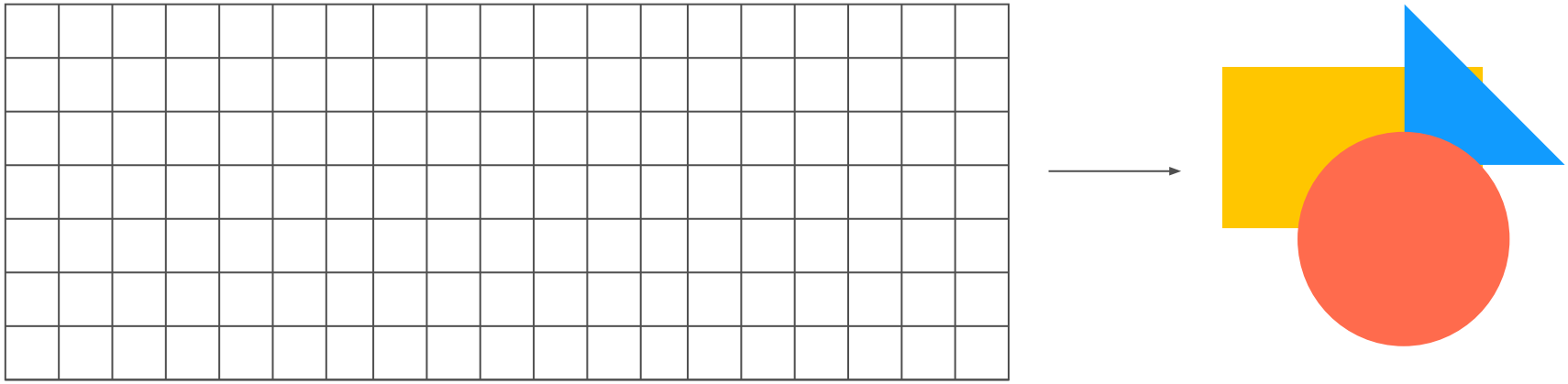
From Counts to Insights:

Analyzing Single-Cell Gene Expression using Public Datasets



About today's talk

Background

Single-cell RNA-seq Studies

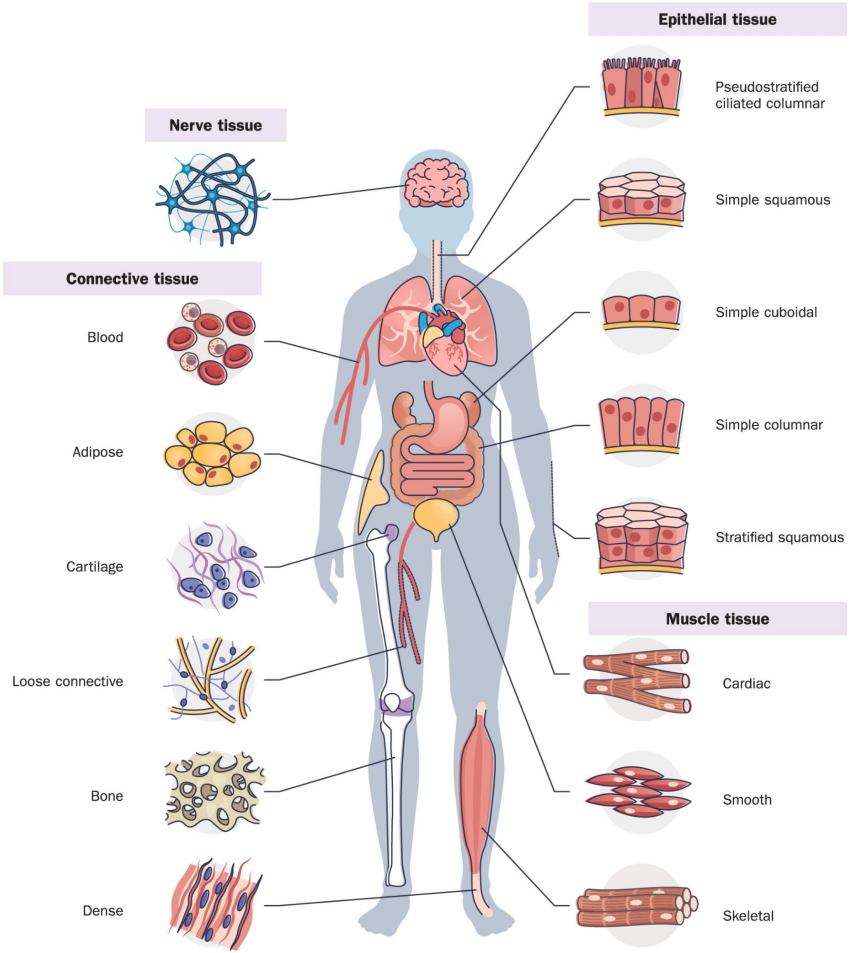
Single-cell Analysis Demo

Takeaways

About cells

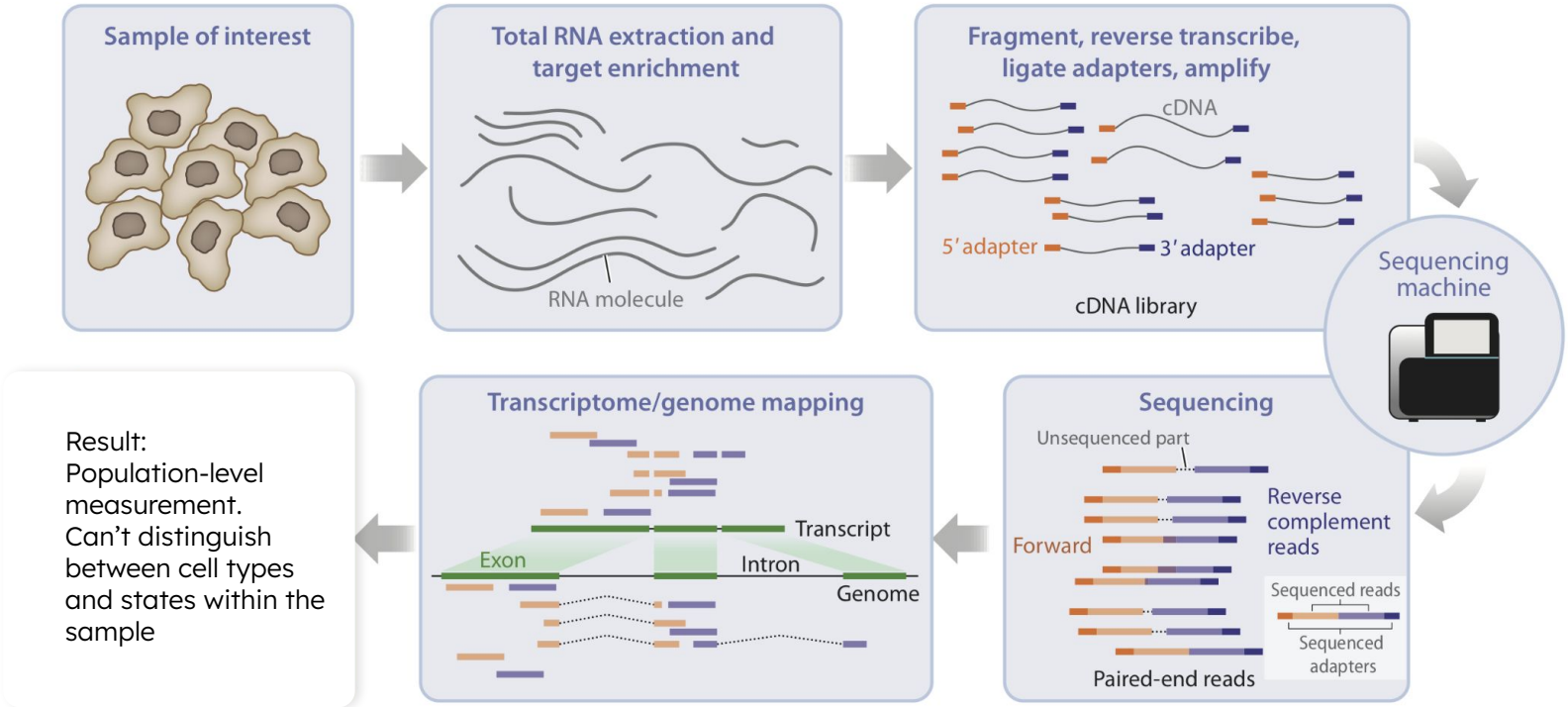
~ 37 trillion cells in the human body

- All cells contain (basically) the same DNA
- But, each cell expresses a different subset of genes
- Goal: Measure gene expression across different cells



British Journal of Nursing

RNA Sequencing

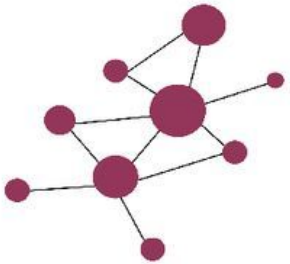


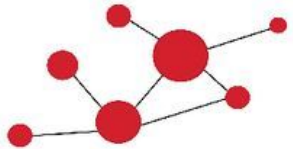
British Journal of Nursing

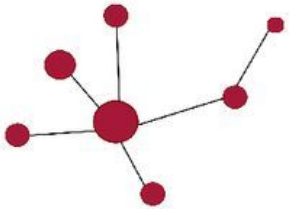
organ/ tissue

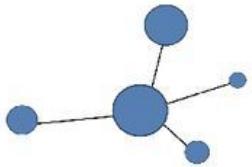
bulk RNA seq

single-cell RNA seq





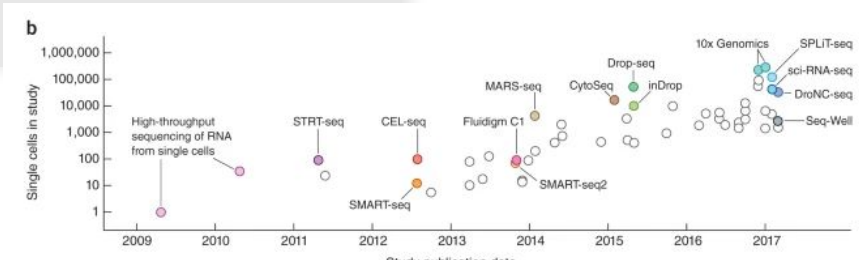




Steinheuer et al., 2021

Bulk → Single-cell RNA-seq

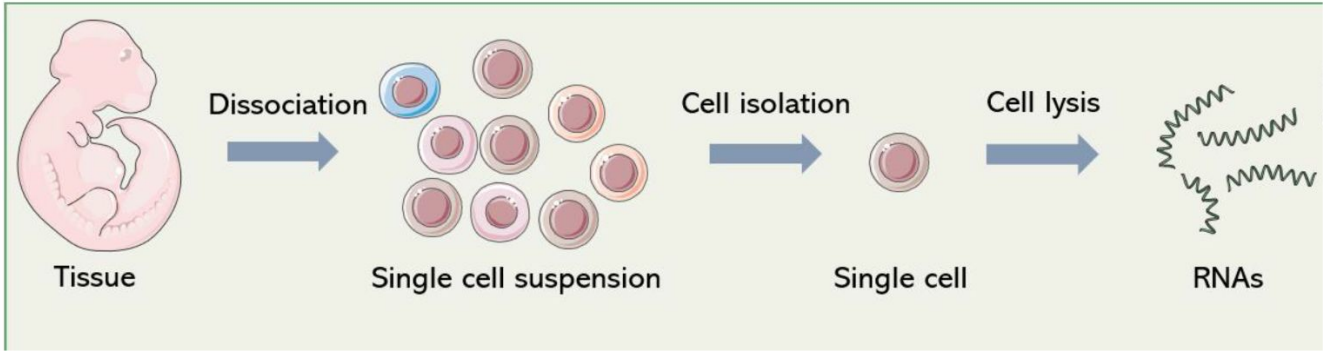
- 1992: Expression of genes measured from single cells for the first time (Eberwine et al.)
- 2009: First untargeted, transcriptome-wide single-cell RNA-seq (Tage et al.)
- 2010: Cell types can be identified without pre-sorting (Guo et al.)
- Last 10 years: New methods enable massive throughput
    - Over time, single-cell tech has become more high throughput and cheaper



Svensson et al., 2018

Single-cell RNA-Sequencing

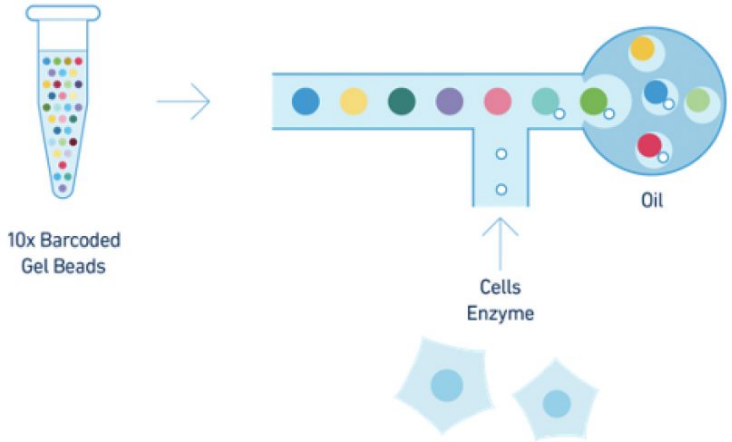
Basic Workflow

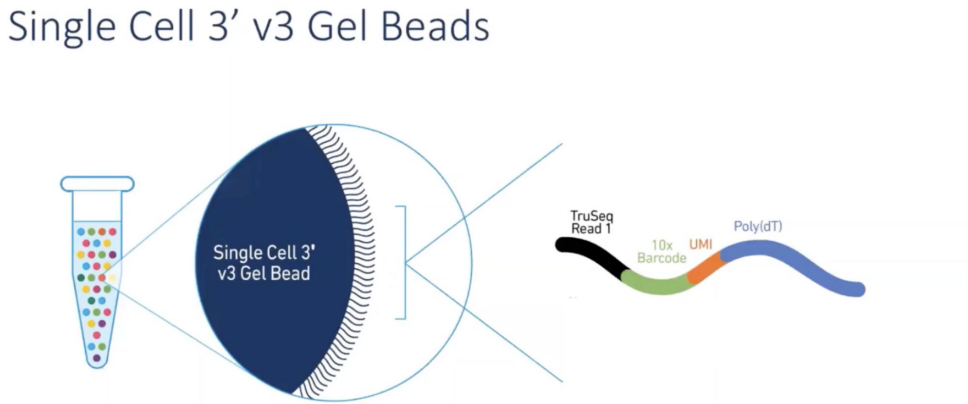


Pan et al., 2022

In scRNA-seq, each cell gets a unique barcode

10x Genomics Example

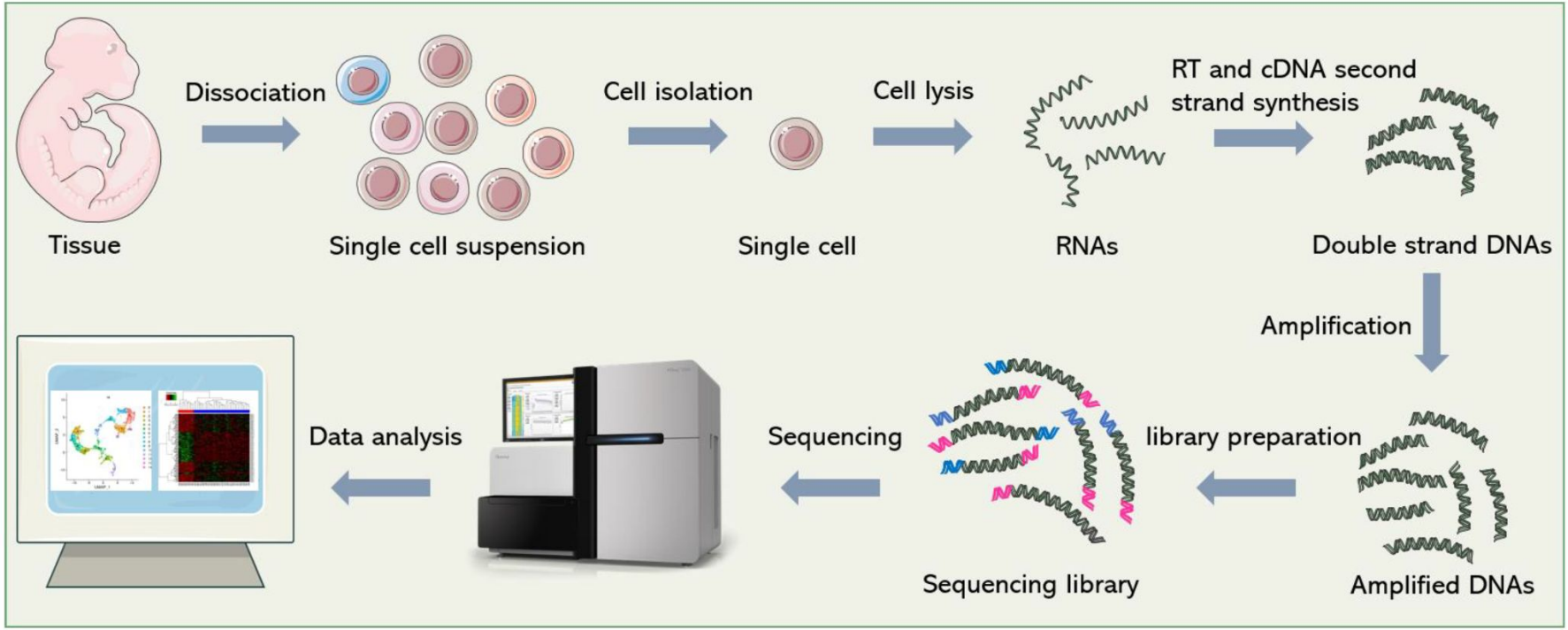




10x Genomics

Single-cell RNA-Sequencing

Basic Workflow

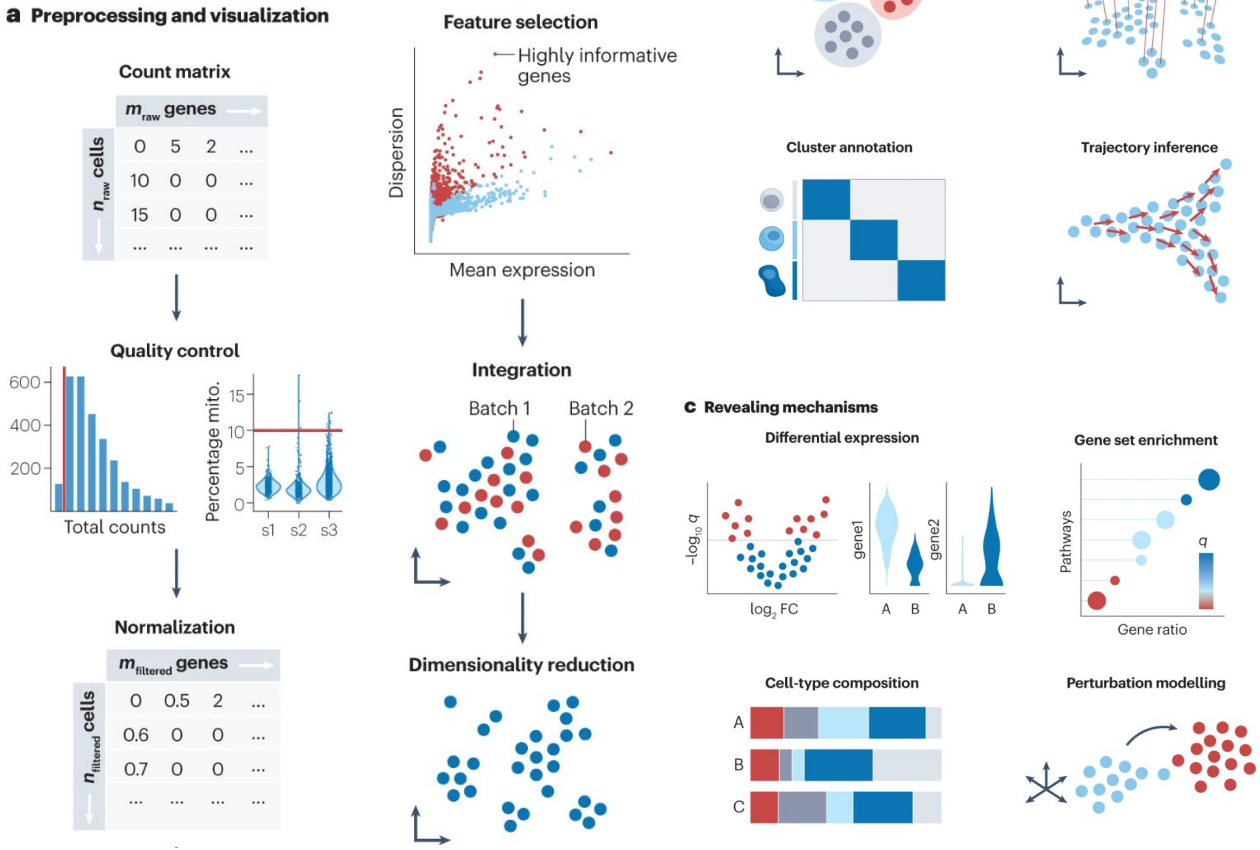


Pan et al., 2022

Single-cell RNA Sequencing

Data Analysis

1. Preprocessing the data
2. Identifying cell types and states
3. Extracting biological insights

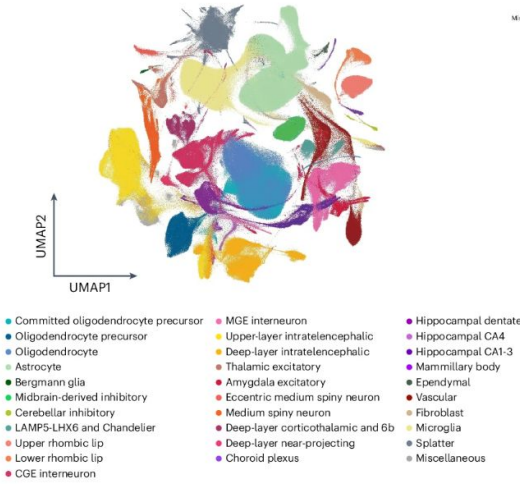


Heumos et al., 2023

What have we learned from scRNA-seq studies?

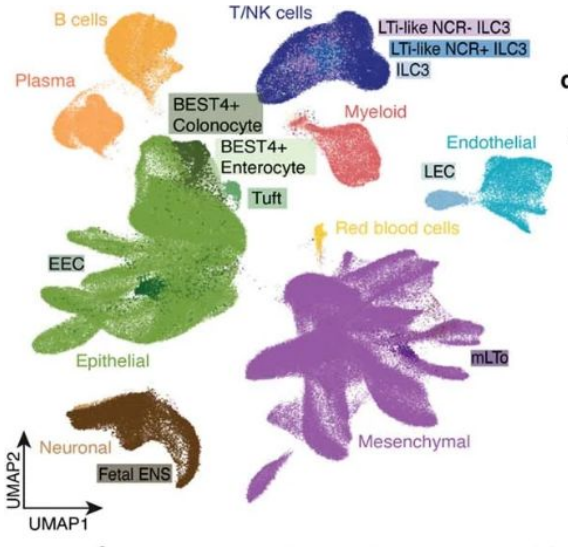
Cell Type Atlases: Mapping tissues

Brain



Chen et al. (2024)

Intestine

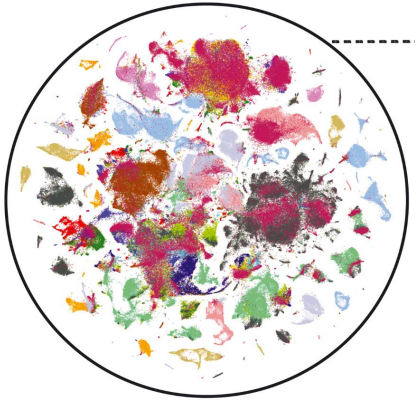


Elmentaite et al. (2021)

Entire Fly

Fly cell atlas

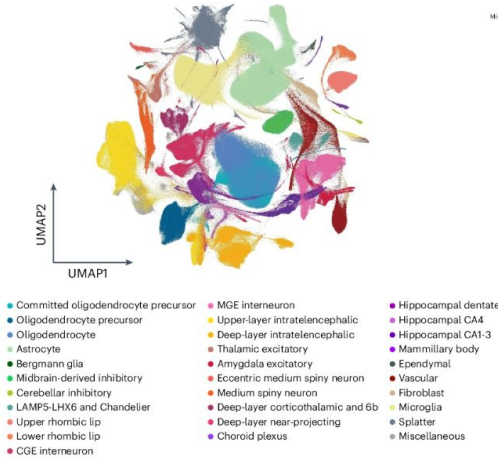
Adult fly
17 tissues
580,000 cells
>250 cell types

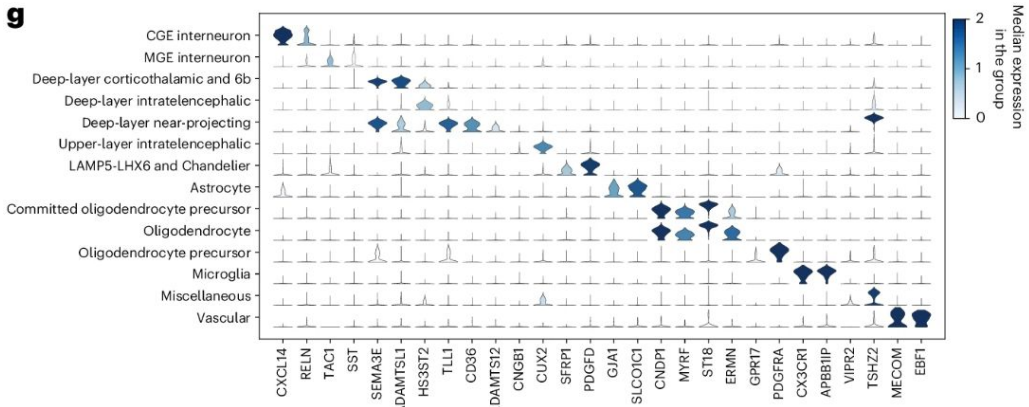


Li et al. (2022)

Differential Expression across Cell Types

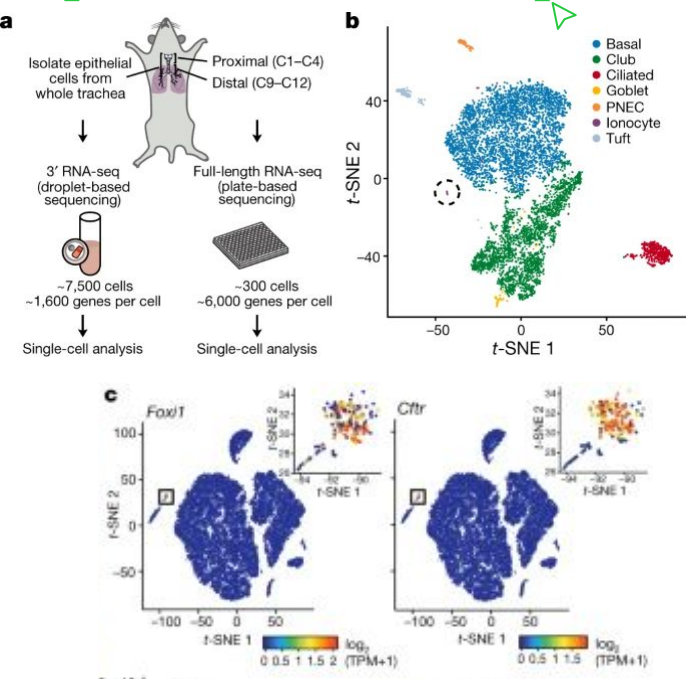
Brain





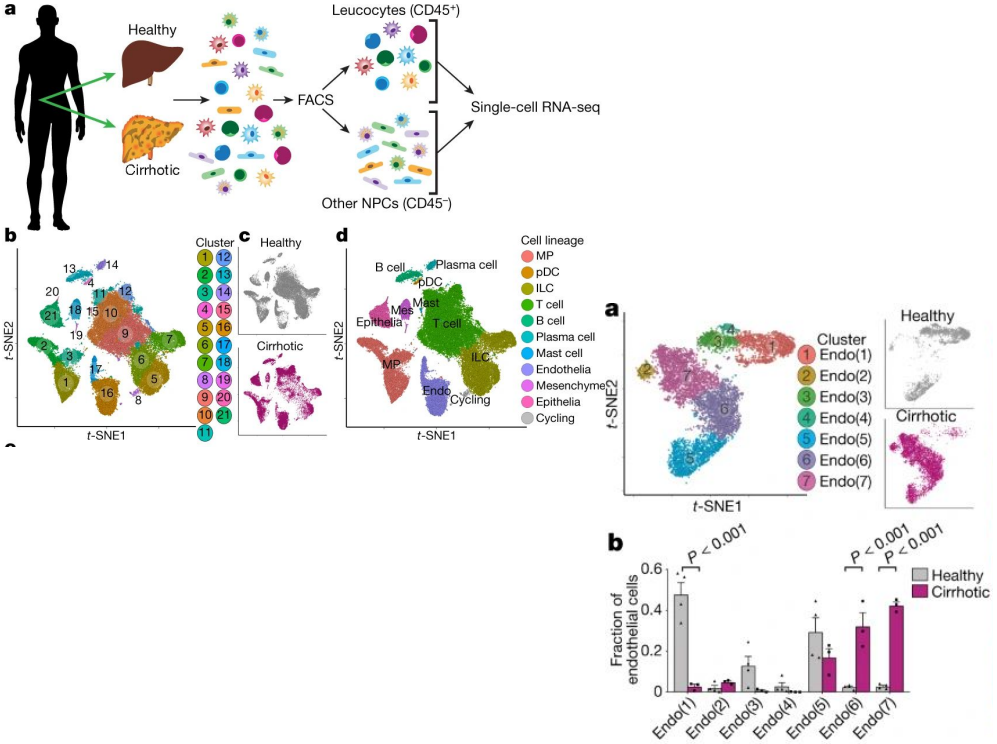
New cell types

Ionocytes in mouse and human



Montoro et al (2018). Plasschaert, et al. (2018),

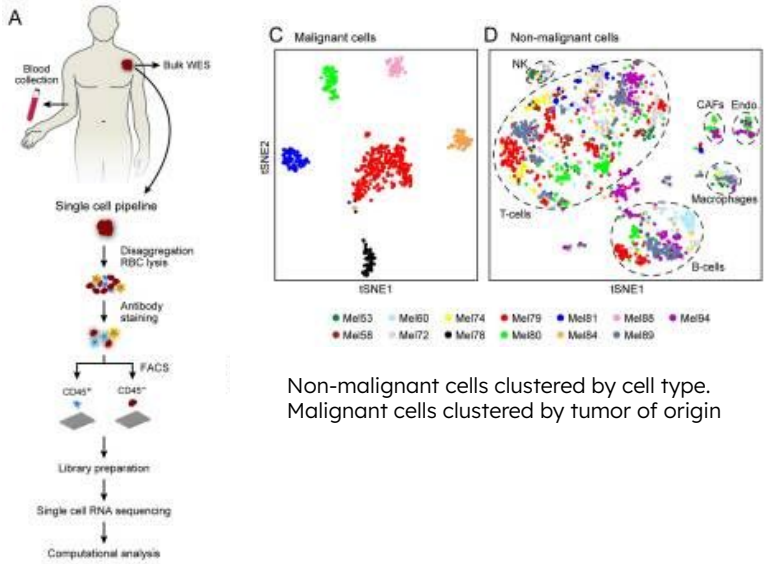
Scar-associated endothelial subpopulations in liver



Ramachandran et al. (2019)

Disease heterogeneity

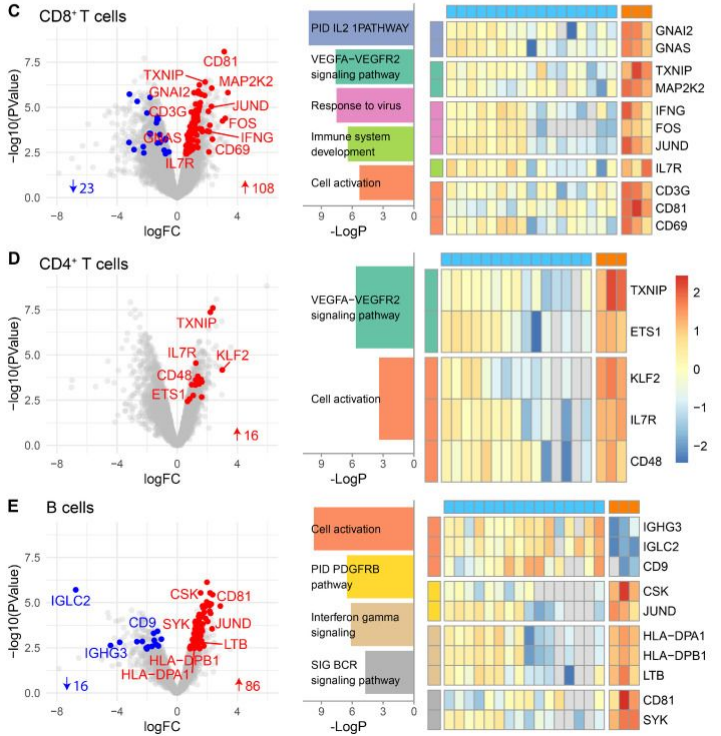
Tumor Biology



Non-malignant cells clustered by cell type.
Malignant cells clustered by tumor of origin

Tirosh et al., 2025

Cystic Fibrosis leads to altered immune cell profiles

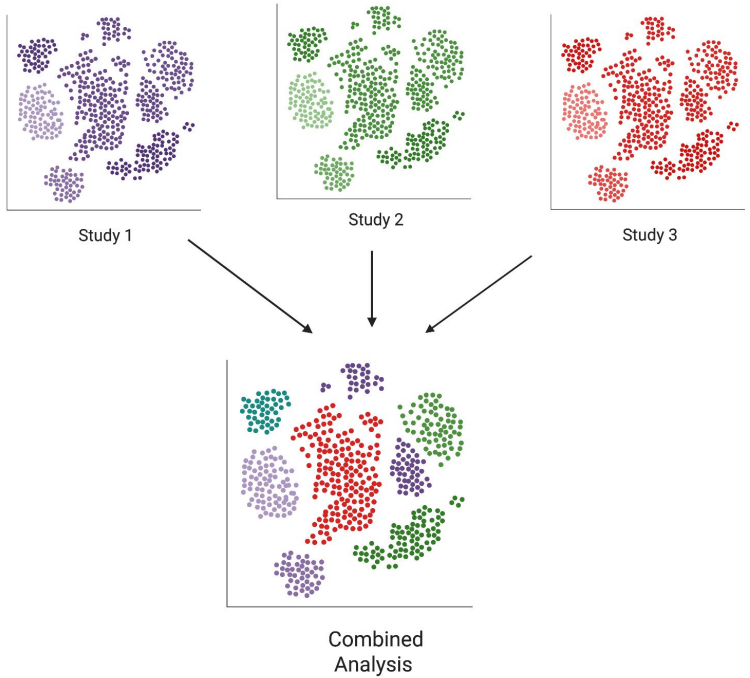


Berg et al., 2025

No Sequencer, No problem...

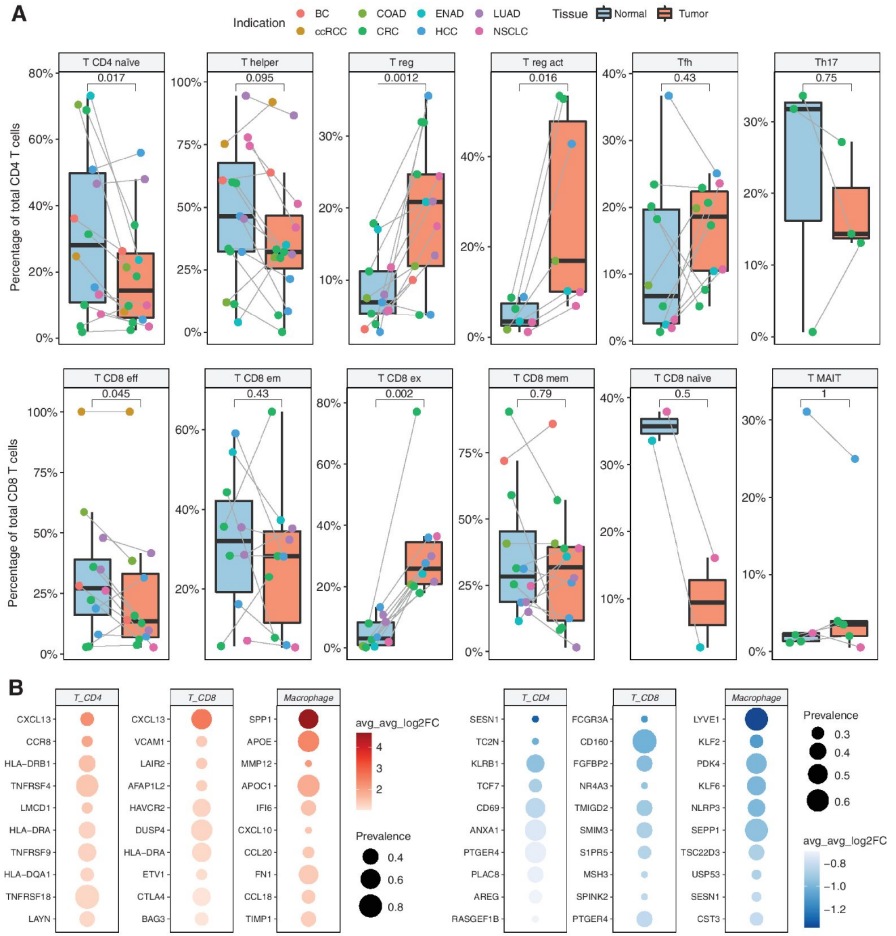
Many large-scale single cell datasets are publicly available, enabling anyone to reanalyze the data, or combine different datasets, for their question of interest.

Meta analyses let you answer questions beyond the scope of any one individual study



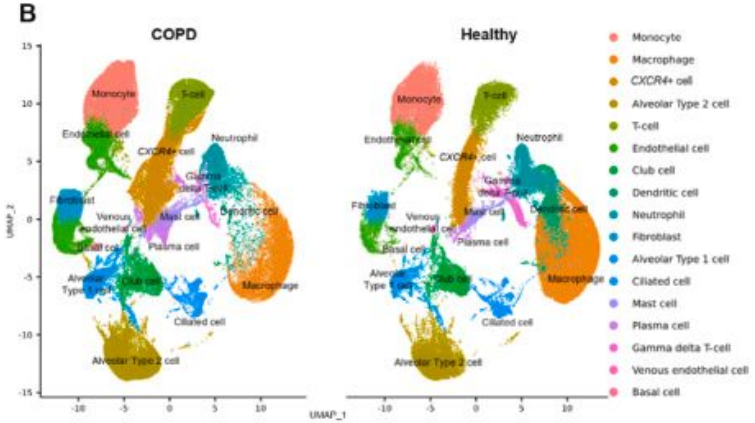
Example: IMMUcan Cancer Database

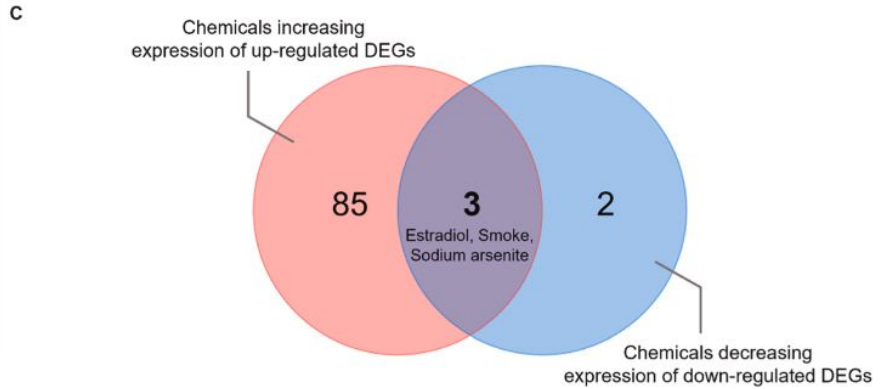
- Integrated 24 cancer datasets across tumor types
- Created a database
- Highlight identification of cell type composition and differential expression conserved across tumor types



Camps et al., 2022

Example: COPD Meta Analysis



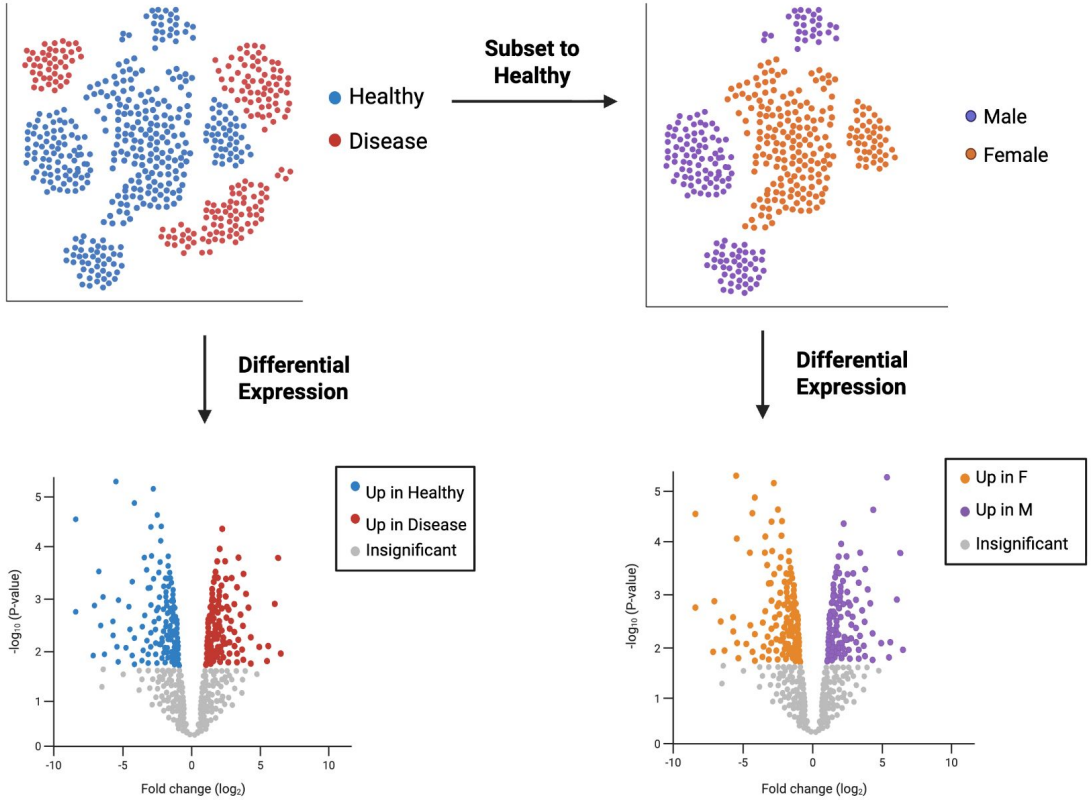


"Three lung scRNA-seq datasets obtained from COPD patients and healthy subjects were integrated."

"Our study suggests the single-cell level mechanisms underlying the pathogenesis of COPD and may provide information on toxic compounds that could be potential risk factors for COPD."

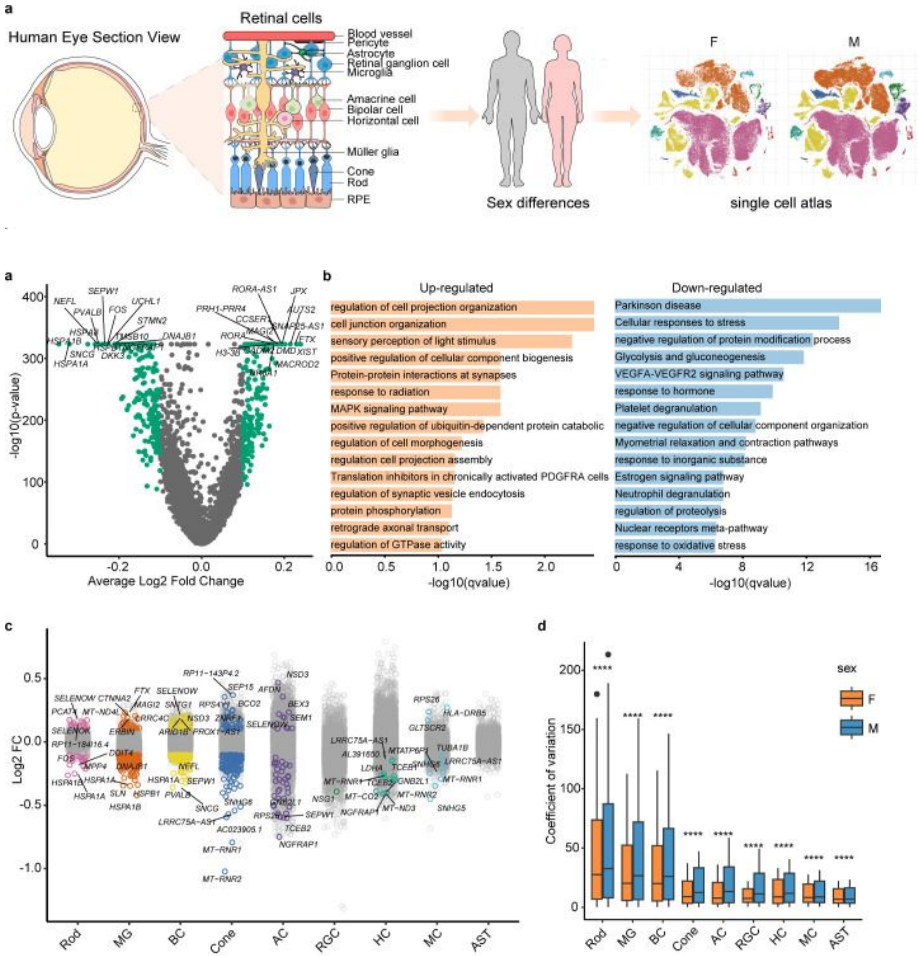
Lee et al., 2023

Same data, New question: Data Reanalysis



Example: Sex-biased gene expression in aging human retina

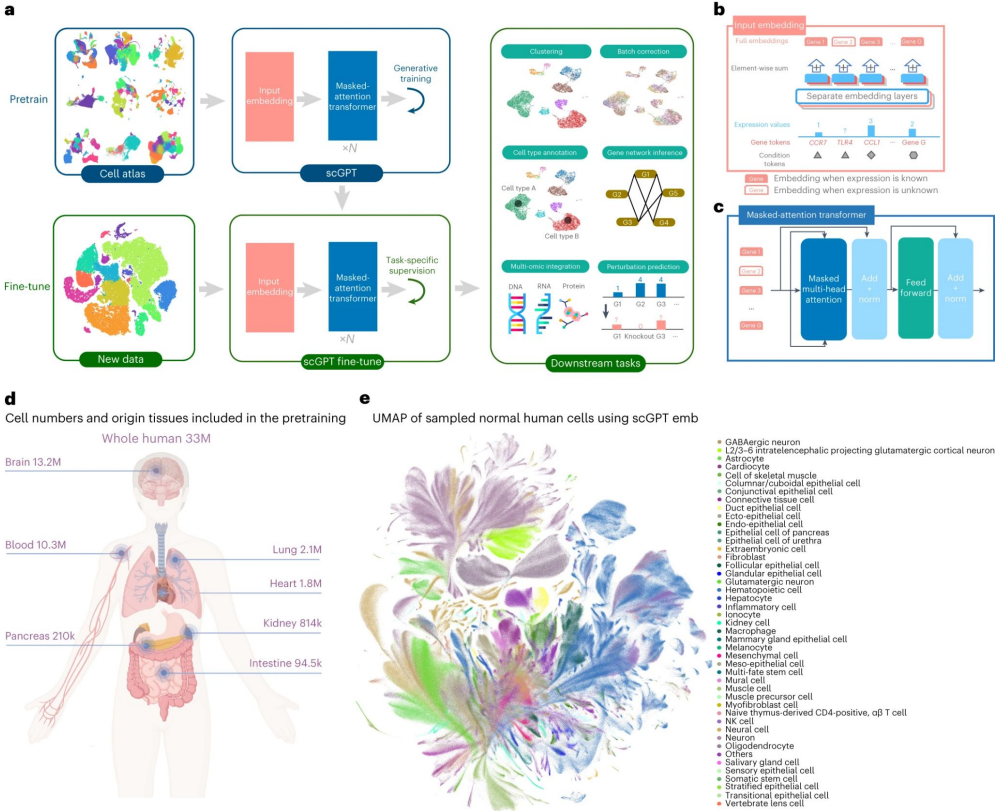
- Original datasets did not stratify by sex
- Reanalysis: compared female vs male retina
- Found female retina have higher expression of immune/inflammatory genes
- Male retina have higher metabolic pathway activity



Liu et al., 2025

Machine learning with single cells

- scGPT (2023): ChatGPT for cells
- Trained on 33 million cells across tissues
- Learnes universal gene expression patterns
- Applications: cell type annotation, gene network prediction, disease response



Cui et al., 2025

Where to get publicly available data

CZ CELL×GENE DISCOVER

Discover the mechanisms of human health

Download and visually explore data to understand the functionality of human tissues at the cellular level with Chan Zuckerberg CELL by GENE Discover (CZ CELLxGENE Discover).

UNIQUE CELLS	DATASETS	CELL TYPES
142M	2001	1085

Deeply Integrated human Single-Cell Omics data

Input gene, tissue or disease name

Statistics

Sample	Cell Type
21,415	461
Cells	**Atlas**
135,482,903	39

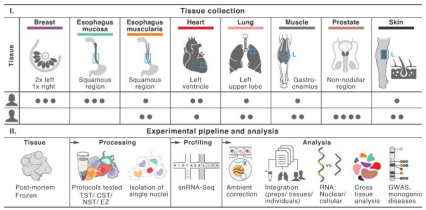
GTEx Portal

Overview

Using 25 archived, frozen tissue samples from 16 donors previously collected as part of the GTEx project, we have begun profiling tissues across the human body at single-cell resolution using snRNA-Seq.

The 8 tissues we have profiled and generated data for are the following:

1. breast
2. esophagus mucosa
3. esophagus muscularis
4. heart
5. lung
6. skeletal muscle
7. prostate
8. skin



A visual overview of the single cell data collection and analysis process

scRNA-seq Analysis Demo

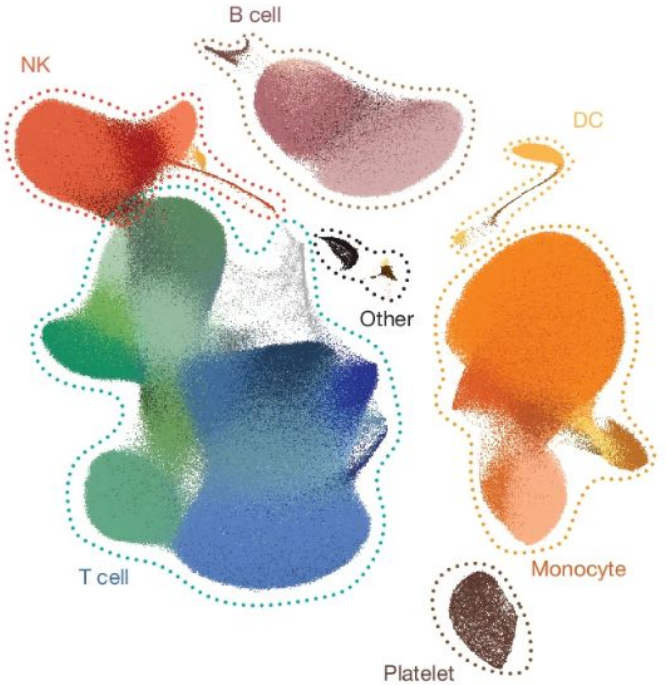
Article | Open access | Published: 29 October 2025

Multi-omic profiling reveals age-related immune dynamics in healthy adults

Qiuyu Gong, Mehul Sharma, Marla C. Glass, Emma L. Kuan, Aishwarya Chander, Mansi Singh, Lucas T. Graybuck, Zachary J. Thomson, Christian M. LaFrance, Samir Rachid Zaim, Tao Peng, Lauren Y. Okada, Palak C. Genge, Katherine E. Henderson, Elisabeth M. Dornisch, Erik D. Layton, Peter J. Wittig, Alexander T. Heubeck, Nelson M. Mukuka, Julian Reading, Garrett Strawn, Teminijesu Titus-Adewunmi, Kathleen Abadie, Charles R. Roll, ... Claire E. Gustafson

+ Show authors

Nature (2025) | Cite this article



scRNA-seq Analysis Demo

2 colab notebooks

- Biodata_human_immune_atlas.ipynb
- Biodata_human_immune_atlas_CD8.ipynb

To run yourself, make a copy in colab (google drive) or download to your python workspace of choice

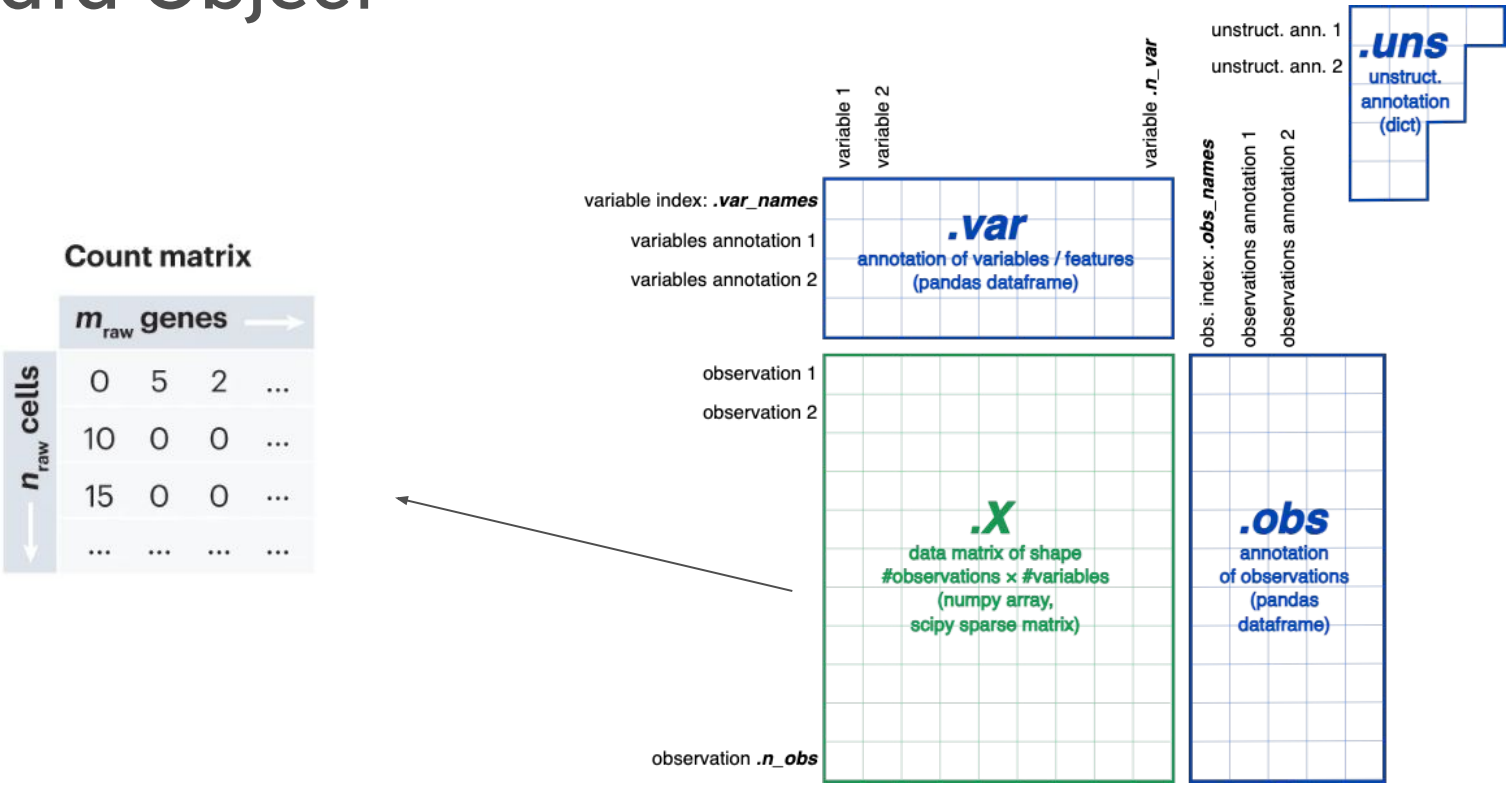
Get Data

- CELLxGENE Census to fetch public data directly into colab notebook
- Offers ability to search entire cellxgene metadata
- Since we know our dataset id of interest, we'll just download a subsmapled version
  - Original # of cells: 1.8 Million
  - Downsampled: 50k

```
# Get object (will take a few minutes)
with cellxgene_census.open_soma() as census:
  adata = cellxgene_census.get_anndata(
      census,
      organism="Homo sapiens",
      obs_coords=sampled_cell_ids,      # randomly subsampled 50k cells
      obs_column_names=[
          "soma_joinid",
          "dataset_id",
          "assay",
          "cell_type",
          "sex",
          "donor_id",
          "disease",
          "tissue",
          "tissue_general",
      ],
      var_column_names=["soma_joinid", "feature_id",
                        "feature_name", "feature_length"],
  )
```

Data Structure

Anndata Object

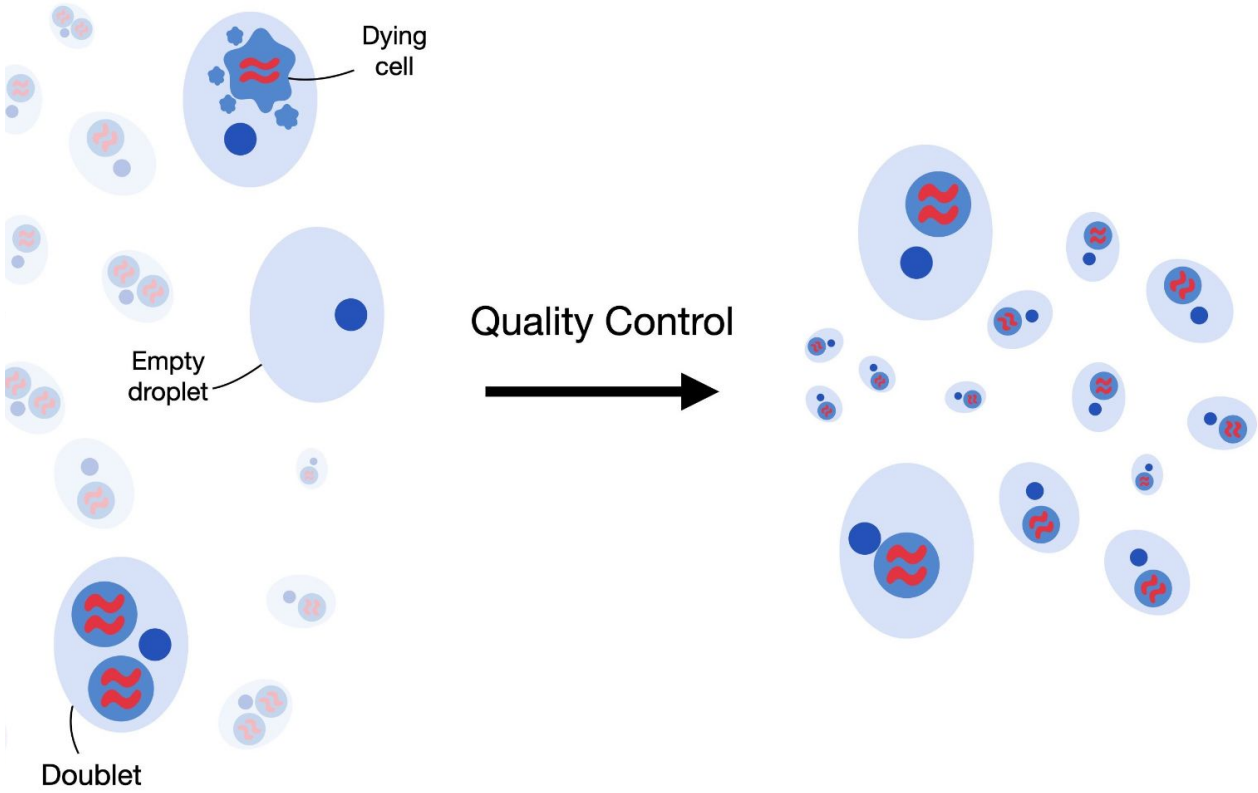


Preprocessing: Quality Control

High mt percentage

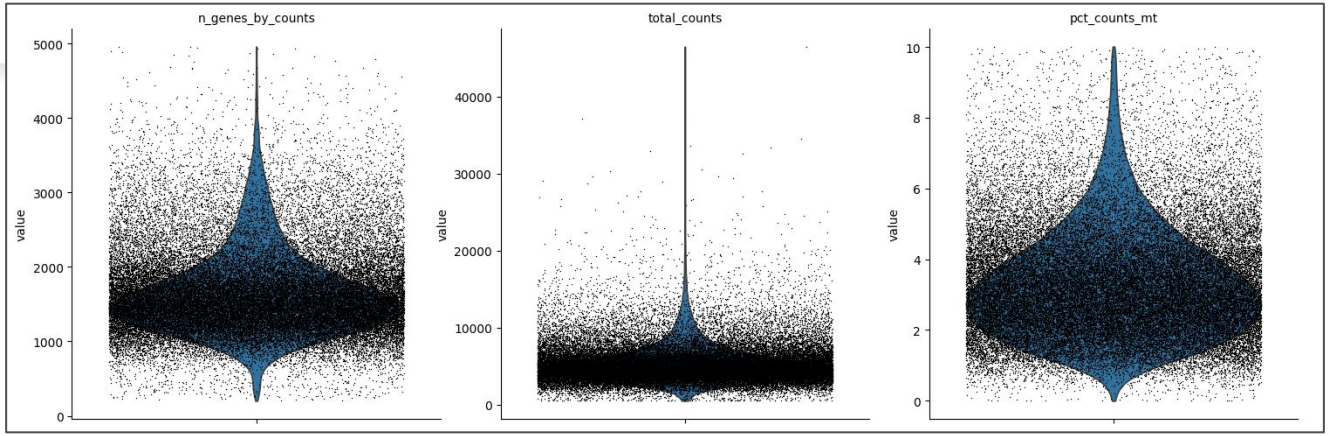
Low counts

High counts



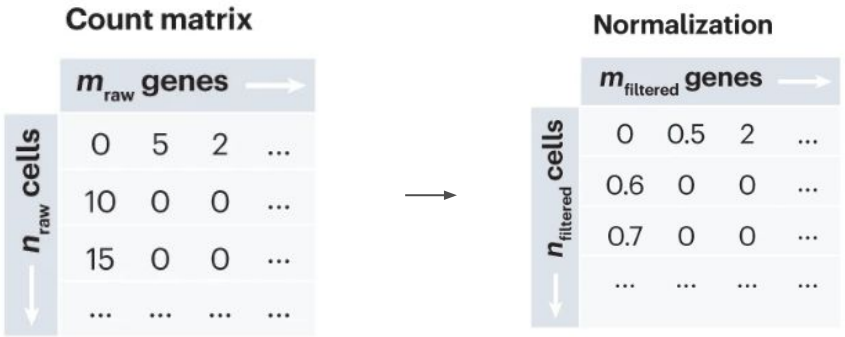
Preprocessing: Quality Control

```
# mark mitochondrial genes
adata.var["mt"] = adata.var_names.str.startswith("MT-")

# calculate qc metrics
sc.pp.calculate_qc_metrics(adata, qc_vars=["mt"], inplace=True, log1p=True)

# plot
sc.pl.violin(
    adata,
    ["n_genes_by_counts", "total_counts", "pct_counts_mt"],
    jitter=0.4,
    multi_panel=True,
)
```



Preprocessing: Normalization

```
# Save counts
adata.layers["counts"] = adata.X.copy()
# Normalize library size
sc.pp.normalize_total(adata)
# Logarithmize the data
sc.pp.log1p(adata)
```



Library Size Normalization:
Scales each cell to have the same total counts

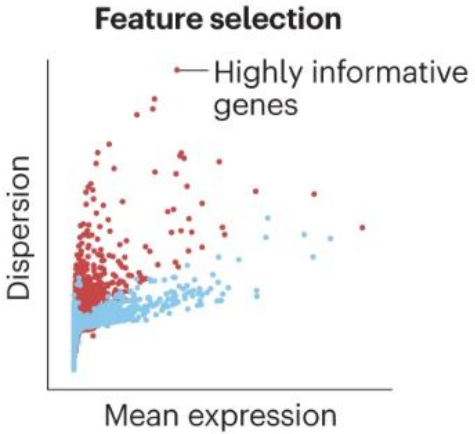
- normalized_count = (raw_count / total_counts_in_cell) × target_sum
- Cells that were deeply sequenced get scaled down and vice versa.

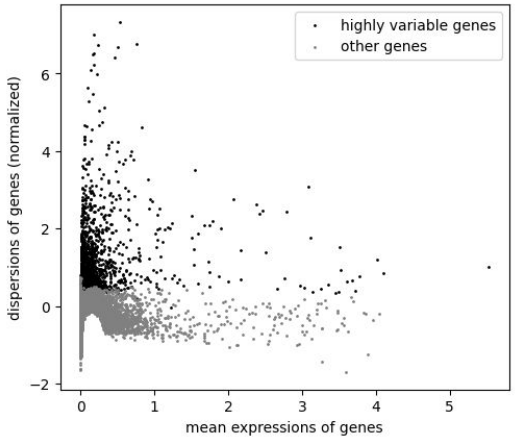
Natural Log:
log(x + 1)

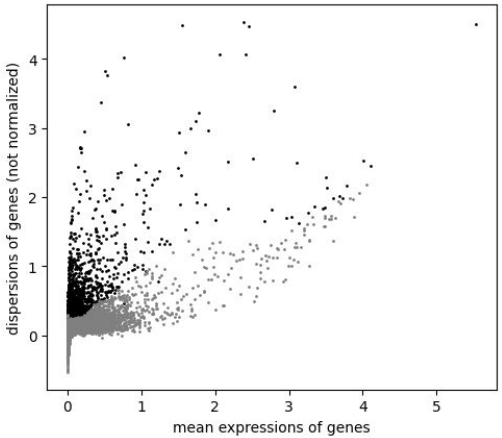
- Raw/normalized counts have a skewed distribution (a few genes have large values, most are near 0.)
- Log compresses large values and spreads out small ones

Preprocessing: Feature Selection

```
sc.pp.highly_variable_genes(adata, n_top_genes=2000, batch_key="donor_id")
sc.pl.highly_variable_genes(adata)
```





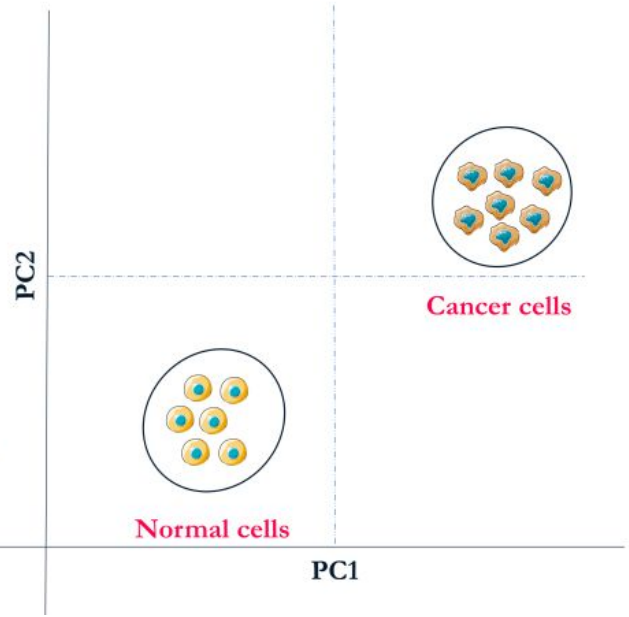


Preprocessing: Dimensionality Reduction with PCA

13 Columns

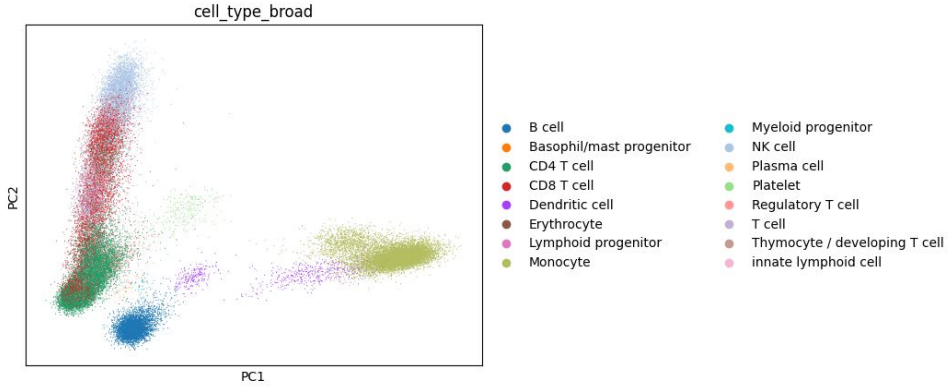
250 rows

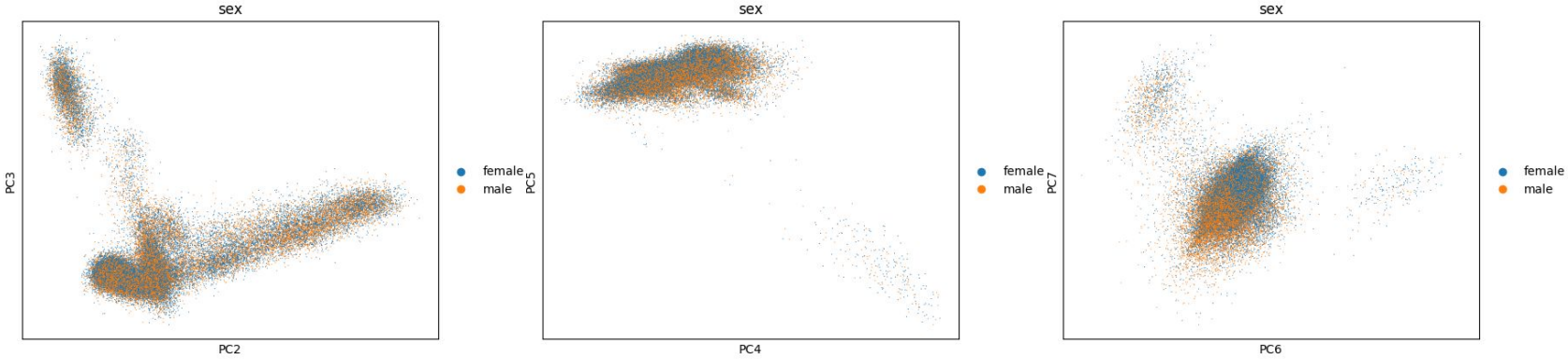
Gene	Cell1	Cell2	Cell3
ACTN1	14.1	3.2	6.5
ADAM15	10.2	2.2	2.5
….	3.6	2.5	1.1
….	9.0	11.2	8.5



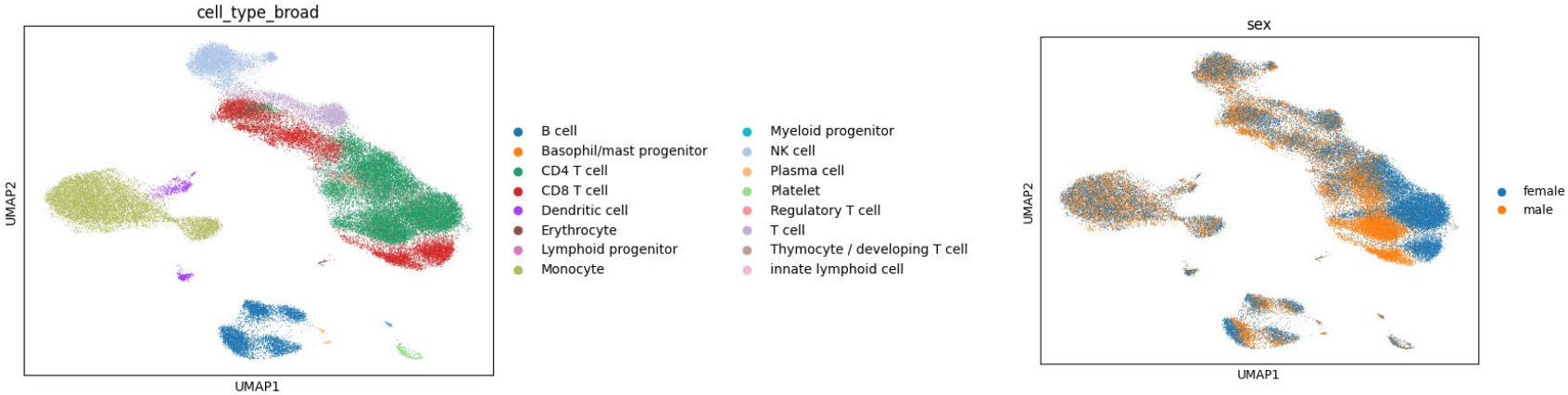
Preprocessing: Dimensionality Reduction with PCA

```
sc.tl.pca(adata)
```





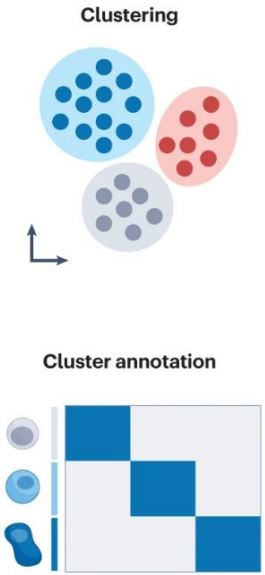
Preprocessing: Dimensionality Reduction with UMAP



- Non-linear dimensionality reduction technique
  - Captures complex relationships in only 2 dimensions
- Often used to visualize scRNA-seq data
- Not directly interpretable like PCA

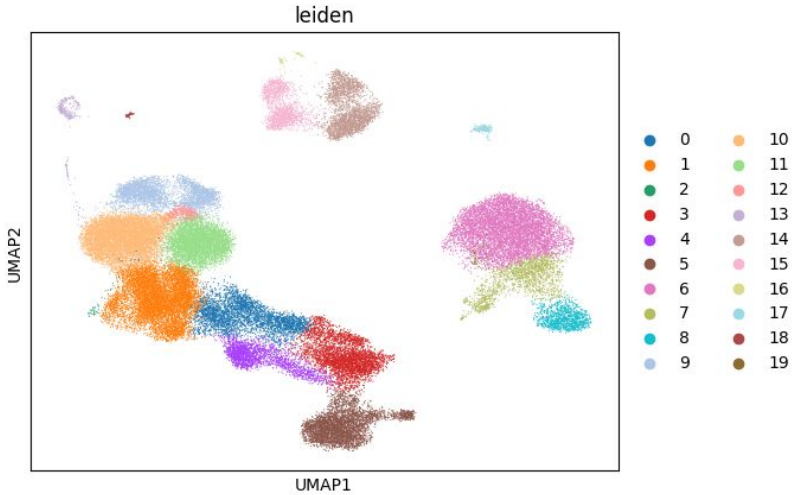
Preprocessing: Clustering

- Intuition: cells of the same type / state should express similar sets of genes
- Build neighborhood graph
  - For each cell, identify its K nearest neighbors in PC space
- Leiden clustering algorithm
  - Identifies groups of cells more densely connected to each other than to cells outside of the group
- Visualize on PCA / UMAP plot



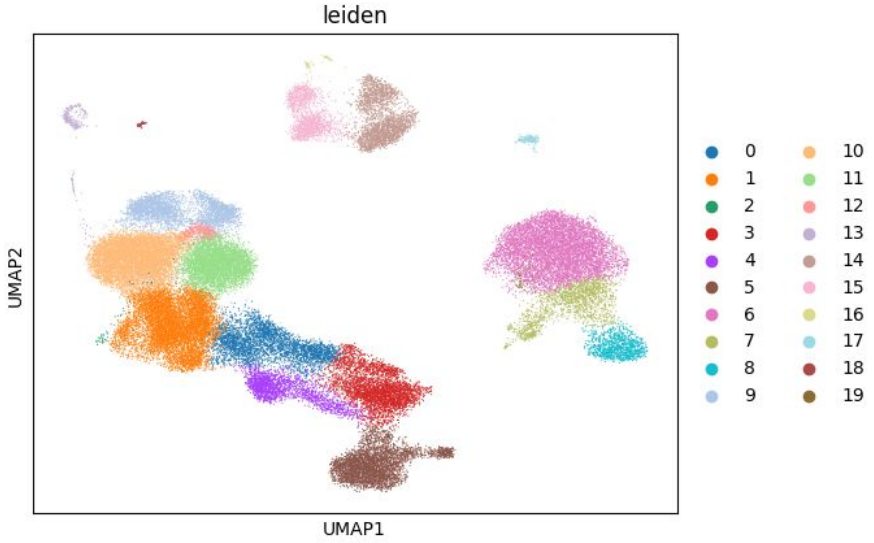
Preprocessing: Clustering

```
# build neighbor graph
sc.pp.neighbors(adata)
# run umap
sc.tl.umap(adata)
# get clusters
sc.tl.leiden(adata, flavor="igraph", n_iterations=2, resolution = 1)
# visualize clusters on umap
sc.pl.umap(adata, color=["leiden"])
```

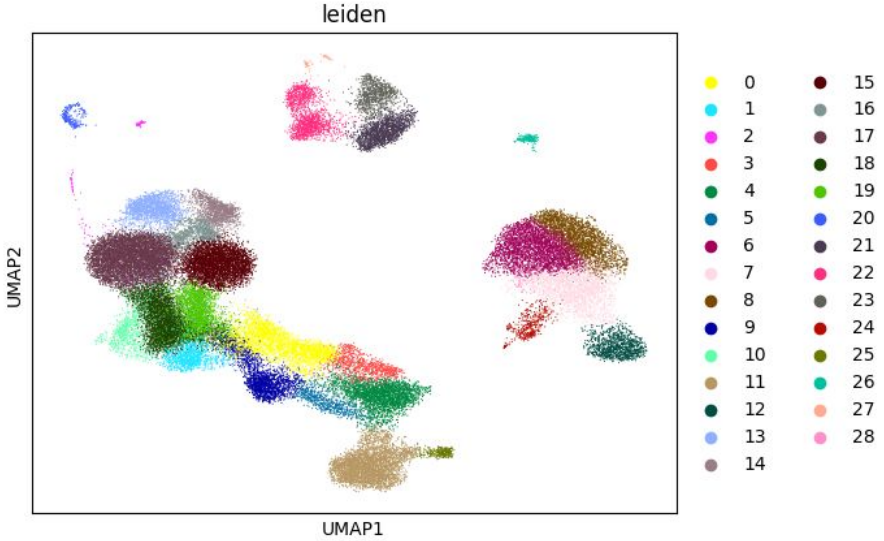


Preprocessing: Clustering

Resolution = 1
Less Clusters

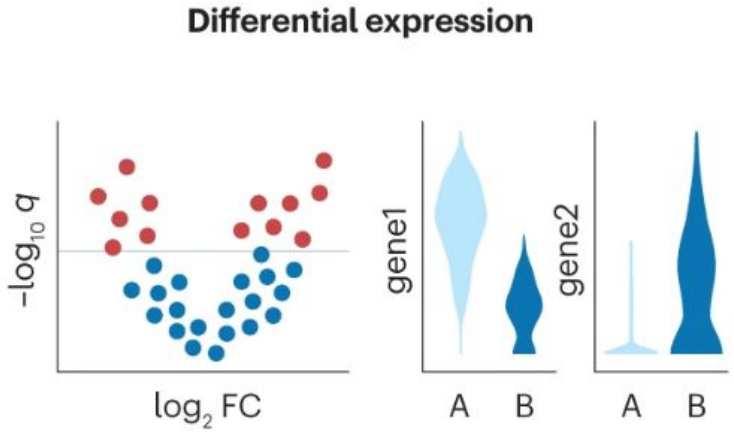


Resolution = 2
More Clusters



Analysis: Differential Expression

- Goal: Identify genes that distinguish cell types or conditions
- Examples:
  - Cluster vs all other clusters
  - Disease vs. Healthy
  - Female vs. Male
- Statistical Method
  - Wilcoxon rank-sum test (non-parametric)
- Outputs:
  - Log-fold change (effect size)
  - P-value (significance)
  - Percentage of cells expressing gene



Analysis: Differential Expression Output

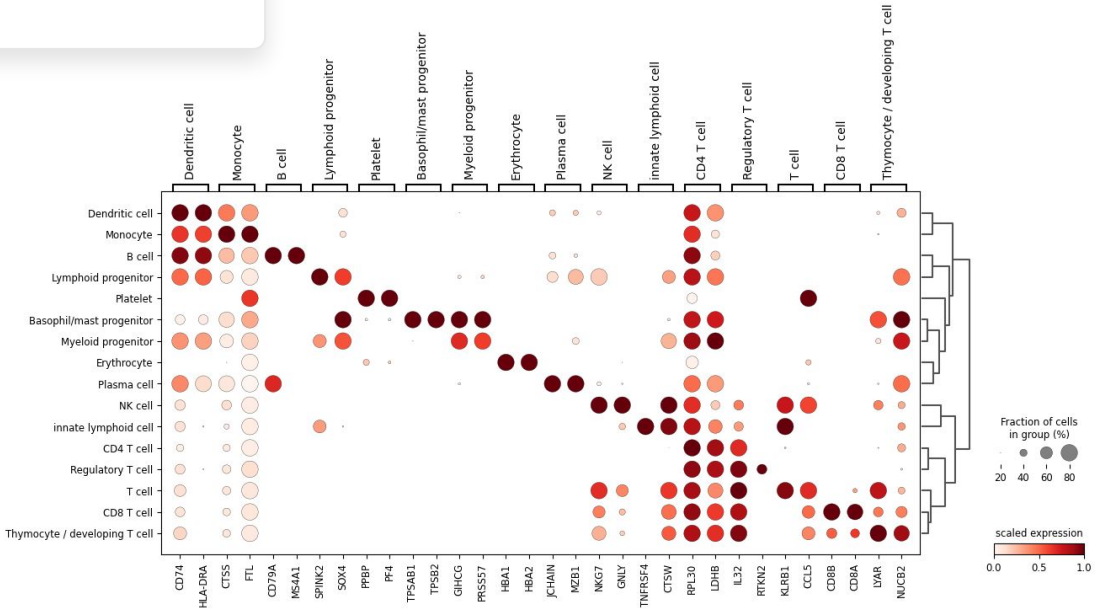
```
# Differential expression between all cell types
sc.tl.rank_genes_groups(adata, groupby='cell_type_broad', method='wilcoxon')
```

Test statistic | Effect size | Significance

	names	scores	logfoldchanges	pvals	pvals_adj
0	CD8B	116.128029	5.977100	0.0	0.0
1	CD8A	104.005249	4.815226	0.0	0.0
2	CD3D	71.395920	1.555180	0.0	0.0
3	CD3G	67.991516	1.497581	0.0	0.0
4	IL32	67.835747	1.686253	0.0	0.0

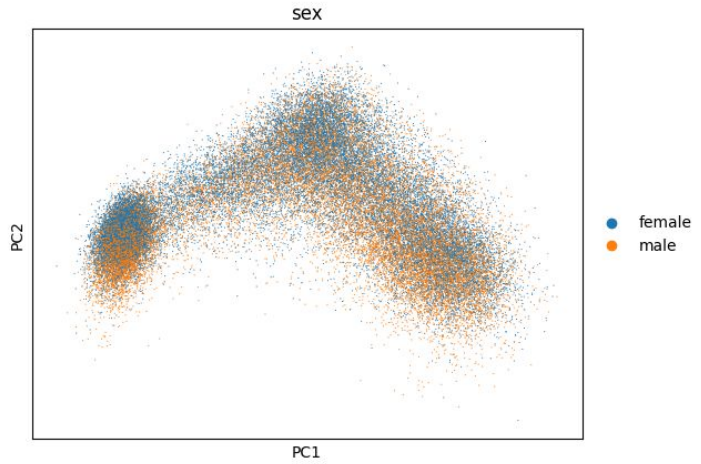
Analysis: Differential Expression

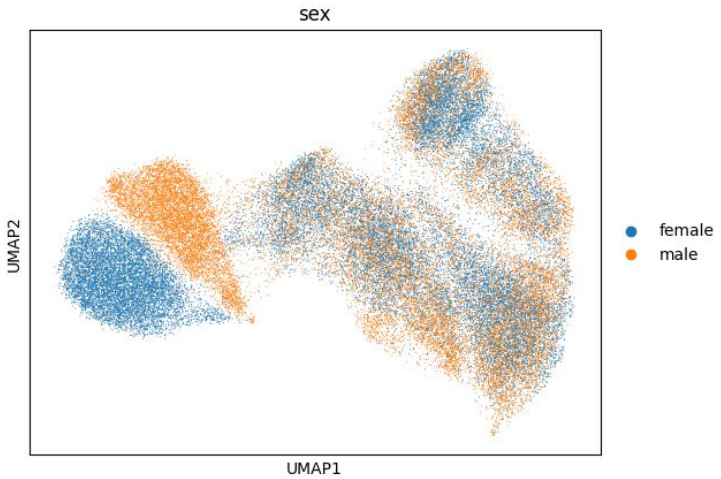
```
# Dot plot of top 2 genes per cell type
sc.pl.rank_genes_groups_dotplot(adata, groupby='cell_type_broad',
                                standard_scale='var',
                                colorbar_title='scaled expression',
                                dot_max=0.8, dot_min=0.2,
                                n_genes=2)
```



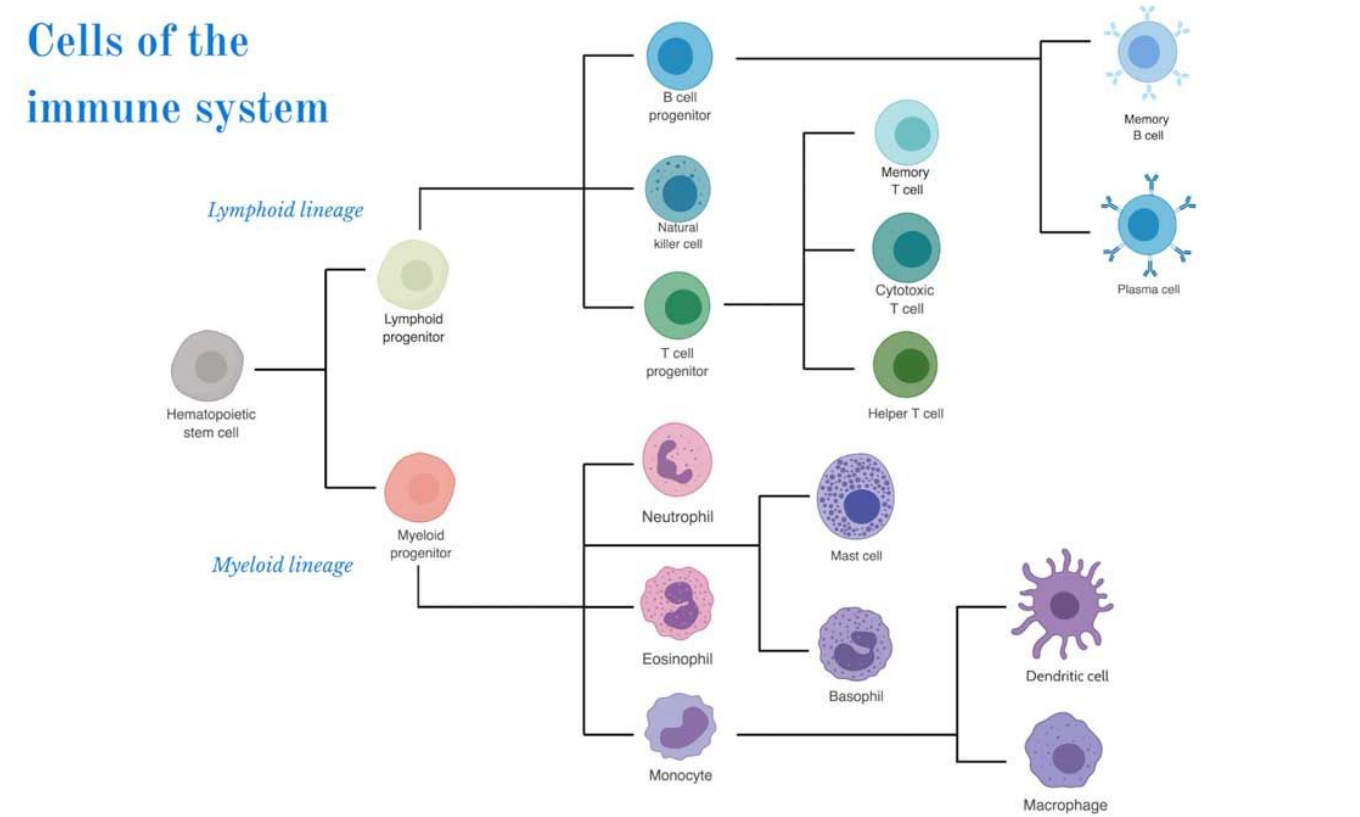
New Analysis: Sex differences in CD8 T Cells

- Download just CD8 T Cells and repeat preprocessing steps
  - Quality control
  - Normalization
  - Find Variable Features
  - Dimensionality Reduction



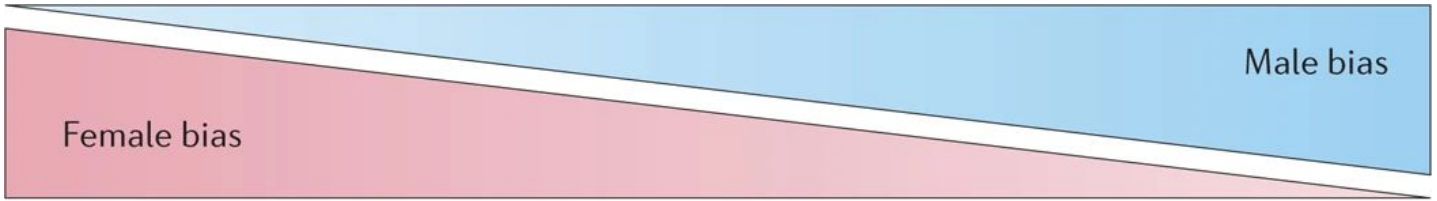


About immune cell types



iBiology, 2019

Sex Differences in Immune responses



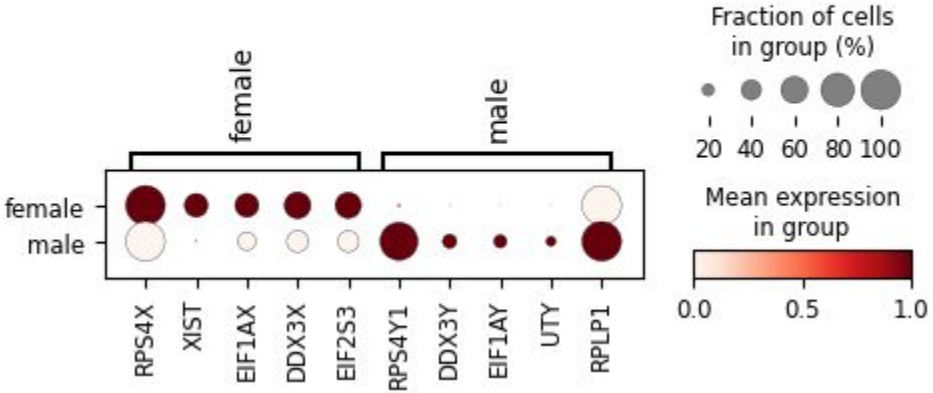
Autoimmune diseases	Infectious diseases		Non-reproductive cancers
• Graves disease • Hashimoto thyroiditis • Multiple sclerosis • Rheumatoid arthritis • Systemic lupus erythematosis • Type 1 diabetes	• HIV • Influenza • Toxoplasmosis • Legionella • Malaria • Zika	• Ebola • MERS • Hepatitis B • Tuberculosis • Leptospirosis • Campylobacter • Schistosomiasis • Amebiasis • Aspergillosis	• Bladder • Bowel • Kidney • Leukaemia • Liver • Lung • Malignant melanoma • Oesophagus • Stomach

Nature Reviews | Immunology

Klein & Flanagan, 2016

Analysis: Differential Expression between Males and Females

Top hits are X and Y chr genes



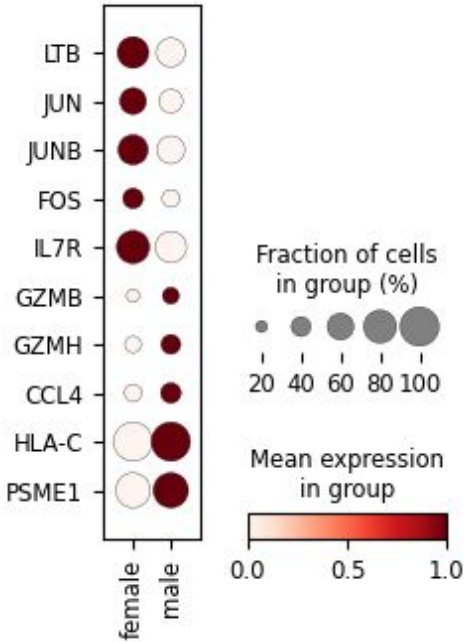
Graves, 2014

Analysis: Differential Expression between Males and Females

Immune-related genes that came up in top 30 F or M DE genes

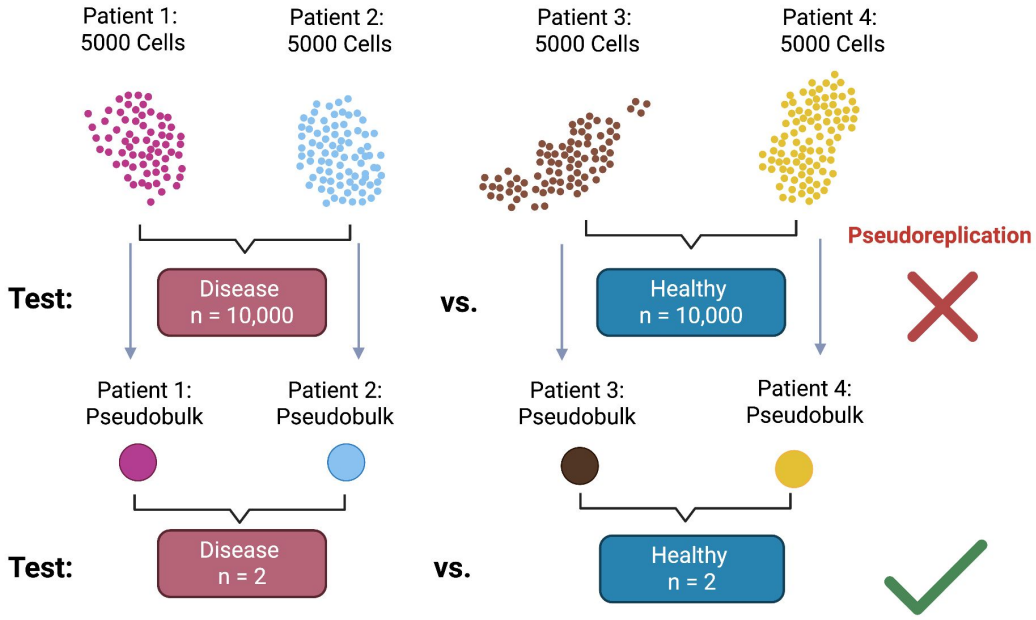
- Females: IL7r - marks memory, long-lived t cells
- Males: Genes related to cytotoxic cd8 signature
  - GZM: capable of killing target cells
  - HLA-C, PSME1 are related to antigen presentation (activated cd8s)

This is an initial exploratory analysis, so we shouldn't make any biological conclusions on this yet. This result can be confounded by one or two donors with high expression, cell composition differences in F vs. M samples, and pseudoreplication.



A note about differential expression methods

- Cells from the same patient are not independent observations, but standard DE tests treat them like they are → pseudoreplication
  - Inflates statistical significance
- Solution: Pseudobulk
  - Aggregate cells by sample within each cell type



Takeaways

Statistical Concepts

- Data Normalization
- Dimensionality reduction
  - PCA
  - UMAP
- Graphs and clustering
- Differential expression tests

scRNA-seq Analysis

- Single cell methods have enabled atlas generation, new cell types, and insights into disease and treatment effects
- Lots of public data out there!
- Differential expression analysis on pseudobulked cells is more statistically reliable

Questions?