KSBi-BIML 2024

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

생명정보학 & 머신러닝 워크샵 (온라인)

Big Data for RNA Informatics

임수빈 _ 아주대학교

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 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

강의개요

Big data for RNA informatics

최근 생성되는 전사체 데이터셋들은 다양한 open repository 데이터베이스들을 통하여 scientific community와 공유되어지고 있는 실정임에도 불구하고 제한된 patient cohort 크기와 QC-passed 된 세포의 수, 임상 정보와 cell metadata 정보의 부재 등으로 인하여 새로운 결과를 도출해 내기에 현실적으로 많은 한계들이 있다. 이를 극복하기 위하여 하나의 큰 big data, 즉 통합된 데이터를 생성하여 uniform 한 파이프라인을 적용하여 효율적이고 효과적인 분석을 할 수 있는 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- Bulk RNA-Seq 개요
- Single-Cell RNA-Seq 개요
- Data Integration for RNA Informatics
- Deep Learning for scRNA-seq
- Spatial RNA informatics

* 강의 난이도: 초급

* 강의: 임수빈 교수 (아주대학교 의과대학)

Curriculum Vitae

Speaker Name: Su Bin Lim, Ph.D.

▶ **Personal Info**

Name	Su Bin Lim
Title	Assistant Professor
Affiliation	Ajou University School of Medicine

▶ **Contact Information**

Address	Worldcup-Ro 164, Yeongtong-Gu, Suwon 16499, South Korea
Email	sblim@ajou.ac.kr
Phone Number	031-219-5056

Research Interest

RNA informatics, computational genomics, systems biology, single-cell analysis

Educational Experience

2015	B.S. in Biomedical Engineering, National University of Singapore, Singapore
2019	Ph.D. in Integrative Sciences and Engineering, National University of Singapore, Singapore

Professional Experience

2020-2021	Postdoctoral Fellow, Johns Hopkins University School of Medicine, USA
2021-	Assistant Professor, Ajou University School of Medicine, South Korea
2022-	Nature Scientific Data, Editorial Board Member
2023-	Frontiers in Cell and Developmental Biology, Editorial Board Member

Selected Publications (5 maximum)

1. SB Lim et al. An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. Nature Communications 8, 1736, 2017.
2. SB Lim et al. Addressing cellular heterogeneity in tumor and circulation for refined prognostication. PNAS 116(36), 2019.
3. KY Goh et al. Matrisomal genes in squamous cell carcinoma of head and neck influence tumor cell motility and response to cetuximab treatment. Cancer Communications 42(4), 355-359, 2022.
4. SB Lim et al. Macrophage-derived TNF-enriched tumor microenvironment shapes pancreatic ductal adenocarcinoma into the basal-like molecular phenotype through upregulating TAp63. Clinical and Translational Medicine 13, 12, 2023
5. Hong J et al. SRSF7 downregulation induces cellular senescence through generation of MDM2 variants. Aging 15, 14591-14606, 2023.

KSBI | 한국생명정보학회

KSBi-BIML 2024

Big Data for RNA Informatics

Su Bin Lim, PhD

Ajou Univ. School of Medicine

sblim@ajou.ac.kr

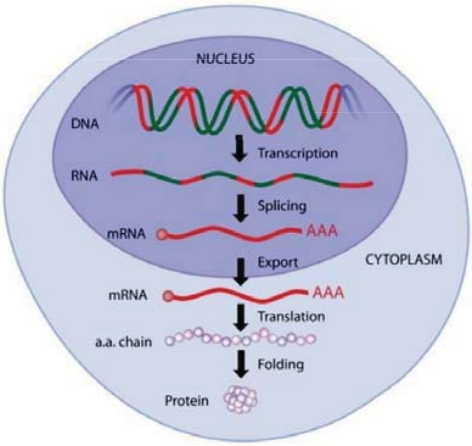
Lecture Outline

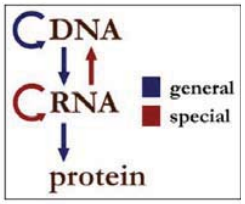
- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

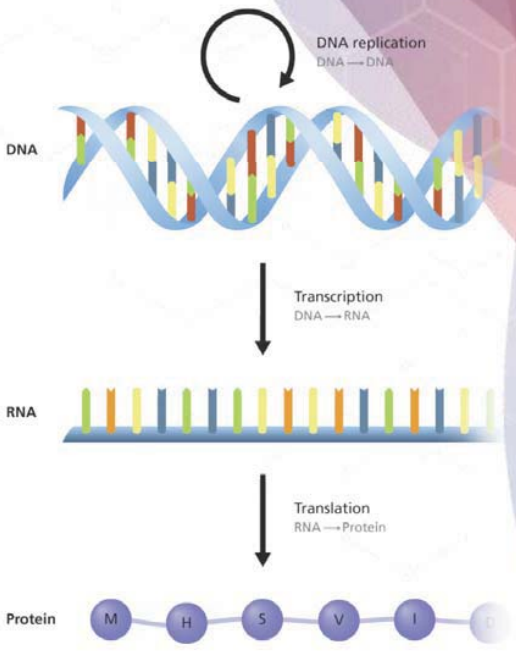
Lecture Outline

- **Bulk** transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

Central Dogma of Biology

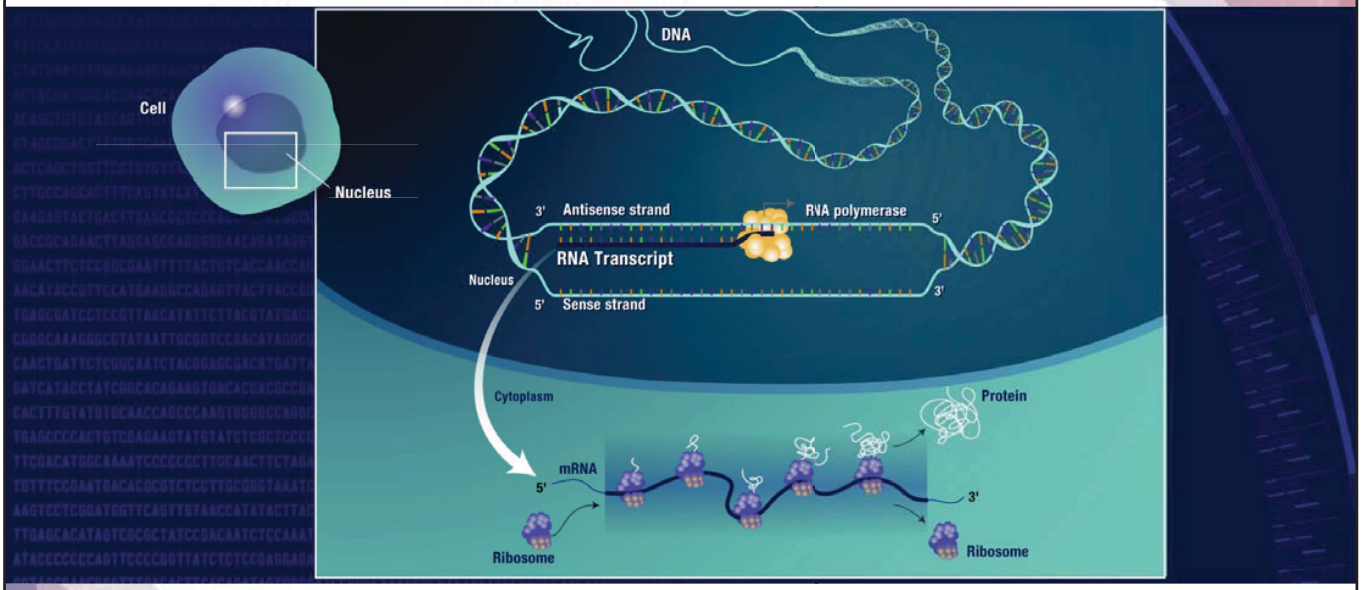






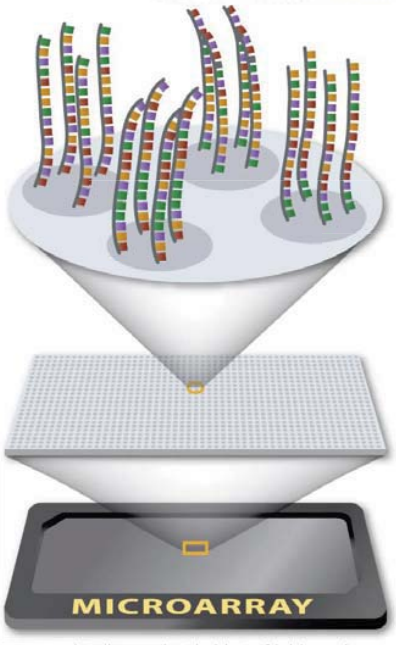
An illustration showing the flow of information between DNA, RNA and protein.
Image credit: Gnome Research Limited

전사체(Transcriptome)란?

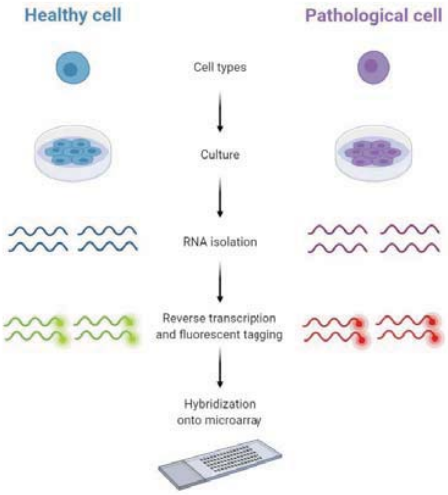


NIH-National Human Genome Research Institute

전사체 분석 방법 – (1) DNA microarray



https://learn.genetics.utah.edu/content/labs/microarray/



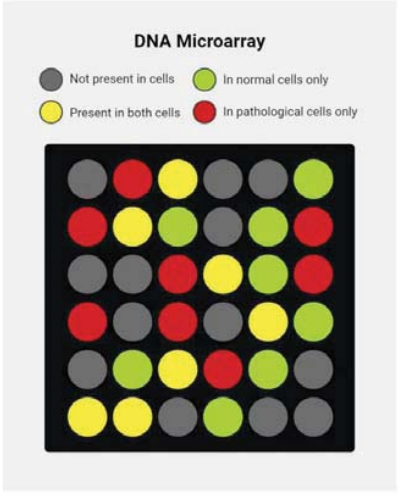
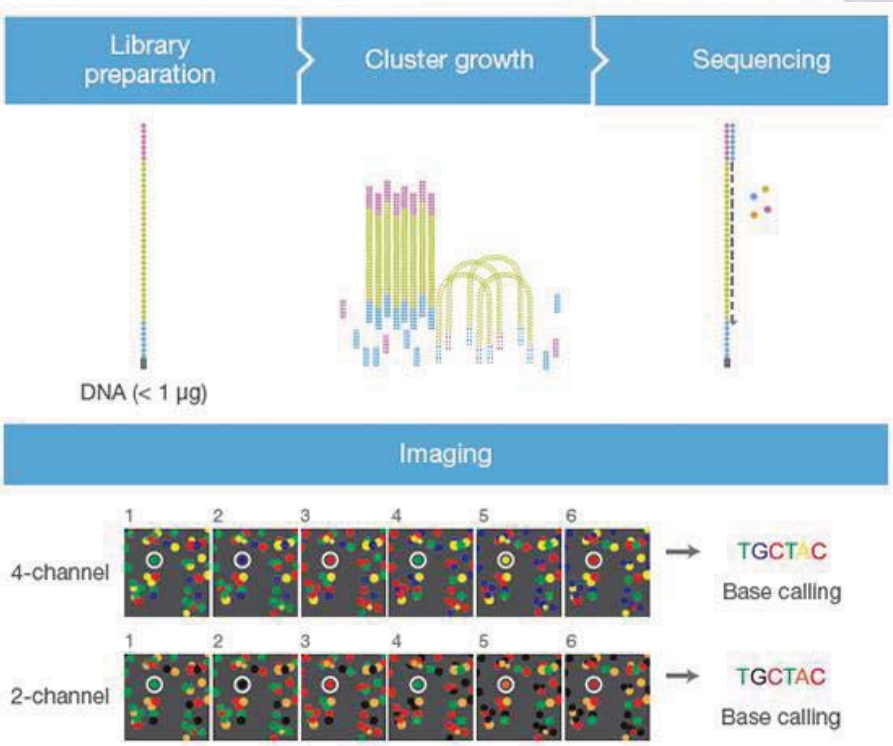


Image By Sagar Aryal, created using biorender.com

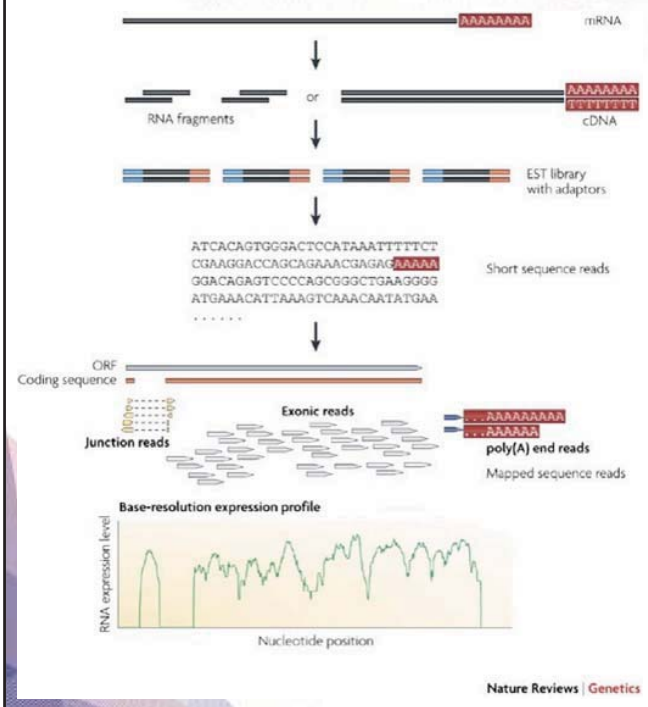
전사체 분석 방법 – (2) NGS Platform



https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

7

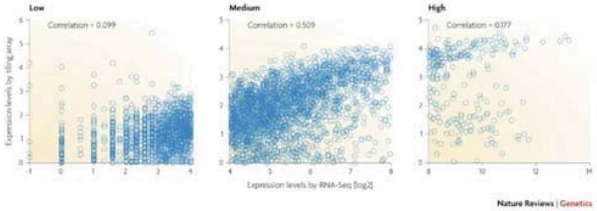
RNA-seq 기본 원리



Advantages of RNA-seq

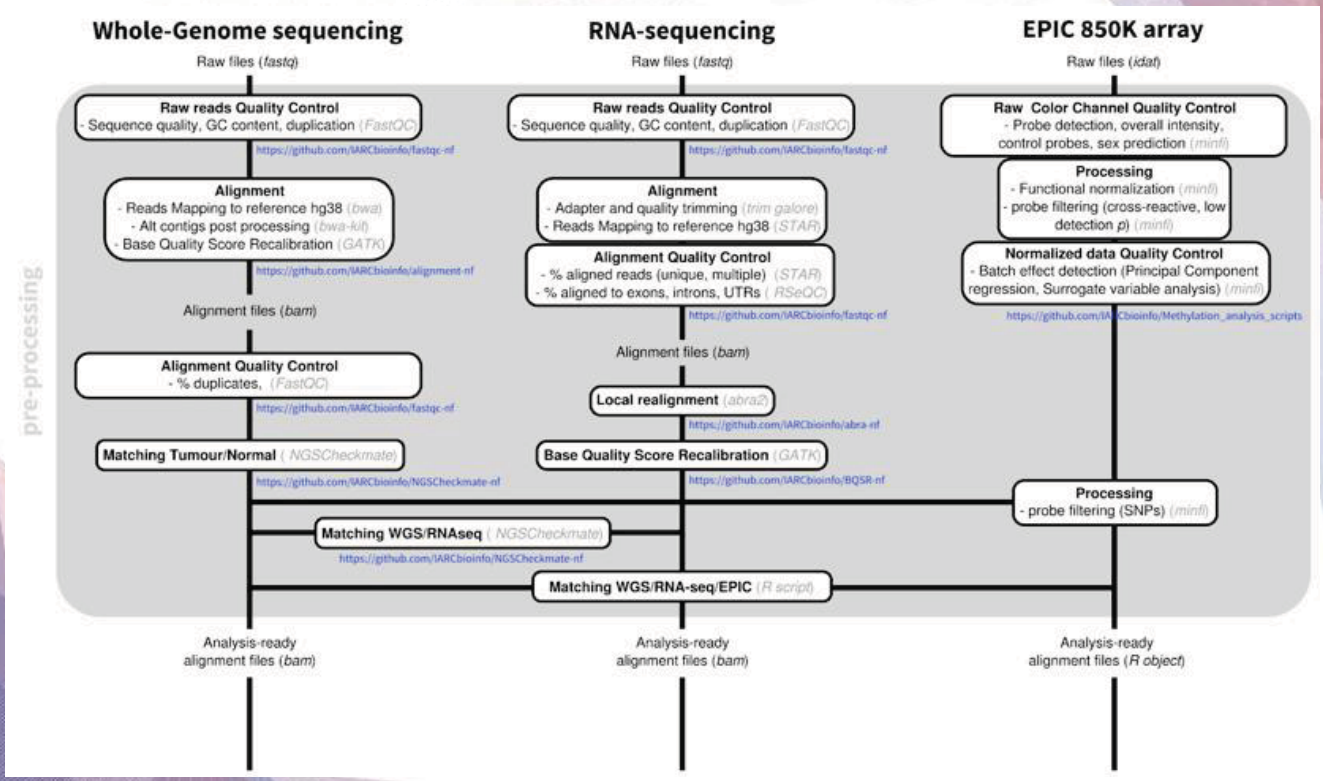
Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
	Technology specifications		
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
	Application		
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
	Practical issues		
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

RNA-seq vs. microarray

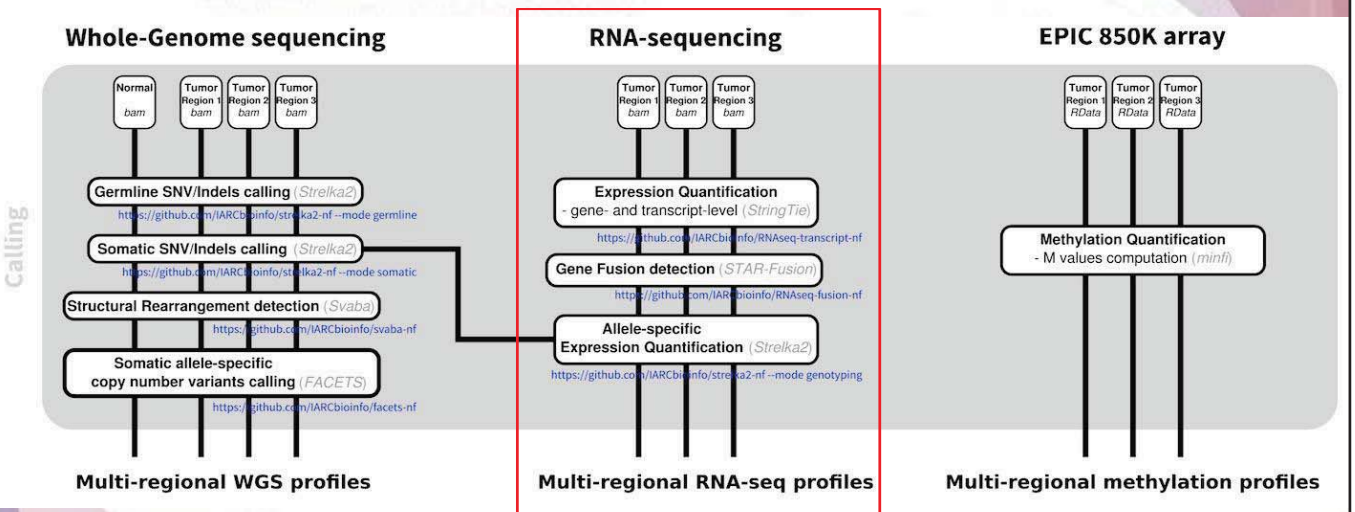


8

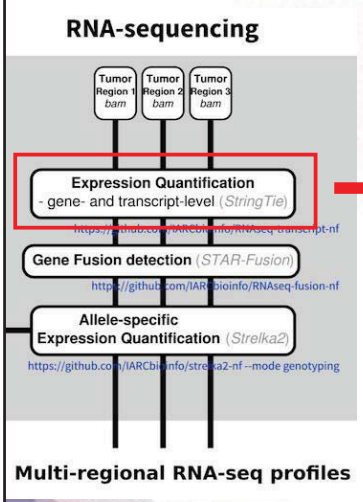
Bioinformatics pipeline for multi-omic data processing: (1) Mapping (alignment)

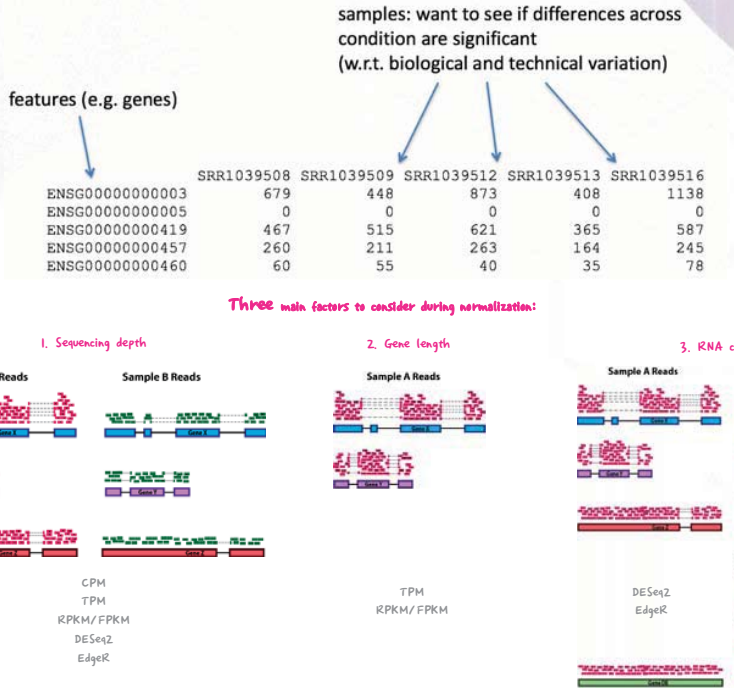


Bioinformatics pipeline for multi-omic data processing: (2) Counting (quantification)



Bioinformatics pipeline for multi-omic data processing: (3) Normalization



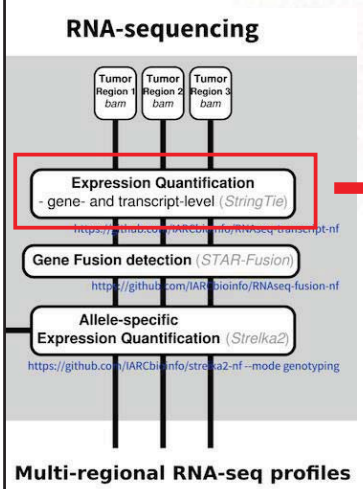


https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html

Lecture Outline

- **Bulk** transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

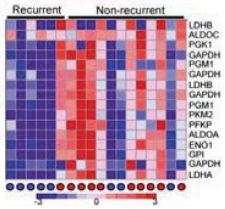
Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (1-4)



1. Expression heatmap

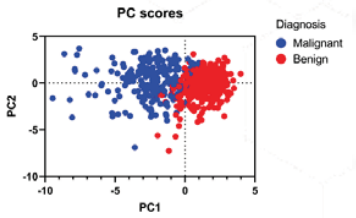
(Morpheus, Broad Institute)

https://software.broadinstitute.org



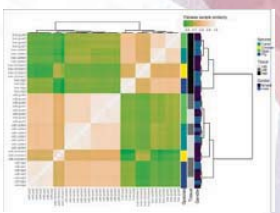
2. PCA

(Principal component analysis)



3. Hierarchical clustering

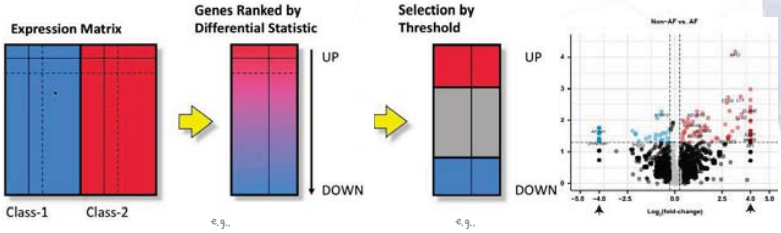
(Principal component analysis)



4. Differential expression (DE) analysis

Tools:

Ballgown
baySeq
Cuffdiff
DESeq2
EBSeq
edgeR + exact test
edgeR + GLM
limma trend
limma voom
NOISeq
SAMseq
...



e.g.,

Fold change (log2 ratio)

t values from t-test

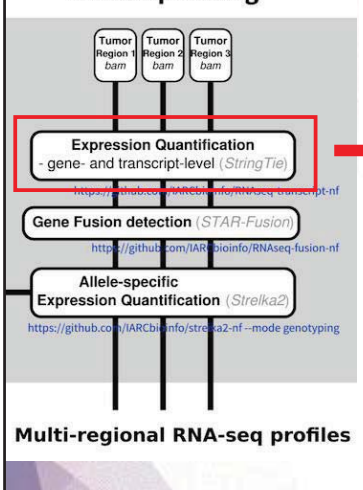
sign(logFC)-log10(pval)

e.g.,

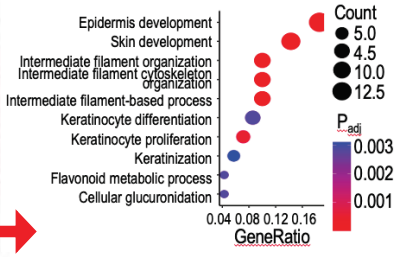
|log2FC| > 0.66 (50% change)

adjusted pval < 0.05

Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (5-7)



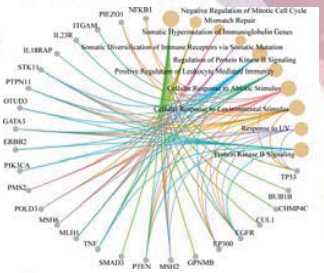
5. GO (gene ontology) / enrichment analysis



Tools:

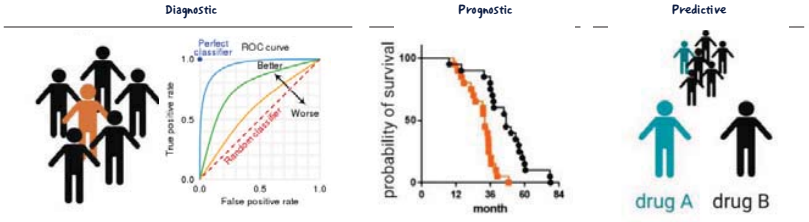
DAVID
GOrilla
QuickGO
GeneGO MetaCore
GOnet
GOATOOLS
GOLEM
AmiGO
GOEAST
GOFFA
ClsuterProfiler
...

6. Gene-concept network

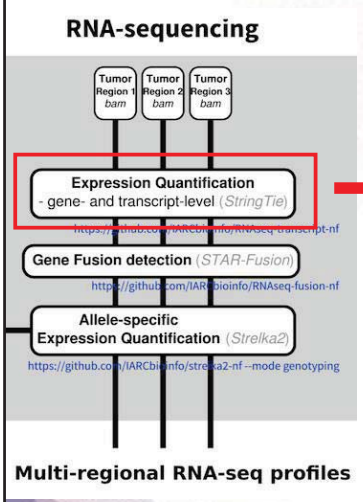


7. Tumor biomarker discovery:

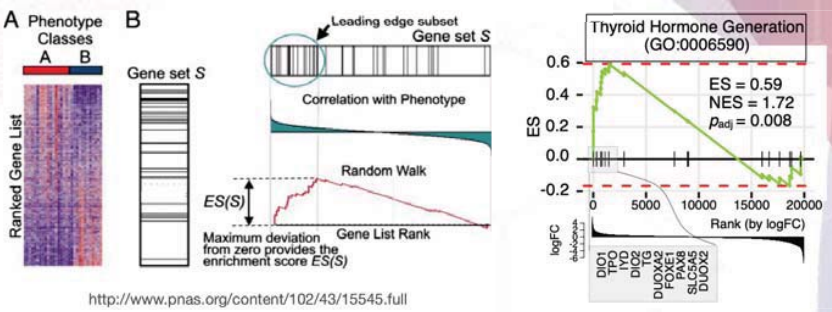
diagnostic, prognostic, and predictive



Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (8-9)



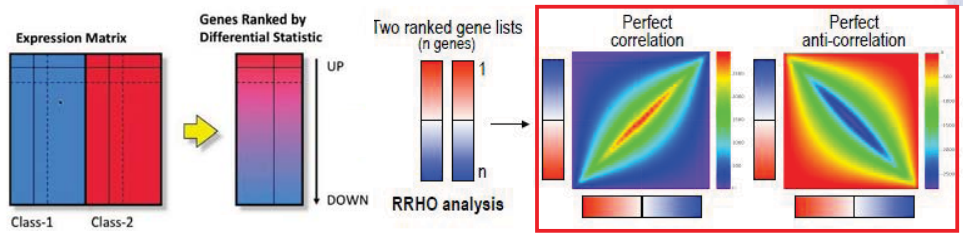
8. Gene set enrichment analysis (GSEA)



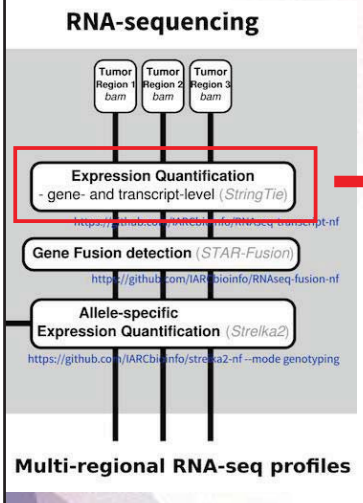
http://www.pnas.org/content/102/43/15545.full

9. RRHO analysis

https://systems.crump.ucla.edu/rankrank/rankranksimple.php

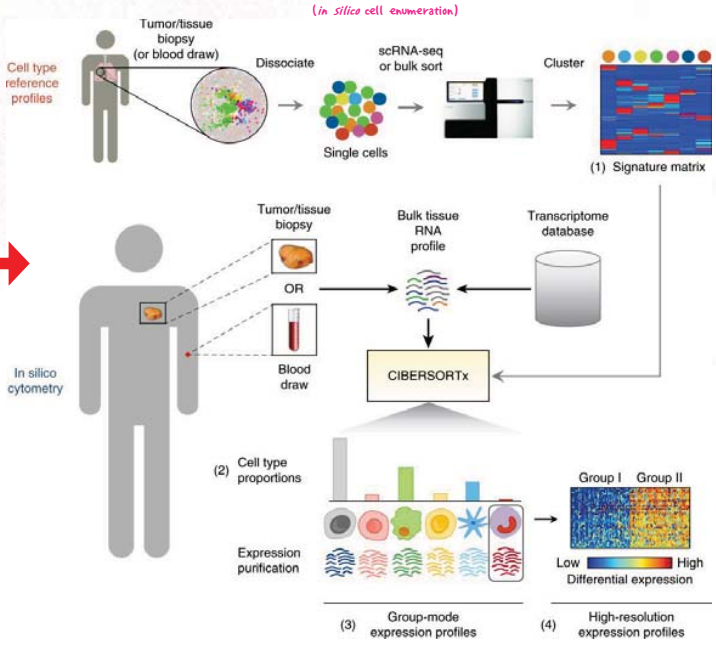


Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (10)



10. Deconvolution

(*in Silico* cell enumeration)



Deconvolution tools:

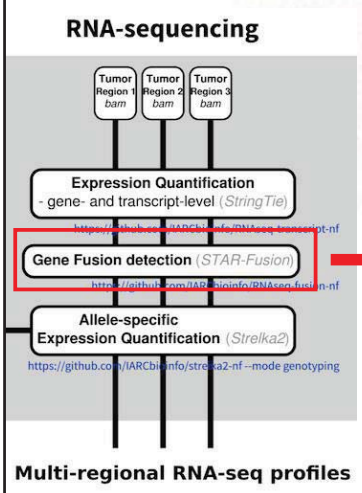
CIBERSORT
OLS
NNLS
FARDEEP
RLR
Lasso
Ridge
DCQ
Elastic net
DSA
EPIC
dtangle
ssFrobenius
ssKL
DeconRNASeq
...

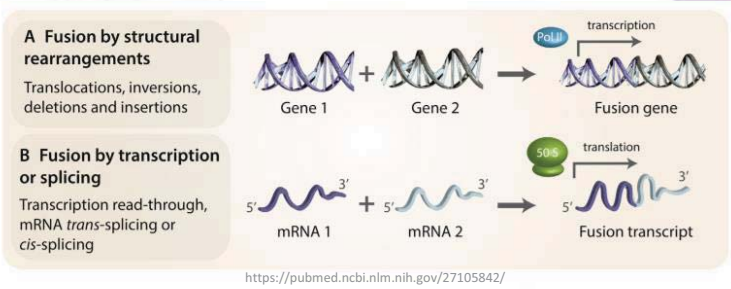
Using scRNA-seq data as reference:

CIBERSORTx
Bisque
deconvSeq
DWLS
MuSiC
SCDC
..

https://www.nature.com/articles/s41587-019-0114-2

Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (11)





https://pubmed.ncbi.nlm.nih.gov/27105842/

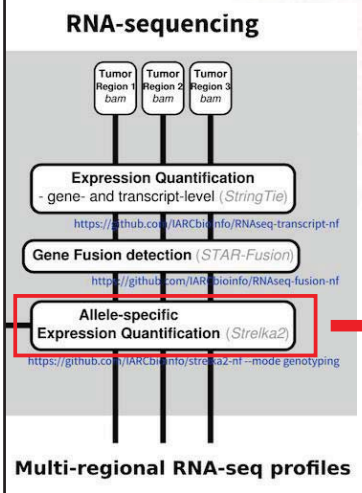
Structural variant detection (WGS as input):

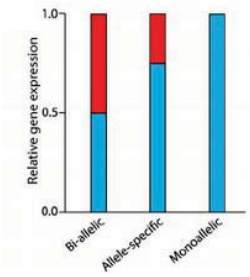
- BreakDancer
- CREST
- GASV
- HYDRA
- PEMer
- R453PlusToolBox
- SVDetect
- VariationHunter
- ...

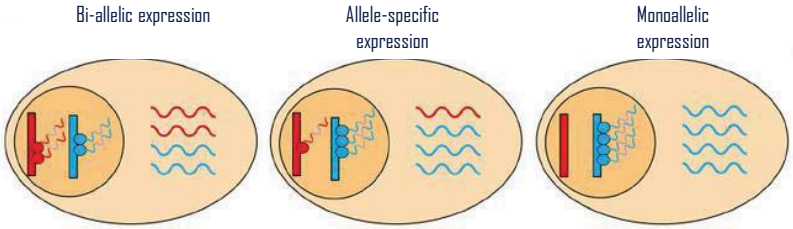
Fusion detection specific (RNA-seq as input):

- BreakFusion
- ChimeraScan
- Comrad
- FusionAnalyser
- defuse
- FusionMap
- FusionHunter
- FusionSeq
- ShortFuse
- SnowShoes-FTD
- SOAPfusion
- Tophat-Fusion
- ...

Analytical tools for bulk (pooled cells, tissues, or biopsies) RNA-seq (12)





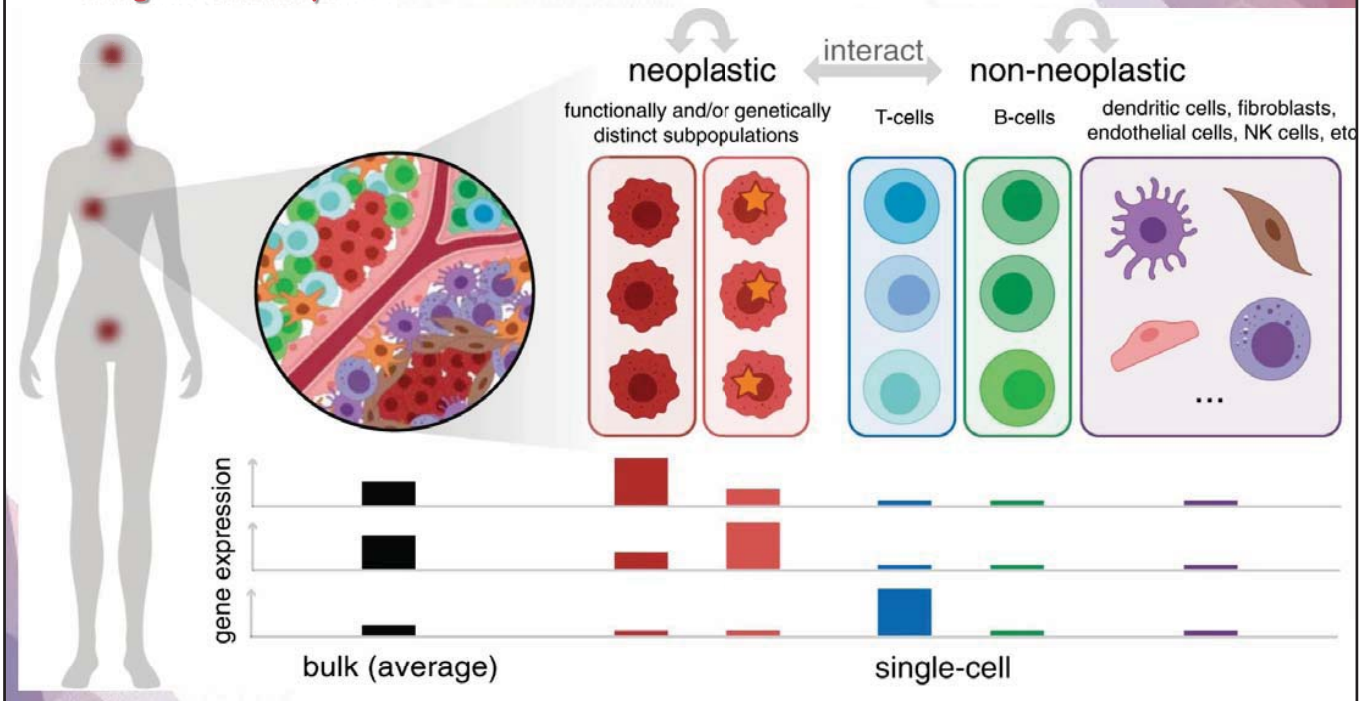


https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004304

Lecture Outline

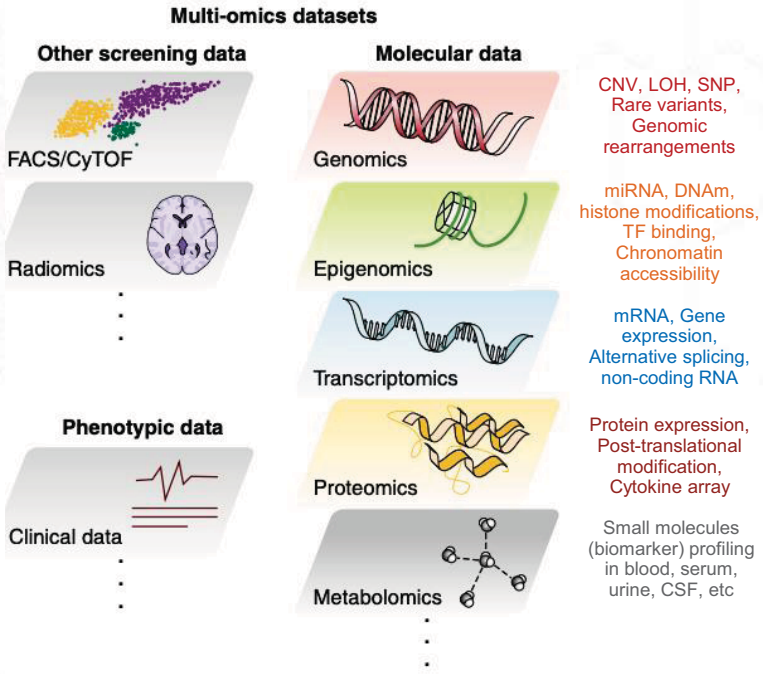
- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

Single-cell analysis



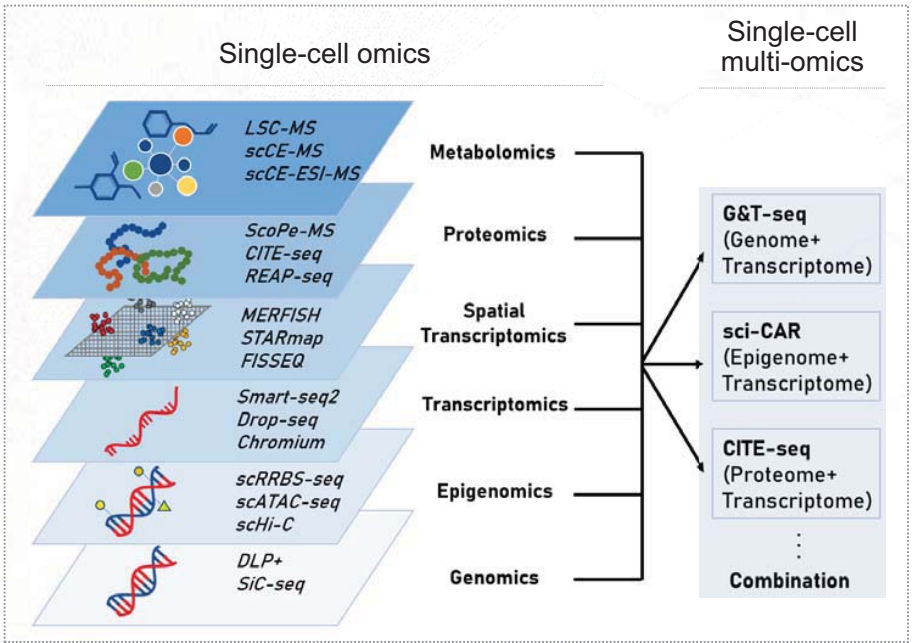
Experimental & Molecular Medicine 52, 1452-1465 (2020)

Single-cell omics



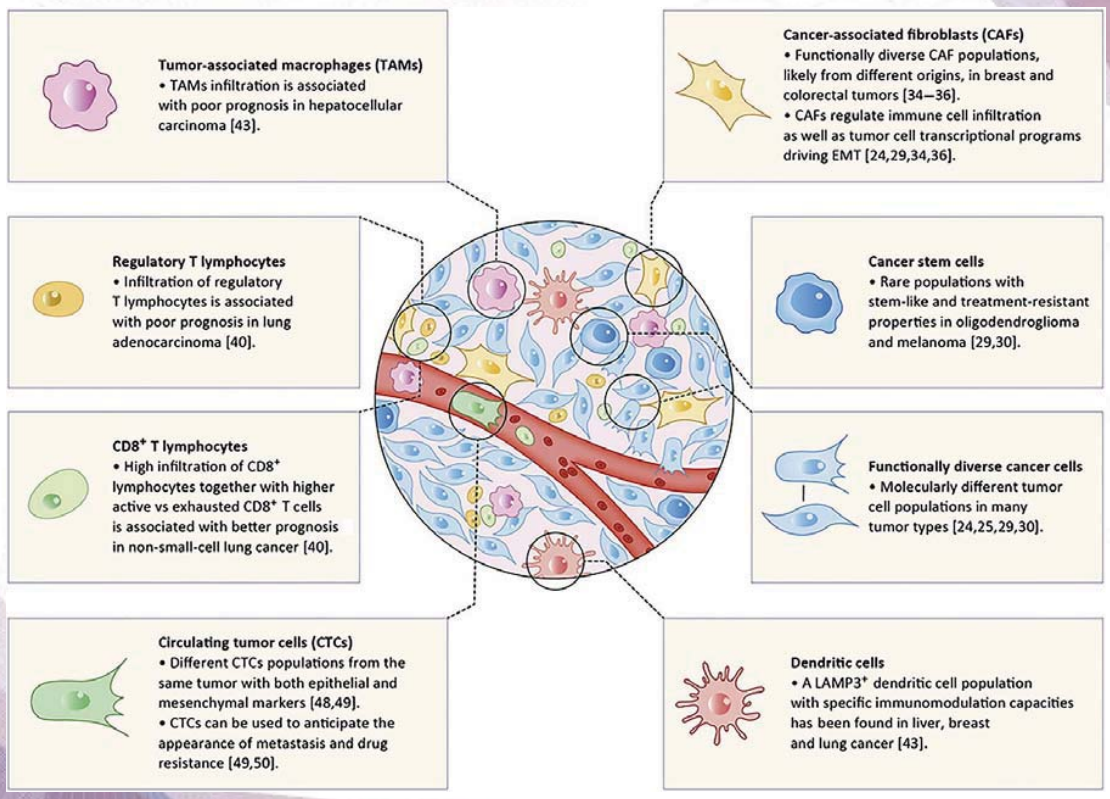
Nature Computational Science 1, 395-402, 2021

Single-cell omics



https://new.ksbmb.or.kr/html/?pmode=webzine&smode=viewDetail&id=201601&menu=379&seq=7762

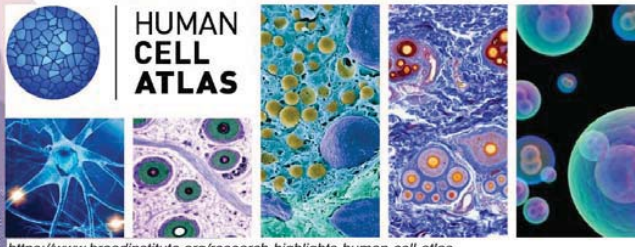
Single-cell analysis in cancer



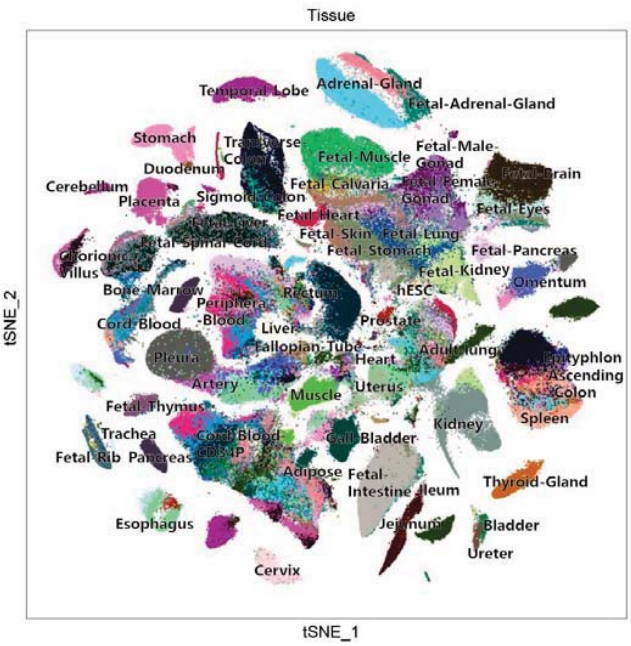
Trends in Cancer 6(1), P13-19, 2020

Single-cell analysis

A zebrafish embryo at an early stage of development. Fluorescent markers highlight cells expressing genes that help determine the type of cell they will become. (JEFFREY FARRELL, SCHIER LAB/HARVARD UNIVERSITY) - https://vis.sciencemag.org/breakthrough2018/finalists/

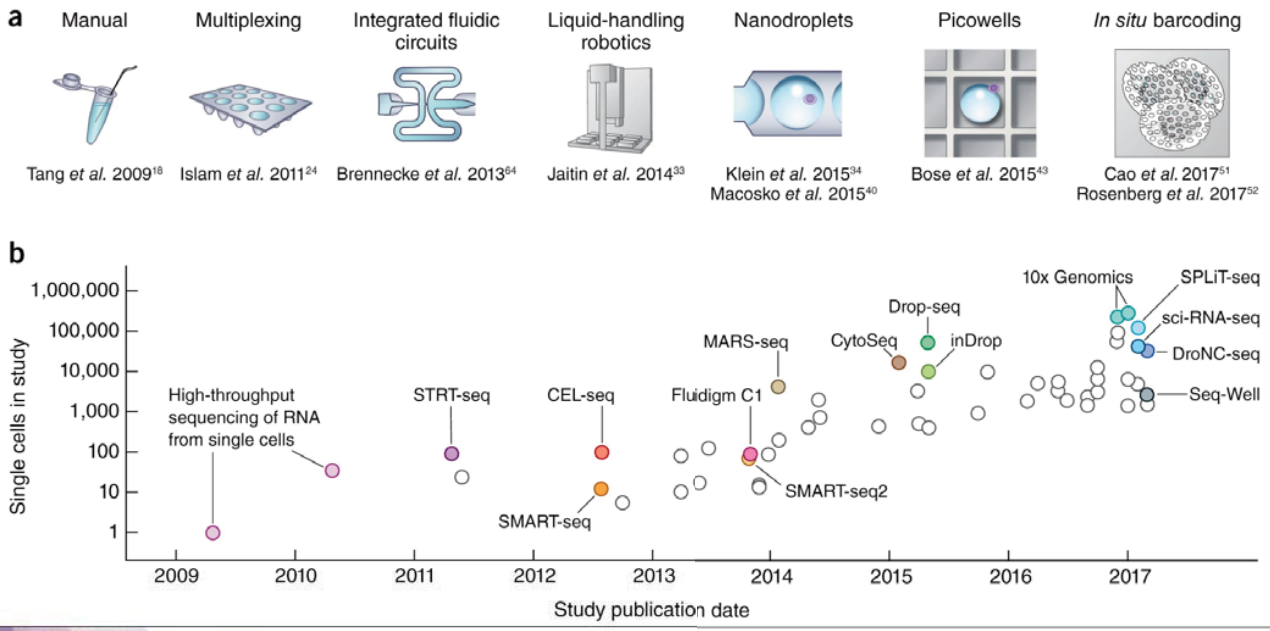


https://www.broadinstitute.org/research-highlights-human-cell-atlas



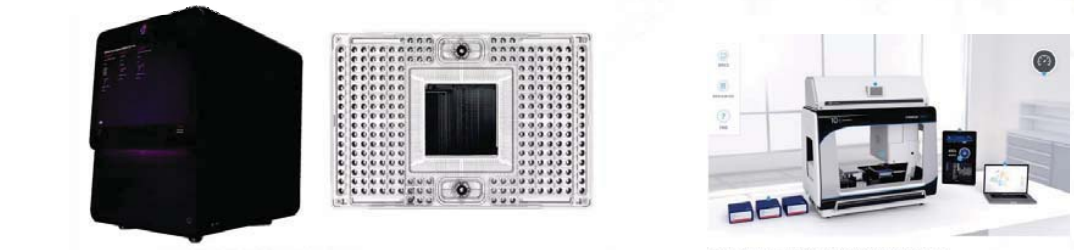
Han X et al. (2020) Construction of a human cell landscape at single-cell level. Nature

Single-cell transcriptomic analysis (scRNA-seq)



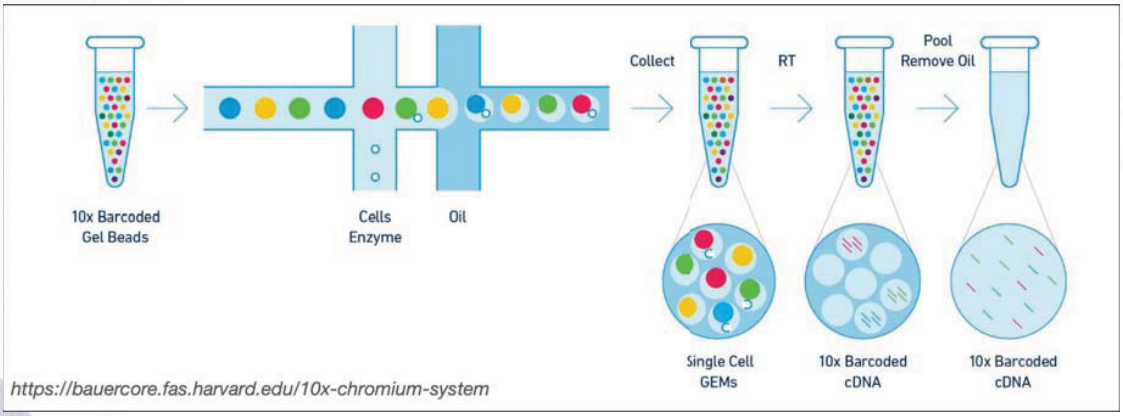
Nature Protocols 13, 559-604 (2018)

Commercialized products for scRNA-seq



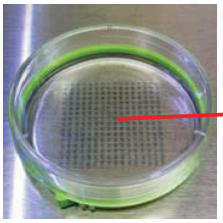
Fluidigm's Polaris System (left) and associated chip (right) with precisely designed integrated fluidic circuit (www.fluidigm.com)

https://www.facebook.com/10xGenomics/photos/a.384002715443493/876620926181667/?type=3&theater

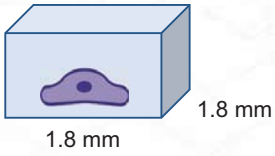


Microfluidic based automated single-cell sorter (iota Sciences)

isoCell

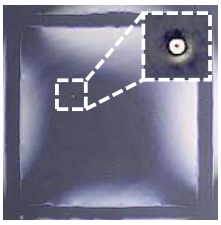


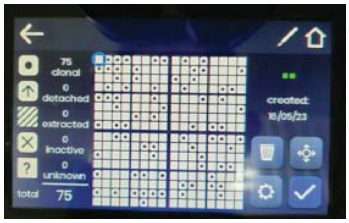
Area per chamber : 3.24 mm^2
Volume per chamber : 600 ~ 800 nL



256 culture chambers on 60-mm petri dish

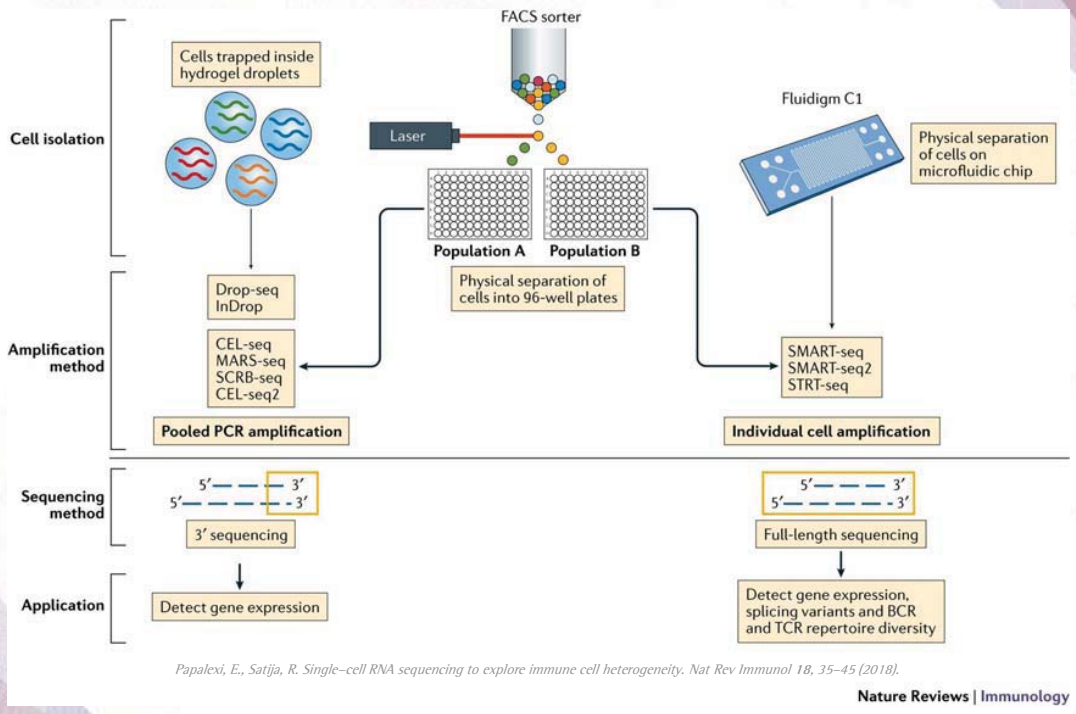
isoHub





Up to 94 single cell chambers per dish
(out of 256). Limited by Poisson distribution

scRNA-seq data generation



Papalexi, E., Satija, R. Single–cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 18, 35–45 (2018).

Nature Reviews | Immunology

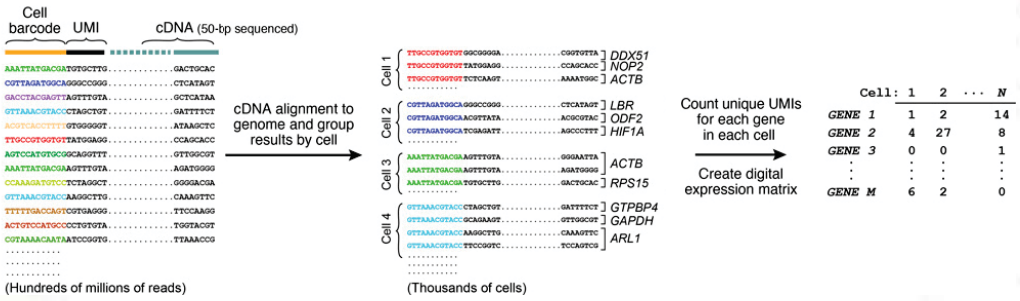
Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

Raw data to count matrix (gene expression)

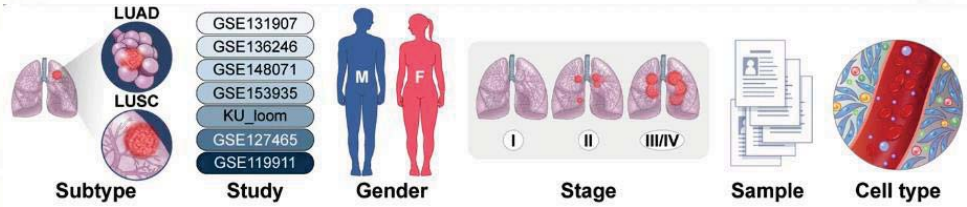
A typical processed scRNA-seq dataset has
(1) count matrix and (2) cell-level metadata

(1) Count matrix (i.e., gene expression profiles)

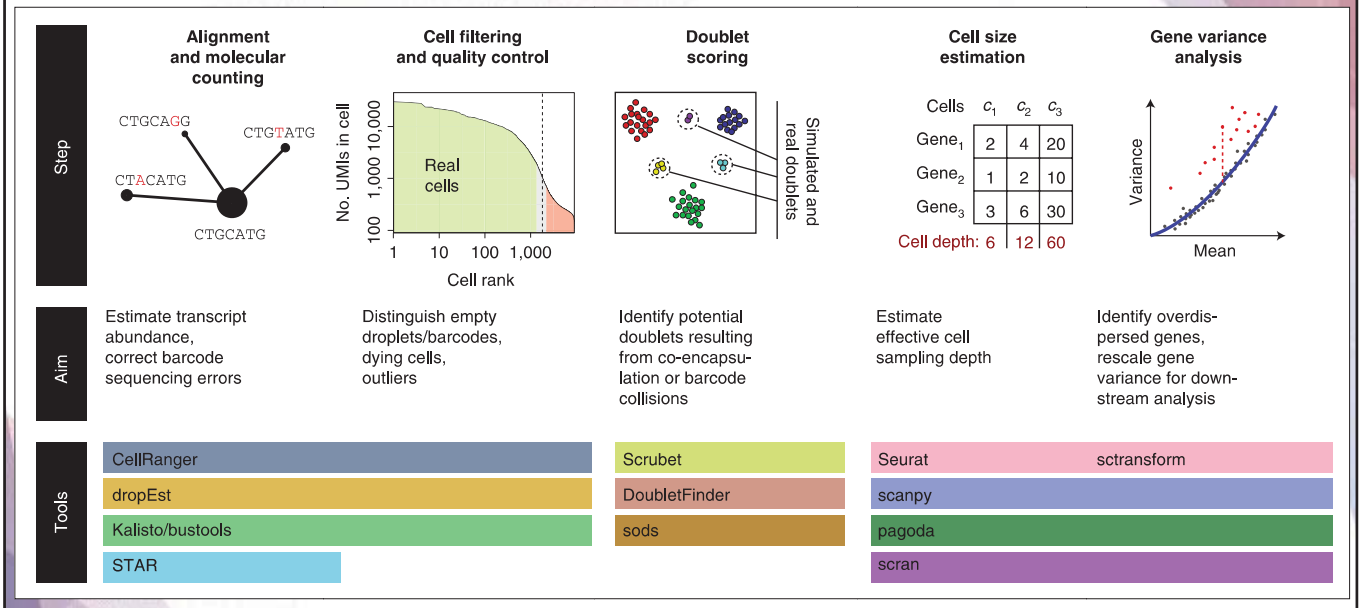


https://www.elveflow.com/microfluidic-reviews/droplet-digital-microfluidics/drop-seq/

(2) Metadata

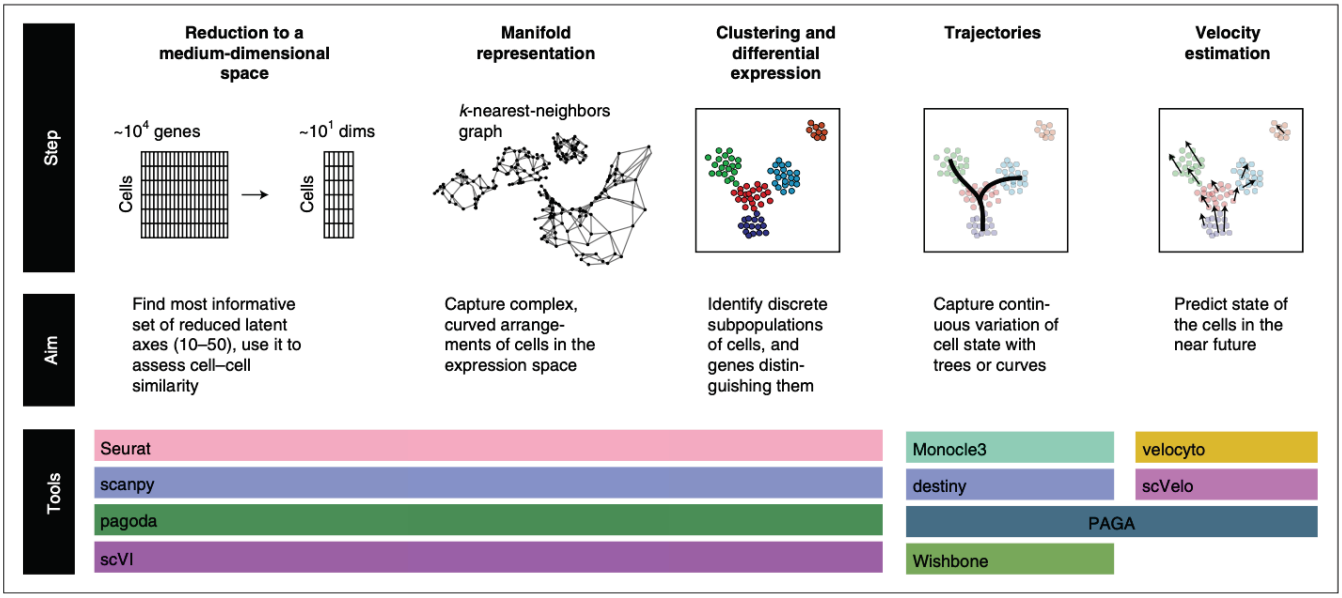


Workflow for scRNA-seq data analysis (preprocessing)



Nature Methods 18(7), 723-732, 2021

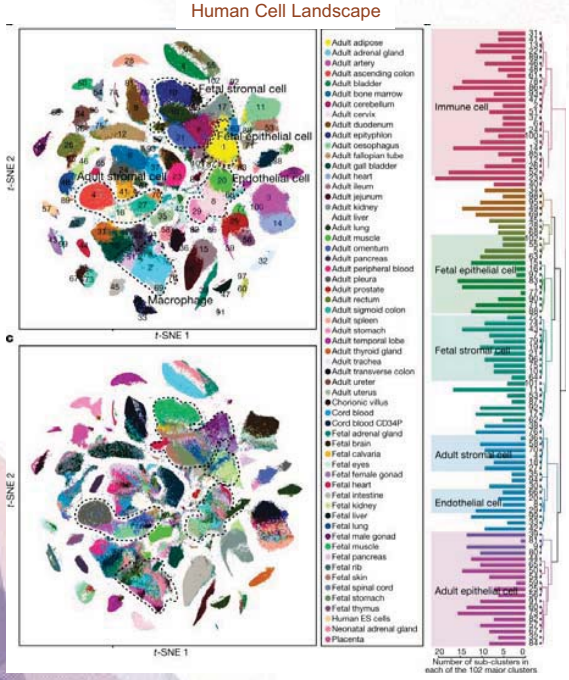
Workflow for scRNA-seq data analysis



Nature Methods 18(7), 723-732, 2021

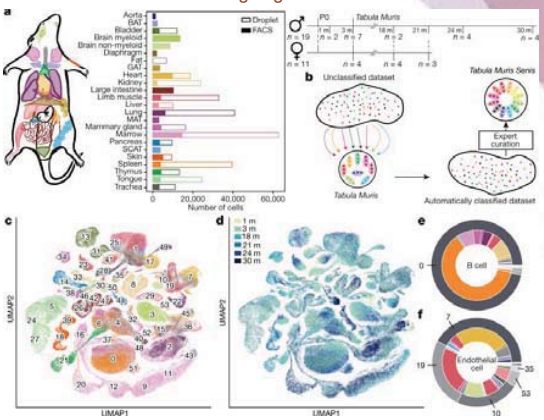
Single-cell atlas

Human Cell Landscape



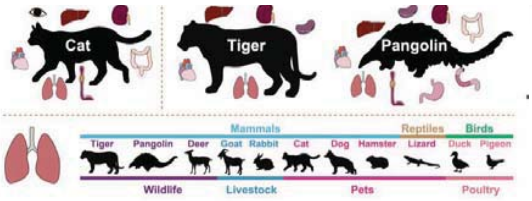
Nature 581, 303-309, 2020

Mouse Ageing Cell Atlas



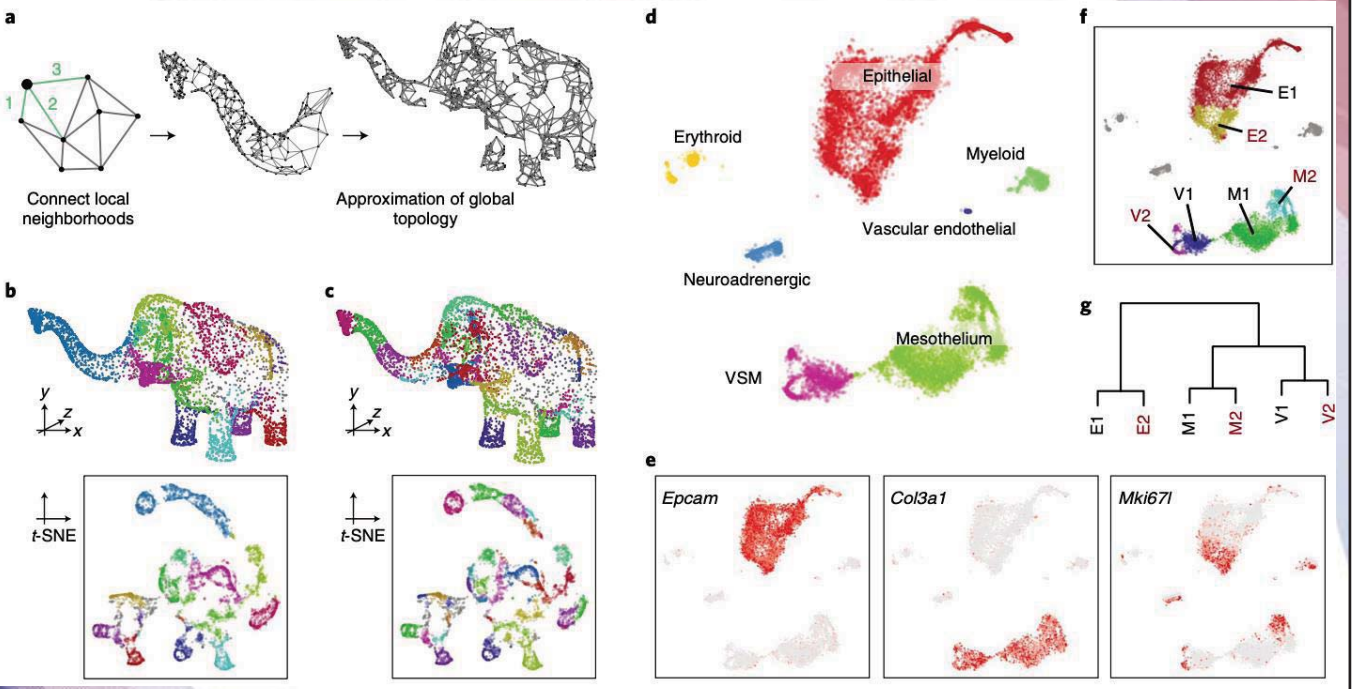
Nature 583, 590-595, 2020

Single-Cell Atlas for 11 non-model mammals, reptiles and birds



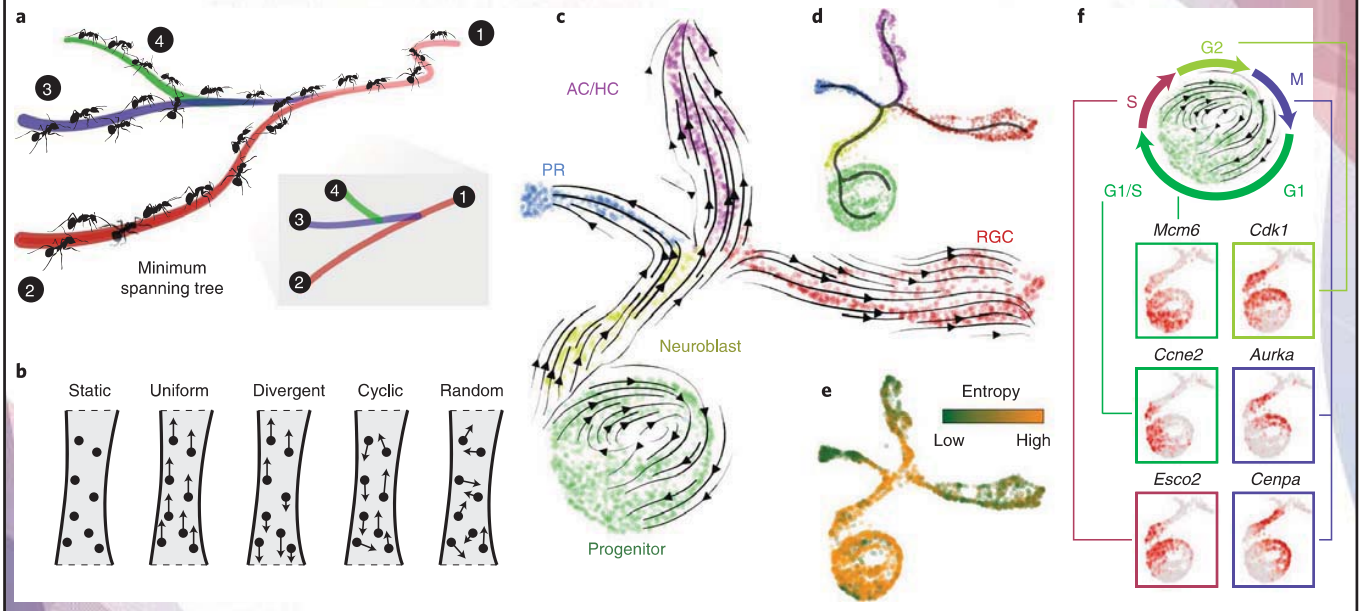
Nature Communications 12, 7083, 2021

Approximating and partitioning complex manifolds



Nature Methods 18(7), 723-732, 2021

Approximating dynamical processes



Nature Methods 18(7), 723-732, 2021

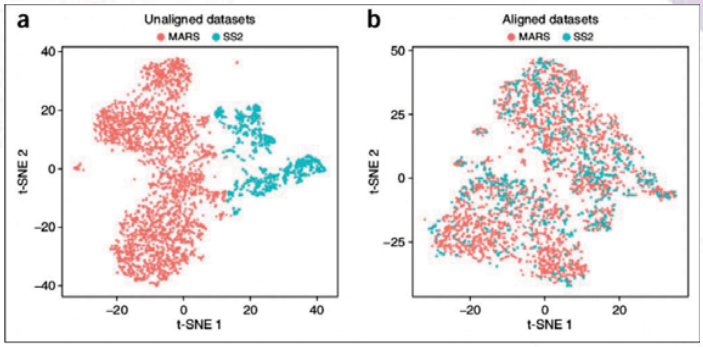
Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- **Data integration and batch effect correction**
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

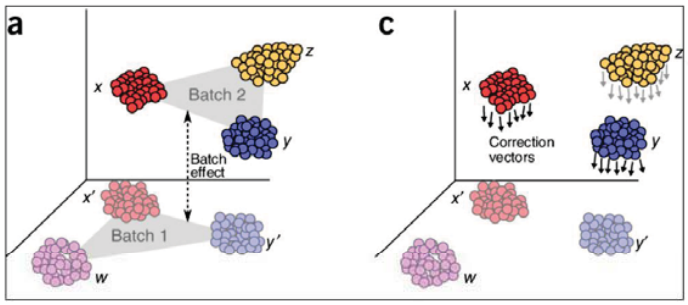
Batch effect correction

Computational methods

1. Seurat
2. Harmony
3. fastMNN
4. MNN Correct
5. ComBat
6. Limma
7. Scene
8. Scanorama
9. MMD-ResNet
10. ZINB-WaVE
11. scMerge
12. LIGER
13. BBKNN



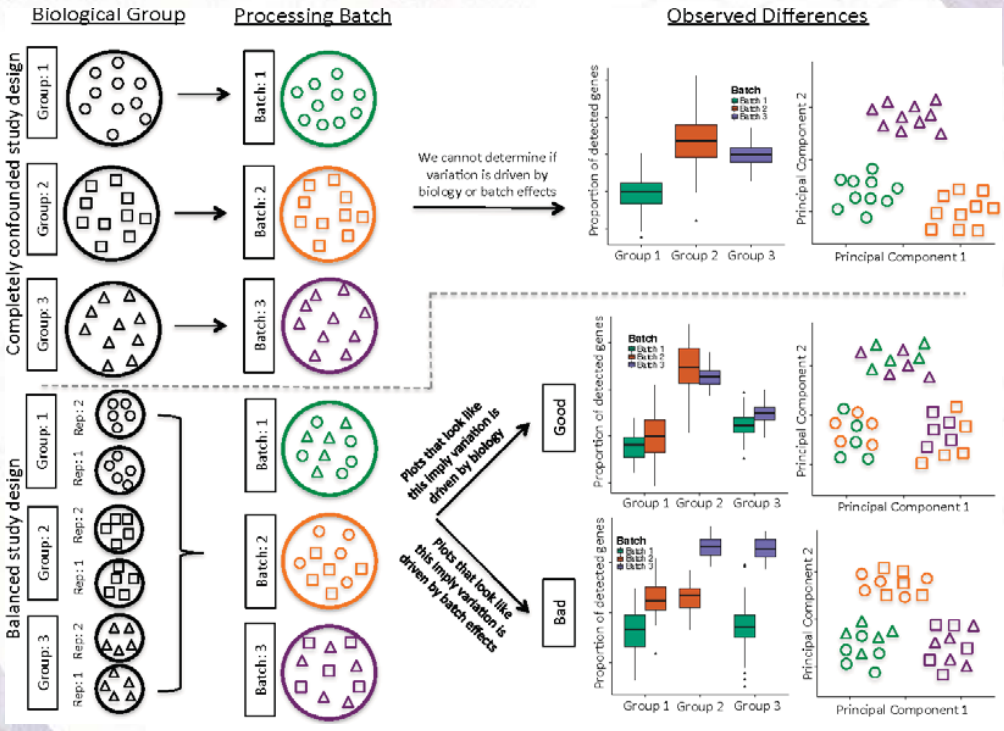
https://www.nature.com/articles/nbt.4096



https://www.nature.com/articles/nbt.4091.pdf

37

Batch effect correction (실험적 기법)



doi: 10.1101/025528 (2015)

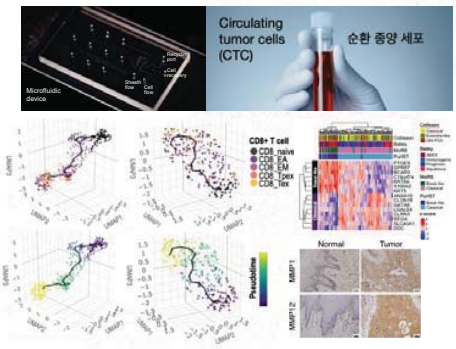
38

Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

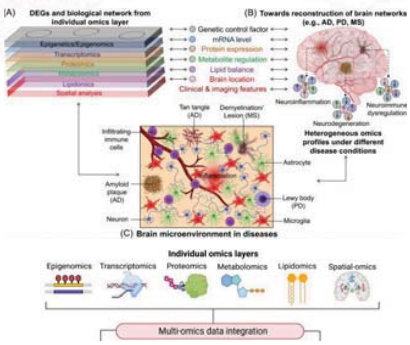
Generation of single-cell atlas

Cancer (CTC, tissue, PBMC)



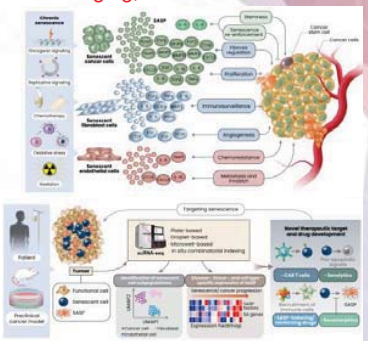
Nature Communications 8 (1), 1-11, 2017
Scientific Data 5, 180136, 2018
Scientific Data 6, 194, 2019
PNAS 116 (36), 17957-17962, 2019
npj Precision Oncology 3, 23, 2019
EMBO Reports 21 (2), e49749, 2020
Cancer Communications 42 (4), p.355-359, 2022
Scientific Data 10, 167, 2023
Cancer Communications, 43 (4), p.455-479, 2023
Advanced Science, 2201663, 2023
Clinical and Translational Medicine, 13(12), 2023

Inflammation, neuroscience



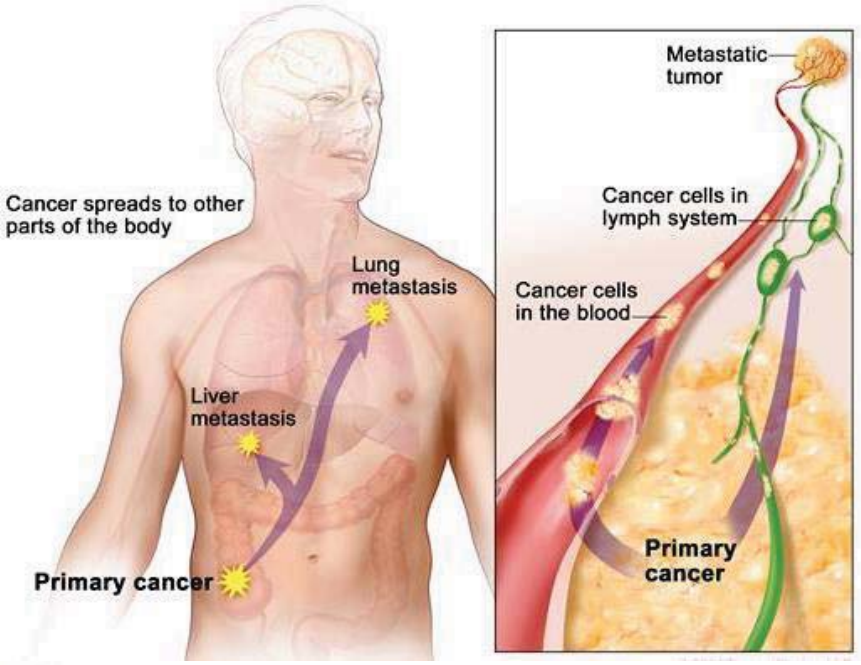
Science Advances 7 (21), eabg9614, 2021
Nature Neuroscience 24, pages1673–1685, 2021
npj Precision Oncology 3, 15, 2019
Cell Stem Cell , Vol. 29 Issue 4 Pages 610-619, 2022
Journal of Pharmaceutical analysis 13(8):1816-1821, 2023
Scientific Data 10, 861, 2023
Journal of Brain Research 3(3), 2020

Aging, cellular senescence



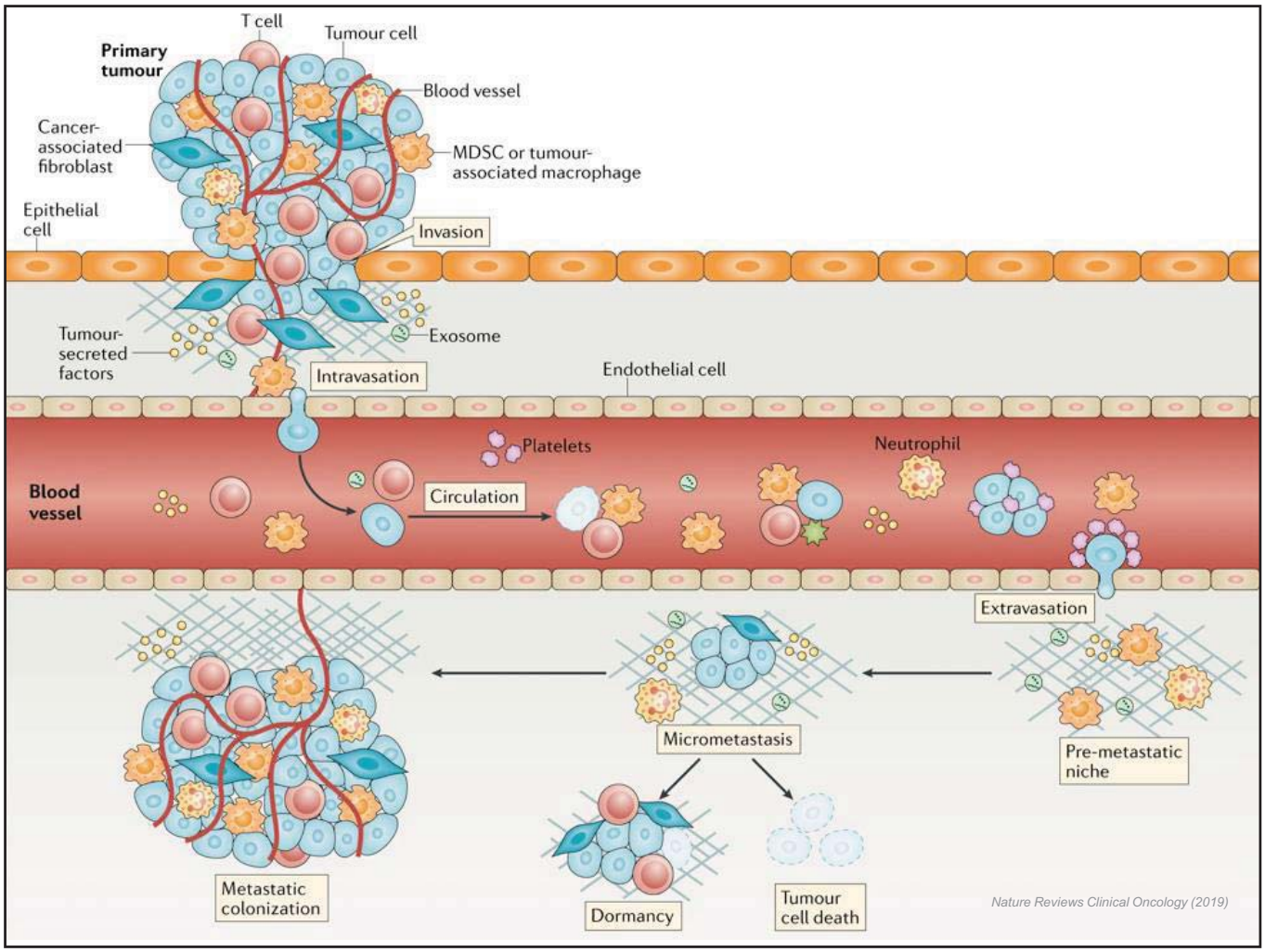
Aging-US 15(24),p.14591-14606, 2023
Molecules and Cell 45(9), 610-619, 2022
Cells, 11(13), 2079, 2022
Heliyon e13170, 2023
Nature Communications, accepted in principle

Metastasis: how cancer spreads



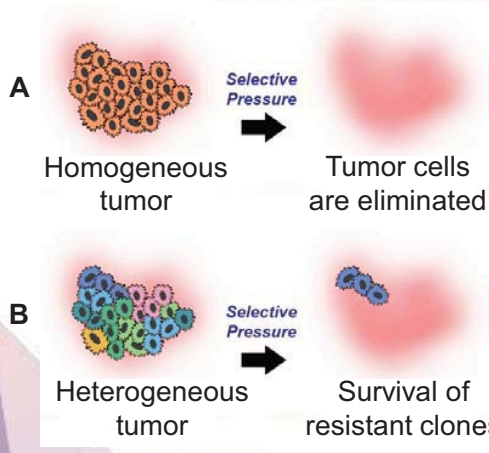
National Cancer Institute; https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis

41

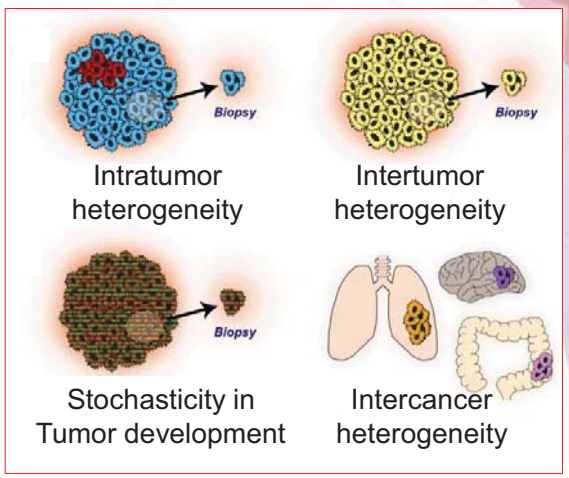


Nature Reviews Clinical Oncology (2019)

Tumor heterogeneity: a major challenge



Impact on prognostication



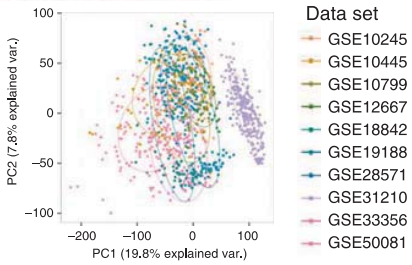
F1000 Research 2016, 5(F1000 Faculty Rev):238

43

"Big data" analytics: deriving prognostic genes

GEO Gene Expression Omnibus · NIH National Cancer Institute National Human Genome Research Institute · THE CANCER GENOME ATLAS · insilico DB · ArrayExpress · GTEx Portal · cBioPortal FOR CANCER GENOMICS

1. Integration



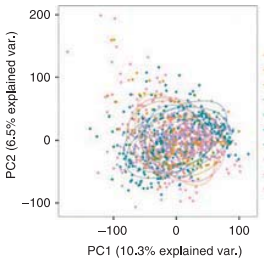
Principal Component Analysis (PCA)

↓

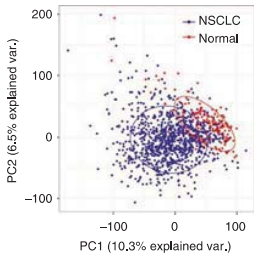
Batch-effect (Technical variation)

Merged Microarray Dataset (MMD)

2. Statistical Correction



Colored by dataset

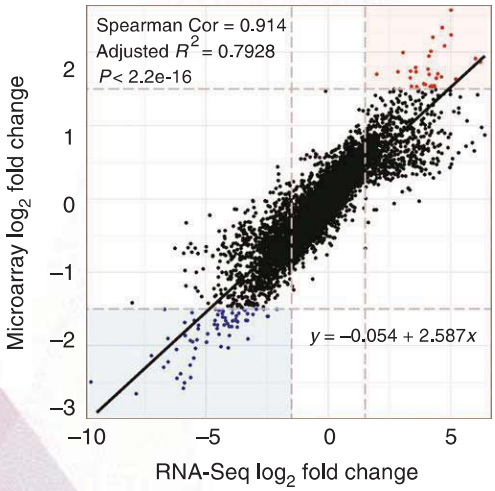


Colored by disease status

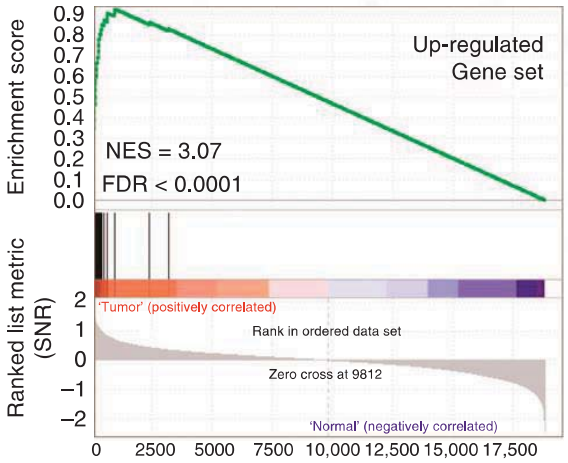
44

MMD validation

1. Comparative genome-wide expression analysis with TCGA

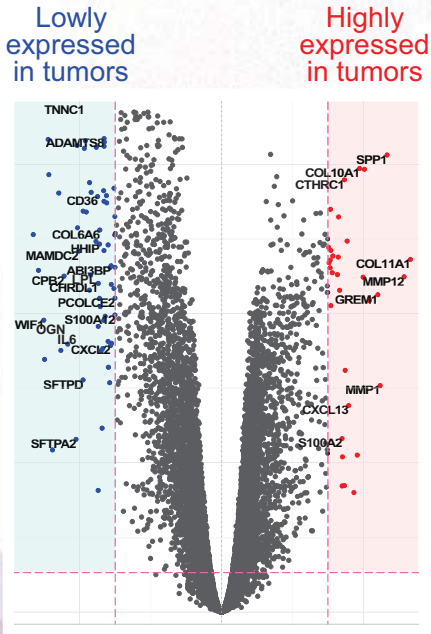


2. Gene set enrichment analysis (GSEA)

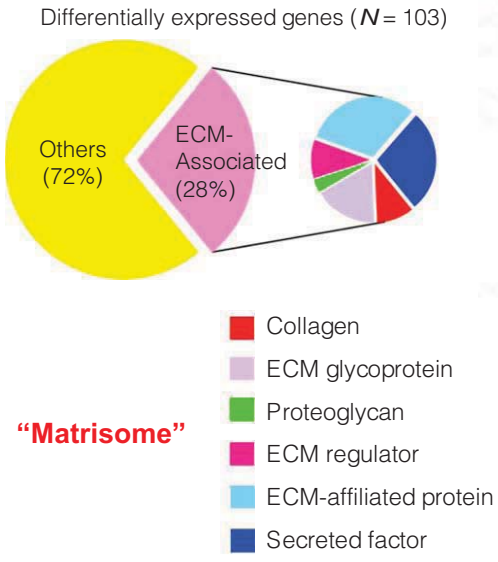


MMD application

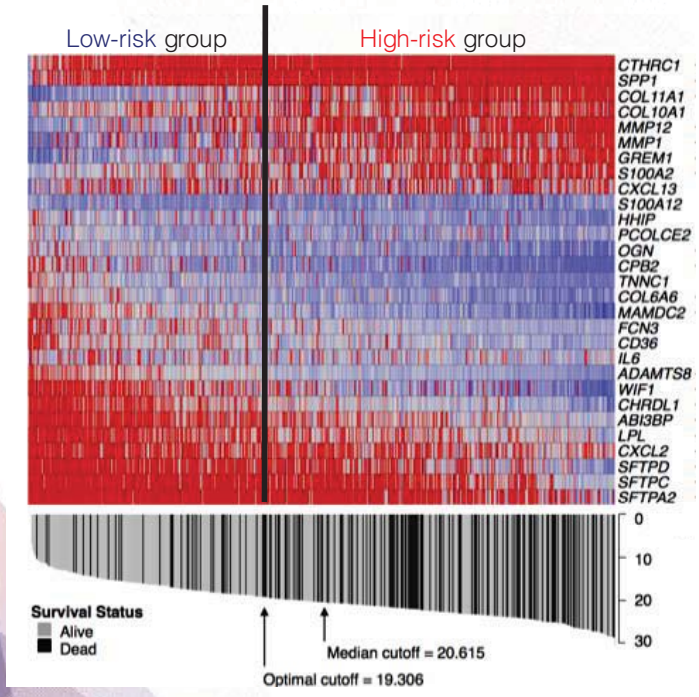
1. Differential Expression Analysis

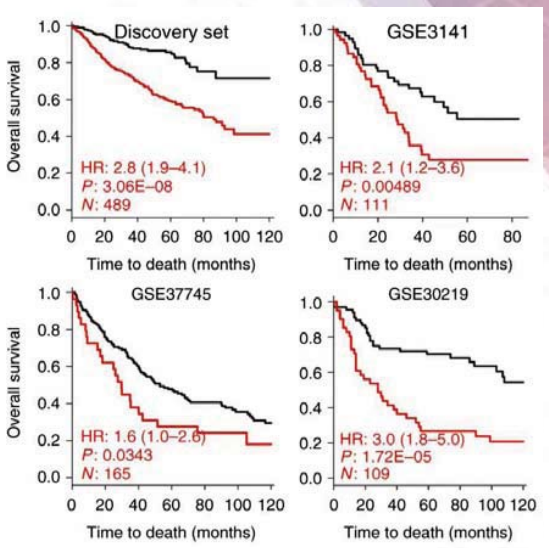


2. Gene Ontology Enrichment Analysis



A 29-gene tumor matrisome index (TMI)





1. Prognostic of overall survival (OS) and recurrence-free survival (RFS)
2. Predictive of adjuvant chemotherapy response

Intertumor heterogeneity

Intratumor heterogeneity (ITH)

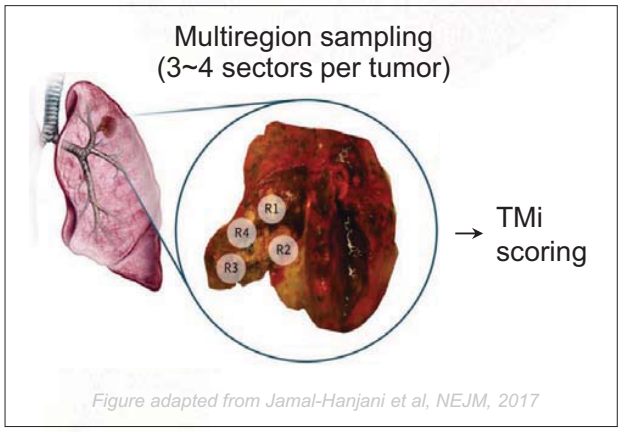
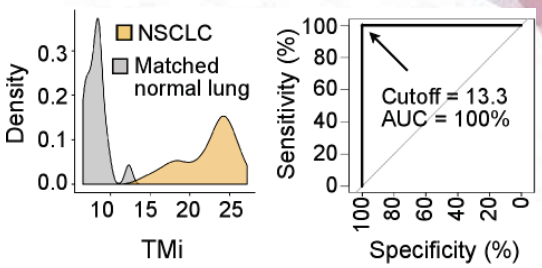
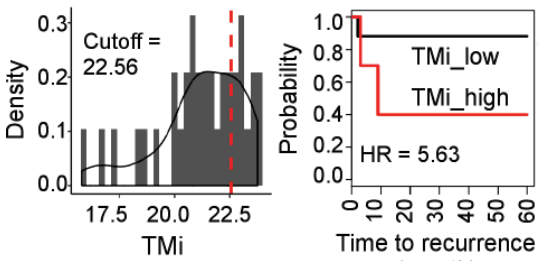


Figure adapted from Jamal-Hanjani et al, NEJM, 2017

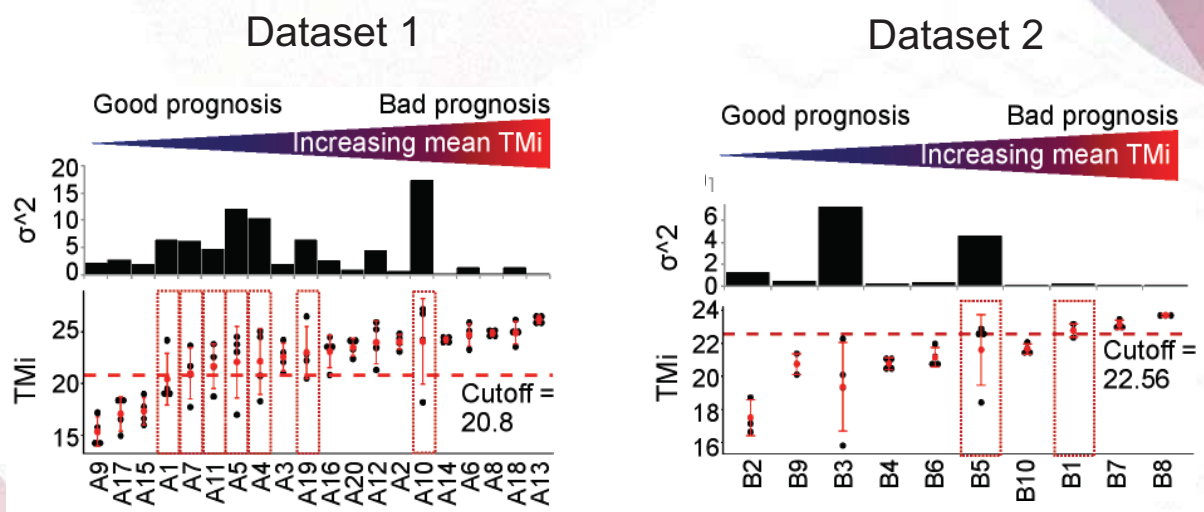
1. Diagnostic accuracy



2. Prognostic accuracy

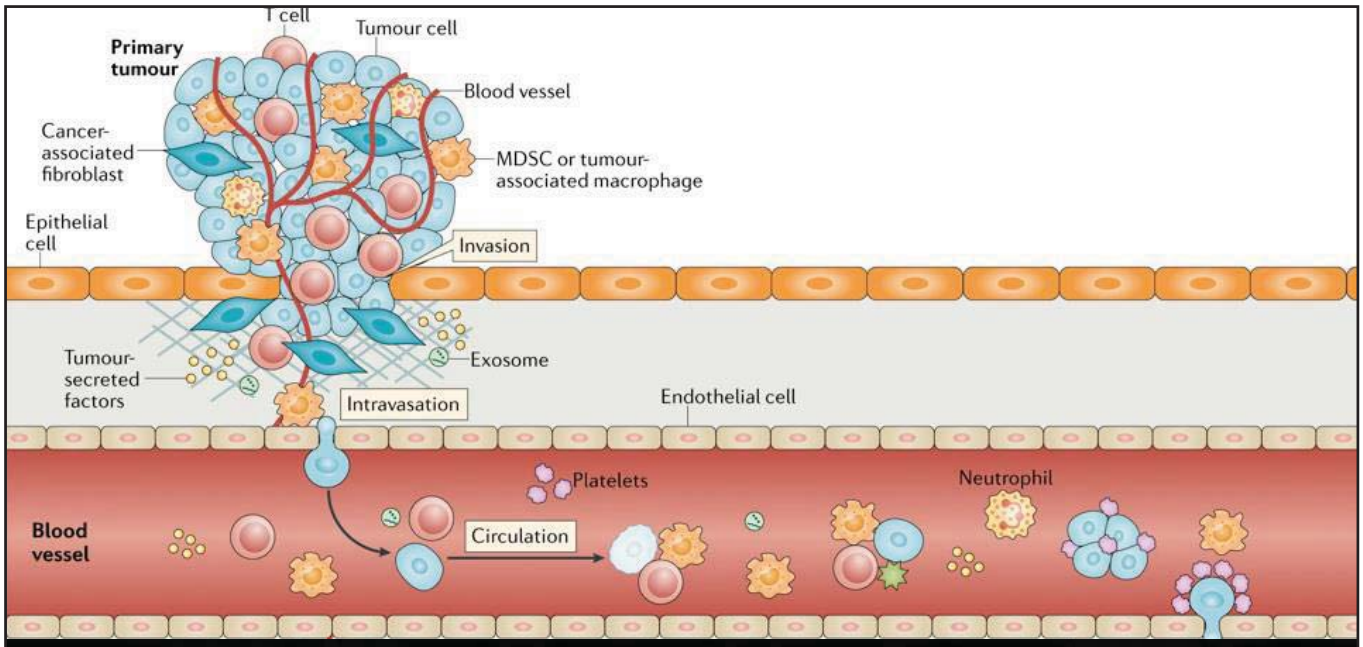


Impact of ITH on patient prognostication



A better strategy needed to refine prognostication

49



Hypothesis:
Abnormal matrisome expression patterns observed in primary tumors might be reflected at later steps of metastasis – i.e., during circulation.

50

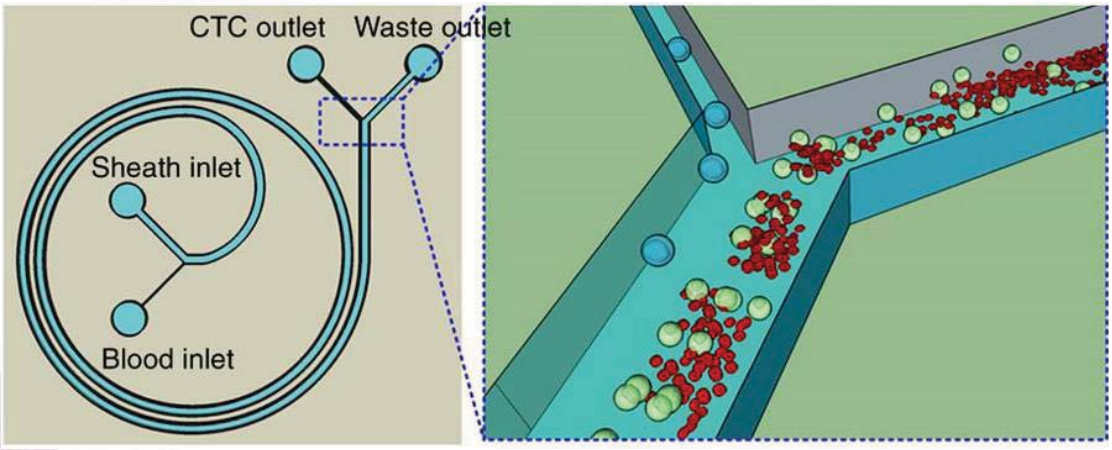
Spiral Microfluidics

"Dean flow fractionation"

(1) Inertial force

(2) Dean flow force

Both are dependent on **size**.

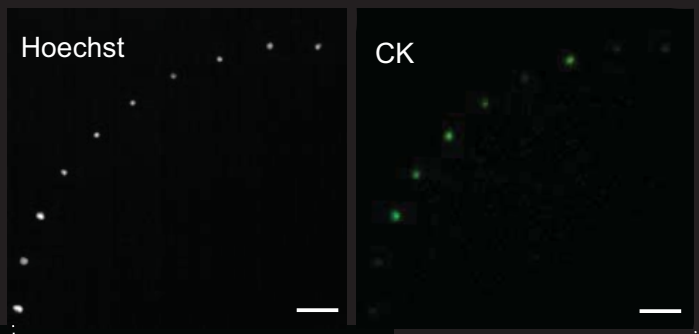


Spiral Microfluidics for isolating CTCs:

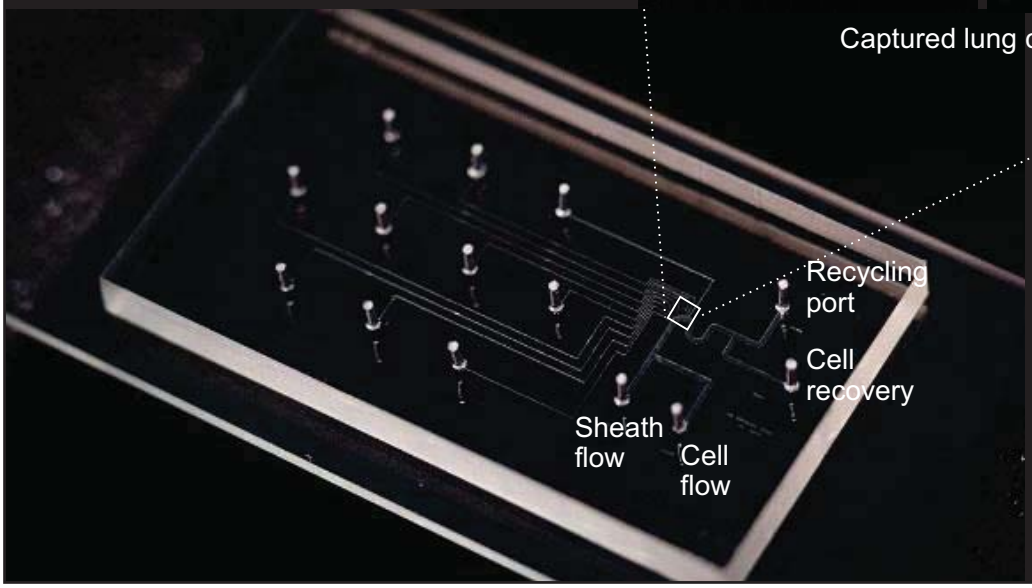
Nature Protocols, 14, 1, 128-37, 2016; Lab on a Chip, 14, 1, 128-137, 2014; Physics Today, 67, 2, 26-30, 2014; Journal of Clinical Oncology, 32, 15, 2014; Lab on a Chip, 14, 1, 128-137, 2014; Cancer Cell, 23, 3, 272-273, 2013.; Scientific Reports, 3, 1259, 2013; European Journal of Cancer, 47, S1, S48 2011; Biosensors & Bioelectronics, 26, 4, 1701-1705, 2010; Lab on a Chip, 11, 11, 1870-1878, 2011.

51

Microfluidic Single-Cell Isolation

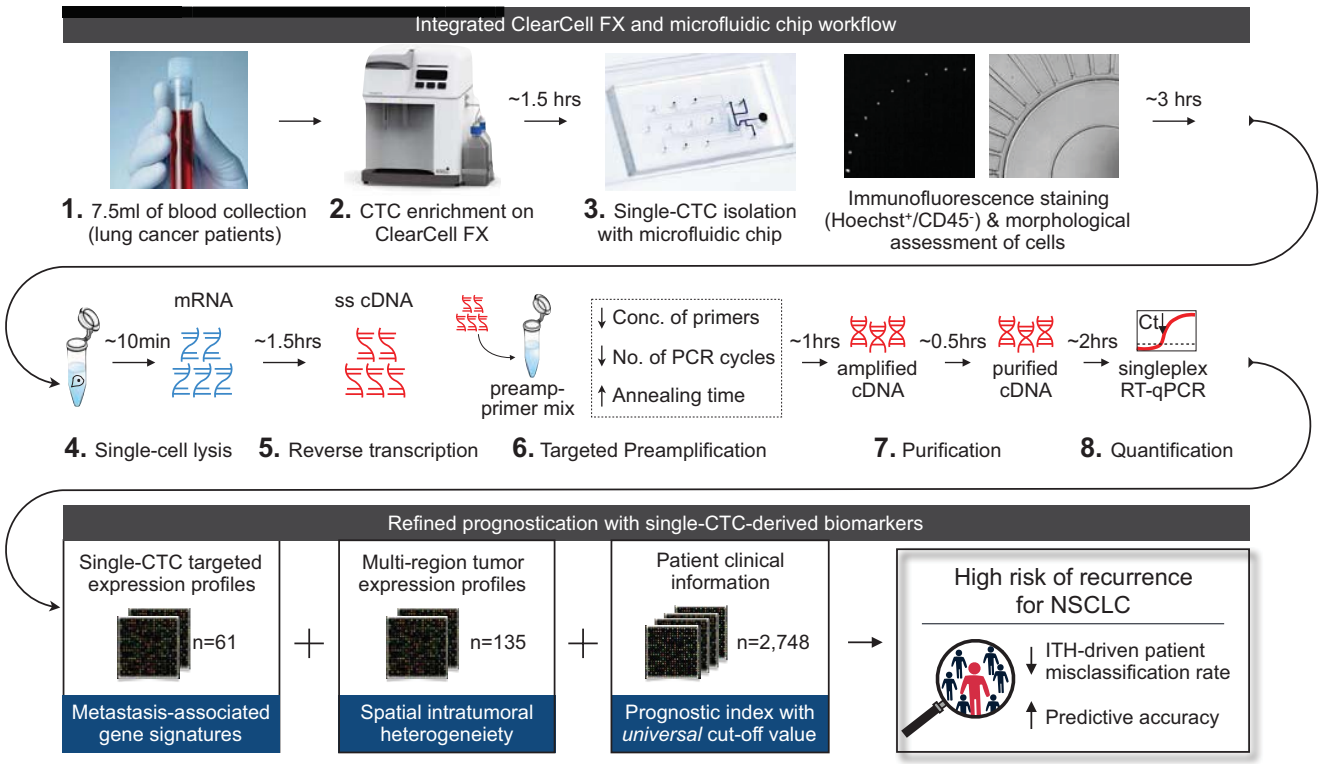


Captured lung cancer cells

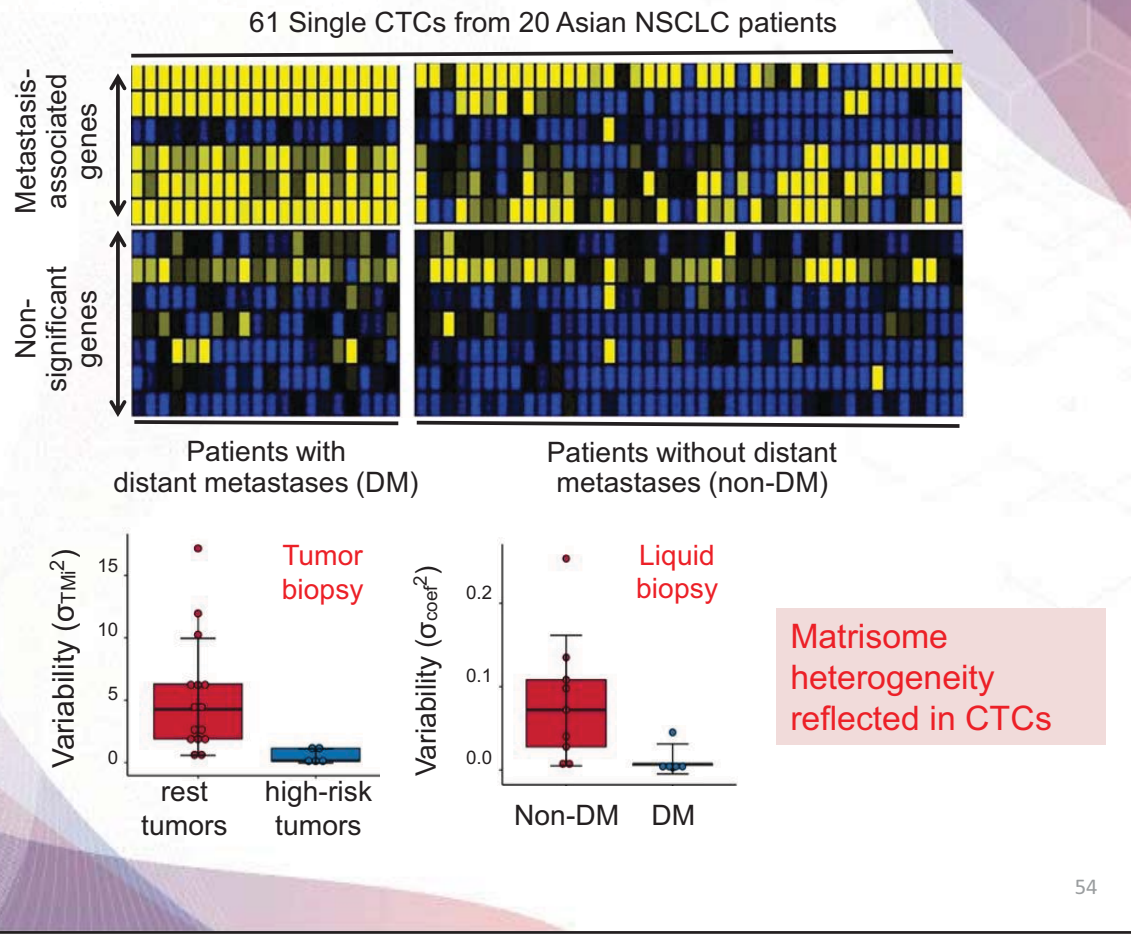


52

Single-Cell Profiling of CTCs



Metastasis-Associated Genes



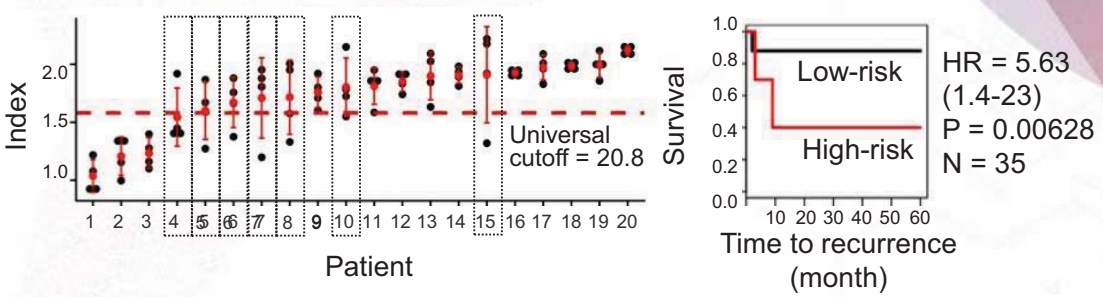
Addressing Tumor Heterogeneity to Refine Prognostic Classifier

Normal lung tissue	vs.	Lung tumor	Poor prognosis	Low intratumor heterogeneity	CTCs from metastatic disease
					
Refined features		*COL11A1* *COL10A1* *CTHRC1* *CXCL13* *GREM1* *MMP1* *MMP12* *S100A2* *SPP1*	*COL11A1* *CTHRC1* *GREM1* *MMP1* *MMP12* *S100A2* *SPP1*	*COL11A1* *COL10A1* *CTHRC1* *CXCL13* *MMP1* *MMP12*	*CXCL13* *GREM1* *MMP1* *MMP12*

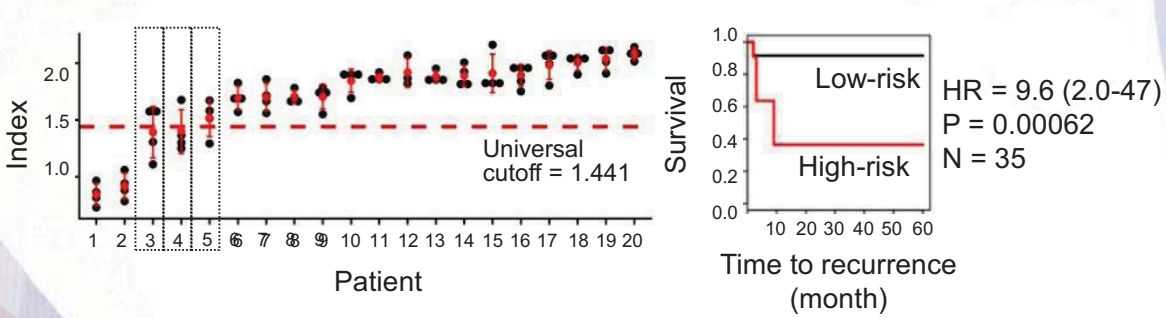
55

Improved Patient Classification

Initial classifier

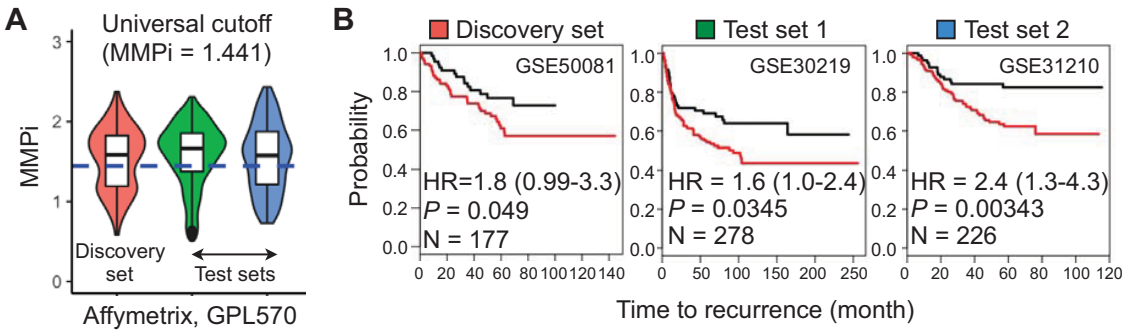


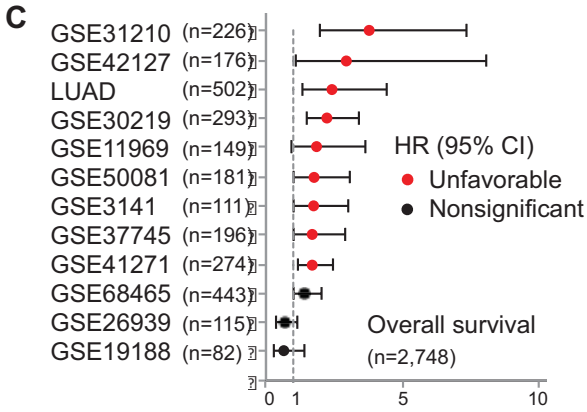
Refined classifier

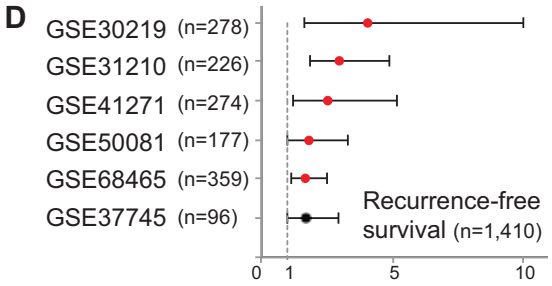


56

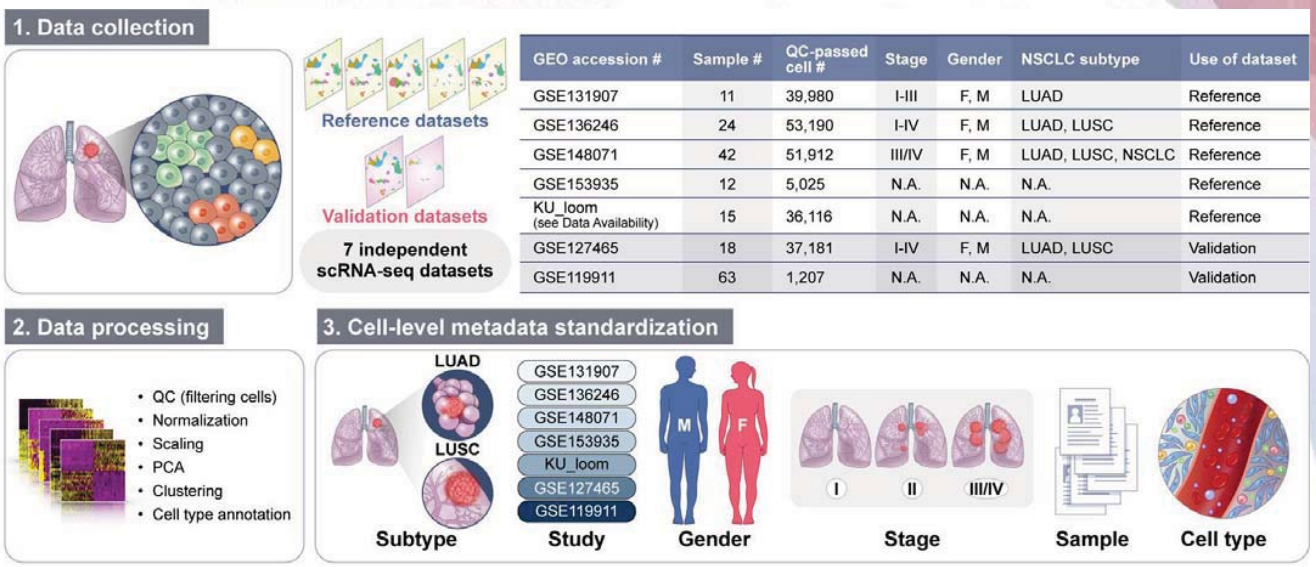
Predefined Cutoff for Patient Stratification





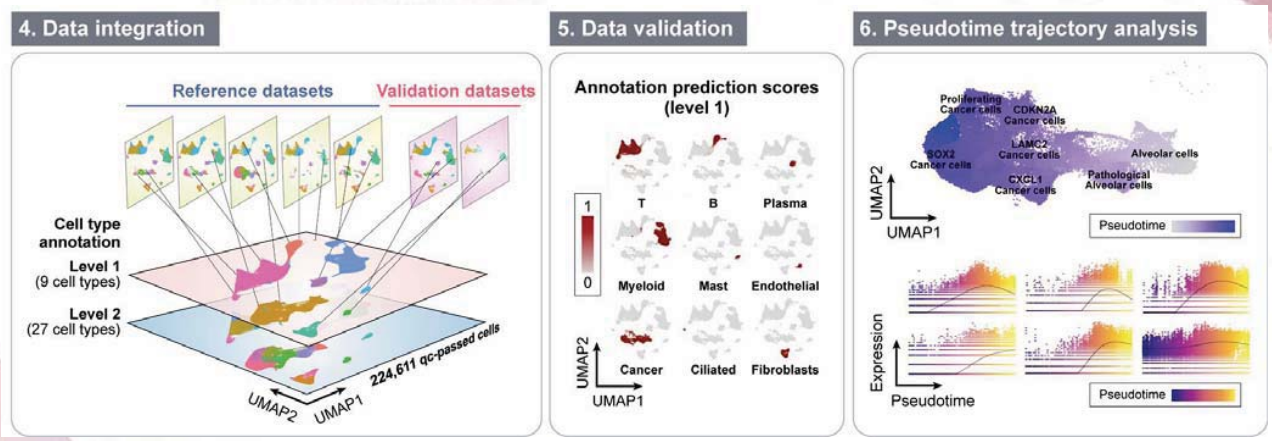


A single-cell atlas of the human lung in non-small cell lung cancer

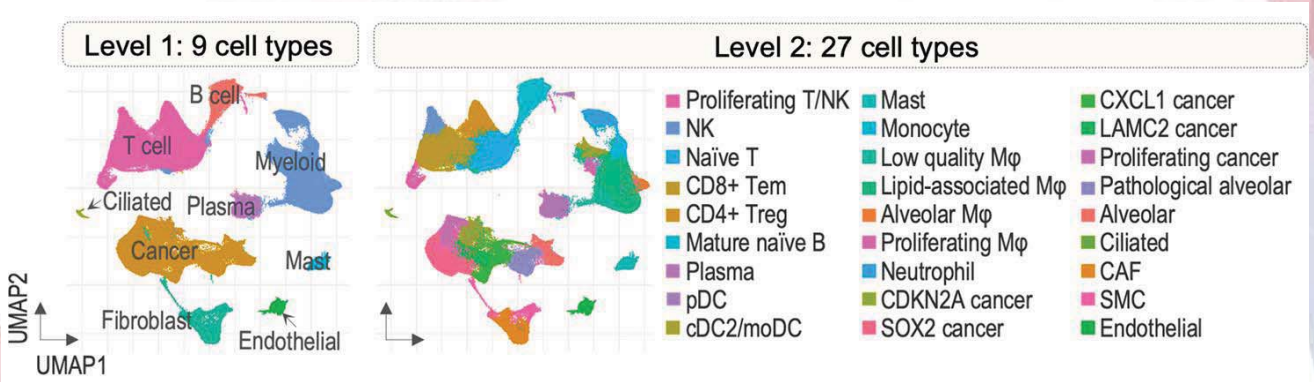


GEO accession #	Sample #	QC-passed cell #	Stage	Gender	NSCLC subtype	Use of dataset
GSE131907	11	39,980	I-III	F, M	LUAD	Reference
GSE136246	24	53,190	I-IV	F, M	LUAD, LUSC	Reference
GSE148071	42	51,912	III/IV	F, M	LUAD, LUSC, NSCLC	Reference
GSE153935	12	5,025	N.A.	N.A.	N.A.	Reference
KU_loom (see Data Availability)	15	36,116	N.A.	N.A.	N.A.	Reference
GSE127465	18	37,181	I-IV	F, M	LUAD, LUSC	Validation
GSE119911	63	1,207	N.A.	N.A.	N.A.	Validation

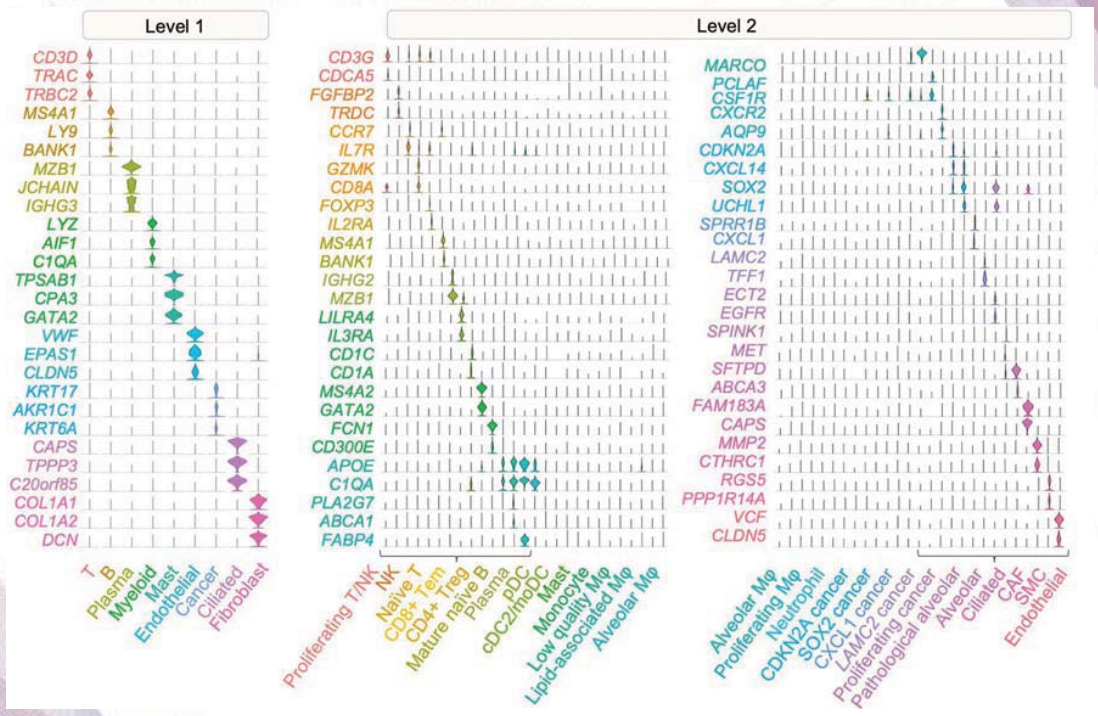
A single-cell atlas of the human lung in non-small cell lung cancer



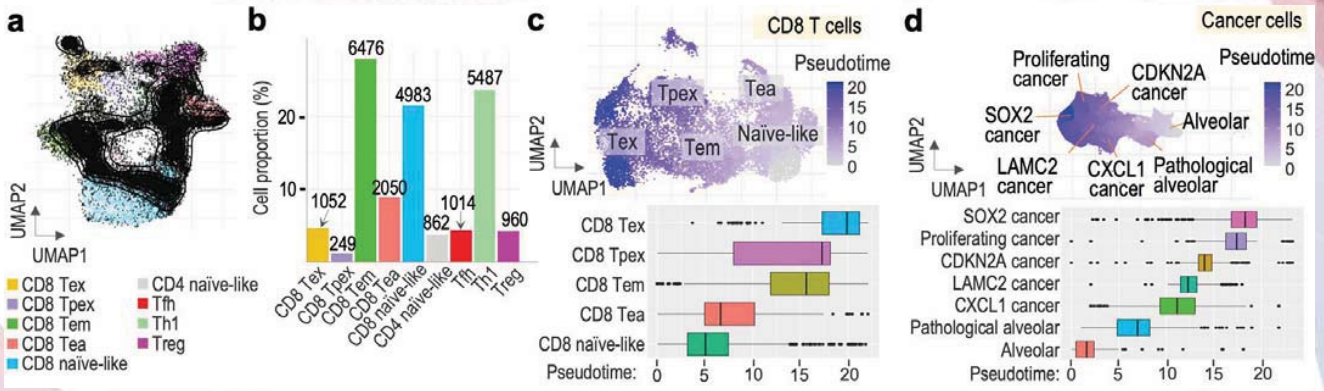
Identification of subpopulations of various cell types



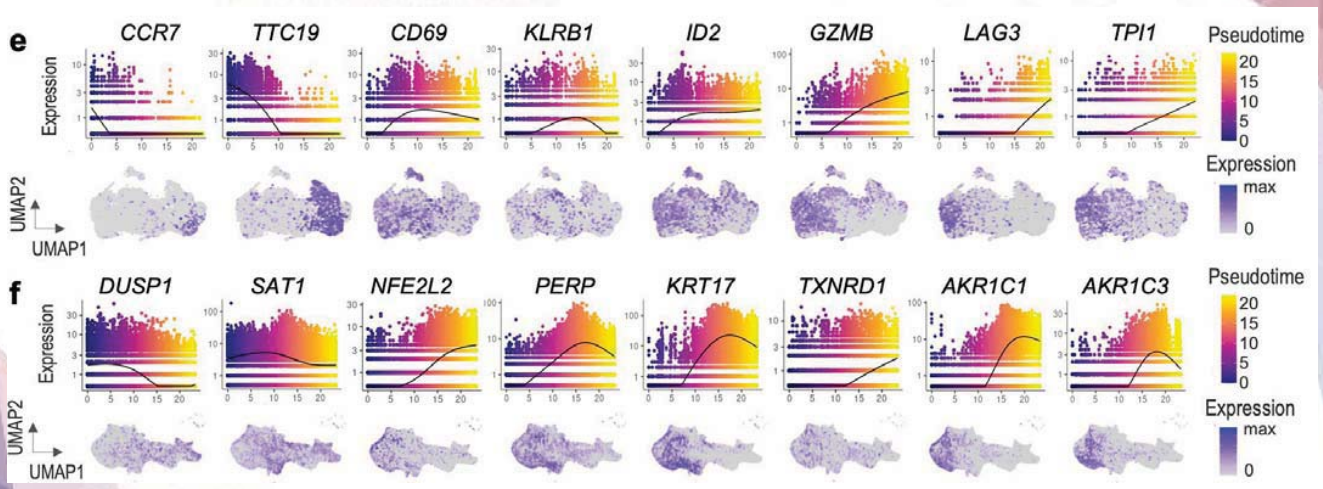
Identification of subpopulations of various cell types



Pseudotime analyses of CD8 T cells and cancer cells



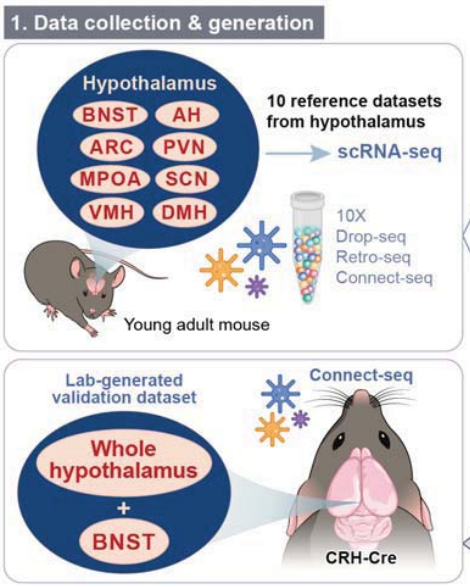
Biological insights and novel biomarker discovery



Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

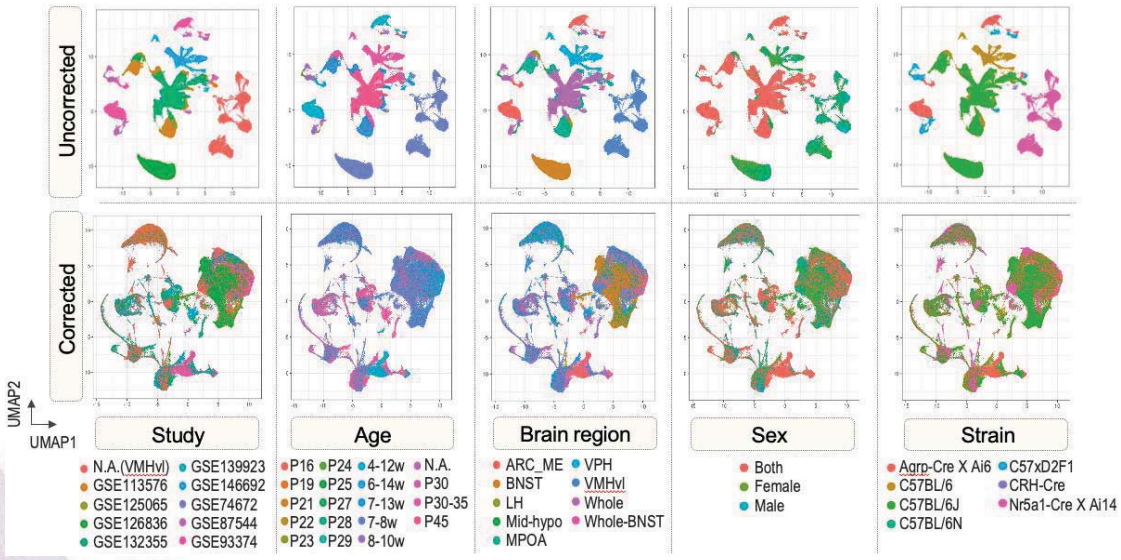
An integrated single-cell transcriptome landscape of postnatal mouse hypothalamus



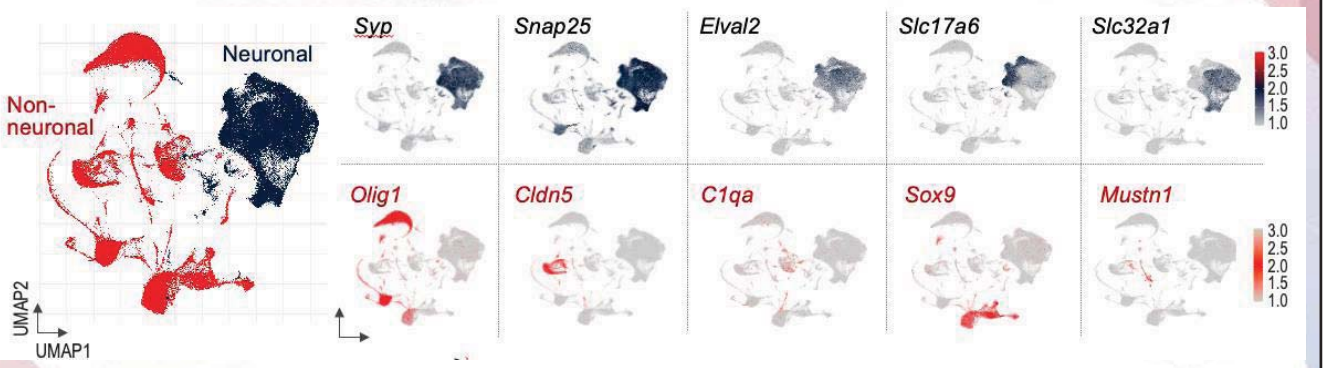
(Dataset information)

Brain region	Cell #	GEO accession #	Strain	Age	Sex	Platform
Whole hypothalamus	6,507	GSE87544	C57 X D2F1	8-10 weeks	Both	Drop-seq
Whole hypothalamus	70,248	GSE132355	C57BL/6J	P45	Both	10X
ARC-ME	19,760	GSE93374	Agrp-Cre X Ai6	4-12 weeks	Both	Drop-seq
MPOA	24,572	GSE113576	C57BL/6J	7-13 weeks	Both	10X
VMHvl	45,561	see Data Availability	Nr5a1-Cre X Ai14	7-8 weeks	Both	Retro-seq
BNST	83,524	GSE126836	C57BL/6J	7-8 weeks	Both	10X
Midline hypothalamus (ARC, VMH, DMH, AH, PVN, SCN)	1,785	GSE74672	C57BL/6N	P14-P28	Both	Drop-seq
LH	5,912	GSE125065	C57	P25-P32	Male	10X
Posterior hypothalamus	36,518	GSE146692	C57	P30-P34	Both	10X
Whole hypopthalamus + BNST	362	GSE139923	CRH-Cre	6-14 weeks	Both	Connect-seq
Whole hypopthalamus + BNST	1,533	see Data Availability	CRH-Cre	6-14 weeks	Both	Connect-seq

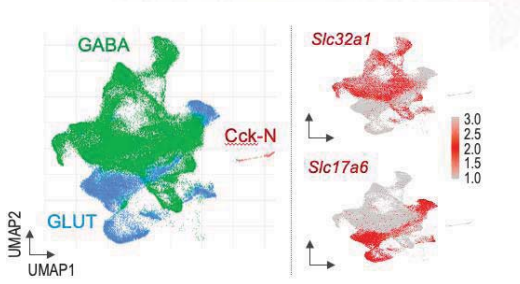
An integrated single-cell transcriptome landscape of postnatal mouse hypothalamus

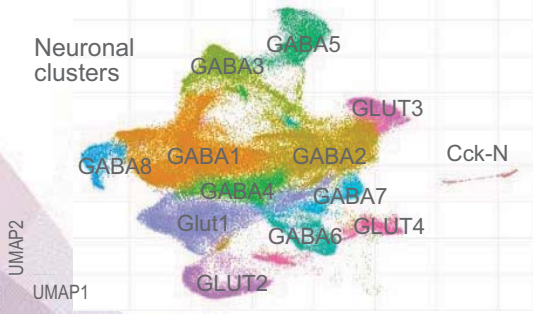


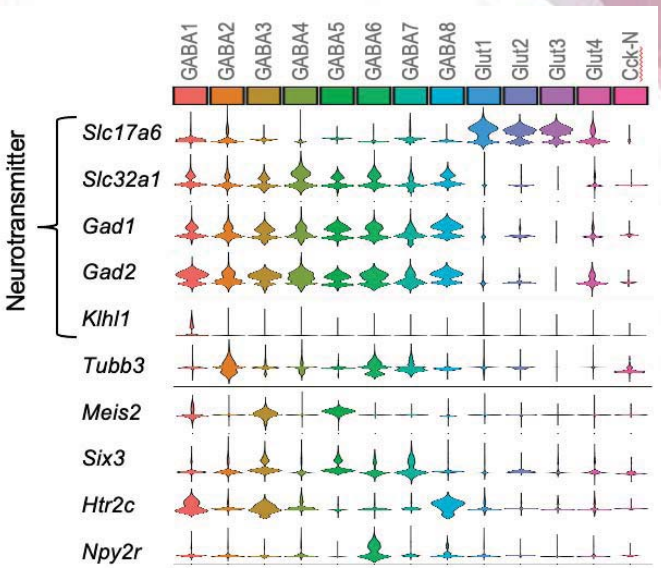
Systematic analysis of neurotransmitters in neuronal subpopulations



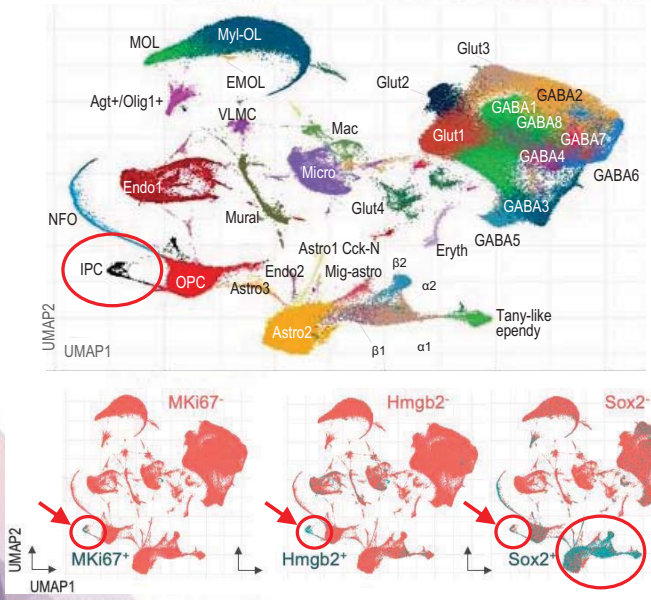
Systematic analysis of neurotransmitters in neuronal subpopulations

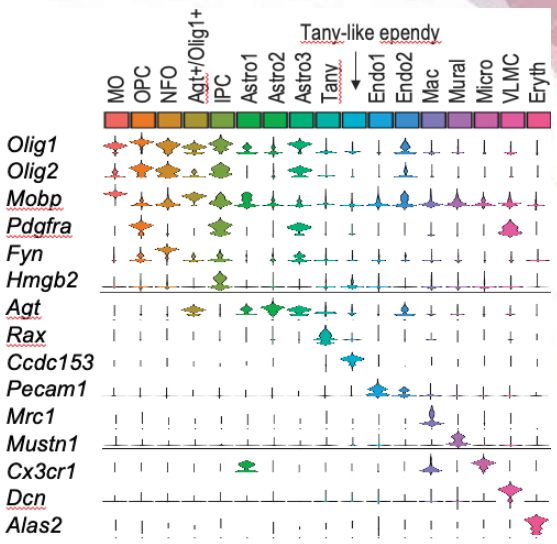




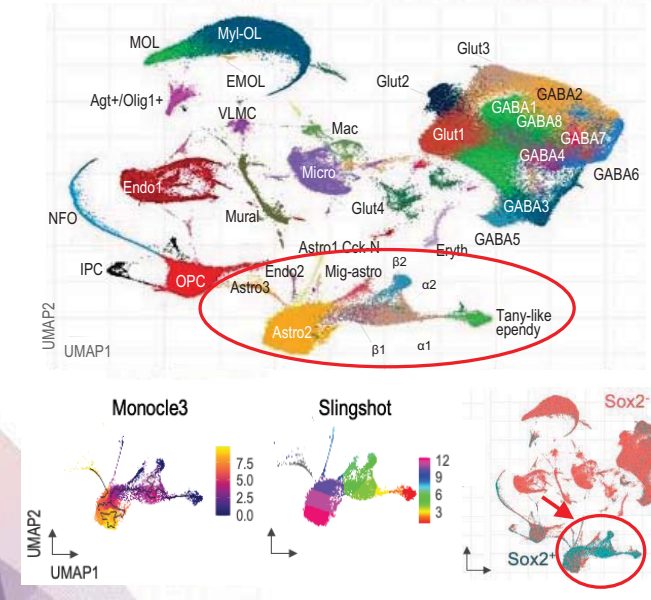


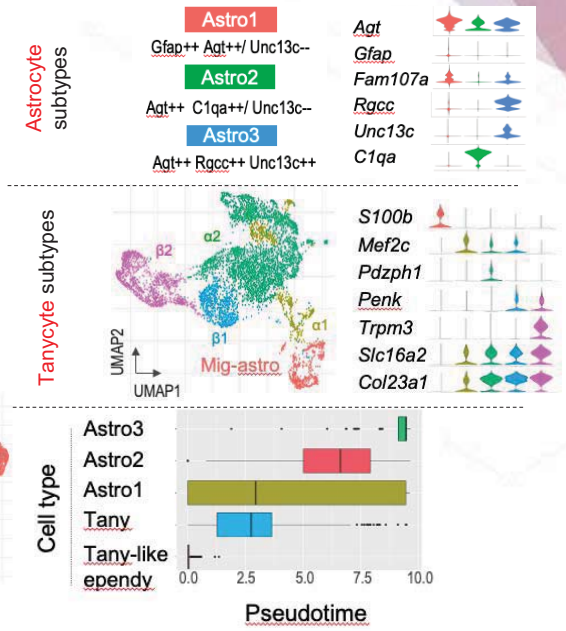
Identification and characterization of intermediate progenitor cells (IPCs)



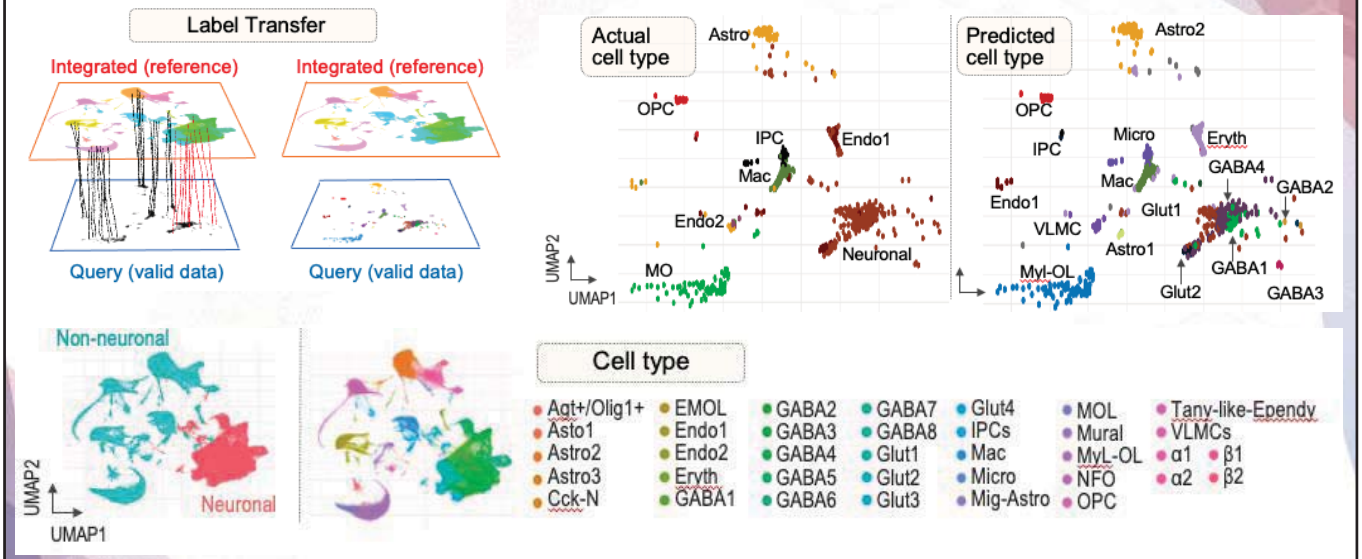


Stem cell phenotype of tanycyte-like ependymal cells giving rise to astrocytes





Validation using lab-generated Connect-seq-derived single nuclei RNA-seq data



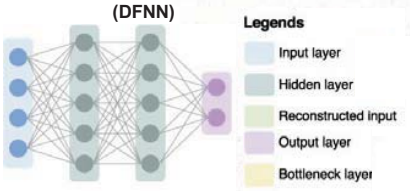
Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- **Deep learning for scRNA-seq**
- Spatial multi-omics
- Multi-omics data analysis

Deep learning for scRNA-seq data analysis

"Deep" = multilayer network structure

1. Deep Feed-Forward Neural network (DFNN)

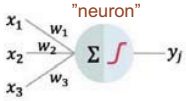


layer layer

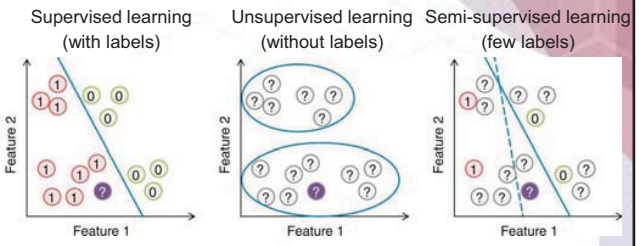
for supervised learning

A neural network

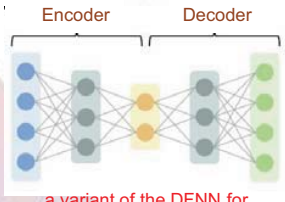
"neuron"



Non-linear activation function (e.g,. sigmoid, rectifier (ReLU), hyperbolic tangent)

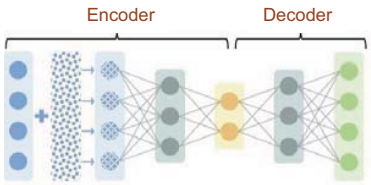


2. Deep autoencoder ("autoencoder")



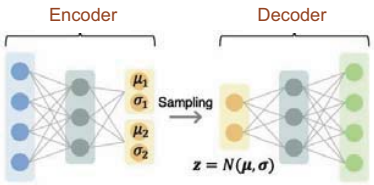
a variant of the DFNN for unsupervised learning

3. Denoising autoencoder (DAE)



a variant of autoencoder, aiming at addressing overfitting problems of autoencoder

4. Variational autoencoder (VAE)

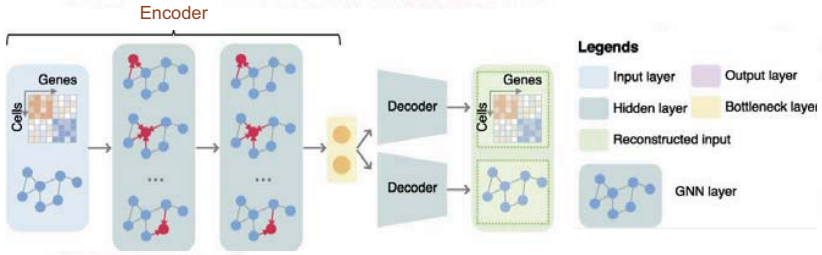


a variant of autoencoder, aiming at generating examples of the latent representations of cells

Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022; Genome Biology 14, 205, 2013

73

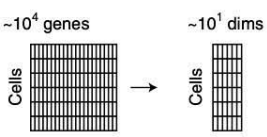
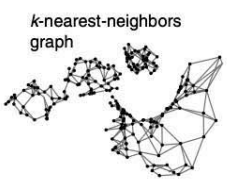
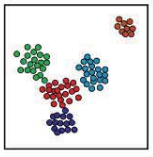
Deep learning for scRNA-seq data analysis



5. Graph autoencoder (GAE)

a variant of the DFNN for unsupervised learning

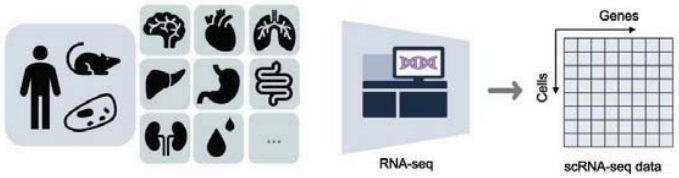
Deep learing (DFNN, autoencoder, DAE, VAE, GAE, etc)

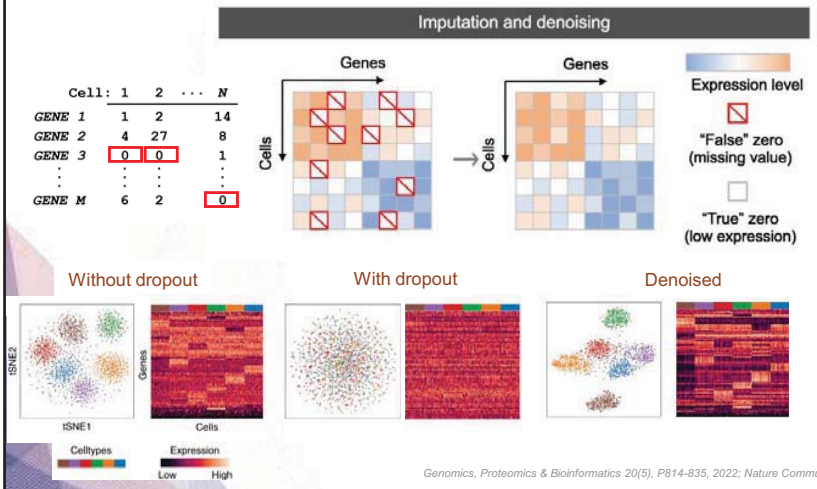
Step	Reduction to a medium-dimensional space	Manifold representation	Clustering and differential expression	Trajectories	Velocity estimation
					

Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022; Nature Methods 18(7), 723-732, 2021

74

Deep learning for (1) imputation and denoising



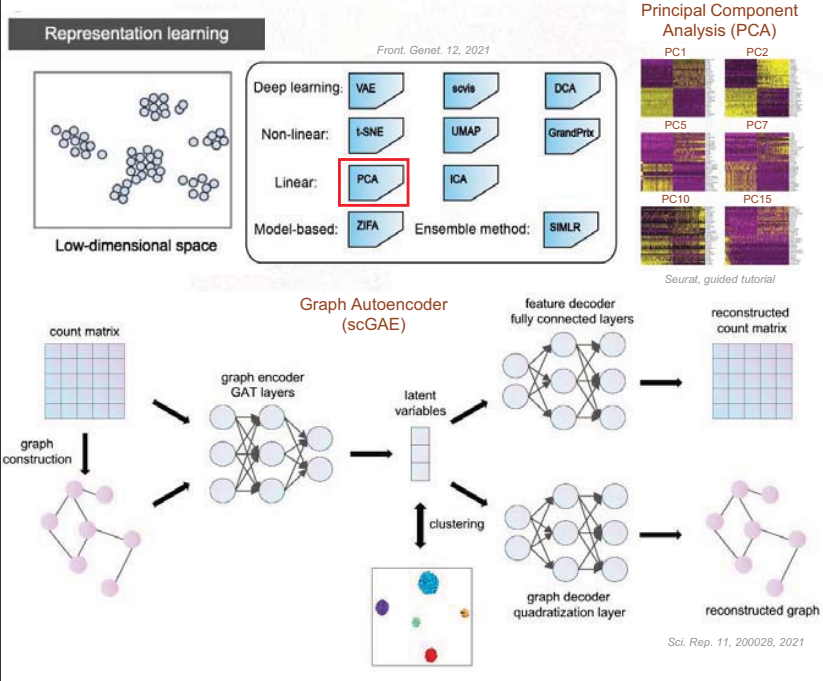


Model Name	Model Type	Code availability	Year
DeepImpute	AE	https://github.com/lanagarmire/deepimpute (Python)	2019
scIGAN	GAN	https://github.com/bm2-lab/mtSC	2020
scGMAI	AE	https://github.com/QUST-AIBBDRC/scGMAI	2021
SAVER-X	AE	https://github.com/jingshuw/SAVERX	2019
DCA	AE	https://github.com/theislab/dca	2019
ZIMBAE	AE	https://github.com/ttgump/ZINBAE	2021
ssSDAE	DAE	https://github.com/klovbe/scSDAE	2020
GraphSCI	AE/GAE	https://github.com/biomed-AI/GraphSCI	2021
SAVERCAT	VAE	-	2020
SEDIM	AE/DFN	https://github.com/li-shaochuan/SEDIM	2021
AdImpute	AE	-	2021
GNNImpute	GAE	https://github.com/Lav-i/GNNImpute	2021
scGAIN	GAN	https://github.com/mgunady/scGAIN	2019
LATE/ TRANSLATE	AE	https://github.com/audreyqyfu/LATE	2020

Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022; Nature Communications 10, 390, 2019

75

Deep learning for (2) dimensionality reduction

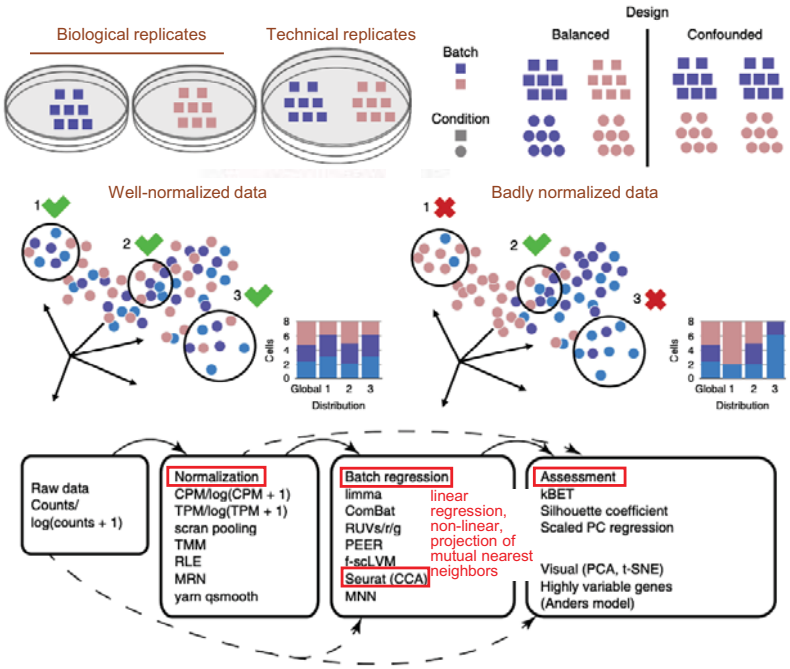


Model Name	Model Type	Code availability	Year
scScope	AE	https://github.com/AltschulerWu-Lab/scScope	2019
VASC	VAE	https://github.com/wang-research/VASC	2018
net-SNE	DFNN	https://github.com/hhcho/netsne	2018
scVI	VAE	https://github.com/YosefLab/scvi-tools	2018
scDHA	AE/VAE	https://github.com/duct317/scDHA	2021
scGSLC	GCN	https://github.com/sharpwei/GCN_sc_cluster	2021
scVAE	VAE	https://github.com/scvae/scvae	2020
scPhere	VAE	https://github.com/klarman-cell-observatory/scPhere	2021
DiffVAE/ GraphVAE	VAE	https://github.com/ioanabica/DiffVAE	2020
MMD-VAE	VAE	https://mmd-vae.hi-it.org/	2019
DR-A	AAE	https://github.com/eugenelin1/DRA	2020
scRAE	AAE	https://github.com/arnabkmondal/scRAE	2021
scRAE	VAE/β-VAE	-	2020
scGAE	GAE	https://github.com/ZixiangLuo1161/scGAE	2021
SCA	AE	https://github.com/kendomaniac/SCAtutorial	2021
GOAE	AE	-	2019
DeepAE	AE	https://github.com/sourcescodes/DeepAE	2020
pmVAE	VAE	https://github.com/ratschlab/pmvae	2021
VEGA	VAE	https://github.com/LucasESBS/vega-reproducibility	2021
Interpretable Autoencoder	AE	https://github.com/theislab/intercode	2020
LDVAE	VAE	https://github.com/YosefLab/scvi-tools	2020
SCDRHA	GAE	https://github.com/WHY-17/SCDRHA	2021
scCDG	DAE/GAE	https://github.com/WHY-17/scCDG	2021
CellVGAE	GAE	https://github.com/davidbuterez/CellVGAE	2022
graph-sc	GAE	https://github.com/ciortanmadalina/graph-sc	2021
contrastive-sc	DFNN	https://github.com/ciortanmadalina/contrastive-sc	2021
resVAE	VAE	https://github.com/lab-conrad/resVAE	2020
HD Spot	AE	-	2020
KPNN	DFNN	https://github.com/epigen/KPNN	2020
SSCA/ SSCVA	AE/VAE	-	2019
MichiGAN	VAE/GAN	https://github.com/welch-lab/MichiGAN	2021

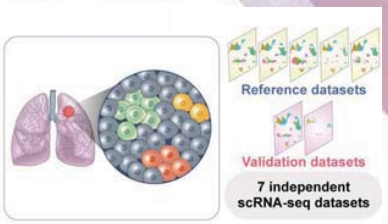
Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022

76

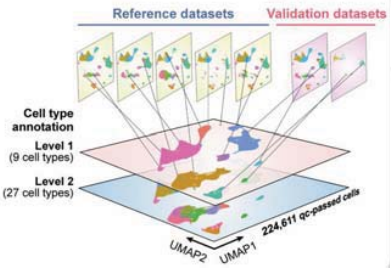
Deep learning for (3) batch effect removal



Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022



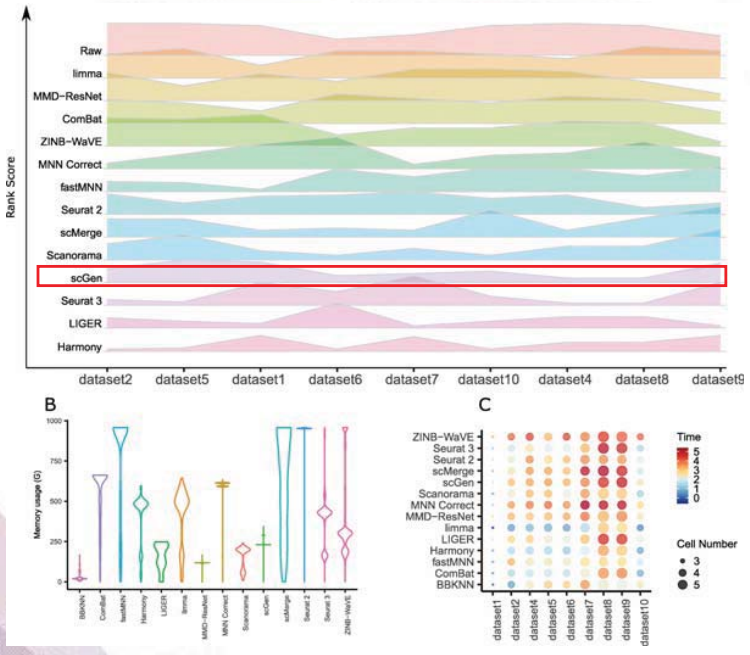
GEO accession #	Sample #	QC-passed cell #	Stage	Gender	NSCLC subtype	Use of dataset
GSE131907	11	39,980	I-III	F, M	LUAD	Reference
GSE136246	24	53,190	I-IV	F, M	LUAD, LUSC	Reference
GSE148071	42	51,912	III/IV	F, M	LUAD, LUSC, NSCLC	Reference
GSE153935	12	5,025	N.A.	N.A.	N.A.	Reference
KU_loom (see Data Availability)	15	36,116	N.A.	N.A.	N.A.	Reference
GSE127465	18	37,181	I-IV	F, M	LUAD, LUSC	Validation
GSE119911	63	1,207	N.A.	N.A.	N.A.	Validation



Nature Scientific Data 10, 167, 2023

Deep learning for (3) batch effect removal

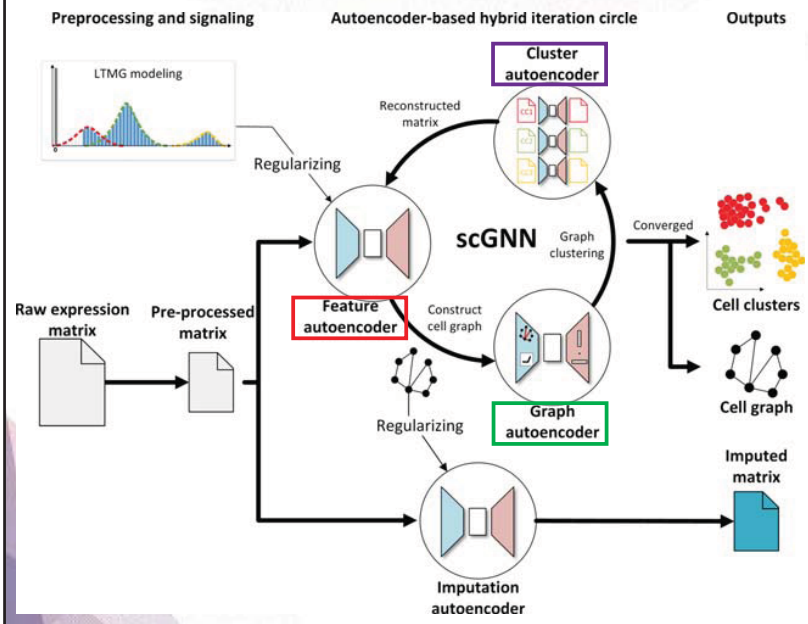
ASW, ARI, LISI, kBET

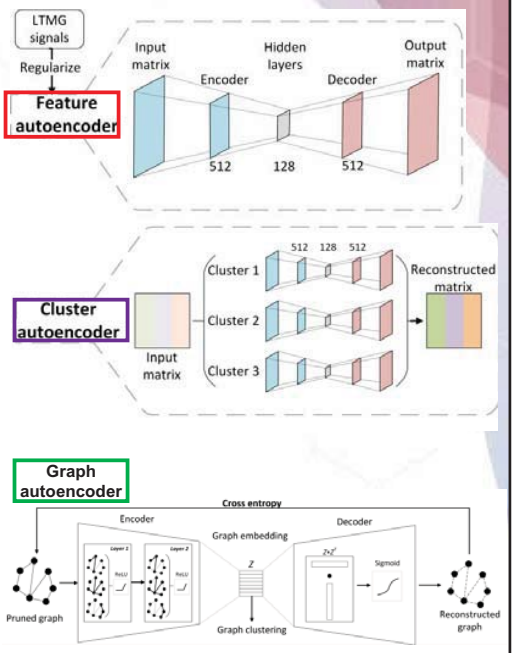


Model Name	Model Type	Code availability	Year
SMILE	DFNN	https://github.com/rpmccordlab/SMILE	2021
DAVAE	VAE	https://github.com/jhu99/davae_paper	2021
SCALEX	VAE	https://github.com/jsxlei/SCALEX	2021
AD-AE	AE	https://gitlab.cs.washington.edu/abdincer/ad-ae	2020
scGAN	VAE	https://github.com/li-lab-mcgill/singlecell-deepfeature	2021
iMAP	AE/GAN	https://github.com/Svvord/iMAP	2021
BERMUDA	AE	https://github.com/txWang/BERMUDA	2019
trVAE	VAE	https://github.com/theislab/trVAE	2020
scDGN	DFNN	https://github.com/SongweiGe/scDGN	2021
scETM	VAE	https://github.com/hui2000ji/scETM	2021
-	BERT Transformer	-	2021
deepMNN	DFNN	https://github.com/zoubin-ai/deepMNN	2020
HDMC	AE	https://github.com/zhanglabNKU/HDMC	2021
CBA	AE	https://github.com/GEOBIOyw b/CBA	2021

Genome Biology 21, 12, 2020; Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022

Deep learning for (4) cell clustering

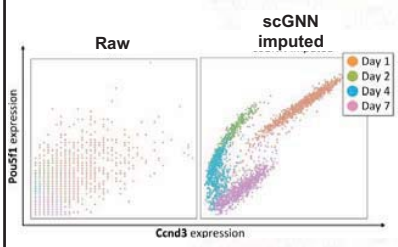




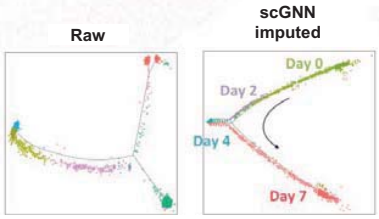
Nature Communications 12, 1882 (2021)

Deep learning for (4) cell clustering

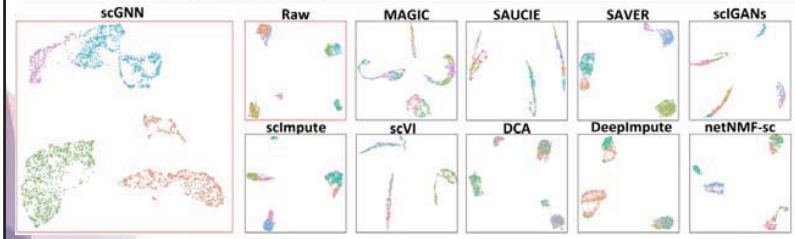
Imputation performace



Trajectory performace



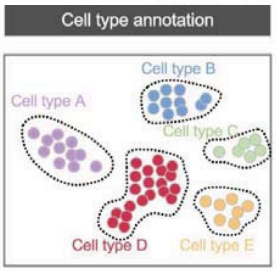
Cell clustering performace



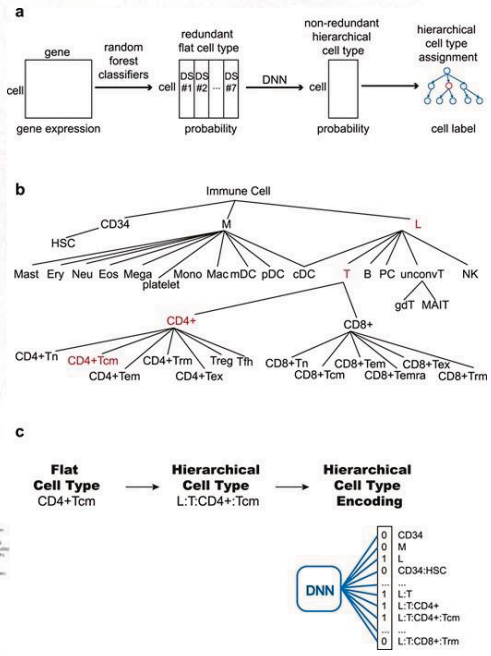
Nature Communications 12, 1882 (2021)

Model Name	Model Type	Code availability	Year
scAIDE	AE/DFNN	https://github.com/tinglabs/scAIDE	2020
scDMFK	AE	https://github.com/xuebaliang/scDMFK	2020
scCCESS	AE	https://github.com/gedcom/scCCESS	2019
DESC	AE	https://github.com/eleozzr/desc	2020
CarDEC	AE	https://github.com/jlakkis/CarDEC	2021
scziDesk	AE	https://github.com/xuebaliang/scziDesk	2020
scGNN	AE/GAE	https://github.com/juexinwang/scGNN	2021
DUSC	DAE	https://github.com/KorkinLab/DUSC	2020
GraphSCC	GCN/DAE	https://github.com/GeniusYx/GraphSCC	2021
SAUCIE	AE	https://github.com/KrishnaswamyLab/SAUCIE	2019
EMDEC	AE	-	2021
MoE-Sim-VAE	VAE	https://github.com/andkopf/MoESimVAE	2020
scvis	VAE	https://bitbucket.org/jerry00/scvis-dev	2018

Deep learning for (5) cell type annotation



101 cell subtypes

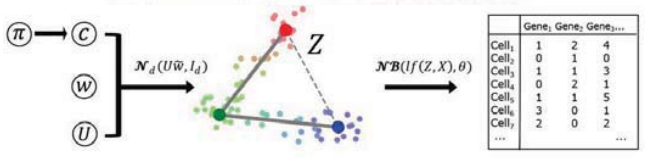


Briefings in Bioinformatics 22(5), bbab039, 2021

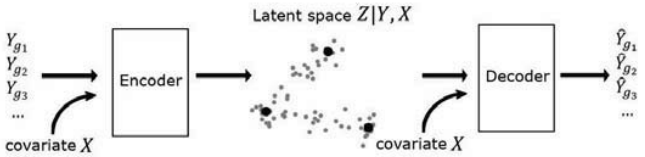
Model Name	Model Type	Code availability	Year
scAnCluster	AE	https://github.com/xuebaliang/scAnCluster	2020
JIND	DFNN	https://github.com/mohit1997/JIND	2022
ItClust	GAE	https://github.com/jianhuupenn/ItClust	2020
scDeepSort	AE	https://github.com/ZJUFanLab/scDeepSort	2021
AutoClass	VAE	https://github.com/datapplab/AutoClass	2022
scANVI	AE	https://github.com/YosefLab/scvi-tools	2021
scSemiCluster	AE	https://github.com/xuebaliang/scSemiCluster	2020
scAdapt	GAN	https://github.com/zhoux85/scAdapt	2021
scArches	VAE	https://github.com/theislab/scarches	2021
MARS	AE	https://github.com/snap-stanford/mars	2020
MAT²	AE	https://github.com/Zhang-Jinglong/MAT2	2021
scNym	DFNN	https://github.com/calico/scnym	2021
scGCN	GCN	https://github.com/QSong-github/scGCN	2021
scMRA	AE | GCN	https://github.com/ddb-qiwang/scMRA-torch	2021
MapCell	DFNN	https://github.com/lianchye/mapcell	2021
sigGCN	GAE /DFNN	https://github.com/NabaviLab/sigGCN	2021
scIAE	AE	https://github.com/JGuan-lab/scIAE	2021

Deep learning for (6) trajectory analysis

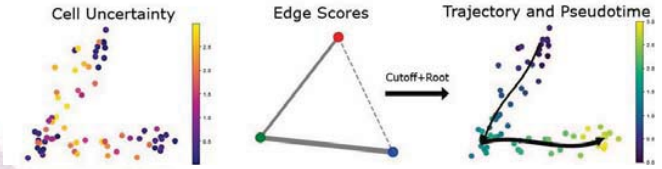
Model



Training



Inference

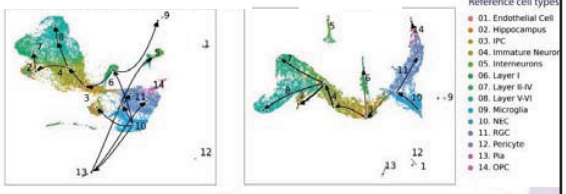


https://doi.org/10.1101/2020.12.26.424452

Developing mouse neocortex

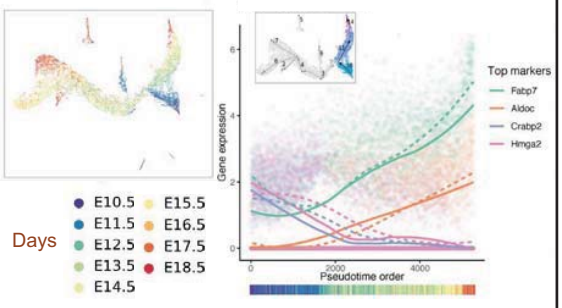
Seurat integration + Slingshot

VITAE



Collection Days

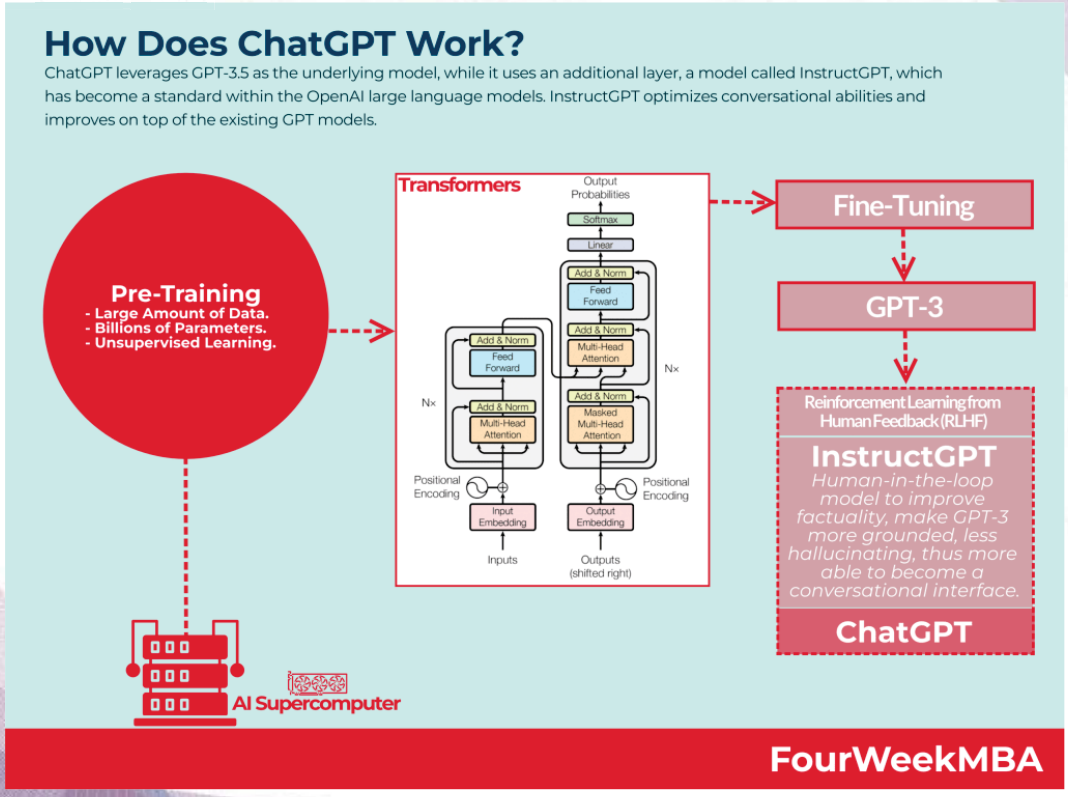
NEC -> RGC -> OPC



Model Name	Model Type	Code availability	Year
VITAE	VAE	https://github.com/jaydu1/VITAE	2020

Genomics, Proteomics & Bioinformatics 20(5), P814-835, 2022

GPT의 개요



https://fourweekmba.com/how-does-chatgpt-work/

83

Transformer: Attention Is All You Need

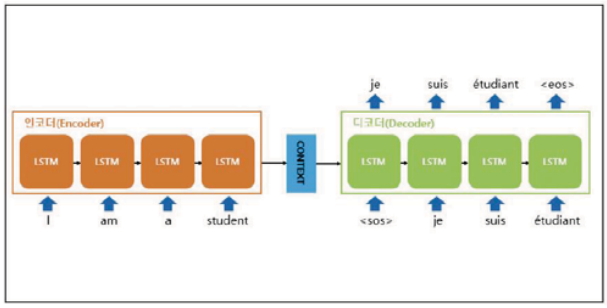
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

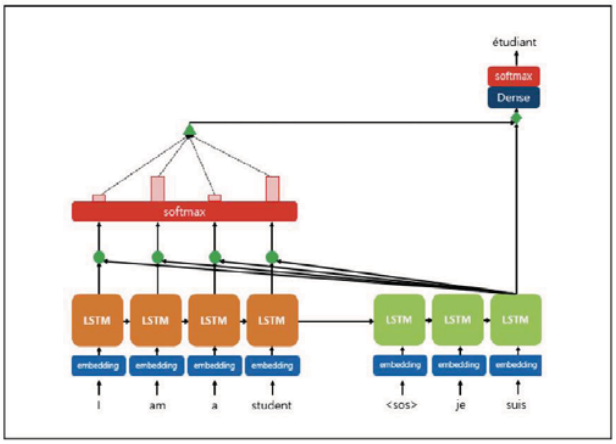
https://arxiv.org/abs/1706.03762



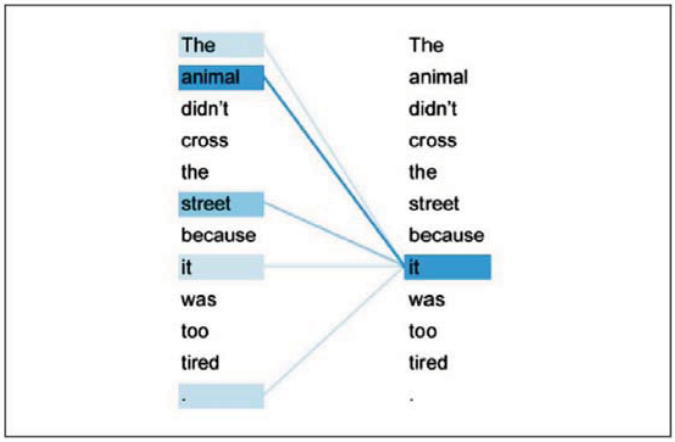
https://wikidocs.net/24996

84

Attention & Self-Attention

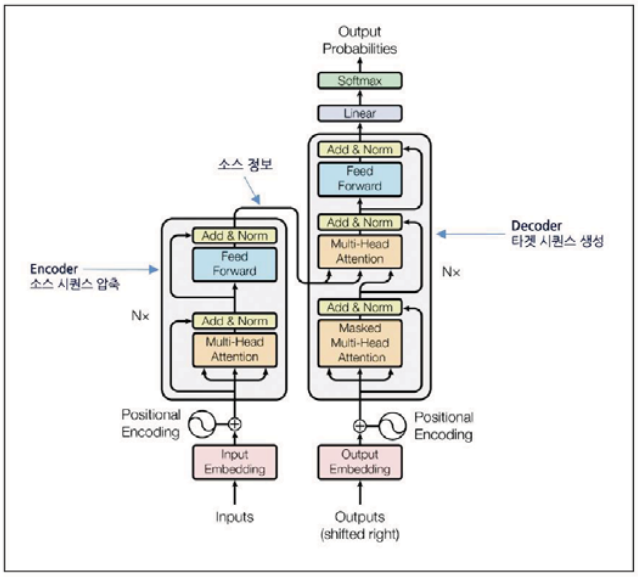


https://wikidocs.net/22893

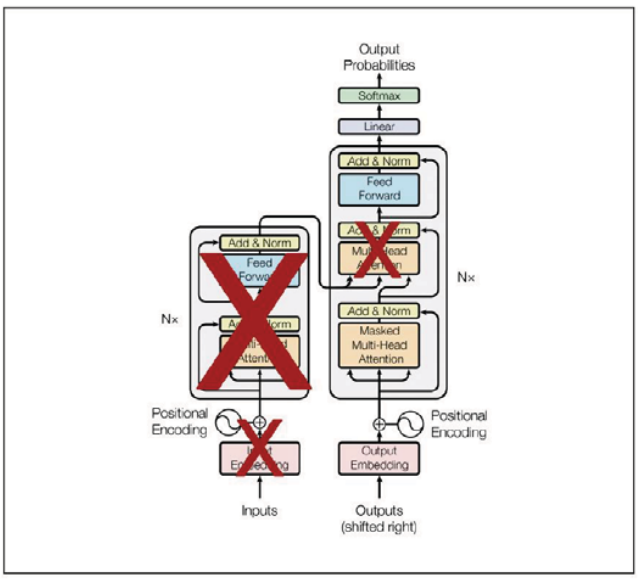


https://wikidocs.net/3137

Masted Multi-Head Attention



https://ratsgo.github.io/nlpbook/docs/language_model/bert_gpt/



https://ratsgo.github.io/nlpbook/docs/language_model/bert_gpt/

Geneformer: transfer learning for exploring network biology

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 31 May 2023

Transfer learning enables predictions in network biology

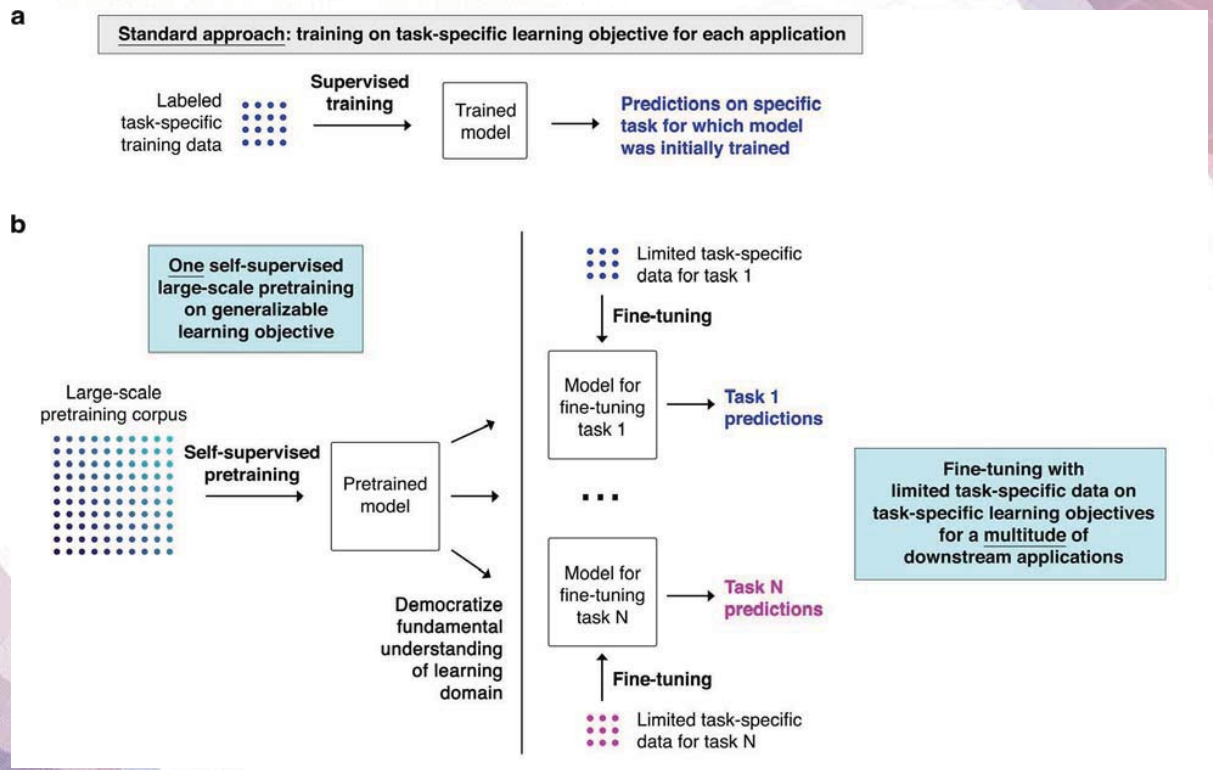
Christina V. Theodoris ✉, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu & Patrick T. Ellinor ✉

Nature **618**, 616–624 (2023) | Cite this article

78k Accesses | **17** Citations | **539** Altmetric | Metrics

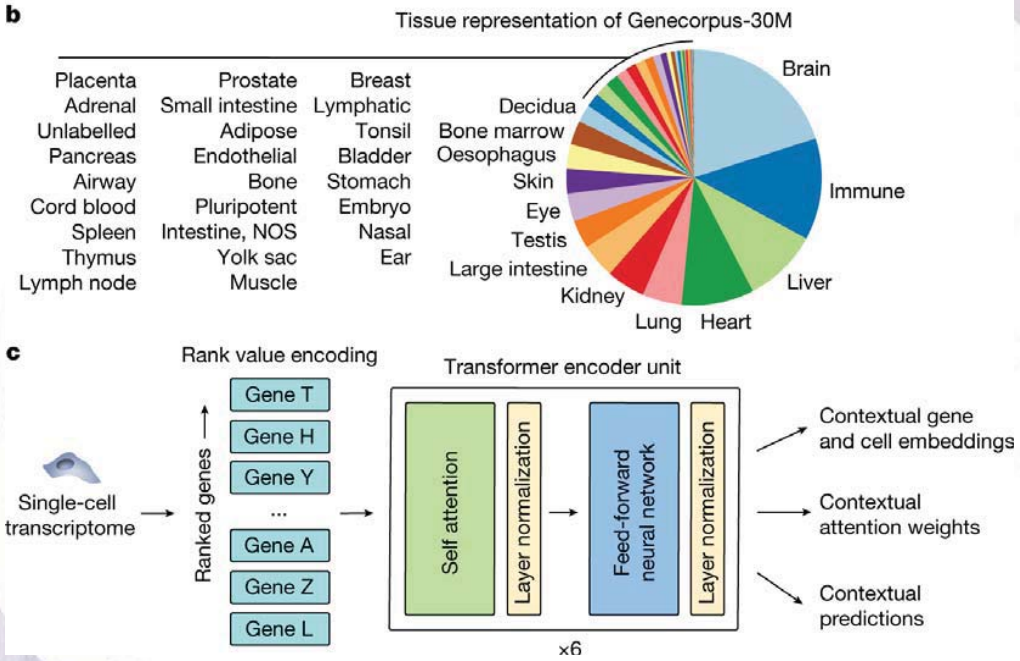
https://www.nature.com/articles/s41586-023-06139-9

Standard learning vs. transfer learning



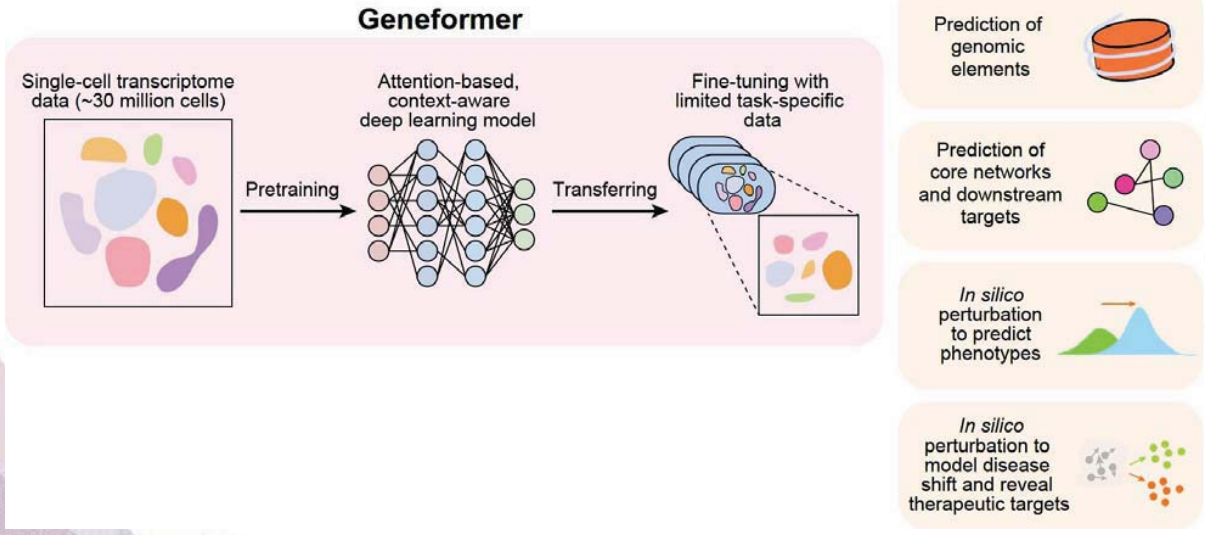
https://www.nature.com/articles/s41586-023-06139-9

Geneformer: transfer learning for exploring network biology



https://www.nature.com/articles/s41586-023-06139-9

Geneformer: transfer learning for exploring network biology



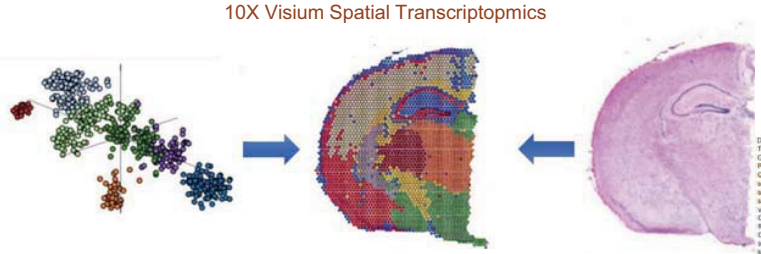
https://link.springer.com/article/10.1007/s11427-023-2431-x

Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- **Spatial multi-omics**
- Multi-omics data analysis

Timeline of single-cell and spatial multimodal methods

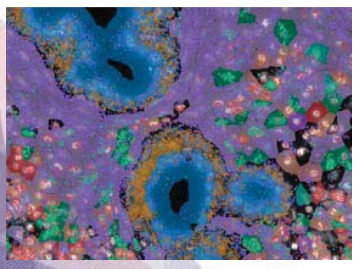
10X Visium Spatial Transcriptopmics



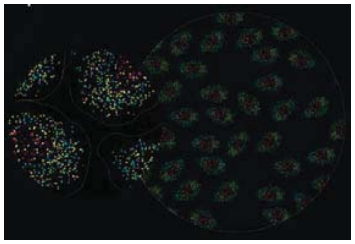
Single Cell Gene Expression | Spatially Resolved Gene Expression | Tissue Section

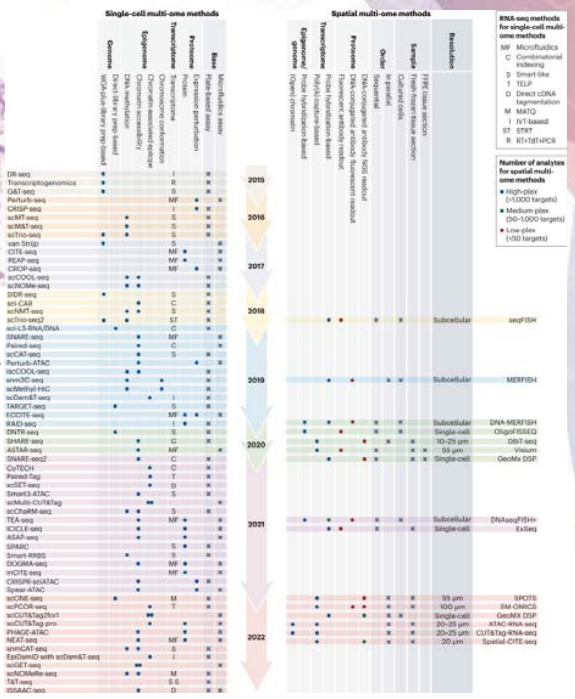
https://sites.dartmouth.edu/cqb/2020/11/24/new-services-sample-multiplexing-spatial-transcriptomics-and-multiomics/

10X In Situ Gene Expression (subcellular)



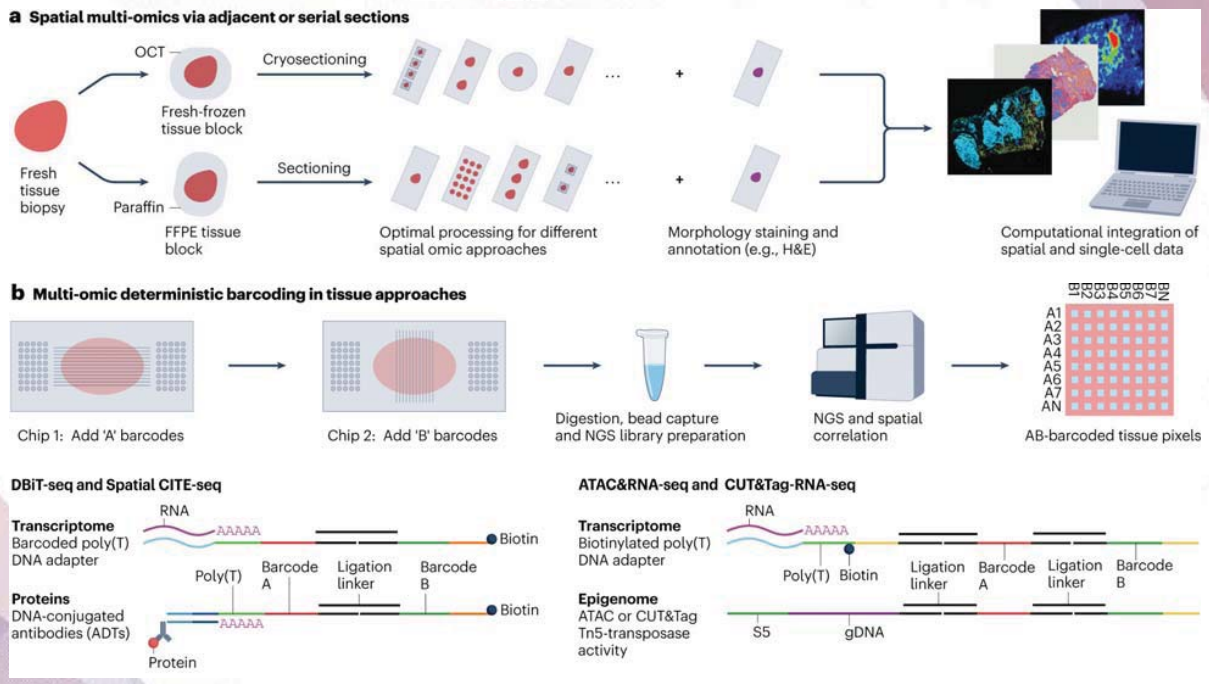
MERFISH In Situ Gene Expression (subcellular)





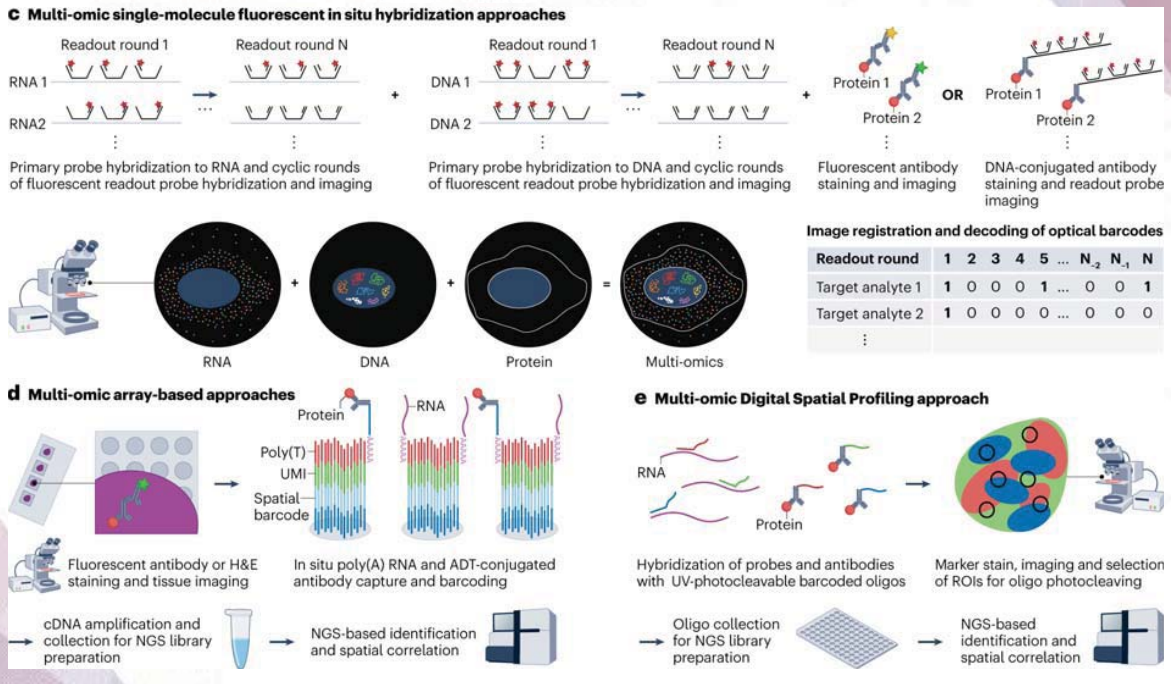
Nature Methods 17, p.11–14 (2020)

Methods for spatial multi-omics



Nature Reviews Genetics 24, 494-515 (2023)

Methods for spatial multi-omics



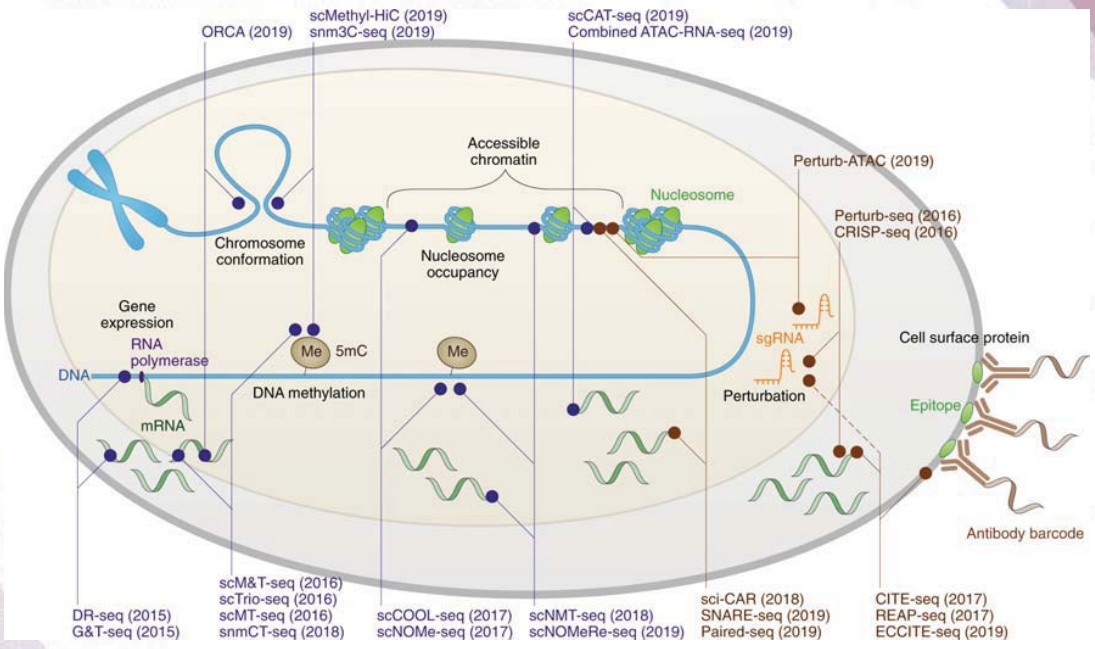
Readout round	1	2	3	4	5	...	N_{-2}	N_{-1}	N
Target analyte 1	1	0	0	0	1	...	0	0	1
Target analyte 2	1	0	0	0	0	...	0	0	0

Nature Reviews Genetics 24, 494-515 (2023)

Lecture Outline

- Bulk transcriptomics
  - Bioinformatics pipeline
  - Application in medicine
- Single-cell transcriptomics
  - Bioinformatics pipeline
- Data integration and batch effect correction
- How can we leverage "big data" for research?
  - Cancer
  - Neuroscience
- Deep learning for scRNA-seq
- Spatial multi-omics
- Multi-omics data analysis

Methods for single-cell multimodal omics analysis



Nature Methods volume 17, pages11–14 (2020)

THANK YOU

Ajou Precision Medicine Lab

아주대학교 의과대학 AJOU UNIVERSITY School of Medicine

APML members

Aejin Lee, PhD

Karolina Prazanowska

Junaid Muhammad

Jiwon Hong

Jae Hyun Shim

Yunjin Go

Jestlin Ng

Research Supports

National Research Foundation of Korea

(2020R1A6A1A03043539, 2020M3A9D8037604, 2022R1C1C1004756)

Ministry of Health & Welfare & Korea Health Industry Development Institute (HR22C1734)

아주대학교 의과대학 생화학교실 임수빈 교수 sblim@ajou.ac.kr