Nov 12th, 2:40 PM - 3:00 PM

Rediscovering Publicly Available Single-cell Data with the DISCO Platform

Jinmiao CHEN
*Duke-NUS and A*STAR*

Rediscovering publicly available single-cell data with the DISCO platform

Jinmiao CHEN
Duke-NUS/NUS/A*STAR

12 November 2024, 2:40 - 3:00 pm (including Q&A)

NUS Shaw Foundation Alumni House

Our vision and mission: Making data and data analysis accessible to everyone.

Our open research on single-cell biology

- DISCO: a single-cell database and knowledgebase
- DISCO-toolkit: effort-less analysis with graphical user interface
- DISCO-GPT: access & analyze data via chatting with AI agent

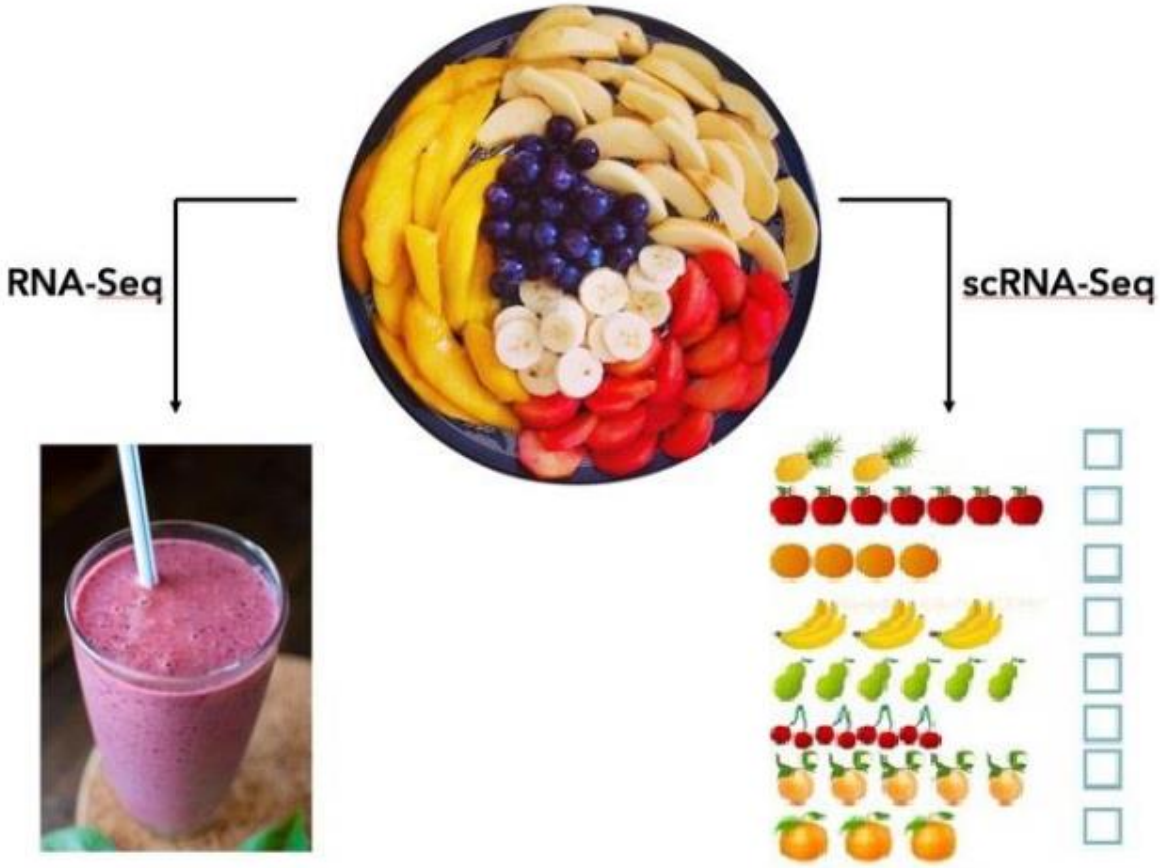
What is single-cell analysis ?

Single-cell analysis is the study of genomics, transcriptomics, proteomics, metabolomics and cell–cell interactions at the single cell level.

Adapted from wikipedia

Why single cell analysis ?

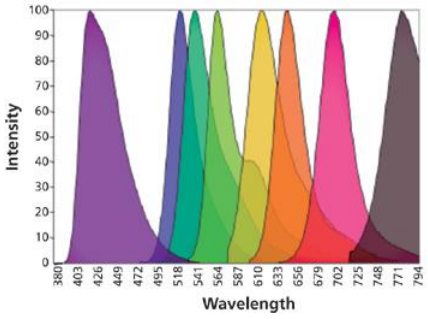
Every cell is special



In the early days of single-cell analysis

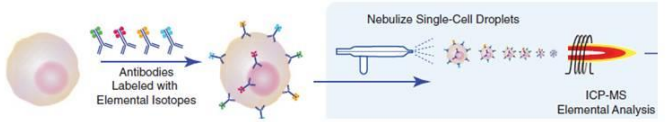
Low dimensionality, measures expression levels of multiple proteins on single cells

- Flow cytometry
  - 16-18 channels

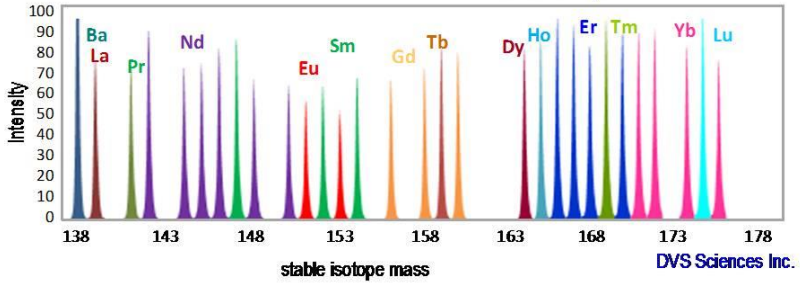


- BD FACSymphony™
  - up to 50 different characteristics per cell

- Mass cytometry
  - > 40 channels
  - Minimal crosstalk between channels



Ornatsky et al., JAAS 2008 (and others)
Bendall et al., 2011



Single cell analysis has moved towards the omics scale

- **Single-cell sequencing** (Method of the Year 2013) measures genome (scDNA-seq), transcriptome (scRNA-seq), epigenome (scATAC-seq), etc at the single cell level

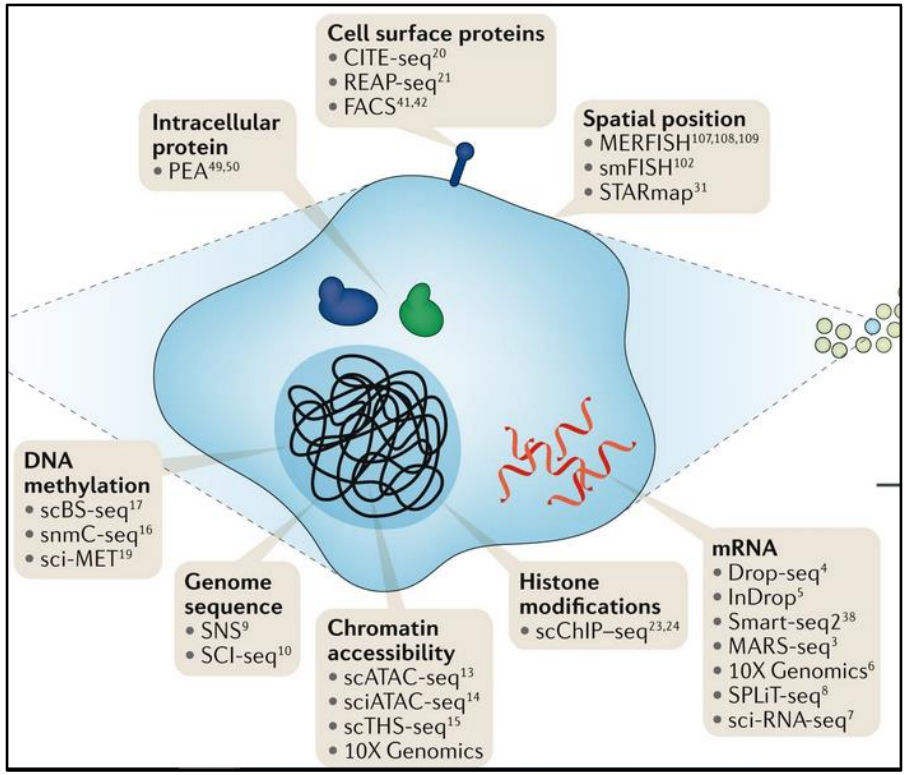
High dimensionality

- **Single-cell multimodal omics** (Method of the Year 2019) simultaneously measures multiple-omes of a cell

High dimensionality + Multi-modality

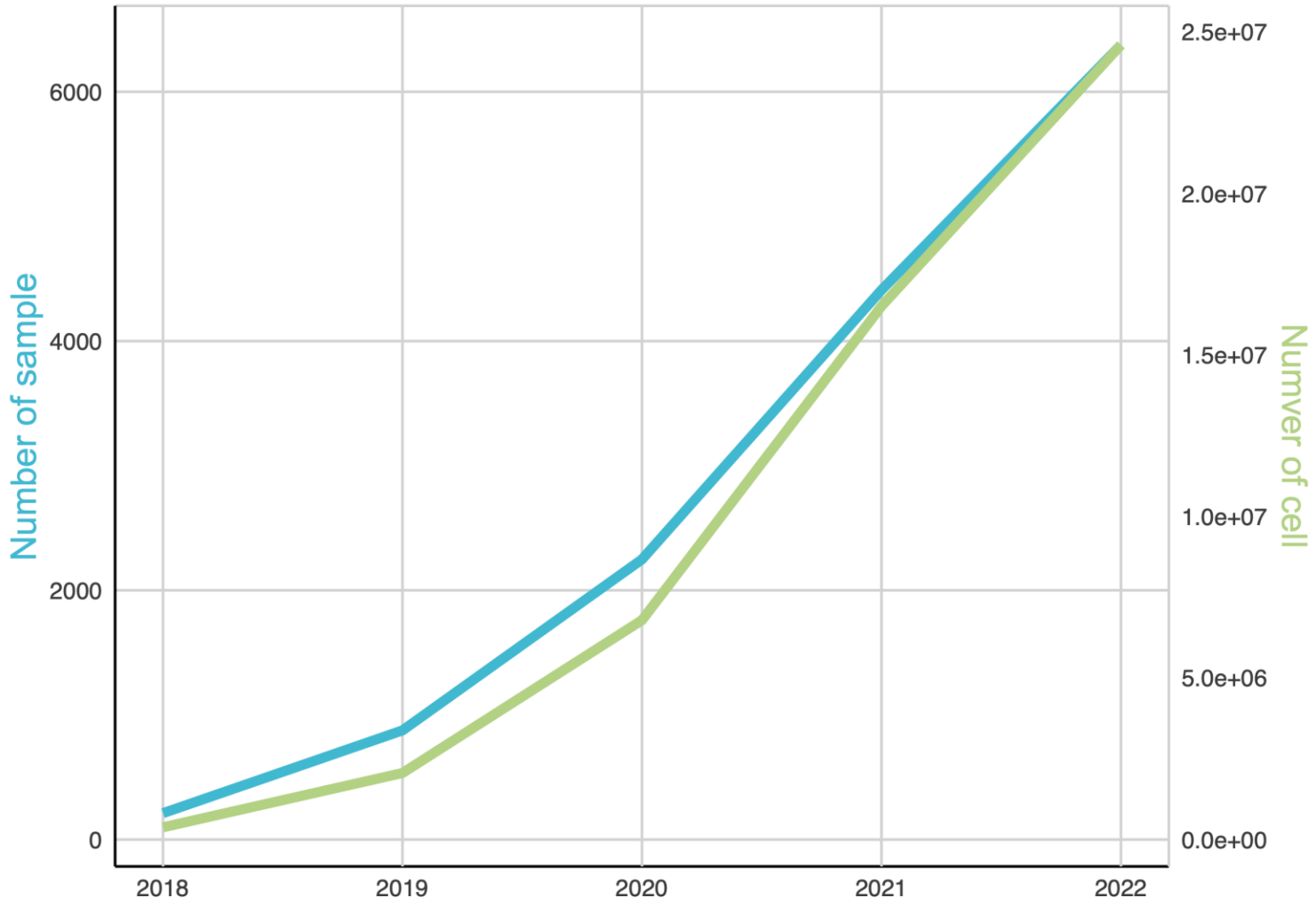
- **Spatial omics** (Method of the Year 2020) simultaneously measures gene/protein/chromatin and cell locations

High dimensionality + Multi-modality + Space



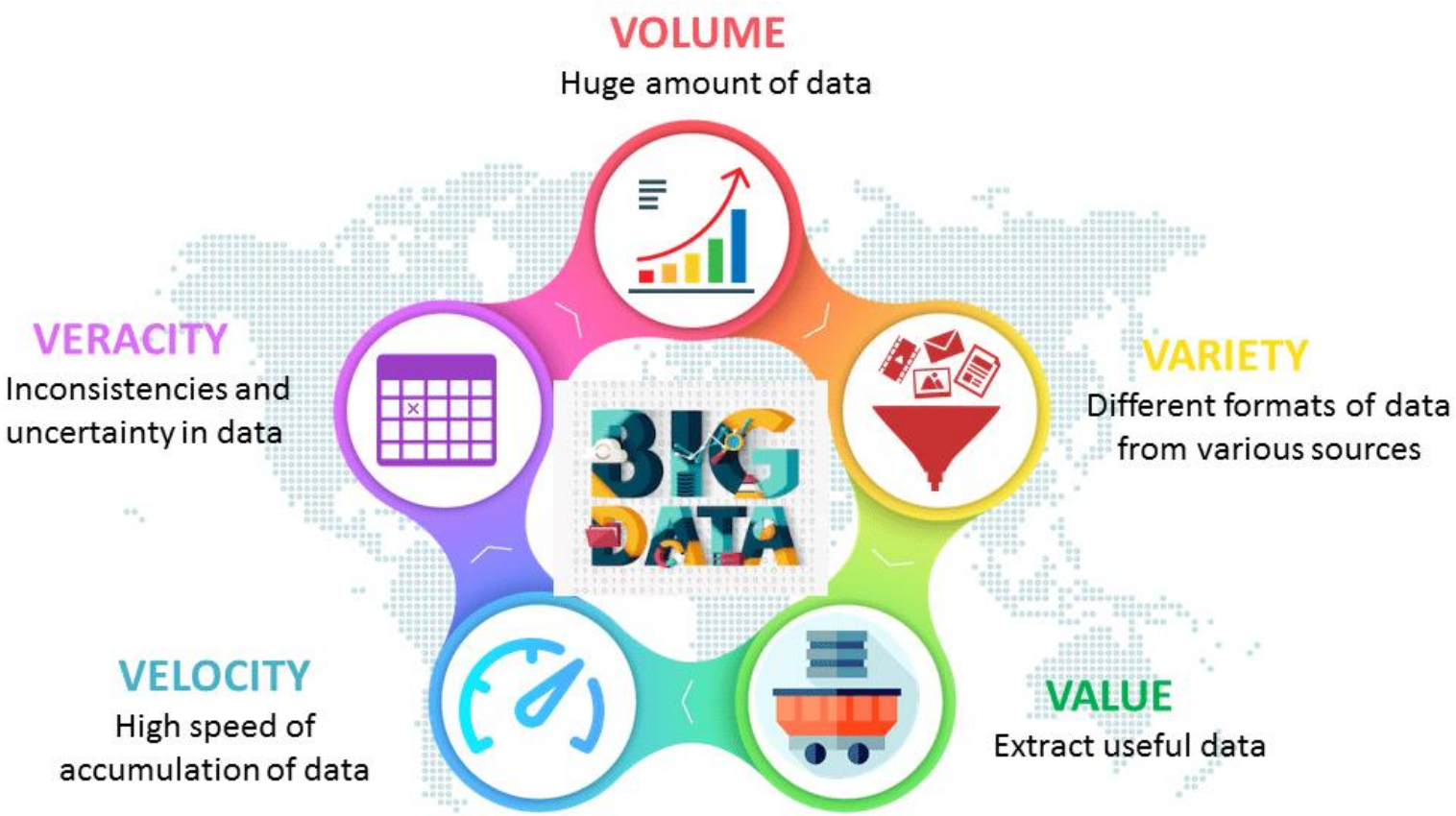
Stuart & Satija, Nat Rev Genet 2019

Exponential growth in single-cell data



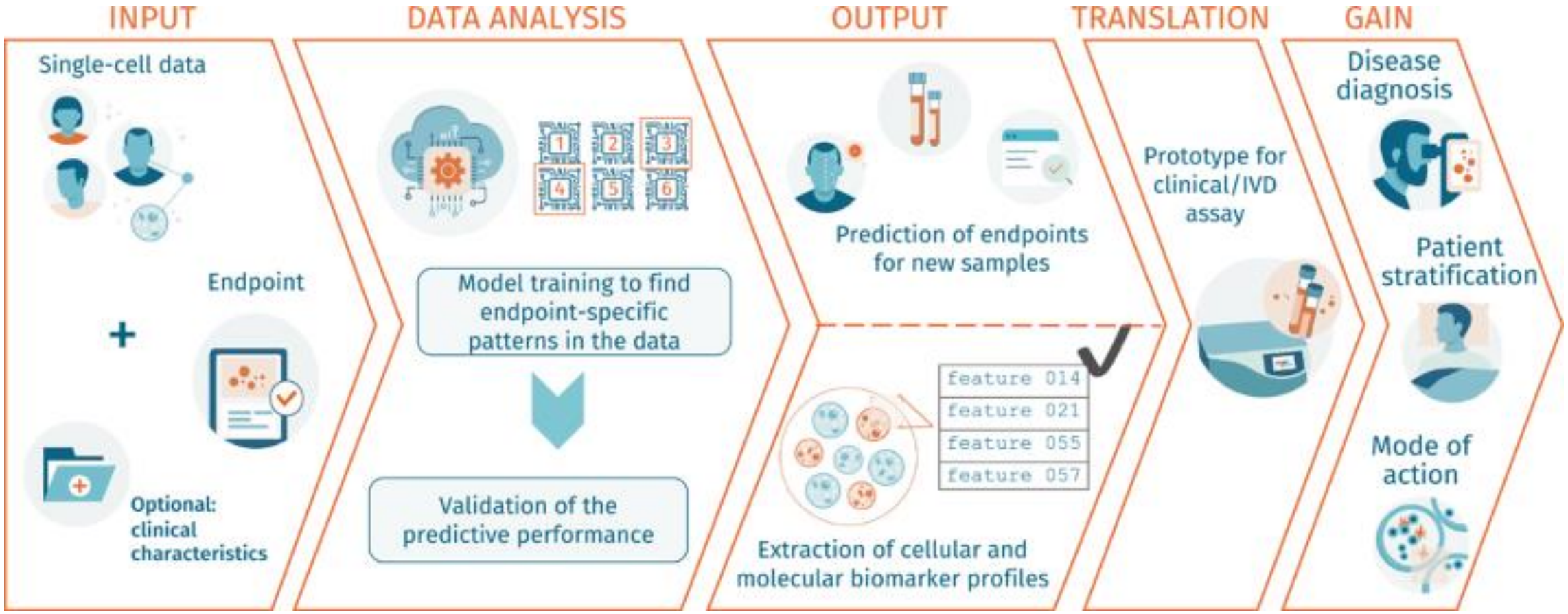
Li et al. NAR 2024

Single-cell analysis produce big data



Single-cell big data comes with big value

True precision medicine through single-cell science

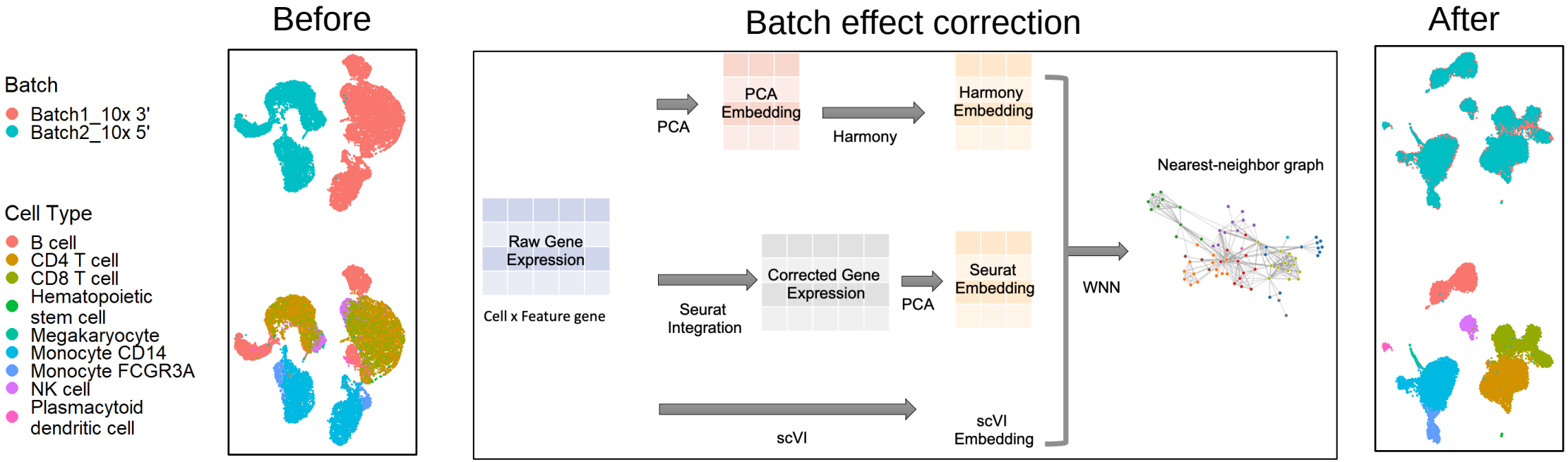


Scailyte AG, Biopharma Dealmakers

However, re-utilizing public data remains challenging

- Meta-data are not curated or harmonized
- Cell type annotation is not standardized
- **Batch effects across studies**
- Lack of user-friendly analysis tools

Batch effects



Criteria of effective batch effect correction:

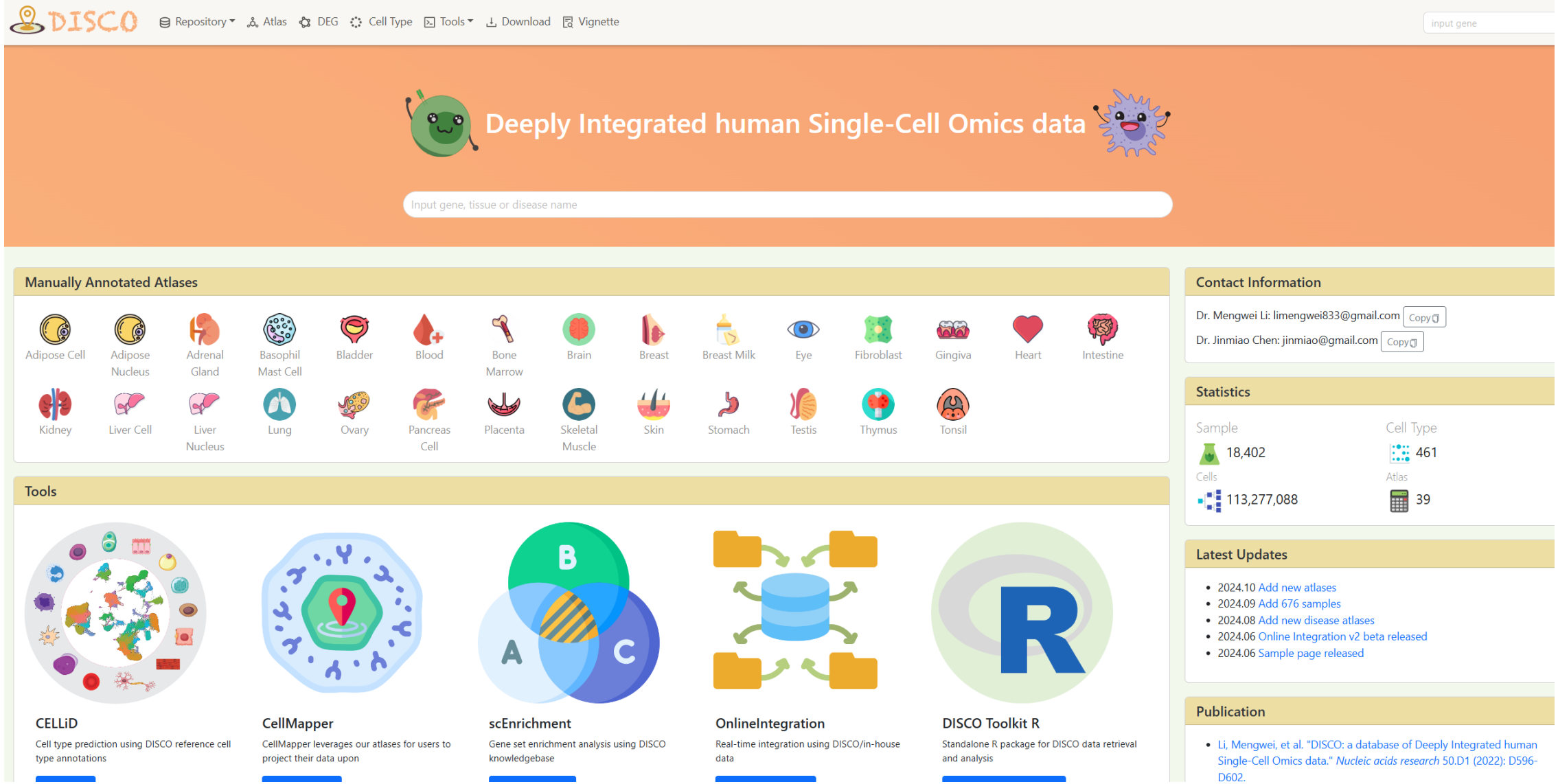
- Batches are well mixed
- Cell types are well separated

Batch effect: unwanted variations across datasets

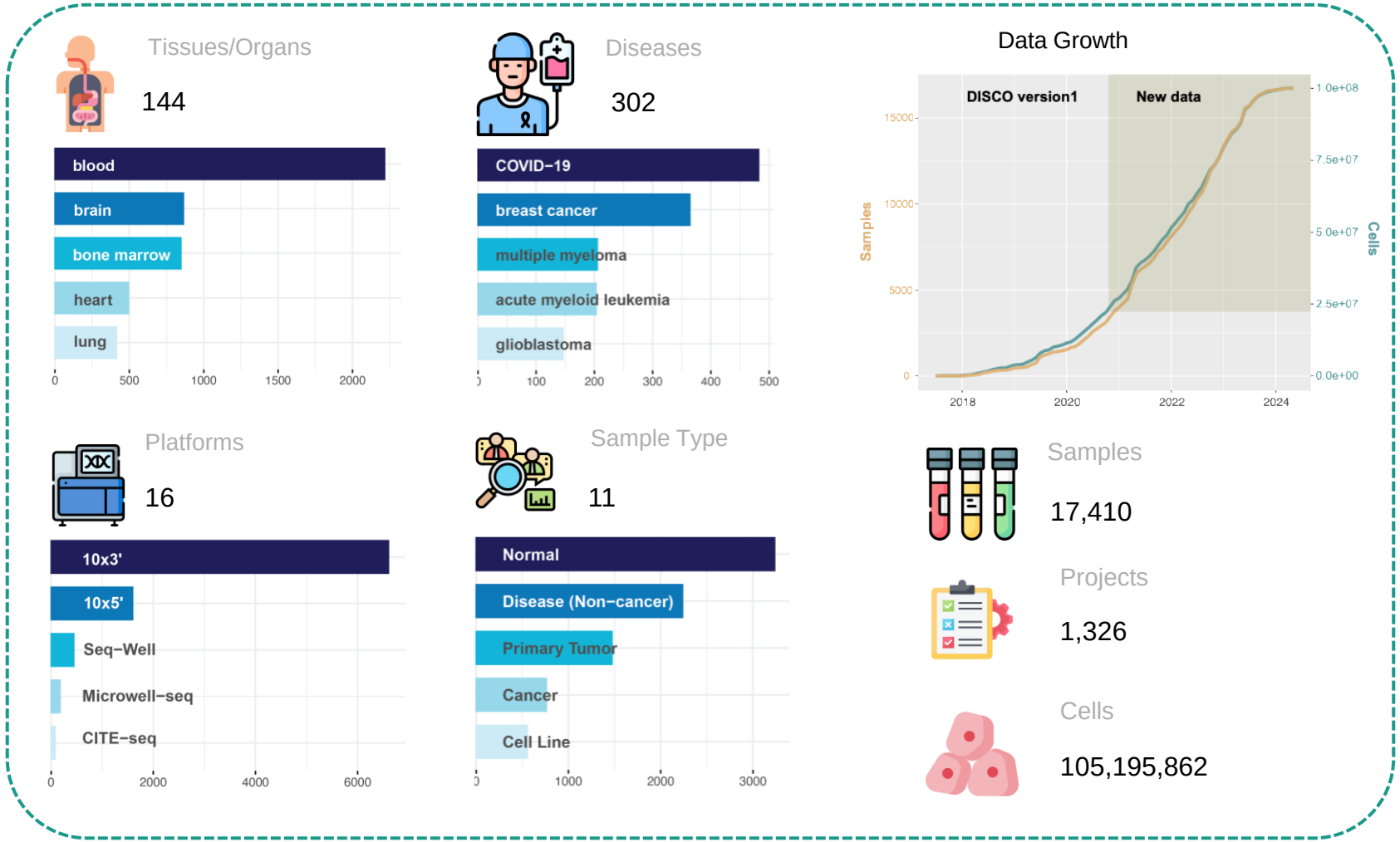
Source of batch effect: capturing times, handling personnel, reagent lots, equipments, animals, technology platforms

DISCO: Deep Integration of Single-Cell Omics

https://www.immunesinglecell.org/



DISCO is currently the largest single-cell database



Users can query, visualize, and download data

https://www.immunesinglecell.org/repository/tissue

DISCO Repository Atlas DEG Cell Type Tools Download Vignette

Filter

tissue: Please select

disease: Please select

sample type: Please select

project id: Please select

platform: Please select

rna source: Please select

Tip

* Batch download requires selecting at least 2 samples and at most 100 samples. Generating the compressed file may take some time.

* Online integration supports a maximum of 100,000 cells per integration.

Batch Download(.rds) Batch Download(.h5) Batch Download(cell type) Online Integration

	Sample ID	Project ID	Sample type	Tissue	Disease	Platform	RNA Source	#Cell	Median UMI	Others
☐	1823_BA24_10x dgcMatrix(.rds) 10Xh5(.h5) Cell type(.txt)	PRJNA434002	control	brain	control	10x3'	nucleus	981	1110	**anatomical site:** anterior cingulate cortex **subject id:** 1823 **age:** 15 **gender:** M
☐	4341_BA24_10x dgcMatrix(.rds) 10Xh5(.h5) Cell type(.txt)	PRJNA434002	control	brain	control	10x3'	nucleus	3780	1958	**anatomical site:** anterior cingulate cortex **subject id:** 4341 **age:** 13 **gender:** M
☐	4341_BA46_10x dgcMatrix(.rds) 10Xh5(.h5) Cell type(.txt)	PRJNA434002	control	brain	control	10x3'	nucleus	4243	1794	**anatomical site:** prefrontal cortex **subject id:** 4341 **age:** 13 **gender:** M
☐	4849_BA24_10x dgcMatrix(.rds) 10Xh5(.h5) Cell type(.txt)	PRJNA434002	disease tissue (non-cancer)	brain	ASD	10x3'	nucleus	4420	1411	**anatomical site:** anterior cingulate cortex **subject id:** 4849 **age:** 7 **gender:** M
☐	4899_BA24_10x dgcMatrix(.rds) 10Xh5(.h5) Cell type(.txt)	PRJNA434002	disease tissue (non-cancer)	brain	ASD	10x3'	nucleus	2707	2715	**anatomical site:** anterior cingulate cortex **subject id:** 4899 **age:** 14 **gender:** M

DISCO is also a knowledge base

- Integrated cell atlases and associated knowledge
- Cell type reference & ontology
- Differentially expressed genes (DEGs)

Integrated atlases for specific tissues, diseases, or cell types

Tissue Atlas

Adipose	cell	36 cell types, 190K cells
Adipose	nucleus	12 cell types, 34K cells
Adrenal gland	cell	10 cell types, 41K cells
Bladder	cell	36 cell types, 41K cells
Blood	cell	25 cell types, 170K cells
Bone marrow	cell	45 cell types, 674K cells
Brain	nucleus	21 cell types, 286K cells
Breast	cell	39 cell types, 175K cells
Breast milk	cell	18 cell types, 108K cells
Eye	cell	26 cell types, 144K cells
Gingiva	cell	34 cell types, 40K cells
Heart	cell/nucleus	29 cell types, 624K cells
Intestine	cell	52 cell types, 414K cells
Kidney	cell/nucleus	25 cell types, 104K cells
liver	cell	45 cell types, 368K cells
Liver	nucleus	19 cell types, 104K cells
Lung	cell	45 cell types, 221K cells
Ovary	cell	44 cell types, 124K cells
Pancreas	cell	36 cell types, 188K cells
Placenta	cell	33 cell types, 91K cells
Skeletal muscle	cell	30 cell types, 68K cells
Skin	cell	40 cell types, 395K cells
Stomach	cell	29 cell types, 31K cells
Testis	cell	40 cell types, 159K cells
Thymus	cell	49 cell types, 422K cells
Tonsil	cell	39 cell types, 274K cells

Integrated atlases for specific tissues, diseases, or cell types

Disease Atlas

Alzheimer's disease — nucleus
Parenchyma, frontal cortex
21 cell types, 98K cells

COVID-19 — cell
Blood
19 cell types, 284K cells

Crohn's Disease — cell
Ileum
23 cell types, 69K cells

Dengue fever — cell
Blood
16 cell types, 134K cells

HIV infection — cell
Blood
18 cell types, 48K cells

HIV infection — cell
Cerebrospinal fluid
13 cell types, 46K cells

HNSCC — cell
Blood
17 cell types, 54K cells

PDAC — cell
Pancreas
47 cell types, 168K cells

Sarcoidosis — cell
Blood
21 cell types, 57K cells

Type 1 diabetes — cell
Pancreas
12 cell types, 30K cells

Type 2 diabetes — cell
Pancreas
13 cell types, 132K cells

Cell Type Atlas

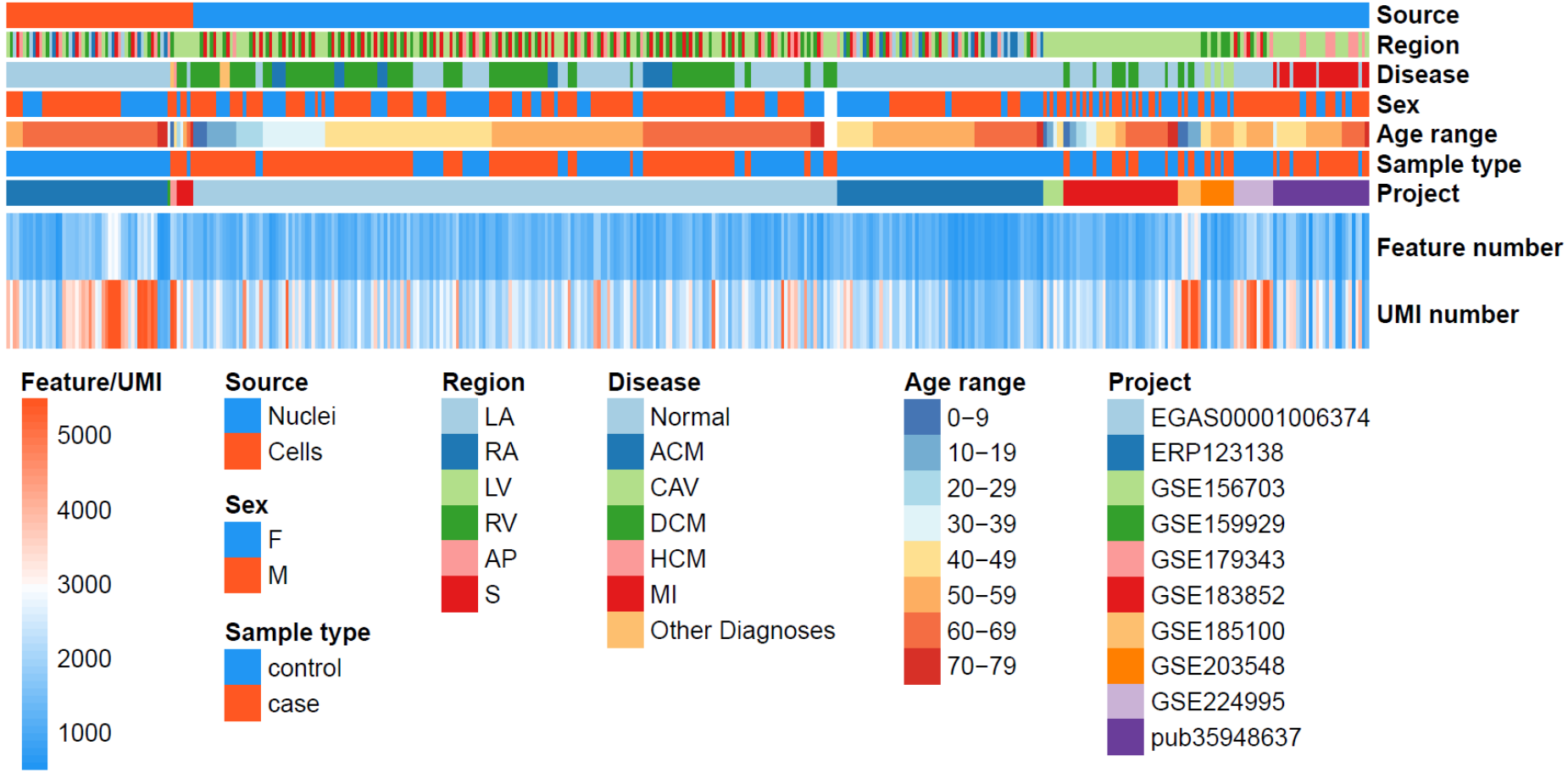
Basophil, Mast cell — cell
5 cell types, 157K cells

Fibroblast — cell
20 cell types, 473K cells

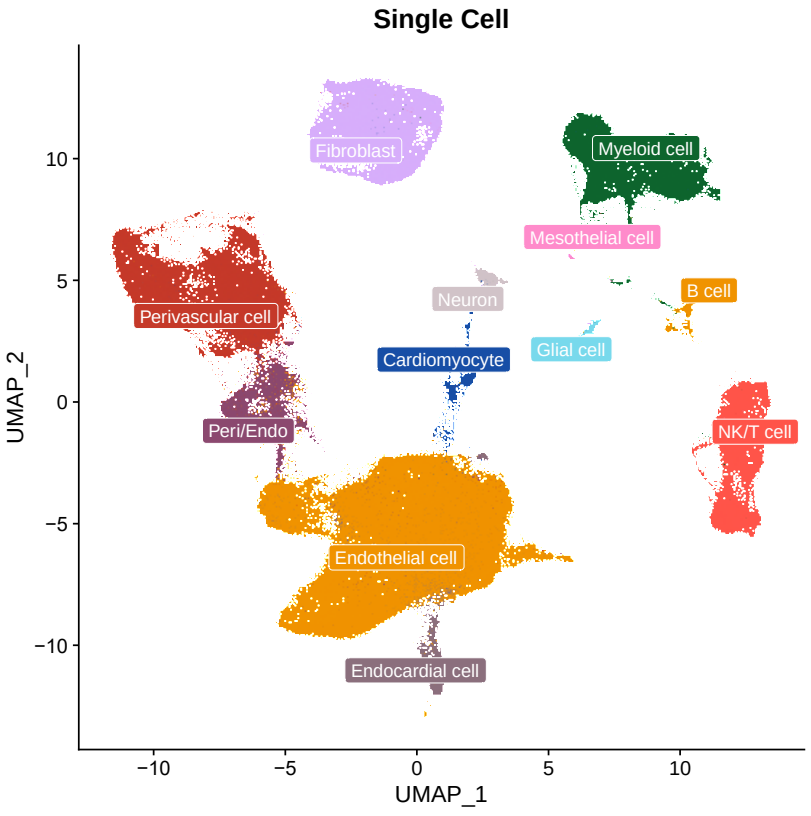
DISCO's collection of human heart data

DISCO data repository

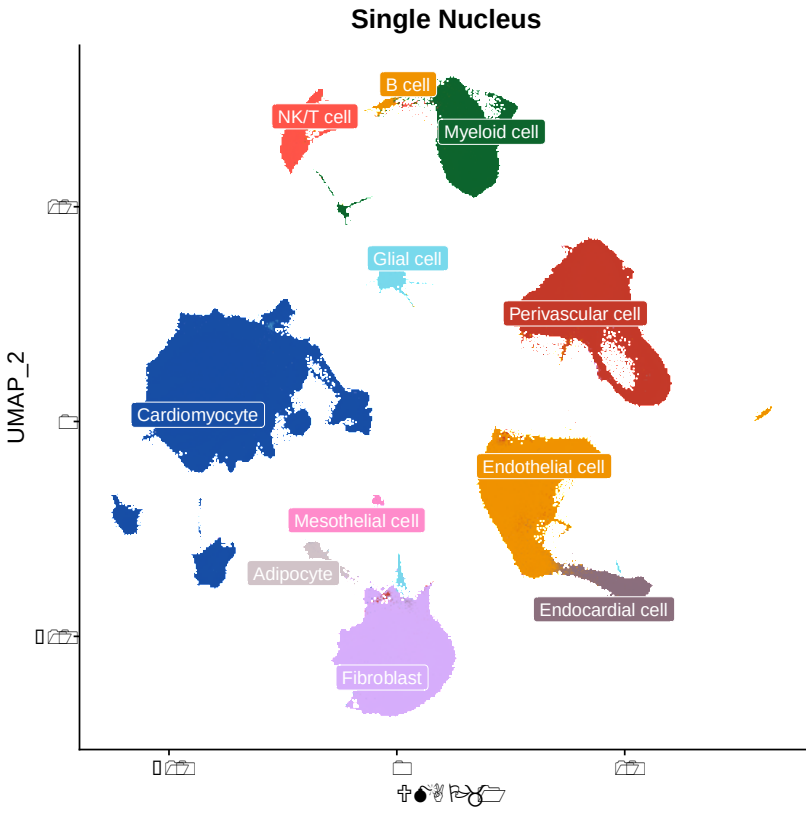
heart



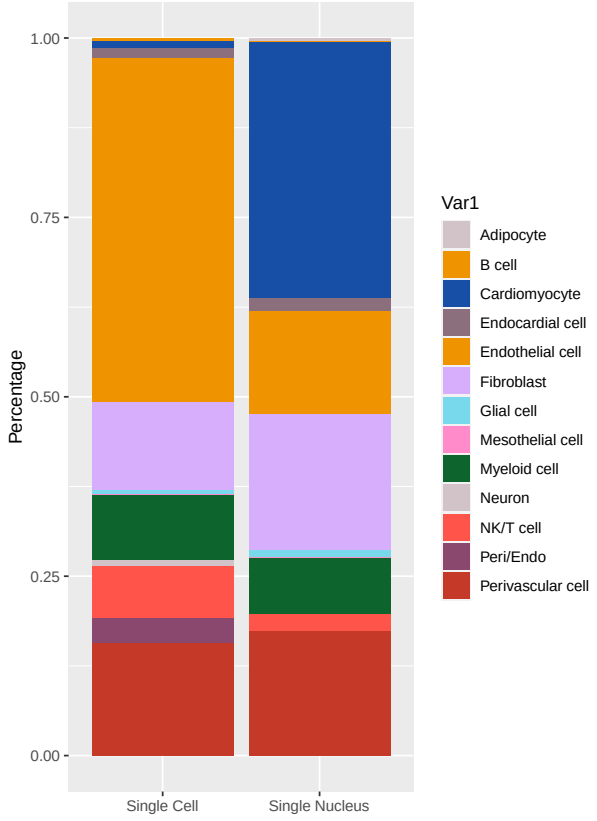
Heart atlas



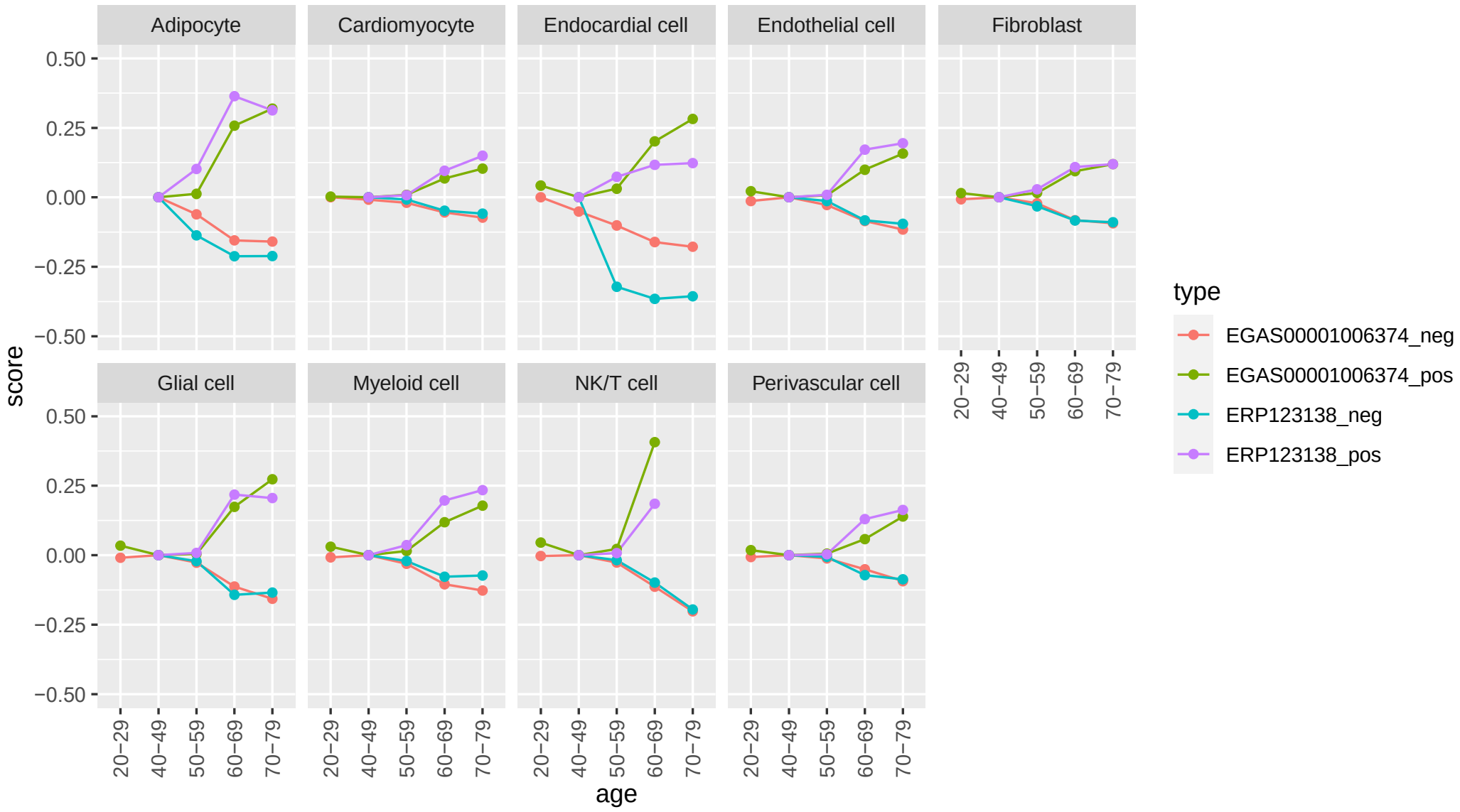
214,928 cells, 57 samples



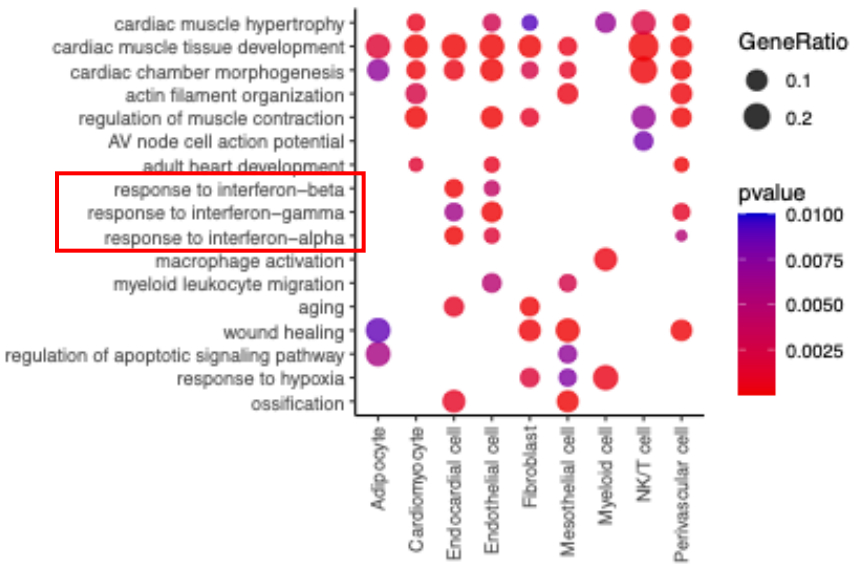
1,445,941 cells, 331 samples

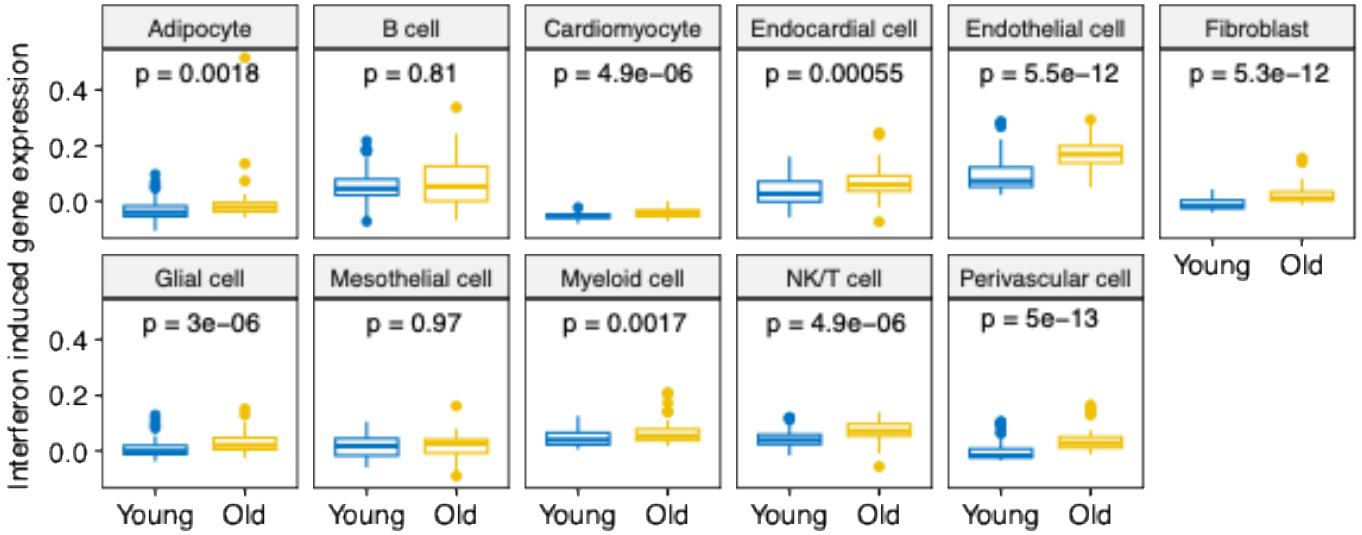


Substantial gene expression changes occur at the age of 60



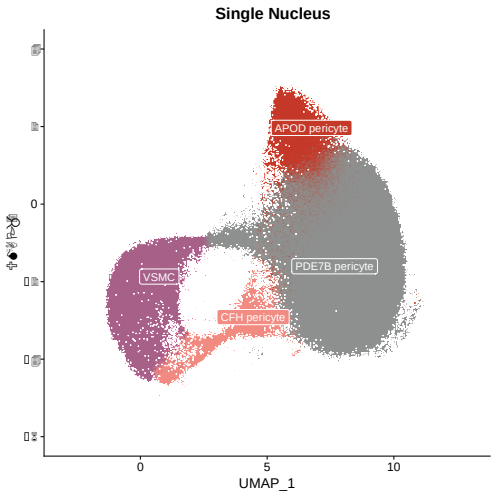
Aging-related DEGs & pathways: response to interferon increases with age

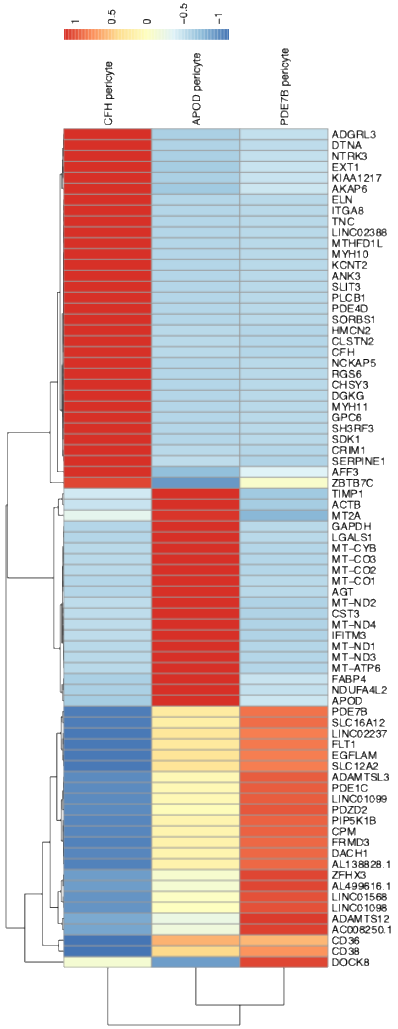


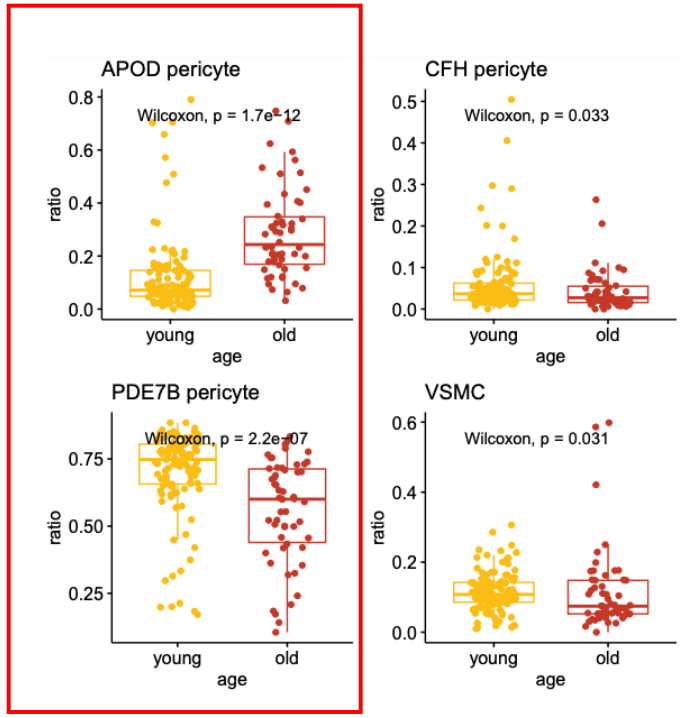


Yong: <60 Old: >=60

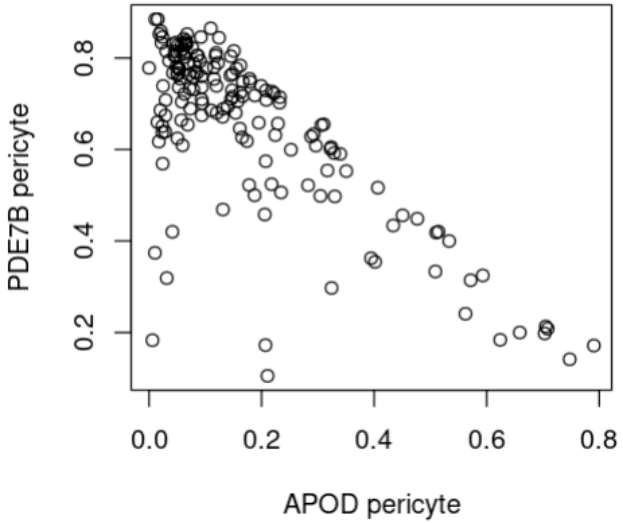
Three subsets of pericytes, APOD pericytes increase with age







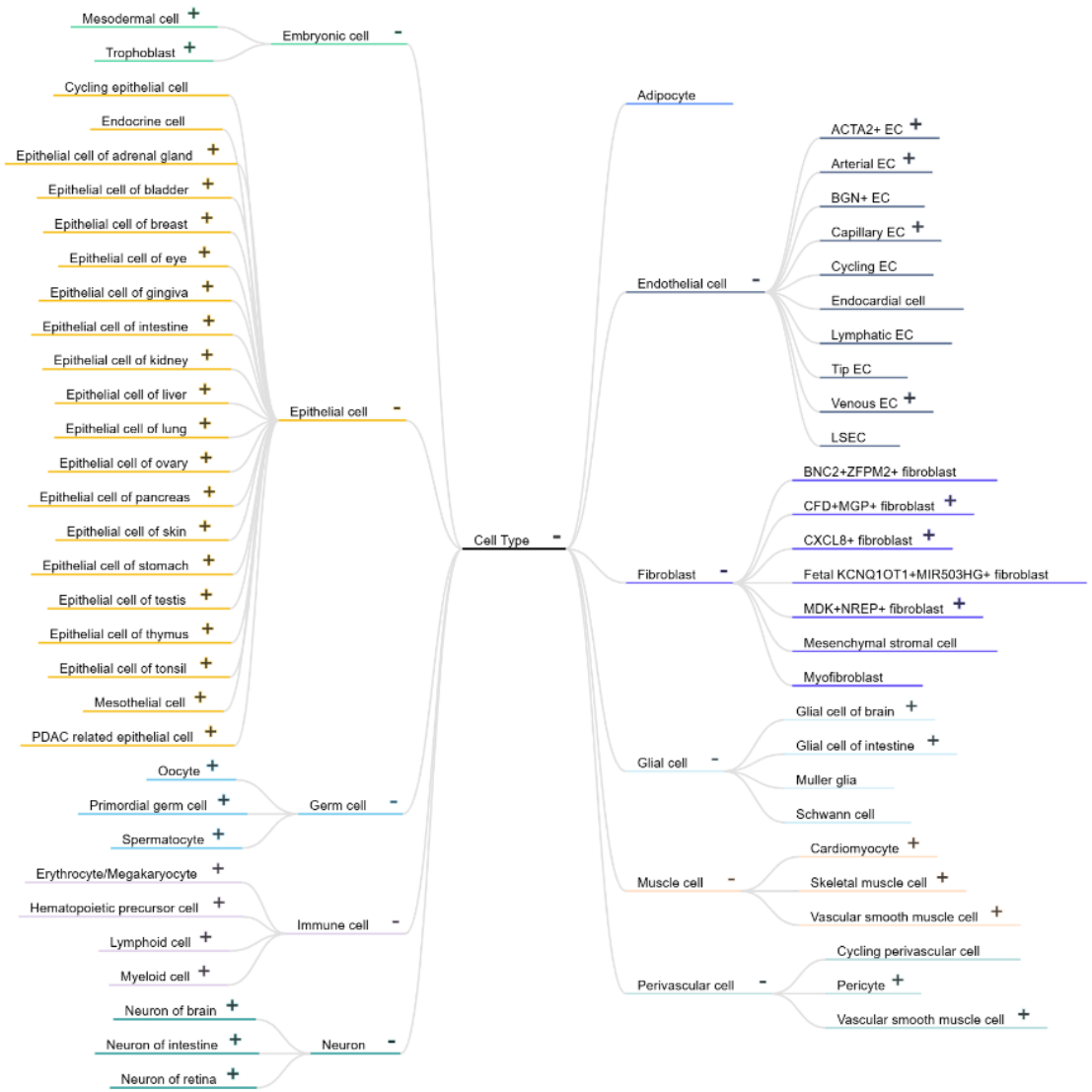
Yong: <60 Old: >=60



A single-cell enhanced knowledgebase of cell type reference & ontology

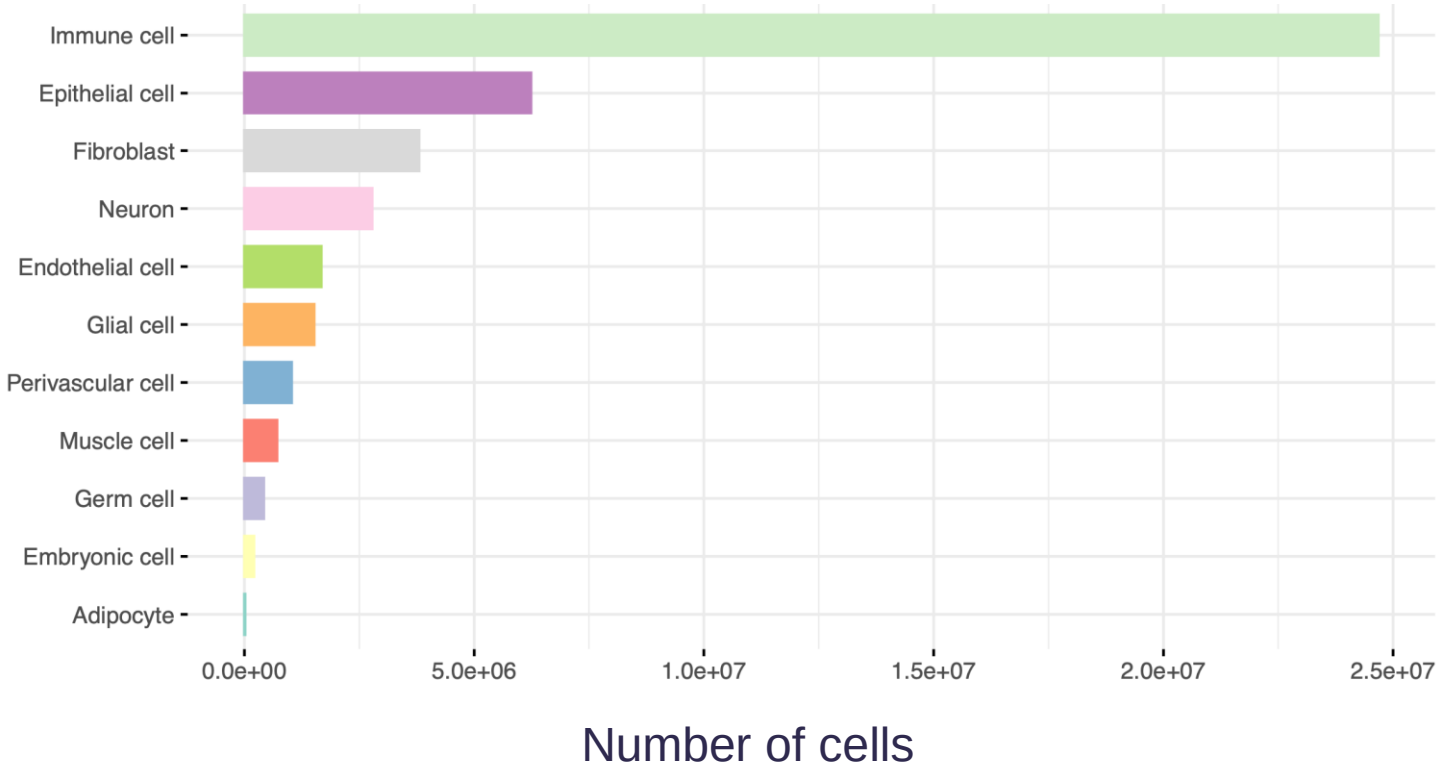
463 cell types https://www.immunesinglecell.org/cell_type

Cell Type	Alias	Parent	Marker (curated)
Enter Cell Type...	Enter Alias...	Enter Parent...	Enter Marker (curated)...
abT (entry) cell	abT (entry) cell	T cell	SATB1 CD1A CCR9 CD1E CD1C
Acinar cell	Acinar cell	Epithelial cell of pancreas	PRSS1 CPA1 CTRB1
ACTA2+ arterial EC	ACTA2+ arterial EC	ACTA2+ EC	ACTA2 RAMP2 TAGLN NEBL
ACTA2+ capillary EC	ACTA2+ capillary EC	ACTA2+ EC	ACTA2 RAMP2 TAGLN AQP1
ACTA2+ EC	ACTA2+ EC ACTA2+ endothelial cell	Endothelial cell	RAMP2 PECAM1 ACTA2
ACTG2+ contractile pericyte	ACTG2+ contractile pericyte	Pericyte	ACTA2 RGS5 ACTG2
ACTG2+ contractile VSMC	ACTG2+ contractile VSMC	Vascular smooth muscle cell	ACTA2 PLN ACTG2
ADAM12+ fibroblast	ADAM12+ fibroblast	CFD+MGP+ fibroblast	DCN LUM FN1 ADAM12
ADAMDEC1+ADAM28+ fibroblast	ADAMDEC1+ADAM28+ fibroblast	MDK+NREP+ fibroblast	DCN LUM ADAMDEC1 ADAM28
Adipocyte	Adipocyte Lipocyte Fat cell	Cell Type	ADIPOQ PNPLA2



Differentially expressed genes (DEGs)

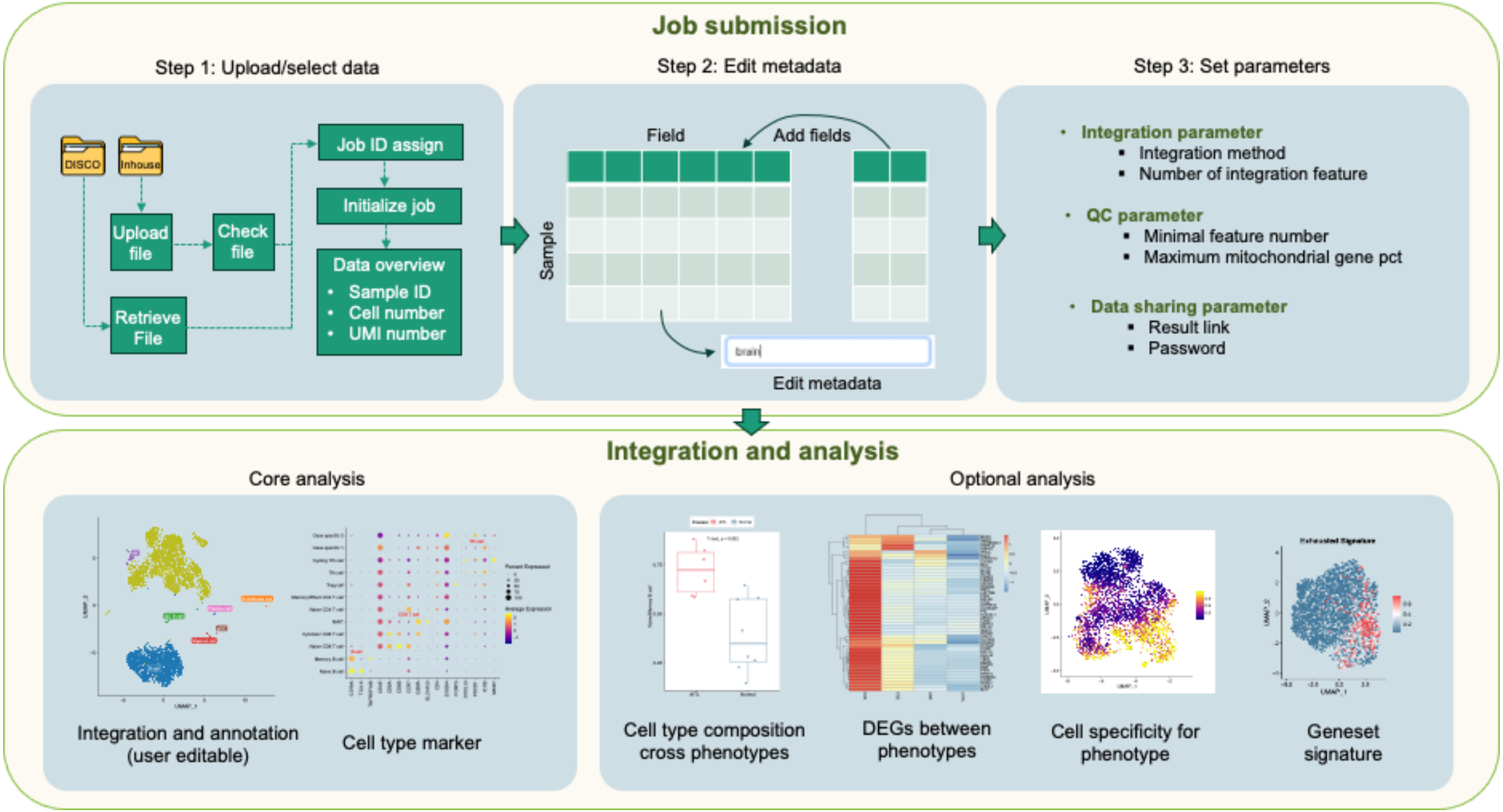
Cell type DEGs	Phenotype DEGs	Gene sets
329,104	18,372	1,268



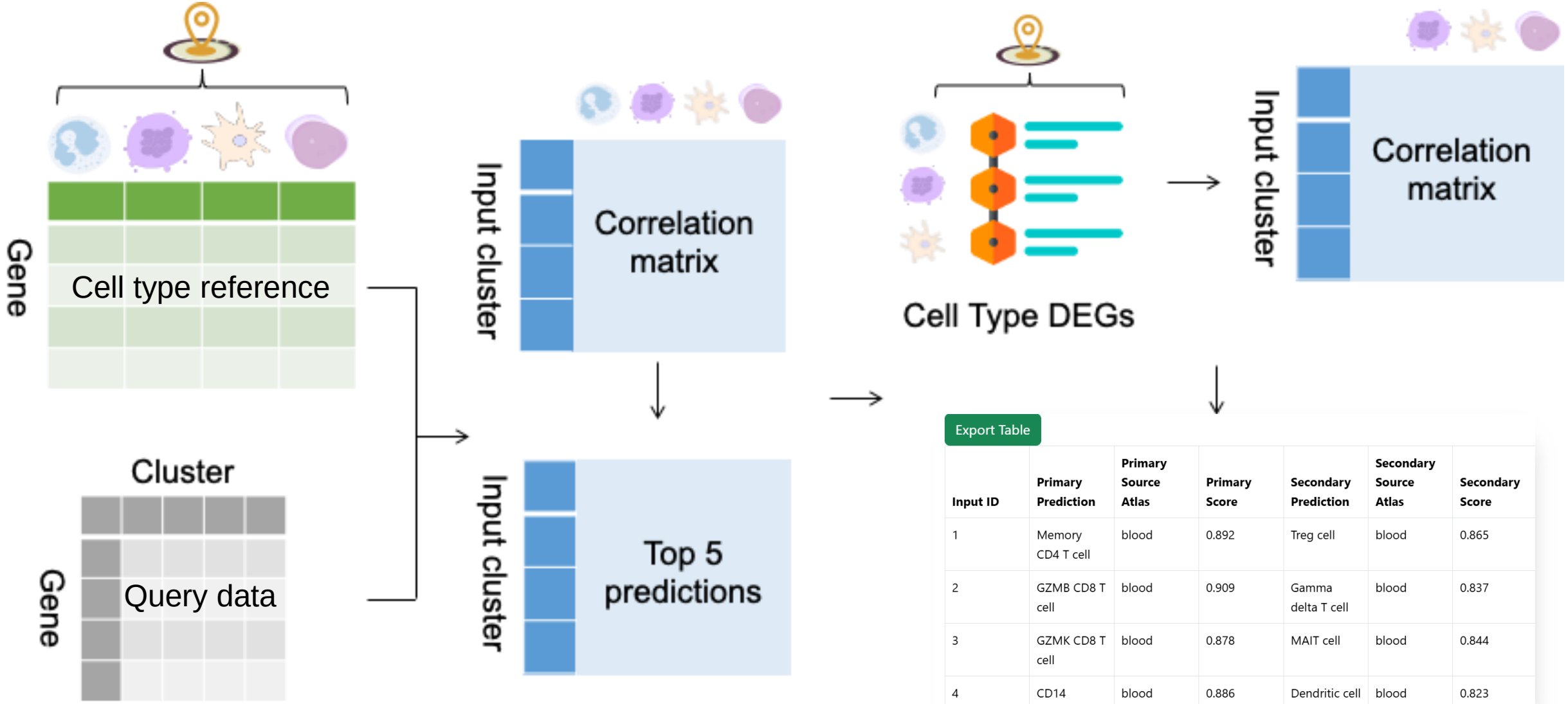
Our open research on single-cell biology

- **DISCO:** a database and knowledgebase
- **DISCO-toolkit:** effort-less analysis with graphical user interface
- **DISCO-GPT:** access & analyze data via chatting with AI agent

Online data integration



CELLiD (cell type annotation)

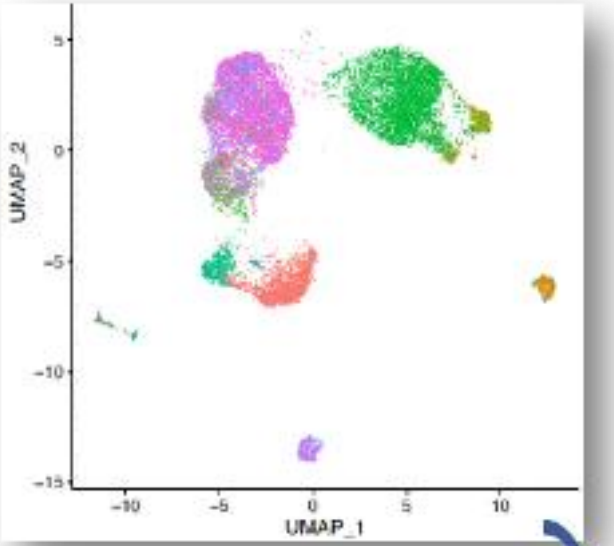


Export Table

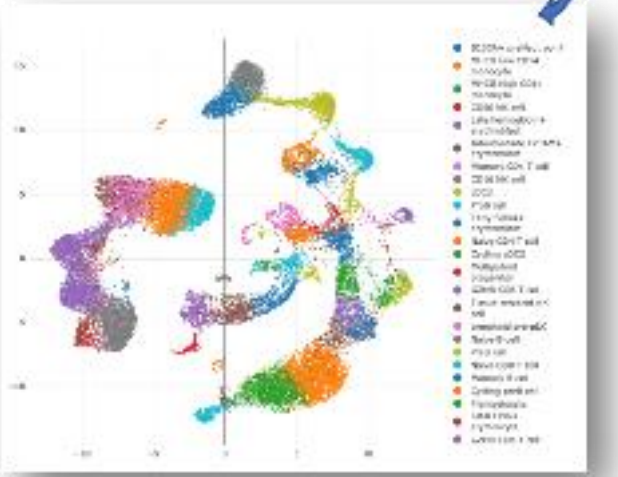
Input ID	Primary Prediction	Primary Source Atlas	Primary Score	Secondary Prediction	Secondary Source Atlas	Secondary Score
1	Memory CD4 T cell	blood	0.892	Treg cell	blood	0.865
2	GZMB CD8 T cell	blood	0.909	Gamma delta T cell	blood	0.837
3	GZMK CD8 T cell	blood	0.878	MAIT cell	blood	0.844
4	CD14 monocyte	blood	0.886	Dendritic cell	blood	0.823
5	CD16 NK cell	blood	0.882	CD56 NK cell	blood	0.825
6	Memory B cell	blood	0.865	Naive B cell	blood	0.831

CellMapper (project query data to reference atlas)

Input data

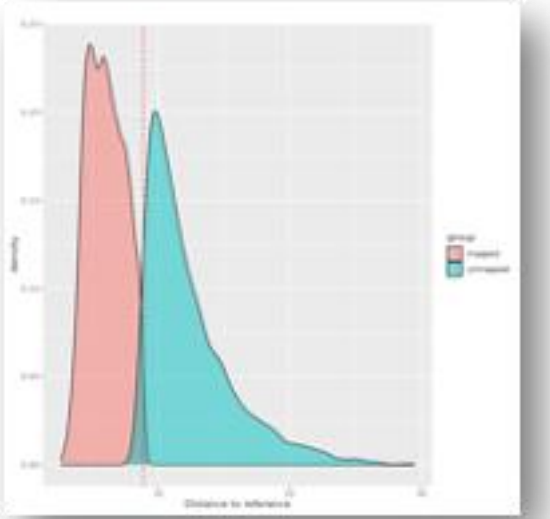


Reference data

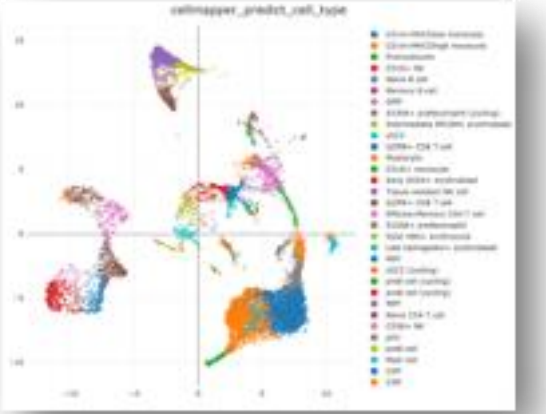


Projection

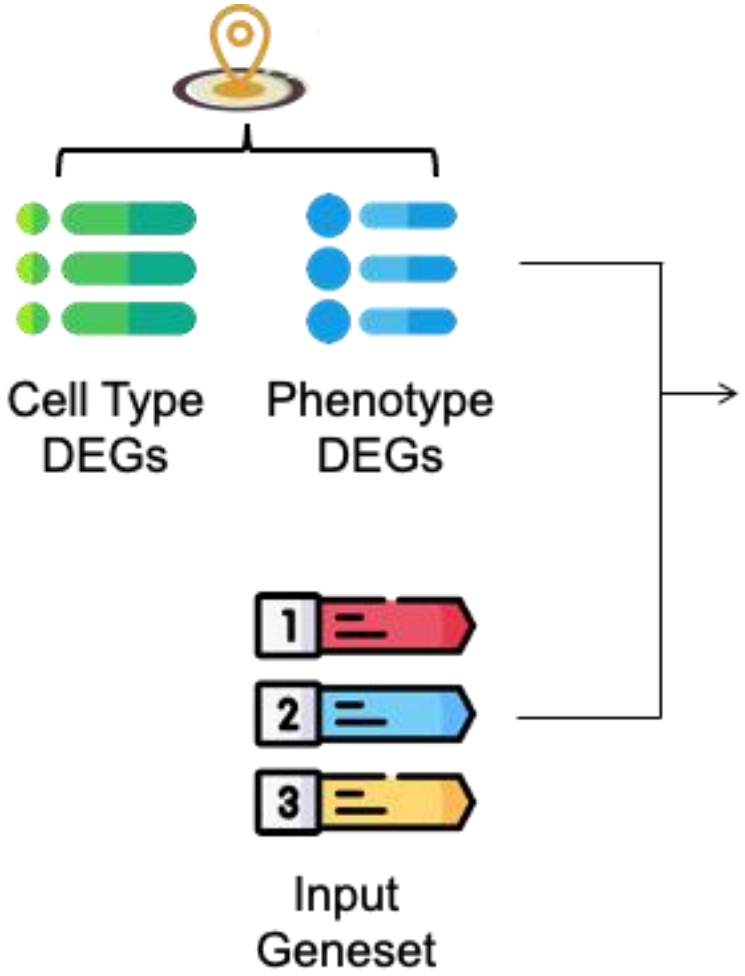
Tumor cell identification

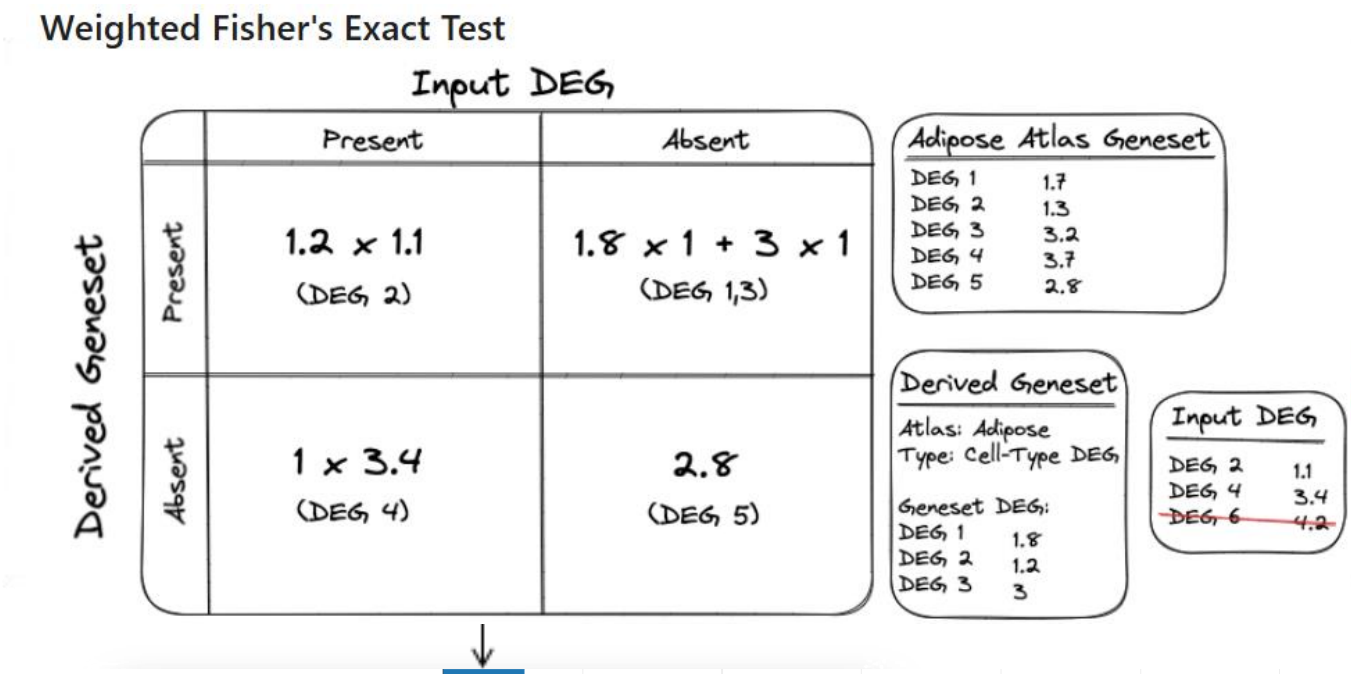


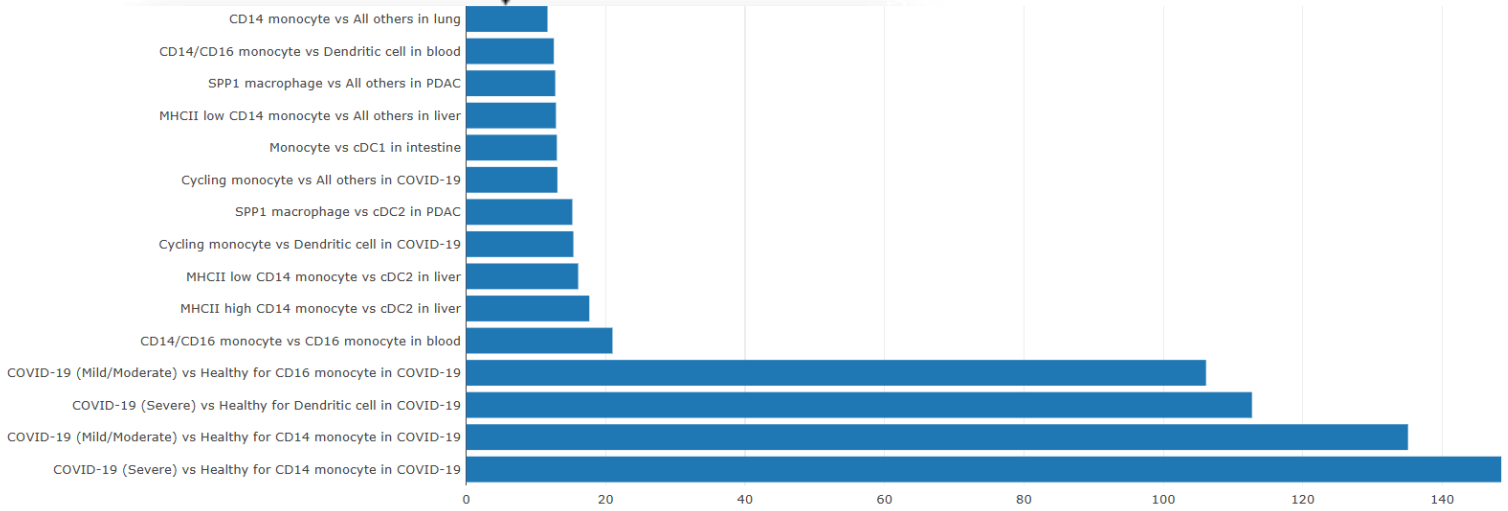
Predicted cell type



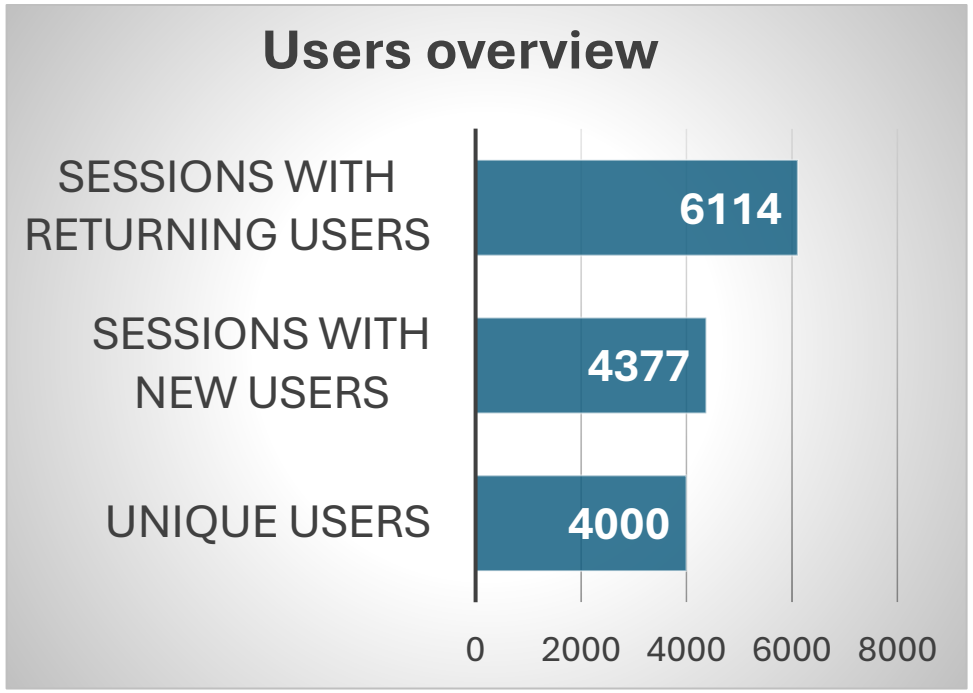
ScEnrichment (gene set enrichment for interpretation)

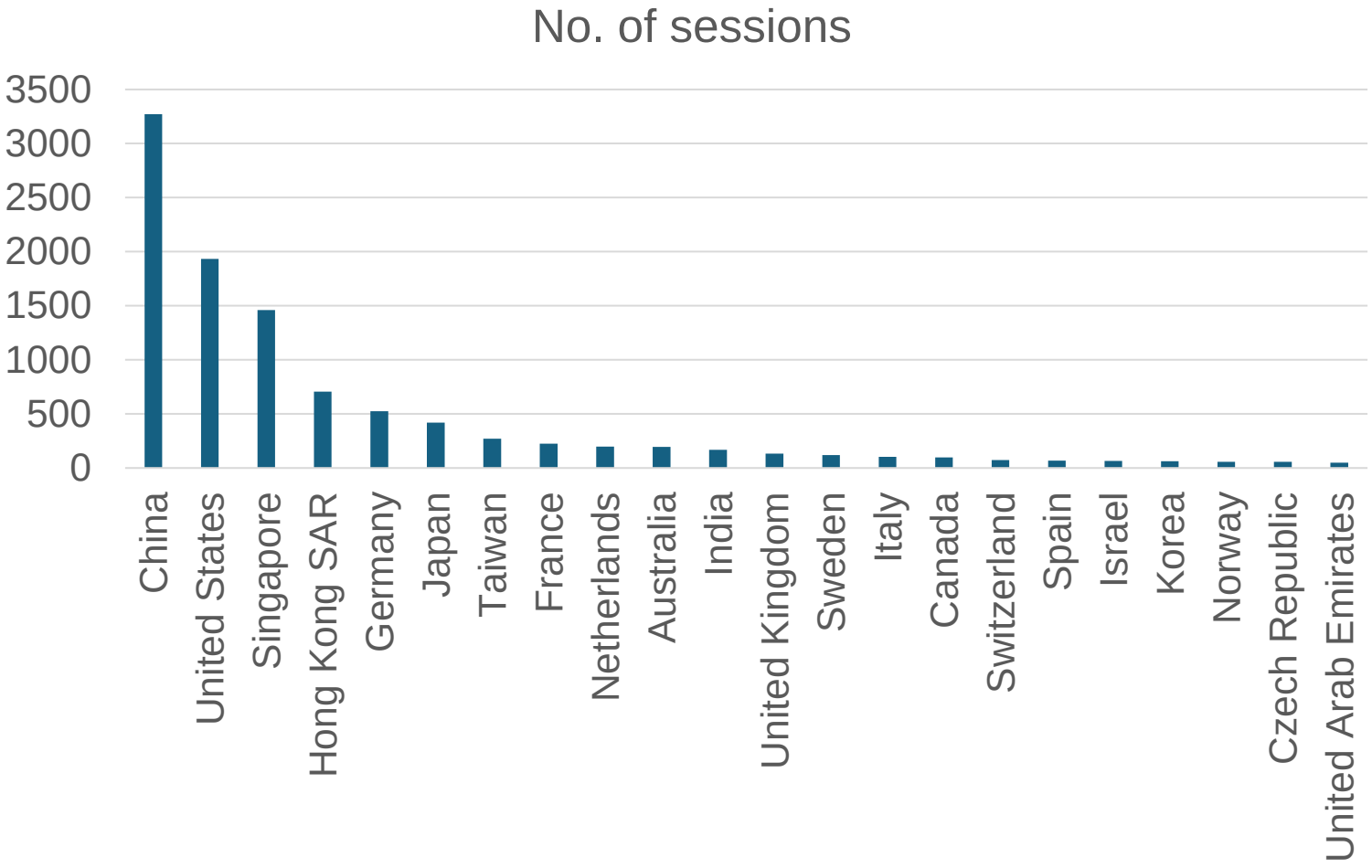




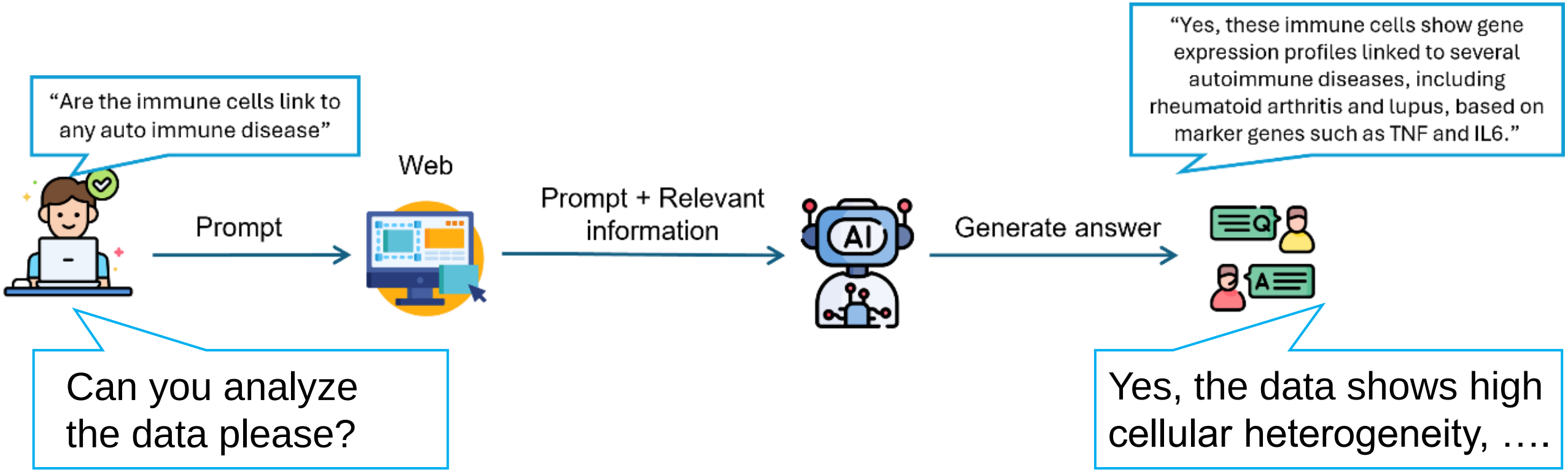


Usage for the past 90 days





DISCO-GPT: access & analyze data via chatting to AI agent (ongoing)



Challenges encountered in maintaining & upgrading DISCO

- Resource:
  - Data server
  - Computing server
  - Hosting server
- Manpower

Acknowledgements

Lab members
Yahui Long
Chengwei Zhong
Hang Xu
Kok Siong Ang
Raman Sethi
Mengwei Li
Nicole Lee
others

Duke-NUS
Valerie Chew

NNI
Li Zeng

NUS
Nicholas Gascoigne
Lina Lim

SIgN
Florent Ginhoux
Lai Guan Ng

IMCB
Vinay TERGAONKAR
Jonathan LOH

I2R
Min Wu
Xiaoli Li

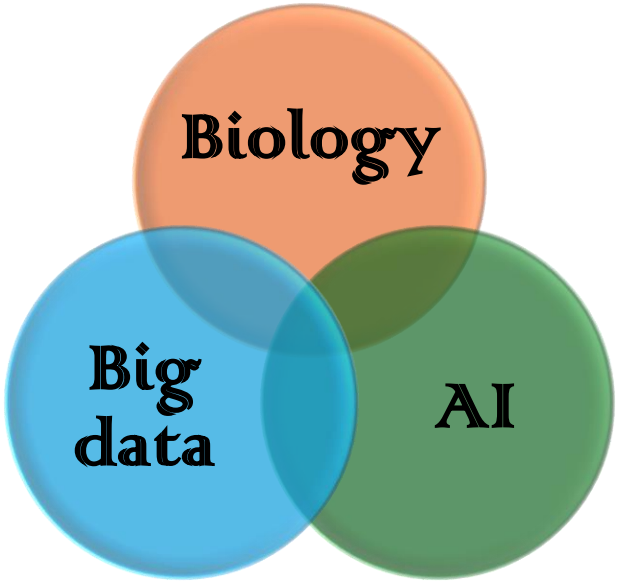
IHPC
Huazhu Fu
Yong Liu

Yale University
Rong Fan

Northwestern University
David Gate

BGI
Longqi Liu
Min Jian
Ao Chen
Xun Xu

A lot of fun being a computational biologist



We are recruiting

- PhD students
- Research assistants
- Postdocs

- jinmiao@gmail.com
- micchenj@nus.edu.sg
- chen_jinmiao@bii.a-star.edu.sg