Tues Jan 17th

Computational Health Informatics Program

Mapping the Trajectory of Tumor Cell Transitions with Manifold & Deep Learning

RESEARCHER

Aniket Dey
Dougherty Valley High School

ADVISOR

Dr. Felix Dietlein
Harvard Medical School

Table of Contents

		Page
I	Paper Introduction	3
II	Methodology	6
III	Reproduction & Demonstration	18
IV	Predictive Models	23
V	Future Work & Discussions	27

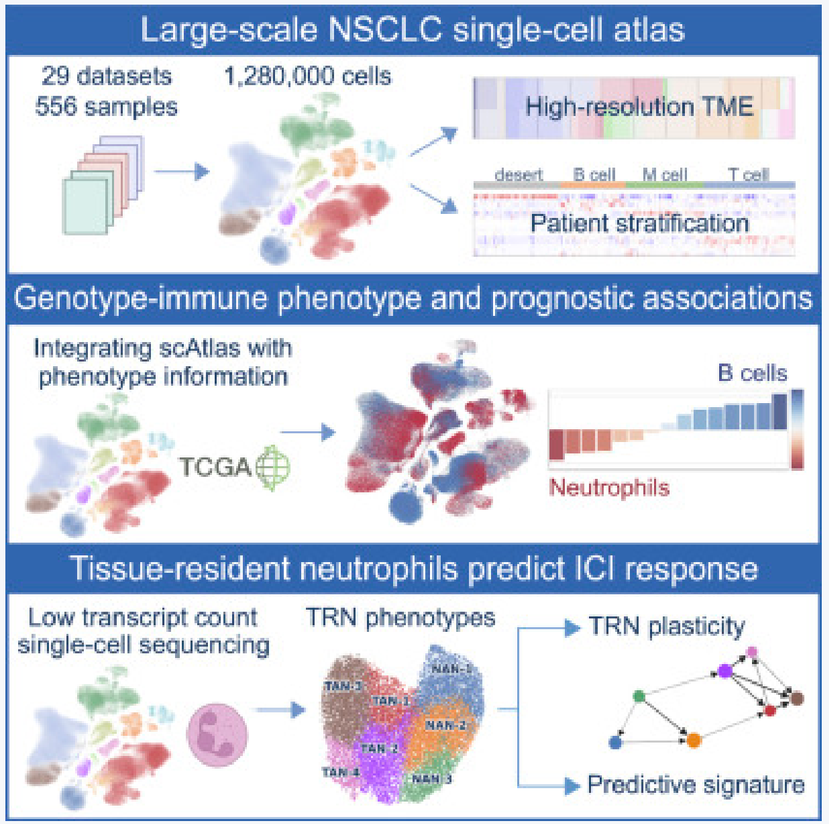
Chapter I

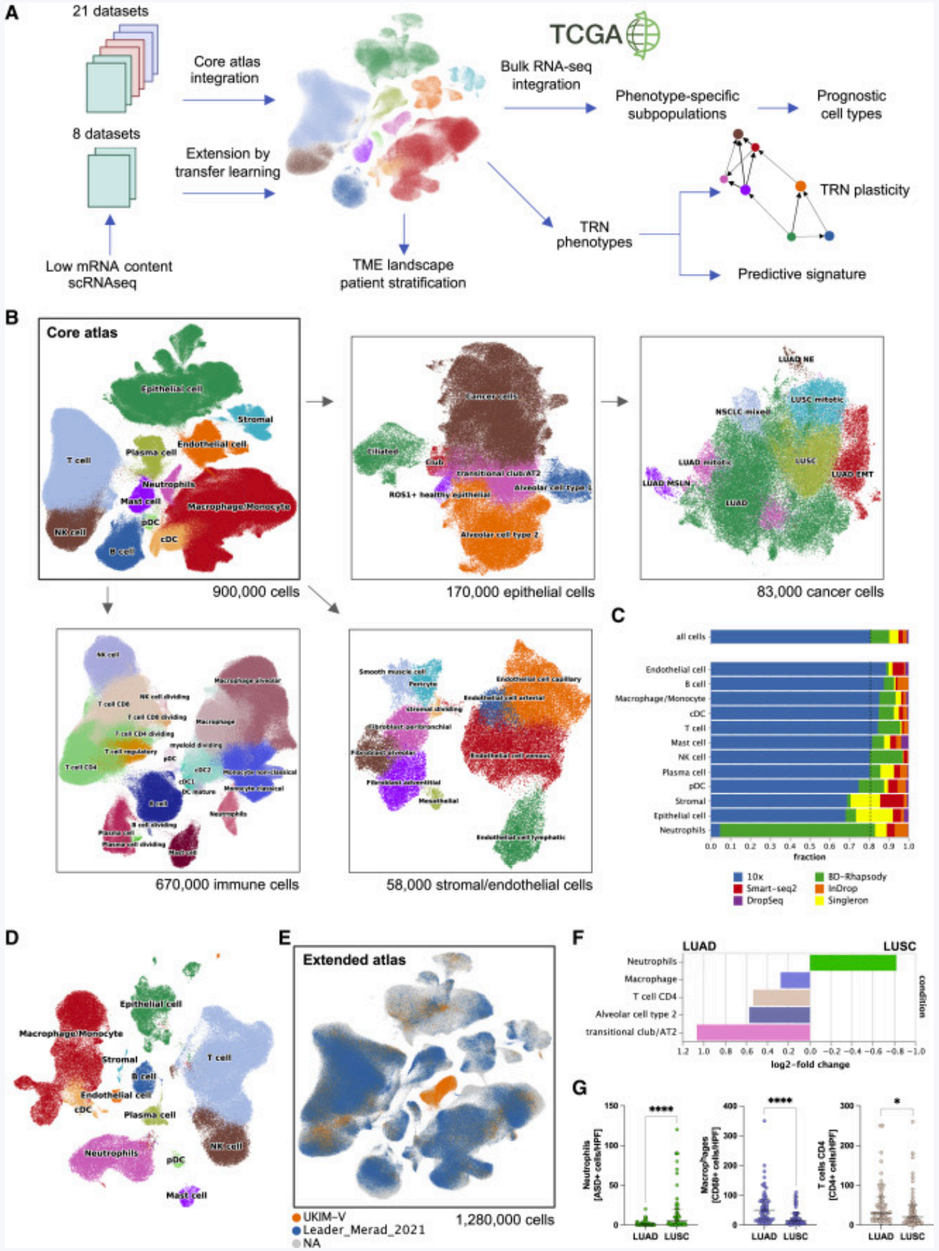
Paper Introduction

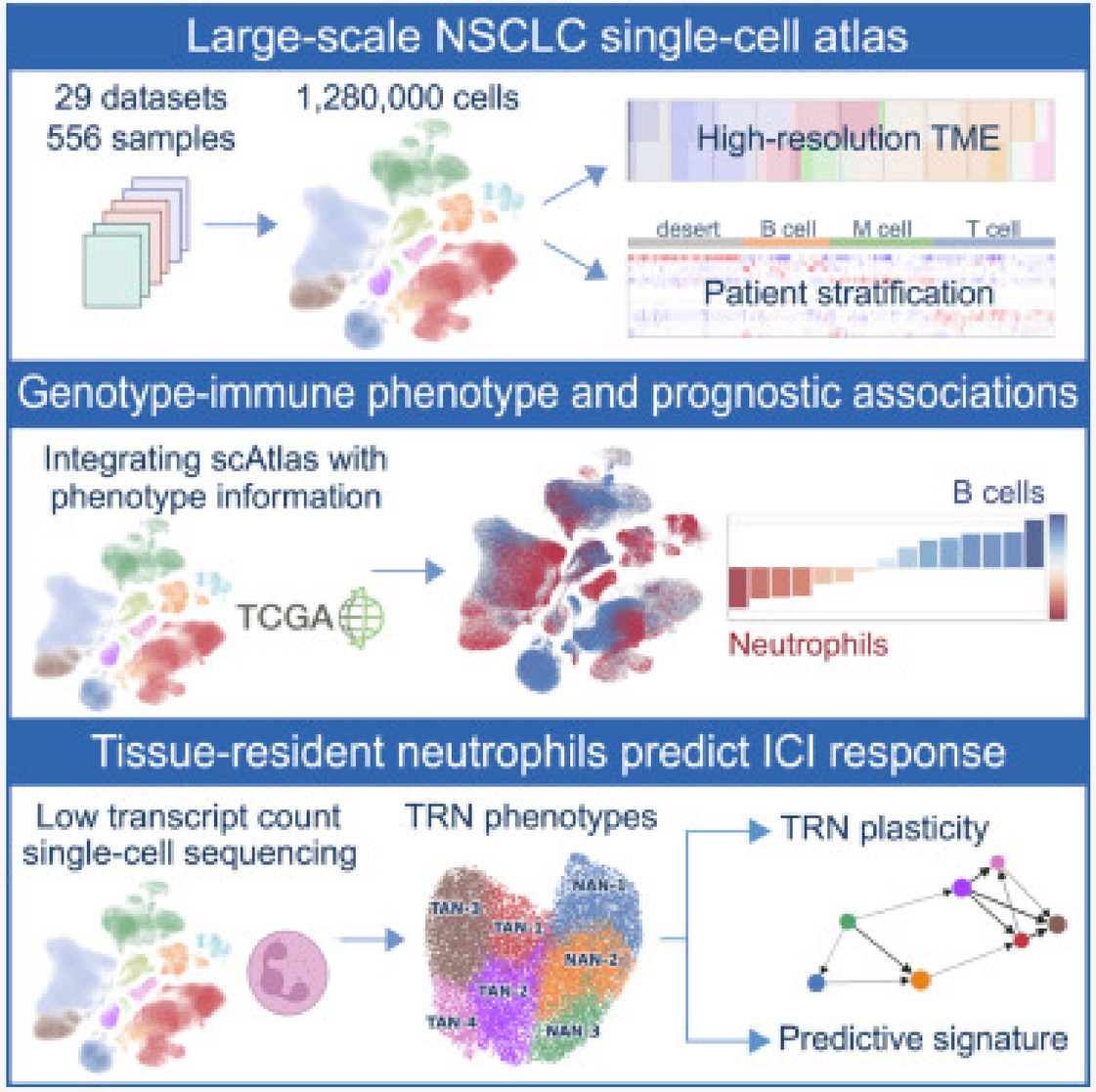
High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer

High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer - Dec 12, 2022

- Non-small cell lung cancer (NSCLC) characterized by molecular heterogeneity with diverse immune cell infiltration patterns which has been linked to therapy sensitivity and resistance.
- Using bulk samples with genomic and clinical information, the study identifies cellular components associated with tumor histology and genotypes
- The study then focuses on the analysis of tissue-resident neutrophils (TRNs) and uncover distinct subpopulations (TAN & NAN) that acquire new functional properties in the tissue microenvironment, providing evidence for the plasticity of TRNs
- Finally, the study shows that a TRN-derived gene signature is associated with anti-programmed cell death ligand 1 (PD-L1) treatment failure




A
21 datasets
8 datasets
Low mRNA content scRNAseq
Core atlas integration
Extension by transfer learning
TME landscape patient stratification
TCGA
Bulk RNA-seq integration
Phenotype-specific subpopulations
Prognostic cell types
TRN phenotypes
TRN plasticity
Predictive signature
B
Core atlas
900,000 cells
170,000 epithelial cells
83,000 cancer cells
670,000 immune cells
58,000 stromal/endothelial cells
C
fraction
10x
Smart-seq2
DropSeq
BD-Rhapsody
InDrop
Singleron
D
E
Extended atlas
UKIM-V
Leader_Merad_2021
NA
1,280,000 cells
F
LUAD
LUSC
log2-fold change
G
LUAD LUSC


Large-scale NSCLC single-cell atlas
29 datasets
556 samples
1,280,000 cells
High-resolution TME
desert
B cell
M cell
T cell
Patient stratification
Genotype-immune phenotype and prognostic associations
Integrating scAtlas with phenotype information
TCGA
B cells
Neutrophils
Tissue-resident neutrophils predict ICI response
Low transcript count single-cell sequencing
TRN phenotypes
TAN-3
TAN-1
NAN-1
NAN-2
TAN-2
TAN-4
NAN-3
TRN plasticity
Predictive signature

Chapter II

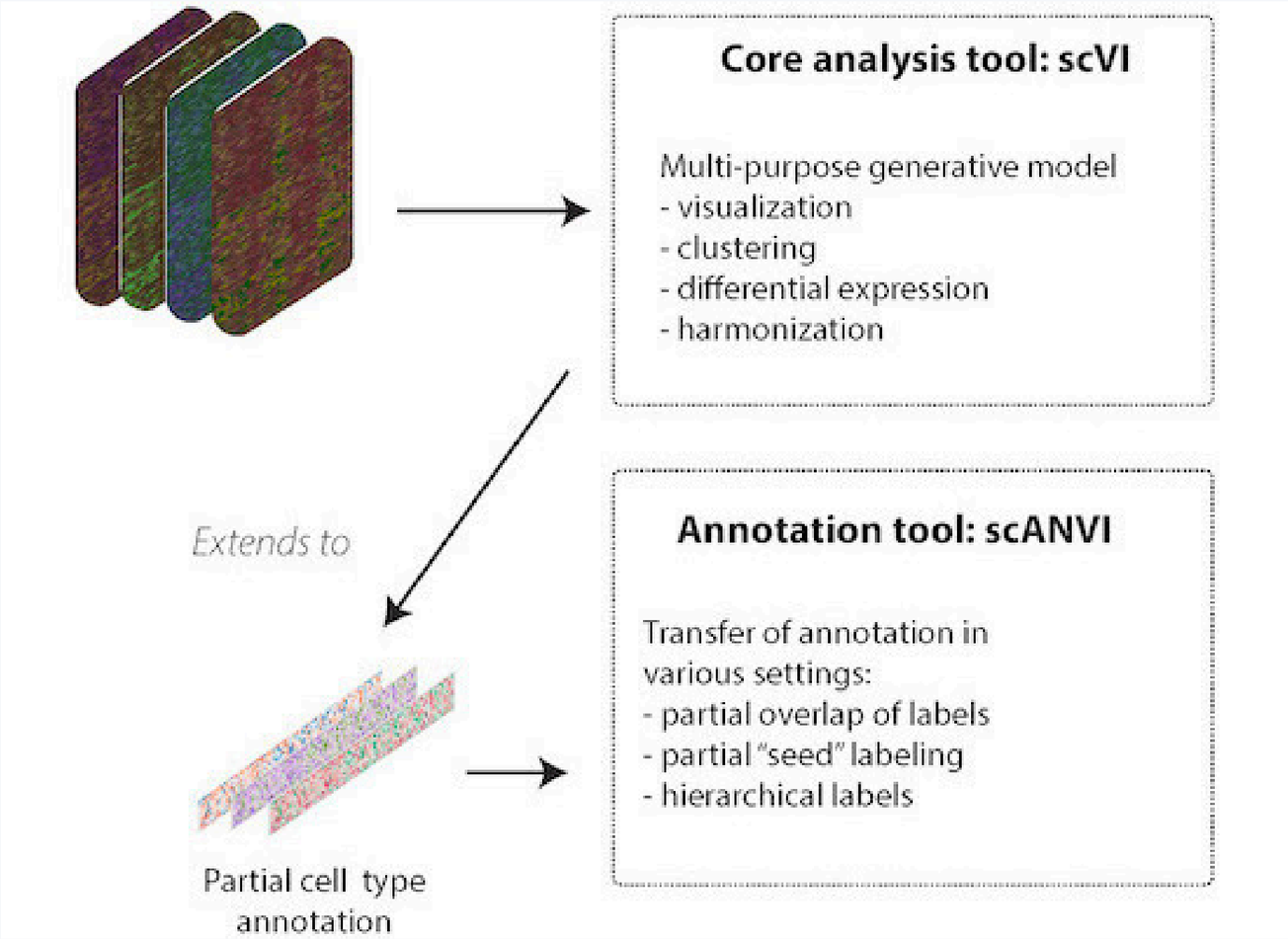
Methodology

SCANPY, Leiden/Louvain Clustering, sciVI, scanVI, Immunochemistry

Immunochemistry & Lab Work

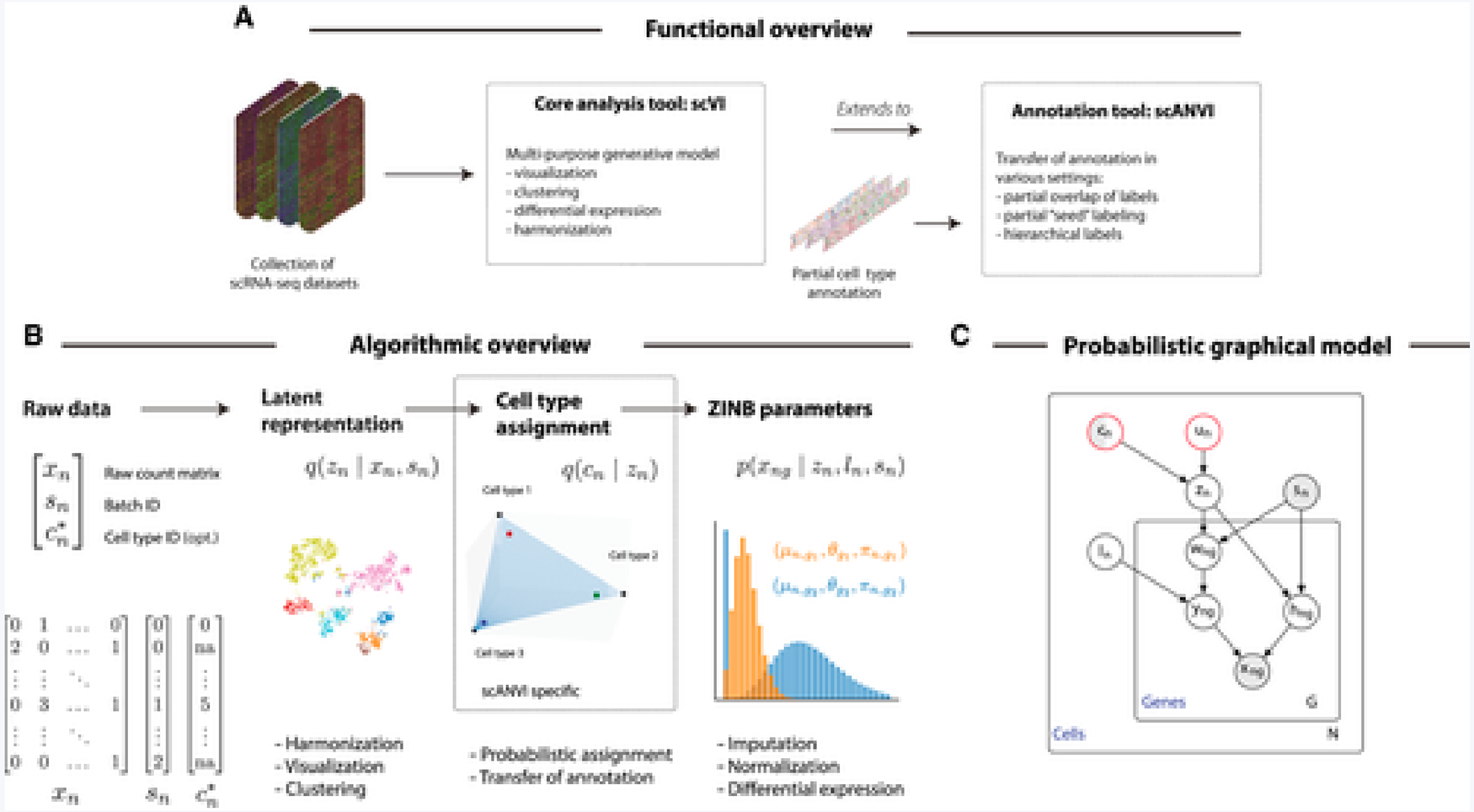
- Samples taken of NSCLC tumor tissues and matched adjacent normal lung tissues (more than 5 cm distance to the tumor)
- Prepared NSCLC tissue and normal lung tissue to create single call solution, RBCs removed, Cells counted
- BD Rhapsody library preparation and sequencing - Single-cell isolation, libraries created and sequenced
- Flow cytometry – detail characterization of neutrophils, gating strategy to define cellpopulations from NSCLC tumor tissue and normal adjacent tissue.
- Multiplex immunofluorescence – cell nuclei and perinuclear area to define cell cytoplasm, total area evaluated, cell phenotyping proven by presence or absence of the marker
- Immunohistochemistry – existence of lymphocytes, neutrophils, macrophages; counted cells as well

scVI & scANVI: Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models - Jan 25, 2021

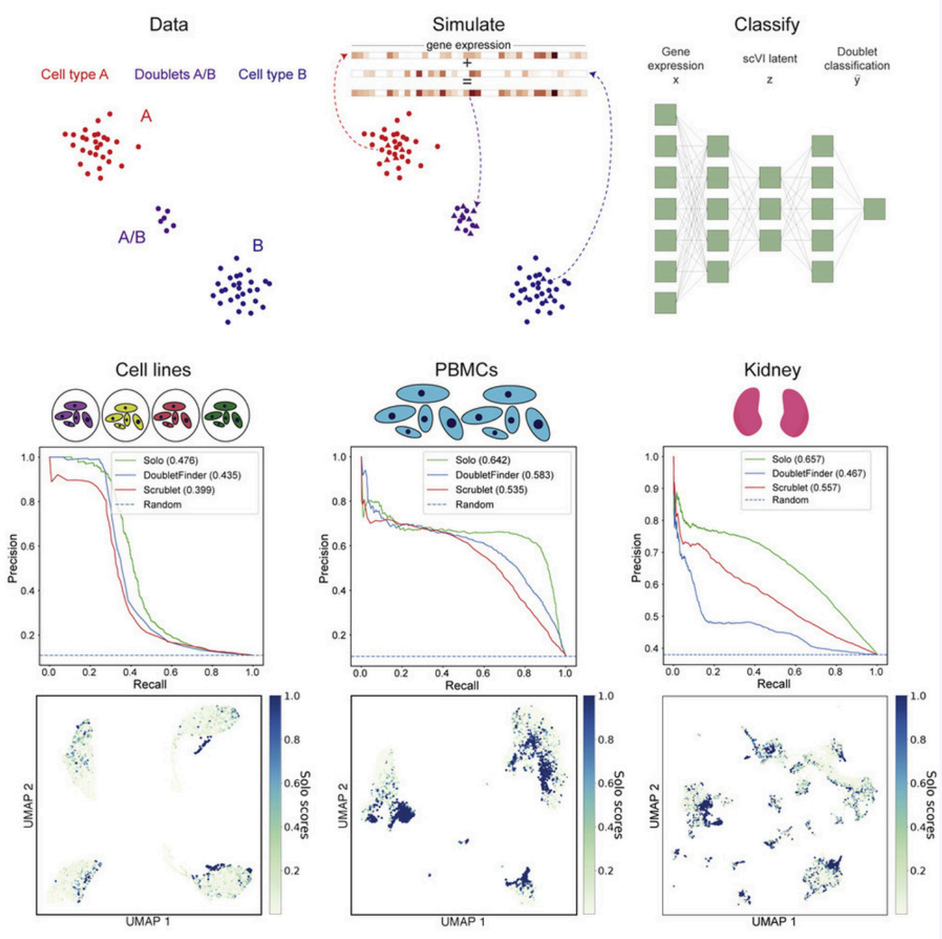


- scVI integrates single-cell RNA-seq datasets in a variety of settings and scANVI, a new development based on scVI, presents automated annotation of cell types and states.
- In scVI, datasets from different labs and technologies are integrated in a joint latent space.
- In scANVI, cell type annotations are transferred between datasets and across different scenarios.
- The performance of scVI and scANVI in data integration and cell state annotation is superior to other related methods.

scVI & scANVI: Multifaceted Overview



Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning - Jun 26, 2020



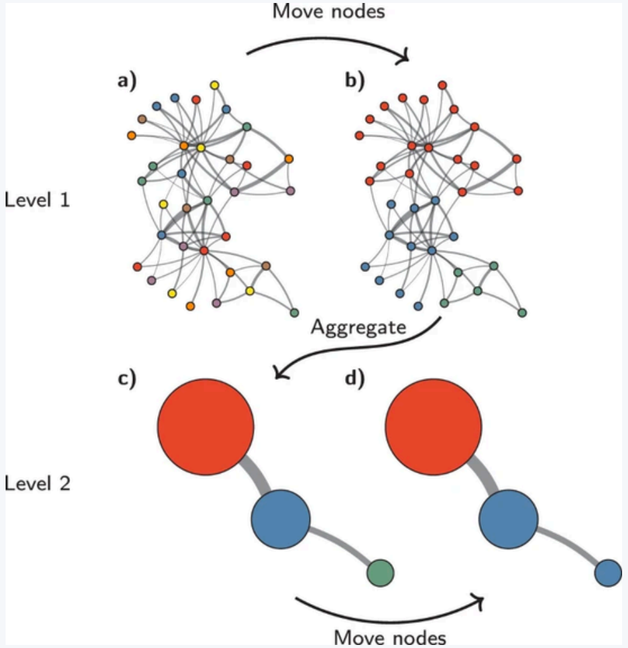
- Current single-cell RNA sequencing methods often result in two or more cells that share the same cell-identifying barcode
- These “doublets” violate the fundamental premise of single-cell technology and can lead to incorrect inferences.
- **Solo is a semi-supervised deep learning approach** that identifies doublets with greater accuracy than existing methods
- Solo embeds cells unsupervised using a variational autoencoder and then appends a feed-forward neural network layer to the encoder to form a supervised classifier. We train this classifier to distinguish simulated doublets from the observed data.

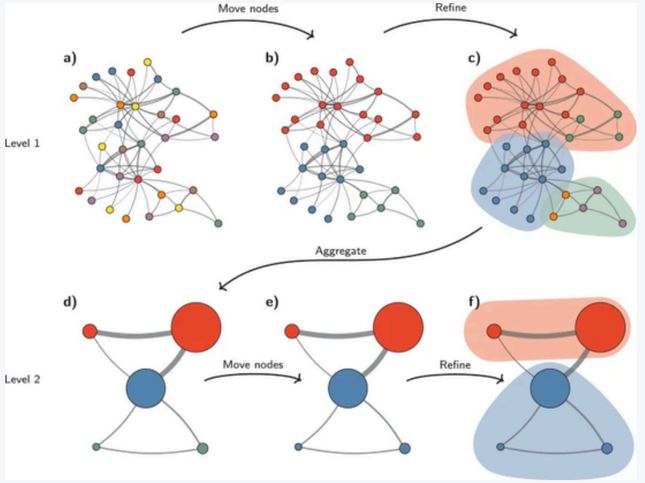
Generation of Core Atlas & Integration of scRNA-seq Datasets

- **Generation of the core atlas** – research datasets (n=3) + published NSCLC studies (2018 – 2021) + non-NSCLC lung samples with normal lungs & COPD
- Preprocessing and quality control of scRNA-seq data – curated, loaded in AnnData containers, QC w scanpy by threshold numbering
- **Integration of scRNA-seq datasets**
  - AnnData object container - adds new levels of further curation, removing duplicates, zeros to fill gene symbols, exclusion of missing symbols
  - scANVI algorithm integration – annotated “seed” datasets based on unsupervised clustering from “raw” data
  - scANVI were pre-trained w scVI model as recommended by the scvi-tools
  - scVI were then trained with 6000 most highly variable genes as determined with scanpy’s104 pp.highly_variable_genes with parameters flavor=”seurat_v3″ and batch_key=”dataset”.
  - Each sample was considered as an individual batch for both scVI and scANVI
- Doublet-detection – SOLO algorithm to computationally detect multiplets, solo is integrated in scvi-tools

From Louvain to Leiden: guaranteeing well-connected communities - Mar 26 2019

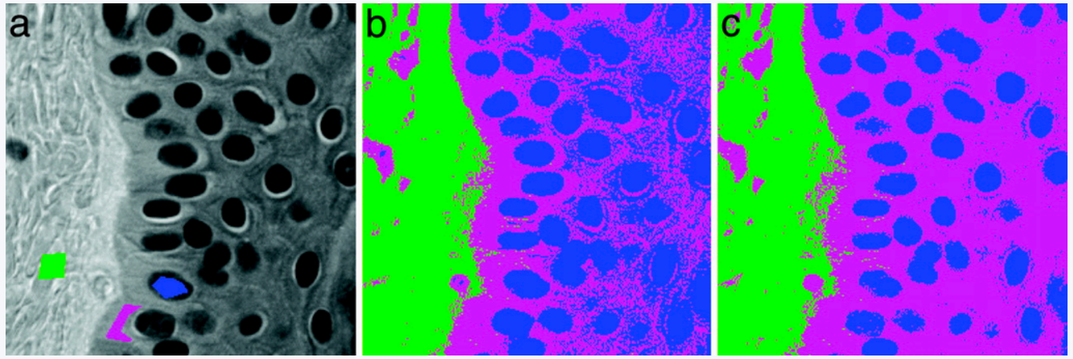
- Leiden algorithm overcomes the problem of arbitrarily disconnected clusters
- Yields communities that are guaranteed to be connected
  - when the algorithm is applied iteratively, it converges to a partition in which all subsets of all communities are guaranteed to be locally optimally assigned
- the Leiden algorithm convincingly outperforms the Louvain algorithm, both in terms of speed and in terms of quality of the results





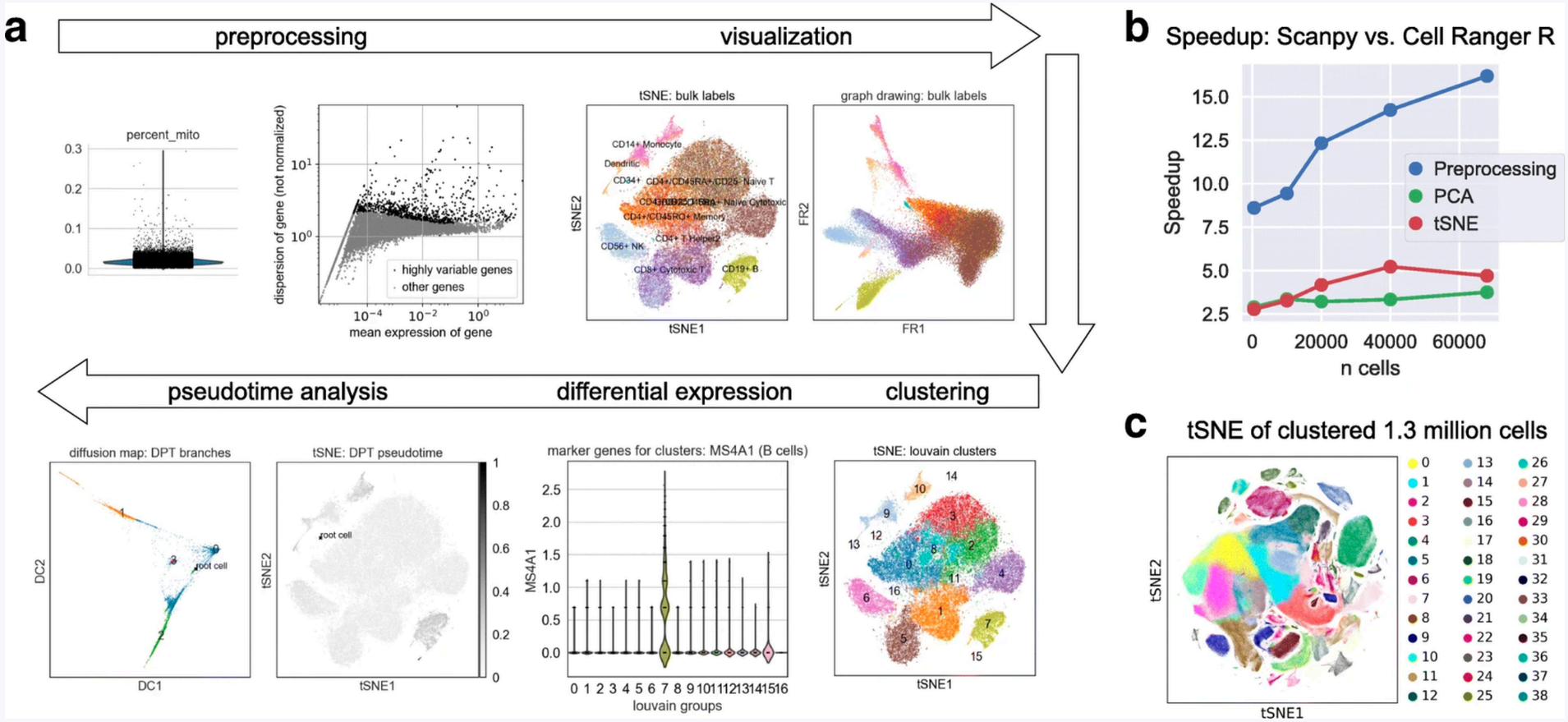
SCANPY: large-scale single-cell gene expression data analysis - Feb 6 2018

- SCANPY is a scalable toolkit for analyzing single-cell gene expression data
  - includes methods for preprocessing, visualization, clustering, pseudotime and trajectory inference, differential expression testing, and simulation of gene regulatory networks
- Python-based implementation efficiently deals with data sets of more than one million cells
- TSNE and graph-drawing (Fruchterman–Reingold) visualizations show cell-type annotations obtained by comparisons with bulk expression.



Pathology slice with partially labeled data (a), tissue classification from spectra by using 1-nearest neighbors (b), and tissue classification from spectra by using geometric diffusion (c). The three tissue classes are marked with blue, green, and pink.

SCANPY - Workflow



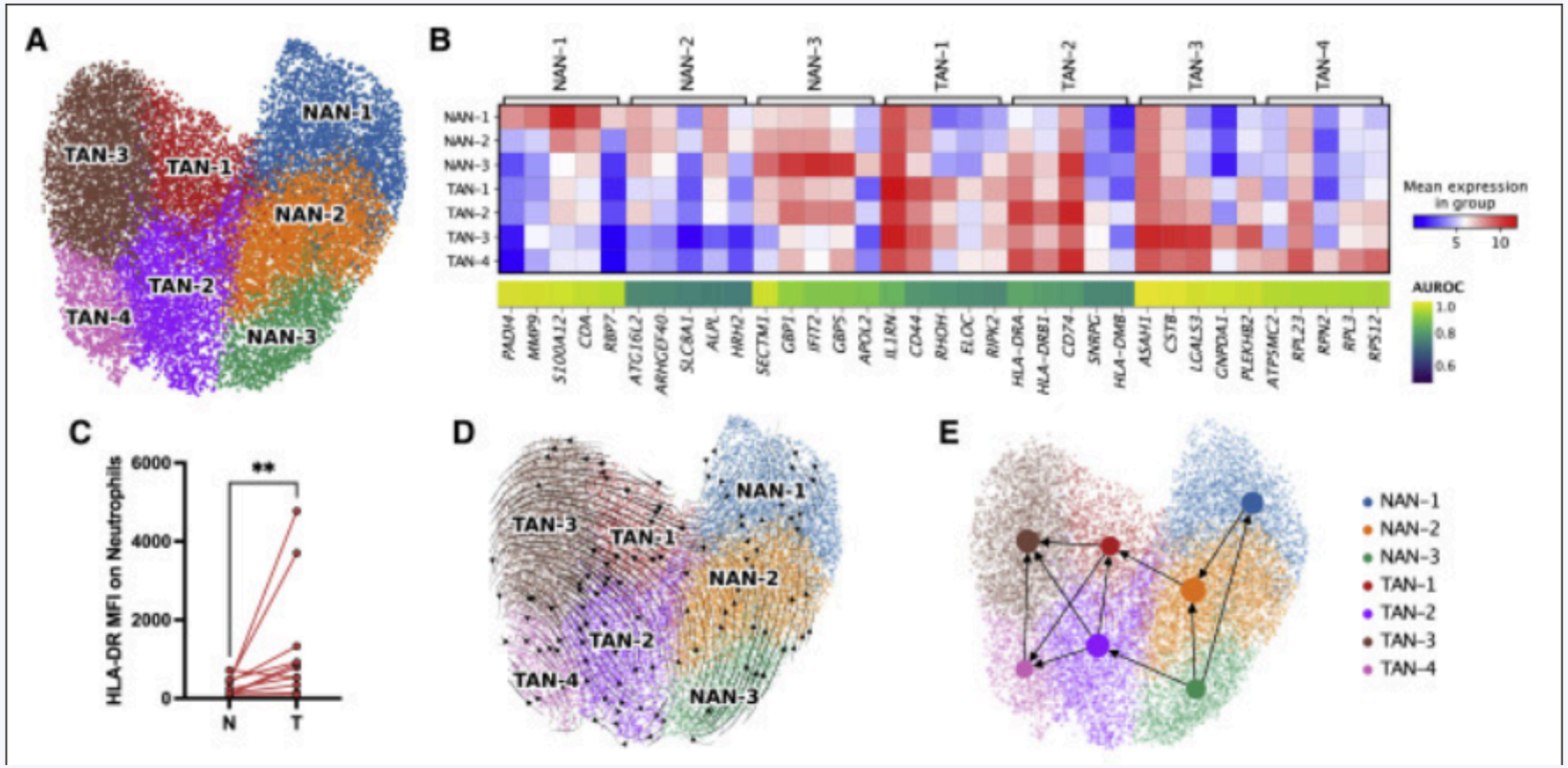
Clustering and Annotation of Transitions

- **Unsupervised clustering and cell-type annotation**
  - Computed UMAP embeddings and unsupervised Leiden-clustering with scanpy based on a cell-cell neighborhood graph derived from scANVI latent space.
  - Coarse, lineage-specific clusters were iteratively sub-clustered to identify cell-types at a more fine-grained resolution
  - Cell type clusters were annotated based on previously reported marker genes; CD8+ T cell subclusters were annotated based on gene sets
- **Comparing cell-type abundances**
  - Cell-type fractions between groups is challenging due to different characteristics of the datasets and the inherent compositional nature of cell-type fractions; applied the scCODA model to determine between LUAD vs LUSC

Data Stratification and Analyses

- **Patient Stratification**
  - Patients were clustered using graph-based Leiden clustering with the “correlation” distance metric for computing the neighborhood graph
  - Dataset-specific batch-effects were removed using a linear model - scanpy.pp.regress_out
  - Patient clusters were labeled according to their predominant cell-types
- **RNA velocity analysis -** Velocity graph visualized as a graph showing the transition confidences as directed edges
- **Differential gene expression testing -** NAN vs TAN
- **Cellphonedb analysis**
  - **Cell-to-cell inter**actions - determined interaction partners that are potentially affected by that change, as those that are expressed in at least 10% of the cells in a certain cell-type.
- **SCISSOR analysis**
  - to associate phenotypic data from bulk RNA-seq experiments with our single-cell data.
  -

Mapping and Results



Chapter III

Reproduction & Demonstration

Utilizing methodologies presented in the primary paper

Work Flow for Reproduction

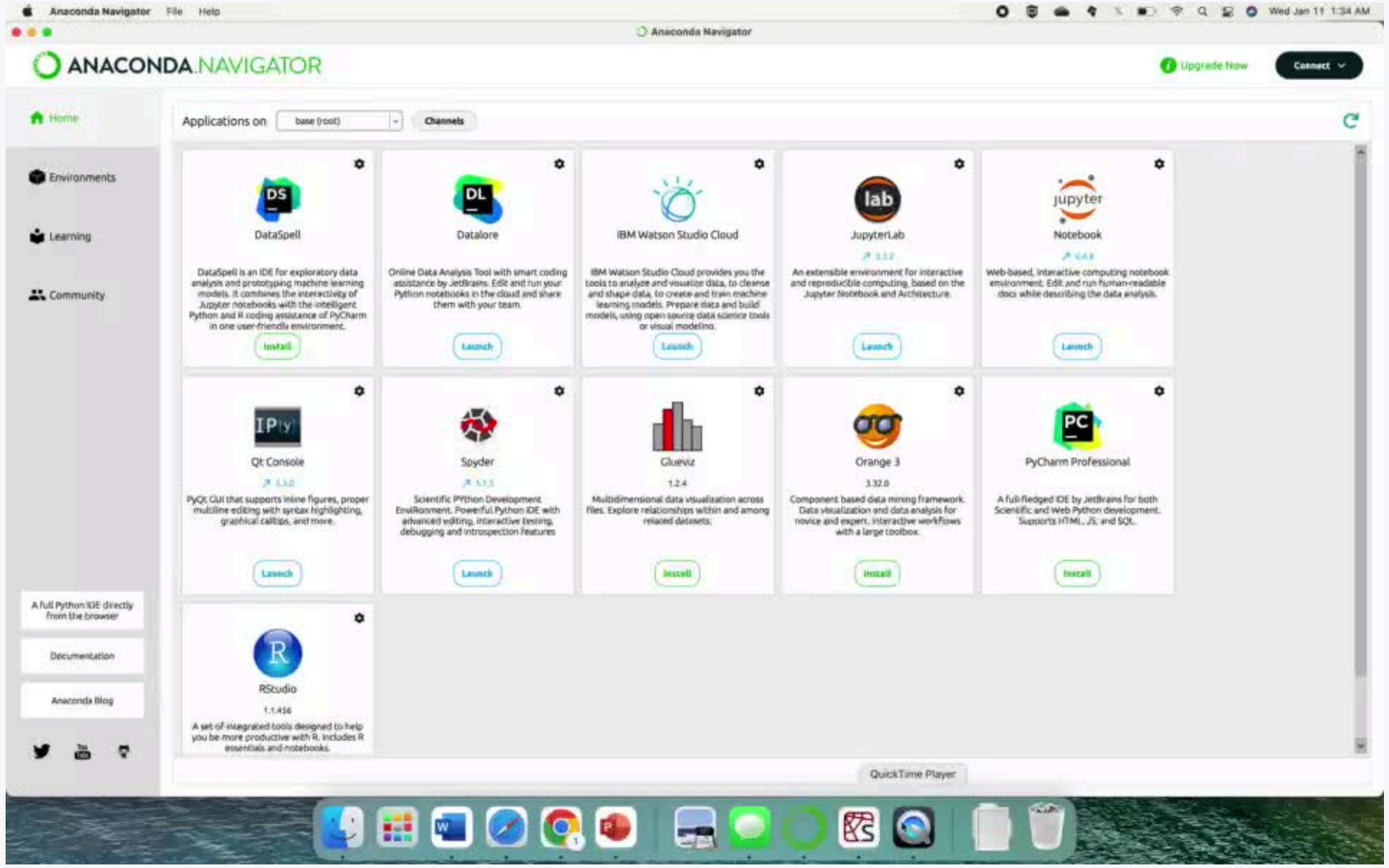
Build Atlas Workflow

- QC of the individual datasets based on detected genes, read counts and mitochondrial fractions
- Merging of all datasets into a single AnnData object. Harmonization of gene symbols.
- Annotation of two "seed" datasets as input for scANVI.
- Integration of datasets with scANVI
- Doublet removal with SOLO
- Annotation of cell-types based on marker genes and unsupervised leiden clustering.
- Integration of additional datasets with transfer learning using scArches.

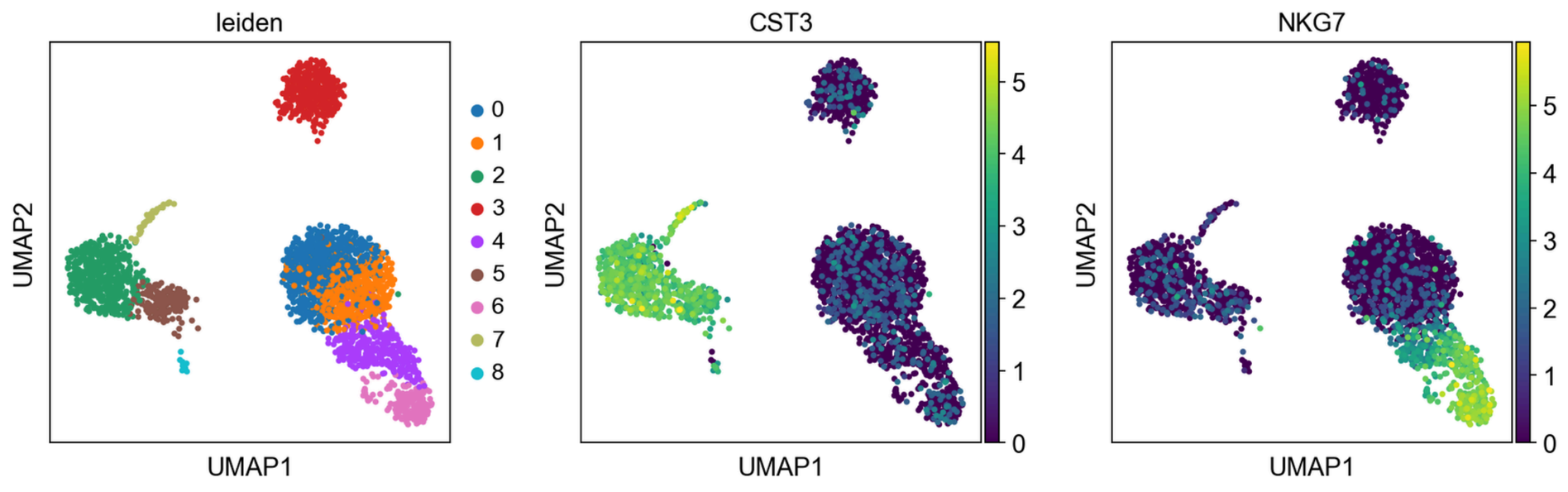
Downstream Analysis Workflow

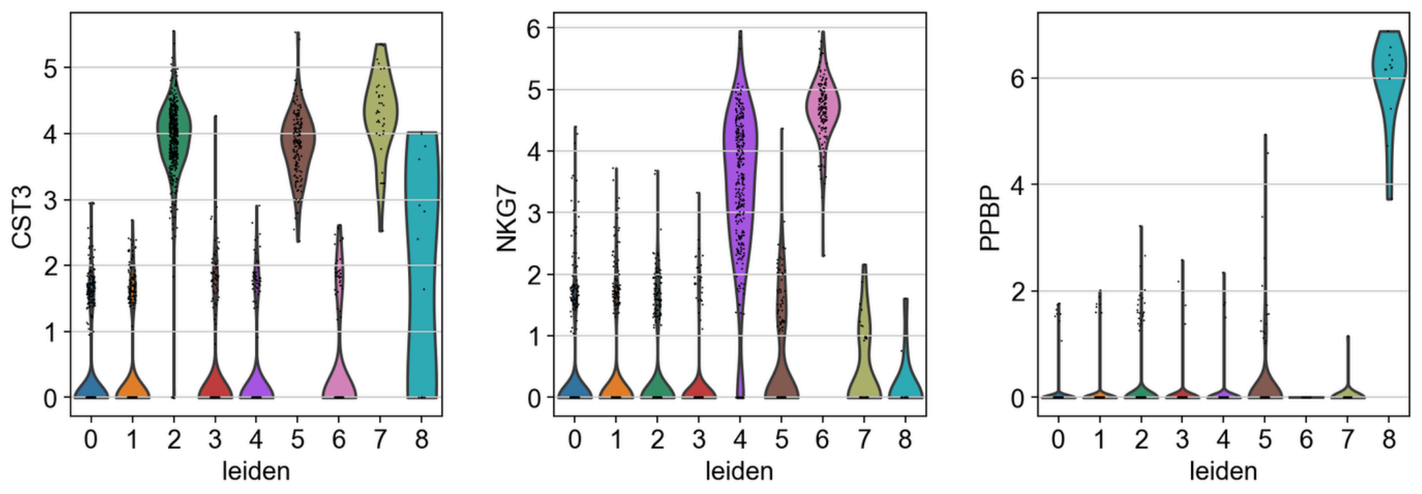
- Patient stratification into immune phenotypes
- Subclustering and analysis of the neutrophil cluster
- Differential gene expression analysis using pseudobulk + DESeq2
- Differential analysis of transcription factors, cancer pathways and cytokine signalling using Dorothea, progeny, and CytoSig.
- Copy number variation analysis using SCEVAN
- Cell-type composition analysis using scCODA
- Association of single cells with phenotypes from bulk RNA-seq datasets with Scissor
- Cell2cell communication based on differential gene expression and the CellphoneDB database.

Methodology Demonstration

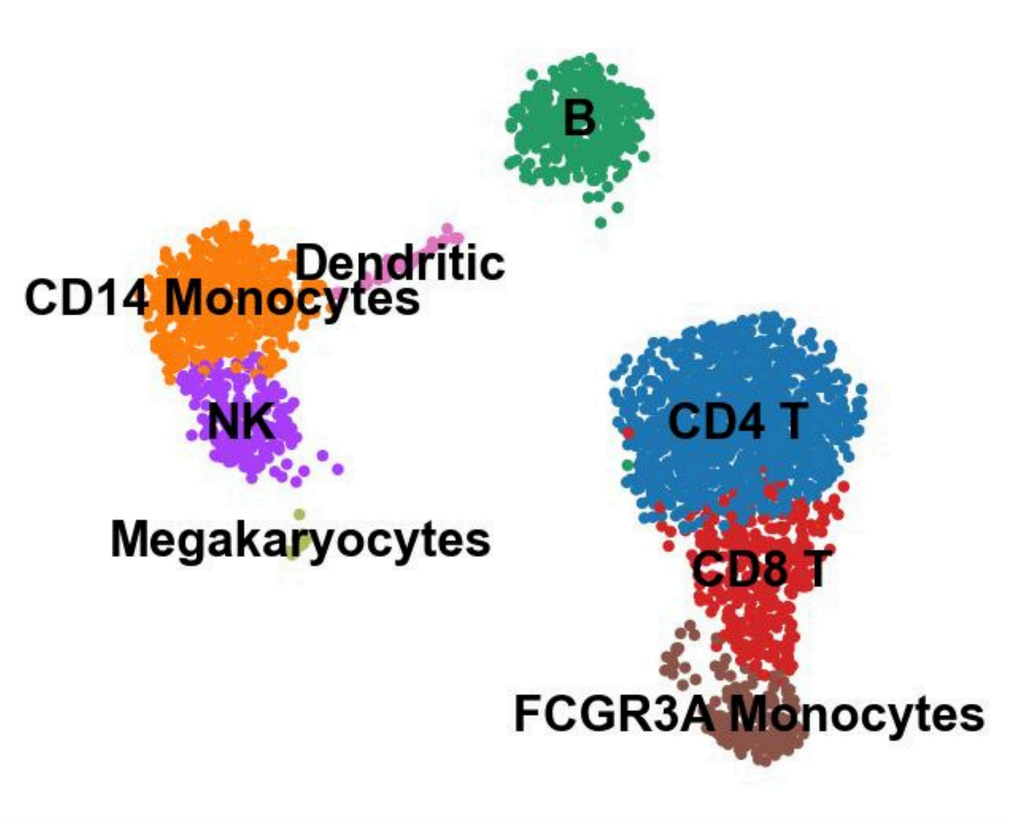


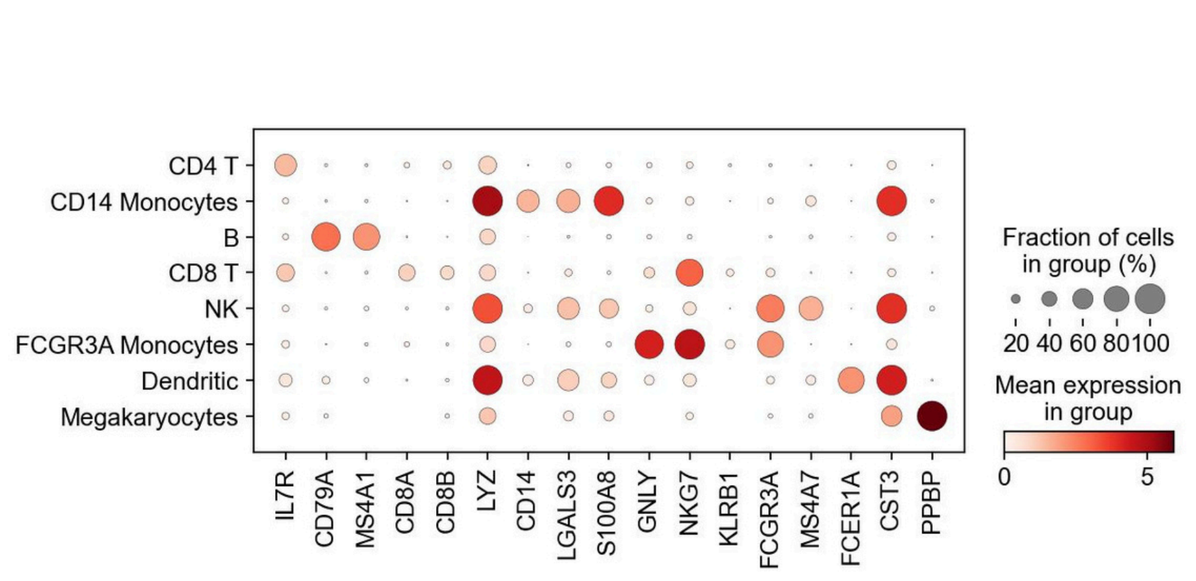
Reproduction Results





Reproduction Results (Cont.)





Chapter IV

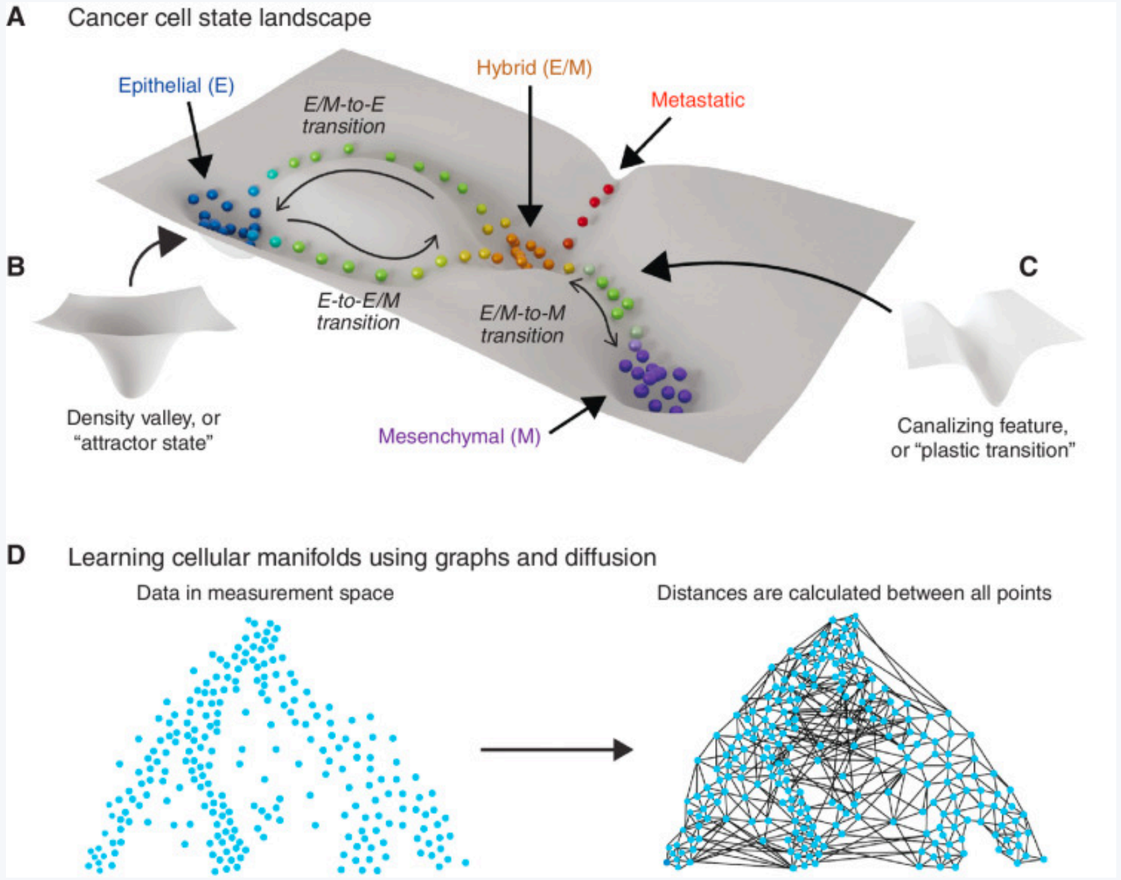
Predictive Models

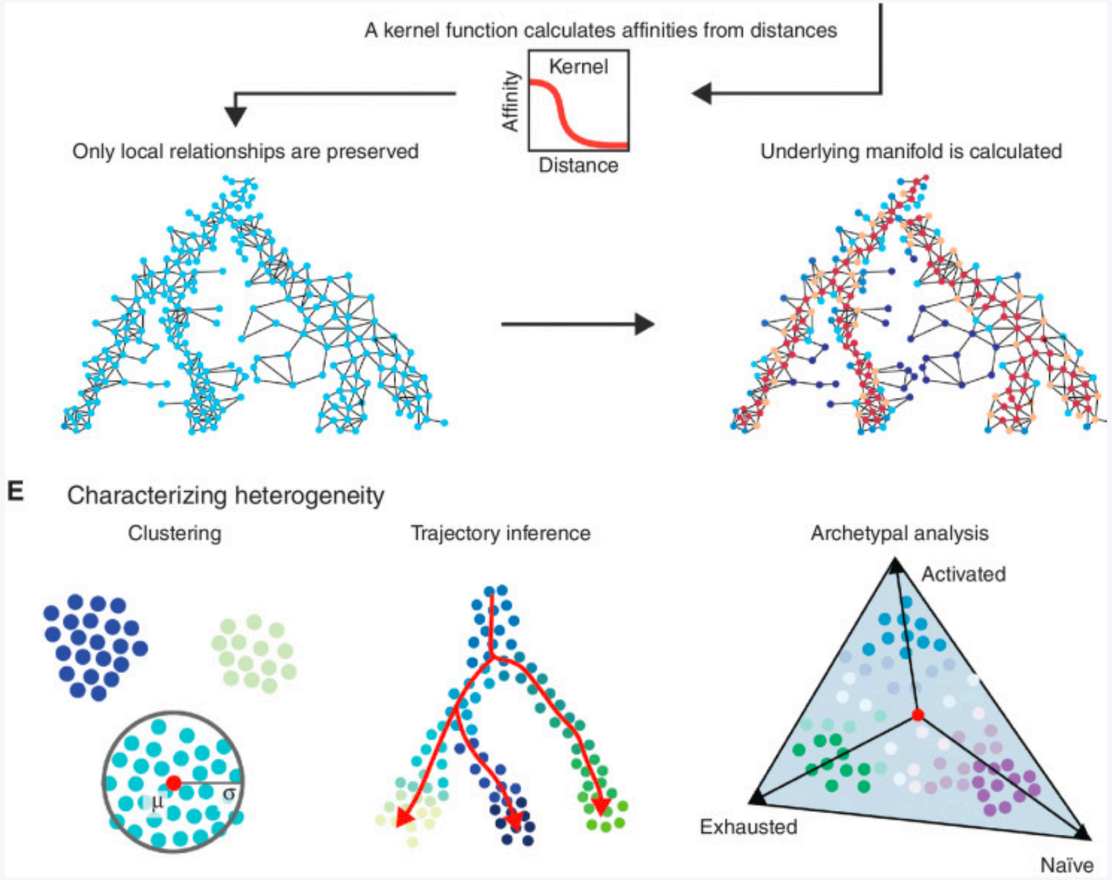
Incorporating Manifold Learning to Map Transitions (EMT)

Mapping Phenotypic Plasticity upon the Cancer Cell State Landscape Using Manifold Learning- Aug 5, 2022

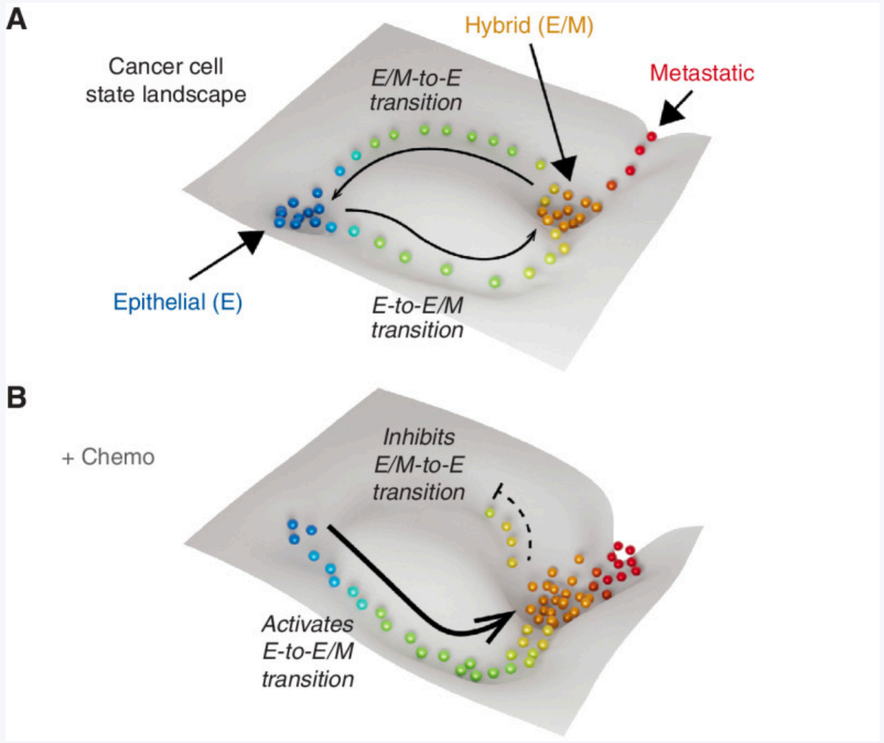
- Phenotypic plasticity describes the ability of cancer cells to undergo **dynamic, nongenetic cell state changes that amplify cancer heterogeneity** to promote metastasis and therapy evasion
- Cancer cells occupy a continuous spectrum of phenotypic states connected by trajectories defining dynamic transitions upon a cancer cell state landscape
- **Manifold learning techniques** are emerging as computational tools to effectively model cell state dynamics in a way that mimics our understanding of the cell state landscape
- “State-gating” therapies targeting phenotypic plasticity **will limit cancer heterogeneity, metastasis, and therapy resistance -** significant scientific challenge because this heterogeneity is clinically associated with aggressive disease, resistance to conventional chemotherapies, and poor overall survival
- Yet this implies that targeting specific cell populations within a tumor, or preventing the emergence of particularly aggressive subpopulations, could lead to enhanced therapies and improved survival outcomes for patients.
- **To reach this goal, strategies to identify and therapeutically target intratumoral cell populations are vital**

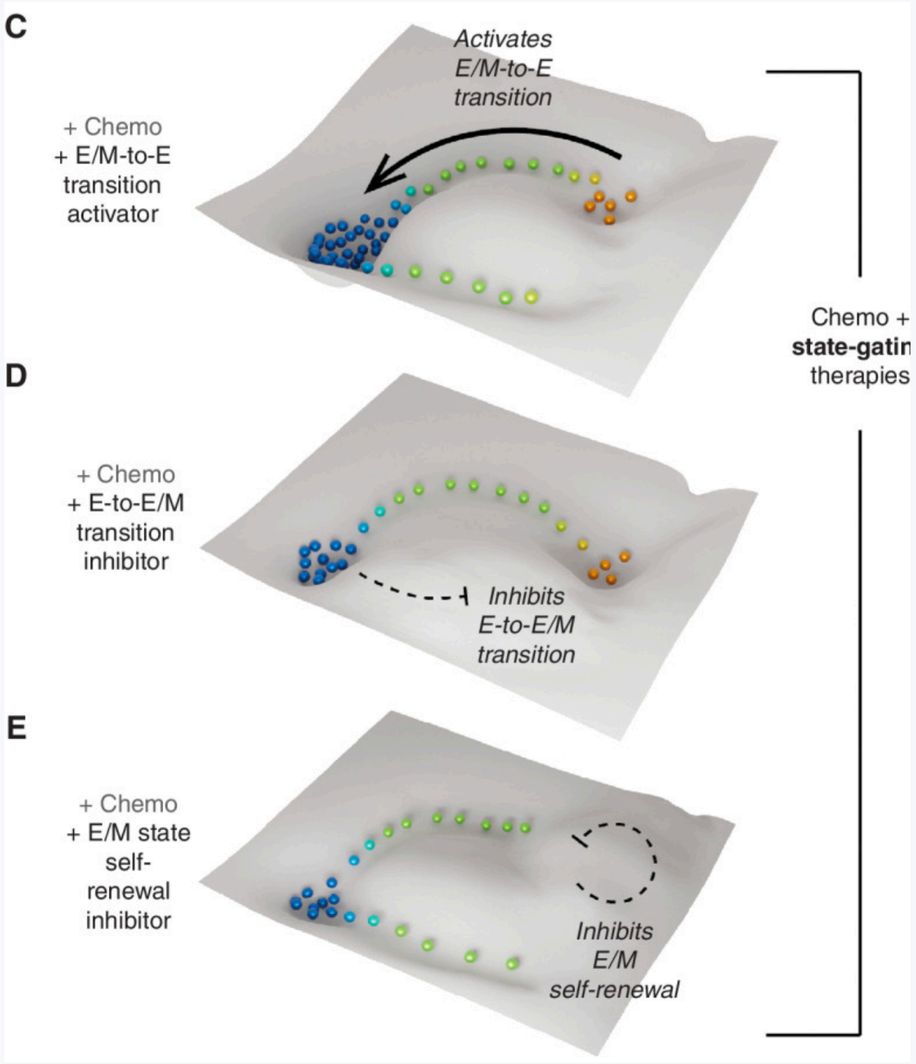
Manifold Learning Techniques - Diagram





Manifold Learning Techniques - Therapeutic Implementation





Chapter IV

Future Work & Discussion

Next Steps and Q&A Session

Conclusions & Next Steps

- Predictive approaches can be successfully incorporated into the mapping of tumor cell transitions in order to advise & guide therapeutic approaches; manifold learning offers an untapped reservoir to investigate plasticity
  - Embedding predictive models into transitional analysis is at the **forefront** of this field
- Next steps include this incorporation of manifold learning into the existing reproduction
  - Focus in on EMT, MET, or related transitions
  - Analyze further papers for deeper understanding of cell-intrinsic and cell-extrinsic trajectory influencers
  - Dissect further related transitions, as well as how this methodology applies to other parts of the TME
- Furthermore, **how do we find the drivers that lead to state changes so we can understand more oof how the TME leads to Phenotypic plasticity?**
  - Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer - Dec 9, 202

Tues Jan 17th

Computational Health Informatics Program

Thank You!

Any Questions?

RESEARCHER

Aniket Dey
Dougherty Valley High School

ADVISOR

Dr. Felix Dietlein
Harvard Medical School