Asthma MAP: Computational Tools and Clustering for the Study of Asthma Heterogeneity

1. Specific Aims

The goal of this project is to derive clinically informative clusters of genes, cells, and patient attributes that elucidate mechanisms of different asthma phenotypes. We hypothesize that endotype clusters from a combination of transcriptional and protein profiling of a well-characterized cohort of asthmatic patients can differentiate asthma severities in a way that identifies new mechanisms and facilitates improved patient care. We will use RNA sequencing and CyTOF data from the Precision Profiling Core C to develop an integrative model of asthma to better understand key aspects of asthma heterogeneity and severity. To this end, we will use our expertise in RNA sequencing to develop and distribute "pipelines" to the Precision Profiling Core C for the processing of bulk 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 disease phenotype datasets will be characterized to define pathobiologically meaningful endotypes of disease. This project will enrich our knowledge of this disease, particularly of the regulatory pathways that contribute to the heterogeneity of patients' asthma experiences. In addition, we will discover new gene networks important to the pathogenesis of asthma and define those that best correlate with clinical outcomes. Finally, this project will generalize asthma disease observations to a systems level so that we might speak to the underlying mechanisms by which various clinical outcomes occur and create a publicly accessible, searchable, integrated asthma MAP.

Aim 1: Bulk-cell RNA-seq processing pipeline and transcript clustering

We will adapt a comprehensive suite of human RNA-Seq tools to generate pipelines for the uniform processing of bulk-cell RNA-Seq data isolated from sputum cell populations. 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, observe the extent to which the transcripts 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 1 and 2 in this cooperative agreement. We will use these data to generate bulk-cell clusters by patients and by genes (BP and BG-clusters, \cite{...}). The patient clustering will define asthma endotypes and the gene clustering will define co-expression networks that will speak to the mechanism of asthma disease.

Aim 2: Single-cell analysis of asthma sputum

We will utilize single-cell measurements of protein and mRNA abundances using mass cytometry (CyTOF) and RNA-seq technologies to deeply characterize rare and novel populations of cells produced in the airways of individuals with asthma. We will develop a signature from multidimensional CyTOF measurements of signaling and surface marker molecules based on an unsupervised community detection method. Further, we will develop a method to match such results across samples such that the populations are validated through repeated detection. These cell-signatures will be used to generate protein and cell specific clusters (CC and CG-clusters) to stratify cell types and signaling pathways. 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 use that as the input to the integrated logic model of Aim 3. For transcripional profiling, we will employ single-cell RNA-Seq, a newer technology that is 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 in coordination with Core C 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 results into metagenes corresponding to pathways (SCG-clusters). 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. With the understanding of signaling and gene-interaction networks we can characterize the pathways involved in the immune reactions to asthma, and understand their involvement in emerging phenotype. Further, these could be exploited as putative drug targets for treatment of patients.

Aim 3: Integrative clustering, interfacing with other projects and the creation of the AsthmaMAP

We will use the data described in Aims 1 and 2 of this project along with the clinical data generated by Core B and external datasets to make integrative clusters by patients, cells and genes (IP, IC and IG-clusters). Cell-

type signatures defined in Aim 2 will be used to de-convolve 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 (e.g. GTEx for tissue-specific context and ENCODE for transcription factor data). We will use logical modeling within the IP clusters to build regulatory differences between the asthma endotypes. Each of these clusters, networks and models will be evaluated in the context of established clinical measurements (e.g. FEV1 and FeNO) to identify effective measures to stratify patients and how they might give insight to the mechanisms of asthma disease and heterogeneity. To effectively disseminate the data, pipelines and analyses generated by this and the other driving projects in this cooperative proposal, we will create the Asthma MAP website. This website will facilitate interaction within this U19 as well as provide publicly accessible, searchable resource for the asthma research community.

2. Significance

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. The YCAAD research team (see Project 1) used a novel hierarchical clustering approach to identify three transcriptional endotypes of asthma (sputum TEA clusters) using sputum microarray data that successfully stratified patients with severe disease characteristics, including inflammation, airway remodeling and severe attacks This represented the first non-invasive transcriptomic stratification of asthma disease severity with the potential to successfully identify high-risk patients and reduce hospitalization. However, the TEA clusters do 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 in a broader and deeper context. To arrive at this goal, we plan to perform RNA-Seq, as well as single-cell measurements with state-of-the-art single-cell technologies including single-cell RNA-Seq and CyTOF, on RNA and cells isolated from sputum from a heterogeneous cohort of individuals with asthma. RNA-Seq is a new but established technology for genome-wide transcriptomic analysis. It has been widely applied for understanding various diseases and enables discovery of gene clusters associated with common functions, as well as identification of 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. At the same time, single-cell technologies can greatly expand upon the sensitivity and cell-type specificity of asthma research. On the transcriptomic level, single-cell RNA-Seq offers 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 and provide the most complete profiling resource to date of a well characterized cohort of asthmatic patients. Single-cell RNASeq transcriptomics will be complemented by single-cell mass cytometry (CyTOF), which provides in depth measurement of protein abundances that define cell activity and function in a lineage specific manner. Currently, CyTOF detects >42 markers of protein abundance measurements and allows us to examine signaling responses within minutes and hours of exposure to relevant antigens, such as house dust-mites. CyTOF is able to probe the response of various cell types and provides a dynamical element to our study.

3. Innovation

Recent efforts have shown that the complex and heterogeneous patterns of asthma can be sub-typed into categories using microarray expression data. This proposed 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 has sufficient resolution to dissect the activities of cells and how they are perturbed in severe disease. The data will be integrated into a model that has the potential to bring personalized medicine to asthma care and provided to the community on a publicly accessible, searchable, integrated asthma MAP.

4. Research Plan

4.A. Plan for Aim 1:Bulk-cell RNA-seq processing pipeline and transcript clustering

4.A.i Rationale

Previous transcriptional analysis of airway cells using microarray technology had some success in clustering endotypes of asthma, but currently RNA-seq has become a standard tool for understanding transcriptional activities in cell populations. Sample collection (Core B) and RNA-Seq (Core C) will be carried out as detailed in other parts of the proposal. We will use our extensive analytic experience in this technique as the foundation to generating asthma clusters and the interpretation of our analyses using single-cell techniques described in Aim 2.

4.A.ii Preliminary results

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

Of critical importance to the shared use of large datasets is uniform processing. In order to provide a resource to the other projects in this proposal and our own clustering aims, we will build an RNA-seq processing pipeline based on the software suite, RSEQtools, that we have largely developed. 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. Also, implemented within RSEQtools are more specialized analysis pipelines that we have developed (e.g. 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?}), as well as thoroughly validated tools such as Bowtie and Tophat \cite{}. These tools are implemented using 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. With this set of tools, we will provide a custom processing pipeline to the Precision Profiling Core C to generate well annotated and consistently processed data to the Driving Projects.

4.A.ii.b: Non-coding RNA (ncRNA) and pseudogene analysis

Other types of transcripts will be important to annotate for analysis of asthma, particularly for the identification of the cell-type signatures that are described in later sections. A fraction of the transcription comes from genomic regions not associated with standard annotations, representing 'non-canonical transcription'. These transcripts are observable even when experimental protocols use poly-A enrichment \cite{}, as will be performed for the samples in this study. A class of non-canonical transcripts of particular interest is the pseudogene, which recent studies have shown are 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. However, the quantification of pseudogene expression is challenging because of the sequence similarity with its 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}.

4.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 ways to understand expression data is clustering analysis. A group of genes in a co-expression cluster have often been demonstrated to be responsible for a common function \cite{}. While there are well known algorithms for expression clustering such as hierarchical clustering, spectral clustering and K-means, we developed several novel methods. 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 of various ncRNAs. This will be the first analysis of this kind for airway transcriptomics.

The functional relationships between co-expressed genes can further be understood in terms of various molecular networks. Over the past decade, we 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 of information flow in these networks. We found that gene-regulatory networks are composed of hierarchical structures dominated by downward information flow and that some transcription factors (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

KKY: SHRINK BY 50%

\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}. ~~This will be the first analysis of this kind for airway transcriptomics.~~

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

We have worked on the development and analysis of multiple RNA-Seq flows in the context of large consortia, including the implementation of tools we developed and other popular tools such as Bowtie and Tophat. 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 Extracellular RNA consortium (http://exrna.org/) 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/).

4.A.iii Approach

4.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 it is here that batch effects and sample processing variation can drive artificial organizations of the data. Technical details to minimize experimental variation are in place, see Core C for details. 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, 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 supercomputing cluster, or through the RSEQtools container image suitable to cloud computing.

4.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.

4.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 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 humans. 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.

4.B Plan for Aim 2: Single-cell analysis of asthma sputum

4.B.i Rationale

Severe asthma is a heterogeneous disease with multiple underlying molecular mechanisms and endotypes. The manifestation of each endotype is the 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 that contribute to the heterogeneity of asthma in patients.

Mutations or expression levels can drive differences in signaling and downstream gene expression in different cell types that can contribute to the overall symptoms of severe asthma. As outlined in Project 2, one relevant pathway in a subset of asthmatic patients may be a Th2 inflammatory response to environmental antigen such as dust mites, that stimulate DKK1, then drives naïve CD4+ T cells towards the Th2 lineage. Th2 cells then secrete IL4, IL5, IL-13 and a variety of pro-inflammatory cytokines that mobilize the response of the immune system, including IL-4-producing follicular T helper cells that produce IgE.Therefore, in depth single cell examination of diverse cell types and their functional responses will provide an in-depth picture of relevant triggers that lead to disease heterogeneity.

In this study, we analyze data generated by Precision Profiling Core C, consisting of high-throughput, multi-dimensional single-cell measurements of gene expression and signaling in sputum cells derived from the airways of patients. By analyzing this mixture of inflammatory and epithelial cell types at the single-cell level we will be able to (1) Dissect the phenotypes of immune and other cell types that are present in cohorts of mild and severe asthmatic patients, with particular power to identify rare phenotypes with large effect; (2) understand signaling logic by utilizing cell-to-cell heterogeneity within each phenotype using single cell functional CyTOF data; and (3) understand gene regulatory network and pathways involved downstream of signaling using single-cell RNA sequencing.

Bulk RNA sequencing as detailed in Aim 1 identifies gene expression from cell samples, and the single-cell technology proposed here has possibility of uncovering the unique transcriptional program of each cell. This will be particular powerful for analysis of samples from the same patient by both platforms. 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

4.B.ii Preliminary Data

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}, and (3) DREVI for characterizing and visualizing relationships between proteins in signaling networks \cite{PMID:25342659}.

One of the advantages of multi-dimensional data is the ability to resolve subtle progression of cell populations within a sample. However, it is hard to directly consider all of the dimensions due to visual and computational problems with high dimensions. For the multidimensional data produced by CyTOF, programs developed for flow cytometry (FlowJo) are not adequate and more advanced software infrastructure is required. Therefore, we developed a dimensionality reduction method known as viSNE cite{PMID: 23685480} that derives an optimal low-dimensional embedding that is able to preserve distances between cells in high-dimensions. This enables the efficient resolution of populations of cells and unsupervised clustering.

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 either 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 to be 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 to 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}.

4.B.iii Approach

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

4.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, and (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 leukemia 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 emerges 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 wire length 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.

4.B.iii.a.1 Subpopulation Characterization and Matching

We propose to find key signaling differences between heterogeneous asthmatic patients and also to identify signaling differences in rare phenotypes to elucidate mechanisms underlying disease and to identify targets for novel therapeutics.

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 cell populations. We propose to develop an approach based on distances between multidimensional distributions in clusters to find matching clusters across individuals with asthma. 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 subject 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 \cite{REFERENCE}.

4.B.iii.a.2 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.

Our goal is to understand how various populations of cells invoke signaling responses to the stimulations detailed in the Precision Profiling Core C.. Cells from the sputum of 6 subjects were tested by stimulation with LPS for 6 hours. Future experiments will involve additional types of stimulation such as PMA, and house dust mite antigen. It has been reported previously \cite{22902532} that triggering the TLR4 receptor on monocytes with its ligand LPS activates several canonical signaling pathways including ERK and NF-KB. 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. Using an unbiased approach, we will curate a panel of signaling pathways from the results of the bulk RNA sequencing data and examine 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.

4.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; and (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 the literature, include (1) debarcoding and error correction, (2) Aligning reads from each unique molecular identifier (UMI), and (3) quantifying the biological noise in genes.

4.B.i.a 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}.

4.B.i.b 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. Previous works (Klein et al., 2015) tend to have very specific recommendations for processing the sequencing such as excluding reads mapping to 400 base pairs distance from end of transcript, or identify a minimal set of genes that explain all reads using the hitting set problem (Klein et al., 2005).

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, probabilistic procedure where each gene is given a probabilistic alignability score that represents how well the collection of reads align to the particular gene. For each read *r,* the probabilistic score incorporates (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; and (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. A starting probabilistic score could be $\Sigma\, PtPg$, the sum of the product of the values for every read that aligns to the gene. This score would need to

be optimized as we generate more data. The UMI would then be assigned to the gene that explains the set with the highest probability.

4.B.i.c 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.

4.B.i.c.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 to some extent. However, we wish to explore a generalized linear model to regress the library size variation against each gene individually, in order to normalize scores in a manner robust to outlier genes that are highly expressed.

4.B.i.c.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%. 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.

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

After the pipeline steps are completed then we can analyze asthma phenotypes, endotypes, 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-impute based on cell clusters and meta-genes, and (3) use the value-imputed data to study gene-gene interactions

4.B.ii.a 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 reduce the number of dimensions non-linearly 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.

4.B.ii.b 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 meta-gene 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.

4.B.ii.c 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 understanding of the molecular mechanisms underlying asthma heterogeneity.

4.C Plan for Aim 3: Integrative clustering and interfacing with other projects

4.C.i Rationale[[DS to improve]]

Each of the methods and data types described in aims 1 and 2 yields unique and valuable information about the biological heterogeneity of asthma. Transforming these analyses into knowledge that can affect patient care requires integrating the data so that each's lessons can be applied to the larger problem. We will integrate the analyses of bulk RNA-seq, single cell RNA-seq and CyTOF measurements with clinical data from Core B to model asthma heterogeneity. This will define the data that best correspond to clinical endotypes in a way that identifies the relevant pathways and identifies potential therapeutic targets.

4.C.ii Preliminary Results

4.C.ii.a Building logical models to characterize clusters

Gene expression is controlled by various gene regulatory factors. Those factors work cooperatively forming a complex regulatory logical circuit on a genome wide scale. 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 and were able to use human ENCODE ChIP-Seq and TCGA RNA-Seq data 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). This will be the first analysis of this kind to address heterogeneity of asthma endotypes.

4.C.ii.b Further experience developing Statistical models of data integration

We have 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 (TF) 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}.

4.C.iii Approach

4.C.iii.a Interrelation with external datasets and creation of the Asthma MAP

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. These resources will permit rapid analysis of the airway signaling landscape and provide valuable detailed understanding of factors contributing to asthma heterogeneity.

Drawing from our experience with other large consortia, we will produce a publicly accessible, searchable, integrated asthma MAP. This website will be populated with links to the data stored on the SRA and in Immport, the custom pipelines developed for the processing of the bulk-cell RNA-seq, single-cell RNA-seq and CyTOF data, as well as the tools used for their analysis. In addition, the clusters produced from the previous aims and as described further in aim 3 will be available and interactive to facilitate data exploration.

4.C.iii.b Deconvolution of cell-type signatures from bulk RNA-seq data

In this aim, we want to identify the cell type signatures in terms of gene expression, and find the gene biomarkers from the signatures that can most discriminate asthma patients; e.g., different TEA clusters. We assume that the mixed effects from various related cell types determine the gene expression from each patient's sputum; i.e., mixtures of various cell type signatures. We then try to use both linear and nonlinear approaches to capture the mixed effects as follows.

We first try the linear models that will be computationally efficient. Given the gene expression levels and cell type fractions for each patient, we can use a linear matrix model to identify cell type gene expression signatures. For instance, the patient's ith gene expression level can be modeled as a linear superposition of the same gene's expression levels of multiple cell type signatures; i.e., 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 Precision Profiling 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.

We then try to apply advanced models to capture nonlinear effects from different cells to gene expression. For example, we can use machine-learning methods to investigate the gene markers from cell type

gene expression signatures for both bulk data and single-cell type. In particular, we would like to use the Denoising Autoencoder (DA), an unsupervised machine-learning framework to extract and characterize cell type signatures. DA is able to discover non-linear expression features from gene expression data using sigmoid transformation. We will apply DA to different patients clusters and compare their non-linear features, and find the genes that have features to most discriminate clusters.

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.

4.C.iii.c Identification of clinical and CyTOF features of clusters and cell type signatures

Clinical information can be used to classify endotypes of asthma and provide valuable guidelines for diagnosis. Some features like FEV1/FVC have been widely used in endotype clustering. However, the quantitative link between the gene clusters and clinical variables is largely unknown. As some clinical variables represent a certain kind of phenotype of asthma and are reflected by distinct syndromes, distinguishing genes or pathways associated with distinct clinical features may identify targets for novel therapeutic approaches. We will build a regression or classification model using highly scored gene signatures, submodules and pathways in different clusters as the predictor, and clinical information as the target. By means of information gain or gini index, we will characterize the highest associated factors for each clinical phenotype. Finally, we will build a functional representation cluster of clinical variables.

In collaboration with Project 2, three asthma associated pathways will be used to validate and extend the cell responses and gene signatures characterized by CyTOF and RNA-Seq. We will begin with the experimentally validated pathways and expand to the whole network, using belief propagation based on experimental results to update and optimize the weight between gene-gene interaction edges. We will apply the Orthoclust framework to identify common and specific regulation or signaling pathways for different cell types and endotypes. Specific modules in signaling response pathways from CyTOF and logic gate analysis will address the dynamic regulation and cascaded signaling transduction in asthma heterogeneity.

4.C.iii.d Logical model-building

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 is a mechanism at the molecular level, and follows certain logical behaviors to give rise to the phenotypes. We plan to use logical modeling approaches to identify gene regulatory logics to characterize the asthma clusters such as severe and mild 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 is disrupted, it can give rise to abnormal gene expression such as those present in asthma. In many cases, the regulatory factors controlling gene expression behave in a discrete fashion and can be modeled using Boolean logical models [147-153]. Additionally, the simple binary operations in the Boolean model do not need large amounts of data and are very computationally efficient. Therefore, we will develop computational algorithms based on Boolean models to study and compare the cooperative logics between various regulatory factors. 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 logics for regulatory modules. Furthermore, we will connect logic circuits for all regulatory modules to build a Boolean regulatory network at system level. Last, we will analyze the Boolean network to predict novel regulatory pathways, and identify asthma cluster's specific pathway logics.

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 integrate data on TF binding from the asthma-related cell types such as eosinophils, lymphocytes, and neutrophils from the ENCODE project and Epigenomics Roadmaps [16, 55]. Second, given a cluster, we want to identify the regulatory logics in the constructed gene regulatory network to drive the cluster's expression patterns. We will use data from regulatory networks 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 look at gene regulatory modules; e.g., the simple triplets consisting of two regulatory factors (RFs) and a common target gene T. 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, the triplet (RF1, RF2, T), where RF1 and RF2 regulate a gene T, follows an AND logic; i.e., both RF1 and RF2 need to express high to turn on the gene T.

In addition, we will also find the logic circuits consisting of the cascaded 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, (RF1, RF2, T) may follow AND logic in severe asthma patients, but OR logic in mild patients. We will also assess the changes of regulatory logics of the same biological pathways across clusters. In addition to identifying logics, these studies may predict solutions that may guide iterative in vitro studies such as gene knockdowns to modulate the regulatory logics.

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

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

This research will include 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 C 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 inform the other driving projects use data from them. For example, Project 1 Aim 3 will use the IP-clusters from Aim 3 to determine cell activities in stimulation assays, and project 2B will be informed by our logic gate regulation of proteins such as DKK1. We will use the data generated from project 2C's patients' microbiomes with information about which organisms are coated with IgA. One approach to integrating these data with RNA-seq is to use the microbiome patient clusters as seeds for our IG-clusters, which will reveal how patient groups separated by their microbial communities are responding differently in their transcriptional activities. These findings will be communicated in monthly meetings of the group and more frequent interactions between subgroups and will form the basis for iterative investigation.

5. Project Deliverables

The deliverables from this project will be clinically informative clusters of genes, cells and patients that characterize different asthma phenotypes. These clusters, detailed below, are described in Table 1 and speak to a variety of hypotheses from ours and the other projects. The tools and results will be made available to the other members of this project and the research community in a publicly accessible, searchable, integrated asthma MAP website http://asthmaMAP.gersteinlab.org, as we have done for other multi-investigator research efforts (e.g. https://www.encodeproject.org/comparative/). This website will serve as a repository for the pipelines, derived datasets and analyses that are the deliverables from each aim of this research proposal.

Aim 1 will produce 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 BP and BG-clusters using global transcription and co-expression of the genes, respectively. Non-coding RNAs, psuedogenes and other transcripts will be mapped onto BG clusters to suggest possible functions. This unrefined clustering will speak to the global transcriptional activity of the sputum and will be the framework refined by integrating other methods. The BP-clusters, where each patient is clustered by his or her global transcription, will define asthma endotypes similarly to methods used to generate TEA clusters in previous reports and will speak to whether the response from RNA-seq is similar to previous work using microarrays.

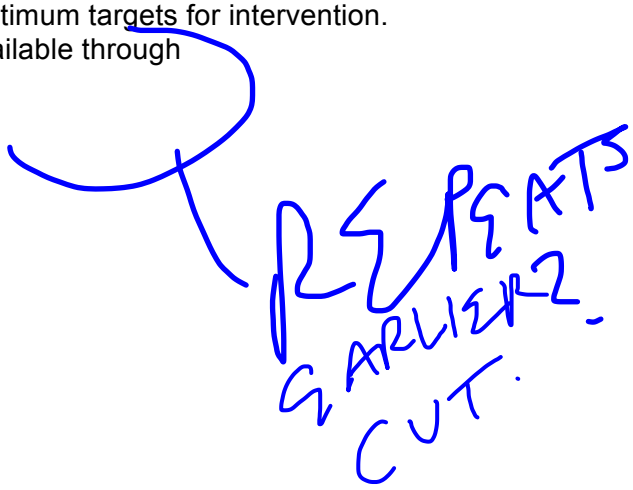
Aim 2 will produce software and pipelines for the analysis of both CyTOF and single-cell RNA-sequencing data as well as results of the analysis of data generated by the Precision Profiling core. Each data type will be used to generate clusters by cell signatures and gene networks. The CyTOF analysis will generate CC-clusters from unsupervised clustering of surface markers. These clusters will define the cell population with a new level of precision, including stratifying lymphocytes into component cells types including Th2 and Tfh cells, as is being explored in detail by Project 2. We will produce a method that identifies these subpopulations of cells and tracks changes in their abundances across patients and across the longitudinal sampling of individuals. Moreover, the CyTOF data will be used to generate CPr-clusters, which will define signaling interactions in and between cells using DREVI and fuzzy logic methods.

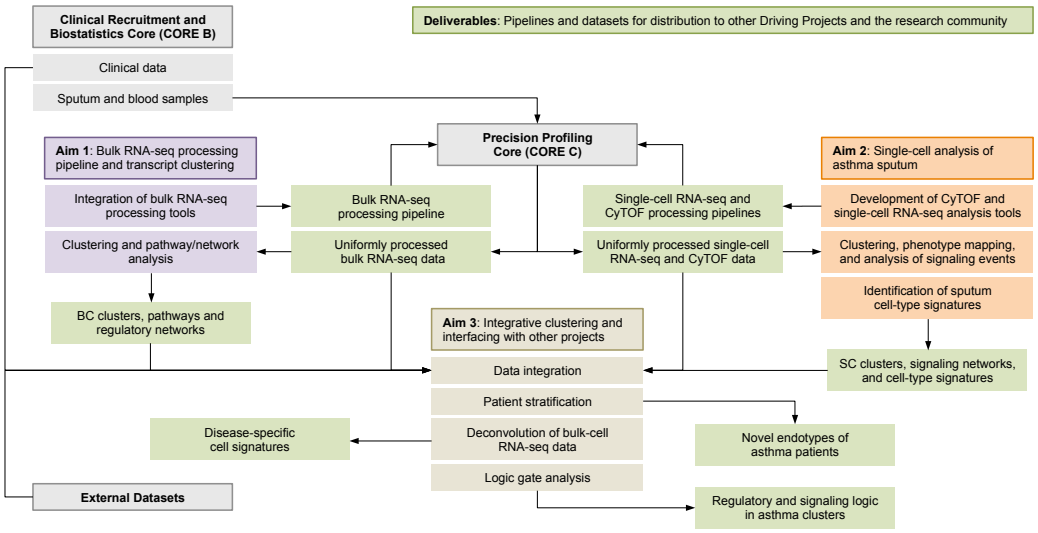
The single-cell RNA-seq data processing will include debarcoding, quantifying noise, and imputing missing values from low abundance genes. The processed data will deliver SC-clusters that define

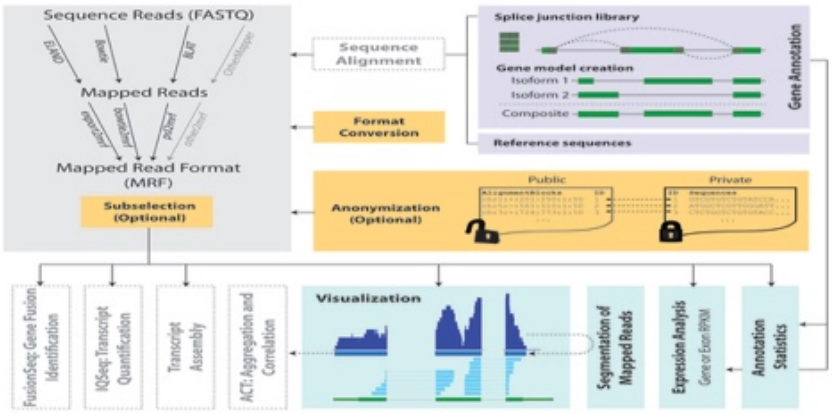
subpopulations of cells by their transcriptional activities. At the gene level, SG-clusters will identify co-expression networks within specific cell types and generate an output of a DREMI analysis of gene-gene interactions and resultant gene modules.

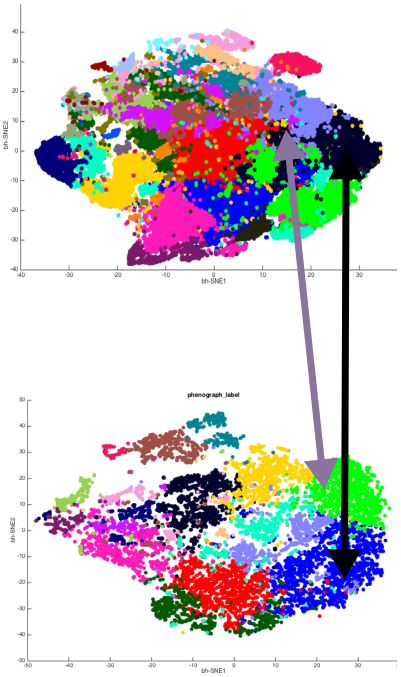
Aim 3 will integrate the above data with clinical data from Core B, the data generated by other Driving Projects and external datasets to produce integrative clusters by patients, cells and genes (IP, IC and IG-clusters). Bulk-cell RNA sequencing data that has been deconvolved with single-cell RNA-seq and CyTOF cell signature data will be merged with other data such as Driving Project 2C's microbiome clusters to yield Patient-level IP-clusters to define novel asthma endotypes. The cell-level IC-clusters will show the populations of cells that are important for disease. Gene-level IG-clusters will define the mechanisms by which those cell populations are different. Specifically, we will produce logic gate models for the different IG-clusters to define the regulatory logic in each asthma subpopulation and identify the optimum targets for intervention.

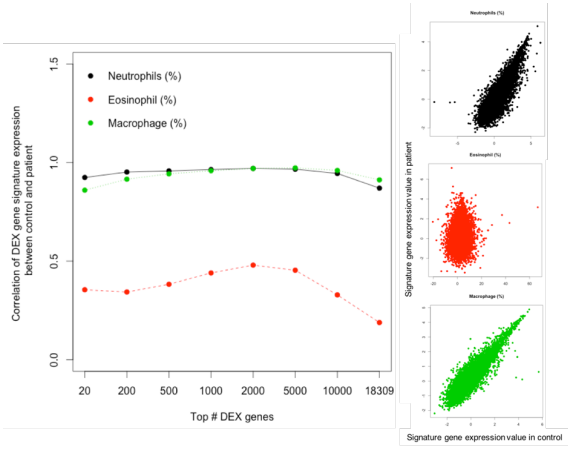
The suite of tools and analyses defined here will be made available through http://asthmaMAP.gersteinlab.org for the research community.











Aim	Cluster Name	Data	Clustering Method	Utility
1	BP	Bulk-cell RNA-seq	by Patient	Novel endotypes of asthma by RNA-seq
	BG		by Gene	Gene networks for understanding novel endotypes
2	SC	Single-cell RNA-seq	by Cell	Cell signatures by transcription
	SG		by Gene	Consistent expression between cells
	CC	Single-cell CyTOF	by Cell	Cell signatures by protein levels
	CPr		by Protein	Signaling network analysis
3	IP	All integrated	by Patient	Novel endotypes of asthma by integrated analysis
	IC		by Cell	Cell type signatures of asthma
	IG		by Gene	Logic modeling for disease mechanism

6. References

\bibliography{}

Figure caption. This figure shows the results of unsupervised clustering using the Phenograph software on two patient sputum samples. The multidimensional CyTOF measurements are reduced to two dimensions using the tSNE algorithm and each cell is rendered as a point in this space. Additionally, the color given to each point indicates the cluster to which the cell belongs. Two matching clusters are shown using arrows. One of the clusters represents an eosinophil population and another represents a neutrophil population, matched using distribution distances (XXX. will this part of the figure be shown).