AN INTEGRATIVE COMPUTATIONAL FRAMEWORK FOR DEFINING ASTHMA ENDOTYPES

by

J. A. Howrylak, MD

Submitted to the Graduate Faculty of
the University of Pittsburgh School of Medicine, Department of
Computational Biology in partial fulfillment
of the requirements for the degree of
M.D., University of Michigan Medical School, 2003

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF COMPUTATIONAL BIOLOGY

This dissertation was presented

by

J. A. Howrylak, MD

It was defended on

November 22nd 2013

and approved by

Naftali Kaminski, Department of Medicine, Pulmonary, Critical Care and Sleep Medicine, Yale University

Eric P. Xing, Department of Computer Science, Machine Learning, Language Technology Institute School of Computer Science, Carnegie Mellon University

Panayiotis V. Benos, Department of Computational Biology University of Pittsburgh School of Medicine

Benjamin A. Raby, Department of Medicine, Channing Division of Network Medicine, Division of Pulmonary and Critical Care Medicine Brigham and Women's Hospital, Harvard Medical School

Augustine M.K. Choi, Department of Medicine Weill Cornell Medical College

Dissertation Advisors: Naftali Kaminski, Department of Medicine, Pulmonary, Critical Care and Sleep Medicine, Yale University,

Eric P. Xing, School of Computer Science, Carnegie Mellon University, Department of Computer Science, Machine Learning, Language Technology Institute

AN INTEGRATIVE COMPUTATIONAL FRAMEWORK FOR DEFINING ASTHMA ENDOTYPES

J. A. Howrylak, MD, PhD

University of Pittsburgh, 2013

The rapid pace of drug development in recent years has led to the recognition that new pharmacotherapies do not have the same effect on all patients. This is particularly true in the case of complex common diseases such as hypertension, diabetes and asthma, where a diversity of pathogenetic factors may interact to produce the same disease, resulting in a large degree of heterogeneity in the response to medical therapy. For this reason, the ability to differentiate between different disease endotypes is of increasing importance to clinical medicine.

In the case of asthma, initial studies have hinted at the presence of multiple disease endotypes with different clinical characteristics. Additional studies have identified novel genetic risk factors and differences in gene expression among asthmatic patients with different disease endotypes. Despite the presence of large-scale clinical and molecular datasets from asthmatic patients, limited efforts have been made to integrate these different formats to develop a systems-level understanding of disease mechanism.

In this thesis, we develop a computational framework for addressing the problem of disease heterogeneity by integrating data from multiple sources, including the genome, phenome and transcriptome in order to define clinically-relevant disease subtypes, and we demonstrate its application in a cohort of asthmatic children. First we perform a cluster analysis of clinical phenotypic data and detect the presence of multiple disease endotypes in a cohort of children with mild-to-moderate asthma. We evaluate the clinical significance of these endotypes by demonstrating their longtudinal stability and association with differential response

to pharmacotherapy. Next, we develop a transcriptional network from the gene expression profiles of these patients and identify the relationship between discrete patterns of expression and asthma endotypes. Finally, we combine longitudinally-derived clinical phenotypes with genetic data to uncover novel genetic associations corresponding to changes in gene expression and the expression of longitudinal clinical traits.

Keywords: computational biology, genomics, bioinformatics, machine learning, cluster analysis, genome-wide association study, gene expression profiling, disease endotypes, childhood asthma.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 MOTIVATION

The rapid development of computational methods in the field of statistical machine learning has provided us with the ability to recognize the distinctive patterns inherent in large electronic datasets, with applicability to fields as diverse as high finance, text mining and biomedicine. However, the complexity of biological systems, and in particular the human body, has made it difficult to link disease patterns to systems-level pathogenesis until very recently. There have been several high-profile developments in both computational methods and biomedical applications that have led to the unprecedented opportunity to leverage massive amounts of electronic medical data for widespread clinical use. For example, the electronic medical record (EMR) makes it possible to perform large-scale data mining of patient clinical characteristics. Similarly, the development of high-throughput technologies to assay the human genome and transcriptome through next generation sequencing and oligonucleotide microarrays provide a wealth of genetic and genomic data. Co-incident with these biomedical advances has been development of powerful data-mining algorithms, made possible due to the increased computational ability of modern computers. The integration of these algorithms with the vast array of biomedical data creates the ability to develop detailed clinical and molecular profiles for individual patients that could be accessed through the EMR.

One application for advanced data-mining techniques has been in disease phenotyping, where cluster analysis has recently become popular [169] in a model-free setting [136]. Clustering techniques were initially described as an application for disease diagnosis over 20 years ago [50], however such methods were not successfully applied to large clinical datasets until

recent advances in both computing power and algorithmic efficiency made these methods more scaleable, and their use more widely applicable. Eisen and colleagues were the first to use clustering as a method of disease phenotyping [46]. Many subsequent papers have used a similar approach to disease phenotyping, most notably for identifying asthma endophenotypes [60, 118, 48], and all have been primarily descriptive in nature, using clustering as a form of exploratory data analysis. The widespread adoption of these methods in the clinical setting has been limited for several reasons, including the problems of quality control and cluster validation, which make it difficult to move beyond pattern recognition toward using the results of cluster analysis for risk-stratification and clinical decision-making [85].

An additional application for machine learning methods is relating genetic associations to disease phenotypes. In recent years, much progress has been made toward understanding the genetic underpinnings of complex diseases, such as asthma. For asthma, studies of disease concordance among twins suggesting that asthma is a highly heritable condition [43, 73, 35, 163, 175, 42, 36, 126, 127, 99, 154]. This observation has led to a large-scale interrogation of the human genome, resulting in the discovery of novel asthma-related genes, replicated in multiple populations [143, 69, 47, 71, 117, 147, 63]. Yet, despite the preponderance of evidence for the heritability of asthma, the previous genome-wide association studies (GWAS) have determined only a small fraction of the total estimated heritability. A major issue leading to this missing heritability is related to the fact that asthma is an enormously dynamic disease, often with inter-individual differences manifesting as changes in severity measures over time. Genetics studies that evaluate for associations at a single point in time are underpowered to capture genetic effects that contribute to a disease trajectory, and not simply a quantitative trait measurement at a specific point in time. The natural history of asthma has an enormous amount of variability, and different individuals have different clinical trajectories. Future efforts to identify associations with significant effects will be more successful if we can incorporate dynamic changes in phenotypic traits into GWAS analysis. Further, although asthma GWAS have discovered many novel associations, they have not done much to explain or reveal the mechanisms behind this complex disease.

The integration of multiple large-scale datasets is another application for machine learning in biomedicine. There have been several early efforts to integrate transcriptional data,

obtained from microarray analysis with genetic association analysis to determine the genes associated with changes in gene expression. One advantage to this approach is that it utilizes intermediate phenotypes, and by integrating molecular data into association analyses, brings us closer to discovering the genetic underpinnings of disease mechanisms. Several studies have utilized this approach to better understand the disease mechanisms behind asthma. Raby and colleagues integrated gene expression profiles from $CD4^+$ T lymphocytes with genetic data to identify several novel genetic determinants of gene expression levels in asthmatic subjects [119]. Hao and colleagues performed large-scale genotyping and gene expression profiling on a cohort of over 1,000 subjects and identified several novel genetic variants significantly associated with changes in the expression of multiple asthma-related gene transcripts [62]. In the same analysis, Hao et. al., integrated the gene expression with *cis*-acting eSNPs to create a Bayesian network that allowed them to identify several genes that were"key drivers" behind the molecular mechanisms involved in asthma. The utilization of novel machine learning methods to further integrate multiple data types should provide additional insights into asthma pathogenesis and novel therapeutic targets.

1.2 DISSERTATION OVERVIEW

The goal of this work is to create a computational framework for the integration of multiple types of clinical and molecular data to better understand complex diseases. We apply this framework to clinical, genetic and gene expression data from a cohort of children with mild-moderate asthma to create predictive models for identifying disease endotypes with different molecular mechanisms and responses to pharmacotherapy. A schematic of this framework is depicted in Figure 1.1.

In Chapter 1, we provide background on the development cluster analysis as a computational tool. We describe its use in the biological and medical settings and its current use as a tool for identifying clinical phenotypes. We describe the computational limitations of cluster analysis as a clinical tool. In Chapter 2, we demonstrate an application of cluster analysis to a childhood asthma dataset, with attention to limitations outlined earlier. We identify

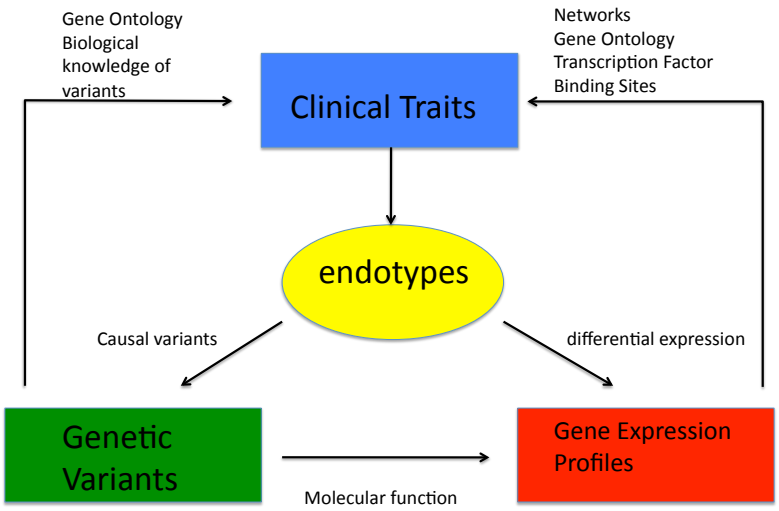


Figure 1.1: Graphical overview of integrative computational framework.

five distinct phenotypic clusters of asthmatics with discrete baseline clinical characteristics. We validate our clusters by demonstrating their longitudinal consistency and demonstrating marked between-cluster differences in long-term asthma control rates, pulmonary function and response to anti-inflammatory asthma therapy, suggesting an potential role for cluster analysis in the classification and clinical management of asthma.

In Chapter 3, we begin by describing biotechnological developments in gene expression profiling, ending with an introduction to high-throughput assays. We highlight the statistical challenges involved in the analysis of high-throughput datasets, and review the application of this technology to disease phenotyping, with an emphasis on asthma. We end by introducing recent computational advances in network analysis and potential applications to transcriptional datasets. In Chapter 4, we demonstrate differential gene expression profiles among asthmatic patients with different endotypes. Through a subsequent co-expression network analysis, we identify a common motif within a module of highly correlated gene transcripts. We also show capability of this module to predict the presence of atopy in an independent

cohort of asthmatic patients.

In Chapter 5, we introduce the current state-of-the-art in genome wide association studies (GWAS) and the statistical challenges present in such analyses. We describe computational techniques such as time-series analysis and sparse regression for feature selection that may be used to increase the power and improve the results and implications of GWAS. We end by introducing a novel computational approach to GWAS, temporally-smoothed lasso (TESL) used to leverage the time-dependencies present in longitudinal clinical data to increase the power of GWAS. In Chapter 6, we apply TESL to a longitudinal asthma endotype and demonstrate the presence of several novel and confirmatory asthma associations. We integrate gene expression profiles obtained from GWAS subjects to show that several associations correspond to novel expression quantitative trait loci (eQTLs).

The results of this work demonstrate the advantages of data integration in the biomedical setting. Using the available data, we demonstrate the presence of multiple disease endotypes among asthmatic children and show that different endotypes correspond to long-term differences in response to several well-known asthma medications. We also identify relationships between these disease endotypes and molecular mechanisms through the integration of genetic associations and transcriptional profiles.

2.0 CLUSTER ANALYSIS AND ENDOTYPES

2.1 CLUSTER ANALYSIS

2.1.1 Definition

Clustering, or unsupervised learning, is a form of exploratory data analysis that involves identifying subgroups within a set of objects. Clustering methods are useful for identifying secondary patterns within a dataset. Mathematically, clustering involves iteratively optimizing an objective function such that the objects within a cluster subgroup are more similar to each other than the objects outside of that cluster.

2.1.2 Methods of Clustering

There are multiple different methods of cluster analysis, with most variation occurring in the formulation of the objective function. There are two general approaches to clustering. One approach begins with each set of objects being assigned to its own cluster. As the algorithm progresses, objects are iteratively assigned to larger and larger clusters. This method of clustering is known as *agglomerative*, or "bottom up". An alternative approach begins with all objects being assigned to the same cluster. As the algorithm progresses, objects are iteratively partitioned to smaller and smaller clusters. This method of clustering is known as *divisive*, or "top down".

2.1.2.1 Hierarchical Clustering Hierarchical clustering is characterized by a hierarchy of subsets within a set of objects, and represents an agglomerative style of clustering. For hierarchical clustering, there are two necessary metrics. The first is a distance metric, which

is used to determine the degree of dissimilarity between two objects. There are multiple distance metrics that may be used for clustering. One example is the Euclidean distance metric, which is defined for two objects (a and b) as: $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$. Another commonly used example, is the Manhattan distance metric: $|a-b|_1 = \sum_i |(a_i - b_i)|$. The second is a linkage metric, which is used to determine the distance between objects in different clusters. For example, for average linkage clustering, at each iteration, the distance between any two clusters, A and B is considered to be the average over all the distances between all pairs of objects (a in A and b in B), or the mean distance between elements within each cluster. Average linkage clustering is seen with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [148], popular in cluster gene expression data generated from microarrays [37]. An alternative linkage metric is minimum, or single-linkage clustering, in which the distance between two clusters A and B is determined by the distance between the two objects in each cluster (a in A and b in B) that are closest to each other. Limitations of hierarchical clustering include difficulty in interpreting the hierarchy, the deterministic nature of the method, which prevents reevaluation after data points are grouped into a node [5]. Furthermore, the tree structure can frequently lock in irrelevant features, reflecting idiosyncrasies of the clustering rules, and early errors can lead to larger and larger systematic clustering errors [159].

2.1.2.2 K-means Clustering K-means clustering is a divisive clustering method that partitions a set of n objects into k clusters. In k-means clustering algorithms, k cluster centers are initialized. Next each object is assigned to the nearest center based upon a chosen distance metric. Finally the cluster centers are re-centered such that they become the centroid of the set of points assigned to their respective cluster. These steps are iterated until either the distance between the points in each cluster and the center is minimized, or the maximal number of iterations has occurred. K-means optimizes the following error function: $F(\mu, C) = \sum_{j=1}^{m} \|\mu_{C(j)} - x_j\|^2$, where $\mu_{C(j)}$ is the center of cluster C and x_j is a datapoint (out of m datapoints) assigned to cluster C. K-means clustering is popular due to its intuitive simplicity and speed. However, there are several drawbacks to this method. First, the number of k clusters that are initialized must be determined in advance.

Because clustering frequently serves as a method for hypothesis generation, the optimal number of clusters is often not known *a priori*. Second, because the initial location for cluster centers is random, the results are not deterministic and may vary between different clustering runs [122]. Third, k-means clustering is sensitive to outliers, and performs poorly when clusters have a non-convex shape, or are non-linearly separable [84].

2.1.2.3 Spectral Clustering Spectral clustering is a relatively recent development in the field of pattern recognition that has become popular due to its ability to outperform alternative forms of clustering, such as k-means. In spectral clustering, data are encoded in the form of an undirected graph $G = (V, E)$, with data points encoded as vertices, v_i for $i = 1...N$ where N is the total number of data points, and the relationship between data points is encoded as edges, with weights w_{ij} encoding the strength of pairwise interaction between two data points. The set of pairwise comparisons between data points forms an affinity matrix, understood in spectral clustering as a weighted, undirected and fully-connected graph. Because G is undirected, the weights are symmetric, $w_{ij} = w_{ji}$. The spectral terminology comes from graph theory, and the spectral analysis of graphs. The calculation of affinities between data points may vary depending on the structure of the data and the goals of cluster analysis [110].

The details of clustering based upon the graph vary depending on the particular algorithm used. One well-known spectral clustering algorithm continues by determining the degree matrix, D [122]. D is a diagonal matrix, defined to be $D_{ii} = \sum_i^n G_{ij}$ where the i^{th} diagonal element of D is the sum of the i^{th} row of G. The next step is to calculate the *Laplacian* graph $L = D - G$. The Laplacian is semi-definite, and may be decomposed into its eigenvalues and eigenvectors, and thus encodes the connectedness of pairwise interactions between different data points in the dataset. For example, if there are three dominant clusters in a set of data points, the Laplacian will demonstrate three dominant eigenvectors. Next, the Laplacian is normalized to scale the entries to a similar range $\hat{L} = D^{-1/2}LD^{-1/2}$ and the eigenvectors are extracted from the normalized Laplacian. A new matrix E is constructed that is composed of the top k eigenvectors in the dataset, where k represents the putative number of clusters present in the dataset. K-means clustering is then used to cluster the data points in E into

k clusters.

The method of spectral clustering has many advantages over other clustering methods, such as hierarchical and k-means clustering. One of the main advantages to spectral clustering is that it does not make strong assumptions about the shapes of the clusters, which is a major limitation of k-means clustering. In other words, the process of determining the eigenvectors of the Laplacian matrix and clustering those eigenvectors allows the data points to be assigned to the cluster to which they have the most connected relationship as opposed to the most proximal cluster [122]. Thus, spectral clustering has the capacity to cluster many different shapes and sizes of clusters that cannot be accurately cluster by other methods. The main disadvantages of spectral clustering are related to the use of k-means clustering and include sensitivity to initial parameter choices, such as the choice of k and the random initialization of cluster centers for k-means clustering [38].

2.1.3 Defining the Optimal Number of Clusters

An open problem in cluster analysis is defining the optimum number of clusters present in a dataset. Although cluster analysis is a pattern recognition tool that is of benefit in learning more about the properties present in a dataset, most clustering method require the user to define the number of clusters to be found within the dataset prior to beginning any data analysis. This is problematic because early in the process of pattern recognition and evaluation of a dataset, the optimum number of clusters is unknown, and little prior knowledge is available. To address this problem, several techniques for determining the optimal number of clusters in a dataset have recently been developed.

2.1.3.1 The Elbow Point One such method for determining the optimal number of clusters involves evaluating the percentage of variance explained as a function of the number of clusters. The number of clusters should be such that adding one more cluster does not lead to measurable improvement in the percent of variance explained by the number of clusters. In practice this method involves evaluating a plot of percent variance as a function of cluster number. The point in the plot where the marginal improvement in percent variance begins

to levels off, or the so-called "elbow" of the plot should represent the point where the number of clusters is optimal [95]. However, it can often be difficult to pinpoint exactly where this "elbow" occurs.

2.1.3.2 The Silhouette Width The optimal number of cluster may also be estimated by using the silhouette width. The silhouette width is a measure of how closely the data points within a cluster are related to each other as opposed to how they are related to data points in other clusters [141]. To calculate the silhouette width, for each data point i, we let $a(i)$ be the average dissimilarity of i with all other data within the same cluster. We can interpret $a(i)$ based on how well matched i is to the cluster to which it is assigned. We then find the average dissimilarity of i using the data from another cluster. We repeat this process for every cluster for which i is not a member and consider the lowest similarity to i to be $b(i)$. The cluster with the lowest similarity is considered to be the cluster that neighbors i. Measure of similarity and dissimilarity are most commonly based upon an appropriately chosen distance metric. After $a(i)$ and $b(i)$ have been found, the silhouette width can be calculated by the following equation:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.1}$$

Where $s(i)$ values close to 1 imply more accurate clusters.

2.1.3.3 The Gap Statistic A newer method for calculating the optimal number of clusters involves using the gap statistic. The way the gap statistic is calculated is by selecting a range of values $1...N$ for the number of clusters present. Then, clustering is performed on the original data points to find k clusters. A dispersion sum is calculated by summing the distance between all data points and the mean of their assigned cluster. Next, a set of reference data points similar in size to the original is created. This is typically done by sampling from a rectangle formed from the original dataset's principal components. Then the dispersion sum of the reference set is calculated. Finally the gap statistic is calculated, which is the log (mean dispersion of reference data points) - log (dispersion of original data

points). This process is iterated over each range of values $1...N$, and the optimal number of clusters is determined from the value of k that leads to the largest gap statistic [165].

2.2 APPLICATIONS TO ASTHMA PHENOTYPING

In the past several decades, there has been a significant amount of progress in the treatment of childhood asthma. The Global Initiative for Asthma (GINA) guidelines [10] and multiple large-scale clinical trials [82, 150, 106], have helped to guide current evidence-based treatments for this complex syndrome. The current stepwise therapeutic approach is designed to maximize the overall level of asthma control and medication compliance while minimizing treatment cost and adverse side effects. However, it is widely recognized that clusters of asthmatic children not only respond differently to medications [158, 130], but also exhibit markedly different disease trajectories, with many children outgrowing their asthma by early adolescence, while others (often with more frequent exacerbations) show disease progression [133, 93, 125, 144] or decreased lung function in adulthood [155]. The lack of common or distinct histological features or reliable quantitative biomarkers suggests that asthma may represent a collection of discrete disorders with some shared phenotypic characteristics, but with distinct etiologies and natural histories. Such heterogeneity poses significant clinical challenges, particularly in regards to long-term prognostication and treatment decision-making.

Clustering methods have become popular in medicine as a way to explore the heterogeneity that is increasingly recognized to be present among patients with complex diseases. For example, asthma has increasingly been recognized as a heterogeneous disease [173], and several cluster analyses of patients with asthma have been performed to explore the presence of disease-relevant subgroups within diverse cohorts of asthmatic patients. Numerous classification schemes based on specific presenting [140, 26, 27] or etiological [130, 120] features, have been proposed. Though tailored treatment strategies are suggested for patients with distinct forms (for example, the timing of long-acting bronchodilator use for the management of exercise-induced bronchoconstriction), most classification schemes have limited utility in

guiding management strategies or reliably predicting long-term morbidity. Recognition of these limitations has motivated the development of multivariate models that consider many patient characteristics simultaneously [120, 30, 49, 65]. More recently, a newer generation of data mining procedures that leverage unsupervised machine learning approaches, have been applied to large asthma cohorts, with early success in defining previously unrecognized clusters of asthma patients [48, 60, 118]. However, these studies were limited by their lack of prospective follow-up data, precluding assessment of the utility of these classification schemes in informing treatment decision-making or disease prognostication.

2.3 LIMITATIONS OF PRIOR CLUSTER ANALYSES

There have been limitations to prior cluster analyses of asthmatic patients. One limitation involved the use of naive clustering methods, such as k-means [60] and hierarchical clustering [118]. The limitations of these methods were described in the preceding sections, and we sought to address this limitation by using spectral clustering in our approach. As described above, spectral clustering has many advantages over hierarchical and k-means clustering, including the ability to discern clusters of differing size and shape, which represents an advantage over k-means clustering. In addition, spectral clustering does not possess the same vulnerability to outliers as hierarchical clustering.

In additional limitation of prior cluster analyses was the lack of a principled approached for determining the optimal number of clusters. Moore and colleagues used hierarchical clustering to partition a cohort of asthmatic patients into smaller subgroups [118]. However, they did not specify a procedure for determining the cut points for different clusters in the hierarchy. Haldar and colleagues used k-means clustering to cluster patients, but likewise did not specify their procedure for optimizing the number of clusters. In the following analysis, we sought to address this limitation by incorporating a formal procedure for optimizing the cluster number into our workflow. We used the gap statistic over a range of cluster numbers to determine the optimal number of clusters for our dataset [165].

Although cluster analysis is useful for identifying patterns within a dataset, it does not provide explanation of why such patterns might exist. Thus, the utility of phenotypic clusters must be based upon the clinical relevance of observed clinical differences between clusters to either risk-stratification or the more rational use of pharmacologic therapies. We address this in our analysis by exploring the cross-sectional clinical differences between clusters, which was done in earlier analyses. We also build upon earlier studies by exploring the longitudinal consistency of clusters over time and differential response to medical therapy between clusters.

3.0 MULTIVARIATE ASTHMA ENDOTYPES

3.1 INTRODUCTION

We performed cluster analysis on a heterogeneous dataset of children with asthma in order to explore the presence of distinct phenotypic cluster corresponding to clinically meaningful differences between patients. In this analysis, we attempt to address many of the computational and clinical limitations of earlier cluster analyses performed on asthmatic children. For this analysis, we evaluated participants in the Childhood Asthma Management Program (CAMP) study [81, 82]. CAMP was a 4.5-year multi-center randomized, double-masked clinical trial evaluating the long-term effects of inhaled budesonide vs. inhaled nedocromil vs. placebo in 1,041 children with mild to moderate childhood asthma. In the primary analysis, no differences in lung function improvement (the primary outcome) were observed between treatment arms, though participants randomized to inhaled budesonide demonstrated markedly improved long-term symptom control and reduced exacerbation rates as compared to participants randomized to either inhaled nedocromil or placebo. In contrast to prior phenotype clustering efforts [48, 60, 118], the availability of both extensive baseline phenotypic data (collected following a 28-day screening period when participants were off all asthma controller medications) and 48 months of prospective follow-up clinical trial data, CAMP provides an opportunity to evaluate whether computational approaches can define meaningful clusters with distinct clinical trajectories and/or treatment responses.

3.2 METHODS

3.2.1 Study Population

The CAMP study design and primary outcomes have been described [81,82]. Subjects aged 5 to 12 years were deemed eligible for enrollment if they (i) had mild-to-moderate persistent asthma, defined by the presence of symptoms, the use of an inhaled bronchodilator at least twice weekly, or the use of daily medication for asthma; (ii) exhibited airway responsiveness to methacholine; and (iii) they had no other clinically significant conditions. Participants were randomized blindly to receive budesonide 200 μg twice daily (Pulmicort, AstraZeneca, Westborough, MA; n = 311), nedocromil sodium 8 mg twice daily (Tilade, RhonePoulenc Rorer, Collegeville, Pa.; n = 312), or matching placebo (n = 418). Subjects were evaluated every four months, for a total of 48 months. Asthma exacerbations were treated by short courses of oral prednisone. The addition of beclomethasone dipropionate (168 μg twice daily; Vanceril, Schering-Plough, Kenilworth, N.J.) was allowed if asthma control was inadequate. If control remained unsatisfactory, replacement or addition of medications was allowed.

3.2.2 Selection of Phenotypes

Phenotypic characteristics, including measurements of lung function, laboratory values, asthma symptoms, and exacerbating factors were measured on each subject prior to and at the time of randomization. From an initial list of 48 clinical variables (Table 3.1), we selected a set of variables, representative of each childs degree of asthma burden, as inputs to a clustering algorithm. Ten variables were excluded from consideration due to an excess of missing data (greater than 10% of values missing). Due to the inherent strong correlation between pre- and post- bronchodilator spirometric measures, we considered only one of the two measurements (either pre- or post-bronchodilator) for FEV_1, FVC, their ratio, and peak flow. In addition, due to the subjective nature of many of the provocative variables, such as animal dander worsens asthma, these variables were excluded from further analysis (Table 3.1). The resultant list for consideration included 18 variables. In contrast to prior studies, potential asthma risk factors, such as gender, ethnicity and environmental exposure were

purposely excluded from consideration for model building.

3.2.3 Preprocessing of Phenotypic Variables

Following variable selection, missing values were imputed using a k-nearest neighbor algorithm from the *pamr* package of Bioconductor 2.51 [164]. Due to the fact that clustering results may be affected by differences in scale among variables [33], vector normalization was performed to scale each variable to a unit vector.

3.2.4 Cluster and Classification Analysis

We used spectral clustering [122], as implemented by the *spec* function of the *kernlab* package [89] of R 2.10.1, to partition the cohort into phenotypic clusters. In recognition that the outcome of clustering methods is dependent on user-defined inputs, we used a data-driven, iterative approach to define both the number of clinical variables to consider and the optimal number of clusters to form. We considered a range of 1 to 10 for the number of clusters to be constructed and, for each iteration, a set of initial cluster centers was generated from a random set of rows in the data eigenvector matrix. We used the gap statistic [165] to select the optimal number of clusters. We used the decision tree method [19, 138] to grow a classification tree by binary recursive partitioning using the 18 variables from the above clustering model to predict the phenotype cluster assignments.

3.2.5 Cluster Validation

3.2.5.1 Comparison to a Univariate Approach to Clustering We assessed the effect of using a single variable for cluster analysis on determining the final cluster assignments. That is, Instead of using a multivariate model with 18 variables for cluster analysis, we used each variable separately to perform the clustering. For the continuous variables, we specified the formation of five clusters, to allow for comparison to the multivariate model. Using each categorical variable independently led to the formation of two clusters. In order to compare the univariate and multivariate approaches to clustering, we evaluated the ability of both approaches to predict future exacerbations (ie. the time to first use of oral prednisone).

Medical History	**Reason for Exclusion**
*Age of asthma onset	included in model
Primary Caregiver's assessment of asthma	too subjective for model
*Atopic dermatitis	included in model
*Positive allergy skin test	included in model
*Prior hospitalizations for asthma	included in model
*Emergency room visits for asthma	included in model
*Hay fever	included in model
Factors worsening Asthma	**Reason for Exclusion**
House dust or Animals or Tobacco smoke	$< 10\%$ missing values (cold air, aspirin),
or Emotional factors or Exercise or Certain foods	remainder considered too subjective
or Respiratory infections or Dampness	
or Changes in the weather or Cold air or Aspirin	
Clinical Presentation	**Reason for Exclusion**
Provoked by exercise	These variables were considered
Provoked by allergy	to be too subjective for use
Age at first symptoms	in the clustering model
Symptom Burden	**Reason for Exclusion**
Age when child began wheezing with shortness of breath	$> 10\%$ missing values
Prior awakening from sleep due to cough or wheeze	$> 10\%$ missing values
Awakening from sleep due in the past 6 months OR in past week	$> 10\%$ missing values
Cough or wheeze during the day unrelated to exercise	too subjective
Cough or wheeze during the day due to exercise	$> 10\%$ missing values
Cough or phlegm with or without an upper respiratory infection	$> 10\%$ missing values
Wheezing present on most days	too subjective
Wheezing present with or without an upper respiratory infection	too subjective
Wheezing present with shortness of breath	too subjective
Two or more episodes of wheezing with shortness of breath	$> 10\%$ missing values
Received a prescription medication for wheezing with shortness of breath	$> 10\%$ missing values
Normal breathing between attacks of wheezing with shortness of breath	$> 10\%$ missing values
Anthropomorphic Measurements	**Reason for Exclusion**
*Body Mass Index	All variables were included
* Waist / hip ratio	in the model
Pulmonary Function	**Reason for Exclusion**
*FEV_1/ FVC	All variables were
*post-BD FEV_1 pre-BD FEV_1/	included in the model
pre-BD FEV_1 (BDR)	
*Methacholine PC_{20} (natural log)	
*Baseline peak expiratory flow rate	
*Post-BD FVC as a percentage of	
the predicted value (post-BD FVC % predicted)	
*Pre-PD FEV_1 as a percentage of	
the predicted value (pre-BD FEV_1 % predicted)	
Peripheral blood measures	**Reason for Exclusion**
*Total serum IgE level ($\log_{10}$)	All variables were
*Absolute serum eosinophils ($\log_{10}$)	included in the model
*Absolute serum lymphocyte count	
*Absolute serum neutrophil count	

Baseline clinical variables considered for cluster analysis. From an initial list of 48 variables shown in the table, we selected 18 clinical variables (denoted by $*$) as inputs to the spectral clustering algorithm.

Table 3.1: Baseline clinical variables considered for cluster analysis.

3.2.5.2 Comparison to an Alternative Clustering Algorithm

In order to evaluate the reproducibility of our cluster assignments we repeated the unsupervised analysis in the CAMP cohort using a different clustering algorithm. As an alternative clustering method, we selected hierarchical clustering because this was the method used in two well-known studies of clustering in asthmatics patients, the Severe Asthma Research Program (SARP) adult [118] and childhood [48] cohorts. These prior studies used hierarchical clustering with Wards minimum distance as an agglomeration method, and we chose to use this clustering method for validation of our clustering results. We used the *hclust* function of the *stats* package in R 2.10.1 to generate five specified clusters and compared the composition of these new clusters to our original cluster assignments. We also performed an outcomes analysis with these new clusters and compared this to our original outcomes analysis.

3.3 RESULTS OF CLUSTER ANALYSIS

3.3.1 Phenotypes

Clinical phenotype data was available for all 1,041 participants. The baseline characteristics assessed following a 28-day screening period off all anti-inflammatory asthma medications, are presented in Table 3.2. As previously reported [81, 82], the demographic composition of the CAMP cohort is consistent with that of childhood asthma in North America, including a higher proportion of boys, early age of onset, and high prevalence of atopic features.

3.3.2 Cluster Analysis

Spectral clustering with 18 asthma-related baseline phenotypic characteristics observed many high gap statistics; however, the maximum statistic was observed when considering five clusters (Figure 3.1). Testing the model by leaving out one variable for each iteration confirmed the importance of all 18 variables in the final model, as exclusion of any one resulted in substantial subgroup fragmentation and inferior model performance (as measured by the gap statistic, Table 3.3).

Variable	**Count or Mean**
Sex	
Male (%)	621 (59.7)
Female (%)	420 (40.3)
Age (years)	8.94 ± 2.12
Self-reported race	
White (%)	711 (68.3)
Black (%)	138 (13.3)
Hispanic (%)	98 (9.41)
Other (%)	94 (9.03)
Family history of asthma (%)	
Yes	574 (55.1)
No	444 (42.7)
Missing	23 (2.21)
Family history of atopy (%)	724 (69.5)
History of tobacco smoke exposure (%)	439 (42.2)
Household income	
< $30,000	242 (23.2)
≥ $30,000	758 (72.8)
Missing	41 (3.94)
Age of asthma onset, years	3.07 ± 2.44
Hospitalized for asthma (%)	320 (30.7)
ER visits for asthma, no./100 person-year	648 ± 62.2
History of atopic dermatitis (%)	298 (28.6)
History of hay fever (%)	557 (53.5)
History of positive skin test (%)	914 (87.8)
Pre-bronchodilator FEV_1, L (range)	1.65 (0.42-3.31)
Pre-bronchodilator FEV_1/FVC ratio (range)	80 (52-100)
FEV_1 bronchodilator response, L (range)	0.10 (-3.77-2.59)
Total serum IgE levels, IU/L (range)	484 (0-5304)
Peripheral blood eosinophil count, $\log_{10}$ /L (range)	2.50 (0-3.72)
Waist to hip ratio	0.88 ± 0.061
Body Mass Index, kg/m^2	18.2 ± 3.52

Table 3.2: Baseline features of 1,041 CAMP asthmatics.

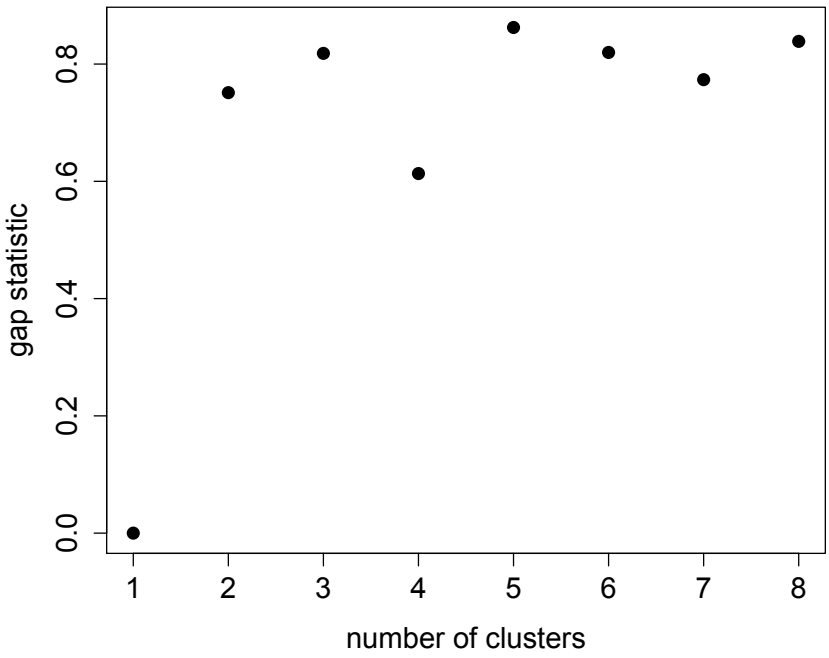
We also repeated the clustering analysis with consideration of subjects of self-reported white ethnicity only (the largest ethnic group in the study), and found no differences in cluster assignment (data not shown). Hence, our final model is optimal with respect to the number of variables and clusters, and does not appear to be confounded by systematic ethnicity-specific phenotypic differences.

Figure 3.2 presents a heat map of the clinical phenotypes grouped by cluster. Though several variables segregate rather discretely by cluster grouping, all but five of the variables included in the final model demonstrated significant differences in distribution across the clusters (all $p < 0.0001$, denoted by * in Figure 3.2 and Table 3.4). Despite their prominence in previously described asthma classification schemes, neither anthropomorphic measures (BMI & waist:hip ratio) nor circulating leukocyte levels (neutrophil or lymphocyte levels) were differentially distributed across the five clusters in this study.

3.3.3 Phenotypic characterization of the asthma clusters

Table 3.5 presents the distribution of asthma-related phenotypes across the observed clusters and Table 3.5 provides a summary of the characteristic features of each cluster. The clusters can be characterized best with respect to three groups of factors: (i) atopic burden (prevalence of atopic dermatitis, allergic rhinitis or skin test reactivity, total serum IgE and peripheral blood eosinophil levels); (ii) lung function and airway lability (pre-bronchodilator FEV_1, FEV_1/FVC, bronchodilator response and methacholine airways hyper-responsiveness); and (iii) baseline exacerbation rates (hospitalization and ED visit rates). Using these three groups of variables, we constructed an Atopy-Obstruction-Exacerbation (AOE) classification scheme by scoring each phenotype group as Low, Medium, or High. For clarity of subsequent discussion, although prospective long-term asthma control was not considered in the clustering procedures (only baseline variables were considered), the cluster groups are also numbered in ascending rank order of poor long-term asthma control (i.e. 1 = best control and 5 = worst, as defined by need for oral steroid therapy during the ~4.5 years of follow-up observation, see below).

The largest group of patients (Cluster 1, 28.8% of cohort) represents the mildest cases,



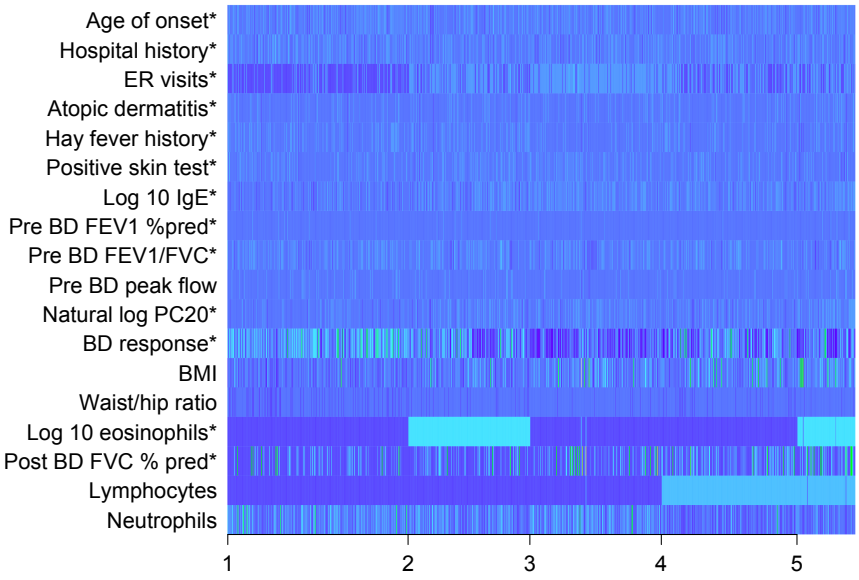
Higher gap statistic indicates greater between-cluster separation.

Figure 3.1: The gap statistic as a function of the number of clusters.

Variable	**Range of Values for each variable**	**Optimal number of clusters with variable removed**[†]
Age of onset (years)	0-12	10
Hospitalized for asthma		9
Missing	0	
Yes	320	
No	721	
ER visits for asthma (number in past year)	0-12	3
History of atopic dermatitis		7
Missing	1	
Yes	298	
No	742	
History of a positive skin test		10
Missing	0	
Yes	914	
No	127	
FEV_1 bronchodilator response (L)		
Methacholine PC_{20} (natural log)	-3.77-2.59	3
Peripheral blood eosinophils (log_{10})	0-3.72	6
Pre-bronchodilator FEV_1/FVC (% predicted)	52-100	6
Pre-bronchodilator peak flow (L/min)	100-550	6
Waist to hip ratio	0.62-1.48	3
Total serum IgE levels (log_{10})	0.30-4.61	9
Body Mass Index (kg/m^2)	12.7-34.3	3
Pre-bronchodilator FEV_1 (% predicted)	44-148	4
Post-bronchodilator FVC (% predicted)	69-162	2
History of hay fever		10
Missing	5	
Yes	557	
No	479	
Lymphocytes (%)	4-78	6
Neutrophils (%)	14-86	6

[†]Clustering robustness analysis. Each of the 18 variables used to perform the clustering analysis were removed from the model one by one and the remaining 17 variables were used to perform the clustering. Shown are the number of optimal clusters as determined by the gap statistic when a particular variable was removed from the model.

Table 3.3: Range of Baseline Features of CAMP Asthmatics.



The above heatmap depicts the differences among normalized clinical variables used for clustering and the different cluster sub-groups. The cluster assignments are grouped along the horizontal axis and the variables used to determine the cluster assignments appear along the vertical axis.

Figure 3.2: Heatmap of phenotypic trait distribution by cluster.

	Cluster 1 (n=300)	Cluster 2 (n=202)	Cluster 3 (n=218)	Cluster 4 (n=225)	Cluster 5 (n=96)	P-value
AOE Classification	**LLL**	**HLM**	**HHM**	**MHH**	**HHH**	
Asthma history						
Age of asthma onset (years)	3.52 ± 2.63	3.09 ± 2.40	3.66 ± 2.62	2.21 ± 1.89	2.27 ± 1.80	< 0.001
Total hospitalized for asthma (%)	0 (0)	0 (0)	1 (0.46)	225 (100)	94 (97.9)	< 0.001
ER visits for asthma (visits / 100 person-years)	44.3	47.0	70.2	75.6	101	< 0.001
Atopic Features						
History of atopic dermatitis (%)	0 (0%)	202 (100%)	2 (0.1%)	0 (0%)	94 (97.9%)	< 0.001
History of hay fever (%)	61 (20.3%)	132 (65.3%)	191 (87.6%)	119 (52.9%)	54 (56.3%)	< 0.001
History of positive skin test (%)	230 (76.7%)	185 (91.6%)	209 (95.9%)	198 (88%)	92 (95.8%)	< 0.001
Total serum IgE levels ($\log_{10}$)	2.37 ± 0.70	2.72 ± 0.72	2.79 ± 0.58	2.64 ± 0.61	2.81 ± 0.63	< 0.001
Spirometry						
Pre-bronchodilator FEV_1 (% predicted)	96.4 ± 12.7	97.7 ± 14.8	89.7 ± 13.9	91.4 ± 13.8	92.0 ± 16.1	< 0.001
Pre-bronchodilator FEV_1/FVC (% predicted)	81.8 ± 7.68	81.5 ± 7.59	77.6 ± 8.54	77.8 ± 8.24	78.6 ± 9.60	< 0.001
Pre-bronchodilator peak flow	276.1 ± 67.3	274.3 ± 73.3	276.7 ± 69.1	276.4 ± 70.8	255.6 ± 73.3	0.12
Airway responsiveness						
Methacholine PC_{20} (natural log)	0.71 ± 1.03	0.14 ± 1.11	-0.54 ± 1.00	0.038 ± 1.14	-0.23 ± 1.17	< 0.001
FEV_1 bronchodilator response (L)	0.077 ± 0.07	0.097 ± 0.08	0.12 ± 0.11	0.12 ± 0.11	0.16 ± 0.14	< 0.001
Anthropomorphic features						
BMI	18.1 ± 3.46	18.6 ± 3.83	18.5 ± 3.66	17.8 ± 3.19	17.6 ± 3.38	0.07
Waist/hip ratio	0.882 ± 0.06	0.885 ± 0.07	0.881 ± 0.06	0.874 ± 0.05	0.877 ± 0.07	0.80
Peripheral blood counts						
Eosinophils ($\log_{10}$)	2.35 ± 0.55	2.54 ± 0.53	2.57 ± 0.52	2.50 ± 0.49	2.71 ± 0.41	< 0.001
Lymphocytes (%)	42.1 ± 11.5	40.9 ± 9.82	41.1 ± 10.9	41.7 ± 10.5	40.8 ± 9.78	0.67
Neutrophils (%)	45.5 ± 12.2	44.5 ± 11.0	44.6 ± 11.5	44.7 ± 11.5	43.1 ± 11.0	0.62

Table 3.4: Distribution of Traits Across Phenotypic Clusters.

with the lowest baseline exacerbation rates, lowest prevalence of atopic features, and preserved lung function (AOE classification LLL). The smallest cluster, Cluster 5 (9.3%), consists of the most severe cases, with the highest baseline exacerbation rates, a very high atopic burden, and reduced lung function (AOE group HHH). The three remaining clusters reflect subsets with intermediate levels of severity and more heterogeneous clinical features. Cluster 2 (19.3%) includes those subjects with high atopic burden but preserved lung function (relative to the other groups) and intermediate airways hyperresponsiveness. This group has an intermediate baseline exacerbation rate, with no reports of hospitalization (AOE group HLM). Patients in Cluster 3 (20.9%) have high atopy burden, the most compromised lung function, and extreme airways hyperresponsiveness, but have intermediate baseline exacerbation rates (AOE group HHM). In contrast, though patients in Cluster 4 (21.6%) are less atopic to those of Cluster 2 (including lower rates of allergic rhinitis and skin test reactivity, and lower serum IgE and peripheral blood eosinophil levels), and have reduced lung function at levels similar to individuals in Cluster 3, they have very high exacerbation rates, particularly with respect to hospitalizations (AOE group MHH). It is clear that no one feature is sufficient to characterize these groups.

3.3.4 Phenotypic clusters, long-term asthma control and response to specific inhaled anti-inflammatory controller medications

To assess whether the derived cluster designations have clinical relevance regarding subsequent risk of exacerbations, we performed survival analysis of time to asthma exacerbation over approximately 4.5 years follow-up. Consistent with the effective randomization of treatment assignment, treatment group did not differ between clusters ($p = 0.91$, Table 3.6, Table 3.7), enabling unbiased assessment of the relationship between cluster grouping and long-term asthma control (Figure 3.3). Kaplan-Meier analysis confirmed that cluster grouping was strongly predictive of time to first course of oral prednisone (Figure 3.10a, Kaplan-Meier log-rank $p < 0.0001$) and time to initiation of additional asthma controller therapies (Figure 3.10b, $p = 0.001$). The most striking differences were observed within the first 12 months post-randomization, by which time the majority of subjects in Clusters 4 and 5

Cluster 1: Relatively mild asthmatics with a low atopic burden (LLL)

- The largest subgroup of patients (28.8%)
- No history of atopic dermatitis, lowest prevalence of hay fever or skin prick test reactivity, lowest IgE levels
- Preserved lung function(highest FEV_1/FVC ratio)
- Lowest bronchodilator response, intermediate airway hyperresponsiveness.
- No prior hospitalization for asthma and the lowest reported prevalence of ED visits
- Lowest risk of poor long-term asthma control**

Cluster 2: Highly atopic asthmatics with preserved lung function (HLM)

- Universally report atopic dermatitis, high prevalence of allergic rhinitis and skin test reactivity
- Preserved lung function (highest FEV_1)
- Intermediate bronchodilator response and airways hyper responsiveness
- No prior hospitalization, but intermediate rates of prior ED visits
- Low-intermediate risk of poor long-term asthma control**

Cluster 3: Highly atopic asthmatics with reduced lung function and severe airways hyperresponsiveness (HHM)

- Rarely report atopic dermatitis (in contrast to HLM cluster),but highest prevalence of allergic rhinitis and skin test reactivity
- Most reduced lung function (low FEV_1/FVC ratio)
- High bronchodilator response and most severe airways hyper responsiveness
- Few prior hospitalizations, but intermediate rates of prior ED visits (similar to HLM cluster)
- Intermediate risk of poor long-term asthma control**

Cluster 4: Asthmatics with reduced lung function and high exacerbation rates, but lower atopic burden (MHH)

- No history of atopic dermatitis, intermediate prevalence of hay fever (52.9%), lower IgE levels
- Most reduced lung function (low FEV_1/FVC ratio, similar to HHH cluster)
- High bronchodilator response and most severe airways hyper responsiveness
- Most reports of prior hospitalization
- Intermediate risk of poor long-term asthma control**

Cluster 5: Asthmatics with most severe disease at baseline, high atopic burden, highest exacerbation rates (HHH)

- Smallest subgroup of patients (9.3%)
- Nearly universal atopic dermatitis, highest prevalence of skin test reactivity, highest IgE levels, highest eosinophilia, intermediate prevalence of allergic rhinitis
- Reduced lung function (low FEV_1/FVC ratio) (similar to MHH cluster)
- Highest bronchodilator response and severe airways hyperresponsiveness
- Most reports of prior hospitalization and highest rate of ER visits
- Highest risk of poor long-term asthma control**

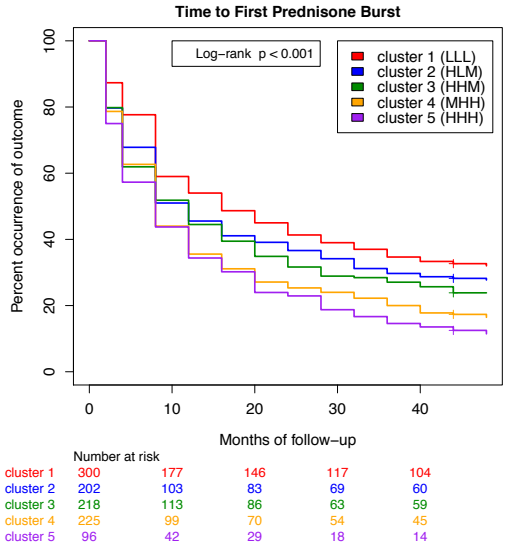
* Atopy-Obstruction-Exacerbation classification denoted in parenthesis.

** Poor long-term asthma control risk is defined from prospective survival analysis of time to first course of oral prednisone. This variable was derived using the defined cluster groupings and was therefore not considered in the spectral cluster analyses used to define the clusters.

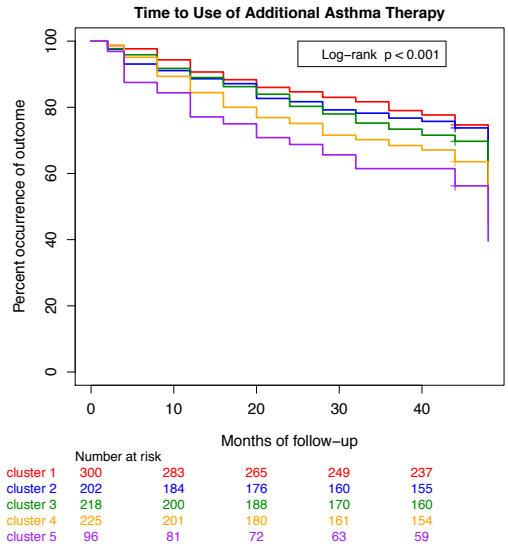
Table 3.5: Summary of clinical characteristics of phenotypic clusters.

(66% and 64%, respectively) had required at least one course of oral steroids, as compared to only 56% and 54% of subjects in Clusters 3 and 2, and only 46% of patients in Cluster 1. These established trends persisted for the remainder of the trial, with greater separation of cluster groupings over time. At the end of the 4-year trial observation period, about a 3-fold difference in the percentage of subjects not requiring oral prednisone was observed between the two most extreme groups (32% in Cluster 1 vs. 11% in cluster 5, $p < 0.0001$). Similar relationships were noted for time to initiation of additional asthma controller therapies (Figure 3.10b).

We next assessed whether treatment response to specific inhaled anti-inflammatory controller medications differed by cluster group. As originally reported in the primary outcomes assessment of the CAMP trial [82], use of inhaled budesonide, compared to placebo, significantly reduced the number of asthma exacerbations Further, it was found that for the entire cohort nedocromil did not significantly reduce exacerbation rates or additional controller therapies compared to placebo. However, in a post hoc evaluation stratified by cluster grouping, significant heterogeneity in treatment response rates to both medications is found (Figure 3.4, Table 3.6, Table 3.7, 3.8, 3.9, 3.10): whereas subjects stratified to the three more mild clusters demonstrated treatment response patterns similar to those reported in the cohort as a whole, the therapeutic efficacy of nedocromil was similar to that of budesonide (as significantly greater than placebo) among subjects in the two most severe clusters (Clusters 4 and 5) - those with the highest risk of exacerbation. Subjects in Cluster 4 those with the lowest atopic burden, worst lung function, and high baseline exacerbation rates demonstrated significant reductions in exacerbation rates when randomized to nedocromil (1.7 fold reduction compared to placebo at 12 months, 1.6 fold reduction at 4 years) that was similar to the reduction in exacerbation observed among those randomized to budesonide (1.6 fold reduction compared to placebo at 12 months, 1.4 fold reduction at 4 years). In this group, there was no difference in exacerbation reduction between those randomized to nedocromil or budesonide ($p = 0.96$). Similar effects were noted in Cluster 5 ($p = 0.22$ for difference between nedocromil and budesonide groups), though the magnitude of treatment effect (compared to placebo) was substantially lower than for subjects in Cluster 4. For subjects in Cluster 5- those with a high atopic burden, low lung function and the highest



(a) Prednisone use



(b) Need for additional therapy

Kaplan-Meier plots by cluster of the cumulative probability of a first course of prednisone **A:** or initiation of additional asthma controller therapies (beclomethasone or other) **B:** during the four-year follow-up period of the CAMP trial.

Figure 3.3: Survival analysis for phenotypic clusters.

baseline exacerbations there was no decrease in exacerbation rate for subjects randomized to either nedocromil (p = 0.56) or budesonide (p = 0.12).

3.3.5 Demographic, environmental, and familial determinants of phenotypic clusters

We next assessed for associations between the observed clusters and known demographic, environmental and familial features implicated in asthma pathogenesis that were not considered during clustering. Descriptions of demographic, environmental and familial clinical variables across phenotypic clusters are presented in Table 3.11. Although trends for higher proportions of non-Hispanic white subjects in the mildest group, and blacks in the most severe groups were noted, these differences were not statistically significant. In contrast, enlightening differences across clusters were observed for numerous environmental and familial factors. For example, though environmental tobacco smoke exposure was reported by subjects in all five clusters, the prevalence was greatest among individuals in Clusters 4 and 5 those with the highest baseline exacerbation rates. However, among subjects in the less severe clusters, a direct relationship between severity and smoke exposure was not observed: those with the lowest childhood smoke exposure (Cluster 2, 30.2%) had higher baseline exacerbation and greater airways hyperresponsiveness than subjects in Cluster 1 who had significantly higher childhood smoke exposure (39.7%),and exacerbation rates, lung function, and airways responsiveness were markedly different between Clusters 1 and 3 despite very similar childhood smoke exposure rates (39.7% vs. 37.6%, respectively).

Similarly, though differences in aeroallergen exposure and in familial burden of both asthma and atopy were observed across the five phenotypic clusters, obvious linear correlations between risk factor exposure and severity of disease were not observed. Thus, with the exception of age, where statistically significant differences were observed across cluster groups, demographic variables including sex or socioeconomic indicators did not differ between clusters, suggesting that although environmental and genetic factors likely contribute to the underlying pathobiological processes that determine cluster designation, none of these variables are sole etiological determinants.

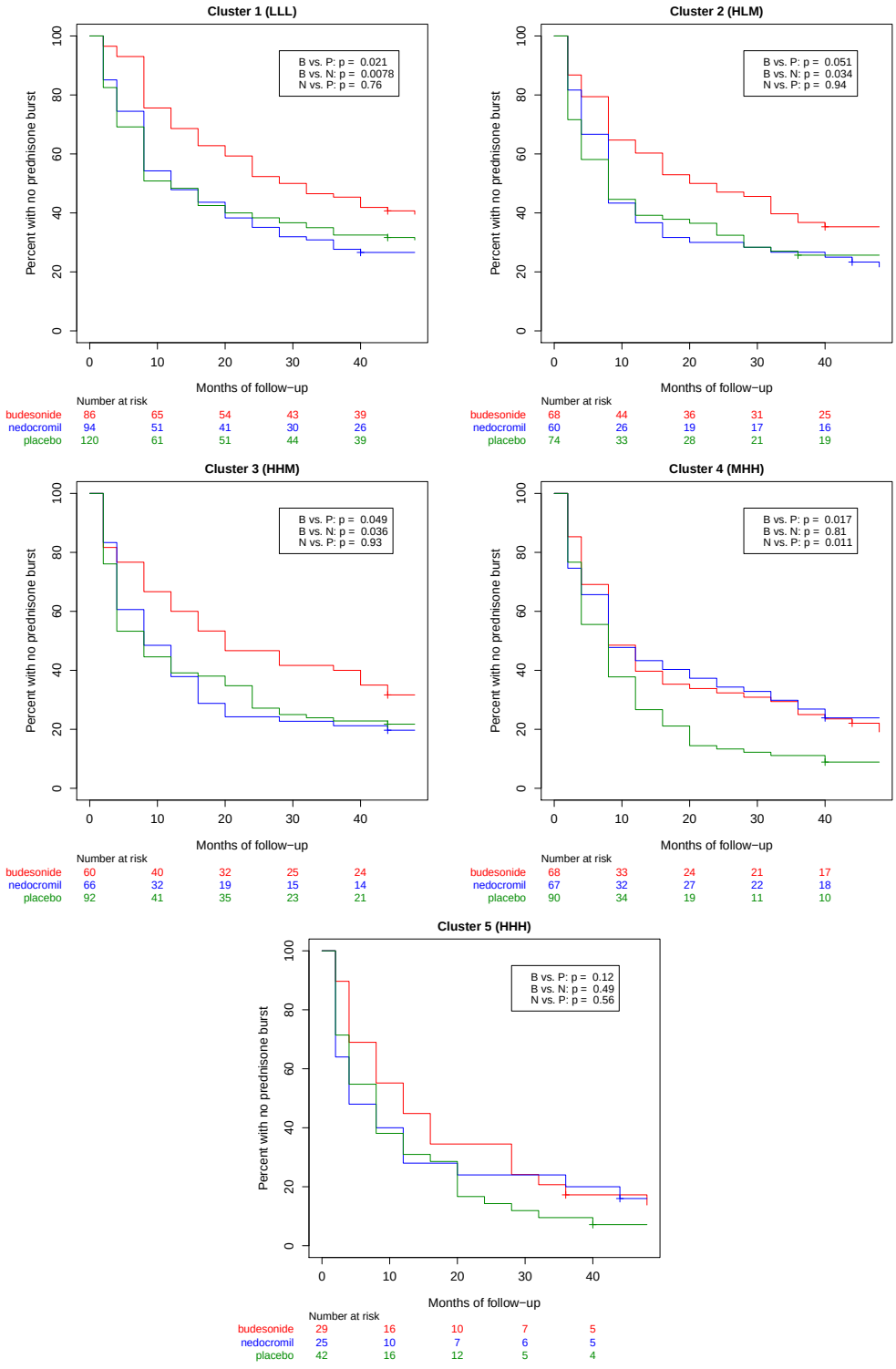


Figure 3.4: Kaplan-Meier estimate by treatment group of the cumulative probability of prednisone use during four years of follow-up, stratified by asthma cluster.

	Cluster 1 (n=300)	Cluster 2 (n=202)	Cluster 3 (n=218)	Cluster 4 (n=225)	Cluster 5 (n=96)	P-value
AOE Classification	**LLL**	**HLM**	**HHM**	**MHH**	**HHH**	
Budesonide (n = 311, 29.9%)						
2 months	0.06 ± 0.34	0.19 ± 0.61	0.18 ± 0.39	0.30 ± 0.90	0.19 ± 0.19	0.13
4 months	0.12 ± 0.22	0.30 ± 0.31	0.34 ± 0.41	0.54 ± 0.46	0.58 ± 0.36	0.003
8 months	0.39 ± 0.63	0.53 ± 0.52	0.71 ± 0.72	1.09 ± 0.82	0.92 ± 0.36	0.01
12 months	0.61 ± 0.67	0.72 ± 0.54	0.91 ± 0.58	1.44 ± 0.64	1.29 ± 0.31	0.004
16 months	0.78 ± 0.44	0.95 ± 0.65	1.21 ± 0.76	1.63 ± 0.42	2.22 ± 0.78	0.01
20 months	0.85 ± 0.38	1.15 ± 0.38	1.46 ± 0.58	1.87 ± 0.69	2.70 ± 0.44	0.01
24 months	1.07 ± 0.89	1.35 ± 0.72	1.57 ± 0.48	2.07 ± 0.44	3.04 ± 0.37	0.04
28 months	1.24 ± 0.46	1.47 ± 0.36	1.76 ± 0.50	2.39 ± 0.93	3.40 ± 0.38	0.04
32 months	1.39 ± 0.45	1.69 ± 0.58	1.96 ± 0.68	2.54 ± 0.47	3.63 ± 0.27	0.05
36 months	1.61 ± 0.62	1.98 ± 0.54	2.19 ± 0.48	2.85 ± 0.73	4.05 ± 0.42	0.03
40 months	1.75 ± 0.46	2.13 ± 0.39	2.33 ± 0.48	3.14 ± 0.74	4.52 ± 0.28	0.03
44 months	1.78 ± 0.31	2.23 ± 0.36	2.56 ± 0.49	3.03 ± 0.59	4.71 ± 0.20	0.02
48 months	1.94 ± 0.46	2.42 ± 0.63	2.86 ± 0.48	3.38 ± 0.52	5.00 ± 0.28	0.04
Nedocromil (n = 312, 30.0%)						
2 months	0.21 ± 0.55	0.29 ± 0.65	0.22 ± 0.60	0.35 ± 0.65	0.54 ± 0.83	0.91
4 months	0.37 ± 0.43	0.55 ± 0.64	0.54 ± 0.56	0.61 ± 0.61	1.13 ± 1.18	0.06
8 months	1.03 ± 1.03	0.94 ± 0.52	1.12 ± 1.14	1.02 ± 0.68	1.59 ± 0.73	0.78
12 months	1.38 ± 0.70	1.26 ± 0.95	1.52 ± 0.78	1.46 ± 0.95	2.14 ± 0.99	0.66
16 months	1.71 ± 0.92	1.76 ± 1.11	1.94 ± 0.70	1.89 ± 0.80	2.78 ± 1.10	0.60
20 months	2.05 ± 0.51	2.10 ± 0.69	2.38 ± 0.79	2.15 ± 0.61	3.00 ± 0.52	0.55
24 months	2.33 ± 0.41	2.42 ± 0.69	2.92 ± 1.93	3.37 ± 0.60	3.50 ± 1.15	0.55
28 months	2.48 ± 0.47	2.74 ± 0.95	3.42 ± 2.07	2.53 ± 0.47	3.91 ± 0.83	0.62
32 months	2.65 ± 0.45	3.20 ± 0.94	3.74 ± 0.68	2.82 ± 0.65	4.29 ± 0.51	0.56
36 months	2.87 ± 0.60	3.27 ± 0.39	4.09 ± 0.65	3.29 ± 1.26	4.85 ± 0.58	0.28
40 months	3.14 ± 0.63	3.51 ± 0.40	4.29 ± 0.60	2.53 ± 0.63	5.15 ± 0.76	0.26
44 months	3.25 ± 0.35	3.78 ± 0.69	4.46 ± 0.67	3.80 ± 0.64	5.40 ± 0.64	0.22
48 months	3.56 ± 0.50	4.17 ± 0.67	4.67 ± 0.36	4.02 ± 0.82	5.42 ± 0.63	0.34
Placebo (n = 418, 40.1%)						
2 months	0.30 ± 0.77	0.45 ± 0.86	0.37 ± 0.72	0.34 ± 0.71	0.41 ± 0.72	0.49
4 months	0.53 ± 0.52	0.84 ± 0.75	0.77 ± 0.65	0.83 ± 0.84	0.76 ± 0.69	0.16
8 months	1.07 ± 0.92	1.32 ± 0.81	1.21 ± 0.73	1.45 ± 0.96	1.61 ± 1.25	0.29
12 months	1.47 ± 0.93	1.80 ± 0.95	1.56 ± 0.66	2.12 ± 0.88	2.29 ± 1.47	0.04
16 months	1.97 ± 0.81	1.92 ± 0.50	1.71 ± 0.43	2.85 ± 0.97	2.77 ± 0.78	0.01
20 months	2.37 ± 1.03	2.30 ± 0.88	1.99 ± 0.58	3.65 ± 1.02	3.80 ± 1.72	<0.001
24 months	2.81 ± 0.95	2.84 ± 0.98	2.54 ± 1.06	4.23 ± 0.93	4.26 ± 0.78	<0.001
28 months	3.38 ± 2.47	3.14 ± 0.55	2.82 ± 0.51	4.75 ± 0.95	4.86 ± 0.92	<0.001
28 months	3.38 ± 2.47	3.14 ± 0.55	2.82 ± 0.51	4.75 ± 0.95	4.86 ± 0.92	<0.001
32 months	3.79 ± 0.96	3.42 ± 0.54	3.10 ± 0.70	5.30 ± 0.96	5.46 ± 0.92	<0.001
36 months	3.98 ± 0.51	4.09 ± 2.78	3.30 ± 0.43	6.01 ± 2.98	5.89 ± 0.70	<0.001
40 months	4.18 ± 0.49	4.34 ± 0.69	3.53 ± 0.53	6.32 ± 0.65	6.40 ± 0.71	<0.001
44 months	4.42 ± 0.63	4.64 ± 0.67	3.75 ± 0.55	6.65 ± 0.71	6.82 ± 0.78	<0.001
48 months	4.69 ± 0.52	4.72 ± 0.78	3.91 ± 0.46	6.89 ± 0.53	7.42 ± 0.86	<0.001
P-values						
Budesonide vs. Nedocromil	0.0006	0.008	0.054	0.96	0.22	
Budesonide vs. Placebo	0.0007	0.001	0.022	0.005	0.13	
Nedocromil vs. Placebo	0.39	0.45	0.65	0.006	0.82	

The cumulative number of subjects experiencing asthma exacerbations as demonstrated by the need for oral prednisone therapy, at each study time point, stratified by treatment group. Shown are the mean (±sd) cumulative number of prednisone bursts per person from the onset of the study period, p-values for between-cluster differences in outcome (far right), and pairwise comparisons of within-cluster differences in outcomes (bottom level).

Table 3.6: Number of prednisone bursts.

	Cluster 1 (n=300)	Cluster 2 (n=202)	Cluster 3 (n=218)	Cluster 4 (n=225)	Cluster 5 (n=96)	P-value
AOE Classification	**LLL**	**HLM**	**HHM**	**MHH**	**HHH**	
Budesonide (n = 311, 29.9%)						
2 months	0.01 ± 0.11	0.04 ± 0.21	0.03 ± 0.18	0.00 ± 0.71	0.00 ± 0.72	0.32
4 months	0.01 ± 0.11	0.05 ± 0.21	0.03 ± 0.18	0.00 ± 0.84	0.04 ± 0.69	0.43
8 months	0.01 ± 0.12	0.06 ± 0.30	0.03 ± 0.18	0.06 ± 0.96	0.04 ± 1.25	0.66
12 months	0.03 ± 0.16	0.06 ± 0.39	0.07 ± 0.26	0.11 ± 0.88	0.13 ± 1.47	0.24
16 months	0.04 ± 0.20	0.06 ± 0.40	0.11 ± 0.31	0.13 ± 0.97	0.17 ± 0.78	0.21
20 months	0.05 ± 0.28	0.10 ± 0.43	0.13 ± 0.33	0.16 ± 1.02	0.17 ± 1.72	0.43
24 months	0.06 ± 0.29	0.10 ± 0.43	0.13 ± 0.33	0.18 ± 0.93	0.17 ± 0.78	0.38
28 months	0.10 ± 0.34	0.10 ± 0.43	0.15 ± 0.36	0.25 ± 0.95	0.26 ± 0.92	0.46
32 months	0.11 ± 0.36	0.10 ± 0.43	0.19 ± 0.44	0.26 ± 0.96	0.26 ± 0.92	0.43
36 months	0.18 ± 0.52	0.11 ± 0.45	0.21 ± 0.50	0.35 ± 2.98	0.30 ± 0.70	0.52
40 months	0.24 ± 0.64	0.11 ± 0.45	0.27 ± 0.63	0.51 ± 0.65	0.36 ± 0.71	0.27
44 months	0.25 ± 0.65	0.13 ± 0.47	0.31 ± 0.67	0.54 ± 0.71	0.38 ± 0.78	0.18
48 months	0.66 ± 0.92	0.37 ± 0.64	0.50 ± 0.71	0.82 ± 0.53	0.70 ± 0.86	0.27
Nedocromil (n = 312, 30.0%)						
2 months	0.00 ± 0.00	0.02 ± 0.13	0.00 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.33
4 months	0.00 ± 0.00	0.09 ± 0.35	0.02 ± 0.12	0.05 ± 0.00	0.17 ± 0.20	0.002
8 months	0.06 ± 0.23	0.19 ± 0.62	0.12 ± 0.38	0.11 ± 0.24	0.23 ± 0.20	0.20
12 months	0.13 ± 0.43	0.28 ± 0.94	0.14 ± 0.39	0.17 ± 0.36	0.36 ± 0.34	0.17
16 months	0.22 ± 0.60	0.41 ± 1.42	0.18 ± 0.46	0.26 ± 0.38	0.59 ± 0.48	0.15
20 months	0.30 ± 0.78	0.48 ± 1.58	0.23 ± 0.56	0.30 ± 0.55	0.68 ± 0.48	0.24
24 months	0.35 ± 0.93	0.52 ± 1.61	0.45 ± 0.94	0.40 ± 0.56	0.82 ± 0.48	0.27
28 months	0.38 ± 0.99	0.64 ± 1.80	0.62 ± 1.12	0.37 ± 0.67	0.86 ± 0.75	0.29
32 months	0.42 ± 1.11	0.69 ± 1.85	0.79 ± 1.36	0.49 ± 0.75	1.00 ± 0.75	0.19
36 months	0.48 ± 1.72	0.76 ± 1.95	0.86 ± 1.46	0.60 ± 0.88	1.15 ± 0.82	0.23
40 months	0.59 ± 1.44	0.84 ± 2.10	0.95 ± 1.59	0.69 ± 1.15	1.15 ± 1.00	0.29
44 months	0.65 ± 1.50	1.00 ± 2.45	1.05 ± 1.87	0.75 ± 1.13	1.15 ± 1.02	0.48
48 months	0.94 ± 1.66	1.27 ± 2.59	1.20 ± 1.79	1.00 ± 1.20	1.26 ± 1.22	0.92
Placebo (n = 418, 40.1%)						
2 months	0.30 ± 0.77	0.45 ± 0.86	0.37 ± 0.72	0.34 ± 0.71	0.41 ± 0.72	0.49
4 months	0.53 ± 0.52	0.84 ± 0.75	0.77 ± 0.65	0.83 ± 0.84	0.76 ± 0.69	0.16
8 months	1.07 ± 0.92	1.32 ± 0.81	1.21 ± 0.73	1.45 ± 0.96	1.61 ± 1.25	0.29
12 months	1.47 ± 0.93	1.80 ± 0.95	1.56 ± 0.66	2.12 ± 0.88	2.29 ± 1.47	0.04
16 months	1.97 ± 0.81	1.92 ± 0.50	1.71 ± 0.43	2.85 ± 0.97	2.77 ± 0.78	0.01
20 months	2.37 ± 1.03	2.30 ± 0.88	1.99 ± 0.58	3.65 ± 1.02	3.80 ± 1.72	<0.001
24 months	2.81 ± 0.95	2.84 ± 0.98	2.54 ± 1.06	4.23 ± 0.93	4.26 ± 0.78	<0.001
28 months	3.38 ± 2.47	3.14 ± 0.55	2.82 ± 0.51	4.75 ± 0.95	4.86 ± 0.92	<0.001
28 months	3.38 ± 2.47	3.14 ± 0.55	2.82 ± 0.51	4.75 ± 0.95	4.86 ± 0.92	<0.001
32 months	3.79 ± 0.96	3.42 ± 0.54	3.10 ± 0.70	5.30 ± 0.96	5.46 ± 0.92	<0.001
36 months	3.98 ± 0.51	4.09 ± 2.78	3.30 ± 0.43	6.01 ± 2.98	5.89 ± 0.70	<0.001
40 months	4.18 ± 0.49	4.34 ± 0.69	3.53 ± 0.53	6.32 ± 0.65	6.40 ± 0.71	<0.001
44 months	4.42 ± 0.63	4.64 ± 0.67	3.75 ± 0.55	6.65 ± 0.71	6.82 ± 0.78	<0.001
48 months	4.69 ± 0.52	4.72 ± 0.78	3.91 ± 0.46	6.89 ± 0.53	7.42 ± 0.86	<0.001
P-values						
Budesonide vs. Nedocromil	0.0006	0.008	0.054	0.96	0.22	
Budesonide vs. Placebo	0.0007	0.001	0.022	0.005	0.13	
Nedocromil vs. Placebo	0.39	0.45	0.65	0.006	0.82	

The cumulative number of subjects experiencing asthma exacerbations as demonstrated by the need for additional asthma controller therapy in the form of beclomethasone for each study time point stratified by treatment group. Shown are mean values with standard deviations. Shown are p-values for between-cluster differences in outcome (far right) and pairwise comparisons of within-cluster differences in outcomes (bottom level). Mean (±sd) cumulative number of exacerbations per person from the onset of the study period

Table 3.7: Need for additional asthma controller medications.

	Initiation of oral prednisone	**Initiation of additional asthma controller therapies**
Cluster (reference is Cluster 1)		
Cluster 2	0.03	0.50
Cluster 3	0.001	0.98
Cluster 4	< 0.0001	0.03
Cluster 5	< 0.0001	0.003
Treatment group (reference is placebo)		
Budesonide	< 0.0001	< 0.0001
Nedocromil	0.32	0.53
Overall Interaction of Cluster and Treatment Group	0.17	0.16
Age	0.08	0.35
Sex	0.50	0.53
Height	0.45	0.93

Cox proportional hazards models for risk of future asthma exacerbations using cluster assignment, age, sex, height, and treatment group as predictor variables under an additive model. Cluster 1 was used as the reference for cluster assignment. Shown are the p-values for the degree of risk each variable contributes to the model.

Table 3.8: Summary of p-values for Cox proportional hazards modeling of risk of asthma exacerbation.

Initiation of oral prednisone	**Budesonide (ref = placebo)**	**Nedocromil (ref = placebo)**	**Budesonide (ref = nedocromil)**
Drug	< 0.001	0.28	< 0.001
Cluster 2	0.12	0.12	0.17
Cluster 3	0.03	0.03	0.06
Cluster 4	< 0.001	< 0.001	0.006
Cluster 5	< 0.001	< 0.001	< 0.001

Cox proportional hazards models for risk of future asthma exacerbations using cluster assignment and treatment group as predictor variables under an additive model. Cluster 1 was used as the reference group for cluster assignment. Shown are the p-values for the degree of risk each variable contributes to the model.

Table 3.9: Summary of p-values for Cox proportional hazards modeling of risk of asthma exacerbation.

Initiation of oral prednisone	Budesonide (ref = placebo)	Nedocromil (ref = placebo)	Budesonide (ref = nedocromil)
Drug*Cluster 2	0.94	0.93	0.99
Drug*Cluster 3	0.93	0.87	0.97
Drug*Cluster 4	0.98	0.05	0.05
Drug*Cluster 5	0.92	0.60	0.50

Cox proportional hazards models for risk of future asthma exacerbations using cluster assignment and treatment group as predictor variables under an interaction model. Cluster 1 was used as the reference group for cluster assignment. Shown are the p-values for the interaction terms.

Table 3.10: Summary of p-values for Cox proportional hazards modeling of drug by cluster interaction.

3.3.6 Decision-tree Algorithm for Efficient Patient Classification

To facilitate implementation of the AOE classification clinically without the need of machine learning software, we screened combinations of the 18 cluster-building variables as candidates to build a simple-to-use classifier. At least one variable from each of the three AOE classes was considered in each model. The most accurate classifier (97% accuracy) included a history of previously being hospitalized for asthma (exacerbation), history of atopic dermatitis (atopy), history of hay fever (atopy) and the natural log of the FEV_1 PC_{20} (Figure 3.5. Given that PC_{20} is not routinely obtained clinically in the pediatric setting, we assessed the performance of the model by either removing PC_{20}, or substituting PC_{20} with other variables (including bronchodilator responsiveness). These maneuvers worsened model performance (second best classification accuracy of 91.4%), with most misclassification occurring between Clusters 1 and Cluster 3 (Figure 3.6), suggesting that airway hyperresponsiveness, is a key factor in distinguishing cluster membership.

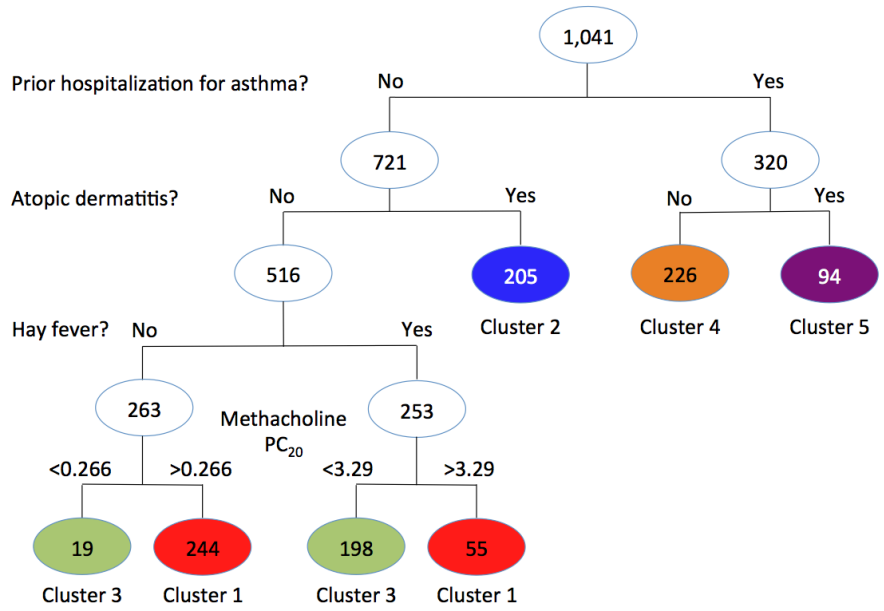
3.3.7 Cluster Validation

3.3.7.1 Longitudinal consistency in phenotype clusters With four years of prospective follow-up as part of the CAMP clinical trial, we were able to assess the consistency over time in variable distribution across the five identified clusters. As demonstrated in Figure 3.7), quantitative measures of lung function and airways responsiveness were consistently

	Cluster 1 (n=300)	Cluster 2 (n=202)	Cluster 3 (n=218)	Cluster 4 (n=225)	Cluster 5 (n=96)	P-value
AOE Classification	**LLL**	**HLM**	**HHM**	**MHH**	**HHH**	
Sex						0.45
Male	173 (57.7%)	115 (56.9%)	130 (59.6%)	146 (64.9%)	57 (59.4%)	
Female	127 (42.3%)	87 (43.1%)	88 (40.4%)	79 (35.1%)	39 (40.6%)	
Age						0.003
Years, mean(± SD)	8.79 (±2.05)	8.83 (±2.12)	9.38 (±2.13)	9.03 (±2.08)	8.46 (±2.25)	
Self-reported race						0.19
White	217 (72.3%)	137 (67.8%)	140 (64.2%)	150 (66.7%)	67 (69.8%)	
Black	36 (12.0%)	27 (13.4%)	26 (11.9%)	35 (15.6%)	14 (14.6%)	
Hispanic	27 (9.0%)	13 (6.4%)	31 (14.2%)	21 (9.3%)	6 (6.3%)	
Other	20 (6.7%)	25 (12.4%)	21 (9.6%)	19 (8.4%)	9 (9.4%)	
Annual Household Income						0.31
Less than $30,000	70 (23.3%)	40 (19.8%)	45 (20.6%)	66 (29.3%)	21 (21.9%)	
Highest Household Education						0.13
Less than high school	1 (0.33%)	1 (0.50%)	2 (0.92%)	1 (0.44%)	0 (0.00%)	
High school	5 (1.7%)	5 (2.5%)	4 (1.8%)	5 (2.2%)	4 (4.2%)	
Higher education	125 (41.7%)	78 (38.6%)	87 (39.9%)	101 (44.9%)	48 (50.0%)	
Family history						
Asthma (Any)	154 (51.3%)	114 (56.4%)	137 (62.8%)	110 (48.9%)	59 (61.5%)	0.07
Asthma (Maternal)	64 (21.3%)	46 (22.8%)	70 (32.1%)	50 (22.2%)	32 (33.3%)	0.02
Asthma (Paternal)	40 (13.3%)	49 (24.3%)	58 (26.6%)	46 (20.4%)	15 (15.6%)	0.0009
Atopy (Any)	189 (63.0%)	158 (78.2%)	162 (74.3%)	139 (61.8%)	76 (79.2%)	0.0004
Atopy (Maternal)	114 (38.0%)	110 (54.5%)	112 (51.4%)	98 (43.6%)	53 (55.2%)	0.0009
Atopy (Paternal)	87 (29.0%)	90 (44.6%)	84 (38.5%)	73 (32.4%)	37 (38.5%)	0.05
Environmental Exposures						
Tobacco smoke	119 (39.7%)	61 (30.2%)	82 (37.6%)	105 (46.7%)	46 (47.9%)	0.01
Dust mite	60 (20.0%)	36 (17.8%)	61 (28.0%)	44 (19.6%)	16 (16.7%)	0.30
Cockroach	1 (0.33%)	3 (1.49%)	1 (0.46%)	0 (0.00%)	0 (0.00%)	0.04
Randomized treatment arm in CAMP clinical trial						0.091
Budesonide	86 (28.7%)	68 (33.7%)	60 (27.5%)	68 (30.2%)	29 (30.2%)	
Nedocromil	94 (31.3%)	60 (29.7%)	66 (30.3%)	67 (29.8%)	25 (26.0%)	
Placebo	120 (40.0%)	74 (36.6%)	92 (42.2%)	90 (40.0%)	42 (43.8%)	

Table 3.11: Distribution of Non-classifying Features Across Phenotypic Clusters.

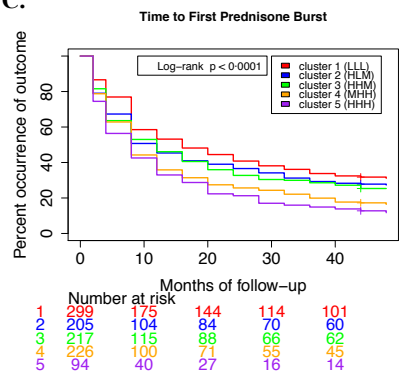
A.



B.

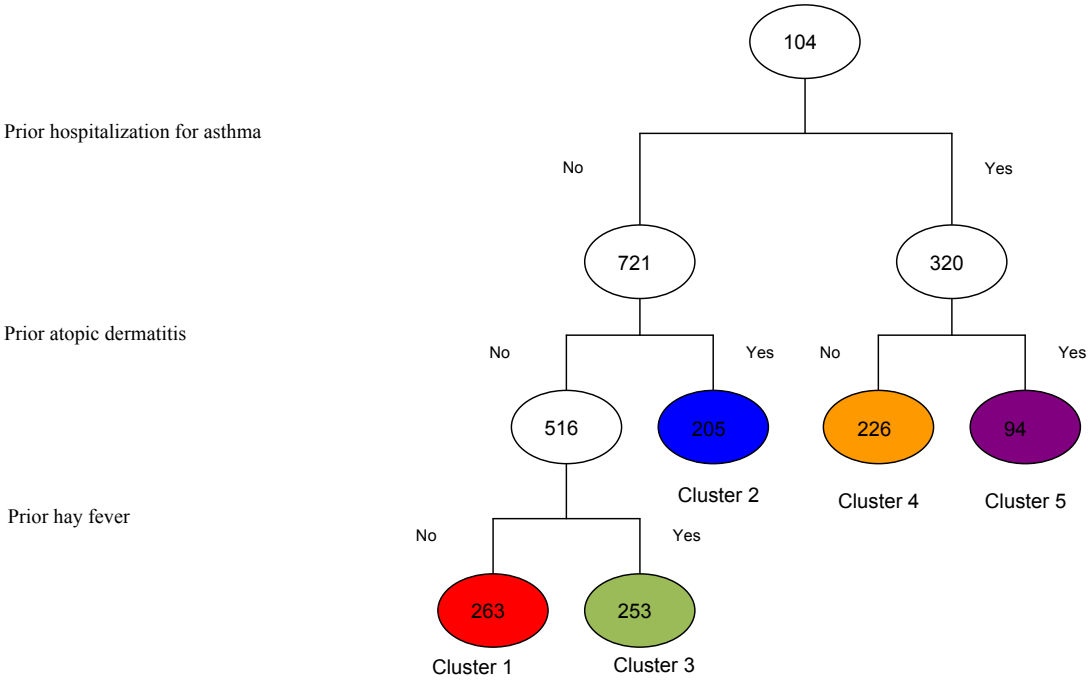
	Cluster				
	1	2	3	4	5
Sensitivity	97.0	100	95.4	100	96.9
Specificity	98.9	99.6	98.9	99.9	99.9
PPV	97.3	98.5	95.8	99.6	98.9
NPV	98.8	100	98.8	100	99.7

C.



A: Decision tree model for asthma classification. Nodes represent numbers of study subjects, branches represent cut-points for clinical variables used in the model (shown at left). End-nodes are colored corresponding to the cluster groupings in Figures 1 and 2. **B:** Performance of decision tree model in classifying subjects into asthma clusters. **C:** Kaplan-Meier plot of the cumulative probability of a first course of prednisone for the cluster assignments, as defined by the decision tree.

Figure 3.5: Asthma classification model.



	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Sensitivity	79.7	100	88.1	100	96.9
Specificity	96.8	99.6	92.6	99.9	99.9
PPV	90.9	98.5	75.9	99.6	98.9
NPV	92.2	100	96.7	100	99.7

Nodes represent numbers of study subjects, branches represent cut-points for clinical variables used in the model (shown at left). Decisions trees was constructed without the PC_20 variable.

Figure 3.6: Decision tree model for asthma classification.

separated across cluster groupings. These patterns persisted over the four-year period of observation. For example, although methacholine PC_{20} is known to demonstrate marked within-subject variability over time (for instance, the intra-class correlation of methacholine PC_{20} during the CAMP clinical trial was 0.426, 95% confidence interval, 0.391 to 0.463), the clear separation of methacholine PC_{20} across many of the clusters persisted over the

four years of observation. A notable exception was pre-bronchodilator FEV_1, which demonstrated marked instability in cluster ranking, particularly towards the end of the observation period. In contrast, post-bronchodilator FEV_1 exhibited more consistent separation, similar to other spirometric measures.

3.3.7.2 Comparison of univariate vs. multivariate cluster analysis We compared our multivariate cluster analysis using 18 variables to a univariate approach, using each of the 18 variables by itself. To evaluate the difference in the ability of each of these methods to predict future exacerbations, as measured by the time to first use of prednisone, we performed a survival analysis for each of the models. We found that the single variable with the best predictive accuracy for future exacerbations was that of history of prior hospitalizations for asthma exacerbations (Figures 3.8, 3.9). This variable was the only single variable to outperform the multivariate phenotypic clusters in terms of its ability to predict future exacerbations. However, despite this finding, we believe that the phenotypic clusters serve as an improvement on the single variable approach in terms of their ability to capture the total constellation of symptoms involved in asthma, for which exacerbation symptoms are one of many important disease features, and atopic features and obstructive symptoms serve as other important descriptors.

3.3.7.3 Reproducibility of cluster assignments using different clustering algorithms Using hierarchical clustering, we were able to generate clusters quite similar in composition to our original clusters in terms of AOE grouping (Table 3.12). To assess whether the new cluster assignments also demonstrated longitudinal consistency similar to the original clusters, we repeated our survival analysis of time to asthma exacerbation using the four years of follow-up data generated as part of the CAMP clinical trial. For the survival analysis, we found that the clusters generated using hierarchical clustering demonstrated a similar natural history to our original clusters (Figure 3.10).

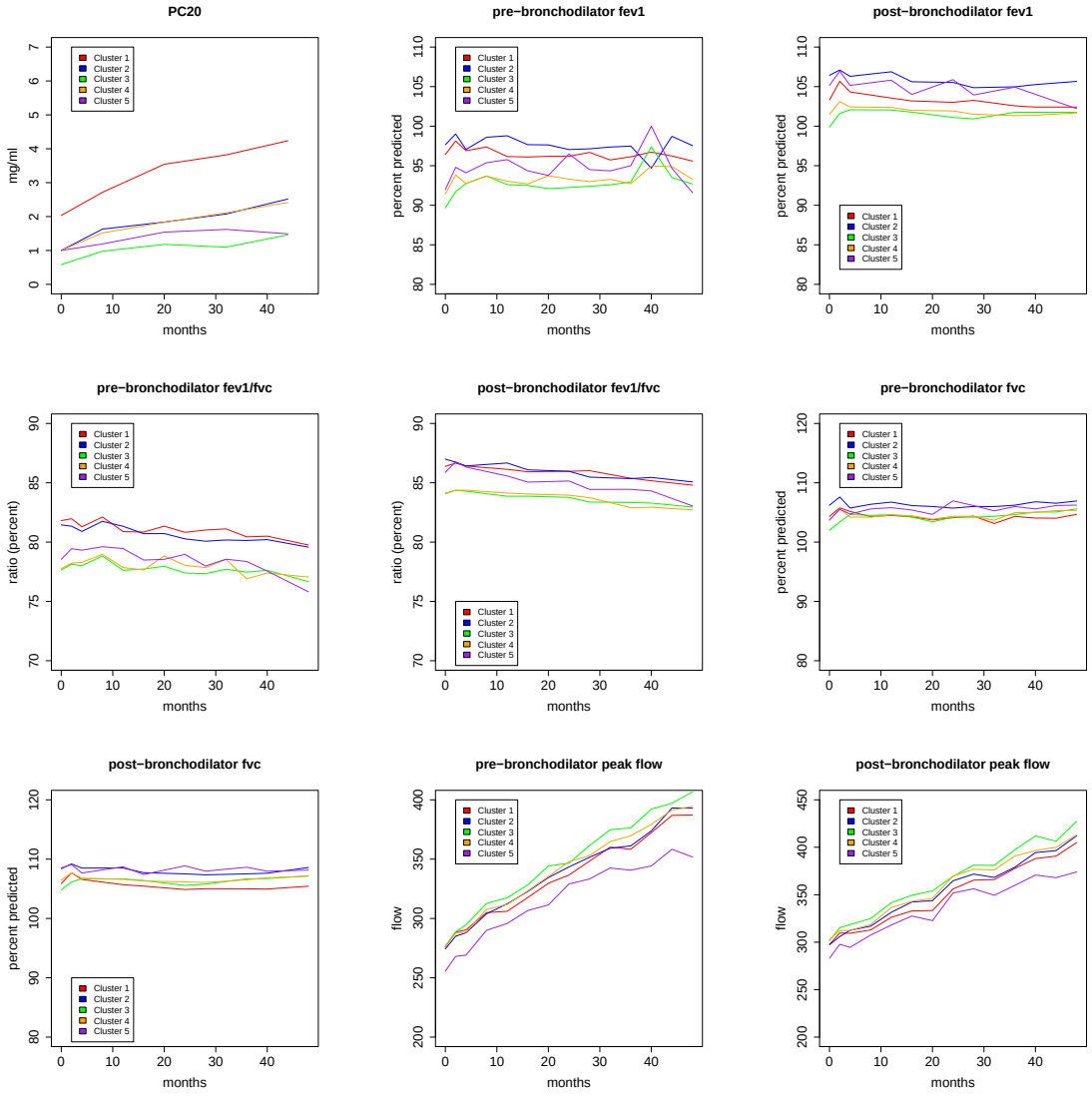
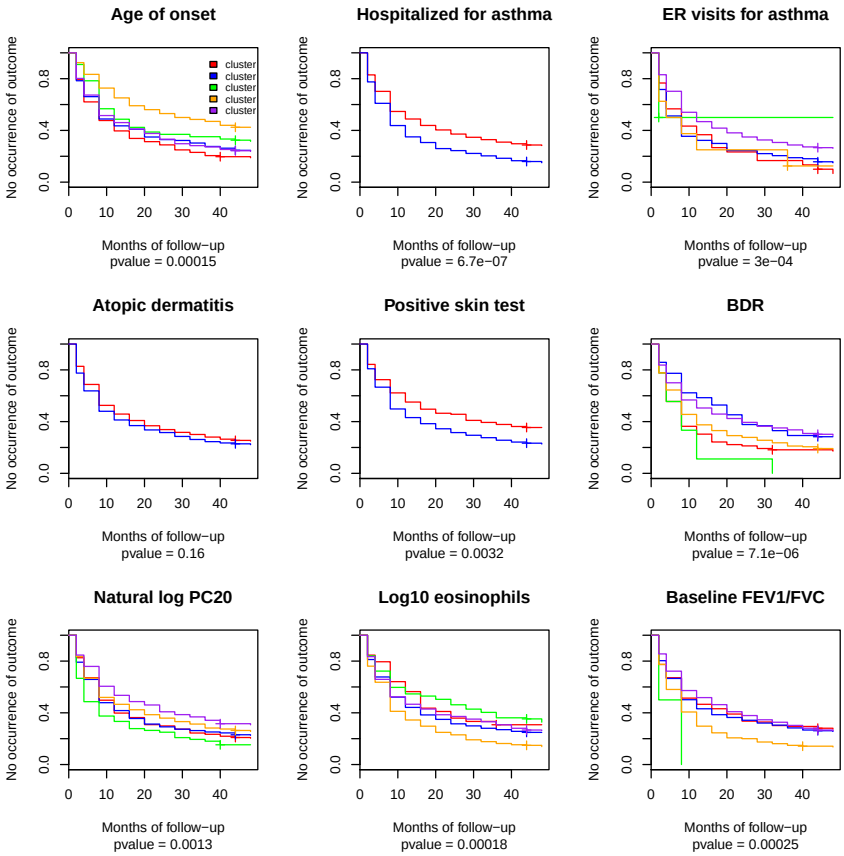
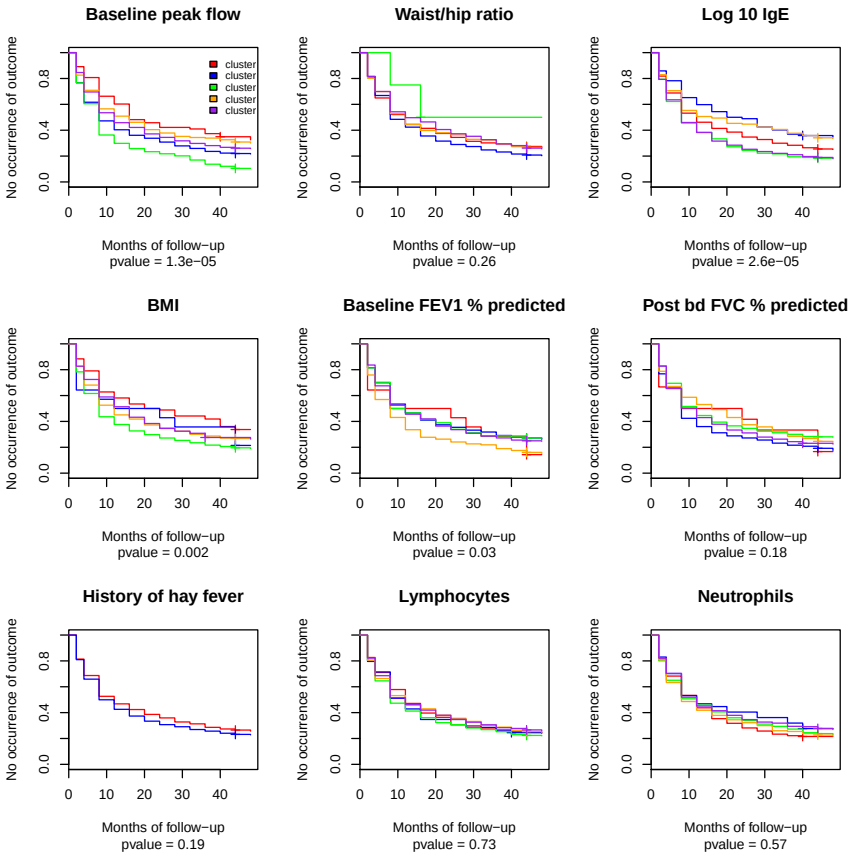


Figure 3.7: Mean pulmonary function measurements by asthma cluster over four years of follow up. P-values < 0.0001 calculated using linear mixed-effects models.



Kaplan-Meier plots by cluster of the cumulative probability of a first course of prednisone during the four-year follow-up period of the CAMP trial. Clusters were determined based upon a single clinical variable, indicated at the top of each figure.

Figure 3.8: Survival analysis for single variable cluster analysis.

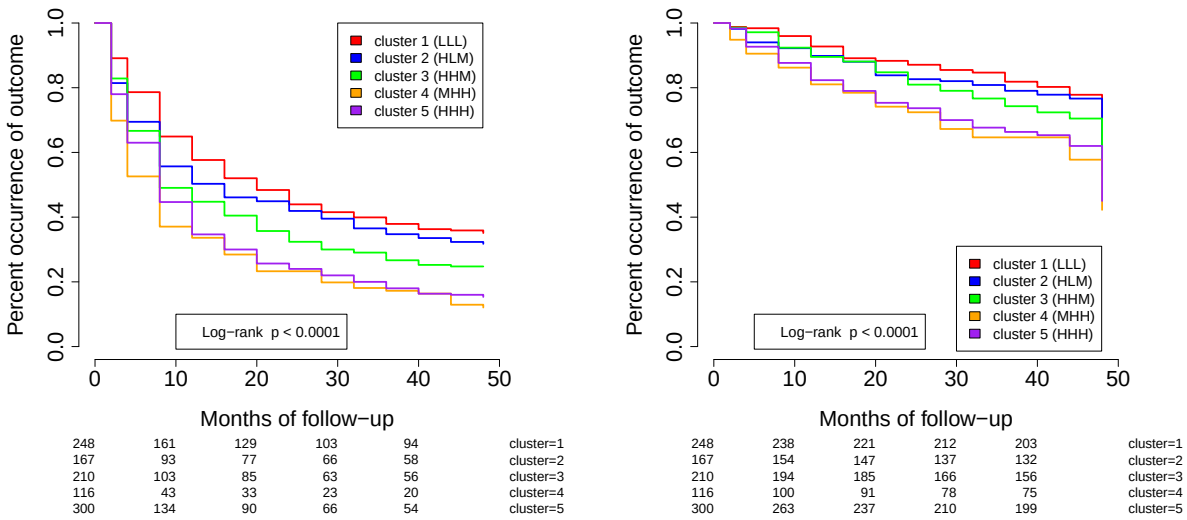


Kaplan-Meier plots by cluster of the cumulative probability of a first course of prednisone during the four-year follow-up period of the CAMP trial. Clusters were determined based upon a single clinical variable, indicated at the top of each figure.

Figure 3.9: Survival analysis for single variable cluster analysis.

		New				
		1	2	3	4	5
Old	1	188	2	7	29	2
	2	0	164	0	37	1
	3	57	0	130	27	4
	4	3	1	1	3	217
	5	0	0	0	20	76

Table 3.12: Comparison of phenotypic clusters generated by hierarchical clustering (new clusters) vs. spectral clustering (old clusters).



(a) Prednisone use

(b) Need for additional therapy

Kaplan-Meier plots by cluster of the cumulative probability of a first course of prednisone **A:** or initiation of additional asthma controller therapies (beclomethasone or other) **B:** during the four-year follow-up period of the CAMP trial for hierarchical clusters.

Figure 3.10: Survival analysis for hierarchical clusters.

3.4 DISCUSSION

The clinical heterogeneity of asthma has motivated the use of machine-learning algorithms for the classification of patients using data-driven, unbiased criteria. While earlier work [60, 118, 48] established the feasibility of this approach, many important questions remain unaddressed, including issues of reproducibility, generalizability and clinical relevance. Without longitudinal follow-up, which was unavailable in prior reports, it is unclear whether the defined clusters have clinical utility. It is in this context that the results of our phenotypic clustering efforts and longitudinal analysis can be summarized.

First, we demonstrate the longitudinal consistency of our phenotypic clusters. When we developed the clusters, we limited ourselves to the clinical data obtained during the baseline assessment of CAMP participants. Next we evaluated for changes in cluster membership over the 48 month study-period, and found remarkable consistency in phenotypic distributions over time, particularly with regard to airway hyperresponsiveness, obstruction and exacerbation rates. These findings echo those of a recent longitudinal cluster analysis that found membership in phenotypic clusters to be extremely stable over time [17]. An additional finding of our study was that different inhaled anti-inflammatory medications appeared to have no statistically significant effect on cluster membership over time, suggesting although these medications may affect day-to-day symptoms, they have minimal effect on the natural history of childhood asthma.

Second, we demonstrate the clinical utility of our phenotypic clusters. We found important between-cluster differences in response to inhaled asthma therapies, with one cluster (Cluster 4) showing decreased rates of exacerbations with both budesonide and nedocromil therapy, while another cluster (Cluster 5) showed poor response with both budesonide and nedocromil therapy. Our data suggest that although inhaled corticosteroids such as budesonide should serve as the primary treatment choice for asthma control in children with mild to moderate asthma, there are several subgroups of patients, including those with the poorest level of baseline asthma control, who appear to respond to nedocromil at levels similar to budesonide. Given safety concerns, particularly in children, regarding the long-term exposure to inhaled glucocorticoid therapy, identification of phenotypic clusters that could benefit

similarly from non-steroidal therapies would be of great value. While retrospective nature of the current study and the small size of several of the clusters limits our ability to draw firm clinical conclusions about the current results, our findings serve as the foundation for future prospective clinical trials investigating personalized responses to inhaled anti-inflammatory medications.

Finally, despite notable differences in the compositions of the patient populations, computational algorithms employed, and the variables considered in generating the clusters, our results show remarkable consistency with those obtained in the pediatric and adult SARP populations, both with respect to the number of phenotypic clusters identified (5-6 clusters in CAMP and SARP cohorts) and the patient characteristics of individual clusters. The similarity of our phenotypic clusters to those of other cohorts provides further evidence for the potential generalizability of clustering as a method of phenotyping asthmatic patients. Observed differences in the degree of atopy and airway obstruction present in the pediatric compared to the adult clusters lend further support to the hypothesis of etiological differences between childhood and adult asthma.

Our study had several limitations. First, we evaluated only children, and reports have shown that pediatric and adult asthma may represent two different disease states, with different pathogenic mechanisms and natural histories [105]. For this reason, the clinical implications of this cluster analysis may not be widely applicable to an adult asthmatic population. Second, our study did not include severe childhood asthmatics. Because our original population was ascertained for the purposes of a clinical trial, it included children with mild-moderate persistent asthma, and specifically excluded those with more severe asthma. Thus, there is a possibility that there is a severe childhood asthmatic phenotypic that was missed with our analysis, although the strong similarities in observed clusters with the childhood SARP study (which included a broader spectrum of disease severity) provides reassurance that the results of our cluster analysis are more widely applicable. Third, the conclusions that we can draw from the clinical outcomes of our clusters are limited due to their small sample size. For example, the children in Cluster 5 had a limited response to inhaled budesonide and nedocromil compared to placebo, suggesting that the children in this cluster may have some resistance to corticosteroid therapy. However, because there were

only 96 children in this cluster, we were underpowered to draw more clinically meaningful conclusions from this particular analysis. It will be necessary to validate some of these preliminary findings in future prospective studies to determine whether the children in this cluster are truly steroid-resistant.

In conclusion, our results suggest support the use of computationally-inferred phenotypic classifications of asthma as having clinical utility. These models define subsets of patients with unique clinical attributes, discrete clinical trajectories, and variable responsiveness to anti-asthma controller medications. Recognition of these clusters, and their clinical relevance should motivate novel strategies in both the research and clinical settings. More refined phenotypic classification may better inform treatment decisions: as suggested by the results of our treatment responsiveness analysis, cluster assignment identifies two subsets of patients who respond similarly to both budesonide and nedocromil, providing clinicians with viable treatment options for patients at risk for corticosteroid-related complications. The observed between-cluster differences in environmental and genetic factors suggest that important etiological differences underlie the configuration of different asthma subgroups. Future studies that consider more homogenous subsets of patients should improve research precision in characterizing the genetic and environmental etiologies. Thus, in addition to helping inform clinical management, these more refined phenotypic classification schemes should help accelerate research efforts in defining the molecular and environmental underpinnings of this complex airways disease.

4.0 GENE EXPRESSION AND ASTHMA ENDOTYPES

4.1 MEASUREMENT OF GENE EXPRESSION

In linking clinical phenotypes to mechanisms of disease, it is helpful to have biological as well as clinical data to formulate hypotheses and make inferences. For human studies, this involves obtaining either genetic material, proteins or other metabolites to explore associations between clinical symptoms and biomarkers. Gene transcripts, which may be stabilized and isolated from biological specimens are useful because they represent the genes which are actively being transcribed and are thus "turned on". Although they do not provide the amount of functional information provided by proteins and metabolites, they provide a larger sense of the active processes involved in cellular metabolism than that provide by exclusively studying the DNA genetic blueprint.

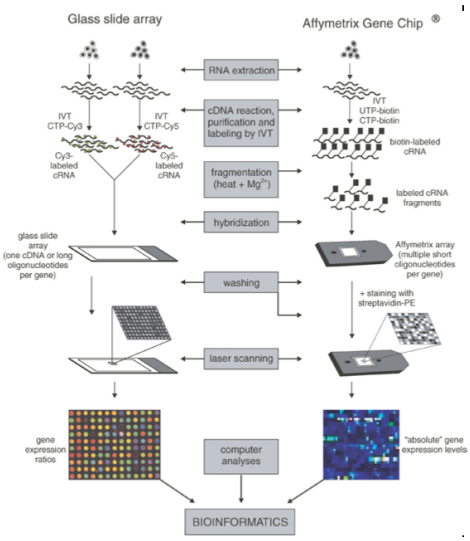
4.1.1 Molecular Biology Techniques

The ability to simultaneously profile large numbers of genes from patient samples is a relatively recent development. Traditionally, levels of gene expression have been measured by northern blotting. Northern blotting is a molecular biology technique that involves using electrophoresis to separate RNA samples by size and detection with a hybridization probe that is complementary to all or part of the target RNA sequence. After an RNA sample is separated on an electrophoresis gel, capillary transfer is used to transfer the separated RNA sequences to a blotting membrane [166]. The name for northern blotting is derived from its relationship to Southern blotting, developed by Edwin Southern as an assay for DNA sequences [151].

Northern blotting continues to be in widespread use, however one of the limitations of this approach is its relatively low throughput. Northern blotting is time-consuming, and also unable to assay more than a few RNA sequences at any one time. In order to address these limitations, Southern began researching techniques for multiplexing the assay of RNA sequences. The results of his efforts led to the development of DNA microarray technology. DNA Microarrays are used to measure the expression of large quantities of gene transcripts simultaneously. Although the specifics vary depending on the particular microarray platform used, DNA microarrays involve assembling a series of DNA probes onto a solid surface. Then RNA is isolated from a biological specimen and is used to synthesized its complementary cDNA sequence, which is more stable than RNA, which degrades rapidly after isolation. The cDNA sequence is then hybridized to the DNA attached to the solid surface, and the hybridization is detected and quantified using either fluorphore, or chemiluminescence-labeled targets to determine the relative abundance of nucleic acid sequences that have hybridized to the surface, (Figure 4.1) [152].

4.1.2 Adjustment for Multiple Testing

Along with the development of microarray technology, which has allowed for high throughput gene expression profiling, has come the development of new statistical techniques to analyze the vast quantities of data available. Although microarray results may be assessed using traditional statistical techniques, such as the parametric ANOVA and Student's t-test, and non-parametric Kruskal-Wallis and Wilcoxon rank sum, adjustments must be made due to the large number of statistical tests performed simultaneously when thousands of genes are assayed at once. In statistical inference, when multiple statistical tests are performed simultaneously, hypothesis tests that incorrectly reject the null hypothesis become increasingly likely to occur. To address this issue, new statistical techniques have been developed to prevent this occurrence, and allow significance levels for single and multiple comparisons to be compared.



Reprinted from "DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers," by F.J.T Staal et. al, 2003, *Leukemia, 17*, p. 1324-32. Copyright 2003 by the Nature Publishing Group. Reprinted with permission.

Figure 4.1: Depiction of DNA microarray workflow.

4.1.2.1 Control of the Familywise Error Rate

One of the first techniques to address the issue of multiple comparisons was the Bonferroni method [44]. This method involves control of the familywise error rate (FWER). The FWER is the probability of making one or more false discoveries, or type I errors when performing multiple hypothesis testing, that is, $FWER = Pr(V \geq 1)$ where V is the number of false positive results. The goal of methods like the Bonferroni method is to assure that the FWER is less than or equal some value, α, such that the probability of making even one type I error within a family is controlled at level α. In statistical inference, a family is understood to be the smallest set of items in an analysis from which statistical inferences may be made.

Methods that control the FWER may control it in the weak sense, such as when control of the FWER at level α is guaranteed only when all null hypotheses are true, or when $m = m_0$, where m is the total number of hypotheses and m_0 is the number of null hypotheses. In this case, the global null hypothesis is true, or may control the FWER in the strong sense, such as when control of the FWER at levelα is guaranteed for any distribution of null hypotheses, including the global null hypothesis.

For the Bonferroni method, let $H_1...H_m$ be a family of hypotheses and $p_1...p_m$ be the corresponding p-values resulting from significance testing, and let I_0 be the subset of the true null hypotheses, having m_0 members. The FWER is the probability of rejecting at least one of the members in I_0. The Bonferroni correction demonstrates that rejecting all $p_i < \frac{\alpha}{m}$ will control the FWER $\leq \alpha$, where m is the total number of hypotheses. The proof follows directly from Boole's inequality:

$$FWER = Pr\left\{\bigcup_{i_0}(p_i \leq \frac{\alpha}{m})\right\} \leq \sum_{i_0}\left\{Pr(p_i \leq \frac{\alpha}{m})\right\} \leq m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha \tag{4.1}$$

The Bonferroni method of correction is useful in situations where it is unsuitable to have just one false positive value. However, in practice, this is often not the case. Furthermore, this particular method of correction controls the rate of false positives at the expense of increasing the number of false negative results.

4.1.2.2 Control of the False Discovery Rate

For the analysis of microarrays, control of the familywise error rate is often too stringent, and results in an unacceptable level of false negative results. In order to address this problem and increase the power of statistical analyses, Benjamini and Hochberg developed a method to control the false discovery rate (FDR), as opposed to the familywise error rate. This approach involves controlling the expected proportion of false positives [12]. For the FDR method, Q is defined as the proportion of false discoveries among the total discoveries ($Q = \frac{V}{R}$), where V is the number of false positive results, and R is the number of rejected null hypotheses, or discoveries. Assuming that S is the number of true positive discoveries, the FDR is:

$$FDR = Q_e = E[Q] = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right] \tag{4.2}$$

Where $\frac{V}{R}$ is defined to be 0 when $R = 0$. Thus, the FDR may be controlled at a level α, or q, where the q-value is equivalent to the p-value in the FDR setting. The q-value of an individual hypothesis test is the minimum FDR at which the test may be considered significant. For the analysis of microarray data, it is common to directly estimate q-values as opposed to fixing the level at which to control the FDR [153].

For the Benjamini-Hochberg (BH) method to control the FDR at level α:

1. For a given α, we find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$
2. Reject all $H_{(i)}$ for $i = 1, ..., k$

The BH method is valid when the m hypothesis tests are independent as well as situations of dependence that satisfy the following inequality [13]:

$$E(Q) \leq \frac{m_0}{m}\alpha \leq \alpha \tag{4.3}$$

4.2 GENE EXPRESSION APPLICATIONS TO ASTHMA PHENOTYPING

Gene expression profiling and microarray technology have become popular in recent years as an unbiased way to understanding relationships between disease states and the levels of gene transcripts. This technology has been widely adopted in the study of asthma due to its potential to uncover novel genes and pathways involved in disease pathogenesis. One of the main applications of gene expression profiling to the study of asthma has been through profiling gene transcripts in different cell types in order to understand the mechanisms of asthma pathogenesis in different compartments.

Several early studies in asthmatic patients obtained bronchial biopsy specimens from patients with asthma. One of the earliest gene expression profiling studies was one by Dolganov and colleagues using real-time polymerase chain reaction (PCR) to quantify transcripts [39]. In another study, Laprise and colleagues performed a microarray analysis on bronchial biopsy specimens obtained from patients and compared levels of gene expression between those with mild asthma and non-asthmatic controls [104]. They found differential levels of gene expression among multiple genes, including some like nitric oxide synthase 2A (NOS2A), that had been previously implicated in asthma, and others, like arachidonate 15-lipoxygenase (ALOX15), not previously implicated in asthma pathogenesis. Other gene expression profiling studies performed on bronchial biopsy specimens include a recent study by Choy and colleagues that evaluated a large cohort of asthmatic and non-asthmatic subjects [25], and an RNA sequencing study by Yick and colleagues that uncovered differential expression of multiple novel and confirmatory asthma-related gene transcripts [186].

Gene expression profiling among asthmatic subjects has included multiple leukocyte cell types, including polymorphonuclear leukocytes [59, 3, 156, 16, 145], neutrophils [6], basophils [187], alveolar macrophages [112], $CD8^+$ T lymphocytes [167] and $CD4^+$ T lymphocytes [61, 88, 78]. Other physiologic compartments that have been studied include airway epithelial cells [177, 178] and airway smooth muscle [157]. All of the aforementioned studies describe both novel and confirmatory asthma-related changes in gene expression profiles, as well as imply the clinical utility of gene expression profiles as biomarkers for asthma phenotypes, particularly in situations where gene transcripts are assayed by non-invasive means,

such as induced sputum [7].

In addition to demonstrating differential gene expression between patients with and without asthma and with different asthma phenotypes, several studies have used classification and clustering methods to identify predictive signatures for asthma phenotypes. For example, Shin and colleagues used multiple logistic regression to optimize a gene expression signature with the power to discriminate between asthmatic and non-asthmatic subjects with high sensitivity and specificity [145]. Woodruff and colleagues performed an unsupervised hierarchical cluster analysis of gene expression profiles of asthmatic patients and evaluated for correlations between gene expression clusters and clinical characteristics [178]. Similarly, Baines et. al, performed hierarchical clustering on gene expression profiles from induced sputum specimens of asthmatic patients and correlated several clusters with clinical characteristics, such as the degree of airway obstruction present among study subjects [7]. However, all of the above studies stop short of validating their gene profiles in an independent population, limiting their clinical utility.

4.3 LIMITATIONS OF PRIOR GENE EXPRESSION ANALYSIS

As described in the previous section, there have been a large number of studies evaluating differential gene expression levels among asthmatic patients. However, there is limited information with respect to the relationships between the gene transcripts. The gene expression profiles of thousands of genes afford the unique opportunity to evaluate gene-gene interactions using computational algorithms to "reverse-engineer" gene networks. Such networks could be used to identify sets of gene transcripts with similar expression patterns, which could then be used to explore the presence of common regulatory transcription factors (gene-sequence interactions). Gene expression networks could also be used to explore common molecular pathways among gene transcripts (gene-gene interactions) [53]. Prior studies have evaluated gene clusters and signatures for enrichment in different molecular pathways, but none have expressly studied the relationships between profiled transcripts and associations with asthma phenotypes.

4.4 GENE EXPRESSION NETWORKS

When studying gene expression profiles of large numbers of genes in patient samples, it is useful to move beyond the detection differential expression levels between different patient groups and toward the exploration of interactions between sets of genes. An advantage to this approach is that it is more closely related to physiologic processes in which networks of genes transcribe proteins that are active in metabolic pathways. The hypothesis is that variation in gene expression leads to differences in these gene networks that is related to the development of disease. Gene networks thus help to provide insight into the physiology of cellular processes at the mRNA level. To this end, multiple methods have been developed in an attempt to identify functionally related gene co-expression networks.

Gene networks display causal relationships between gene transcripts and are often represented by undirected graphs where the nodes of the graph are genes and the edges are the causal relationships between genes. Gene networks can also be represented by adjacency matrices. An adjacency matrix of a finite graph, G on n vertices is the nxn matrix where the non-diagonal entry a_{ij} is the number of edges from vertex i to vertex j, and the diagonal entry, a_{jj} is the number of edges from vertex j to itself.

4.4.1 Early Methods of Inferring Gene Networks

Some of the earliest methods developed to identify gene networks were developed by the work of Kauffman [91] and Thomas [162] on random Boolean gene networks. However, the assumption that gene networks have a random topology has increasingly been questioned and more recent assumptions are that gene networks possess scale-free [9] and small-world [170] topologies, with a power law distribution for node connectivities.

Experimental mRNA levels obtained through microarray technology can provide a "snapshot" of the molecular state of cell populations at the transcript level, and are thus rich in information that may be used to reverse engineer gene networks [18]. Understanding connections between genes with similar expression patterns may help to reveal the structure behind the process of transcriptional regulation. A relatively simple method for identify-

ing potentially interacting genes has been termed the "guilt by association" method. In this method, gene associations are explored through clustering algorithms [46] and principal component analysis [68] and genes with similar expression patterns are grouped together. These methods provide a corse-grained approximation of gene-gene interactions that may then be followed up with more extensive analysis of the function of genes within the networks. These methods work well in networks with a small number of connections, but the interpretation becomes more ambiguous in the case of heavily connected networks.

4.4.2 Bayesian Networks

Other methods that have been used to develop gene networks include Bayesian belief networks [51,132,80]. Bayesian belief networks are probabilistic graphical models that represent a set of random variables and their conditional dependencies in the form of a directed acyclic graph (DAG), G [131]. In a Bayesian network graph, the nodes, $X_1, ..., X_n$, represent variables, and the edges represent conditional dependencies between nodes. The component, θ, describes a conditional distribution for each variable, given its parents in G. Thus, two nodes that are not connected by an edge are considered to be conditionally independent. Each node is associated with a probability function that takes as input a set of values from the node's parent variables and provides as output the probability of the variable represented by the node.

The graph, G, encodes the Markov assumption, such that each variable X_i is independent of its non-descendants, given its parents in G. Because G represents a series of conditional independencies, any joint distribution that satisfies the Markov assumption may be decomposed into its product form. For example, consider a finite set $\chi = \{X_1, ..., X_n\}$ of random variables where each variable X_i can take on a value x_i from the domain $Val(X_i)$, then, we may say:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}^G(X_i)), \tag{4.4}$$

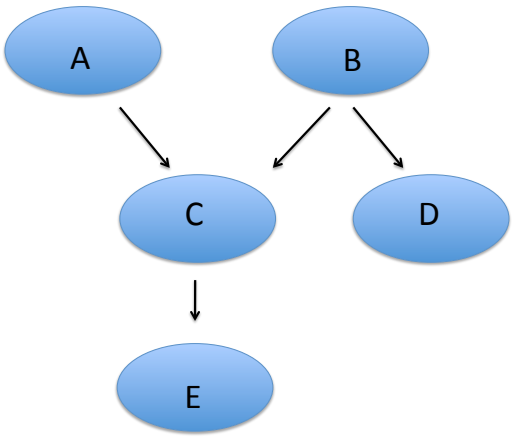
Where $\mathbf{Pa}^G(X_i)$ is the set of parents of X_i in G. To specify the joint distribution, the product form of each of the conditional probabilities must also be specified, such that

$P(X_i|\mathbf{Pa}^G(X_i))$ for each variable X_i. The parameters that specify these distributions are denoted by θ. Figure 4.2 depicts a sample Bayesian network structure and its conditional independence relationships.

In the case of gene expression, the gene network can be used to compute the probabilities of gene-gene interactions. The probability distribution is considered over all possible experimental conditions, and the state of the system is described using random variables, where the random variables represent the expression levels of individual genes. The complexity of these models may be increased by adding additional attributes, such as experimental conditions, temporal indicators, cellular locations, etc. Such models can be used to answer questions related to the dependence of a particular gene on an experimental condition and the genes that mediate direct and indirect dependencies.

Modeling Bayesian networks requires two stages: model selection (structure learning) and parameter learning. Model selection involves learning a network structure and parameter learning involves estimating the probability values associated with each network node. Bayesian networks may be learned from gene expression data by splitting the dataset into a training set, $D = x[1], ..., x[M]$ of independent samples from an unknown distribution, $P(\mathbf{X})$, and estimating this distribution using G [132]. G may be learned by introducing a statistically motivated scoring function to evaluate each network with respect to the training data and searching for the optimal network according to this score [67]. One scoring function is based upon Bayesian reasoning scores candidate graphs by their posterior probability given the data. From this network, it is possible to infer subnetworks of closely related genes and to model perturbations in experimental conditions and effects on the genes within the network.

Learning the Bayesian network structure may be reduced to an optimization problem in the space of all DAGs. The number of such graphs is super exponential in the number of variables and an exhaustive search of this space is thus intractable (NP-hard). For this reason, in practice several heuristics must be employed, such as local optimization algorithms. For example, for the sparse candidate algorithm, one identifies a small number of candidate parents for each gene from simple correlations. The one restricts the search to networks in which only the candidate parents of a variable can be its parents [51]. However, this



This network demonstrates the following conditional independence relationships: $I(A;B)$, $I(C;D|A,B)$, $I(E;A,B,D|C)$, $I(D;A,C,E|B)$, $I(A;B,D)$, and the following joint distribution: $P(A,B,C,D,E) = P(A)P(B)P(C|A,B)P(D|B)P(E|C)$.

Figure 4.2: A simple Bayesian network.

method may result in an overly restricted search space. An additional limitation of Bayesian networks when applied to gene expression datasets is that the number of genes is often much larger than the number of samples. The consequence of this is that it leads to a diffused posterior probability over an extremely large model space that cannot possibly list all of the plausible networks, given the data. To address this difficulty, smaller sets of features from within the network may be sampled using Monte Carlo strategies to estimate the posterior probability of those features, given the data [194, 174, 96].

4.4.3 Correlation Networks

An alternative method of inferring relationships between genes assayed using high-throughput methods, such as microarrays is to infer the pair-wise correlations between genes [4]. The pair-wise correlations between genes can be used to construct a gene relevance network, where the network nodes correspond to gene expression, and the i^{th} gene expression profile, x_i, is a vector whose components represent the gene expression values across m microarrays [76]. The co-expression similarity s_{ij} between genes i and j is defined as the absolute value of the correlation coefficient between their expression profiles:

$$s_{ij} = |cor(x_i, x_j)| \tag{4.5}$$

A thresholding procedure may be used to transform the co-expression similarity into a measure of connection strength (adjacency) such that an unweighted network adjacency a_{ij} between gene expression profiles is defined by hard thresholding the co-expression similarity s_{ij} as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

where τ is the hard threshold parameter. Two genes are related ($a_{ij} = 1$) if the absolute correlation between their expression profiles exceeds the hard threshold τ. The advantage

of this approach is that it leads to simple network concepts, but it may lead to the loss of information when genes with correlations below τ are labeled as ($a_{ij} = 0$), even though there is some weak correlation between them. To address this potential loss of information, it is possible to define a weighted co-expression network that preserves the continuous nature of the co-expression information. A weighted co-expression network may be defined as:

$$a_{ij} = s_{ij}^{\beta} \tag{4.7}$$

with $\beta \geq 1$. This soft threshold leads to a weighted gene co-expression network that emphasizes strong correlations while punishing weak correlations [189, 74]. A limitation of this method is that it may lead to highly connected and dense networks of gene expression values, which may be difficult to interpret. Therefore, with this method, after learning a co-expression network it is often necessary to cluster the network into smaller more highly correlated modules of genes. Genes within modules may be explored for functional relationships and involvement in similar molecular pathways.

5.0 USING GENE CO-EXPRESSION NETWORKS TO DEFINE ASTHMA ENDOTYPES

5.1 INTRODUCTION

An important implication of recent advances in our understanding of asthma phenotypic clusters is that we may use these clusters to uncover associated differences in pathogenetic mechanism, and thus have the potential to identify new therapeutic targets, with increased treatment specificity as well as new molecular biomarkers for improved clinical detection. Several studies have furthered our current understanding of the relationship between phenotypic clusters and molecular mechanism. Woodruff and colleagues profiled a selected subset of gene expression levels in asthmatic subjects and found that differences in gene expression corresponded to differences in multiple clinical measures of asthma severity, demonstrating a link between clinical phenotype and molecular mechanism [178]. Baines and colleagues subsequently found a correspondence between transcriptional profiles and different clinical characteristics in an asthmatic population. However, the cross-sectional nature of these studies limits the clinical applicability of the findings [7].

In the current analysis, our goal was to link differences in gene expression levels to longitudinally stable clinical phenotypes with demonstrated differences in response to medical therapy. In a prior analysis, we evaluated for the presence of phenotypic clusters in a cohort of children with mild-moderate persistent asthma obtained from the Childhood Asthma Management Program (CAMP) study [82]. Among these children, we identified 5 distinct phenotypic clusters with different degrees of airflow obstruction, rates of exacerbation and atopic characteristics. We further found that these clusters demonstrated both longitudinal consistency over the 48 month study period and differences in response to medical therapy. In

the current study we extend our earlier analysis through an exploration of differences in gene expression between different phenotypic clusters, with the goals of identifying novel molecular biomarkers corresponding to different phenotypes and further elucidating the differences in molecular mechanism between subjects in different clusters [77]. We uncovered the presence of distinct gene co-expression modules in $CD4^+$ lymphocytes isolated from the peripheral blood of a subset of CAMP participants. Gene expression levels within these modules were associated with different phenotypic clusters and were highly predictive of multiple clinical characteristics, such as levels of atopy and asthma control. We validated these results in an independent population, and evaluated for the presence of shared transcription factor binding sites among the genes of each module.

5.2 METHODS

5.2.1 Study Population

$CD4^+$ lymphocytes were isolated from peripheral blood samples collected from 299 subjects from four clinical centers (Baltimore, Boston, Denver, St. Louis) participating in the Childhood Asthma Management Program (CAMP) Continuation Study, part 2 (CAMPCS/2). CAMP was a multi-center randomized, double-masked clinical trial of the long-term effects of three inhaled treatments for mild to moderate childhood asthma, with 1041 subjects enrolled [81]. Two subsequent 4-year observational follow-up studies of CAMP participants, CAMPCS/1 and CAMPCS/2 were carried out upon completion of the original CAMP study. Blood samples and clinical data for the current study were obtained during a routine CAMPCS/2 clinical visit between May 1, 2004 and July 31, 2007. The study visit included questionnaire assessments of asthma symptoms and medication use.

5.2.2 RNA Extraction and Microarray Preprocessing

We isolated $CD4^+$ T cells from the collected mononuclear cell layer using anti-CD4+ microbeads by column separation (Miltenyi Biotec, Auburn, CA) [86, 193]. Total RNA was

extracted using the RNeasy Mini Protocol (Qiagen, Gaithersburg, MD) [22, 57, 58]. Expression profiles were generated with the Illumina HumanRef8 v2 BeadChip arrays (Illumina, San Diego, CA) according to protocol. Arrays were read using the Illumina BeadArray scanner and analyzed using BeadStudio (version 3.1.7) without background correction. Raw expression intensities were processed using the *lumi* package [40] of Bioconductor with background adjustment with Robust Multi-Array Average (RMA) convolution [83] and $\log_2$ transformation of each array. The combined samples were quantile normalized. The complete raw and normalized microarray data are available through the GeneExpression Omnibus of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/geo/, accesion ID GSE22324).

5.2.3 Identification of differentially expressed genes

In order to classify gene expression levels from multiple phenotypic clusters into differential expression patterns, we used an empirical Bayes hierarchical modeling approach to calculate the posterior probability of each gene expression value fitting a particular pattern of expression [121, 94, 188]. For example, for this analysis, we were interested in patterns of differential expression of genes across different phenotypic clusters. We developed a set of 49 theoretical pattern assumptions (Table 5.1), such as the assumption of the null hypothesis of no differential expression across clusters for a gene, or the assumption of differential expression across all cluster for a gene, and then calculated the posterior probability of each gene fitting a particular pattern of expression. We assigned genes to the gene pattern with maximum posterior probability.

5.2.4 Identification of gene co-expression modules

To identify dense subnetworks of genes with highly correlated expression levels, we used the gene transcripts to model a weighted co-expression network [102]. The advantage of the weighted co-expression network is that it does not rely upon arbitrary thresholds to determine the presence of a connection between two transcripts. Instead, the weighted network uses all correlations to develop a soft threshold. For transcripts that were identified as differentially

Pattern number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Interpretation
1	1	1	1	1	1	null hypothesis (none different)
2	1	2	3	4	5	all different
3	2	1	1	1	1	cluster 1 different, others similar
4	1	2	1	1	1	cluster 2 different, others similar
5	1	1	2	1	1	cluster 3 different, others similar
6	1	1	1	2	1	cluster 4 different, others similar
7	1	1	1	1	2	cluster 5 different, others similar
8	2	2	1	1	1	cluster 1, 2 similar, others similar
9	2	1	2	1	1	cluster 1, 3 similar, others similar
10	2	1	1	2	1	cluster 1, 4 similar, others similar
11	2	1	1	1	2	cluster 1, 5 similar, others similar
12	1	2	2	1	1	cluster 2, 3 similar, others similar
13	1	2	1	2	1	cluster 2, 4 similar, others similar
14	1	2	1	1	2	cluster 2, 5 similar, others similar
15	1	1	2	2	1	cluster 3, 4 similar, others similar
16	1	1	2	1	2	cluster 3, 5 similar, others similar
17	1	1	1	2	2	cluster 4, 5 similar, others similar
18	1	2	3	1	1	cluster 1, 4, 5 similar, others different
19	1	1	2	3	1	cluster 1, 2, 5 similar, others different
20	1	1	1	2	3	cluster 1, 2, 3 similar, others different
21	2	3	1	1	1	cluster 3, 4, 5 similar, others different
22	2	2	3	1	1	cluster 1, 2, similar, cluster 4, 5 similar
23	2	2	1	3	1	cluster 1, 2, similar, cluster 3, 5 similar
24	2	2	1	1	3	cluster 1, 2, similar, cluster 3, 4 similar
25	2	3	2	1	1	cluster 1, 3, similar, cluster 4, 5 similar
26	2	1	2	3	1	cluster 1, 3, similar, cluster 2, 5 similar
27	2	1	2	1	3	cluster 1, 3, similar, cluster 2, 4 similar
28	2	1	1	2	3	cluster 1, 4, similar, cluster 2, 3 similar
29	2	3	1	2	1	cluster 2, 4, similar, cluster 3, 5 similar
30	2	1	3	2	1	cluster 2, 4, similar, cluster 2, 5 similar
31	2	1	1	3	2	cluster 1, 5, similar, cluster 2, 3 similar
32	2	1	3	1	2	cluster 1, 5, similar, cluster 2, 4 similar
33	2	3	1	1	2	cluster 1, 5, similar, cluster 3, 4 similar
34	1	1	2	3	4	cluster 1, 2, similar, others different
35	1	2	1	3	4	cluster 1, 3, similar, others different
36	1	2	3	1	4	cluster 1, 4 similar, others different
37	1	2	3	4	1	cluster 1, 5, similar, others different
38	2	1	1	3	4	cluster 2, 3, similar, others different
39	2	1	3	1	4	cluster 2, 4, similar, others different
40	2	1	3	4	1	cluster 2, 5, similar, others different
41	2	3	1	1	4	cluster 3, 4, similar, others different
42	2	3	1	4	1	cluster 3, 5, similar, others different
43	2	3	4	1	1	cluster 4, 5, similar, others different
44	1	1	2	1	3	cluster 1, 2, 4 similar, others different
45	1	2	1	3	1	cluster 1, 3, 5 similar, others different
46	1	2	1	1	3	cluster 1, 3, 4 similar, others different
47	2	1	3	1	1	cluster 2, 4, 5 similar, others different
48	2	1	1	3	1	cluster 2, 3, 5 similar, others different
49	2	1	1	1	3	cluster 2, 3, 4 similar, others different

Table 5.1: Description of Gene Pattern Interpretations.

expressed (DE), we constructed an adjacency matrix. Each entry in the adjacency matrix was determined by the absolute value of the Pearson's correlation coefficient between two gene transcripts (x_i, x_j), adjusted so that the overall network was scale-free. The pairwise connection strength between different transcripts (x_i, x_j) was calculated by the adjacency function $a_{ij} = |cor(x_i, x_j)|^{\beta}$, using the estimated power parameter, β to form a weighted co-expression network.

We developed a soft threshold by selecting the parameters leading to a scale-free network. We used the top transcripts from this network in a topological overlap matrix (TOM) calculation, and 1-TOM was used as a distance matrix for subsequent hierarchical clustering to form highly correlated co-expression modules.

5.2.5 Identification of Shared Regulatory Regions within Gene Co-Expression Modules

In order to identify regulatory motifs among the genes within highly correlated co-expression modules, we obtained the genomic coordinates for each transcript within each module and the corresponding promoter sequence by querying the UCSC Genome Bioinformatics [168] and BioMart data resources [15]. We searched position frequency matrices (PFMs) corresponding to transcription factor motif matches within each promoter sequence using data from publicly available sources, including the Human Protein DNA Interactome (hPDI) database [184], and JASPAR. [142] We next mapped the PFMs corresponding to each binding motif back to our set of promoter sequences to obtain sequence matches to the binding motif among our set of promoters. We used a multinomial model with a Dirichlet conjugate prior to calculate a probability score for each promoter-motif match, and considered a match successful if the minimum score was greater than 90%.

5.2.6 Gene Ontology Enrichment Analysis

We performed a Gene Ontology (GO) enrichment analysis on the differentially expressed transcripts in each gene co-expression modules. For each gene co-expression module, we calculated all enrichments in the specified ontologies (CC = cellular component, MF =

molecular function, BP = biological process), and collected information about the terms with highest enrichment. We calculated an enrichment p-value for each GO term identified using a Fisher exact test to evaluate the number of co-expression module genes present in a particular GO ontology compared to the total number of background genes in that GO category. As background we used all genes present in all of the GO categories (in any of the ontologies).

5.2.7 Validation in an Independent Cohort

To assess the generalizability of the association between the gene co-expression modules and atopy, we evaluated whether the genes in the blue module could be used to predict atopic status in an independent cohort (N = 88) of atopic (N = 72) and non-atopic (N = 16) subjects with (N = 68) and without asthma (N = 20). We used a gene expression dataset that was publicly available on the GEO website (GSE473) and has been previously described [116]. We used the genes present in the blue module to grow a binary recursive partitioning decision tree to predict phenotype cluster assignments within our patient population [20, 139].

5.3 RESULTS

5.3.1 Distribution of phenotypic traits

Clinical phenotype data was available for all 299 participants. The characteristics assessed within one month of the time blood was obtained for microarray analysis are presented in Table 5.2. The clinical characteristics presented in the table represent follow up data obtained between 9-14 years after the onset of the original CAMP study. The mean age of study subjects was 20.4 years of age, compared to 5-12 years in the original study. The ethnic and gender distributions were similar to those of the original study (Table 5.2).

Several measures of atopic burden were obtained at the time of sample collection, including serum IgE and eosinophil levels. As was observed at the onset of the original study, the degree of atopy was highest in Clusters 2,3 and 5 and lowest in Cluster 1 and Cluster

4. Similarly, the spirometric values show a similar distribution to that obtained at baseline, with Cluster 4 and Cluster 5 showing the highest levels of airway obstruction.

The number of active smokers represented a minority of this cohort (11.4%), with the highest percentage of smokers in Cluster 3 (16.3%) and the lowest percentage of smokers in Cluster 4 (4.3%).

5.3.2 Gene Transcripts Demonstrate Atopic Patterns of Expression

In order to understand the relative contribution of different genes to the formation of the asthma phenotypic clusters, we performed gene expression profiling of individuals from different phenotypic clusters to detect patterns of expression. For the set of phenotypic clusters, gene expression levels could be sorted into 49 distinct theoretical patterns (Table 5.1).

For each transcript in each phenotypic cluster, we calculated the posterior probability for each of the 49 patterns and assigned the transcript to the expression pattern with maximum posterior probability (MPP). Differentially expressed (DE) transcripts were defined as those with MPP greater than a specific threshold set to limit the false discovery rate (FDR) to < 0.05 for each of the DE patterns (2-49 in the table). Using this approach, we found that 99.7 percent of the DE transcripts were confined to 2 of the 49 possible DE patterns.

The expression pattern containing the highest number differentially expressed transcripts (22,119 of 22,184 total transcripts) was the null hypothesis expression pattern, or the pattern of no difference in expression between the different phenotypic clusters. The expression pattern containing the second highest number differentially expressed transcripts (501 of 22,184 total transcripts) was the pattern of similar expression between Clusters 1 and 4 and between Clusters 2, 3 and 5. Our earlier analysis of the clinical data from the CAMP study demonstrated that Cluster 1 and Cluster 4 had the lowest atopic burden of the phenotypic clusters. That is, at the time of initial recruitment to the CAMP study, clusters 1 and 4 had the lowest levels of atopic dermatitis (Cluster 1 = Cluster 4 = 0%), the lowest history of hay fever (Cluster 1 = 20.3%, Cluster 4 = 52.9%), the lowest history of a positive skin test (Cluster 1 = 76.7%, Cluster 4 = 88%), and the lowest $\log_{10}$ total serum IgE levels (Cluster 1 = 2.37, Cluster 4 = 2.64). Thus, a large number of differentially expressed genes

AOE Classification	Cluster 1 N=102 LLL	Cluster 2 N=50 HLM	Cluster 3 N=49 HHM	Cluster 4 N=70 MHH	Cluster 5 N=28 HHH
Demographics					
Age (years)	20.1 ± 2.0	20.3 ± 2.2	21.0 ± 2.2	20.7 ± 2.1	19.9 ± 2.4
Male (%)	59 (57.8)	37 (74.0)	24 (49.0)	44 (62.9)	20 (71.4)
Female (%)	43 (42.2)	13 (26.0)	25 (51.0)	26 (37 .1)	8 (28.6)
White (%)	84 (82.4)	35 (70.0)	34 (69.4)	53 (75.7)	21 (75.0)
African American (%)	11 (10.8)	13 (26.0)	10 (20.4)	16 (22.9)	6 (21.4)
Hispanic (%)	7 (6.9)	2 (4.0)	5 (10.2)	1 (1.4)	1 (3.6)
Atopic Features					
Serum IgE (log_{10})	2.33 ± 0.66	2.85 ± 0.54	2.53 ± 0.60	2.60 ± 0.54	2.69 ± 0.52
Serum Eosinophils (log_{10})	2.17 ± 0.50	2.37 ± 0.46	2.42 ± 0.35	2.34 ± 0.32	2.44 ± 0.31
Spirometry					
Pre-bronchodilator FEV_1 (% predicted)	98.2 ± 13.2	98.5 ± 10.9	98.4 ± 11.3	94.1 ± 13.7	97.0 ± 11.1
Pre-bronchodilator FVC_1/FVC (% predicted)	78.5 ± 7.67	78.5 ± 7.89	78.6 ± 7.55	76.2 ± 7.57	75.3 ± 9.64
Pre-bronchodilator peak flow	576.6 ± 144.5	634.0 ± 157.1	564.0 ± 147.0	579.8 ± 137.8	598.8 ± 162.7
Airway responsiveness					
Methacholine PC_{20} (natural log)	1.09 ± 0.56	0.91 ± 0.53	0.75 ± 0.47	0.92 ± 0.52	0.96 ± 0.58
Environmental Exposures					
Tobacco Smoking (%)					
Yes	14 (13.7)	7 (14.0)	8 (16.3)	3 (4.3)	2 (7.1)
No	76 (74.5)	38 (76.0)	34 (69.4)	60 (85.7)	21 (75.0)
Average cigarettes smoked per day	1.3 ± 4.1	1.2 ± 3.7	1.9 ± 4.7	0.5 ± 2.8	0.5 ± 2.1

Table 5.2: Characteristics of Study Subjects.

demonstrated an expression pattern that was associated with the degree of atopic burden present among study subjects at the time of enrollment in the CAMP study, suggesting atopic status was the primary driver of the change in gene expression for these transcripts.

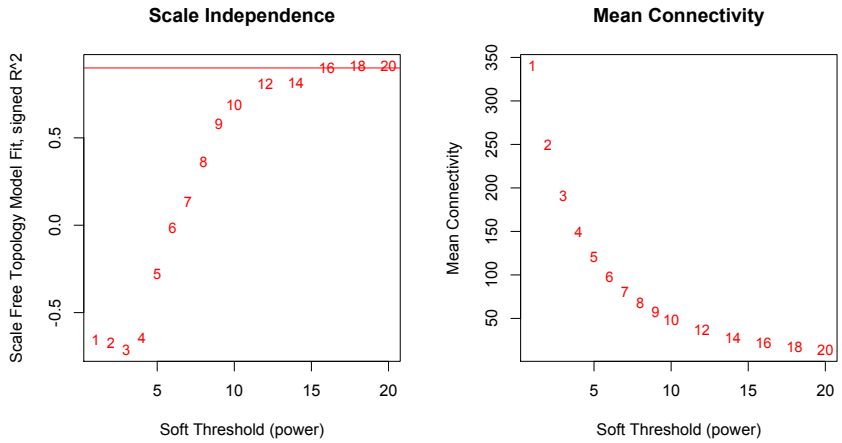
5.3.3 Atopic Patterns of Expression form Highly Correlated Co-Expression Modules

We explored the relationship between the 501 genes associated with atopic status by determining whether this set of genes could be clustered into tightly correlated co-expression modules with similar function and regulatory mechanisms. We calculated the correlation coefficient among all of the transcripts that were differentially expressed for the atopic pattern. We next partitioned these transcripts into co-expression network modules using the previously described method [102,92] because previous work has demonstrated such modules to be related to biological function [21,54,55,75].

In order to construct our weighted co-expression network, we determined the appropriate soft-thresholding power β to which co-expression similarity is raised to calculate adjacency [190] based on the criterion of approximate scale-free topology. We chose a set of candidate powers and examined the scale independence and mean connectivity for different power thresholds. The result is shown in Figure 5.1. We chose a power of 16, which was the lowest power for which the scale-free topology fit index reached 0.90.

Next we calculated the adjacencies between transcripts, using the soft-thresholding power of 16. To minimize the effects of noise and spurious associations, we transformed the adjacency matrix into a Topological Overlap Matrix (TOM), and calculated the corresponding dissimilarity. We used hierarchical clustering to produce a hierarchical clustering tree dendrogram of gene transcripts.

In the dendrogram, each leaf (short vertical line) corresponds to a specific transcript. Branches of the dendrogram group together densely interconnected, highly co-expressed genes. In order to identify network modules, shown qualitatively as branches of the dendrogram, we needed to partition the branches of the dendrogram, which we did by using the Dynamic Tree Cut algorithm [103]. A representation of the gene co-expression network



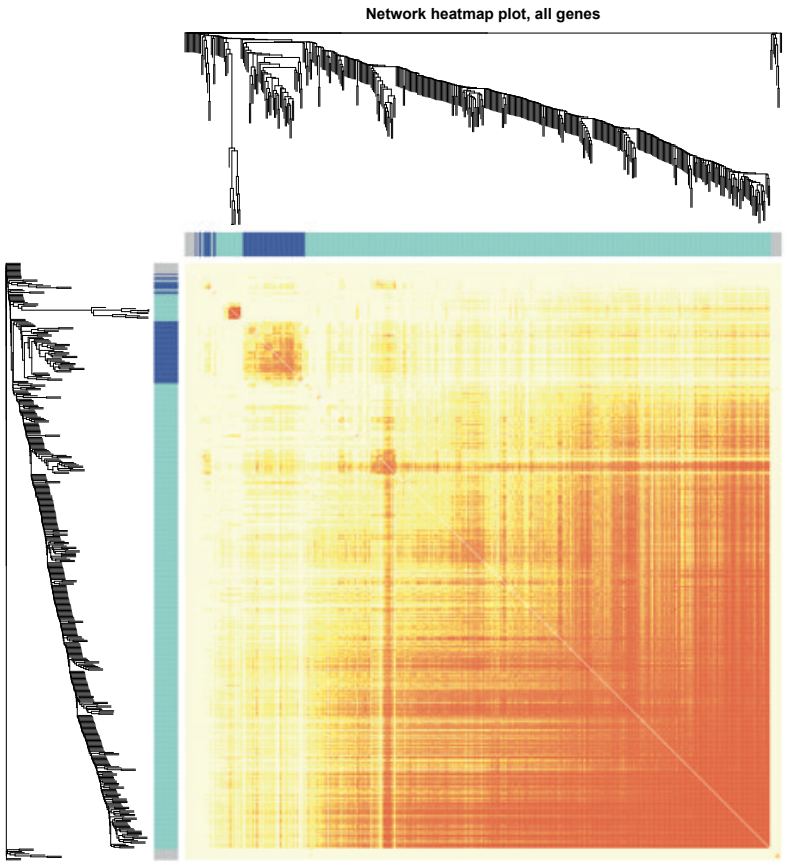
The left panel shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The right panel shows the mean connectivity (degree, y-axis) as a function of the soft thresholding power (x-axis).

Figure 5.1: Analysis of network topology for different soft-thresholding powers.

clustered into highly correlated modules within that network is shown in Figure 5.2. Cluster analysis grouped the gene co-expression network into 3 modules, shown in blue, turquoise and grey. The most notable module of highly correlated gene transcript is the blue module, while the relationship between genes within the grey and turquoise modules are more diffuse.

5.3.4 Gene Co-expression Modules Have Similar Regulatory Domains

To explore the presence of a common regulatory molecule for the genes in each co-expression module, we evaluated for common motifs in the promoter region of genes within each module. For the blue module, we identified 10 binding motifs with matches to the promoter region of genes within that module. We used the PWMs for each binding motif to find sequences matches within the promoter region of module genes. We found two motifs located within the Zinc finger protein 3 gene (ZNF3) with 373 and 308 matches to promoter sequences, depicted in Figure 5.3. These two motifs were similar in composition and each had multiple matches to the promoter regions of the 63 gene transcripts within the co-expression module, suggesting the possibility that either ZNF3 or a related protein may be a dimeric or



Higher levels of correlated expression are shown in red. The axes display the clustering map with a color bar for each module. Within the network are several modules, with the blue module showing a highly correlated cluster of gene transcripts. A small highly correlated cluster is also present within the turquoise module.

Figure 5.2: Representation of the atopic gene co-expression network and its modules.

multimeric transcription factor.

5.3.5 Enrichment Analysis of Module Genes

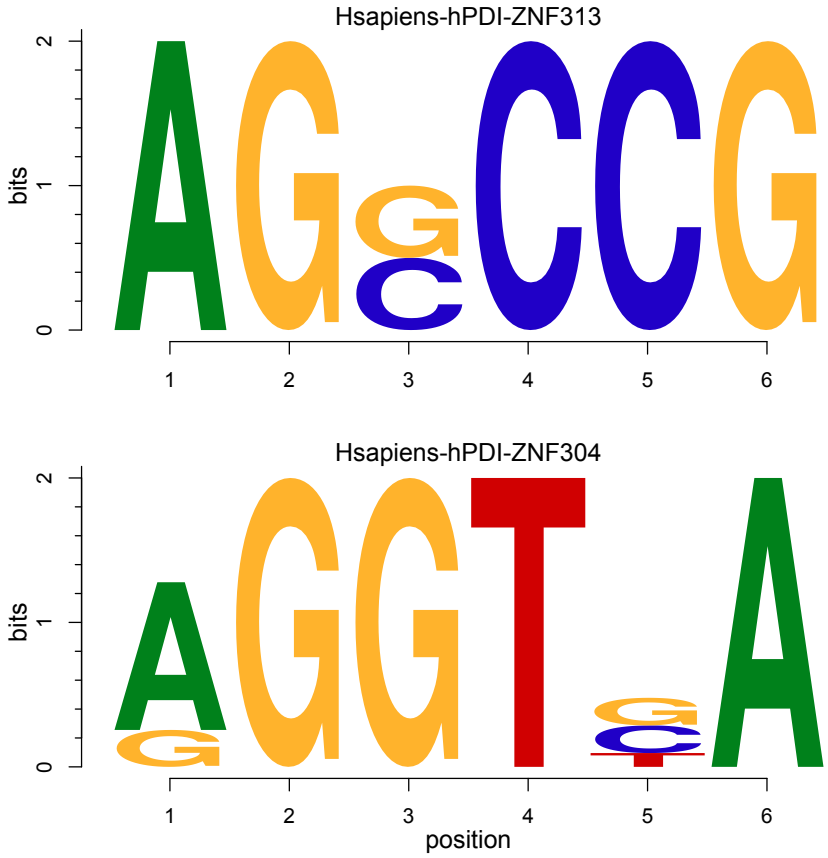
Among the gene transcripts represented in our atopic expression pattern, we identified 3 distinct co-expression modules. The module assignment is plotted under the gene dendrogram in Figure 5.4. In order to better characterize the function of the 501 genes associated with atopic burden in these 3 co-expression modules, we performed an enrichment analysis to determine whether the set of genes in each module was enriched in different biological processes, molecular functions and cellular locations. We assessed the Gene Ontology (GO) annotations for the genes present in each of the modules. Table 5.3 depicts each module and the primary GO categories for which each one was enriched. The blue and grey modules were enriched for multiple immunologic categories, and we performed further analysis on these modules because we hypothesized that the functional categories suggested potential mechanisms of asthma pathogenesis.

5.3.6 Differentially Expressed Genes are Associated with Different Clinical Outcomes

To explore the clinical relevance of these gene co-expression modules, we used linear model to detect associations between genes within each co-expression module and a range of clinical outcomes. We found that for the blue co-expression module, there was a subset of genes that were significantly associated with self-reported activity limitation. (see Table 5.4)

5.3.7 Gene Co-expression Modules are Predictive of Atopic Status

The atopic signature developed from the blue module of gene expression profiles in the CAMP dataset was predictive of atopic status in an independent dataset (see Table 5.5, Figure 5.5). Of the 63 genes present in the blue module, five select genes successfully classified all of the atopic patients in an independent cohort with a sensitivity of 100%. The genes in the atopic signature were highly enriched for multiple immunologic pathways (see Table 5.6). Evaluation of the genes present within the blue module revealed that of the 63 genes present



The relative size of the letters indicates their frequency in the sequences. The total height of the letters depicts the information content of the position, in bits.

Figure 5.3: Sequence logo for the two motifs with the largest number of matches to module promoter sequences.

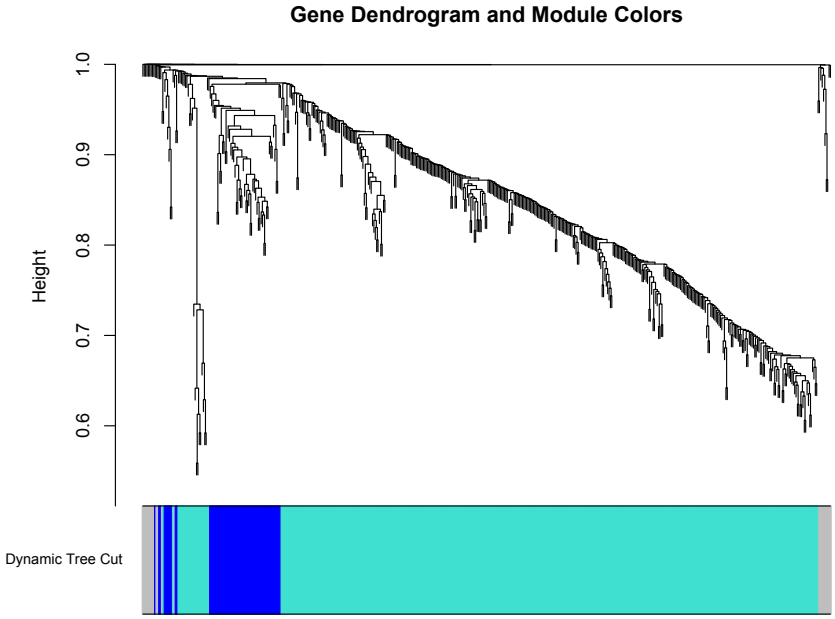


Figure 5.4: Clustering dendrogram of genes, with dissimilarity based on topological overlap, with assigned module colors.

Module	Size	P-value	Bonferroni Correction	Genes in Term	Ontology	GO Term Name
blue	62	2.2e-05	2.9e-01	21	CC	nuclear part
blue	62	6.8e-05	9.1e-01	5	BP	pattern recognition receptor signaling p
blue	62	7.2e-05	9.6e-01	4	BP	TRIF-dependent toll-like receptor signal
blue	62	7.9e-05	1.0e+00	5	BP	innate immune response-activating signal
blue	62	8.0e-05	1.0e+00	4	BP	MyD88-independent toll-like receptor
blue	62	1.0e-04	1.0e+00	4	BP	toll-like receptor 3 signaling pathway
blue	62	1.0e-04	1.0e+00	5	BP	activation of innate immune response
blue	62	1.2e-04	1.0e+00	54	CC	intracellular
blue	62	1.5e-04	1.0e+00	4	BP	Toll signaling pathway
blue	62	1.8e-04	1.0e+00	6	BP	immune response-regulating signaling
grey	22	2.4e-04	1.0e+00	2	BP	response to gonadotropin stimulus
grey	22	3.4e-04	1.0e+00	2	MF	tumor necrosis factor receptor binding
grey	22	5.3e-04	1.0e+00	2	BP	defense response to Gram-positive bacteria
grey	22	5.6e-04	1.0e+00	2	BP	response to activity
grey	22	5.8e-04	1.0e+00	9	BP	phosphate-containing compound metabolic
grey	22	7.2e-04	1.0e+00	2	MF	tumor necrosis factor receptor superfamily
grey	22	8.0e-04	1.0e+00	2	BP	leukocyte cell-cell adhesion
grey	22	1.1e-03	1.0e+00	1	BP	negative regulation of L-glutamate
grey	22	1.1e-03	1.0e+00	1	BP	positive regulation of translational
grey	22	1.1e-03	1.0e+00	1	BP	negative regulation of branching
turquoise	413	1.0e-16	1.3e-12	342	CC	intracellular
turquoise	413	3.7e-15	5.0e-11	305	CC	intracellular organelle
turquoise	413	7.6e-15	1.0e-10	285	CC	intracellular membrane-bounded organelle

Table 5.3: GO Enrichment Analysis.

Name	Symbol	p-value	BH adjusted p-value
phospholipase C $\beta 2$	PLCB2	0.01	0.22
mannosidase, α, class 2B, member 1	MAN2B1	0.02	0.22
SH3KBP1 binding protein 1	SHKBP1	0.02	0.22
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2	SMARCC2	0.02	0.22
chromosome 19 open reading frame 24	C19orf24	0.02	0.22
mediator complex subunit 25	MED25	0.03	0.22
adrenocortical dysplasia homolog (mouse)	ACD	0.03	0.22
ubiquinol-cytochrome c reductase core protein I	UQCRC1	0.03	0.22
calcineurin binding protein 1	CABIN1	0.04	0.22
poly (ADP-ribose) polymerase family, member 10	PARP10	0.04	0.22
proline-serine-threonine phosphatase interacting protein 1	PSTPIP1	0.04	0.24

BH = Benjamini-Hochberg adjustment for multiple testing.

Table 5.4: Blue Module Genes Associated with Activity Limitation.

Metric	(%)
Sensitivity (%)	100
Specificity (%)	87.5
Positive Predictive Value (PPV) (%)	100
Negative Predictive Value (NPV) (%)	87.5

Table 5.5: Accuracy of Atopic Gene Signature in an Independent Population.

in the module, a greater proportion of genes were underrepresented (Table 5.7) by the atopic clusters than overrepresented Table 5.8.

5.4 DISCUSSION

Our analysis of gene expression profiles obtained from CAMP participants 9-14 years from the study onset was notable for several key findings. First, regarding our previous cluster analysis, in which we detected 5 phenotypic clusters using the baseline clinical data from study subjects, we found that even after 9-14 years there continued to be longitudinal consistency in the clinical characteristics of subjects within different phenotypic clusters. Second, subjects with a higher degree of atopic features demonstrated differential expression in a subset of genes enriched for immunologic processes closely tied to asthma pathogenesis. A subset of these differentially expressed genes formed an atopic signature that we used to successfully determine atopic status from gene expression profiles obtained from an independent population of asthmatic subjects. Third, from our co-expression module of atopic genes, we were able to define several highly correlated subnetworks with similar expression levels. Evaluation of a selected subnetwork revealed the presence of a common motif among the promoter regions of genes within the network that corresponded to a binding site for a zinc-finger transcription factor.

In an earlier cluster analysis we performed using clinical data from participants in the CAMP study, we found that children could be characterized in terms of 5 distinct pheno-

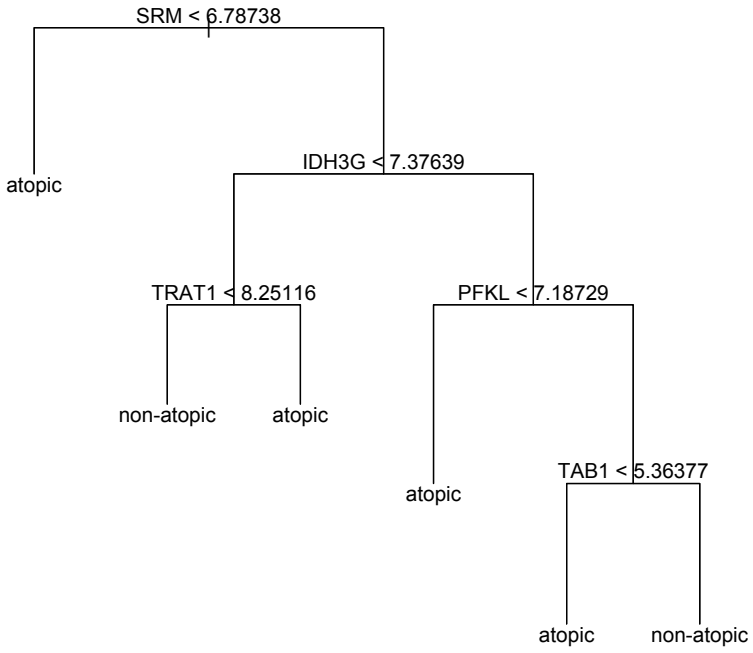


Figure 5.5: Decision Tree Classification Model for Atopic Status.

Name	Symbol	Atopic expression relative to non-atopic	GO Annotation
spermidine synthase	SRM	down-regulated	GO:0003824:catalytic activity
isocitrate dehydrogenase 3 (NAD+) gamma	IDH3G	down-regulated	GO:0006099:tricarboxylic acid cycle GO:0004449:IDH (NAD+) activity GO:0000287:magnesium ion binding
T cell receptor associated transmembrane adaptor 1	TRAT1	up-regulated	GO:0006968:cellular defense response GO:0007173:EGFR receptor signaling GO:0038095:Fc-epsilon receptor signaling pathway GO:0008543:fibroblast growth factor receptor signaling GO:0008543:innate immune response GO:0001920:negative regulation of receptor recycling GO:0051051:negative regulation of transport GO:0048011:neurotrophin TRK receptor signaling pathway GO:0048015:phosphatidylinositol-mediated signaling GO:0050850:positive regulation of calcium-mediated signaling GO:0050862:positive regulation of T cell receptor signaling pathway
phosphofructokinase, liver	PFKL	down-regulated	
TGF-beta activated kinase 1/ MAP3K7 binding protein 1	TAB1	down-regulated	GO:0000187:activation of MAPK activity GO:0000185:activation of MAPKKK activity GO:0038095:Fc-epsilon receptor signaling pathway GO:0003007:heart morphogenesis GO:0007249:I-kappaB kinase/ NF-kappaB cascade GO:0001701:in-utero embryonic development GO:0045087: innate immune response GO:0007254:JNK cascade GO:0030324:lung development GO:0002755:MyD88-dependent toll-like receptor signaling pathway GO:0002756:MyD88-independent toll-like receptor signaling pathway GO:0035872:nucleotide-binding domain, leucine rich repeat containing receptor signaling pathway GO:0070423:nucleotide-binding oligomerization domain containing signaling pathway GO:0051092:positive regulation of NF-kappaB transcription factor activity GO:0051403:stress-activated MAPK cascade GO:0034166:toll-like receptor 10 signaling pathway GO:0034134:toll-like receptor 2 signaling pathway GO:0034138:toll-like receptor 3 signaling pathway

Table 5.6: Gene Signature Predictive of Atopy.

Name	Symbol	Fold Change	p-value	BH adjusted p-value
RAB4B, member RAS oncogene family	RAB4B	0.89	0.002	0.04
ITGA5 integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	ITGA5	0.87	0.003	0.04
ubiquinol-cytochrome c reductase core protein I	UQCRC1	0.85	0.004	0.04
isocitrate dehydrogenase 3 (NAD+) gamma	IDH3G	0.85	0.004	0.04
solute carrier family 22, member 18	SLC22A18	0.86	0.004	0.04
proline-serine-threonine phosphatase interacting protein 1	PSTPIP1	0.84	0.005	0.04
chromosome 19 open reading frame 24	C19orf24	0.87	0.006	0.04
flotillin 1	FLOT1	0.86	0.007	0.04
canopy 3 homolog (zebrafish)	CNPY3	0.87	0.007	0.04
mediator complex subunit 25	MED25	0.88	0.007	0.04
unc-45 homolog A (C. elegans)	UNC45A	0.85	0.008	0.04
DEAD (Asp-Glu-Ala-Asp) box polypeptide 41	DDX41	0.85	0.01	0.04
cytochrome b561 family, member A3	CYB561A3	0.87	0.01	0.04
microspherule protein 1	MCRS1	0.87	0.01	0.04
phenylalanyl-tRNA synthetase, alpha subunit	FARSA	0.88	0.01	0.04
interleukin 17 receptor A	IL17RA	0.85	0.01	0.04
TGF-beta activated kinase 1/ MAP3K7 binding protein 1	TAB1	0.87	0.01	0.04
nudix (nucleoside diphosphate linked moiety X)-type motif 16-like 1	NUDT16L1	0.87	0.02	0.04
cleavage and polyadenylation specific factor 3-like	CPSF3L	0.87	0.02	0.04
exocyst complex component 3	EXOC3	0.89	0.02	0.04
UPF1 regulator of nonsense transcripts homolog (yeast)	UPF1	0.86	0.02	0.04
spermidine synthase	SRM	0.86	0.02	0.04
ribosomal protein S6 kinase, 70kDa, polypeptide 2	RPS6KB2	0.88	0.02	0.04
nuclear prelamin A recognition factor	NARF	0.87	0.02	0.04
DENN/MADD domain containing 1C	DENND1C	0.88	0.02	0.04
HECT domain containing E3 ubiquitin protein ligase 3	HECTD3	0.89	0.02	0.04
nuclear protein localization 4 homolog (S. cerevisiae)	NPLOC4	0.86	0.02	0.04
adrenocortical dysplasia homolog (mouse)	ACD	0.89	0.02	0.04
mannosidase, alpha, class 2B, member 1	MAN2B1	0.84	0.02	0.04
ATPase type 13A1	ATP13A1	0.88	0.02	0.04
phospholipase C, beta 2	PLCB2	0.85	0.02	0.04
zinc finger and BTB domain containing 17	ZBTB17	0.89	0.02	0.044
solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1	SLC25A1	0.88	0.026	0.044
coiled-coil domain containing 124	CCDC124	0.89	0.027	0.045
coiled-coil domain containing 124	CCDC124	0.89	0.027	0.045
VPS9 domain containing 1	VPS9D1	0.89	0.027	0.045
zinc finger protein 3	ZNF3	0.88	0.03	0.047
zinc finger protein 692	ZNF692	0.88	0.033	0.048
zinc finger protein 692	ZNF692	0.89	0.033	0.048
BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	BAP1	0.89	0.035	0.049
phosphofructokinase, liver	PFKL	0.87	0.036	0.049
A kinase (PRKA) anchor protein 8-like	AKAP8L	0.88	0.037	0.049
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2	SMARCC2	0.86	0.037	0.049
ATP-binding cassette, sub-family F (GCN20), member 3	ABCF3	0.88	0.038	0.049

BH = Benjamini-Hochberg adjustment for multiple testing.

Table 5.7: Blue Module Genes Under-expressed by Atopic Clusters.

Name	Symbol	Fold Change	p-value	BH adjusted p-value
T cell receptor associated transmembrane adaptor 1	TRAT1	1.23	0.008	0.04
Mdm4 p53 binding protein homolog (mouse)	MDM4	1.15	0.02	0.04
MyoD family inhibitor domain containing	MDFIC	1.20	0.02	0.04
ribosomal protein S6 kinase, 90kDa, polypeptide 3	RPS6KA3	1.14	0.029	0.047
caspase 8, apoptosis-related cysteine peptidase	CASP8	1.19	0.03	0.047
myeloid cell leukemia sequence 1 (BCL2-related)	MCL1	1.18	0.038	0.049

BH = Benjamini-Hochberg adjustment for multiple testing.

Table 5.8: Blue Module Genes Over-expressed by Atopic Clusters.

typic clusters, which differed in terms of atopic burden, airway obstruction and rates of exacerbation. In this analysis, we found that an average of 12 years after the original study, participants continued to exhibit cluster-specific differences in several clinical characteristics, including differences in atopic features, spirometry and airway responsiveness. This is notable because at this time, in contrast to the physiologic differences, many of these patients described relatively mild symptoms. Furthermore, this finding, from the largest randomized, placebo-controlled clinical trial with extended follow-up for children with mild-moderate asthma, [29] suggests that the decreased symptoms many childhood asthmatics describe as they age does not correspond to disease remission. In fact, the original pathogenetic mechanisms appear to persist into young adulthood, albeit with minimal subjective symptoms.

Gene expression profiles from $CD4^+$ T cells collected from the same set of patients further confirm the persistence of physiologic differences between asthmatics assigned to different clusters. A set of 501 out of 22,184 total genes assayed displayed an expression pattern that correlated with atopic status, and a subset of these genes formed a signature that was highly predictive of atopy in an independent population of mild-moderate asthmatics, a finding that lends further credibility to the hypothesis that different phenotypic clusters correspond to differences in the underlying pathobiological mechanisms of asthma, as well as validating the biological relevance of our longitudinal phenotypic clusters.

Within our set of atopic genes, we were able to distinguish 3 highly-correlated gene co-expression modules, corresponding to differences in multiple gene ontologic process groups. Within one of these co-expression modules, we identified a transcription-factor binding site motif that was present in the promoter region of most of the genes within the co-expression module. The transcription factor motif was complementary to a binding site located in the protein encoded by the zinc-finger 3 gene (ZNF3), or a protein with a similar structure. Although the ZNF3 gene has not previously been linked to asthma, other zinc-finger proteins have been associated with modulation of Th2 cell differentiation [98] and secretion of the IL-17 cytokine, [109] both of which have documented associations to asthma pathogenesis. Furthermore, the family of zinc-finger proteins forms a large class of transcription factors with the potential to serve as therapeutic targets. [56,179] This finding has potential implications for the future of asthma drug development pipelines, that could focus efforts toward developing drugs to target transcription factors that modulate widespread changes in the gene expression levels of genes within a co-expression modules.

Our study had several limitations. One limitation was that while we identified a predictive biomarker signature for atopy, this particular population of patients was largely asymptomatic, reducing its potential for immediate clinical use. The availability of longitudinal gene expression data from prospective studies could help to elucidate temporal changes in the expression of these genes and their role in the pathogenesis of childhood asthma. An additional limitation is that although we were able to identify a transcription factor binding site motif in the promoter region of multiple genes within a co-expression module, we have no experimental evidence validating this finding. Further work is necessary to validate the existence of this motif within the co-expression module and its binding affinity for the ZNF3 protein.

In summary, our findings lend further support to the hypothesis that asthma phenotypic clusters are associated with differences in the underlying molecular mechanisms of asthma pathogenesis. This finding has implications for drug development and personalized approaches to the treatment of this complex disease. Further work will be necessary to validate these early findings and explore the mechanistic differences between different clusters.

6.0 GENETIC ASSOCIATIONS AND ENDOTYPES

6.1 QUANTITATIVE TRAITS

Variability in disease endotypes is hypothesized to be the result of multiple factors that are both genetic and environmental. To improve risk-stratification and refine treatment options, it is useful to identify genetic variants associated with disease endotypes. Defining variants associated with specific endotypes will allow us to screen patients at risk for specific disease manifestations. Disease endotypes may be described by variety of clinical traits that are both discrete (hospitalized vs. not-hospitalized) and continuous (body weight and blood pressure). Continuous traits often show a wide range of variation across a population that may be attributed to genetic factors. These traits are referred to as quantitative traits, and the causal genetic variants are referred to as quantitative trait loci (QTLs). In order to model the genetic contribution to quantitative traits, we often assume that the phenotypic value, y is a simple summation of genetic and environmental factors:

$$y = m + G + E \tag{6.1}$$

where m is a constant, G denotes the effect of all genes contributing to the phenotype, and E denotes the environmental effects [146]. Both E and G are assumed to have $mean = 0$, so that m represents the average phenotypic value. We may also assume that both E and G are uncorrelated, and that the variance of G is σ_G^2, and the variance of E is σ_E^2. The variance of y is the sum of σ_G^2 and σ_E^2. The heritability, h, is defined as the percentage of phenotypic variation with a genetic origin and is defined as the ratio of $\frac{\sigma_G^2}{\sigma_y^2}$.

If we consider a specific locus as bi-allelic, with alleles A and a, at any given locus, the genotype is defined as either AA, Aa or aa. If we let x_M denote the number of A alleles (0 or 1) inherited from the mother, and x_F denote the number of A alleles (0 or 1) inherited from the father, then we may state that $x = x_M + x_F$. Thus, the total number of A alleles in a particular genotype may be either 0, 1 or 2. We may thus model a quantitative trait as follows:

$$y = \mu + \alpha(x_M + x_F) + \delta|x_M + x_F| + e \tag{6.2}$$

where e is a random variable with $mean = 0$. The μ term is the mean value of y when an individual is an $aa - homozygote$. The mean level of y for an $AA - homozygote$ is $\mu + 2a$, and the mean level of y for an $Aa - heterozygote$ is $\mu + \alpha + \delta$. If $\delta = 0$, the locus is considered to be *additive*, which means that each A allele adds to the average value of the phenotype, y. If $\delta = \alpha$, the locus is considered to be *dominant*, and a single A allele produces the full genetic effect on the phenotype, y. Conversely, if $\delta = -\alpha$, the locus is considered to be *recessive*, and a single A allele produces no effect on the phenotype, y, while two A alleles are necessary to produce a full effect.

The number of A alleles in an individual is $x = x_M + x_F$, with x_M and x_F being Bernoulli random variables. If we assume that each allele is independent with the same probability p of an A allele, then the distribution of x is binomial, such that:

$$Pr(x = 2) = p^2, Pr(x = 1) = 2p(1 - p), Pr(x = 0) = (1 - p)^2 \tag{6.3}$$

where p is the frequency of A in the population. Thus, the mean value of x is $2p^2 + 2p(1 - p) + 0 = 2p$, and the mean value of the phenotype, y, is $m = \mu + 2p\alpha + 2p(1 - p)\delta$. The variable e is assumed to incorporate all other factors contributing to trait variability. We may rewrite the value of the phenotype as:

$$y = m + \{\alpha + (1 - 2p)\delta\}x[(x_M - p) + (x_F - p)] - \{2\delta\}x[(x_M - p)(x_F - p)] + e \tag{6.4}$$

where $[(x_M - p) + (x_F - p)]$ and $[(x_M - p)(x_F - p)]$ are uncorrelated because x_M and x_F are independent. Thus, the variance of y is:

$$\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2 \tag{6.5}$$

where

$$\sigma_y^2 = sp(1-p)[\alpha + (1-2p)\delta]^2, \sigma_D^2 = 4p^2(1-p)^2\delta^2, \sigma_e^2 = var(e) \tag{6.6}$$

where σ_A^2 is considered to be the additive variance, σ_D^2 is the dominance variance and the locus specific heritability, h^2 is $\frac{\sigma_A^2+\sigma_D^2}{\sigma_y^2}$.

6.2 MAPPING QUANTITATIVE TRAIT LOCI

6.2.1 Early Methods

In practice, we consider the contribution of multiple potential loci to a particular phenotype. To model the contribution of a genotype, g to a complex phenotype, y, we may consider a small number of QTLs with genotypes $g_1, ..., g_p$, such that there are 2^p distinct genotypes. In this case, the mean value for the phenotype, y is:

$$E(y|g) = \mu_{g_1,...,g_p} \tag{6.7}$$

and the variance for y is:

$$var(y|g) = \sigma^2_{g_1,...,g_p} \tag{6.8}$$

If we assume additivity:

$$\mu_{g_1,...,g_p} = \mu + \sum_{j=1}^{p} \Delta_j g_j \tag{6.9}$$

where ($g_j = 1$ or 0) We also assume a constant variance, $\sigma_g^2 \equiv \sigma^2$ and normally distributed residual variation, $y|g \sim N(\mu_g, \sigma^2)$. Under this model, the simplest method of detecting genetic associations to complex traits is through marker regression. That is, we split individuals

into groups according to the genotype at each marker. We next perform a hypothesis test, either a t-test or an ANOVA, depending on the number of genotypes, to determine whether there are differences in the mean value of the phenotype, y between individuals in different genotype groups. This process may be repeated for each locus in the association analysis. This approach is relatively simple, and easily incorporates additional covariates. However, when using this approach, it is necessary to exclude individuals with missing data. In addition, this method does not take into account the location of a particular locus in relation to other loci and only considers one QTL at a time.

6.2.2 Interval Methods

Interval methods improve upon earlier methods for identifying QTL associations. The major innovation of interval methods was to move beyond the evaluation of single loci in isolation and incorporates the spatial relationships between genetic loci into association mapping. Interval methods take missing genotype data into account and also incorporate the marker location and relationships by making use of a genetic map of typed loci and interpolating between them [100]. For this model, we assume a single causal QTL. For each position in the genome, one locus at a time is posited as the putative QTL, and we assume the phenotype $y \sim N(\mu_z, \sigma)$. If the QTL genotype is BB/AB, we let $x = 1/0$. Given genotypes at linked markers, we assume y is distributed as a mixture of normal distributions, with mixing proportion $Pr(z = 1|\text{marker data})$, such that:

		QTL genotype	
M_1	M_2	BB	AB
BB	BB	$(1 - r_L)(1 - r_R)/(1 - r)$	$r_L r_R/(1 - r)$
BB	AB	$(1 - r_L)/r_r/r$	$r_L(1 - r_R)/r$
AB	BB	$r_L(1 - r_R)/r$	$(1 - r_L)r_R/r$
AB	AB	$r_L r_R/(1 - r)$	$(1 - r_L)(1 - r_R)/(1 - r)$

M_1 and M_2 represent two different markers, and r_L and r_R represent two different alleles.

If we let $p_i = Pr(z_i = 1|\text{marker data})$ and assume $y_i|z_i \sim N(\mu_{zi}, \sigma^2)$, we may conclude that:

$$Pr(y_i|\text{marker data}, \mu_0, \mu_1, \sigma) = p_i f(y_i; \mu_1, \sigma) + (1 - p_i) f(y_i; \mu + 0, \sigma) \tag{6.10}$$

where $f(y; \mu, \sigma) = exp[-(y - \mu)^2\sigma^2]/\sqrt{2\pi\sigma^2}$ and the log likelihood is:

$$l(m_0, \mu_1, \sigma) = \sum_i \log Pr(y_i|\text{marker data}, \mu_0, \mu_1, \sigma) \tag{6.11}$$

The maximum likelihood estimates (MLEs) of μ_0, mu_1, σ are values for which the log likelihood function is maximized. The maximum values for μ_0, mu_1, σ may be determined by utilizing the expectation-maximum algorithm [34].

The strength of evidence for the presence of a QTL at a particular location is determined by a LOD score, where:

$$\text{LOD}(\gamma) = log_{10} \text{ likelihood ratio of QTL at position } \gamma \text{ to no QTL} \tag{6.12}$$

$$= log_{10}\{\frac{Pr(y|\text{QTL at}\gamma, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma})}{Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})}\} \tag{6.13}$$

Assuming a single QTL at position γ, $\hat{\mu}_{0\gamma}$, $\hat{\mu}_{1\gamma}$, $\hat{\sigma}_\gamma$ are the MLEs.

The advantages to interval mapping are that this method takes into account missing data, allows examination of positions between markers, and gives improved estimates of QTL effects. Disadvantages include increased computation time and that QTLs are still considered one at a time.

6.3 GENOME WIDE ASSOCIATION STUDIES (GWAS)

The development of cost-effective methods for genotyping larger and larger numbers of single nucleotide polymorphism (SNP) markers together with the development of dense maps of genetic loci within the human genome have made possible large-scale genetic association studies involving comprehensive genome-wide surveys, referred to as genome-wide association studies (GWAS). GWAS further exploit the spatial relationships between genetic loci that are spaced in close proximity along the chromosome. Such loci are said to be in linkage disequilibrium (LD), with a higher probability of being inherited together.

Several high-profile genome-wide association studies (GWAS) has been widely used to detect novel genetic associations in complex diseases [124, 41, 115]. The GWAS approach

is particularly advantageous because it provides greater power to find disease-associated genetic variants than conventional linkage studies [64], and also provides a way to identify previously unsuspected potentially causal genetic loci, compared to earlier candidate gene studies [113]. Along with the development of GWAS technology, new statistical methods have arisen to analyze GWAS output.

6.3.1 Multiple Linear Regression for GWAS

In quantitative genetics, a multiple linear regression model is often used to describe the relationship between phenotypes and genetic markers. One advantage of this model is that all marker effects may be estimated simultaneously and then used to perform hypothesis testing to identify the QTL signal. A typical regression model is:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \tag{6.14}$$

where y_i $(i = 1, ..., n)$ is the phenotypic value of the i^{th} individual in the mapping population, β_0 is the intercept, x_{ij} is the genotypic value of the j^{th} marker for individual i, β_j is the effect of marker j, and e_i is the random error assumed to follow a normal distribution $N(0, \sigma_e^2)$ with mean zero and variance σ_e^2 independently for $i = 1, ..., n$. The genotype, x_{ij}, is defined as:

$$x_{ij} = \begin{cases} 1 & \text{if genotype is AA} \\ 0 & \text{if genotype is AB} \\ -1 & \text{if genotype is BB} \end{cases} \tag{6.15}$$

For QTL mapping with a dense set of markers, as in GWAS, we are interested in estimating the effects of markers, $\beta = \{\beta_1, ..., \beta_p\}$ and identifying markers in linkage disequilibrium (LD) with QTLs.

6.3.2 Limitations of Multiple Linear Regression for GWAS

Ordinary least squares (OLS) is a statistical method used to estimate regression coefficients. In the case of QTL mapping, there are several problems with this method. Although OLS gives an unbiased estimate of regression coefficients, there is often a great deal of variation, which leads to inaccuracy in predicted values. In addition, OLS is not available in situations where the number of exploratory variables is larger than the number of observations, the $p > n$ problem. In GWAS, it is common to have a dense set of SNP markers that is much larger than the number of individuals ascertained for analysis. In this case, most of the SNP markers assayed have minimal effects on the phenotypic trait studied. This results in a loss of power, when all SNP markers are considered in a multiple linear regression QTL mapping analysis. In practice, multiple linear regression is seldom used for all but the simplest genetic association studies. Instead, univariate linear regression is used repeatedly on each marker across the genome to avoid the $p > n$ problem and analytical intractability inherent in estimating the for such a large regression analysis. However, splitting the multivariate regression problem into a series of univariate regression analyses results in a large number of statistical tests necessitating adjustment for multiple testing. For this reason, it is often useful to obtain a sparse model to increase the analytical power of the QTL mapping and avoid the computational pitfalls involved in multivariate regression to solve for a large number of .

6.3.3 Sparse Regression for GWAS

6.3.3.1 Ridge Regression Ridge regression is a form of penalized multiple linear regression [72]. Ridge regression adopts the l_2 norm penalty function, $\lambda \sum_{j=1}^{p} \beta_j^2$. The regression coefficients, $\beta = \{\beta_1, ..., \beta_p\}$, are estimated by minimizing the penalized sum of squares $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$. The penalty function, $\lambda \sum_{j=1}^{p} \beta_j^2$, Ridge regression is a specialized case of regularized regression that puts constraints on the size of coefficients to control large variances associated with resulting estimates. It works by "shrinking" the effect of redundant variables, such as redundant genetic markers, by imposing a penalty on the size of their coefficients. In addition, ridge regression prevents any one regression coef-

ficient from getting very large and thus protects against overfitting and the high variance that often result from multiple linear regression with highly correlated variables.

6.3.3.2 LASSO regression Lasso regression was developed as a method of variable selection, and is particularly useful in addressing the $p > n$ problem. In lasso regression, the regression coefficients, $\beta = \{\beta_1, ..., \beta_p\}$, are estimated by minimizing the penalized sum of squares $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$. The sum of absolute values, the so-called l_1 norm of the regression coefficients, $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penalty function, and $\lambda \geq 0$ is the shrinkage factor, specified at the onset. When the penalty function is added to the residual sum of squares and $\lambda > 0$, lasso is able to shrink the least squares estimators toward zero and thereby decrease their variance. This approach is useful for GWAS, where only a small number of SNP markers are related to the phenotype of interest. However, lasso has several drawbacks compared to ridge regression. First, when the explanatory variables are collinear, lasso will select only a single variable at random from a group of highly correlated variables. This is a problem as GWAS are performed with densely spaced SNP markers, that are often highly correlated and in linkage disequilibrium. Second, when $p > n$, the largest number of explanatory variables that may be selected from the model is n, which may lead to the loss of important variables and interacting loci in genetic association studies.

6.4 INTRODUCTION TO TEMPORALLY-SMOOTHED LASSO (TESL)

6.4.1 A Local Autoregressive Model for Dynamic Traits

While standard lasso addresses the problem of selecting disease-associated SNPs from a large number of SNPs, it does not account for the correlation across d-traits at many time points. To address this critical gap, we first use a local autoregressive model to characterize the temporal correlation across the T time points in the d-trait. We then incorporate this information within our algorithm, termed temporally-smoothed lasso (TESL) [97]. That is, we incorporate the parameters of the local autoregressive model into a structured penalized

regression method in order to find the regression coefficients.

The assumption of the local autoregressive model is that each d-trait is modeled locally, as a function of the trait value at the previous time point. Assuming a linear model for local temporal dependency, and given measurements of the d-trait at two adjacent time points t and $t+1$, $1 \leq t < T$, we have:

$$\mathbf{y}_{t+1} = \alpha_t \mathbf{y}_t + \alpha_t^0 \mathbf{1} + \epsilon_t, \tag{6.16}$$

where the α_t represents the slope for the multiplicative dependency of d-trait value at time $t+1$ on the value at the previous time point, α_t^0 corresponds to the intercept, $\mathbf{1}$ is a vector of 1's of length N, and ϵ_t is an N-vector of the noise terms with mean 0 and constant variance.

The parameters α_t in the above model describes the local linear dependency, and the set $\{\alpha_t$'s for $t = 1, \ldots, T-1\}$ represents the overall shape in the d-trait. We note that although a linear relationship is assumed locally for adjacent time points, the set of models described by α_t's does not assume any particular functional form globally for the d-traits, and can capture a wide variety of non-stationary dynamic trends, including locally-linear, logistic, and cyclic trends. This flexibility is an improvement upon methods that assume parametric functions for the shape of the trajectory [181].

After centering the d-trait data $\mathbf{y}_t$ by subtracting the mean $\bar{\mathbf{y}}_t$ to obtain $\mathbf{y}_t^c = \mathbf{y}_t - \bar{\mathbf{y}}_t$, the parameters α_t and α_t^0 in Eq. (6.16) can be estimated from the mean-centered data, using the standard least square method as follows:

$$\hat{\alpha}_t = \text{argmin } (\mathbf{y}_{t+1}^c - \alpha_t \mathbf{y}_t^c)^\top \cdot (\mathbf{y}_{t+1}^c - \alpha_t \mathbf{y}_t^c).$$

We are only concerned with the slope parameters α_t's since they represent the temporal dependency of the magnitudes of a d-trait between two time points. The estimate for α_t is given as:

$$\hat{\alpha}_t = \frac{(\mathbf{y}_t^c)^\top \cdot \mathbf{y}_{t+1}^c}{(\mathbf{y}_t^c)^\top \cdot \mathbf{y}_t^c}. \tag{6.17}$$

The estimates of α_t's describe the local temporal dependencies for each d-trait. The α_t's are then incorporated into the penalty in TESL. Recall that the α_t's encode the strength of dependence of a d-trait between two adjacent time points.

6.4.2 Formulation and Parameter Estimation for the Temporally-Smoothed Lasso

In TESL, we use the α_t's to design a penalty for regularized regression that enforces a temporal structure in the estimated association strengths β_t's. The additional penalty is added to the lasso objective function as shown below:

$$\hat{\boldsymbol{B}} = \text{argmin} \sum_t (\mathbf{y}_t - \boldsymbol{X}\beta_t)^\top (\mathbf{y}_t - \boldsymbol{X}\beta_t) + \lambda \sum_t \parallel \beta_t \parallel_1 + \gamma T(\boldsymbol{B}), \tag{6.18}$$

where $T(\cdot)$ denotes the autoregressive fusion penalty:

$$T(\boldsymbol{B}) = \sum_{t=1}^{T-1} \parallel \beta_{t+1} - \hat{\alpha}_t \beta_t \parallel_1 . \tag{6.19}$$

The regularization parameter λ controls the sparsity of the estimated β_t. A larger value of λ yields a solution that is more sparse while a smaller value of λ yields a solution that has fewer non-zero elements. The regularization parameter γ on the autoregressive fusion penalty controls the difference in association strengths between β_{t+1} and $\hat{\alpha}_t \beta_t$. A large value of γ encourages this difference to be zero. This implies that the non-stationary linear dependencies between the d-traits at two adjacent time points t and $t+1$, as described by the local autoregressive model, are reflected in the temporal dependencies between the association strengths β_{t+1} and β_t through the parameter $\hat{\alpha}_t$. For instance, if there is a large increase between the d-trait values at time t and $t+1$, then α_t will be a large positive value, and there will be a noticeable difference in the association strengths β_{t+1} and β_t. Both regularization parameters λ and γ can to be selected via cross-validation.

We require a fast optimization procedure for this problem since a genome-wide scan contains a huge number of SNPs. While this problem is convex, the autoregressive fusion penalty poses a challenge to attaining efficient optimization because it is not smooth. Thus, we use the smoothing proximal gradient method [23] whereby the strategy is to reformulate the autoregressive fusion penalty via the dual norm to decouple the non-separable terms. Then, a smooth approximation to the reformulated penalty is derived. The result is an

objective function with a smooth square loss term, the smooth approximation to the autoregressive fusion penalty and the non-smooth lasso penalty. Since the reformulated objective consists of simply a smooth component and the non-smooth lasso penalty, we can optimize this efficiently with the fast iterative shrinkage-thresholding algorithm (FISTA) [11].

6.4.3 Selection of Regularization Parameters

To select the best model, we need to determine the optimal values for the two regularization parameters: λ for the lasso penalty and γ for the autoregressive fusion penalty. The standard procedure for selecting these regularization parameters is to do cross-validation. This has also been described in the setting of penalized regression for genome-wide association mapping [183]. We split the data into a training set (90%) and a test set (10%) and do cross-validation to select the (λ, γ) regularization parameter set with the lowest test error. Once we obtain the β_t's from the best model, the selected SNPs were those with a non-zero regression coefficient (association strength). We then reestimate the regression coefficients using only the selected variables with standard least squares, which removes the bias imposed by the penalties.

6.4.4 Simulation Study

We demonstrate our method on simulated datasets, and compare its performance with those from the standard univariate association test and lasso that do not take into account temporal structures in the d-traits.

6.4.4.1 Experimental Setup of Simulations We simulated datasets as follows. In order to generate the genotype data, we first selected a segment of 50 SNPs randomly from chromosome 7 of HapMap CEU panel [161] after removing the SNPs with MAF less than 0.10. In addition to the 60 individuals in the HapMap CEU panel, we generated additional 90 individuals by randomly mating the original 60 individuals to obtain genotypes for 150 individuals in total. Assuming that observations for a d-trait are obtained over 10 time points, we generated the true association strengths by randomly selecting three SNPs and

setting the association strengths for these SNPs to non-zero values that vary linearly or cyclically over time. We set the association strengths for all of the other SNPs to zero. For the linear d-trait, the association strength β_1^j of each association SNP at the first time point was randomly drawn from a uniform distribution [0.05, 0.1], and the association strengths of the same SNP at subsequent time points were set to $\beta_1^j \times t$, where $t = 2, \ldots, 10$. For the cyclical dynamic trait, we assume that the association strength of each SNP changes according to a sinusoid functional form with 1.5 cycles and with the peak amplitude randomly sampled from a uniform distribution [0.80, 0.85]. Given these genotypes and true association strengths, we generated the d-trait values using a linear model with noise distributed as $N(0, 1.0)$. In our simulation study, we assume that all of the temporal correlation in the d-trait is induced by the temporally changing genetic effects.

6.4.4.2 Illustrative Examples of Dynamic-Trait Associations In order to illustrate the genetic effects on a d-trait, the simulated d-trait data under the scenario of linear growth are shown in Figures 6.1A-D. While Figure 6.1A shows the d-trait trajectories for all of the 150 individuals in a single simulated dataset, those same individuals are grouped into three subsets according to the genotypes of the true association SNPs in Figures 6.1B-D. The thick blue curves in Figures 6.1B-D indicate the mean trajectories within the group. Figure 6.1B shows individuals with no mutations in any of the three association loci. Since these individuals do not possess the genetic variants that can drive a temporal change in the d-trait, the trajectories do not show any trend over time. As we introduce one or two mutations in the genetic loci, a trend of linear growth starts emerging as shown in Figure 6.1C. Figure 6.1D shows those individuals with more than three mutations, and this pattern of linear growth becomes more apparent.

Given the d-trait data in Figure 6.1A simulated from the genotype data and the true association strengths in Figure 6.1E, we perform a d-trait association mapping using a univariate association test and lasso applied to each time point separately as well as our proposed method, and show the results in Figures 6.1F-H, respectively. When the true association signals are very weak at the early time points, all of the methods are unsuccessful in detecting the signal. However, the results from univariate analysis and lasso contain many

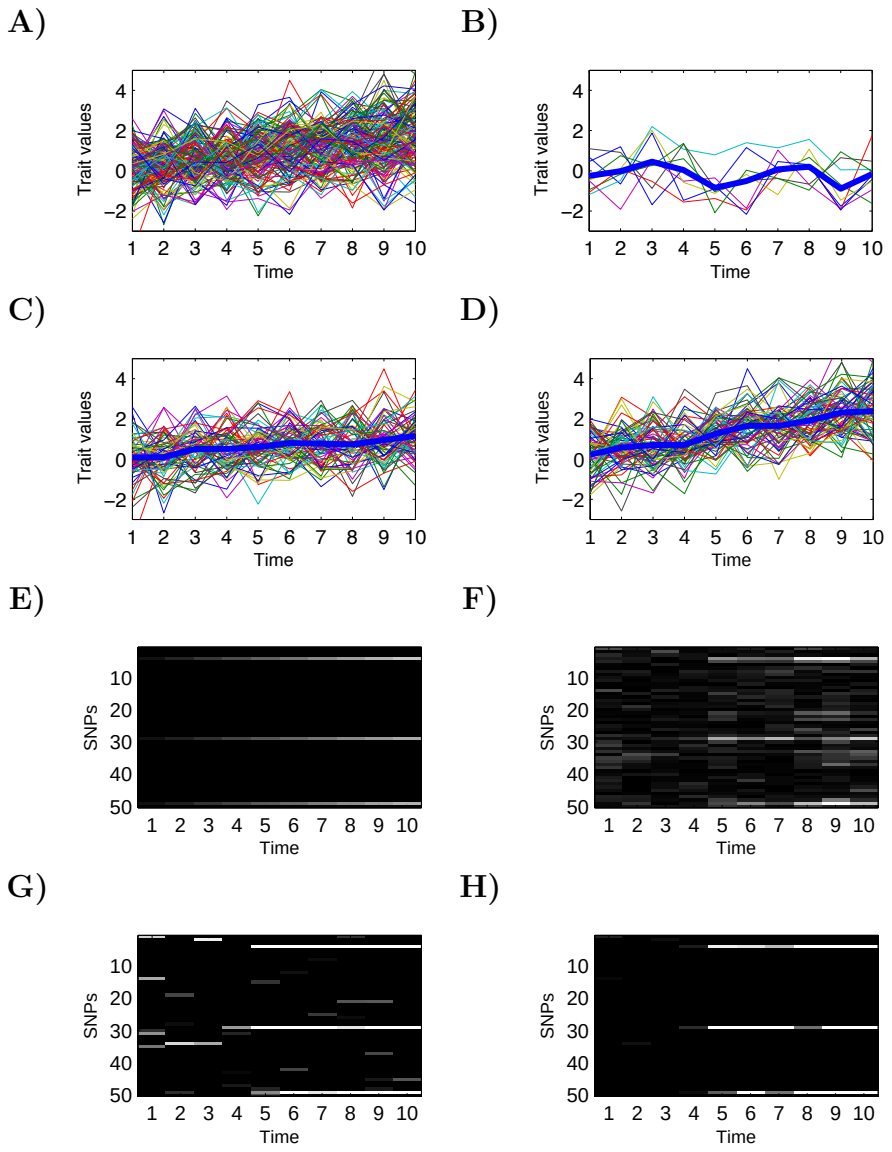
false positives at these early time points, whereas our method has significantly fewer false positives. As the true association signals increase over time, our method starts detecting it at earlier time points for the SNP on the third row than other methods, because it can take into account the temporal correlation and infer from the observations at the adjacent time points that the weak signal is indeed a true association.

Similarly, an illustration of the simulation scenario with a cyclical dynamic is shown in Figure 6.2. The d-trait values have a cyclical trajectory over time in Figures 6.2C and D, because of the periodic change in the genetic effects of the association variants as shown in Figure 6.2E. When the individuals have no mutations at any of the three association SNPs, the d-trait measurements do not show a cyclic trend, as can be seen in Figure 6.2B. Overall, the association results show that our method in Figure 6.2H finds fewer false positives than the univariate test in Figure 6.2F and lasso in Figure 6.2G. When the association signal is weak between the peak and valley of the cycle, our method is able to borrow information from the adjacent time points and infer the presence of associations, whereas the other methods simply miss the signals.

6.4.4.3 Accuracies for Detecting True Associations We perform a systematic and quantitative comparison of the performance of different association analysis methods in terms of type I errors and powers for detecting true associations by generating 50 simulated datasets as described above and averaging the results over these datasets. Assuming that association strengths are indicated by -log(p-values) for univariate tests and the absolute values of the estimated regression coefficients for lasso and our proposed method, we sort the estimated association strengths and compare the top 5% values of the sorted list with the list of known true associations to compute type I errors and powers. The average type I errors and powers over time are shown in Figures 6.3A and B for the scenario of linear dynamic. Across all time points, our proposed method has significantly lower type I errors and higher powers than other methods that do not take advantage of the temporal correlation structure in the d-traits. This gap in the performance is significantly greater at early time points when the association signals are weak. Thus, our method can be potentially used to detect the genetic effects of the associated variant on a d-trait at an early stage for a diagnostic purpose.

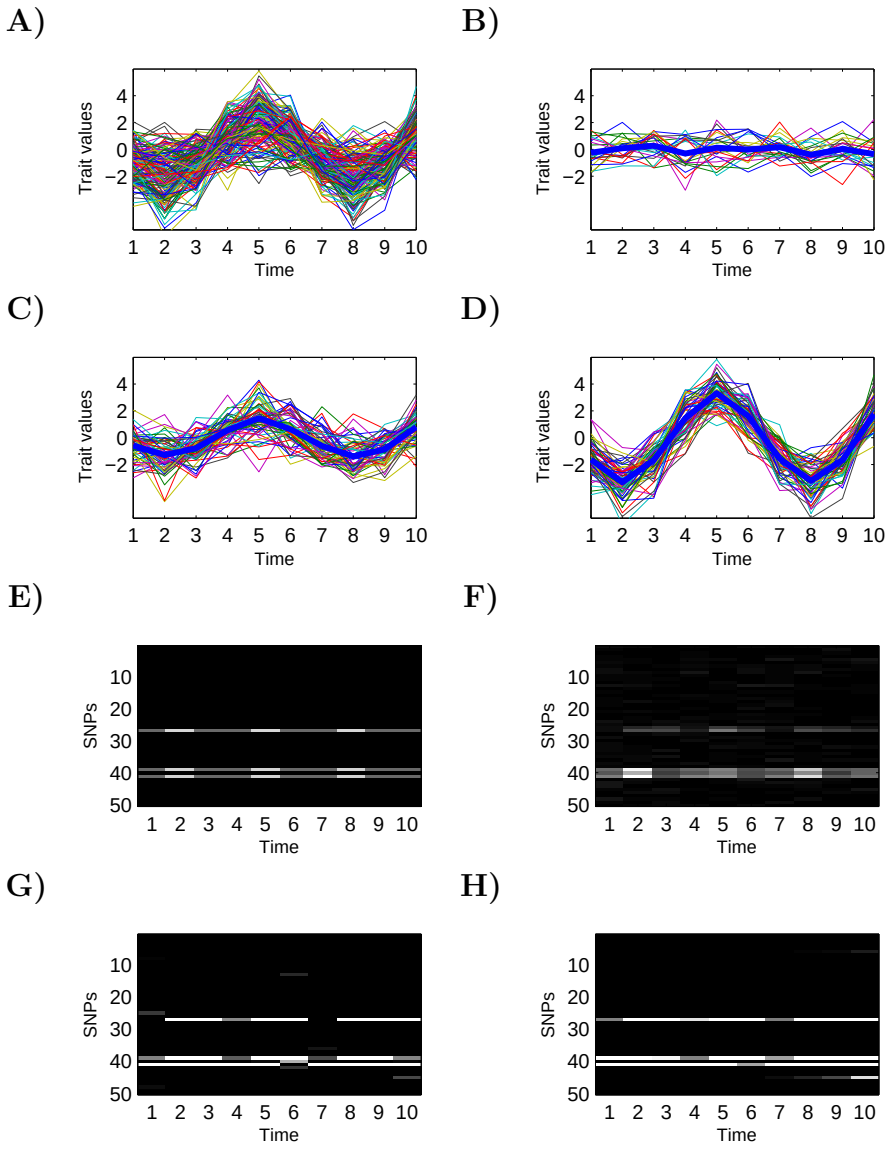
For the scenario of a cyclic dynamic, we show the type I errors and powers averaged over 50 simulated datasets in Figure 6.4. Again, our method outperforms other methods across all time points. Especially when the association signals are weak between the peak and valley of each cycle, the performance gap is significantly greater. This is because our method can learn the presence of associations between peaks and valleys by exploiting the temporal correlation in the d-trait.

We vary the number of association SNPs to 3, 7, and 10, and compare the performance of different association methods in Figure 6.5, under the scenario of linear dynamic. Type I errors are shown in Figures 6.5A-C for the number of association SNPs 3, 7, and 10, respectively, and powers are shown in Figures 6.5D-F. The results are averaged over 50 simulated datasets, and top 10% with the highest estimated association strengths are used. Our method has significantly lower type I errors and powers than any other methods across all time points. Although the performance of all methods tends to increase over time as the association strengths of the association SNPs increase, our method can still detect the associations significantly better than other methods at the early time points.



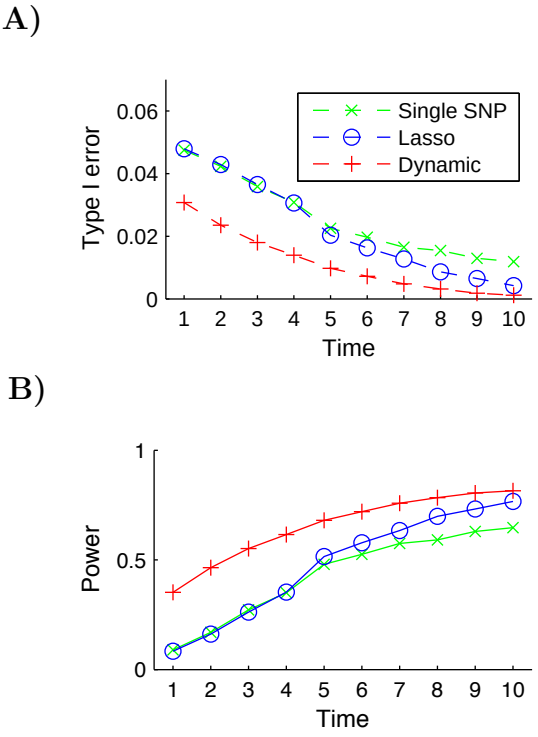
Using a simulated dataset for a d-trait with a linear dynamic. **A:** Trait measurements over 10 time points for 150 individuals. **B:** Trait measurements for the individuals with no mutations in all of the three causal genetic loci. These are unaffected individuals with no specific trend in the trajectories over time. The thick blue curve indicates the mean trajectory among these individuals. **C:** Trait measurements for the individuals with one or two mutations on causal genetic loci. The trend of a linear increase starts to emerge as these mutations drive the temporal change in the d-trait. **D:** Trait measurements for the individuals with more than three mutations on causal genetic loci. As more causal mutations are introduced, the linear trend is stronger with steeper slope than in Panel C. **E:** True association strengths used to generate this simulation dataset. **F:** -log(p-values) from single-SNP association tests. **G:** The absolute values of the estimated association strengths from lasso. **H:** The absolute values of the estimated association strengths from d-trait lasso.

Figure 6.1: Illustration of d-trait association mapping.



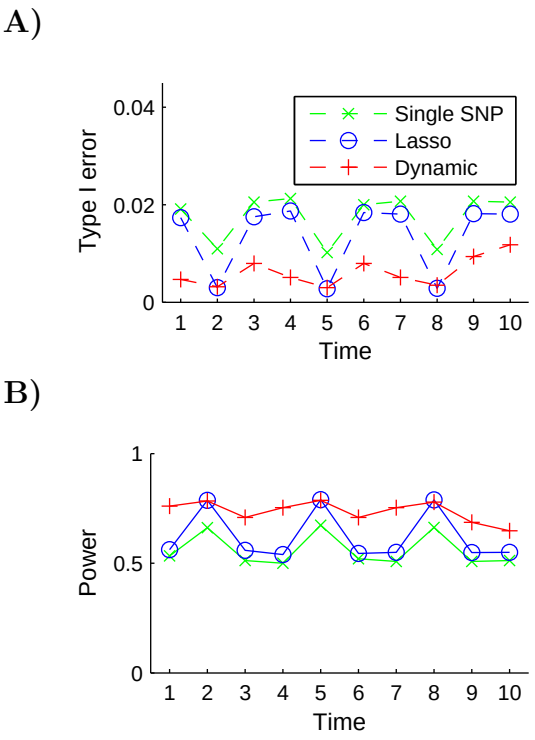
Using a simulated dataset for a d-trait with a cyclic dynamic. **A:** Trait measurements over 10 time points for 150 individuals. **B:** Trait measurements for the individuals with no mutations in all of the three causal genetic loci. These are unaffected individuals with no specific trend in the trajectories over time. The thick blue curve indicates the mean trajectory among these individuals. **C:** Trait measurements for the individuals with one or two mutations on causal genetic loci. The cyclic trajectory starts to emerge as these mutations drive the temporal change in the d-trait. **D:** Trait measurements for the individuals with more than three mutations on causal genetic loci. As more causal mutations are introduced, the cyclic trend is stronger with higher peaks and lower valleys than in Panel C. **E:** True association strengths used to generate this simulation dataset. **F:** -log(p-values) from single-SNP association tests. **G:** The absolute values of the estimated association strengths from lasso. **H:** The absolute values of the estimated association strengths from d-trait lasso.

Figure 6.2: Illustration of d-trait association mapping.



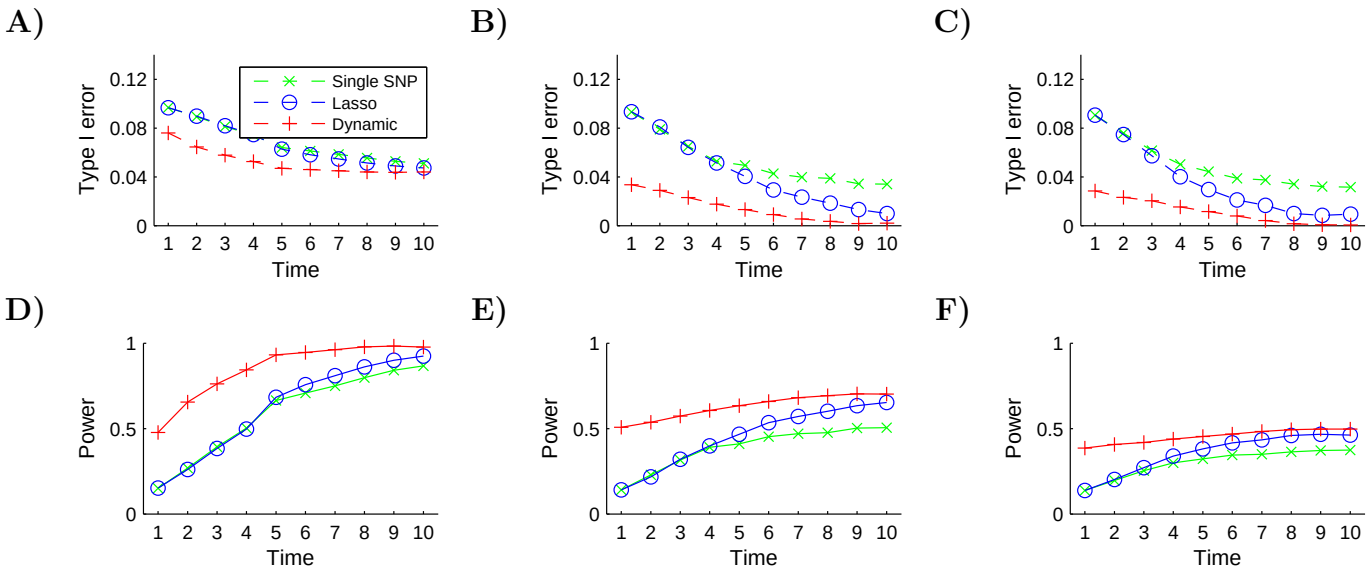
Data were simulated assuming a linear dynamic in d-traits. **A:** Type I errors over the 10 time points. **B:** Powers. Results were averaged over 50 simulated datasets.

Figure 6.3: Comparisons of different methods for a d-trait association analysis using simulated datasets.



Data were simulated assuming a cyclic dynamic in d-traits. **A:** Type I errors over the 10 time points. **B:** Powers. Results were averaged over 50 simulated datasets.

Figure 6.4: Comparisons of different methods for a d-trait association analysis using simulated datasets.

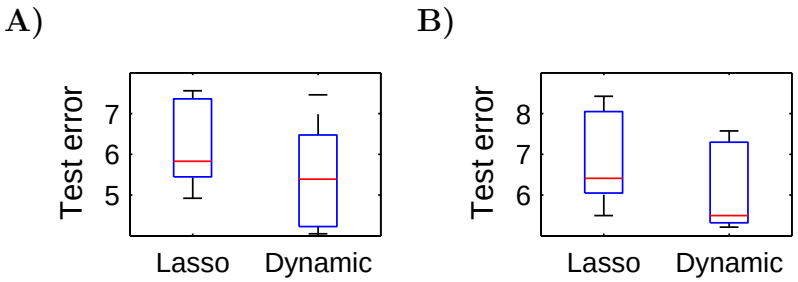


Data were simulated assuming a linear dynamic in d-traits. Type I errors over the 10 time points are shown for **A:** S=3, **B:** S=7, **C:** S=10. Powers over the 10 time points are shown for **D:** S=3, **E:** S=7, **F:** S=10. Results were averaged over 50 simulated datasets.

Figure 6.5: Comparisons of different methods for a d-trait association analysis using simulated datasets when the number of causal loci S varies.

6.4.4.4 Prediction Accuracy Finally, we compare the performance of the different methods in terms of the prediction errors in Figure 6.6. We generate test data for additional 100 individuals, and compute the squared differences between the predicted values based on the estimated regression coefficients and the observed values. The prediction error is defined as an average of these squared differences over 100 individuals in the test set. We repeat this process for 50 simulated datasets, and show the prediction errors averaged over these datasets in Figures 6.6A and B for the two scenarios of dynamic growth and cyclic trend, respectively. We find that our method has significantly lower prediction errors than lasso that does not combine information across time.

6.4.4.5 Non-dynamic Genetic Effects in Dynamic Trait Association While the simulation studies above considered the d-trait association scenarios with dynamic genetic effects, where the size of genetic effects on d-traits changes over time, we now consider



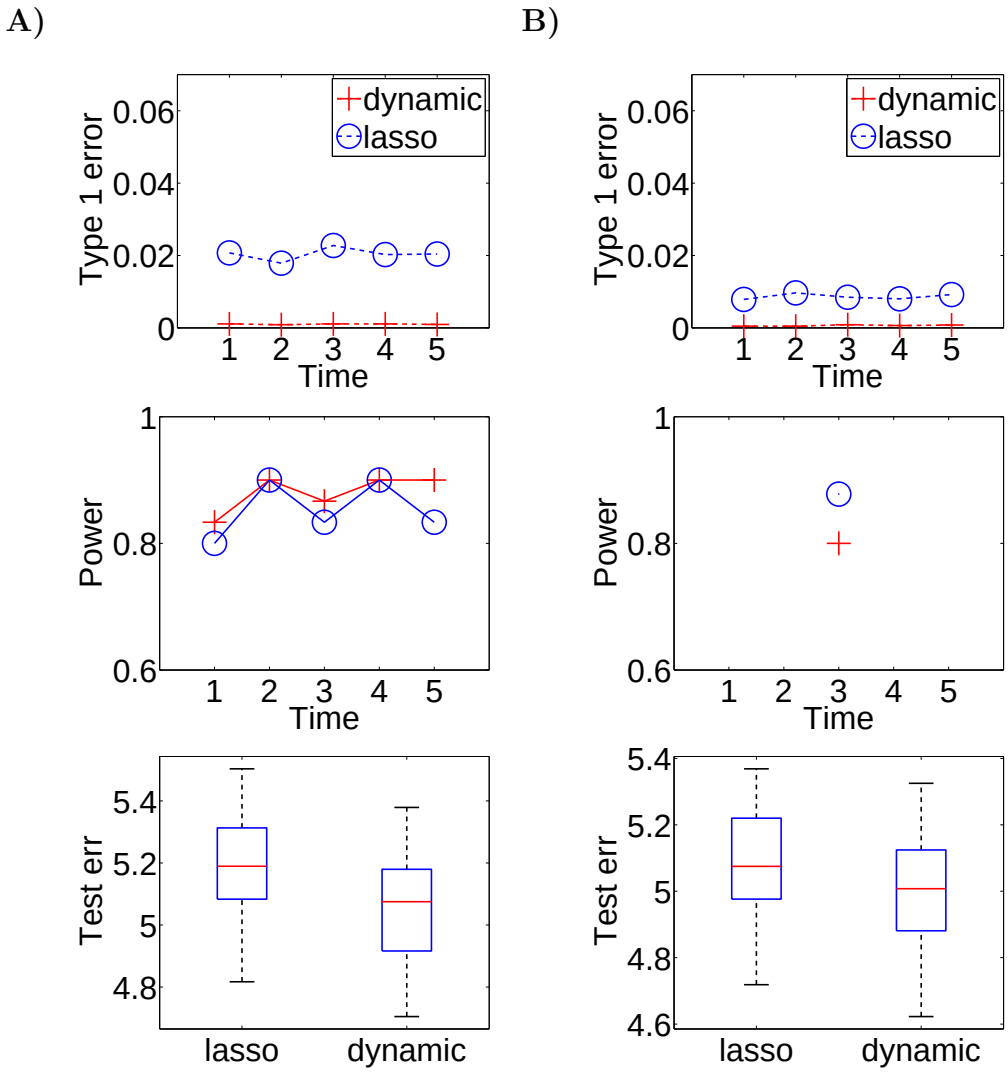
A: A linear dynamic in d-traits. **B:** A cyclic dynamic in d-traits. Results were averaged over 50 simulated datasets.

Figure 6.6: Test errors using simulated datasets.

association scenarios with non-dynamic effects of SNPs on d-traits such as constant genetic effect sizes over time or genetic effects only at a single time point. In order to simulate data, we obtained 721 SNPs on chromosome 22 for 721 individuals from HapMap 3 data [160] after filtering out SNPs with MAF less that 0.05 and SNPs with pairwise correlation greater than 0.88 between neighboring SNPs. Then, assuming three association SNPs and the number of time points $T = 5$, we set the true values for association strengths as follows. For the case of constant genetic effects over time, the strength of each of the three association SNPs was sampled randomly from the uniform distribution [0.1, 0.5]. For the case of a genetic effect at a single time point, the association SNPs were assumed to influence the d-trait only at $t = 3$, and the association strengths were randomly sampled from a uniform distribution [0.1, 0.8]. All of the results were obtained by choosing 360 samples randomly as a training set, performing 10-fold cross-validation to select the regularization parameters, and computing the prediction errors on the remaining 361 samples.

In Figure 6.7, we compare the performance of our method with the standard lasso in terms of the type I errors, powers, and prediction errors averaged over 50 simulated datasets. The left and right columns in Figure 6.7 show results from each of the two simulation scenarios, constant and single-time-point genetic effects, respectively. We find that our new method gives lower type I errors than lasso in both of the scenarios (the top row in Figure 6.7). This shows the benefit of the smoothing effect of the autoregressive fusion penalty that sets the association strengths of SNPs with no associations to zeros across time points. For the

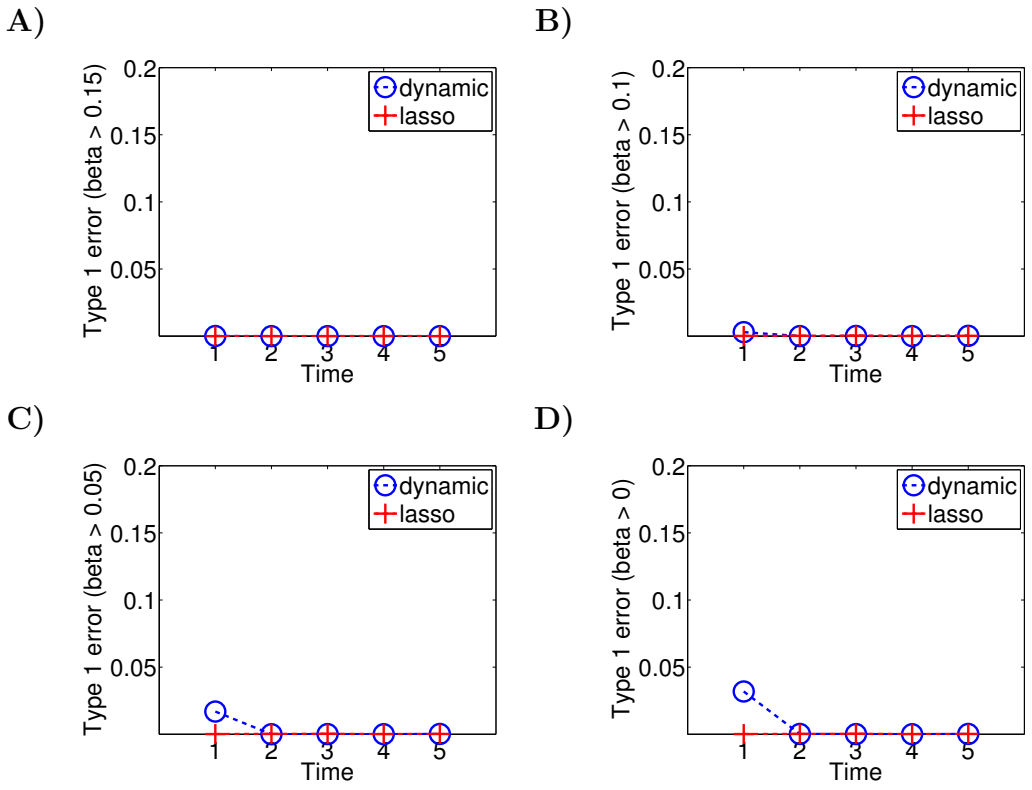
scenario of the constant genetic effects, this smoothing effect of our penalty also slightly increases the power, whereas it slightly decreases the power for the case of genetic effects at a single-time point (the middle row in Figure 6.7). For both of the scenarios, the TESL method gave lower prediction errors than lasso, showing that the benefit of the smoothing effect of our method generally outweighs their potential disadvantages.



A: Genetic effects of association SNPs were assumed to be constant across all time points. **B:** Association SNPs were assumed to influence the d-trait only at a single time point (t=3). Type I errors, powers, and prediction errors are shown in the top, middle, and bottom rows, respectively.

Figure 6.7: Results on simulated datasets under scenarios for non-dynamic genetic effects.

6.4.4.6 Null Distribution To assess the type I error of our method, we have performed null simulations and computed the type I error for both TESL and standard lasso. For this experiment, we used the 721 individuals with 721 SNPs on chromosome 22 from HapMap 3 data [160] with MAF > 0.05 previously described in Section 6.4.4.5. To simulate the null distribution, we set the association strengths for all SNPs at all time points to zero so that there are no true associations. We then generated the d-trait values using a linear model with noise distributed as $N(0, 1.0)$. We compare the type I error between our method and standard lasso. After thresholding the association strengths at different levels 0.15, 0.1, 0.05, and 0, the type I error is shown in Figures 6.8A-D, respectively. These results were averaged over 50 simulated data sets. As can be observed from Figure 6.8D, the type I error < 0.05. In addition, the magnitude of the association strengths β_t of the false positives are very small. At $|\beta_t| > 0.15$, as shown in Figure 6.8A, the type I error is zero at all time points.



A: regression β=0.15 **B:** regression β=0.1 **C:** regression β=0.05 **D:** regression β=0.

Figure 6.8: Type I error for different regression coefficient thresholds.

6.4.4.7 Computation Time In order to assess the feasibility of running TESL on a large dataset, we measure the computation time for running our method and the standard lasso on datasets of varying sizes and show the results in Figure 6.9. In Figure 6.9A, we vary the number of SNPs while fixing the number of time points at 10. In Figure 6.9B, we vary the number of time points, fixing the number of SNPs at 1000. The results show that our method can efficiently handle datasets of thousands of SNPs and up to a hundred time points in a few minutes. For example, a dataset with 8000 SNPs measured over 10 time points could be handled in less than 7 minutes. The project page along with available code for TESL may be accessed at: http://cogito-b.ml.cmu.edu/dynamictraits/.

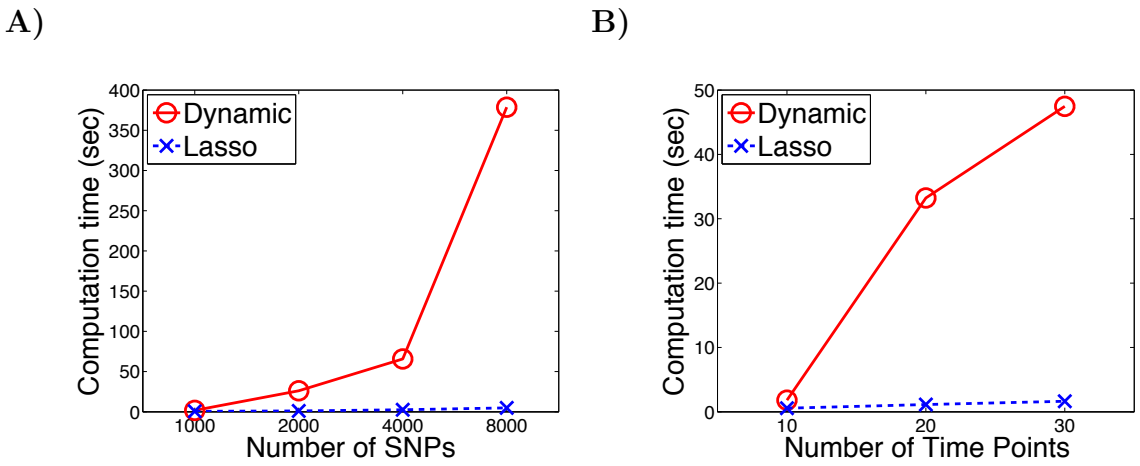


Figure 6.9

7.0 GENETIC ASSOCIATIONS AND DYNAMIC ASTHMA ENDOTYPES

7.1 INTRODUCTION

Current GWAS are limited in that they only consider associations between genotypes and static phenotypes — snapshots of traits at a single time point — when more information is often available. For example, clinical trials typically record longitudinal traits such as BMI [31], triglycerides [192] and blood glucose levels [129]. We refer to a longitudinal trait that changes value over time as a *dynamic trait* or *d-trait*. A particular genotype can cause a trait not only to vary across individual, but also to change over time for the same individual; and the potential for such information, if available, to identify disease-related genes can be significant. If d-trait data is accessible, it is desirable to use this information to identify genetic loci that drive the temporal evolution of the trait. Such a method of association analysis can utilize the complete trajectory for a particular clinical trait which increases the power of GWAS. To provide some insight into the intuition behind this problem, consider the situation where two individuals with different genotypes have the same measurement of a trait at a specific time point (Fig. 7.1A). In this situation, GWAS would not detect any contributing genetic variations. However, these individuals may have completely different trajectories for this trait (Fig. 7.1B). Thus, using dynamic information, it is more probable that these variations will be detected.

The concept of finding associations between regions of the genome and longitudinal traits has been previously investigated. Techniques such as repeated measures analysis [135, 28, 32, 129], averaging of traits across time points [79] and identifying principal components [101] to capture the temporal information of complex traits have been used with a modest degree of success. However, such techniques may result in loss of information, particularly

the information contained in the trajectory of a particular trait. To address this problem, more sophisticated methods have been proposed, such as non-parametric approaches based on Legendre orthogonal polynomials [31] and adaptive splines [192], to model longitudinal data. This is a promising direction, but these methods are different from our proposed method in that they do not explicitly impose the sparsity of associations between genotype and phenotype, which is crucial in elevating the stringency of detection in high-dimensional (large number of candidate predictors) and high-noise cases. In the literature of quantitative trait locus (QTL) mapping for linkage analysis, the problem of finding a linkage between a genetic locus and a d-trait has been broadly termed as functional mapping [111,182,172] and the standard mixture model for QTL mapping for a static trait has been extended to address this problem. However, the mixture model for functional mapping was designed for linkage analysis, and does not extend to the general problem of association mapping of unrelated individuals with a relatively dense set of genetic markers.

7.2 APPLICATION OF TESL TO A COHORT OF ASTHMATIC CHILDREN

7.2.1 Description of Study Subjects and Dynamic Trait

The current study utilized GWAS data from 466 non-Hispanic white children who participated in the Childhood Asthma Management Program (CAMP). Characteristics of the study subjects are presented in Table 7.1. CAMP was a multi-center randomized, double-masked clinical trial of the long-term effects of three inhaled treatments for mild to moderate asthma, with 1,041 subjects enrolled and followed for a period of 48 months. Inclusion criteria and protocols for collection of baseline phenotypic data have been described in detail elsewhere [82, 81].

As part of the study, several measures of pulmonary function were evaluated at regular intervals. We chose one of these measures, the forced expiratory volume measured after the administration of an inhaled bronchodilator medication (post-bronchodilator FEV_1) to use

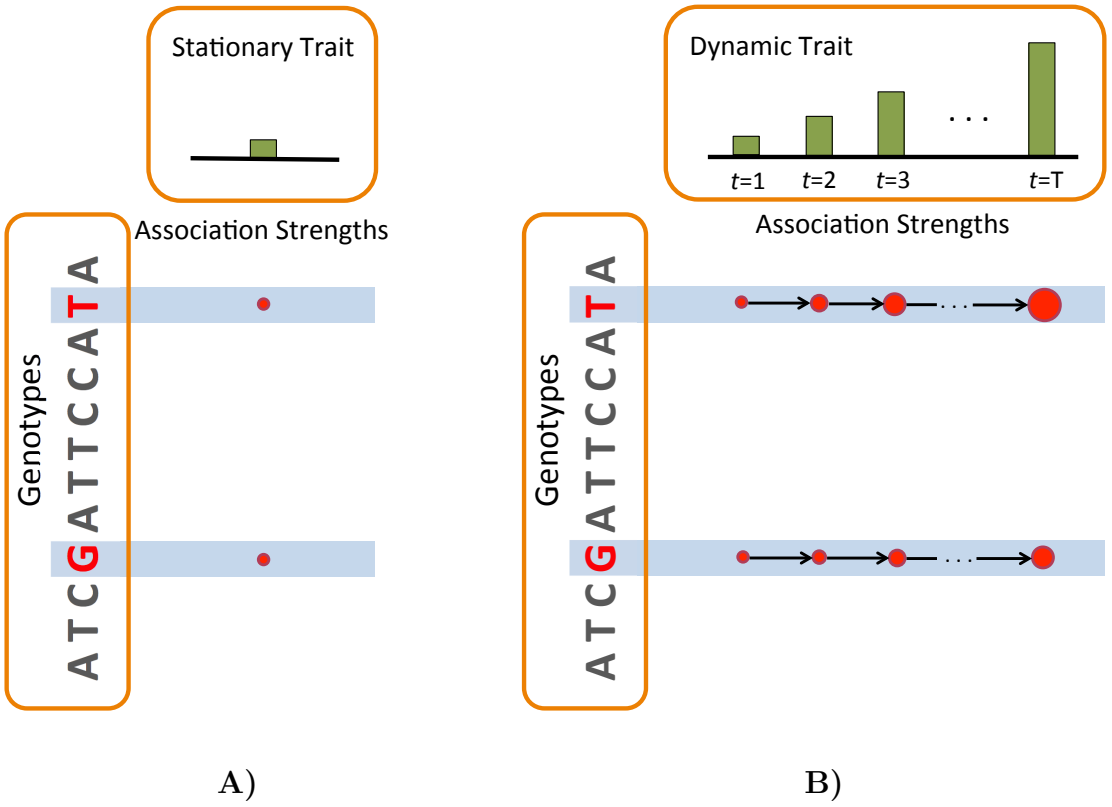


Figure 7.1: Illustration of association analysis.

study month	FEV_1 Mean (SD) [Range]
0	103.0 (12.8) [62, 154]
2	104.9 (13.1) [67, 145]
4	104.3 (12.8) [46, 151]
12	104.4 (12.6) [50, 151]
16	103.0 (12.6) [46, 136]
24	103.3 (12.4) [66, 137]
28	103.0 (12.2) [62, 141]
36	103.1 (13.0) [48, 142]
40	103.2 (12.2) [66, 143]
48	102.7 (12.9) [56, 142]

FEV_1 = One-second forced expiratory volume. SD = Standard deviation.

Table 7.1: Characteristics of FEV_1 D-trait Over Time.

for association analysis. The post-bronchodilator FEV_1 is a sensitive measure of the degree of airway obstruction present in asthmatics after the administration of a corrective medication. For this reason it is thought to be representative of an asthmatic child's intrinsic level of lung function that is not responsive to medical therapy [137]. The units of measurement for the FEV_1 are typically in liters. However, as the expected value for the FEV_1 can vary with height, the FEV_1 is typically reported as a percentage of the value predicted for a certain age and height, which is what we used for the current study. In the original CAMP study, study subjects were evaluated at regular intervals and measures of pulmonary function, including the FEV_1, were obtained. Over the 48-month period of the study, each subject's FEV_1 was measured a total of 10 times, approximately once every four months. The exact study months are listed in Table 7.1. We adjusted the FEV_1 trait for baseline imbalances in age, sex, height, height squared and the clinical location from which the subjects were recruited. The residuals obtained after making these adjustments were used as inputs to TESL.

7.2.2 Preprocessing of Genetic Data

Of the original CAMP participants, 968 provided DNA and 299 provided RNA for genetic studies as part of the CAMP Genetics Ancillary Study. Of the 968 participants who pro-

vided DNA, 466 were non-Hispanic white children. The Institutional Review Boards of the Brigham and Women's Hospital (Boston, MA) and of the other CAMP study centers approved this study. Informed assent and consent were obtained from the study participants and their parents to collect material for genetic studies.

Details of genotyping and quality control have been described previously [70]. Genome-wide genotyping for CAMP subjects was performed on the HumanHap550 Genotyping BeadChip or Infinium HD Human610-Quad BeadChip by Illumina, Inc (San Diego, CA) at the Channing Laboratory. Data from those subjects genotyped using Illumina technologies was combined into a primary dataset of SNPs having missingness $< 1\%$, passing Hardy-Weinberg Equilibrium (HWE) (p-value threshold of $1\text{x}10^{-3}$), and having minor allele frequency (MAF)>0.05.

To expand the association results, imputation of all SNPs available in the June 2010 release of the 1000 Genome Project (1000GP) data using MaCH [176] was performed for the genotype data. A set of 6,216,972 imputed SNPs had a MAF>0.05 and a ratio of empirically observed dosage variance to the expected (binomial) dosage variance greater than 0.5, indicating good quality of imputation.

We performed pruning of highly correlated SNP markers. PLINK was used to prune SNPs based upon the degree of linkage disequilibrium, with the goal of generating a pruned subset of SNPs in approximate linkage equilibrium to each other. The multiple correlation coefficient for each SNP being regressed simultaneously on all other SNPs within a sliding window of 150 SNPs was calculated, and SNPs were pruned if the calculated multiple correlation coefficient was below the variance inflation factor (VIF) threshold (1.85). The total number of SNPs remaining after this step was 306,025. A genomic inflation factor (GIF) was obtained for each time point of the FEV_1 from unadjusted Chi-square values for allelic association. The GIF was 1.000, indicating minimal population stratification. To improve the computational time required for the genome-wide analysis, we divided each chromosome into sections of approximately 3500 SNPS and then ran our method on each chromosome section. All genetic and phenotypic data are available in the database of Genotypes and Phenotypes (dbGaP) accession phs000166.v2.p1.

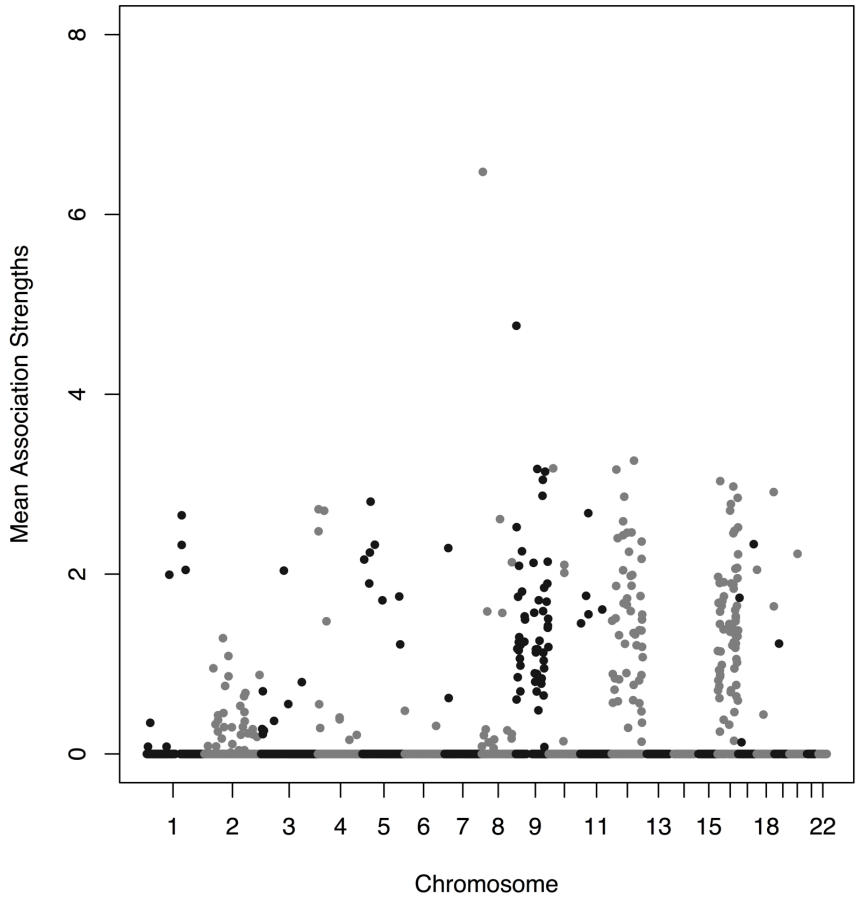
7.3 ANALYSIS OF THE CAMP COHORT WITH TESL

We applied the TESL to a GWAS of SNPs and FEV_1 values from non-Hispanic white participants in the CAMP study. To select the regularization parameters for TESL, we chose a range of λ and γ parameters and examined the cross-validation test error. We calculated the cross-validation test error for each pair (λ, γ) of the regularization parameters, and selected the values resulting in the minimum cross-validation test error. Running TESL on each section for a single regularization parameter set (λ, γ) took approximately 1.3 minutes. Thus, cross-validating over a possible 200 parameter sets yielded a runtime of approximately 5 hours per section.

For the optimal regularization parameter set (λ, γ), TESL identified 271 non-zero associations (SNPs with non-zero regression coefficients for at least one time point) from 306,025 SNPs tested. Although the 271 associations were spread across the genome, several chromosomes contained multiple non-zero associations, including chromosomes 2, 8, 9, 12 and 16. To prioritize the non-zero associations, we calculated the mean of the absolute values of the association strengths (regression coefficients) across all time points to generate a single association strength value for each SNP. We ranked these values in order of magnitude and used the values to prioritize the non-zero associations. A Manhattan plot of mean association strengths over time for TESL is shown in Figure 7.2. Because we used a sparse regression method, the majority of the association strengths in the Manhattan plot were zero, while 271 associations were non-zero. The top 10 associations as ranked by association strength are displayed in Table 7.2.

7.3.1 Functional Analysis of Temporally-Smoothed Lasso Associations

In order to evaluate the functional significance of the non-zero associations, we performed a gene ontology (GO) enrichment analysis on all genes neighboring the associated genomic regions. The 271 non-zero associations were within 100 kb of 554 genes. We performed a hypergeometric test to determine whether the set of genes neighboring significant associations was enriched for specific biological processes (BP) compared to that expected for a randomly



The black and gray striations denote different chromosomes.

Figure 7.2: Manhattan plot of mean association strengths for each chromosome.

selected set of genes. Genes annotated to GO terms received weights based on the scores of neighboring GO terms, and significance scores of connected components were compared to detect the most significant local terms within the GO hierarchy [1]. We found that the genes neighboring non-zero associations were functionally enriched in multiple biological processes of known relevance to asthma, including multiple pathways related to innate immunity and inflammation (Table 7.3). The observation that our non-zero associations were in close proximity to genes with functional relevance to asthma lends further support to the validity of our method and its ability to detect asthma-relevant associations.

SNP	CHR	BP	MAF	Mean $\|(\beta^j)\|$
chr8:4233416	8	4233416	0.068	6.47
rs576967	9	1701524	0.061	4.76
chr12:94963945	12	94963945	0.075	3.26
rs10828784	10	18828279	0.29	3.18
chr9:91083694	9	91083694	0.058	3.17
chr12:19567609	12	19567609	0.070	3.16
chr9:124655498	9	124655498	0.10	3.14
rs7023886	9	114767890	0.13	3.05
rs2229320	16	10909244	0.14	3.03
chr16:66927184	16	66927184	0.10	2.97

Table 7.2: Top 10 significant SNPs from in CAMP dataset identified by TESL.

We next performed a targeted enrichment analysis on the genes neighboring the (*cis*-acting) non-zero associations that demonstrated differential expression between different alleles. We used a linear model to detect differences in gene expression levels between subjects with different alleles, and select the genes with the highest level of differential expression (p-value < 0.10) for enrichment analysis. Of 22,184 genes assayed on the microarray chip, 554 were within 100 kb of non-zero SNP associations and 43 of these were differentially expressed between subjects with different alleles, with 37 having available GO annotations that could be used for analysis.

We evaluated these 37 genes for functional enrichment in the BP ontology as described above, and the top GO annotations detected by this method are presented in Table 7.4. We detected several significant GO terms with established relevance to asthma pathogenesis, such as the response to glucocorticoid stimulus (GO:0051384), which was underrepresented in our set of genes. Glucocorticoid medications are one of the mainstays of asthma therapy and numerous clinical trials highlight the benefits of these medications in children with asthma [24]. The genes associated with this GO term were G protein-coupled receptor 83 (GPR83) on chromosome 11, and transcription factor AP-4 (TFAP4) on chromosome 16. Both of these genes were located in *cis* to SNPs with non-zero associations and demonstrated differential expression between subjects with different alleles. Another GO term relevant to asthma pathogenesis was positive regulation of TGF beta production (GO:0071636), which

was also underrepresented for our set of genes. The gene associated with this GO term was glucosidase, beta (bile acid) 2 (GBA2), located on chromosome 9 and also differentially expressed between subjects with different alleles. TGF beta has been liked to airway remodeling and the pathogenesis of asthma [45], so the presence of this GO term among the significant annotations suggests a functional link between this genomic region and allelic differences in asthma pathogenesis.

GO ID	**Term**	**Annotated**	**Significant**	**Expected**	***p*-value**
GO:0034122	negative regulation of toll-like receptor signaling	13	4	0.24	0.00014
GO:0050777	negative regulation of immune response	53	7	0.96	0.00025
GO:0002062	chondrocyte differentiation	69	7	1.26	0.00079
GO:0032695	negative regulation of interleukin-12 production	11	3	0.2	0.00153
GO:0009581	detection of external stimulus	101	8	1.84	0.00266
GO:0006953	acute-phase response	45	5	0.82	0.00299
GO:0006925	inflammatory cell apoptotic process	14	3	0.25	0.00322
GO:0034260	negative regulation of GTPase activity	14	3	0.25	0.00322
GO:0007062	sister chromatid cohesion	34	4	0.62	0.00636
GO:0030199	collagen fibril organization	36	4	0.66	0.00781
GO:0032088	negative regulation of NF-kappaB transcription	57	5	1.04	0.00826
GO:0006833	water transport	38	4	0.69	0.00946
GO:0071385	cellular response to glucocorticoid stimulus	22	3	0.4	0.01195
GO:1900449	regulation of glutamate receptor signaling	22	3	0.4	0.01195
GO:0021983	pituitary gland development	43	4	0.78	0.01452
GO:0048333	mesodermal cell differentiation	24	3	0.44	0.01521
GO:0048147	negative regulation of fibroblast proliferation	24	3	0.44	0.01521
GO:0019319	hexose biosynthetic process	67	5	1.22	0.01596
GO:0032873	negative regulation of stress-activated MAPK cascade	25	3	0.46	0.017
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	26	3	0.47	0.01891

The 'Significant' column refers to the number of significant genes associated with a particular GO term. The 'Expected' column refers to the percentage of significant genes expected to be associated with a GO term assuming a random distribution.

Table 7.3: Statistics for Significant GO Terms for Genes Neighboring Non-zero Associations.

7.3.2 Description and Functional Significance of Top SNP Association

The SNP with the largest association strength was located at chr8:4233416 (Table 7.2). The FEV_1 trajectory for this SNP is shown in Figure **??**. For this SNP, subjects with different alleles had very different trajectories over the 48-month duration of the study. Subjects possessing two copies of the minor allele had relatively high values of the FEV_1 (mean FEV_1

GO ID	Term	Annotated	Significant	Expected	p-value
GO:2000278	regulation of DNA biosynthetic process	15	2	0.03	0.00056
GO:0050679	positive regulation of epithelial cell proliferation	114	3	0.24	0.00247
GO:0071695	anatomical structure maturation	40	2	0.09	0.00403
GO:0002040	sprouting angiogenesis	48	2	0.10	0.00577
GO:0006987	activation of signaling protein activity...	63	2	0.13	0.00976
GO:0030198	extracellular matrix organization	202	3	0.43	0.01206
GO:0051329	interphase of mitotic cell cycle	397	4	0.85	0.01415
GO:0045766	positive regulation of angiogenesis	89	2	0.19	0.01883
GO:0072203	cell proliferation involved in metanephros development	10	1	0.02	0.02345
GO:0071157	negative regulation of cell cycle arrest	10	1	0.02	0.02345
GO:0060041	retina development in camera-type eye	103	2	0.22	0.02474
GO:0008652	cellular amino acid biosynthetic process	104	2	0.22	0.02519
GO:0046479	glycosphingolipid catabolic process	11	1	0.02	0.02576
GO:0045601	regulation of endothelial cell differentiation	11	1	0.02	0.02576
GO:0071636	positive regulation of TGF-beta production	11	1	0.02	0.02576
GO:0090136	epithelial cell-cell adhesion	11	1	0.02	0.02576
GO:0043923	positive regulation by host of viral transcription	11	1	0.02	0.02576
GO:0051384	response to glucocorticoid stimulus	108	2	0.23	0.02701
GO:0015669	gas transport	12	1	0.03	0.02807
GO:0044319	wound healing, spreading of cells	12	1	0.03	0.02807

The 'Significant' column refers to the number of significant genes associated with a particular GO term. The 'Expected' column refers to the percentage of significant genes expected to be associated with a GO term assuming a random distribution.

Table 7.4: Statistics for Significant GO Terms for Differentially Expressed Genes.

= 140%) at the beginning compared to subjects with two copies of the major allele (mean FEV_1 = 103%) and heterozygotes (mean FEV_1 = 100%). However, for subjects with two copies of the minor allele, the FEV_1 declined linearly during the study to levels similar to those of subjects possessing two copies of the major allele (FEV_1 = 115%, minor allele, FEV_1 = 103%, major allele) and heterozygotes (mean FEV_1 = 101%). Conversely, the subjects with two copies of the major allele and the heterozygotes had relatively stable values of FEV_1 over the study period, that were consistently lower than the FEV_1 values for subjects with two copies of the minor allele.

The SNP chr8:4233416 is located on the short arm of chromosome 8, and upstream from the gene for defensin β-1 (DEFB1). The DEFB1 protein is expressed in airway epithelial tissue and has been found to play a role in host defense against respiratory pathogens [52]. Prior studies have linked polymorphisms in the gene for DEFB1 with an increased risk of

asthma [108, 107]. Although we did not observe differential expression of DEFB1 between subjects with different alleles, it is plausible that variation in the upstream region leads to regulatory changes with an effect on DEFB1, such as enhanced degradation. Further work will be necessary to validate these early findings in an independent population and to characterize the functional significance of this variation.

This SNP and the associated d-trait trajectory serve to illustrate the advantage of TESL. If we were to perform a cross-sectional analysis exploring associations between this SNP and the FEV_1 at discrete points in time, the potential to detect a significant association would be reduced. However, because we were able to leverage the dynamic change in the trajectory across time, we had increased power to detect this association. Furthermore, not only did the value of the FEV_1 vary between subjects with different allelic variants, but the trajectory of the trait also varied among different subjects. Heterozygotes and subjects with two copies of the major allele had a relatively stable FEV_1 trajectory, while subjects with two copies of the minor allele had an FEV_1 trajectory that declined linearly over the duration of the study period, suggesting that the variant exerts an effect on asthma pathophysiology that occurs over an extended period of time.

7.3.3 Non-Zero Associations Correspond to Differences in Gene Expression

In order to investigate the functional significance of the associations identified by TESL, we evaluated for the presence of differential gene expression levels between different allele types, and associations with the FEV_1 trajectory. Of the 554 genes we investigated, we found 6 genes with significant (adjusted p-value < 0.05) differences in expression after adjustment for multiple testing (Table 7.5). Two of these genes were of particular interest due to their putative role in asthma pathogenesis, and we have provided a detailed review of our findings below.

One gene of interest was DENN/MADD domain containing 5B (DENND5B), located within 100 kb of SNP marker rs7313158 on chromosome 12. For this gene, the SNP marker rs7313158 was located approximately 100 kb downstream from the coding region of the DENND5B gene. We found that subjects possessing two copies of the major allele had

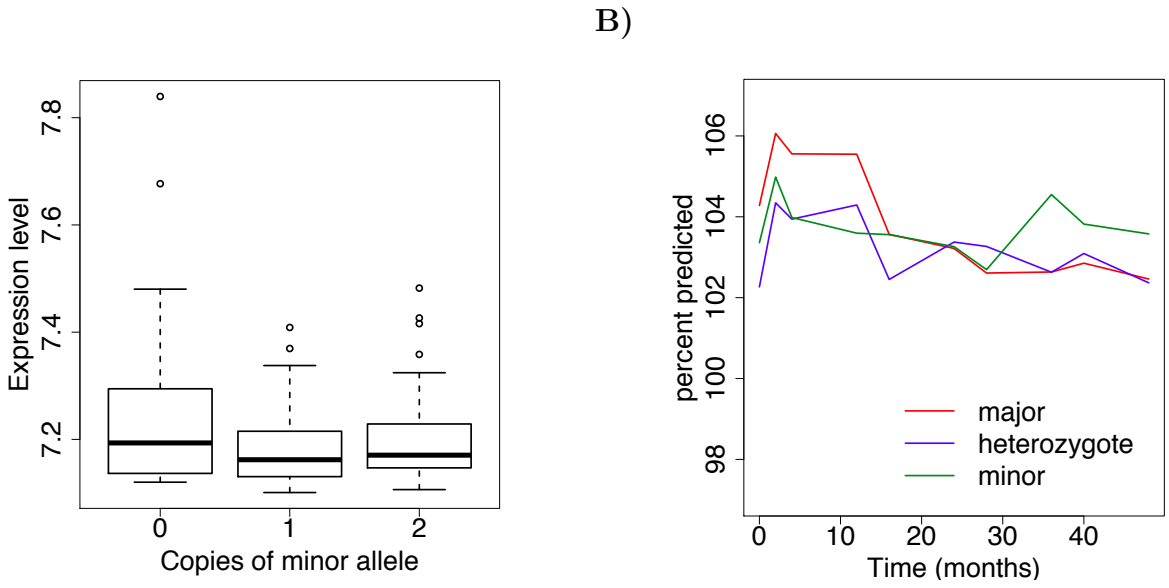
SNP	Gene	Chr	Minor Allele Av. Expression	Heterozygote Av. Expression	Major Allele Av. Expression	p-value	Adj. p-value
rs1752380	SELENBP1	1	7.24	7.23	7.32	0.0029	0.032
rs1752380	TUF1	1	7.87	7.84	7.77	0.016	0.089
rs2728436	PEX2	8	9.78	9.64	9.77	0.0027	0.0053
rs4741755	KIAA0020	9	9.42	9.39	9.46	0.012	0.037
rs7313158	DENNDB5	12	7.20	7.18	7.24	0.004	0.008
rs4026608	IRAK3	12	7.85	7.85	8.16	0.0075	0.03

Table 7.5: List of Genes With Differential Allelic Expression Patterns.

higher levels of expression of this gene compared to subjects with two copies of the minor allele and heterozygotes (Benjamini-Hochberg (BH) adjusted p-value = 0.008) as shown in Figure 7.4). Subjects with different alleles also had differences in their FEV_1 trajectory over time. Those with two copies of the major allele had relatively higher FEV_1 levels early in the course of the study (mean FEV_1 = 104%, major allele, FEV_1 = 103%, minor allele, FEV_1 = 102%, heterozygote), but FEV_1 levels for the major allele decreased over time (mean FEV_1 = 102% at 48 months) (Fig. 7.4). Conversely, the FEV_1 of subjects with two copies of the minor allele and heterozygotes did not demonstrate this decline in FEV_1 over time.

The rs7313158 SNP locus was associated with the DENND5B gene, which is located on the reverse strand of chromosome 12. The eQTL locus is downstream from DENND5B, and suggests the presence of a functional regulatory locus within this region. There are no existing associations between the DENND5B gene and asthma or atopy, although variants in the DENND1B gene have been associated with childhood asthma [147]. Like DENND1B, DENND5B contains a GTPase binding domain, however it's function in normal physiology is not completely characterized [114].

Another gene that demonstrated differential expression for different alleles was the interleukin-1 receptor-associated kinase 3 (IRAK3) gene, located within 100 kb of SNP marker rs4026608. For the IRAK3 gene, subjects with the major allele had relatively higher levels of expression of this gene compared to subjects with two copies of the minor allele and heterozygotes (BH adjusted p-value = 0.03) (Fig. 7.5). Subjects with different alleles also had corresponding differences in their FEV_1 trajectory. Those with two copies of the major allele had con-

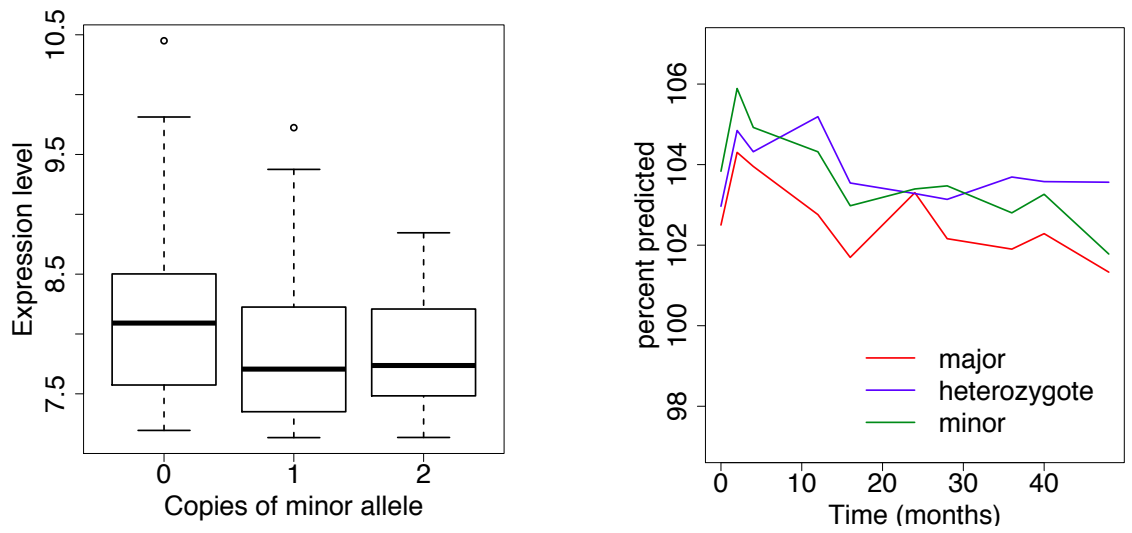


A: Gene expression levels of DENND5B for subjects with different alleles for SNP rs7313158. Subjects with two copies of the major allele (0) have relatively higher levels of DENND5B expression than subjects with two copies of the minor allele (2) and heterozygotes (1), Benjamini-Hochberg (BH) adjusted p = 0.008. **B:** Post-bronchodilator FEV_1 trajectory for subjects with different alleles for SNP marker rs7313158. At the beginning of the time course, the FEV_1 for subjects with the major allele is relatively higher than the FEV_1 for subjects with the minor allele and heterozygotes. However, as time progresses, the FEV_1 for subjects with the major allele decreases and becomes more similar to that of subjects with the minor allele and heterozygotes.

Figure 7.4: Gene expression for DENND5B and FEV_1 trajectory for *cis*-associated eQTL rs7313158.

sistently lower FEV_1 values than those with two copies of the minor allele (mean FEV_1 = 103%, major allele, mean FEV_1 = 104%, minor allele at study onset; mean FEV_1 = 101%, major allele, mean FEV_1 = 102%, minor allele at 48 months) (Fig. 7.5).

The rs4026608 SNP locus was associated with the IRAK3 gene, also located on chromosome 12. This SNP locus is located upstream from the IRAK3 gene, and may also indicate the presence of a nearby regulatory factor for this gene. IRAK3 variants have previously been implicated in asthma pathogenesis [8, 134]. A recent study has shown that IRAK3 can inhibit Toll-like receptor 2 (TLR2), leading to increased susceptibility to infections in asthmatic patients [180], suggesting a role for this protein in mucosal immunity and the



A: Gene expression levels of IRAK3 for subjects with different alleles for SNP rs4026608. Subjects with two copies of the major allele (0) have relatively higher levels of IRAK3 expression than subjects with two copies of the minor allele (2) and heterozygotes (1), Benjamini-Hochberg (BH) adjusted p = 0.03. **B:** Post-bronchodilator FEV_1 trajectory for subjects with different alleles for SNP marker rs4026608. Throughout the duration of the study period, the FEV_1 for subjects with the major allele is consistently lower than the FEV_1 for subjects with the minor allele and heterozygotes.

Figure 7.5: Gene expression for IRAK3 and FEV_1 trajectory for *cis*-associated eQTL rs4026608.

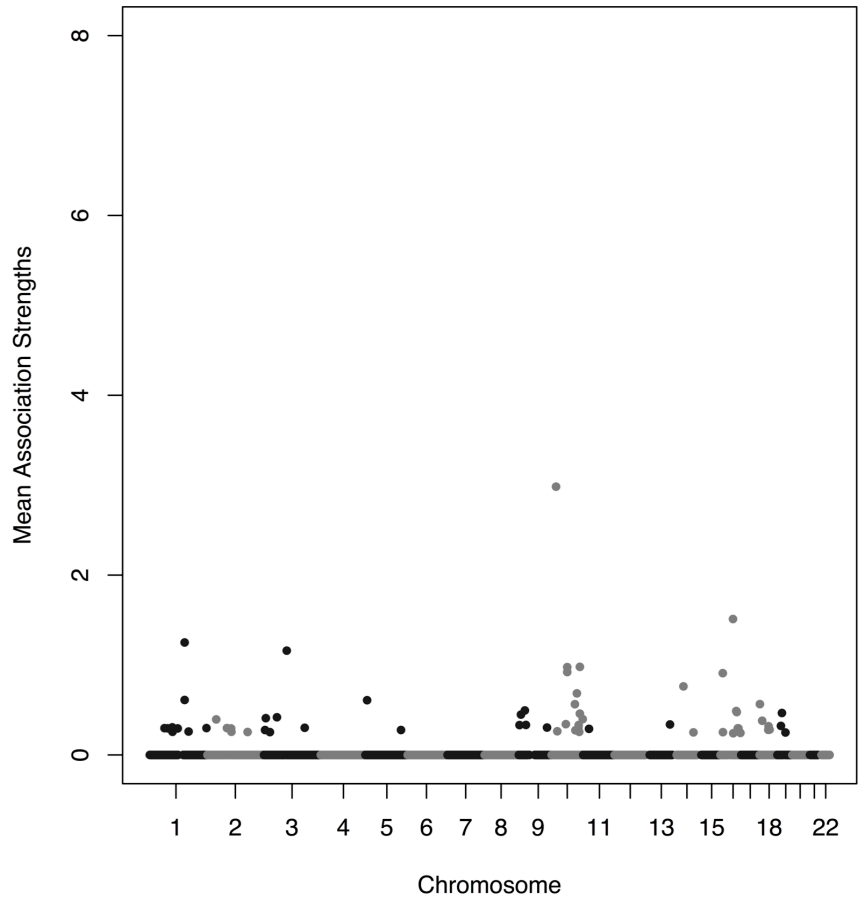
development of asthma and atopy.

7.3.4 Comparison of Temporally-Smoothed Lasso with Univariate and Lasso Association Methods

We applied two other GWAS methods, univariate regression and standard lasso for comparison. We first used univariate regression to identify associations between SNPs and FEV_1 values. We evaluated for associations at each time point individually, and did not identify any SNPs that reached genome wide significance (p-value $< 5 \times 10^{-8}$) [128, 2, 149, 90, 87].

We next used standard lasso to identify associations between SNPs and FEV_1 values. A Manhattan plot of the mean association strengths for lasso regression across time is shown

in Figure 7.6. Standard lasso found fewer associations than TESL (92 vs. 271 non-zero associations), indicating that lasso had a reduced sensitivity for detecting associations, compared to TESL. However, there were 28 SNPs that were identified by both standard lasso and TESL, depicted in Table 7.6.



The black and gray striations denote the different chromosomes.

Figure 7.6: Manhattan plot of all the mean association strengths across time for every chromosome using lasso.

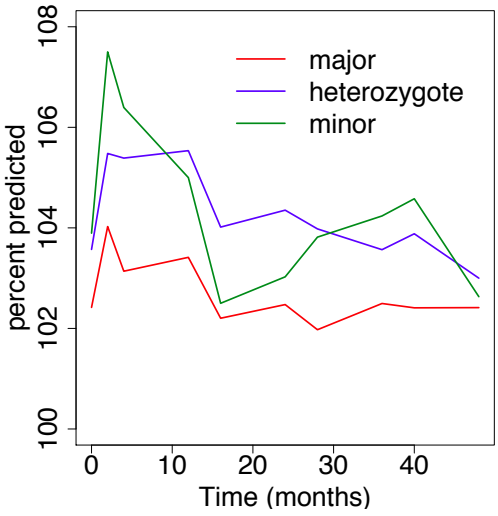
Among the SNPs identified by both lasso and TESL, rs10828784 on chromosome 10 had the largest association strength. The FEV_1 trajectory for this marker is shown in Figure 7.7A. However, lasso was unable to identify any of the top 10 associations identified by TESL, such as that for rs576967 on chromosome 9 and shown in Figure 7.7B. We believe that lasso was

SNP	CHR	BP	Mean $\|(\beta)\|$ lasso	Mean $\|(\beta)\|$ d-trait
rs12741361	1	149612957	1.25	2.65
rs1752380	1	149614370	0.61	2.32
rs1980769	1	96660789	0.31	1.99
rs9726107	1	166451806	0.26	2.05
rs6721668	2	37906399	0.40	0.95
rs11693474	2	83722702	0.30	0.30
rs956966	2	102954007	0.26	0.86
rs6745725	2	172163179	0.26	0.36
chr2:4100198	2	4100198	0.28	0.28
rs62134811	2	55032629	0.42	0.37
chr2:96998163	2	96998163	1.16	2.04
chr2:173923930	2	173923930	0.30	0.80
rs275510	5	6912895	0.61	2.16
rs4741755	9	2657929	0.33	2.52
chr9:25554159	9	25554159	0.33	2.25
rs13290170	9	8882736	0.45	1.75
chr9:120368912	9	120368912	0.30	0.95
rs10828784	10	18828279	2.98	3.18
rs12411340	10	66707498	0.98	2.01
rs11818488	10	66708116	0.92	2.10
chr11:24580358	11	24580358	0.29	1.76
rs1946151	16	53302612	1.51	2.70
rs1728778	16	67186352	0.49	1.88
rs7194083	16	9497354	0.91	1.44
rs13335305	16	10263717	0.25	0.86
rs1881220	16	84999403	0.24	1.95
rs1940289	18	951825	0.57	2.05
chr19:19145018	19	19145018	0.47	1.23

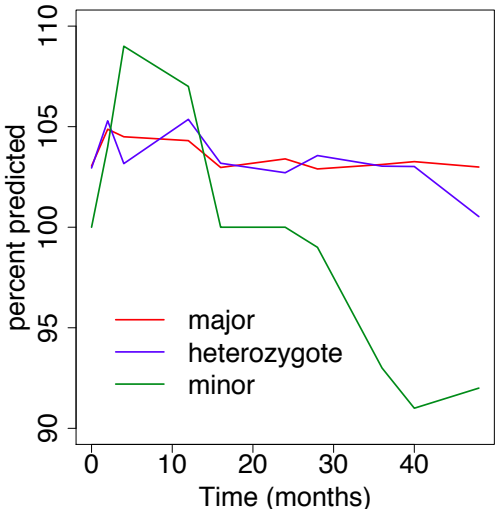
Table 7.6: List of positive associations identified by both lasso and d-trait.

able to detect an association with rs10828784 but not rs576967 due to differences in the rate of change of the FEV_1 trajectories. Although the FEV_1 trajectories for subjects with different variants of rs10828784 and rs576967 both demonstrate allelic variation, the trajectory for subjects with different variants of rs576967 shows a much steeper rate of change between time points than the trajectory for subjects with different variants of rs10828784. TESL was able to capture this temporal fluctuation whereas standard lasso was not.

A)



B)



A: Post-bronchodilator FEV_1 trajectory for subjects with different alleles for SNP marker rs10828784. This marker was identified as positively associated with FEV_1 by both standard lasso and temporally smoothed lasso. Across time, the subjects with two copies of the major allele have relatively stable values of FEV_1, that are lower than the heterozygotes and those with two copies of the minor allele. The subjects with two copies of the minor allele have significantly more variability in the value of their FEV_1, and the trait experiences many more fluctuations over time compared to subjects possessing the other alleles. **B:** Post-bronchodilator FEV_1 trajectory for subjects with different alleles for SNP marker rs576967. This marker was identified as positively associated with FEV_1 only by temporally smoothed lasso and not by standard lasso. Across time, the subjects with two copies of the major allele and the heterozygotes have relatively stable values of FEV_1. Conversely, the subjects with two copies of the minor allele have a much greater point-to-point fluctuation in the FEV_1 values that is identified by the temporally-smooth lasso method.

Figure 7.7: Comparison of d-traits with associations detected by TESL and standard lasso.

7.4 DISCUSSION

Asthma is a disease that is characterized by chronic airway inflammation, airway hyperresponsiveness and reversible airflow limitation [66]. Although progress has been made toward understanding the genetic underpinnings of childhood asthma [123], the task of identifying the genes contributing to asthma's missing heritability and linking associated genes to disease mechanisms still remains. The current paper strives to address these goals through the recognition that many genotype to phenotype associations remain undetected because

GWAS only map to trait measurements at a single time point. By using TESL, we were able to identify genomic regions that would not have been detected with traditional statistical methods by exploiting the information contained in the trajectory of phenotype values. In addition, the positive associations we identified using TESL were linked to statistically significant changes in the trajectory of the FEV_1 trait in patients with different alleles, and in several cases these associations were *cis* eQTLs, with different alleles showing different patterns of gene expression.

We demonstrated several methodological advantages to TESL. First, using this method with longitudinal measurements, we were able to leverage the joint effects of SNPs across time and increase our power to identify associations that could not be identified using either conventional univariate regression or standard lasso. Second, by identifying positive associations that correspond to allelic differences in a d-trait, the dynamic trait association analysis provides some insight into the potential function of associations. For example, we identified several positive associations between genetic loci and a relative decrease in the value of the FEV_1, suggesting that the associated genetic region was involved in the development of increased airway obstruction and worsening asthma.

In addition, the quantity of genome-wide genotype data available for analysis continues to increase. Current datasets contain millions of SNPs, more densely sampled from the genome than ever before. However, many of the SNPs are in LD with each other, and thus are highly correlated. In our model, we use the lasso (l_1 penalty) to encourage sparsity in the selected predictors. However, while lasso is known to perform well in high-dimensional settings where there is minimal correlation among predictor variables [14, 191, 171], it has been shown that it can be unstable when a high degree of correlation exists. This means that when two predictors are highly correlated, lasso cannot distinguish between them, leading to issues of repeatability and generalizability [185]. This is potentially problematic for association analyses where there is a preference for large numbers of correlated SNPs that allow for increased precision in mapping associations to particular regions of the genome.

For this particular analysis, we made several modifications to the original data set in order to avoid any stability problems caused by the structure inherent in the data. We first pruned the SNPs in LD. However, the disadvantage associated with heavily pruning

SNPs is that it becomes more difficult to map associations to a particular location of the genome and reduces the overall likelihood of identifying potentially causative SNPs. Thus, to increase the number of SNPs we include in our analysis, we create multiple data sets of the same chromosome, each data set containing near independent SNPs. Moving forward, investigating how to account for this structure in the model and algorithm would contribute significantly to achieving robust and generalizable association results that may be replicated in similar populations.

Finding associations between genetic variants and phenotypic measurements given the high-dimensional nature of GWAS data is a challenging problem. While the strength of the method presented is the ability to accommodate d-traits, there are also certain limitations. For example, similar to other GWAS methods, we assume a common disease, common variant model. Thus, we are limited to markers with a MAF>5%. Future work should address how to increase the sensitivity of the algorithm to detect rare variants in a setting where dynamic trait data is available. A second limitation is that a d-trait may be influenced by multiple genetic loci, each having different types of trajectories in its genetic effect on the d-trait, such as a mixture of cyclic and linear genetic effects on d-traits. In such a case, the local autoregressive model would capture only the mixture of different types of genetic effects, and TESL would not be able to separate the different types of genetic effects. A third limitation is that our model assumes the d-trait measurements and the SNP effects on these d-traits are synchronized in time across subjects. As longitudinal data collected for one individual can be shifted in time compared to the data for a different individual, it is important to take into account such shifts during analysis. Future work can also address whether using higher order autoregressive models would be advantageous for this application. A method that can overcome these limitations and reliably detect d-trait associations will be key to understanding the genetic basis of d-traits.

In conclusion, we have introduced a novel method for identifying associations between genetic loci and d-traits that vary over time. We have also demonstrated through simulations that the d-trait method has a low type I error rate and higher power to detect significant associations than lasso or univariate regression methods. Finally, we applied the d-trait method to a population of children with persistent asthma and identified several loci associated with

allelic differences in FEV_1 trajectories over time. Multiple loci were also associated with statistically significant allelic differences between levels of gene expression. Thus, we believe this method represents a new standard for association analysis of longitudinal dynamic traits.

8.0 CONCLUSIONS

8.1 SUMMARY

In this thesis, we introduced an integrative computational framework that may be used to identify complex disease endotypes. We began by using cluster analysis to define five endotypes among children with mild-moderate asthma using a selected set of clinical traits. We next explored the relationship between these endotypes and levels of gene expression in $CD4^+$ T-lymphocytes, and found a set of genes with expression patters that were closely associated with our endotypes, particularly the degree of atopy present. We grouped this set of genes into highly correlated modules, and found that one module could be used to predict the degree of atopy present in an independent cohort of asthmatic patients. Within this module, we also identified a common motif corresponding to a common regulatory molecule. Finally, among our endotypes, we found that there was one that was strongly linked to low lung function. To further investigate the genetic basis of this endotype, we used a novel method to identify genetic associations with this longitudinal trait. We identified several novel and confirmatory associations, and found that several associations were *(*cis)-eQTLs, corresponding to changes in gene expression that varied among subjects with different alleles.

Our results support the use of an integrative computational framework to identify endotypes of complex diseases. We have developed predictive models that defines discrete patient subsets with unique clinical attributes, discrete clinical trajectories, and variable responsiveness to anti-asthma controller medications. Recognition of these endotypes and their clinical relevance should motivate novel strategies in both the research and clinical settings. Though asthma has long been recognized as a clinically and etiologically heterogeneous disorder, our results, together support the presence of no more than 5-6 asthma phenotypic clusters, and

suggest that the heterogeneity is perhaps not as daunting as previously anticipated. The observed between-cluster differences in the prevalence of specific environmental and genetic factors suggest that important etiological differences underlie the configuration of different asthma phenotypic clusters. Integration of gene expression profiles and genetic data allowed us to identify multiple genes and genetic variants with plausible links to molecular mechanisms underlying asthma pathogenesis. Thus, in addition to helping inform clinical management, these integrated multivariable phenotypic classification schemes should help accelerate research efforts in defining the molecular and environmental underpinnings of this complex airways disease.

8.2 FUTURE DIRECTIONS

Moving forward, it will be necessary to validate the findings described in this thesis. Although we identified several clinical asthma endotypes in a large cohort of asthmatic children, prospective validation will be necessary before this framework is ready for widespread use in the clinical setting. Further validation of gene expression and genetic endotypes will also need to be performed in larger and more heterogeneous populations. In addition, further investigation of genetic variants and gene expression profiles corresponding to different endotypes should be performed to identify common regulatory molecules as potential novel drug targets.

The integrative computational framework described in this thesis suggests the possibility that multiple types of data may one day be incorporated into routine clinical care. In order for this to happen, it will be crucial to develop user-friendly, web-based applications that can be administered and interpreted easily. In clinic visits, patients specific clinical information could be obtained from clinicians, along with genetic and gene expression profiles. Targeted software could be used to assign patients to specific disease endotypes based upon this information, and everything could be integrated into the EMR for appropriate risk-stratification and selection of customized pharmacologic treatments. Such a program could move us closer to a more efficient and personalized model of medical practice.

BIBLIOGRAPHY

[1] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–7, Jul 2006.

[2] Verneri Anttila, Hreinn Stefansson, Mikko Kallela, Unda Todt, Gisela M Terwindt, M Stella Calafato, Dale R Nyholt, Antigone S Dimas, Tobias Freilinger, Bertram Müller-Myhsok, Ville Artto, Michael Inouye, Kirsi Alakurtti, Mari A Kaunisto, Eija Hämäläinen, Boukje de Vries, Anine H Stam, Claudia M Weller, Axel Heinze, Katja Heinze-Kuhn, Ingrid Goebel, Guntram Borck, Hartmut Göbel, Stacy Steinberg, Christiane Wolf, Asgeir Björnsson, Gretar Gudmundsson, Malene Kirchmann, Anne Hauge, Thomas Werge, Jean Schoenen, Johan G Eriksson, Knut Hagen, Lars Stovner, H-Erich Wichmann, Thomas Meitinger, Michael Alexander, Susanne Moebus, Stefan Schreiber, Yurii S Aulchenko, Monique M B Breteler, Andre G Uitterlinden, Albert Hofman, Cornelia M van Duijn, Päivi Tikka-Kleemola, Salli Vepsäläinen, Susanne Lucae, Federica Tozzi, Pierandrea Muglia, Jeffrey Barrett, Jaakko Kaprio, Markus Färkkilä, Leena Peltonen, Kari Stefansson, John-Anker Zwart, Michel D Ferrari, Jes Olesen, Mark Daly, Maija Wessman, Arn M J M van den Maagdenberg, Martin Dichgans, Christian Kubisch, Emmanouil T Dermitzakis, Rune R Frants, Aarno Palotie, and International Headache Genetics Consortium. Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nat Genet*, 42(10):869–73, Oct 2010.

[3] T Aoki, Y Matsumoto, K Hirata, K Ochiai, M Okada, K Ichikawa, M Shibasaki, T Arinami, R Sumazaki, and E Noguchi. Expression profiling of genes related to asthma exacerbations. *Clin Exp Allergy*, 39(2):213–21, Feb 2009.

[4] A Arkin, J Ross, and H H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149(4):1633–48, Aug 1998.

[5] Jeff Augen. *Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine*. Addison-Wesley Professional, 2005.

[6] K J Baines, J L Simpson, N A Bowden, R J Scott, and P G Gibson. Differential gene expression and cytokine production from neutrophils in asthma phenotypes. *Eur Respir J*, 35(3):522–31, Mar 2010.

[7] Katherine J Baines, Jodie L Simpson, Lisa G Wood, Rodney J Scott, and Peter G Gibson. Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. *J Allergy Clin Immunol*, 127(1):153–60, 160.e1–9, Jan 2011.

[8] Lenuta Balaci, Maria Cristina Spada, Nazario Olla, Gabriella Sole, Laura Loddo, Francesca Anedda, Silvia Naitza, Maria Antonietta Zuncheddu, Andrea Maschio, Daniele Altea, Manuela Uda, Sabrina Pilia, Serena Sanna, Marco Masala, Laura Crisponi, Matilde Fattori, Marcella Devoto, Silvia Doratiotto, Stefania Rassu, Simonetta Mereu, Enrico Giua, Natalina Graziella Cadeddu, Roberto Atzeni, Umberto Pelosi, Adriano Corrias, Roberto Perra, Pier Luigi Torrazza, Pietro Pirina, Francesco Ginesu, Silvano Marcias, Maria Grazia Schintu, Gennaro Sergio Del Giacco, Paolo Emilio Manconi, Giovanni Malerba, Andrea Bisognin, Elisabetta Trabetti, Attilio Boner, Lydia Pescollderungg, Pier Franco Pignatti, David Schlessinger, Antonio Cao, and Giuseppe Pilia. Irak-m is involved in the pathogenesis of early-onset persistent asthma. *Am J Hum Genet*, 80(6):1103–14, Jun 2007.

[9] Barabasi and Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, Oct 1999.

[10] E D Bateman, S S Hurd, P J Barnes, J Bousquet, J M Drazen, M FitzGerald, P Gibson, K Ohta, P O'Byrne, S E Pedersen, E Pizzichini, S D Sullivan, S E Wenzel, and H J Zar. Global strategy for asthma management and prevention: Gina executive summary. *Eur Respir J*, 31(1):143–78, Jan 2008.

[11] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[13] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188, 2001.

[14] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[15] Biomart database. http://biomart.org/.

[16] Unnur S Bjornsdottir, Stephen T Holgate, Padmalatha S Reddy, Andrew A Hill, Charlotte M McKee, Cristina I Csimma, Amy A Weaver, Holly M Legault, Clayton G Small, Renee C Ramsey, Debra K Ellis, Conor M Burke, Philip J Thompson, Peter H Howarth, Andrew J Wardlaw, Phillip G Bardin, David I Bernstein, Louis B Irving, Geoffrey L Chupp, George W Bensch, Gregory W Bensch, Jon E Stahlman, Monroe Karetzky, James W Baker, Rachel L Miller, Brad H Goodman, Donald G Raible, Samuel J Goldman, Douglas K Miller, John L Ryan, Andrew J Dorner, Frederick W Immermann, and

Margot O'Toole. Pathways activated during human asthma exacerbation as revealed by gene expression patterns in blood. *PLoS One*, 6(7):e21902, 2011.

[17] Anne Boudier, Ivan Curjuric, Xavier Basagaña, Hana Hazgui, Josep M Anto, Jean Bousquet, Pierre O Bridevaux, Elise Dupuis-Lozeron, Judith Garcia-Aymerich, Joachim Heinrich, Christer Janson, Nino Künzli, Bénédicte Leynaert, Roberto de Marco, Thierry Rochat, Christian Schindler, Raphaëlle Varraso, Isabelle Pin, Nicole Probst-Hensch, Jordi Sunyer, Francine Kauffmann, and Valérie Siroux. Ten-year follow-up of cluster-based asthma phenotypes in adults. a pooled analysis of three cohorts. *Am J Respir Crit Care Med*, 188(5):550–60, Sep 2013.

[18] Paul Brazhnik, Alberto de la Fuente, and Pedro Mendes. Gene networks: how to put the function in genomics. *Trends Biotechnol*, 20(11):467–72, Nov 2002.

[19] Leo Breiman. *Classification and regression trees*. Wadsworth International Group, Belmont, Calif., 1984.

[20] Leo Breiman. *Classification and regression trees*. Wadsworth International Group, Belmont, Calif., 1984.

[21] Marc R J Carlson, Bin Zhang, Zixing Fang, Paul S Mischel, Steve Horvath, and Stanley F Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7:40, 2006.

[22] J Chambers, A Angulo, D Amaratunga, H Guo, Y Jiang, J S Wan, A Bittner, K Frueh, M R Jackson, P A Peterson, M G Erlander, and P Ghazal. Dna microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J Virol*, 73(7):5757–66, Jul 1999.

[23] X. Chen, Q. Lin, S. Kim, J.G. Carbonell, and E.P. Xing. Smoothing proximal gradient method for general structured sparse learning. *Arxiv preprint arXiv:1202.3708*, 2012.

[24] Bradley E Chipps. Inhaled corticosteroid therapy for patients with persistent asthma: learnings from studies of inhaled budesonide. *Allergy Asthma Proc*, 30(3):217–28, 2009.

[25] David F Choy, Barmak Modrek, Alexander R Abbas, Sarah Kummerfeld, Hilary F Clark, Lawren C Wu, Grazyna Fedorowicz, Zora Modrusan, John V Fahy, Prescott G Woodruff, and Joseph R Arron. Gene expression patterns of th2 inflammation and intercellular communication in asthmatic airways. *J Immunol*, 186(3):1861–9, Feb 2011.

[26] G L Colice, J V Burgt, J Song, P Stampone, and P J Thompson. Categorizing asthma severity. *Am J Respir Crit Care Med*, 160(6):1962–7, Dec 1999.

[27] W M Corrao, S S Braman, and R S Irwin. Chronic cough as the sole presenting manifestation of bronchial asthma. *N Engl J Med*, 300(12):633–7, Mar 1979.

[28] Michael C Costanza, Sigrid Beer-Borst, Richard W James, Jean-Michel Gaspoz, and Alfredo Morabia. Consistency between cross-sectional and longitudinal snp: blood lipid associations. *Eur J Epidemiol*, 27(2):131–8, Feb 2012.

[29] Ronina A Covar, Anne L Fuhlbrigge, Paul Williams, H William Kelly, and the Childhood Asthma Management Program Research Group. The childhood asthma management program (camp): Contributions to the understanding of therapy and the natural history of childhood asthma. *Curr Respir Care Rep*, 1(4):243–250, Dec 2012.

[30] Ronina A Covar, Robert Strunk, Robert S Zeiger, Laura A Wilson, Andrew H Liu, Scott Weiss, James Tonascia, Joseph D Spahn, Stanley J Szefler, and Childhood Asthma Management Program Research Group. Predictors of remitting, periodic, and persistent childhood asthma. *J Allergy Clin Immunol*, 125(2):359–366.e3, Feb 2010.

[31] Kiranmoy Das, Jiahan Li, Zhong Wang, Chunfa Tong, Guifang Fu, Yao Li, Meng Xu, Kwangmi Ahn, David Mauger, Runze Li, and Rongling Wu. A dynamic model for genome-wide association studies. *Hum Genet*, 129(6):629–39, Jun 2011.

[32] Philip L De Jager, Joshua M Shulman, Lori B Chibnik, Brendan T Keenan, Towfique Raj, Robert S. Wilson, L. Yu, S.E. Leurgans, D. Tran, C.D. Anderson, A. Biffi, J.J. Corneveaux, M.J. Huentelman, J. Rosand, M.J. Daly, A.J. Myers, E.M. Reiman, D.A. Bennett, and Denis A Evans. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiol Aging*, 33(5):1017.e1–15, May 2012.

[33] Marcilio Carlos Pereira de Soutolio, Daniel S. A. de Araujo, Ivan G. Costa, Rodrigo G. F. Soares, Teresa Bernarda Ludermir, and Alexander Schliep. Comparative study on normalization procedures for cluster analysis of gene expression datasets. In *IJCNN*, pages 2792–2798, 2008.

[34] AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39:1–38, 1977.

[35] E J Devor and M H Crawford. Family resemblance for normal pulmonary function. *Ann Hum Biol*, 11(5):439–48, 1984.

[36] J C Dewar and A P Wheatley. The heritability of allergic disease. *Monogr Allergy*, 33:4–34, 1996.

[37] Patrik D'haeseleer. How does gene expression clustering work? *Nat Biotechnol*, 23(12):1499–501, Dec 2005.

[38] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means, spectral clustering and normalized cuts. *Knowledge Discovery and Data Mining (KDD)*, 2004.

[39] G M Dolganov, P G Woodruff, A A Novikov, Y Zhang, R E Ferrando, R Szubin, and J V Fahy. A novel method of gene transcript profiling in airway biopsy homogenates reveals increased expression of a na+-k+-cl- cotransporter (nkcc1) in asthmatic subjects. *Genome Res*, 11(9):1473–83, Sep 2001.

[40] Pan Du, Warren A Kibbe, and Simon M Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–8, Jul 2008.

[41] R.H. Duerr, K.D. Taylor, S.R. Brant, J.D. Rioux, M.S. Silverberg, M.J. Daly, A.H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Britton, H. Yang, S. Targan, L.W. Datta, E.O. Kistner, L.P. Schumm, A.T. Lee, P.K. Gregersen, M.M. Barmada, J.I. Rotter, D.L. Nicolae, and J.H. Cho. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, 314(5804):1461–3, Dec 2006.

[42] D L Duffy. A population-based study of bronchial asthma in adult twin pairs. *Chest*, 102(2):654, Aug 1992.

[43] D L Duffy, N G Martin, D Battistutta, J L Hopper, and J D Mathews. Genetics of asthma and hay fever in australian twins. *Am Rev Respir Dis*, 142(6 Pt 1):1351–8, Dec 1990.

[44] O.J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[45] Catherine Duvernelle, Véronique Freund, and Nelly Frossard. Transforming growth factor-beta and its role in asthma. *Pulm Pharmacol Ther*, 16(4):181–196, 2003.

[46] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, Dec 1998.

[47] Manuel A R Ferreira, Melanie C Matheson, David L Duffy, Guy B Marks, Jennie Hui, Peter Le Souëf, Patrick Danoy, Svetlana Baltic, Dale R Nyholt, Mark Jenkins, Catherine Hayden, Gonneke Willemsen, Wei Ang, Mikko Kuokkanen, John Beilby, Faang Cheah, Eco J C de Geus, Adaikalavan Ramasamy, Sailaja Vedantam, Veikko Salomaa, Pamela A Madden, Andrew C Heath, John L Hopper, Peter M Visscher, Bill Musk, Stephen R Leeder, Marjo-Riitta Jarvelin, Craig Pennell, Dorret I Boomsma, Joel N Hirschhorn, Haydn Walters, Nicholas G Martin, Alan James, Graham Jones, Michael J Abramson, Colin F Robertson, Shyamali C Dharmage, Matthew A Brown, Grant W Montgomery, Philip J Thompson, and Australian Asthma Genetics Consortium. Identification of il6r and chromosome 11q13.5 as risk loci for asthma. *Lancet*, 378(9795):1006–14, Sep 2011.

[48] Anne M Fitzpatrick, W Gerald Teague, Deborah A Meyers, Stephen P Peters, Xingnan Li, Huashi Li, Sally E Wenzel, Shean Aujla, Mario Castro, Leonard B Bacharier, Benjamin M Gaston, Eugene R Bleecker, Wendy C Moore, and National Institutes

of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the national institutes of health/national heart, lung, and blood institute severe asthma research program. *J Allergy Clin Immunol*, 127(2):382–389.e1–13, Feb 2011.

[49] Erick Forno, Anne Fuhlbrigge, Manuel E Soto-Quirós, Lydiana Avila, Benjamin A Raby, John Brehm, Jody M Sylvia, Scott T Weiss, and Juan C Celedón. Risk factors and predictive clinical scores for asthma exacerbations in childhood. *Chest*, 138(5):1156–65, Nov 2010.

[50] M F Frecker, M Preus, L Kozma, E Krasrits, V Stenszky, C Balazś, and N R Farid. Heterogeneity by cluster analysis techniques of graves' patients typed for hla dr and igg heavy chain markers. *Mol Biol Med*, 3(1):63–71, Feb 1986.

[51] N Friedman, M Linial, I Nachman, and D Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–20, 2000.

[52] T Ganz. Defensins and host defense. *Science*, 286(5439):420–1, Oct 1999.

[53] Timothy S Gardner, Diego di Bernardo, David Lorenz, and James J Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–5, Jul 2003.

[54] Peter S Gargalovic, Minori Imura, Bin Zhang, Nima M Gharavi, Michael J Clark, Joanne Pagnon, Wen-Pin Yang, Aiqing He, Amy Truong, Shilpa Patel, Stanley F Nelson, Steve Horvath, Judith A Berliner, Todd G Kirchgessner, and Aldons J Lusis. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc Natl Acad Sci U S A*, 103(34):12741–6, Aug 2006.

[55] Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos, Alec Brozell, Eric E Schadt, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*, 2(8):e130, Aug 2006.

[56] Willemijn M Gommans, Hidde J Haisma, and Marianne G Rots. Engineering zinc finger protein transcription factors: the therapeutic relevance of switching endogenous gene expression on or off at command. *J Mol Biol*, 354(3):507–19, Dec 2005.

[57] P Gonzalez, J S Zigler, Jr, D L Epstein, and T Borrás. Identification and isolation of differentially expressed genes from very small tissue samples. *Biotechniques*, 26(5):884–6, 888–92, May 1999.

[58] L Gu, S Tseng, R M Horner, C Tam, M Loda, and B J Rollins. Control of th2 polarization by the chemokine monocyte chemoattractant protein-1. *Nature*, 404(6776):407–11, Mar 2000.

[59] Hakon Hakonarson, Unnur S Bjornsdottir, Eva Halapi, Jonathan Bradfield, Florian Zink, Magali Mouy, Hildur Helgadottir, Asta S Gudmundsdottir, Hjalti Andrason, Asdis E Adalsteinsdottir, Kristleifur Kristjansson, Illugi Birkisson, Thor Arnason, Margret Andresdottir, David Gislason, Thorarinn Gislason, Jeffrey R Gulcher, and Kari Stefansson. Profiling of genes expressed in peripheral blood mononuclear cells predicts glucocorticoid sensitivity in asthma patients. *Proc Natl Acad Sci U S A*, 102(41):14789–94, Oct 2005.

[60] Pranab Haldar, Ian D Pavord, Dominic E Shaw, Michael A Berry, Michael Thomas, Christopher E Brightling, Andrew J Wardlaw, and Ruth H Green. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med*, 178(3):218–24, Aug 2008.

[61] Nadia N Hansel, Sara C Hilmer, Steve N Georas, Leslie M Cope, Jia Guo, Rafael A Irizarry, and Gregory B Diette. Oligonucleotide-microarray analysis of peripheral-blood lymphocytes in severe asthma. *J Lab Clin Med*, 145(5):263–74, May 2005.

[62] Ke Hao, Yohan Bossé, David C Nickle, Peter D Paré, Dirkje S Postma, Michel Laviolette, Andrew Sandford, Tillie L Hackett, Denise Daley, James C Hogg, W Mark Elliott, Christian Couture, Maxime Lamontagne, Corry-Anke Brandsma, Maarten van den Berge, Gerard Koppelman, Alise S Reicin, Donald W Nicholson, Vladislav Malkov, Jonathan M Derry, Christine Suver, Jeffrey A Tsou, Amit Kulkarni, Chunsheng Zhang, Rupert Vessey, Greg J Opiteck, Sean P Curtis, Wim Timens, and Don D Sin. Lung eqtls to help reveal the molecular underpinnings of asthma. *PLoS Genet*, 8(11):e1003029, 2012.

[63] Michishige Harada, Tomomitsu Hirota, Aya I Jodo, Yuki Hitomi, Masafumi Sakashita, Tatsuhiko Tsunoda, Takehiko Miyagawa, Satoru Doi, Makoto Kameda, Kimie Fujita, Akihiko Miyatake, Tadao Enomoto, Emiko Noguchi, Hironori Masuko, Tohru Sakamoto, Nobuyuki Hizawa, Yoichi Suzuki, Shigemi Yoshihara, Mitsuru Adachi, Motohiro Ebisawa, Hirohisa Saito, Kenji Matsumoto, Toshiharu Nakajima, Rasika A Mathias, Nicholas Rafaels, Kathleen C Barnes, Blanca E Himes, Qing Ling Duan, Kelan G Tantisira, Scott T Weiss, Yusuke Nakamura, Steven F Ziegler, and Mayumi Tamari. Thymic stromal lymphopoietin gene promoter polymorphisms are associated with susceptibility to bronchial asthma. *Am J Respir Cell Mol Biol*, 44(6):787–93, Jun 2011.

[64] John Hardy and Andrew Singleton. Genomewide association studies and human disease. *N Engl J Med*, 360(17):1759–68, Apr 2009.

[65] Tmirah Haselkorn, Robert S Zeiger, Bradley E Chipps, David R Mink, Stanley J Szefler, F Estelle R Simons, Marc Massanari, and James E Fish. Recent asthma exacerbations predict future exacerbations in children with severe or difficult-to-treat asthma. *J Allergy Clin Immunol*, 124(5):921–7, Nov 2009.

[66] National heart, lung, and blood institute. Expert panel report 3: Guidelines for the diagnosis and treatment of asthma. Technical report, National heart, lung and blood institute, 2007.

[67] D Heckerman. *Learning in Graphical Models*, chapter A tutorial on learning with Bayesian networks. Kluwer, 1998.

[68] S G Hilsenbeck, W E Friedrichs, R Schiff, P O'Connell, R K Hansen, C K Osborne, and S A Fuqua. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst*, 91(5):453–9, Mar 1999.

[69] Blanca E Himes, Gary M Hunninghake, James W Baurley, Nicholas M Rafaels, Patrick Sleiman, David P Strachan, Jemma B Wilk, Saffron A G Willis-Owen, Barbara Klanderman, Jessica Lasky-Su, Ross Lazarus, Amy J Murphy, Manuel E Soto-Quiros, Lydiana Avila, Terri Beaty, Rasika A Mathias, Ingo Ruczinski, Kathleen C Barnes, Juan C Celedón, William O C Cookson, W James Gauderman, Frank D Gilliland, Hakon Hakonarson, Christoph Lange, Miriam F Moffatt, George T O'Connor, Benjamin A Raby, Edwin K Silverman, and Scott T Weiss. Genome-wide association analysis identifies pde4d as an asthma-susceptibility gene. *Am J Hum Genet*, 84(5):581–93, May 2009.

[70] Blanca E Himes, Xiaofeng Jiang, Ruoxi Hu, Ann C Wu, Jessica A Lasky-Su, Barbara J Klanderman, John Ziniti, Jody Senter-Sylvia, John J Lima, Charles G Irvin, Stephen P Peters, Deborah A Meyers, Eugene R Bleecker, Michiaki Kubo, Mayumi Tamari, Yusuke Nakamura, Stanley J Szefler, Robert F Lemanske, Jr, Robert S Zeiger, Robert C Strunk, Fernando D Martinez, John P Hanrahan, Gerard H Koppelman, Dirkje S Postma, Maartje A E Nieuwenhuis, Judith M Vonk, Reynold A Panettieri, Jr, Amy Markezich, Elliot Israel, Vincent J Carey, Kelan G Tantisira, Augusto A Litonjua, Quan Lu, and Scott T Weiss. Genome-wide association analysis in asthma subjects identifies spats2l as a novel bronchodilator response gene. *PLoS Genet*, 8(7):e1002824, Jul 2012.

[71] Blanca E Himes, Barbara Klanderman, John Ziniti, Jody Senter-Sylvia, Manuel E Soto-Quiros, Lydiana Avila, Juan C Celedón, Christoph Lange, Thomas J Mariani, Jessica Lasky-Su, Craig P Hersh, Benjamin A Raby, Edwin K Silverman, Scott T Weiss, and Dawn L Demeo. Association of serpine2 with asthma. *Chest*, 140(3):667–74, Sep 2011.

[72] A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[73] R J Hopp, A K Bewtra, G D Watt, N M Nair, and R G Townley. Genetic analysis of allergic disease in twins. *J Allergy Clin Immunol*, 73(2):265–70, Feb 1984.

[74] S Horvath, B Zhang, M Carlson, K V Lu, S Zhu, R M Felciano, M F Laurance, W Zhao, S Qi, Z Chen, Y Lee, A C Scheck, L M Liau, H Wu, D H Geschwind, P G Febbo, H I Kornblum, T F Cloughesy, S F Nelson, and P S Mischel. Analysis of oncogenic

signaling networks in glioblastoma identifies aspm as a molecular target. *Proc Natl Acad Sci U S A*, 103(46):17402–7, Nov 2006.

[75] S Horvath, B Zhang, M Carlson, K V Lu, S Zhu, R M Felciano, M F Laurance, W Zhao, S Qi, Z Chen, Y Lee, A C Scheck, L M Liau, H Wu, D H Geschwind, P G Febbo, H I Kornblum, T F Cloughesy, S F Nelson, and P S Mischel. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proc Natl Acad Sci U S A*, 103(46):17402–7, Nov 2006.

[76] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol*, 4(8), 2008.

[77] Judie A Howrylak, Anne L Fuhlbrigge, Robert C Strunk, Robert S Zeiger, Scott T Weiss, and Benjamin A Raby. Classification of childhood asthma phenotypes and clinical responses to inhaled anti-inflammatory medications. *J Allergy and Clin Immunol* (submitted), 2013.

[78] Gary M Hunninghake, Jen-hwa Chu, Sunita S Sharma, Michael H Cho, Blanca E Himes, Angela J Rogers, Amy Murphy, Vincent J Carey, and Benjamin A Raby. The cd4+ t-cell transcriptome and serum ige in asthma: Il17rb and the role of sex. *BMC Pulm Med*, 11:17, 2011.

[79] Medea Imboden, Emmanuelle Bouzigon, Ivan Curjuric, Adaikalavan Ramasamy, Ashish Kumar, Dana B. Hancock, J.B. Wilk, J.M. Vonk, G.A. Thun, V. Siroux, R. Nadif, F. Monier, J.R. Gonzalez, M. Wjst, J. Heinrich, L.R Loehr, N. Franceschini, K.E. North, J. Altmueller, G.H. Koppelman, S. Guerra, F. Kronenberg, M. Lathrop, M.F. Moffatt, G.T. OConnor, D.P. Strachan, D.S. Postma, S.J. London, C. Schindler, M. Kogevinas, F. Kauffmann, D.L. Jarvis, F. Demenais, and Nicole M Probst-Hensch. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*, 129(5):1218–28, May 2012.

[80] Seiya Imoto, Takao Goto, and Satoru Miyano. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. *Pac Symp Biocomput*, pages 175–86, 2002.

[81] The CAMP Investigators. The childhood asthma management program (camp): design, rationale, and methods. childhood asthma management program research group. *Control Clin Trials*, 20(1):91–120, Feb 1999.

[82] The CAMP Investigators. Long-term effects of budesonide or nedocromil in children with asthma. the childhood asthma management program research group. *N Engl J Med*, 343(15):1054–63, Oct 2000.

[83] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, Apr 2003.

[84] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 2010.

[85] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[86] H Jonuleit, E Schmitt, G Schuler, J Knop, and A H Enk. Induction of interleukin 10-producing, nonproliferating cd4(+) t cells with regulatory properties by repetitive stimulation with allogeneic immature human dendritic cells. *J Exp Med*, 192(9):1213–22, Nov 2000.

[87] Yoichiro Kamatani, Koichi Matsuda, Yukinori Okada, Michiaki Kubo, Naoya Hosono, Yataro Daigo, Yusuke Nakamura, and Naoyuki Kamatani. Genome-wide association study of hematological and biochemical traits in a japanese population. *Nat Genet*, 42(3):210–5, Mar 2010.

[88] B Kapitein, M O Hoekstra, E H J Nijhuis, D J Hijnen, H G M Arets, J L L Kimpen, and E F Knol. Gene expression in cd4+ t-cells reflects heterogeneity in infant wheezing phenotypes. *Eur Respir J*, 32(5):1203–12, Nov 2008.

[89] A. Karatzoglou, K. Hornik, A. Smola, and A. Zeileis. Kernlab - an s4 package for kernel methods in r. *Journal of Statistical Software*, 11, 2004.

[90] Sekar Kathiresan, Olle Melander, Candace Guiducci, Aarti Surti, Noël P Burtt, Mark J Rieder, Gregory M Cooper, Charlotta Roos, Benjamin F Voight, Aki S Havulinna, Björn Wahlstrand, Thomas Hedner, Dolores Corella, E Shyong Tai, Jose M Ordovas, Göran Berglund, Erkki Vartiainen, Pekka Jousilahti, Bo Hedblad, Marja-Riitta Taskinen, Christopher Newton-Cheh, Veikko Salomaa, Leena Peltonen, Leif Groop, David M Altshuler, and Marju Orho-Melander. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet*, 40(2):189–97, Feb 2008.

[91] S Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–8, Oct 1969.

[92] Mark P Keller, YounJeong Choi, Ping Wang, Dawn Belt Davis, Mary E Rabaglia, Angie T Oler, Donald S Stapleton, Carmen Argmann, Kathy L Schueler, Steve Edwards, H Adam Steinberg, Elias Chaibub Neto, Robert Kleinhanz, Scott Turner, Marc K Hellerstein, Eric E Schadt, Brian S Yandell, Christina Kendziorski, and Alan D Attie. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res*, 18(5):706–16, May 2008.

[93] W J Kelly, I Hudson, J Raven, P D Phelan, M C Pain, and A Olinsky. Childhood asthma and adult lung function. *Am Rev Respir Dis*, 138(1):26–30, Jul 1988.

[94] C M Kendziorski, M Chen, M Yuan, H Lan, and A D Attie. Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*, 62(1):19–27, Mar 2006.

[95] David J. Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and a critique. *Strategic Management Journal*, 17(6):441–458, 1996.

[96] Chang Sik Kim. Bayesian orthogonal least squares (bols) algorithm for reverse engineering of gene regulatory networks. *BMC Bioinformatics*, 8:251, 2007.

[97] Seyoung Kim, Judie Howrylak, Kyung Ah Sohn, Scott T Weiss, Benjamin A Raby, and Eric P Xing. Association analysis of dynamic traits via temporally-smoothed lasso. *Manuscript under Review*, 2013.

[98] Masayuki Kitajima, Chiaki Iwamura, Takako Miki-Hosokawa, Kenta Shinoda, Yusuke Endo, Yukiko Watanabe, Ryo Shinnakasu, Hiroyuki Hosokawa, Kahoko Hashimoto, Shinichiro Motohashi, Haruhiko Koseki, Osamu Ohara, Masakatsu Yamashita, and Toshinori Nakayama. Enhanced th2 cell differentiation and allergen-induced airway inflammation in zfp35-deficient mice. *J Immunol*, 183(8):5388–96, Oct 2009.

[99] G Koeppen-Schomerus, J Stevenson, and R Plomin. Genes and environment in asthma: a study of 4 year old twins. *Arch Dis Child*, 85(5):398–400, Nov 2001.

[100] E S Lander and D Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–99, Jan 1989.

[101] Christoph Lange, Kristel K Van Steen, T Andrew, H Lyon, DL DeMeo, Benjamin A Raby, Amy Murphy, Edwin K Silverman, A MacGregor, Scott T Weiss, and Nan M Laird. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Mol Biol*, 3(17), 2004.

[102] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[103] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–20, Mar 2008.

[104] Catherine Laprise, Robert Sladek, André Ponton, Marie-Claude Bernier, Thomas J Hudson, and Michel Laviolette. Functional classes of bronchial mucosa genes that are differentially expressed in asthma. *BMC Genomics*, 5(1):21, Mar 2004.

[105] G L Larsen. Differences between adult and childhood asthma. *J Allergy Clin Immunol*, 106(3 Suppl):S153–7, Sep 2000.

[106] Robert F Lemanske, Jr, David T Mauger, Christine A Sorkness, Daniel J Jackson, Susan J Boehmer, Fernando D Martinez, Robert C Strunk, Stanley J Szefler, Robert S Zeiger, Leonard B Bacharier, Ronina A Covar, Theresa W Guilbert, Gary Larsen, Wayne J Morgan, Mark H Moss, Joseph D Spahn, Lynn M Taussig, and Childhood

Asthma Research and Education (CARE) Network of the National Heart, Lung, and Blood Institute. Step-up therapy for children with uncontrolled asthma receiving inhaled corticosteroids. *N Engl J Med*, 362(11):975–85, Mar 2010.

[107] T F Leung, C Y Li, E K H Liu, N L S Tang, I H S Chan, E Yung, G W K Wong, and C W K Lam. Asthma and atopy are associated with defb1 polymorphisms in chinese children. *Genes Immun*, 7(1):59–64, Jan 2006.

[108] Hara Levy, Benjamin A Raby, Stephen Lake, Kelan G Tantisira, David Kwiatkowski, Ross Lazarus, Edwin K Silverman, Brent Richter, Walter T Klimecki, Donata Vercelli, Fernando D Martinez, and Scott T Weiss. Association of defensin beta-1 gene polymorphisms with asthma. *J Allergy Clin Immunol*, 115(2):252–8, Feb 2005.

[109] Anna Lluis, Michaela Schedel, Jing Liu, Sabina Illi, Martin Depner, Erika von Mutius, Michael Kabesch, and Bianca Schaub. Asthma-associated polymorphisms in 17q21 influence cord blood ormdl3 and gsdma gene expression and il-17 secretion. *J Allergy Clin Immunol*, 127(6):1587–94.e6, Jun 2011.

[110] U. V. Luxburg, M Belkin, O Bousquet, and Pertinince A. A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 2007.

[111] Chang-Xing Ma, George Casella, and Rongling Wu. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics*, 161:1751–1762, 2002.

[112] Anne-Marie Madore, Stéphanie Perron, Véronique Turmel, Michel Laviolette, Elyse Y Bissonnette, and Catherine Laprise. Alveolar macrophages in allergic asthma: an expression signature characterized by heat shock protein pathways. *Hum Immunol*, 71(2):144–50, Feb 2010.

[113] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*, 363(2):166–76, Jul 2010.

[114] Andrea L Marat, Hatem Dokainish, and Peter S McPherson. Denn domain proteins: regulators of rab gtpases. *J Biol Chem*, 286(16):13791–800, Apr 2011.

[115] Ruth McPherson, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R. Cox, D.A. Hinds, L.A. Pennachio, A. Tybjaerg-Hansen, A.R. Folsom, E. Boerwinkle, H.H. Hobbs, and Jonathan C Cohen. A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316(5830):1488–91, Jun 2007.

[116] Reza Mobini, Bengt A Andersson, Jonas Erjefält, Mirjana Hahn-Zoric, Michael A Langston, Andy D Perkins, Lars Olaf Cardell, and Mikael Benson. A module-based analytical strategy to identify novel disease-associated genes shows an inhibitory role for interleukin 7 receptor in allergic inflammation. *BMC Syst Biol*, 3:19, 2009.

[117] Miriam F Moffatt, Ivo G Gut, Florence Demenais, David P Strachan, Emmanuelle Bouzigon, Simon Heath, Erika von Mutius, Martin Farrall, Mark Lathrop, William O C M Cookson, and GABRIEL Consortium. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*, 363(13):1211–21, Sep 2010.

[118] Wendy C Moore, Deborah A Meyers, Sally E Wenzel, W Gerald Teague, Huashi Li, Xingnan Li, Ralph D'Agostino, Jr, Mario Castro, Douglas Curran-Everett, Anne M Fitzpatrick, Benjamin Gaston, Nizar N Jarjour, Ronald Sorkness, William J Calhoun, Kian Fan Chung, Suzy A A Comhair, Raed A Dweik, Elliot Israel, Stephen P Peters, William W Busse, Serpil C Erzurum, Eugene R Bleecker, and National Heart, Lung, and Blood Institute's Severe Asthma Research Program. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med*, 181(4):315–23, Feb 2010.

[119] Amy Murphy, Jen-Hwa Chu, Mousheng Xu, Vincent J Carey, Ross Lazarus, Andy Liu, Stanley J Szefler, Robert Strunk, Karen Demuth, Mario Castro, Nadia N Hansel, Gregory B Diette, Becky M Vonakis, N Franklin Adkinson, Jr, Barbara J Klanderman, Jody Senter-Sylvia, John Ziniti, Christoph Lange, Tomi Pastinen, and Benjamin A Raby. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood cd4+ lymphocytes. *Hum Mol Genet*, 19(23):4745–57, Dec 2010.

[120] National Asthma Education and Prevention Program. Expert panel report 3 (epr-3): Guidelines for the diagnosis and management of asthma-summary report 2007. *J Allergy Clin Immunol*, 120(5 Suppl):S94–138, Nov 2007.

[121] M A Newton, C M Kendziorski, C S Richmond, F R Blattner, and K W Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol*, 8(1):37–52, 2001.

[122] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[123] Carole Ober and Tsung-Chieh Yao. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev*, 242(1):10–30, Jul 2011.

[124] Diabetes Genetics Initiative of Broad Institute of Harvard, MIT, Lund University, and Novartis Institutes of BioMedical Research et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6, Jun 2007.

[125] H Oswald, P D Phelan, A Lanigan, M Hibbert, J B Carlin, G Bowes, and A Olinsky. Childhood asthma and lung function in mid-adult life. *Pediatr Pulmonol*, 23(1):14–20, Jan 1997.

[126] L J Palmer, P R Burton, A L James, A W Musk, and W O Cookson. Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families. *Eur J Hum Genet*, 8(11):853–60, Nov 2000.

[127] L J Palmer, M W Knuiman, M L Divitini, P R Burton, A L James, H C Bartholomew, G Ryan, and A W Musk. Familial aggregation and heritability of adult lung function: results from the busselton health study. *Eur Respir J*, 17(4):696–702, Apr 2001.

[128] Orestis A Panagiotou, John P A Ioannidis, and Genome-Wide Significance Project. What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *Int J Epidemiol*, 41(1):273–86, Feb 2012.

[129] Andrew D Paterson, Daryl Waggott, Andrew P Boright, S Mohsen Hosseini, Enqing Shen, Marie-Pierre Sylvestre, I. Wong, B. Bharaj, P.A. Cleary, J.M. Lachin, J.E. Below, D. Nicolae, N.J. Cox, A.J. Canty, L. Sun, and Shelley B Bull. A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both a1c and glucose. *Diabetes*, 59(2):539–49, Feb 2010.

[130] I D Pavord, C E Brightling, G Woltmann, and A J Wardlaw. Non-eosinophilic corticosteroid unresponsive asthma. *Lancet*, 353(9171):2213–4, Jun 1999.

[131] J Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[132] D Pe'er, A Regev, G Elidan, and N Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1:S215–24, 2001.

[133] Peter D Phelan, Colin F Robertson, and Anthony Olinsky. The melbourne asthma study: 1964-1999. *J Allergy Clin Immunol*, 109(2):189–94, Feb 2002.

[134] María Pino-Yanes, Inmaculada Sánchez-Machín, José Cumplido, Javier Figueroa, María José Torres-Galván, Ruperto González, Almudena Corrales, Orlando Acosta-Fernández, José Carlos García-Robaina, Teresa Carrillo, Anselmo Sánchez-Palacios, Jesús Villar, Mariano Hernández, and Carlos Flores. Il-1 receptor-associated kinase 3 gene (irak3) variants associate with asthma in a replication study in the spanish population. *J Allergy Clin Immunol*, 129(2):573–5, 575.e1–10, Feb 2012.

[135] Laura J Rasmussen-Torvik, Alvaro Alonso, Man Li, Wen Kao, Anna Köttgen, Yu Yan, David Couper, Eric Boerwinkle, Suzette J Bielinski, and James S Pankow. Impact of repeated measures and sample selection on genome-wide association studies of fasting glucose. *Genet Epidemiol*, 34(7):665–73, Nov 2010.

[136] N Ressler. Computer-assisted diagnosis by a model-free system of direct data analysis. *Perspect Biol Med*, 19(1):101–17, 1975.

[137] Daphne C Richter, James R Joubert, Haylene Nell, Mace M Schuurmans, and Elvis M Irusen. Diagnostic value of post-bronchodilator pulmonary function testing to distinguish between stable, moderate to severe copd and asthma. *Int J Chron Obstruct Pulmon Dis*, 3(4):693–9, 2008.

[138] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, 1996.

[139] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, 1996.

[140] Stéphanie Romanet-Manent, D Charpin, A Magnan, A Lanteaume, D Vervloet, and EGEA Cooperative Group. Allergic vs nonallergic asthma: what makes the difference? *Allergy*, 57(7):607–13, Jul 2002.

[141] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[142] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–4, Jan 2004.

[143] Eric M Schauberger, Susan L Ewart, Syed H Arshad, Marianne Huebner, Wilfried Karmaus, John W Holloway, Karen H Friderici, Julie T Ziegler, Hongmei Zhang, Matthew J Rose-Zerilli, Sheila J Barton, Stephen T Holgate, Jeffrey R Kilpatrick, John B Harley, Stephane Lajoie-Kadoch, Isaac T W Harley, Qutayba Hamid, Ramesh J Kurukulaaratchy, Max A Seibold, Pedro C Avila, William Rodriguez-Cintrón, Jose R Rodriguez-Santana, Donglei Hu, Christopher Gignoux, Isabelle Romieu, Stephanie J London, Esteban G Burchard, Carl D Langefeld, and Marsha Wills-Karp. Identification of atpaf1 as a novel candidate gene for asthma in children. *J Allergy Clin Immunol*, 128(4):753–760.e11, Oct 2011.

[144] Malcolm R Sears, Justina M Greene, Andrew R Willan, Elizabeth M Wiecek, D Robin Taylor, Erin M Flannery, Jan O Cowan, G Peter Herbison, Phil A Silva, and Richie Poulton. A longitudinal, population-based, cohort study of childhood asthma followed to adulthood. *N Engl J Med*, 349(15):1414–22, Oct 2003.

[145] Seung Woo Shin, Tae Jeong Oh, Se-Min Park, Jong Sook Park, An Soo Jang, Sung Woo Park, Soo Taek Uh, Sungwhan An, and Choon-Sik Park. Asthma-predictive genetic markers in gene expression profiling of peripheral blood mononuclear cells. *Allergy Asthma Immunol Res*, 3(4):265–72, Oct 2011.

[146] David Siegmund and Benjamin Yakir. *The Statistics of Gene Mapping*. Springer, 2007.

[147] Patrick M A Sleiman, James Flory, Marcin Imielinski, Jonathan P Bradfield, Kiran Annaiah, Saffron A G Willis-Owen, Kai Wang, Nicholas M Rafaels, Sven Michel, Klaus Bonnelykke, Haitao Zhang, Cecilia E Kim, Edward C Frackelton, Joseph T Glessner, Cuiping Hou, F George Otieno, Erin Santa, Kelly Thomas, Ryan M Smith, Wendy R Glaberson, Maria Garris, Rosetta M Chiavacci, Terri H Beaty, Ingo Ruczinski, Jordan S Orange, Jordan M Orange, Julian Allen, Jonathan M Spergel, Robert Grundmeier, Rasika A Mathias, Jason D Christie, Erika von Mutius, William O C Cookson, Michael Kabesch, Miriam F Moffatt, Michael M Grunstein, Kathleen C Barnes, Marcella Devoto, Mark Magnusson, Hongzhe Li, Struan F A Grant, Hans Bisgaard, and Hakon Hakonarson. Variants of dennd1b associated with asthma in children. *N Engl J Med*, 362(1):36–44, Jan 2010.

[148] R. Sokal and R. Michener. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, 28, 1958.

[149] Nicole Soranzo, Tim D Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, Perttu Salo, Benjamin F Voight, Philippa Burns, Roman A Laskowski, Yali Xue, Stephan Menzel, David Altshuler, John R Bradley, Suzannah Bumpstead, Mary-Susan Burnett, Joseph Devaney, Angela Döring, Roberto Elosua, Stephen E Epstein, Wendy Erber, Mario Falchi, Stephen F Garner, Mohammed J R Ghori, Alison H Goodall, Rhian Gwilliam, Hakon H Hakonarson, Alistair S Hall, Naomi Hammond, Christian Hengstenberg, Thomas Illig, Inke R König, Christopher W Knouff, Ruth McPherson, Olle Melander, Vincent Mooser, Matthias Nauck, Markku S Nieminen, Christopher J O'Donnell, Leena Peltonen, Simon C Potter, Holger Prokisch, Daniel J Rader, Catherine M Rice, Robert Roberts, Veikko Salomaa, Jennifer Sambrook, Stefan Schreiber, Heribert Schunkert, Stephen M Schwartz, Jovana Serbanovic-Canic, Juha Sinisalo, David S Siscovick, Klaus Stark, Ida Surakka, Jonathan Stephens, John R Thompson, Uwe Völker, Henry Völzke, Nicholas A Watkins, George A Wells, H-Erich Wichmann, David A Van Heel, Chris Tyler-Smith, Swee Lay Thein, Sekar Kathiresan, Markus Perola, Muredach P Reilly, Alexandre F R Stewart, Jeanette Erdmann, Nilesh J Samani, Christa Meisinger, Andreas Greinacher, Panos Deloukas, Willem H Ouwehand, and Christian Gieger. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the haemgen consortium. *Nat Genet*, 41(11):1182–90, Nov 2009.

[150] Christine A Sorkness, Robert F Lemanske, Jr, David T Mauger, Susan J Boehmer, Vernon M Chinchilli, Fernando D Martinez, Robert C Strunk, Stanley J Szefler, Robert S Zeiger, Leonard B Bacharier, Gordon R Bloomberg, Ronina A Covar, Theresa W Guilbert, Gregory Heldt, Gary Larsen, Michael H Mellon, Wayne J Morgan, Mark H Moss, Joseph D Spahn, Lynn M Taussig, and Childhood Asthma Research and Education Network of the National Heart, Lung, and Blood Institute. Long-term comparison of 3 controller regimens for mild-moderate persistent childhood asthma: the pediatric asthma controller trial. *J Allergy Clin Immunol*, 119(1):64–72, Jan 2007.

[151] E M Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–17, Nov 1975.

[152] F J T Staal, M van der Burg, L F A Wessels, B H Barendregt, M R M Baert, C M M van den Burg, C van Huffel, A W Langerak, V H J van der Velden, M J T Reinders, and J J M van Dongen. Dna microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-b acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324–32, Jul 2003.

[153] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 3(479-498), 64.

[154] D P Strachan, H J Wong, and T D Spector. Concordance and interrelationship of atopic diseases and markers of allergic sensitization among adult female twins. *J Allergy Clin Immunol*, 108(6):901–7, Dec 2001.

[155] Robert C Strunk, Scott T Weiss, Katherine P Yates, James Tonascia, Robert S Zeiger, Stanley J Szefler, and for the CAMP Research Group. Mild to moderate asthma affects lung growth in children and adolescents. *J Allergy Clin Immunol*, 118(5):1040–7, Nov 2006.

[156] Lily S Subrata, Joelene Bizzintino, Emilie Mamessier, Anthony Bosco, Katherine L McKenna, Matthew E Wikström, Jack Goldblatt, Peter D Sly, Belinda J Hales, Wayne R Thomas, Ingrid A Laing, Peter N LeSouëf, and Patrick G Holt. Interactions between innate antiviral and atopic immunoinflammatory pathways precipitate and sustain asthma exacerbations in children. *J Immunol*, 183(4):2793–800, Aug 2009.

[157] Amanda Sutcliffe, Fay Hollins, Edith Gomez, Ruth Saunders, Camille Doe, Marcus Cooke, R A John Challiss, and Chris E Brightling. Increased nicotinamide adenine dinucleotide phosphate oxidase 4 expression mediates intrinsic airway smooth muscle hypercontractility in asthma. *Am J Respir Crit Care Med*, 185(3):267–74, Feb 2012.

[158] S J Szefler and D Y Leung. Glucocorticoid-resistant asthma: pathogenesis and clinical implications for management. *Eur Respir J*, 10(7):1640–7, Jul 1997.

[159] P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, E S Lander, and T R Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12, Mar 1999.

[160] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

[161] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1399–1320, 2005.

[162] R Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–85, Dec 1973.

[163] S F Thomsen, S van der Sluis, K O Kyvik, A Skytthe, and V Backer. Estimates of asthma heritability in a large twin sample. *Clin Exp Allergy*, 40(7):1054–61, Jul 2010.

[164] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, May 2002.

[165] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2), 2001.

[166] P Trayhurn. Northern blotting. *Proc Nutr Soc*, 55(1B):583–9, Mar 1996.

[167] Eleni Tsitsiou, Andrew E Williams, Sterghios A Moschos, Ketan Patel, Christos Rossios, Xiaoying Jiang, Oona-Delpuech Adams, Patricia Macedo, Richard Booton, David Gibeon, Kian Fan Chung, and Mark A Lindsay. Transcriptome analysis shows activation of circulating cd8+ t cells in patients with severe asthma. *J Allergy Clin Immunol*, 129(1):95–103, Jan 2012.

[168] Ucsc genome browser. http://genome.ucsc.edu/.

[169] W Vogt and D Nagel. Cluster analysis in diagnosis. *Clin Chem*, 38(2):182–98, Feb 1992.

[170] A Wagner and D A Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803–10, Sep 2001.

[171] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[172] Z. Wang and R. Wu. A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics. *Statistics in Medicine*, 23(19):3033–3051, 2004.

[173] Sally E Wenzel. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med*, 18(5):716–25, May 2012.

[174] Adriano V Werhli and Dirk Husmeier. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6:Article15, 2007.

[175] Gonneke Willemsen, Toos C E M van Beijsterveldt, Caroline G C M van Baal, Dirkje Postma, and Dorret I Boomsma. Heritability of self-reported asthma and allergy: a study in adult dutch twins, siblings and parents. *Twin Res Hum Genet*, 11(2):132–42, Apr 2008.

[176] Cristen J Willer, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, Simon C Heath, Nicholas J Timpson, Samer S Najjar, Heather M Stringham, James Strait, William L Duren, Andrea Maschio, Fabio Busonero, Antonella Mulas, Giuseppe Albai, Amy J Swift, Mario A Morken, Narisu Narisu, Derrick Bennett, Sarah Parish, Haiqing Shen, Pilar Galan, Pierre Meneton, Serge Hercberg, Diana Zelenika, Wei-Min Chen, Yun Li, Laura J Scott, Paul A Scheet, Jouko Sundvall, Richard M Watanabe, Ramaiah Nagaraja, Shah Ebrahim, Debbie A Lawlor, Yoav Ben-Shlomo, George Davey-Smith, Alan R Shuldiner, Rory Collins, Richard N Bergman, Manuela Uda, Jaakko Tuomilehto, Antonio Cao, Francis S Collins, Edward Lakatta, G Mark Lathrop, Michael Boehnke, David Schlessinger, Karen L Mohlke, and Gonçalo R Abecasis. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–9, Feb 2008.

[177] Prescott G Woodruff, Homer A Boushey, Gregory M Dolganov, Chris S Barker, Yee Hwa Yang, Samantha Donnelly, Almut Ellwanger, Sukhvinder S Sidhu, Trang P Dao-Pick, Carlos Pantoja, David J Erle, Keith R Yamamoto, and John V Fahy. Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. *Proc Natl Acad Sci U S A*, 104(40):15858–63, Oct 2007.

[178] Prescott G Woodruff, Barmak Modrek, David F Choy, Guiquan Jia, Alexander R Abbas, Almut Ellwanger, Laura L Koth, Joseph R Arron, and John V Fahy. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med*, 180(5):388–95, Sep 2009.

[179] H Wu, W P Yang, and C F Barbas, 3rd. Building zinc fingers by selection: toward a therapeutic application. *Proc Natl Acad Sci U S A*, 92(2):344–8, Jan 1995.

[180] Qun Wu, Di Jiang, Sean Smith, Jyoti Thaikoottathil, Richard J Martin, Russell P Bowler, and Hong Wei Chu. Il-13 dampens human airway epithelial innate immunity through induction of il-1 receptor-associated kinase m. *J Allergy Clin Immunol*, 129(3):825–833.e2, Mar 2012.

[181] R. Wu and M. Lin. Functional mapping how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*, 7(3):229–237, 2006.

[182] R. L. Wu, C.X. Ma, M. Lin, and G. Casella. A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics*, 166:1541–1551, 2004.

[183] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–21, Mar 2009.

[184] Zhi Xie, Shaohui Hu, Seth Blackshaw, Heng Zhu, and Jiang Qian. hpdi: a database of experimental human protein-dna interactions. *Bioinformatics*, 26(2):287–9, Jan 2010.

[185] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):187–193, 2012.

[186] Ching Yong Yick, Aeilko H Zwinderman, Peter W Kunst, Katrien Grünberg, Thais Mauad, Annemiek Dijkhuis, Elisabeth H Bel, Frank Baas, René Lutter, and Peter J Sterk. Transcriptome sequencing (rna-seq) of human endobronchial biopsies: asthma versus controls. *Eur Respir J*, 42(3):662–70, Sep 2013.

[187] Lama A Youssef, Mark Schuyler, Laura Gilmartin, Gavin Pickett, Julie D J Bard, Christy A Tarleton, Tereassa Archibeque, Clifford Qualls, Bridget S Wilson, and Janet M Oliver. Histamine release from the basophils of control and asthmatic subjects and a comparison of gene expression between "releaser" and "nonreleaser" basophils. *J Immunol*, 178(7):4584–94, Apr 2007.

[188] Ming Yuan and Christina Kendziorski. A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics*, 62(4):1089–98, Dec 2006.

[189] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4:Article17, 2005.

[190] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4, 2005.

[191] P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[192] Wensheng Zhu, Kelly Cho, Xiang Chen, Meizhuo Zhang, Minghui Wang, and Heping Zhang. A genome-wide association analysis of framingham heart study longitudinal data using multivariate adaptive splines. *BMC Proc*, 3 Suppl 7:S119, 2009.

[193] Emmanuel Zorn, David B Miklos, Blair H Floyd, Alex Mattes-Ritz, Luxuan Guo, Robert J Soiffer, Joseph H Antin, and Jerome Ritz. Minor histocompatibility antigen dby elicits a coordinated b and t cell response after allogeneic stem cell transplantation. *J Exp Med*, 199(8):1133–42, Apr 2004.

[194] Min Zou and Suzanne D Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–9, Jan 2005.