KSBi-BIML 2023

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

생물정보학&머신러닝 워크샵(온라인)

Single-cell RNA-sequencing analysis of cancer

이세민 _ UNIST

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 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 2023

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

안녕하십니까?

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

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

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

2023년 2월

한국생명정보학회장 이 인 석

강의개요

Single-cell RNA-sequencing analysis of cancer

본 강좌에서는 최근 각광 받고 있는 단세포 전사체 데이터 분석 기술에 대한 소개와 실제 데이터에 대한 분석 실습을 병행하고자 한다. 단세포 전사체 분석 기술은 세포의 분화, 암의 진화, 면역 세포 프로파일링 및 종양 내 이질성 분석 등에 활용되고 있으며, 관련 기술과 응용 사례에 대한 소개 및 현재 가장 널리 사용되고 있는 10x Genomics사의 Chromium Single Cell Gene Expression Solution을 사용하여 생산된 암샘플 단세포 전사체 데이터를 위주로 다양한 분석 방법에 대한 실습을 진행하고자 한다. 강의는 다음의 내용을 포함한다.

- Single-cell RNA-sequencing(scRNA-seq)의 소개 및 개요
- 암 scRNA-seq 연구 동향
- 암 scRNA-seq 데이터 분석 실습

* 교육생준비물:
  노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 중급

* 강의: 이세민 교수 (울산과학기술원 바이오메디컬공학과)

Curriculum Vitae

Speaker Name: Semin Lee, Ph.D.

▶ **Personal Info**

Name	Semin Lee
Title	Associate Professor
Affiliation	Ulsan National Institute of Science and Technology

▶ **Contact Information**

Address	UNIST-gil 50, Bldg #110, Room #301-7, Ulsan, 44919
Email	seminlee@unist.ac.kr

Research Interest

Cancer genomics & single-cell genomics

Educational Experience

2003	B.S. in Biological Sciences, Seoul National University, Korea
2004	M.S. in Bioinformatics, KAIST, Korea
2007	Ph.D. in Bioinformatics, University of Cambridge, UK

Professional Experience

2011-2016	Research Fellow, Department of Biomedical Informatics, Harvard Medical School USA
2016-	Associate Professor, Department of Biomedical Engineering, UNIST, Korea

Selected Publications (5 maximum)

1. Ji Hoon Phi, Ae Kyung Park, Semin Lee, et al. Genomic analysis reveals secondary glioblastoma after radiotherapy in a subset of recurrent medulloblastomas, Acta Neuropathologica, 2018, Jun;135(6):939-953.
2. Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017 Jan 23.
3. Xi R, Lee S, Xia Y, Kim TM, Park PJ. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. 2016 Jul 27;44(13):6274-86.
4. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, Cai X, Luquette LJ, Lee E, Park PJ, Walsh CA. Mosaic Mutations Trace Developmental and Transcriptional Histories of Single Human Neurons. Science. 2015 Oct 2;350(6256):94-8.
5. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330-7.

KSBi-BIML 2021

Single-cell RNA-sequencing analysis of cancer

Semin Lee, PhD

Email : seminlee@unist.ac.kr

Ulsan National Institute of Science and Technology

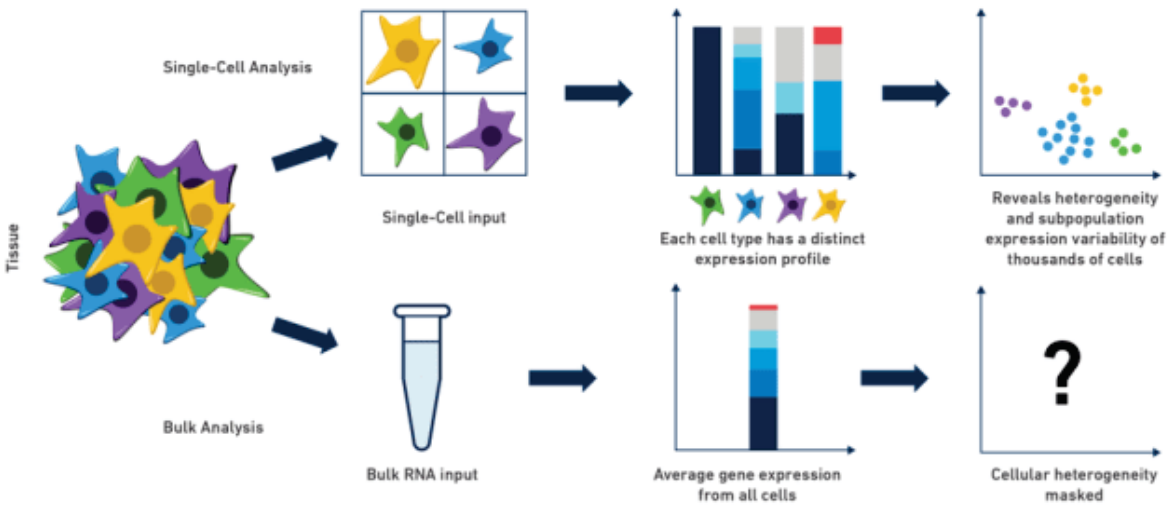
Organized by

한국생명정보학회 Korean Society for Bioinformatics

한국생명정보학회 Korean Society for Bioinformatics

Introduction

Why single-cell RNA sequencing?



https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Limitations of bulk RNA sequencing

- Bulk sequencing methods are limited to reporting an average signal from a complex population of cells.
- So, it is difficult to resolve cell-to-cell variations in tumor and identify the complex nature of tumor microenvironment (TME).
- Computational deconvolution techniques could help infer the cellular composition of tumors, but such analyses are limited to a few known cell types.

Advantages of single-cell RNA sequencing

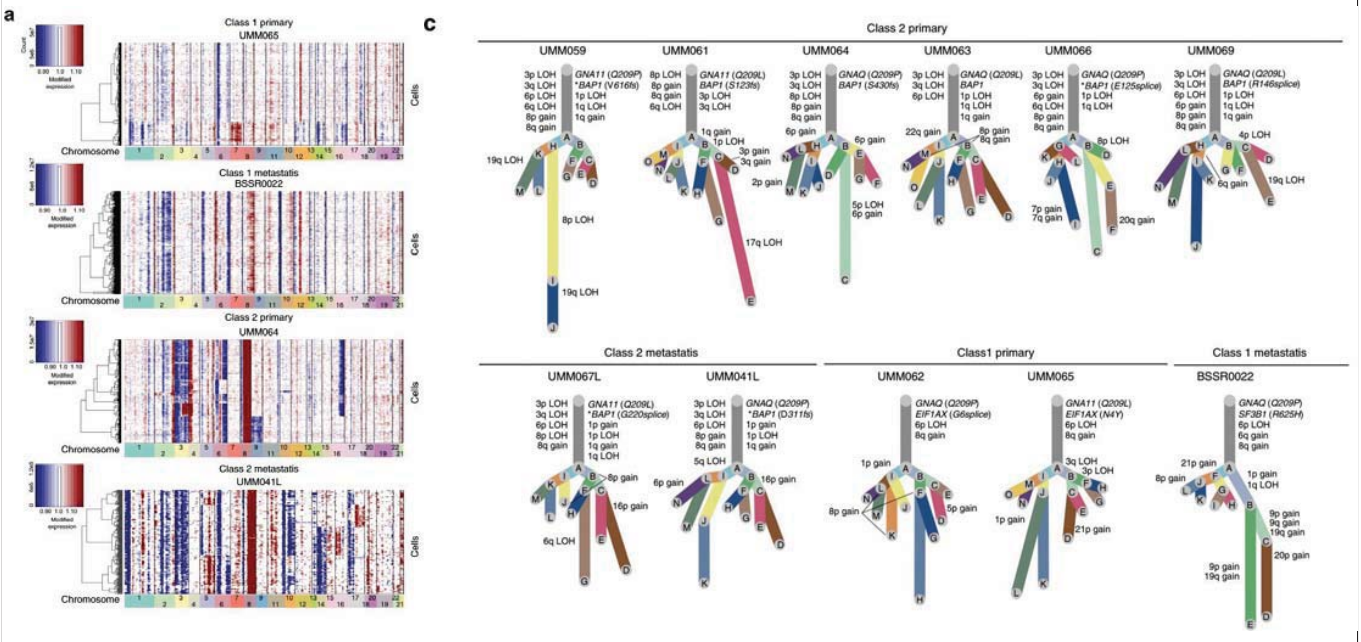
- Single-cell **RNA** sequencing
  - not only identifies intratumoral heterogeneity but also reconstruct a high-resolution map of the TME.
  - identifies cell-specific genetic variants and reconstruct tumor clonality and evolution.
- Single-cell sequencing does not rely on known cell type-specific gene signatures or surface markers.

Single-cell RNA sequencing in cancer

- Tumor heterogeneity
- Clonal evolution of cancer
- Circulating tumor cells
- Tumor microenvironment

SBi 한국생명정보학회 Korean Society for Bioinformatics

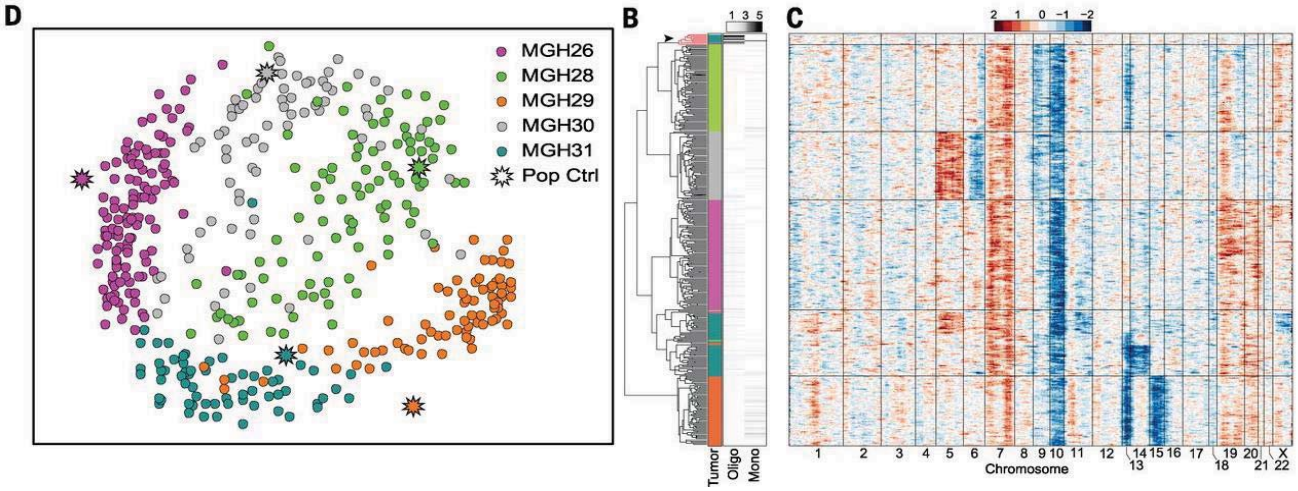
Clonal evolution of cancer



Durante et al. Nat Commun. 2020

SBi 한국생명정보학회 Korean Society for Bioinformatics

Tumor heterogeneity



Patel et al. Science. 2014

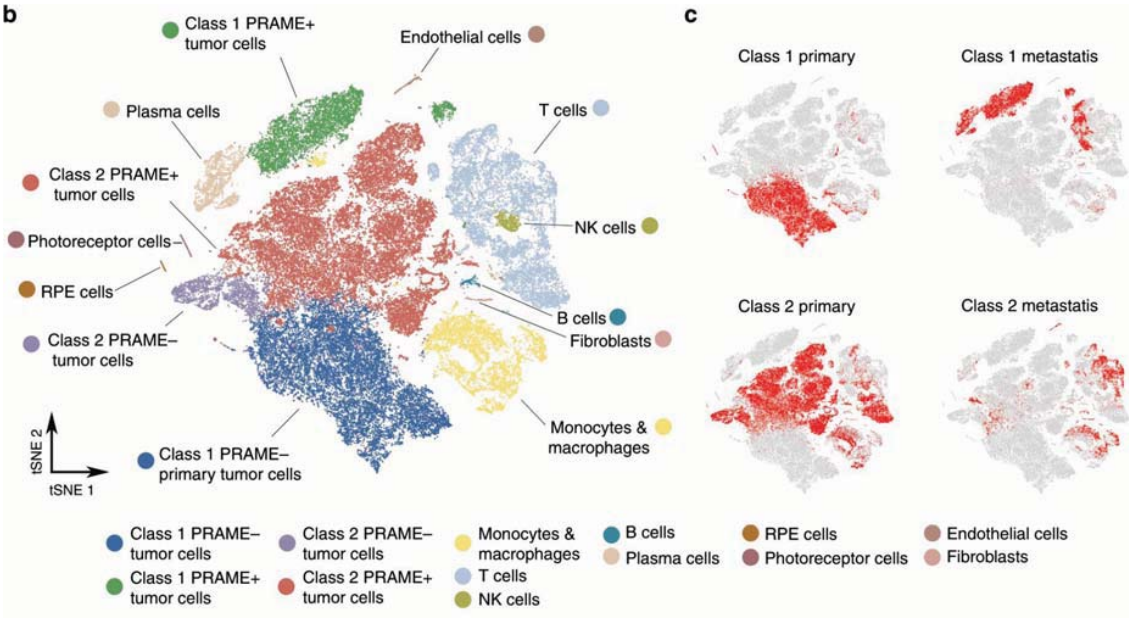
Circulating tumor cells

Cancer type	CTC enrichment	CTC criteria (micromanipulation)	Single-cell profiling	Number of CTCs (number of patients)[a]
Multiple myeloma	FACS with serial dilution	$CD45^-$, $CD138^+$	SMART-seq2	21 (2)
Colon	CellSearch®	$CD45^-$, $EpCAM^+$	Multiplex PCR	11 (8)
Ovary	Biocoll separation, Dynabeads® CD45 depletion	$DAPI^+$, $CK/EpCAM^+$, $CD45^-$	Multiplex PCR	15 (3)
Breast	MagSweeper®	$EpCAM^+$	Microfluidic RT-PCR[b]	105 (50)
	Microfluidic ^{neg}CTC-iChip	$EpCAM/HER2/CDH11^+$, $CD45/CD16/CD14^-$	Optimized Tang's method	15 (10)
	Microfluidic CTC-iChip	$EpCAM/HER2/EGFR^+$, $CD45^-$	SMART-seq v4[c]	15 (10)
	Microfluidic ClearCell® FX	$CD45/CD31^-$, $Calcein^+$ [d]	Polaris™ IFC	68 (4)
Melanoma	MagSweeper®	$CD45^-$, $Calcein^+$	SMART-seq	6 (1)
Prostate	MagSweeper®	$CD45^-$, $EpCAM^+$, $DAPI^-$	SMART-seq, Advantage 2 PCR (Clontech)	20 (4)
	ScreenCell®	$CD45^-$	Microfluidic RT-PCR[e]	38 (9)
	Microfluidic CTC-iChip	$CD45^-$, $EpCAM/CDH11^+$	Modified Tang's method	77 (13)
Lung	Integrated nanoplatform	$EpCAM^+$	Multiplex PCR	8 (1), 18 (1), 74 (1)
	Microfluidic ClearCell® FX	$CD45^-$ [f]	Multiplex PCR	61 (20)
Prostate, breast	CellSearch®, Parsortix™	$EpCAM/pan\text{-}keratins^+$	Multiplex PCR	13 (1), 8 (1)
Pancreas, breast, prostate	Microfluidic CTC-iChip	$CD45^-$	Modified Tang's method	7 (–), 29 (–), 77 (–)

Lim et al. NPJ Precis Oncol. 2019

SBi 한국생명정보학회 Korean Society for Bioinformatics

Tumor microenvironment

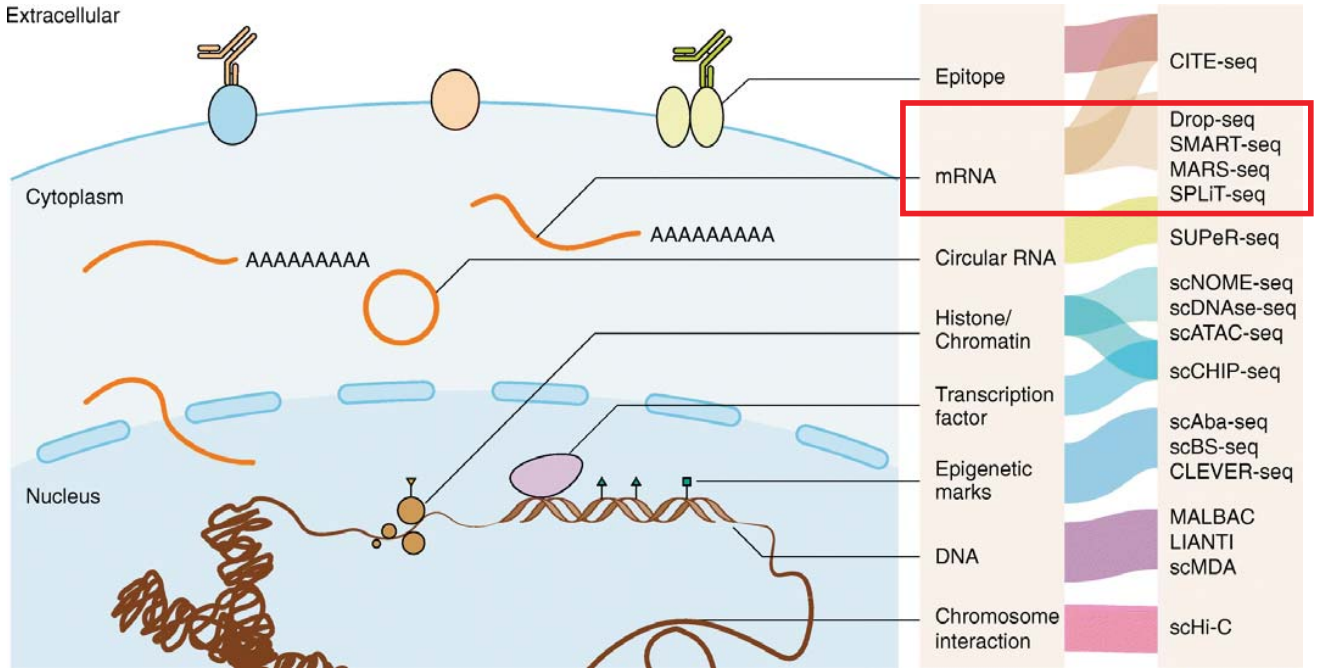


Durante et al. Nat Commun. 2020

Single-cell sequencing

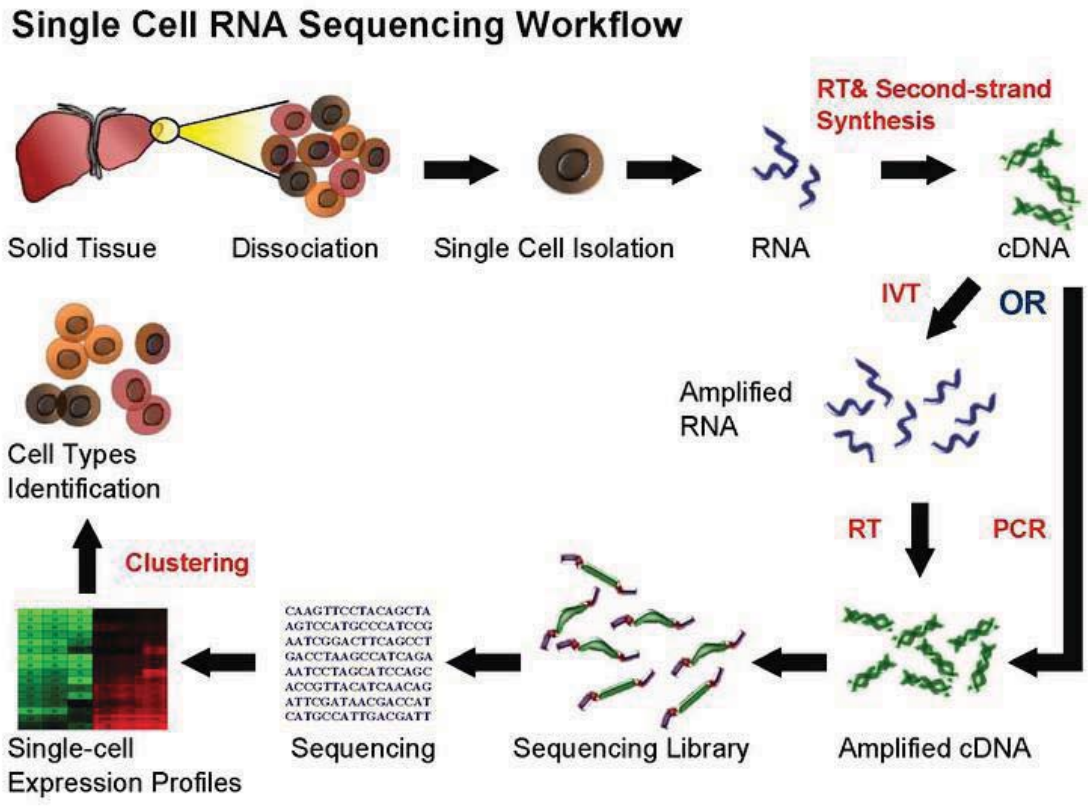
SBi 한국생명정보학회 Korean Society for Bioinformatics

State of the art of single-cell sequencing technologies



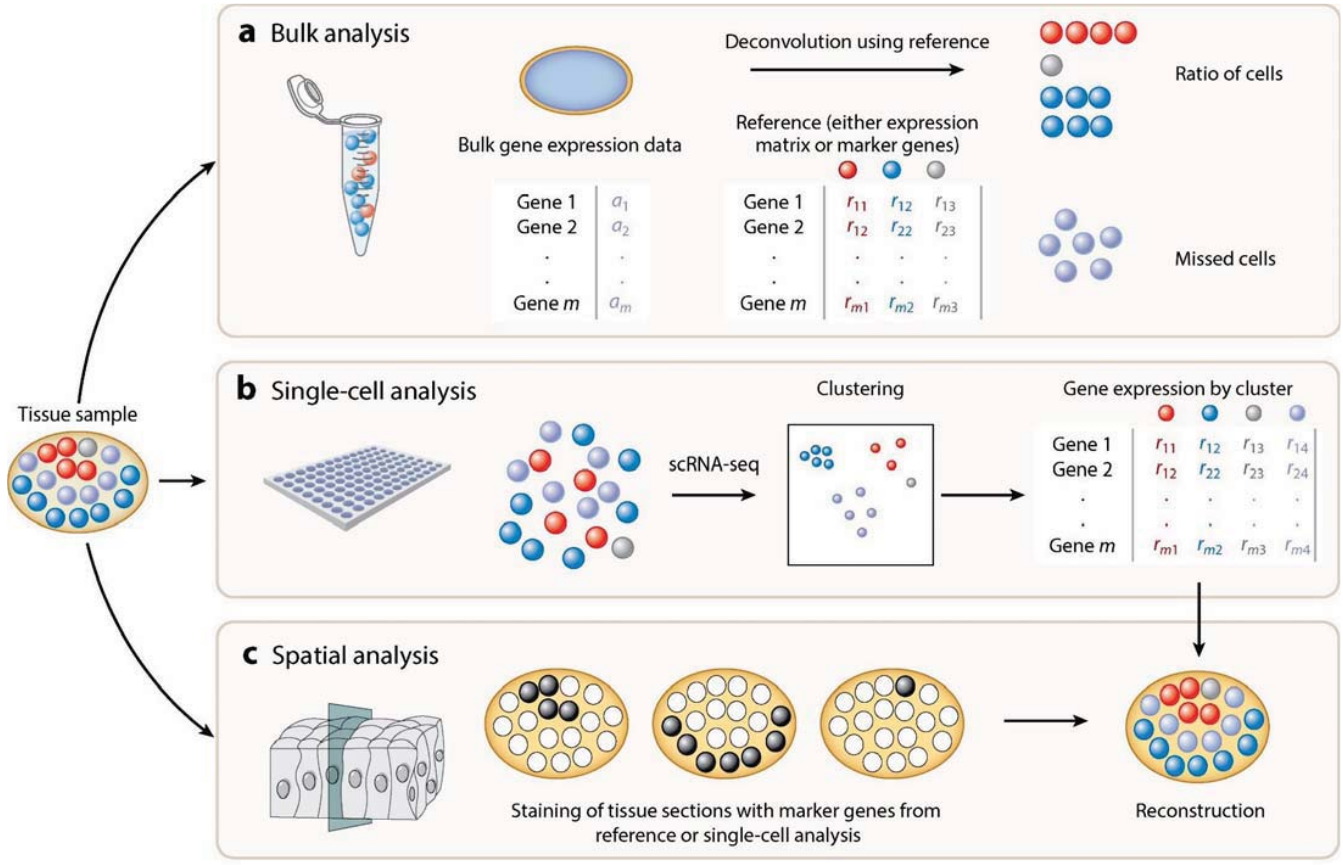
SBi 한국생명정보학회 Korean Society for Bioinformatics

Ren et al. Genome Biol. 2018



SBi 한국생명정보학회 Korean Society for Bioinformatics

https://en.wikipedia.org/wiki/Single_cell_sequencing

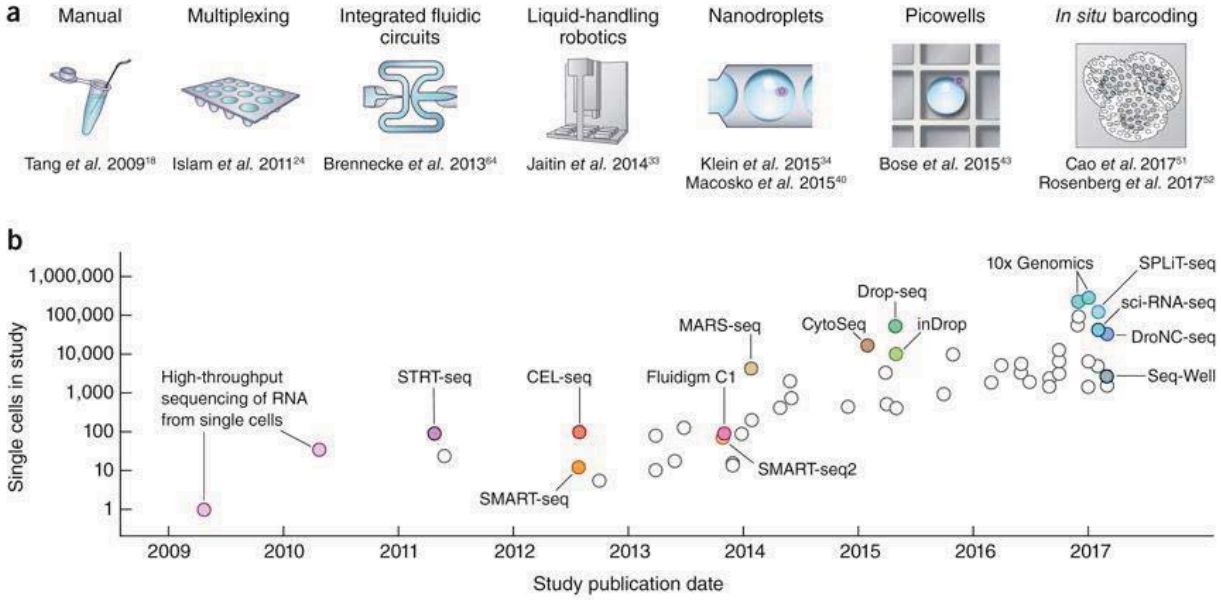


Chen X, et al. 2018. Annu. Rev. Biomed. Data Sci. 1:29–51

SBi 한국생명정보학회 Korean Society for Bioinformatics

Chen et al. Annual review of biomedical data science. 2018

Scaling of scRNA-seq experiments



SBi 한국생명정보학회 Korean Society for Bioinformatics

Svensson et al. Nature Protocols. 2018

REPORT

Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq

Fabienne Lescroart[1,*], Xiaonan Wang[2,3,*], Xionghui Lin[1,*], Benjamin Swedlund[1], Souhir Gargouri[1], Adriana Sànchez-Dànes[1], ...
+ See all authors and affiliations

Science 09 Mar 2018:
Vol. 359, Issue 6380, pp. 1177-1181
DOI: 10.1126/science.aao4174

REPORT

Single-cell multiomics sequencing and analyses of human colorectal cancer

Shuhui Bian[1,2,3,*], Yu Hou[1,2,*], Xin Zhou[4,*], Xianlong Li[1,2,*], Jun Yong[1,5,*], Yicheng Wang[1,2,*], Wendong Wang[4], Jia Yan[1,2], Bo...
+ See all authors and affiliations

Science 30 Nov 2018:
Vol. 362, Issue 6418, pp. 1060-1063
DOI: 10.1126/science.aao3791

RESEARCH ARTICLE

Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq

Andrew S. Venteicher[1,2,3,*], Itay Tirosh[2,*,†], Christine Hebert[1,2], Keren Yizhak[1,2], Cyril Neftel[1,2,4], Mariella G. Filbin[1,2,5], Volk...
+ See all authors and affiliations

Science 31 Mar 2017:
Vol. 355, Issue 6332, eaai8478
DOI: 10.1126/science.aai8478

nature genetics

Article | Published: 11 March 2019

Interrogation of human hematopoiesis at single-cell and single-variant resolution

Jacob C. Ulirsch, Caleb A. Lareau, Erik L. Bao, Leif S. Ludwig, Michael H. Guo, Christian Benner, Ansuman T. Satpathy, Vinay K. Kartha, Rany M. Salem, Joel N. Hirschhorn, Hilary K. Finucane, Martin J. Aryee, Jason D. Buenrostro & Vijay G. Sankaran

Nature Genetics **51**, 683–693 (2019) | Download Citation

nature
International Journal of Science

Letter | Published: 14 March 2018

A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex

Suijuan Zhong, Shu Zhang, Xiaoying Fan, Qian Wu, Liying Yan, Ji Dong, Haofeng Zhang, Long Li, Le Sun, Na Pan, Xiaohui Xu, Fuchou Tang, Jun Zhang, Jie Qiao & Xiaoqun Wang

Nature **555**, 524–528 (22 March 2018) | Download Citation

nature medicine

Letter | Published: 25 June 2018

Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis

Peter Savas, Balaji Virassamy, [...] Sherene Loi

Cell Stem Cell

Volume 17, Issue 3, 3 September 2015, Pages 360-372

CellPress

Resource

Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis

nature immunology

Resource | Published: 15 February 2016

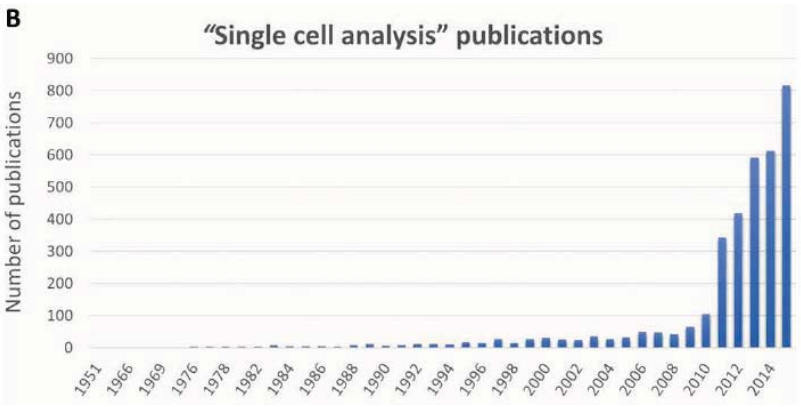
The heterogeneity of human CD127$^+$ innate lymphoid cells revealed by single-cell RNA sequencing

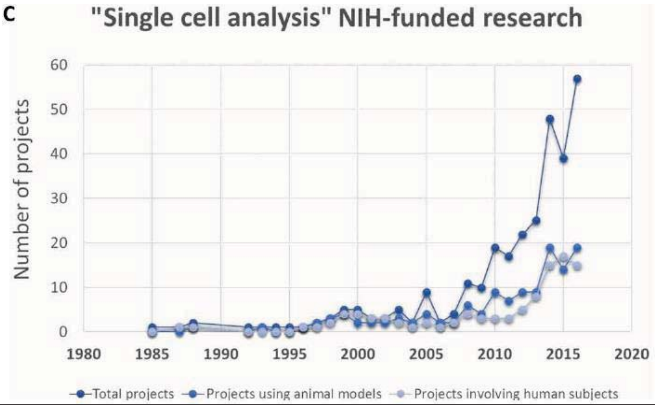
Åsa K Björklund, Marianne Forkel, Simone Picelli, Viktoria Konya, Jakob Theorell, Danielle Friberg, Rickard Sandberg & Jenny Mjösberg

Nature Immunology **17**, 451–460 (2016) | Download Citation

SBi 한국생명정보학회 Korean Society for Bioinformatics

Historical trends of single cell analysis

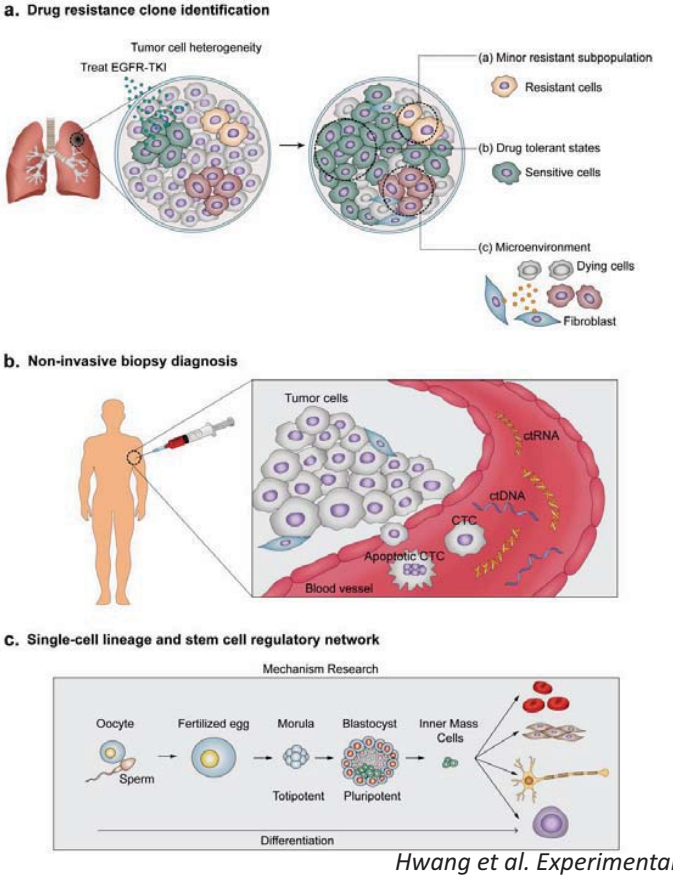




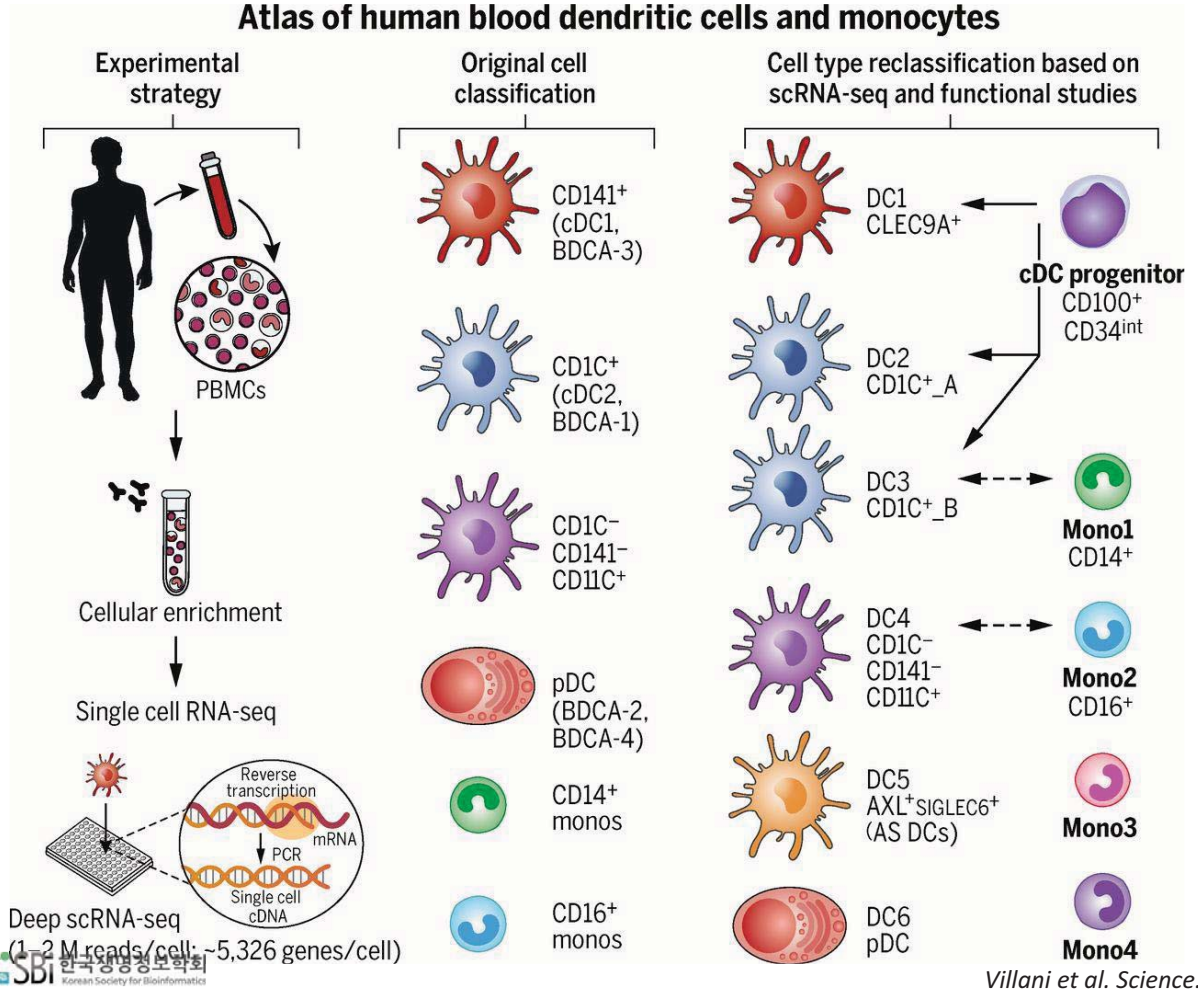
SBi 한국생명정보학회 Korean Society for Bioinformatics

Roy et al., Sci. Adv. 2018

Many facets of scRNA-seq applications

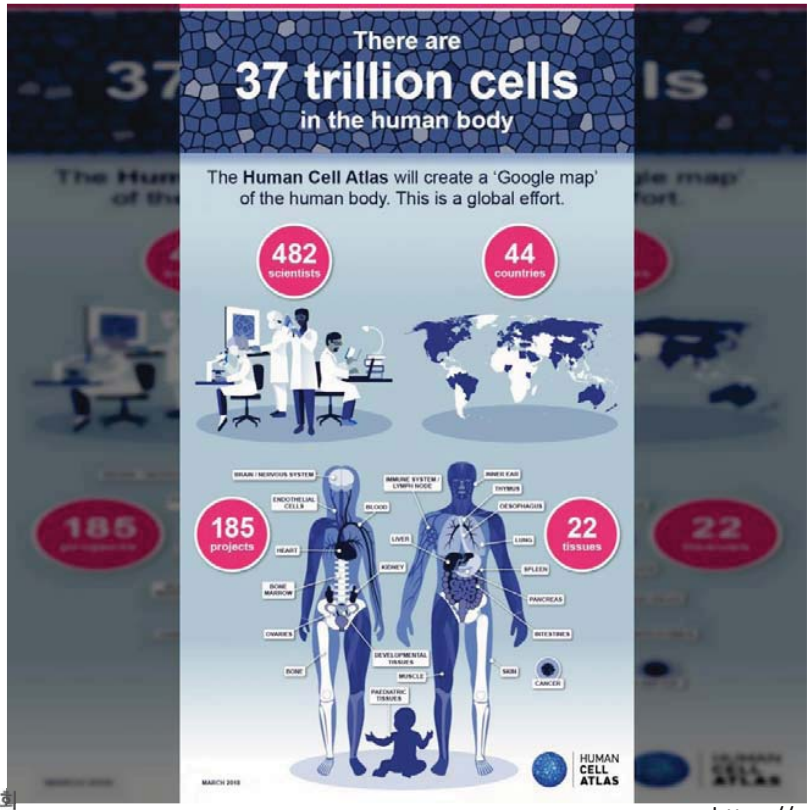


Hwang et al. Experimental & Molecular Medicine. 2018



Villani et al. Science. 2017

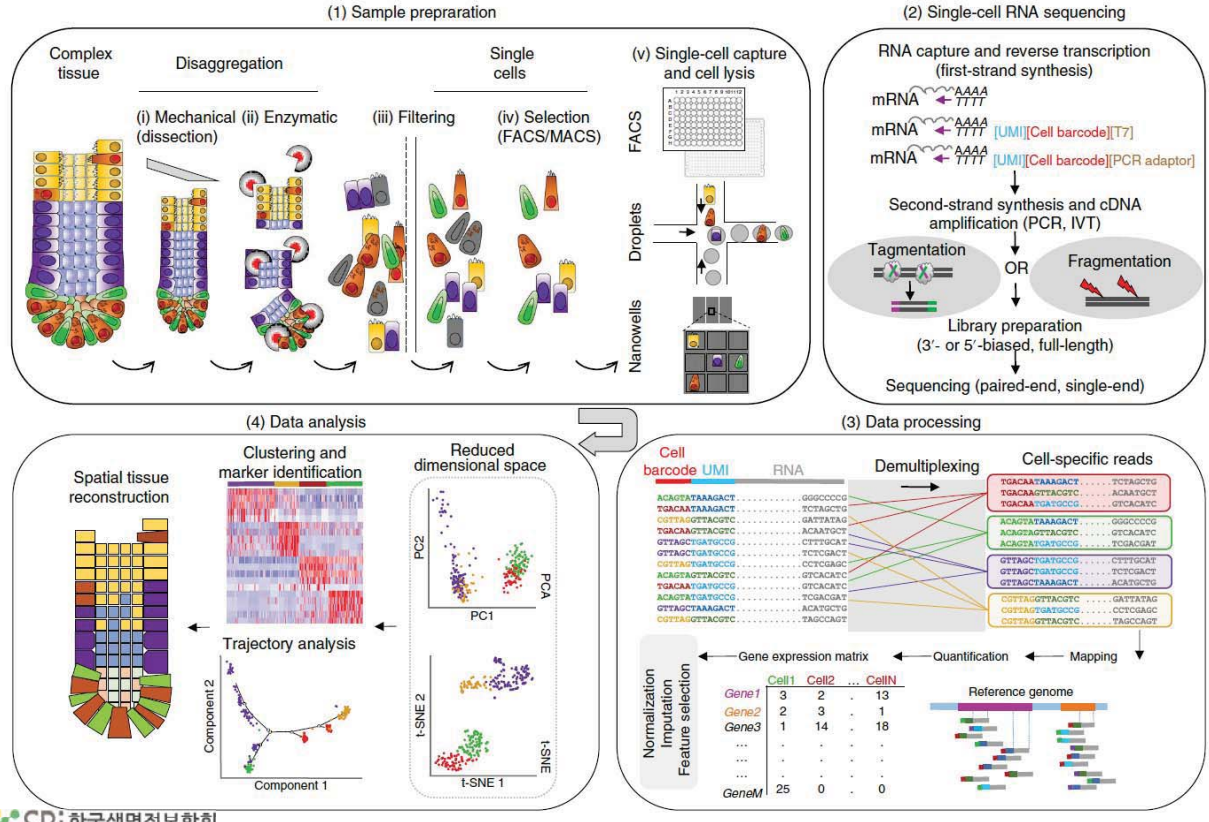
The Human Cell Atlas



SBi 한국생명정보학회 Korean Society for Bioinformatics

https://www.humancellatlas.org/

Single-cell RNA sequencing process

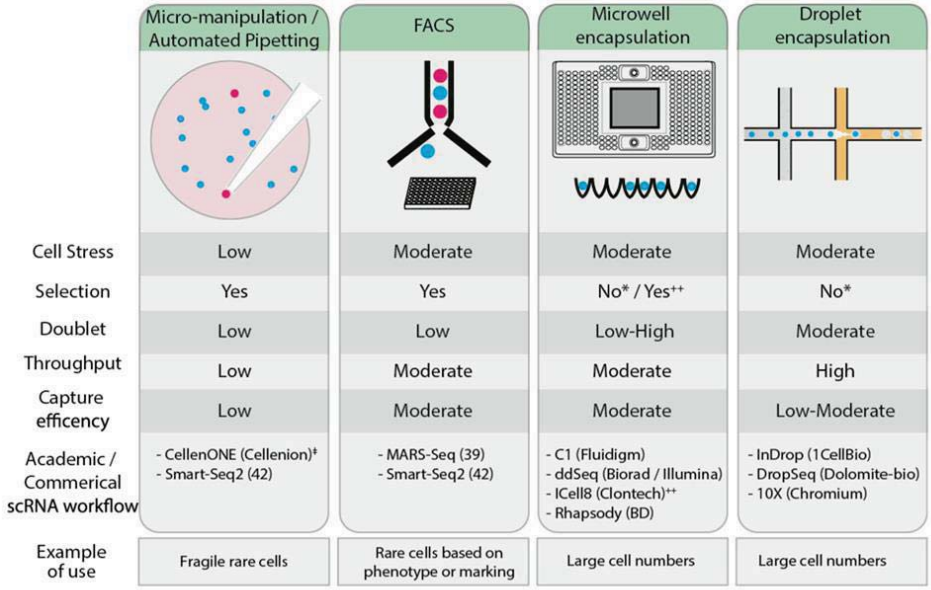


SBi 한국생명정보학회 Korean Society for Bioinformatics

Lafzi et al. Nat Prodoc. 2018

Single-cell RNA sequencing platforms

SBi 한국생명정보학회 Korean Society for Bioinformatics



	Micro-manipulation / Automated Pipetting	FACS	Microwell encapsulation	Droplet encapsulation
Cell Stress	Low	Moderate	Moderate	Moderate
Selection	Yes	Yes	No* / Yes++	No*
Doublet	Low	Low	Low-High	Moderate
Throughput	Low	Moderate	Moderate	High
Capture efficency	Low	Moderate	Moderate	Low-Moderate
Academic / Commerical scRNA workflow	- CellenONE (Cellenion)‡ - Smart-Seq2 (42)	- MARS-Seq (39) - Smart-Seq2 (42)	- C1 (Fluidigm) - ddSeq (Biorad / Illumina) - ICell8 (Clontech)++ - Rhapsody (BD)	- InDrop (1CellBio) - DropSeq (Dolomite-bio) - 10X (Chromium)
Example of use	Fragile rare cells	Rare cells based on phenotype or marking	Large cell numbers	Large cell numbers

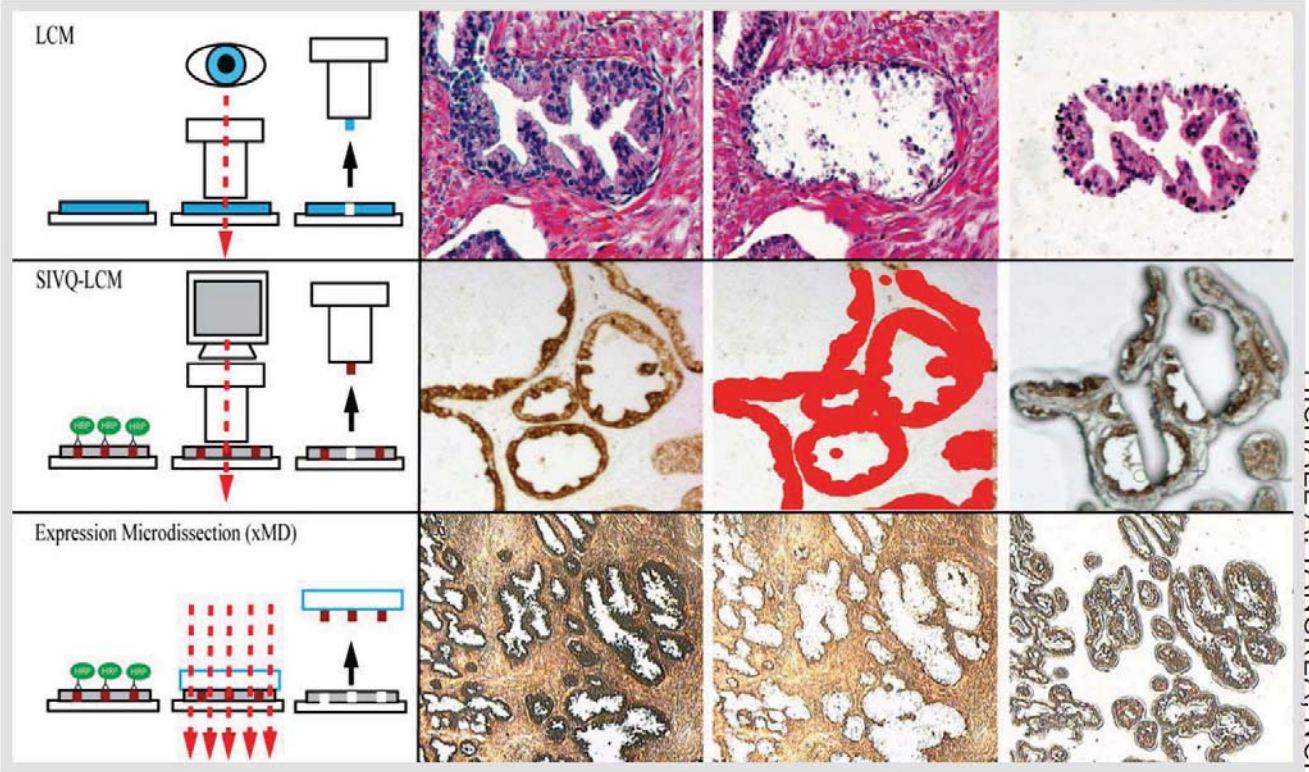
	FACS		Microwell encapsulation				Droplet encapsulation		
	Smart-Seq2	MARS-Seq	C1	ddSeq	ICell8	Rhapsody	InDrop	DropSeq	10X
Singlet Capture efficency	82%	92%	39%	2.6%	37%++	Not reported	7%	Not reported	50%
Doublet rate	Not reported	2.27%	3-30%	5.8%	1.3-4%	0.6%	4%	0.36-11.3	1.6-3%
Reference	42	39	37 FWP	PB	PB	PB	36	37	26

‡Automated pipetting system
*Preselection or enrichment can be performed prior
++Only reagents added to wells containing singlets, determined by system
FWP: Fluidigm white paper
PB: Product brochure / manual

SBi 한국생명정보학회 Korean Society for Bioinformatics

Nguyen et al. Front. Immunol. 2018

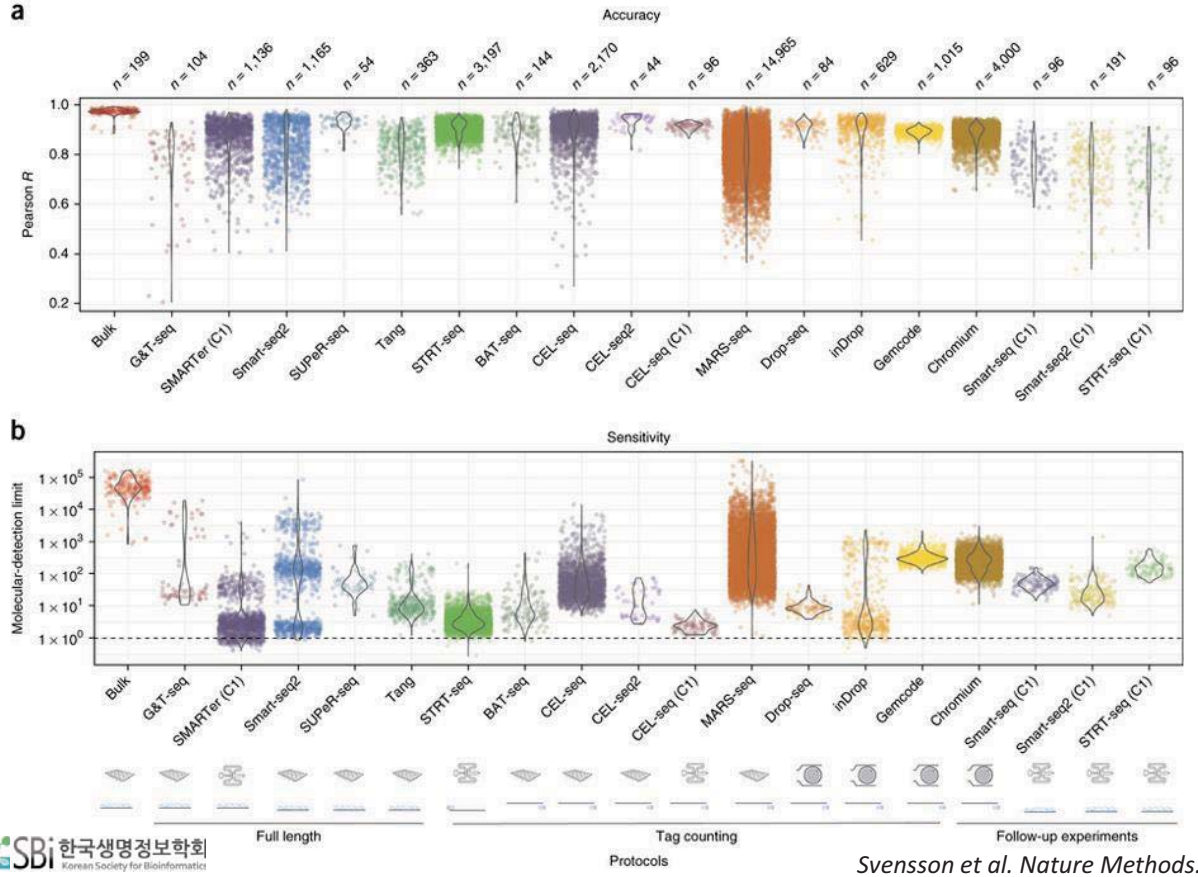
Laser capture microdissection (LCM)



SBi 한국생명정보학회 Korean Society for Bioinformatics

https://irp.nih.gov/catalyst/v19i6/new-methods

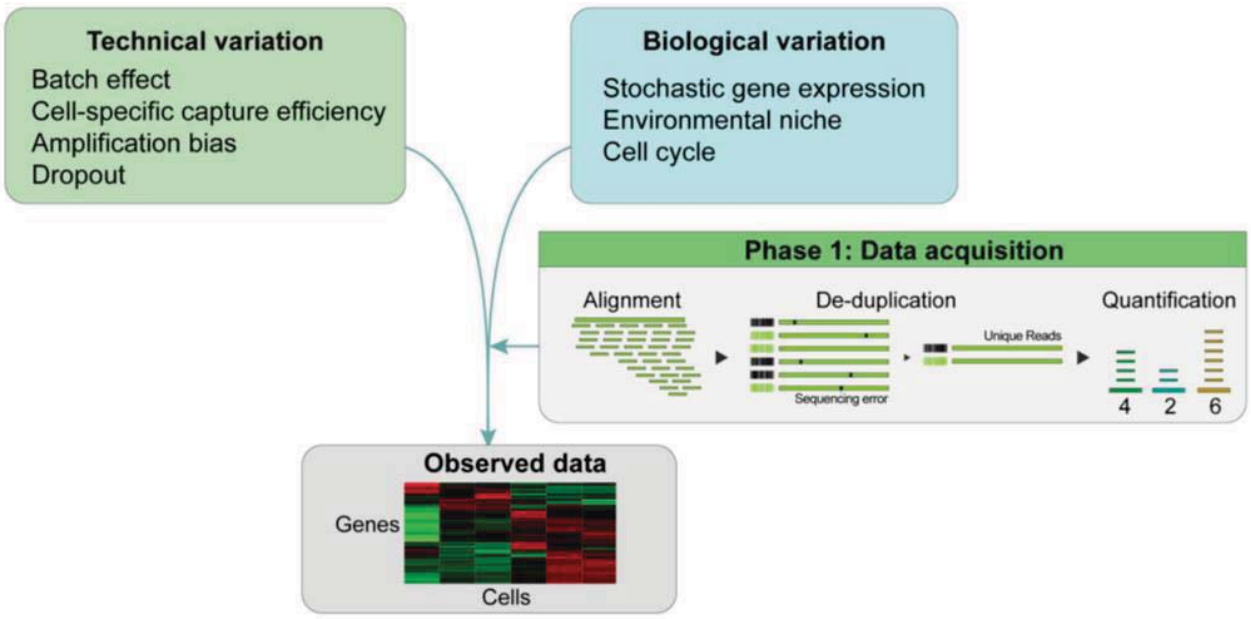
Performance metrics for scRNA-seq protocols



SBi 한국생명정보학회 Korean Society for Bioinformatics

Svensson et al. Nature Methods. 2017

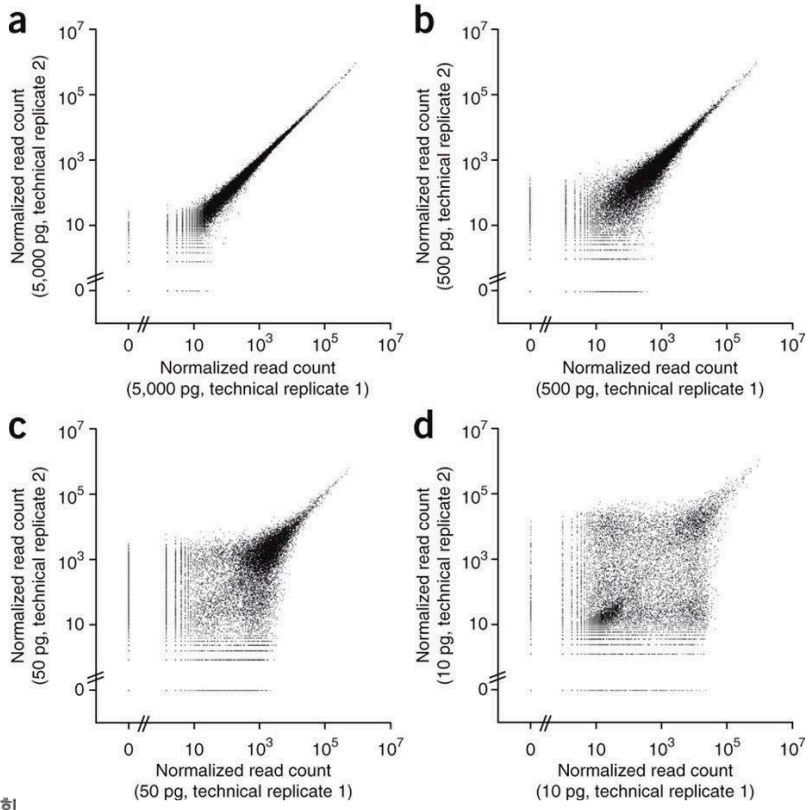
Technical noise in scRNA-seq



SBi 한국생명정보학회 Korean Society for Bioinformatics

Hwang et al. Experimental & Molecular Medicine. 2018

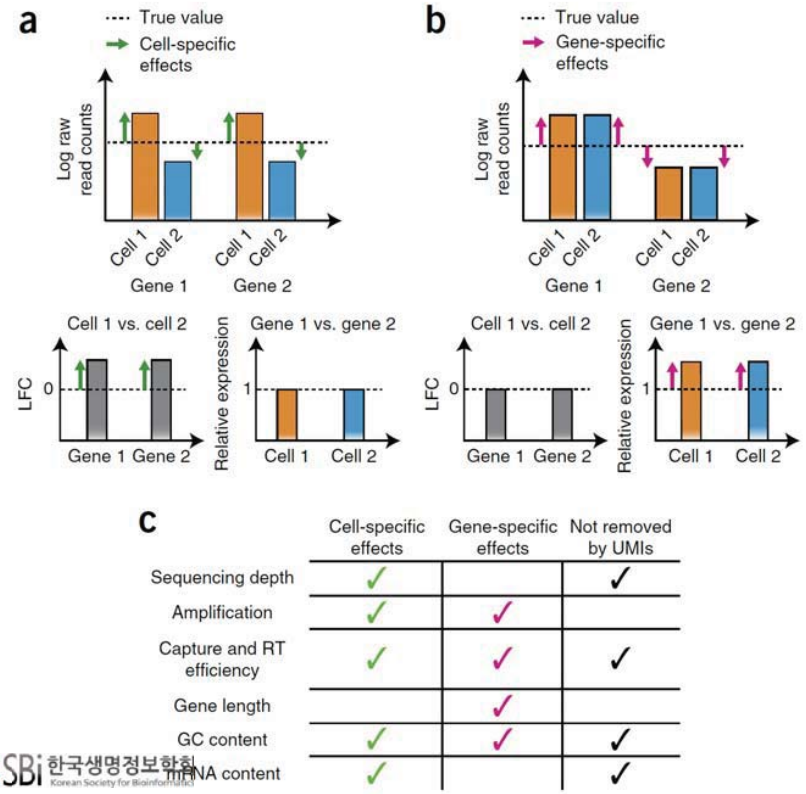
Technical noise in scRNA-seq



SBi 한국생명정보학회 Korean Society for Bioinformatics

Brennecke et al. Nature Methods. 2013

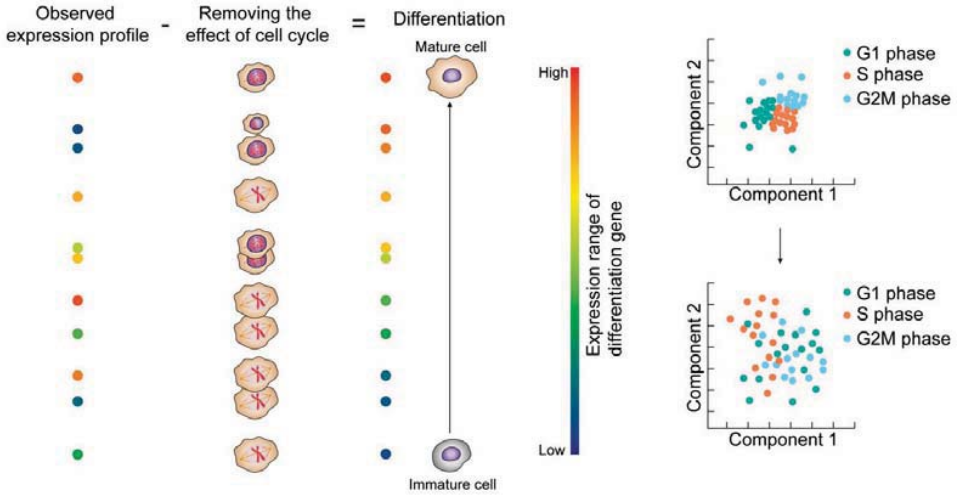
Cell and gene-specific effects in RNA-seq experiments

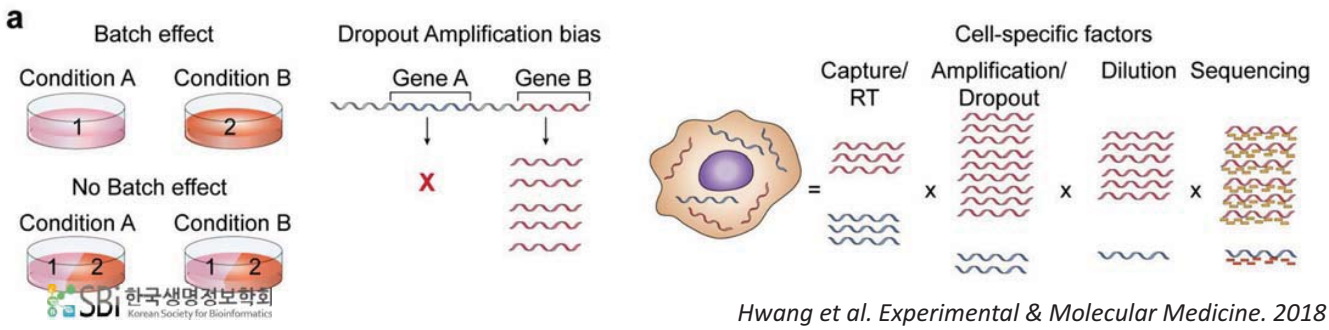


There are several experimental sources of systematic biases that can affect measurements of gene expression, including gene- and cell-specific features

Vallejos et al. Nature Methods. 2017

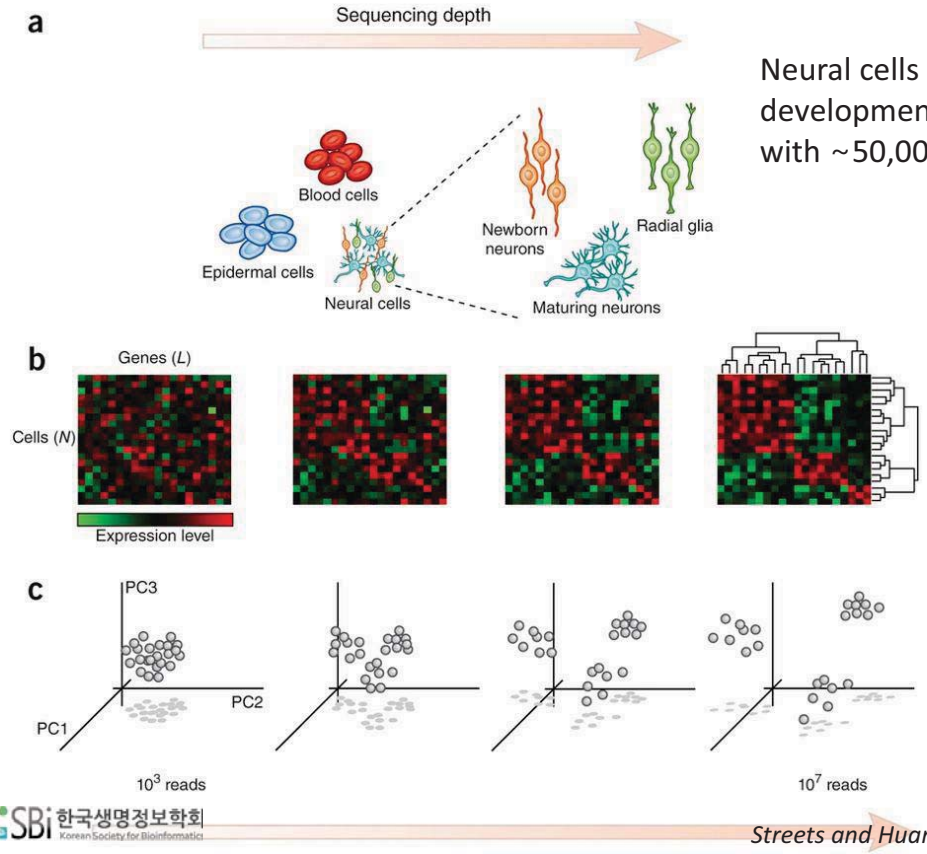
Technical noise in scRNA-seq





Hwang et al. Experimental & Molecular Medicine. 2018

The effect of sequencing depth

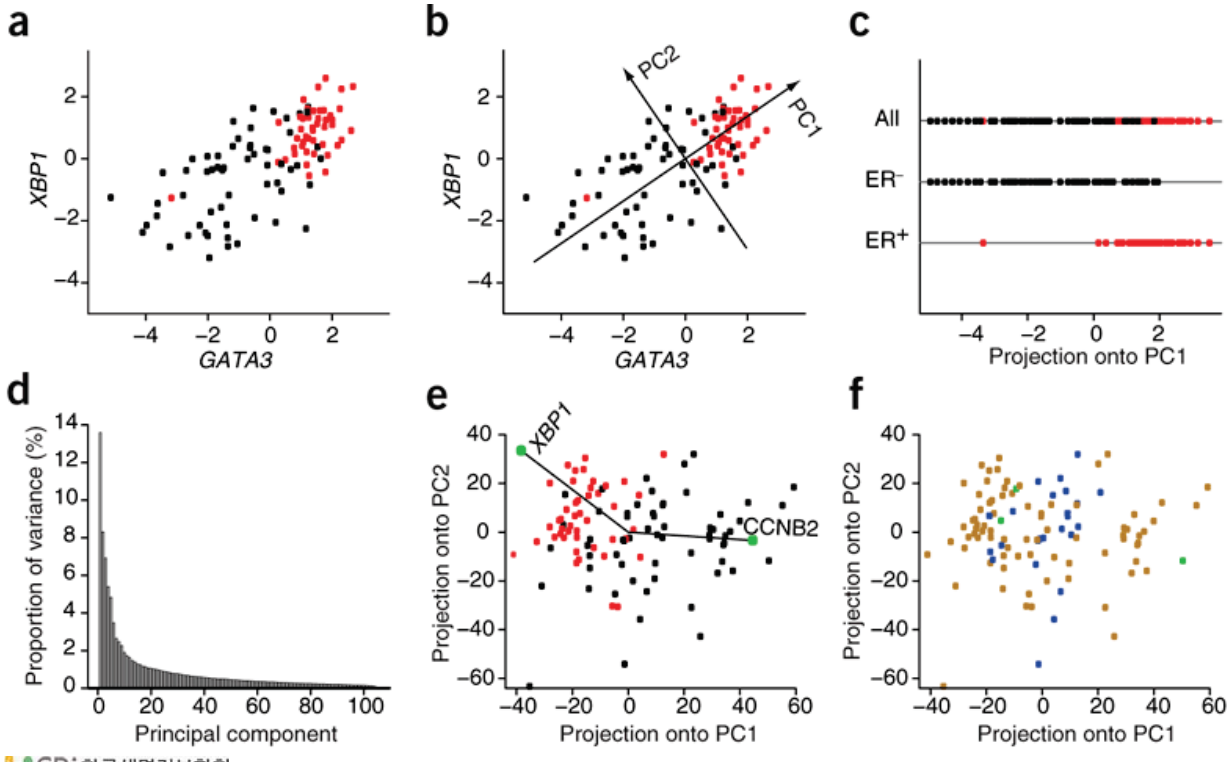


Neural cells at various stages of development can be identified with ~50,000 reads per cell

SBi 한국생명정보학회 Korean Society for Bioinformatics

Streets and Huang. Nature Biotechnology. 2014

Principal Component Analysis (PCA)



SBi 한국생명정보학회 Korean Society for Bioinformatics

Ringnér et al. Nature Biotechnology. 2008

T-distributed stochastic neighborhood embedding (t-SNE)

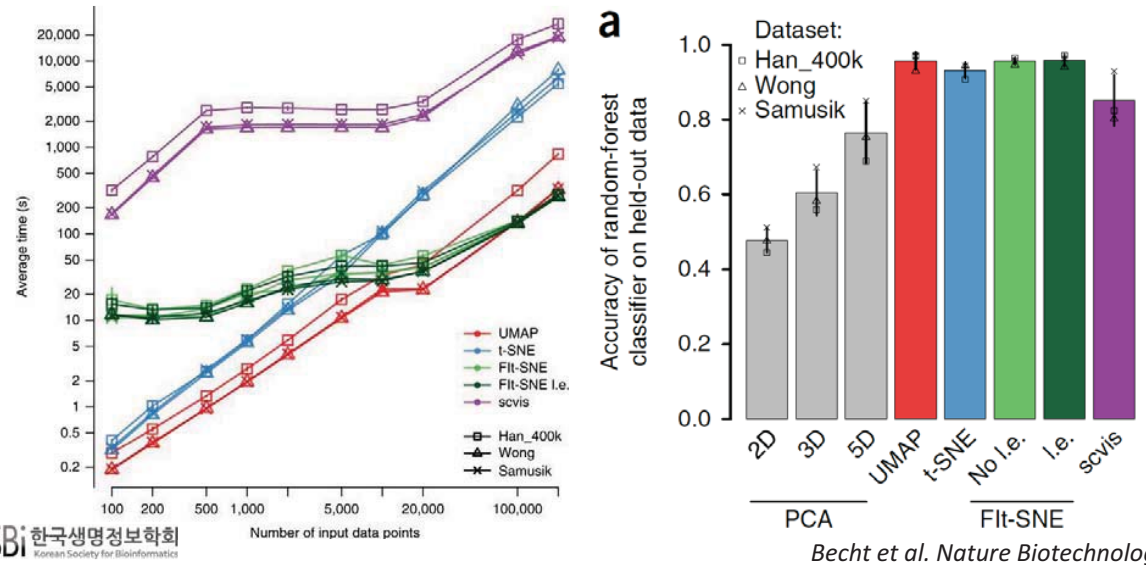
- PCA has historically been the most commonly used method for dimensionality reduction.
- The importance of nonlinear dimensionality reduction techniques has recently been recognized.
  - able to avoid overcrowding of the representation
  - Isomap, Diffusion Map and t-SNE
- t-SNE is currently the most commonly used technique in single-cell analysis
  - t-SNE suffers from limitations such as loss of large-scale information
  - slow computation time
  - inability to meaningfully represent very large datasets

SBi 한국생명정보학회 Korean Society for Bioinformatics

Becht et al. Nature Biotechnology. 2019

Uniform manifold approximation and projection (UMAP)

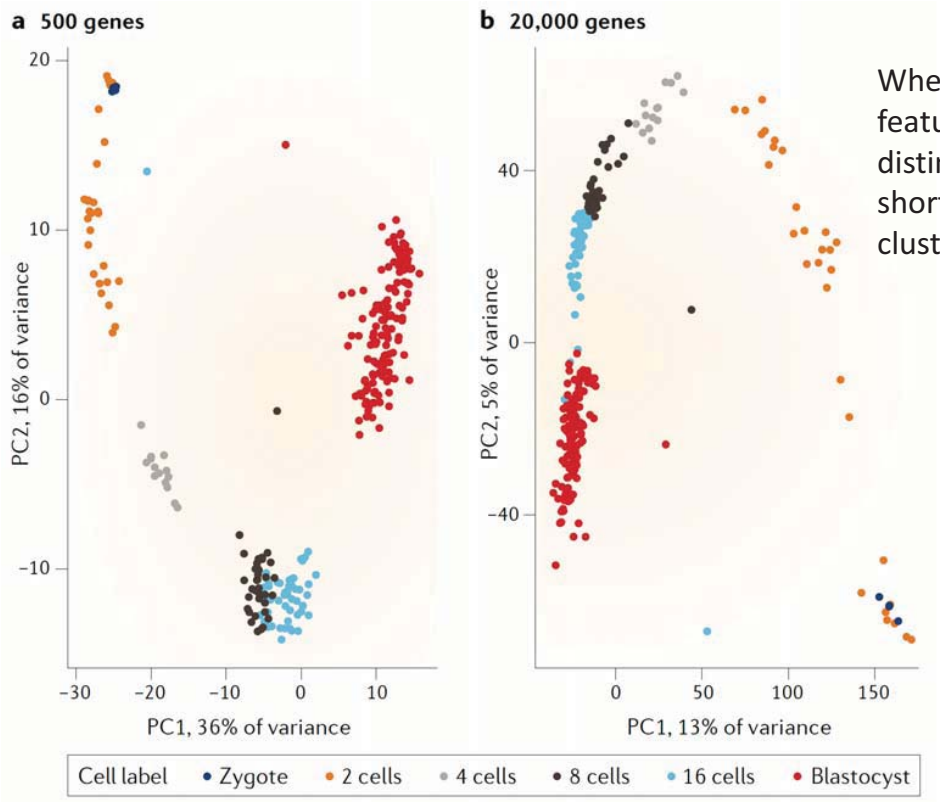
- Preserve as much of the local and more of the global data structure than t-SNE, with a shorter run time



SBi 한국생명정보학회 Korean Society for Bioinformatics

Becht et al. Nature Biotechnology. 2019

The curse of dimensionality



When using a large number of features, clusters are less distinct, as indicated by the shorter distances between clusters

SBi 한국생명정보학회 Korean Society for Bioinformatics

Kiselev et al. Nature Reviews. 2019

Clustering analysis

Table 1 | **Clustering methods for scRNA-seq**

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2 (REF.[115]), RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

DBSCAN, density-based spatial clustering of applications with noise; PCA, principal component analysis; scRNA-seq, single-cell RNA sequencing.

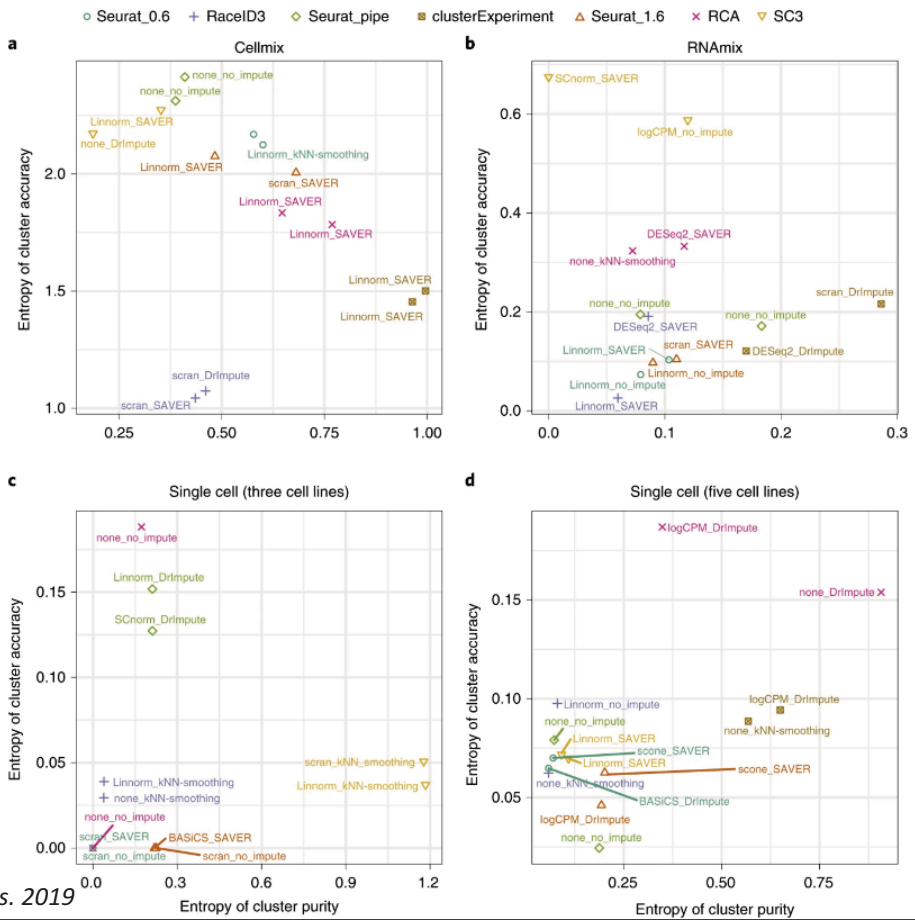
SBi 한국생명정보학회 Korean Society for Bioinformatics

Kiselev et al. Nature Reviews. 2019

Computational challenges

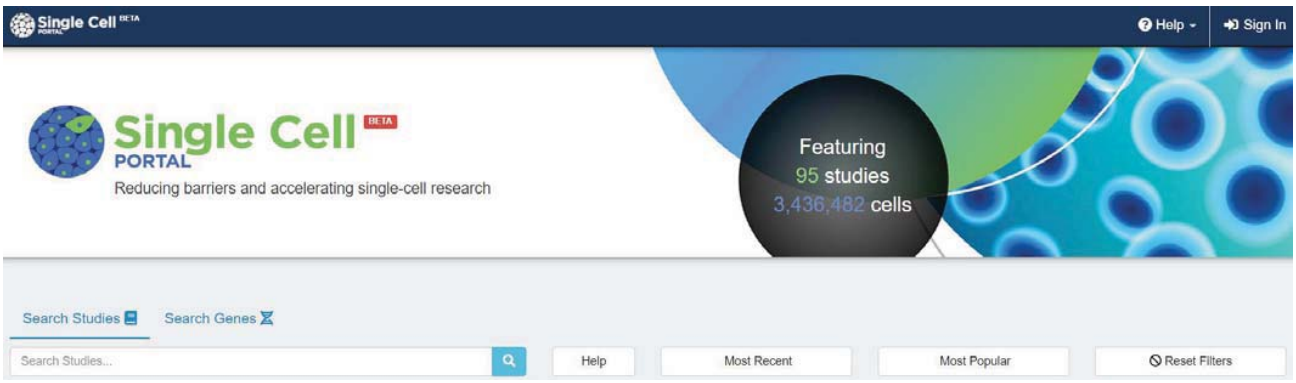
- A large data set ensures that analyses will have high power and improves the ability to detect rare cell types.
- Although it is possible to cluster such large data sets in a time span of hours, visualizing and interpreting the clustering results is difficult.
- Linear transformations, such as PCA, are unable to accurately capture relationships between cells because of the high levels of dropout and noise.
- Nonlinear techniques, such as tSNE and UMAP, can provide outcomes that are often more aesthetically pleasing and easier to interpret by visual inspection.
- They contain parameters that are required to be manually defined by the user and can strongly affect the visualization.

Comparisons of scRNA-seq clustering methods



Tian et al. Nature Methods. 2019

Databases

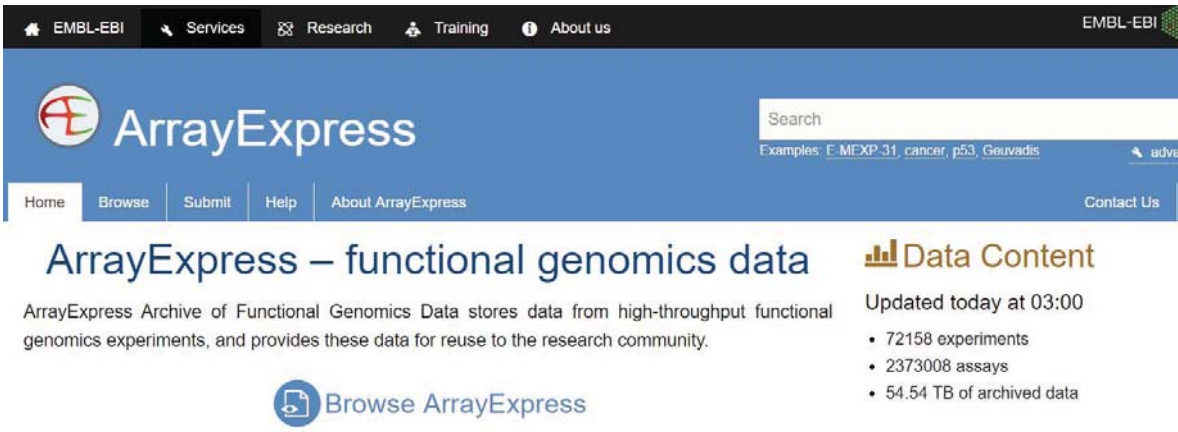


https://portals.broadinstitute.org/single_cell

SBi 한국생명정보학회 Korean Society for Bioinformatics

Data availability. All raw sequencing data are available in ArrayExpress under accessions E-MTAB-6149 and E-MTAB-6653. Also, Rds files were uploaded. These can be imported in CellView to visualise clusters, scroll through tSNE projections and explore gene expression. Moreover, scRNA-seq source data were formatted as .loom files, which can be visualized in an interactive manner through SCope (https://gbiomed.kuleuven.be/scRNAseq-NSCLC)[54]. Finally, gene expression data for all 52 clusters are available in Supplementary Table 4, and cluster-specific gene expression data for tumor-derived and non-malignant lung-tissue-derived cells are available in Supplementary Table 5 (only for clusters having >100 cells from both sources).

Lambrechts et al. Nat Med. 2018



https://www.ebi.ac.uk/arrayexpress/

E-MTAB-6149 - Single cell sequencing of lung carcinoma

E-MTAB-6149.processed.1.zip	6.7 MB	14 March 2018, 13:36
E-MTAB-6149.processed.2.zip	13.2 MB	14 March 2018, 13:36
E-MTAB-6149.processed.3.zip	157.6 MB	14 March 2018, 13:36
E-MTAB-6149.processed.4.zip	57.4 MB	14 March 2018, 13:36
E-MTAB-6149.processed.5.zip	40.4 MB	14 March 2018, 13:36
E-MTAB-6149.processed.6.zip	5.8 MB	14 March 2018, 13:36
E-MTAB-6149.processed.7.zip	7.6 MB	14 March 2018, 13:36
E-MTAB-6149.idf.txt	6 KB	9 July 2018, 16:39
E-MTAB-6149.sdrf.txt	66 KB	15 March 2018, 08:18
1247.R1.fastq.gz	7.94 GB	14 March 2018, 13:45
1247.R2.fastq.gz	1.72 GB	14 March 2018, 13:46
1247.R3.fastq.gz	5.35 GB	14 March 2018, 13:46
BT1249.R1.fastq.gz	5.47 GB	14 March 2018, 13:47
BT1249.R2.fastq.gz	1.14 GB	14 March 2018, 13:47
BT1249.R3.fastq.gz	3.69 GB	14 March 2018, 13:47
BT1290_R1.fastq.gz	1.48 GB	14 March 2018, 13:47
BT1290_R2.fastq.gz	2.94 GB	14 March 2018, 13:47
BT1291_R1.fastq.gz	6.92 GB	14 March 2018, 13:48
BT1291_R2.fastq.gz	14.61 GB	14 March 2018, 13:49

Processed data

Investigation description

Sample and data relationship

Raw data

SBi 한국생명정보학회 Korean Society for Bioinformatics

Page bottom

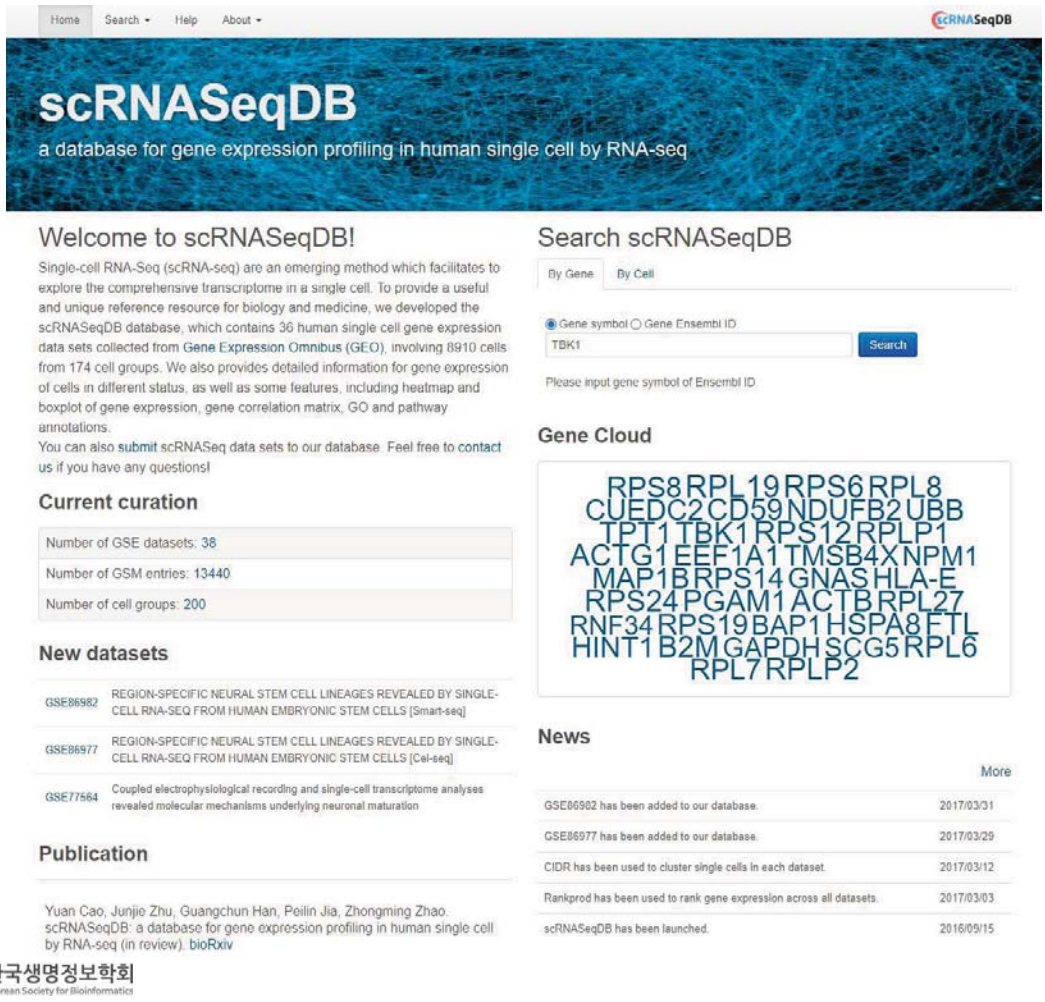
Samples

This page shows the samples included in PanglaoDB. The database currently has 1368 scRNA-seq dataset samples. The controls below can be used to filter by species and sequencing protocol.

Filter by species	Filter by protocol	Sort on
Human	all protocols	Most recent

Refresh

Status	SRA	SRS	Tissue/Site	Protocol	Species	No. Cells	Action
	SRA553822	SRS2119548	Cultured embryonic stem cells	10x chromium	Homo sapiens	6501	view
	SRA598936	SRS2428405	Kidney cortex	10x chromium	Homo sapiens	3759	view
	SRA608611	SRS2517316	Lung progenitors	10x chromium	Homo sapiens	1077	view
	SRA608611	SRS2517317	Lung progenitors	10x chromium	Homo sapiens	906	view
	SRA711898	SRS3349280	Tonsil	10x chromium	Homo sapiens	3503	view
	SRA605557	SRS2493625	Cd14+ monocytes	10x chromium	Homo sapiens	411	view
	SRA605557	SRS2493626	Cd14+ monocytes	10x chromium	Homo sapiens	409	view
	SRA678017	SRS3112005	Embryonic kidney cortex	10x chromium	Homo sapiens	4231	view
	SRA678017	SRS3112006	Embryonic kidney cortex	10x chromium	Homo sapiens	3800	view
	SRA695134	SRS3218229	Alveolar rhabdomyosarcoma	10x chromium	Homo sapiens	8992	view
	SRA550660	SRS2089635	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	1860	view
	SRA550660	SRS2089636	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	1580	view
	SRA550660	SRS2089637	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	10,292	view
	SRA550660	SRS2089638	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	1818	view
	SRA550660	SRS2089639	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	10,940	view
	SRA646572	SRS2833946	Human embryo forebrain	10x chromium	Homo sapiens	6943	view
	SRA646572	SRS2833947	Human embryo forebrain	10x chromium	Homo sapiens	4127	view
	SRA650215	SRS2853842	Fetal kidney	drop-seq	Homo sapiens	1372	view
	SRA650215	SRS2853843	Fetal kidney	drop-seq	Homo sapiens	789	view
	SRA650215	SRS2857943	Fetal kidney	drop-seq	Homo sapiens	362	view
	SRA628554	SRS2664364	Peripheral blood mononuclear cell	10x chromium	Homo sapiens	3352	view
	SRA628554	SRS2664365	Peripheral blood mononuclear cell	10x chromium	Homo sapiens	3081	view



Journal papers on scRNA-seq analysis of cancer

SBi 한국생명정보학회
Korean Society for Bioinformatics

nature medicine

RESOURCE

https://doi.org/10.1038/s41591-018-0096-5

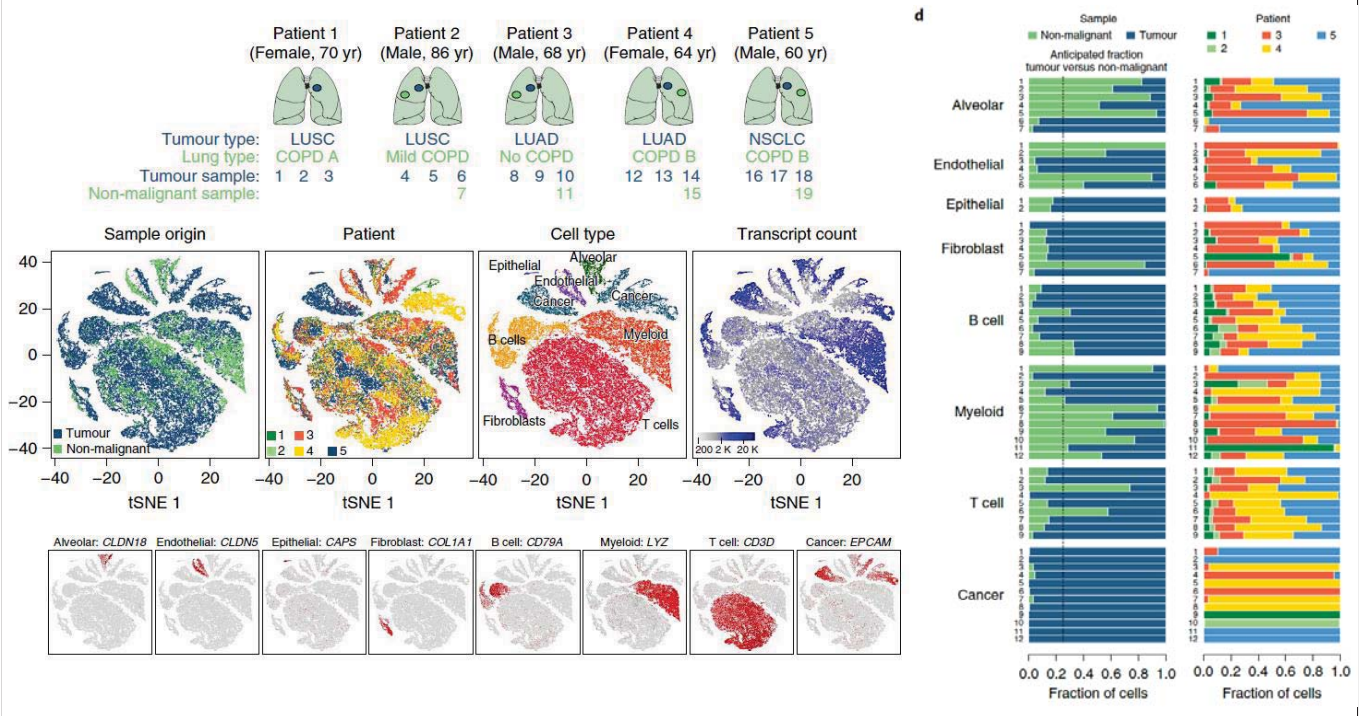
Phenotype molding of stromal cells in the lung tumor microenvironment

Diether Lambrechts[1,2*], Els Wauters[3,4], Bram Boeckx[1,2], Sara Aibar[5,6], David Nittner[7,8], Oliver Burton[6,9], Ayse Bassez[1,2], Herbert Decaluwé[10,11], Andreas Pircher[1,12], Kathleen Van den Eynde[13], Birgit Weynand[13], Erik Verbeken[13], Paul De Leyn[11], Adrian Liston[6,9], Johan Vansteenkiste[3,4], Peter Carmeliet[1,12,14], Stein Aerts[5,6] and Bernard Thienpont[1,15*]

Cancer cells are embedded in the tumor microenvironment (TME), a complex ecosystem of stromal cells. Here, we present a 52,698-cell catalog of the TME transcriptome in human lung tumors at single-cell resolution, validated in independent samples where 40,250 additional cells were sequenced. By comparing with matching non-malignant lung samples, we reveal a highly complex TME that profoundly molds stromal cells. We identify 52 stromal cell subtypes, including novel subpopulations in cell types hitherto considered to be homogeneous, as well as transcription factors underlying their heterogeneity. For instance, we discover fibroblasts expressing different collagen sets, endothelial cells downregulating immune cell homing and genes coregulated with established immune checkpoint transcripts and correlating with T-cell activity. By assessing marker genes for these cell subtypes in bulk RNA-sequencing data from 1,572 patients, we illustrate how these correlate with survival, while immunohistochemistry for selected markers validates them as separate cellular entities in an independent series of lung tumors. Hence, in providing a comprehensive catalog of stromal cells types and by characterizing their phenotype and co-optive behavior, this resource provides deeper insights into lung cancer biology that will be helpful in advancing lung cancer diagnosis and therapy.

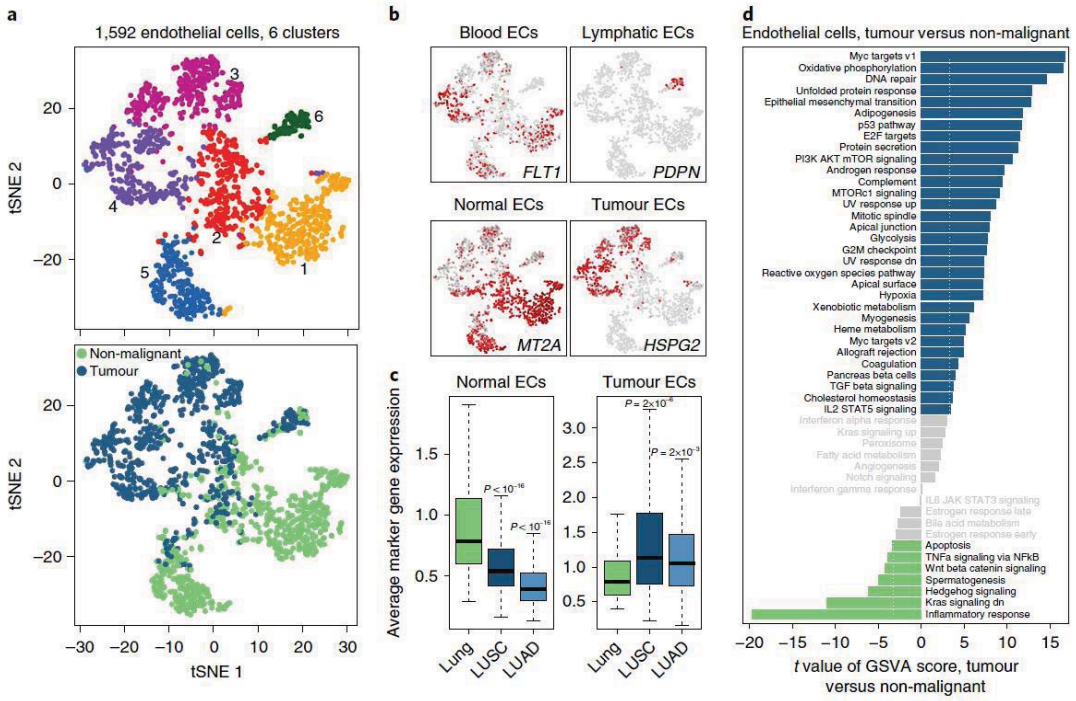
NATURE MEDICINE | VOL 24 | AUGUST 2018 | 1277–1289 | www.nature.com/naturemedicine

Many stromal cell subclusters were enriched for either tumor-derived or lung tissue-derived cells



Lambrechts et al. Nat Med. 2018

Myc targets as the top enriched signature in tumor endothelial cells

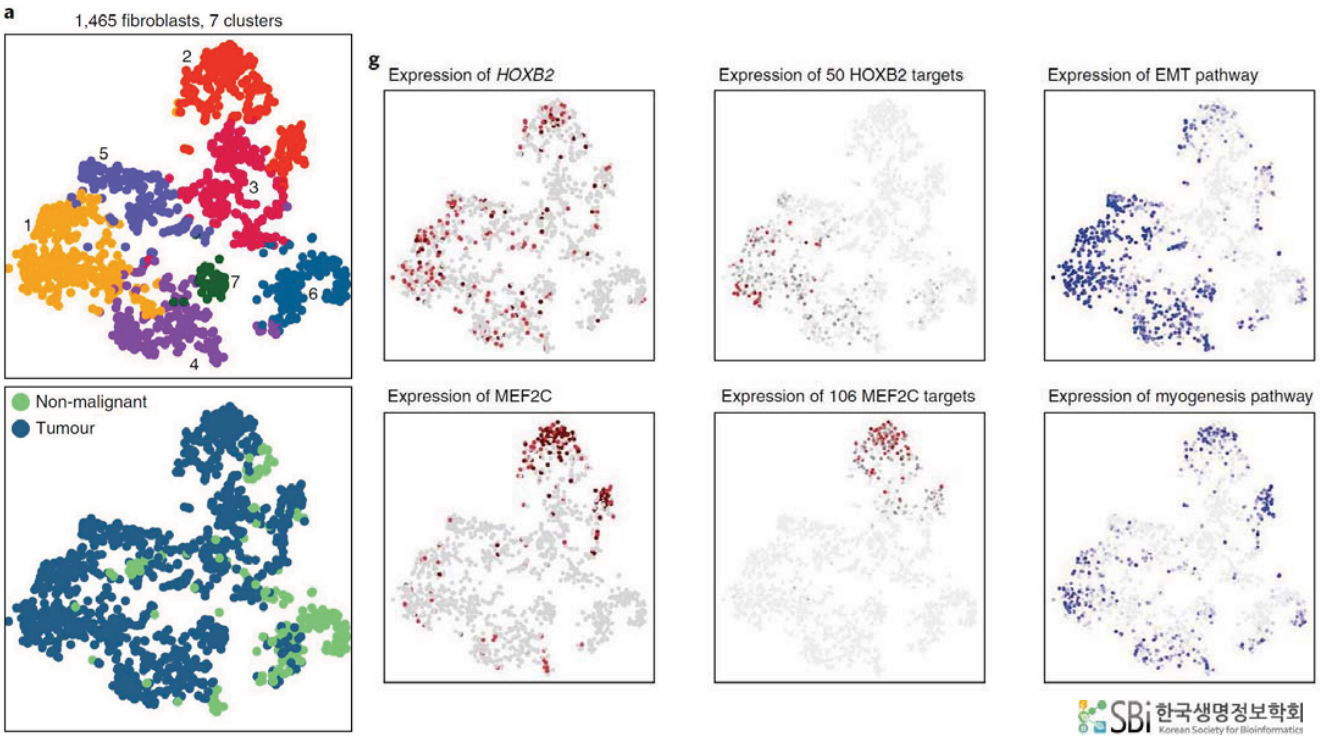


Most significantly downregulated pathway was involved in inflammatory responses.

SBi 한국생명정보학회 Korean Society for Bioinformatics

Lambrechts et al. Nat Med. 2018

Lung tumors are enriched with fibroblasts with expression of EMT pathway



SBi 한국생명정보학회 Korean Society for Bioinformatics

Preparation of single-cell suspensions. Following resection in the operating room, samples from the tumor and adjacent non-malignant lung tissue from the same resection specimen at maximal distance (>5 cm) from the tumor were isolated and transported rapidly to the research facility. On arrival, samples were rinsed with PBS and the tumor sample macroscopically examined for tumor positioning. The tumor sample was subsequently divided into three pieces, with one piece containing mainly tissue derived from the tumor core, one piece containing tissue mainly derived from the tumor edge and a third piece originating from the position intermediate to the other two samples. Each sample was subsequently minced on ice to smaller pieces of less than 1 mm^3 and transferred to 10 ml digestion medium containing 0.2% collagenase I/II (ThermoFisher Scientific), DNAse I (Sigma) and 25 units dispase (Invitrogen) in DMEM (ThermoFisher Scientific). Samples were incubated for 15 min at 37 °C, with manual shaking every 5 min. Samples were then vortexed for 10 s and pipetted up and down for 1 min using pipettes of descending sizes (25 ml, 10 ml and 5 ml). Next, 30 ml ice-cold PBS, pH 7.4, (ThermoFisher Scientific) containing 2% fetal bovine serum (ThermoFisher Scientific) was added and samples were filtered using a 40-µm nylon mesh (ThermoFisher Scientific). Following centrifugation at 120× *g* and 4 °C for 5 min, the supernatant was decanted and discarded, and the cell pellet was resuspended in 2 ml red blood cell lysis buffer and transferred to a 2-ml DNA low bind tube. Following a 5-min incubation at room temperature, samples were centrifuged (120×*g*, 4 °C, 5 min) using a swing-out rotor. Samples were next resuspended in 1 ml PBS containing 8 µl UltraPure BSA (50 mg ml^{-1}; AM2616, ThermoFisher Scientific) and filtered over Scienceware Flowmi 40-µm cell strainers (VWR) using wide-bore 1 ml low-retention filter tips (Mettler-Toledo). Next, 10 µl of this cell suspension was counted using an automated cell counter (Luna) to determine the concentration of live cells. Throughout the dissociation procedure, cells were maintained on ice whenever possible, and the entire procedure was completed in less than 1 h (typically ~45 min) to avoid dissociation-associated artefacts recently described[13]. By using a dissociation signature[13] to detect dissociation-associated changes in gene expression, a positive signal for less than 2% of cells was detected (Supplementary Fig. 2a).

Droplet-based scRNA-seq. Single-cell suspensions were converted to barcoded scRNA-seq libraries by using the Chromium Single Cell 3' Library, Gel Bead & Multiplex Kit and Chip Kit (10x Genomics), aiming for an estimated 4,000 cells per library and following the manufacturer's instructions. Samples were processed using kits pertaining to either the V1 or V2 barcoding chemistry of 10x Genomics (Supplementary Table 2). Single samples are always processed in a single well of a PCR plate, allowing all cells from a sample to be treated with the same master mix and in the same reaction vessel. For each patient, all samples (non-malignant and tumor) were processed in parallel in the same thermal cycler. Libraries were sequenced on an Illumina HiSeq4000, and mapped to the human genome (build hg19) using CellRanger (10x Genomics). Gene positions were annotated as per Ensembl build 85 and filtered for biotype (only protein-coding, long intergenic non-coding RNA, antisense, immunoglobulin or T-cell receptor).

Single-cell gene expression quantification and determination of the major cell types. Raw gene expression matrices generated per sample using CellRanger (version 2.0.0) were combined in R (version 3.3.2—*Sincere Pumpkin Patch*), and converted to a Seurat object using the Seurat R package (version 1.4.0.7)[11]. From this, all cells were removed that had either fewer than 201 UMIs, over 6,000 or below 101 expressed genes, or over 10% UMIs derived from mitochondrial genome. From the remaining 52,698 cells, gene expression matrices were normalized to total cellular read count and to mitochondrial read count using linear regression as implemented in Seurat's *RegressOut* function. As a result, none of the principle components subsequently identified were correlated with transcript count (data not shown). From the remaining 52,698 cells, variably expressed genes were selected as having a normalized expression between 0.125 and 3, and a quantile-normalized variance exceeding 0.5. To reduce dimensionality of this dataset, the resulting 2,192 variably expressed genes were summarized by principle component analysis, and the first 8 principle components further summarized using tSNE dimensionality reduction using the default settings of the *RunTSNE* function. Cell clusters in the resulting two-dimensional representation were annotated to known biological cell types using canonical marker genes (Supplementary Fig. 1). Of note, very few stromal cells (~2%) were positive for cell proliferation markers (Supplementary Fig. 4). We therefore opted not to correct our gene expression matrices for effects of cell cycle.

Subclustering of the major cell types. To identify subclusters within these eight cell types, we reanalyzed cells belonging to each of these eight cell types separately. Specifically, we applied dimensionality reduction using principle component analysis in each cell type on variably expressed genes as described above. To identify which principle components were informative, we applied Horn's parallel analysis for principle component analysis[44] as implemented in the R *paran* package (version 1.5.1.), selecting those principle components having eigenvalues that exceed the eigenvalues generated using ten random permutations by >50%. Using the graph-based clustering approach implemented in the *FindClusters* function of the Seurat package, with a conservative resolution of 0.5 and otherwise default parameters, each cell type was reclustered by its principle components. Notably, subclustering was robust to alterations in the number of principle components, in the resolution or in the *K* parameter (Supplementary Fig. 3a–c). Moreover, few of the subclusters identified contained many cells wherein less than 300 genes were detected, indicating that increasing the threshold of 100 genes will not affect our results (Supplementary Fig. 20). This yielded 64 subclusters (52 stromal subclusters) in total, as listed in Supplementary Table 3. For visualization purposes, these informative principle components were converted into tSNE plots as above.

Identification of marker genes. To identify marker genes for each of these 64 subclusters within these 8 cell types, we contrasted cells from that subcluster to all other cells of that subcluster using the Seurat *FindMarkers* function. Marker genes were required to have an average expression in that subcluster that was >2.5-fold higher than the average expression in the other subclusters from that cell type, and a detectable expression in >15% of all cells from that subcluster. Additionally, marker genes were required to have the highest mean expression in that subcluster, out of all 64 subclusters. This yielded a list of in total 402 marker genes (Supplementary Table 3) for 51 subclusters (42 stromal cell subclusters), whereas for 13 subclusters we failed to identify marker genes. When analyzing marker genes for several subclusters in aggregate, such as for tumor endothelial cells (endothelial cell clusters 3 and 4) or for macrophages (myeloid clusters 1–4, 6–8, 10 and 11), we simply combined the marker genes for all associated subclusters.

nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0045-3

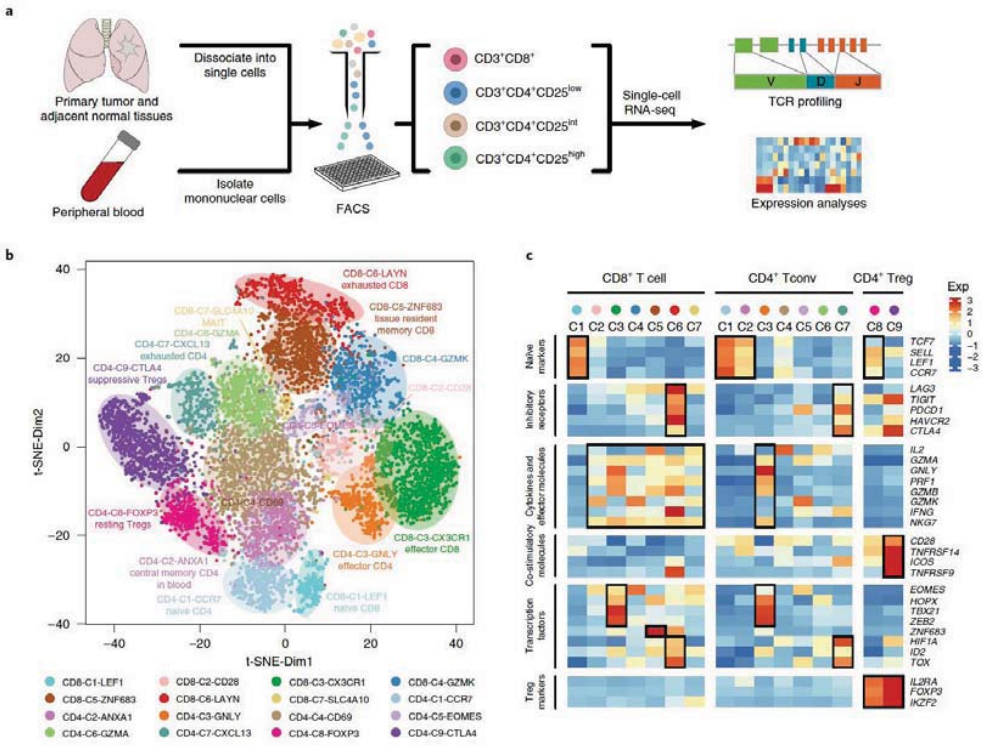
Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing

Xinyi Guo[1,6], Yuanyuan Zhang[1,6], Liangtao Zheng[2,6], Chunhong Zheng[1,6], Jintao Song[3,6], Qiming Zhang[1], Boxi Kang[1], Zhouzerui Liu[1], Liang Jin[3], Rui Xing[4], Ranran Gao[1], Lei Zhang[2], Minghui Dong[1], Xueda Hu[1], Xianwen Ren[1], Dennis Kirchhoff[5], Helge Gottfried Roider[5], Tiansheng Yan[3]* and Zemin Zhang[1,2]*

Cancer immunotherapies have shown sustained clinical responses in treating non-small-cell lung cancer[1–3], but efficacy varies and depends in part on the amount and properties of tumor infiltrating lymphocytes[4–6]. To depict the baseline landscape of the composition, lineage and functional states of tumor infiltrating lymphocytes, here we performed deep single-cell RNA sequencing for 12,346 T cells from 14 treatment-naïve non-small-cell lung cancer patients. Combined expression and T cell antigen receptor based lineage tracking revealed a significant proportion of inter-tissue effector T cells with a highly migratory nature. As well as tumor-infiltrating CD8+ T cells undergoing exhaustion, we observed two clusters of cells exhibiting states preceding exhaustion, and a high ratio of "pre-exhausted" to exhausted T cells was associated with better prognosis of lung adenocarcinoma. Additionally, we observed further heterogeneity within the tumor regulatory T cells (Tregs), characterized by the bimodal distribution of *TNFRSF9*, an activation marker for antigen-specific Tregs. The gene signature of those activated tumor Tregs, which included *IL1R2*, correlated with poor prognosis in lung adenocarcinoma. Our study provides a new approach for patient stratification and will help further understand the functional states and dynamics of T cells in lung cancer.

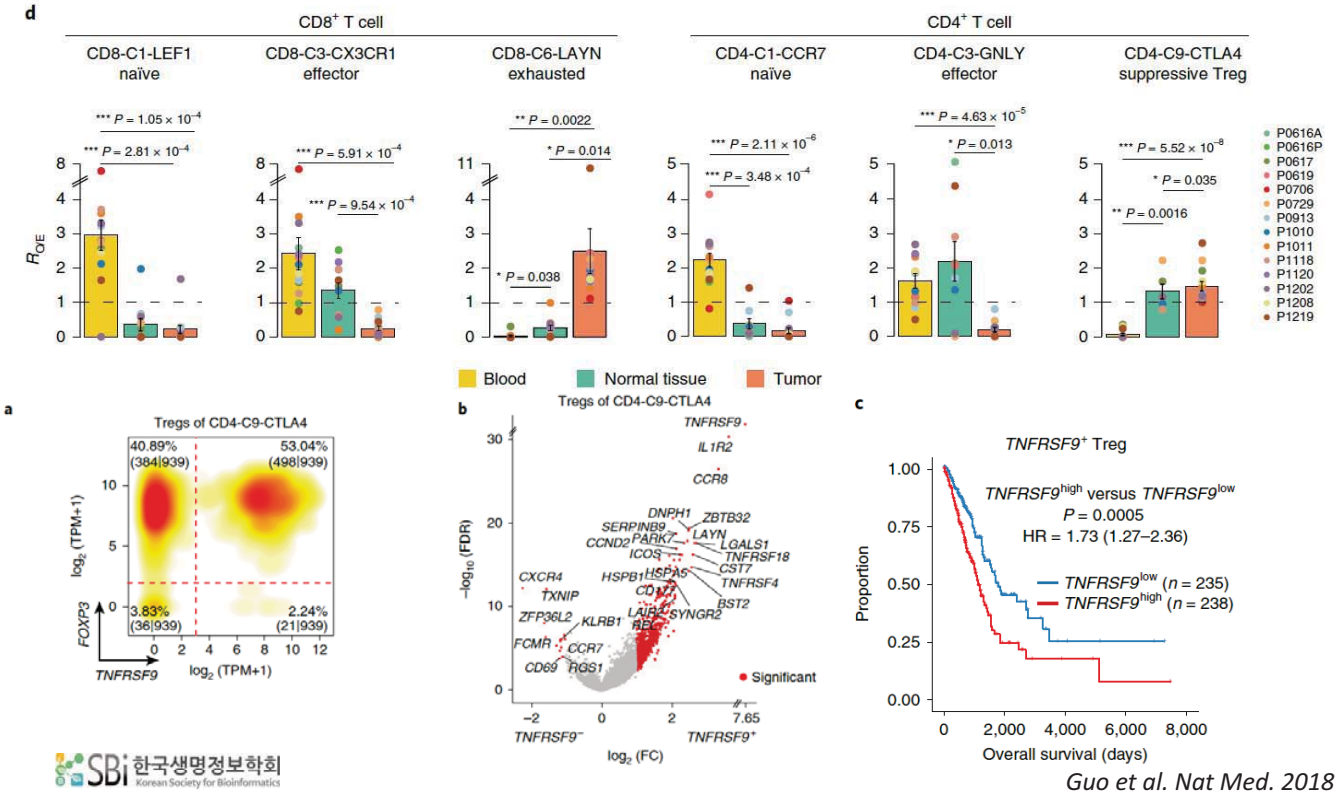
NATURE MEDICINE | VOL 24 | JULY 2018 | 978–985 | www.nature.com/naturemedicine

Seven CD8 and nine CD4 clusters were identified



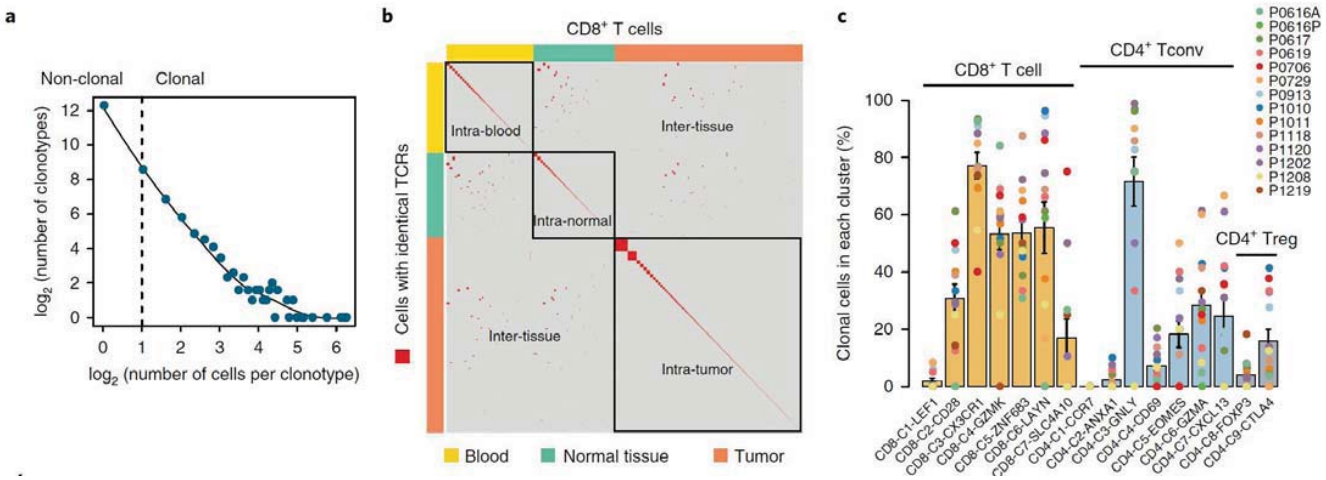
Guo et al. Nat Med. 2018

T cells clustered primarily based on their tissue origins and subtypes



Guo et al. Nat Med. 2018

T cell clusters CD8-C3-CX3CR1 and CD4-C3-GNLY showed the highest proportions of clonal cells



SBi 한국생명정보학회 Korean Society for Bioinformatics

Guo et al. Nat Med. 2018

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-020-1125-8

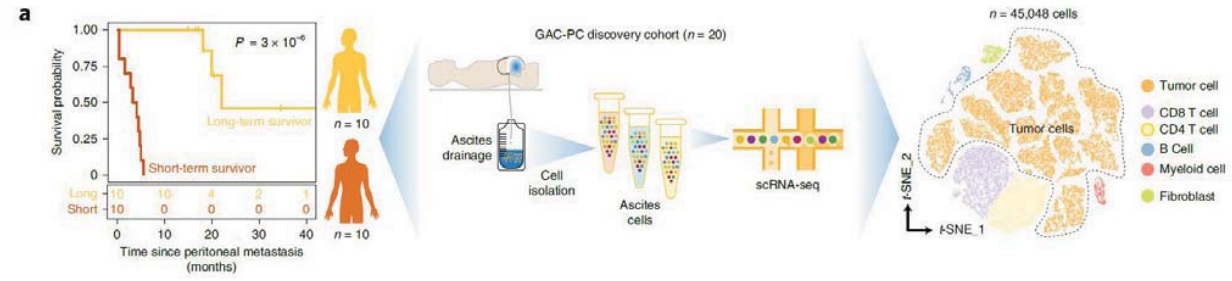
Check for updates

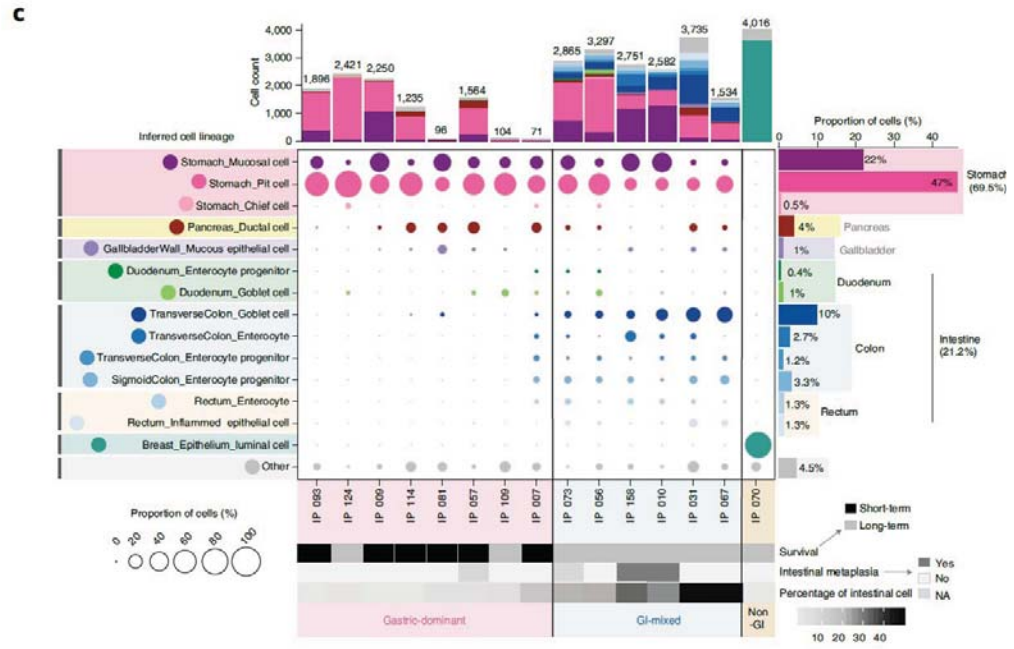
Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma

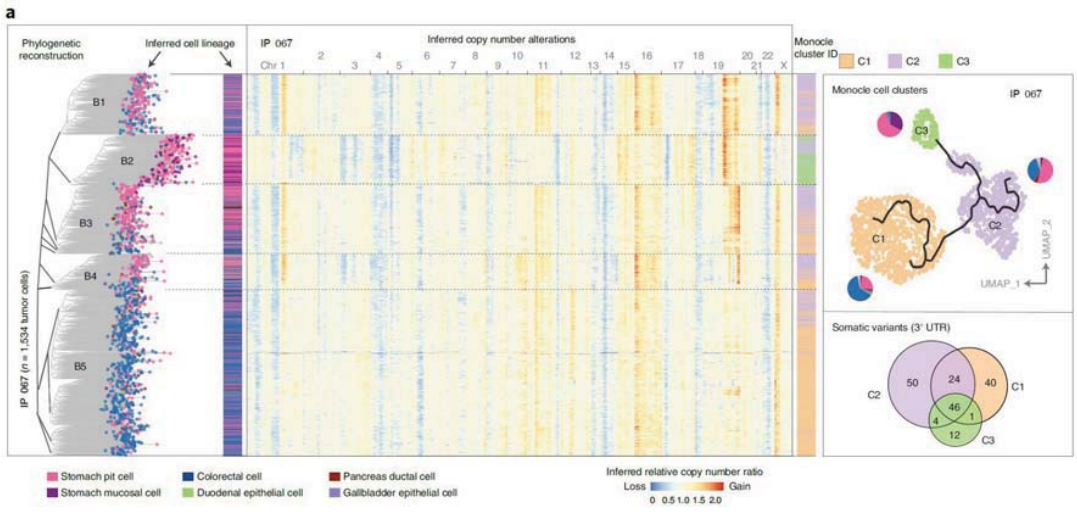
Ruiping Wang[1], Minghao Dang[1], Kazuto Harada[2,12], Guangchun Han[1], Fang Wang[3], Melissa Pool Pizzi[2], Meina Zhao[2], Ghia Tatlonghari[2], Shaojun Zhang[1], Dapeng Hao[1], Yang Lu[4], Shuangtao Zhao[1], Brian D. Badgwell[5], Mariela Blum Murphy[2], Namita Shanbhag[2], Jeannelyn S. Estrella[6], Sinchita Roy-Chowdhuri[6], Ahmed Adel Fouad Abdelhakeem[2], Yuanxin Wang[1], Guang Peng[7], Samir Hanash[7], George A. Calin[8], Xingzhi Song[1], Yanshuo Chu[1], Jianhua Zhang[1], Mingyao Li[9], Ken Chen[3], Alexander J. Lazar[6,10], Andrew Futreal[1], Shumei Song[2], Jaffer A. Ajani[2] ✉ and Linghua Wang[1,11] ✉

Intratumoral heterogeneity (ITH) is a fundamental property of cancer; however, the origins of ITH remain poorly understood. We performed single-cell transcriptome profiling of peritoneal carcinomatosis (PC) from 15 patients with gastric adenocarcinoma (GAC), constructed a map of 45,048 PC cells, profiled the transcriptome states of tumor cell populations, incisively explored ITH of malignant PC cells and identified significant correlates with patient survival. The links between tumor cell lineage/state compositions and ITH were illustrated at transcriptomic, genotypic, molecular and phenotypic levels. We uncovered the diversity in tumor cell lineage/state compositions in PC specimens and defined it as a key contributor to ITH. Single-cell analysis of ITH classified PC specimens into two subtypes that were prognostically independent of clinical variables, and a 12-gene prognostic signature was derived and validated in multiple large-scale GAC cohorts. The prognostic signature appears fundamental to GAC carcinogenesis and progression and could be practical for patient stratification.

SBi 한국생명정보학회 Korean Society for Bioinformatics


a
Survival probability
P = 3 × 10⁻⁶
Long-term survivor
Short-term survivor
Long 10 10 4 2 1
Short 10 0 0 0 0
Time since peritoneal metastasis (months)
n = 10
n = 10
GAC-PC discovery cohort (n = 20)
Ascites drainage
Cell isolation
Ascites cells
scRNA-seq
n = 45,048 cells
Tumor cells
t-SNE_2
t-SNE_1
Tumor cell
CD8 T cell
CD4 T cell
B Cell
Myeloid cell
Fibroblast


c
Cell count
Inferred cell lineage
Stomach_Mucosal cell
Stomach_Pit cell
Stomach_Chief cell
Pancreas_Ductal cell
GallbladderWall_Mucous epithelial cell
Duodenum_Enterocyte progenitor
Duodenum_Goblet cell
TransverseColon_Goblet cell
TransverseColon_Enterocyte
TransverseColon_Enterocyte progenitor
SigmoidColon_Enterocyte progenitor
Rectum_Enterocyte
Rectum_Inflammed epithelial cell
Breast_Epithelium_luminal cell
Other
Proportion of cells (%)
Stomach (69.5%)
Pancreas
Gallbladder
Duodenum
Colon
Rectum
Intestine (21.2%)
IP 093 IP 124 IP 009 IP 114 IP 081 IP 057 IP 109 IP 007 IP 073 IP 056 IP 158 IP 010 IP 031 IP 067 IP 070
Short-term
Long-term
Survival
Intestinal metaplasia
Yes
No
NA
Percentage of intestinal cell
Gastric-dominant
GI-mixed
Non-GI


a
Phylogenetic reconstruction
Inferred cell lineage
IP 067
Inferred copy number alterations
Monocle cluster ID
C1 C2 C3
B1 B2 B3 B4 B5
IP 067 (n = 1,534 tumor cells)
Monocle cell clusters
IP 067
UMAP_1
UMAP_2
Somatic variants (3' UTR)
50 24 40 46 1 4 12
Stomach pit cell
Stomach mucosal cell
Colorectal cell
Duodenal epithelial cell
Pancreas ductal cell
Gallbladder epithelial cell
Inferred relative copy number ratio
Loss Gain
b
IP-009
IP 009 (n = 2,250 tumor cells)
26 69 21


Thank you
Contact
Semin Lee
Email: seminlee@unist.ac.kr
Organized by
SBi 한국생명정보학회
Korean Society for Bioinformatics