The power of single cells: Building a tumor immune atlas

Dana Pe’er

Department of Biological Science
Department of Systems Biology
Columbia University
“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

The Precision Medicine Initiative

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

**DNA**

**RNA**

**NETWORK**

**PHENOTYPE**
Cells: Key intermediates from genotype to phenotype

When a genetic association is found:
Which tissue dysfunctions?
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

Skin Cells | Immune Cells | Brain Cells | Muscle Cells
---|---|---|---
Fibroblasts | Megakaryocytes | Neurons | Smooth Muscle
Adipocytes | Dendritic Cells | Ependymal Cells | Skeletal Muscle
Epithelial Cells | T Cells | Astrocytes | Cardiac Muscle
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
Cell phenotypes accumulate in complex non-convex manifolds
A “social network” for cells

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

Levine*, Simonds* et al. Cell 2015
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

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

- **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.**

Need thousands of CD45+ cells per tumor
In-Drop Parallel Processing of RNA-Seq Libraries from >10,000 Individual Cells

- Microfluidic device can do 30,000 cells in one 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

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

- Entire regions on the map are tumor specific
- Are these differences real biology or technical effects?
Single cell RNA-seq as imagined

Count Matrix

2D projection of cells (tSNE)

Cells

Genes

Cluster 1  Cluster 2  Cluster 3
**Problem:** Single-cell RNA-seq data involves significant dropouts and library size variation.

**Observed Count Matrix**

Cells

Genes

Need to impute dropouts & normalize data

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

- Normalizing independent of cell types
  - To mean/median library size
  - BASiC/ERCCs

Clustering Cells

Problems:
- Dropouts not resolved
  - Zeros remain zero!
- Removes biological stochasticity specific to cell type
- Leads to improper clustering
- Biased results in downstream analysis

Observed Count Matrix

Cells → Normalization → Clustering Cells → Downstream Analysis

Cells: Genes
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

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

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

After Normalization

Dropout not resolved

Spurious Differential Expression
**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

![TSNE plot with clusters and cell counts](image)
**Approach:** Simultaneous inference of clusters and imputing parameters

- Parameters characterizing each cluster
- **iterative learning**
- Assignment of cells to clusters
- Cell-specific Parameters for Imputing & Normalization

![Diagram](image-url)
Modeling: Clusters of Cells using a Bayesian Mixture Model

Ideal Count Matrix (normalized)

Cells

Genes

Cluster 1  Cluster 2  Cluster 3
Modeling: Clusters of Cells using a Bayesian Mixture Model

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

Each gene:

Mixture of Log-Normal Models
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

Without Technical Variation

With Technical Variation

Latent counts which we want to recover

\[ Y = [y_1, \ldots, y_{10}] \]

\[ y_j \sim N(\mu_k, \Sigma_k) \]

Observation
BISCUIT
(Bayesian Inference for Single Cell ClUstering and ImpuTing)
BISCUIT
(Bayesian Inference for Single Cell ClUstering and ImpuTing)

Hyper-priors
- \( I, I, \mu', \Sigma', d, \Sigma, \theta \)
- \( \phi, \mu', \Sigma', H, \Sigma, \sigma' \)

Hyper-parameters
- \( \pi_k, \mu_k, \Sigma_k^{-1} \)
- \( \nu, \delta, \omega, \theta \)

Cluster-specific parameters
- \( z_j, y_j \)

Cell-specific parameters
- \( x_j \)

Observed gene expression per cell \( j \)

Global Distributions of Observed Data

Priors set based on observed lib size distribution
Inference Algorithm

Parallel Sampling from derived conditional posterior distributions: $P(\text{parameter} | \text{data, other parameters})$

- Estimate hyper-priors based on Data
- Sample hyper-parameters
- Sampling technical variation parameters
- Sampling cluster-specific parameters
- Sampling assignment of cells to clusters using Chinese Restaurant Process (CRP)
  
  Also allows estimating the number of clusters

Gibbs iterations

scaling mean, cov per cell
Cluster-dependent **Imputing & Normalizing**

[Diagram showing genes, observed data, cells, and BISCUIT clusters]

- Cells
- Observed Data
- Genes
- BISCUIT clusters
Cluster-dependent **Imputing & Normalizing**

[Diagram showing observed data, cells, and BISCUIT clusters]

- **Cells**
- **Genes**
- **Observed Data**
- **BISCUIT clusters**
- **Sort cells into clusters**
Cluster-dependent **Imputing & Normalizing**

- **Cells**
- **Observed Data**

**BISCUIT clusters**

- **Genes**

- **Library Size**
  - **Gene A**
Cluster-dependent **Imputing & Normalizing**

- Cells
- Genes
- Observed Data

**BISCUIT clusters & parameters**

- TSNE 1
- TSNE 2

- Library Size for Gene A
Cluster-dependent **Imputing & Normalizing**

From Observed Data, cells are imputed and normalized using BISCUIT clusters & parameters. This is achieved through a linear transformation.
Cluster-dependent **Imputing & Normalizing**

- **Cells**
  - **Observed Data**

- **BISCUIT clusters & parameters**

- **Impute & Normalize**
  - With a linear transformation

- **Imputed Data**

- **Clusters**
  - Cluster 1
  - Cluster 2
  - Cluster 3

**Gene A**

- Normalized
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: 0.74

F-score: 0.61

F-score: 0.5 for 67% of cells
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

- Tregs (FOXP3)
- Exhausted T-cells
- NKs (NKG7)
- Monocytes (CD14, CD68)
- Mast cells
- Bcells (CD79A, CD19)
- DCs (CD11C)
- Myeloid (CD14+, CD81+, APOE-)
- Myeloid (CD14+, CD81+, APOE+)
- T-cells CD4+ (CD3, CD4)
- T-cells CD8+ (CD3D, CD8A, CD8B)
- CD4+ T-cells
- CD8+ T-cells
- NKs
Patient specific differences in co-variation structure

Stronger covariance between ICOS and CTLA4

Weaker covariance between OX40 and GITR

These are the co-receptors that are targeted by immunotherapy
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 system:

- 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
Columbia Systems Biology
NIH Director’s Pioneer Award
David & Lucile Packard Foundation
NSF