Leveraging large-scale single cell RNA-seq data to understand collective cellular behavior in the human immune system

Word count: 2486

Summary:

In the human immune system, high-level physiological functions, such as eradicating foreign bacterial or viral pathogens, emerge from collective cell behavior [Kidd et al., Stanisic et al.]. A better understanding of how immune system functionality emerges from a heterogeneous population of interacting cells can lead to more precise diagnoses and more effective treatments. However, current methods analyze cellular behavior within isolated cellular populations or at a single scale [Davis et al.]. Historically, a lack of data compounded with the sheer complexity of the human immune system has prevented researchers from developing a global model that is rooted in single cell behavior.

Due to massive growth in single cell measurement capabilities, we have new tools to study the complexity across populations and scales in heterogeneous tissues [Shapiro et al., Zheng et al.]. With single cell RNA-sequencing (scRNA-seq), we can simultaneously query the transcriptional state of each individual cell and the abundance of each cellular subpopulation. This data allows us to connect single cell changes to tissue-level processes, providing a new lens to better understand how cells in the immune system carry out global functions. My goal is to build new, multiscale models of the immune system that capture how diseases engage, disrupt, and evade the immune system from the cellular level up.

Comparing scRNA-seq profiling of tissue samples from different time points or patients will help us understand how single cells work together to execute immune functions and how the system breaks down in disease. Yet, this analysis will require a new scale of data, thousands of samples each consisting of thousands of individual cells and and new tools that will operate over vast quantities of high dimensional data.

In my postdoc, I will extend ideas developed in my PhD and gain expertise in machine learning, signal processing and distributed computing in order to leverage large-scale scRNA-seq data to understand how cells coordinate each of these responses through three aims.

1. **A vocabulary for describing cellular behavior: interpretable, minimal, and comprehensive**
2. **Efficient, scalable, and dynamic pipeline for scRNA-seq sample comparisons**
3. **Multi-scale immunology: identifying covariances across distinct cellular populations**

Past research: Exploiting low dimensionality in gene expression

During my PhD, I have applied a central idea from signal processing to develop methods for denoising data and analyzing scRNA-seq data. Gene expression, similar to natural signals such

as audio and images, has an inherent low dimensionality to it. While natural images can be broken down into a combination of a relatively small number of wavelets, gene expression can be well represented as a linear combination of “transcriptional programs” or groups of covarying genes.

My work focused on developing analytical methods that exploit the inherent low dimensionality in gene expression for two specific applications: enabling inexpensive, higher throughput RNA-seq measurements and developing methods to compare heterogeneous tissue samples.

Exploiting low dimensionality in gene expression to enable massively multiplexed experiments

RNA-seq users are faced with a choice of how to allocate limited sequencing resources. They are faced with a tradeoff between sequencing depth and sample throughput (Figure 1). To characterize this tradeoff and understand how increasing measurement throughput affected information that could ultimately be extracted from the data, we derived an analytical model that identified which variables are important in accurately computing transcriptional programs. Our model revealed an unintuitive quantity that enabled extraction of useful biological information even when sample multiplexing was increased 100-fold.

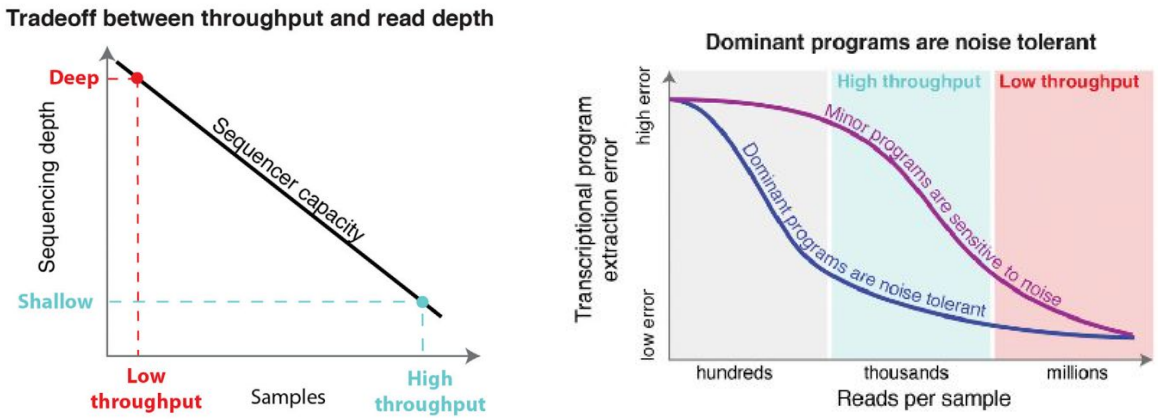

***Figure 1. Left**: Due to the fixed capacity of sequencing machines there is a tradeoff between sample multiplexing and sequencing depth. **Right**: Dominant transcriptional programs, which capture much of the total variance within a gene expression matrix, are accurately extracted at low sequencing depths while transcriptional programs that capture less variance are extracted with higher error at the same depth. [Heimberg et al. 2016]*

We used a perturbation theory model to identify how sampling noise affects the gene expression programs, such as principal components, extracted from a dataset. The analytical framework identified the spectral properties of the data, the eigenvalues of the gene expression covariance matrix, to directly determined how accurately the transcriptional programs could be extracted. The more total variance captured within a single transcriptional program, the less sequencing depth required to reliably extract the transcriptional program (Figure 1).

Simulating sequencing noise on gene expression datasets revealed that in many cases you could increase sample measurement throughput 100-fold. Out of the ~350 gene expression datasets tested, the vast majority of the variance of each one was captured within a few transcriptional programs and at 1% of conventional sequencing depths. By combining the analytical model with typical parameters found from surveying public data, we accurately predicted the sequencing depth to recover the transcriptional programs that represent even minor changes in gene expression.

Impact: Our analytical framework identified the factors which determine the appropriate RNA-seq multiplexing throughput for a desired analytical resolution.

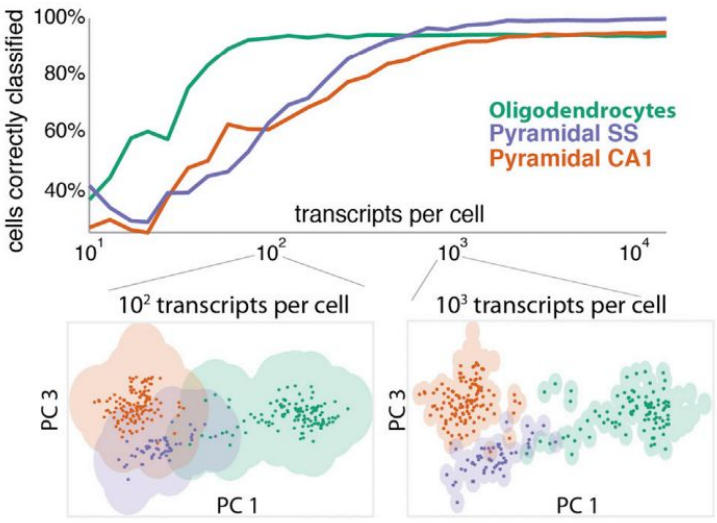

***Figure 2. Top**: Cell type classification accuracy in neurons as a function of transcripts counted per cell with scRNA-seq. **Bottom**: Cell types are classified by regions in principal component space. At low depths, many cells are incorrectly assigned to regions (denoted by overlap of colors). At 1000 transcripts per cell, the precision in mapping cells into principal component space is sufficient for high accuracy classification. [Heimberg et al. 2016]*

Enabling comparisons between patient samples with scRNA-seq with new analyses

scRNA-seq provides a powerful lens to study physiological processes taking place in heterogeneous tissues between patient samples. However, due to the complexity and multi-scale nature of immune system, we currently lack the methods to compare macroscopic properties of the immune system across samples. My current project is to develop computational tools that enable us to identify which biological processes differ in specific cellular populations.

Using scRNA-seq to compare patient samples comes with challenges from measurement noise, high dimensionality, and a lack of mapping between cells from different samples. To overcome these challenges, we developed a probabilistic “feature space” analysis to construct a statistical

portrait of the immune system (Figure 2). We first map cells to a common low dimensional space using a dictionary of gene expression features collected from thousands of publicly available gene expression profiles. Once in a common low dimensional space, we can apply machine learning methods, such as principal component analysis, random forests and neural networks to find regions in feature space that are enriched for different patient samples.

When we applied this analysis to publicly available patient samples containing bone marrow samples from Acute Myeloid Leukemia (AML) patients [Zheng et al.], we were able to find patient specific signatures of cancer initiation and cancer progression. Our analyses found that in one patient, a population of red blood cell precursors had dramatically expanded and each cell shared a nearly identical gene expression profile, most likely as a consequence of a cancerous clonal expansion. Many of these cells had turned off expression exterior protein identification markers, Human Leukocyte Antigens, helping these cancerous cells to evade immune detection [Hicklin et al.]. In another patient, their hematopoietic stem cells were not terminally differentiating into appropriate myeloid cell types. Only by linking cellular population counts with each cell's' gene expression profile, were we able to extract biological insight on the differences of these two AML patients.

Impact: The feature-space framework will be able to take on these data-intensive questions because will be able to scale well with the data as it does not require any large matrix decompositions and will be amenable to distributing its most computational intensive steps.

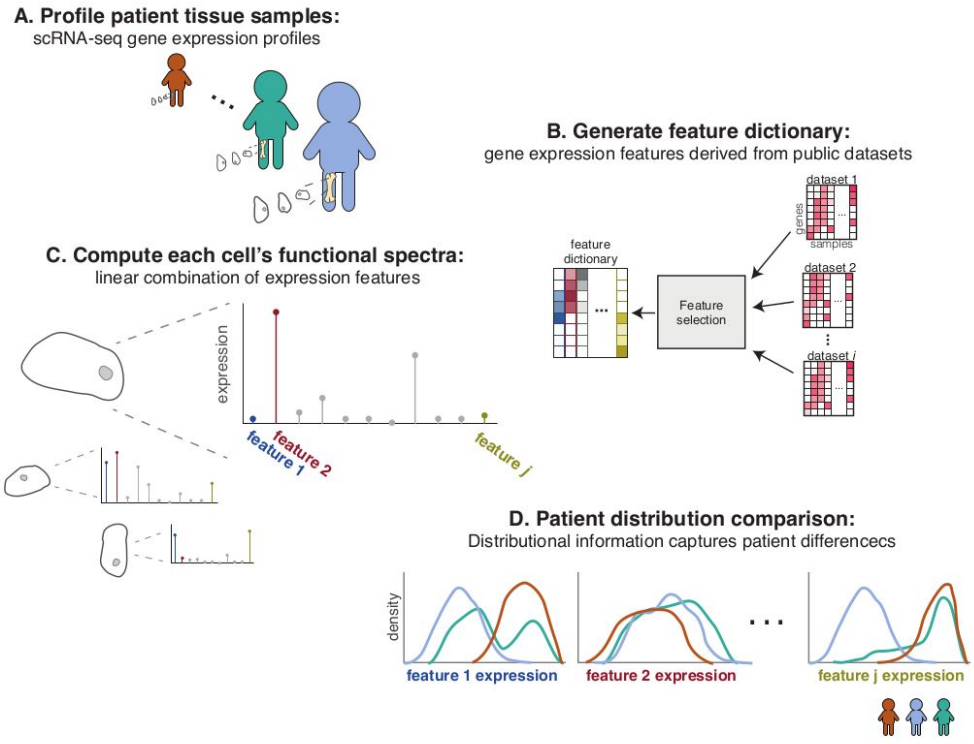

Figure 3. A) *A series of scRNA-seq datasets sampled from human bone marrow.* ***B)*** *A feature dictionary is generated by extracting and collecting features from hundreds of individual gene expression datasets.* ***C)*** *Individual cells are projected into a subspace of the feature dictionary. The features which determine*

the subspace is are identified through a feature selection step. ***D)*** *Donors have unique distributions of feature expression levels of all of their cells.*

Future research: Towards multiscale modeling of the human immune system

Researchers have been isolating populations of cells and searching for differences between patients that explain disease heterogeneity for decades. While some of these efforts have been successful, variability in patient response to treatment remains unexplained. In Rheumatoid Arthritis, for example, different drugs target different aspects of the immune system, yet we are unable to predict which ones will be effective for a patient. By studying drug response in the context of the collective behavior of the immune system, we may be able to gain broader understanding of what makes these drugs effective for specific patient populations. In my post-doc, I will generate systems level models of the immune system to identify predictors for patient responses and disease status.

scRNA-seq now enables us to characterize cellular behavior within heterogeneity within tissue samples. My previous research serves as a proof of concept, demonstrating a method that is able to automatically identify differences between individuals. Scaling the analysis with data will allow us to address new questions, such as how cellular populations in the immune system coordinate gene expression to perform sophisticated tasks.

As publicly available scRNA-seq data accumulates, the scope and scale of data will be sufficient to generate systems level immune models for interpreting variability in health and disease. To generate such models, I will describe the complete vocabulary of cell's transcriptional states, create a public resource of aggregated patient samples for comparisons, and analyze cellular populations across patients to see how coordinated activity between immune cell types affects disease status and patient outcomes.

1. A new vocabulary for describing cellular behavior: interpretable, minimal, and comprehensive

As developed in my PhD work, in order to analyze scRNA-data from multiple patient samples, we compare them in a common reduced dimensionality subspace. We select a small set of vectors from a gene expression feature dictionary to generate a low dimensional approximation of the data. In Aim 1, I will create a dictionary of gene expression features, with particular focus on three aspects that are fundamental to making this dictionary useful for downstream analyses.

Comprehensive: To generate a complete dictionary, I will extract transcriptional features from massive public gene expression repositories such as Gene Expression Omnibus [Edgar et al.] and the Cell Atlas projects [Regev et al.]. Some of these features will represent cellular identity,

allowing me to break out individual cellular populations by conditioning on these feature levels, while others will represent transient transcriptional processes.

Interpretable: As data scales, the richness of interpretation of results is paramount. I will compare the top genes in features to curated gene lists, searching for known functionalities [Subramaniam et al.]. By mining experimental metadata from genetic and chemical perturbation studies, I will be able to add functional information to each transcriptional program.

Minimal: To enable unambiguous downstream interpretation, I will remove redundant and uninformative features. I will downsample dictionary features, to identify which features can be removed without introducing error into the low dimensional representation of each cell's gene expression profile. Only features that represent a large amount of variance within other experiments will be kept in the dictionary.

Impact: Instead of re-interpreting the results for each new dataset, we can employ this universal cellular vocabulary to describe each cell's gene expression laying the foundation for new, large scale analyses of scRNA-seq of heterogeneous tissues.

2. Efficient, scalable, and dynamic pipeline for scRNA-seq sample comparisons

Because each scRNA-seq dataset is 10^4-10^5 times bigger than one from bulk RNA-seq, new technical challenges arise. In Aim 2 I will develop a computational pipeline which will enable me to use thousands of public scRNA-seq data in my analysis. One of the challenges will be in efficiently implementing these algorithms, scaling smoothly when I need to distribute computation, and dynamically incorporating new data as it becomes available.

Data: Through an established collaboration with the Marson and Ye labs here at UCSF, we will collect our own patient samples. I will complement this data with published data from other research labs. If growth of scRNA-seq datasets is similar to microarrays, public repositories will grow to contain tens of terabytes of processed data within the next few years.

Algorithm: The algorithm that we have already developed projects all cells into a common low dimensional space. To find a suitable representative space we use LASSO [Tibshirani], to find a small set of feature vectors from the dictionary which accurately represent all single cells. This reduced data will contain a features by cells matrix for each sample.

Implementation: Each cell's representation can be handled independently. Therefore, this task is highly parallelizable. As the data grows, we can use Spark, MPI or other platforms to scale data processing. Once reduced to a low dimensionality, the data can be stored in memory and hosted on a server for researchers' benefit.

Impact: A pipeline to reduce the large and high dimensional data into space small enough to be stored into memory will allow for further analyses to be run. Once aggregated, the data will become available to the community as a useful resource for their own studies.

3. Multi-scale immunology: identifying covariances across distinct cellular populations

In the immune system cells physical interactions and cytokine signalling between cells are essential for proper immune function. Using the low dimensional representations of individual cells established in Aim 1 and 2, I will train machine learning models to identify combinations of gene expression features that covary across cell types as a result of these interactions and study how they behave in disease.

Machine learning model: I will use machine learning models to identify how distributions of cellular features within populations combine to affect the immune response to disease and drugs. After projecting each cell into a low dimensional space, I can compile feature distributions for each individual patient for each cell type. I will feed these into decision trees and neural networks and train them to predict disease activity scores and whether a treatment was effective for a patient.

Interpretable models such as decision trees can be analyzed to understand how gene expression features between cell types are combined to have predictive power. Insight gained from the models can be used to identify combinations of gene expression markers to be used as complex biomarkers. Conversely, neural networks performance, will most likely be better and therefore may be useful as a diagnostic tool for patients.

Validation: To validate this model I will use data from patients who have received treatment for the disease. I will test whether in response to drugs, the patient samples have gained or lost key covariates that are predictive of disease activity. This will help us understand the effect that drugs targeting specific gene expression pathways have on the immune system as a whole.

Impact: Identifying modes communication between cellular populations are known within the immune system and understanding how these change in disease will provide new insight into how the immune system is overcome in disease and can lead to new biomarkers for diagnosis and patient stratification.

References:

1. Davis, Mark M. Systems immunology: just getting started. Nature Immunology, 2017
2. Edgar, Ron et al. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 2002
3. Heimberg, Graham et al. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. Cell Systems, 2016
4. Hicklin, Daniel J. et al. HLA class I antigen downregulation in human cancers: T-cell immunotherapy revives an old story. Molecular Medicine Today, 1999
5. Kidd, Brian A et al. Unifying immunology with informatics and multiscale biology. Nature Immunology, 2014
6. Regev, Aviv, et al. The human cell atlas. BioRxiv, 2017

7. Shapiro, Ehud. et al. Single-cell sequencing-based technologies will revolutionize whole-organism science, 2013
8. Stanisic, Danielle I. et al. Escaping the immune system: How the malaria parasite makes vaccine development a challenge. Trends in Parasitology, 2013
9. Subramaniam, Aravind et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 2005
10. Tibshirani, Robert. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B, 1996
11. Zheng, Grace et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications, 2017