Driving Project 3: Asthma pathway modeling for understanding severity and heterogeneity

[[DW: figures are https://docs.google.com/presentation/d/1N-7SazJH9Gn7qCVOjn7yic3RJHBKo8vI5SixlbpFzGc/edit?usp=sharing]]

Alt1: Specific molecular interactions between the innate (YKL-40) and adaptive (IGA to microbiome and IGE responses) immune systems underlie the specific endotypes of asthma

Alt2: Understanding asthma heterogeneity and severity through modeling of the interactions between the innate and adaptive immune responses

Table of Contents

1. Specific Aims

We aim to use RNA sequencing and CyTOF experiments performed by the Precision Profiling Core to develop an integrative model of asthma to better understand key aspects of its heterogeneity and differential severity. [[In particular, we will model the effects of IgA and IgE-mediated adaptive immunity responses in a cell-specific manner with emphasis on effects of YKL-40 and DKK1 levels.--should delete, too much overlap with other projects(DS)]] These analyses will yield mechanistic insight to the transcripts and pathways associated with asthma lung dysfunction clusters which is necessary for translating these findings to patient care. To this end, we will use our expertise in RNA sequencing to develop and distribute pipelines to the Precision Profiling Core for the processing of bulk-cell and single-cell RNA-Seq and CyTOF data. These cleaned and uniformly-processed data will be clustered, built into regulatory networks, integrated with external datasets and patient clinical data, and modeled. This analysis will give a new understanding of the regulatory pathways by which patients' asthma experiences vary, define the features that best correlate with clinical outcomes, and speak to the underlying mechanisms by which those outcomes occur. [[delete(DS)-- Further, these models will inform, direct and iteratively learn from experiments performed by the other Driving Projects in this proposal to identify the pathways that lead to asthma heterogeneity and severity.]][[DS/TG: send around pic,,, interface diagram & para on interfaces]]

[[MG: need to say that the point of this project is generate many different clusters of genes that characterize diff. patient groups. Our hypothesis is that different endophenotypes (which are evident in the different clusters) characterize different patient group (diff types of asthma)]]
[[enrich our knowledge of this disease, generalize to a systems level, identify new features]]

Aim 1: Develop a bulk RNA-seq processing pipeline and cluster transcripts

We will adapt a comprehensive suite of human RNA-Seq tools to generate pipelines for the uniform processing bulk-cell RNA-Seq data. This will build on a considerable body of preliminary results that we have from developing human RNA-Seq pipelines, both for long and short RNA. We will create a workflow to quantify transcript abundances, determine the degree to which they have been spliced and modified, see the extent to which they correspond to annotated portions of the genome, as well as identify non-coding RNAs and transcribed pseudogenes. These pipelines will be passed to Core C for use in generating a uniformly-processed dataset for use by each of the Driving Projects in this cooperative agreement. We will use these processed data to generate co-expression clusters and networks and compare their performance at stratifying patients by disease phenotype to established methods, including clinical measurements (e.g. FEV1 and FeNO) and TEA clusters. These clusters will be made available to the other Driving Projects for use in selecting targets for pathway analyses and will be refined using the data generated by the other groups. [[MG: move clinical stuff go into aim 3]]
[[create BC clusters]]

Aim 2: Single-cell Analysis of Asthma Sputum

We will utilize single-cell measurements of protein and mRNA abundances using (1) CyTOF and (2) RNA-sequencing technologies: (1) CyTOF enables the measurement of signaling and surface marker molecules from which we will employ an unsupervised community detection method to deeply characterize rare and novel populations of cells produced in the airways of asthmatic patients. Further, we will develop a method to match such results across samples such that the populations are validated through repeated detection across patients. Building on our previously developed information theoretic techniques DREMI and DREVI, we will also characterize signaling relationships between proteins and cytokine responses in subpopulations of sputum cells and give that as input to the integrated logic model of Aim 3. (2) Single-cell RNA sequencing is a newer technology that theoretically capable of giving measurements of the entire complement of expressed genes within a cell. However, current single cell technologies suffer from a high degree of technical noise due to mRNA loss during sampling, cell-to-cell variations in sequencing efficiency and amplification biases. We propose to develop a pipeline that quantifies technical variability for each gene and converts from raw reads to gene counts in a biologically meaningful manner. This processing pipeline will be handed to the profiling Core in order to process data from Projects 1 and 2. After processing, we will employ our previously developed dimensionality reduction methods to reduce

the data into a few robust dimensions, and cluster the result into metagenes corresponding to pathways. These metagenes will subsequently be analyzed by DREMI and DREVI in order to characterize novel pathways and their interactions in the variety of cell populations present in the sputum.
[[create SCP, SCG, SSC clusters]

Aim 3: Integrative clustering & interfacing w other projects [[DS]]

We will use the single-cell RNA-seq to identify cell-type signatures of sputum from patients with varying asthma severity. This will allow for more precise cell-sorting to identify patterns of cell enrichment or depletion that are associated with clinical data. Moreover, these cell-type signatures that can be used to deconvolve the bulk-cell RNA-seq data to its component cell transcripts, increasing the effective dynamic range of the cell-specific transcriptional data and facilitating integration with the abundance of bulk-RNA-seq datasets. These data will be used to refine clusters and networks from Aim 1 and combined with the clinical data to identify the precise, cell-specific transcriptional program associated with clinical measurements and TEA clusters. In addition, these data will be integrated with existing datasets (e.g. GTEx for tissue-specific context and ENCODE for transcription factor data) to build regulatory networks that will model the signaling activities of the cells in the sputum. The regulatory networks will be refined by the CyTOF signaling networks from Aim 2 and a regulatory logic gate model created. This model will give unparalleled definition of cellular responses and pathway dysfunctions that are associated with patient phenotypes and identify optimal targets for manipulating cellular and patient responses.
[[MG: mention the integrative clusters in aim 3, make a table 3, wehre we list all the clusters & we say we’ll put tehm on the a website]]

2. Significance [[KKY fix of signif & background .75 to 1pg]]

Asthma is a chronic inflammatory disease of the airways which afflicts ~7% of the U.S. population \cite{21430629}. In most individuals, symptoms are easily controlled by treatment with bronchodilators and relatively low doses of inhaled corticosteroids, but as many as 30% of asthmatics do not respond adequately to standard therapies, and approximately 5% of asthmatics have a severe, refractory form of the disease. Previous work used a novel hierarchical clustering approach to identify three transcriptional endophenotypes of asthma (sputum TEA clusters) that successfully stratified patient phenotypes including the amount of airway inflammation (by FeNO), lung function and cytokine levels in the airways. This represented the first non-invasive stratification of asthma disease severity with the potential to successfully identify high-risk patients and reduce hospitalization. However, the TEA clusters did not have the resolution or dynamic range to elucidate molecular mechanisms by which the individuals responded differently.

The goal of this project is to expand our horizon by characterizing asthma is a broader and deeper context. To arrive at this goal, we plan to perform RNA-Seq, as well as state-of-the-art single-cell technologies including single-cell RNA-Seq and CyTOF, on a cohort of patient samples. RNA-Seq is a new but established technology for genome-wide transcriptomic analysis. It has been widely applied for understanding various diseases such as cancer. Transcriptomic analysis using RNA-Seq enables one to discover gene clusters responsible for common functions, as well as to identify novel transcripts with the same functions. The data will widen our current knowledge on asthma from a few specific pathways to a system-wide level. On the other hand, single-cell technologies can greatly expand upon the sensitivity and cell-type specificity of asthma research. On the transcriptomic level, single-cell RNA-Seq offer an unbiased measurement of the entire collection of mRNA transcripts produced in each cell. Despite its inherent sparsity, it complements bulk RNA-seq in characterizing the heterogeneity among different cell types. A promising approach for discovering new cell types is to perform unsupervised clustering on single-cell RNA-Seq data. Performing both bulk RNA-Seq and single-cell RNA-Seq will therefore synergize our research. Another single-cell technology we are going to employ is CyTOF. CyTOF complements RNA-Seq because it provides information on a proteomic level. Currently, CyTOF data consists of dozens of dimensions (around 45 presently) of protein abundance measurements. It allows us to examine signaling responses within minutes and hours of exposure to antigen, for instance dust-mites. CyTOF is able to probe the response of various cell types and provides a dynamical element to our study.

In summary, the methods described in this project which will more clearly define the molecular mechanisms that drive severe disease, and therefore help us to characterize the endophenotypes of asthma. These data are critical to pass beyond reactive medicine common in asthma treatment to more personalized methods with specific cellular mechanisms that can be targeted for therapies.

3. Innovation

Recent efforts have shown that the complex and heterogeneous disease that is asthma can be sub-typed into categories using microarray expression data. This work goes several steps further, not only offering transcriptional clustering with unprecedented sensitivity and dynamic range, but does so in a way that will likely offer mechanistic insight and novel therapeutic targets. By using single-cell techniques to interrogate the transcriptional and signaling responses this work will give resolution to observe the activities of cells and how that is perturbed in severe disease. The data will be integrated into a model that has the potential to bring personalized medicine to asthma care.

5. Research Plan

5.A. Plan for Aim 1

5.A.i Rationale

By uniform samples processing and extensive genome wide data integration, we aim to develop a resource for identifying novel asthma-related genetic elements. This genome wide knowledge base of asthma sample transcription will provide a necessary foundation for single-cell sequencing and protein methods in Aim 2 and the modeling of dynamics and response of specific pathways in Aim 3.

5.A.ii Preliminary results

5.A.ii.a Application of RNA-seq processing tools

We plan to build a RNA-seq processing pipeline based on a software suite, RSEQtools, we have largely developed in the past. We have developed a number of tools and data formats to handle the increasingly large quantities for data generated by RNA-Seq experiments. For example, we have developed the Mapped Read Format (MRF), a compact data summary format for short, long and paired-end read alignments that enables the anonymization of confidential sequence information. RSEQtools use this format for the analysis of RNA-Seq experiments. \cite{21134889}. These tools consist of a set of modules that perform common tasks such as calculating gene and exon expression values, generating signal tracks of mapped reads and segmenting that signal into actively transcribed regions. RSEQtools is implemented in C and the source code is available at http://rseqtools.gersteinlab.org/. Along with RSEQtools, we have developed three different RNA analysis pipelines: FusionSeq for fusion transcript detection \cite{20964841}, IQSeq for transcript quantification \cite{22238592}, and DupSeq for analyzing expression patterns of highly homologous genomic regions \cite{25157146?}. All three have been compiled into the RSEQtools framework for customizable workflow.

5.A.ii.b: Non-coding RNA and pseudogene analysis

A fraction of the transcription comes from genomic regions not associated with standard annotations, representing 'non-canonical transcription'. A class of non-coding transcripts of particular interest is the pseudogene. Although pseudogenes have long been considered as nonfunctional genomic loci, recent studies have shown that pseudogenes may serve as useful biomarkers to distinguish different cell types. Despite their low abundance, pseudogenes and ncRNAs have been shown to exhibit a greater degree of cell-type specific expression than mRNAs \cite{25157146} and are therefore useful in several aspects of this study, including the assignment of single-cell RNA seq cell type in Aim 2. In addition, pseudogenes have been shown to perform regulatory roles in in cancer, X-chromosome inactivation and intercellular signaling \cite{??}, and therefore should be taken into account for a regulatory model of asthma, as we will produce in Aim 3. However, the quantification of pseudogene expression is challenging because of the sequence similarity with the pseudogene's of parent genes. To address the issue, we developed DupSeq, which solves this problem by focusing only on those reads and regions that are uniquely mappable \cite{25157146?}.

Several other classes of non-coding RNAs have been shown to play regulatory or other roles in the cell. To identify these loci we will apply incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a gold standard training set and a minimum-run–maximum-gap algorithm to process reads mapping outside of protein-coding transcripts, pseudogenes and annotated non-coding RNAs \cite{21177971, 25164755}.

5.A.ii.c: Functional annotation through clustering and network analyses

We have extensive experience in characterizing the functions of genes and non-coding elements via expression data through clustering and network analyses. One of the important way to understand expression data is clustering analysis. A group of genes in a co-expression cluster are in general presumably responsible for a common function. While there are well known algorithms for expression clustering such as hierarchical clustering, spectral clustering and K-means, we developed several novel methods. In the microarray era, we developed a spectral biclustering method for co-clustering genes and conditions. More recently, we developed a new clustering framework, OrthoClust, for simultaneously clustering network data across different contexts \cite{25249401}. OrthoClust is able to identify conserved and specific components across different networks. We applied OrthoClust in the comparative transcriptome analysis, and discovered co-expression modules shared in animals and enriched in their developmental genes. Furthermore, expression clusters can be used for annotating functions of unknown transcripts. For example, in modENCODE analysis, by mapping the expression profiles of various ncRNAs to expression clusters, we have used identified functions various ncRNAs.

The functional relationships between co-expressed genes can further be understood in terms of various molecular networks. Over the past decade, the Gerstein lab have developed a number of tools to analyze the organization and structure of biological networks. We have identified many relationships between topological properties of genes in networks and their functional genomics features. For instance, we identified that a node's tendency to act as a hub or bottleneck with various forms of "essentiality" (i.e., the degree to which a given node is essential for various functions in a network) \cite{15145574, 17447836}. Another important topological feature is the so-called network hierarchy, which is essentially the direction information flow in these networks. We found that gene-regulatory networks are composed of hierarchical structures dominated by downward information flow and that some TFs act as top master regulators to govern the transcription of downstream TFs. We developed methods to determine the hierarchical organization of regulatory networks and applied them to analyze the regulatory networks of a variety of species from yeast to human, including networks constructed from ENCODE, modENCODE and MCF7 data \cite{25880651,22955619,22125477,21177976}. In addition, we introduced a framework to quantify differences between networks and found a consistent ordering of rewiring rates of different network types. \cite{21253555}.

5.A.ii.d : RNA-seq pipeline development for large-scale projects

In this project, we purpose to do process a cohort of RNA-Seq samples. We have worked on the development and analysis of multiple RNA-Seq flows, including tools we developed as well as other popular tools such as Bowtie and Tophat, in the context of large consortium. For example, we have been playing a role in such activities for the ENCODE consortium \cite{17568003}, including a recent publication involving the processing and integration of all ENCODE and modENCODE data, which involved 575 experiments and more than 65 billion reads from three organisms. \cite{25164755}. We are the data integration hub in the exRNA consortium that generates hundreds of RNA-Seq and small RNA-Seq samples. Other notable consortia for which we have processed large quantities of data include the BrainSpan project (http://www.brainspan.org/) which collected RNA-seq data for 8-16 brain structures in each of 13 developmental stages \cite{24695229}, as well as the PsychENCODE project (http://psychencode.org/) and Extracellular RNA (http://exrna.org/).

5.A.iii Approach

5.A.iii.a Process all the RNA-Seq data in a uniform fashion

A critical component to projects that involve a large number of samples sequenced over time is the uniform processing of the data. This is particularly true in cases where clustering will play a role in a generation of conclusions, as in this project, as it is here that batch effects and sample processing variation can drive artificial organizations of the data. We will process bulk RNA-Seq samples in a uniform fashion using the RSEQtools pipeline that we developed, and where appropriate we will combine this with tools like Tophat and Cufflinks. These tools and pipelines have been used extensively by large consortia \cite{25164755,Rseqtools figure }.

Briefly, sequencing reads with quality scores are mapped to references using several alignment algorithms. The mapped reads are converted to a format that facilitates anonymization and are then processed through a variety of tools including the assembly and quantification of transcripts, generation of sequence tracks and annotation. In addition to so-called standard gene annotation, such as we performed for the GENCODE project \cite{22955987}, other features such as functional RNA structures can be annotated using our tools \cite{17568003}. Moreover, this process is iterative, in that the exon transcripts are re-aligned to more accurately quantify different gene isoforms. As the components of RSEQtools can be readily assembled and extended to build customizable RNA-Seq workflows additional components like single cell analysis developed in aim 2 as well as sample deconvolution developed in aim 3 can be easily incorporated into the pipeline. This pipeline can be easily ported to the core for the universal processing of the data through Yale's dedicated next-generation sequencing dedicated Yale supercomputing cluster, or through the RSEQtools container image suitable to cloud computing.

[[DS - suggested Figure 1 RSeqTools pipeline]]

5.A.iii.b Finding ncRNAs and transcribed pseudogenes

We will utilize a statistical approach that compares the levels of expression in the known exon regions to threshold the RNA-seq signal and identify the intergenic and intronic regions that show significant expression. Next, we will utilize the methods we developed (e.g., incRNA \cite{21177971}) to further classify and characterize these regions. Specifically, we will use the known coding sequences, UTRs, and non-coding RNAs to train a random forest algorithm and apply the trained algorithm to classify the novel transcript regions to one of the classes. Next we will assign targets to the classified regions by comparing them both with the annotated cis-regulatory elements (e.g. enhancers) and with proximal genes. We will also utilize statistical methods to identify antisense transcripts that have roles in regulating the overlapping transcript.

We will employ our pipeline to identify the transcriptional activity. The essence of the pipeline is to focus on reads and pseudogene regions that are uniquely mappable for the calculation of RPKM. Given previously published results on human pseudogenes with small-scale validation \cite{102,103??} which imply that ~15% of human pseudogenes are transcribed, we can set an RPKM threshold for human analysis such that it gives an approximate agreement with the previous validation. Furthermore, we can generalize our work on comparing pseudogenes expression across organisms to the comparison of pseudogenes expression across a variety of samples in a uniform fashion.

5.A.iii.c Functional annotation through clustering and network analyses

We aim to develop an asthma resource for identifying novel asthma-related genetic elements. Toward this goal, we will perform various clustering and network analysis. We will employ various clustering algorithms to group transcripts based on purely the RNA-Seq data. The clusters will further be validated using biological features such as sequence similarity, genomic distance, and co-regulation. Moreover, we will attempt to predict biological significance of transcripts from biological associations of the modules (e.g. GO terms). As the functions of protein coding genes are more widely known, we will use such clusters to annotate the functions of novel transcripts such as ncRNAs and potentially functional pseudogenes. The clusters will also be used to relate some of the well-known asthma pathways and modules to other less characterized components. The analysis enables us to explore novel asthma-related elements and to examine the relationship between asthma and other pathways in human. Apart from clustering data, we will perform bi-clustering to obtain samples/patients clusters. Certain clusters provide another dimension of information. They will be used for annotating other clinical information.

We plan to extend the OrthoClust framework we developed to compare networks constructed by using samples from patients and samples from control, as well as samples in various cell types. For instance, the quantification on the addition and removal of nodes and edges in cross-species analysis can be easily generalized for comparing signaling pathways for asthma study. Furthermore, as a general formalism, OrthoClust can be used to study specific modules contributed to asthma.

5.A.iv Deliverables [[ds: condense here & move more later]]

The primary deliverable from this aim is the pipeline for the processing of bulk RNA-seq data which will be delivered to Core C for execution and made available to the research community. This process will include detailed annotation of transcripts including structural information, ncRNAs and psuedogenes. We will then take these rigorously and uniformly processed data and from Core C and generate novel co-expression clusters. Clinical data will be mapped onto these clusters to determine the extent to which the bulk-cellular expression can segregate patient phenotypes. In addition, the ncRNAs, psuedogenes and other transcripts will be mapped onto these clusters to suggest hypotheses for their functions. [[MG2DS: move to aim3]] These clusters (and those from later aims) will be made available to the research community through a website dedicated to this purpose, as we have done for other multi-investigator research efforts (e.g. https://www.encodeproject.org/comparative/).

0) TEA clusters phase 0 (transcriptional endophenotypes of asthma)
1) R clusters from pure RNA-Seq data bcTEA clusters? vs scTEA clusters vs LgTEA clusters
2. Mapping of ncRNAs, pseudogenes and other transcripts to potential clusters
2) software

5.A.v Potential Pitfalls

A potential problem in large scale sequencing efforts are the so-called batch effects caused by technical variation between runs. While extensive effort will be taken by the Precision Profiling Core to mitigate such effects (see XXX), processing steps including principle component clusters will be used to check for associations based on sequencing runs.

5.B Plan for Aim 2

5.B.i Rationale

Severe asthma is a heterogeneous disease with multiple underlying mechanisms and endotypes. The manifestation of each endotype is cumulative result of the coordinated and collective behavior of multiple cell types, leading to the phenotypic symptoms. With, single-cell technology we can measure with great precision the cell types involved in asthmatic response and in the particular modes of signaling employed by these cell types and their differences from healthy patients.

Mutations can drive defects in signaling and downstream gene expression in different cell types that can lead to the overall symptoms of severe asthma. For instance, a subset of asthmatic patients demonstrate a Th2 inflammatory response that starts with overreaction of innate immune cells (macrophages) to environmental antigens such as dust mites, that then drive Naïve CD4+ T cells towards the Th2 lineage. Th2 cells then secrete IL4, IL5, IL-13 and a variety of pro-inflammatory cytokines which mobilize the response of the immune system. Therefore, examination of diverse cell types and their responses to cytokines and stimuli can give us a picture of how the disease is triggered and how it progresses.

In this study, we analyze data generated by the profiling core. This data consists of high-throughput, multi-dimensional single-cell measurements of gene expression and signaling in sputum cells derived from the airways of patients with severe asthma. Sputum contains a mixture of blood and epithelial cell types which are derived from the airways of the lung. By analyzing this data at the single-cell level we will be able to:

1. Discover the phenotypes of immune and other cell types that are present in severely asthmatic patients, particularly rare phenotypes with large effect.
2. To understand signaling logic by utilizing cell-to-cell heterogeneity within each phenotype using CyTOF data.
3. To understand gene regulatory network and pathways involved downstream of signaling using single-cell RNA sequencing.

While bulk RNA sequencing is established and technology for understanding gene expression from cell samples, single-cell technology has possibility of uncovering the unique transcriptional program of each cell. Additionally, differences between cells can be informative of the underlying relationship or network between proteins and genes. This gives an understanding of both the heterogeneity that exists within cell populations and the cellular logic that generates the heterogeneity in cellular decision-making. Results from the bulk analysis of Aim 1 can be used to validate the populations and relationships found in Aim 2.

5.B.ii Preliminary Results

We have previously developed methods for analyzing single-cell data. Our methods are:

1) viSNE which is a dimensionality reduction and visualization algorithm for single-cell data analysis \cite{PMID: 23685480}.
2) DREMI for quantifying signaling interactions in single-cell data \cite{PMID:25342659}.
3) DREVI for characterizing and visualizing relationships between proteins in signaling networks \cite{PMID:25342659}.

One of the advantages of multi-dimensional data are ability to resolve subtle populations progression of cells within a sample. However, it is hard to directly consider all of the dimensions due to visual and computational problems with high dimensions. Therefore, we developed a dimensionality reduction method known as viSNE cite{PMID: 23685480} that preserves distances between cells in high-dimensions optimally in low (⅔ dimensions). This enables the resolution of populations of cells and unsupervised clustering efficiently.

We have also developed methods for characterizing signaling in populations of cells. A major problem in quantifying signaling relationships is highly biased sampling arising from many cells (especially immune cells) that do not respond to stimuli or respond stochastically. In such cases the joint density is very peaked and any statistic that is computed from the joint density considers dense regions more important than sparse regions, even though dependencies and signal transfer can only be inferred when looking at the system under a whole range of conditions. DREVI is based on conditional density estimation between the independent and dependent variable, and reveals the functional shape of the dependency between molecules as well as the stochastic spread in the function along the full dynamic range of molecular operation. Along with DREVI, we developed an information theoretic dependency metric (conditional-Density Resampled Estimate of Mutual Information) for scoring the strength of relationships based on the conditional probability. With DREVI and DREMI, one can quantitatively determine the strength of information transfer and the functions computed by these networks.

The quantitative, behavioral descriptions offered by DREVI and DREMI allow us tease out subtly altered signaling functionality in closely related cell types (Th1 vs Th2 CD4+ helper cells) or between distinct cohorts of subjects (mild vs severe asthma). Such differences are important because related cell types often contain similarly wired circuits, which reuse the same molecules, but behave phenotypically differently. DREMI and DREVI found differences in activation thresholds and shapes of response functions between the signaling networks of naïve and activated T cells. In comparing signaling between naive and antigen-exposed CD4(+) T lymphocytes, we find that although these two cell subtypes had similarly wired networks, naive cells transmitted more information along a key signaling cascade than did antigen-exposed cells [20] (See Fig. 8). These methods were also used to track differences in signaling response between T cells from healthy mice and from non-obese diabetic (NOD) mice, which are prone to developing Type 1 diabetes \cite{PMID:25362052}.

5.B.iii Approach

We use two key technologies (1) CyTOF or mass cytometry and (2) Fluidigm C1 microfluidic device for single-cell RNA-sequencing.

5.B.iii.a CyTOF Analysis

The main aims of CyTOF analysis for asthma sputum samples are

(1) Determination of heterogeneous cell subpopulations present in patients
(2) Matching of subpopulations and quantification of heterogeneity between patients.
(3) Characterization of signaling responses by higher-dimensional DREVI with a fuzzy logic model for integration with RNA-sequencing data.

Determination of cell populations: In order to determine cell types within a sample of single-cells, we propose to utilize our previously developed dimensionality reduction methods in conjunction with newly developed unsupervised clustering. Several unsupervised clustering algorithms have been developed in other fields for tackling related problems. Community detection algorithms from social network research seem particularly promising given their speed and utilization of a cell-similarity graph rather than spatial embedding of the data. Recently, the software tool phenograph \cite{PMID: 26095251} was developed which heavily utilizes the Louvain Community detection method to discover immune cell types present in AML patients. The Louvain method repeatedly and sequentially merges nodes in a cell-similarity graph based on the increase in a measure known as modularity, which quantifies cluster quality. Preliminary results utilizing Phenograph on this data is shown in Fig XXX.

Another class of algorithms for unsupervised clustering emerge from literature in VLSI physical placement, where clusters of network elements (logic gates, buffers etcetera) are placed nearby on chips in an attempt to minimize wirelength and crowding. Algorithms in this class utilize recursive bisection \cite{http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=855358&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F6899%2F18566%2F00855358.pdf%3Farnumber%3D855358}, and spectral methods for clustering \cite{http://www.sliponline.org/Publications/Journals/j37.pdf}. In this project, we will evaluate the robustness of a variety of unsupervised clustering algorithms and utilize the most robust combination of methods to discover novel populations.

Subpopulation Characterization and Matching: We propose to find key signaling differences between mild and severe asthmatic patients and also to identify signaling differences in rare phenotypes in order to find potential targets for drug treatment.

Although Phenograph is able to produce clusters, it does not have the capability of matching clusters between patients in order to find consistently repeating rare populations. We propose to develop an approach based on distances between multidimensional distributions in clusters to find matching clusters across patients. Each cluster is essentially defined by the multi-dimensional probability density function of its markers. We propose to use kernel density estimation to compute a set of marginal densities and for each cluster in patient X, to find the matching cluster in patient Y by finding the cluster that minimizes the distance between these marginal densities. There are several methods of computing distances between densities including a simple L1-norm, KL-divergence, as well as hellinger divergence.

Analyzing Signaling Relationships in subpopulations: Once the clusters or phenotypes of cells are established then we can gauge signaling response within each cluster with previously developed information theoretic techniques for analyzing signaling interactions, DREMI and DREVI, described in the significance section.

[Figure 3 here]

Our goals in asthma sputum samples are to understand how various populations of cells invoke signaling responses to the stimulations given in the analysis by the profiling core. Cells from the sputum of 6 subjects was tested by stimulation with LPS for 6 hours. Future experiments will involve additional types of stimulation such as PMA, and household antigen. It has been reported in literature \cite{22902532} that monocytes which express the TLR4 receptor respond to stimulation by LPS with activation of several canonical signaling pathways including ERK, NFKB. Additionally cells which do not express much TLR also respond, but more slowly with a STAT3 and ITK response. Additional pathways downstream of TLR such as the RIP and TRAF pathways, leading to interferon responses have been reported to be involved. We will take a more unbiased view of this by curating a panel of signaling pathways from the results of the bulk RNA sequencing data and examining these pathways with a time course.

In order to study signal integration along various pathways we will study them using a higher-dimensional extension to DREMI/DREVI. With higher-dimensional DREMI/DREVI we can study how signals from various pathways converge together to form resultant responses in cytokine and transcription factor production. Additionally, higher dimensional DREVI can also be utilized to understand signaling logic.

Logic Model for Signaling For signaling interaction, it can often be seen in signaling that cellular logic can be primarily digital in nature. Indeed many of the signaling response functions examined in [ref] show a sigmoidal relationship, where the level of the Y molecule abruptly increases to a higher stable state upon increase in the X molecule. In Figure 4, we see that this is the case also for multi-parent interactions. Here, if the sum of the levels of two driver molecules is above a threshold, then the level of pGSK3b changes to a higher state. Thus, we can apply the logic-gate models in gene regulation to identify signaling logic-gate functions. Moreover, due to the relatively high noise in signaling, we will also use advanced logical models such as fuzzy logic models. As shown in Figure 4 this can be modeled as a fuzzy logical-OR. The advantage of this form of modeling is that it can make the creation of an integrated model consisting of signaling and gene expression components seamless. Furthermore, logical models are known to scale to large circuits (such as computer chip networks) and can be useful for simulation and prediction of perturbation/drug responses. Hence, we propose to fit signaling interactions found using DREVI and DREMI to suitable logical forms, with parameters for noise and thresholding. Signaling interactions tend to be AND/OR/NOT at a simple level:

1) OR gates model two signaling molecules that can phosphorylate the same residue on a child protein;
2) AND gates model protein complexes or other dual-residue modifications that are necessary for the activation of a protein;
3) NOT gates indicate an inhibition of a molecule by another.

5.B.i. Processing of Single-Cell RNA Sequencing Data

Single-cell RNA sequencing has the possibility of offering an unbiased view of the pathways that are transcriptionally activated upon immune-system activation at a single-cell resolution, even when cells seem phenotypically similar. However, single-cell sequencing suffers from more technical noise as compared to bulk RNA-sequencing, arising largely from three sources 1) sampling inefficiencies which result in only a small fraction of the total number of transcripts being captured, 2) cell-to-cell variations in sequencing efficiency, potentially due to differences in lysis between cells, 3) amplification bias owing to the small amount of starting material for the RNA-sequencing. Attempts have been made to address these concerns (Grun, Kester, & van Oudenaarden, 2014) (Brennecke et al., 2013). However, there is no standard pipeline in place that addresses all of the concerns in going from raw reads from a sequencer (such as the Illumina Hi-Seq) to robust transcript counts.

The main steps of such a pipeline, which have been investigated in literature, are as follows:
1. Debarcoding and error correction
2. Aligning reads from each UMI
3. Quantifying the biological noise in genes

Debarcoding and Error Correction Cell-specific barcodes are the key to identification of the particular collection of transcript sequenced from a single cell. However, these barcodes can be erroneously sequenced, leaving many transcripts unassociated with particular cells. Therefore, an error correction scheme that considers the closest hamming distance barcode from a given barcode could help associate more reads to cells. The design of barcodes with a minimal hamming distance of 3 would allow for the correction of a single error whose probability is estimated by Illumina to be 10^-6.

Aligning Reads from each UMI After splitting reads into their cell of origin, reads can be further divided into their molecules of origin using the unique molecular identifier or UMI tag. Similar to the barcode, the UMI is sequenced along with the read. UMIs are essential to both controlling bias and in identifying the closest element of the transcriptome. The following is a general set of steps for assigning UMI-read-collections to genes.

Previous works (Klein et al., 2015) tend to have very specific recommendations for processing the sequencing. For instance Klein et al propose the following steps.

1. Align reads with standard software such as Bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009).
2. Due to the 3' bias in the library, exclude reads mapping to 400 base pairs distance from end of transcript
3. Exclude reads that map to 10 or more genes
5. Collect reads with similar UMI tags
5. Identify a minimal set of genes that explain all reads using the hitting set problem
6. Attempt to find a member of the hitting set that explains all of the reads.

However, this type of highly deterministic procedure with many thresholds is unlikely to yield good results in all situations. Furthermore, the reason for minimizing the set of genes to explain all reads is unclear and could end up missing many valid alignments. Therefore, we propose to develop an alternative robabilistic procedure where each gene is given a probabilistic alignability score that represents how well the collection of reads align to the particular gene. The probabilistic score incorporates for each read r that aligns to the gene the following:

1. Pt : How far the read aligns from the end of the transcript, with genes aligning close to the end having a high distribution, modeled as a skewed lognormal distribution.
2. Pg: How many other genes the read itself aligns to, which is a distribution peaking at 0 with a thin tail such as a Gaussian distribution.

The probabilistic score is the sum of $\Sigma\ PtPg$ of the product of the values for every read that aligns to the gene. The UMI would then be assigned to the gene that explains the set with the highest probability.

Quantifying the biological noise in Genes Quantifying the biological noise of each gene involves separating components for technical variation from biological variation in cell-cell gene abundances. There are generally thought to be two sources of technical variation.

1. Cell-to-cell variability in RNA sequencing efficiency: This essentially means that many RNA molecules are captured from some cells whereas few are captured from other cells. Therefore transcript abundances are sensitive to variations to changes in sequencing efficiency resulting from processing steps such as lysis efficiency. Therefore, normalizing by the library size or total number of transcripts sequenced from a cell can mitigate this type of variation.
2. RNA sampling from cells: Previous work has quantified the fraction of transcripts that are sequenced using ERCC spike-ins and found the efficiency to be about 3.6%. This implies that the RNA-sequencing reaction only samples approximately this amount of transcripts from the entire complement within the cell. Further, Grun et al. find that this sampling probability is distributed such that the variance is equal to the mean of the distribution and therefore can be described as a Poisson distribution.

If the complete variation in measured gene expression is due to Poisson sampling then the fano factor of the gene expression should be equal to 1, higher fano factors indicate the presence of actual biological variability rather than simply technical variability. Therefore the amount of information in each gene measurement can be quantified by its fano factor and utilized in selecting genes to analyze.

5.B.ii Single-cell RNA-sequencing Analysis

After the pipeline steps are completed then we can analyze phenotypes and gene-gene interactions in a similar way as we analyzed CyTOF data. However, one of the keys to successfully extracting information from single-cell RNA sequencing data is to be able to use the high-dimensionality of the data, to bolster individual (especially low-abundance) gene dimensions that can suffer from dropout. We propose the following steps in order to be able to analyze and cluster single-cell RNA-sequencing data.

(1) Use non-linear dimensionality reduction and clustering on genes to form meta-genes
(2) Value-imputation based on cell clusters and meta genes.
(3) Use the value-imputed data to study gene-gene interactions

Non-linear dimensionality-reduction and clustering Some genes are naturally expressed at low abundances and these can be especially affected by the poisson sampling process by which RNA is captured from single cells. However, since single-cell RNA sequencing data involves measuring thousands of gene dimensions, it is possible to impute values for dropout dimensions using information from a combination of higher-fidelity dimensions. In order to tackle this problem, we propose to non-linearly reduce the number of dimensions by utilizing a method such as bh-SNE \cite{25449901, ACM link: http://dl.acm.org/citation.cfm?id=2697068} or non-linear PCA \cite{16109748}. After this reduction, we will cluster genes based on the dimensionality-reduced embedding of each cell. We call the resultant cell groupings metagenes. Such metagenes may represent pathways or other functional groupings, which can be examined by enrichment analysis.

Cell clustering and value imputation based on meta-genes: Once meta-genes are derived, cells can be clustered based on the average expression of meta-genes. Each metagene is essentially a cluster of genes that have similar co-occurrences in the population of cells. Therefore we can use cell clusters derived from meta-genes in order to impute missing values for low-abundance genes. If a cell expresses many members of a metagene, then it can infer a missing value for a gene within the meta-gene by taking a weighted average of cells in its cluster.

Use the value imputed data to study gene-gene interactions through DREMI: Once values are imputed into the cell-gene matrix, then it becomes possible to study pairwise gene-gene interaction strengths once again using techniques such as DREMI. We propose to study pairwise dremi on all pairs of genes exhaustively to derive a gene-gene DREMI matrix. This is essentially an adjacency matrix where the similarity is defined by the mutual information metric DREMI. Next this adjacency matrix can be utilized in graphical or spectral clustering to discover gene modules or pathways through which information is flowing. Note that this is different from the meta-genes because the genes along mutually informative pathways need not have similar expression across cells, they must simply be mutually informative or predictive of one another under probabilistic analysis. In this way we hope to discover new gene-modules or pathways that may be characteristic of cell-subpopulations in asthma patients. These modules can form the basis for additional CyTOF experimentation to discover how signaling is processed along new pathways that have not been studied extensively, and makes for an iterative approach to deepening the molecular mechanisms underlying the disease.

[Figure 5 here?]

5.B.iv Deliverables

The deliverables from this aim include software and pipelines for the analysis of both CyTOF and single-cell RNA-sequencing data as well as results of the analysis on data generated by the profiling core. From CyTOF analysis:

(1) CyTOF phenotypic clusters that result from unsupervised clustering of surface markers.
(2) A method that matches subpopulations and tracks changes in similar subpopulations across patients.
(3) Method for characterization of signaling interactions using DREVI and fuzzy logic in the subpopulations.

From single-cell RNA sequencing:

(1) A pipeline that processes single-cell RNA sequencing data by debarcoding, quantifying noise, and selecting genes.
(2) A method that returns metagenes (reduced gene dimensions) and clusters of cells in single-cell RNA sequencing.
(3) A method that imputes missing values for low abundance genes in single-cell RNA sequencing data.
(4) DREMI analysis of gene-gene interactions and resultant gene modules.

5.B.v Pitfalls

Aim 2.

Potential pitfalls and include

1) Non-robustness of unsupervised clustering methods. Alternative methods can include spectral clustering \cite{Ng et al. NIPS 2001 (I can't find another id for this)}.
2) Difficulty matching clusters between patients, alternatives can include renormalization or matching of post-processing signatures.

3) Iterative deconvolution, single cell transcriptomics (4 pages)
 - relate the clusters (part 1) to the deconvolution
 - relate the CyTOF (part 2) to the deconvolution
==

5.C Plan for Aim 3: integrative model building [1700 words]

5.C.i Rationale

The analyses related to clinical issues from the bulk RNA-seq, single cell RNA-seq and CyTOF measurements come from their integration in the form of a model. This will define the data that best correspond to clinical phenotypes in such a way that the pathways contributing to those phenotypes can be identified.

5.C.ii Preliminary Results [600 words]

5.C.ii.b preliminar results related to budilign logical models characterizations cluster [[DW to fix up]]

LOREGIC- grab text for LOREGIC [1 para]; dreiss [1 line]

Gene expression is controlled by various gene regulatory factors. Those factors work cooperatively forming a complex regulatory logical circuit on genome wide. Recently, an increasing amount of next generation sequencing data provides great resources to study regulatory activity, so it is possible to go beyond this and systematically study regulatory circuits in terms of logic elements. To this end, we developed Loregic, a computational method integrating gene expression and regulatory network data, to characterize the cooperativity of regulatory factors. Loregic uses all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target \cite{ PMID: 25884877}. We attempt to find the gate that best matches each triplet's observed gene expression pattern across many conditions. In Loreigc, we also developed a consistency score based on Laplace's rule of succession and permutation test to measure how a triplet is consistent with a logic gate. We made Loregic available as a general-purpose tool (github.com/gersteinlab/loregic). We validated it with known yeast transcription-factor knockout experiments. Next, using human ENCODE ChIP-Seq and TCGA RNA-Seq data, we were able to demonstrate how Loregic characterizes complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs in human cancer. In addition, we inter-related Loregic's gate logic with other aspects of regulation, such as indirect binding via protein-protein interactions, feed-forward loop motifs and global regulatory hierarchy. Besides the regulatory logics, we also developed continuous model-based approaches such as DREISS for dynamics of gene expression driven by external and internal regulatory modules based on state space model to help dissect the temporal dynamic effects of different regulatory subsystems on gene expression (https://github.com/gersteinlab/Dreiss, PLoS Computational Biology, minor revision).

5.C.ii.d Futher experince developing Statistical models of data integration

- CRIT & fungus paper - grab: sbfuel

We plan to develope many integrative models suing the clusters… we have much experience doing this…

The Gerstein lab has experience integrating diverse data types, including RNA-seq and mass spectrometry data. For example, we used gas-chromatography mass spectrometry profiles of the biofuel-producing fungus *Ascocoryne sarcoides* and its associated RNA-seq data to predict the novel biofuel-production biosynthetic pathway \cite{22396667}.

We also developed a machine learning algorithm using high-order neural networks to predict complex peptide-protein binding, which can greatly help clinical peptide vaccine search and design \cite{PMID: 26206306}. (High-order neural networks and kernel methods for peptide-MHC binding prediction, PP Kuksa, MR Min, R Dugar, M Gerstein. (2015) *Bioinformatics* Jul 23. pii: btv371.)

We have developed statistical predictive models by integrating various omics data types. For instance, transcription factors and histone modifications are two interrelated components that regulate the transcriptional output of a gene. To quantify the relationship between TF binding and gene expression, we have constructed linear and non-linear models that take the binding signals of multiple TFs in the transcription start site (TSS) proximal to genes as the input to “predict” gene expression levels as the output \cite{22955978, 22955616, 21926158}. Similarly, we have also constructed models to predict gene expression levels based on histone modification signals at different positions proximal to the TSS of different genes \cite{22950368, 21324173, 21177976, 22950368}. We constructed TF and histone models for predicting expression levels of protein-coding and non-coding genes \cite{21324173, 21177976, 21926158}. Strikingly, the models trained solely on protein-coding genes also predict the expression levels of non-coding genes, suggesting a common regulatory mechanism is shared between them. In addition, our models indicate that, in different species, the functions of histone modifications are conserved. A universal model trained from histone modification data that contains equal numbers of human, worm and fly genes can predict gene expression level with fairly high accuracy in all three distantly related organisms \cite{25164755}.

5.C.iii Approach

5.C.ii.a Interrelation with external datasets

there’s a wealth of info relatd tot rnaseq … we plan to inter-relate the data generated here w these extrenal resoucle

There are several big-data projects relevant to the analysis and interpretation of the bulk-cell and single-cell RNA-seq data and their interrelation with CyTOF measurements. For example, GTEx (http://www.gtexportal.org/) has tissue-specific transcription data, including lung, which can be used to infer aberrant transcription in the asthma disease states. Data from the ENCODE project (https://genome.ucsc.edu/ENCODE/), particularly the ChIP-Seq data, will give a regulatory framework into which the asthma data can be mapped. We have experience integrating ENCODE data into regulatory networks \cite{22955619} and studying the impact of transcription factor binding and histone modifications on gene expression \cite{21324173}. We will leverage this to embed transcripts into cellular regulatory networks and to provide the context needed to understand the role they may play in intercellular signaling. After that, we will identify the key transcripts with high network centralities, and try to predict their functions using “guilt-by-association” with their neighbors.

Besides ENCODE, several other large consortia are generating data systematically across the human genome, resulting in a wealth of functional information of great value to RNA-Seq integrative analyses. The Epigenomics Roadmap Project and the International Human Epigenome Consortium have generated rich maps of histone modifications, including deep maps of more than 20 modifications in a small number of cell lines, maps of a few modifications in a large number of cell types, as well as maps of DNA methylation and DNA accessibility. Over 1,200 data samples from primary tissues have been collected and analyzed by the NIH Genotype-Tissue Expression (GTEx) Project. By integrating the transcripts with the Human Epigenome Atlas and GTEx data we will examine potential effects of a transcript on chromatin modifications in target cells. This is particularly important for those lncRNAs known to regulate histone marks such as H3K27me3 and H3K9me3 through interactions with the members of the Polycomb complex.

Other sources of complementary, large-scale human data include: the NIMH Brainspan Project, the 1000 Genomes Project, and the NCI Cancer Genome Atlas (TCGA) Project. The DOE kbase (of which we are members) \cite{kbase} provides new genomic toolsets that we will harness.

[[DW's bullets]][[dW to update inverting the text]]

- Signature finding [2-3 page] [500 words - 3 para SKL & DW]

[[ds: aim 3]] In asthma study, certain signaling pathways (e.g. wnt signaling) are of particular interest. We plan to study these pathways via network analysis, for instance, their hierarchical structure. While there is a few well studied pathways closely related to asthma, it is instructive to explore how these pathways interact with other signaling pathways, mediated by the sharing of various molecular components. We plan to study the cross-talks between pathways by integration of various networks like such signaling networks and protein-protein interaction network.

5.C.iii.a Deconvolution of cell-type signatures from bulk RNA-seq data

[[DW: one integrated section where we develop signatures & we make avaialble signature pipeline - merge 5.C.iii.b]][[DW: 5.C.iii.b has been merged to 4C.iii.a]]

[[DW we need ot clarify the Qs that this section will adddress : here we will inter-relate cell type signatures, single cell data , bulk data and proportions of cells i nthe bulk … we use the single cell to help refine cell type signatures, we will the signatures to help determine cell proportions in the builk incl CYTOF]]

In this aim, we want to identify the cell type signatures in terms of gene expression, and the signatures that can most discriminate asthma patients; e.g., different TEA clusters. We plan to apply both linear and non-linear models. We assume that the mixture of multiple cell type signatures determines the gene expression of each patient. For instance, the patient's ith gene expression level can be modeled as a function of the same gene's expression levels of multiple cell type signatures.

[[method 1]]We will start from the linear model, which will be computationally efficient, that the ith gene expression level of kth individual person, x(i,k) is the linear combination of this gene's expression levels of different cell type signatures; i.e., x(i,k)=\Sum_{j=1}^{m} w(j,k) * s(i,j), where s(i,j) is the ith gene's expression level in the jth cell type, and w(i,k) is the contributing weight of jth cell type to kth person, which can be the jth cell type fraction of kth person. If we rewrite this linear model in a matrix form, we have that X=SW, where X is the gene expression matrix whose the rows and columns represent genes and persons, W is the cell type fraction matrix whose rows and columns represent cell types and persons, and S is the cell type signature matrix whose the rows and columns represent genes and cell types.

The single-cell RNA-seq data described in Aim 2 will yield counts of different cell types, providing the data required for matrix W. The bulk RNA-seq data provided by Core C after being processed by the pipelines developed in Aim 1 will provide matrix X, so we need to find the optimal S to minimize ||X-SW||_F given X and W. The optimal solution S=XW*, where W* is pseudo inverse of W s.t., WW*=I identity matrix.

[[ds: we will make this inot a pipeline "signature pipeline that we;ll use here + also apply to the earlier T-clusters for P1]]
[[DW to do run it…. it should read meth1, meth2, runit]]These methods will be compiled into a cell-type signature pipeline

that will be distributed to the other Driving Projects for determining the relative fractions of cell-type expression from bulk RNA-seq data. This will be applied to the novel clusters produced in Aim 1 to elucidate the effect of cell-specific transcription in driving the clustering of different samples. Moreover, we will apply this pipeline to established clustering methods, such as TEA clusters, to observe cell type signatures in these contexts. We integrate the analyses of these different clustering methods to identify the cell and gene specific biomarkers that most discriminate clusters.

[[method 2]] In addition, we will also try to use the advanced models such as a machine learning method to investigate the gene markers from cell type gene expression signatures for both bulk data and single-cell type. For example, the Denoising Autoencoder(DA) is an unsupervised machine learning framework that be used to extract and characterize gene signatures involved in new pathway and principals. Firstly, All the gene expression value are transformed into value with the range [0,1], the input node is defined as x, a N-dimensional vector, and N is the number of genes. Then we will define latent representative or code node, denoted by y with dimension N\prime (N\prime << N). We map x to y by a non-linear sigmoid function, which we call it encoding step; e.g., y=sigmod(Wx+b). The latent node y is used to capture first N\prime principal components of data x. Because of the non-linearity function, DA is capable of capturing multi-model aspects of input distribution. We will later decode y and reconstruct z to be the same shape of x using similar transformation; e.g., z=sigmod(W^T y+b\prime). The reconstruction error is estimated by cross-entropy: L= -\sum x log z + (1-x) log (1-z). Before we encode x to y, we add noise by randomly changing some feature to 0, which are controlled by 'corruption level' parameter. The stochastic gradient descendent algorithm will be used to optimize the reconstruction errors. After converged, we will link the y value of latent node into component activity. By hierarchical clustering, we can characterize nodes highly associated with known asthma subtype/TAE cluster. From these latent nodes, we extract high weighted genes according to weight matrix W. Based on the distribution of weight score for each gene, we can characterize the significant high and low weighted gene as the candidate gene markers.

We will also integrate our modeling and analysis into a pipeline, which inputs gene expression and cell fraction data along with clusters information and outputs cell type signatures and gene biomarkers to discriminate clusters.

5.C.iii.c Identification of clinical and CyTof features of clusters

and association clusters cell type signatures

[[SKL - how we’ll cluster based on clincial … how we’ll incorpate p2 bio clusters as seeds … use them w/ orthoclusts as containst… 3 para]]

Clinical information [[SKL: Do you work we need more information hereIt workedWasWe need more information about what clinical variables are could be clustered for instance we could put in peak airway volume or something like that we need a little bit more from here may we can get it for next Wednesday's meeting or you can get it from talking to Jeff]]

can be used to classify endotype of asthma and provide valuable guidelines for diagnosis. Some features like FEV1 and FVC has been widely used in the endotype clustering. However, the traditional distance-based agglomerative clustering algorithms know nothing about the probabilistic model of the data and cannot infer the number of clusters, so we will use bayesian hierarchical clustering method to infer clinical phenotypic clusters and associate clinical features. clinical data are presented by multinomial distribution with a Drichlet prior and continuous data are assumed to follow a gaussian distribution. Starting from the initiation state that each sample was treated as a trivial cluster and then iteratively merges the most similar cluster, a product of marginal likelihood with unknown parameters will be optimized. We can define clinical phenotypic clusters based on 1000 more samples.

Meanwhile, In project 2, three asthma associated pathways will be used to validate and extend the gene signatures characterized by CytoF and RNA-Seq. We use the experimental validated pathways as seed and expand to whole network. Belief propagation based on experimental results can be used to update and optimize the weight between gene-gene interaction edges.

We can build co-expression network for each cell type and clinical phenotypic clusters (endotypes). Based on these networks, we applied our Orthoclust framework to identify common and specific regulation or signaling pathways for different cell types and endotypes. Firstly, multi-layer network is constructed based on networks for cell type or endotype samples. The individual layers of co-association networks are combined by connecting the same genes on these layers, which will form a super network. After simulated annealing optimization, we will identify network modules that are specific for each cell type and endotypes, which will help us on classification based on gene expression levels. We will also explore common and specific modules for different asthma developmental stage to investigate the potential gene markers that associate with asthma prognosis. Specific module in signaling response pathway from CytoF and logic gate analysis will explain the dynamic regulation and cascaded signaling transduction in Asthma progression.

5.C.iii.e. Logical modelling building

[250 words - dW - para][[DW merge all the logic modelling…. we're going to use logic gates to characgterize teh clusters]][[DW: all logical modeling text integrated]][[Keeping in the logical modeling section is that we're going to use this to characterize the many clusters]][[DW:added transitions, removed chip-seq, moved consistency scores to prelim., updated reg. and signaling logics]]

In addition to identification of clusters as described above, we will also explore the biological mechanisms for the phenotypes of these clusters. The gene regulation and signaling interaction are two major mechanisms at the molecular level, and follow certain logical behaviors to give rise to the phenotypes. We plan to use logical modeling approaches to identify logical functions in gene regulation and signaling interaction, and to use them to characterize the asthma clusters. For example, we can find the different gene regulatory logics between server and mild asthma patients.

For gene regulation, it is noteworthy that various regulatory mechanisms are influential at different levels of the genome including transcriptome and proteome. These gene regulatory factors cooperate in multiple dimensions to facilitate the correct function of the genome as a whole. If their cooperation has some problems, it can give rise to abnormal gene expression such as one in asthma. In many cases, the regulatory factors controlling gene expression behave in a discrete fashion and can be modeled using Boolean models and logic circuits [147-153]. Additionally, the simple binary operations in the Boolean model do not need large amounts of data are therefore very computationally efficient. Therefore, we will develop computational algorithms based on Boolean models to study and compare the logic of combinatorial cooperation between various regulatory factors such as TF-TF and TF-phosphorylation for different patient clusters. First, we will model the regulatory factors along with their targets (regulatory modules) using input-output logic circuits. By integrating gene expression data and regulatory information, we will then identify the behavior of logic circuits for individual regulatory modules. Furthermore, we will connect logic circuits for all regulatory modules to build a Boolean regulatory network, hence providing a system-level view of gene regulation. Last, we will analyze the Boolean network using various algorithms based in network theory to predict novel regulatory pathways, and identify asthma cluster's specific regulatory logical pathways.

We plan to identify the gene regulatory logics based on logic-gate models above for different asthma clusters, and find the specific logics that drive the cluster's expression such its biomarker gene expression. First, we want to construct the gene regulatory networks consisting of various regulatory factors and their target genes. In order to define a more complete set of TF-gene regulatory relationships, we will combine these data with data on TF binding using the asthma-related cell types such as Eosinophils, Lymphocytes, Blood and Neutrophils previously published by the ENCODE project and Epigenomics Roadmaps and described in other studies [16, 55]. Second, we want to identify the regulatory logics in the constructed gene regulatory network to drive the expression patterns for a particular group of patients with similar clinical features such as a TEA cluster. We will use data from regulatory networks (defined by regulatory factors and their target genes) and binarized gene expression datasets across the cluster's patients. The binarized gene expression data (on=1 and off=0) is the direct result of the network's regulatory factors activity on the target genes. Our study will decompose the regulatory network into gene regulatory modules. Those modules can be the simple triplets consisting of two regulatory factors (RFs) and a common target gene T, or the ones with multiple RFs and common targets. The main idea is to describe each module using a particular type of logic gate, i.e. the logic gate that best matches the binarized expression data for that triplet across all samples. For example, RF1 and RF2 regulate a gene T following an AND logic; i.e., both

RF1 and RF2 need to express high to turn on the gene T. We will also assign a consistency score to measure how (RF1, RF2, T) is consistent with AND logic as introduced in Preliminary results.

In addition to the logic gates from regulatory modules, we will also find the logic circuits consisting of the cascaded logic gates for the regulatory pathways. After finding the regulatory logics for different clusters, we will compare the logics across clusters, and find the cluster's specific regulatory logics. For example, the triplet of RF1 and RF2 regulating T may follow AND logic in severe asthma patients, but OR logic in mild patients. We will also check the changes of regulatory logics of the same biological pathways across clusters. In addition to identify logics, we will want to develop theoretic solutions to guide genomic engineering techniques like knockdowns for changing the regulatory logics, especially for severe patients.

Hence, using these basic logical modes, combined with a stochastic noise model, we propose to encapsulate protein and gene interactions in a computationally efficient logic model. Finally, we will also develop a pipeline for this logical modeling and analysis, which outputs the gene regulatory and signaling logics to characterize the clusters.

4. Interactions with the other members of this U19 Cooperative Proposal

This research will be undertaken with extensive interaction and collaboration with the other members of this U19 proposal \cite{interactions figure}. In our first two aims we will be working closely with the Precision Profiling Core using test datasets to generate a processing pipelines for the bulk-RNAseq, single cell RNAseq and CyTOF data. These pipelines will be given to the core for implementation, which they will then use to distribute data to all three Driving Projects. Our final aim will generate a model that will both use data from and inform the other driving projects. It will use data from other projects to refine the clusters of transcripts. For example, microbial community clusters from Project 2 Aim 2 could be used to seed clusters in our model. By this method we will evaluate the strength of the data generated by other groups at stratifying patients into clinically relevant phenotypes. Our model will inform the work in other projects by offering novel clusters to test. Project 1 aim 3 will use the clusters from our model to determine cell activities in a stimulation assay. These findings will be communicated in monthly meetings of the group and more frequent interactions between subgroups.

So Mark today and in this last section we shouldDefine the interface between the same and the other aims we should also make a table describing all the clusters and how these clusters are going to be used throughout the grant and by other subprojects

Merged this bit with the deliverables section

[[

5.C.iv   Deliverables

[[

[DW] bioinformatics tools such as R packages & websites to identify cell type signatures, analyze enriched features like clinical, cytof, dynamics, find regulatory logics…
databases for cell type signatures
preliminary results: https://github.com/gersteinlab/Dreiss, https://github.com/gersteinlab/Loreigc,

[SKL]
Biomarkers for diagnosis and treatment. we will also investigate the molecular mechanism of Asthma.

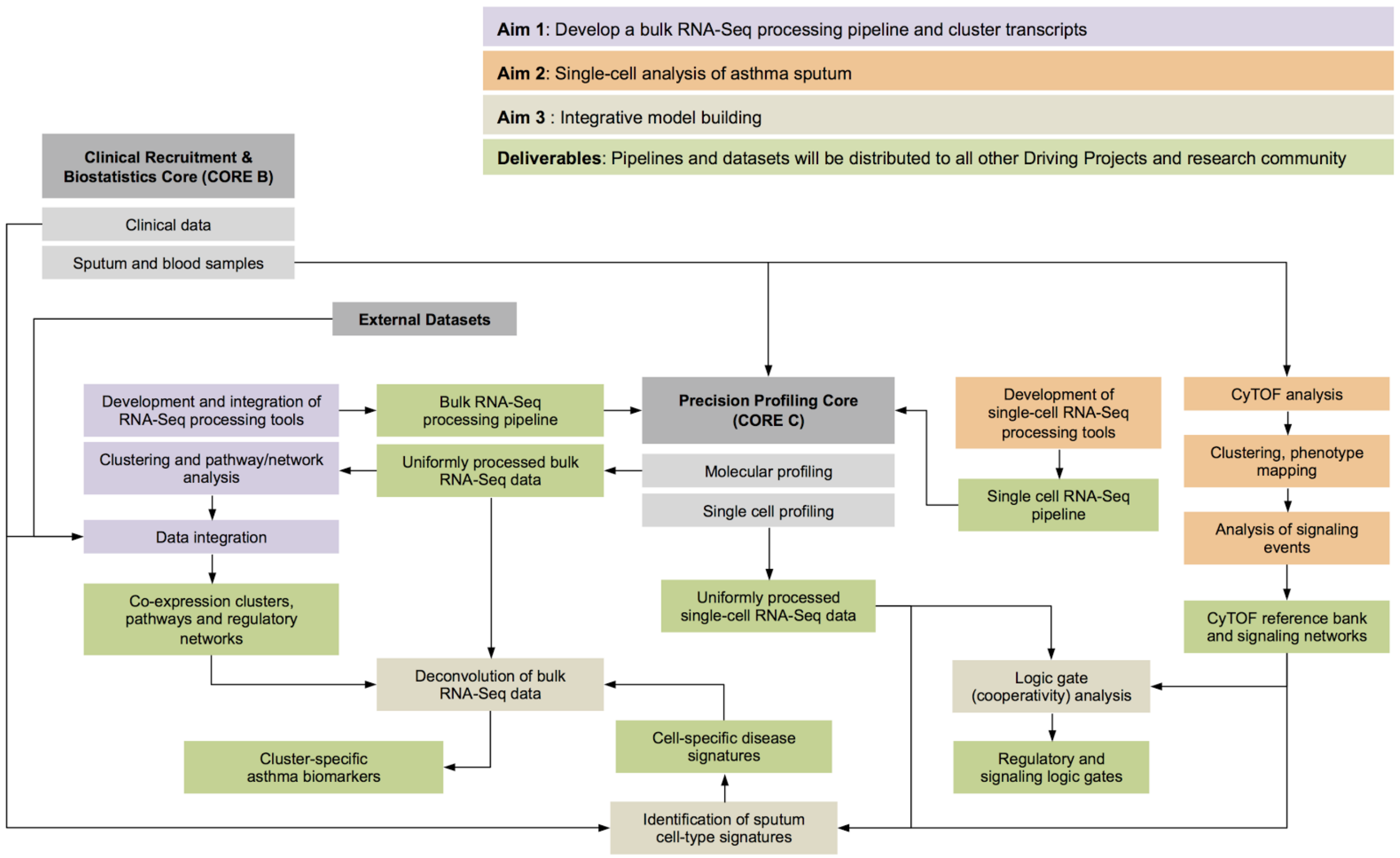
the signaling pathway and logic gate

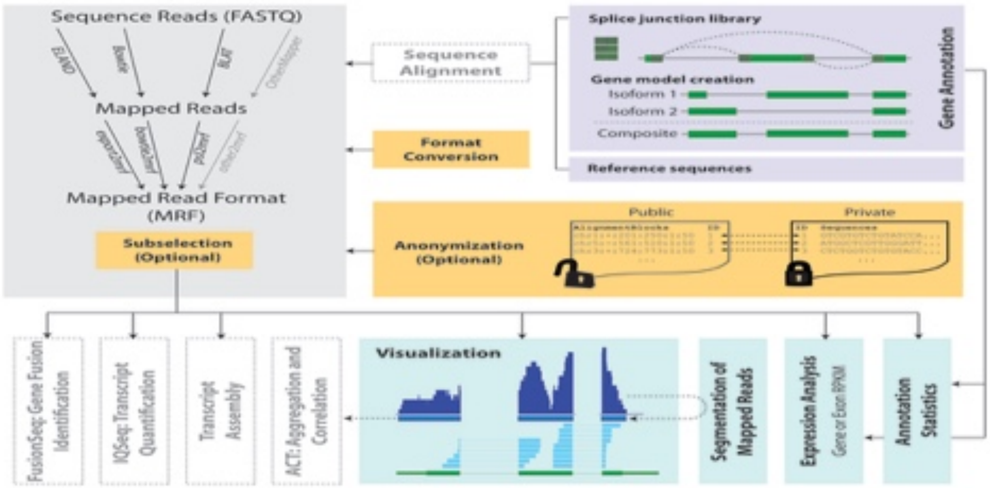
5.C.v Pitfalls

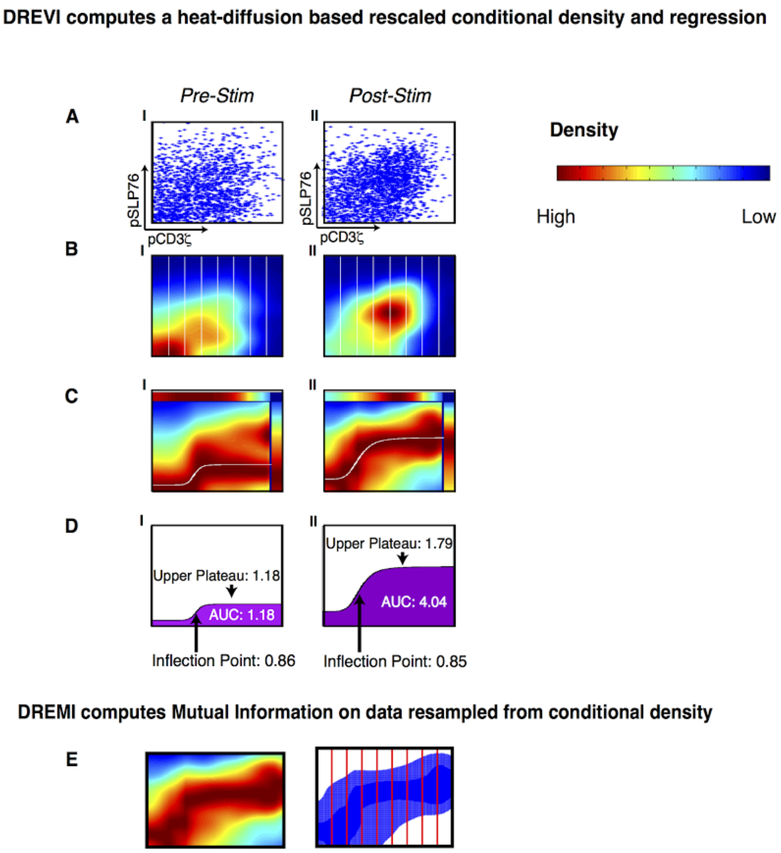
1. limitations of different method: microarray data, RNA-Seq, single cell vs bulk cell, Cytof (limited by known knowledge)
2. limitations of different analysis method: deconvolution method for bulk cell data (microarray and RNA-Seq); single cell, link/variability between transcriptome and proteome
3. clinical versus basic research. heterogeneity of patient samples and limiting of clinical diagnosis (histology versus molecular level).

6. References

\bibliography{}

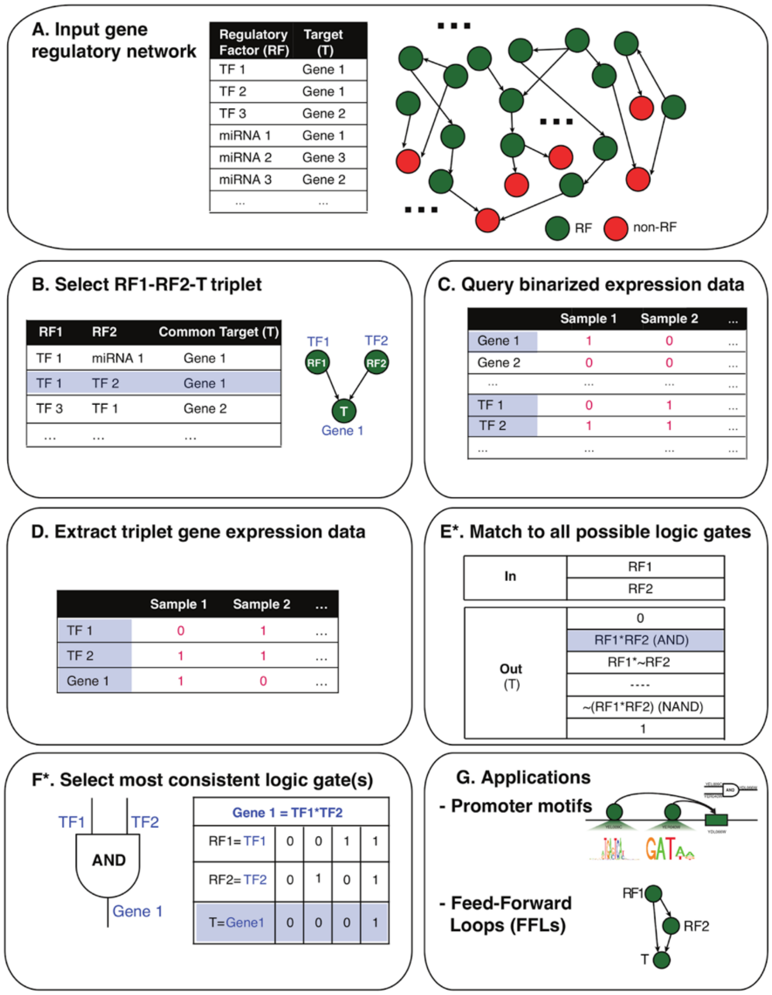

Aim 1: Develop a bulk RNA-Seq processing pipeline and cluster transcripts
Aim 2: Single-cell analysis of asthma sputum
Aim 3 : Integrative model building
Deliverables: Pipelines and datasets will be distributed to all other Driving Projects and research community
Clinical Recruitment & Biostatistics Core (CORE B)
Clinical data
Sputum and blood samples
External Datasets
Development and integration of RNA-Seq processing tools
Bulk RNA-Seq processing pipeline
Precision Profiling Core (CORE C)
Development of single-cell RNA-Seq processing tools
CyTOF analysis
Clustering and pathway/network analysis
Uniformly processed bulk RNA-Seq data
Molecular profiling
Single cell profiling
Single cell RNA-Seq pipeline
Clustering, phenotype mapping
Data integration
Analysis of signaling events
Uniformly processed single-cell RNA-Seq data
Co-expression clusters, pathways and regulatory networks
CyTOF reference bank and signaling networks
Deconvolution of bulk RNA-Seq data
Logic gate (cooperativity) analysis
Cell-specific disease signatures
Cluster-specific asthma biomarkers
Regulatory and signaling logic gates
Identification of sputum cell-type signatures


Sequence Reads (FASTQ)
ELAND
Bowtie
BLAT
OtherMapper
Mapped Reads
eland2mrf
bowtie2mrf
psl2mrf
other2mrf
Mapped Read Format (MRF)
Subselection (Optional)
Sequence Alignment
Format Conversion
Anonymization (Optional)
Public
Private
Splice junction library
Gene model creation
Isoform 1
Isoform 2
Composite
Gene Annotation
Reference sequences
FusionSeq: Gene Fusion Identification
IQSeq: Transcript Quantification
Transcript Assembly
ACT: Aggregation and Correlation
Visualization
Segmentation of Mapped Reads
Expression Analysis
Gene or Exon RPKM
Annotation Statistics



DREVI computes a heat-diffusion based rescaled conditional density and regression

DREMI computes Mutual Information on data resampled from conditional density



A. Input gene regulatory network

Regulatory Factor (RF)	Target (T)
TF 1	Gene 1
TF 2	Gene 1
TF 3	Gene 2
miRNA 1	Gene 1
miRNA 2	Gene 3
miRNA 3	Gene 2
...	...

B. Select RF1-RF2-T triplet

RF1	RF2	Common Target (T)
TF 1	miRNA 1	Gene 1
TF 1	TF 2	Gene 1
TF 3	TF 1	Gene 2
...	...	...

C. Query binarized expression data

	Sample 1	Sample 2	...
Gene 1	1	0	...
Gene 2	0	0	...
...	...	...	...
TF 1	0	1	...
TF 2	1	1	...
...	...	...	...

D. Extract triplet gene expression data

	Sample 1	Sample 2	...
TF 1	0	1	...
TF 2	1	1	...
Gene 1	1	0	...

E*. Match to all possible logic gates

In	RF1 RF2
Out (T)	0 RF1*RF2 (AND) RF1*~RF2 ---- ~(RF1*RF2) (NAND) 1

F*. Select most consistent logic gate(s)

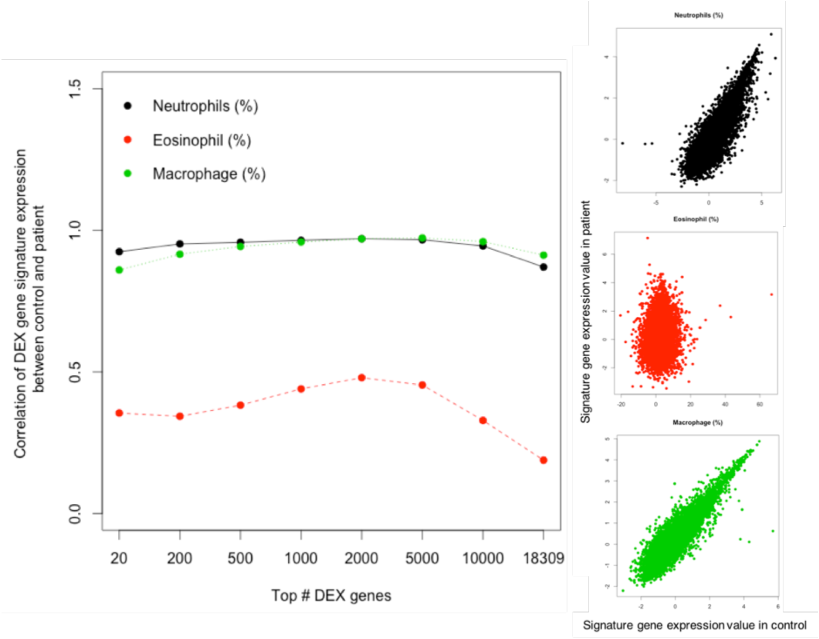
Gene 1 = TF1*TF2				
RF1= TF1	0	0	1	1
RF2= TF2	0	1	0	1
T=Gene1	0	0	0	1

G. Applications

- Promoter motifs

- Feed-Forward Loops (FFLs)

* See Figure 2 for details



Aim	Cluster Name	Data	Clustering Method	Utility
1	BCP	Bulk-cell RNA-seq	by Patient	Unrefined patient stratification by transcription
	BCG		by Gene	Unrefined co-expression networks for holistic response analysis
2	SCC	Single-cell RNA-seq	by Cell	Cell signatures by transcription
	SCG		by Gene	Co-expression within cells
	SCyC	Single-cell CyTOF	by Cell	Cell signatures by protein levels
	SCyG		by Gene	Signaling network analysis
3	ACP	All	by Patient	Novel clusters of asthma endotypes
	ACC		by Cell	Cell signatures for disease stratification
	ACG		by Gene	Logic modeling for disease mechanism