² Pawel Zajac,¹ Jian-Bing Fan,³ Peter Lönnerberg,¹ and Sten Linnarsson^{1,5}

¹Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77 Stockholm, Sweden; ²Department of Neuroscience, Karolinska Institutet, SE-171 77 Stockholm, Sweden; ³illumina inc., San Diego, California 92121, USA

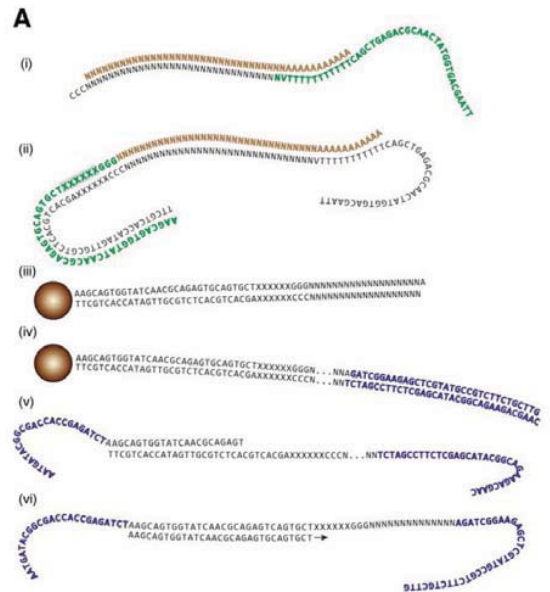
Our understanding of the development and maintenance of tissues has been greatly aided by large-scale gene expression analysis. However, tissues are invariably complex, and expression analysis of a tissue confounds the true expression patterns of its constituent cell types. Here we describe a novel strategy to access such complex samples. Single-cell RNA-seq expression profiles were generated, and clustered to form a two-dimensional cell map onto which expression data were projected. The resulting cell map integrates three levels of organization: the whole population of cells, the functionally distinct subpopulations it contains, and the single cells themselves—all without need for known markers to classify cell types. The feasibility of the strategy was demonstrated by analyzing the transcriptomes of 85 single cells of two distinct types. We believe this strategy will enable the unbiased discovery and analysis of naturally occurring cell types during development, adult physiology, and disease.

Figure 1. Single-cell tagged reverse transcription (STRT). (A) Overview of the method, illustrating the main steps in sample preparation: (i) mRNA (brown) is reverse transcribed using a tailed oligo-dT primer (green), generating a first-strand cDNA with 3-6 added cytosines; (ii) a helper oligo (green) causes template-switching and thereby introduces a barcode (shaded) and a primer sequence into the cDNA; (iii) the product is amplified by single-primer PCR exploiting the template-suppression effect and is then immobilized on beads, fragmented, and A-tailed; (iv) the Illumina P2 adapter (blue) is ligated to the free end; (v) the P1 adapter is introduced in the library PCR step, using a primer tailed with the P1 sequence (blue); and (vi) the final library is sequenced from the P1 side using a custom primer. Each read (arrow) begins by the barcode, followed by three to six Cs, followed by the mRNA insert. (B) Illustration of read mapping and annotation, for a two-exon gene. Reads mapping to the sense strand of exons, as well as to splice junctions, were counted toward the expression of the gene. Reads mapping upstream of, downstream from, or in introns were counted for quality control purposes, and anti-sense hits were used to judge the background level.



Saifu Islam

Sten Linnarsson



Multiplexing using barcode reduces cell-to-cell amplification bias since single-cell cDNA was pooled before amplification

3. Single cell technology: SMART (switching mechanism at the 5' end of RNA templates) (2012)

Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells

Daniel Ramsköld^{1,2,7}, Shujun Luo^{3,7}, Yu-Chieh Wang⁴, Robin Li³, Qiaolin Deng¹, Omid R Faridani¹, Gregory A Daniels⁵, Irina Khrebtkova³, Jeanne F Loring⁴, Louise C Laurent⁶, Gary P Schroth³ & Rickard Sandberg^{1,2}

Genome-wide transcriptome analyses are routinely used to monitor tissue-, disease- and cell type-specific gene expression, but it has been technically challenging to generate expression profiles from single cells. Here we describe a robust mRNA-Seq protocol (Smart-Seq) that is applicable down to single cell levels. Compared with existing methods, Smart-Seq has improved read coverage across transcripts, which enhances detailed analyses of alternative transcript isoforms and identification of single-nucleotide polymorphisms. We determined the sensitivity and quantitative accuracy of Smart-Seq for single-cell transcriptomics by evaluating it on total RNA dilution series. We found that although gene expression estimates from single cells have increased noise, hundreds of differentially expressed genes could be identified using few cells per cell type. Applying Smart-Seq to circulating tumor cells from melanomas, we identified distinct gene expression patterns, including candidate biomarkers for melanoma circulating tumor cells. Our protocol will be useful for addressing fundamental biological problems requiring genome-wide transcriptome profiling in rare cells.

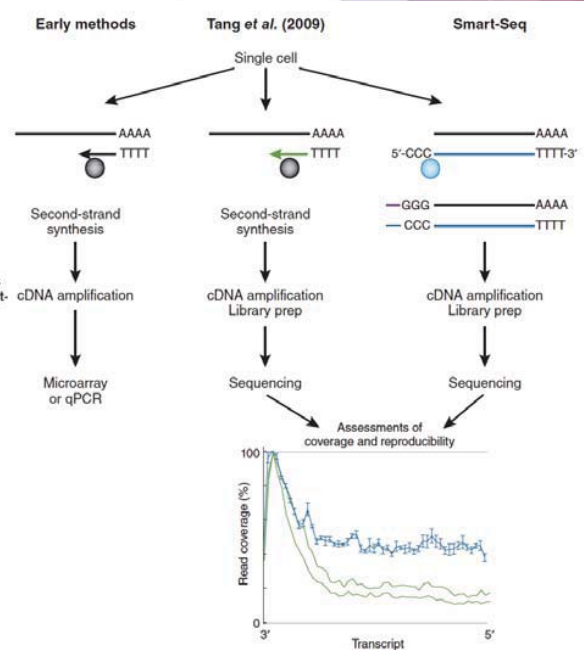
SMART-seq focus on the full-length transcript coverage. SMART-seq can achieve the coverage by using a reverse transcriptase enzyme from the Moloney murine leukemia virus (template switching and terminal transferase activity).



Daniel Ramsköld



Rickard Sandberg

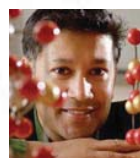


3. Single cell technology: UMI (2013)

Quantitative single-cell RNA-seq with unique molecular identifiers

Saifu Islam¹, Amit Zeisel¹, Simon Joost², Gioele La Manno¹, Pawel Zajac¹, Maria Kasper², Peter Lönnerberg¹ & Sten Linnarsson¹

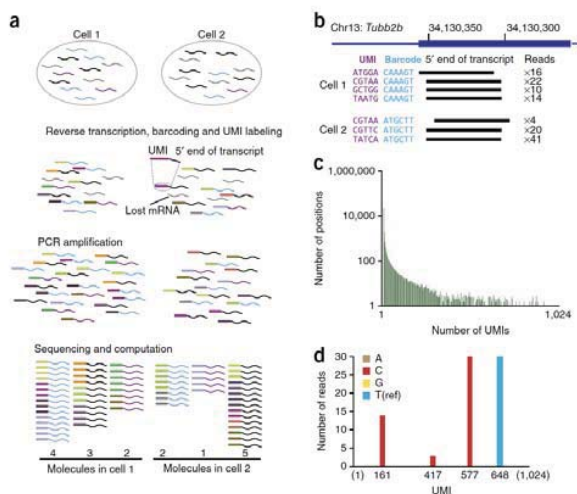
Single-cell RNA sequencing (RNA-seq) is a powerful tool to reveal cellular heterogeneity, discover new cell types and characterize tumor microevolution. However, losses in cDNA synthesis and bias in cDNA amplification lead to severe quantitative errors. We show that molecular labels—random sequences that label individual molecules—can nearly eliminate amplification noise, and that microfluidic sample preparation and optimized reagents produce a fivefold improvement in mRNA capture efficiency.



Saifu Islam



Sten Linnarsson



UMI enables to quantify the mRNA eliminating amplification noise.

15

3. Single cell technology: inDrops, cell (2015), page 1187-1201 / Drop-seq, Cell (2015), page 1202-1214

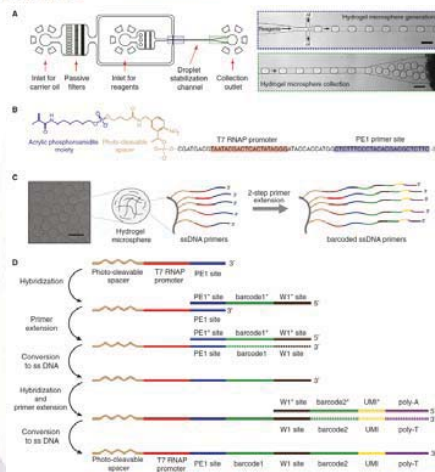
Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells

Allon Klein,^{1,2} Linas Mazutis,^{2,3,4} Ilike Akartuna,^{2,5} Naren Tallapragada,¹ Adrian Veres,^{1,4,6} Victor Li,¹ Leonid Peshkin,¹ David A. Weitz,^{2,7} and Marc W. Kirschner^{1,*}

¹Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
²Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
³School of Engineering and Applied Sciences (SEAS), Harvard University, Cambridge, MA 02138, USA
⁴Vilnius University Institute of Biotechnology, Vilnius LT-02241, Lithuania
⁵Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
⁶Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA
⁷Co-first author
^{*}Correspondence: weitz@seas.harvard.edu (D.A.W.), marc@hms.harvard.edu (M.W.K.)
<http://dx.doi.org/10.1016/j.cell.2015.04.044>



Allon Klein



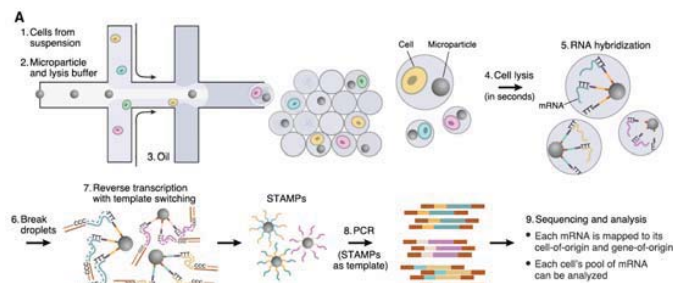
Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan M. Macosko,^{1,2,3} Anindita Basu,^{4,5} Rahul Satija,^{4,5,7} James Nemesh,^{1,2,3} Karthik Shekhar,¹ Melissa Goldman,^{1,2} Itay Trosh,¹ Allison R. Bialas,² Nolan Kamitaki,^{1,2,3} Emily M. Martersteck,^{1,2,3} John J. Trombetta,¹ David A. Weitz,^{1,2,3} Joshua R. Sanes,² Alex K. Shalek,^{4,5,12} Aviv Regev,^{4,5,12} and Steven A. McCarroll^{1,2,3,4,5,6}

¹Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
²Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
³Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
⁴Karman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
⁵School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
⁶New York Genome Center, New York, NY 10013, USA
⁷Department of Biology, New York University, New York, NY 10003, USA
⁸The Program in Cellular and Molecular Medicine, Children's Hospital Boston, Boston, MA 02115, USA
⁹Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
¹⁰Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA
¹¹Institute for Medical Engineering and Science and Department of Chemistry, MIT, Cambridge, MA 02139, USA
¹²Department of Biology, MIT, Cambridge, MA 02139, USA
¹³Howard Hughes Medical Institute, Chevy Chase, MD 20915, USA
^{*}Correspondence: emacosko@genetics.med.harvard.edu (E.M.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)
<http://dx.doi.org/10.1016/j.cell.2015.05.002>



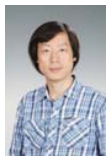
Evan Macosko



Drop-seq enables to measure transcriptome of tens of thousands of cells by using aqueous droplets and barcoding system with UMI.

16

3. Single cell technology (history)



Fuchou Tang



Sten Linnarsson



Rickard Sandberg

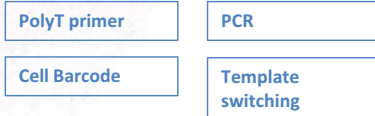


Itai Yanai

2009: First single cell RNA-seq paper



2011: STRT (single cell tagged reverse transcription) - Multiplexing



2012: SMART (switch mechanism at the 5' end of RNA templates)-seq



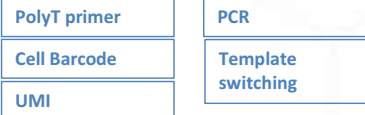
2012: CEL-seq



2013: Smart-seq2



2013: UMI



2014: MARS (Massively parallel RNA single-cell)-seq



2015: inDrops, Drop-seq



Rickard Sandberg



Sten Linnarsson



Ido Amit



Allon Klein

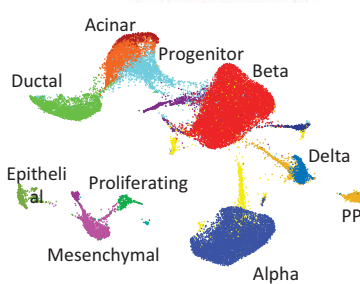


Evan Macosko

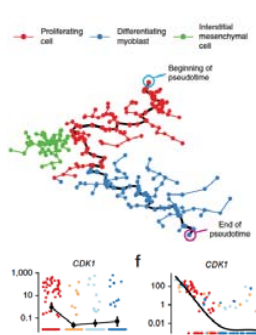
4. What can we do using single cell transcriptomics?

- Heterogeneity → Cell type identification
- Dynamics → Cell type differentiation trajectory / Gene regulatory network reconstruction → Identifying key molecules regulating cellular behavior
- Interactions → Cellular communication

Dimension reduction and Clustering
Pancreatic cells

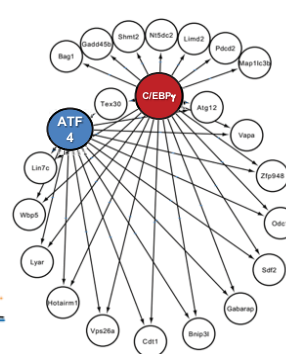


Pseudotime

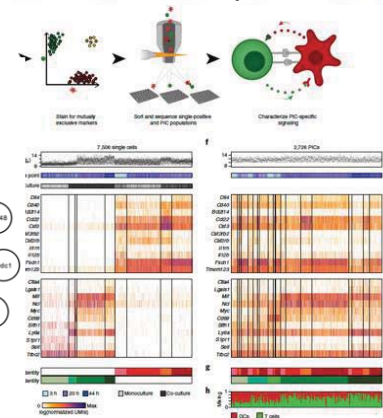


Trapnell et al., Nature Biotech. (2014)

GRN

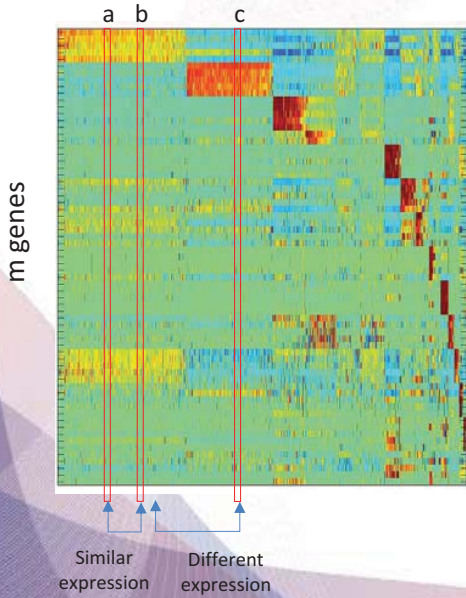


PIC-seq

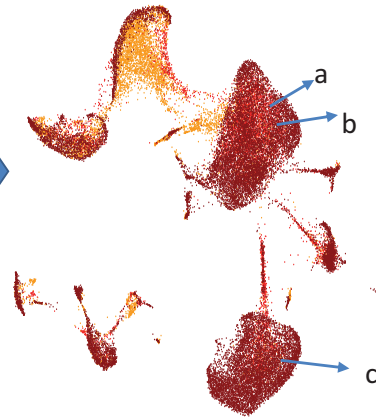


Giladi et al., Nature Biotech. (2020)

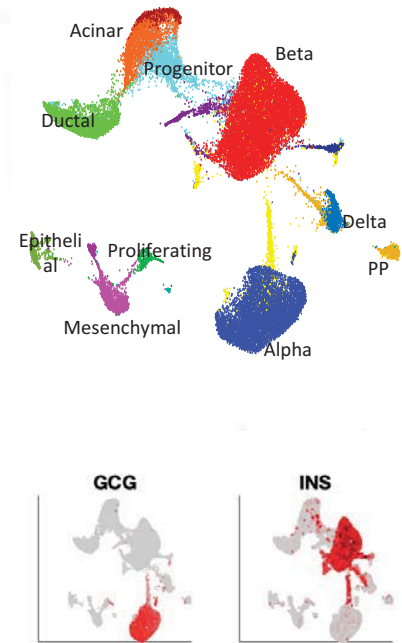
4. What can we do using single cell transcriptomics? – Dimension reduction and clustering



Dimension reduction (PCA, tSNE, UMAP)

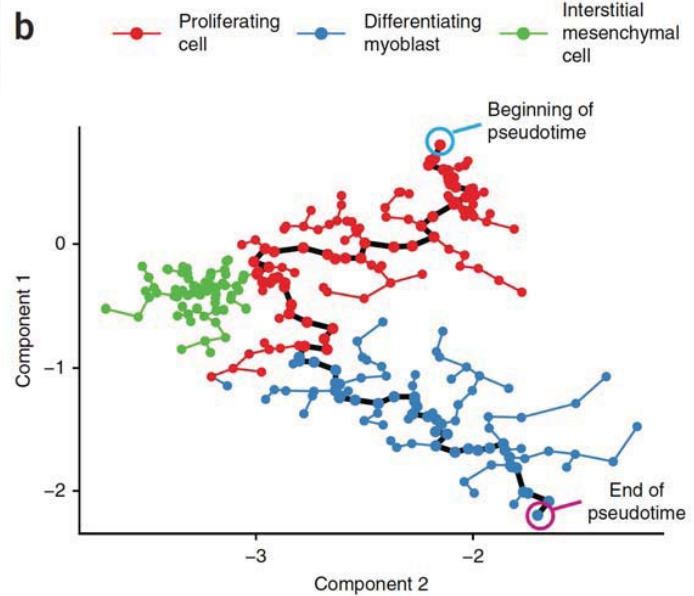
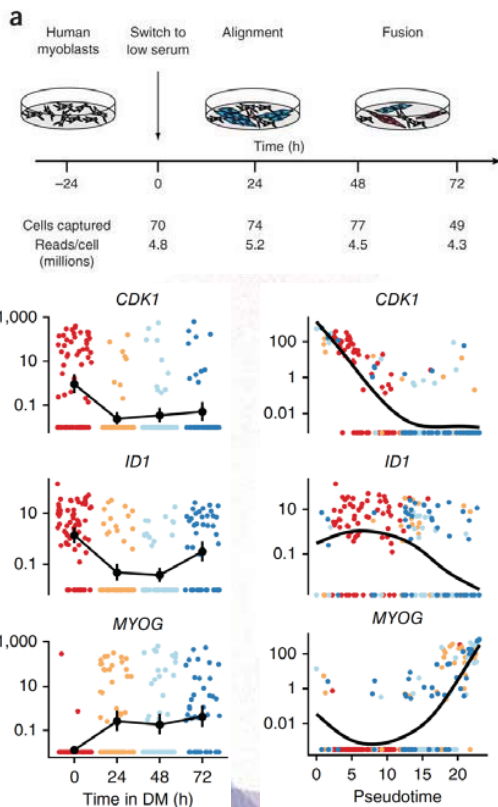


Clustering and cell type annotations



19

4. What can we do using single cell transcriptomics? – Dynamics: Pseudotime analysis

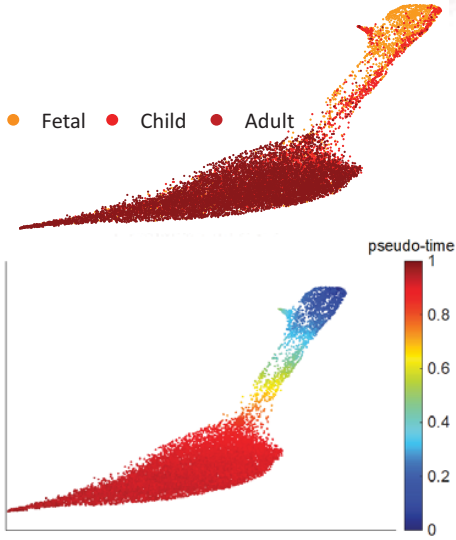


Trapnell et al., Nature Biotech. (2014)

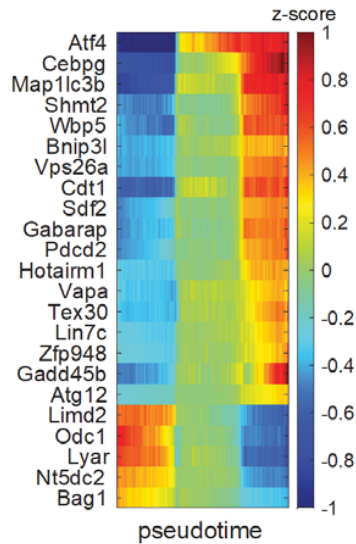
20

4. What can we do using single cell transcriptomics? – Dynamics: Pseudotime analysis

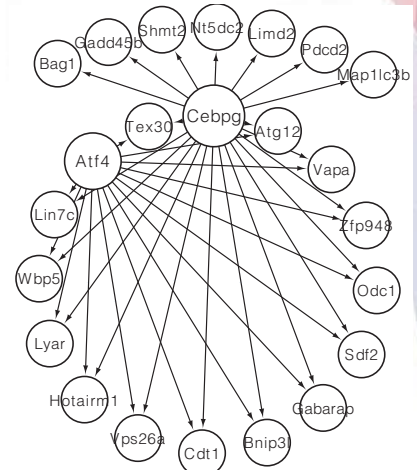
Pseudotime analysis



Gene-gene causal relationships



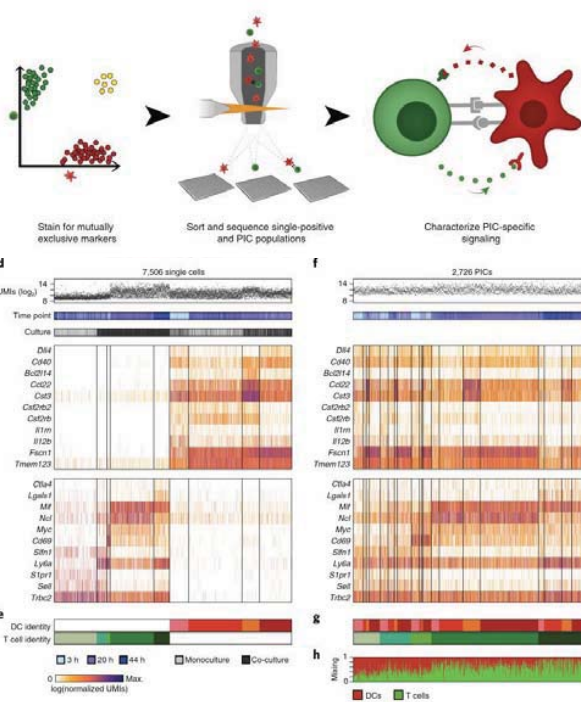
Gene regulatory networks



21

4. What can we do using single cell transcriptomics? – Beyond single cell

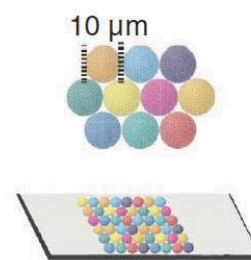
PIC-seq (two-cells)



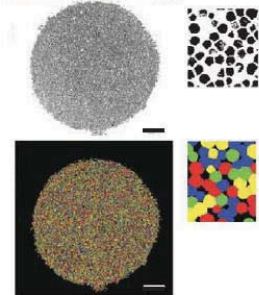
Giladi et al., Nature Biotech. (2020)

Spatial transcriptomics

A Bead deposition

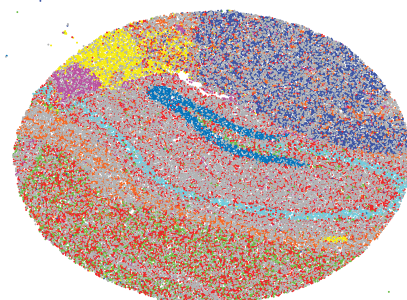


In situ indexing



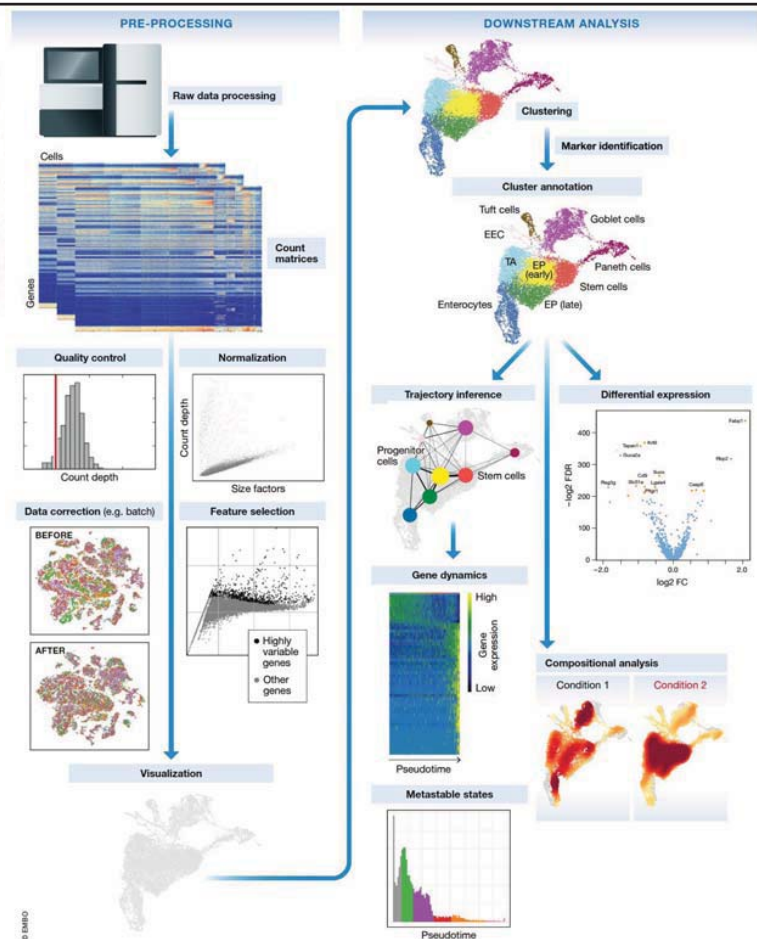
Rodrigues et al., Science (2019)

Mouse hippocampus data



22

5. Single cell transcriptomics data analysis pipeline



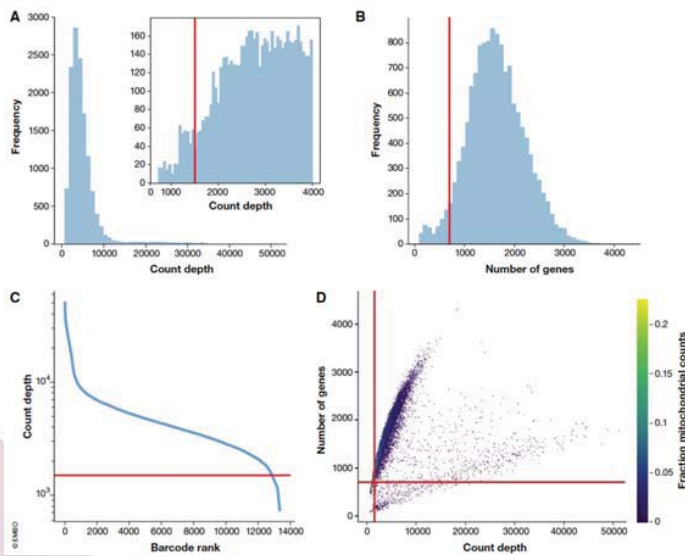
23

5. Single cell transcriptomics data analysis pipeline: quality control

- For cell QC, we must filter out dying cells or doublets.
- Three QC criterion
 1. The **number of counts** per barcode
 2. The **number of genes** per barcode
 3. The fraction of counts from **mitochondrial gene** per barcode
- Example 1: cells with low counts, few detected genes, and a high fraction of mitochondrial genes may represent dying cells
- Example 2: cells with high counts and a large number of detected genes may represent doublets

24

5. Single cell transcriptomics data analysis pipeline: quality control

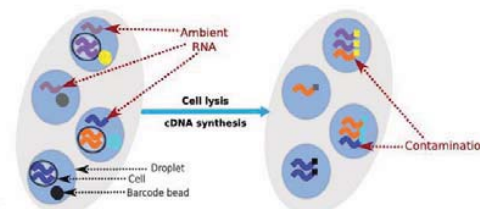


- Considering only one criteria can lead to misinterpretation of cellular signals
 - High fraction of mitochondrial counts may be involved in respiratory processes
 - Low counts and genes may correspond to quiescent cell populations
 - High counts and genes may correspond large cell size (adipocytes)

25

5. Single cell transcriptomics data analysis pipeline: quality control

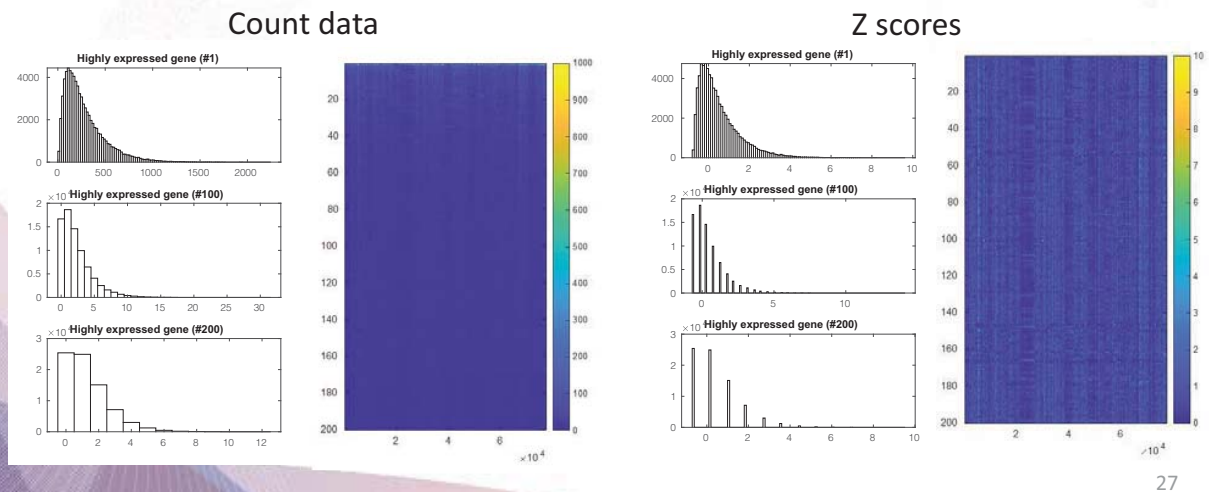
- For **gene QC**, we can filter out genes that are not expressed in more than a few cells
 - You need to consider the minimum number of cell cluster size of your interest when you choose the threshold.
 - For example, you need to lower the threshold, when you tried to find a rare cell type.
- Ambient gene expression** (counts that do not originate from a barcoded cell, but from other lysed cells whose mRNA contaminated the cell suspension prior to library construction)
 - For example, super-highly expressed genes can be detected in an empty droplet. SoupX and DecontX provide a correction for this contamination in droplet-based scRNA-seq datasets



26

5. Single cell transcriptomics data analysis pipeline: normalization

- **Gene normalization / scaling (z scores):** scaling gene counts to have zero mean and unit variance. There is no consensus on whether or not to do this.
 - Should all genes be weighted equally? versus
 - Is the magnitude of expression of a gene an informative for the importance of the gene?



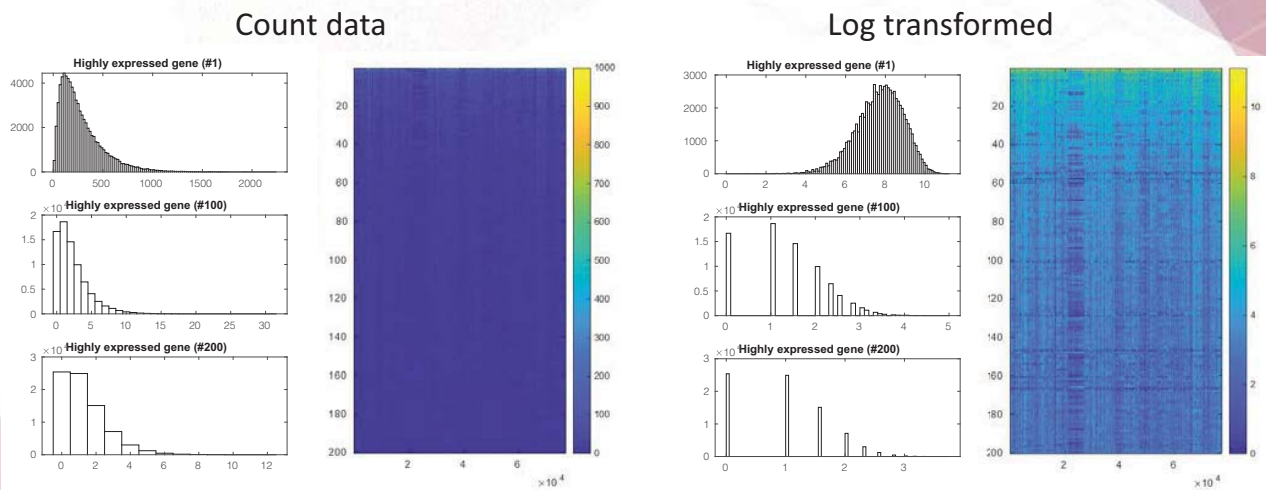
27

5. Single cell transcriptomics data analysis pipeline: normalization

- **Log normalization:** $\log_2(x+1)$ This is a useful tool for the downstream analysis (DEG, clustering, etc.)! →
 1. Distance between log-transformed expression values represent **log fold changes**
 2. Log transformation mitigates the **mean-variance relationship**. Expression variance of a gene can be considered as the importance of the gene. Then, the importance of lowly-expressed genes can be ignored.
 3. Log transformation reduces the **skewness** of the data to approximate the **assumption** of many downstream analysis tools that the data are **normally distributed**

28

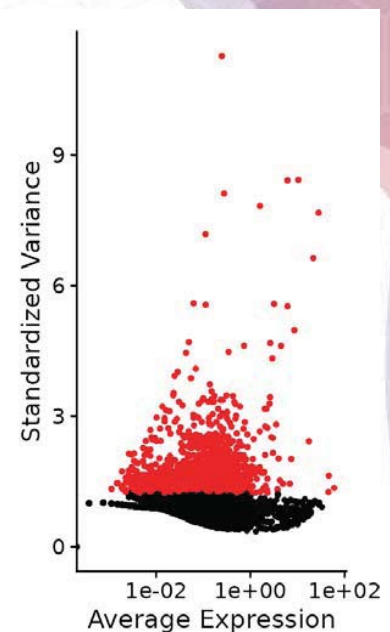
5. Single cell transcriptomics data analysis pipeline: normalization



29

5. Single cell transcriptomics data analysis pipeline: feature selection

- Depending on the task and the complexity of the dataset, typically between 1,000 and 5,000 Highly Variable Genes (HVGs) are selected.
- Seurat procedure
 1. Genes are binned by their **mean expression of count** data
 2. The genes with the highest **variance-to-mean ratio** are selected as HVGs



30

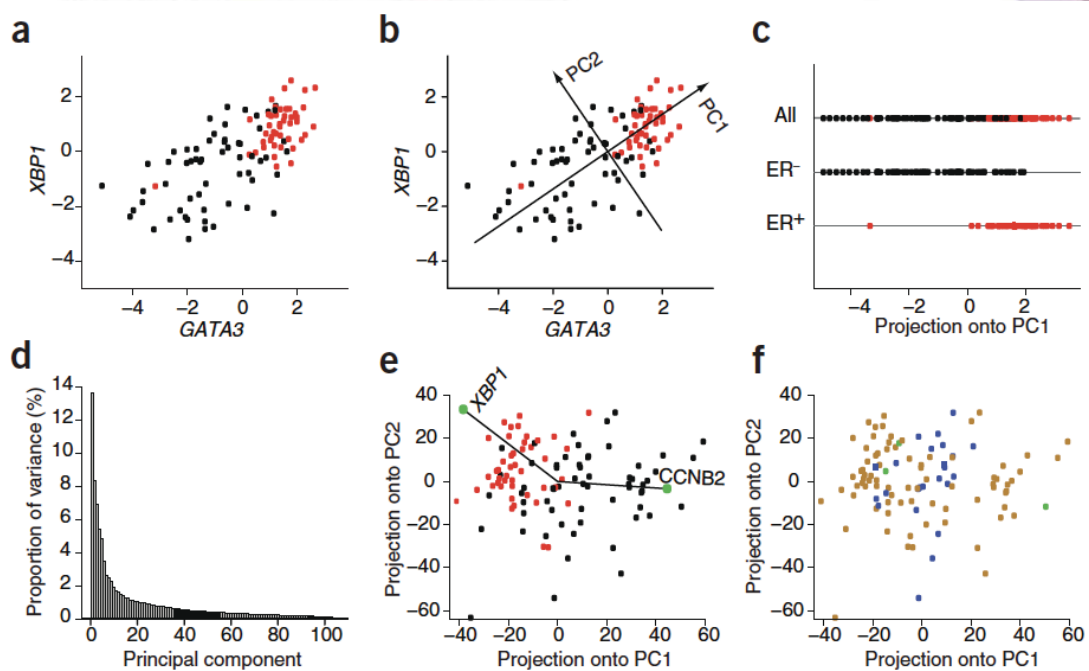
5. Single cell transcriptomics data analysis pipeline: dimension reduction and visualization

- Dimensionality reduction algorithms **embed** the expression matrix into a **low-dimensional space**, which is designed to capture the **underlying structure** in the data in as few dimensions as possible.
- Two main objectives:
 1. Visualization: optimally describe the dataset in 2D or 3D
 2. Summarization: can be used to reduce the data to its essential components by finding the inherent dimensionality
- 2D visualization cannot be used for summarization (**Visualization ≠ Summarization**)
- Linear method: **PCA**
- Non-linear methods: **t-SNE**, **Diffusion maps**, **UMAP**, **SPRING's force-directed layout**.

*Dimensionality reduction
#Visualization

31

5. Single cell transcriptomics data analysis pipeline: dimension reduction and visualization - PCA

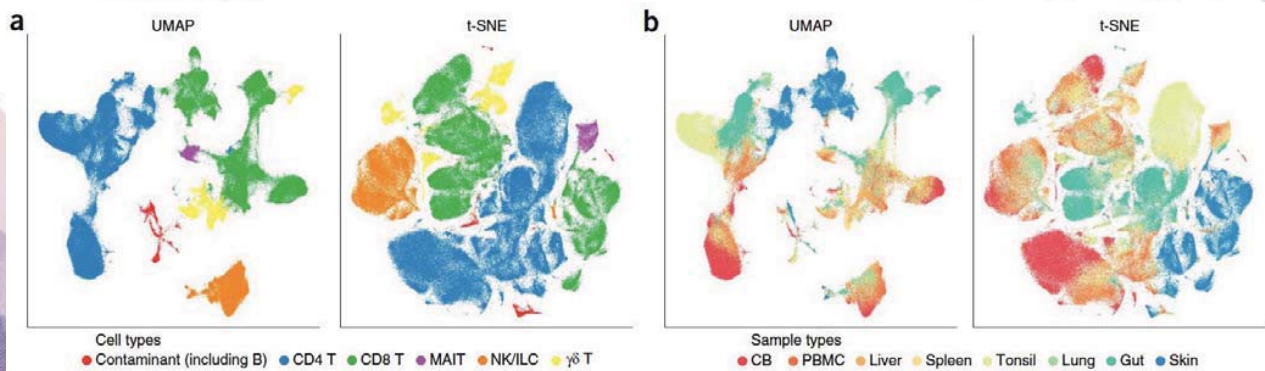


Ringner, "What is principal component analysis?", Nature Biotechnology 2008

32

5. Single cell transcriptomics data analysis pipeline: dimension reduction and visualization - PCA

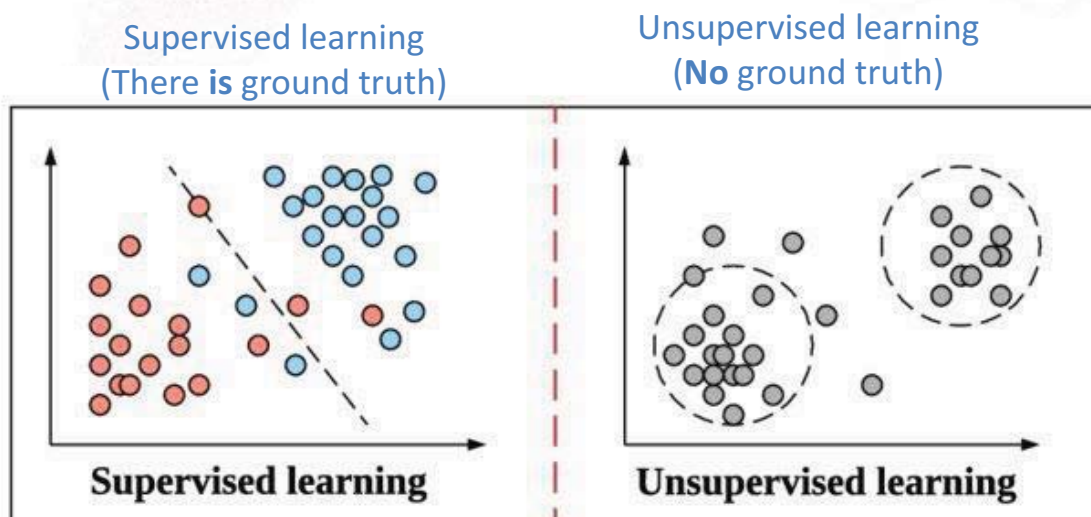
- t-SNE suffers from limitations such as loss of large-scale information (the inter-cluster relationships) and inability to meaningfully represent very large datasets.
- Both UMAP and t-SNE were successfully pulling together only clusters corresponding to similar cell populations.
- However, t-SNE separated cell populations into distinct clusters than UMAP.



33

5. Single cell transcriptomics data analysis pipeline: clustering

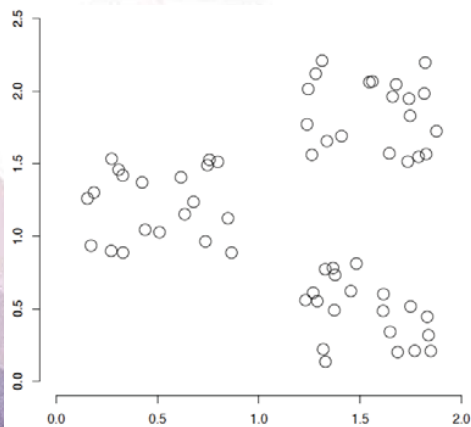
- Clustering allows us to infer the **identity** of member cells.
- Clusters are obtained by grouping cells based on the **similarity** (or **distance**) of their gene expression profiles.
- Clustering is a classical **unsupervised machine learning** problem, based directly on a distance matrix.



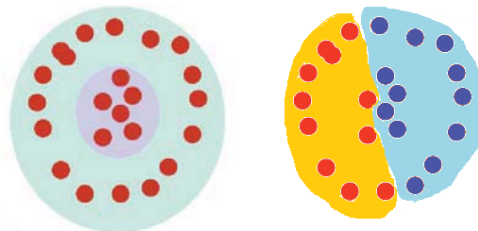
34

5. Single cell transcriptomics data analysis pipeline: clustering

- Two approaches
 1. Conventional clustering algorithms: Cells are assigned to clusters by minimizing intracluster distances or finding dense regions in the reduced expression space.
 2. Community detection methods: graph-partitioning algorithms and thus rely on a graph representation of single-cell data. (KNN graph)

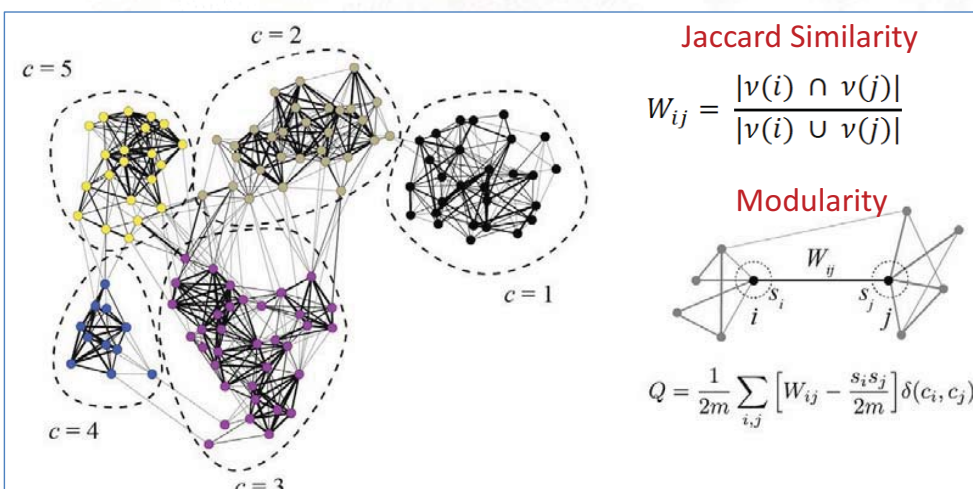


- How would you design an algorithm for finding the three clusters in this case?
- A cluster is a collection of data items which are “similar” between them, and “dissimilar” to data items in other clusters.



35

5. Single cell transcriptomics data analysis pipeline: clustering



Dana Pe'er

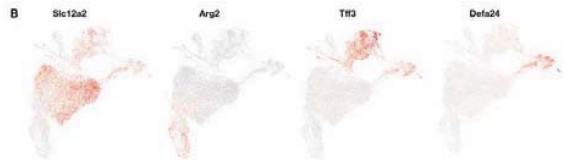
Algorithm

1. Construct a K-nearest neighbor graph using Euclidean distance
2. Construct a weighted shared nearest neighbor graph using Jaccard similarity
3. Maximize the modularity using the Louvain method

36

5. Single cell transcriptomics data analysis pipeline: cell type identification

- The gene signature a.k.a. **marker genes** characterize the cluster and are used to annotate it with a meaningful biological label.
- There are several axes of variation that determine cellular identity
 1. It is **not always clear** what constitutes a cell type. Ex) “T cells” may be a satisfactory label of a cell type to some, others may look for CD4+ and CD8+ T cells.
 2. Cells of the same cell type in **different states** may be detected in separate clusters.
- Identifying and annotating clusters relies on using external sources of information describing the expected expression profiles of individual cell identities.



37

5. Single cell transcriptomics data analysis pipeline: cell type identification

- Two ways to use reference database information to annotate clusters
 1. **Data-derived marker genes**
 - Differential expression (DE) testing between **the cells in one cluster** and **all other cells** in the dataset (typically up-regulated in the cluster of interest) with simple statistical tests (Wilcoxon rank-sum test or the t-test)
 - Null hypothesis: genes have the same distribution of expression values between the two groups.
 - The p-values are often inflated, which can lead overestimation of the number of marker genes. However, the ranking is unaffected. → We can focus on the **top-ranked marker** genes.
 - Differential gene expression not only depends on the cell cluster but also on the **dataset composition**.
 2. **Automated cluster annotation** by directly comparing the gene expression profiles of annotated reference clusters to individual cells.

❖ **The current best practice is a combination of both approaches.**

38

