



Generative AI for scRNA-seq data analysis

Jiadong Mao, Xiaochen Zhang

(Lê Cao Lab)

Sandeep Santhosh Kumar

(Shim Lab)

Melbourne Integrative Genomics
(MIG)

Generative AI for scRNA-seq data analysis



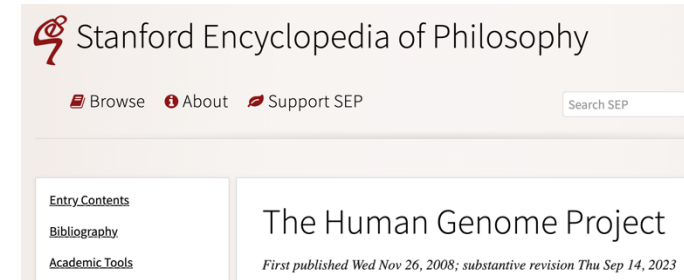
Why RNA?

Why not just DNA?

- The Human Genome Project (HGP) and the cure of cancer
 - Sequence 3 billion base pairs of A, G, C, T
 - ‘(HGP will) enable most individuals to live a ... life without disease.’
 - ‘By 2010 individualised medicine would be a reality; physicians would routinely take check swabs from patients and send their DNA out for testing’
- This, alas, did not happen
 - Genome is a book too complicated to decipher
 - Eg E coli genome and gene regulation
- Multi-cellular organisms
 - ‘Cell types’

Genome: the complete set of genes or genetic materials present in a cell or organism.

Genomics: the branch of molecular biology concerned with the structure, function, evolution and mapping of genomes



Cell

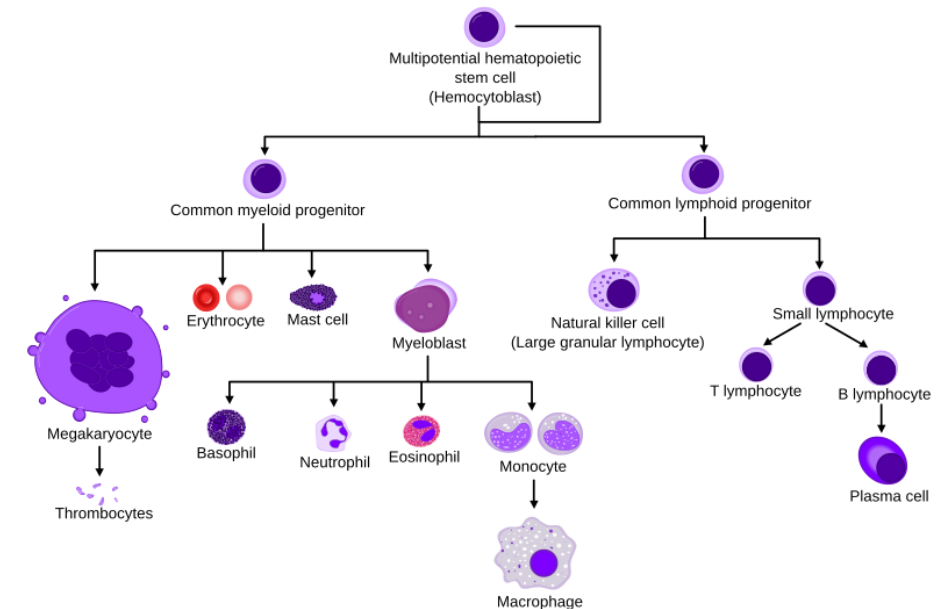
Leading Edge

Commentary The cellular dogma

Stephen R. Quake^{1,2,*}

¹The Chan Zuckerberg Initiative, Redwood City, CA, USA

²Depts of Bioengineering and Applied Physics, Stanford University, Stanford, CA, USA



Haematopoiesis. Wikipedia

From genotype to phenotype

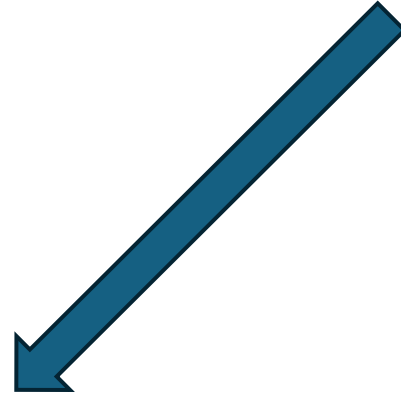
- Crick's central dogma
 - Information flow at molecular level
 - Genomics, epigenomics, **transcriptomics**, proteomics, ...
- mRNA and cell types
- From Human Genome Project to Human Cell Atlas
 - Single-cell omics



Genotype: the genetic constitution of an individual organism

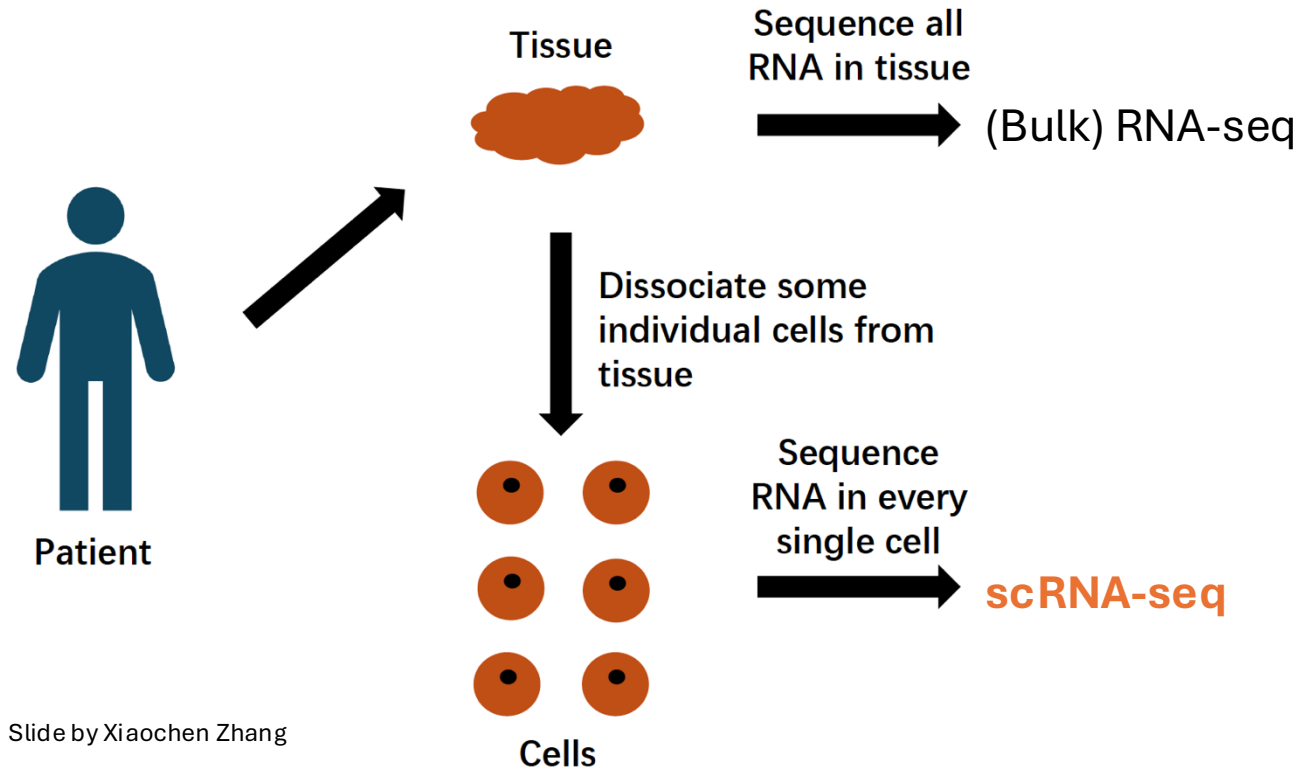
Phenotype: the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

Generative AI for scRNA-seq **data** analysis



How the data are generated?

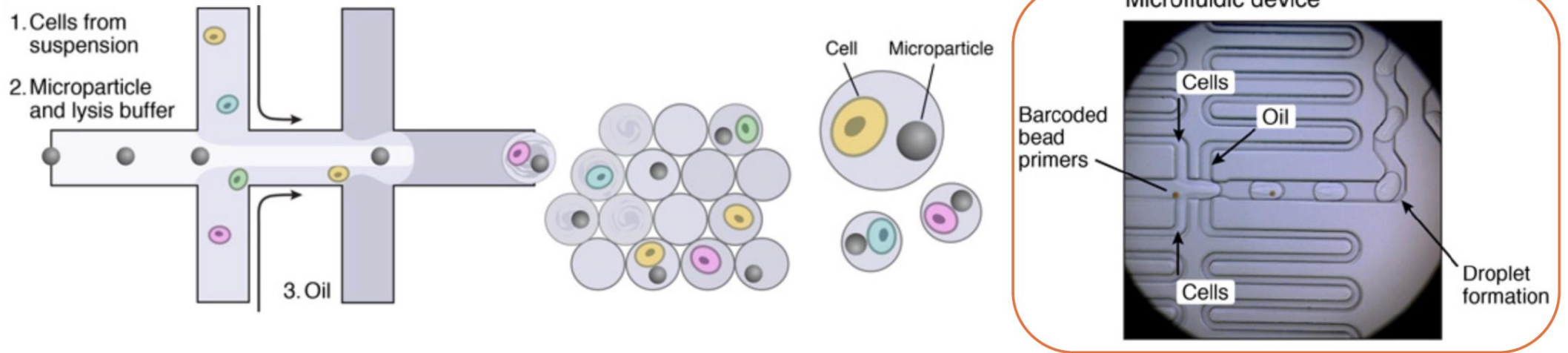
scRNA-seq: overview



Slide by Xiaochen Zhang

What happened in the machine

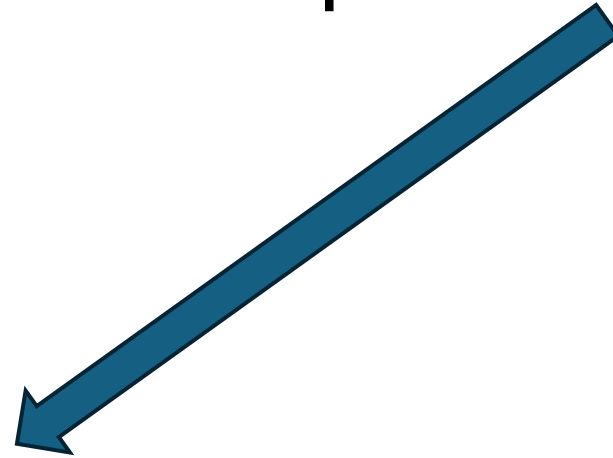
Most scRNA-seq experiments use **droplet-based** platforms for dissociation.



Some scRNA-seq experiments use **plate-based** platforms for dissociation.

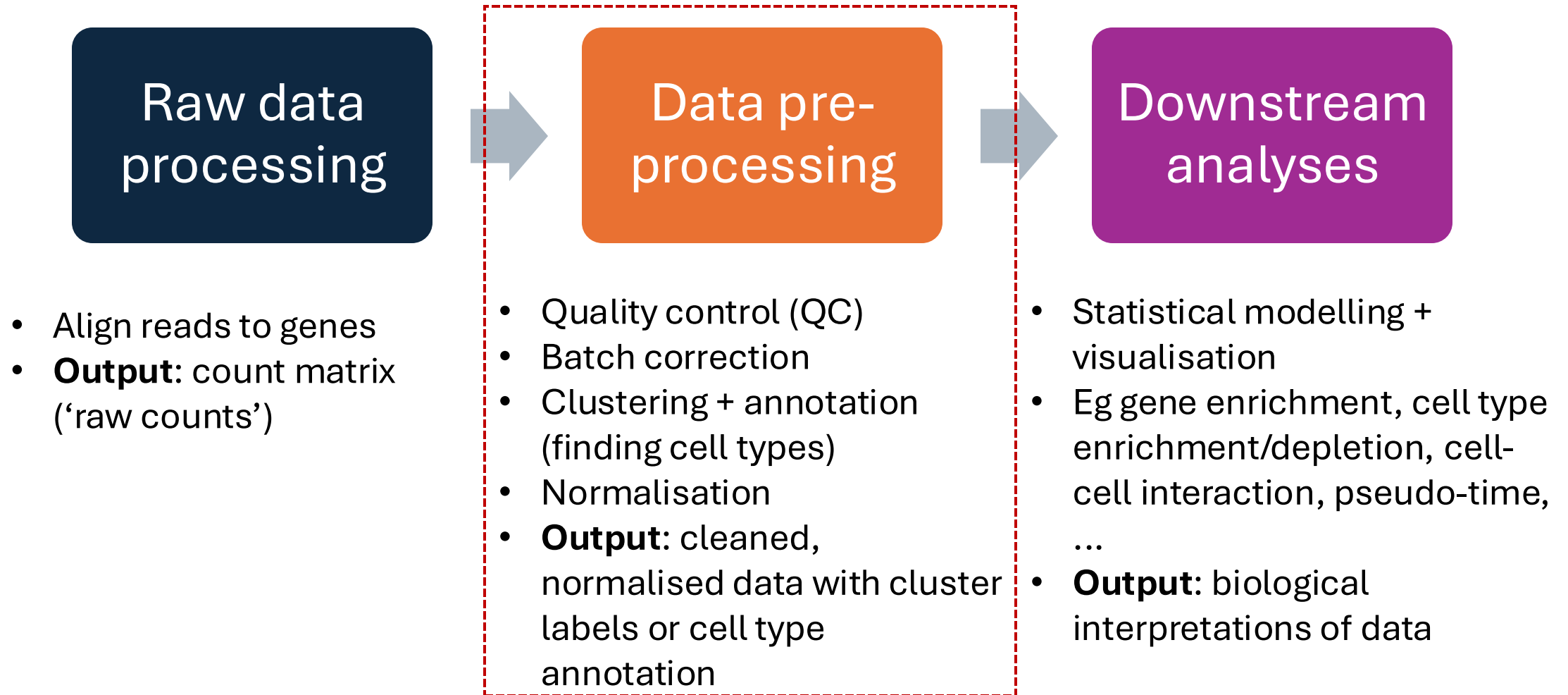
We assign each dissociated cell a barcode consisting of 4-20 bases to determine its identity.

Generative AI for scRNA-seq data **analysis**

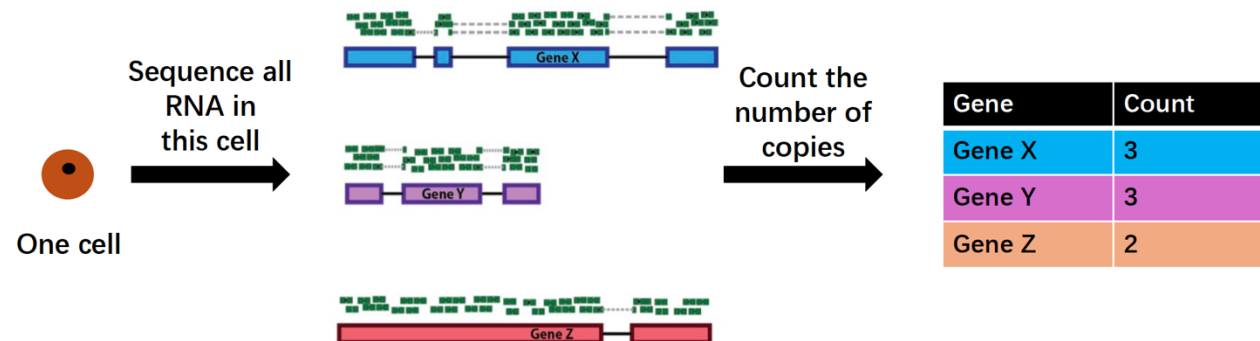
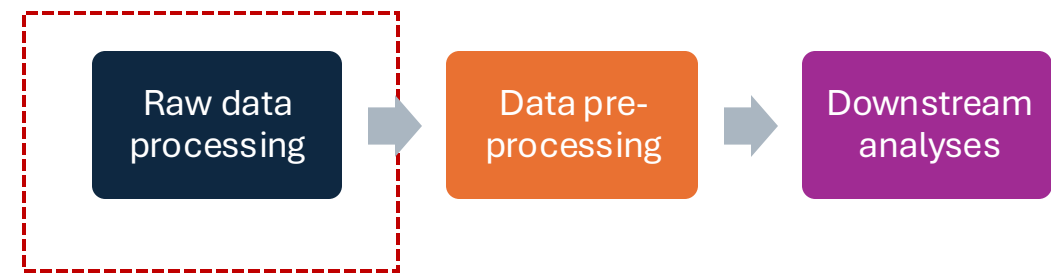


How the analysis is typically done?

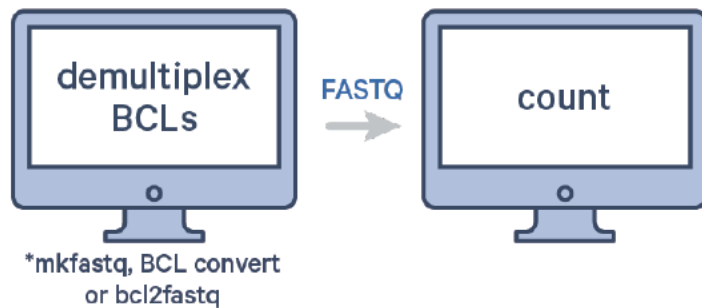
Review of pipeline



From sample to counts



----- Cell Ranger -----



	pbmc2_10X_V2_AAACCTGAGATGGGTC	pbmc2_10X_V2_AAACCTGAGCGTAATA	pbmc2_10X_V2_AAACCTGAGCTAGGCA	pbmc2_10X_V2_AAACCTGAGGGTCTCC
TSPAN6	0	0	0	0
TNMD	0	0	0	0
DPM1	0	0	0	0
SCYL3	0	0	0	0
C1orf112	0	0	0	0
FGR	0	0	2	0
CFH	0	0	0	0
FUCA2	0	0	0	0
GCLC	0	0	0	0
NFYA	0	0	0	0
STPG1	0	0	0	0
NIPAL3	0	1	0	0
LAS1L	0	0	0	0
ENPP4	0	0	0	0

High dimension (typically > 10,000 cells * 20,000 genes)

Dimension reduction tools

Very Sparse (more than 90% of the matrix includes 0)

Feature selection, new computational methods

Quality control

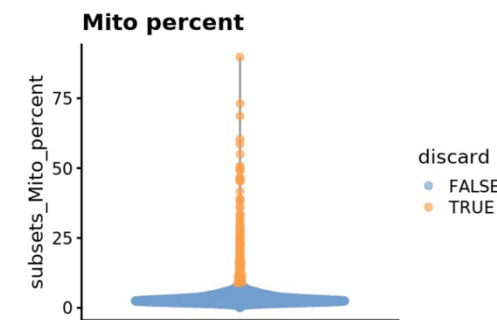
- Low-quality cells
 - Low gene counts (expressing too few genes)
 - High proportion of mitochondria genes
- Low-quality genes
 - Low cell counts (expressed in too few cells)
- Feature selection
 - Highly variable genes
- Not covered:
 - Normalisation, batch correction





```
sc.pp.filter_cells(adata, min_counts=3)
```

```
pbmc.var["mito"] = pbmc.var_names.str.startswith("MT-")
```

```
sc.pp.filter_genes(adata, min_counts=3)
```

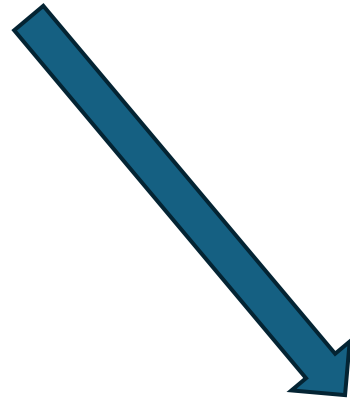
```
sc.pp.highly_variable_genes(  
    adata,  
    n_top_genes=1200,  
    subset=True,  
    layer="counts",  
    flavor="seurat_v3",  
    batch_key="cell_source",  
)
```



	Gene 1	Gene 2	Gene 3
	2	5	0
	1	1	0
	1	1	2
	1	0	4

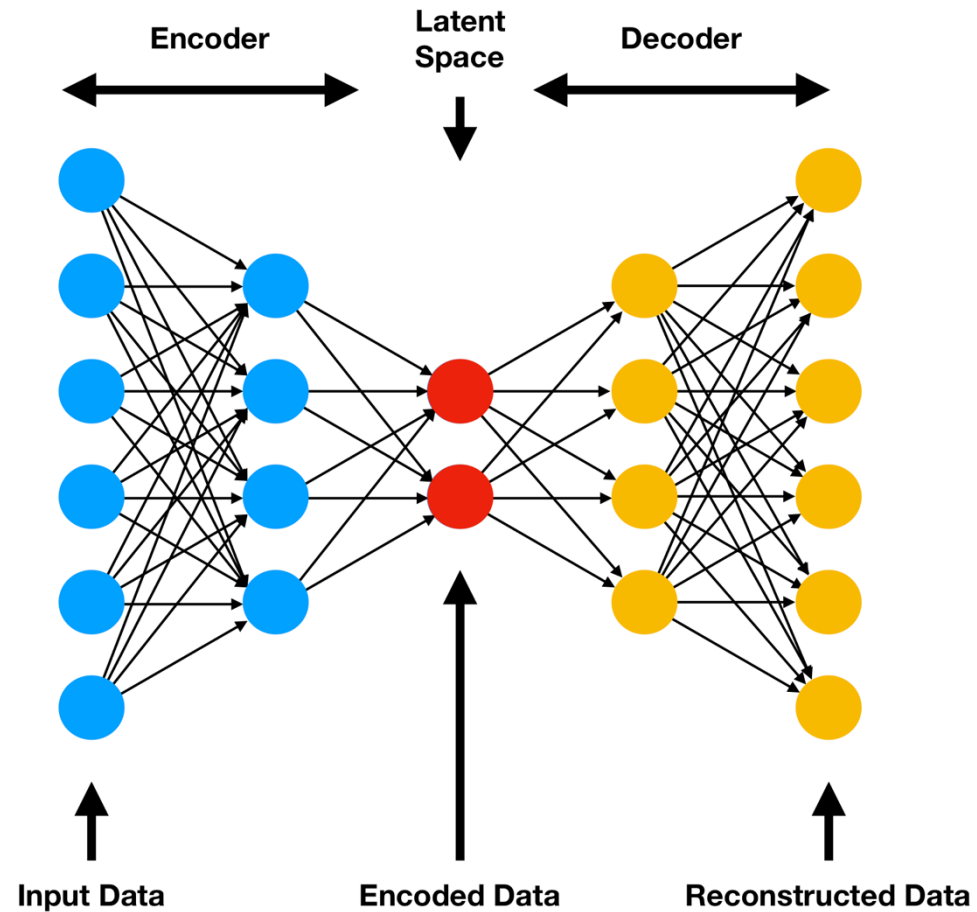
Variance = 0.25 4.92 3.67

Generative AI for scRNA-seq data analysis

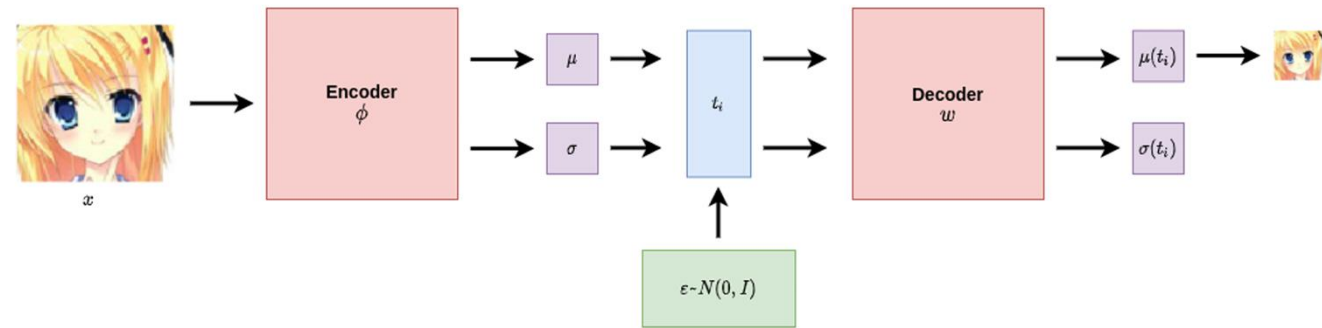


What is generative AI?

Autoencoders



Is PCA an autoencoder?



Variational autoencoder (VAE)

Generative modelling of images

Training



training data

Applications

Corrupted sample with dropout



Clear one



denoising by reconstruction

Real one



Reconstruction



reconstruction
(not a goal in itself)

Original image



Dist=10.768



Dist=11.597



Dist=12.939



image embedding

a=0.00



a=0.17



a=0.33



a=0.50



a=0.67



a=0.83



a=1.00



a=0.00



a=0.17



a=0.33



a=0.50



a=0.67



a=0.83



a=1.00



image interpolation (in latent space)

Key: dimension reduction (encoding) to a space (**manifold**) that is easier to manipulate

Generative AI for scRNA-seq data analysis



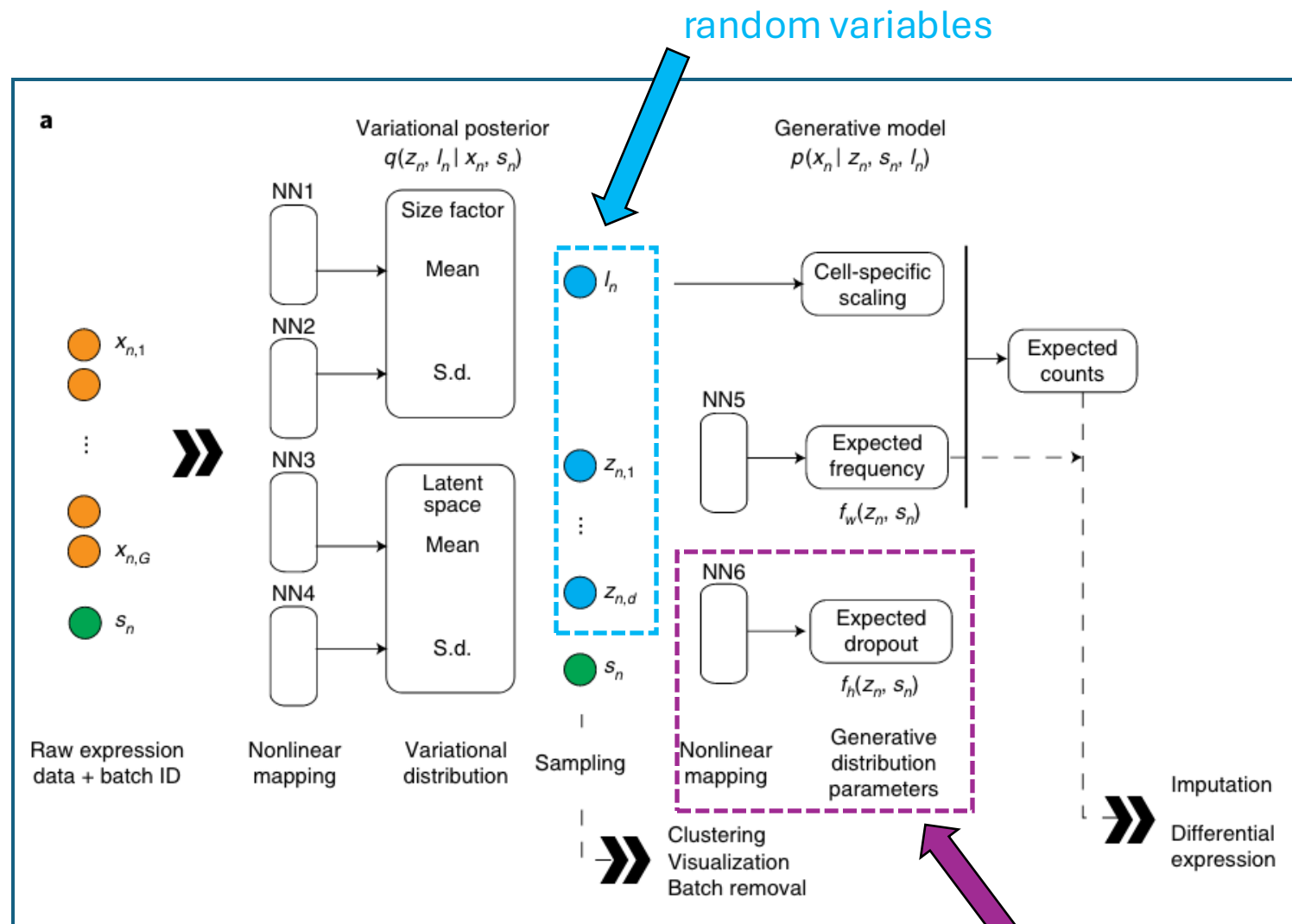
scVI model

Six neural networks (NNs):

- NN1, NN2 **encode** mean & sd of l_n , size factor of cell n
- NN3, NN4 **encode** mean & sd latent variables (default $d = 10$)

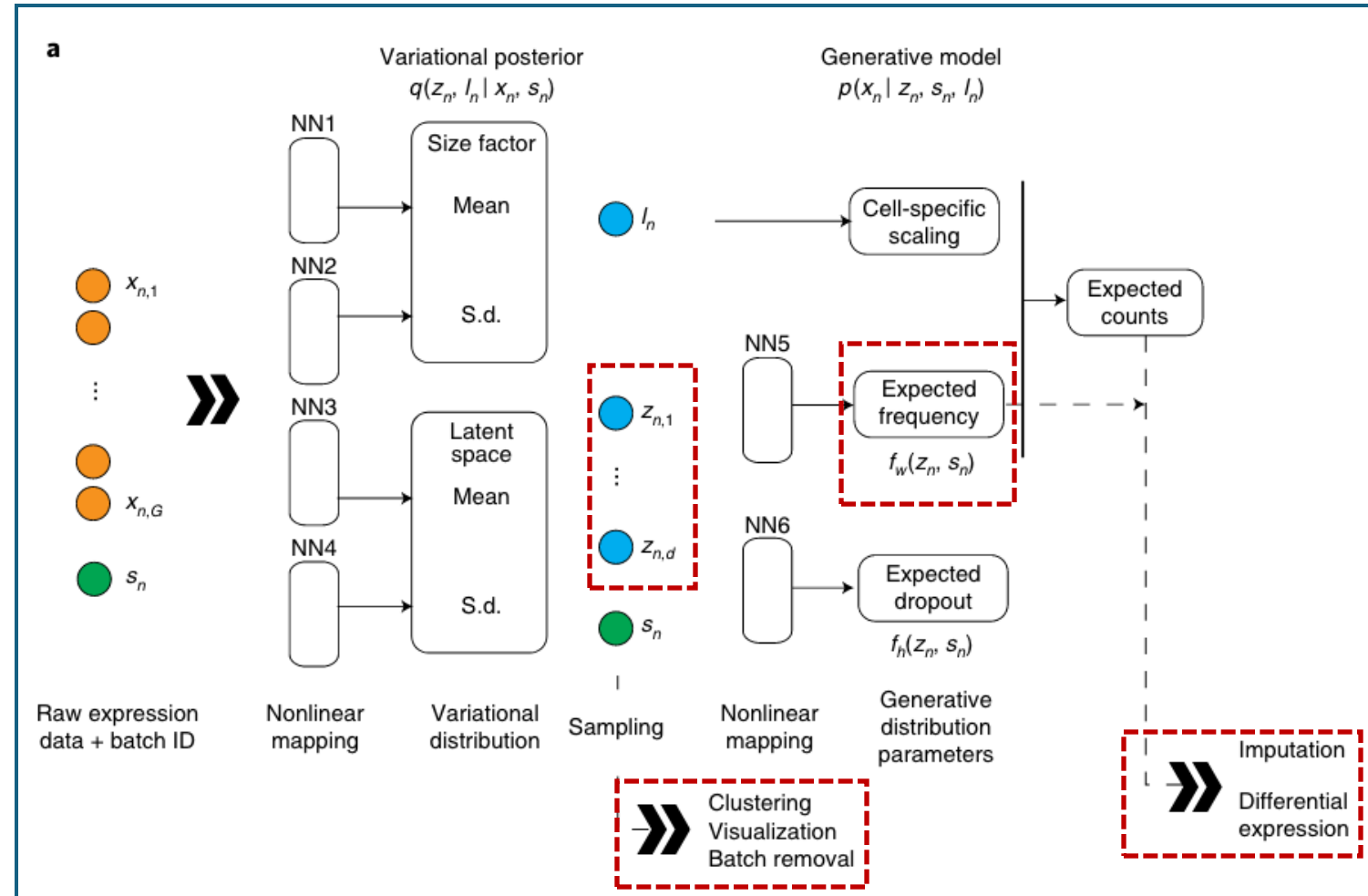
Then, generate values from $z_{n,1}, \dots, z_{n,d}$,

- NN5 **decodes** 'expected frequency', ie normalised counts
- NN6 regenerates dropout events (**optional**)
- Optional: generate l_n and recover counts
- Have to use raw counts as input



What has scVI done?

- Modelled **latent variables**
 - (Nonlinear) dimension reduction
 - Independent from **size factor** & **batch**
 - Suitable for visualisation & clustering
- Generated expected frequencies
 - Normalised (against size factor) & batch corrected counts
- Generated expected counts
 - Batch corrected (not normalised)
 - Useful for DE



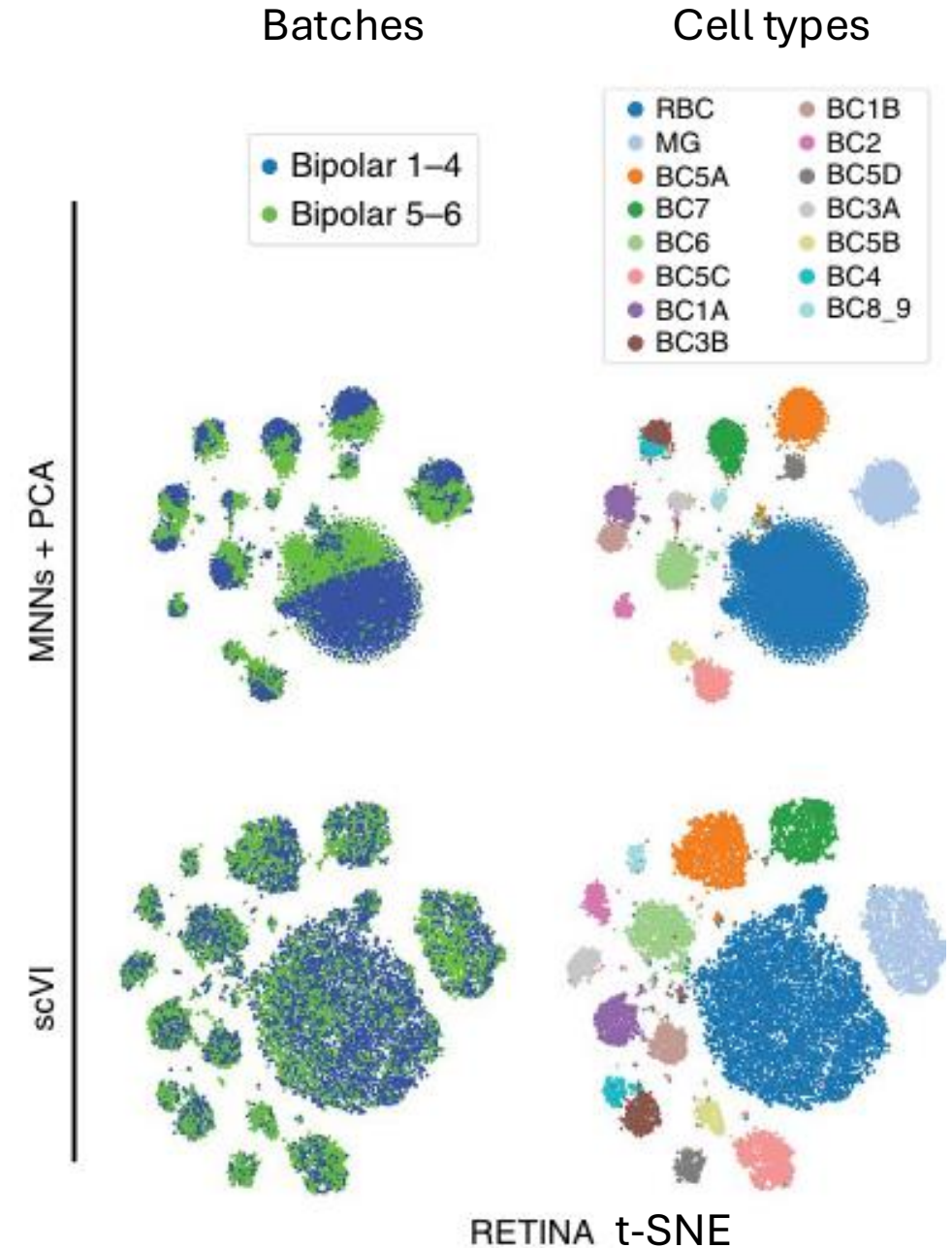
Batch correction

Data

- 27,499 cells
- Cell type labels from original authors
- 2 batches (bipolar 1–4, bipolar 5–6)

Goal

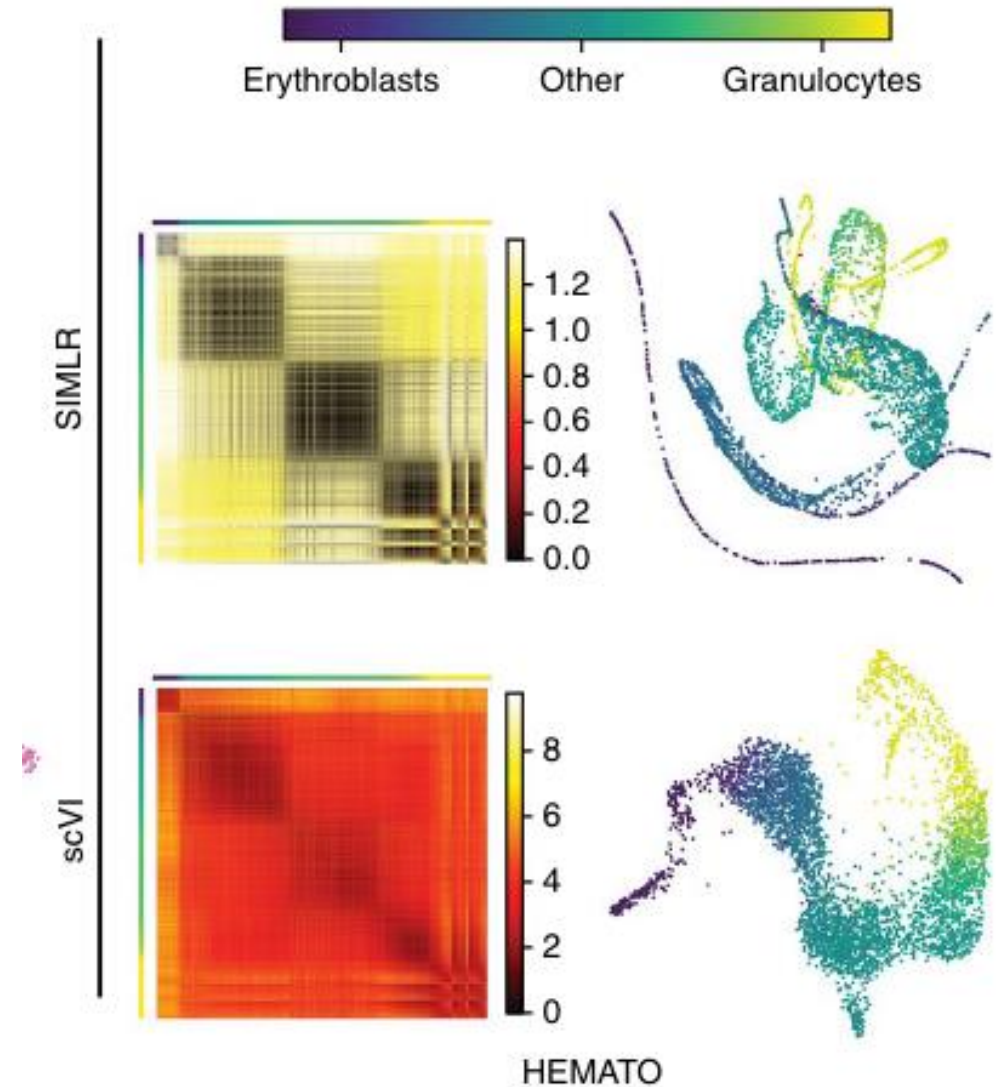
Comparing scVI and MNNs on batch correction



Preservation of continuous biology

HEMATO

- 4,016 hematopoietic progenitor cells
- Cells along a developmental trajectory



Good. But not quite enough.

scANVI: semi-supervised VAE

Article



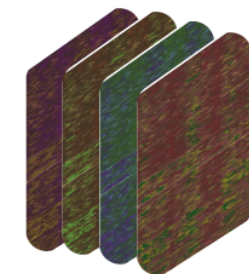
molecular
systems
biology

Motivation

- Unsupervised batch correction is not always appropriate
- Eg when batch cofounds biology
 - Disease samples in batch 1, control in batch 2
 - This happens a lot, due to ...
- Batch effects removed, part of biological difference also removed
- A bit of guidance (supervision) helps
 - Eg control and control should be similar across batches
 - Eg CD4 T cells should stay together on UMAP/t-SNE
- Guidance comes in the form of (partial) labels

Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models

Chenling Xu^{1,†}, Romain Lopez^{2,†}, Edouard Mehlman^{2,3,†}, Jeffrey Regier⁴, Michael I Jordan^{2,5} & Nir Yosef^{1,2,6,7,*}



Collection of
scRNA-seq datasets

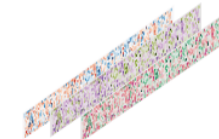


Core analysis tool: scVI

Multi-purpose generative model

- visualization
- clustering
- differential expression
- harmonization

Extends to



Partial cell type
annotation

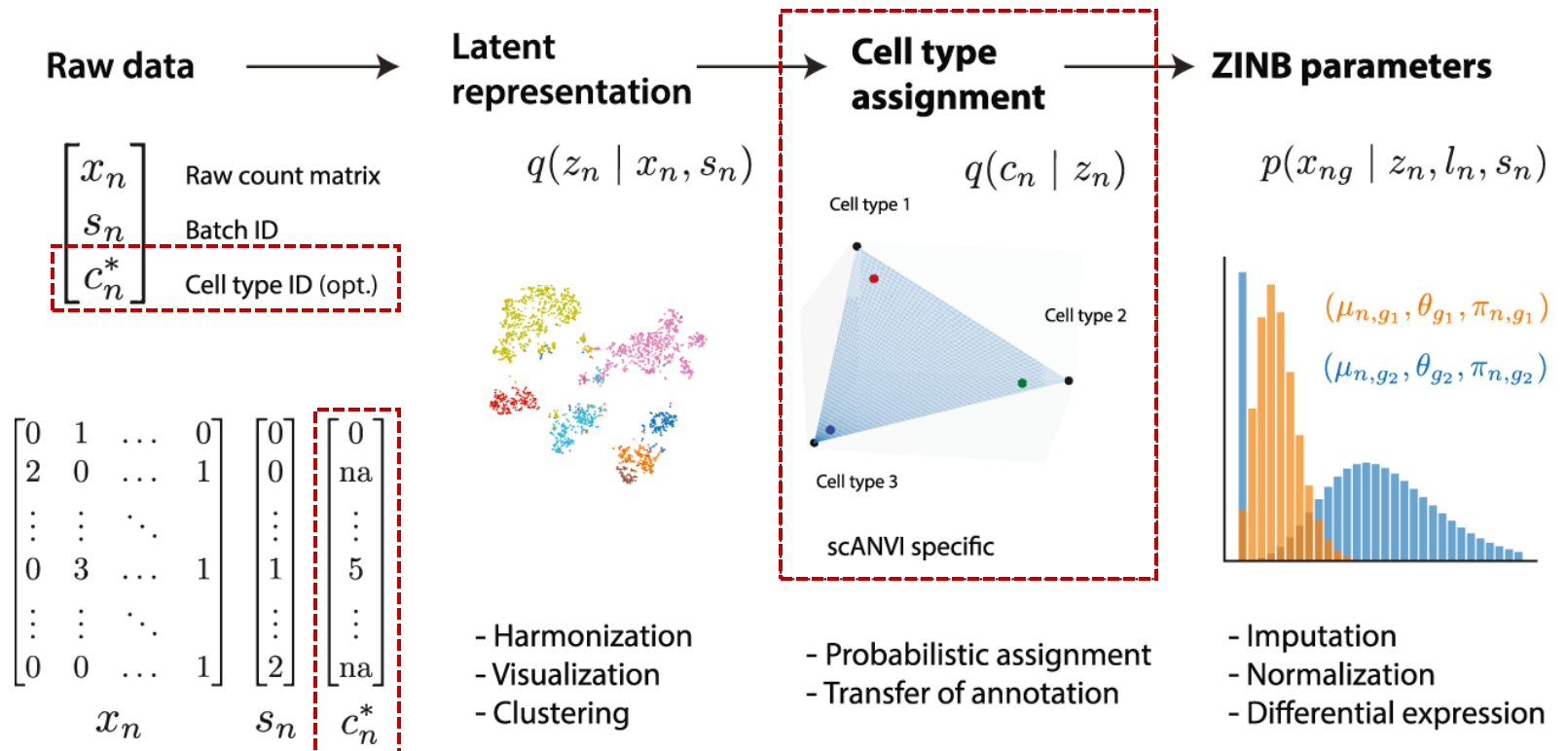
Annotation tool: scANVI

Transfer of annotation in various settings:

- partial overlap of labels
- partial "seed" labeling
- hierarchical labels

scANVI model

- Added cell type labels
 - Could be simply cluster labels
 - Can be partial
- Restriction: cells of same label have similar latent representations
- Result: batch correction by aligning cells from same type



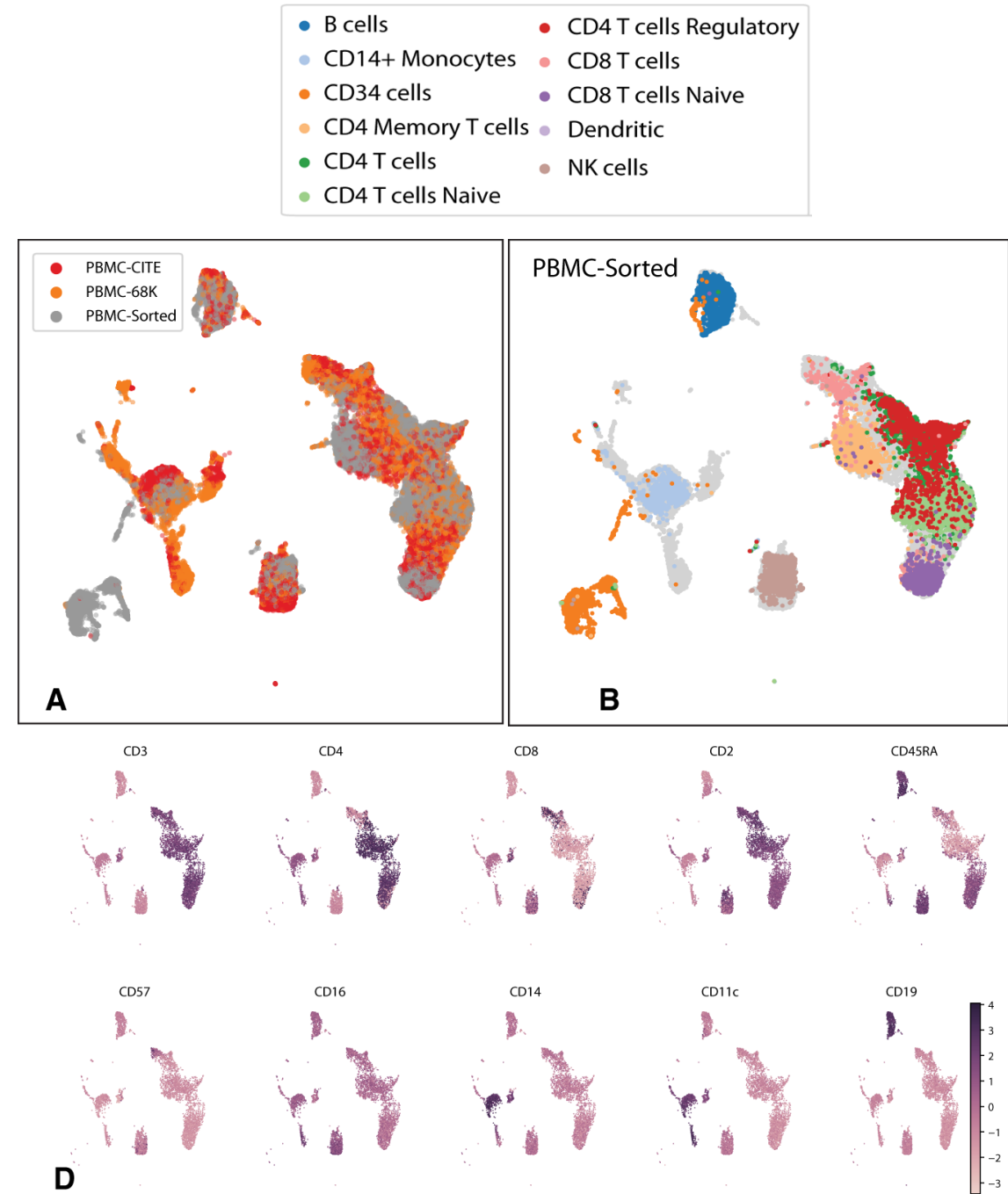
Case study

Data

- Three PBMC (peripheral blood mononuclear cells) datasets
 - 169,850 cells in total
- PBMC-Sorted: annotated
- PBMC-65k: unannotated
- PMBC-CITE: unannotated, also contains 10 surface protein counts
- Protein markers serve as validation

Challenges

- Partially labelled
- Not all cell types available across three datasets (imbalanced design)



Put them into use

Building large-scale cell atlas

Motivation

- Integrate large number of studies from hundreds of patients

Challenges

- Serious batch effects
- Large number of cells (millions), huge computational burden

Why scANVI

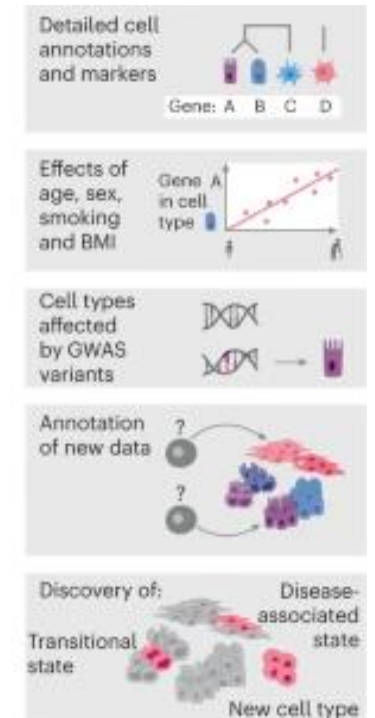
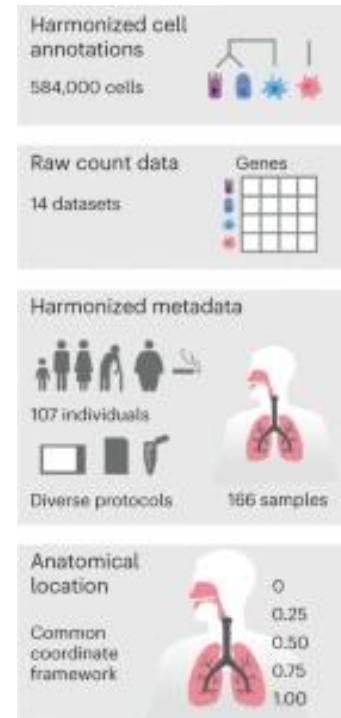
- Flexible and biology-preserving batch correction
- Computationally scalable
 - Takes longer to train compared to other methods, on smaller scale data
 - Computational time increases not fast when sample size increases
 - Reason: back propagation, small batch learning (no need to handle huge matrices)

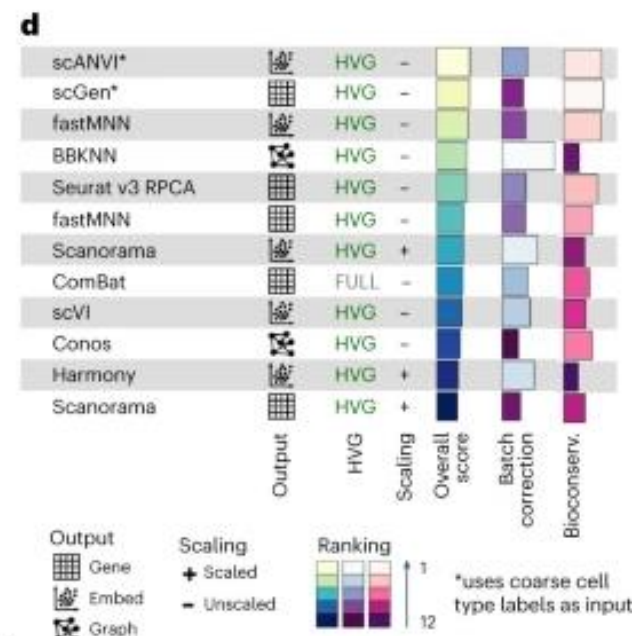
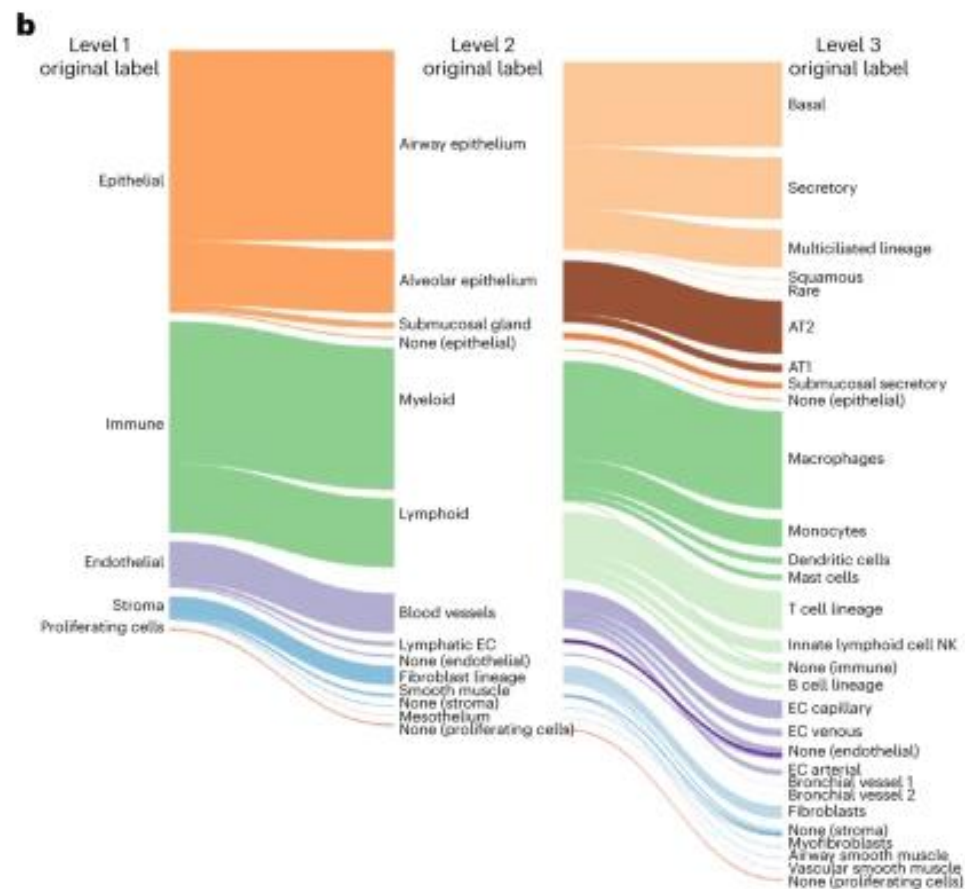
An integrated cell atlas of the lung in health and disease

[Lisa Sikkema](#), [Ciro Ramírez-Suástegui](#), [Daniel C. Strobl](#), [Tessa E. Gillett](#), [Luke Zappia](#), [Elo Madisson](#), [Nikolay S. Markov](#), [Laure-Emmanuelle Zaragosi](#), [Yuge Ji](#), [Meshal Ansari](#), [Marie-Jeanne Arguel](#), [Leonie Apperloo](#), [Martin Banchero](#), [Christophe Bécavin](#), [Marijn Berg](#), [Evgeny Chichelnitskiy](#), [Mei-i Chung](#), [Antoine Collin](#), [Aurore C. A. Gay](#), [Janine Gote-Schniering](#), [Baharak Hooshier Kashani](#), [Kemal Inecik](#), [Manu Jain](#), [Theodore S. Kapellos](#), [Lung Biological Network Consortium](#), ... [Fabian J. Theis](#) 

+ Show authors

[Nature Medicine](#) **29**, 1563–1577 (2023) | [Cite this article](#)





Query your own lung samples

Use of atlas: ‘reference mapping’

- Get cell type annotations
- Compare with your own annotations
- Marker genes
- Levering more samples
 - Eg I don’t have enough control samples in my study

The integrated Human Lung Cell Atlas

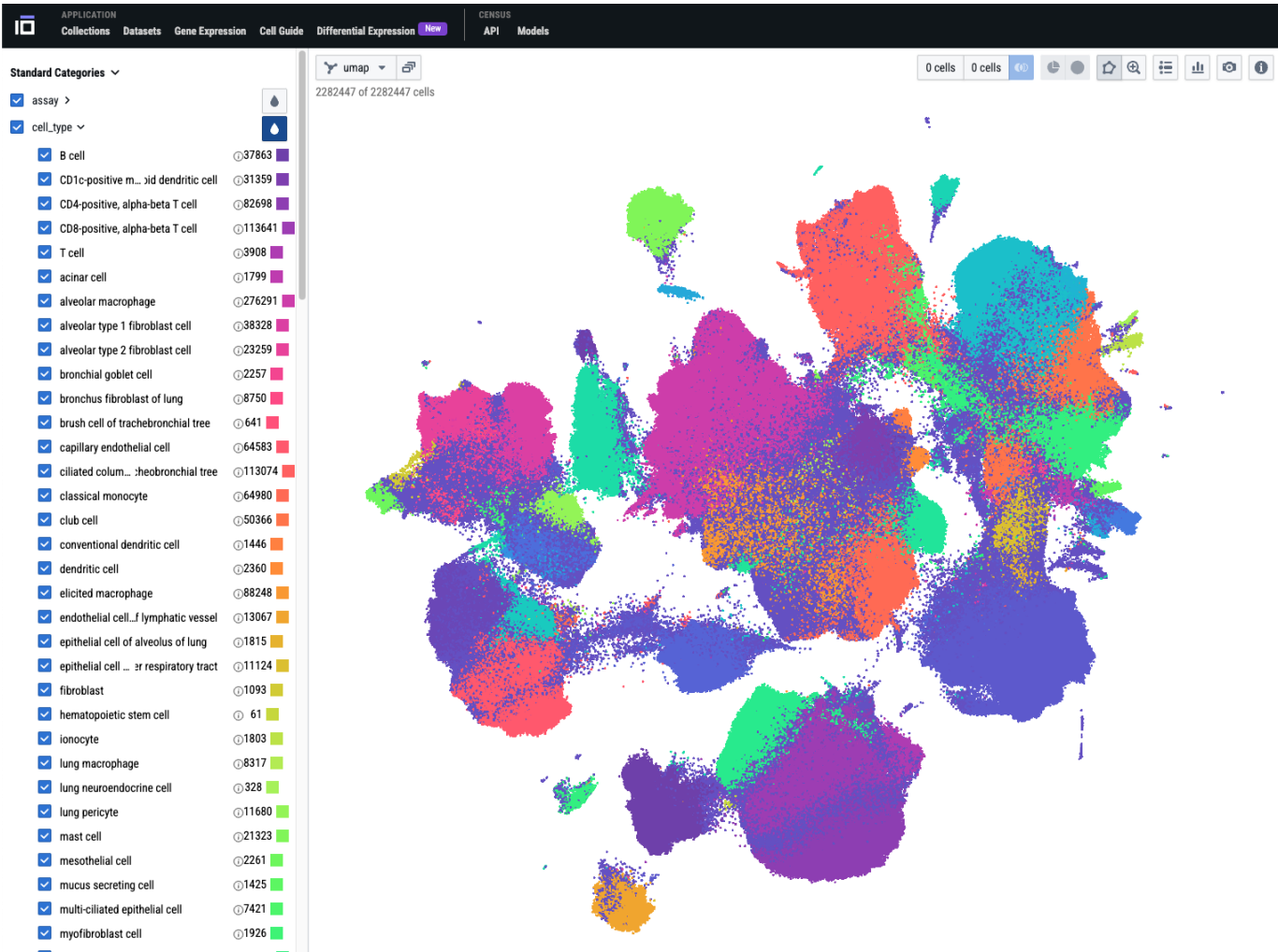
CZI Single-Cell Biology, Human Cell Atlas (HCA)

The integrated Human Lung Cell Atlas (HLCA) represents the first large-scale, integrated single-cell reference atlas of the human lung. It consists of over 2 million cells from the respiratory tract of 486 individuals, and includes 49 different datasets. It is split into the HLCA core, and the extended or full HLCA. The HLCA core includes data of healthy lung tissue from 107 individuals, and includes manual cell type...

[Show More](#)

Publication [Sikkema et al. \(2023\) Nat Med](#)
Contact [Malte D. Luecken](#)
Other [HLCA landing page](#)

Dataset	Tissue	Disease	Assay	Organism	Cells		
An integrated cell atlas of the human lung in health and disease (full)	4 tissues	normal 15 diseases	9 assays	Homo sapiens	2,282,447	Download	Explore
An integrated cell atlas of the human lung in health and disease (core)	3 tissues	normal	5 assays	Homo sapiens	584,944	Download	Explore



Discussions

- scVI & scANVI serve as data pre-processing tools
- Decoded data are ready for downstream analysis
- Scalable and flexible
- But lack interpretability compared to linear methods
 - There is no 'loading' for latent variables
- Suitable for building atlases and transfer learning
 - Train on large scale data
 - Download trained model and fine tune on own data



Mr Atlas on Collins Street