KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Single-cell RNA-seq data analysis for marker/drug target discovery

윤석현 _ 단국대학교

KSBI
KOREAN SOCIETY FOR BIOINFORMATICS
한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 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 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Single-cell RNA-seq data analysis for marker/drug target discovery

단일세포 RNA-seq과 공간전사체 기술은 조직 혹은 종양 미세환경(TME)에서 세포들이 어떤 상태에 있고, 주변 세포들과 어떤 상호작용을 하는지, 정상조직 혹은 기준이 되는 조직과 어떤 차이가 있는지 등을 정밀하게 확인할 수 있는 기술로 이의 데이터의 분석을 통해 진단 혹은 예후과 관련된 마커나 치료 표적 후보들을 발굴하고 이를 기반으로 검증을 위한 실험을 기획함으로써 많은 시간과 비용을 절약할 수 있다. 물론 이를 위해서는 세포유형 식별, 세포간 상호작용 추론, DEG 및 gene set enrichment 분석 등 다양한 분석도구를 이용한 데이터의 처리가 요구된다.

본 프로그램은 생명과학/의학분야 연구자 혹은 석사 수준의 생명정보 연구자가 손쉽게 종양/조직 미세환경에 대한 분석 결과를 얻고 해당 분석 결과의 데이터 마이닝을 직접 수행하여 마커나 치료 표적 후부를 발굴하고 이들의 생물학적 의미를 추론해낼 수 있도록 해보자는 데에 목표를 두고있다. 본 프로그램을 통해 참여자들은 최소한의 (파이썬) 프로그래밍 스킬로 마커/표적 발굴 연구를 위한 다양한 분석 도구로부터 얻어진 결과를 한눈에 확인하고 이 결과들의 데이터 마이닝/활용법 실습을 통해 시스템적 관점에서 발굴된 후보들에 연관된 생명 과정의 통찰을 얻고 검증을 위한 실험을 기획할 수 있는 방법을 체험해 볼 수 있다.

강의는 다음의 내용을 포함한다:

- SCODA 파이프라인을 이용한 단일세포 RNA-seq 데이터의 처리 (세포유형 식별, Copy number variation 및 ploidy 추정, 세포유형간 단백질 상호작용 분석, 조건 간 DEG 분석 및 Gene set enrichment 분석 도구 소개 포함)
- AnnData 포맷 소개 및 SCANPY를 이용한 단일세포 RNA-seq 데이터의 전처리 (실습)
- DEG 분석 결과/세포간 상호작용 결과의 데이터 마이닝 및 마커/표적 발굴 (실습)
- CNV (추정치) 기반의 암세포 표현형 특징 분석 (실습)
- GSEA를 이용한 신호 경로 분석 (실습)

* 참고 강의교재: N/A

* 교육생준비물: 노트북 (무선랜 접속가능, Google Chrome 설치 필요)

* 강의 난이도: 중급 (단일세포 RNA-seq 기술에 대한 이해 및 기초적인 프로그래밍)

* 강의: 윤석현 교수 (단국대학교 전자전기공학과)

Curriculum Vitae

Speaker Name: Seokhyun Yoon, Ph.D.

▶ Personal Info

Name	Seokhyun Yoon
Title	Professor
Affiliation	Dankook University

▶ Contact Information

Address	Rm #310, 2nd engineering building, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, Republic of Korea, 16890
Email	syoon@dku.edu

Research Interest

Bioinformatics, Machine learning and computational transcriptomics

Educational Experience

2003	Ph.D. in Electrical and Computer Engineering, New Jersey Institute of Technology, USA
1996	M.S. in Electronics, Sungkyunkwan University, USA
1992	B.S. in Electronics, Sungkyunkwan University, Korea

Professional Experience

2005-	Professor, Dankook University, Korea
2003-2005	Senior Member of Technical Staff, Samsung Electronics, Korea
1999-1999	Member of Technical Staff, Electronics & Telecom Research Institute (ETRI), Korea

Selected Publications (5 maximum)

1. D. Hong, H. Kim, W. Yang, C. Yoon, M. Kim, CS Yang and S. Yoon, "Integrative analysis of single-cell RNA-seq and gut microbiome metabarcoding data elucidates macrophage dysfunction in mice with DSS-induced ulcerative colitis," Communications Biology, June 2024. https://doi.org/10.1038/s42003-024-06409-w
2. J. Lee, M. Kim, K. Kang, CS Yang and S. Yoon, "Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing," Briefings in Bioinformatics, Jan. 2023. https://doi.org/10.1093/bib/bbad006
3. M. Kim, W. Yang, D. Hong, HS Won, S. Yoon, "A Retrospective View of the Triple-Negative Breast Cancer Microenvironment: Novel Markers, Interactions, and Mechanisms of Tumor-Associated Components Using Public Single-Cell RNA-Seq Datasets," Cancers, Mar. 2024. https://doi.org/10.3390/cancers16061173
4. JS Kim, HK Kim, M. Kim, S. Jang, E. Cho, S. Mun, J. Lee, D. Hong, S. Yoon and CS Yang,, "Colon-Targeted eNAMPT-Specific Peptide Systems for Treatment of DSS-Induced Acute and Chronic Colitis in Mouse," Antioxidants, Nov. 2022. https://doi.org/10.3390/antiox11122376

KSBI | 한국생명정보학회

KSBi-BIML 2025

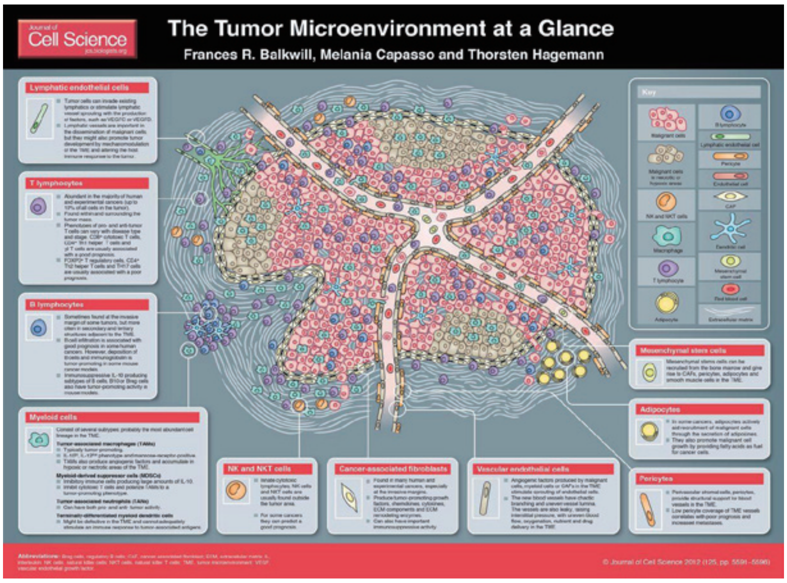
Single-cell RNA-seq data analysis
for marker and druggable target discovery

Single-cell RNA-seq data analysis for marker and druggable target discovery

Seokhyun Yoon

Professor, Dept. of Electronics & Electrical Eng., Dankook Univ
CTO, MLBI Lab. Co. Ltd.

Tumor micro-environment


Journal of Cell Science
The Tumor Microenvironment at a Glance
Frances R. Balkwill, Melania Capasso and Thorsten Hagemann
Lymphatic endothelial cells
T lymphocytes
B lymphocytes
Myeloid cells
NK and NKT cells
Cancer-associated fibroblasts
Vascular endothelial cells
Pericytes
Adipocytes
Mesenchymal stem cells
Key

Balkwill et. al., The tumor microenvironment at a glance. Journal of Cell Science (2012)

Analysis Tools

- Single-cell RNA-seq data
  - Cell type annotation (e.g., HiCAT)
  - DEG and GSA/GSEA
  - Inference of ligand-receptor interactions among cell groups (e.g., CellPhoneDB)
  - CNV estimation and Ploidy inference (e.g., InferCNV)
  - Pseudo-time trajectory analysis for differentiation study (e.g., Monocle)
  - :
  - Integrated analysis pipelines (e.g., SCODA)
- Spatially resolved transcriptomics data
  - Cell-type deconvolution/detection
  - Neighborhood enrichment analysis
  - Inference of ligand-receptor interactions
  - :

Contents & Resources

- Contents
  1. SCODA를 이용한 단일세포 RNA-seq 분석 개요
  2. AnnData 포맷 소개, SCANPY를 이용한 전처리, 세포유형식별
  3. CNV와 Ploidy 추정, 세포간 상호작용 분석
  4. DEG 분석, 마커 탐색, Gene Set Enrichment 분석

- Objective
  - SCODA로 처리된 단일세포 RNA-seq 데이터의 데이터 마이닝을 통해
  - 관련된 다양한 주제(세포 유형식별, CNV 추정 및 암세포 식별, 세포간 상호작용 분석, DEG 및 Gene Set분석)에 대한 이해를 높이고
  - 마커나 치료 표적의 발굴과 함께 검증을 위한 실험을 기획할 수 있도록 한다.

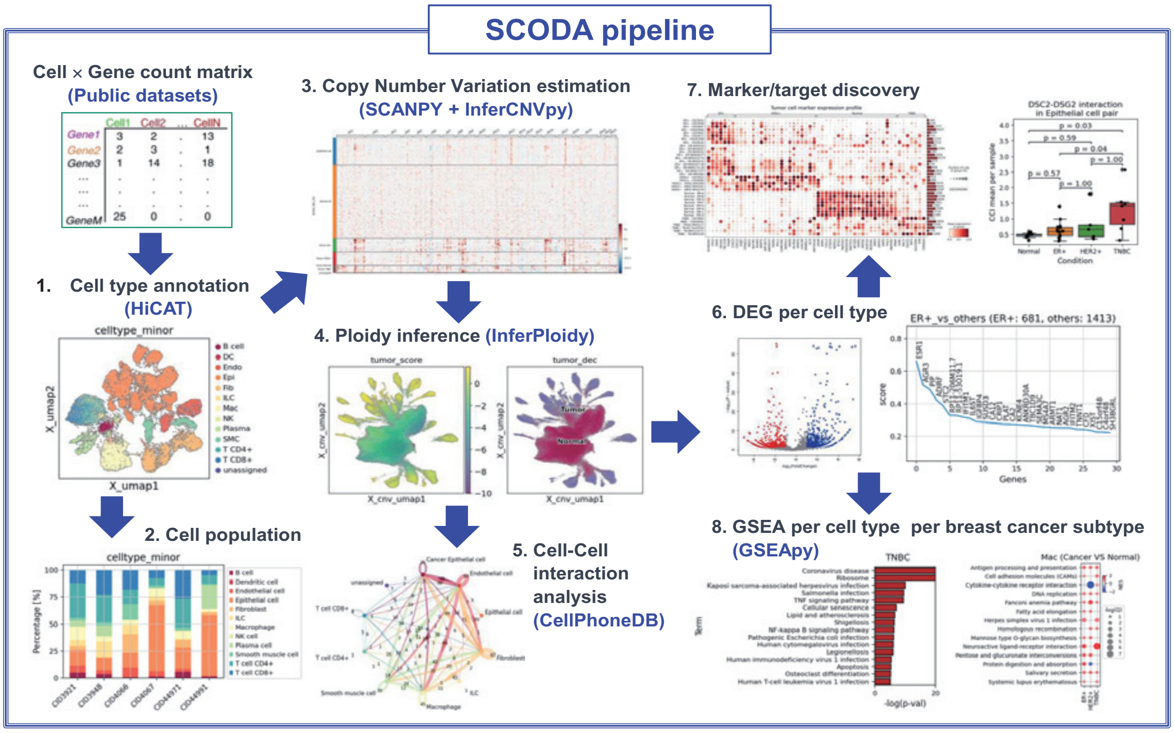
- Practice Workshop homepage: https://github.com/combio-dku/scoda_explorer/blob/main/Workshop/SCODA_worshop_2501.md

SCODA pipeline

an automated, web-based analysis pipeline for single-cell RNA-seq data

- SCODA demo site: https://mlbi-lab.net
- BRIC 기업기술웨비나: https://www.youtube.com/watch?v=ajRnK3QeCWA

SCODA pipeline overview



SCODA pipeline

Cell × Gene count matrix (Public datasets)

1. Cell type annotation (HiCAT)

2. Cell population

3. Copy Number Variation estimation (SCANPY + InferCNVpy)

4. Ploidy inference (InferPloidy)

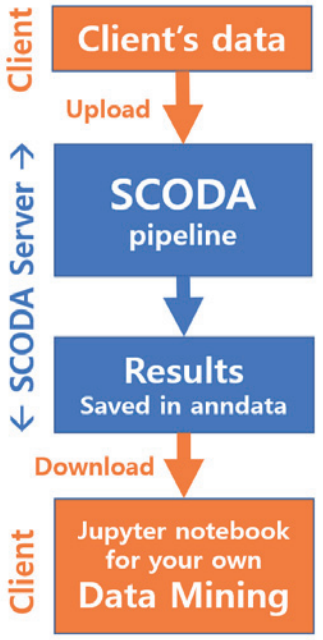
5. Cell-Cell interaction analysis (CellPhoneDB)

6. DEG per cell type

7. Marker/target discovery

8. GSEA per cell type per breast cancer subtype (GSEApy)

SCODA (Free demo site: https://mlbi-lab.net)



(1) Simply upload your data

1. Compressed 10x mtx files + optional meta_data.csv
2. Compressed CSV file + optional meta_data.csv
3. Compressed h5ad file (including meta data in obs field)

(2) Automated pipeline analyzes your data

1. Celltype annotation (HiCAT, MIT license)
2. DEG & GSA/GSEA (GSEApy, MIT license)
3. Cell-cell interaction (CellPhoneDB, MIT license)
4. CNV estimation & ploidy detection (InferCNV, BSD-3c + InferPloidy)

❖ We use proven & standardized analysis tools
❖ The pipeline can be tailored with optional settings

(3) Once it is done, you can download the result

1. Saved in Anndata formatted h5ad file (tar.gz compressed)
2. It contains all the analysis results and the count matrix itself

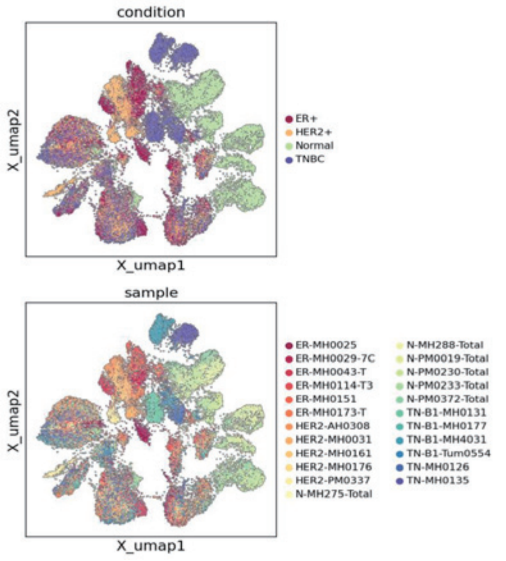
(4) Explore the results by yourself on your web browser

1. We provide "jupyter notebook" with example codes
2. You can freely use SCODA-viz & SCANpy (free open SW)
3. With a little bit of programming, you can get deeper insight into the single-cell RNA-seq data you uploaded
4. We offer free workshop for data mining (upon request)

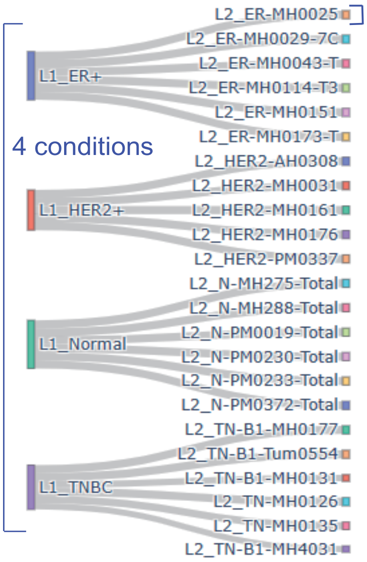
SCODA pipeline: Preparing the dataset

❑ Preparing the dataset

- Need to set conditions to compare
- And collect samples for each condition



23 samples

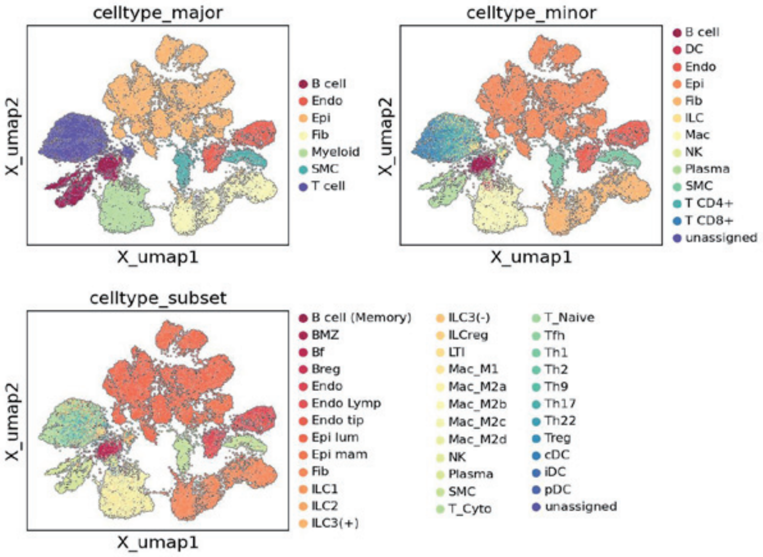


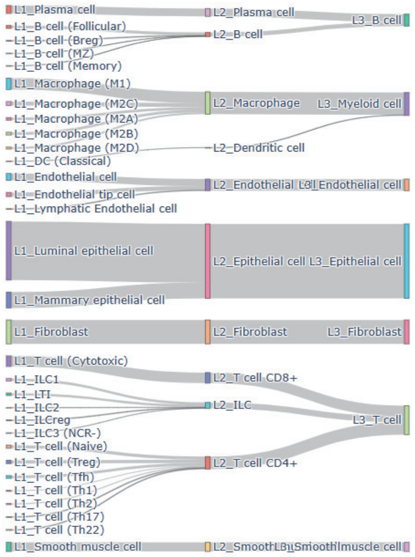
4 conditions

Each sample contains many cells of different types

SCODA pipeline: Cell-type annotation

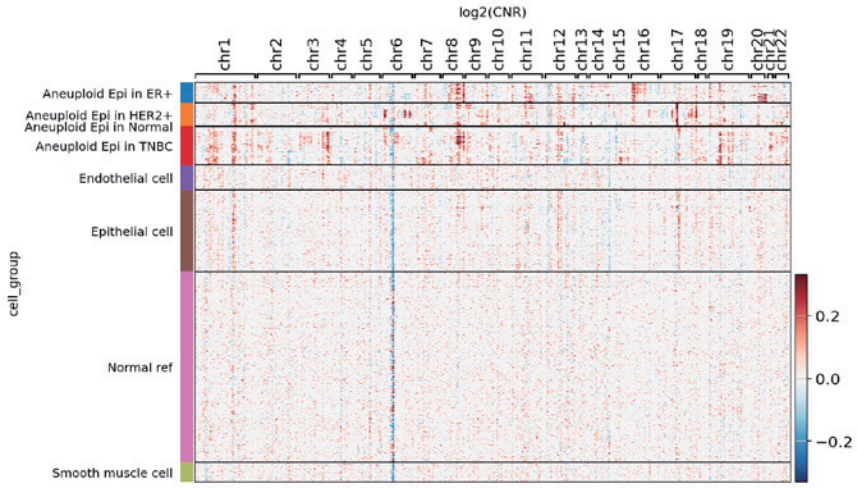
- ❑ Uses HiCAT (github.com/combio-dku/HiCAT)
- ❑ Utilizes cell-type markers from the R&D systems
- ❑ Annotate in 3-level taxonomy (major type, minor type, and subset)

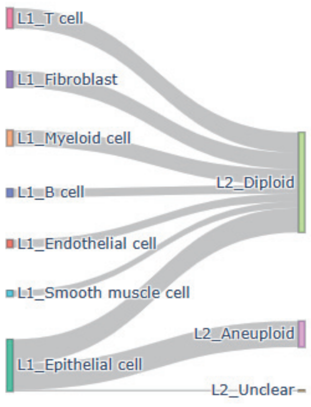




SCODA pipeline: CNV estimation & ploidy inference

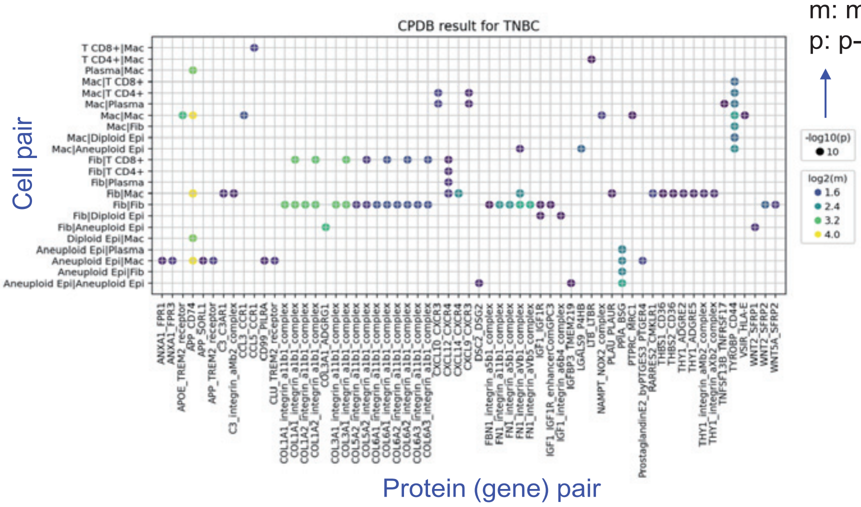
- ❑ Uses InferCNVpy (github.com/broadinstitute/infercnv) + InferPloidy
- ❑ Utilizes T cell, B cell, myeloid cell, and fibroblast as normal reference
- ❑ InferPloidy used to determine whether diploid (normal CNV) or aneuploid (abnormal CNV)



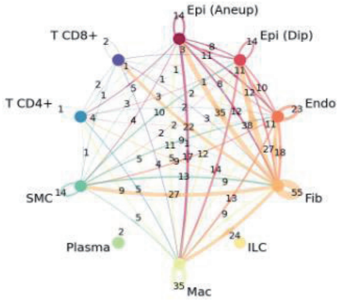


SCODA pipeline: Inferring ligand-receptor interaction

- ❑ Uses CellPhoneDB v4.0 (https://github.com/ventolab/CellphoneDB)
- ❑ First, run CellPhoneDB for each sample separately
- ❑ Then, condition specific interactions are identified by collecting the interactions commonly found in samples of a specific condition



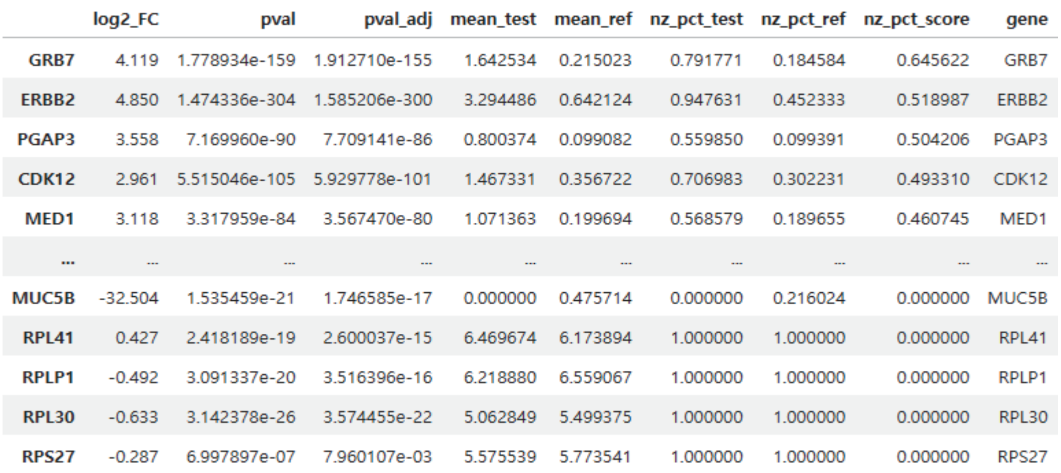
m: mean strength
p: p-value



SCODA pipeline: DEG analysis

- ❑ Uses SCANPY (https://scanpy.readthedocs.io/en/stable/)
- ❑ condition-specific DEG for each cell-type
- ❑ Provides both ‘one versus the rest’ and ‘one versus the reference'

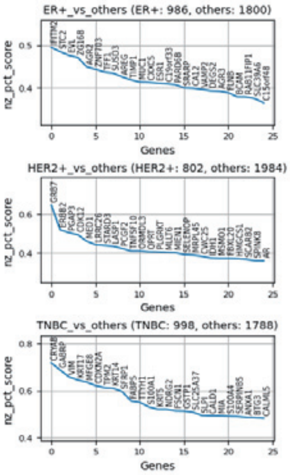
Epithelial cells: HER2+ vs. others

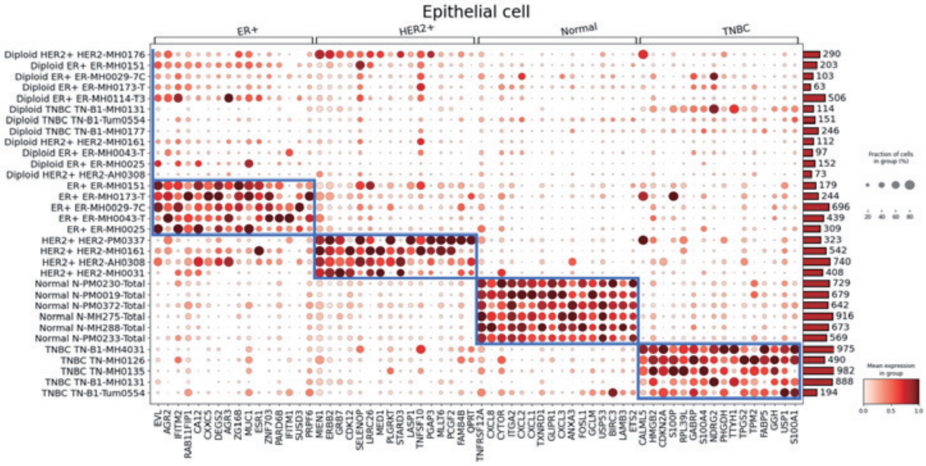


	log2_FC	pval	pval_adj	mean_test	mean_ref	nz_pct_test	nz_pct_ref	nz_pct_score	gene
GRB7	4.119	1.778934e-159	1.912710e-155	1.642534	0.215023	0.791771	0.184584	0.645622	GRB7
ERBB2	4.850	1.474336e-304	1.585206e-300	3.294486	0.642124	0.947631	0.452333	0.518987	ERBB2
PGAP3	3.558	7.169960e-90	7.709141e-86	0.800374	0.099082	0.559850	0.099391	0.504206	PGAP3
CDK12	2.961	5.515046e-105	5.929778e-101	1.467331	0.356722	0.706983	0.302231	0.493310	CDK12
MED1	3.118	3.317959e-84	3.567470e-80	1.071363	0.199694	0.568579	0.189655	0.460745	MED1
...	...	...	...	...	...	...	...	...	...
MUC5B	-32.504	1.535459e-21	1.746585e-17	0.000000	0.475714	0.000000	0.216024	0.000000	MUC5B
RPL41	0.427	2.418189e-19	2.600037e-15	6.469674	6.173894	1.000000	1.000000	0.000000	RPL41
RPLP1	-0.492	3.091337e-20	3.516396e-16	6.218880	6.559067	1.000000	1.000000	0.000000	RPLP1
RPL30	-0.633	3.142378e-26	3.574455e-22	5.062849	5.499375	1.000000	1.000000	0.000000	RPL30
RPS27	-0.287	6.997897e-07	7.960107e-03	5.575539	5.773541	1.000000	1.000000	0.000000	RPS27

SCODA pipeline: Marker discovery

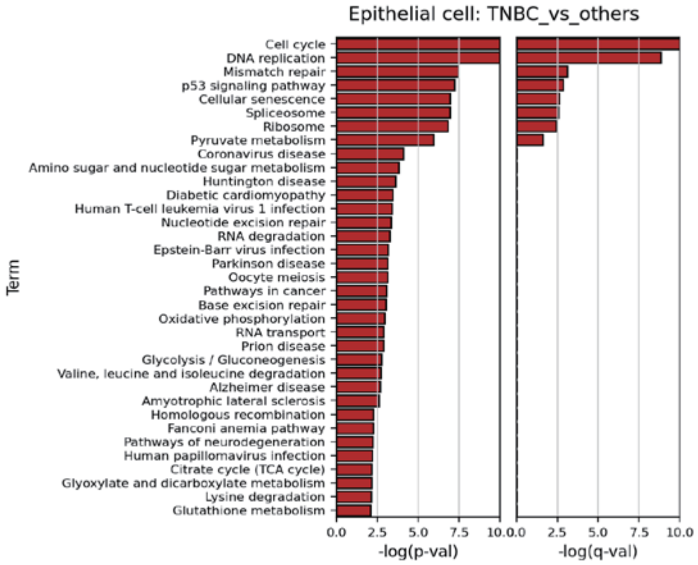
- Marker discovery in SCODA
  - Use DEG results. But not the log fold changes.
  - Marker score = nz_pct_score = nz_pct_test × (1 – nz_pct_ref)

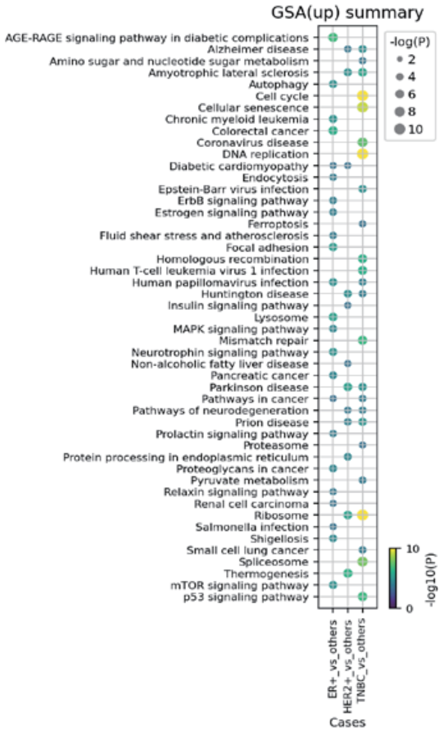




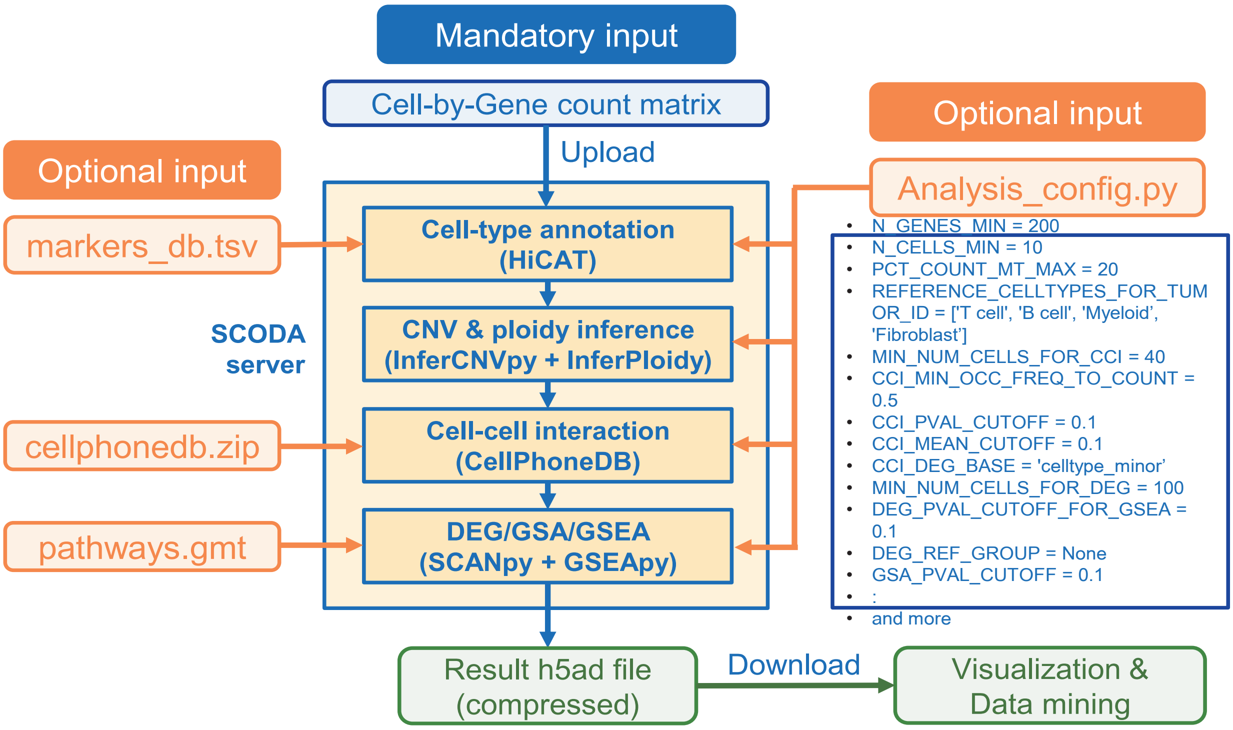
SCODA pipeline: Gene set (enrichment) analysis

- GSEApy (github.com/zqfang/GSEApy)
- Gene set analysis performed with the condition-specific DEGs separately for each cell type





SCODA configuration


Mandatory input
Cell-by-Gene count matrix
Upload
Optional input
markers_db.tsv
cellphonedb.zip
pathways.gmt
SCODA server
Cell-type annotation (HiCAT)
CNV & ploidy inference (InferCNVpy + InferPloidy)
Cell-cell interaction (CellPhoneDB)
DEG/GSA/GSEA (SCANpy + GSEApy)
Optional input
Analysis_config.py
Result h5ad file (compressed)
Download
Visualization & Data mining

- N_GENES_MIN = 200
- N_CELLS_MIN = 10
- PCT_COUNT_MT_MAX = 20
- REFERENCE_CELLTYPES_FOR_TUMOR_ID = ['T cell', 'B cell', 'Myeloid', 'Fibroblast']
- MIN_NUM_CELLS_FOR_CCI = 40
- CCI_MIN_OCC_FREQ_TO_COUNT = 0.5
- CCI_PVAL_CUTOFF = 0.1
- CCI_MEAN_CUTOFF = 0.1
- CCI_DEG_BASE = 'celltype_minor'
- MIN_NUM_CELLS_FOR_DEG = 100
- DEG_PVAL_CUTOFF_FOR_GSEA = 0.1
- DEG_REF_GROUP = None
- GSA_PVAL_CUTOFF = 0.1
- :
- and more

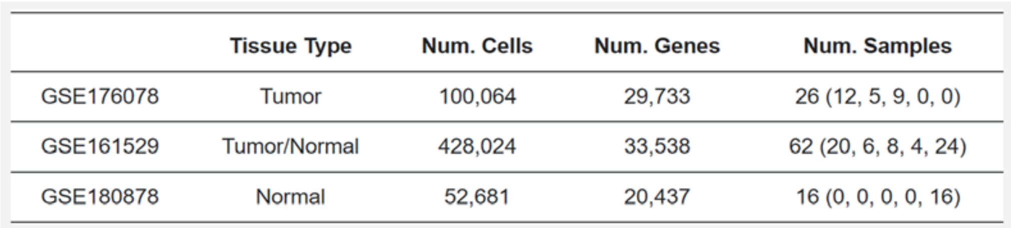
What one can do with this analysis pipeline?

Example study 1:
A study on tumor micro-environment of breast cancer using SCODA

- A Retrospective View on Triple Negative Breast Cancer Microenvironment: Novel Markers, Interactions, and Mechanisms of Tumor-Associated Components using public Single-cell RNA Seq Datasets, **Cancers, Mar. 23**

Background

- ❑ Objectives
  - ❍ Triple-negative breast cancer (TNBC) is a significant clinical challenge due to its aggressive nature and limited treatment options.
  - ❍ In search of new treatment targets, not only single genes but also gene pairs involved in protein interactions, we explored the tumor microenvironment (TME) of TNBC from a retrospective point of view, using public single-cell RNA sequencing datasets.
- ❑ Datasets used



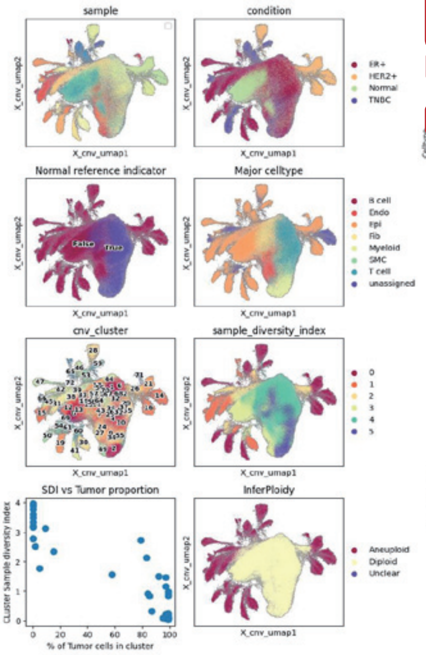
	Tissue Type	Num. Cells	Num. Genes	Num. Samples
GSE176078	Tumor	100,064	29,733	26 (12, 5, 9, 0, 0)
GSE161529	Tumor/Normal	428,024	33,538	62 (20, 6, 8, 4, 24)
GSE180878	Normal	52,681	20,437	16 (0, 0, 0, 0, 16)

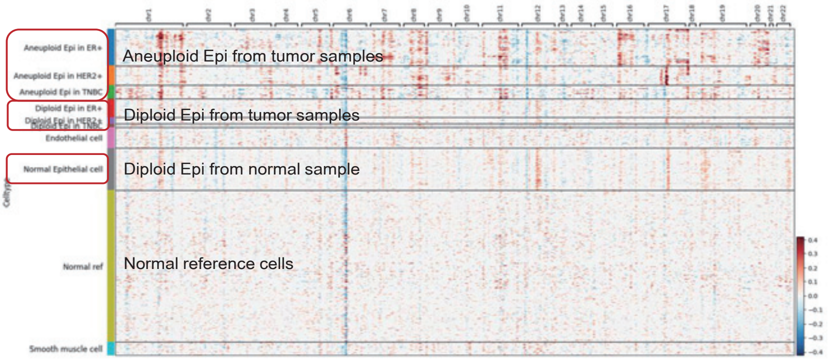
The numbers in parenthesis are the number of samples for ER+, HER2+, TNBC, Preneoplastic and Normal, respectively; In GSE161529, we did not consider 7 lymph node sequencing samples from the same ER+ patients.

Breast Cancer datasets (200K cells)

Samples collected from 3 BC datasets: 42 samples, 4 conditions (ER+, HER2+, TNBC, and Normal)

UMAP projection of CNV's





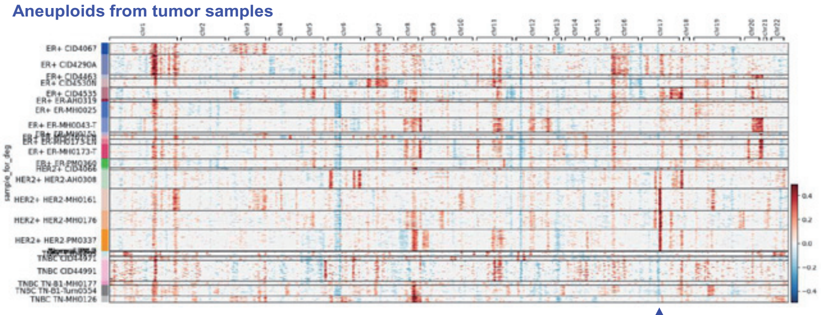
Aneuploid Epi from tumor samples

Diploid Epi from tumor samples

Diploid Epi from normal sample

Normal reference cells

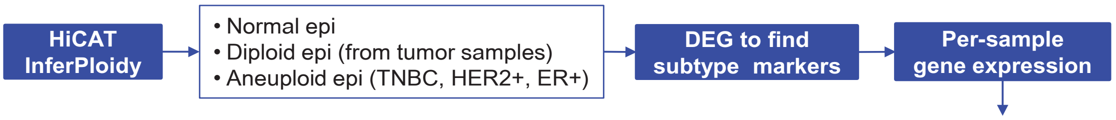
Aneuploids from tumor samples



Genomic spot containing ERBB2 gene in chr17

Breast Cancer Subtype markers

❑ BC subtype markers from DEG analysis (one vs the rest) on epithelial cells

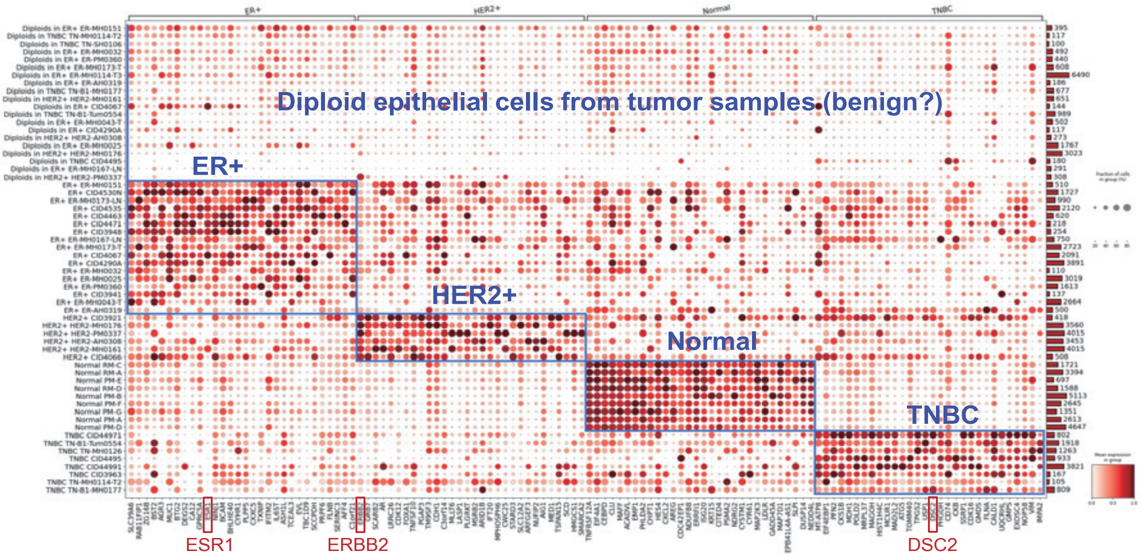


HiCAT InferPloidy

- Normal epi
- Diploid epi (from tumor samples)
- Aneuploid epi (TNBC, HER2+, ER+)

DEG to find subtype markers

Per-sample gene expression



Diploid epithelial cells from tumor samples (benign?)

ER+

HER2+

Normal

TNBC

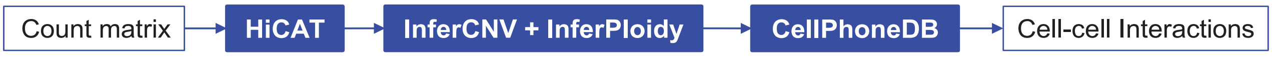
ESR1

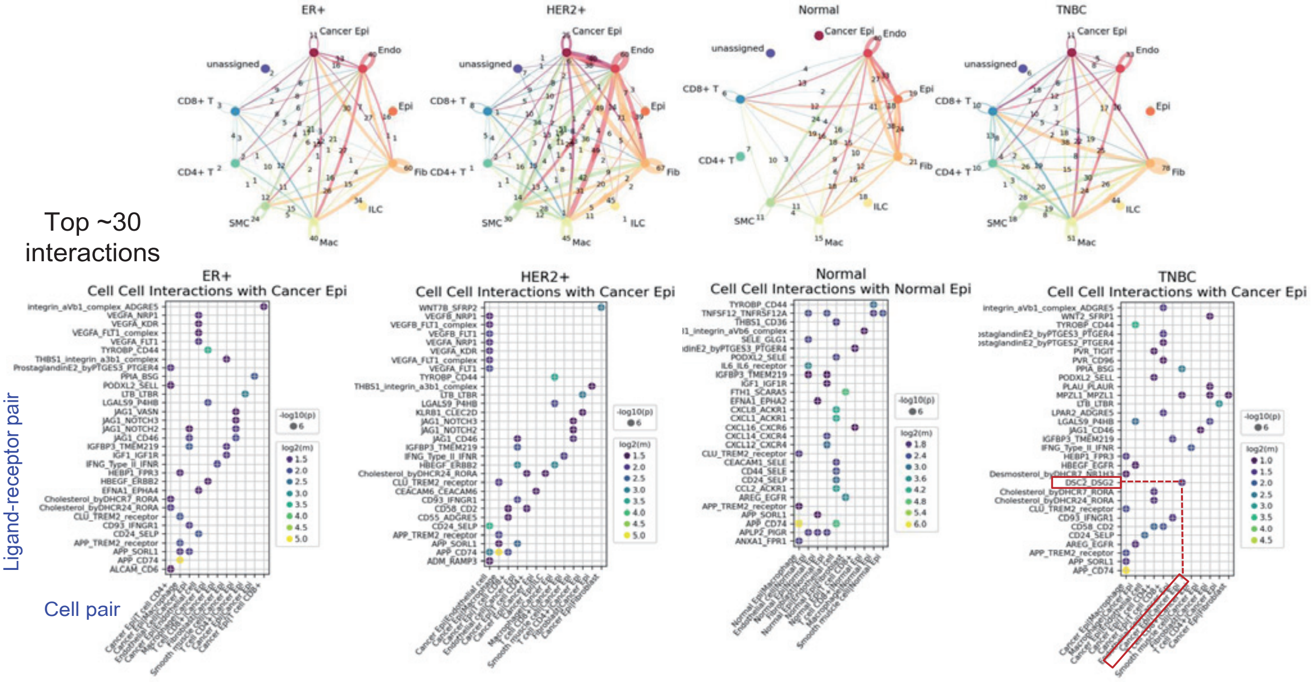
ERBB2

DSC2

Shown up to 30 markers for each condition

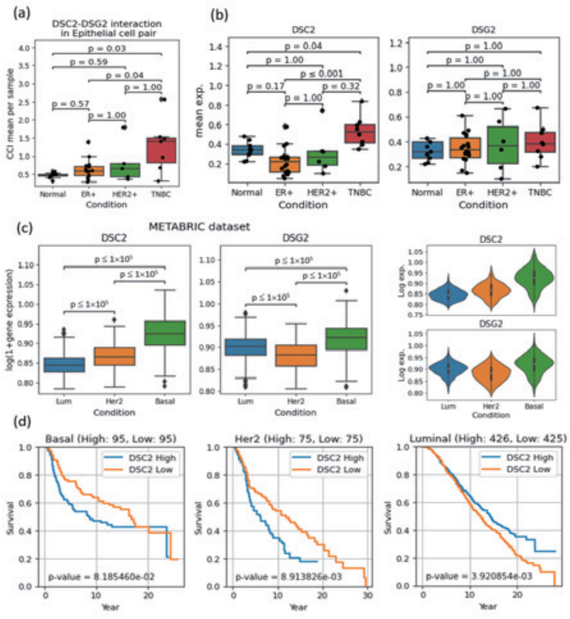
Cell-cell interactions in Breast Cancer


Count matrix → HiCAT → InferCNV + InferPloidy → CellPhoneDB → Cell-cell Interactions



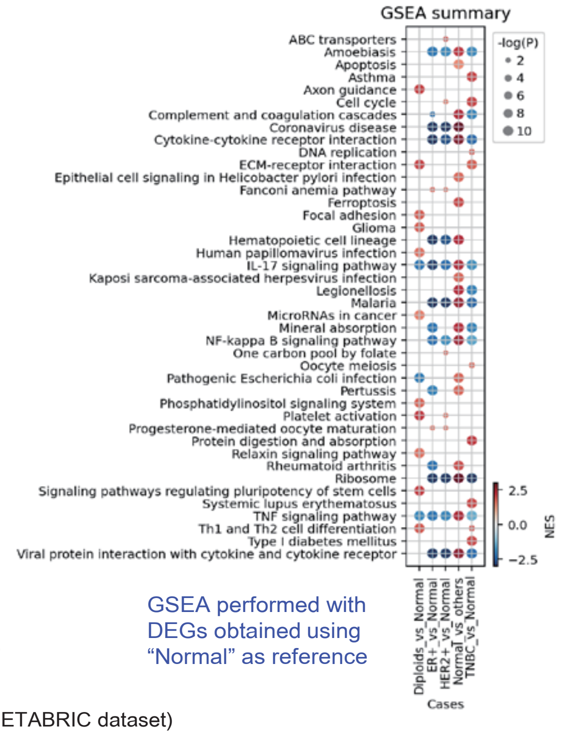
Top ~30 interactions

DSC2 as a prognostic marker

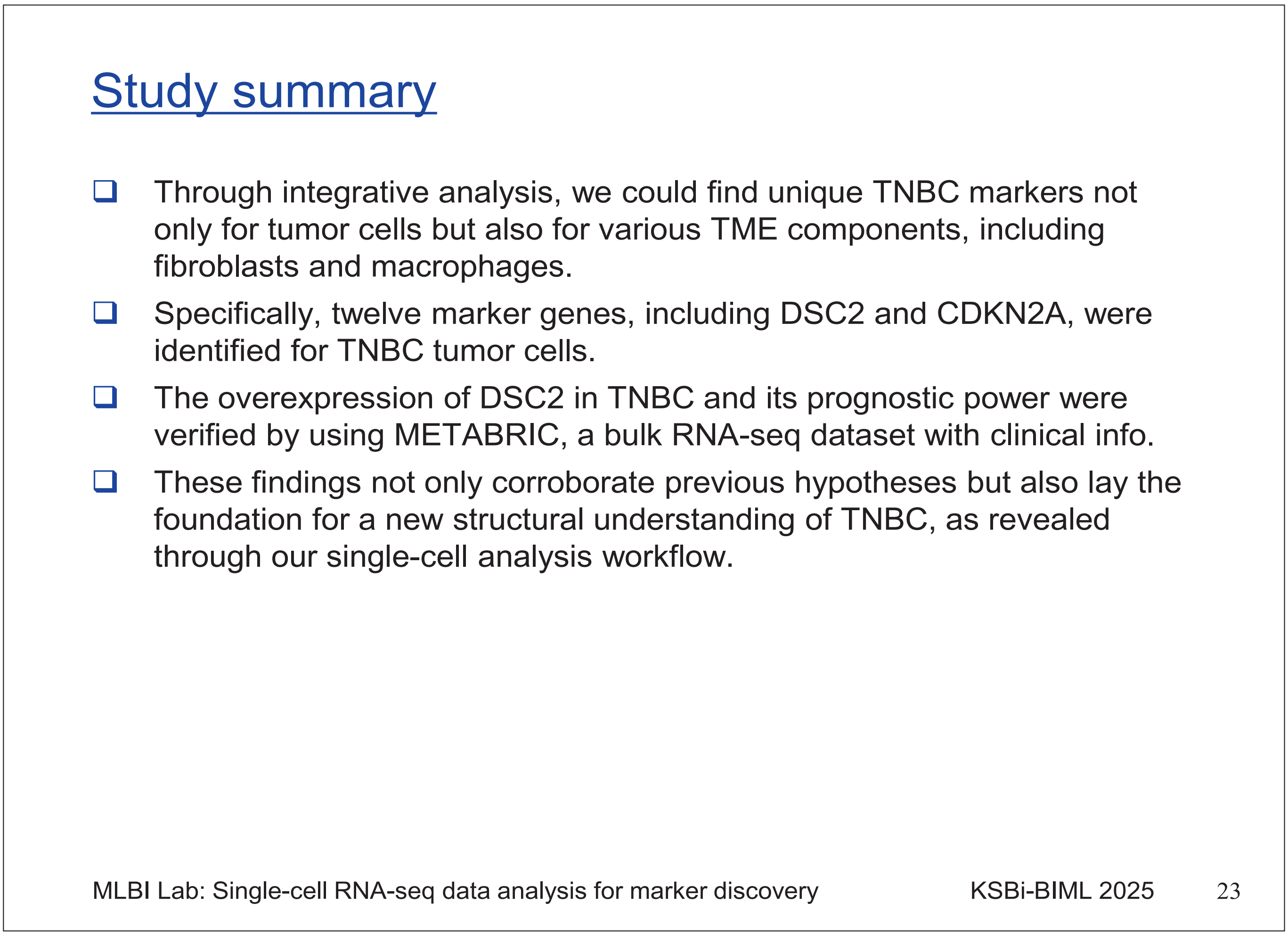


(a) DSC2-DSG2 interaction strength (single-cell RNA-seq datasets)
(b) DSC2, DSG2 gene expressions (single-cell RNA-seq datasets)
(c) DSC2, DSG2 gene expressions (METABRIC dataset)
(d) Survival differences of high/low DSC2 breast cancer patients (METABRIC dataset)

GSEA summary

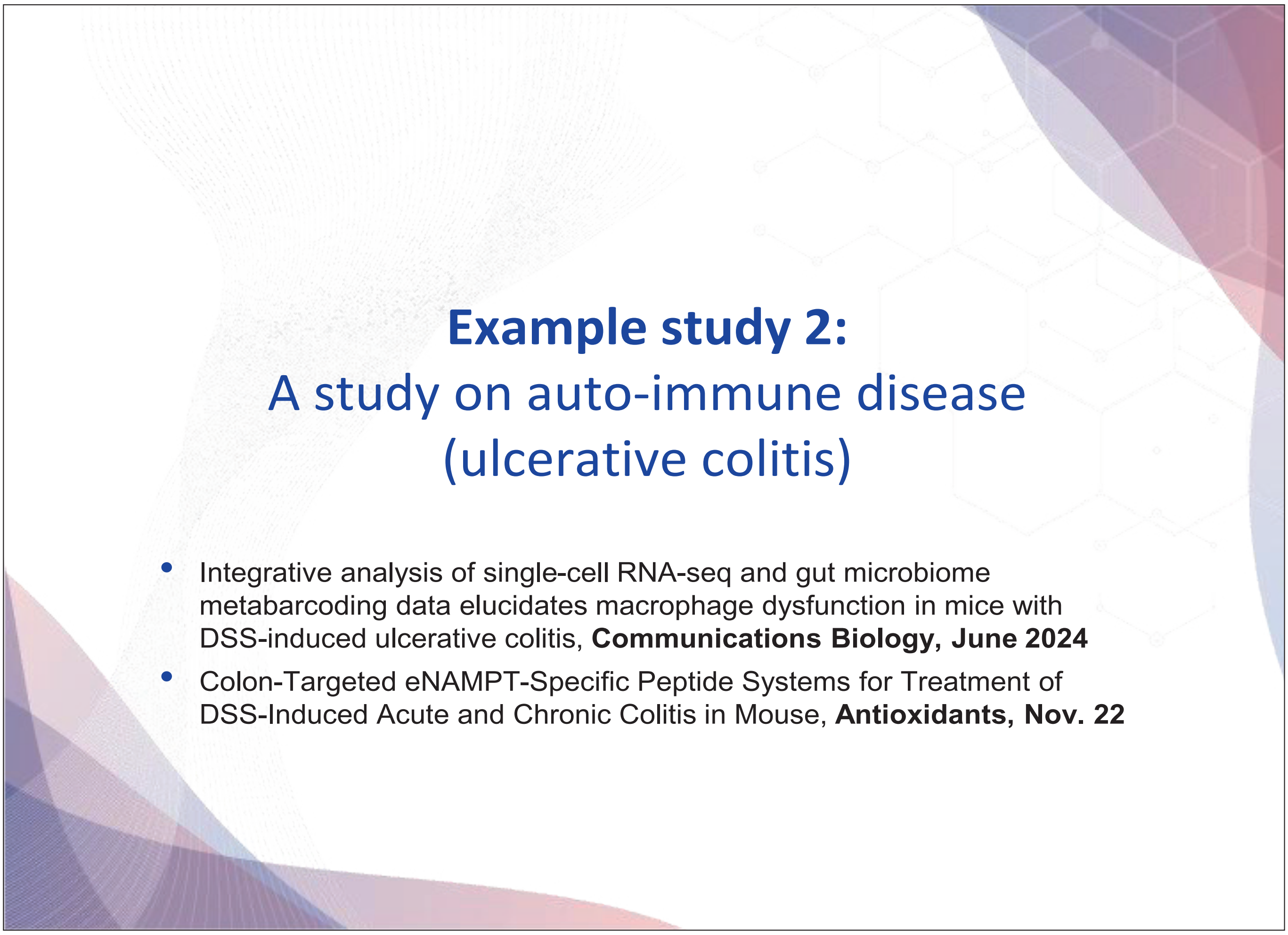


GSEA performed with DEGs obtained using “Normal” as reference



Study summary

- ❑ Through integrative analysis, we could find unique TNBC markers not only for tumor cells but also for various TME components, including fibroblasts and macrophages.
- ❑ Specifically, twelve marker genes, including DSC2 and CDKN2A, were identified for TNBC tumor cells.
- ❑ The overexpression of DSC2 in TNBC and its prognostic power were verified by using METABRIC, a bulk RNA-seq dataset with clinical info.
- ❑ These findings not only corroborate previous hypotheses but also lay the foundation for a new structural understanding of TNBC, as revealed through our single-cell analysis workflow.



Example study 2:
A study on auto-immune disease (ulcerative colitis)

- Integrative analysis of single-cell RNA-seq and gut microbiome metabarcoding data elucidates macrophage dysfunction in mice with DSS-induced ulcerative colitis, **Communications Biology, June 2024**
- Colon-Targeted eNAMPT-Specific Peptide Systems for Treatment of DSS-Induced Acute and Chronic Colitis in Mouse, **Antioxidants, Nov. 22**

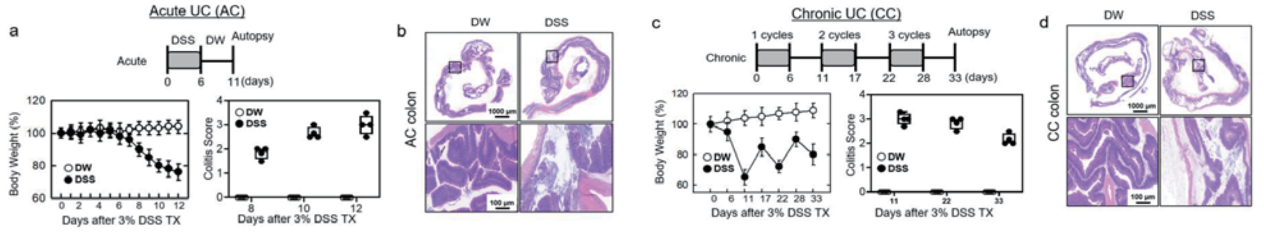
Background

- Objectives
  - Ulcerative colitis (UC) is a significant inflammatory bowel disease caused by an abnormal immune reaction.
  - There are still gaps in our understanding in what immune changes contribute to the chronic inflammation.
  - Our research aims to address this gap by analyzing single-cell RNA-seq datasets for human UC patients and DSS-induced UC mice.

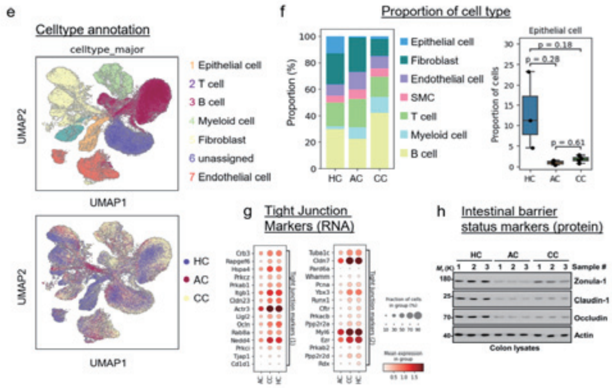
- Human datasets
  - SCP259 consisting of 360,000 cells from healthy, inflamed, and non-inflamed colonic tissues

Experiment & Analysis Overview

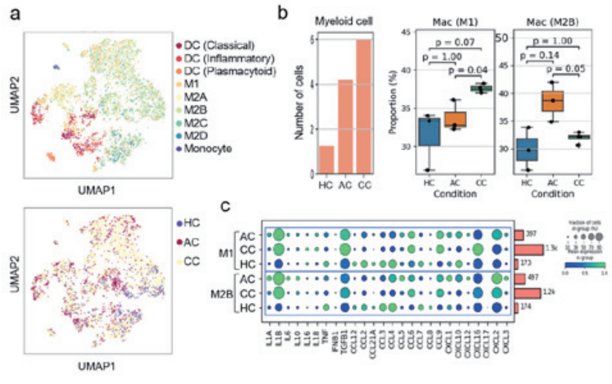
Sample preparation (DSS-induced UC mice)



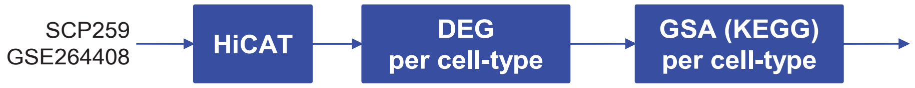
Single-cell RNA-seq data overview



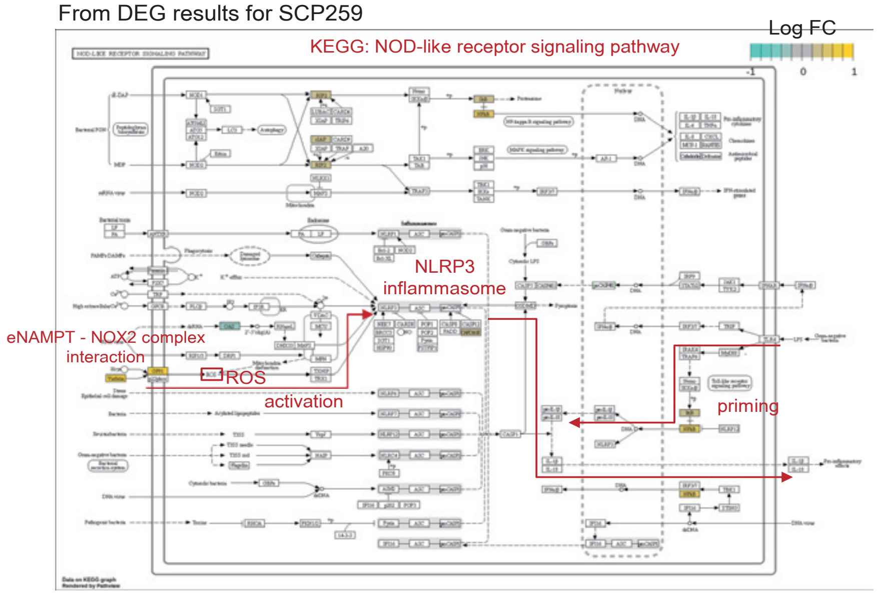
Macrophage subset overview

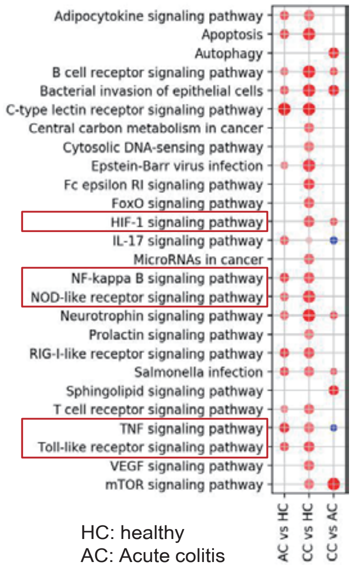


Macrophage dysfunction in ulcerative colitis



From DEG results for SCP259

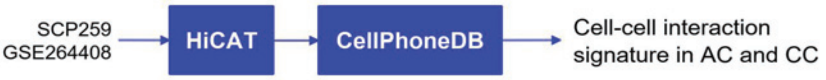


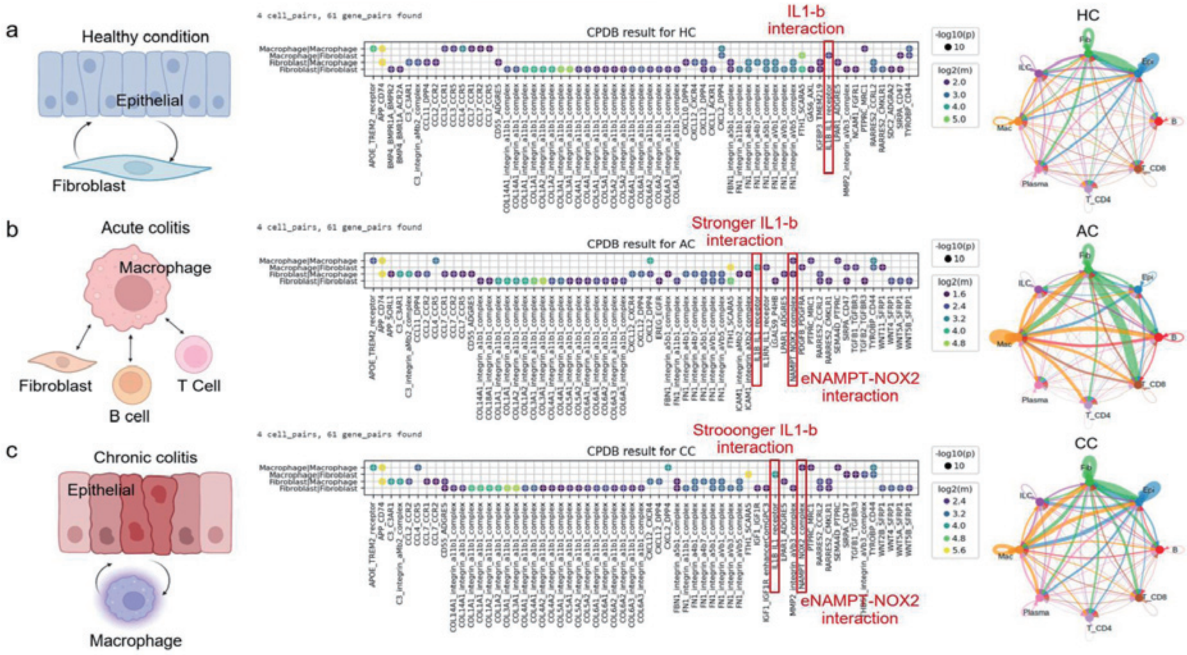


HC: healthy
AC: Acute colitis
CC: Chronic colitis

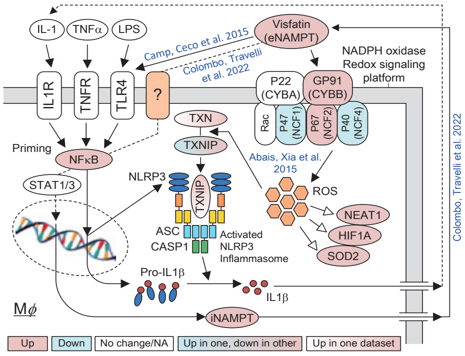
Visfatin(eNAMPT) - NOX2 interaction increases ROS level
→ activate NLRP3 inflammasome → secret IL1β

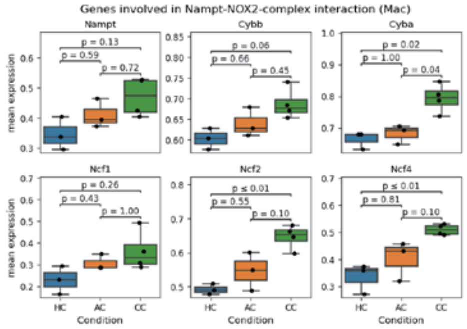
Macrophage dysfunction in chronic ulcerative colitis

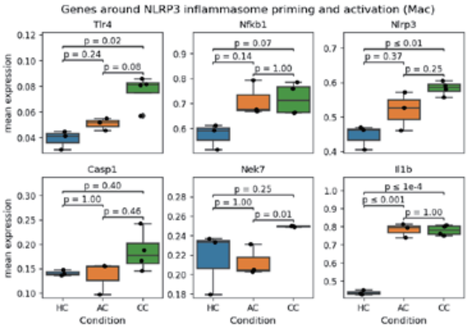


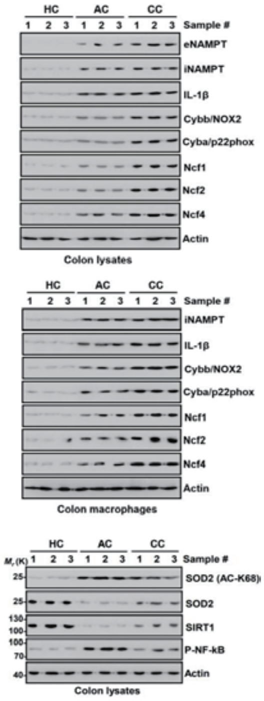


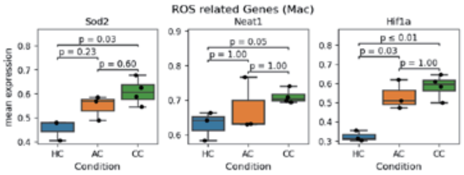
Macrophage dysfunction in chronic ulcerative colitis











Study summary

- ❑ Elucidated macrophage dysfunction and alternative polarization in UC patients and DSS-induced UC mice
- ❑ Identified a specific ligand-receptor interaction (eNAMPT-NOX2 complex) that may cause chronic activation of NLRP3 inflammasome
- ❑ Blockade of eNAMPT-NOX2 complex may alleviate colitis
- ❑ Similar approach can be applied to other autoimmune diseases to discover possible drug-targets

Contributors

이중호 박사
단국대 박사 (23)
(주) 현대중공업

김한별 연구원
단국대 석사 (23)
국립암센터

김민수 연구원
단국대 석사 (24)
국립암센터

양원희
단국대 석사 (24)
(주) 넥슨

성원정
단국대 석사과정
인공지능융합학과

채재영
단국대 석사과정
인공지능융합학과

홍다원 박사
단국대 생명융합학과

정선주 교수
단국대 생명융합학과

강근수 교수
단국대 미생물학과

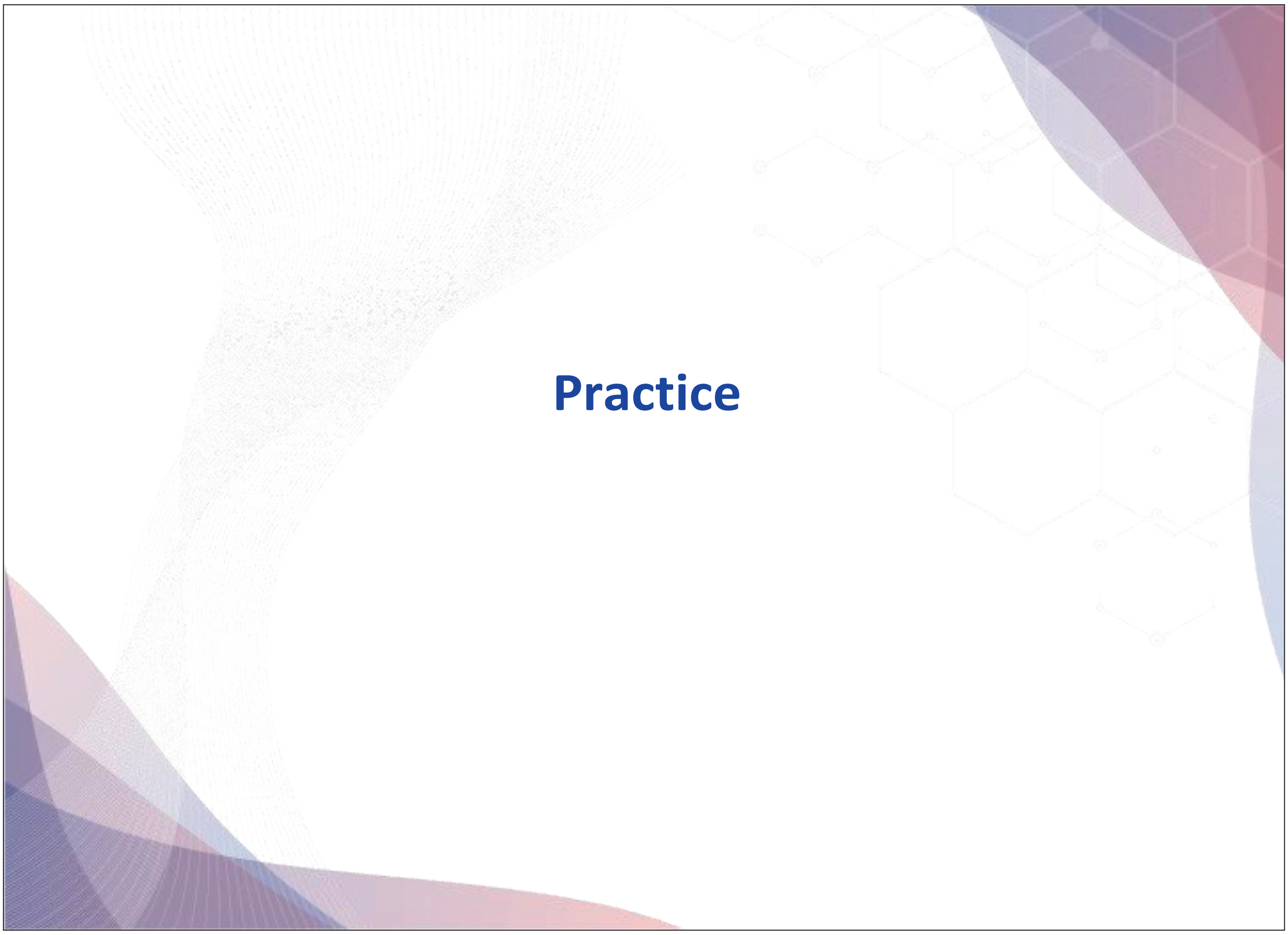
원혜성 교수
카톨릭대학병원 종양내과

양철수 교수
한양대 분자생명과학과

윤석현 교수
단국대 전자전기공학과
CTO, MLBI Lab.

MLBI Lab: Single-cell RNA-seq data analysis for marker discovery

KSBi-BIML 2025



Practice

Contents & Resources

- ❑ Contents
  1. SCODA를 이용한 단일세포 RNA-seq 분석 개요
  2. AnnData 포맷 소개, SCANPY를 이용한 전처리, 세포유형식별
  3. CNV와 Ploidy 추정, 세포간 상호작용 분석
  4. DEG 분석, 마커 탐색, Gene Set Enrichment 분석

- ❑ Objective
  - ❍ SCODA로 처리된 단일세포 RNA-seq 데이터의 데이터 마이닝을 통해
  - ❍ 관련된 다양한 주제(세포 유형식별, CNV 추정 및 암세포 식별, 세포간 상호작용 분석, DEG 및 Gene Set분석)에 대한 이해를 높이고
  - ❍ 마커나 치료 표적의 발굴과 함께 검증을 위한 실험을 기획할 수 있도록 한다.

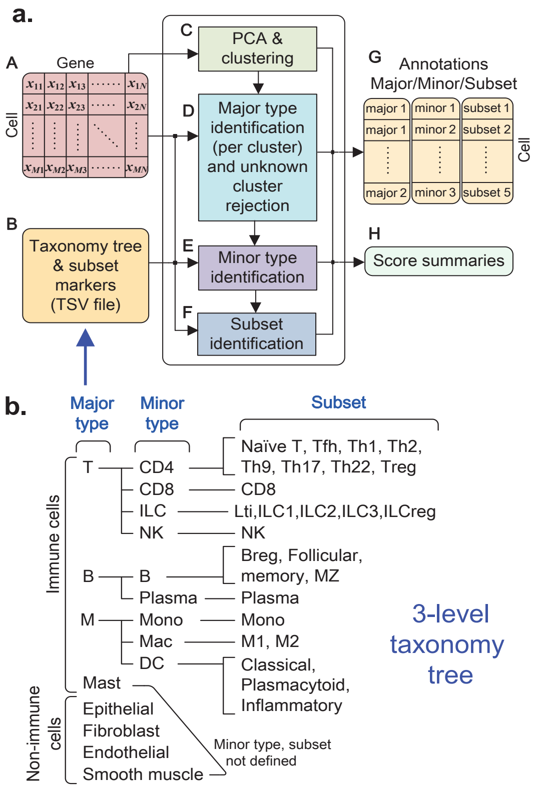
- ❑ Practice Workshop homepage: https://github.com/combio-dku/scoda_explorer/blob/main/Workshop/SCODA_worshop_2501.md

Automatic cell-type annotation using HiCAT

- Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing, **Briefings in bioinformatics, March 23**
- MarkerCount: A stable, count-based cell type identifier for single-cell RNA-seq experiments, **Comput. & Struct. Biotech. Journal, June 22**
- https://github.com/combio-dku/HiCAT
- BRIC Webinar: https://www.youtube.com/watch?v=wGFIzABbTF0

HiCAT: features

- ❑ **Marker-based**: utilizes known sets of cell type markers
- ❑ **Hierarchical annotation**: using 3-level taxonomy tree to successively annotate major type, minor type, and subset
- ❑ Using binary information (either expressed or not)
- ❑ it applies Gene Set Analysis (hyper-geometric function) for cell type scoring
- ❑ It can be used in both python and R environment


a.
A
Gene
Cell
x_{11} x_{12} x_{13} x_{1N}
x_{21} x_{22} x_{23} x_{2N}
x_{M1} x_{M2} x_{M3} x_{MN}
B
Taxonomy tree & subset markers (TSV file)
C
PCA & clustering
D
Major type identification (per cluster) and unknown cluster rejection
E
Minor type identification
F
Subset identification
G
Annotations Major/Minor/Subset
major 1
minor 1
subset 1
major 1
minor 2
subset 2
major 2
minor 3
subset 5
Cell
H
Score summaries
b.
Major type
Minor type
Subset
Immune cells
T
CD4
Naïve T, Tfh, Th1, Th2, Th9, Th17, Th22, Treg
CD8
CD8
ILC
Lti,ILC1,ILC2,ILC3,ILCreg
NK
NK
B
B
Breg, Follicular, memory, MZ
Plasma
Plasma
M
Mono
Mono
Mac
M1, M2
DC
Classical, Plasmacytoid, Inflammatory
Mast
Non-immune cells
Epithelial
Fibroblast
Endothelial
Smooth muscle
Minor type, subset not defined
3-level taxonomy tree

HiCAT: building blocks and procedure

c. Building blocks

A. PCA & Clustering

B. Count markers, get GSA score, Find best & 2nd best candidates

$b_{ik} = 1$ if $x_{ik} > 0$ or 0 otherwise

$n_{ij} = \sum_{k \in M_j} b_{ik}$

$p_{ij} = 1 - \sum_{k=0}^{n_{ij}-1} \frac{\binom{|M_i|}{k}\binom{N_g - |M_i|}{|G_j| - \mathrm{k}}}{\binom{N_g}{|G_j|}}$

$s_{ij} = -\log_{10}(p_{ij}) \rightarrow j_i^* = \underset{j}{\operatorname{argmax}}\, s_{ij}$

C. Unknown cluster detection

q_k: % of usable cells in cluster k

Apply linear fit $a^* k + b^*$ on ordered q_k's

Test if $q_k \lesssim a^* k + b^* - \epsilon$ (margin)

D. Use GMM to correct cell type, Find best & 2nd best

Use EM algorithm to obtain GMM

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{M_i} \pi_{i,k} N(\boldsymbol{x}; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k})$$

$l_{ij} = \log f_i(\boldsymbol{x}_j) \rightarrow j_i^{**} = \underset{j}{\operatorname{argmax}}\, l_{ij}$

E. Rejection of unclear cells

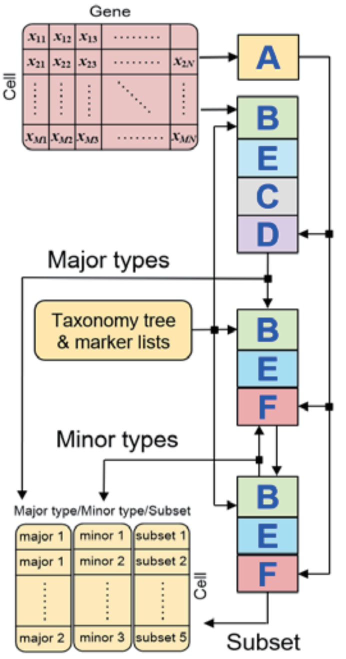
Rejection threshold t is obtained s.t.

$$FPR(t) = 1 - \frac{|C_i^{(2)}(t)|}{|C_i^{(1)}(t)|} < FPR_{\mathrm{th}}$$

F. kNN to correct minor type/subset

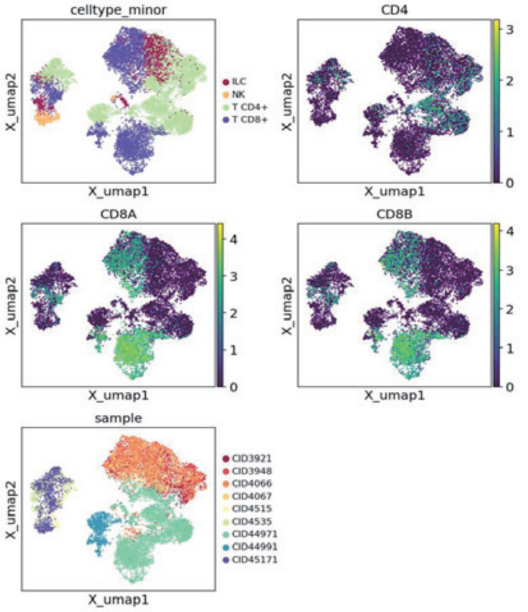
$C_i^{(1)}$ and $C_i^{(2)}$: the set of cells of which the best & the 2nd best type is i

d. Processing steps


Gene
Cell
x_{11} x_{12} x_{13}
x_{21} x_{22} x_{23} x_{2N}
x_{M1} x_{M2} x_{M3} x_{MN}
A
B
E
C
D
Major types
Taxonomy tree & marker lists
B
E
F
Minor types
B
E
F
Major type/Minor type/Subset
major 1
minor 1
subset 1
major 1
minor 2
subset 2
major 2
minor 3
subset 5
Cell
Subset

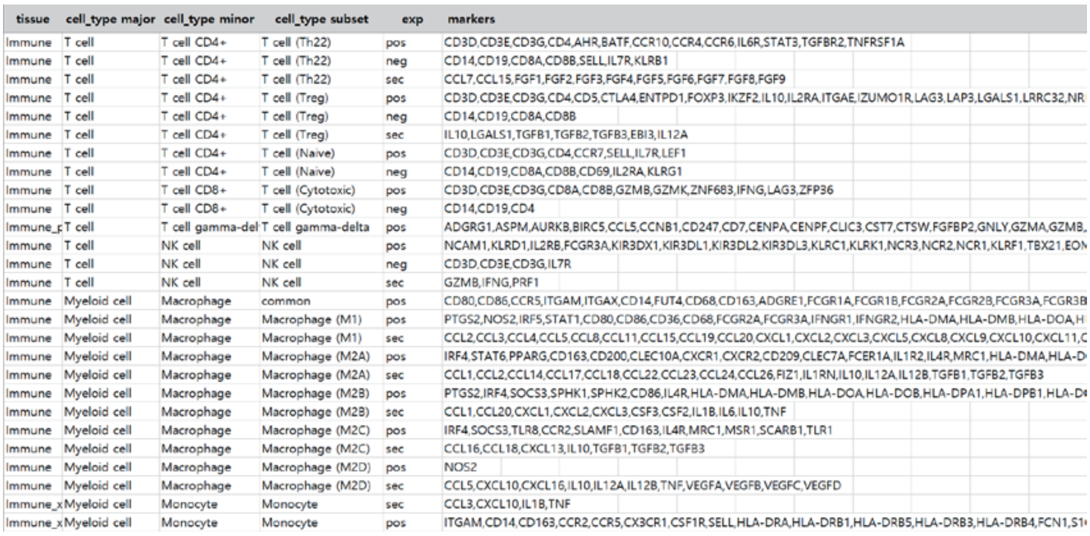
Gene expressions in single-cell RNA-seq data

- ❑ Protein expression versus RNA expression
  - ❍ As you see, not all CD4+ T cells express CD4 gene
  - ❍ Then, are they not really CD4 T cells?
  - ❍ Different from protein, RNA is unstable so that its expression is transient → Even if protein exist, its RNA maybe not.
  - ❍ Single-cell RNA-seq necessarily undergoes random sampling from a pool of RNAs from many cells
  - ❍ CD4 RNAs not selected for some T cells may result in the zero expression of CD4 gene in those CD4+ T cells



HiCAT: marker DB

- ❑ HiCAT marker DB specifies only subset markers
- ❑ Major(or minor) type markers = the union of the markers for the subsets belonging to that major(or minor) cell type

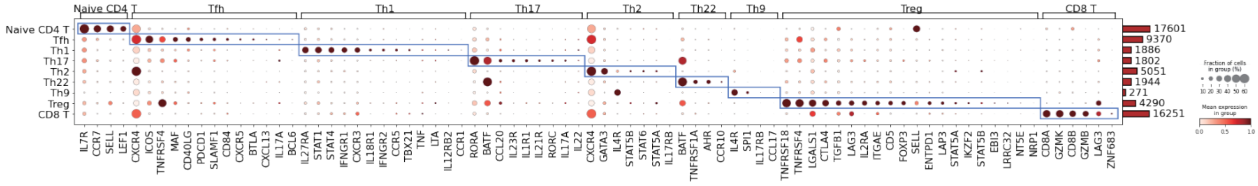


tissue	cell_type major	cell_type minor	cell_type subset	exp	markers
Immune	T cell	T cell CD4+	T cell (Th22)	pos	CD3D,CD3E,CD3G,CD4,AHR,BATF,CCR10,CCR4,CCR6,IL6R,STAT3,TGFBR2,TNFRSF1A
Immune	T cell	T cell CD4+	T cell (Th22)	neg	CD14,CD19,CD8A,CD8B,SELL,IL7R,KLRB1
Immune	T cell	T cell CD4+	T cell (Th22)	sec	CCL7,CCL15,FGF1,FGF2,FGF3,FGF4,FGF5,FGF6,FGF7,FGF8,FGF9
Immune	T cell	T cell CD4+	T cell (Treg)	pos	CD3D,CD3E,CD3G,CD4,CD5,CTLA4,ENTPD1,FOXP3,IKZF2,IL10,IL2RA,ITGAE,IZUMO1R,LAG3,LAP3,LGALS1,LRRC32,NR
Immune	T cell	T cell CD4+	T cell (Treg)	neg	CD14,CD19,CD8A,CD8B
Immune	T cell	T cell CD4+	T cell (Treg)	sec	IL10,LGALS1,TGFB1,TGFB2,TGFB3,EBI3,IL12A
Immune	T cell	T cell CD4+	T cell (Naive)	pos	CD3D,CD3E,CD3G,CD4,CCR7,SELL,IL7R,LEF1
Immune	T cell	T cell CD4+	T cell (Naive)	neg	CD14,CD19,CD8A,CD8B,CD69,IL2RA,KLRG1
Immune	T cell	T cell CD8+	T cell (Cytotoxic)	pos	CD3D,CD3E,CD3G,CD8A,CD8B,GZMB,GZMK,ZNF683,IFNG,LAG3,ZFP36
Immune	T cell	T cell CD8+	T cell (Cytotoxic)	neg	CD14,CD19,CD4
Immune_p	T cell	T cell gamma-del	T cell gamma-delta	pos	ADGRG1,ASPM,AURKB,BIRC5,CCL5,CCNB1,CD247,CD7,CENPA,CENPF,CLIC3,CST7,CTSW,FGFBP2,GNLY,GZMA,GZMB,
Immune	T cell	NK cell	NK cell	pos	NCAM1,KLRD1,IL2RB,FCGR3A,KIR3DX1,KIR3DL1,KIR3DL2,KIR3DL3,KLRC1,KLRK1,NCR3,NCR2,NCR1,KLRF1,TBX21,EOM
Immune	T cell	NK cell	NK cell	neg	CD3D,CD3E,CD3G,IL7R
Immune	T cell	NK cell	NK cell	sec	GZMB,IFNG,PRF1
Immune	Myeloid cell	Macrophage	common	pos	CD80,CD86,CCR5,ITGAM,ITGAX,CD14,FUT4,CD68,CD163,ADGRE1,FCGR1A,FCGR1B,FCGR2A,FCGR2B,FCGR3A,FCGR3B
Immune	Myeloid cell	Macrophage	Macrophage (M1)	pos	PTGS2,NOS2,IRF5,STAT1,CD80,CD86,CD36,CD68,FCGR2A,FCGR3A,IFNGR1,IFNGR2,HLA-DMA,HLA-DMB,HLA-DOA,H
Immune	Myeloid cell	Macrophage	Macrophage (M1)	sec	CCL2,CCL3,CCL4,CCL5,CCL8,CCL11,CCL15,CCL19,CCL20,CXCL1,CXCL2,CXCL3,CXCL5,CXCL8,CXCL9,CXCL10,CXCL11,C
Immune	Myeloid cell	Macrophage	Macrophage (M2A)	pos	IRF4,STAT6,PPARG,CD163,CD200,CLEC10A,CXCR1,CXCR2,CD209,CLEC7A,FCER1A,IL1R2,IL4R,MRC1,HLA-DMA,HLA-D
Immune	Myeloid cell	Macrophage	Macrophage (M2A)	sec	CCL1,CCL2,CCL14,CCL17,CCL18,CCL22,CCL23,CCL24,CCL26,FIZ1,IL1RN,IL10,IL12A,IL12B,TGFB1,TGFB2,TGFB3
Immune	Myeloid cell	Macrophage	Macrophage (M2B)	pos	PTGS2,IRF4,SOCS3,SPHK1,SPHK2,CD86,IL4R,HLA-DMA,HLA-DMB,HLA-DOA,HLA-DOB,HLA-DPA1,HLA-DPB1,HLA-D
Immune	Myeloid cell	Macrophage	Macrophage (M2B)	sec	CCL1,CCL20,CXCL1,CXCL2,CXCL3,CSF3,CSF2,IL1B,IL6,IL10,TNF
Immune	Myeloid cell	Macrophage	Macrophage (M2C)	pos	IRF4,SOCS3,TLR8,CCR2,SLAMF1,CD163,IL4R,MRC1,MSR1,SCARB1,TLR1
Immune	Myeloid cell	Macrophage	Macrophage (M2C)	sec	CCL16,CCL18,CXCL13,IL10,TGFB1,TGFB2,TGFB3
Immune	Myeloid cell	Macrophage	Macrophage (M2D)	pos	NOS2
Immune	Myeloid cell	Macrophage	Macrophage (M2D)	sec	CCL5,CXCL10,CXCL16,IL10,IL12A,IL12B,TNF,VEGFA,VEGFB,VEGFC,VEGFD
Immune_x	Myeloid cell	Monocyte	Monocyte	sec	CCL3,CXCL10,IL1B,TNF
Immune_x	Myeloid cell	Monocyte	Monocyte	pos	ITGAM,CD14,CD163,CCR2,CCR5,CX3CR1,CSF1R,SELL,HLA-DRA,HLA-DRB1,HLA-DRB5,HLA-DRB3,HLA-DRB4,FCN1,S1

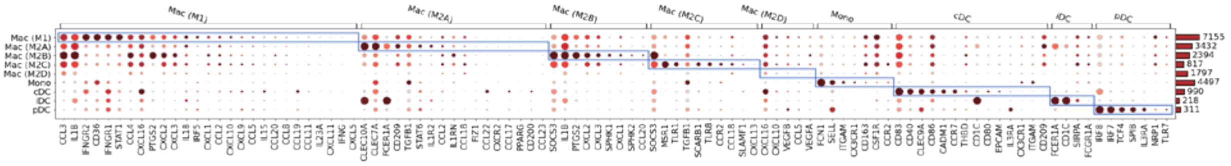
Obtained from https://www.rndsystems.com/resources/cell-markers

Subset marker expression under HiCAT annotation

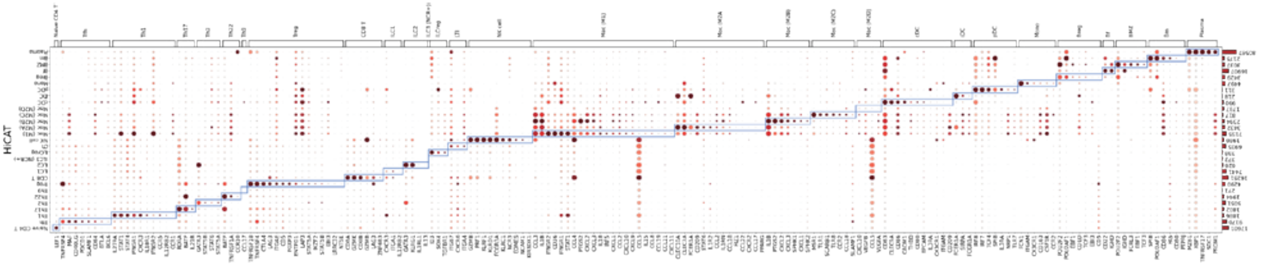
T cell subsets



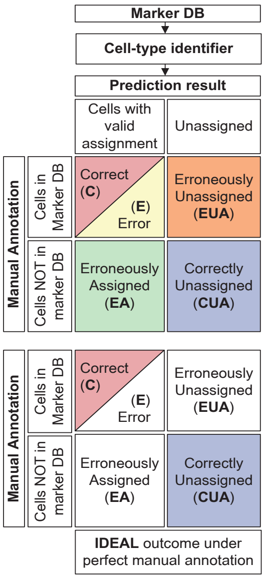
Myeloid cell subsets



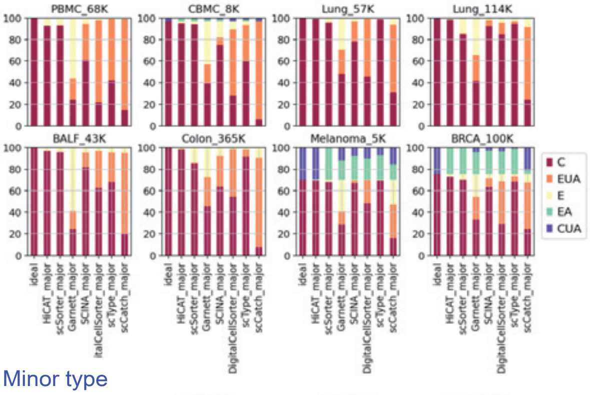
All immune cell subsets



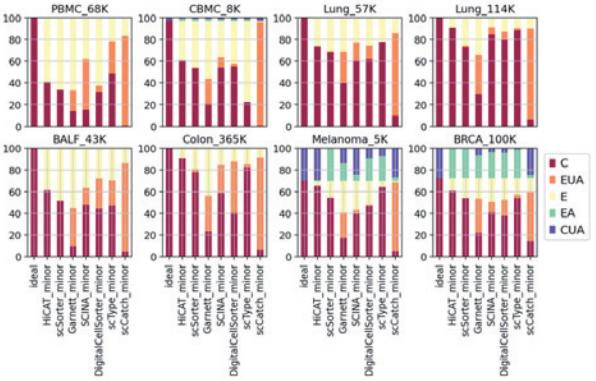
Quantitative comparisons


Marker DB
Cell-type identifier
Prediction result
Cells with valid assignment
Unassigned
Manual Annotation
Cells in Marker DB
Cells NOT in marker DB
Correct (C)
(E) Error
Erroneously Unassigned (EUA)
Erroneously Assigned (EA)
Correctly Unassigned (CUA)
Manual Annotation
Cells in Marker DB
Cells NOT in marker DB
Correct (C)
(E) Error
Erroneously Unassigned (EUA)
Erroneously Assigned (EA)
Correctly Unassigned (CUA)
IDEAL outcome under perfect manual annotation

Major type



Minor type



- ❑ For 8 datasets
- ❑ Compared HiCAT with 6 other marker-based annotation tools
  - ❍ scSorter
  - ❍ Garnett
  - ❍ SCINA
  - ❍ DigitalCellSorter
  - ❍ scType
  - ❍ scCatch
- ❑ Separately for major type and minor type
- ❑ HiCAT shows the best performance in most of the datasets

InferPloidy: a fast and accurate ploidy inference tool for tumor micro-environment study

(Unpublished, on going research)

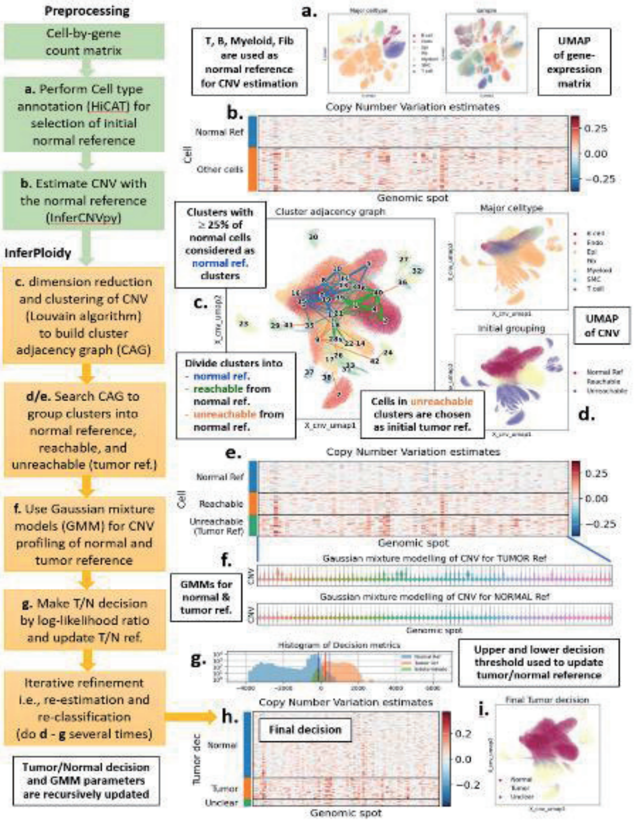
Background

- ❑ Previous works
  - ❍ InferCNV (2017) (CNV estimation only)
  - ❍ CopyKat (2021) (CNV estimation + tumor cell classification)
  - ❍ SCEVAN (2023) (CNV estimation + tumor cell classification)
- ❑ Motivation & Objective
  - ❍ Existing tools take too much time (It takes hours to get the result even for small-sized dataset with a few tens of thousand cells)
  - ❍ Computation time and memory requirement do not scale well with the size of datasets
  - ❍ We tried to develop a much faster and more accurate ploidy inference tool that runs on top of InferCNV.

InferPloidy

Brief processing steps

1. Utilizes Cell type annotation (HiCAT) for selection of normal reference cells
2. CNV estimation using InferCNV
3. Clustering of CNVs using Louvain algorithm to build cluster adjacency graph (CAG)
4. Search graph to test reachability from normal reference clusters
5. Cells in unreachable clusters used as tumor reference cells
6. Gaussian Mixture model is used for profiling of CNVs of normal and tumor reference cells
7. Likelihood ratio is used to make ploidy decision (diploid or aneuploid)
8. Iterative refinement of CNV profiles (do step 4-7 several times)



Preprocessing

Cell-by-gene count matrix

a. Perform Cell type annotation (HiCAT) for selection of initial normal reference

b. Estimate CNV with the normal reference (InferCNVpy)

InferPloidy

c. dimension reduction and clustering of CNV (Louvain algorithm) to build cluster adjacency graph (CAG)

d/e. Search CAG to group clusters into normal reference, reachable, and unreachable (tumor ref.)

f. Use Gaussian mixture models (GMM) for CNV profiling of normal and tumor reference

g. Make T/N decision by log-likelihood ratio and update T/N ref.

Iterative refinement i.e., re-estimation and re-classification (do **d** - **g** several times)

Tumor/Normal decision and GMM parameters are recursively updated

T, B, Myeloid, Fib are used as normal reference for CNV estimation

UMAP of gene-expression matrix

Clusters with ≥ 25% of normal cells considered as normal ref. clusters

Divide clusters into
- normal ref.
- reachable from normal ref.
- unreachable from normal ref.

Cells in unreachable clusters are chosen as initial tumor ref.

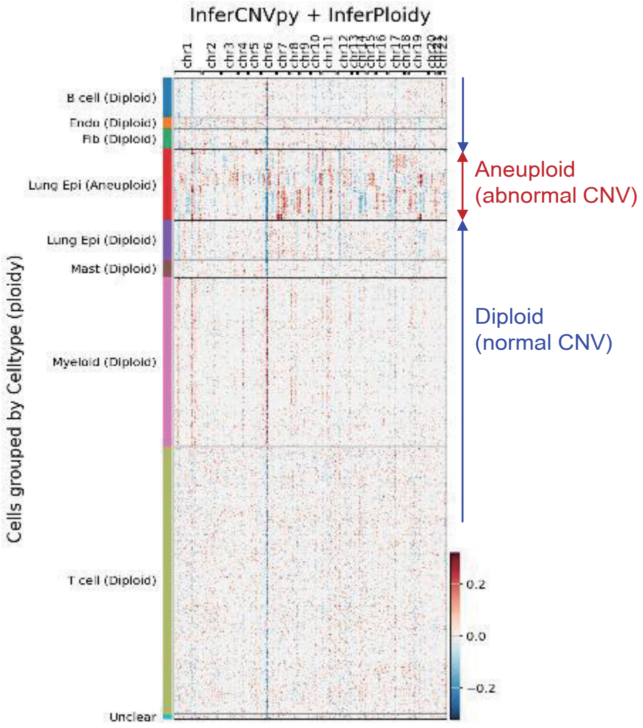
UMAP of CNV

GMMs for normal & tumor ref.

Upper and lower decision threshold used to update tumor/normal reference

Final decision

CNV estimates & inferred ploidy

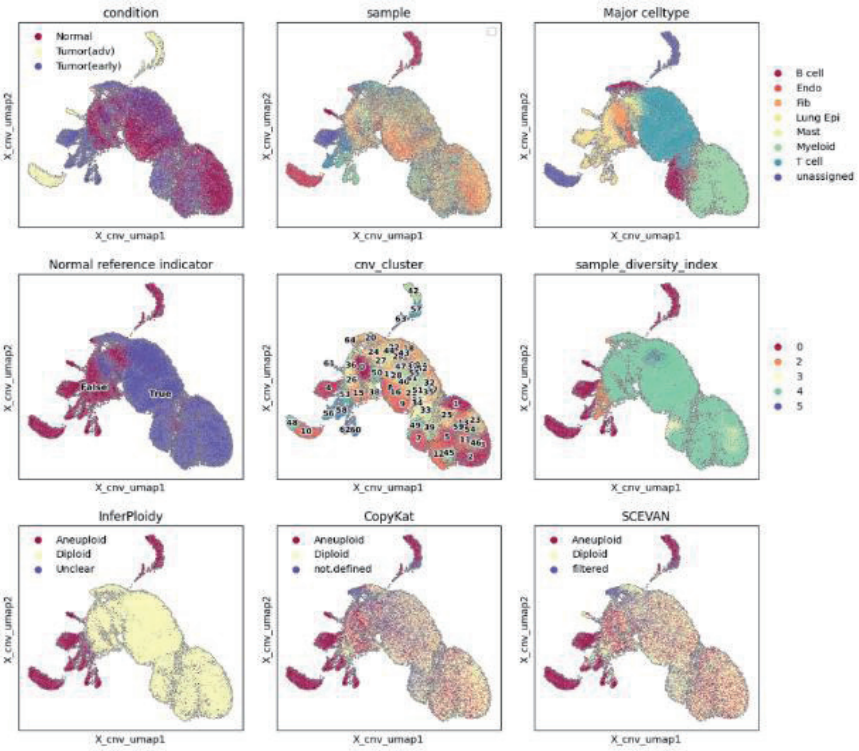


Aneuploid (abnormal CNV)

Diploid (normal CNV)

- ❑ Dataset: GSE131907
  - ❍ Non-small cell lung cancer
  - ❍ 26 samples,
  - ❍ 3 conditions (**distant normal, early tumor, adv tumor**),
  - ❍ containing 100K cells,
  - ❍ except for lymph node samples
- ❑ (to reduce running time) collected 1200 cells randomly from each sample
- ➢ analyzed ~36K cells
- ❑ Utilizing the annotated cell type,
- ❑ we used T/B/myeloid cells and fibroblasts as normal reference for InferCNV, CopyKat, and SCEVAN

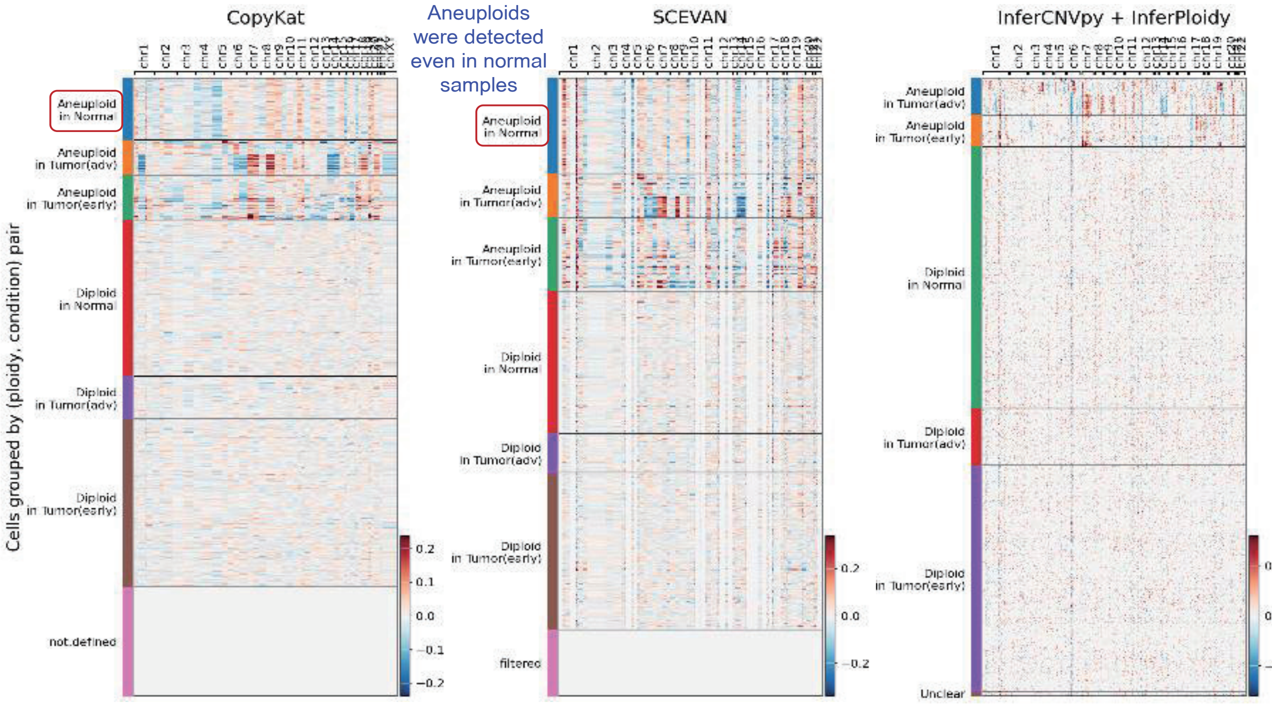
A comparison with other tools



Remarks

- ❑ Normal cell region roughly corresponds to the clusters with high sample diversity index
- ❑ While, tumor cell clusters to that with low sample diversity index
- ❑ With CopyKat and SCEVAN, many cells in normal reference were decided as aneuploid, which are certainly decision errors.
- ❑ Running time for 36K selected cells
  - ❍ InferCNV + InferPloidy: 15 mins
  - ❍ SCEVAN: 14 hours
  - ❍ CopyKat: 1.76 days

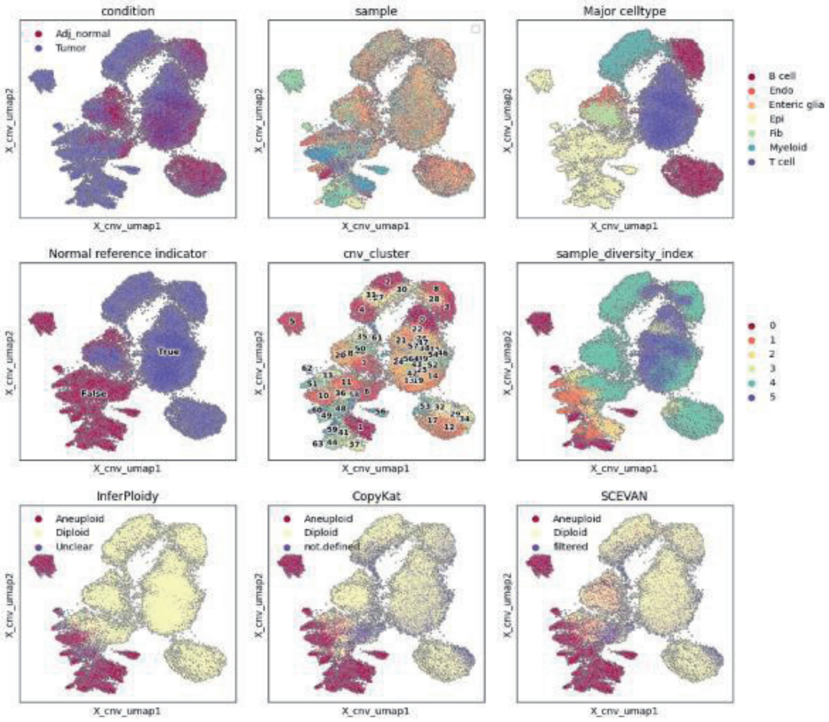
Comparison (GSE131907)



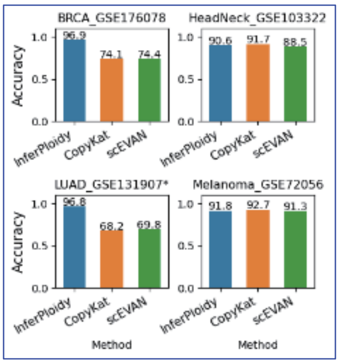
Aneuploids were detected even in normal samples

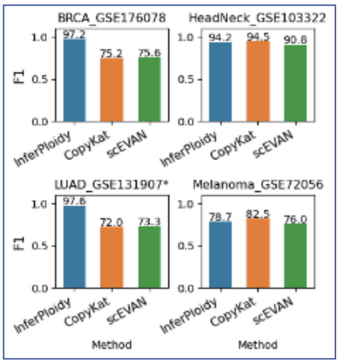
Comparison (other datasets)

GSE132465 Colorectal cancer dataset



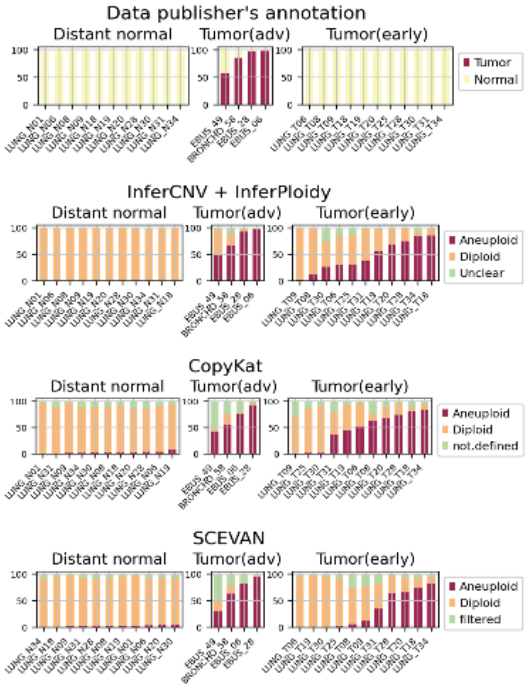
Accuracy and F1 score against publisher's annotation (excluding normal reference)





Aneuploids fraction in epithelial cells

GSE131907 (LUAD)



- ❑ No aneuploids in distant normal samples
- ❑ Tumor samples contains both aneuploid and diploid epithelial cells.
- ❑ Some tumor samples do not contain aneuploid cells
- ❑ Probably, the diploid epithelial cells in tumor samples are benign

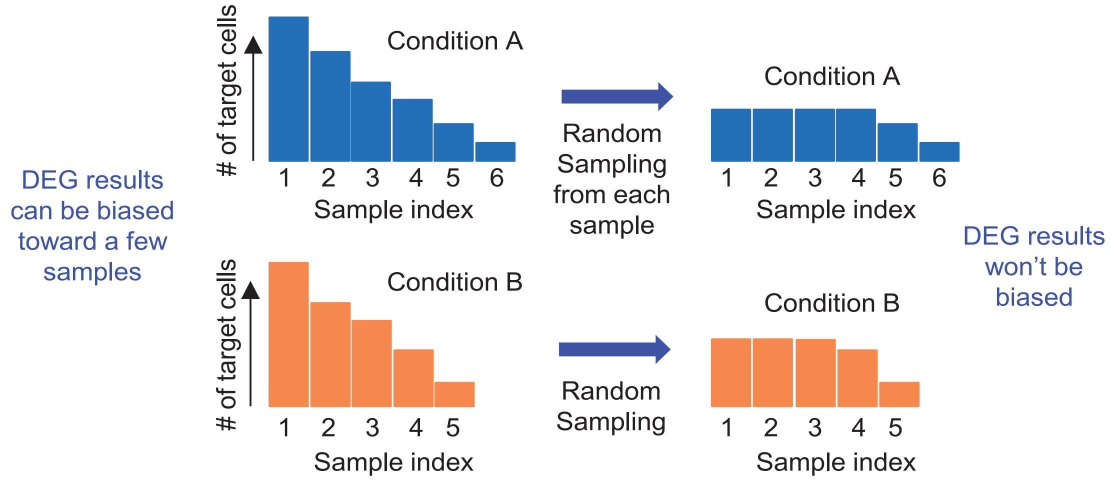
Remarks

- Accurate annotation of cell type and identification of cancer cells (aneuploid cells) are crucial for the down-stream analysis, including CCI, DEG, and GSEA
- HiCAT and InferPloidy are good choices for this purposes as they are faster and more accurate
- Equipped with these tools, SCODA pipeline can be conveniently used for the analysis of your single-cell RNA-seq data, specifically for tissue or tumor microenvironment study
- SCODA-viz package and Jupyter notebook with example codes are freely available for visualization and your own data mining

Features in SCODA DEG analysis

DEG analysis in SCODA

- ❑ Tool: SCANPY (https://scanpy.readthedocs.io/en/stable/)
- ❑ We are interested in condition-specific DEGs for each cell type
- ❑ To avoid sample bias, need to collect cells as evenly among samples as possible, but keeping the total number not too small



DEG results can be biased toward a few samples

Random Sampling from each sample

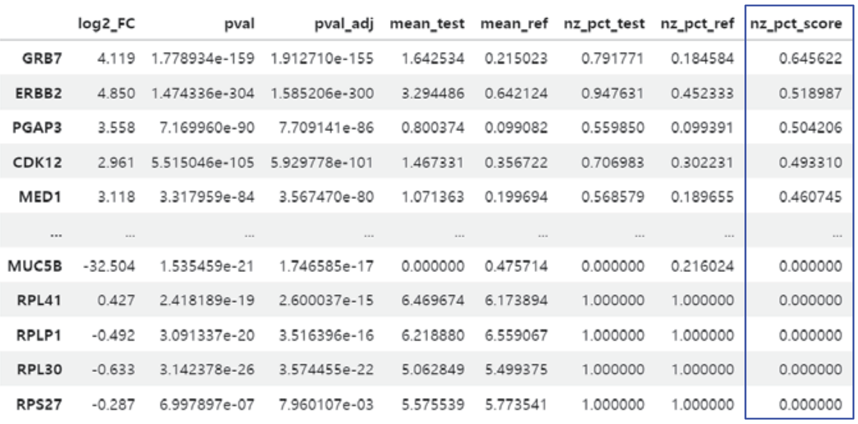
Random Sampling

DEG results won't be biased

DEG analysis in SCODA

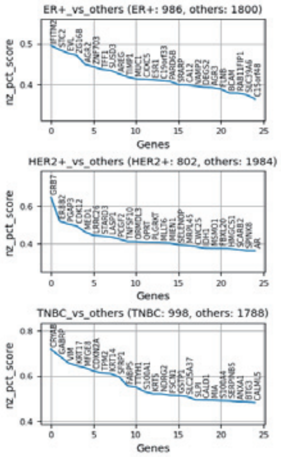
- ❑ Traditional DEG
  - ❍ $\log(FC) = \log\left(\frac{\exp(\log(1+X_{Test}))-1+10^{-10}}{\exp(\log(1+X_{Ref}))-1+10^{-10}}\right)$
  - ❍ Statistical test (e.g., *t*-test) performed assuming $\log(1+X)$ are normally distributed.
  - ❍ This assumption is reasonable for bulk RNA-seq data

DEG result on Epithelial cells (HER2+ vs. others)



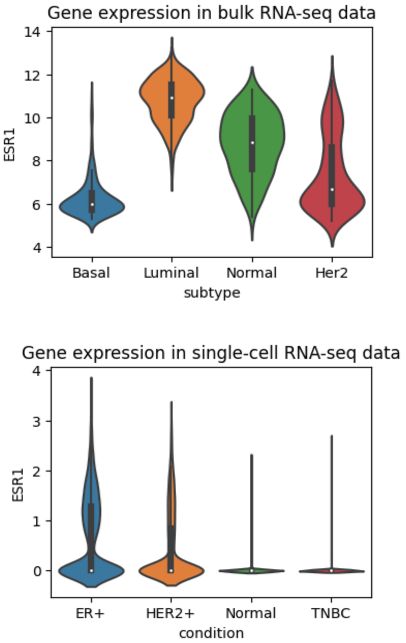
	log2_FC	pval	pval_adj	mean_test	mean_ref	nz_pct_test	nz_pct_ref	nz_pct_score
GRB7	4.119	1.778934e-159	1.912710e-155	1.642534	0.215023	0.791771	0.184584	0.645622
ERBB2	4.850	1.474336e-304	1.585206e-300	3.294486	0.642124	0.947631	0.452333	0.518987
PGAP3	3.558	7.169960e-90	7.709141e-86	0.800374	0.099082	0.559850	0.099391	0.504206
CDK12	2.961	5.515046e-105	5.929778e-101	1.467331	0.356722	0.706983	0.302231	0.493310
MED1	3.118	3.317959e-84	3.567470e-80	1.071363	0.199694	0.568579	0.189655	0.460745
...	...	...	...	...	...	...	...	...
MUC5B	-32.504	1.535459e-21	1.746585e-17	0.000000	0.475714	0.000000	0.216024	0.000000
RPL41	0.427	2.418189e-19	2.600037e-15	6.469674	6.173894	1.000000	1.000000	0.000000
RPLP1	-0.492	3.091337e-20	3.516396e-16	6.218880	6.559067	1.000000	1.000000	0.000000
RPL30	-0.633	3.142378e-26	3.574455e-22	5.062849	5.499375	1.000000	1.000000	0.000000
RPS27	-0.287	6.997897e-07	7.960107e-03	5.575539	5.773541	1.000000	1.000000	0.000000

Selected Epi markers



DEG in single-cell RNA-seq data

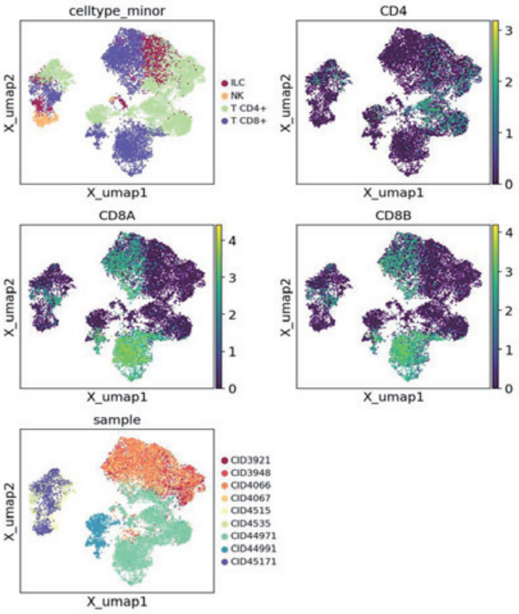
- What's different from bulk RNA-seq data
  - In bulk, gene expressions can be roughly approximated as normal random variable
  - So that t-test (or other test) gives us good insight into the 'relative' difference in gene expressions
  - In single-cell RNA-seq data, many cells have 'zero-expression'.
  - If we include zero-expression cells when computing mean and SD, the fold changes and the statistical test (p-value) under the normality assumption may be meaningless.
  - Then, do we have to exclude them?
    (I don't think so)



ESR1 expression in Epithelial cells

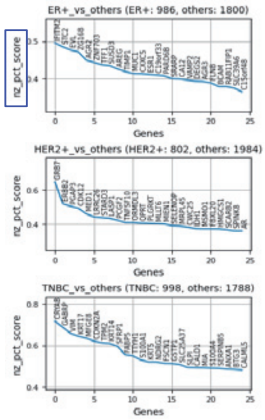
DEG in single-cell RNA-seq data

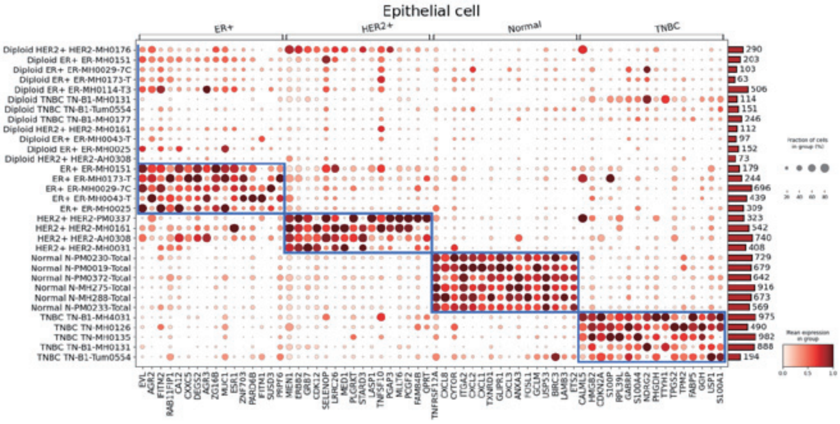
- Zero-expression cells in single-cell RNA-seq data
  - As you see, not all CD4+ T cells express CD4 gene
  - Then, are they not really CD4 T cells?
  - Different from protein, RNA is unstable so that its expression is transient → Even if protein exist, its RNA maybe not.
  - Single-cell RNA-seq necessarily undergoes random sampling from a pool of RNAs from many cells
  - CD4 RNAs not selected for some T cells may result in the zero expression of CD4 gene in those CD4+ T cells



Marker discovery in SCODA

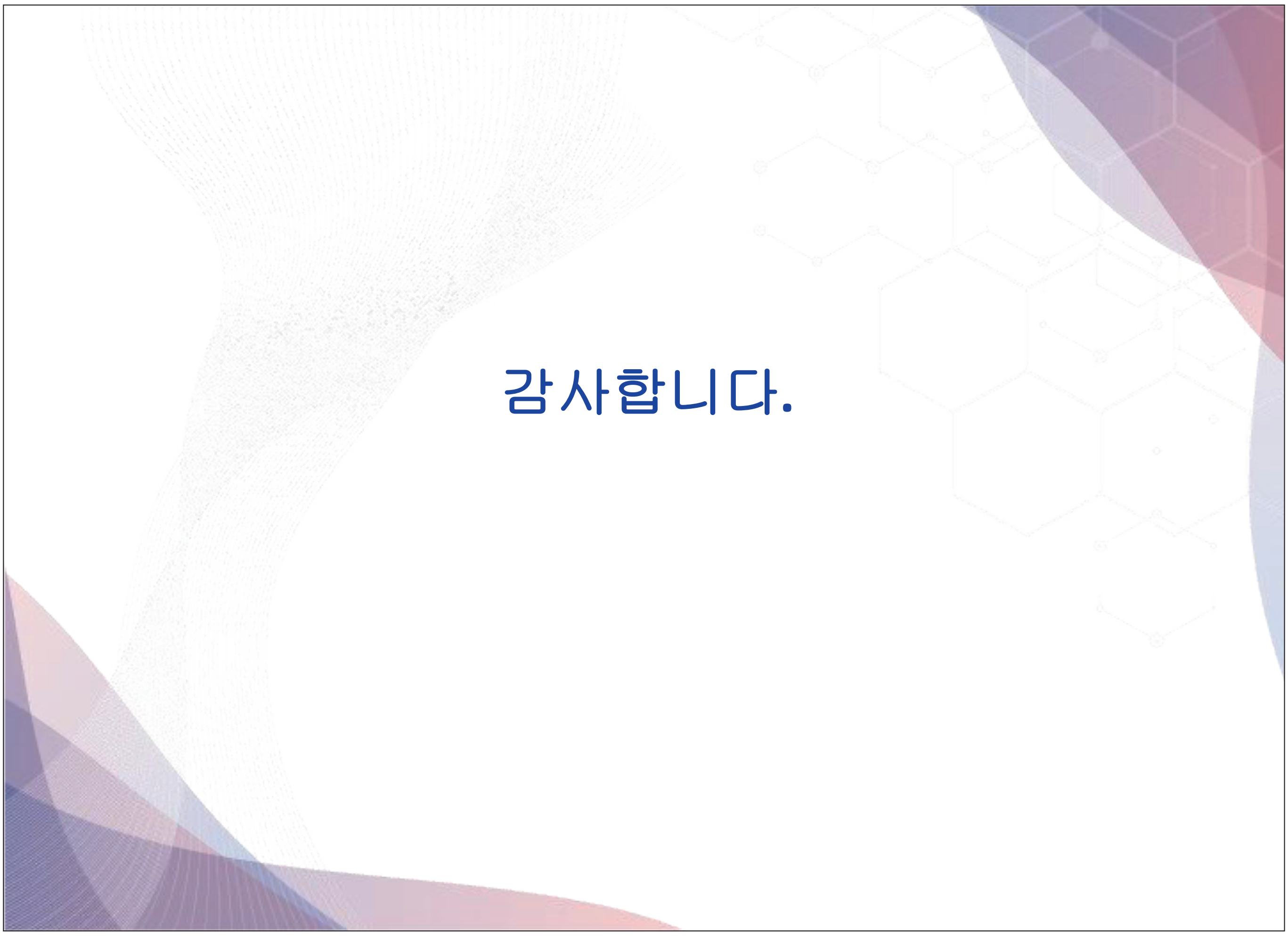
- ❑ Need to distinguish "differential marker" from the "marker" in its original meaning, i.e., it is expressed only in a certain group of cells, but not (or at least 'seldom') elsewhere.
- ❑ Marker discovery in SCODA
  - ❍ Use DEG results. But not the log fold changes.
  - ❍ Marker score = nz_pct_score = nz_pct_test × (1 – nz_pct_ref)





DEG analysis in SCODA

- ❑ Two options for DEG
  - ❍ One versus (all) others
  - ❍ One versus the reference (condition)
  - ❍ (They are the same if there are only two conditions to compare)
- ❑ SCODA perform DEG in one-versus-the-rest
- ❑ If you can specify one of the conditions as DEG reference, it gives DEG result with one-versus-the reference as well



감사합니다.