Bayesian Inference for Single-cell ClUstering and ImpuTing (BISCUIT)

Elham Azizi

Memorial Sloan Kettering Cancer Center™

BioC 2017: Where Software and Biology Connect

Profiling Tumor-Immune Ecosystem in Breast Cancer

- Immunotherapy treatments successful only in a subset of patients and cancer types
- Underlying biology determining success is not known
- Variability in responses suggest a complex immune environment
- **Goal:** Unsupervised characterization of tumor-infiltrating immune subpopulations across subtypes of breast cancer, identify impact of environmental cues
- Understanding the tumor-immune ecosystem can guide development of treatments to activate immune cells against the tumor
- **Pilot Data:** Single-cell RNA-seq 9000 CD45+ immune cells from 4 tumors (patients)

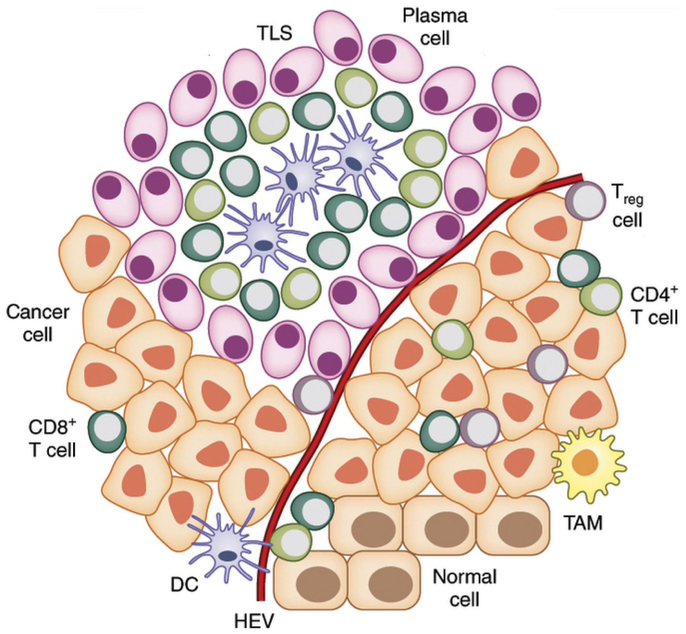
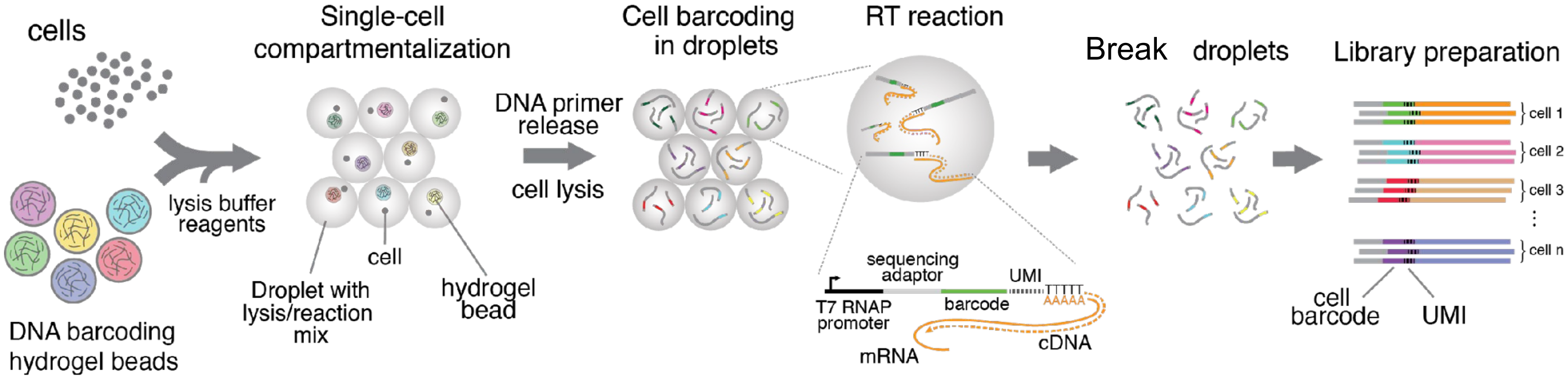


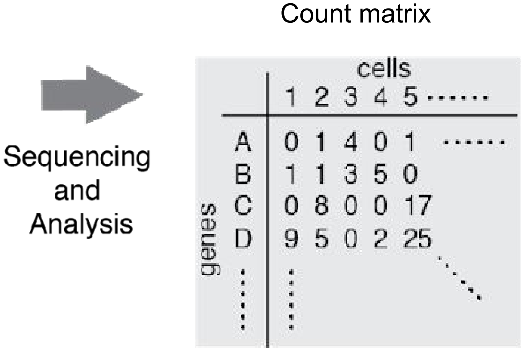
*figure adapted from Kroemer Nat Med 2015

Collaboration with Alexander Rudensky, MSKCC

Single-cell RNA-seq reveals heterogeneity in expression

Measurement of gene expression at resolution of single cells



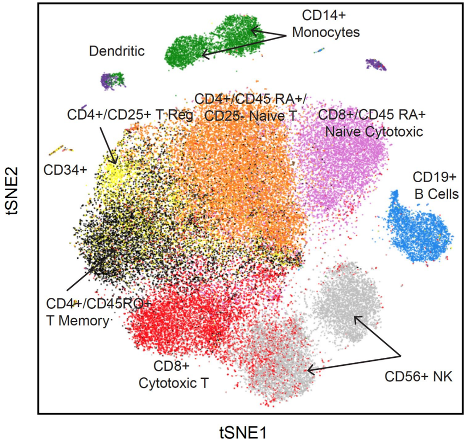


Count matrix

Sequencing and Analysis

genes	cells 1	2	3	4	5	......
A	0	1	4	0	1	......
B	1	1	3	5	0	
C	0	8	0	0	17	
D	9	5	0	2	25	⋱
⋮	⋮					

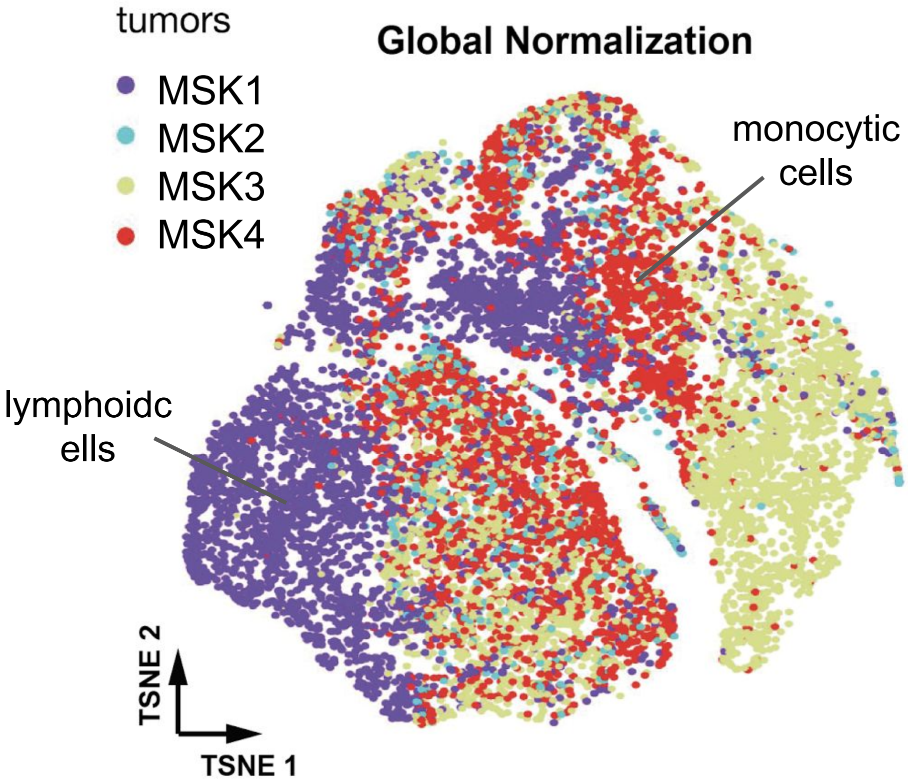
Allows characterizing novel cell types and functions based on heterogeneity



PBMC single-cell data (Zheng et al. bioRXiv 2016)

indrop (Klein et al *Cell* 2015)

Single-cell RNA-seq data for immune cells from 4 breast cancer tumors



Global Normalization

monocytic cells

lymphoidc ells

9000 CD45+ cells from 4 tumors

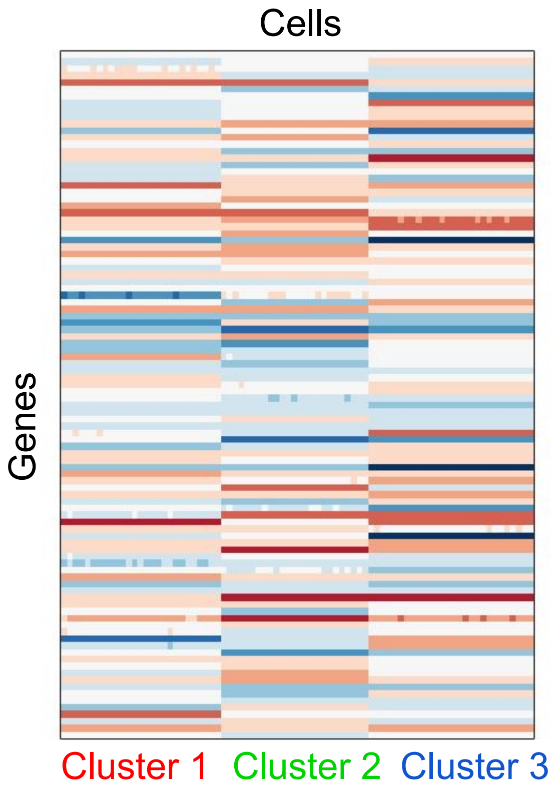
- Normalization by library size
  - Unclear structure of cell types
  - Large patient biases

Problems of scRNA-seq data:

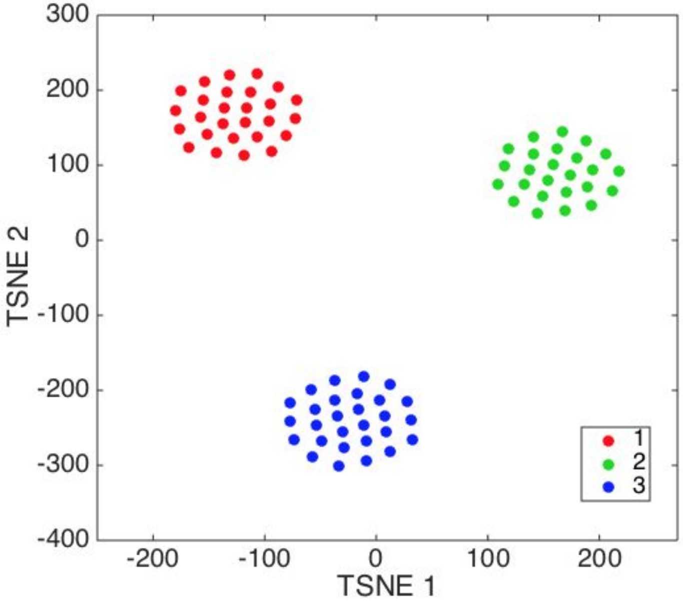
- Sampling sparse amounts of mRNA leads to “Drop-outs”
- Amplification differences
- Cell-type specific capture rates

Goal: Characterizing cell subpopulations using Single-cell RNA-seq data

Count Matrix

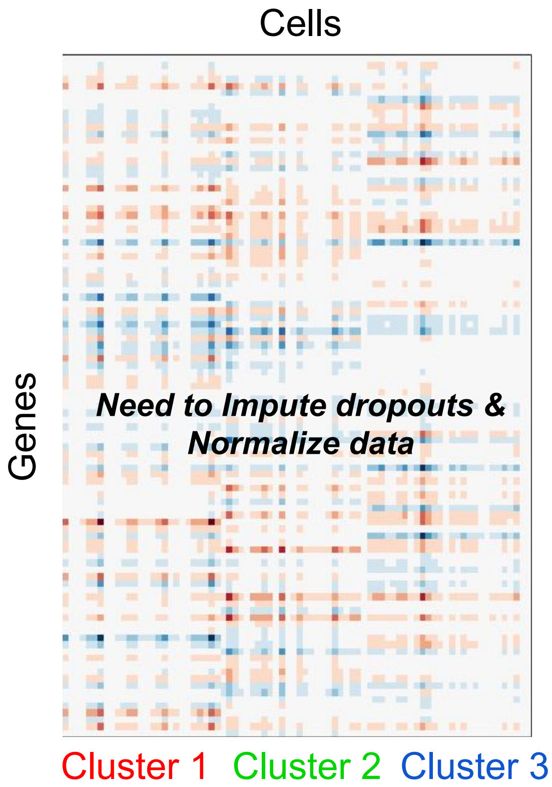


2D projection of cells (TSNE)



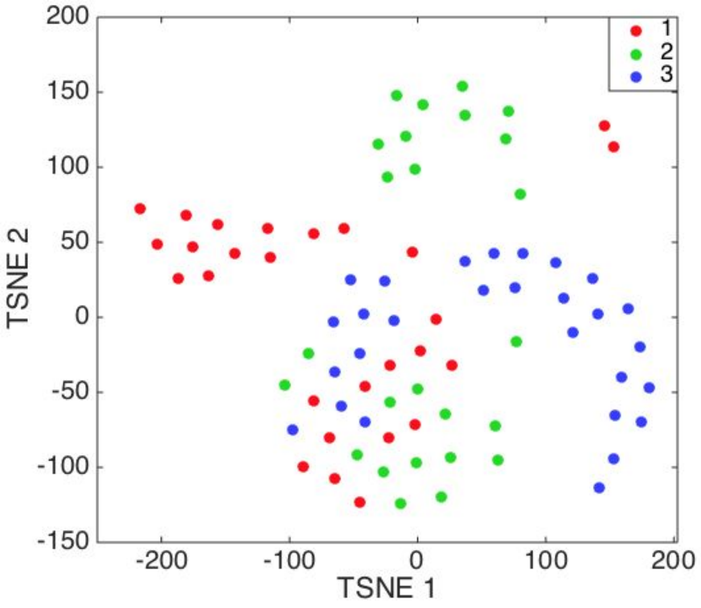
Problems: Single-cell RNA-seq data involves significant dropouts and library size variation

Observed Count Matrix



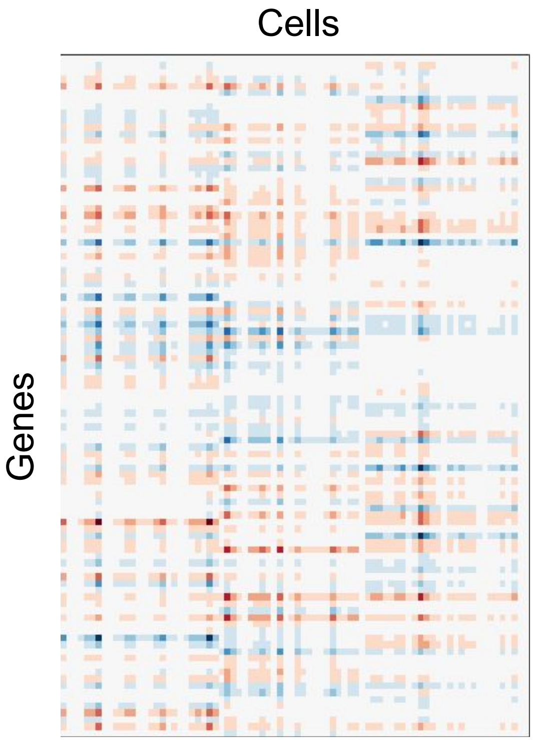
Need to Impute dropouts & Normalize data

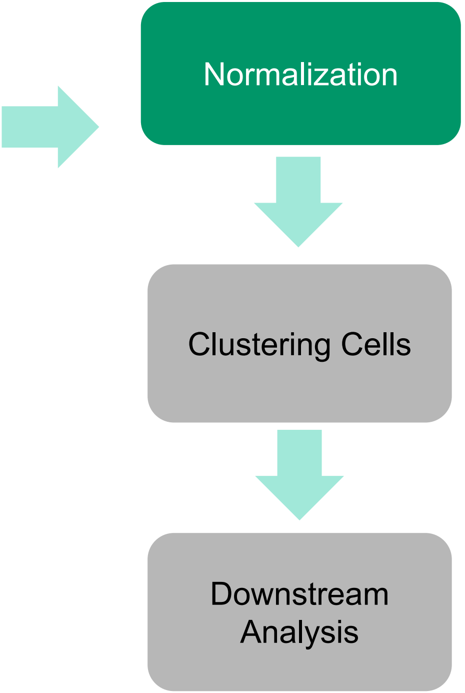
2D projection of cells (TSNE)



Common Approach:
Normalizing independent of cell types

Observed Count Matrix





Normalization

Clustering Cells

Downstream Analysis

To mean/median library size
Downsampling
BASiCS with spike-ins/ERCCs

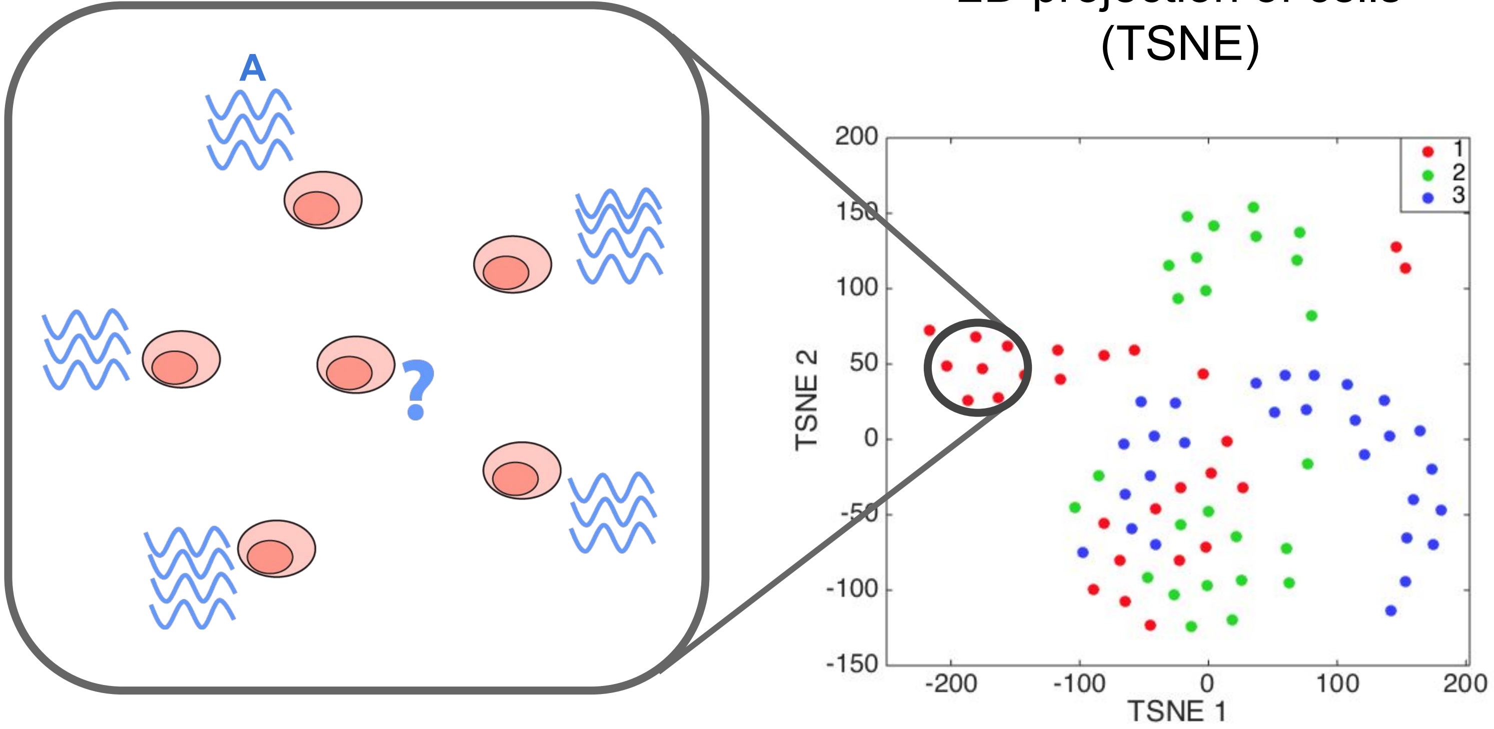
Problems:

- **Dropouts not resolved Zeros remain zero!**
- Removes biological stochasticity specific to cell type
- Leads to improper clustering; Biased downstream analysis

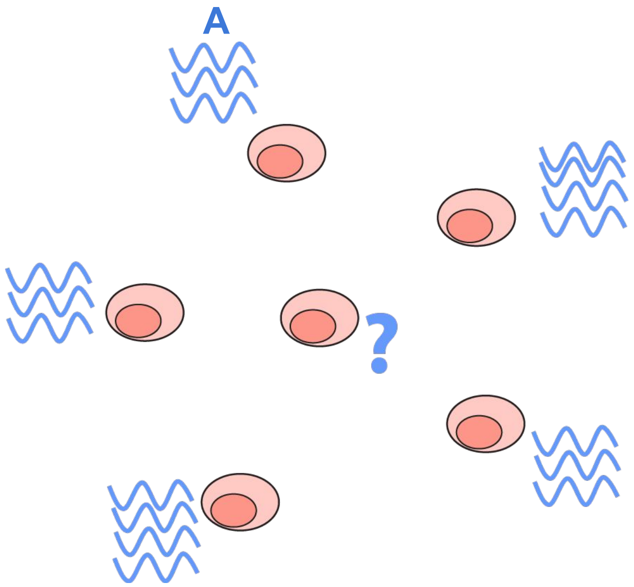
Main Concepts behind Biscuit for Normalization and Imputing

Two ideas for imputing expression in Single-cell RNA-seq data

2D projection of cells (TSNE)



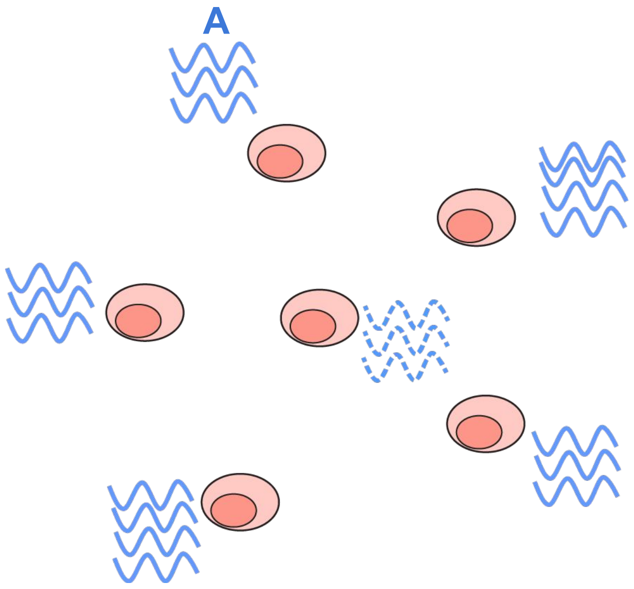
Idea 1: Impute dropouts based on cell type



No expression of
Gene A in a cell

But we observe cells
with same type mostly
have high expression of
Gene A

Idea 1: Impute dropouts based on cell type

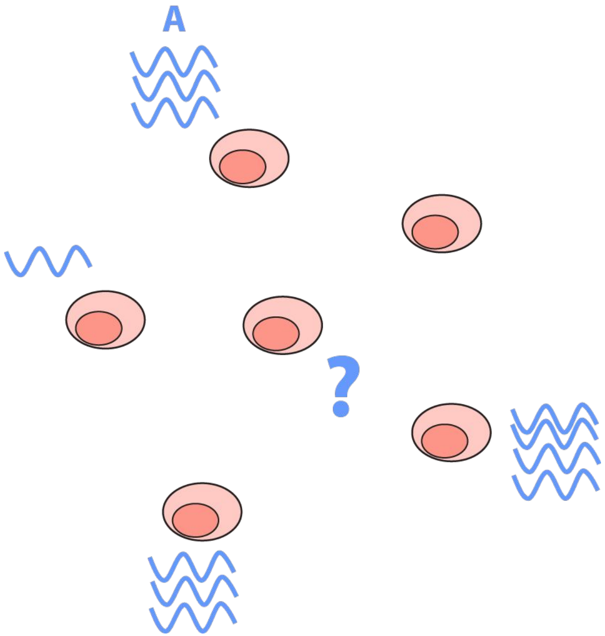


No expression of
Gene A in a cell

But we observe cells
with same type mostly
have high expression of
Gene A

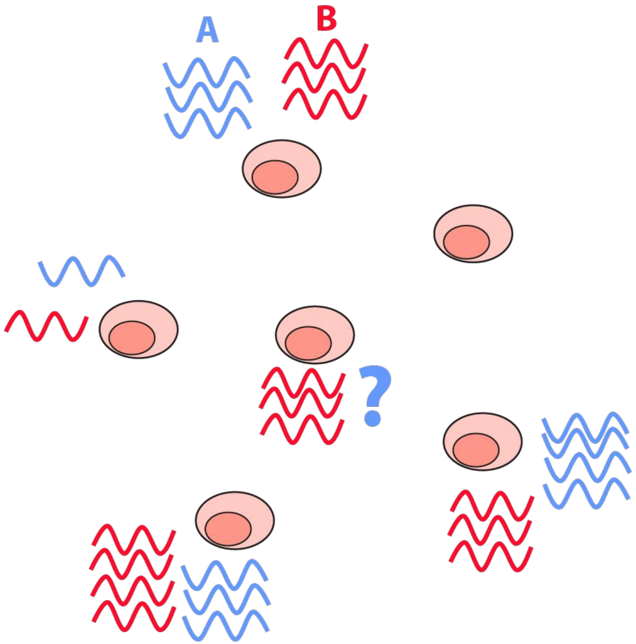
Impute dropout in Gene
A based on similar cells

Idea 2: Impute dropouts based on co-expression patterns



No significant inference based on similar cells

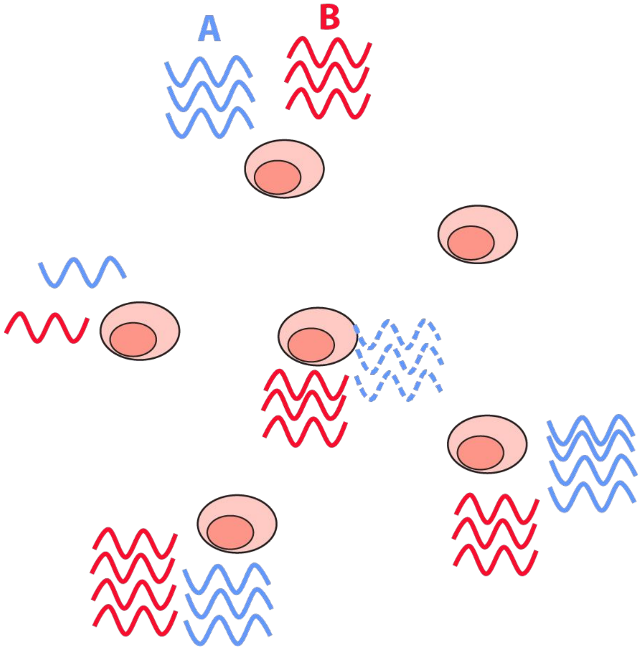
Idea 2: Impute dropouts based on co-expression patterns



No significant inference based on similar cells

However **Gene A** always co-expressed with Gene B in cells of same type

Idea 2: Impute dropouts based on co-expression patterns



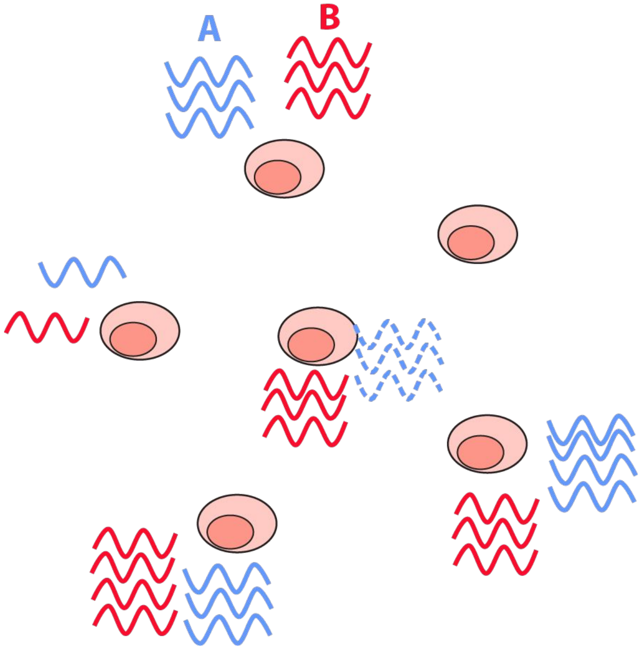
No significant inference based on similar cells

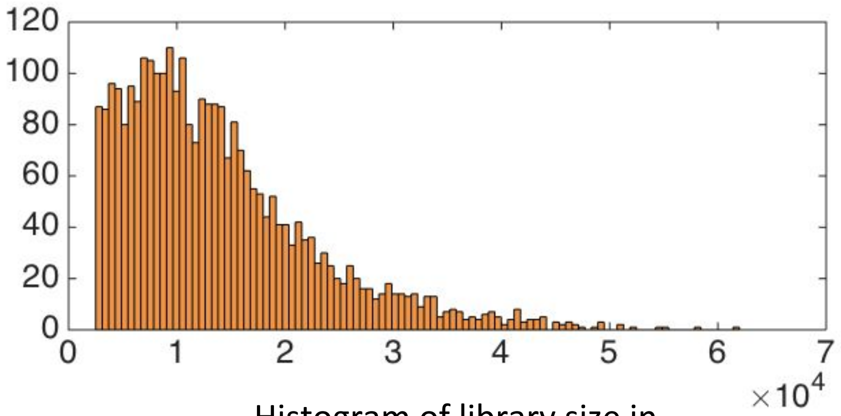
However **Gene A** always co-expressed with Gene B in cells of same type

Impute dropout in Gene A based on Gene B

Problem #2: Normalization

Normalization of Single-cell RNA-seq data

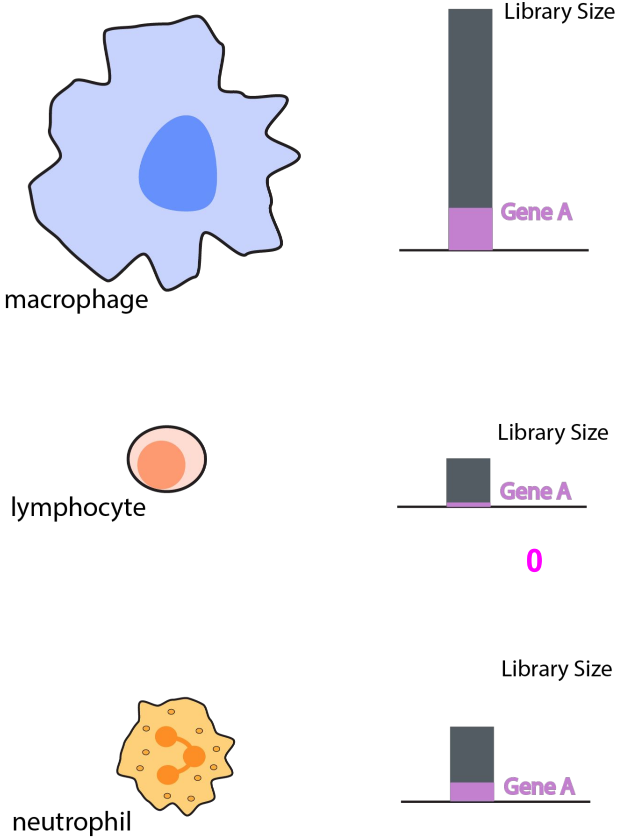




Histogram of library size in example SC dataset From Zeisel, Science 2014

In addition to imputing dropouts, we need to **normalize** data by library size

Problem with Global Normalization

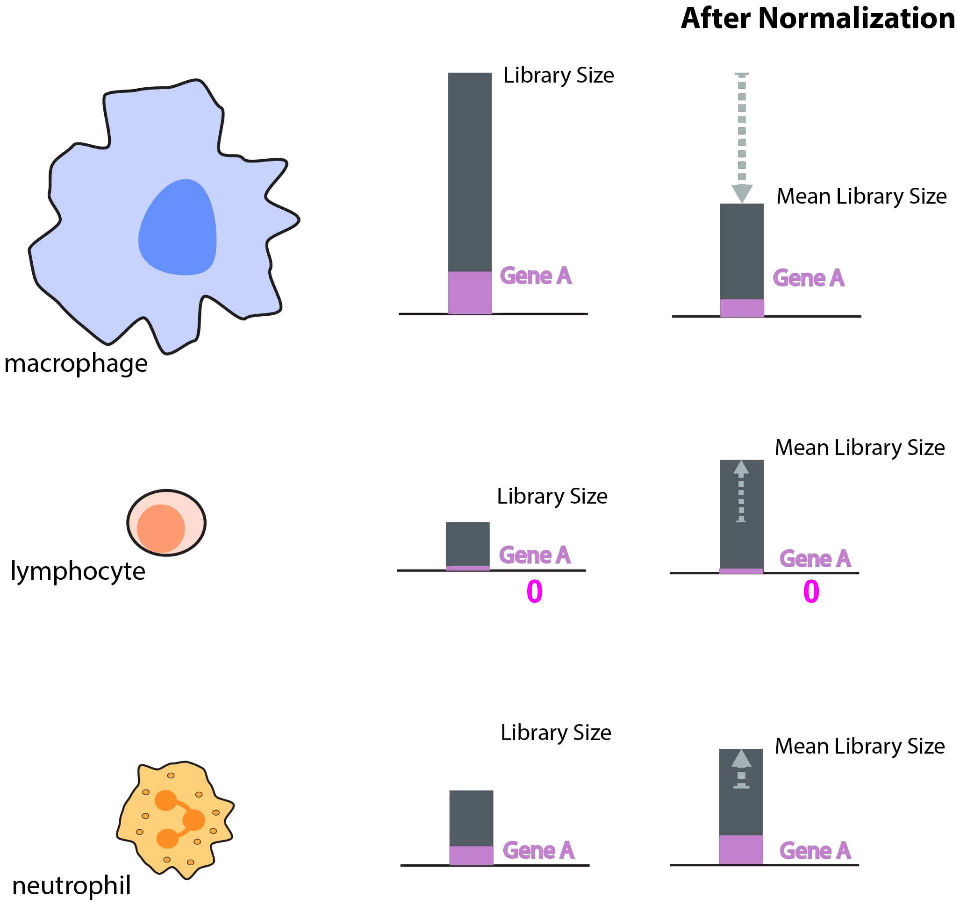


Cells with different sizes have very different total number of transcripts

Example Housekeeping Gene

High chance of Dropouts in smaller cells

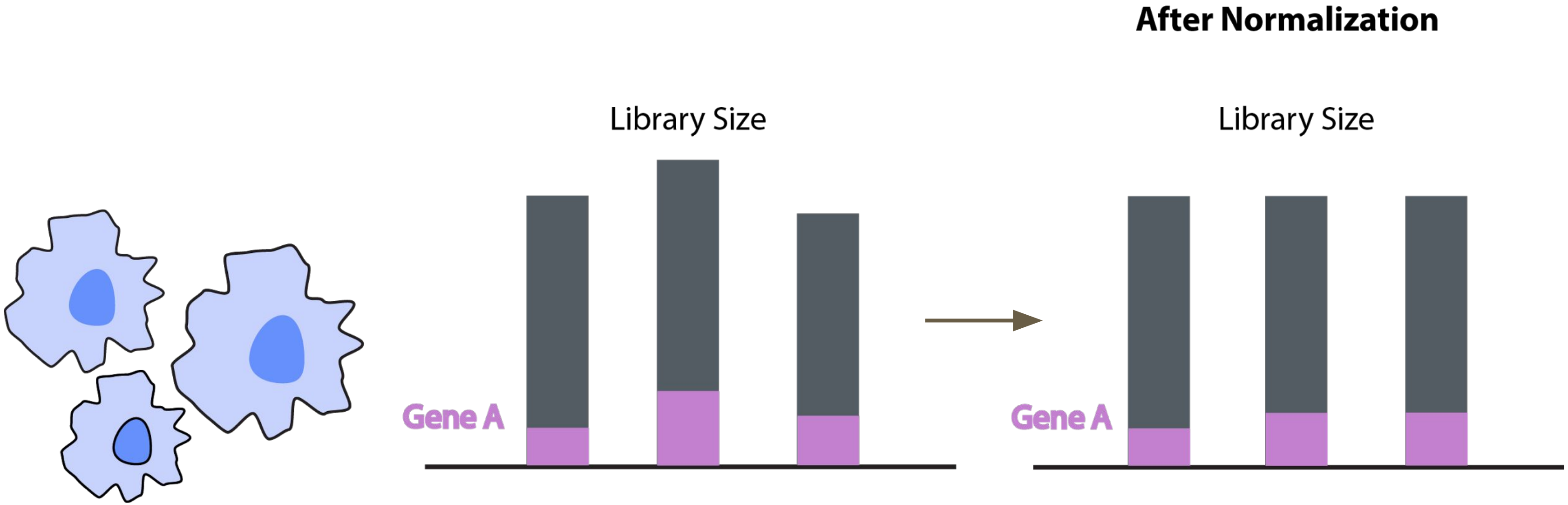
Problem with Global Normalization



Dropout not resolved

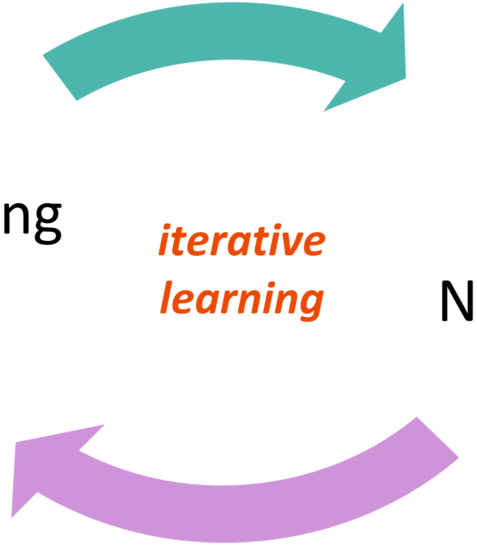
Spurious Differential Expression

Key: Different normalization for each cell type



Chicken and egg problem:
Normalize based on cell types but we
do not know cell types!

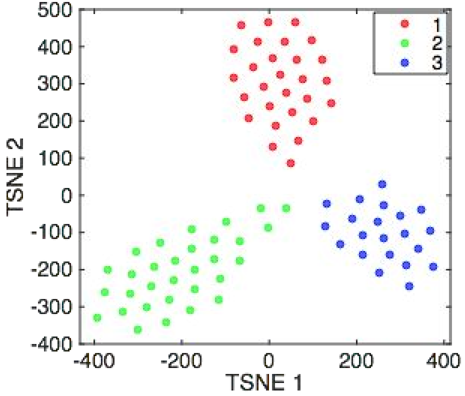
Approach:
Simultaneous inference of clusters and imputing parameters

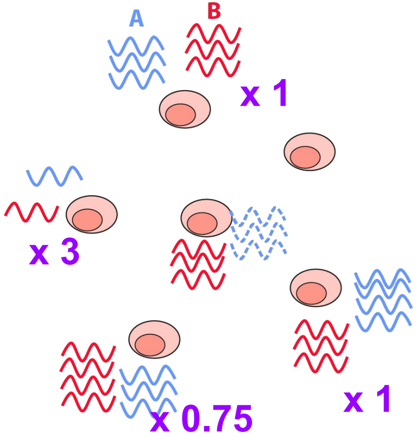


Clustering Cells

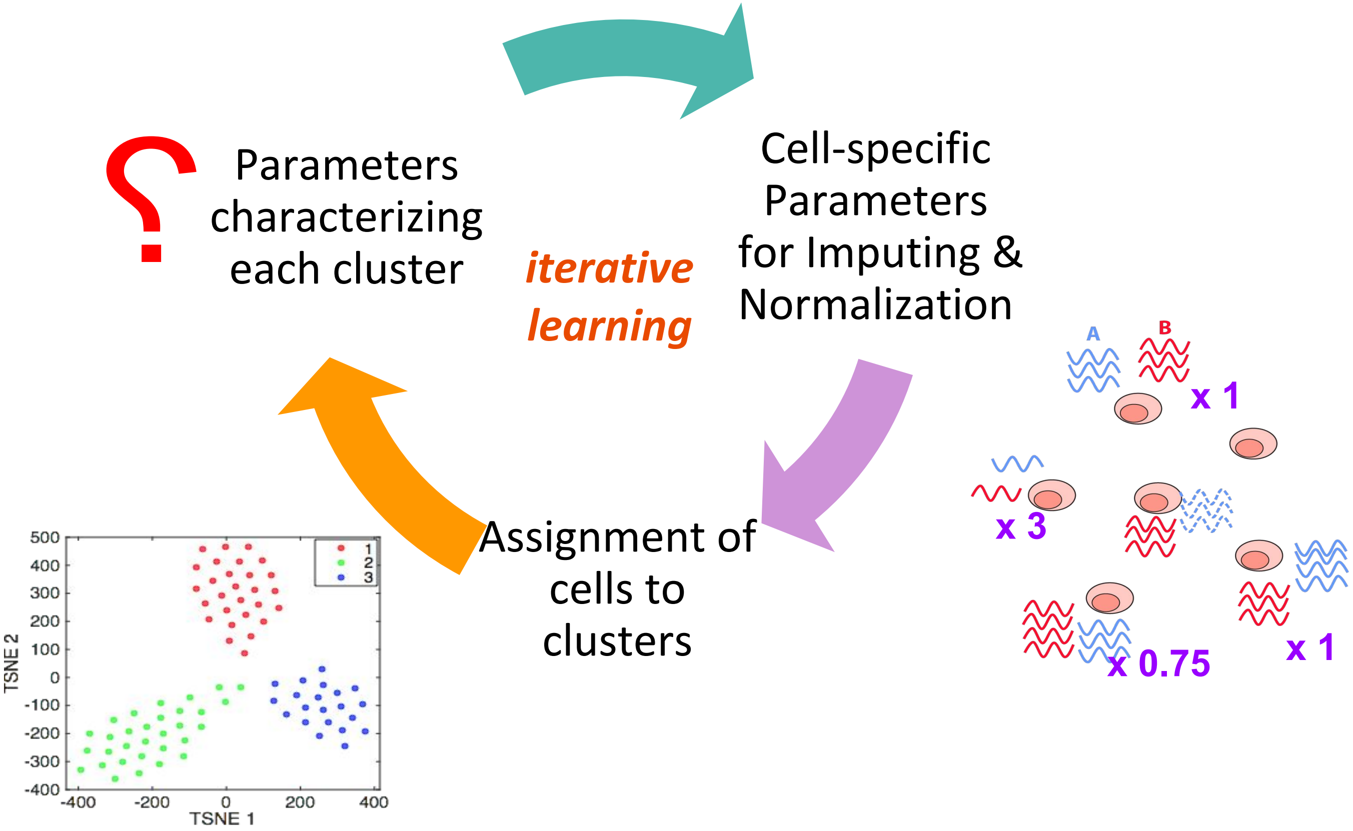
iterative learning

Imputing & Normalization





Approach:
Simultaneous inference of clusters and imputing parameters



Parameters characterizing each cluster

Cell-specific Parameters for Imputing & Normalization

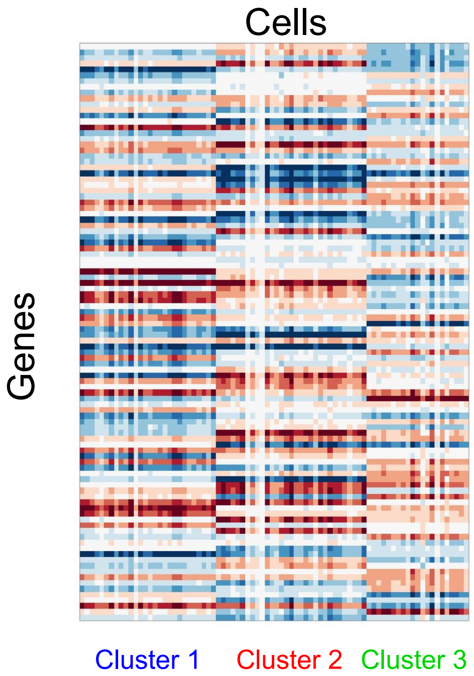
iterative learning

Assignment of cells to clusters

Modeling Single-cell data using a Bayesian Mixture Model

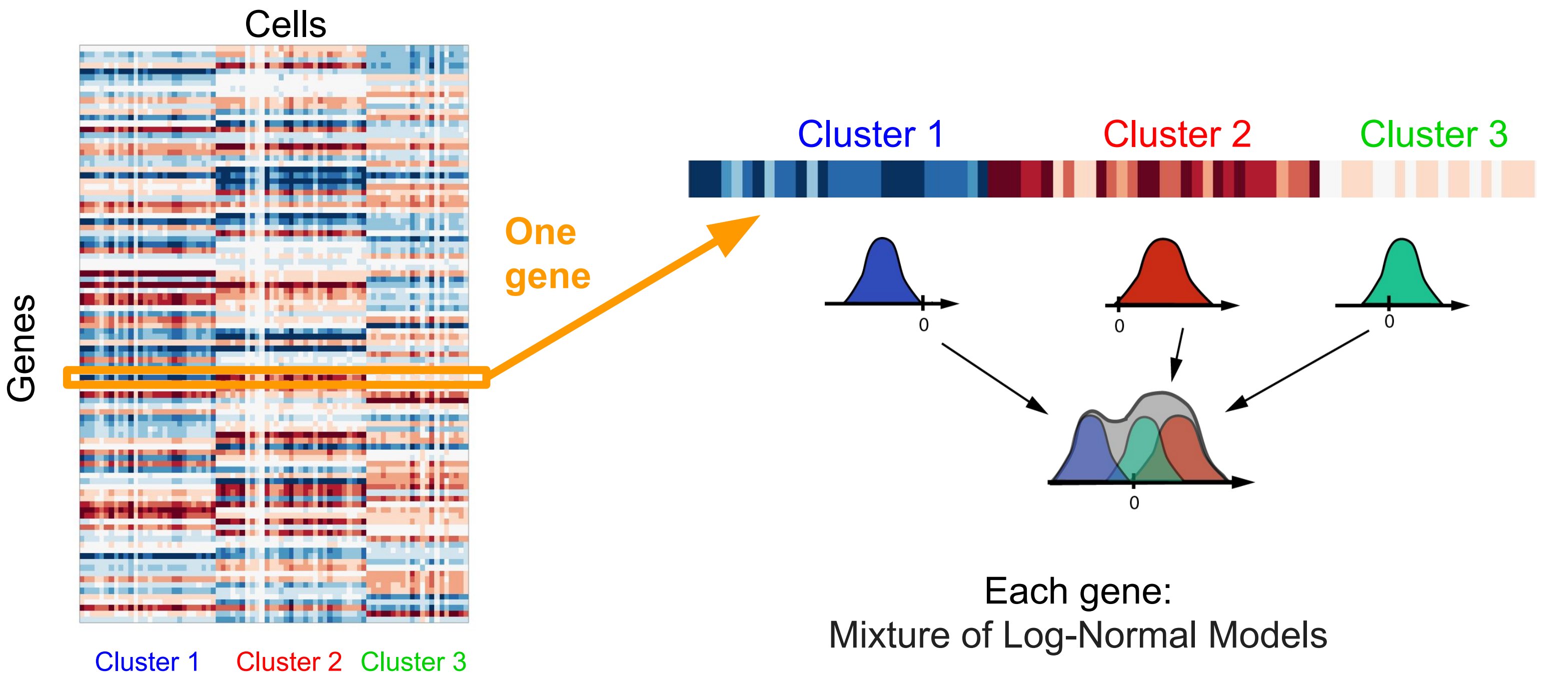
Modeling Clusters of Cells using a Bayesian Mixture Model

Ideal Count Matrix
(normalized)



Modeling Clusters of Cells using a Bayesian Mixture Model

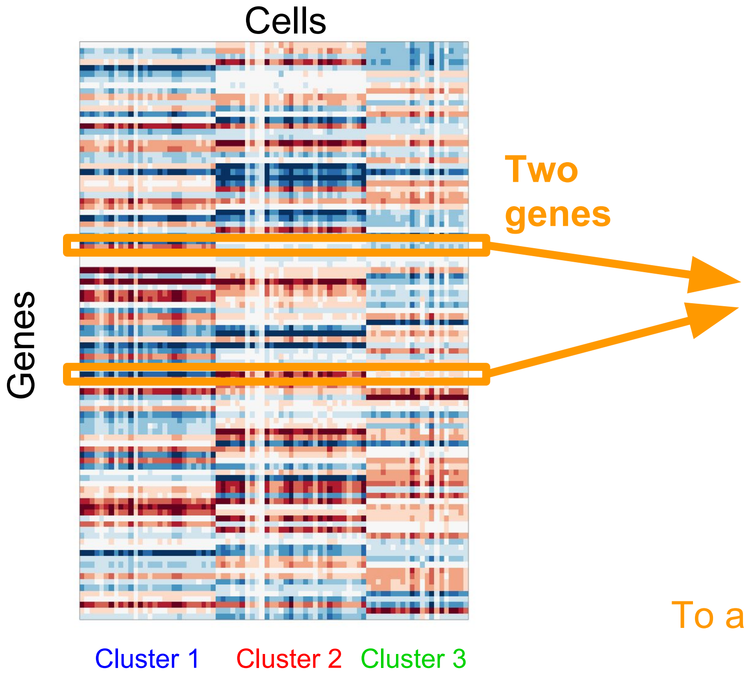
Ideal Count Matrix
(normalized)



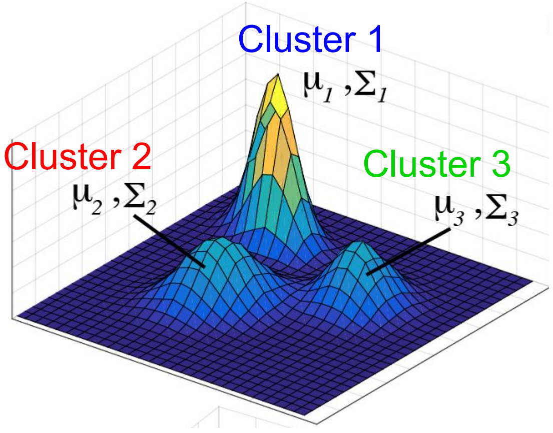
Each gene:
Mixture of Log-Normal Models

Modeling Clusters of Cells using a Bayesian Mixture Model

Ideal Count Matrix
(normalized)



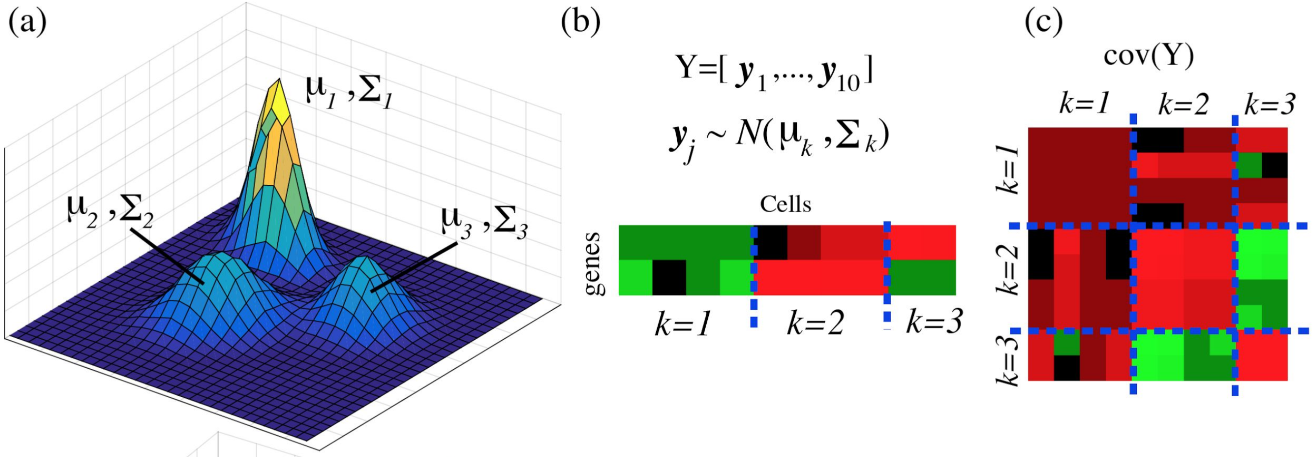
$$\boldsymbol{y}_j \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \qquad z_j = k$$



Modeling all genes together:
Mixture of Multivariate Log-Normals
To also take advantage of co-expression patterns in learning clusters

Generative Model

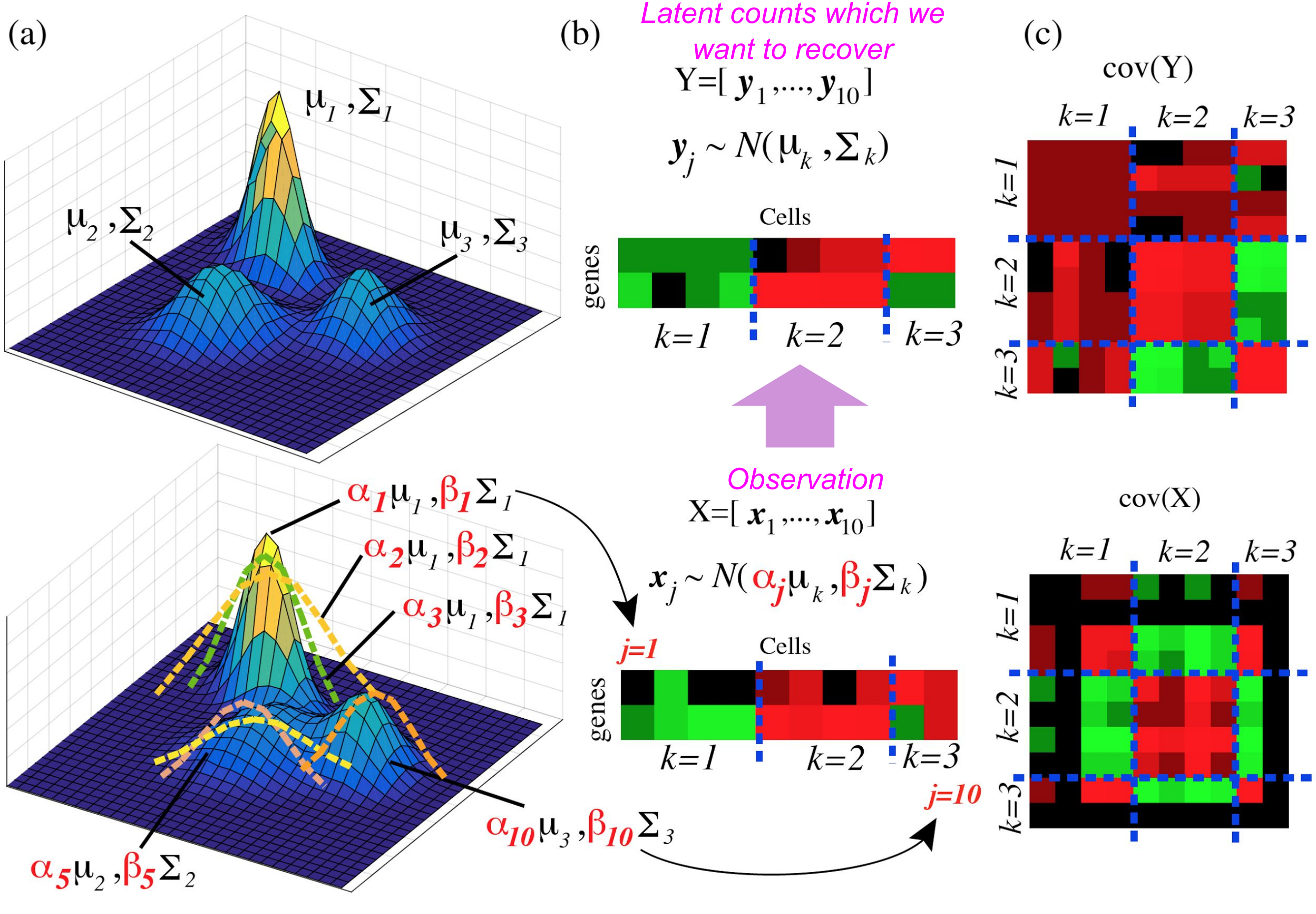
Without Technical Variation



Generative Model with Technical Variation

Without Technical Variation

With Technical Variation



Latent counts which we want to recover

$Y=[\,\boldsymbol{y}_1,\dots,\boldsymbol{y}_{10}]$

$\boldsymbol{y}_j \sim N(\mu_k,\Sigma_k)$

Observation

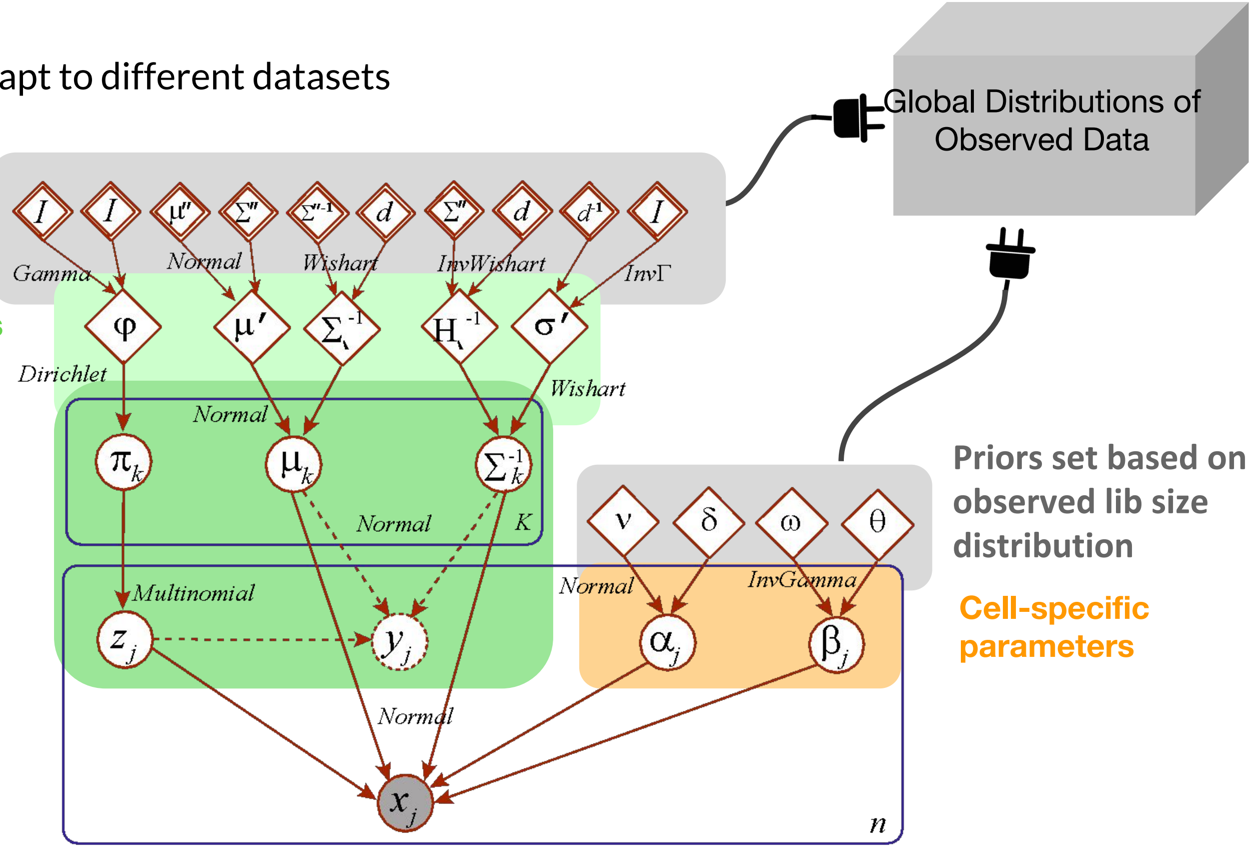
$X=[\,\boldsymbol{x}_1,\dots,\boldsymbol{x}_{10}]$

$\boldsymbol{x}_j \sim N(\alpha_j\mu_k,\beta_j\Sigma_k)$

(a)
(b)
(c)
μ_1,Σ_1
μ_2,Σ_2
μ_3,Σ_3
Cells
genes
k=1
k=2
k=3
cov(Y)
$\alpha_1\mu_1,\beta_1\Sigma_1$
$\alpha_2\mu_1,\beta_2\Sigma_1$
$\alpha_3\mu_1,\beta_3\Sigma_1$
$\alpha_5\mu_2,\beta_5\Sigma_2$
$\alpha_{10}\mu_3,\beta_{10}\Sigma_3$
j=1
j=10
cov(X)

BISCUIT (Bayesian Inference for Single-cell ClUstering and ImpuTing)

Data-driven priors to adapt to different datasets



Global Distributions of Observed Data

Hyper-priors

Hyper-parameters

Cluster-specific parameters

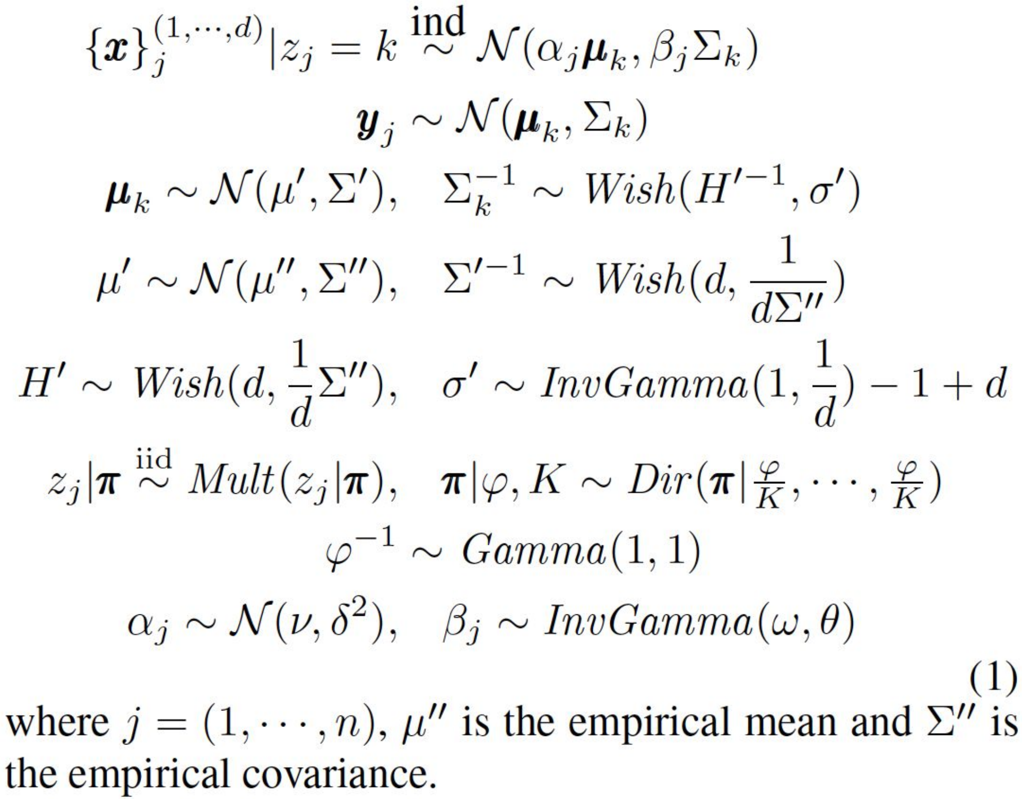
Priors set based on observed lib size distribution

Cell-specific parameters

Observed gene expression per cell *j*

Prabhakaran*, Azizi* *ICML* 2016

Model Specification



$$
\begin{gathered}
\{\boldsymbol{x}\}_j^{(1,\cdots,d)} | z_j = k \overset{\text{ind}}{\sim} \mathcal{N}(\alpha_j \boldsymbol{\mu}_k, \beta_j \Sigma_k) \\
\boldsymbol{y}_j \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \\
\boldsymbol{\mu}_k \sim \mathcal{N}(\mu', \Sigma'), \quad \Sigma_k^{-1} \sim Wish(H'^{-1}, \sigma') \\
\mu' \sim \mathcal{N}(\mu'', \Sigma''), \quad \Sigma'^{-1} \sim Wish(d, \frac{1}{d\Sigma''}) \\
H' \sim Wish(d, \frac{1}{d}\Sigma''), \quad \sigma' \sim InvGamma(1, \frac{1}{d}) - 1 + d \\
z_j | \boldsymbol{\pi} \overset{\text{iid}}{\sim} Mult(z_j | \boldsymbol{\pi}), \quad \boldsymbol{\pi} | \varphi, K \sim Dir(\boldsymbol{\pi} | \tfrac{\varphi}{K}, \cdots, \tfrac{\varphi}{K}) \\
\varphi^{-1} \sim Gamma(1, 1) \\
\alpha_j \sim \mathcal{N}(\nu, \delta^2), \quad \beta_j \sim InvGamma(\omega, \theta)
\end{gathered}
\tag{1}
$$

where $j = (1, \cdots, n)$, μ'' is the empirical mean and Σ'' is the empirical covariance.

Prabhakaran*, Azizi* *ICML* 2016

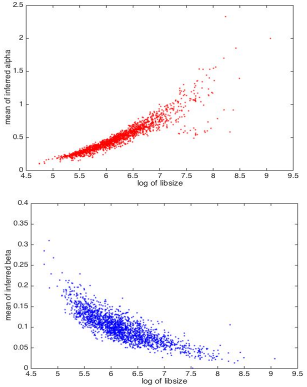
Inference Algorithm

Parallel Sampling from derived conditional posterior distributions:
P(parameter| data, other parameters)

Estimate hyper-priors based on Data

Sample hyper-parameters

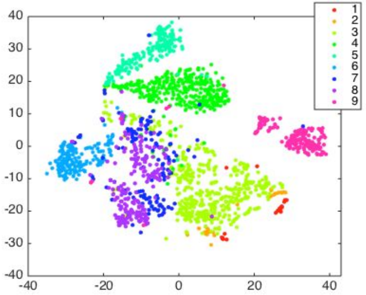
Sampling technica variation parameters



scaling mean, cov per cell

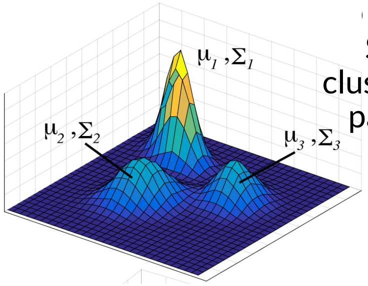
Gibbs iterations

Sampling assignment of cells to clusters using Chinese Restaurant Process (CRP)



Also allows estimating the number of clusters

Sampling cluster-specific parameters

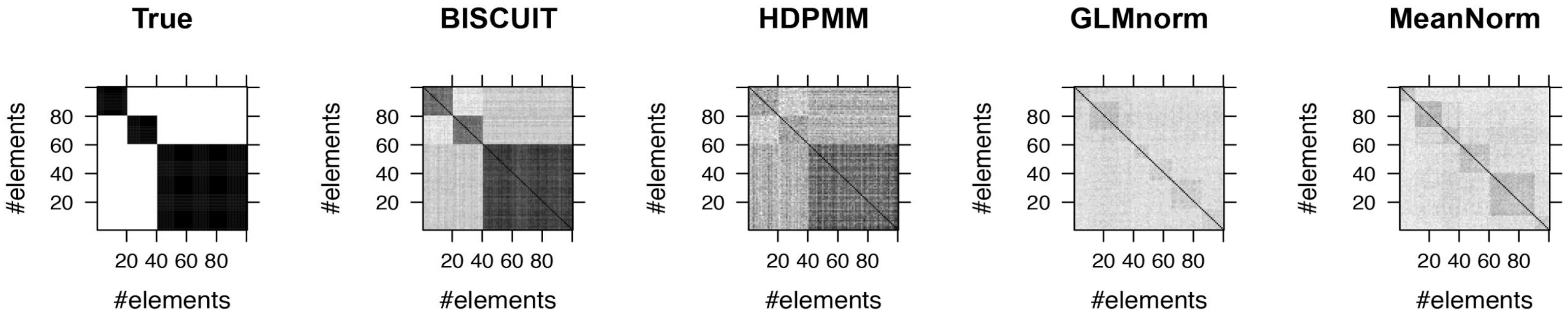


Prabhakaran*, Azizi* *ICML* 2016

Performance on Simulated Data

Data simulated from model for 100 cells, 50 genes in 3 clusters

Confusion matrices showing true labels and those from MCMC-based methods



True

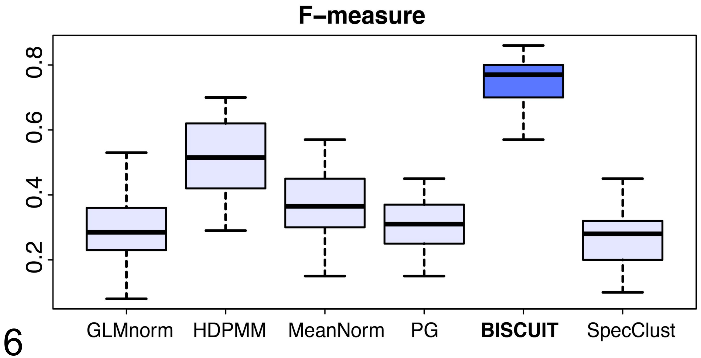
BISCUIT

HDPMM

GLMnorm

MeanNorm

Boxplots of F-scores obtained in 15 experiments with randomly-generated data

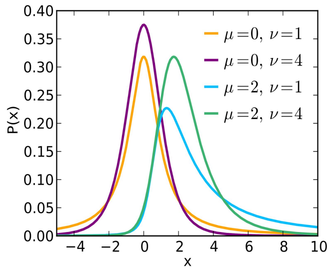


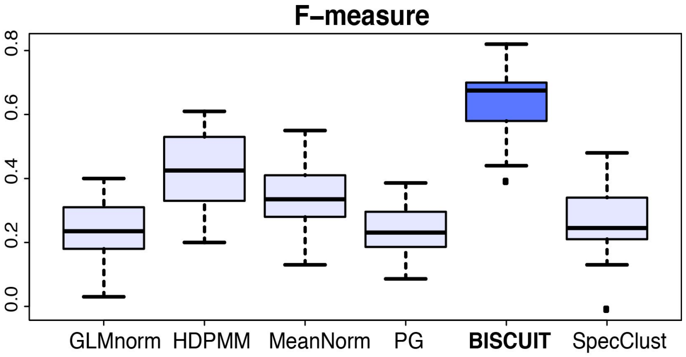
F–measure

Prabhakaran*, Azizi* *ICML* 2016

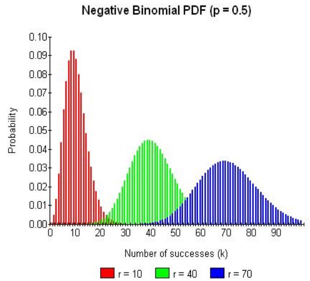
Model Mismatch:
Robustness when counts are not LogNormal

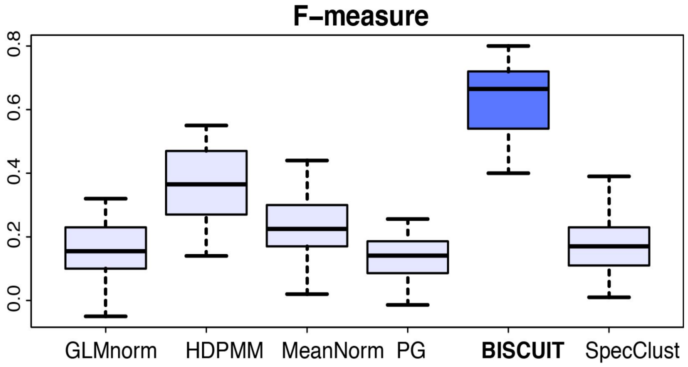
Noncentral Student's t





(Log) Negative binomial





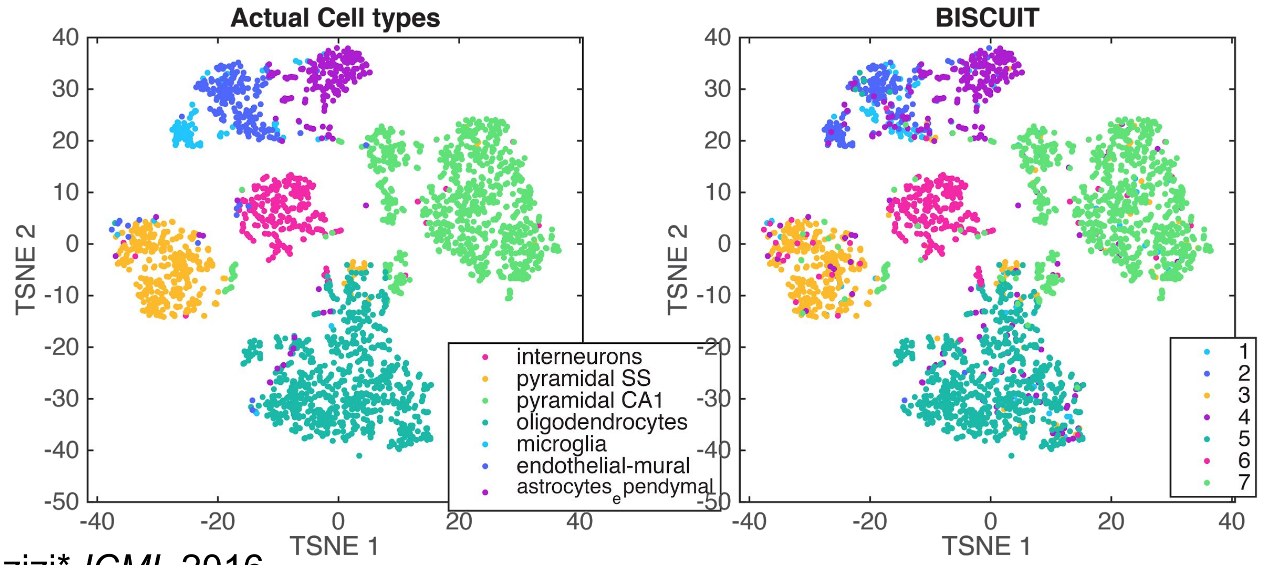
Prabhakaran*, Azizi* *ICML* 2016

Performance on Single-cell Data Zeisel et al., 2015

- 3005 mouse cortex cells, with UMIs
- Deep coverage gives good ground truth for **7 Cell types**
- Selected 558 genes with largest standard deviation across cells

- Fit model to *log(counts+1)*

F-score: 0.91

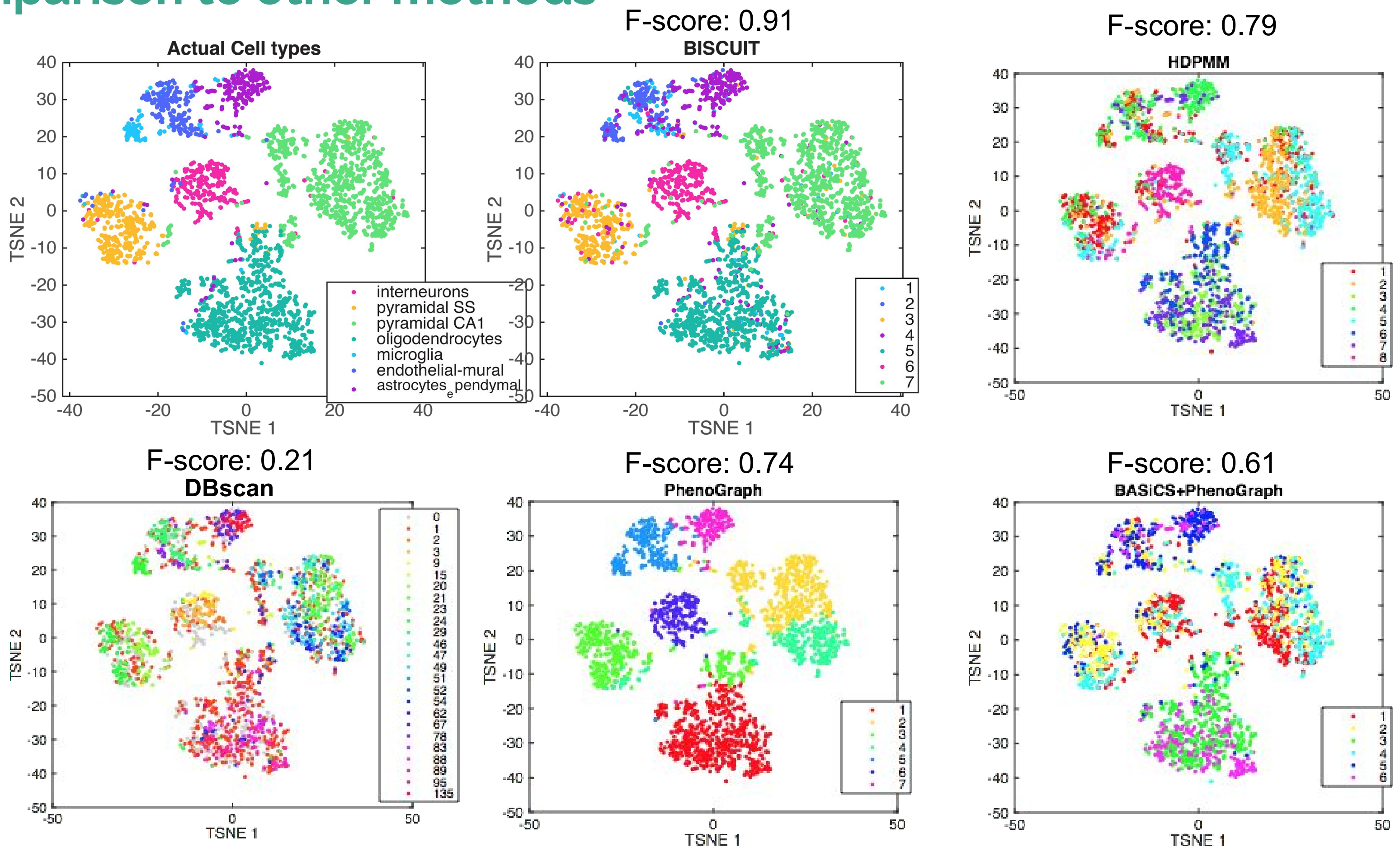


Actual Cell types

BISCUIT

Prabhakaran*, Azizi* *ICML* 2016

Comparison to other methods



F-score: 0.91

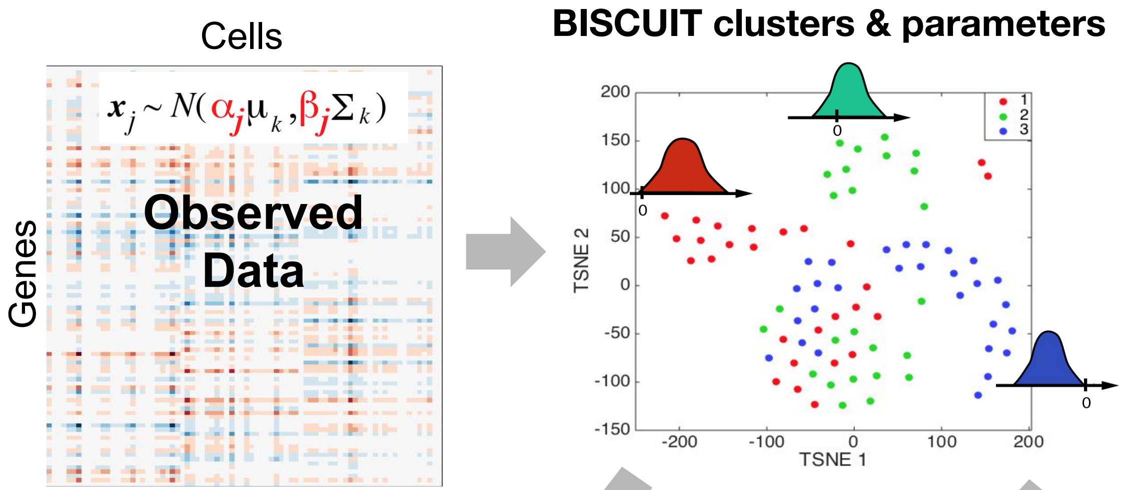
F-score: 0.79

F-score: 0.21

F-score: 0.74

F-score: 0.61

Cluster-dependent Imputing & Normalizing



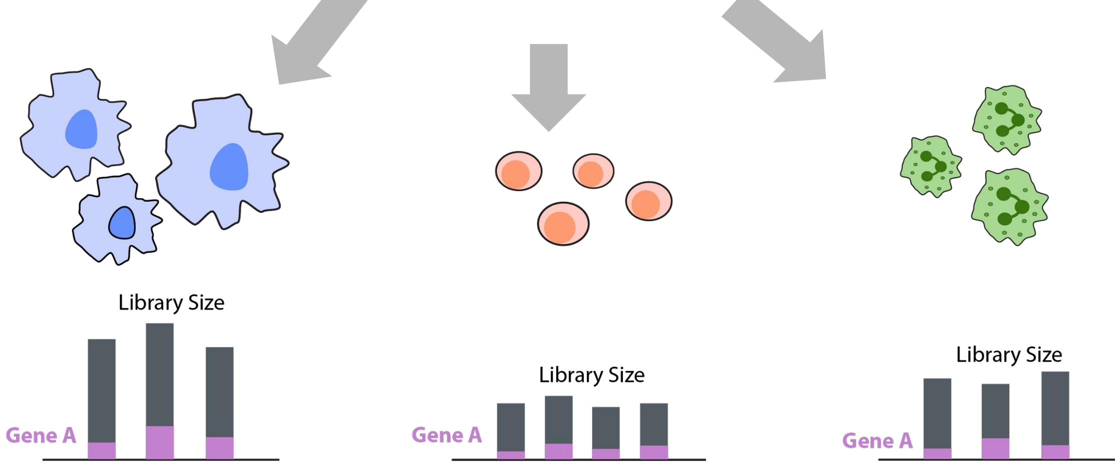
$$y_j \sim N(\mu_k, \Sigma_k)$$

Impute & Normalize
With a linear transformation

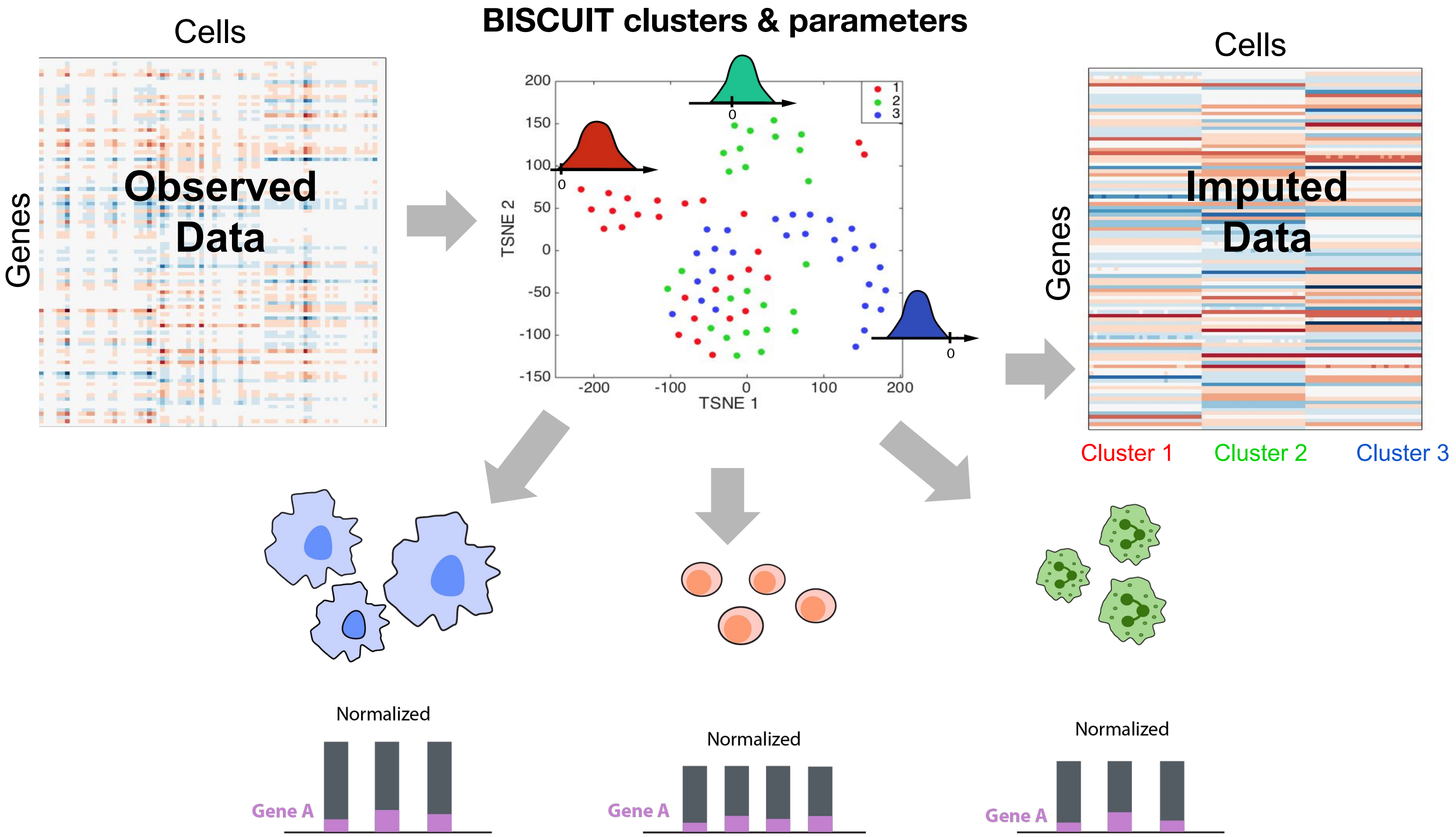
$$\boldsymbol{y}_j = A\boldsymbol{x}_j + b$$

$$A = \frac{1}{\beta} I$$

$$b = (I - \alpha A)\boldsymbol{\mu}_k$$

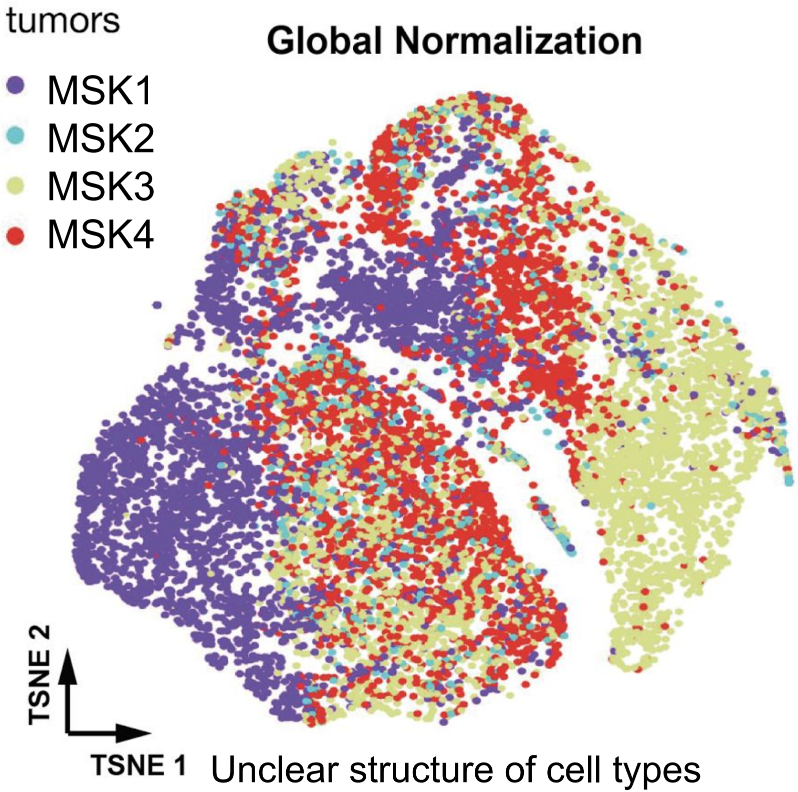


Cluster-dependent Imputing & Normalizing



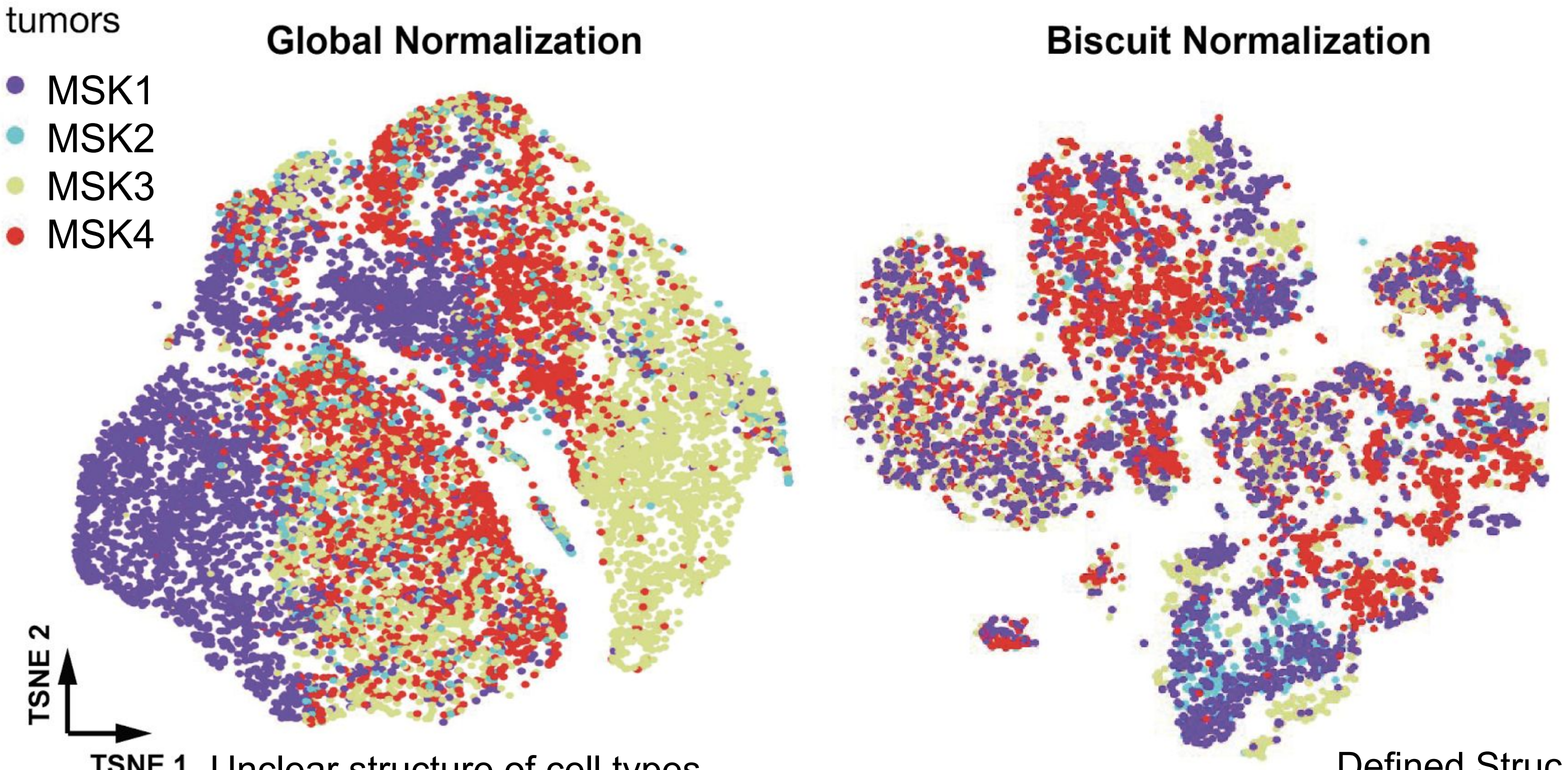
Characterizing tumor-infiltrating immune cells in breast cancer

Single-cell RNA-seq data for immune cells from 4 breast cancer tumors



Unclear structure of cell types
Large patient biases

Single-cell RNA-seq data for immune cells from 4 breast cancer tumors



Unclear structure of cell types
Large patient biases

Defined Structure
Corrected patient biases

Impact of environment in immune cell heterogeneity

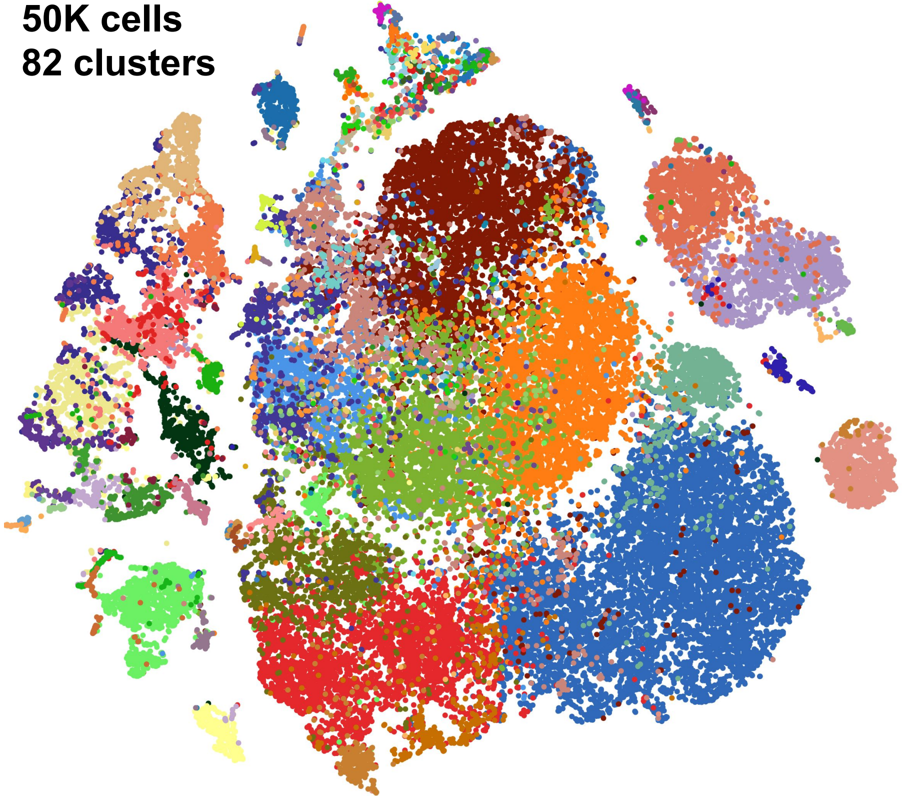
- Little known about how tissue microenvironment modulates anti-tumor immunity

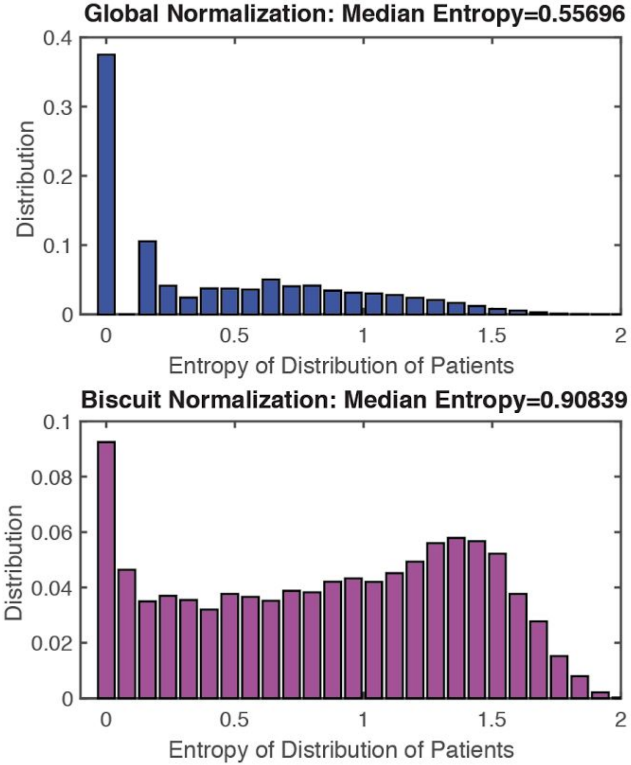
- Collected 50K CD45+ leukocytes from 8 patients
  - Different ranges of tumor grade, ER, PR, Her2, age, two cases of metastases
  - **Different environments (tissues):**
    - Tumor
    - Peripheral blood
    - Lymphnode
    - Normal (prophylactic mastectomies)

- Largest single cell immune map based on tissue residence.

Map of immune cells from 8 breast cancer patients Normalized by BISCUIT

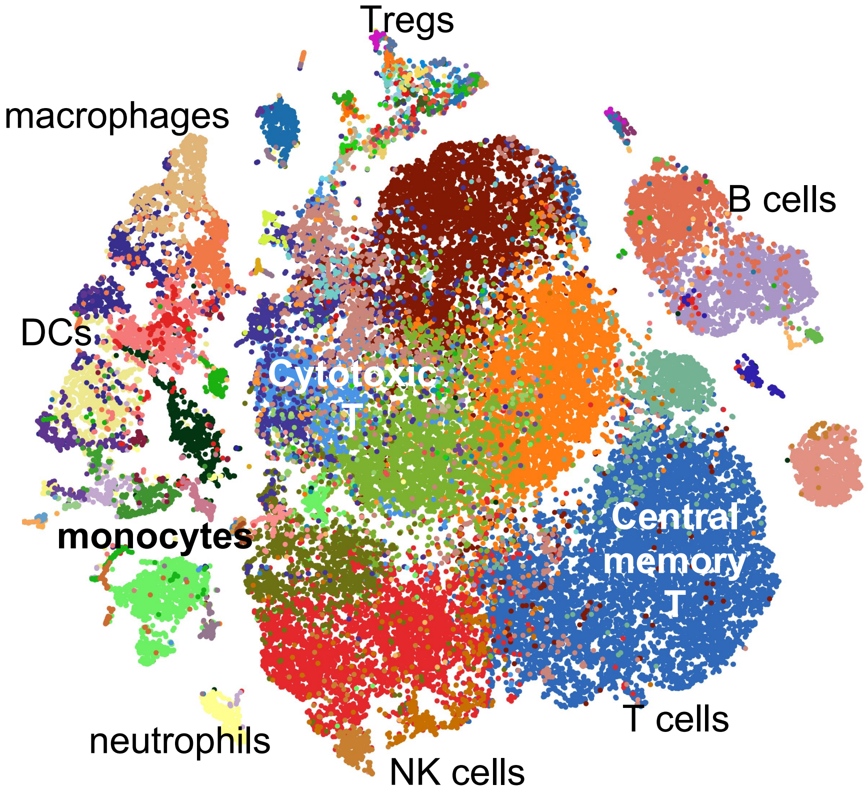
50K cells
82 clusters

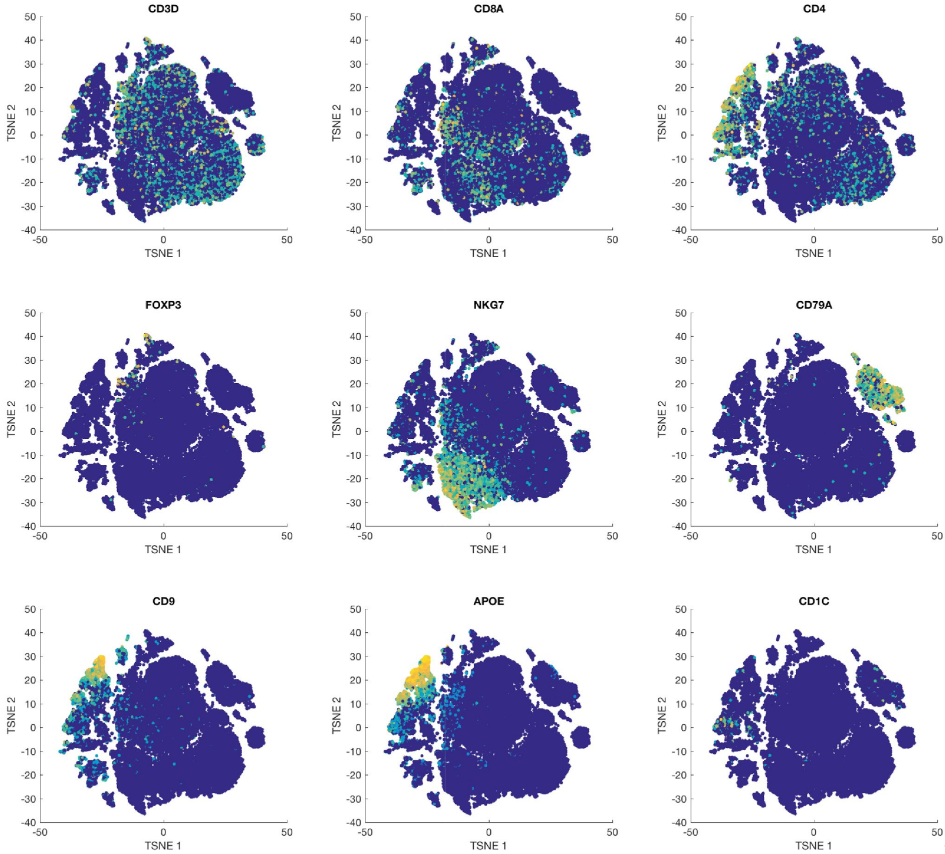




Higher entropy shows more mixing of patients with Biscuit

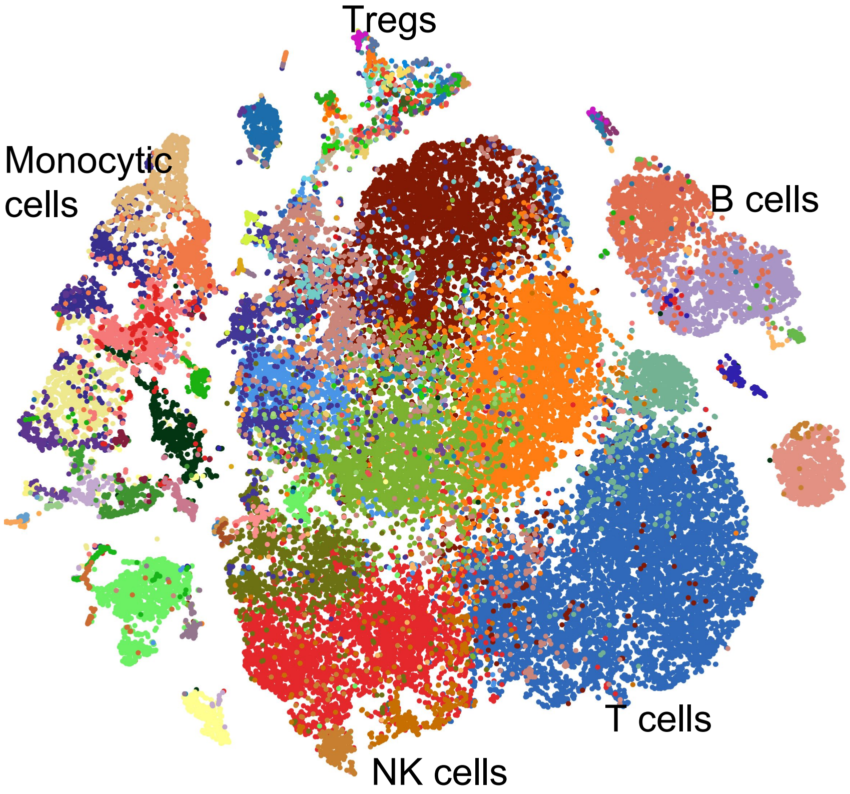
Map of immune cells from 8 breast cancer patients Normalized by BISCUIT

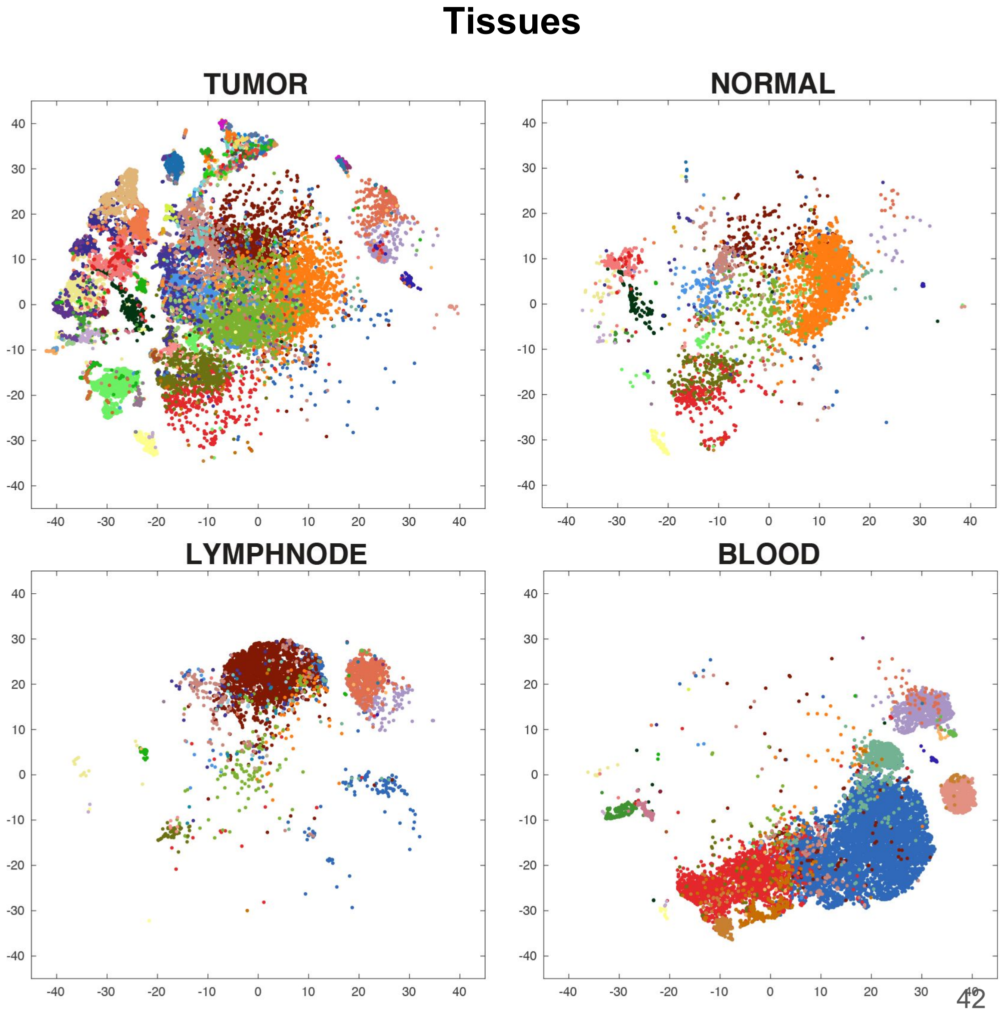


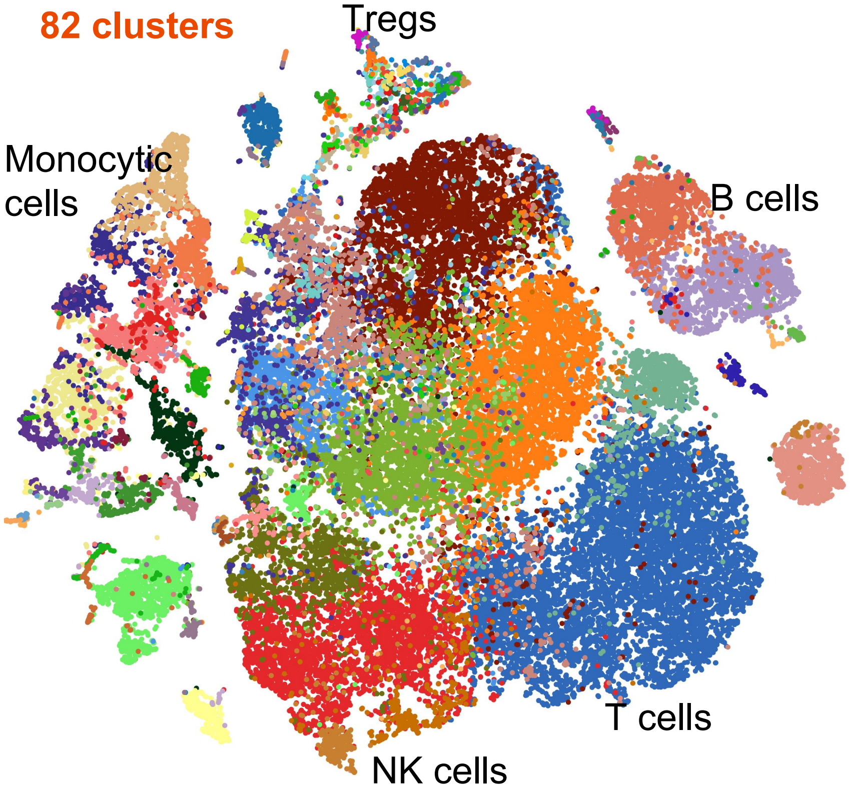


82 clusters

Impact of Environments



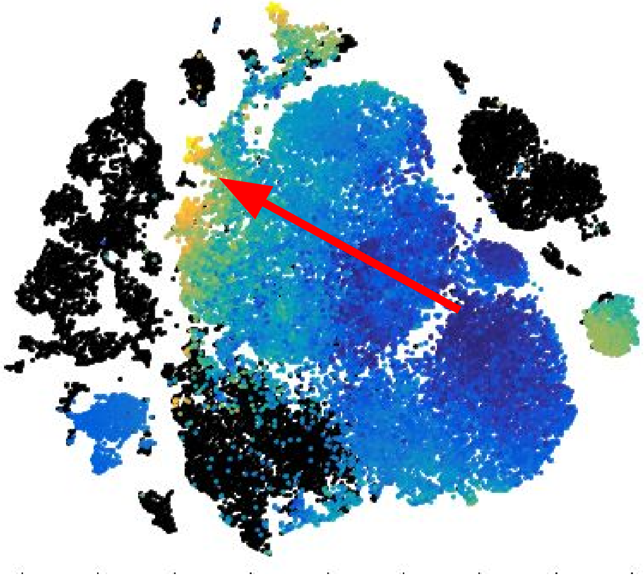


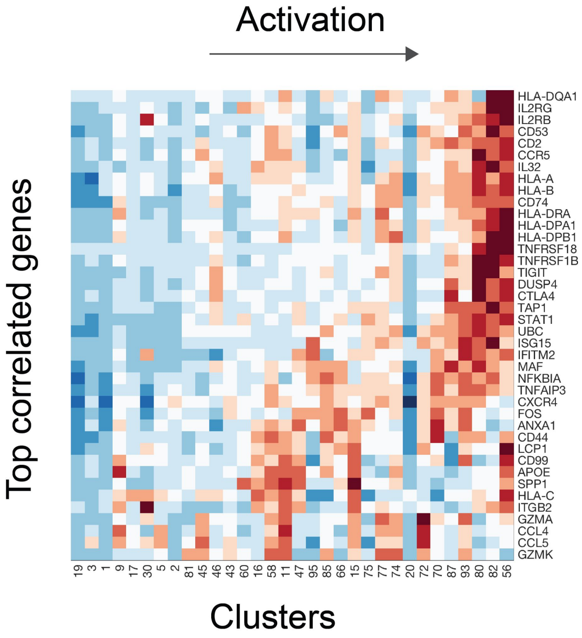


What are the differences across inferred cell subpopulations?

T cell activation: First component of variation

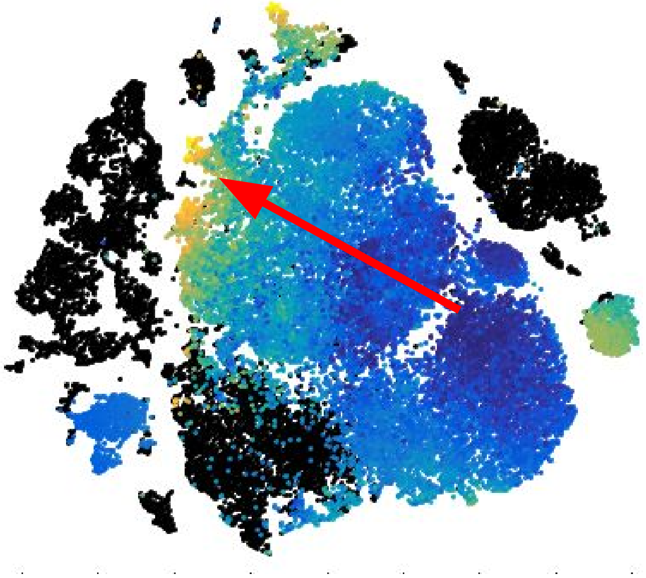
- Correlated genes enriched for cytokine production & signaling, lymphocyte activation, leukocyte differentiation, ligand receptor interaction

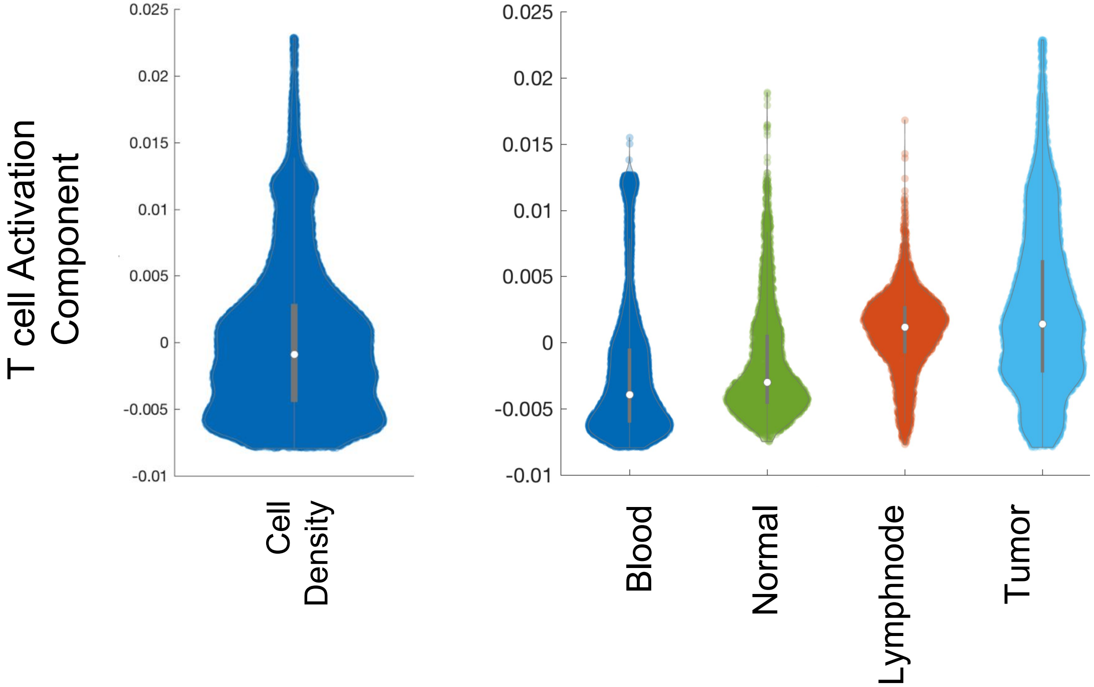




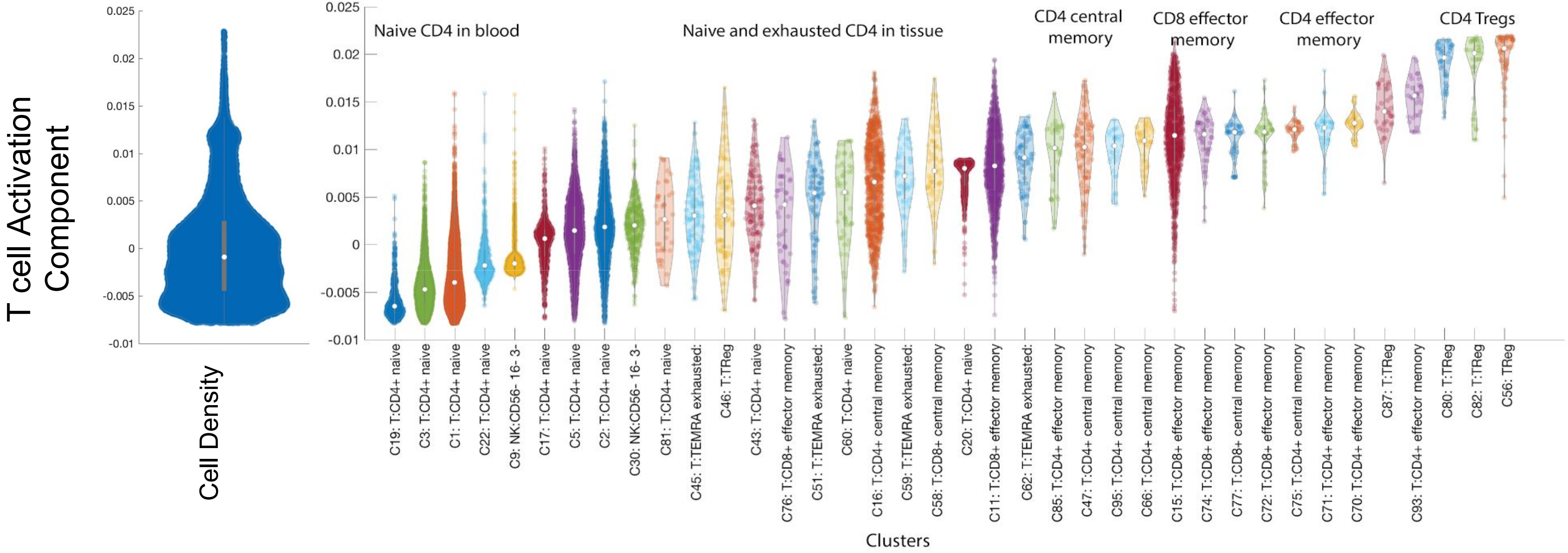
T cell activation: First component of variation

- Comparing distribution of cells along the activation component shows tumor is more activated.

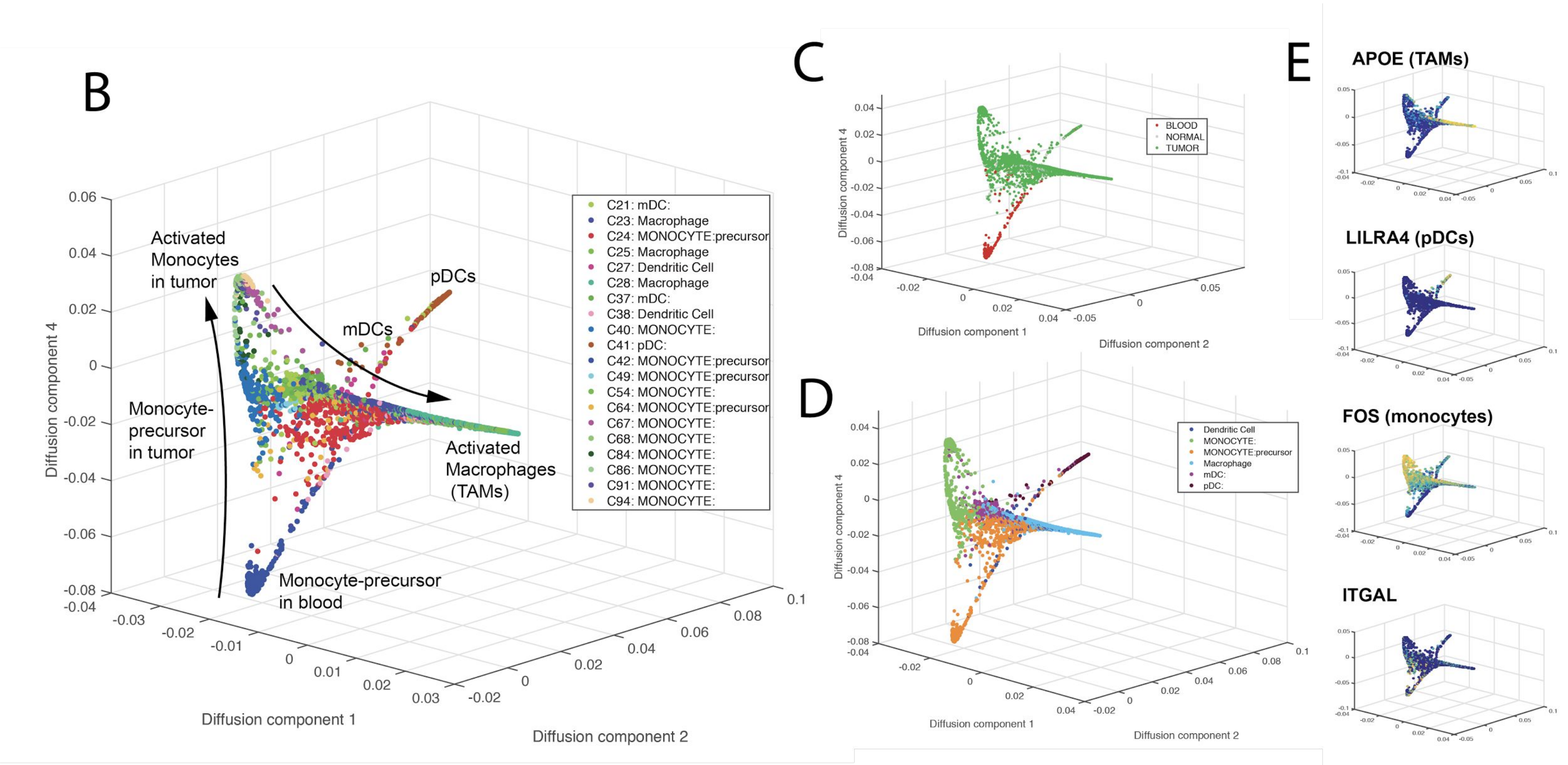




Activation state of each cluster

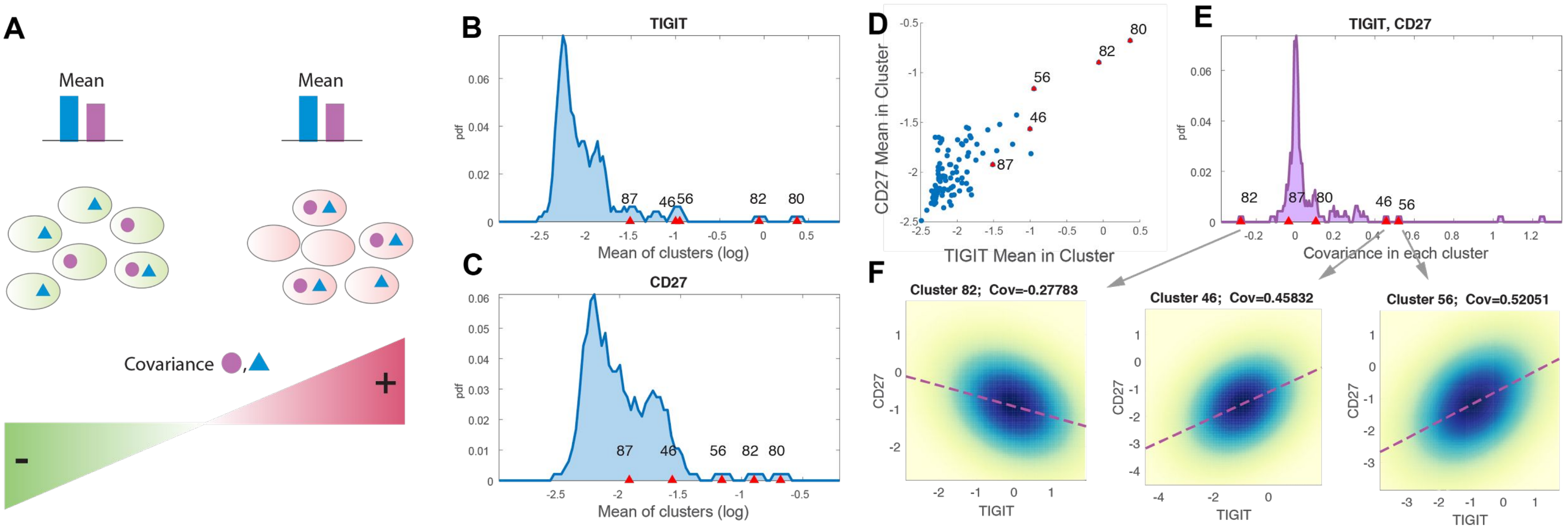


Activation of monocytic cells: first components of variation



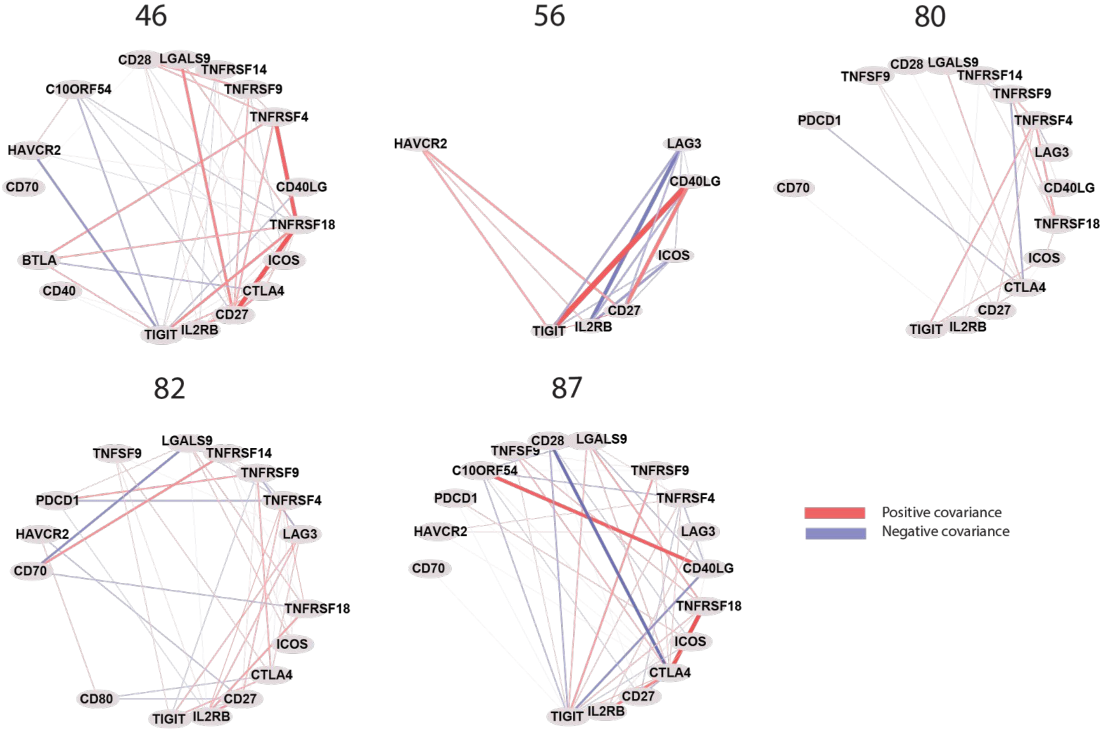
Covariance patterns identify Treg clusters

Markers differentially expressed in mean but differ in covariance patterns

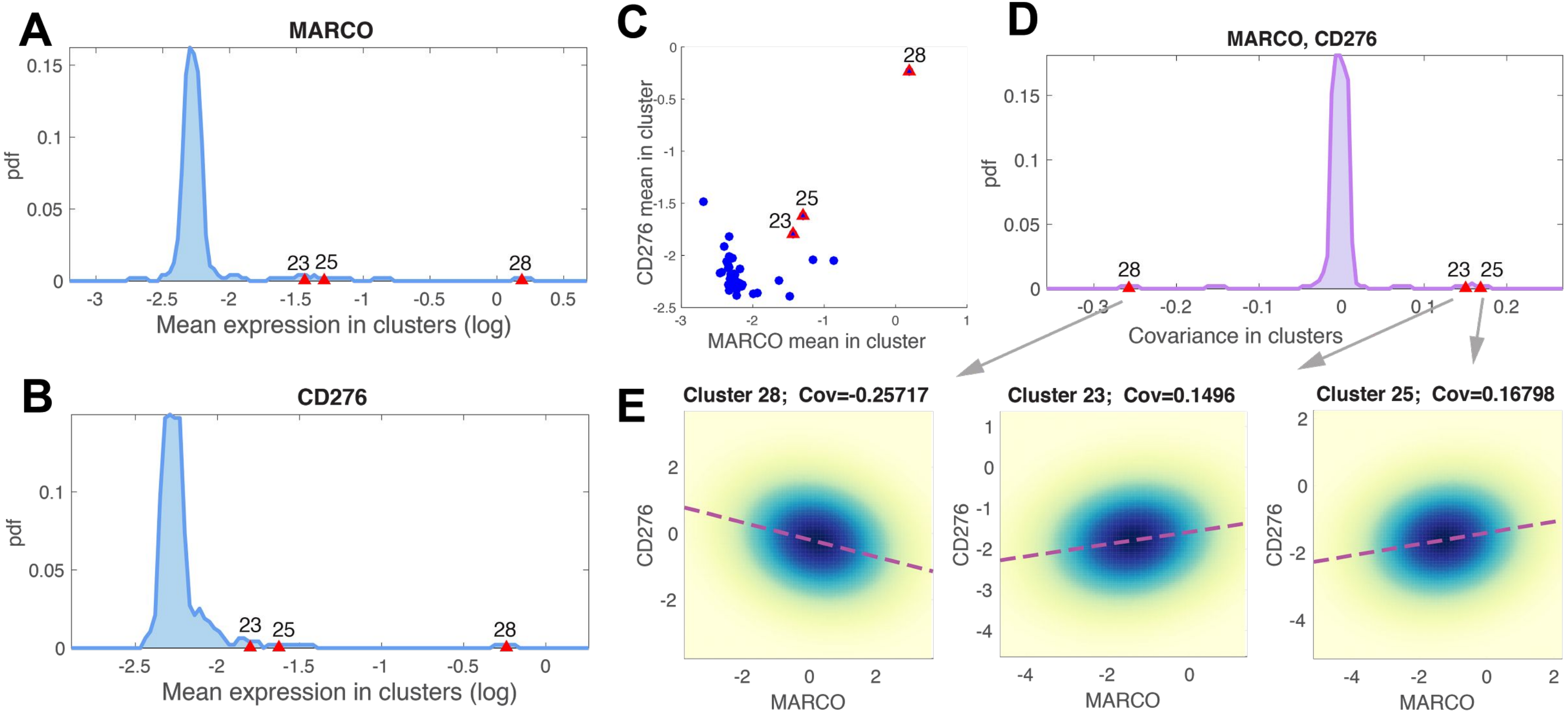


Different covariance patterns of immunotherapy targets in Tregs across patients

- Incorporating personalized co-expression of drug targets can broaden the scope of immunotherapy



Macrophage clusters differ in covariance between M1/M2 markers



Summary

- Analyzing single cell data involves computational challenges: dropouts, technical variation dependent on cell types
- BISCUIT:
  - A bayesian approach for simultaneous clustering and imputing
  - Clusters identified with both mean and gene-gene covariance patterns
  - Incorporating covariance informations improves normalization and imputing
- Map of tumor-immune ecosystem in breast cancer
  - Single cell data for 50K CD45+ cells from 8 patients analyzed with Biscuit
  - Substantial diversity of immune cell types driven by environment
  - Activation of T cells and monocytic cell types explain most of variation
  - Covariance patterns can be informative in characterization of cell types and development of personalized treatments

Acknowledgments

Dana Pe’er
Alexander Rudensky

Barbara Engelhardt (Princeton)

David Blei (Columbia)

Pe’er Lab
Sandhya Prabhakaran
Ambrose Carr
Andrew Cornish
Linas Mazutis
Vaidotas Kiseliovas
Juozas Nainys

Rudensky Lab
George Plitas
Kasia Konopacki

R code:

https://github.com/sandhya212/BISCUIT_SingleCell_IMM_ICML_2016

Email elham.azizi@gmail.com