



Single Cell Sequencing for Drug Discovery: Applications and Challenges

Sarah Middleton
Computational Biologist, GSK





Single Cell Sequencing for Drug Discovery: Applications and Challenges

Sarah Middleton
Computational Biologist, GSK

Computational biology in pharma R&D

Target identification

- Genetic disease association (GWAS)
- Disease vs healthy gene expression
- Pathway/network analysis

Compound screening

- Tractability prediction
- Assay development/analysis
- Rational design

Pre-clinical development

- Biomarker prediction
- Animal model data analysis
- Patient stratification

Post-market research

- Drug repurposing
- Combination therapy



Target validation

- Perturbation 'omics analysis (e.g. CRISPR)
- *In vitro / in vivo* expression analysis

Lead optimization & Candidate selection

- Mechanism of action
- Pharmacodynamics

Clinical trials

- Patient stratification
- Medical image analysis

Computational biology in pharma R&D

Target identification

- Genetic disease association (GWAS)
- Disease vs healthy gene expression
- Pathway/network analysis

Compound screening

- Tractability prediction
- Assay development/analysis
- Rational design

Pre-clinical development

- Biomarker prediction
- Animal model data analysis
- Patient stratification

Post-market research

- Drug repurposing
- Combination therapy



Target validation

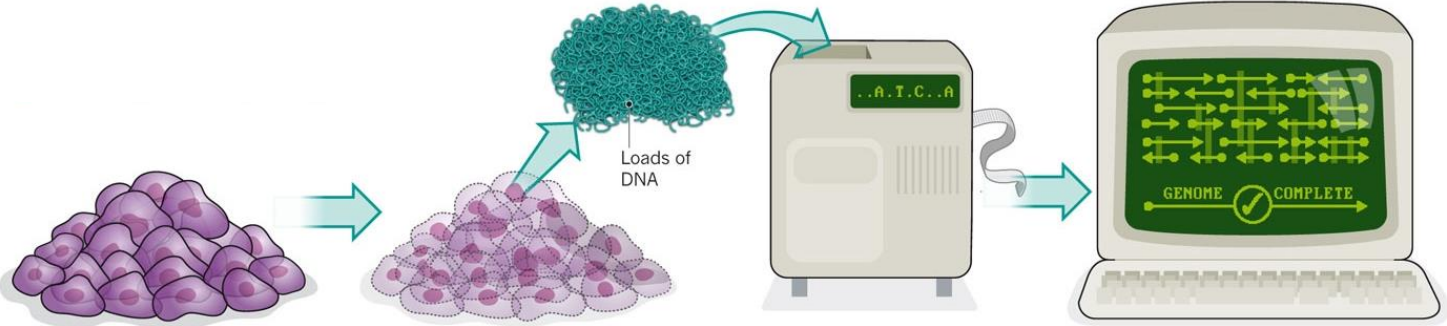
- Perturbation ‘omics analysis (e.g. CRISPR)
- *In vitro* / *in vivo* expression analysis

Lead optimization & Candidate selection

- Mechanism of action
- Pharmacodynamics

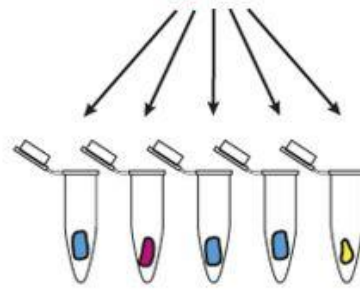
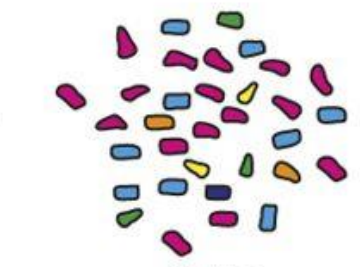
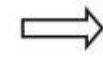
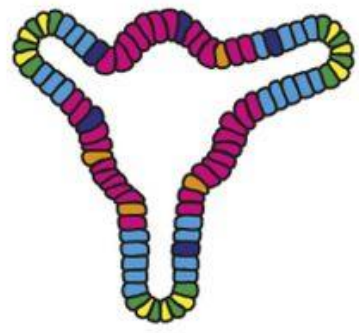
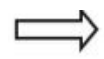
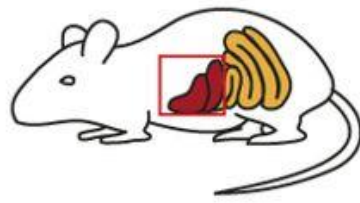
Clinical trials

- Patient stratification
- Medical image analysis









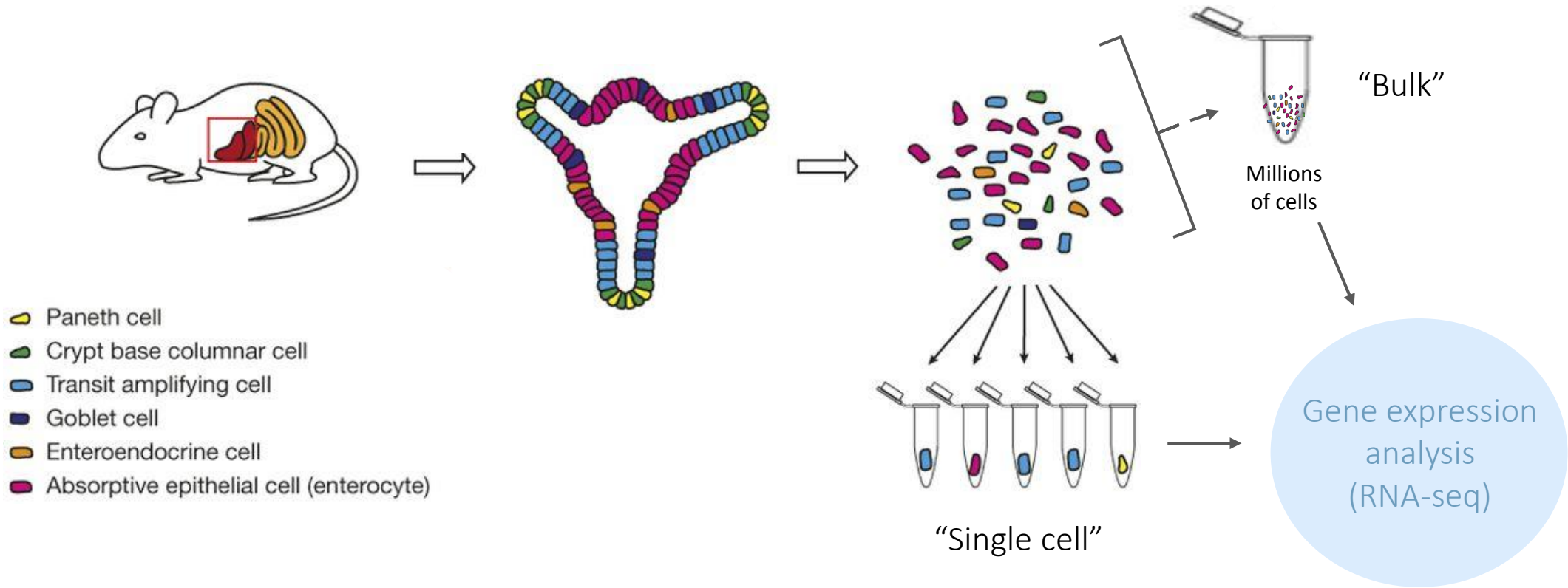
‘Omics data:

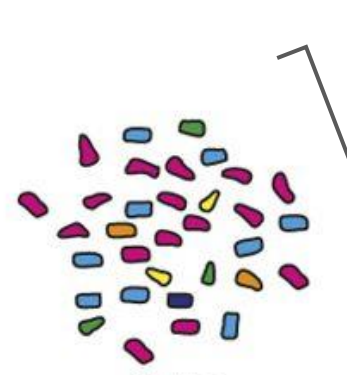
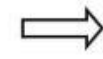
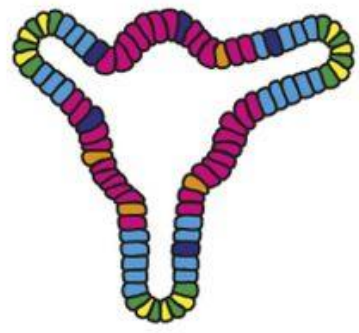
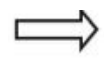
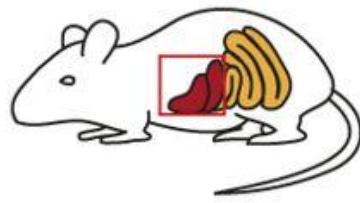
- Genome (DNA-seq)
- Transcriptome (RNA-seq; aka “gene expression”)
- Proteome (mass spec)
- Metabolome, interactome, ...



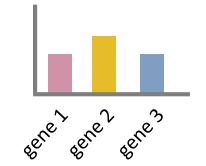
Gene expression analysis (RNA-seq)

-  Paneth cell
-  Crypt base columnar cell
-  Transit amplifying cell
-  Goblet cell
-  Enteroendocrine cell
-  Absorptive epithelial cell (enterocyte)





“Bulk”



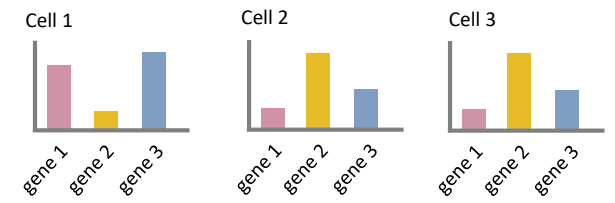
Millions of cells

Gene expression analysis (RNA-seq)

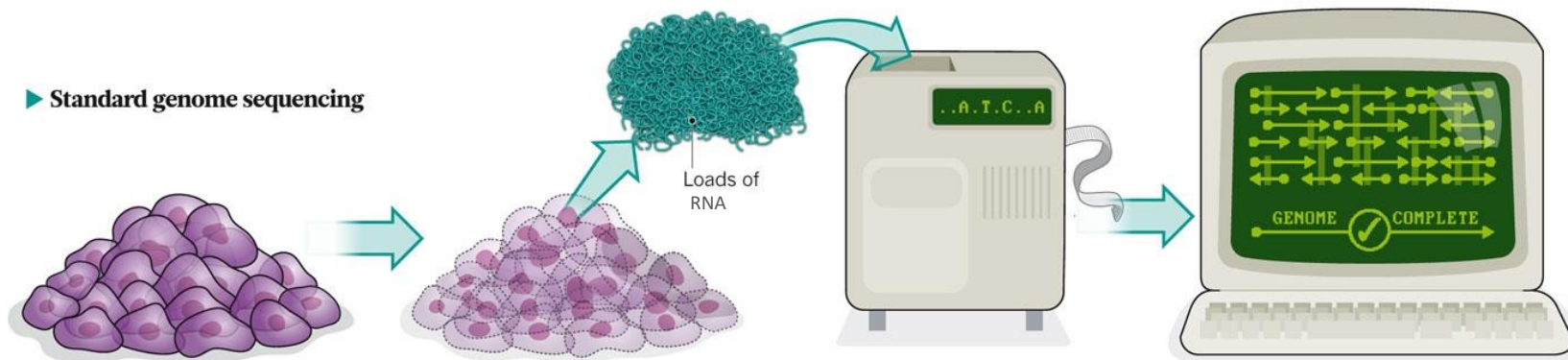
- Paneth cell
- Crypt base columnar cell
- Transit amplifying cell
- Goblet cell
- Enteroendocrine cell
- Absorptive epithelial cell (enterocyte)



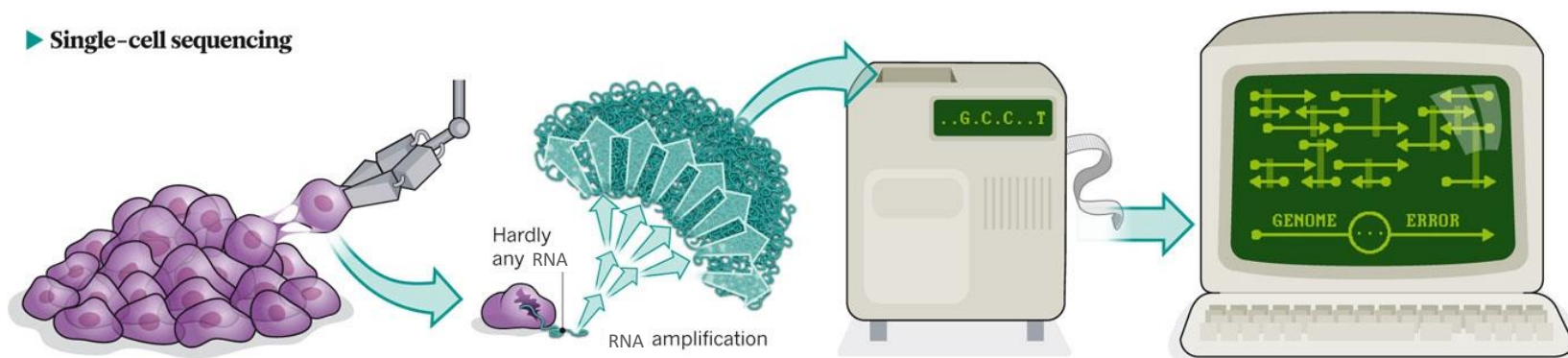
“Single cell”



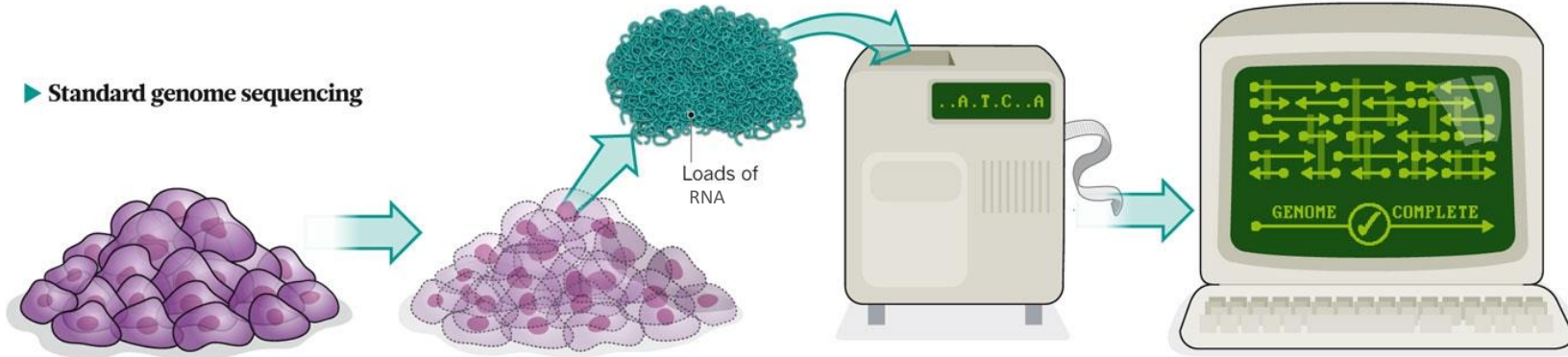
► Standard genome sequencing



► Single-cell sequencing



► **Standard genome sequencing**



RNA molecules
(~100-300k/cell)



Reverse Transcription:
~50% capture efficiency ↓

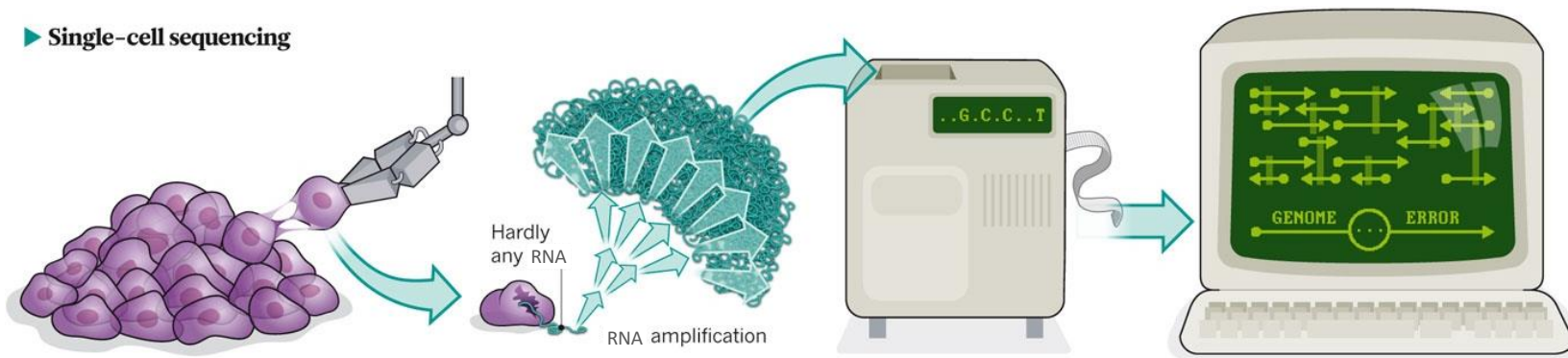


Several other steps:
80-90% efficiency each ↓

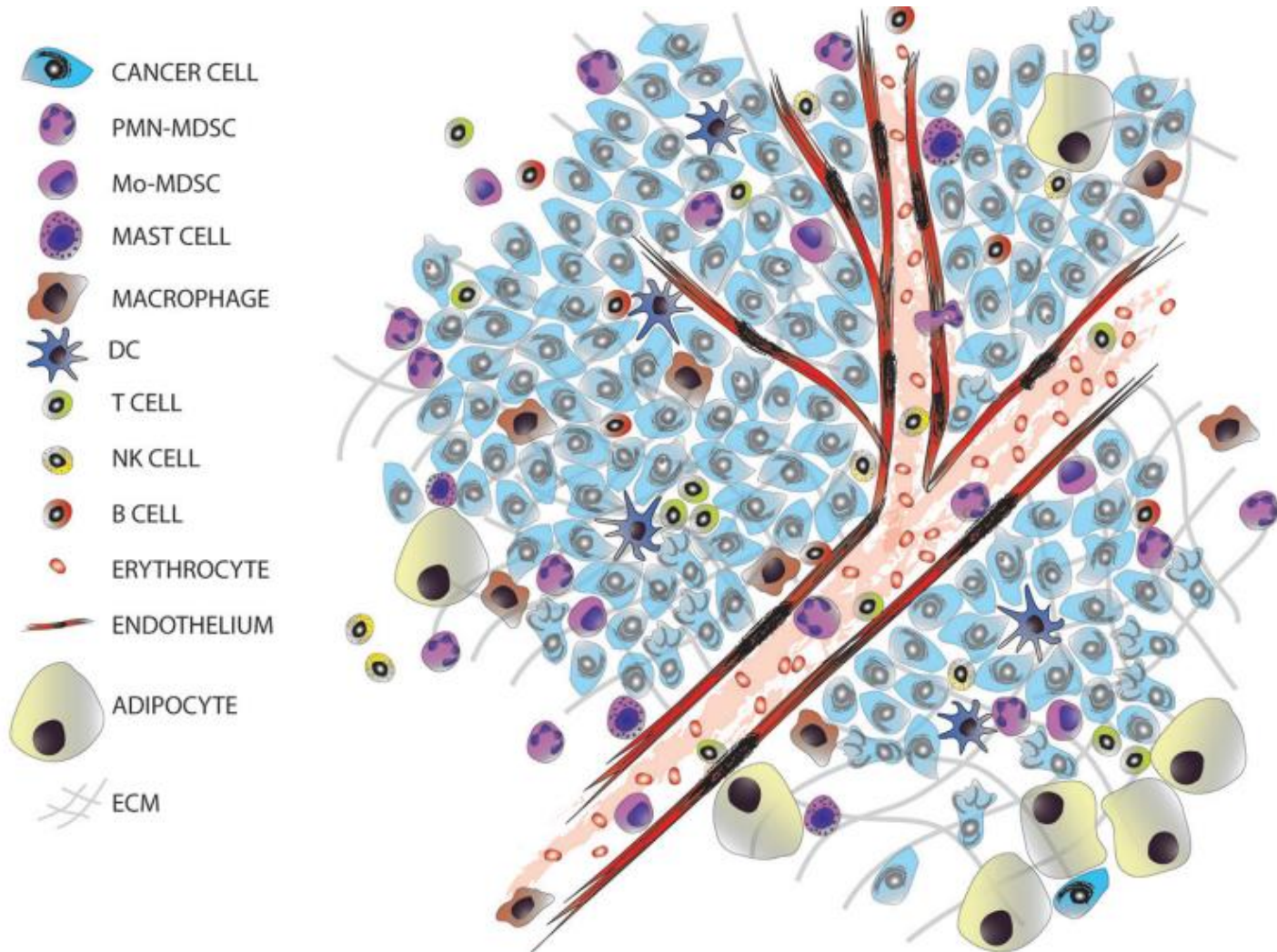


Overall: 10-20%

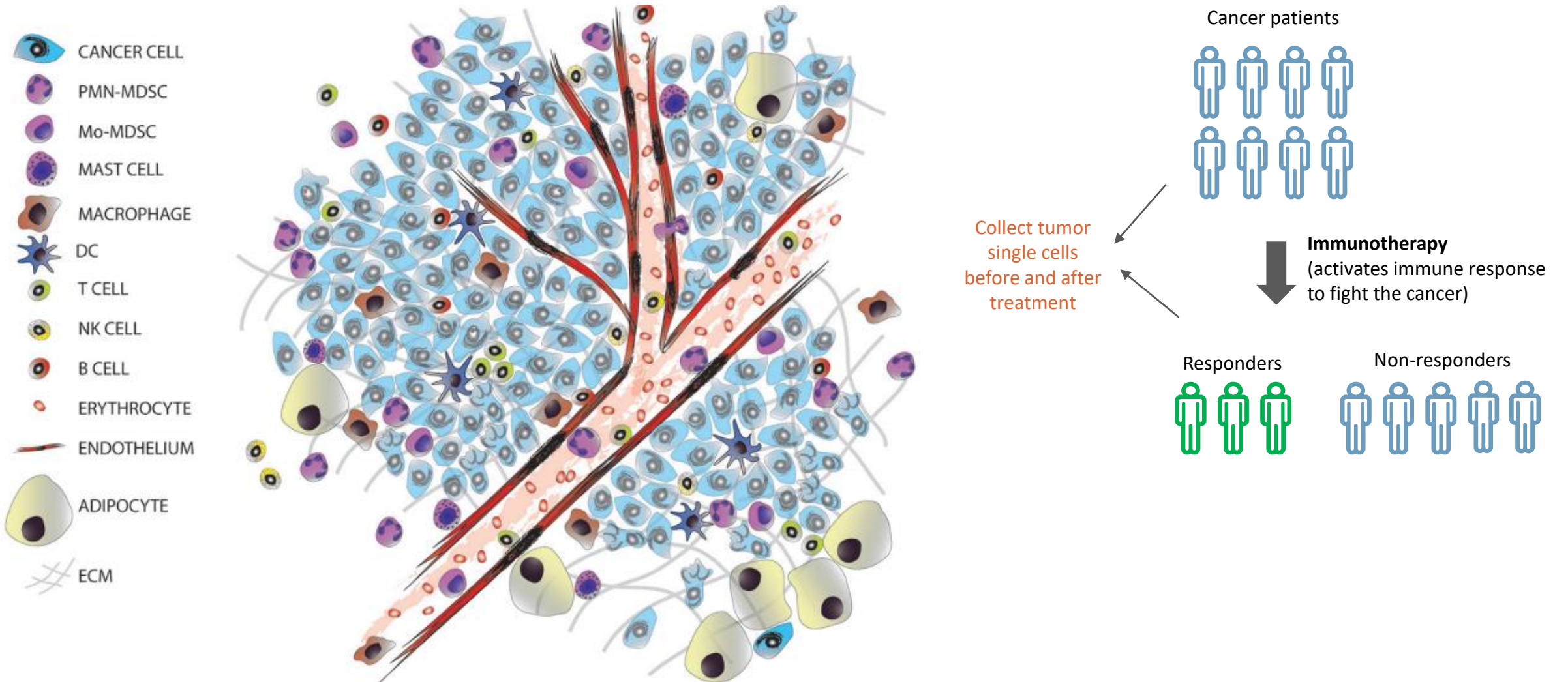
► **Single-cell sequencing**



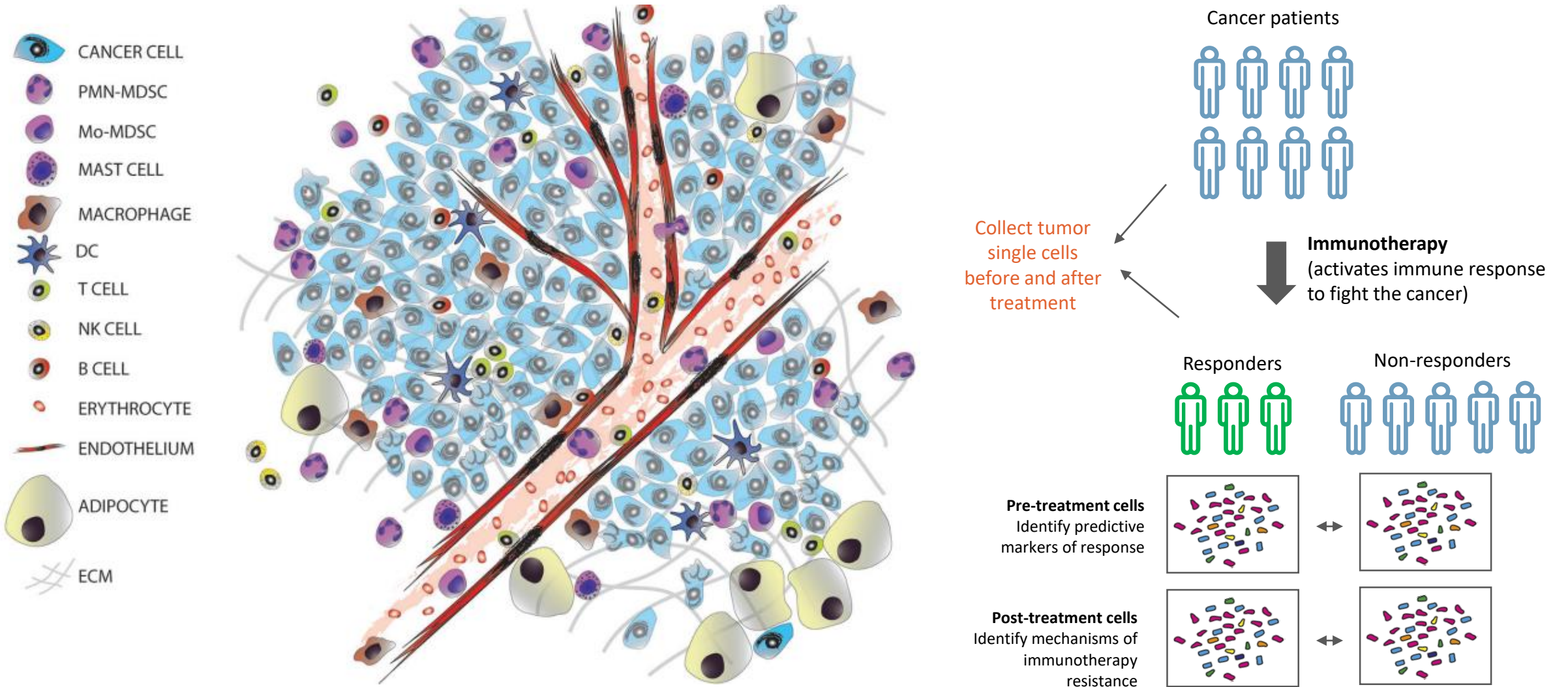
Single cell application: mechanisms of therapy-resistant cancer



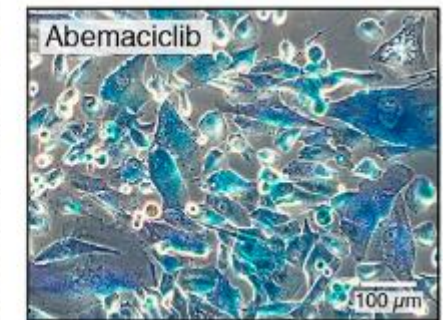
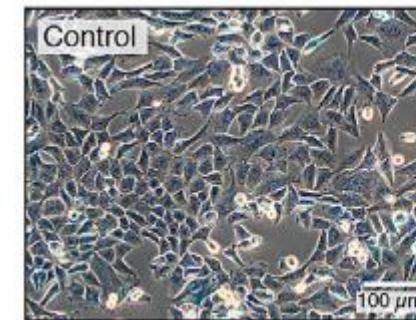
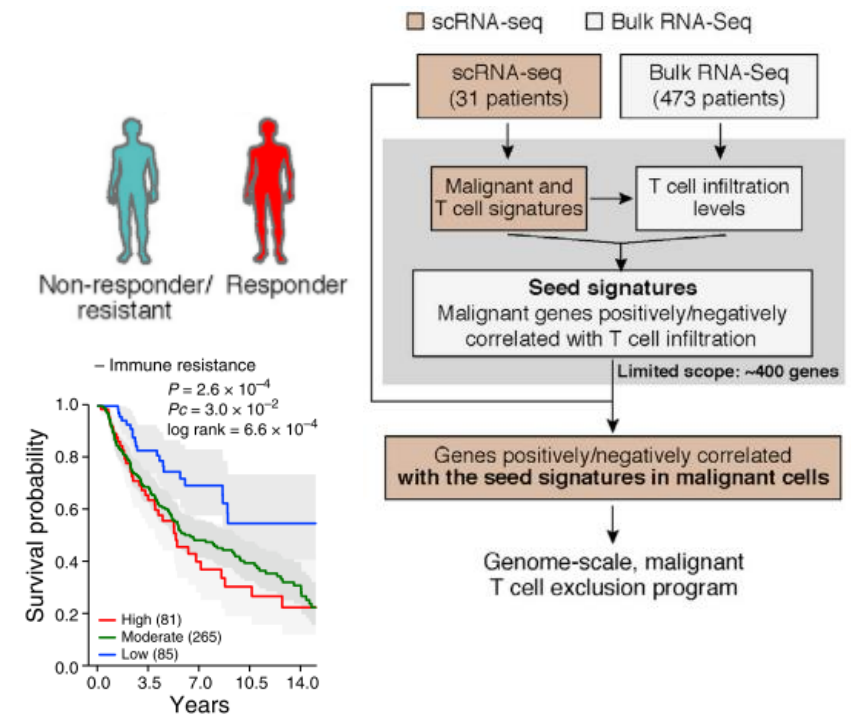
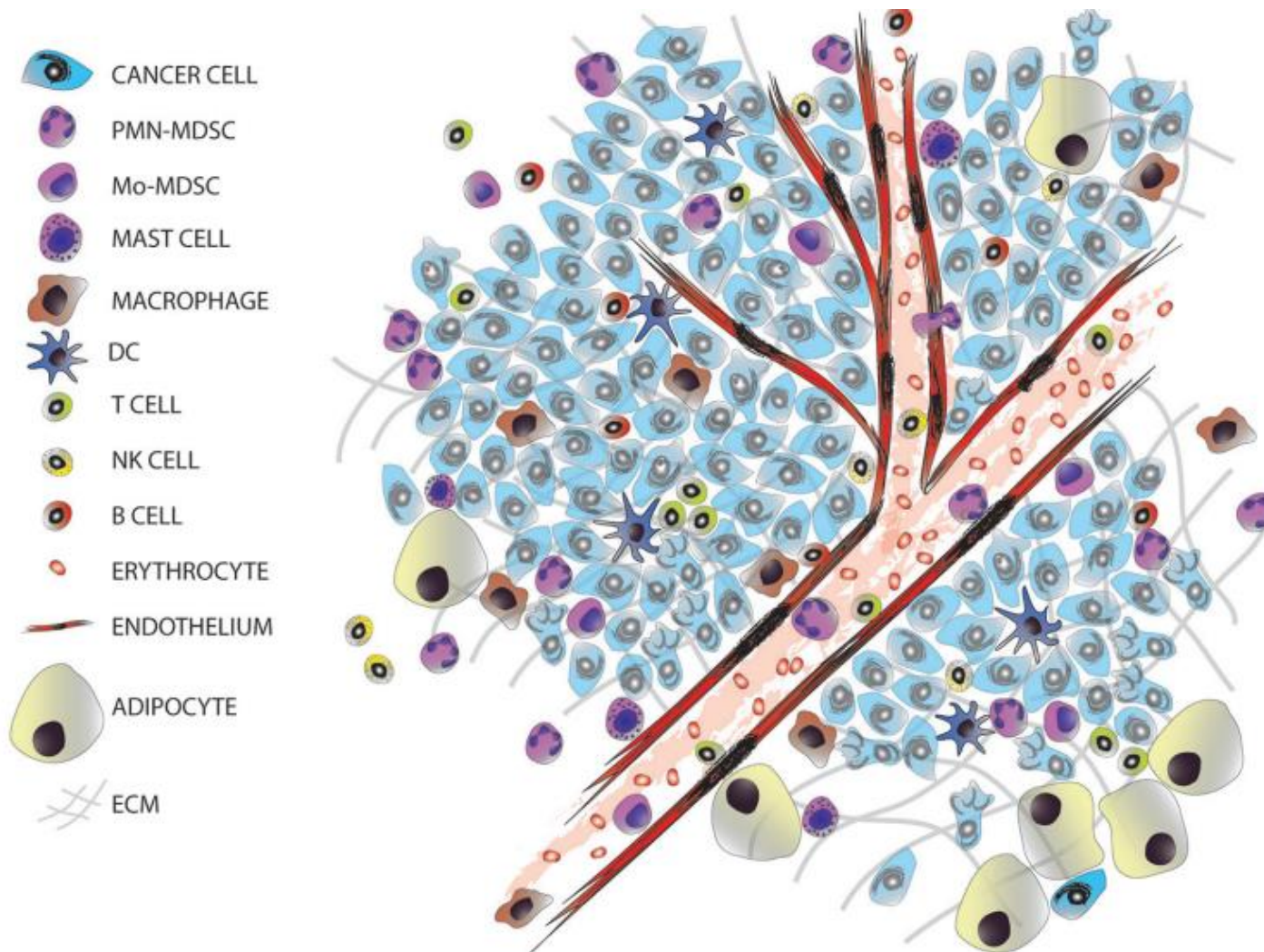
Single cell application: mechanisms of therapy-resistant cancer



Single cell application: mechanisms of therapy-resistant cancer

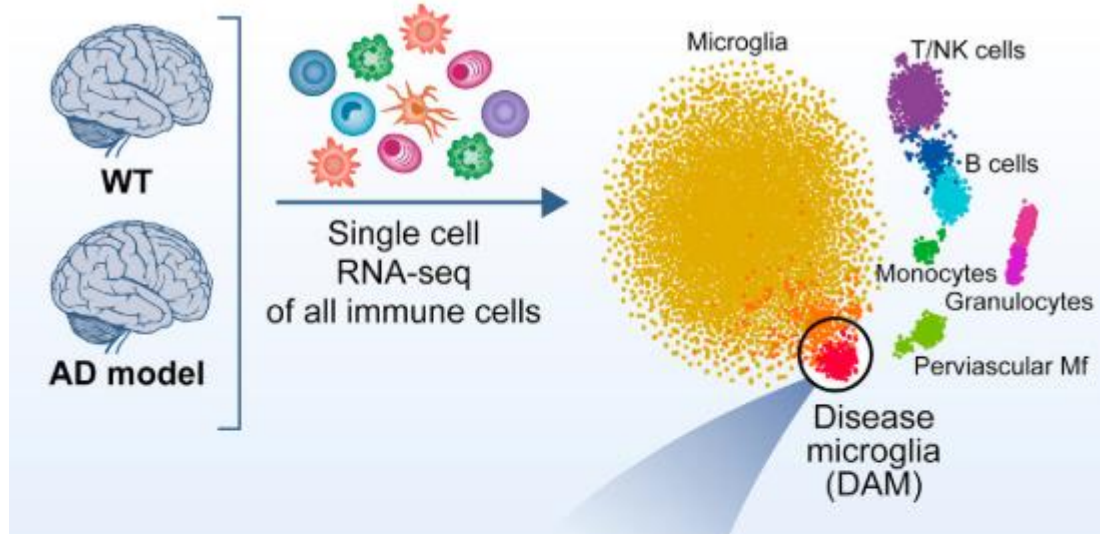


Single cell application: mechanisms of therapy-resistant cancer

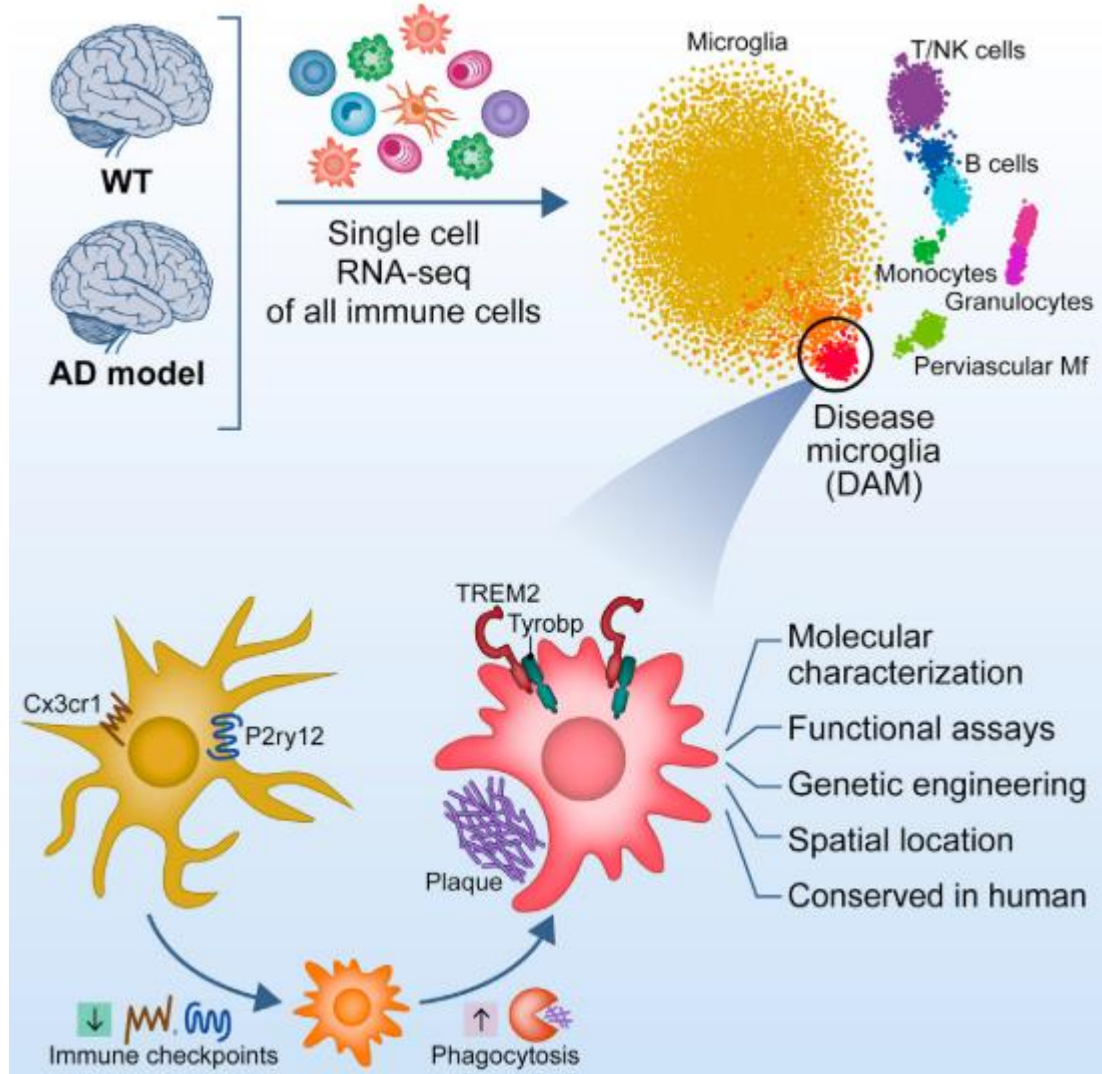


CDK4/6 inhibitors as potential combination therapy to improve response

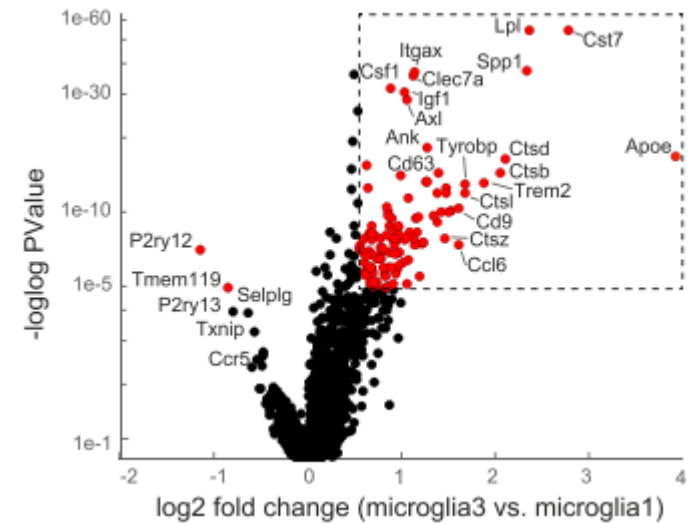
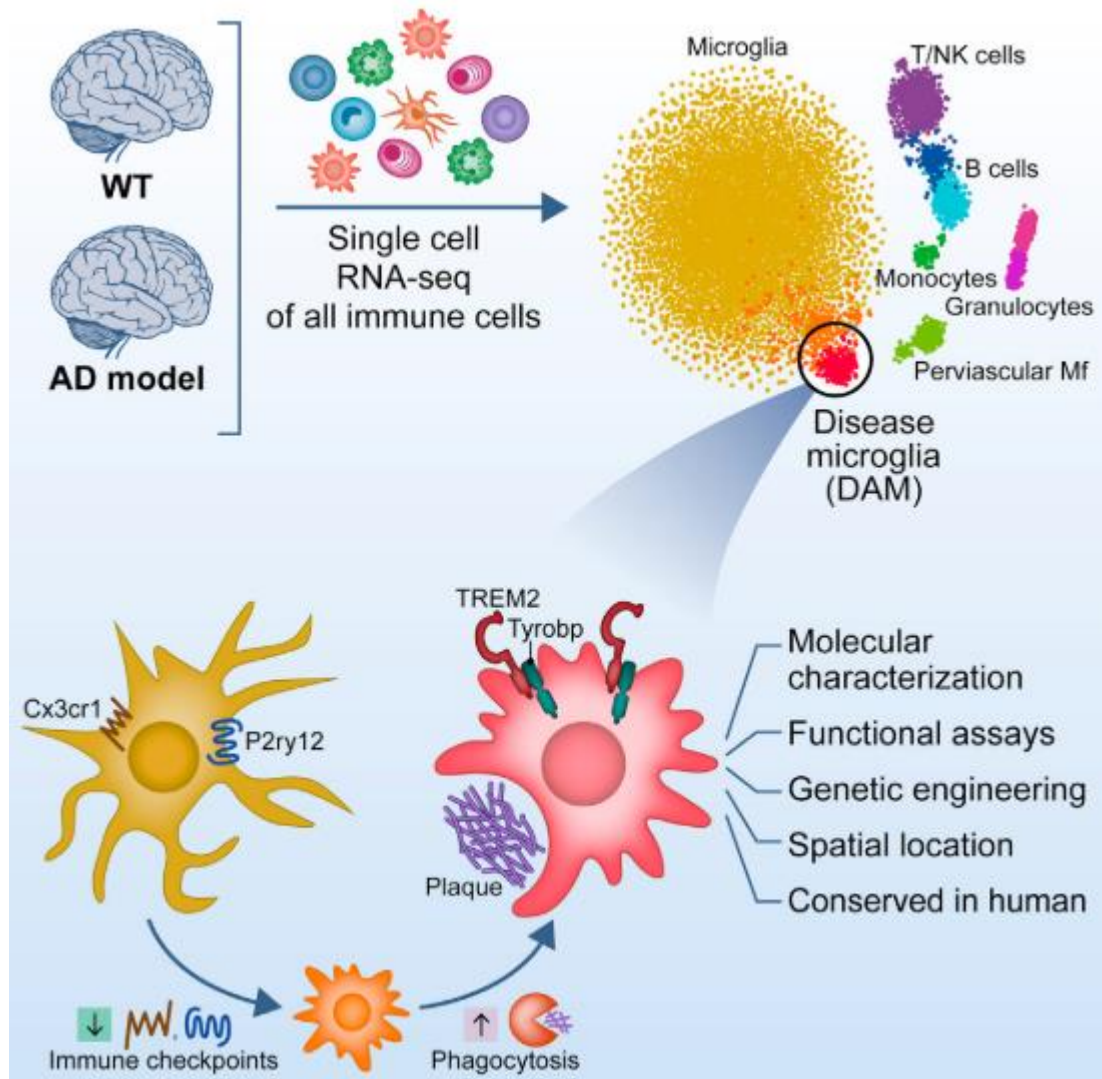
Single cell application: Disease-specific cell types in Alzheimer's



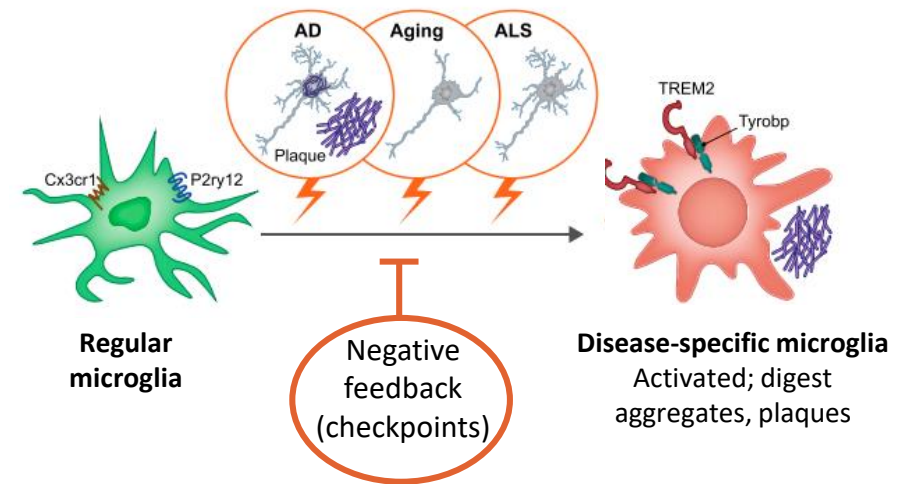
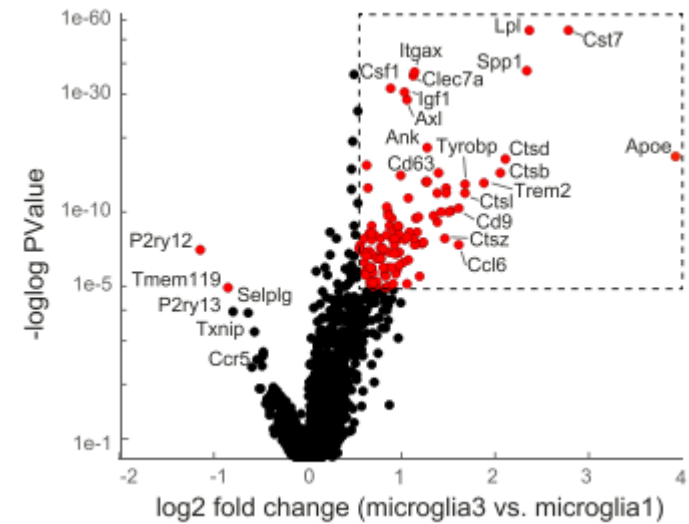
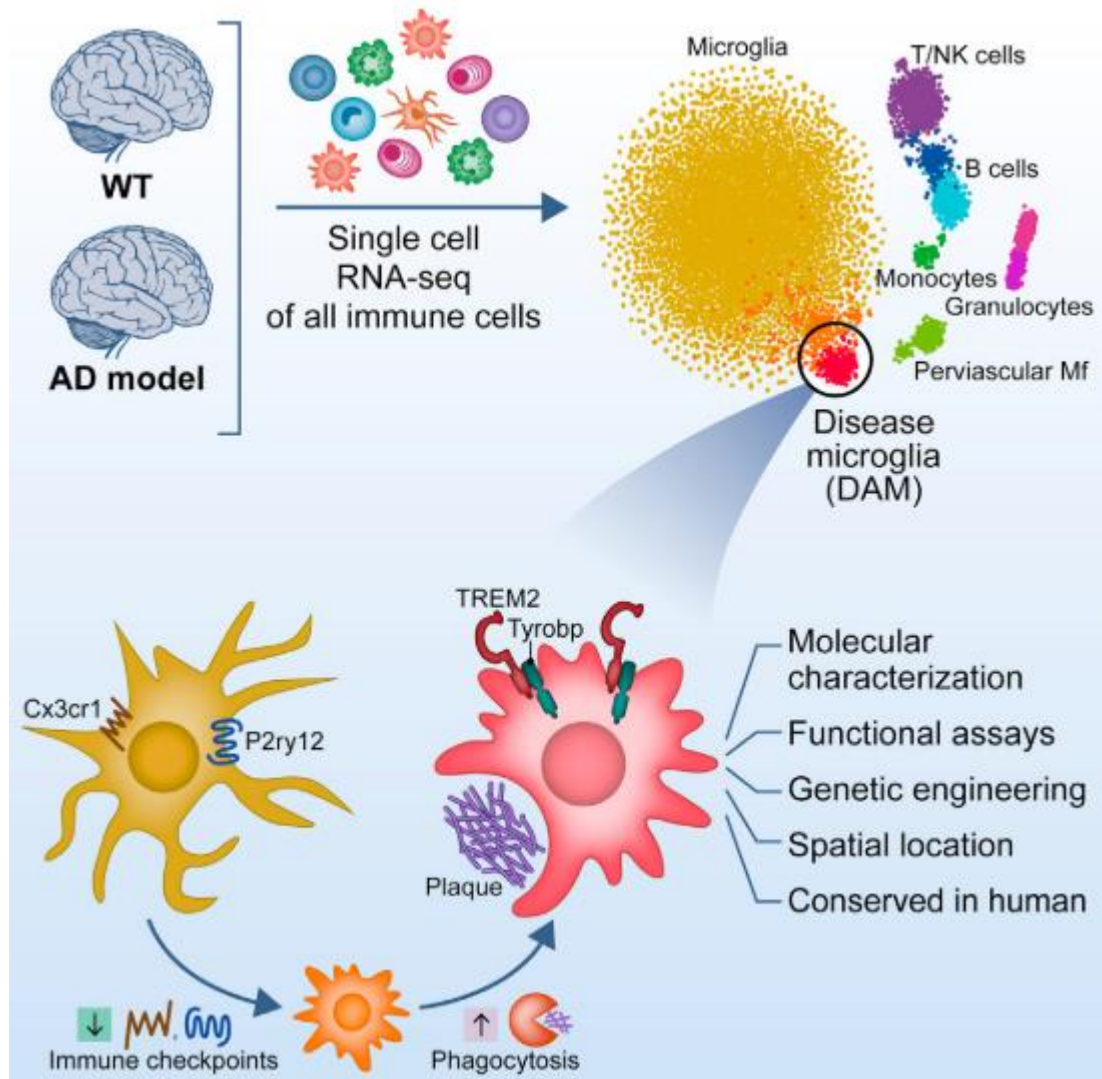
Single cell application: Disease-specific cell types in Alzheimer's



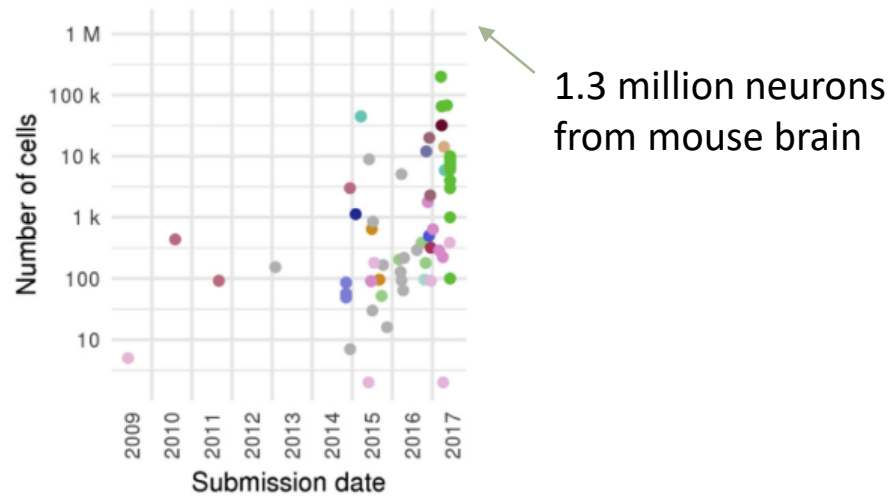
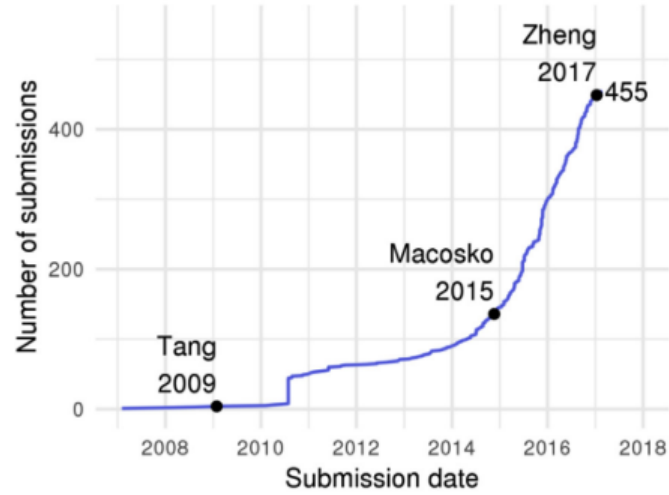
Single cell application: Disease-specific cell types in Alzheimer's



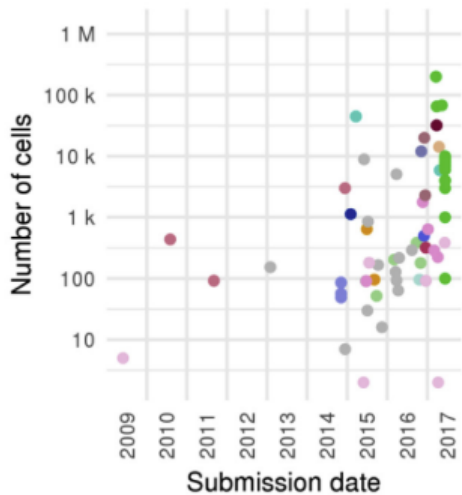
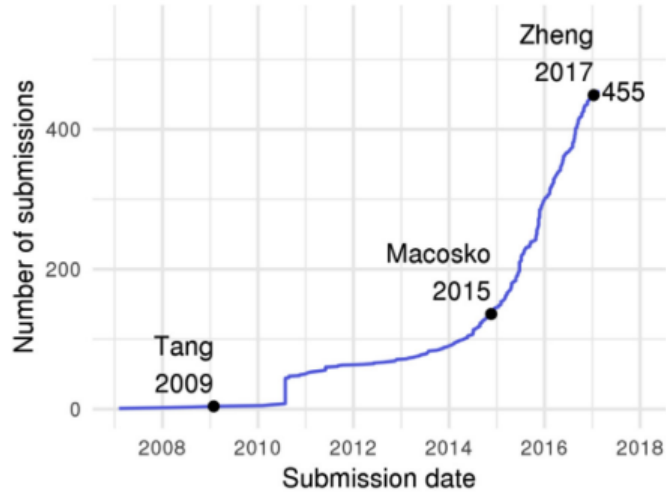
Single cell application: Disease-specific cell types in Alzheimer's



Scaling up single cell analysis



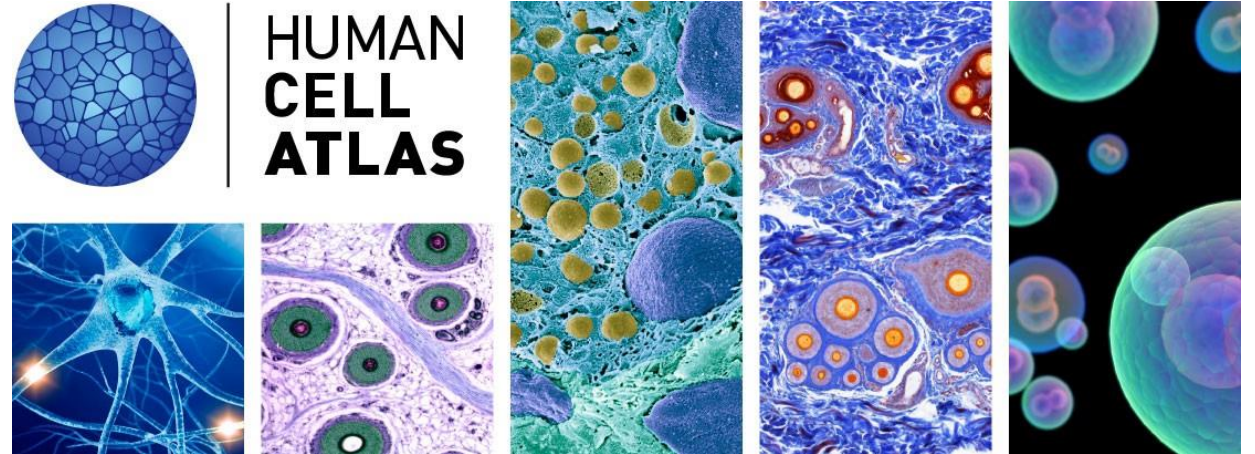
Scaling up single cell analysis



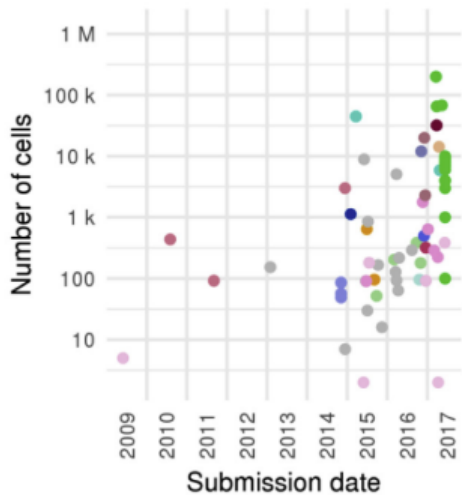
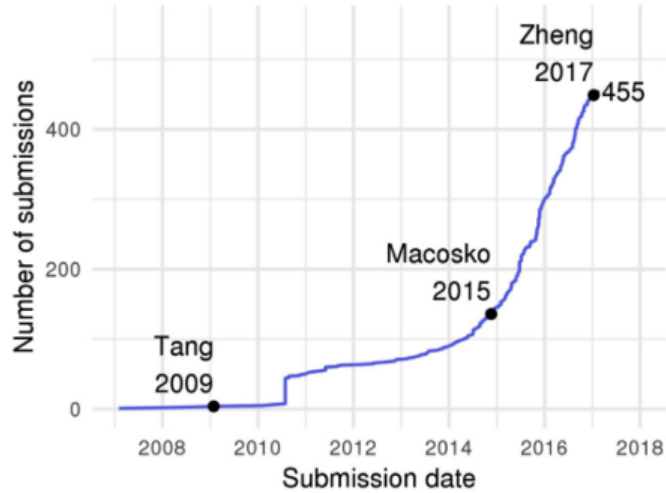
1.3 million neurons from mouse brain

Goal: create a map of all cells in the human body

Several trillion cells; hundreds of cell types



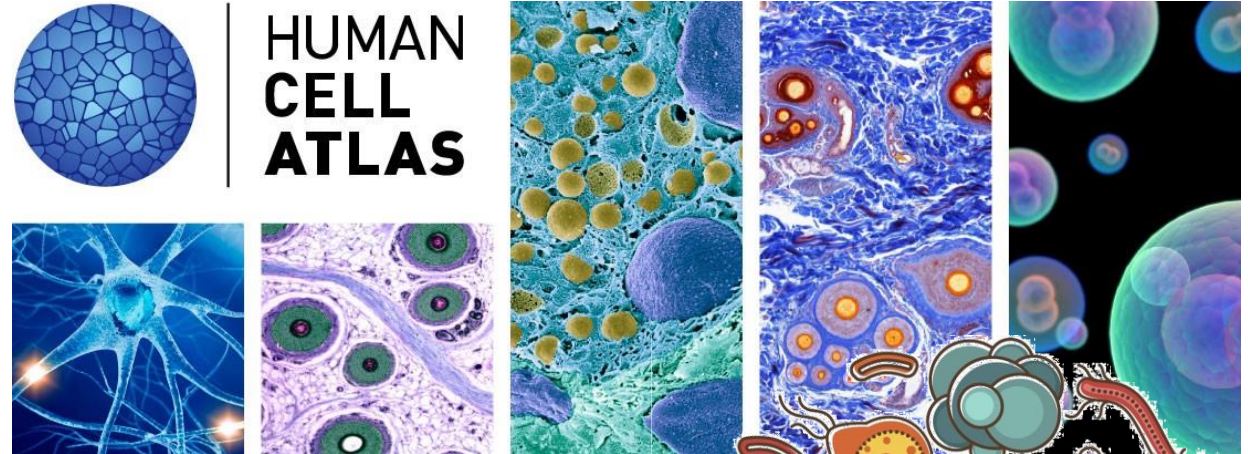
Scaling up single cell analysis



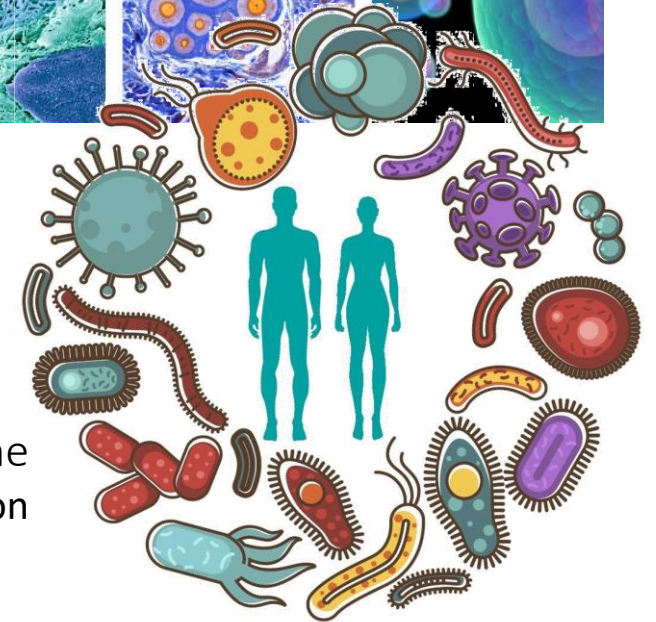
1.3 million neurons from mouse brain

Goal: create a map of all cells in the human body

Several trillion cells; hundreds of cell types

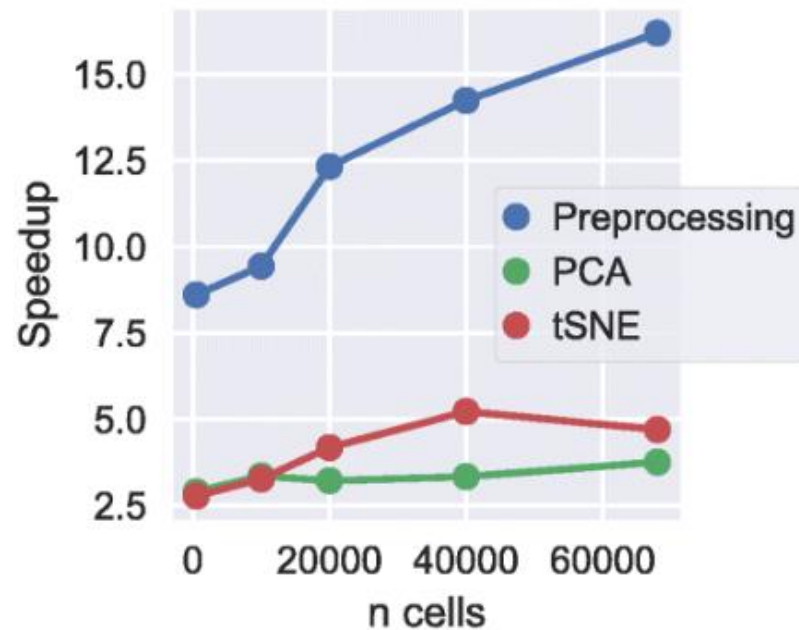


Human microbiome
100 trillion microbes/person

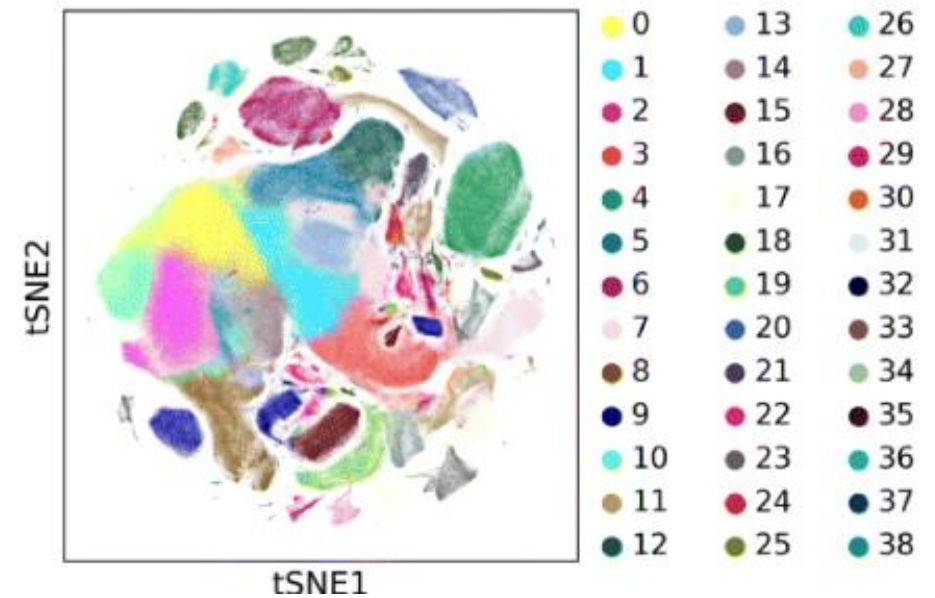


Scaling up single cell analysis

Speedup: Scanpy vs. Cell Ranger R



tSNE of clustered 1.3 million cells



HDF5 files for large 'omics data



$$n > p$$

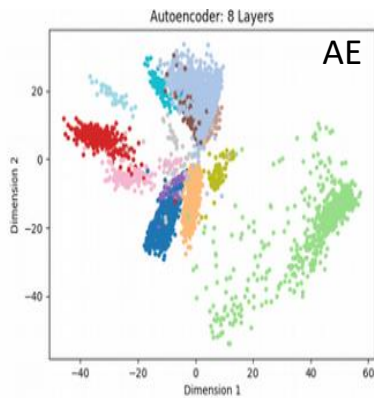
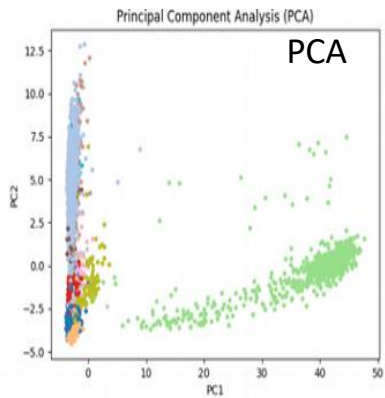
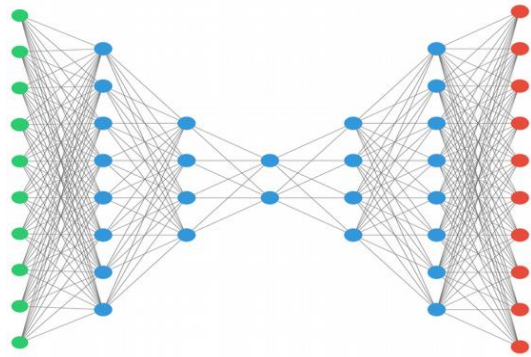
n = observations (cells, 100k-1mil+)

p = features (genes, 20k)

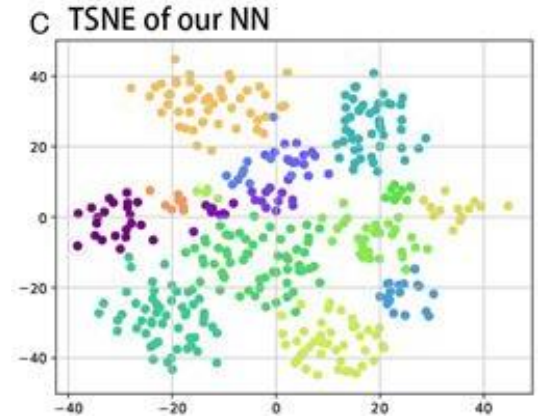
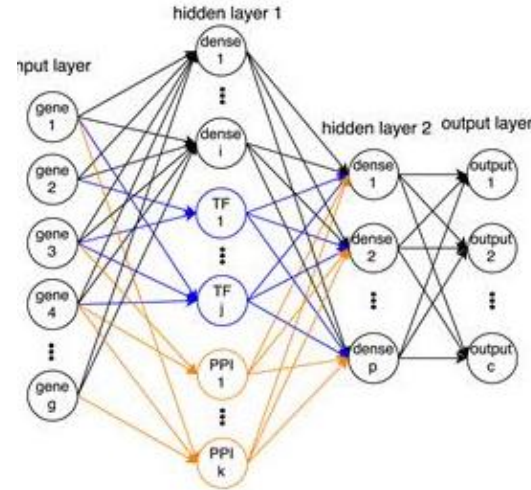


Big data has advantages!

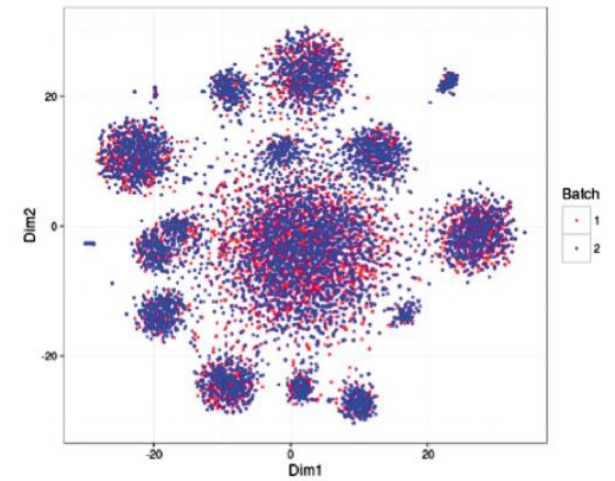
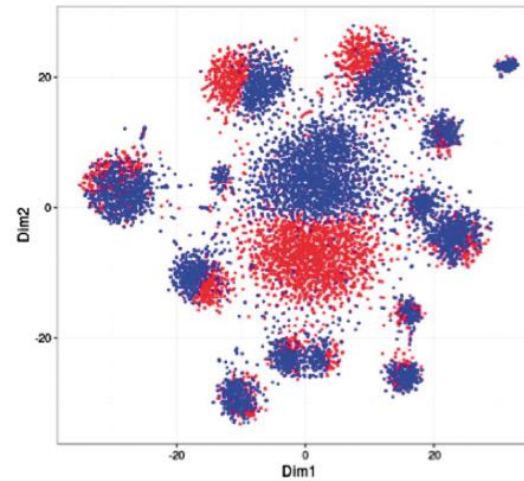
Autoencoder (AE) for non-linear dimensionality reduction



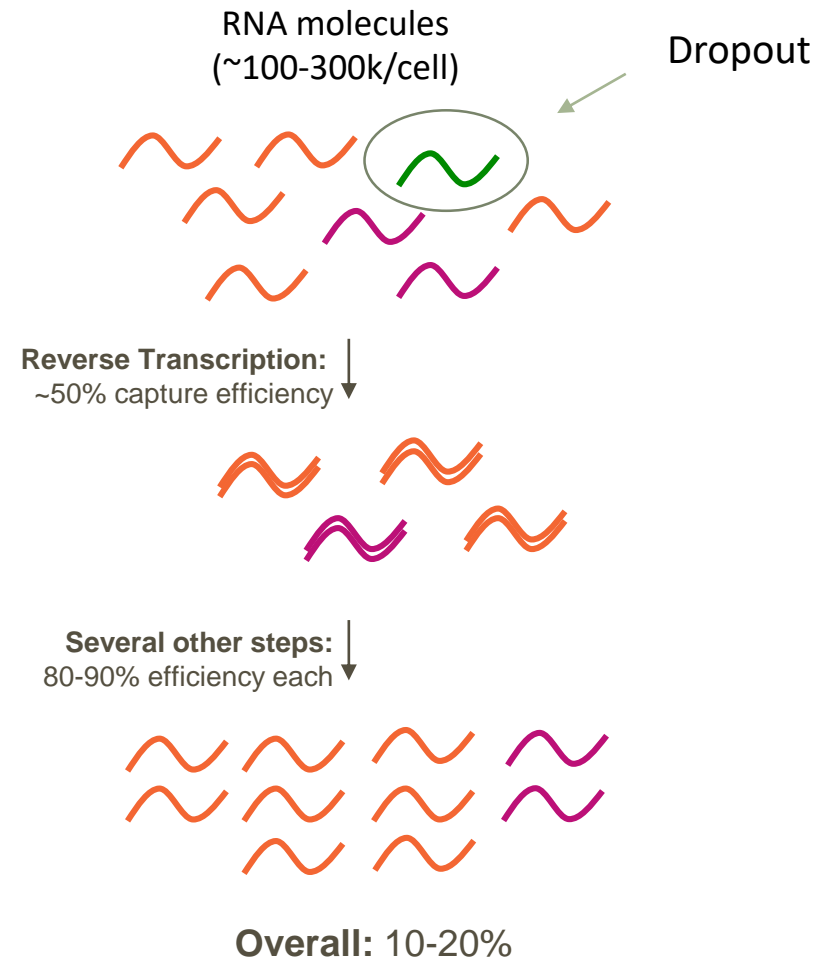
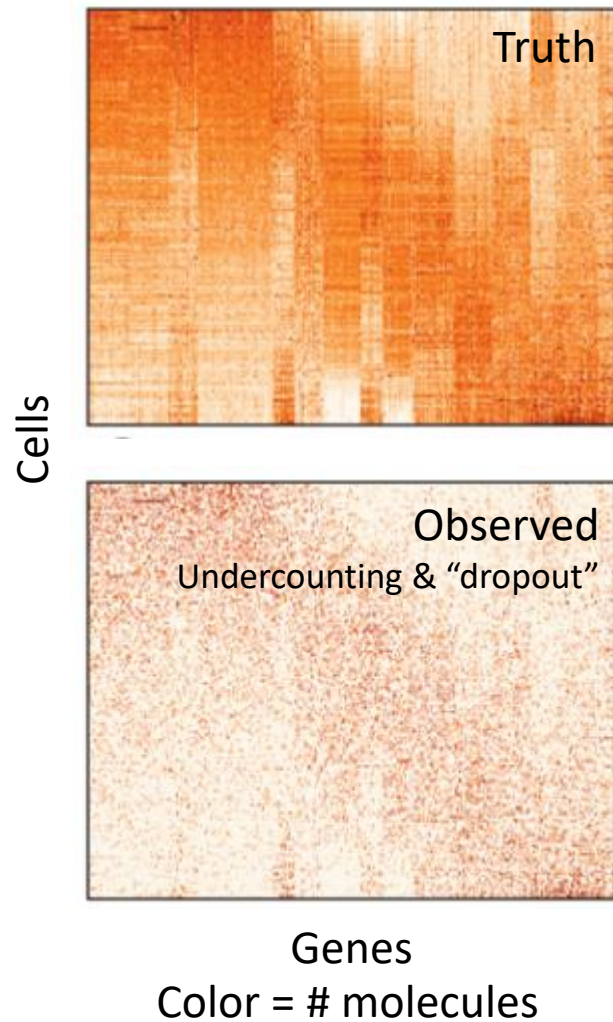
Neural networks for cell type identification



Residual neural networks for batch effect correction

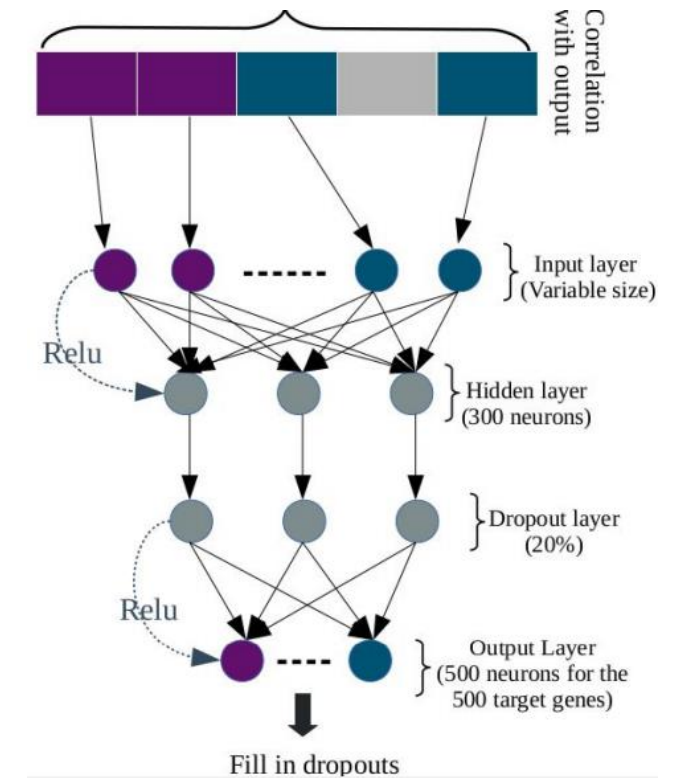
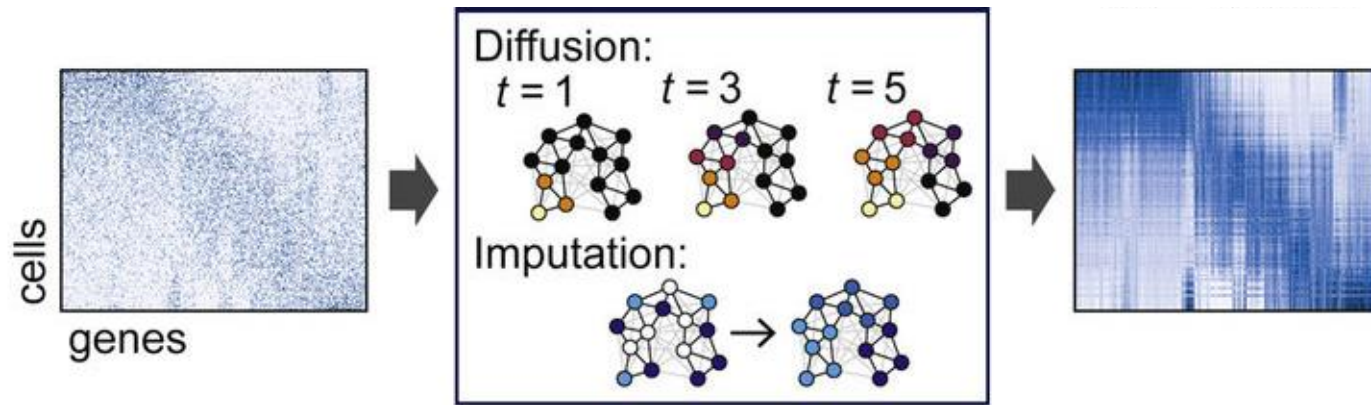


The missing data problem



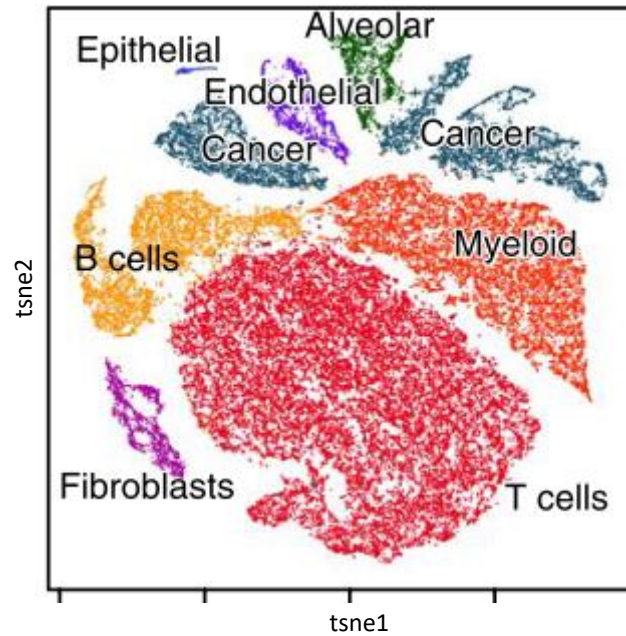
The missing data problem

1. Model missing information
2. Imputation



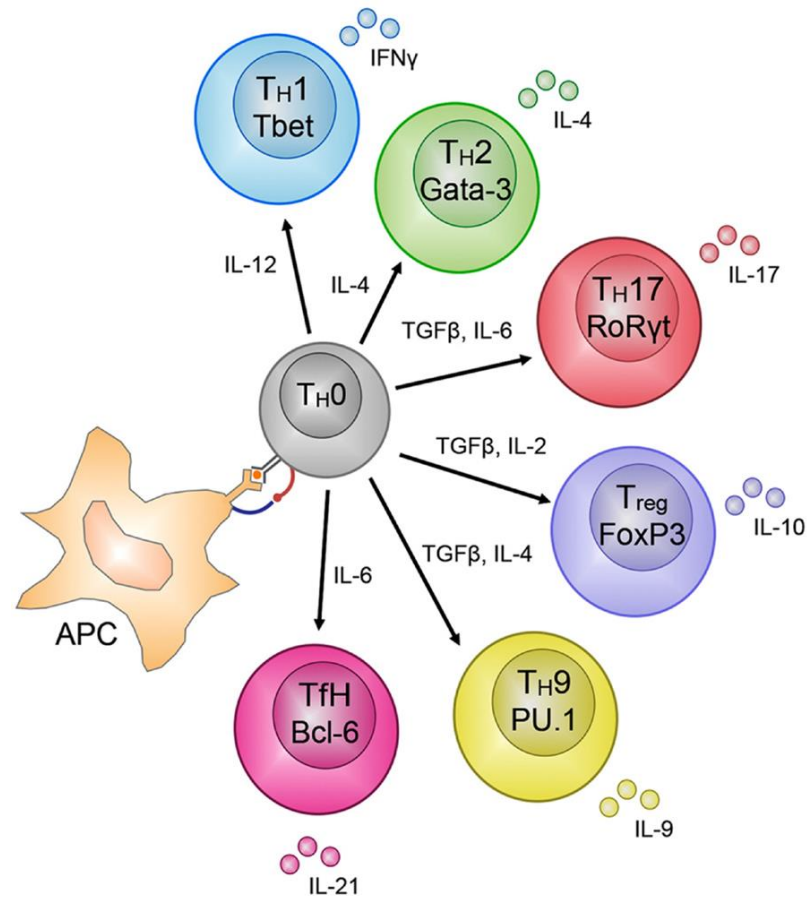
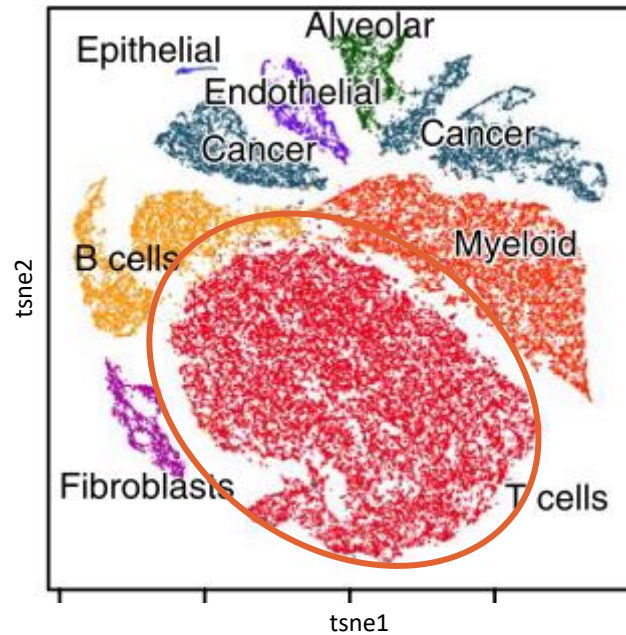
Missing data makes cell subtype identification difficult

Coarse-grained types separate well...



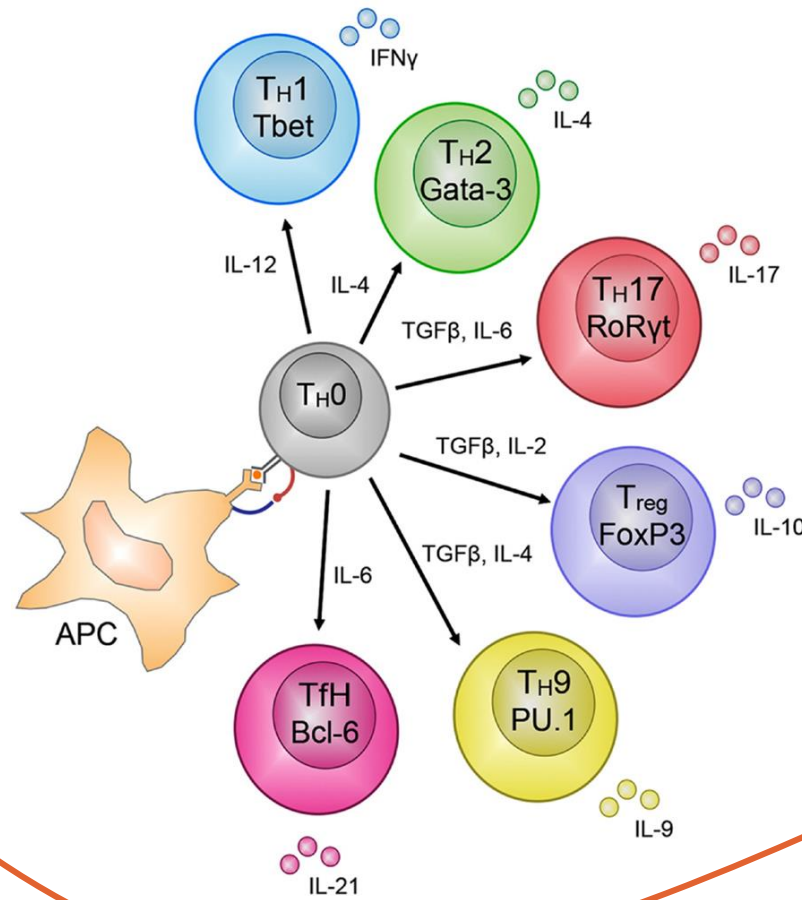
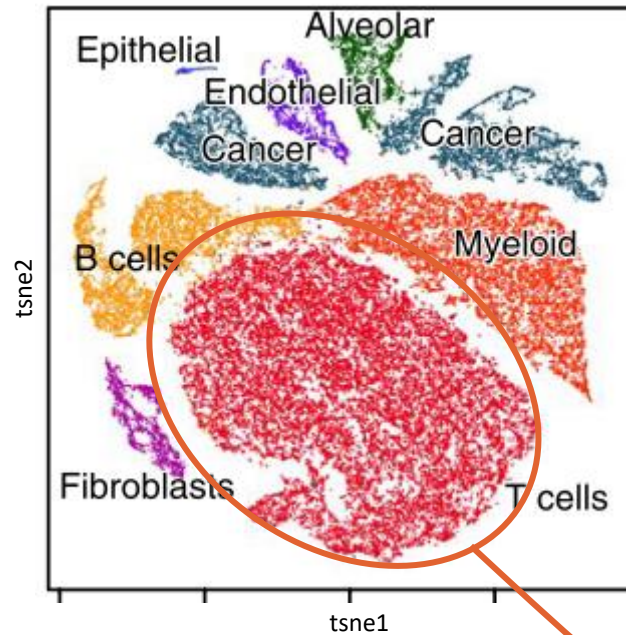
Missing data makes cell subtype identification difficult

Coarse-grained types separate well...

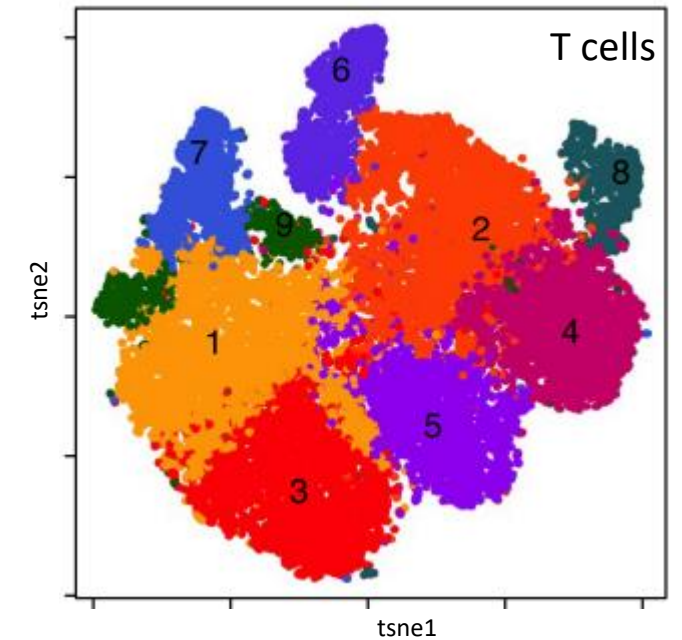


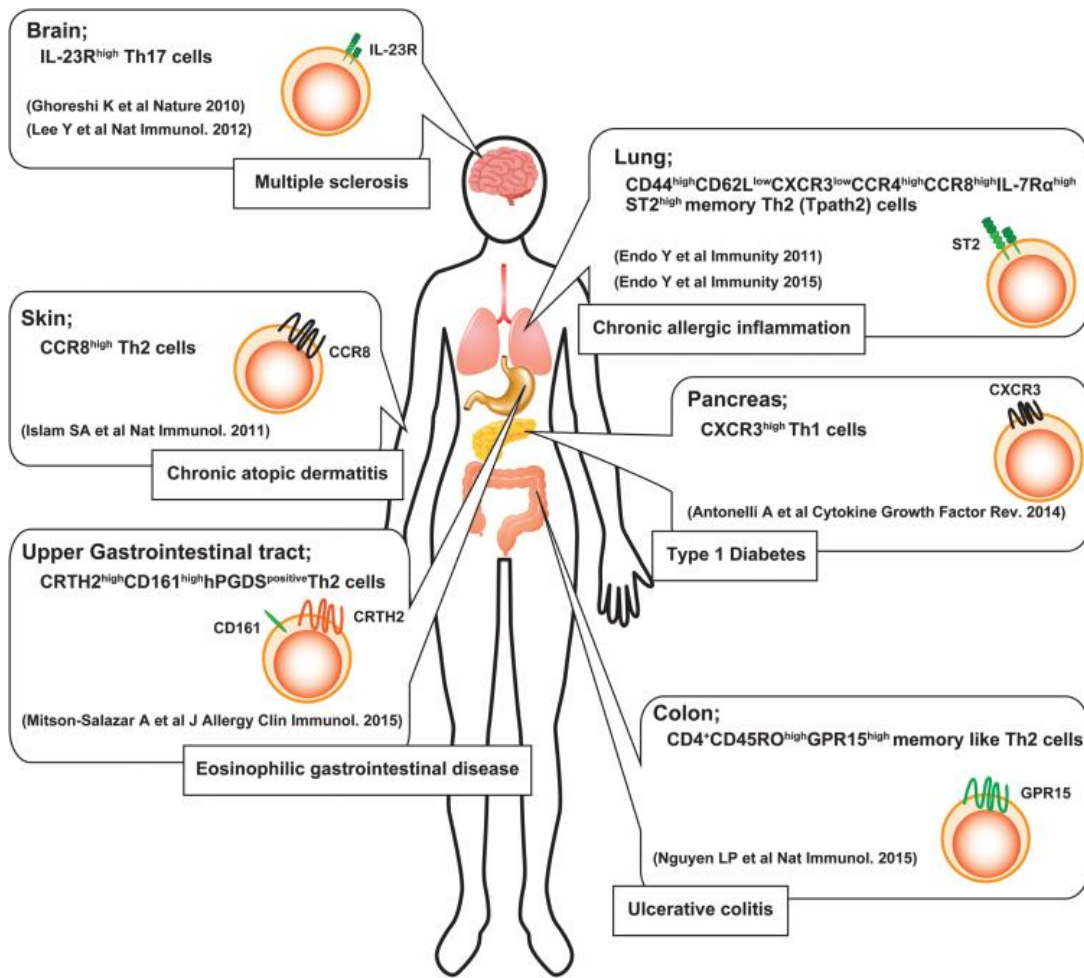
Missing data makes cell subtype identification difficult

Coarse-grained types separate well...

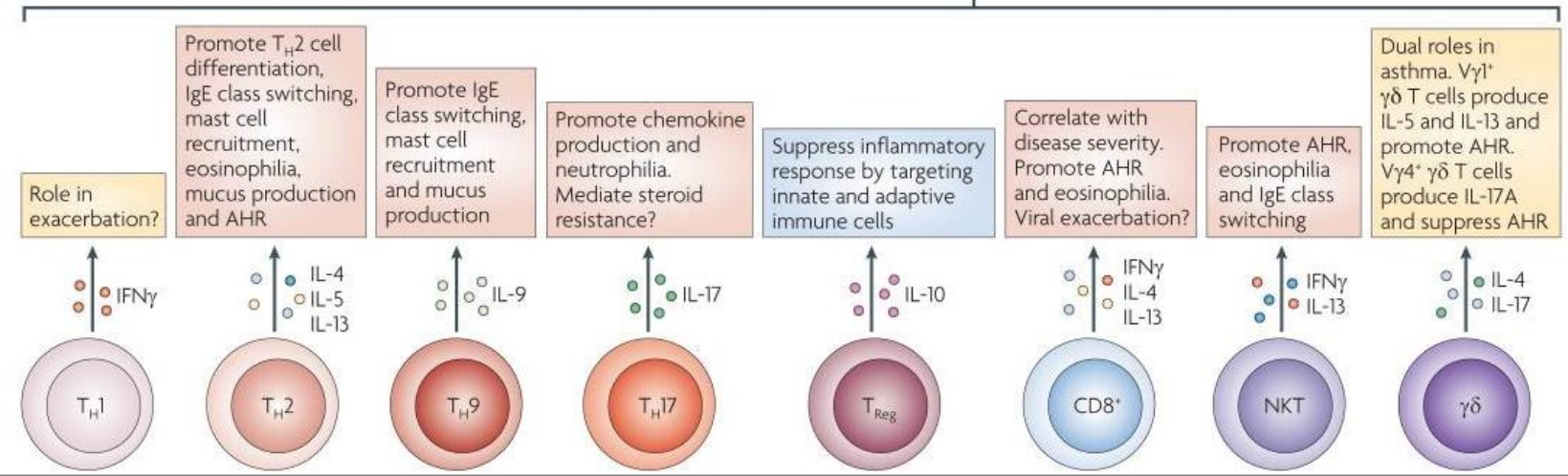
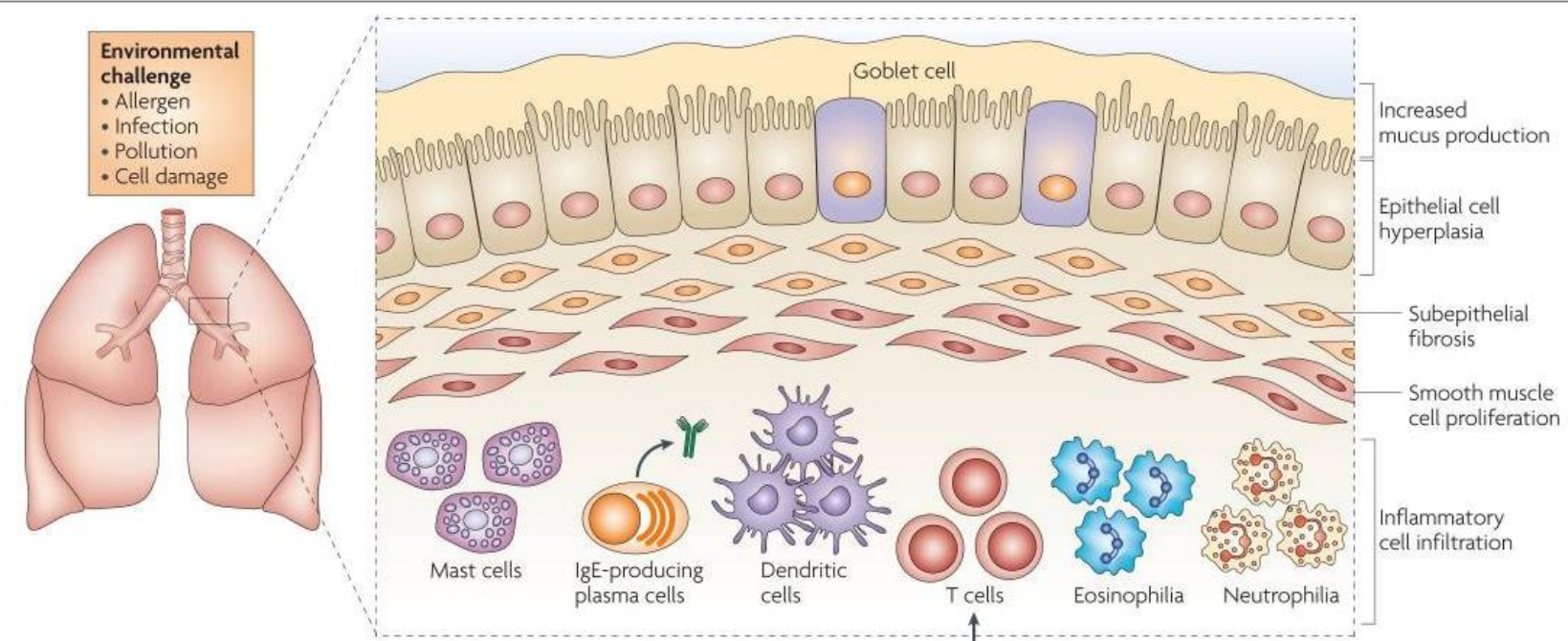
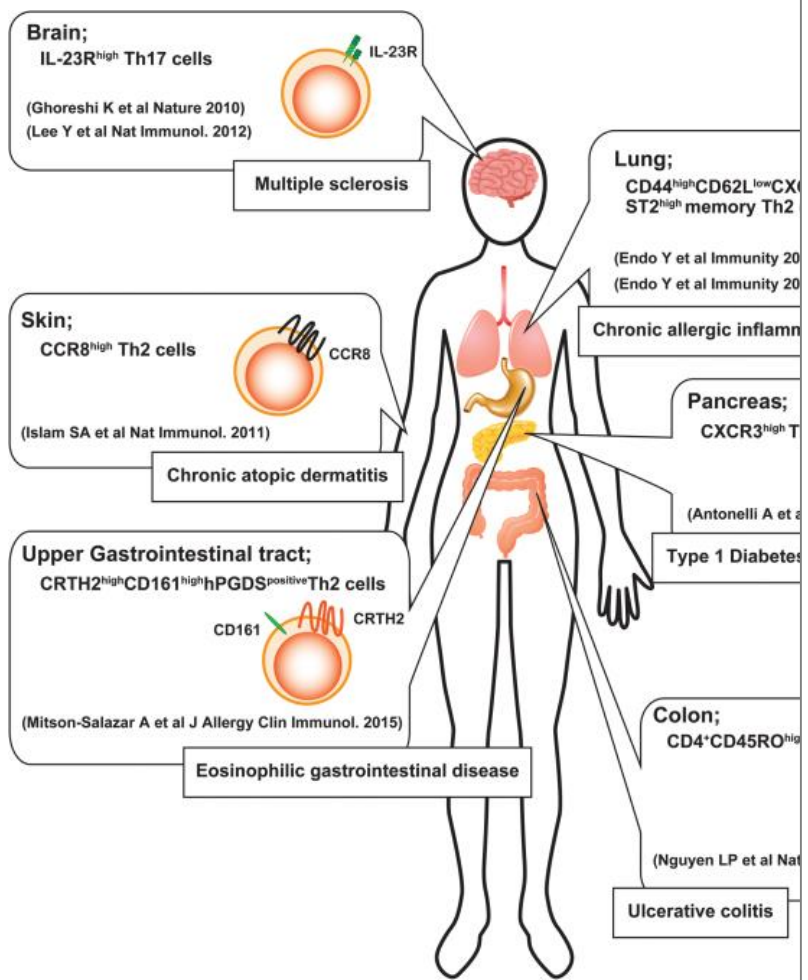


...But finer subdivisions can be murky

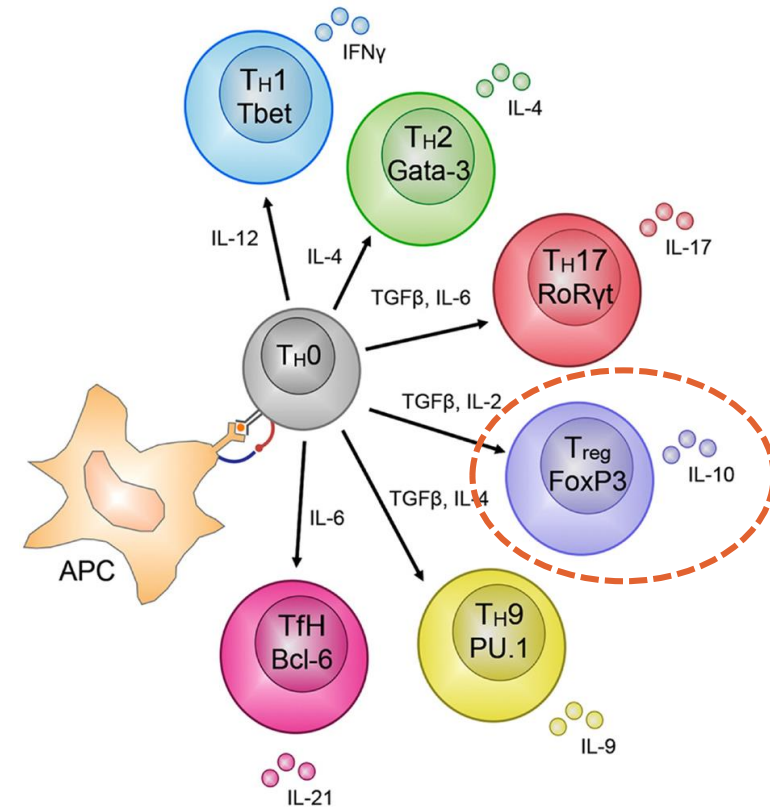
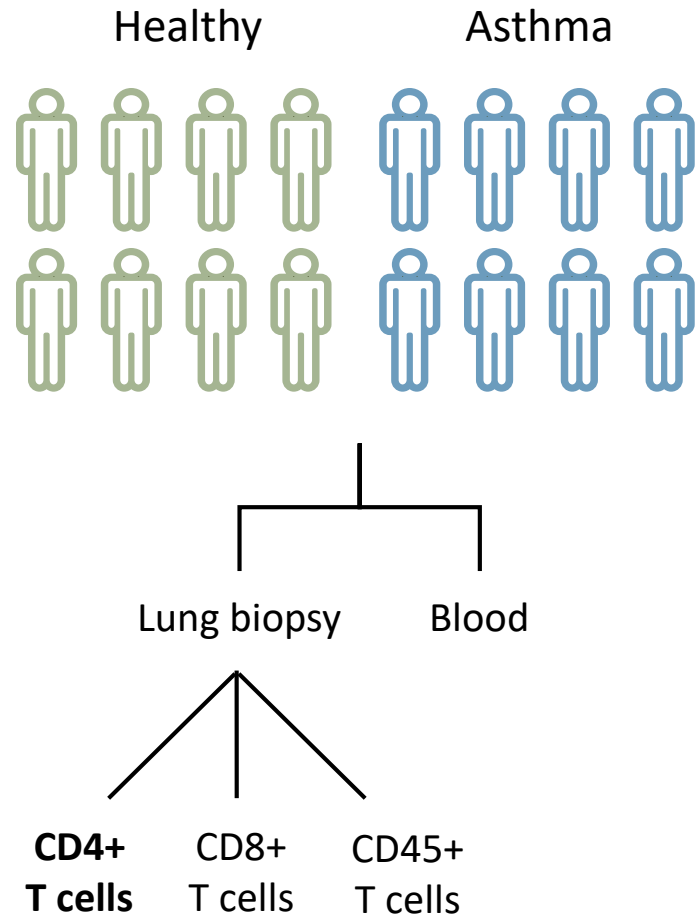




T cell subtypes in Asthma



Example: CD4+ T cells from lung



T regulatory cells ("Tregs")

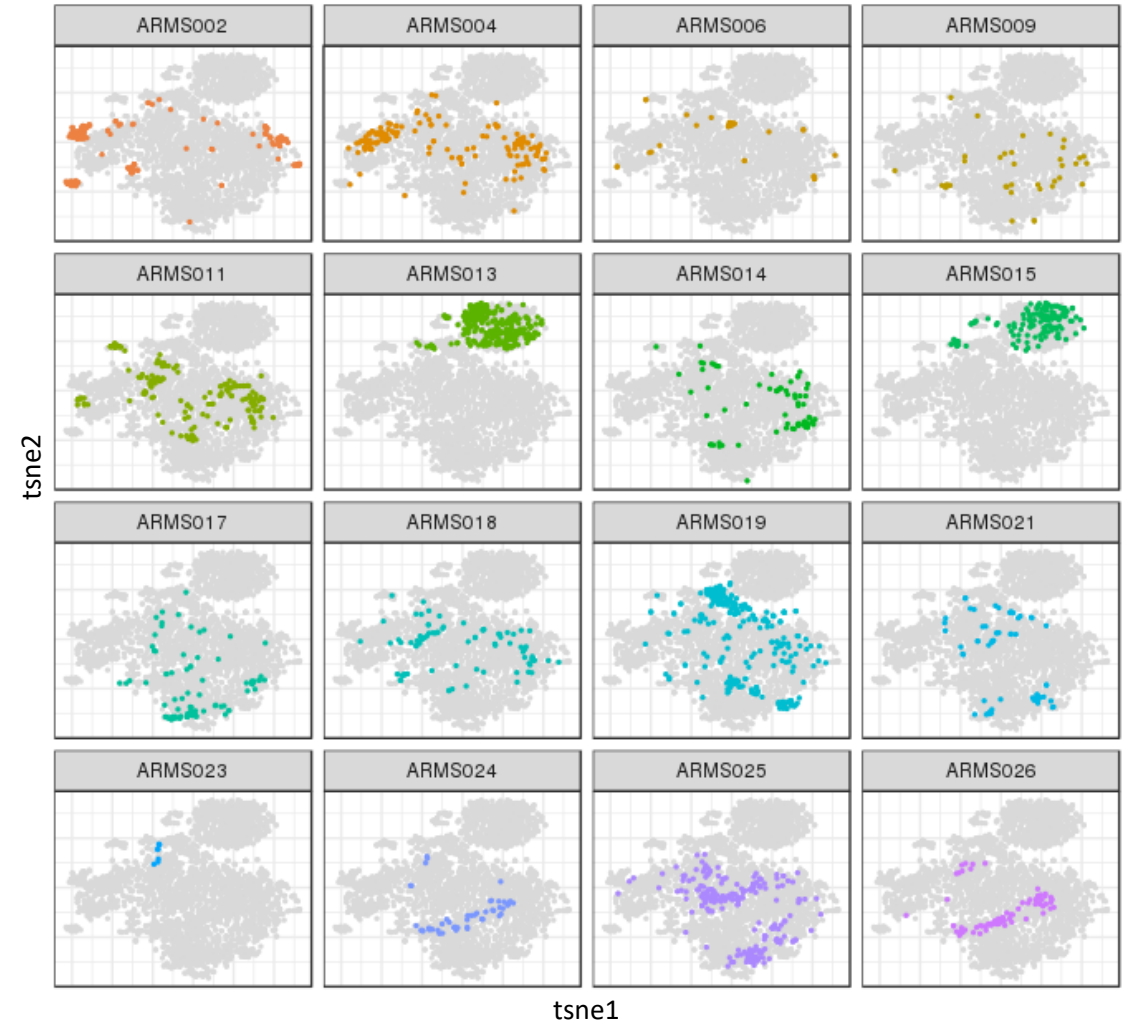
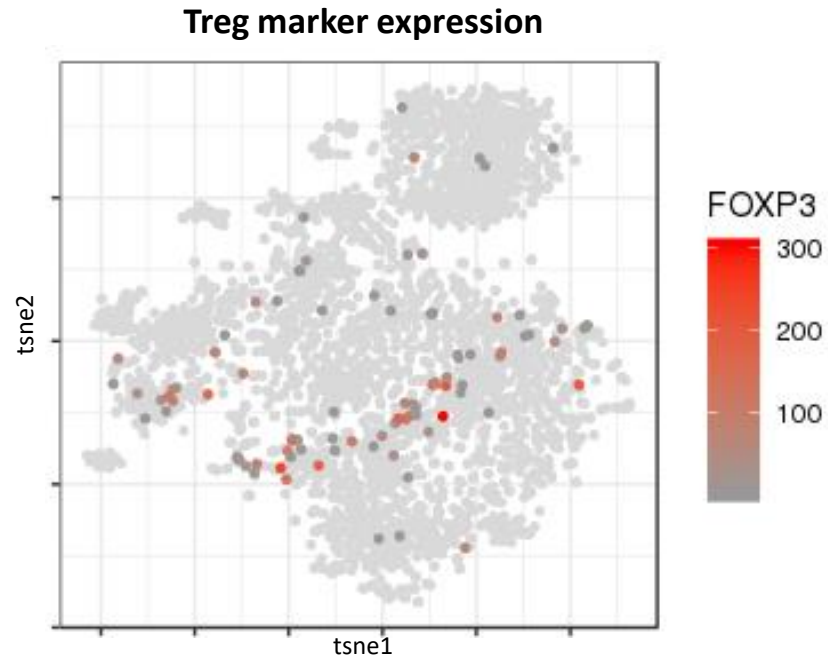
Immune-suppressive functions

→ Difference in **Treg abundance** between healthy and asthma?

→ Difference in **Treg gene expression** between healthy and asthma?

Example: CD4+ T cells from lung

Cells belonging to each donor



Two problems:

1. Marker only detected in a few cells (likely due to dropouts)
2. Clustering is driven by donor (likely due to batch effects)

Result: can't confidently identify sub-types of CD4 T cells

Simple approach to improve cell type identification

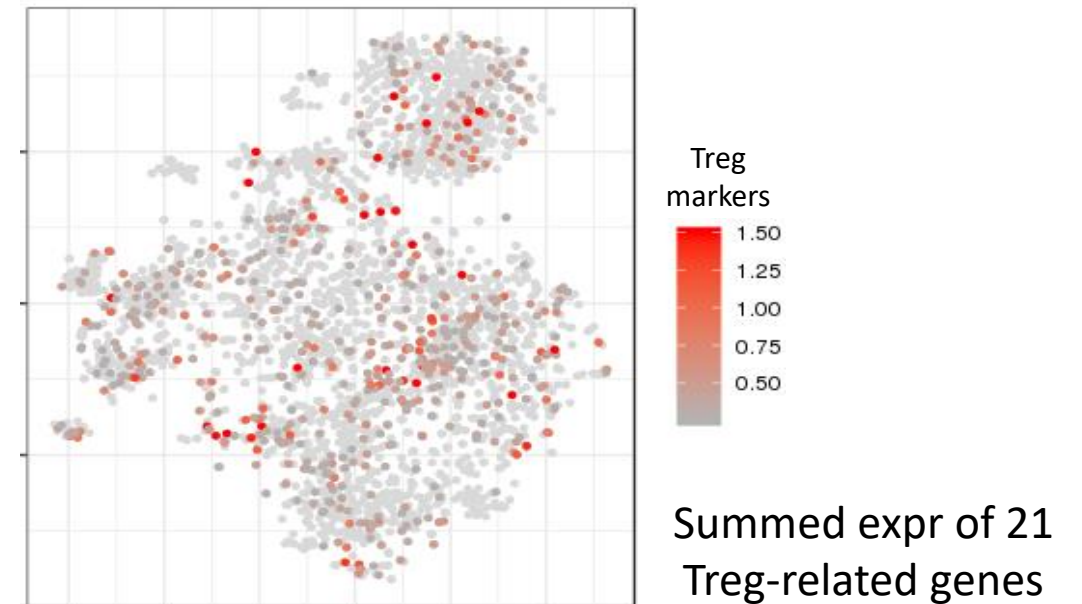
Two concepts that help:

1. Aggregated info gives a clearer picture

- Individual gene expression is noisy & prone to dropout
- Combining signal from multiple genes can be more reliable

2. Biologically relevant genes are better for clustering

- Variation less dominated by technical/batch effects



Simple approach to improve cell type identification

A better feature space for cell type identification

1. Define a set of axes to measure similarity of each cell to various cell types or functional states
2. Place cells into the space according to their scores for each axis & identify clusters

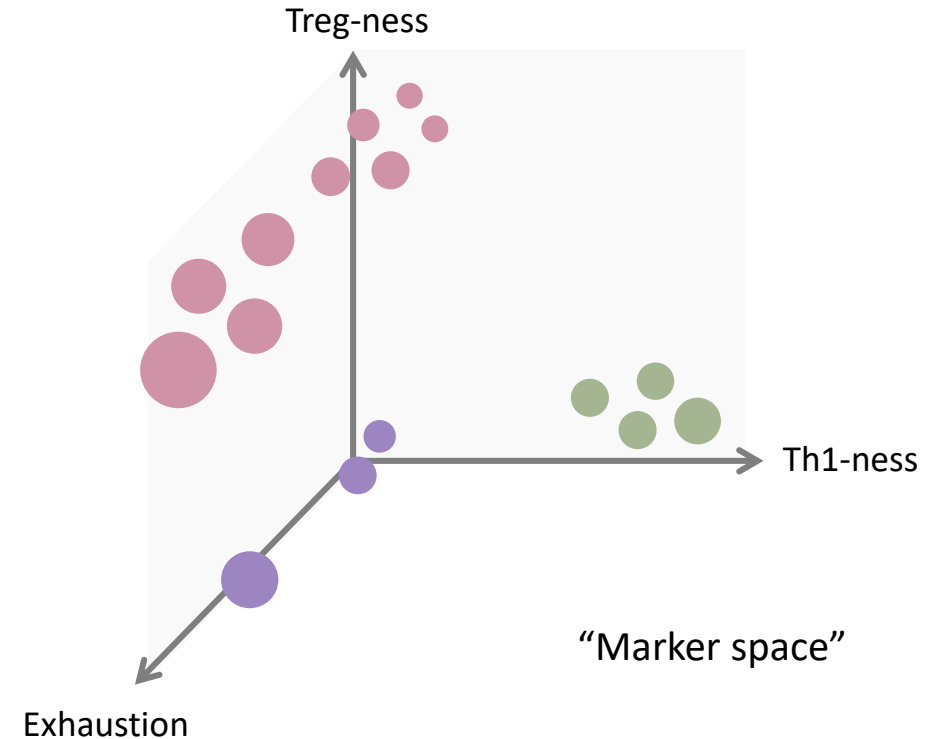
Axis 1: “Treg-ness” = FOXP3 + CTLA4 + IL10 + ...

Axis 2: “Th1-ness” = TBX21 + CXCR3 + CCR5 + ...

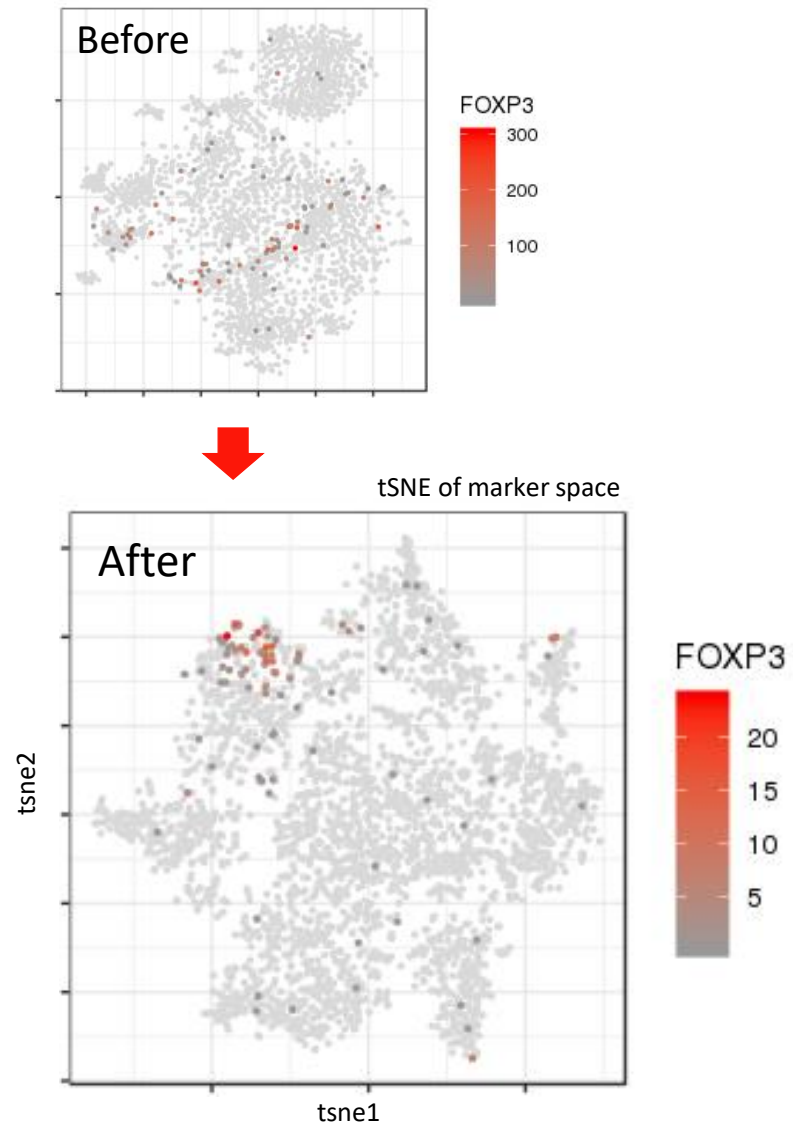
Axis 3: “G2/M Phase” = CDK1 + CCNA2 + CCB1 + ...

Axis 4: “Exhaustion” = PDCD1 + CD244 + CD160 + ...

...

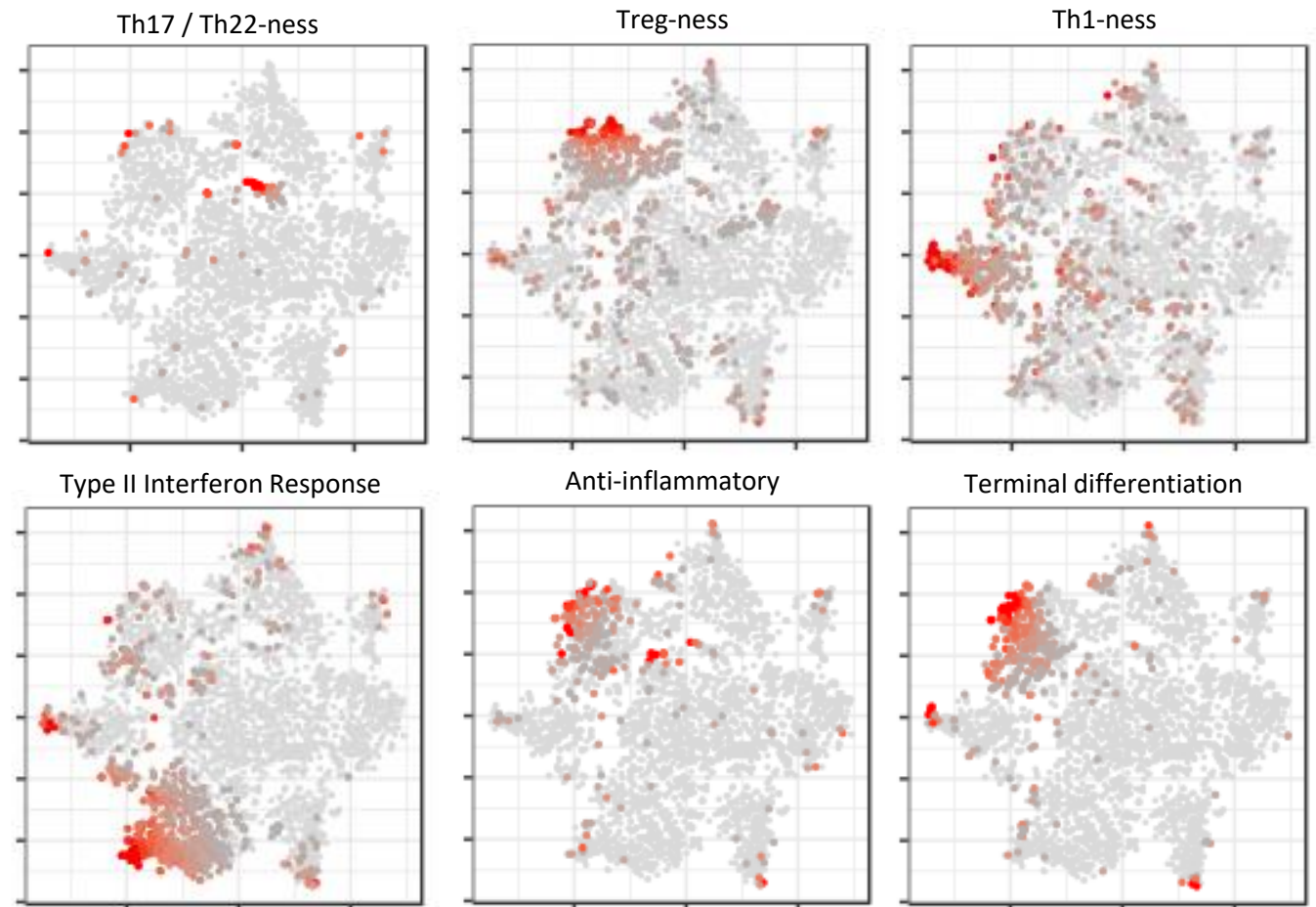
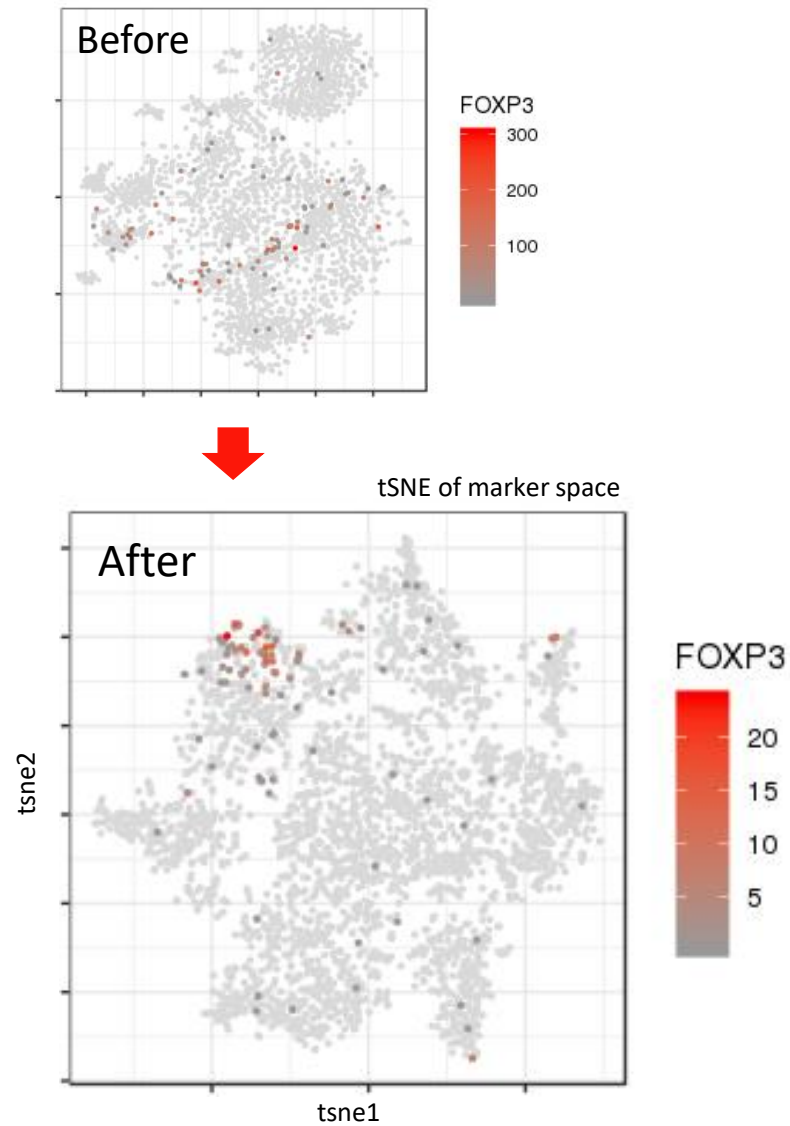


Improved cell type identification



Clusters not driven by patient!

Improved cell type identification



Now we can **compare treatment groups** to identify:

- Changes in cell type composition
- Changes gene expression within a cell type

Future directions

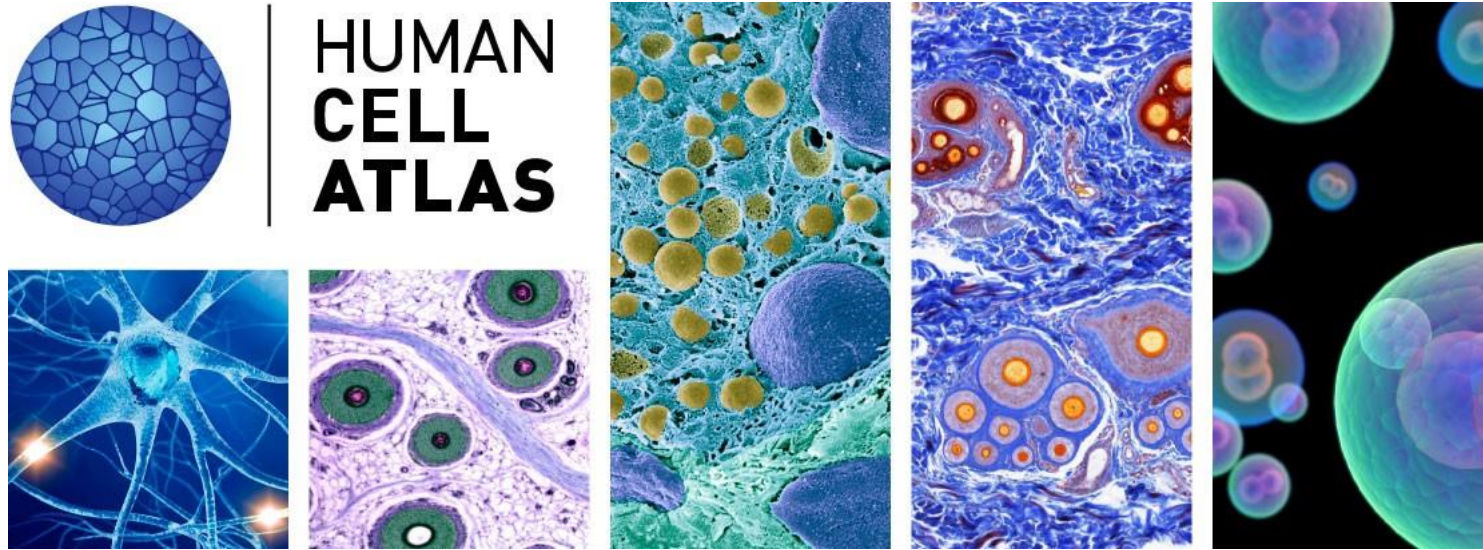


Image: [Broad Inst](#)

