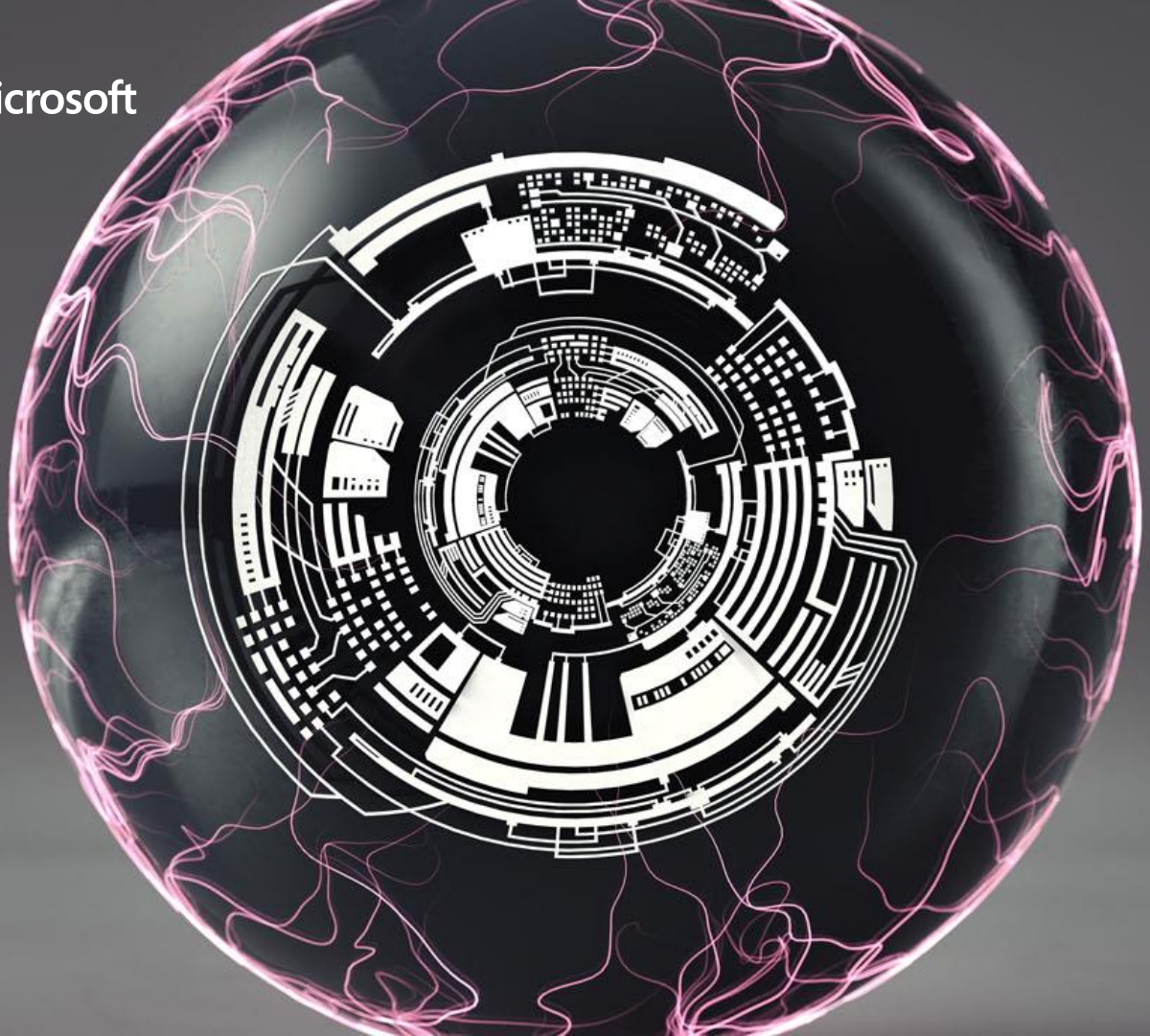
Microsoft



Microsoft Research

Faculty Summit

2016

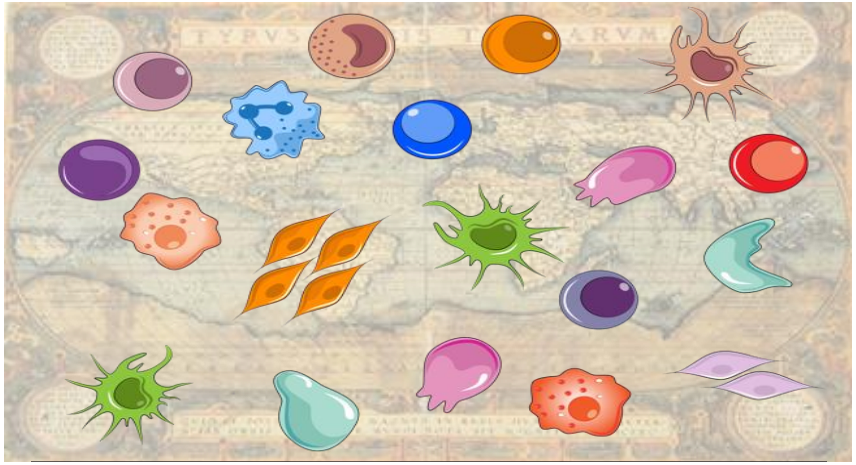
The power of single cells: Building a tumor immune atlas

Dana Pe'er

Department of Biological Science

Department of Systems Biology

Columbia University

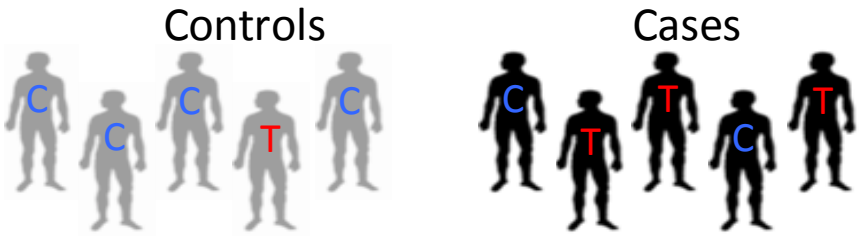


The Precision Medicine Initiative

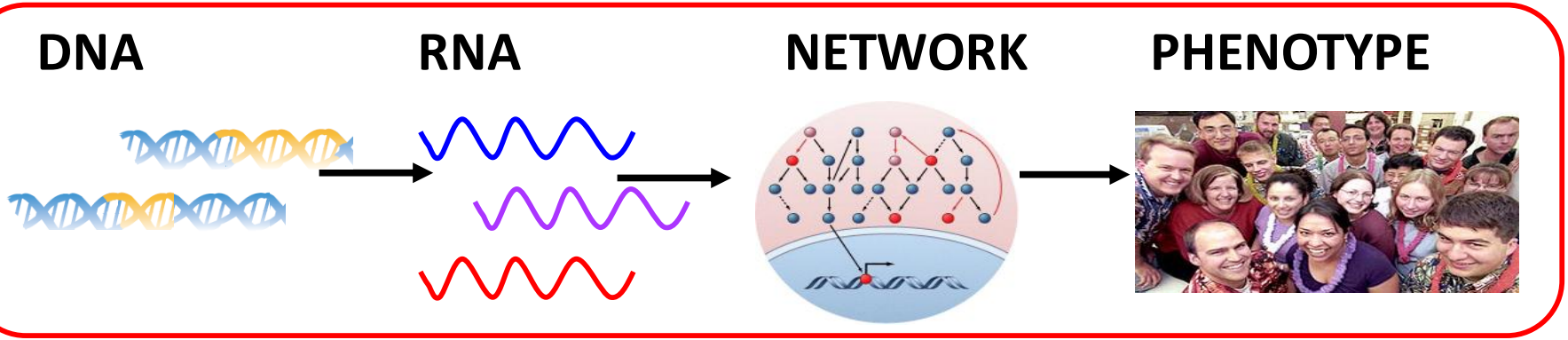
Most “personalized medicine” efforts are focused on DNA, but DNA will not “cut it”

“Doctors have always recognized that every patient is unique. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?” President Obama, 2015

Precision Medicine?

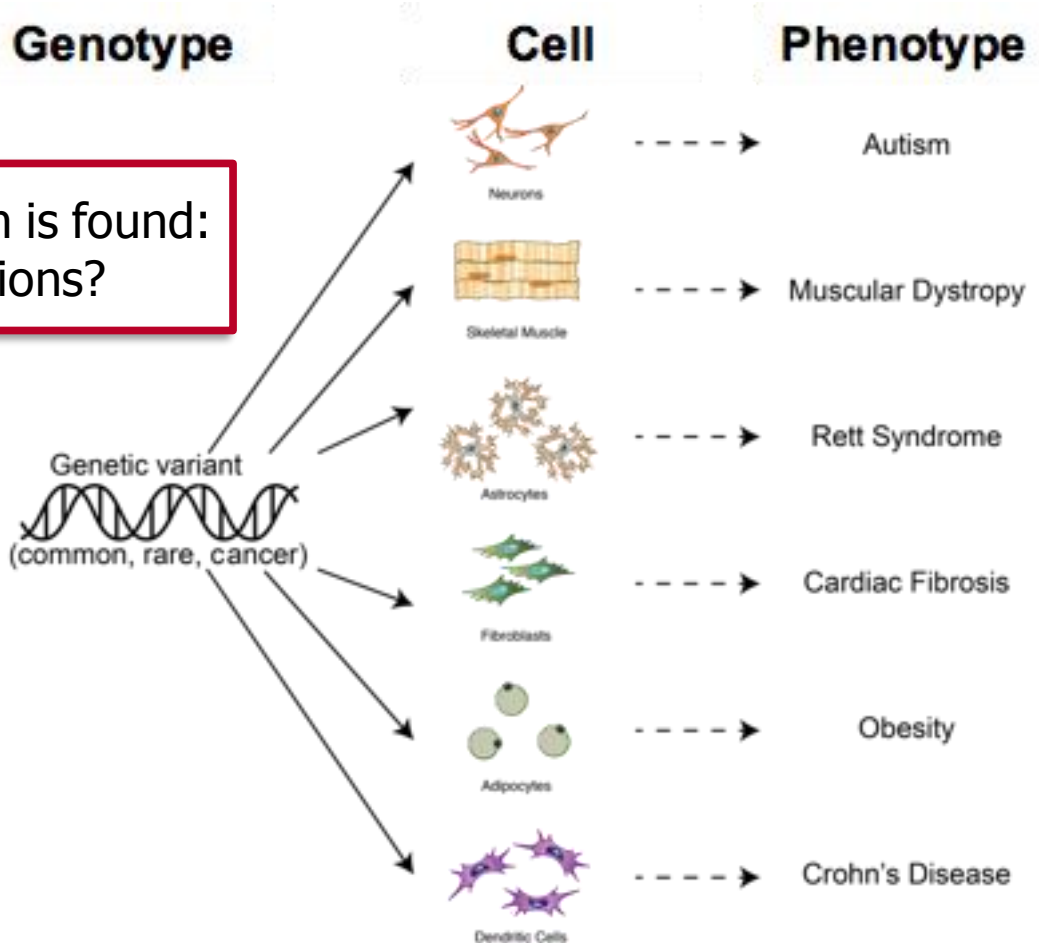


- 90% of the loci that associate with human traits and diseases are outside genes
- Recent evidence supports that these fall in regions that regulate gene expression



Cells: Key intermediates from genotype to phenotype

When a genetic association is found:
Which tissue dysfunctions?



One Genome – Many Cell Types

3 billion letters

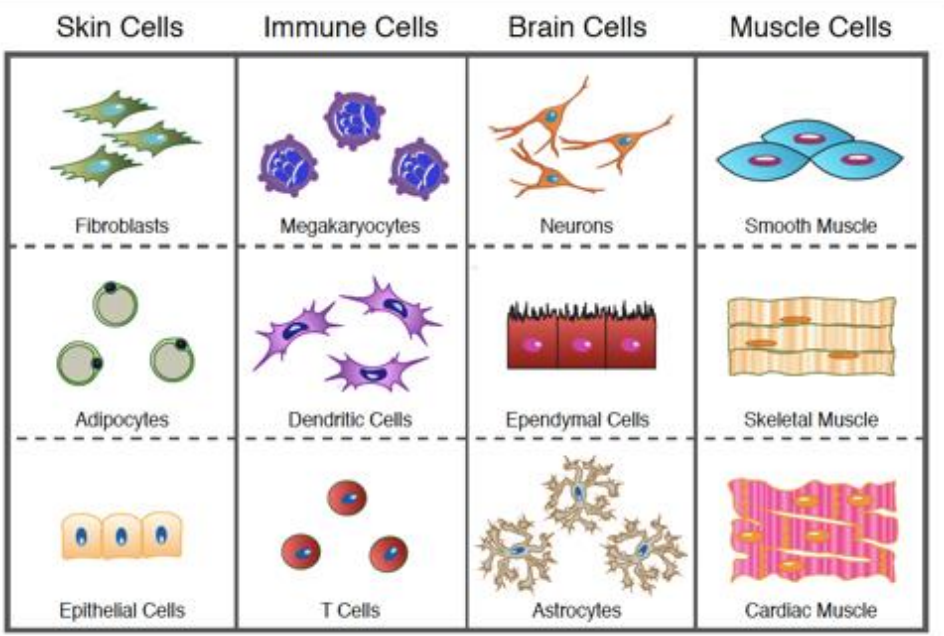
ACCAGTTACGACGGTCA

TTTGAGTTGGTTTTTTC

ACGGTAGAACGTACCGT

TACCAGTA

Gene expression is central to understanding genetics and disease

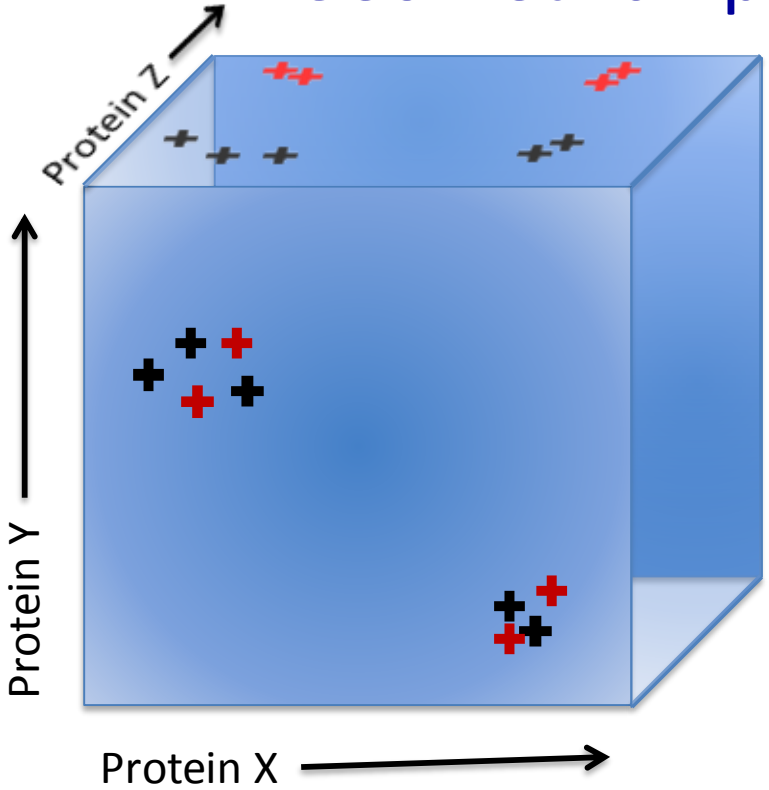


- Expressed genes differ between cell types
- The regulatory region of a gene differs between cell types!
- Tissues contain many different cell types

A cell atlas will be as empowering as the human genome map.

- Our genes are well mapped, but most of cell types remain unknown
- Cells are basic biological units
- Diseases are caused by malfunction of specific cell types.
- Goal: Construct a comprehensive map of all cell types in our body

A Geometric Approach to Phenotype

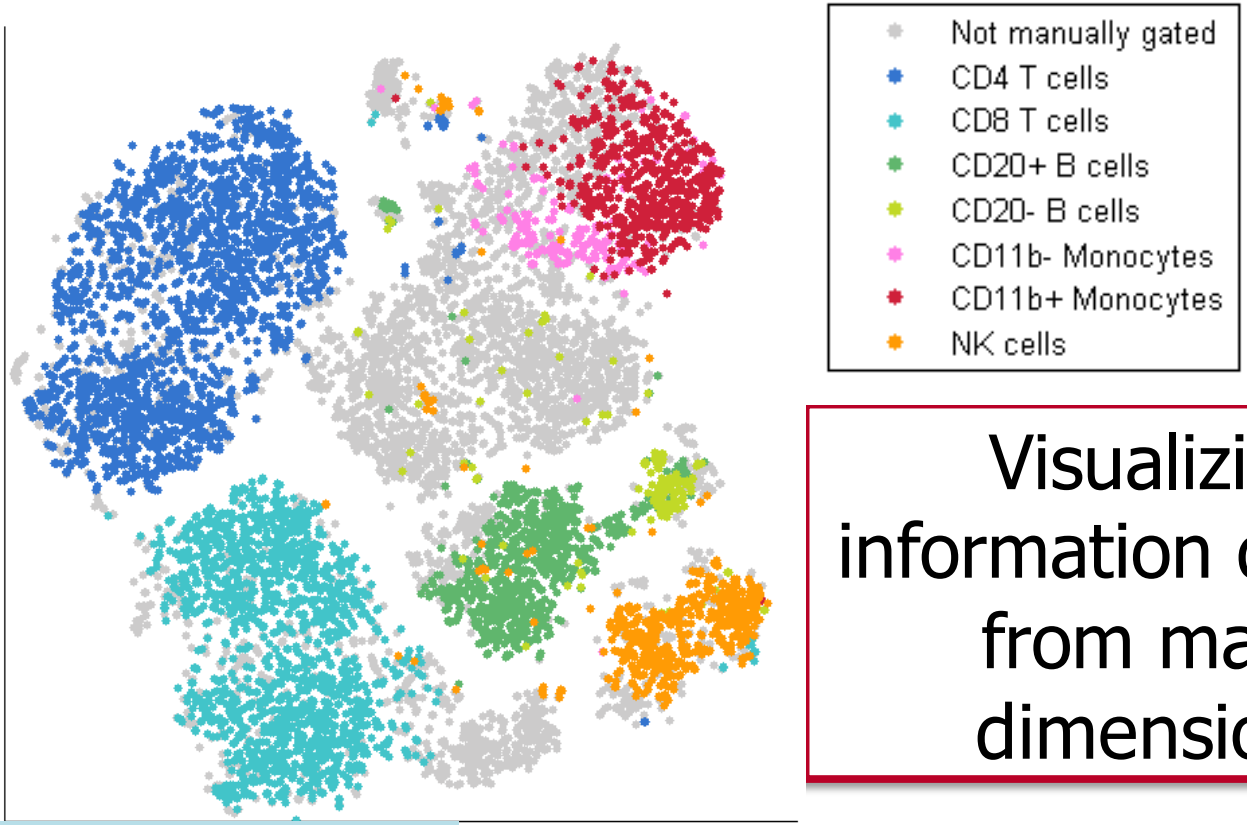


Cell Phenotype: A configuration of multidimensional expression

- Defines a region in "**phenotypic space**"
- Data will consist of **millions** of multi-parameter cells

Emerging high dimensional single cell technologies: CyTOF, single-cell RNA-seq and MIBI allow us to characterize "phenotypic space"

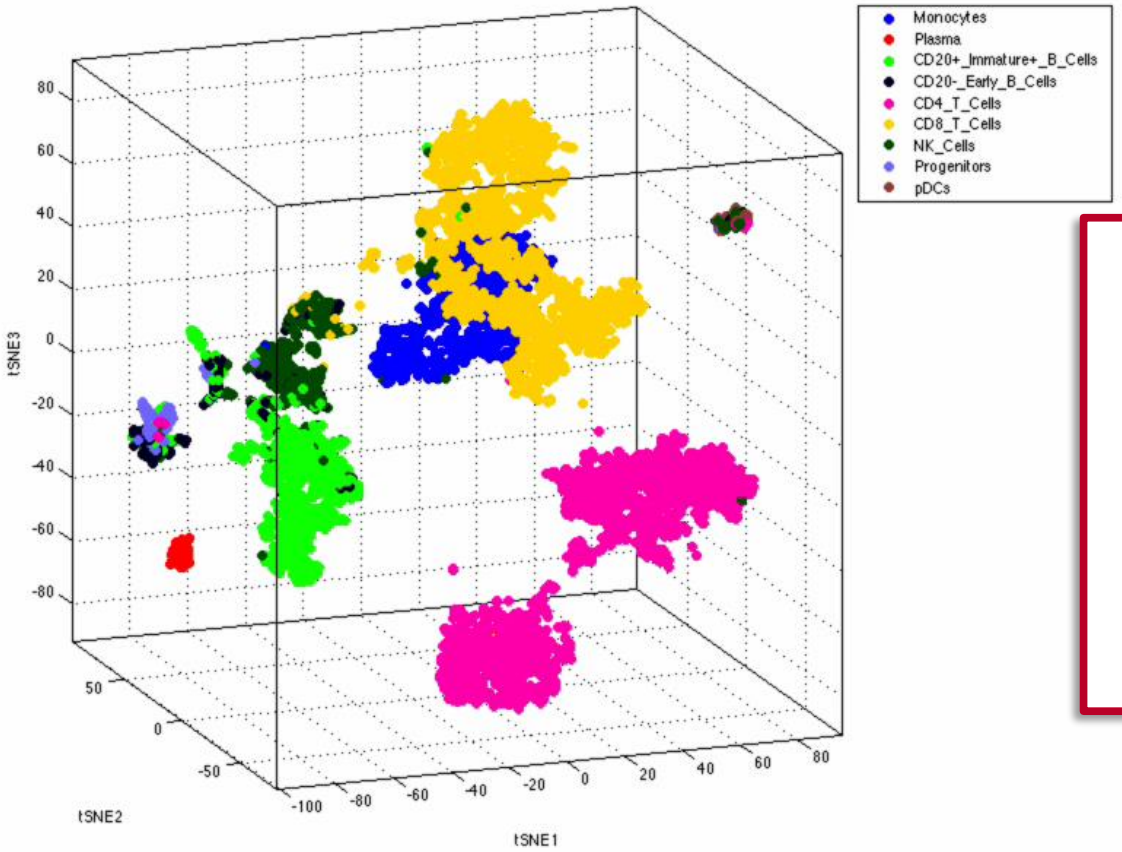
viSNE map of healthy bone-marrow



Visualizing information derived from many dimensions

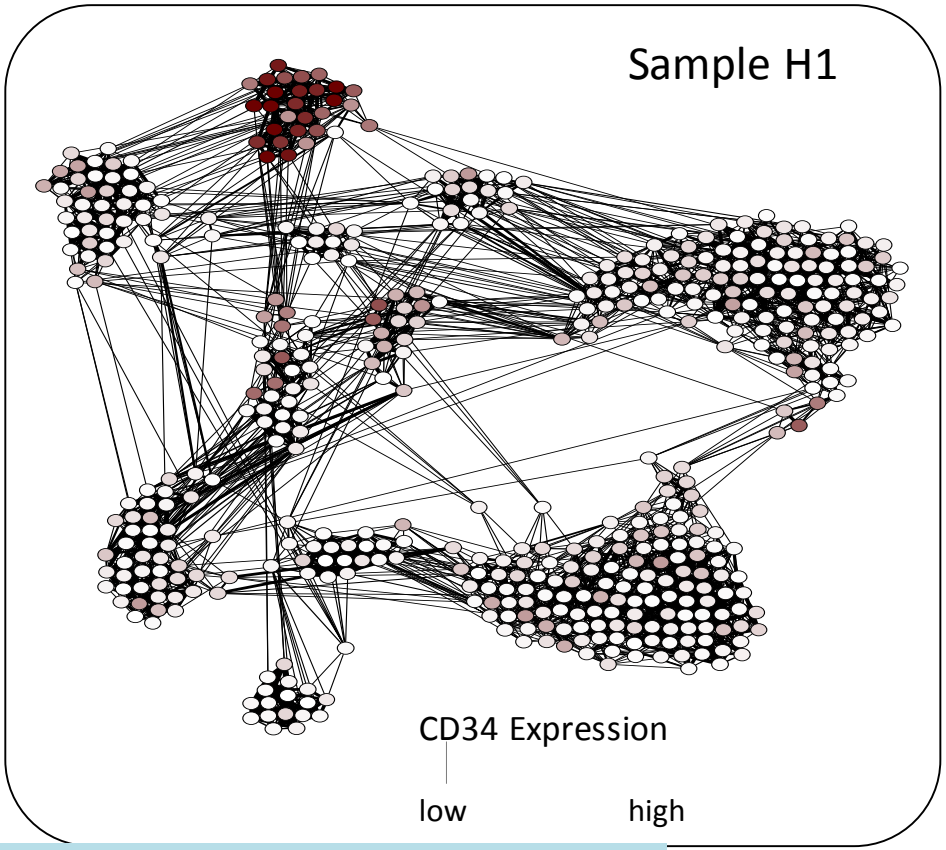
Amir et. al. Nature Biotech 2013

t-SNE based Non-Linear Dimensionality Reduction



Cell phenotypes accumulate in complex non-convex manifolds

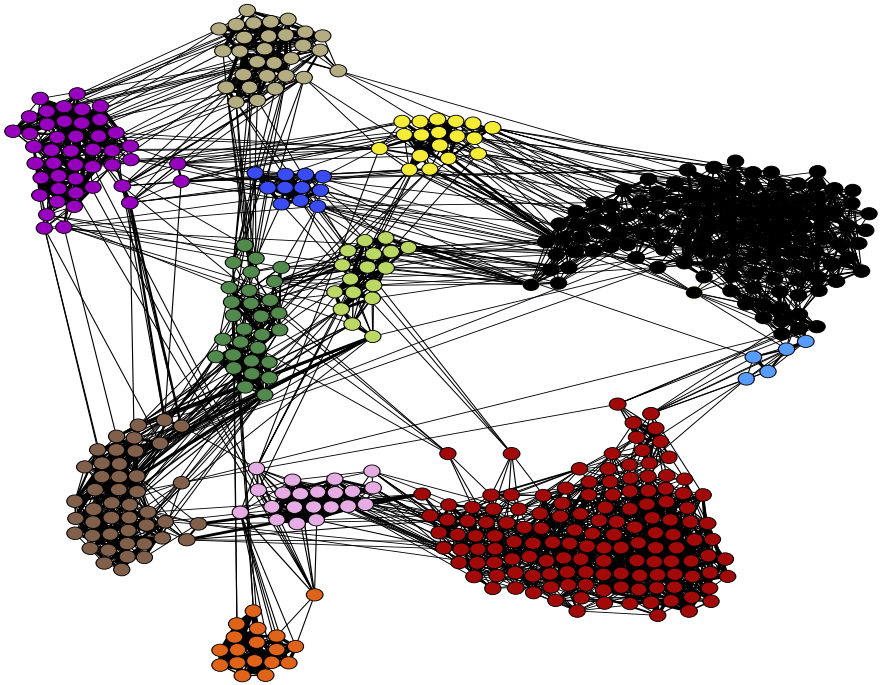
A “social network” for cells



Levine,Simonds* et. al. Cell 2015*

- Convert data to graph using Jaccard metric
  - Graph approximates phenotypic manifold
- Perform density estimation <u>in the graph</u>
  - Identify regions of phenotypic stability
- Produce explicit labeling of subpopulations

Phenograph, community detection



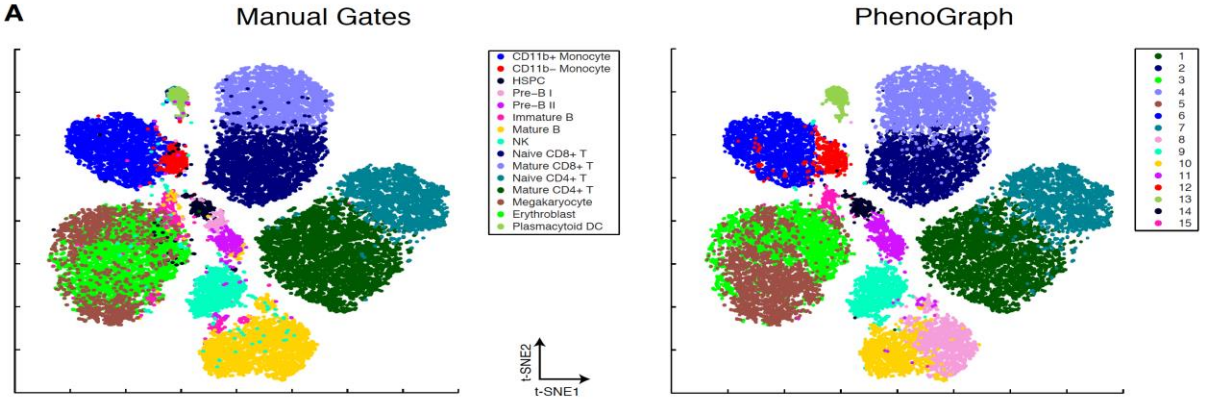
- Community detection identifies densely interconnected node sets

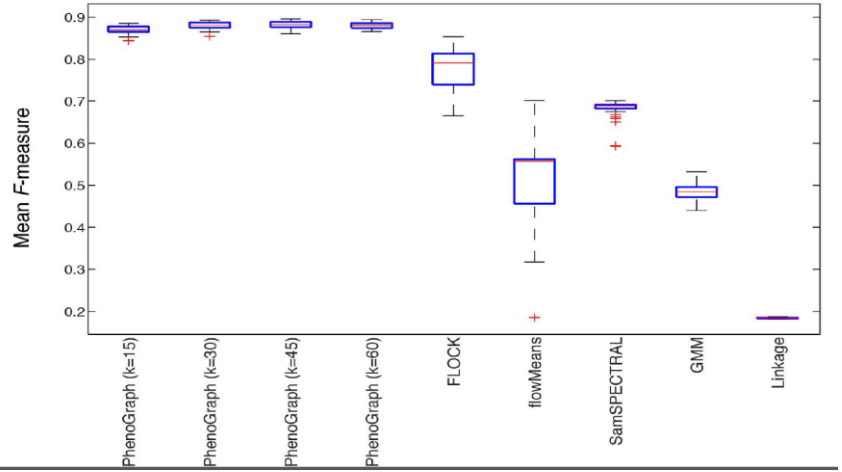
$$Q = \frac{1}{2m} \sum_{i,j} \left[W_{ij} - \frac{s_i s_j}{2m} \right] \delta(c_i, c_j)$$

  - W_{ij}: affinity function [ij coupling]
  - s_i: total affinity of i
  - c_i: community assignment for i
  - 2m: vol(W) [normalization]

- Combinatorial optimization
  - Louvain method provides efficient heuristic (Blondel *et al.* J. Stat. Mech. 2008)

PhenoGraph outperforms leading methods for subpopulation detection





Can run on 1 million cells in the same time it takes competing methods to run on 80,000 cells

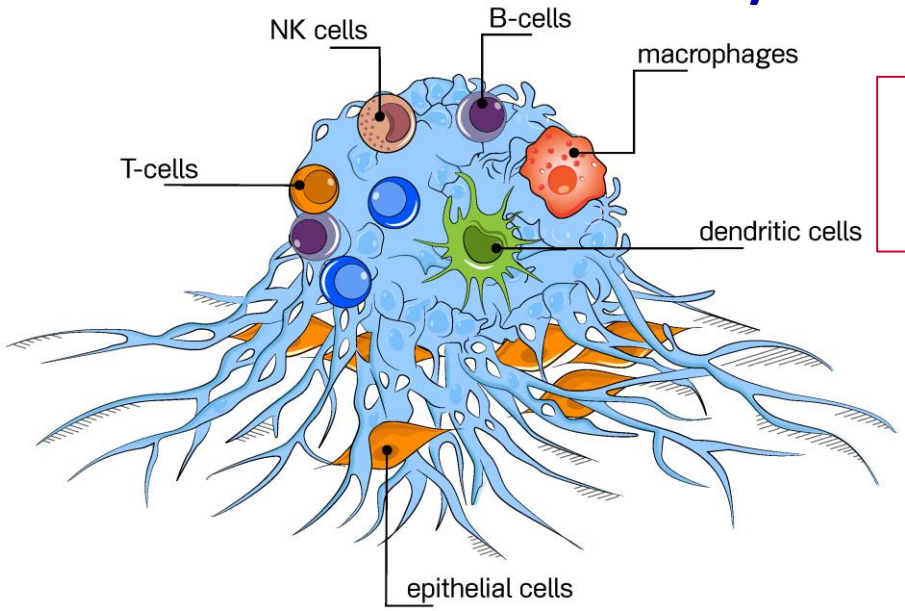
Immunotherapy in Cancer

- **The miracle**: 40% of metastatic melanoma patients showing "durable response" of many years
- Success stories in many additional "bad cancers" including Lung, AML, Bladder, Glio-blastoma
- Immunotherapy works for a small % of cancer patients, but when it works, it works

Precision Cancer Medicine

- Current efforts are based on “targeted therapy”, but
  - Cancer is so “smart and evolving”, simple drugs will not cut it.
  - Preexisting resistant clones present before treatment
- Need smart and adaptive drug like our own immune system
- **Need:** “big data” approaches to understand how immunotherapy can be extended to all patients

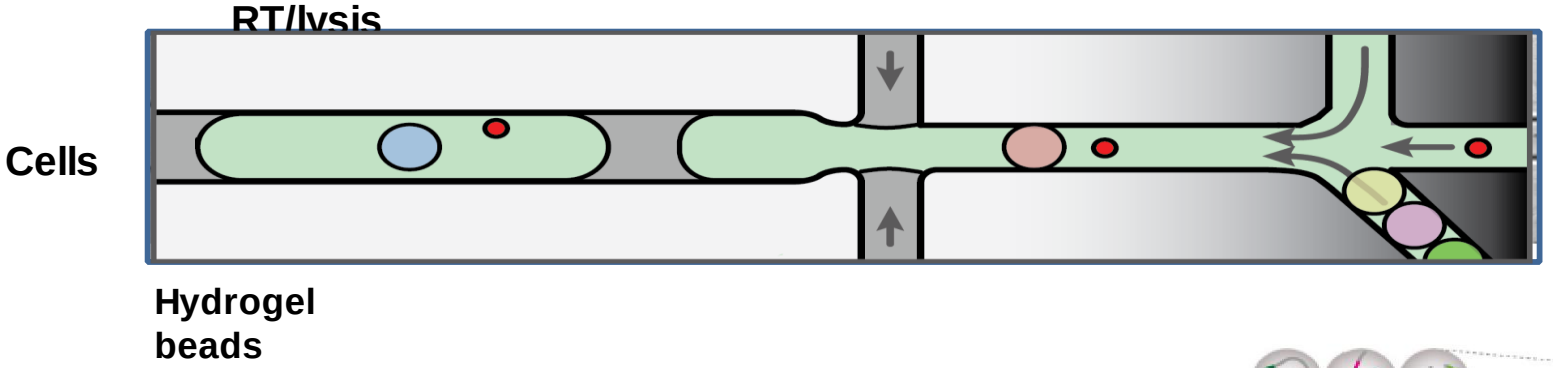
Tumor Immune System Atlas



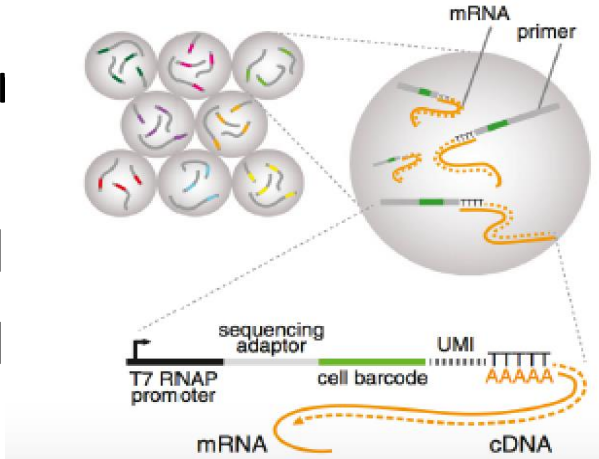
Need thousands of CD45+ cells per tumor

- Goal: Characterize sub-populations in tumor immune ecosystem.
- Challenge: Substantial unknown diversity.
- A better understanding of tumor immune eco-system will aid the development strategies to activate it against the tumor.

In-Drop Parallel Processing of RNA-Seq Libraries from >10,000 Individual Cells

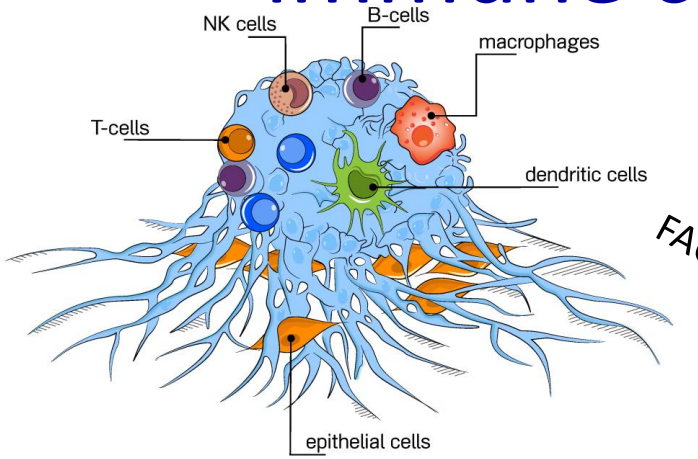


- Microfludic device can do 30,000 cells in o
  experiment
- Tiny wells cut cost of reagents by 1000-fold
- **Highly scalable and inexpensive single-cell seq**

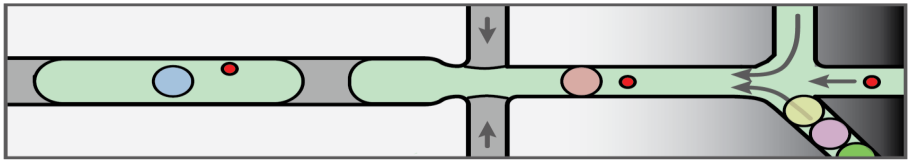


Klein, Mazutis* et. al. Cell 2015*

In-drop characterization of tumor immune cells in breast cancer



FACS for CD45+ cells

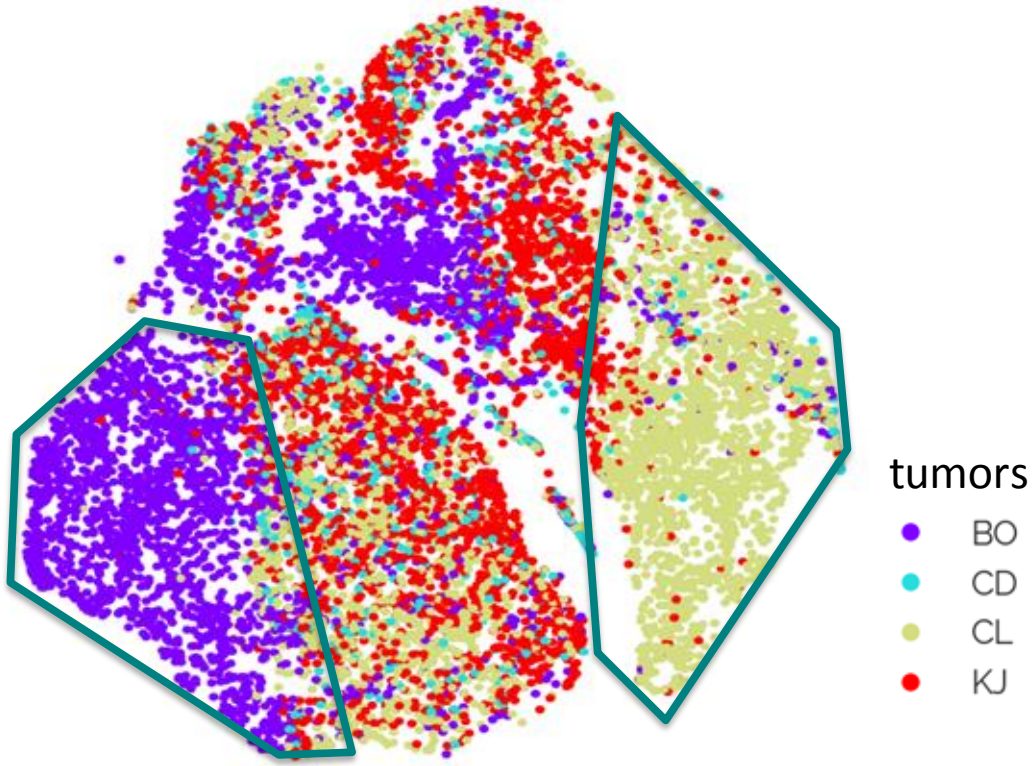


Data-Driven approach:

- > 3000 CD45+ collected per tumor
- Mean molecules per cell > 3500

Carr, Mazutis, Plitas with Rudensky lab, MSKCC

CD45+ TILs from 4 breast cancers



tSNE 2D projection

- Entire regions on the map are tumor specific
- Are these differences real biology or technical effects?

Single cell RNA-seq as imagined

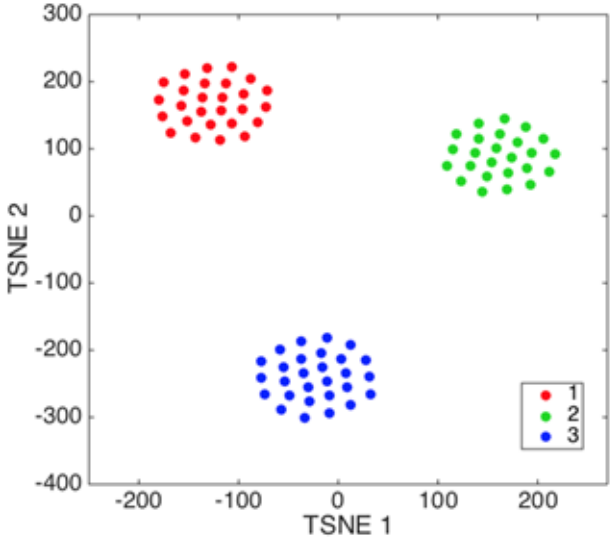
Count Matrix

Cells

Genes

Cluster 1 Cluster 2 Cluster 3

2D projection of cells (tSNE)

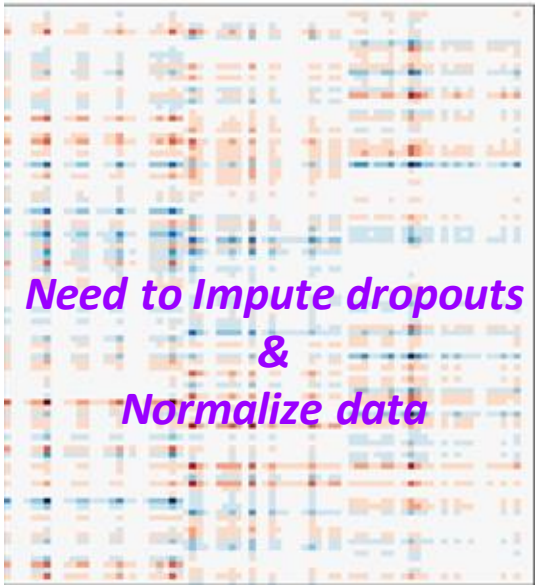


Problem: Single-cell RNA-seq data involves significant dropouts and library size variation

Observed Count Matrix

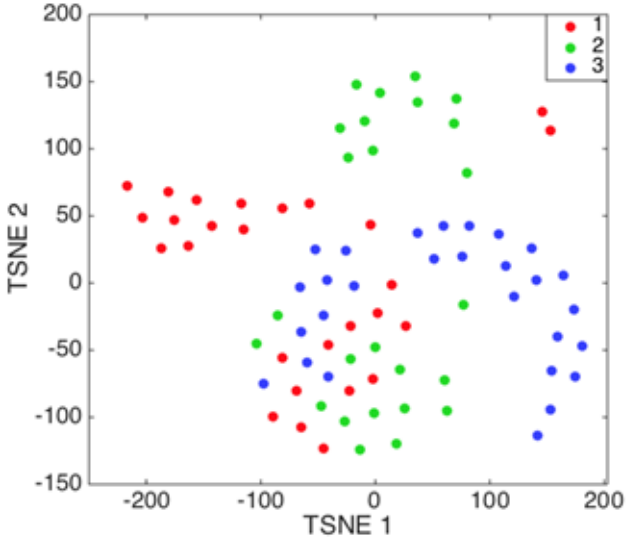
Cells

Genes



Cluster 1 Cluster 2 Cluster 3

2D projection of cells



Typically sample only 5% of a cell's transcriptome

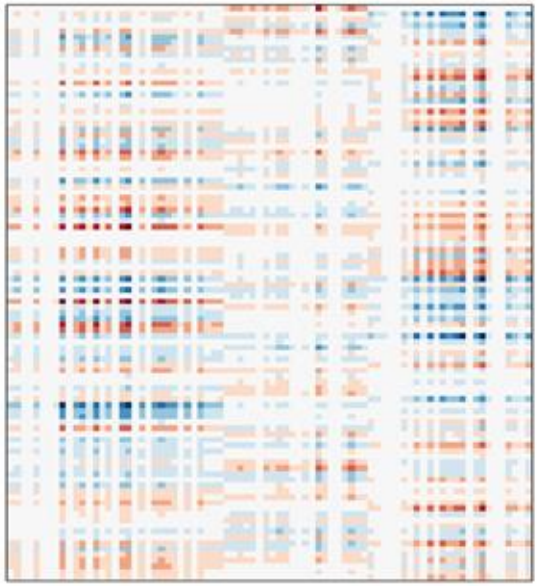
Common Approach:

Normalizing independent of cell types

Observed Count Matrix

Cells

Genes



Normalization

Clustering Cells

Downstream Analysis

To mean/median library size

BASiC/ERCCs

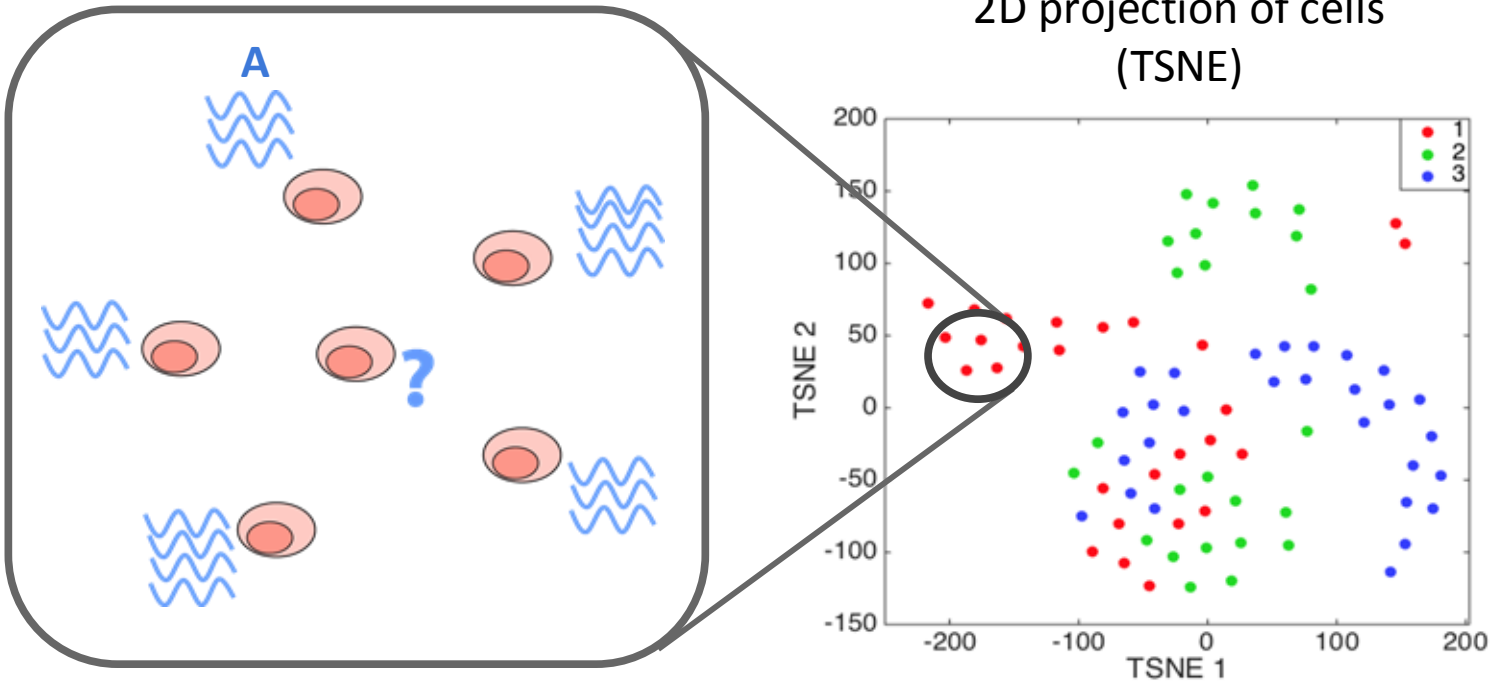
Problems:

- **Dropouts not resolved**

Zeros remain zero!

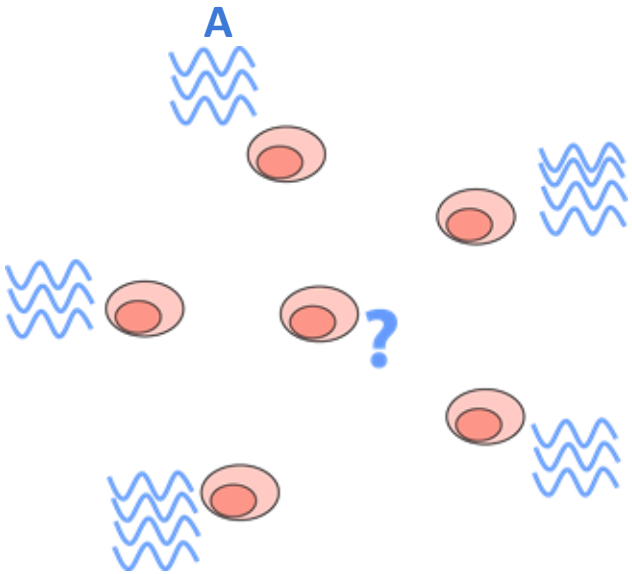
- Removes biological stochasticity specific to cell type
- Leads to improper clustering
- Biased results in downstream analysis

How can we **impute** expression in Single Cell RNA-seq data?



Prabhakaran, Azizi* et.al, ICML 2016*

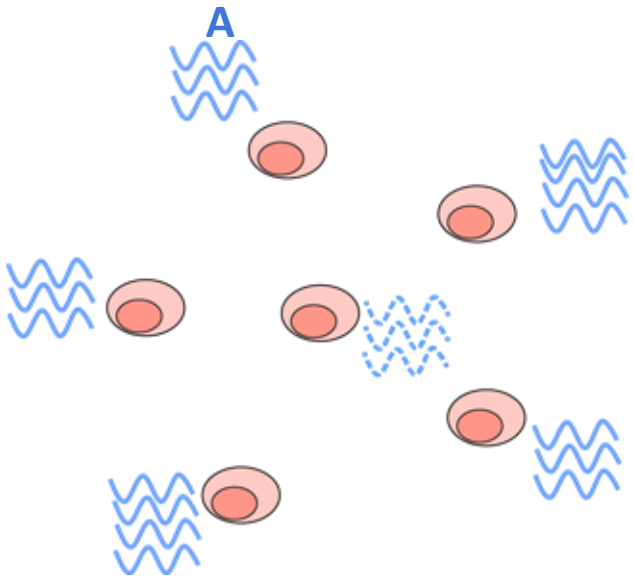
Idea 1: Impute dropouts based on expression in cells with same type



No expression of **Gene A** in a cell

But we observe similar cells mostly express Gene A

Idea 1: Impute dropouts based on expression in cells with same type

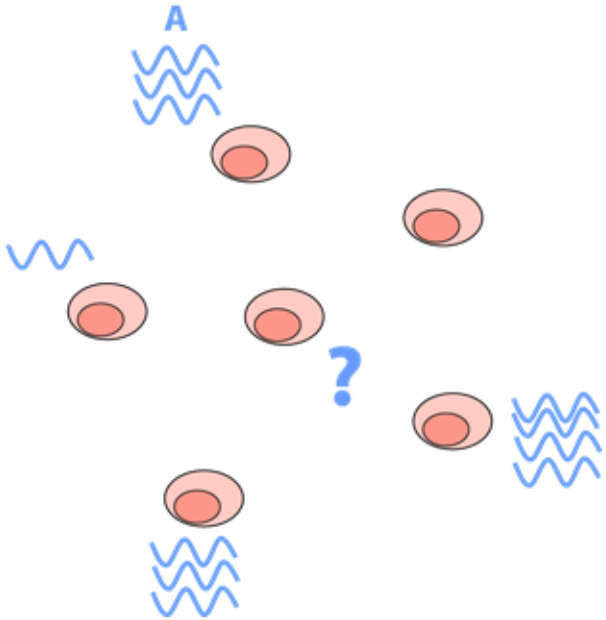


No expression of **Gene A** in a cell

But we observe similar cells mostly express Gene A

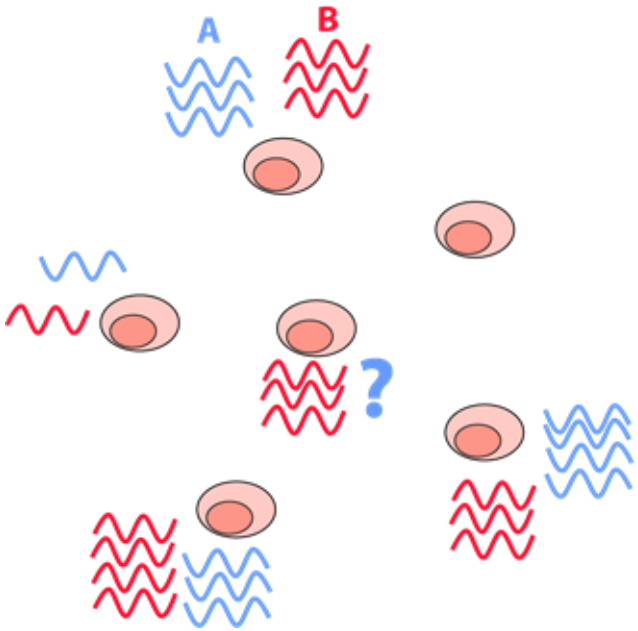
Impute dropout in Gene A based on similar cells

Idea 2: Impute dropouts based on co-expression patterns



No significant inference based on similar cells

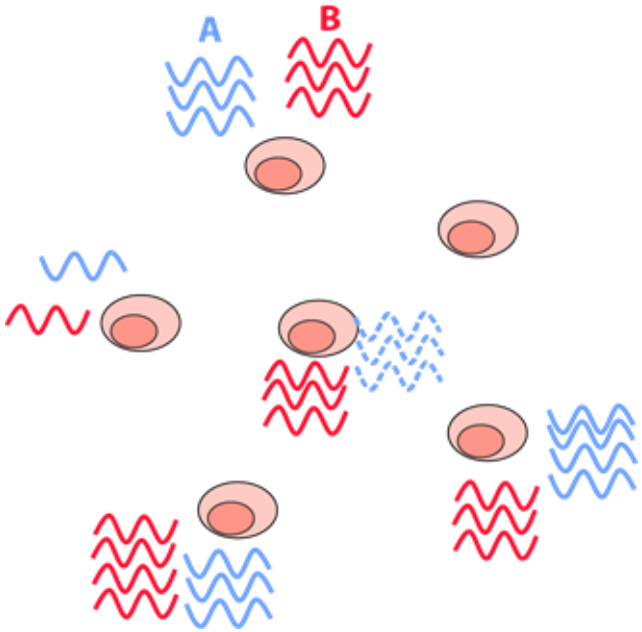
Idea 2: Impute dropouts based on co-expression patterns



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

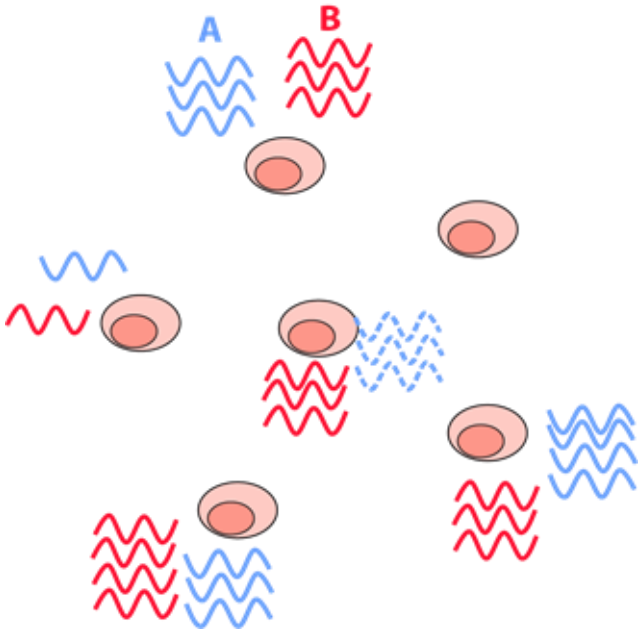


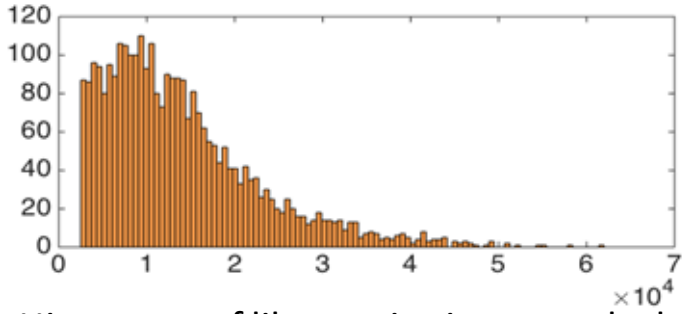
No significant inference based on similar cells

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

Impute dropout in Gene A based on Gene B

Imputing & **Normalization**

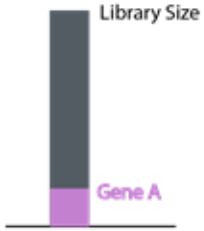




Histogram of library size in example dataset
From Zeisel, Science 2014

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

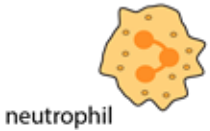
Problem with Global Normalization

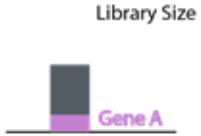


Cells with different sizes have very different total number of transcripts

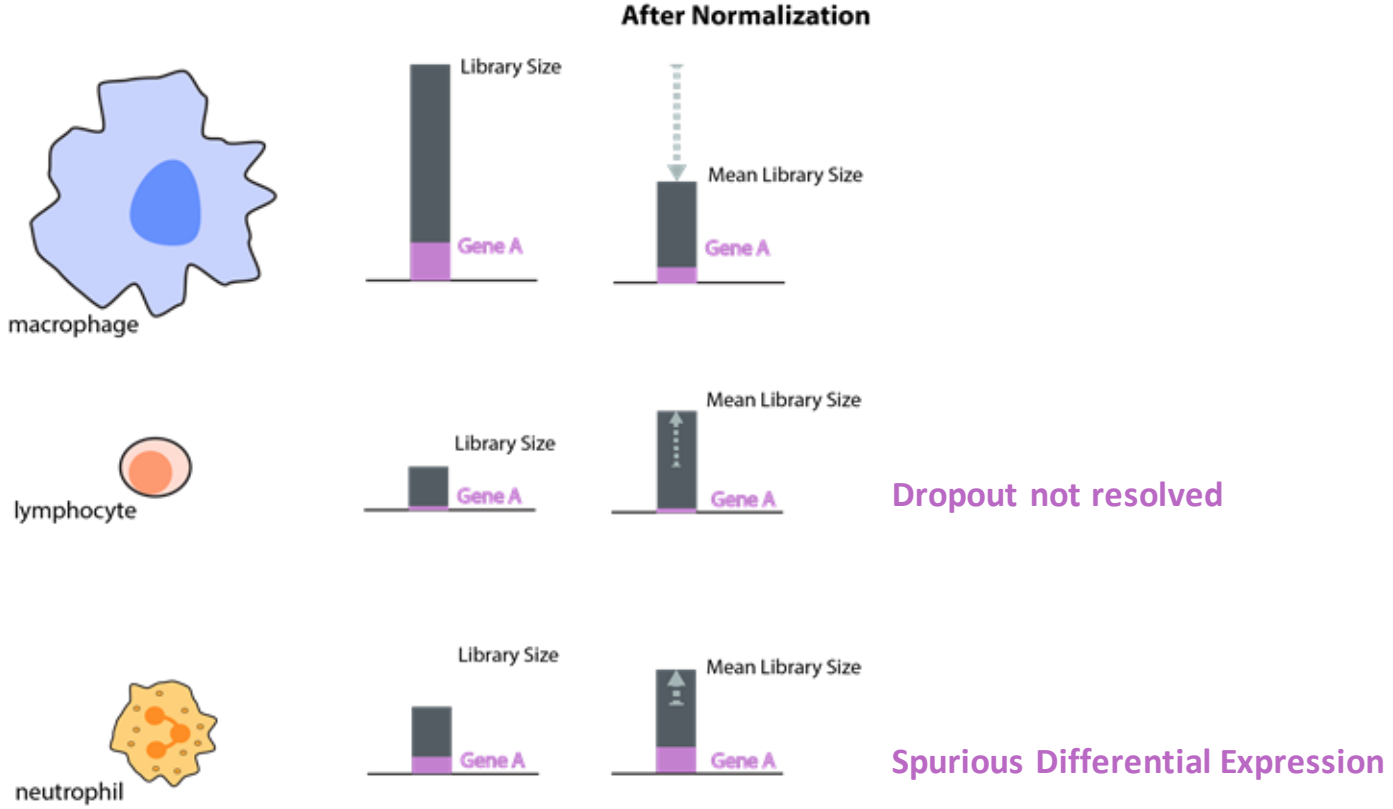
Example Housekeeping Gene

High chance of Dropouts in smaller cells

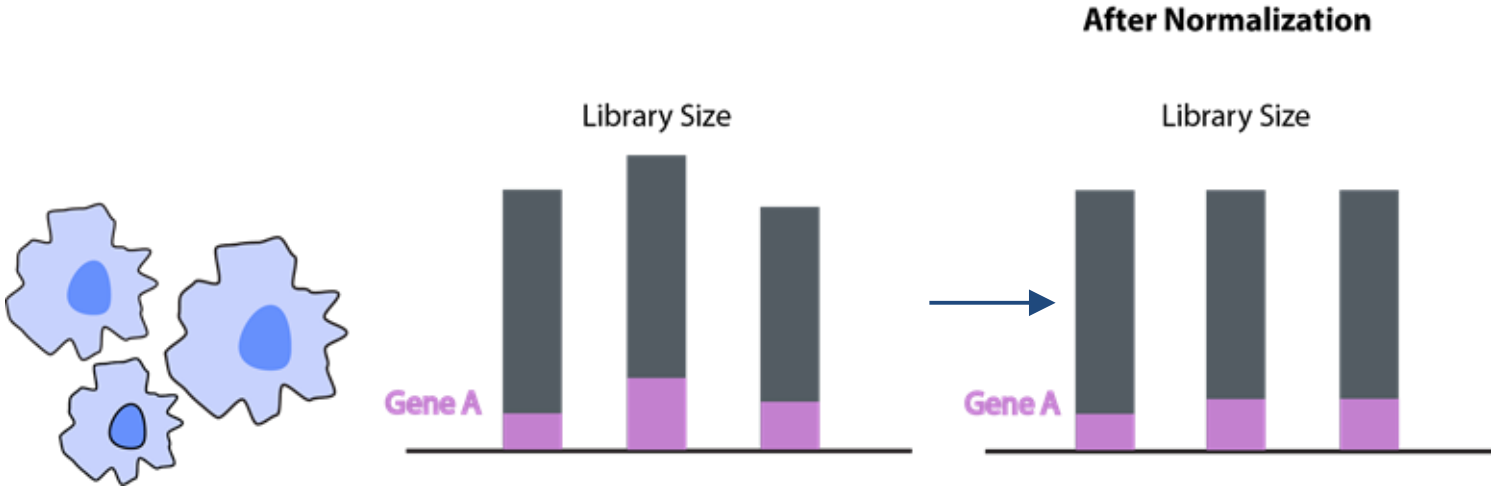




Problem with Global Normalization



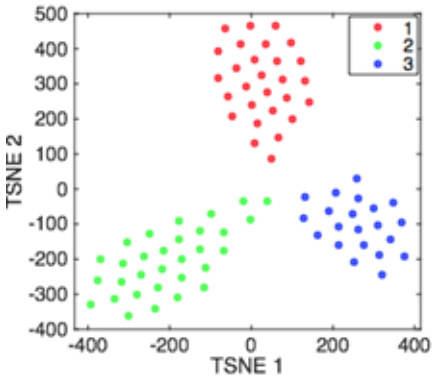
Idea: Different normalization for each cell type



Problem:
We don’t know cell types

Need to infer **cell clusters**

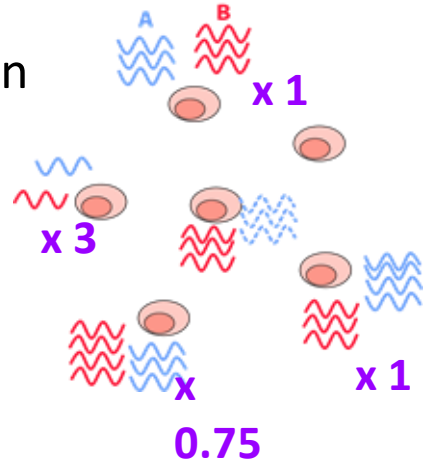
Approach: Simultaneous inference of clusters and imputing parameters



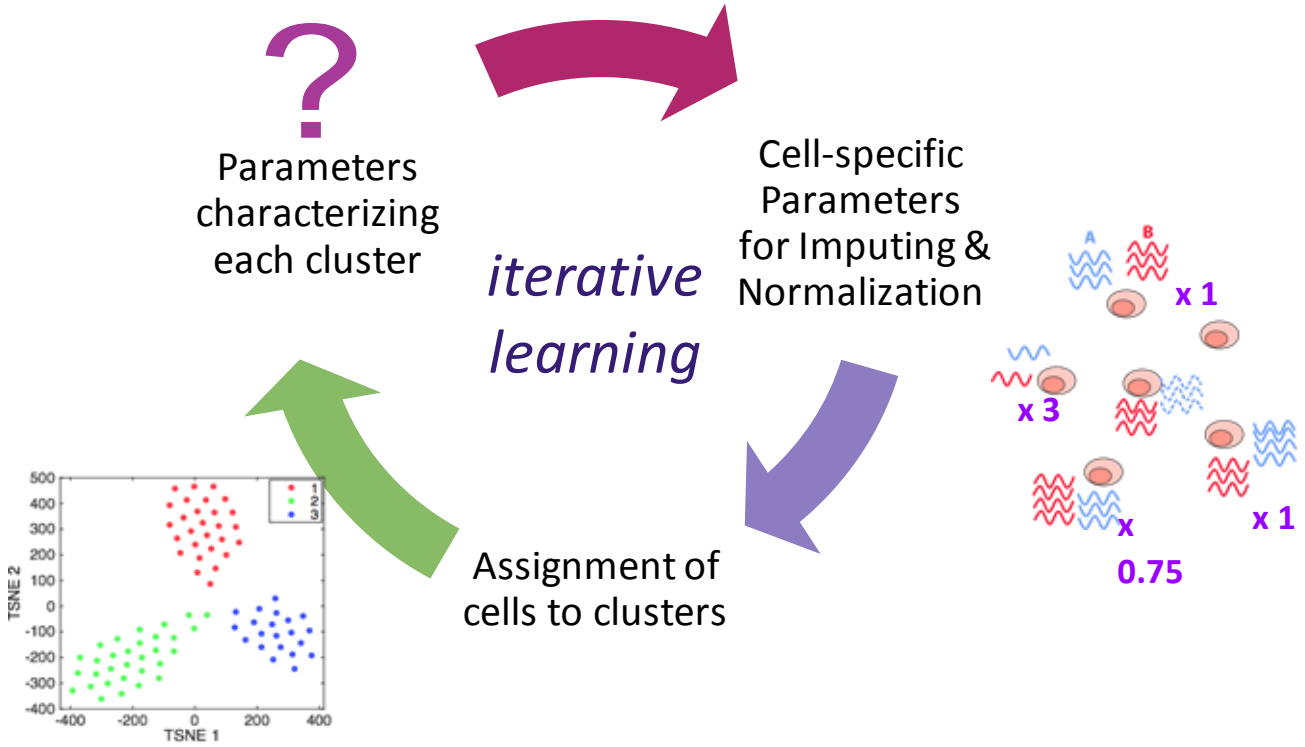
Clustering Cells

iterative learning

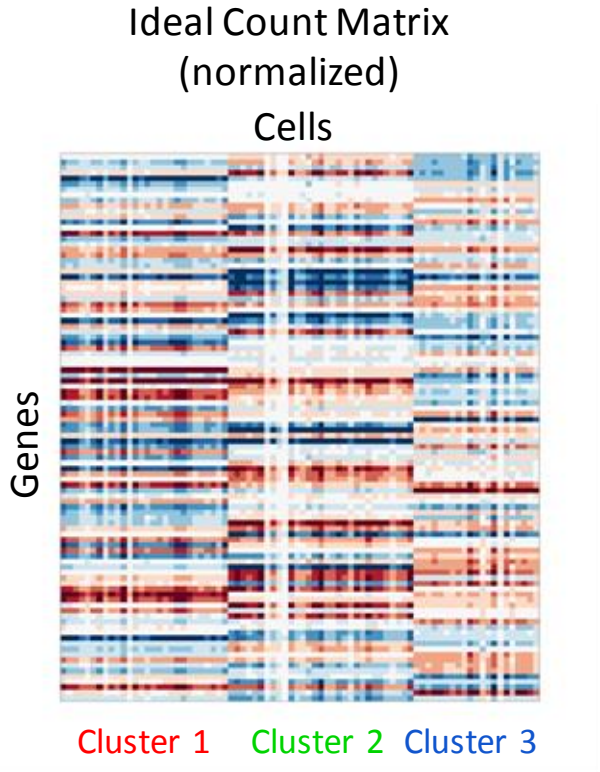
Imputing & Normalization



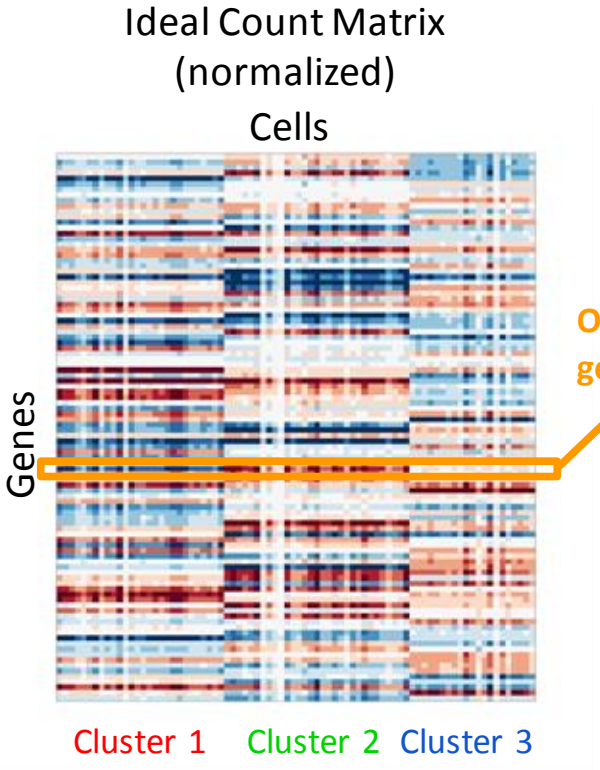
Approach: Simultaneous inference of clusters and imputing parameters



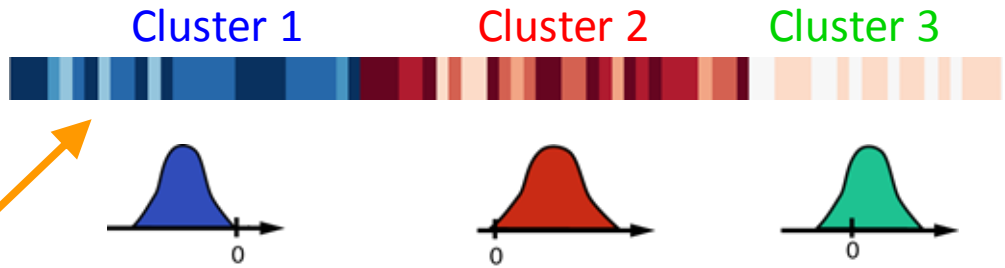
Modeling: Clusters of Cells using a Bayesian Mixture Model



Modeling: Clusters of Cells using a Bayesian Mixture Model

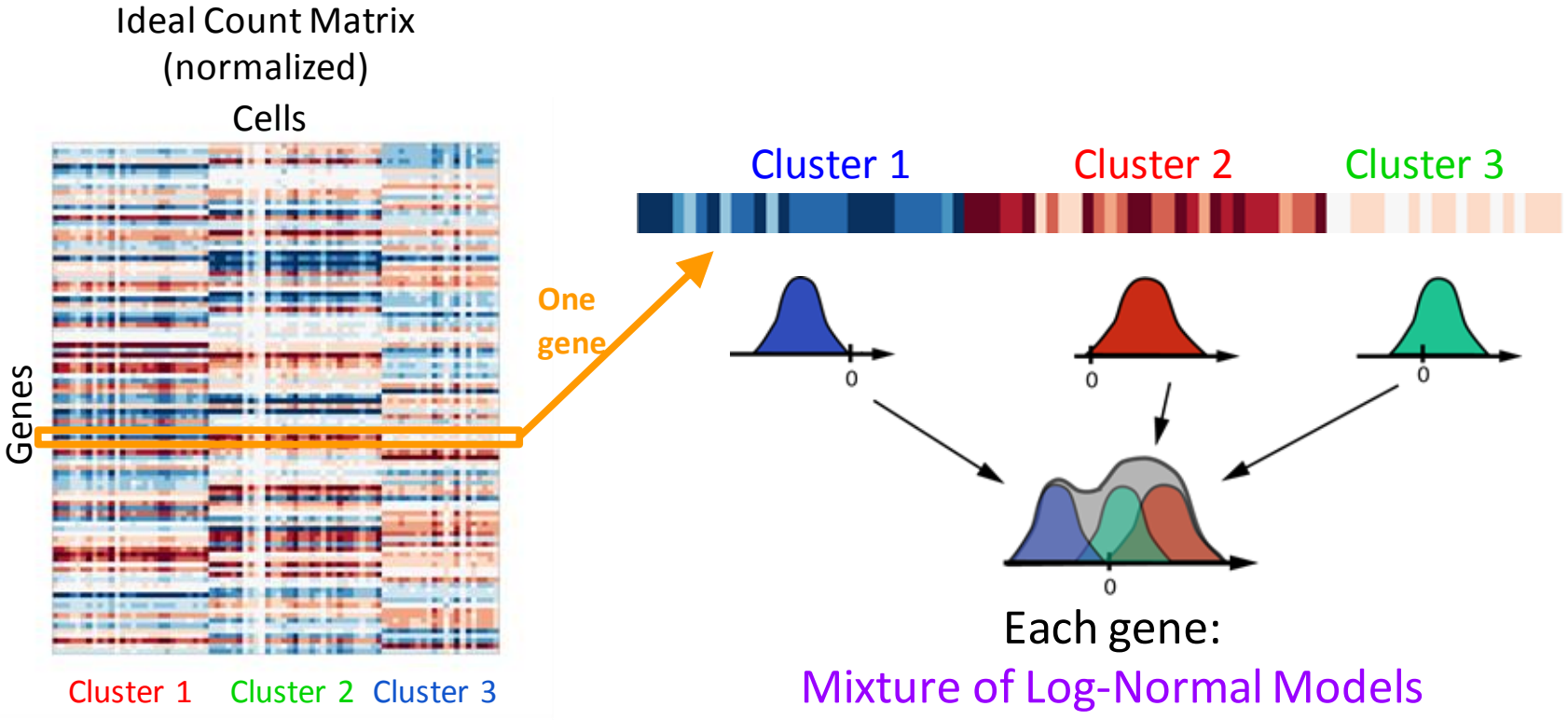


One gene

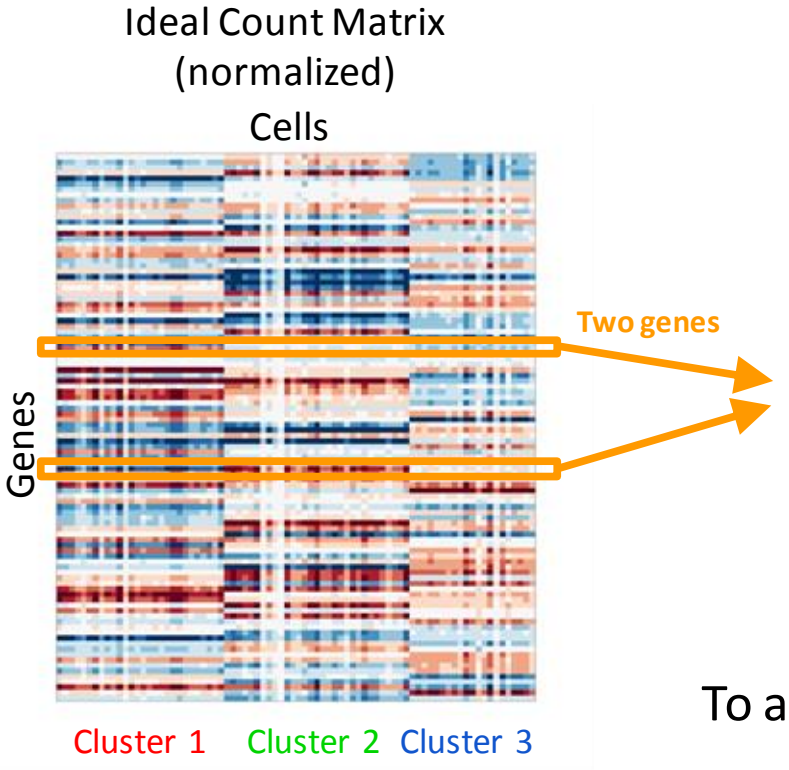


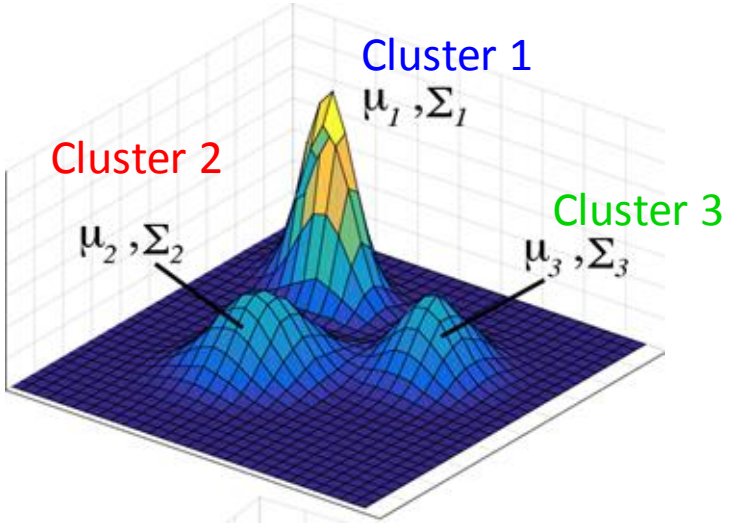
Model distributions of Log of counts for each gene per cell type as a Gaussian distribution

Modeling: Clusters of Cells using a Bayesian Mixture Model



Modeling: Clusters of Cells using a Bayesian Mixture Model



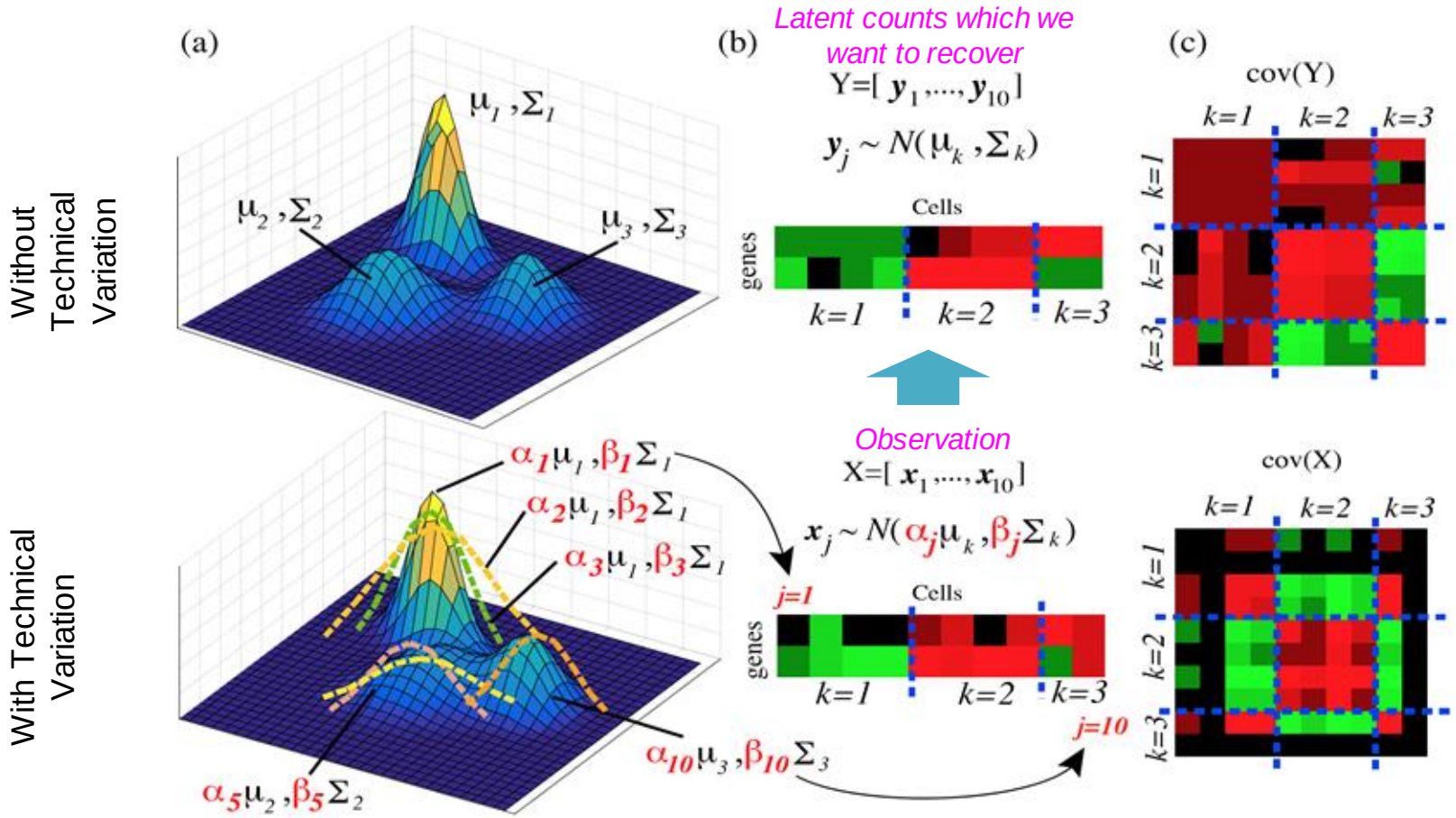


Modeling all genes together:

Mixture of Multivariate Log-Normals

To also take advantage of co-expression patterns in learning clusters

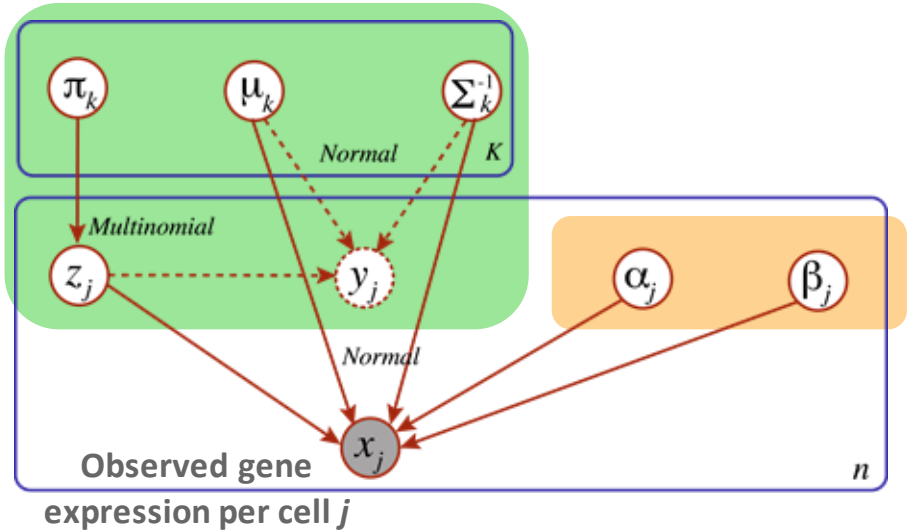
Generative Model with Technical Variation



BISCUIT

(Bayesian Inference for Single Cell ClUstering and ImpuTing)

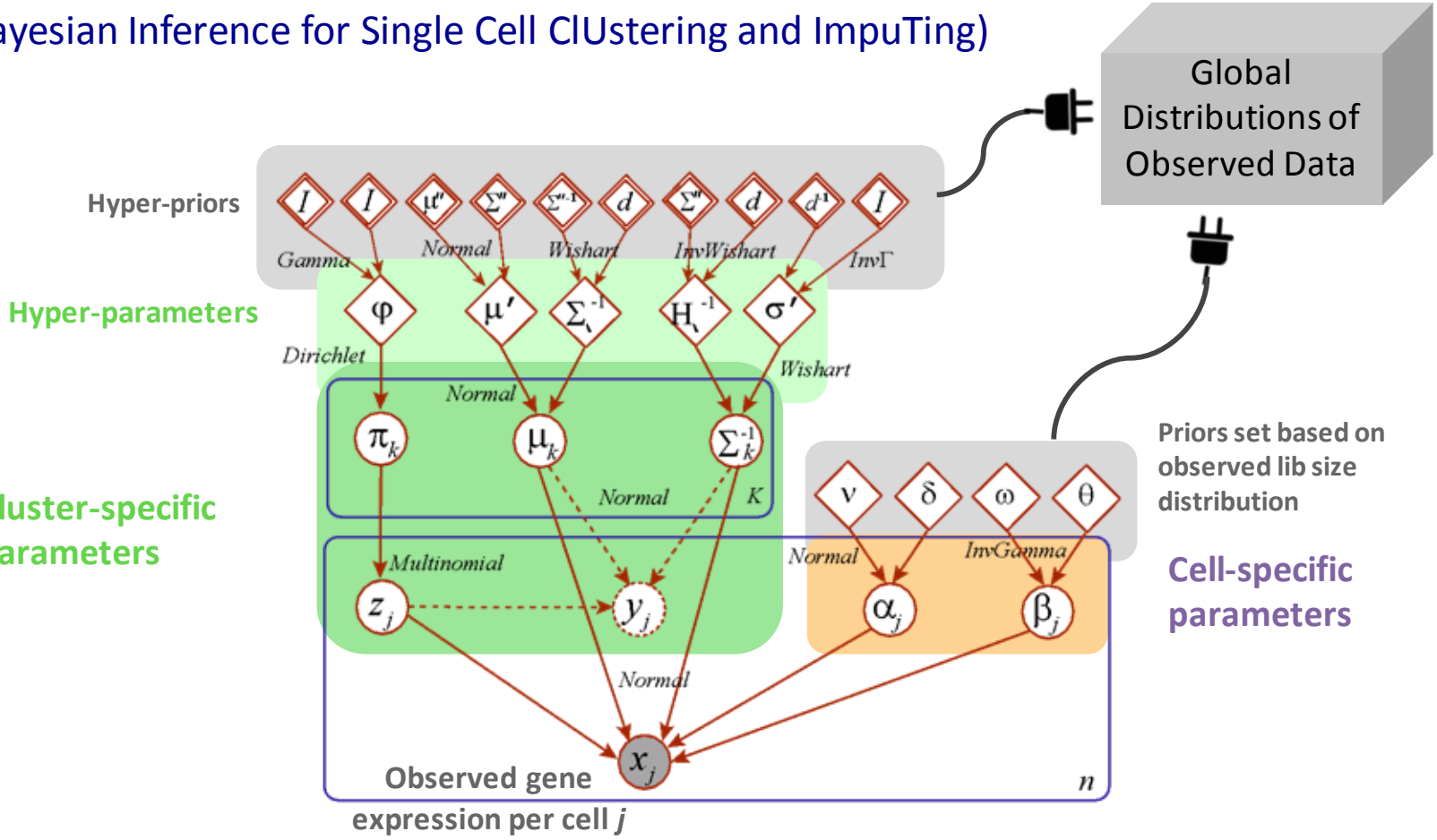
Cluster-specific parameters



Cell-specific parameters

BISCUIT

(Bayesian Inference for Single Cell ClUstering and ImpuTing)



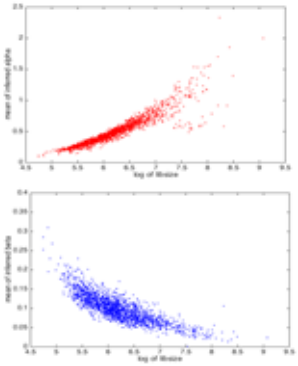
Inference Algorithm

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

Estimate hyper-priors based on Data

Sample hyper-parameters

Sampling technical variation parameters

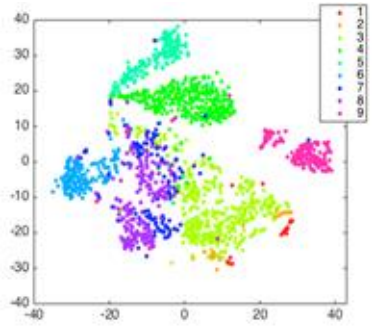


scaling mean, cov per cell

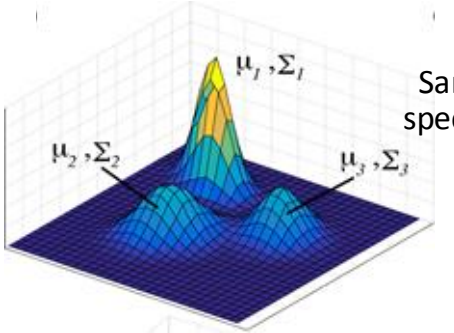
Gibbs iterations

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

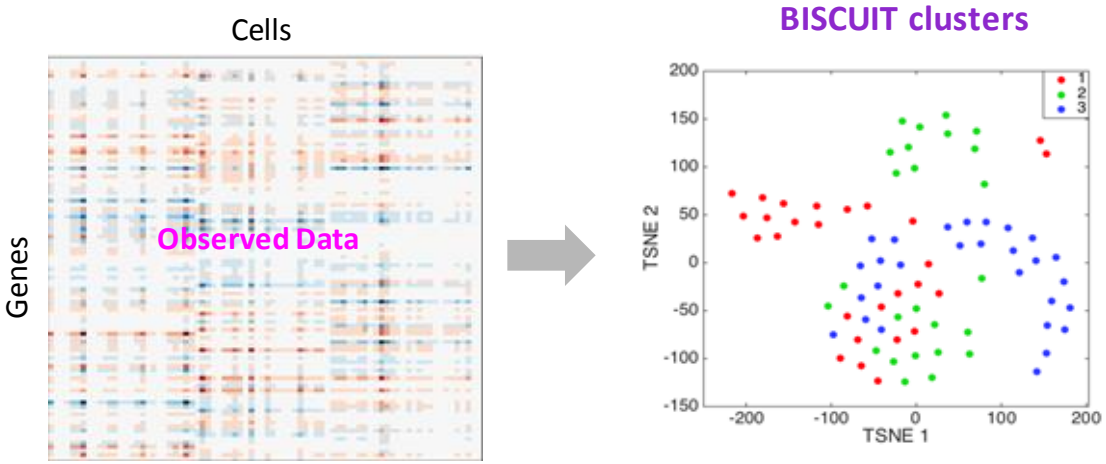
Also allows estimating the number of clusters



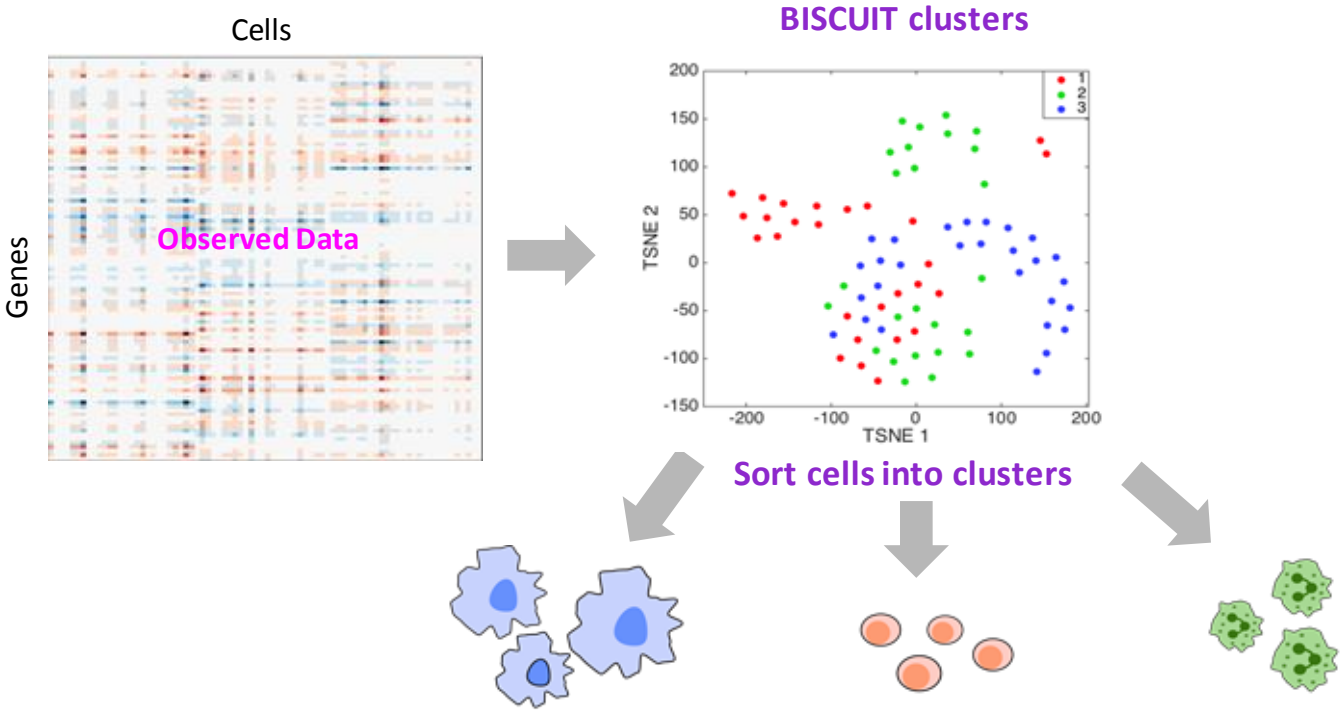
Sampling cluster-specific parameters



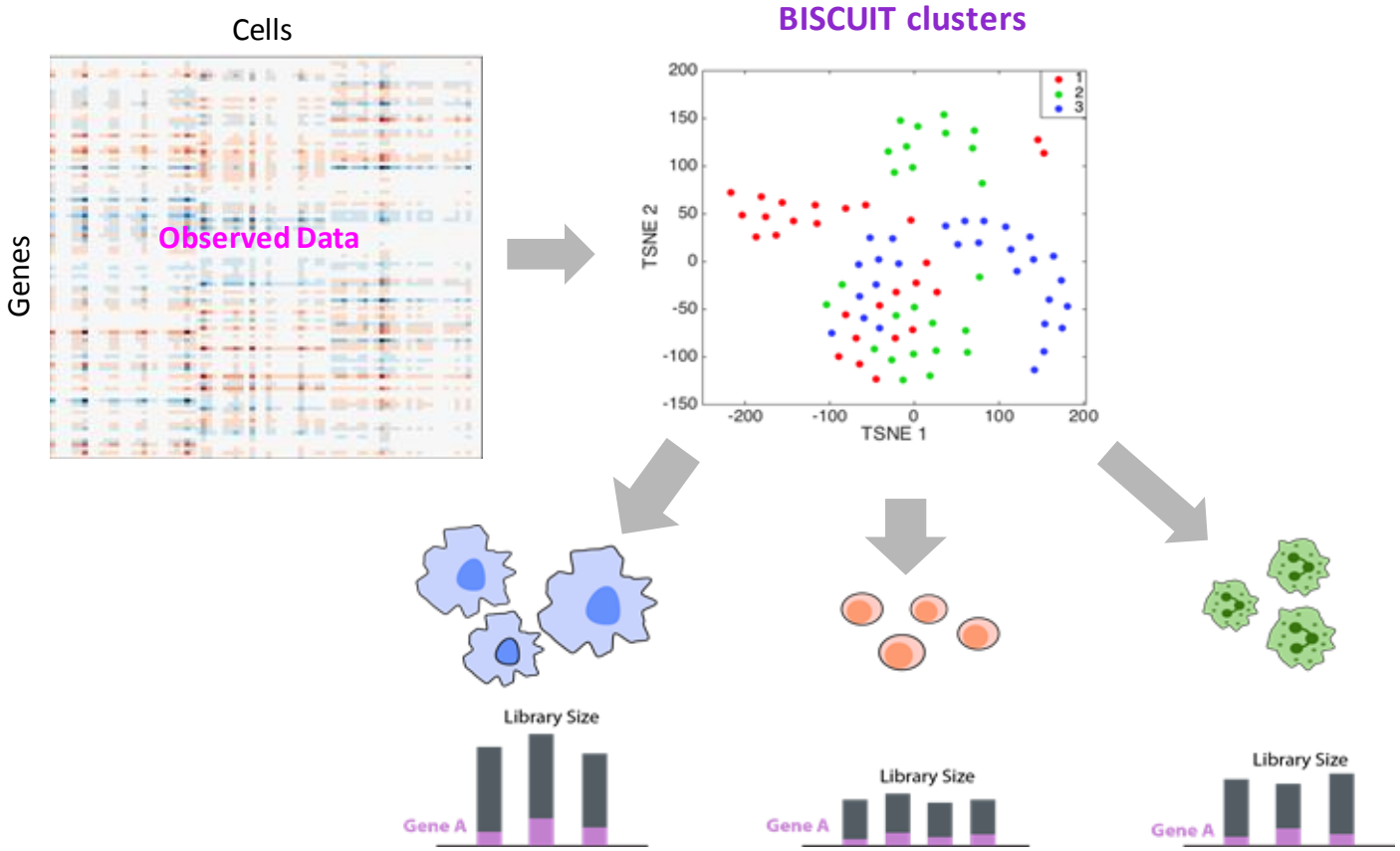
Cluster-dependent Imputing & Normalizing



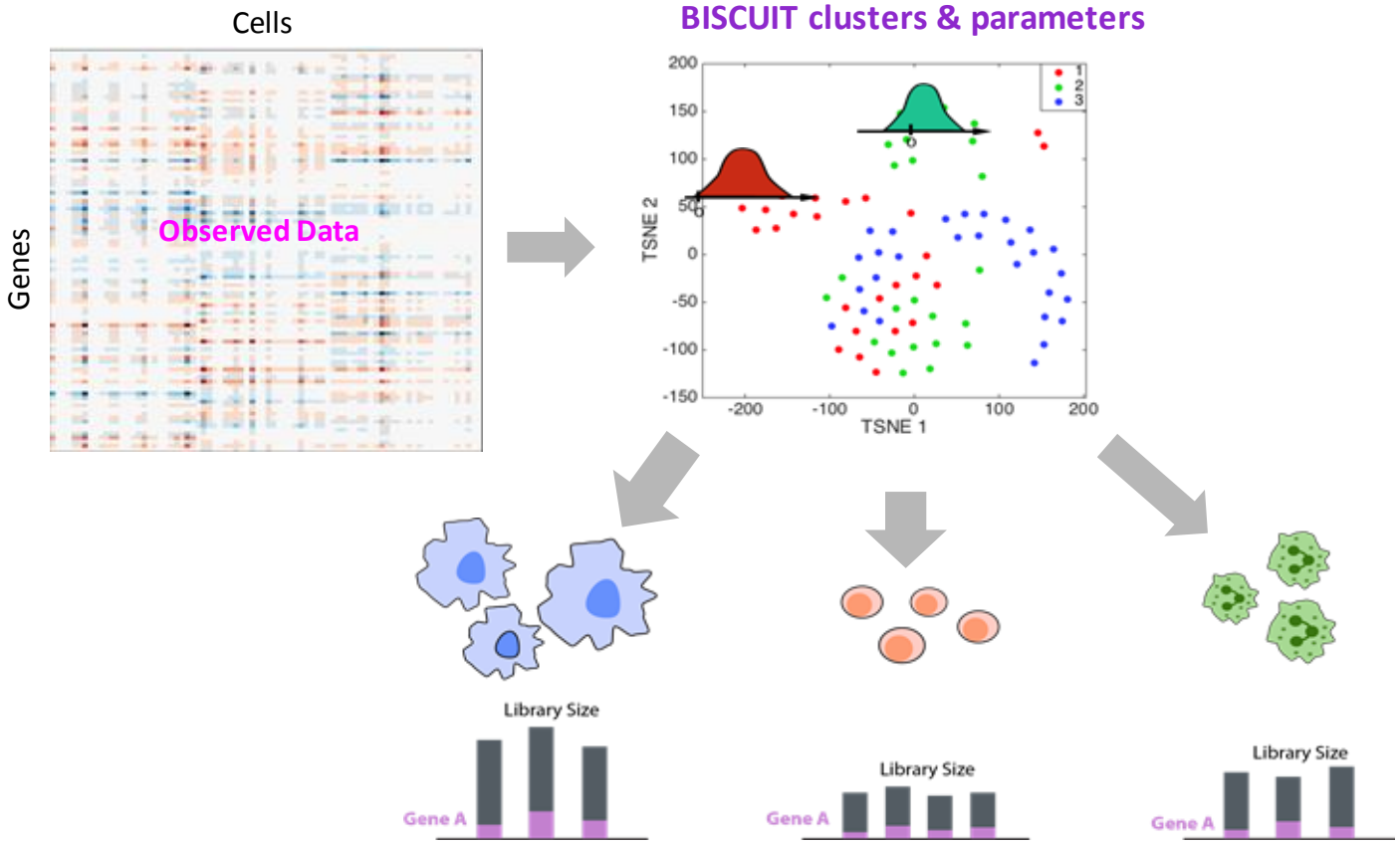
Cluster-dependent Imputing & Normalizing



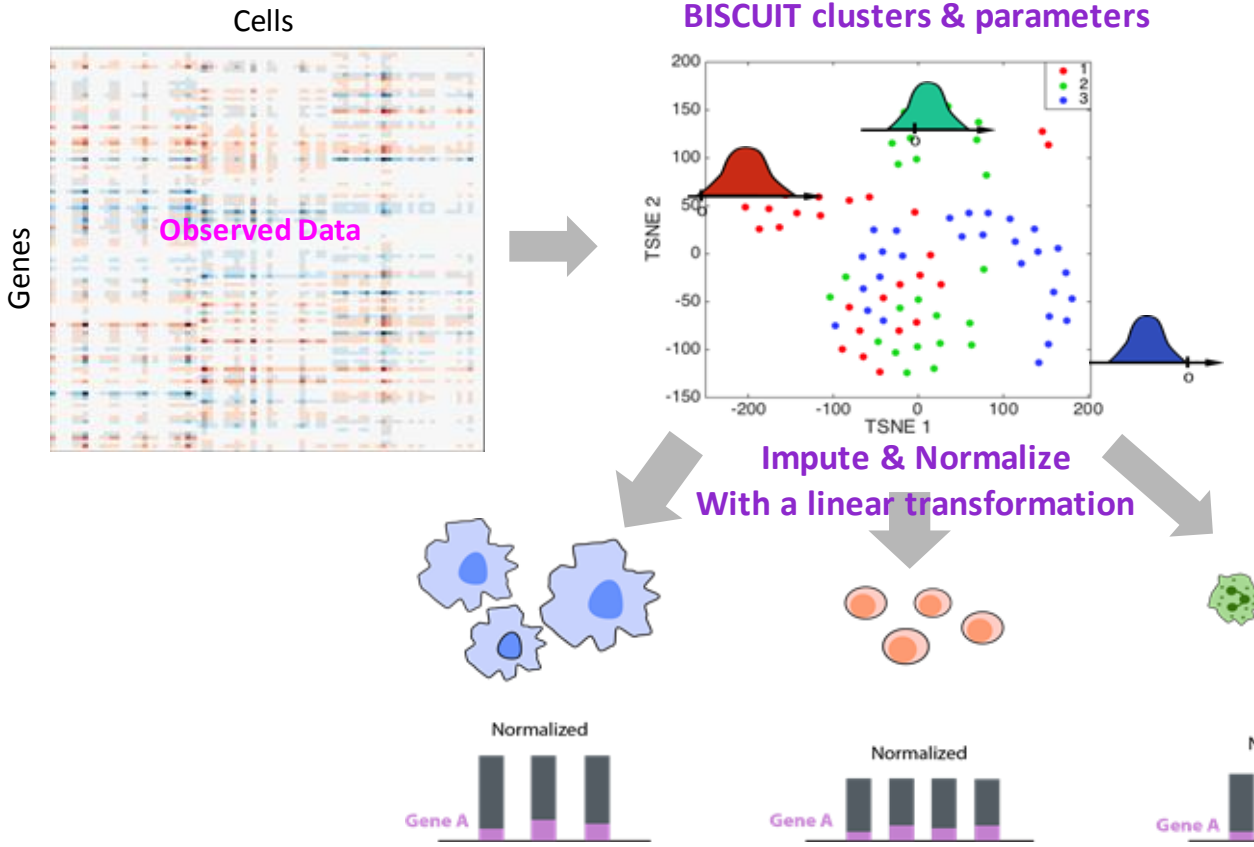
Cluster-dependent Imputing & Normalizing



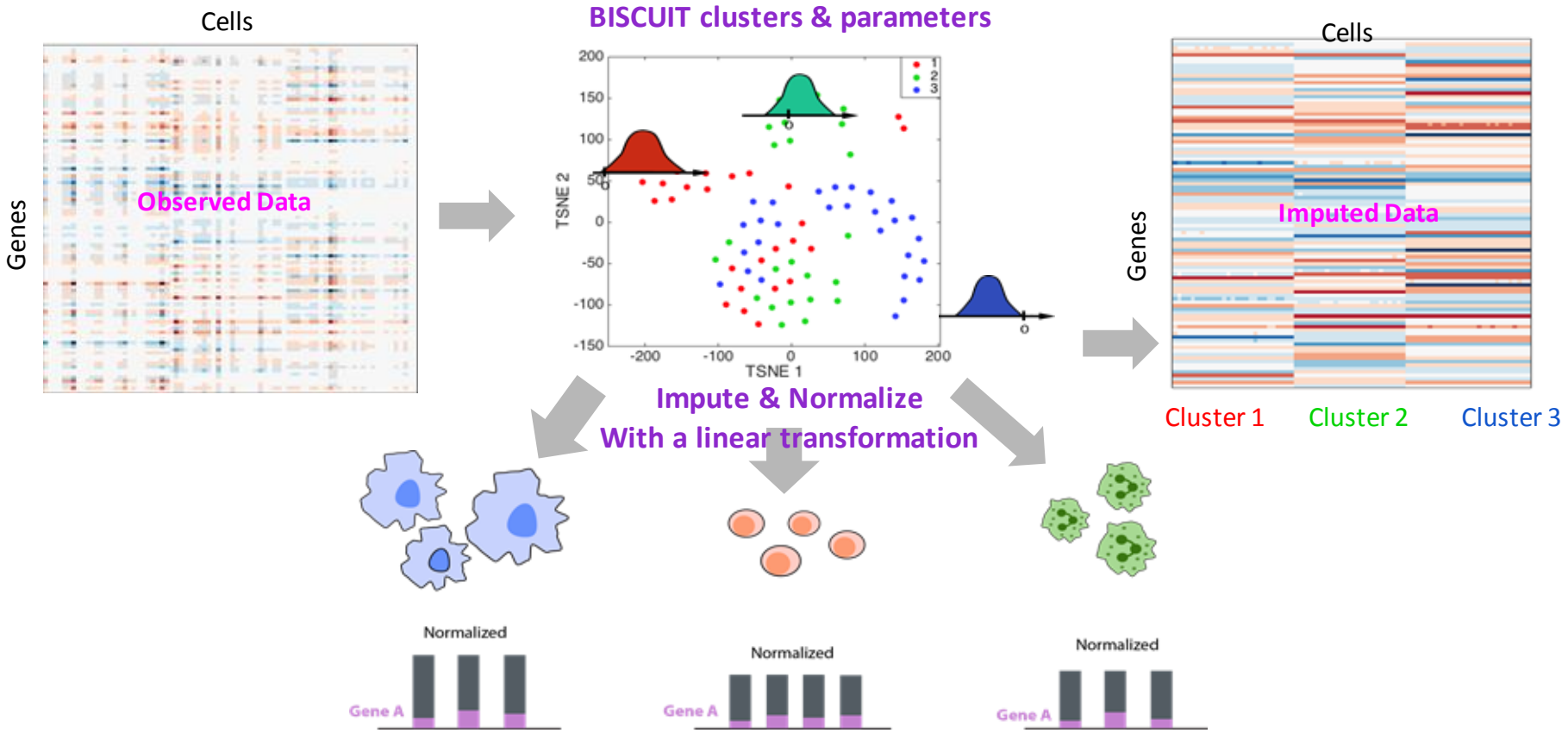
Cluster-dependent Imputing & Normalizing



Cluster-dependent **Imputing & Normalizing**



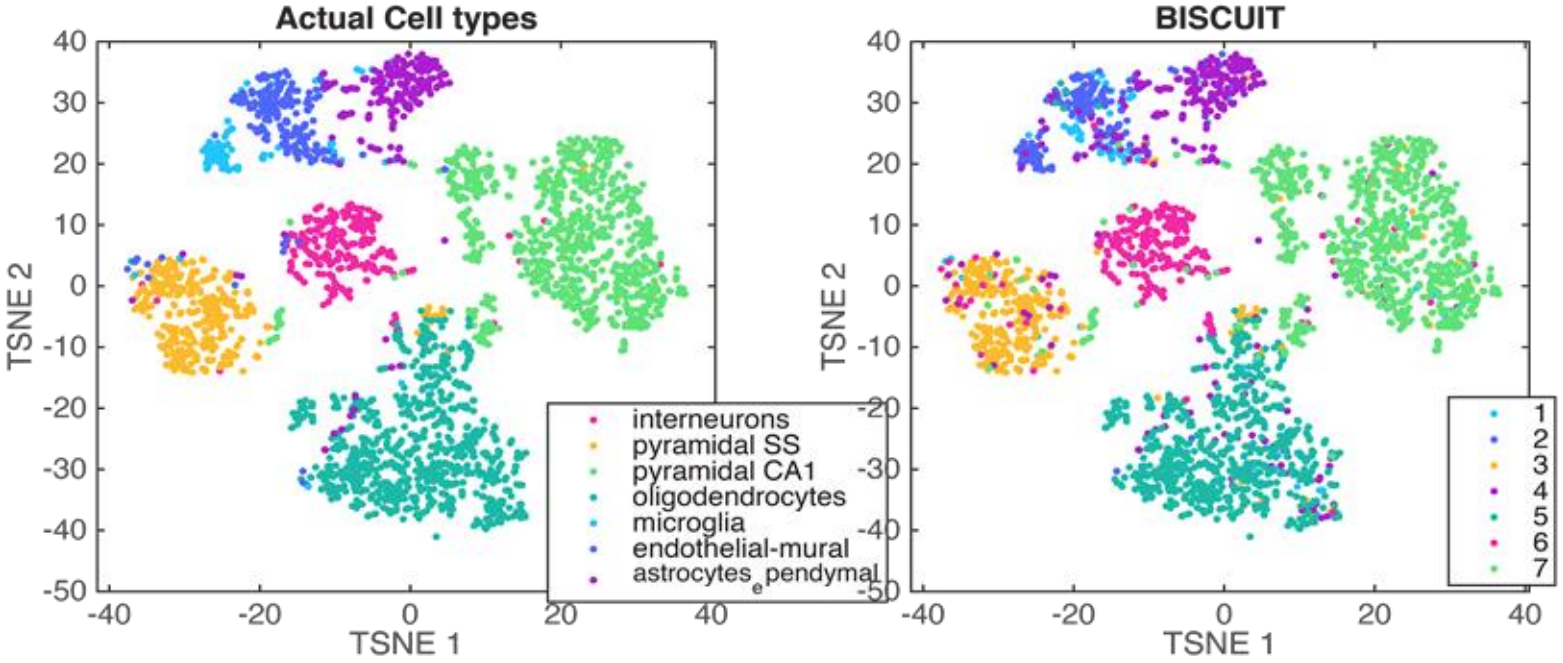
Cluster-dependent Imputing & Normalizing



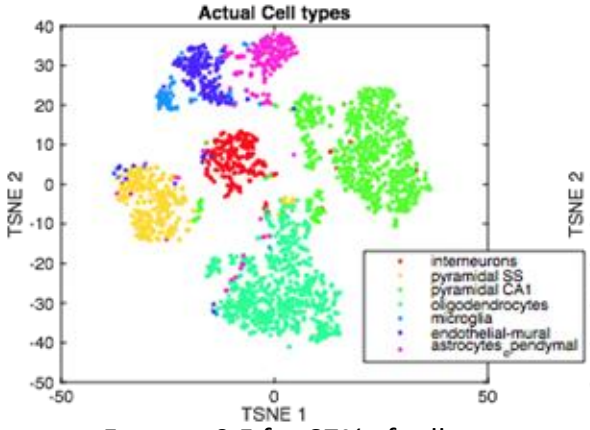
Performance: Testing on neuron single cell data

- 3005 mouse cortex cells from Zeisel et al., *Science* 2015
- Deep coverage (2 million reads per cell) gives good ground truth for 7 Cell types.
- No prior information used: selected 558 genes with largest standard deviation across cells

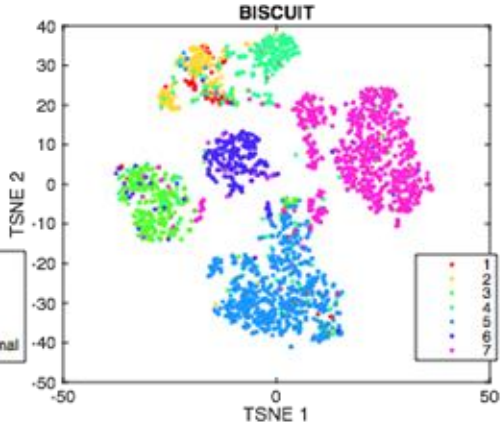
F-score: 0.91



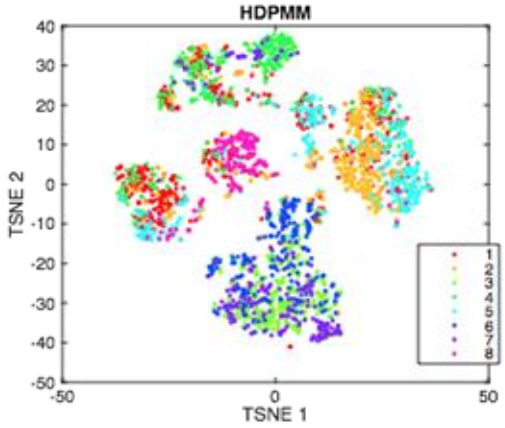
Comparing: Biscuit to other methods



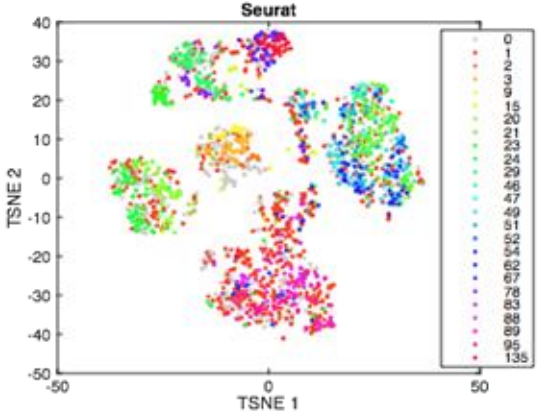
F-score: 0.91



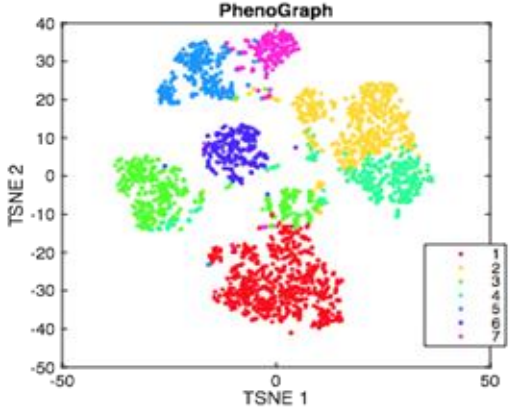
F-score: 0.79



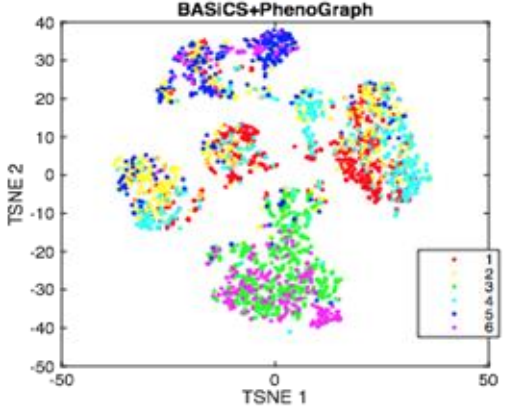
F-score: 0.5 for 67% of cells



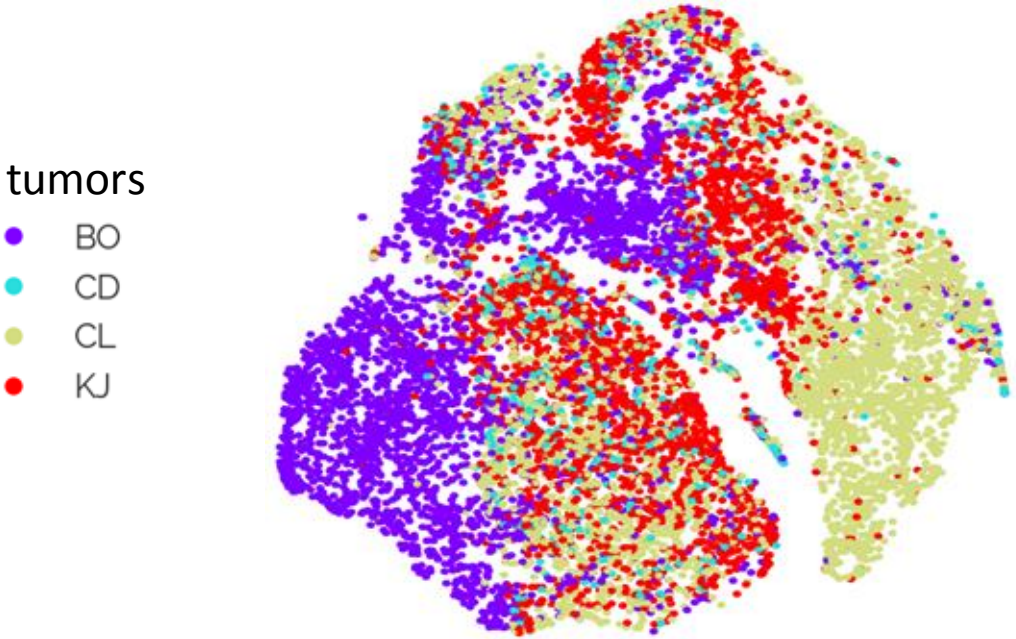
F-score: 0.74



F-score: 0.61



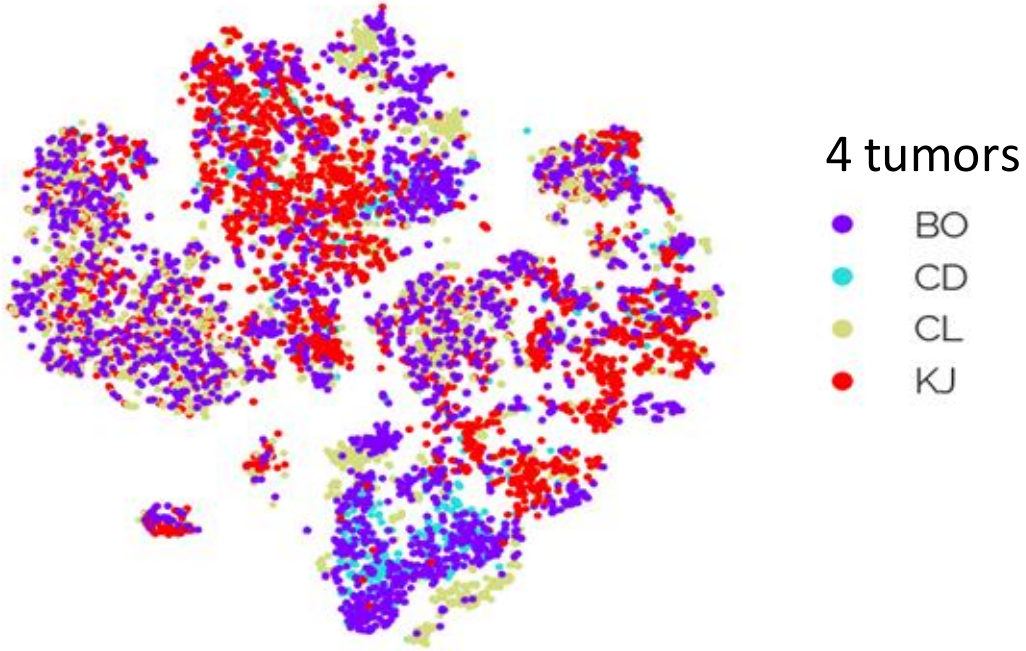
Reminder: Breast TIL data before Biscuit



- Skews data, non-overlapping cells across tumors
- Unclear structure of cell types, mostly distinguishes myeloid from lymphoid cells

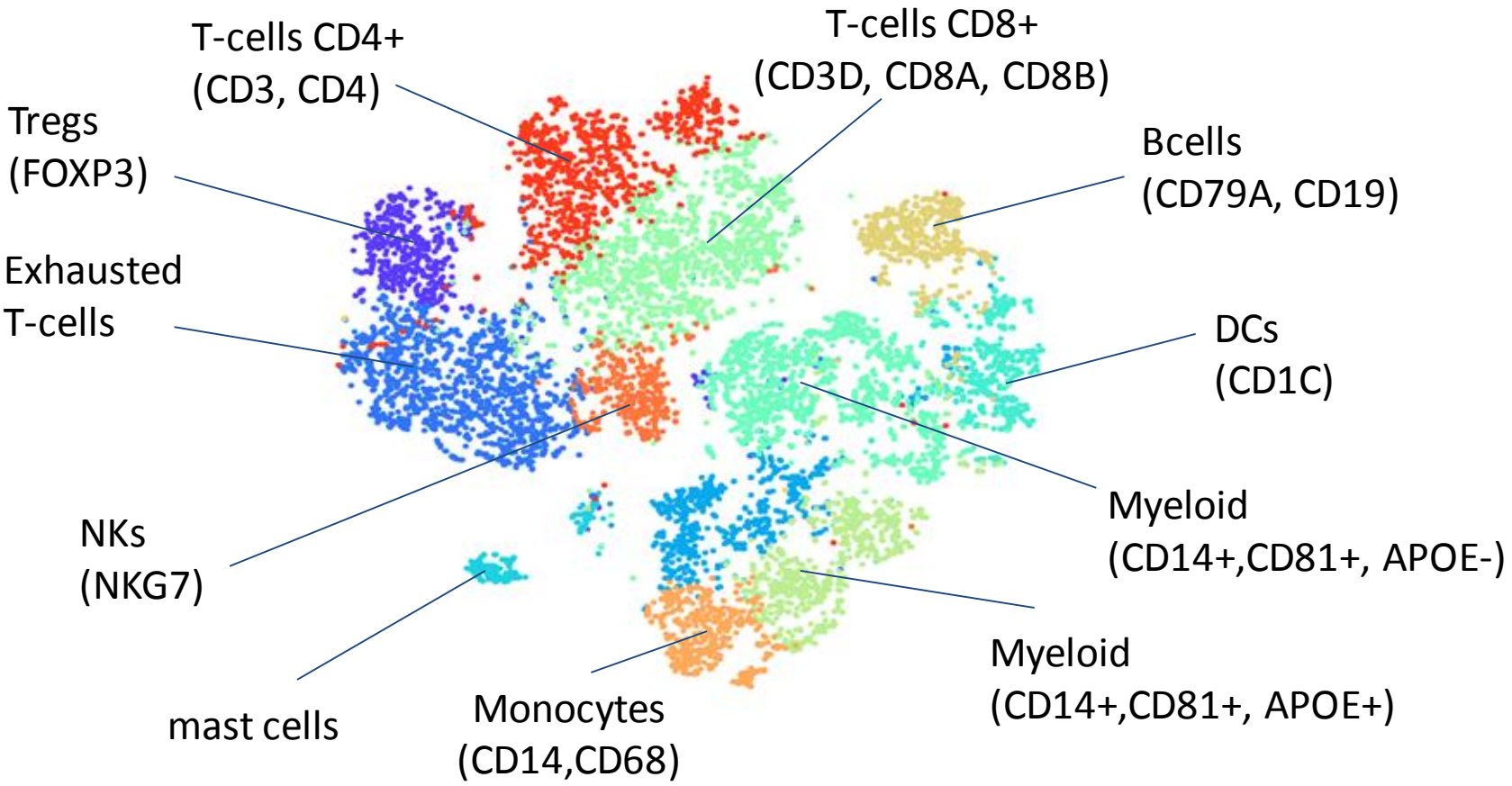
Breast cancer TIL data **after Biscuit**

12,000 Cells, ~3000 molecules per cell



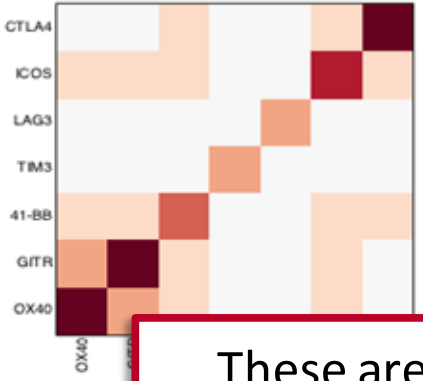
- Most of the tumor specific regions vanish
- Most of the map includes cells from all 4 tumors

Breast cancer TIL data after Biscuit

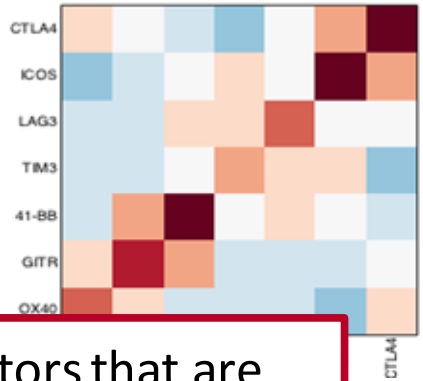


Patient specific differences in co-variation structure

Patient 1



Patient 2

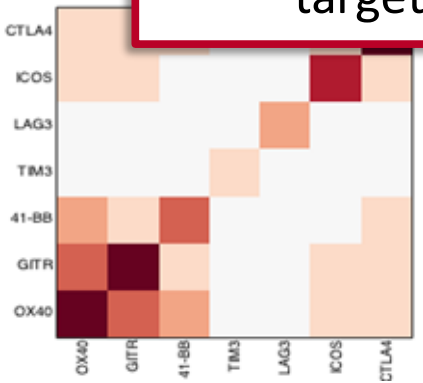


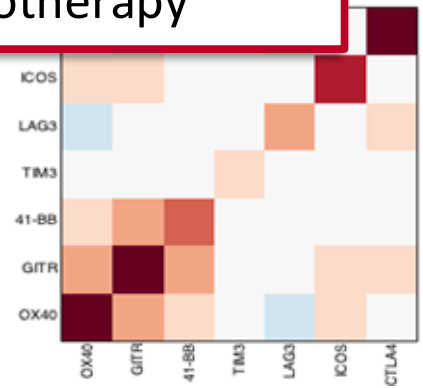
Stronger covariance between ICOS and CTLA4

Weaker covariance between OX40 and GITR

These are the co-receptors that are targeted by immunotherapy

Patient 3

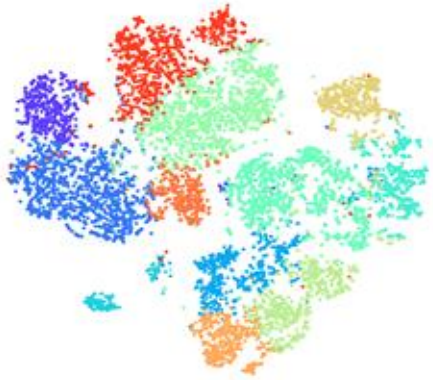




Summary for Biscuit

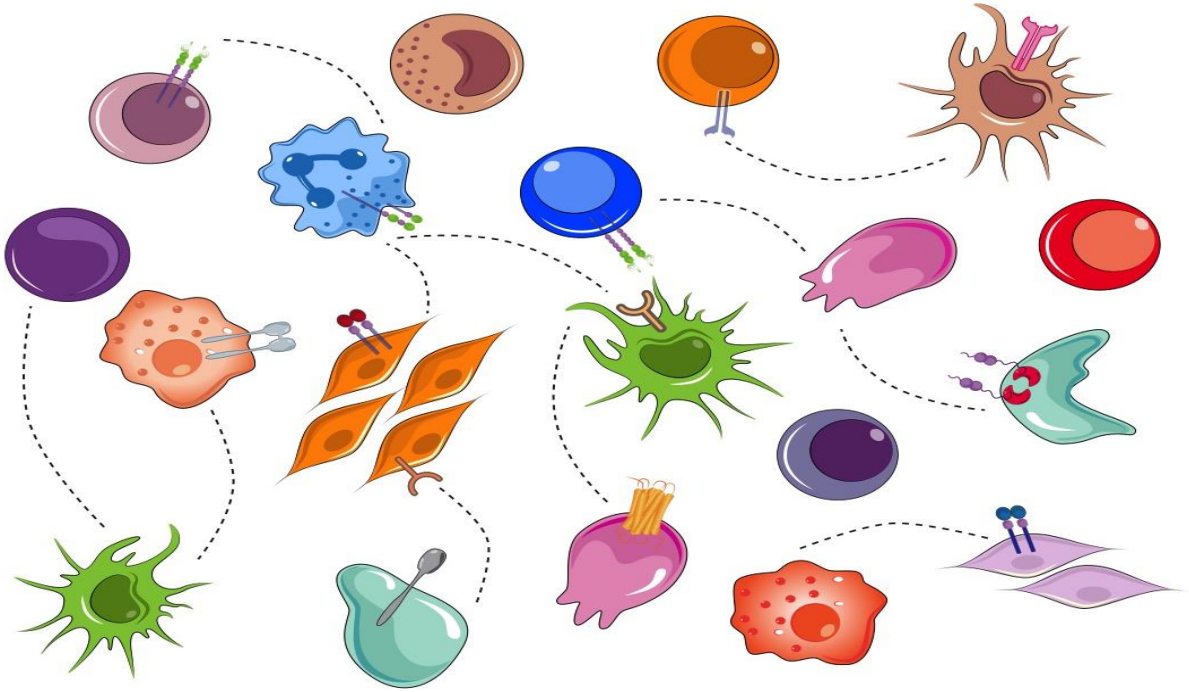
We introduce BISCUIT:

- iteratively clusters and normalizes single-cell RNA-seq data based on different cell types.
- hierarchical Bayesian mixture model with an efficient Gibbs sampler for inferring cell-specific parameters.
- imputes dropout gene expression values.



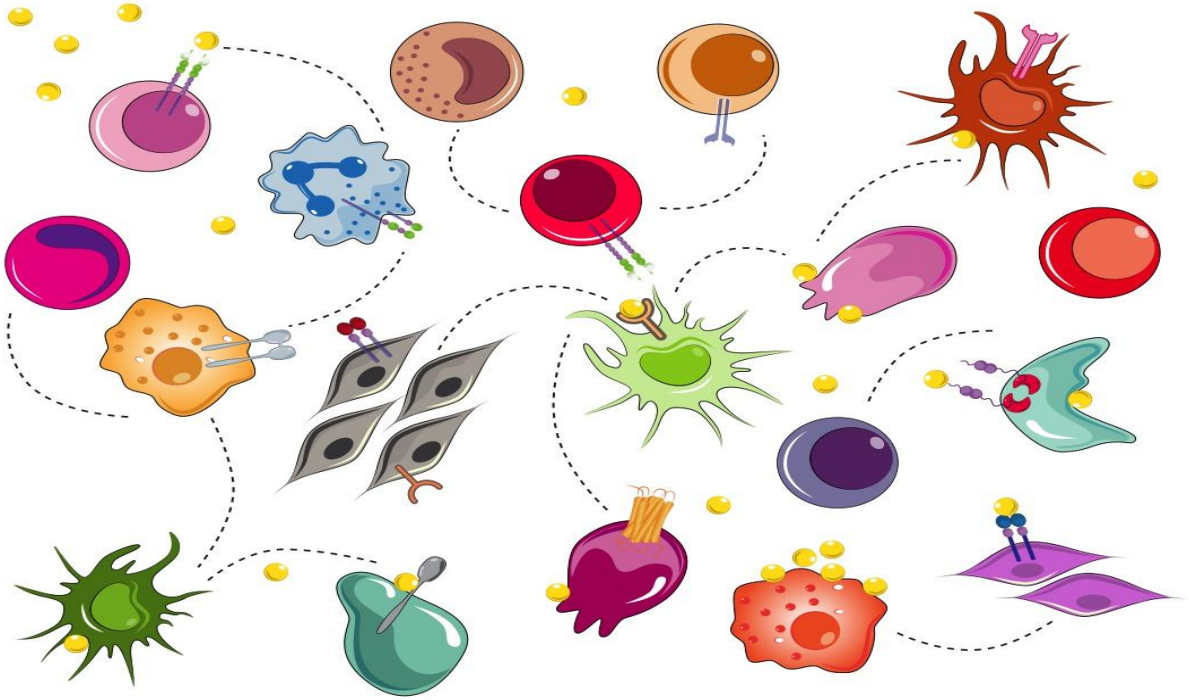
We constructed a cell atlas of the tumor immune s,

- Captured a rich diversity of tumor immune cell types
- Cancer specific differences in co-receptor patterns that can guide combinatorial immunotherapy (releasing multiple breaks).



Once we have the parts we can ask how these interact

- Effect of tissue context, organ site, cancer/healthy?
- What happens in response to drug?



Longitudinal Studies:

- How does the tumor ecosystem change under drug perturbation?
  - both cell state and cell to cell interactions
- How does this differ between responders and non-responders?
- Our measurements are genome-wide, mechanism!!

Acknowledgements

Elham Azizi
Sandhya Prabhakaran
Ambrose Carr
Linas Mazutis
Jacob Levine
Kristy Choi
Josh Nainys
Manu Setty
Vaidas Kiseliovas
Rami Eitan
Zakary Posewitz

Sasha Rudensky (MSKCC)
George Plitas
Casia Konopac

Patients

S2C
NATIONAL CANCER INSTITUTE
NIH DIRECTOR'S PIONEER A-W-A-R-D
COLUMBIA SYSTEMS BIOLOGY
the David & Lucile Packard FOUNDATION
NSF