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

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

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 its heterogeneity and differential severity. In particular, we will model the effects of IgA and IgE-mediated adaptive responses on specific cell types. This will include the effects of YKL-40 and DKK1 levels, which have been shown to be correlated with lung remodeling through the WNT signaling pathway. These analyses will offer used to generate novel insight into transcripts and pathways associated with asthma lung dysfunction and direct validation experiments by the other Driving Projects. We will use our expertise in RNA sequencing to develop and distribute processing pipelines to the Precision Profiling Core for both bulk-cell and single-cell RNA-Seq experiments. We will process, analyze and interpret these data to make predictions of pathways important to asthma. These analyses will inform targets for single-cell protein-level experiments (CyTOF) to give insight into the signaling pathways and activities of proteins that have been predicted to important roles (e.g. network hubs and interaction nodes). These data will be integrated to model the transcripts, proteins and pathways that determine asthma severity and give a new understanding of the mechanisms by which patients' asthma experiences vary. Further, these models will inform, direct and iteratively learn from experiments performed by the other Driving Projects in this proposal to define the causative pathways for near-fatal, severe and mild asthma.

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 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. This will include the effects of YKL-40 and DKK1 levels, which have been shown to be correlated with lung remodeling through the WNT signaling pathway. Building on recent work showing transcriptional clustering by clinical phenotypes, 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. We will use our expertise in RNA sequencing to develop and distribute processing pipelines to the Precision Profiling Core for both bulk-cell and single-cell RNA-Seq experiments. We will process, analyze and interpret these data to make predictions of pathways important to asthma. These analyses will inform targets for single-cell protein-level experiments (CyTOF) to give insight into the signaling pathways and activities of proteins that have been predicted to important roles (e.g. network hubs and interaction nodes). These data will be integrated to model the transcripts, proteins and pathways that determine asthma severity and give a new understanding of the mechanisms by which patients' asthma experiences vary. Further, these models will inform, direct and iteratively learn from experiments performed by the other Driving Projects in this proposal to define the causative pathways for near-fatal, severe and mild asthma.

Aim 1: Develop Bulk RNAseq processing tools and identify transcripts critical to the pathophysiologic heterogeneity of asthma

We will adapt a comprehensive suite of human RNA-Seq tools to process the 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 and see the extent to which they correspond to annotated portions of the genome. We will identify ncRNAs & txn pgenes... These deconvoluted data will be used to build clusters and networks to identify transcripts and pathways that distinguish samples from patients with different asthma severities. Critical features of the networks, especially those with predicted importance in signaling, will be targeted for CyTOF analysis in Aim 2.

Aim 2: Single-cell Analysis of Asthma Sputum

[[Smita to write]]

Aim 3: Integrative model building

We will use the cell-assigned single-cell RNA-seq as sputum-specific cell-type signatures that can be used to deconvolve the bulk-cell RNA-seq data to its component cell transcripts. These deconvolved data will be integrated with the regulatory networks for each cell type to model the logical gate structure evident in the data. [[Smita to write rest]]

2. Significance

Asthma is a chronic inflammatory disease of the airways which afflicts ~7% of the U.S. population.\cite{Moorman JE, Zahran H, Truman BI, Molla MT. Current asthma prevalence - United States, 2006-2008. MMWR Surveill Summ 2011;60 Suppl:84-6} 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 that are independent of disease activity or cell type in the sputum. This represented the first non-invasive stratification of asthma disease severity with the potential to successfully identify high-risk patients and reduce hospitalization. Moreover, the TEA clusters give clues to the molecular mechanisms by which individuals respond differently. The methods described in this project greatly expand upon the sensitivity, dynamic range and cell-type specificity (either by single-cell sequencing or deconvolution) of this research, which wikk more clearly define the molecular mechanisms that drive severe disease. These data are critical to pass beyond reactive medicine in asthma treatment to more predictive methods with specific cellular mechanisms that can be targeted for therapies.

3. Innovation

Asthma is a heterogeneous disease and we lack a clear understanding of the causes and mechanisms that underlie this heterogeneity. The primary deliverable of this driving project will be a model that may elucidate these causes and mechanisms using a novel combination of data that is uniquely able to address questions of signaling network function and disease-associated variation. Our demonstrated expertise in bulk-cell RNA-seq will be leveraged to the powerful emerging technologies of single cell RNA-seq and CyTOF. Not only will these single-cell measurements facilitate the deconvolution of the bulk measurements into its component cell signals, but will also speak to the cell-type variability that will be critical to our modeling efforts. Moreover, the combination of these methods will identify proteins of central importance for the signaling network, which can then be assayed and quantified using CyTOF. Because the active signaling molecule (e.g. phosphorylated form) can be quantified directly, we will have unprecedented resolution into the active signaling network and its effects in asthma. The combination of the two RNA-seq experiment types with CyTOF data will provide a high quality model that will direct and illuminate the experimentation of the other two driving projects and will be of great use to identifying appropriate action to affect patient care.

4. Research Plan

4.A. Plan for Aim 1 [DS, KKY + TG]

4.A.i Rationale

By uniform samples processing and extensive genome wide data integration, we aim to develop an asthma resource for identifying novel asthma-related genetic elements.

4.A.ii Preliminary results

4.A.ii.a Application of RNAseq processing tools

The Gerstein lab has developed a number of tools and data formats to handle the increasingly large quantities for data generated by RNA-Seq experiments. For example, sequence reads from a specific individual can contain sufficient information to potentially identify and genetically characterize that person, raising privacy concerns. In order to address these issues, 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, while allowing one to still carry out many functional genomics studies. We have developed a suite of tools (RSEQtools) that use this format for the analysis of RNA-Seq experiments. 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. In addition to the anonymization afforded by MRF, this format also facilitates the decoupling of the alignment of

reads from downstream analyses. RSEQtools is implemented in C and the source code is available at http://rseqtools.gersteinlab.org/. Along with RSEQtools, we have developed three different analysis pipelines including FusionSeq for fusion transcript detection, IQSeq for transcript quantification, and DupSeq for analyzing expression patterns of highly homologous genomic regions. They were all assembled into the SEQtools framework for customizable workflow.

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

The advance of RNA-Seq greatly enhances the discovery of novel transcripts. To investigate newly transcriptionally active regions further, we developed incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a gold standard training set. We also developed Pseudo-seq, which addresses the issue of quantification of pseudogene expression, which is difficult to separate from the transcription of parent genes with similar sequences. Pseudoseq solves this problem by calculating the expression in terms of RPKM for pseudogenes by focusing only on those reads and regions that are uniquely mappable [16]. Though pseudogenes have long been thought to be non-functional, recent studies have revealed their roles in cancer, X-chromosome inactivation and intercellular signaling [17-19].

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

The Gerstein lab has 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 such as hierarchical clustering, spectral clustering and K-means, for expression clustering, we have been focusing on the development of novel methods for specific purposes. 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 data across multiple species.\cite{25249401} This integrates co-association networks of individual species utilizing the orthology relationships of genes between species. The essence of OrthoClust is 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) (Yu et al., 2004a, Yu et al., 2007). 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 (Cheng et al., 2011b; Gerstein et al., 2012; Gerstein et al., 2010; Yan et al., 2010). Gerstein lab has been developing tools for comparing networks. For instance, we introduced a framework to quantify differences between networks and by comparing matching networks across organisms, found a consistent ordering of rewiring rates of different network types. \cite{21253555} .

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

The Gerstein lab has extensive experience in RNA-Seq pipeline development and analysis \cite{19015660}. In particular, we continue to play a major 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} 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}, and the PsychENCODE project (http://psychencode.org/).

4.A.iii Approach

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

We plan to process bulk RNA-Seq samples in a uniform fashion based on RSEQtools we developed. Moreover, as the components of RSEQtools can readily be assembled and extended to build customizable RNA-Seq workflows. To this end, additional components like samples deconvolution and single cell analysis can be easily incorporated into the pipeline. In addition to 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}.

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{}) 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 compare the identified antisense transcription between different regions of the brain and identify the differentially regulated transcripts.

Although pseudogenes have long been considered as nonfunctional genomic loci, recent studies have shown that, in some cases, pseudogenes are not only transcribed, but perform crucial regulatory roles via their mature RNA products \cite{}. In our previous work, we have shown that pseudogene transcription exhibits tissue specificity103. These observations imply that pseudogenes could have unique biological activities in the sputum.

We propose to identify the transcriptional activity for each pseudogene annotation using the following protocol. First, we will remove pseudogene regions with mappability lower than 1. Second, we will discard the pseudogene regions shorter than 100 nucleotides after the mappability filtering. Only RNA-seq reads mapped to the remaining unique regions will be used to compute a normalized expression value (RPKM). Next, given previously published results on human pseudogenes with small-scale validation 102,103, which imply that ~15% of human pseudogenes are transcribed, we can set an RPKM threshold for human analysis such that it gives anapproximate agreement with the previous validation. With the assumption that the transcription of protein coding genes in human, chimpanzee, and macaque samples have similar distributions, we may apply quantile normalization on the gene transcription data for chimpanzee and macaque samples, using human as a reference. We will then apply this normalization to the pseudogene transcription data, and consistently apply the human threshold across the three species.

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 perform various clustering and network analysis. We will employ various clustering algorithms to group transcripts based on their expression profiles. The clusters will further be validated using biological features such as sequence similarity, genomic distance, and co-regulation. This clustering will allow us to interrelate cellular mRNAs and potentially determine the cellular origin or mechanisms of action. We will perform cross-validation on the robustness of our clusters. 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.

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.

We plan to extend our previous work on cross-species network comparison to compare networks constructed by using samples from patents 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, a cross species clustering algorithm can be used to study specific modules contributed to asthma.

4.A.iii.d Interrelation with external datasets

We will integrate all the RNA-Seq with ENCODE data. By mapping ENCODE annotations onto transcripts, we can look for common patterns of TF regulation and chromatin state to provide insights into regulation. 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.

4.A.iv Deliverables

The primary deliverable from this aim is the rigorous and uniform processing of bulk RNA-seq data from patients. De-identified data will be made available to the other contributors to this proposal through a website interface, as we have done for many other projects (e.g. dart.gersteinlab.org, pseudogene.org). In addition, detailed annotation of the transcripts including structural information, ncRNAs and psuedogenes will be included. Moreover, all clusters and networks will be available, including those through which external datasets were interrelated.

4.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.

4.B Plan for Aim 2

4.B.i Rationale

Severe asthma is a heterogeneous disease with multiple underlying molecular 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 where 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 its progresses.

In this study, we perform 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.

The key advantage of using single-cell technology over bulk technology is that bulk samples give an average reading for each protein or gene being measured. For instance, in standard bulk RNA-sequencing, the mRNA molecules from the entire sample is collected together and sequenced, such that cell-to-cell differences are obscured. However, in single-cell data, the unique transcriptional program of each cell can be uncovered, and differences between cells can be informative of the underlying relationship or network between proteins and genes. At a fundamental level 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.

4.B.ii Preliminary Data

Cells from the sputum of 6 subjects was tested using a minimal panel of 15 surface markers and 2 cytokine antibodies. The samples were stimulated with LPS for 6 hours and resulted in a significant effect for TNF-alpha, indicating that the stimulus elicited the desired effect.
Ruth says more to come …

Preliminary single-cell RNA sequencing data is not yet available.

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

CyTOF data consists of dozens of dimensions (around 45 presently) of protein abundance measurements. Cells are fixed and permeabilized, and then proteins are tagged by antibodies that are chelated to rare metal ions. These tagged cells are then nebulized and sent through a plasma chamber where the contents of the cell are vaporized in a plasma chamber and the rare metal isotopes are sent through a time-of-flight mass spectrometer that measures their abundance by mass. These mass measurements then correspond to the abundance of the associated proteins. The tagged proteins can be of different varieties including (i) surface markers, (ii) internal proteins tagged by either phosphorylated or unphosphorylated residues, (iii) transcription factors, and (iv) cleaved proteins. We plan to utilize information from bulk RNA-sequencing in order to curate a protein-measurement panel for CyTOF that contains pathways that were upregulated in terms of gene expression.

CyTOF allows us to examine signaling responses within minutes and hours of exposure to antigen along these pathways. The start of an asthmatic attack is over-reaction by members of the innate immune system to household antigen. For example, dendritic cells in normal patients are not sensitive to dust mites, but macrophages in severely asthmatic patients are sensitive, recognize household antigen, and present them to members of the innate immune system. The profiling core in this case will measure response with respect to stimulation by dust-mites and also inflammatory cytokines after which, we can analyze the signaling responses in various cell types through CyTOF.
Since CyTOF data is high-throughput, high-dimensional, and single-cell it is necessary to utilize and develop sophisticated algorithmic techniques that can handle the stochasticity, and dimensionality of the data. We propose to develop the following techniques for understanding asthmatic response from CyTOF data:

4.B.iii.a.1 Clustering and phenotype mapping

One of the advantages of single-cell data is that new populations of cells can be discovered by multidimensional analysis of a large set of surface marker proteins. A promising approach for discovering new populations is unsupervised clustering, where the type or number of clusters is left unspecified. 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 (Cell et al. 2015) 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 2.

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, attempt 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.

4.B.iii.a.2. Analysis of signaling events using DREMI and DREVI

Once the clusters or phenotypes of cells are established then we can gauge signaling response within each cluster by utilizing our previously developed information theoretic techniques for analyzing signaling interactions, DREMI and DREVI. These methods characterize relationships in signaling networks by quantifying the strengths of network edges and deriving signaling response functions [ref]. 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.

[Figure 3 here]

We recently extended DREMI and DREVI to higher dimensions where multiple parent molecules are allowed for each child molecule, i.e., models multi-molecule interactions:

[Figure 4 here]

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.

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.

4.B.iii.b Single-cell RNA-sequencing Analysis

Single-cell RNA-seq is a powerful method for uncovering the relationships between genes. Unlike CyTOF, RNA-sequencing is able to offer an unbiased measurement of the entire complement of mRNA transcripts produced in each cell. However, this information is fundamentally limited by the inherent sparsity of the data---the data has far fewer cell measurements currently than the number of dimensions. Additionally, current single-cell RNA sequencing technologies tend to be limited both in terms of the number of cells (hundreds) and in terms of the sampling of transcripts within each cell due to the "dropout" effects of low-abundance genes. Hence, each dimension is in some sense less representative of the true gene-gene relationship than in higher throughput data, and in some sense less trustworthy.

In order to tackle this problem, we propose to non-linearly reduce the number of dimensions by utilizing a method such as bh-SNE [ref] or non-linear PCA [ref]. 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. Each metagene can then impute missing values from its co-cluster members. Then DREMI, DREVI and other information theoretic approaches can be computed on the metagenes in order to understand how different pathways interact with each other. This can help uncover novel pathways that are involved and differentially regulated in severe asthma robustly from non-robust data.

[Figure 5 here?]

[[some text from Mark's cegsbrain grant, if it would helpful to adapt…]]

Given the heterogeneous nature of cell populations throughout the sputum, we will exploit the single-cell RNA-seq data in neurons and glial cells to quantify gene expression estimates in each cell and assess consistency of expression across cells within a tissue, between regions, and within species for mRNAs, long non-coding RNAs, and pseudogenes. Pseudogenes especially, when combined with their parent-gene transcription signals, may serve as useful biomarkers to distinguish different cell types. It has also been reported that despite their low abundance, pseudogenes and ncRNAs exhibit a greater degree of cell-type specific expression than mRNAs17. Compared to the tissue-level, single-cell expression data are more uniquely suited to the detection of poorly expressed RNAs because it is possible to distinguish stochastic experimental noise (resulting from sample preparation and data processing) from genes that are expressed only in a small fraction of the total number of cells. Preliminary data obtained from 27 cells acutely isolated mouse embryonic neocortex (unpublished data) show that even though protein coding mRNAs are best represented and the highest-expressed, the level of pseudogenes and other ncRNAs is sufficient for a meaningful comparison (Figure 6).

4.B.iii.c Logic Modeling

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 are above a threshold, then the level of pGSK3b changes to a higher state. 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.

Hence, using these basic logical modes, combined with a stochastic noise model, we propose to encapsulate protein and gene interactions in a computational efficient logic model.

4.B.iv Deliverables

1. Software that clusters and matches clusters between patients.
2. Software that simulates the logical models and analyzes network differences.
3. Analysis subpopulations in Asthma
4. Analysis of signaling within the subpopulations in Asthma

4.B.v Pitfalls

Aim 2. CyTOF analysis of sputum (3 pages) (Contributed by Smita Krishnaswamy)

1) taking the clusters and selecting a CyTOF
2) .75 page CyTOF proc from the orange machine to a datafile
3) cytof clustering & trajectory mapping
4) signalling network analysis [1-2pg]

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

==========

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

4.C.i Rationale

We would like to integrate the single cell and bulk data to generate signatures of cells and integrative models of the ....

4.C.ii Preliminary Results [600 words]

4.C.ii.a Integration of regulatory network data

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. 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).

4.C.ii.b Integration with of RNA-seq with mass spectrometry data

- CRIT & fungus paper - grab: sbfuel

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.)

4.C.ii.c Statistical models of data integration

grab text for statistical models

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}.

4.C.iii Approach

[[DW’s bullets]]

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

4.C.iii.a 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 the signatures that can most discriminate asthma patients and controls. 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. We will start from the linear model 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.

In Aim 1, we already measured X and W, 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 will apply this model to different groups of patients such as the TEA clusters that were previously identified and that will be explored further in Driving Project 1 and the novel clusters identified in Aim 1, and find the cluster's cell type signatures. We will compare those signatures and identify ones along with its gene biomarkers that most discriminate clusters.

We will also try to use the advanced models such as X=f(S,W) where f is a nonlinear function if they can better discriminate clusters.

4.C.iii.b Comparison of cell type signatures across clusters and public signatures and the identification of cluster-specific biomarkers / signatures

We recognized the TEA cluster was identified based known KEGG pathway, one of the biggest gene knowledge base. But, there are still chances to be limited in a subset of all genes, which might ignore the de novo identification of gene markers. However, discovery based on the whole transcriptome is also a big challenge. It will not only require a huge computing resource if doing global hierarchical clustering, but high noise introduced by uncorrelated genes and bias by highest effect genes will fail to identify clusters. To tackle this, we adopted a deep learning framework to globally investigate the gene markers for both bulk data and single-cell type.
Denosing Autoencoder(DA) is an unsupervised machine learning framework that be used to extract and characterize gene signatures involved in new pathway and principals. DA incorporating nose during training, and can be used to learn compact and representative features from unlabeled data.

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 defined latent representative or code node, denoted by y with dimension N\prime (N\prime << N). We mapped x to y by a non-linear sigmoid function, which we call it encoding step. 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. And then we decode y and reconstruct z to be the same shape of x using similar transformation. The reconstruction error is estimated by cross-entropy.

y=sigmod(Wx+b)
z=sigmod(W^T y+b\prime)

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 descendant algorithm is 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.

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

4.C.iii.d. Logical modelling building [250 words - dW - para]

In addition to the statistical/machine learning approaches described in the Preliminary Results and above, the Gerstein lab will also explore a logical modeling approach to identify gene regulatory logics for asthma. 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 corrupts, they 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 patient cluster's specific regulatory logical pathways.

4.C.iii.e Logic gates analysis of gene expression [DW][1pg]

We plan to identify the gene regulatory logics based on logic-gate models above for different asthma patient 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. As described in the preliminary results, we can use Target Identification from Profiles (TIP) to ChIP-seq and chromatin accessibility data generated by the proposed research program to identify the genes targeted by TFs for the cell types associated with asthma such as Eosinophils, Lymphocytes, Blood and Neutrophils. In order to define a more complete set of TF-gene interactions, we will combine these data with data on TF binding using the same cell lines previously published by the ENCODE project and Epigenomics Roadmaps and described in other studies [16, 55]. We will also use TF protein expression data, along with corresponding published PWM motifs and DNase-seq data, to infer potential binding.

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 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 investigate different options for computing a gate consistency score. For example, for any triplet with m binary input samples and any gate g we could compute the gate consistency score as (n1 +n2 +n3 +n4)/m, where ni is the number of vectors matching one of the possible input/output combinations. We will also study different ways of assessing the statistical significance of the chosen gate. A possible option would be to perform a permutation test, where by repeatedly randomizing the target gene we would estimate the chance of obtaining the same logic gate by chance. If for a given triplet, we find a logic gate that satisfies our criteria, we will call it a consistent logic gate for that triplet. Otherwise, if for a given triplet no consistent logic gate is found, this would suggest that either the activity relationship between the two RFs cannot be described by our model (i.e. the target gene is regulated by more than two RFs) or the available regulatory and gene expression data are too noisy. 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.

Finally, after finding the regulatory logics for various patient 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.

4.C.iii.f Logic gates analysis of gene expression and CyTOF [1pg]

4.C.iii.g Development of a hierarchical cell-cell communication model

4.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

4.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).

5. References

\bibliography{}

6. Budget

180-190K
Smita in this budget
may backload