Query to reference single-cell integration with transfer learning

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Ziga Avsec, Alexander V. Misharin, Fabian J Theis

Problem Setting

- Have a set of large single-cell reference atlases
- Want to learn from references for improved analysis of new data
- Many analysis difficulties:
  - May not have access to reference data
  - Technical batch effects between and within datasets
  - Biological perturbations between and within datasets
  - Tedious to cluster and annotate new data
  - May not have computational resources

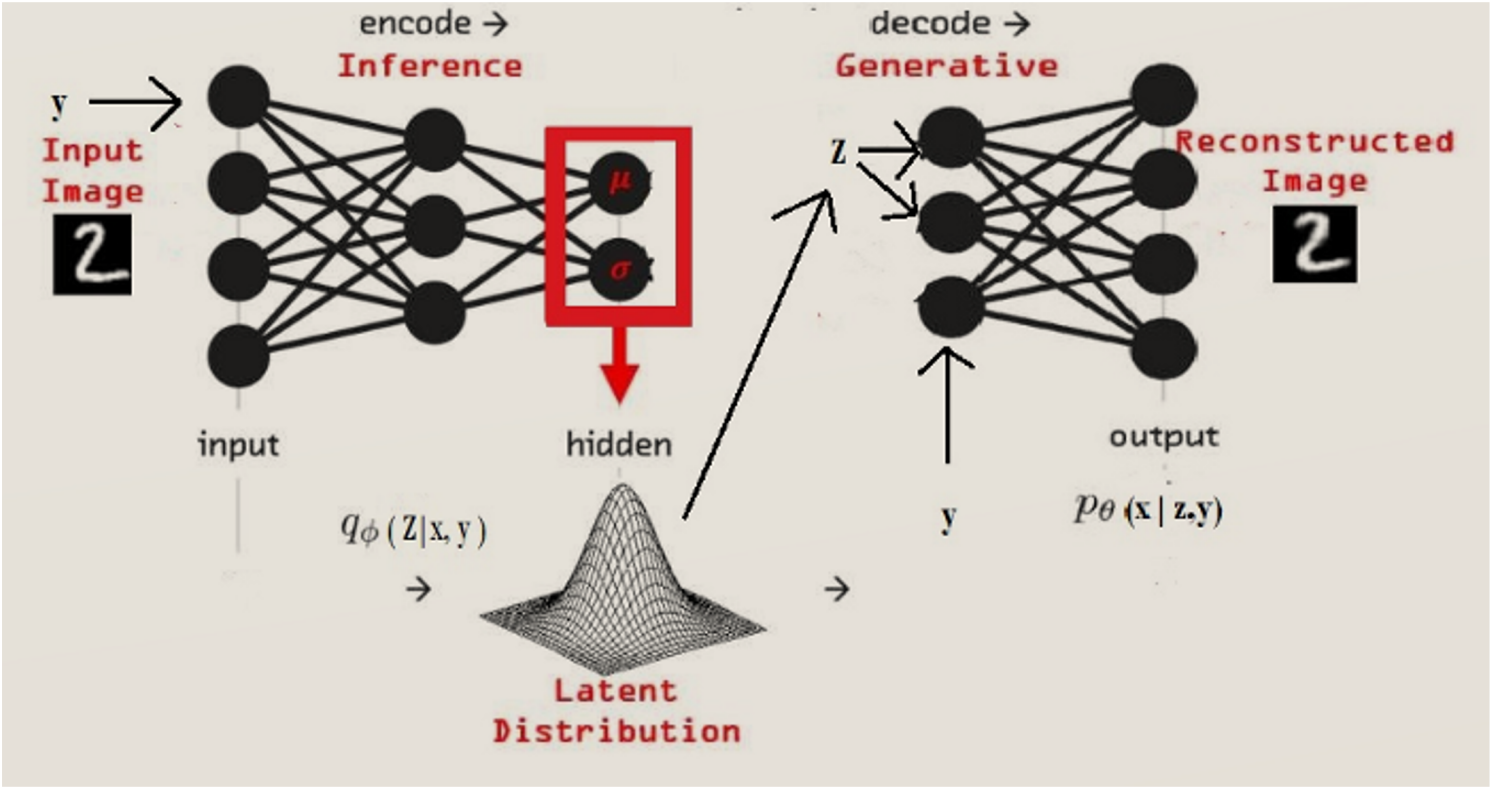
Goals

- Automate clustering and annotation of new datasets
- Enable easy comparison across tissues, species, and disease conditions
- Share knowledge even with data privacy restrictions

Methods

- Transfer learning
- Model sharing
- Architecture surgery
- Deep generative models

Conditional Variational Autoencoder (cVAE)



Plot from https://towardsdatascience.com/understanding-conditional-variational-autoencoders-cd62b4f57bf8

scArches: Model Setup

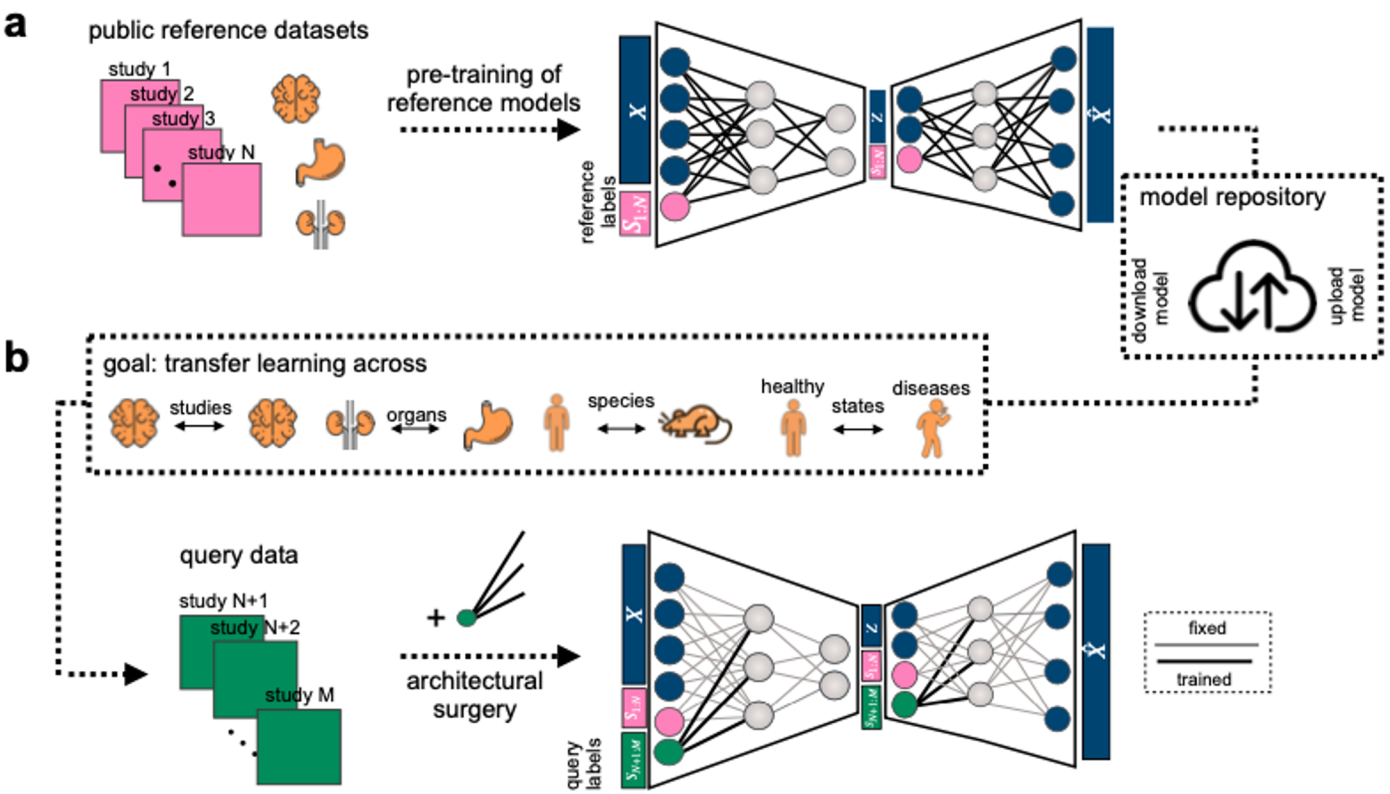
- Have *N* reference datasets
  - Gene expression data X_i
  - Categorical study label S_i
- Pretrain model with $X_{1:N}$, $S_{1:N}$
- Have access to weights from pretrained model

scArches: Model Setup

- Have *N* reference datasets
  - Gene expression data X_i
  - Categorical study label S_i
- Pretrain model with $X_{1:N}$, $S_{1:N}$
- Have access to weights from pretrained model

- Get *M* new query datasets
  - Each has X_i , S_i

scArches

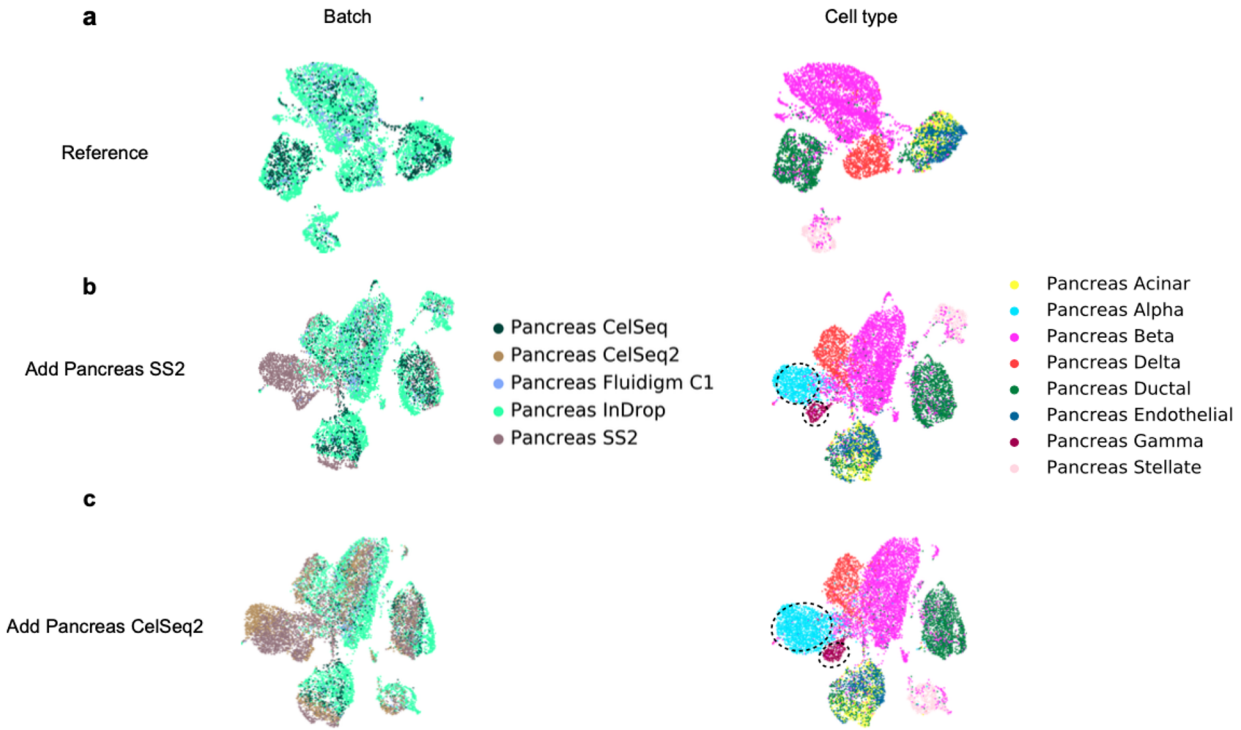

a
public reference datasets
study 1
study 2
study 3
study N
pre-training of
reference models
reference
labels
model repository
download
model
upload
model
b
goal: transfer learning across
studies
organs
species
healthy
states
diseases
query data
study N+1
study N+2
study M
architectural
surgery
query
labels
fixed
trained

scArches: Pancreas Experiment

- Reference atlas of 3 pancreas studies
  - Remove all alpha cells, gamma cells
- Query with 2 new pancreas studies
  - Include alpha cells, gamma cells
- All different sequencing technologies

- Expectations:
  - Shared cell types have similar latent representations to reference
  - New cell types (alpha and gamma cells) have different latent representations

scArches: Pancreas Experiment



Evaluation

- Entropy of batch mixing (EBM)
  - Higher scores -> better mixing of cells across batches in latent space
- K-Nearest Neighbors (KNN) purity
  - Higher scores -> small neighborhood of cells in original data mapped to same neighborhood in latent space
- Tradeoff between the two

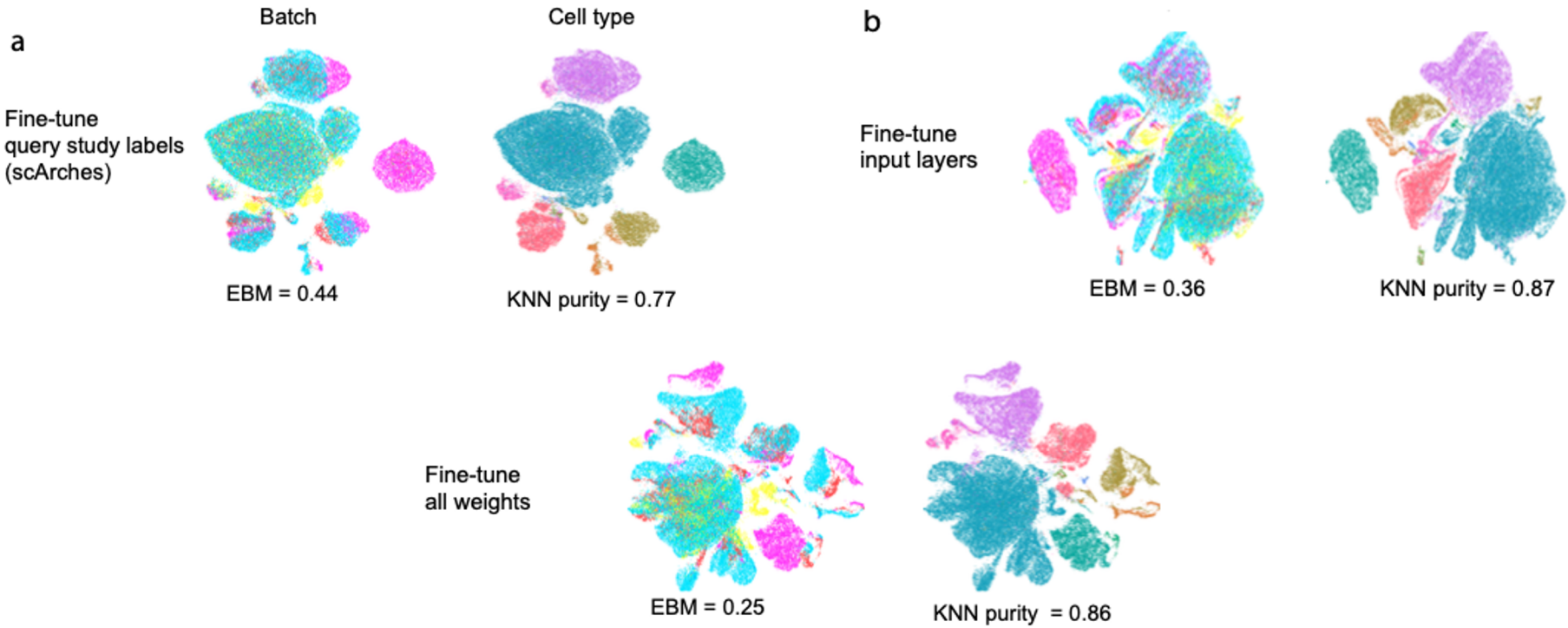
Transfer Learning Approach

- scArches approach
  - Fine-tune newly-introduced weights in first layer of encoder
  - Fine-tune newly-introduced weights in first layer of decoder
- Other considered approaches
  - Fine-tune all weights in first layer of encoder, first layer of decoder
  - Fine-tune all model weights

TL Approach Experiment

- Mouse brain datasets
  - 2 reference
  - 2 query
- Look for:
  - Good batch mixing
  - Preservation of distinct clusters for different cell types

Transfer Learning Approach



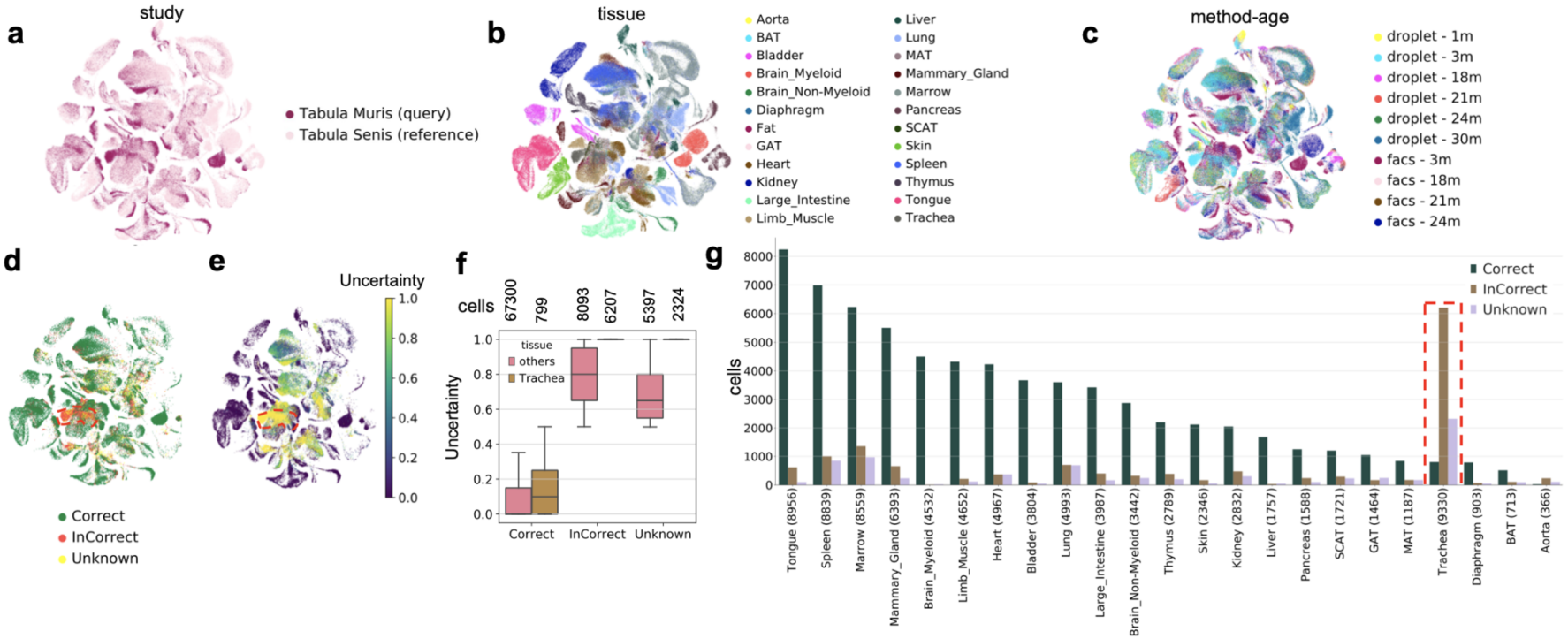
Benchmarking efforts

- Batch correction compared to existing fully-trained methods
  - Seurat v3, Harmony, Liger, Scanorama, MNN correct, Conos, trVAE
  - KNN purity and EBM as performance metrics
  - Tested across 2 organs & 4 data sets
    - scArches + trVAE on par in preserving internal substructures in orig. data
    - Outperformed on mixing across studies
    - Substantially outperforms baseline trVAE without TL
- Effect of dataset size on integration quality
  - Subsamples of varying sizes
  - Increasing sample size → increasing KNN & EBM across datasets & sample sizes for scArches and scArches + trVAE
  - Outperforms all other methods in presence of low cell numbers where TL is beneficial
  - Outperforms other models' integration in large data regimes

scArches: mapping across tissues, trachea experiment

- Reference atlas of 155 cell types across 23 tissues and 5 age groups (1-30 months)
  - Remove tracheal cells
- Query atlas contains 90,120 cells at 3 month time point from 24 tissues
  - Includes tracheal cells
- Reported successful integration across time points and sequencing technologies
  - Distinct cluster of tracheal cells identified (n = 9,330)
- To test transfer of cell type labels from reference:
  - Trained a KNN classifier on reference latent space
  - Each query cell annotated by nearest reference neighbor, given uncertainty score
  - Report 89% label transfer accuracy (except for trachea)
  - Misclassified cells and out of dist. cells received high uncertainty scores

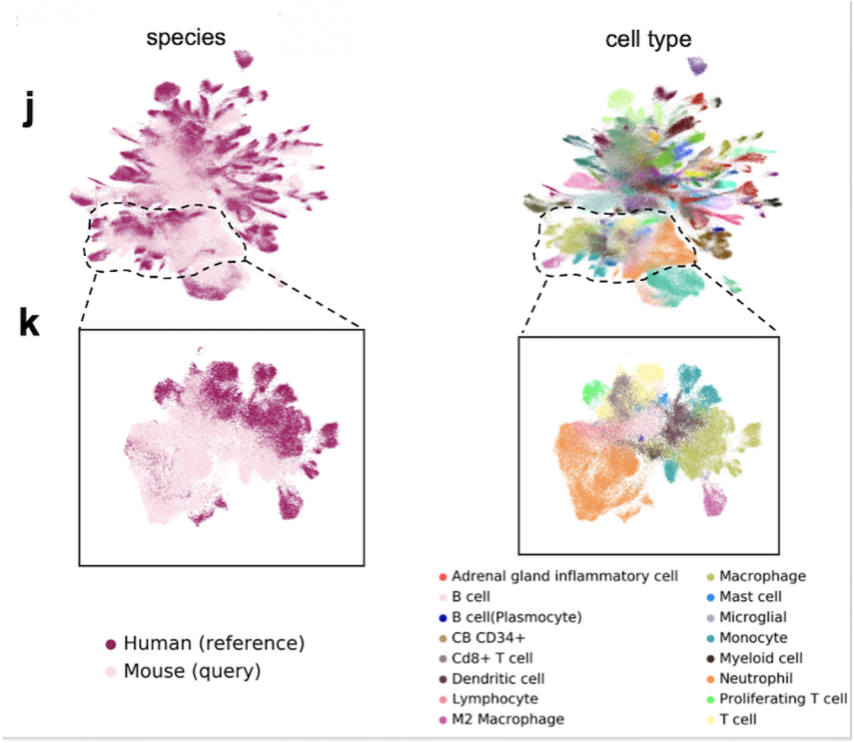
scArches: mapping across tissues, trachea experiment



scArches: mapping across species, human-mouse experiment

- Reference model trained on Human Cell Landscape (HCL)
  - 249,845 cells across 63 human tissues
- After architecture surgery, aligned Mouse Cell Atlas (MCA), n=122,944, into reference human cell atlas
- Different profiling and sequencing technologies
- Expectation: all cell types won’t overlap due to species-specific cell types and functions
- Result:
  - similar immune cell types (e.g. neutrophils, macrophages) clustered together across species
  - species-specific cells placed separately
- Strong regularization of transfer from reference via scArches
  - Overcome strong species biological effect
  - Focus on gene expression similarity across major mammalian cell types

scArches: mapping across species, human-mouse experiment



scArches: mapping across disease states

- Essential to contextualize query data with healthy reference to study disease
- 3 criteria for disease-to-healthy data integration
  - Preservation of biological variation of healthy cell states
  - Integration of matching cell types between healthy reference and disease query
  - Preservation of distinct disease variation, e.g. emergence of new cell types unseen during healthy reference building

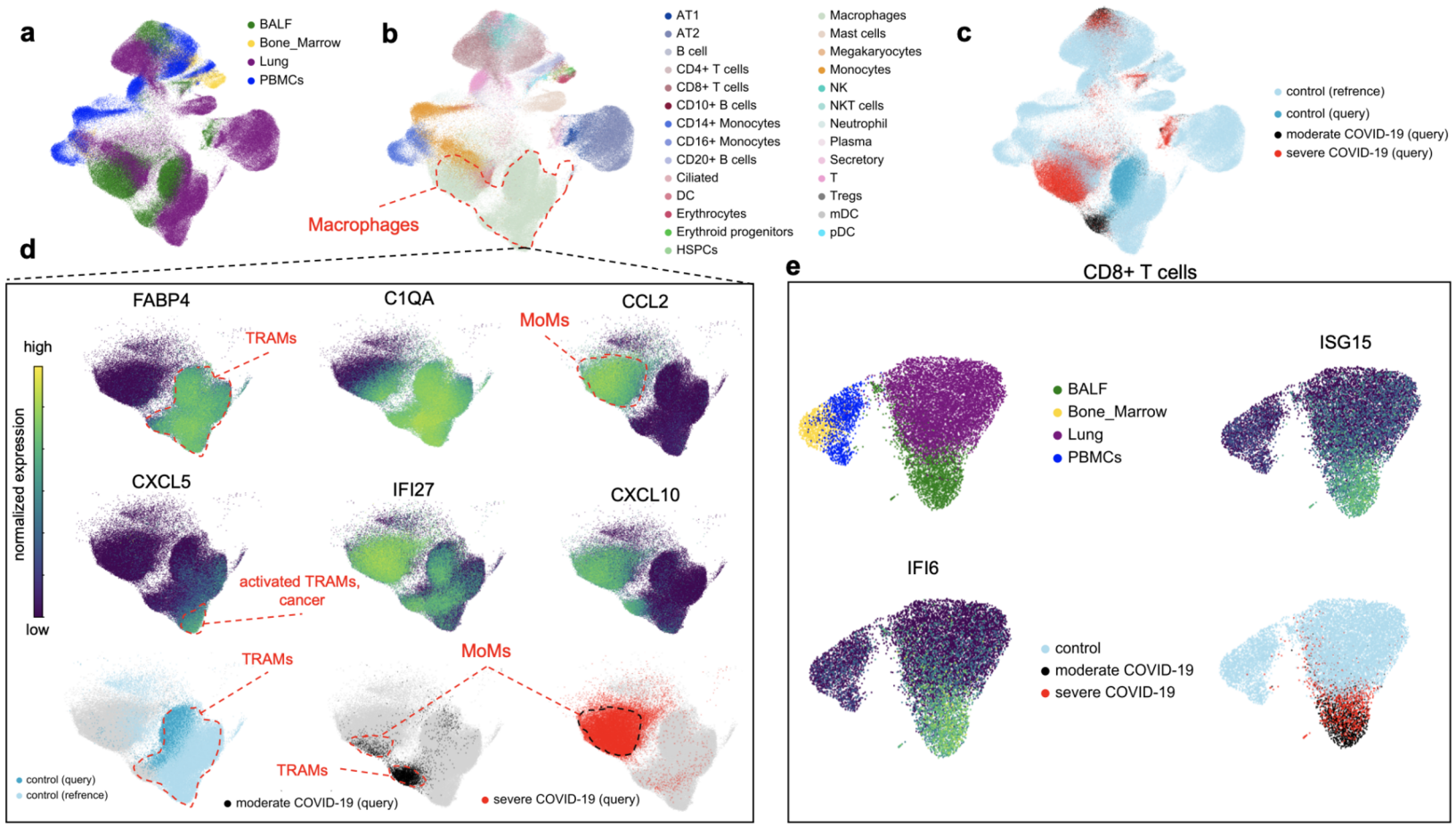
COVID-19 experiment background

- Reference: bone marrow, PBMCs and normal lung tissue (n = 154,723)
- Query: immune & epithelial cells from healthy controls, and patients with moderate & severe COVID-19 (n = 62,469)
  - airway epithelial cells, plasma and B cells, CD8+ T cells, neutrophils, monocytes, mast, natural killer cells, dendritic cells, and macrophages
- Immunology review
  - Monocytes and macrophages
  - CD8+ T cells

COVID-19 experiment findings

- Healthy query data integrates well with healthy reference
- Macrophage cluster: 2 main groups
  - TRAMs (tissue-resident alveolar macrophages), found in healthy tissue
  - MoMs (monocyte-derived inflammatory macrophages), not found in healthy tissue
- Cell activation/expression state can influence data mixing degree
  - Difference in expression of TRAMs in COVID-19 vs. healthy lung tissue
- MoMs placed in closer proximity to monocytes than TRAMs
  - reflects ontological relationship
  - gradient of C1QA expression use to differentiate between monocytes & MoMs
- Activation of CD8+ T cells in immune response also reflected in distinct clusterings of COVID-19 patients and healthy lung references

scArches: COVID-19 experiment



Let’s discuss...

- The importance of choosing the right reference
- How does this model compare to other implementations (was adequate benchmarking done?)?
- What sort of quality control is done on user query data?
- What methods are there for batch-effect differentiation (e.g. lab-to-lab variation vs. healthy-disease variation)
- What are the security and privacy implications of this system?
- In mapping diverse datasets to each other, e.g. mouse and human atlases, how is bias to one species over the other controlled/balanced over time?