12-14-2015

Genetic and Epigenetic Variations in Asthma and Wheeze Illnesses

Todd M. Everson
University of South Carolina - Columbia

GENETIC AND EPIGENETIC VARIATIONS IN ASTHMA AND WHEEZE ILLNESSES

by

Todd M. Everson

Bachelor of Science
Colorado State University, 2006

Master of Public Health
Oregon Health and Science University, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Epidemiology

The Norman J. Arnold School of Public Health

University of South Carolina

2015

Accepted by:

Melinda Forthofer, Major Professor

Hongmei Zhang, Committee Member

John Holloway, Committee Member

Wilfried Karmaus, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

I gratefully acknowledge the guidance I received from my dissertation committee. I am sincerely grateful to my committee chair, Dr. Melinda Forthofer, whose unwavering support guided me through this process and encouraged me to frame my research within the broader context of public health. I thank Dr. John Holloway, Dr. Hongmei Zhang and Dr. Wilfried Karmaus for fostering my interests in epigenetics and for their assistance in exploring unique statistical and methodological approaches for answering difficult epidemiologic questions. Above all, I am truly appreciative of my entire committee for their mentorship and their patience throughout this dissertation. I also would like to thank the Isle of Wight research team, including S. Hasan Arshad, Ramesh Kurukulaaratchy, Susan Ewart, Veeresh Patil, and Gabrielle Lockett, for all of the hard work they did in developing the cohort, generating the data, and contributing to discussions about my research.

I would also like to acknowledge the funding that made this research possible. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z and MRC Hub grant G0900747 91070) for the generation of the methylation data. The research reported in this work was supported by the National Institute of Allergy and Infectious Diseases under award number R01 AI091905 (PI: Wilfried Karmaus) and R01 AI061471 (PI: Susan Ewart); the 18-year follow-up was supported by a grant from the National Heart and Blood Institute (R01 HL082925, PI: S. Hasan Arshad).

ABSTRACT

Asthma, a chronic respiratory condition, is common worldwide with no cure and limited effective prevention strategies. It is well recognized that asthma has a multifaceted etiology, though many of the underlying mechanisms involved in asthma development, persistence and remission are still convoluted. Epigenetic mechanisms, such as DNA methylation, regulate gene-expression but are not related to changes in the actual DNA sequence. Recently, differential patterns of DNA methylation within many genes have been associated with asthma, particularly within genes involved in the differentiation of pro-inflammatory T-helper 2 (Th2) cells. DNA methylation patterns within less known biologic pathways undoubtedly are involved in asthma pathogenesis as well. The purpose of this dissertation was three-fold. First, we explored whether genetic and epigenetic variations within Th2-genes differed among persons with different phenotypic presentations of wheeze illnesses. Second, we conducted an epigenome-wide association study (EWAS) to identify novel DNA methylation loci associated with asthma. Last, we conducted a follow-up study of our top EWAS findings, to investigate whether the expression of the associated genes were predictive of infant wheeze.

We found that DNA-M within *GATA3* and *IL4* varied based on different wheeze-illness phenotypes, suggesting that Th2-genes are under differential epigenetic regulation for different presentations of asthma. We also identified nine novel DNA methylation loci (cg25578728 in *CHD7*, cg16658191 in *HK1*, cg00100703 in *UNC45B*, cg07948085

[intergenic], cg04359558 in *LITAF*, cg20417424 in *ST6GALNAC5*, cg19974715 [intergenic], cg01046943 in *NUP210* and cg14727512 in *DGCR14*) associated with asthma at age 18. For two of those genes (*HK1* and *LITAF*), expression levels in cord blood were predictive of infant wheeze. Interestingly, the observed methylation and expression patterns of *HK1* and *LITAF* could be consistent with increased resistance to apoptotic signaling. Apoptotic-resistance among pro-inflammatory cells can increase the duration of an inflammatory response and is affiliated with asthmatic pathophysiology. Thus we may have identified under-studied genes and their epigenetic regulation, which could play important roles in asthma pathophysiology. These genes may offer new insights into the etiology of asthma, be investigated as potential targets for therapy, or be considered for inclusion in algorithms used to predict early-life wheeze and later-life asthma.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Symbols

β	Value of a parameter estimate from a regression model.
α	Significance level used to determine statistically significant findings.
χ^2	Denoting that the associated test-statistic or p-value came from a Chi squared test.

List of Abbreviations

BDR ..... bronchodilator reversibility

BHR ..... bronchial hyper-responsiveness

CpG ..... cytosine-phosphate-guanine

DNA-M ..... DNA methylation

EWAS ..... epigenome-wide association study

FDR ..... Multidrug Resistance

FeNO ..... fractional exhaled nitric oxide

FEV1 ..... forced expiratory volume in one second

FVC ..... forced vital capacity

GWAS ..... genome-wide association study

Ig ..... immunoglobulin

IOW ..... Isle of Wight

LD ..... linkage disequilibrium

MAF ..... minor allele frequency

RF ..... random forest

SNP ..... single nucleotide polymorphism

SPT ..... skin prick test

Th ..... T-helper cell

TSS ..... transcription start site

UTR ..... untranslated region

VIM ..... variable importance measure

CHAPTER 1

BACKGROUND ON ASTHMA AND EPIGENETICS

1.1 Asthma Background:

Asthma is a chronic respiratory disease often defined by reversible airway obstruction, wheeze, bronchial hyper-responsiveness, and inflammation[1]. Many molecular pathways influence the recruitment and the activity of different immune cells during asthmatic symptoms, thus leading to heterogeneity in what triggers the symptoms to occur as well as their severity and responsiveness to treatments[1]. Given the vast heterogeneity in asthma, many possible phenotypes have been described. In fact, it is unclear whether asthma is a single disease with multiple presentations, or instead several different diseases, all with the common symptom of reversible restriction of the airways[2].

Some characteristics allow for the distinction of common asthma sub-types. For instance, those whose asthma symptoms are triggered by allergens (ie. pollen, cockroach, mold, and dust mite) are often distinguished from those whose symptoms arise with no apparent allergic triggers as "extrinsic" vs. "intrinsic" types[1]. Interestingly, there is also some overlap in triggers for both extrinsic and intrinsic asthma. Both may be symptomatic to exercise, cold air, and inhaled irritants[3], suggesting some overlap in common underlying mechanisms leading to exacerbations. Intrinsic asthma appears to be more severe, as positive skin prick tests (SPTs) are less common in severe asthma compared to less-severe asthma and extrinsic asthma tends to have an earlier onset than

intrinsic[3] but is more transient and responds better to treatment with corticosteroids[1]. The exact distribution of extrinsic vs intrinsic asthma is unclear, although the majority of asthmatics tend have the extrinsic characteristics of allergic sensitization such as high serum immunoglobulin (Ig)E and/or positive SPT results[3].

Aside from differences in triggers and allergic hypersensitivity characterized by extrinsic vs. intrinsic phenotypes, asthmatics also may differ in many other characteristics important in understanding disease progression such as age of symptomatic onset, comorbidity with other allergic diseases (eczema and/or rhinitis), overall lung function, responsiveness to inhaled corticosteroids (ICS) and other treatments, as well as markers of airway eosinophilia such as fractional-exhaled nitric oxide (FeNO)[4]. These measures inform us about the pathophysiology of asthma, but there is also heterogeneity in the morbidity associated with asthma, such as frequency and severity of wheeze attacks, as well as when wheeze disturbs sleep, limits speech, or limits exercise[5]. Thus, it is of great importance to identify risk factors for the development, persistence, remission, and relapse of different asthmatic phenotypes[6].

Asthma Epidemiology:

The prevalence of asthma varies by many demographic characteristics: age, sex, race, and socio-economic status (SES). The Centers for Disease Control and Prevention (CDC) showed that between 2008 and 2010, the average annual prevalence of asthma in the US tended to be higher in children than in adults, higher in females than males, higher in blacks compared to other races, higher among Hispanics compared to non-Hispanic, and higher among those in lower socio-economic groups[7]. The sex difference in asthma is age-dependent. Males typically have higher incidence and prevalence from birth through

puberty, then tend to exhibit higher rates of remission, whereas females tend to have higher incidence and prevalence as young adults[6]. Asthma severity parallels the same sex-age-dependent pattern with males having more severe asthma in childhood and females having more severe asthma in adulthood[6].

Asthma, as well as other allergic diseases, have been increasing in prevalence worldwide for decades and are some of the most common sources of chronic health issues[8,9], the Global Asthma Network estimates that as many as 334 million people are affected by asthma[10]. Despite increasing global prevalence, region specific trends indicate that asthma prevalence in many high-income countries has recently peaked, whereas prevalence in low- or middle-income countries is still on the rise[11]. Despite the recent decline in incidence, mortality, and health care utilization, the prevalence of asthma symptoms remains high in high-income countries, ranging from 8.6-14.4%[11]. A recent study in the UK found that new diagnoses of asthma have plateaued, but that the lifetime-prevalence is still increasing[12]. In the US, the prevalence of asthma, which was only 3% in the 1980s, steadily grew to 8.4% by 2010[7]. The American Lung Association reports that approximately 25.9 million people in the US had asthma in 2011 and an estimated 13.2 million (51% of those with asthma) had at least one asthma attack within the 12 months prior to the survey[13]. Asthma also generates a substantial burden on health-care systems and was responsible for 10.6 million physician office visits and 2.1 million emergency room visits in 2010 in the US alone[13].

The prevalence of asthma symptoms was highly variable in low- or middle-income regions, but in many cases was higher than that of high-income countries, ranging from 6.8-21.7%[11]. In fact, the majority of persons affected by asthma are living in low- or

middle- income countries[10]. The increasing incidence of asthma in low- or middle-income countries appears to be related to urbanization, which is expected to continue to expand in most of these countries[11]. So it is possible that the increasing incidence of asthma in low- and middle-income countries will actually accelerate in the years to come, further increasing the already heavy global burden of asthma. High prevalence worldwide and increasing prevalence in low- and middle- income countries means that millions of people will continue to chronically medicate in order to prevent or relieve asthma symptoms since there is no cure for this disease[14].

Asthma Etiology:

Due to the increasing prevalence worldwide, research into asthmatic risk factors and underlying mechanisms of disease received substantial focus. Yet, the etiology of asthma is only partially understood, likely due to the multiple environmental and genetic risk factors implicated in its pathogenesis and its complex symptomatic presentation[2,15]. Although asthma is frequently defined by reversible airway obstruction, wheeze, bronchial hyper-responsiveness, and/or inflammation, individual cases of asthma may exhibit these characteristics and with varying frequency and severity[2]. Asthma has a large heritable component[16]; yet, known genetic variations and heritable risk factors only account for a small proportion of actual cases[15]. Also, identical twins are more likely to both be asthmatic than fraternal twins, yet most identical twins with an asthmatic co-twin are not affected by asthma, suggesting both genetic and non-genetic etiologies. Twin studies have also shown that approximately one-third of the variation in age at onset of asthma is due to genetic factors while the remaining two thirds are likely due to environmental factors[17]. Thus further studies are necessary to reveal key etiologic

elements in the pathogenesis of asthma, or its various phenotypes, which could then become the targets for prevention or treatment.

Though it is well recognized that asthma has a multifaceted etiology, many of the underlying mechanisms associated with asthma risk factors are still convoluted. Indeed, many environmental and genetic components[6,18] have been implicated as risk factors for, or associated with, asthma. However, the only confirmed "cause" of the "underlying asthma trait", as opposed to triggers of asthmatic symptoms, is exposure to tobacco smoke, prenatally, in childhood or in adulthood[14]. Studies of other possible "causes" have produced conflicting evidence about their roles in the development of an underlying asthma trait.

Potential environmental "causes" of asthma include pet ownership, living on a farm, mold in homes, and antibiotic or paracetamol exposure in early childhood. Persons with furry pets tend to have lower rates of asthma, though at least some of this association appears to be driven by selective avoidance of pet-ownership by persons with allergies[19,20]. Living on a farm has been observed as protective against the development of asthma and other allergic diseases, possibly due to greater frequency and diversity of microbial exposure leading to the development of a robust innate immune system ; however, this is a hypothesized mechanism, the details of which remain unclear[21]. Mold growth is more common in the homes of asthmatic children[22], but most of these children do not exhibit an allergic response to fungal molds, thus obscuring how this mechanism could work[10]. Children that were exposed to antibiotics[23] or paracetamol[24] very early in life are more likely to develop asthma, but this association may be a product of reverse causation[23]. Both antibiotics and paracetamol are often administered to infants as

treatment for wheeze and infants with recurrent wheeze may already be pre-disposed to developing asthma later in life[23,25]. Thus many of the environmental exposures that are recognized as risk factors for asthma also have many questions surrounding whether or not they are causally related to asthma.

Aside from possible environmental contributions to asthma risk, as many as 100 genes have been implicated in asthma etiology[6,18]. Many of these were identified using genome-wide association studies (GWAS), which interrogate a large number potential genetic risk factors with no prior knowledge about their possible relationships with asthma[26]. Replication of these associations has been inconsistent and largely unsuccessful[18,26,27]. A likely reason for the lack of consistent replication among GWAS is that many of these studies were conducted within different populations which had different rates of genetic variation at identified loci[6]. It is also likely that poor replication could be due to lack of *penetrance*, in which some genetic causes may be necessary but not sufficient risk factors for asthma[27]. Similarly, a genetic risk factor may only influence the development of asthma under the correct environmental or epigenetic conditions as in gene-environment and gene-epigenetic interactions[18,26].

1.2 Epigenetics and DNA methylation:

Epigenetic mechanisms, consisting of DNA methylation (DNA-M), histone modifications, and microRNAs (miRNA), are processes which influence gene-expression that are independent of variations to the genetic sequence. The majority of epidemiologic research into epigenetic mechanisms of disease has focused on DNA-M, the covalent addition of a methyl group to a carbon at the 5-position of a cytosine residue that is followed by a guanine residue (CpG site) within the DNA, due its stability and

technological advancement allowing for fairly rapid, accurate and affordable measurements[28]. Many of these studies utilize the Illumina 450K HumanMethylation array[29] because of its high-accuracy, expansive interrogation, and affordability, though it is limited by biased measurement of CpG sites within promoter regions and CpG islands, and that despite having high-throughput scale, it still only measures < 2% of the 28 million total CpG sites in human DNA[30]. Thus, studies with this array may be missing key elements of the DNA-M profile that may still be important in the diseases under study; however, this is currently the most appropriate array for conducting epidemiologic research of DNA-M across the genome[30]. As new technology becomes more affordable and more accurate, the field may switch toward sequencing technologies that measure methylation at more, or even all, CpG sites.

Most early studies of DNA-M considered it to function as a silencer of gene expression, potentially through the inhibition of transcription factor (TF) binding[31]. Since its discovery as a regulator of expression, investigators have uncovered multiple effects of DNA-M on expression depending on where the methylation occurs within the gene[32]: (1) direct gene-silencing via promoter methylation, (2) interacting with or blocking DNA-binding proteins, (3) repression of intragenic repetitive elements, and (4) influencing alternative splicing via differential methylation at intron-exon boundaries. Complicating the epigenetic landscape are the interactions between epigenetic mechanisms, and the interactions between genetic and epigenetic variation. For instance, methylation at particular sites can lead to histone modifications and vice versa[33]. Also, although epigenetic modifications do not alter the genomic sequence, they can be influenced by genetic variation and can be induced by environmental stimuli[18] making for complex

multi-factor interactions. These interactions between genetic variation and levels of DNA methylation are referred to as methylation quantitative trait loci (methQTLs)[34–36]. Despite these complex interacting mechanisms, DNA-M has been shown to have independently important role in the regulation of gene expression and thus likely plays a role in many complex diseases.

Role for Epigenetics in Asthma Etiology, a brief review:

Epigenetic mechanisms are known to play major roles in cellular differentiation and immune cell activation[33]. DNA methylation biomarkers have been implicated in airway inflammation[37], immune cell fate, and asthma[33]. The majority of epidemiologic research in this area has focused on candidate regions of the genome known to be important in asthma pathogenesis, such as genes driving $CD4^{+}$T cells toward Th2 rather than Th1 phenotypes, an important polarization that occurs in asthmatics[8]. Such research has shown that the loss of DNA-M, accompanied by histone modifications at Th2 cytokine loci (*IL4*, *IL5*, and *IL13*), promote a Th2 response whereas the presence of DNA-M within the *IL4* locus has been shown to promote a Th1 response[38]. Statistical interactions between genetic variants and DNA-M within the genes for the IL4-receptor (*IL4R*) and GATA Binding Protein 3 (*GATA3*), are also involved in Th2 polarization, contribute to asthma risk in a potentially age-dependent pattern[35,39]. The majority of strong associations between DNA-M and asthma have been observed within these and other genes involved in T-cell differentiation[40], though investigators have also found associations related to specific characteristics of asthma such as FeNO and bronchodilator responsiveness. Higher levels of FeNO were associated with lower levels of DNA-M the arginase genes (*ARG1* and *ARG2*)[41], *IL6*, and *iNOS*[37] implicating that

DNA-M plays an important role in the production of nitric oxide, which is an important marker of airway inflammation and eosinophilic asthma. Others have shown that higher levels of DNA-M within the promoter of the adrenoceptor β2, Surface (*ADRB2*), which can induce relaxation of smooth muscle in the airways, have been associated with decreased severity of dyspnoea and trended towards improved spirometry measures[42]. Taken together, these findings indicate that DNA-M plays many roles in regulating different biological pathways that are important in asthma etiology.

1.3 Objectives:

Purpose of the dissertation:

The overarching goal of this dissertation was to identify novel genetic and epigenetic loci associated with asthma and complex wheeze phenotypes, and explore whether a novel set of genes associated with young adult wheeze could also be predictive of infant wheeze. We attempted to do this by conducting a genome-wide DNA-M study of physician-diagnosed asthma, a cross-sectional study of Th2-path genetic and epigenetic variations among complex wheeze phenotypes, and a prospective study of the expression and DNA-M of a novel gene-set and infant wheeze.

Specific Aims:

1. Aim 1: To identify CpG loci associated with prevalent asthma at age 18 from a genome-wide DNA methylation dataset with a two-stage design.

    Research Question 1.1: What parameters in the recursive Random Forest (RF) algorithm need to be changed from their default values?

    Research Question 1.2: Stage 1 – What epigenetic loci are selected via recursive RF feature selection?

Research Question 1.3: Stage 1 – Was the site selection from recursive RF feature selection confounded by cell-type proportions?

Research Question 1.4: Stage 2 – Which sites selected from Stage 1 can be corroborated with multivariable logistic regression for their associations with asthma status?

Research Question 1.5: Stage 2 – Are these sites associated with markers of lung dysfunction or allergic inflammation of the lung?

Research Question 1.6: What are the biological functions of the genes that the selected CpG sites are in?

2. AIM 2: To explore whether the combined genetic and epigenetic variation within an asthma-associated pathway (ie. the Th2 path) can improve our classification of asthmatic phenotypes at age 18.

    Research Question 2.1: Are specific genetic and epigenetic variations in the selected path able to classify different asthma phenotypes?

3. AIM 3: To identify whether the gene-set associated with our top DNA-M findings from Aim 1 are differentially expressed in cord blood in relation to infant wheeze.

    Research Question 3.1: How are gene expression levels related within the selected gene-set?

    Research Question 3.2: Which genes within the gene-set are differentially expressed in cord blood samples in relation to wheeze within the one year of birth?

    Research Question 3.3: Which genes within the selected gene-set are correlated (positively or negatively) with DNA-M levels?

CHAPTER 2

GENOME-WIDE DNA METHYLATION ASSOCIATION STUDY OF ASTHMA

2.1 Introduction:

Asthma is a common chronic respiratory disease affecting around 334 million people worldwide[13], causing substantial health care costs and morbidity-related absenteeism[14]. Over the previous two decades, many studies have sought to characterize the underlying mechanisms leading to the development and persistence of asthma. Yet, the etiology of asthma remains only partially understood, likely due to its complex physiology and the multiple environmental and genetic risk factors implicated in its pathogenesis[2,14].

Asthma is defined by multiple symptoms and characteristics such as airflow obstruction, bronchial hyper-responsiveness and airway inflammation. However, individual cases of asthma may exhibit only some of these characteristics and with varying frequencies and severities. This dynamic leads some to hypothesize that asthma is a diagnosis that encompasses multiple diseases, each of which may have its own unique etiology and pathophysiology[2,38]. Also, asthma has a substantial heritable component[14,16], known genetic variations and heritable risk factors only account for a small proportion of actual cases[15].

Previous evidence suggests that asthma has no individual cause[27]; indeed, many environmental risk factors and as many as 100 genes[18] have been implicated in studies of asthma. Epigenetic mechanisms, which control heritable variations in gene-

expression not related to changes in genomic sequence[32], have received recent attention in studies of asthma because these mechanisms are considered to be heritable and can be altered via environmental exposures, particularly exposures that occur *in utero*[14,38,43]. One of the most thoroughly studied epigenetic mechanisms is DNA methylation (DNA-M), which is the covalent addition of a methyl group to the DNA at a cytosine residue that is followed guanine (CpG site); this acts as an important regulator of gene transcription and may influence alternative splicing[32].

Recent work has shown that epigenetic mechanisms are important in regulating the expression of pro- and anti-inflammatory cytokines, which drive T-cell differentiation as well as the severity and duration of an inflammatory response[44]. Also a number of association studies have found variations in DNA-M associated with asthma status[40], lung function[45], and nitric oxide synthesis[41]. These studies have provided strong evidence that epigenetic mechanisms play key roles in the biological processes that result in asthmatic predisposition and the onset of symptoms. Thus further studies are necessary to reveal key epigenetic elements in the pathogenesis and persistence of asthma, or its various phenotypes, which can then become the targets for prevention or treatment.

We conducted an exploratory study aimed at identifying novel DNA-M markers, from a genome-wide screening, which could be effective classifiers of current prevalent asthma and which may reveal some of the underlying role that DNA-M plays in the pathogenesis or persistence of asthma. Because asthma is heterogeneous condition, we were also interested in learning whether some these markers were more strongly associated with asthma among those with and without allergic sensitization, and for

specific phenotypic characteristics of asthma such as airway obstruction (FEV1/FVC Ratio), bronchodilator reversibility (BDR) and fraction of exhaled nitric oxide (FeNO).

2.2 Methods:

The Isle of Wight birth cohort

The Isle of Wight (IOW) birth cohort was established to study the natural history of asthma and allergies in children born between January 1, 1989 and February 28, 1990 in Isle of Wight, UK. The study was approved by the local research ethics committee (now named the National Research Ethics Service, NRES Committee South Central – Southampton B, 06/Q1701/34) and written informed consent was provided by the infants' parents. Details about the birth cohort have been described in detail elsewhere[46,47]. After exclusion of adoptions, prenatal deaths and refusals, 1,456 children were enrolled, and followed-up at 1 (n=1,167; 80.2%), 2 (n=1,174; 80.6%), 4 (n=1,218; 83.7%), 10 (n=1,373; 94.3%), and 18 (n=1,313; 90.2%) years of age. At each follow-up, participants were administered detailed questionnaires and evaluated for manifestations of allergic disease. Questionnaires included the International Study of Asthma and Allergies in Childhood (ISAAC)[5] as well as study-specific questions about allergic disease and relevant risk factors. At the 18 year follow-up, questionnaire data was obtained via in-person interview (n=864; 66%), by telephone (n=421; 32%), or by mail (n=28; 2%). Most of those who attended the follow-up visit in-person were also assessed via spirometry, BDR, FeNO, and skin prick tests (SPTs). At this 18-year follow-up a random subset of female (n=245) and male (n=125) participants were selected to take part in epigenetic screening.

Dependent Variables

The primary dependent variable for this study was dichotomous prevalent asthma status (asthma vs. no asthma), determined by questionnaire. Participants were determined to have asthma if they had an affirmative answer to "Have you ever had asthma?", as well as an affirmative response to either "Have you had wheezing in the last 12 months?" or "Have you had asthma treatment in the last 12 months?"

Continuous measures of lung volume and airway obstruction were assessed via spirometry at age 10 and 18 years. Lung function measurements were performed using a Koko Spirometer and software with a desktop portable device (PDS Instrumentation, Louisville, USA), according to American Thoracic Society guidelines [48,49]. Prior to lung function measurements children were required to be free from respiratory infection for 14 days and to not be taking oral steroids. In addition, they were required to abstain from any beta-agonist medication for 6 hours and from caffeine intake for at least 4 hours. Forced expiratory volume in one second (FEV_1) measured the volume of air (in liters) exhaled over the first second of a forced expiration done with maximal effort. Forced vital capacity (FVC) was the total volume of air that could be expired after full inspiration. The FEV_1/FVC ratio was calculated by dividing FEV_1 by FVC, and represents the proportion of the vital capacity that an individual can expire over one second, given maximal effort. BDR measured the percent change in FEV_1, taken before and after administration of 600 ng/ml salbutamol, which is a bronchodilator. Percent change for BDR was calculated via pre-bronchodilator FEV_1 minus post-bronchodilator FEV_1 divided by the pre-bronchodilator FEV_1 multiplied by 100.

Fraction of exhaled nitric oxide (FeNO) is a marker of airway inflammation. FeNO measurements (Niox mino, Aerocrine AB, Solna, Sweden) were obtained prior to spirometric assessments and in accordance with American Thoracic Society (ATS) guidelines. Expiratory flow against resistance was maintained at 50 ml/s to avoid contamination of the airways. Participants with a current infection, as well as those with asthmatic symptoms or treatment (with antibiotics or oral steroids) within the previous two weeks, were rescheduled for a later date. Because the distribution of FeNO was heavily positively skewed, all parametric analyses were performed with log-transformed FeNO, which better approximated a normal distribution.

Atopy status, was assessed by skin prick tests (SPT) administered via a standard method[50] with a battery of 11 allergens. Food allergens tested were cows' milk, hens' egg, peanut and cod. Inhalant allergens tested were house dust mite, cat, dog, *Alternaria alternata*, *Cladosporium herbarium*, grass pollen mix, and tree pollen mix. Histamine and saline acted as positive and negative controls, respectively (Alk-Abello, Horsholm, Denmark). Positive SPTs were defined as having a mean wheal diameter of 3 mm greater than the negative control; if the positive control yielded a diameter less than 3mm, the test was deemed inconclusive. Participants with at least one positive SPT were defined as atopic, while those with no positive SPTs were defined as not having atopy.

Data Collection and DNA Methylation

Blood samples for epigenetic screening were collected at the 18 year follow-up, DNA was extracted from whole blood using a standard salting out procedure [51]. DNA concentration was determined by the PicoGreen dsDNA quantitation kit (Molecular Probes, Inc., OR, USA). One microgram of DNA was bisulfite-treated for cytosine to

thymine conversion using the EZ 96-DNA methylation kit (Zymo Research, CA, USA), following the manufacturer's standard protocol. Genome-wide DNA methylation was assessed using the Illumina Infinium HumanMethylation450K BeadChip (Illumina, Inc., CA, USA), which interrogates >484,000 CpG sites, regions of DNA where a cytosine nucleotides are followed by a guanine nucleotide, associated with approximately 24,000 genes. The BeadChips were scanned using a BeadStation, and the methylation levels (β value, described below) were calculated for each queried CpG locus using the Methylation Module of BeadStudio software. Arrays were processed using a standard protocol as described elsewhere[52], with multiple identical control samples assigned to each bisulphite conversion batch to assess assay variability and samples were randomly distributed on microarrays to control against batch effects.

Data Cleaning

The program for data cleaning was written in R (R Development Core Team, 2012). Quality control (QC) measures were employed to improve the reliability of data prior to analysis. In our study, the detection *P*-value reported by BeadStudio (Illumina software to process raw intensities) was used as a QC measure of probe performance. Probes whose detection *P*-values > 0.01 in >10% of the samples were removed[53]. The methylation data were then preprocessed and technical variations removed via peak-correction using the Bioconductor IMA (Illumina methylation analyser) package[54]. Dropping control probes and probes with poor detection *P*-values yielded 383,998 remaining probes; since males and females were studied together, CpGs on sex-specific (X and Y) chromosomes were dropped. The arrays were processed in three different batches; batch number was recorded as a categorical variable which was used in ComBat

to adjust for inter-array variation[55]. A very conservative approach was utilized for addressing systematically biased probes. We also excluded all probes with SNPs that had a minor allele frequency (MAF) > 1.0% in European populations (or any population if European-specific MAF was not available) and within 10 nucleotides of, or directly at, single base extension via dbSNP137[56], resulting in a final set of 248,336 CpGs for analysis. Dropping this large set of potentially biased probes was necessary because the variable selection method depends on a conditional variable importance metric, which means that the selection of biased probes could adversely affect the selection of other-unbiased probes.

Methylation levels for each queried CpG were calculated as beta (β) values. These represent the proportions of methylated probes for each specific CpG site and can be interpreted as percent methylation. The β values close to 0 or 1 tend to suffer from severe heteroscedasticity[57]. The β values were utilized for RF, described below, which is a non-parametric method and does not assume a normal distribution. However, for parametric statistical analyses, we utilized M-values which address the issue of heteroscedasticity and were calculated via $\log_2(\beta / (1-\beta))$[57]. Because M-values expand the distribution of the methylation levels, they may emphasize outliers; thus when using M-values, strong outliers were recoded as missing. Potential outliers were identified using adjusted boxplots and a coefficient of 2.5 via the robustbase package in R.

Random Sampling: Stage-1 vs Stage-2

The epigenetic sample was randomly divided into two independent sub-samples. To create the stage-1 sample, 25% of participants with asthma and 25% of participants without asthma were randomly selected from the epigenetic sample (N=370), yielding a

discovery data set (n_{S1}=91). The stage-2 validation data set (n_{S2}=279) included all samples not selected into the discovery data set. To improve the likelihood that findings selected from stage-1 could be validated in stage-2, we compared the prevalence of asthma and gender, as well as the mean values for airway obstruction and inflammation across the two sub-samples. Differences in prevalence of asthma or gender, between the stage-1 and stage-2 samples, were assessed with Pearson's Chi-squared tests. Students T-tests were used to determine whether the mean values of continuous variables differed between the stage-1 and stage-2 samples. The specific analyses utilized within each stage are detailed below.

Stage 1 (n_{S1}=91) – Recursive RF feature Selection

For stage-1 we conducted recursive RF feature selection [58,59] described below. RF is a non-parametric machine learning technique that can deal with substantially large numbers of predictors (p) relative to the number of samples (n), it is robust to outliers and noise, and it naturally incorporates conditional effects into the model via recursive binary partitioning without needing to specify interactions *a priori* [60]. The RF algorithm has been thoroughly described elsewhere[60,61]. Briefly, this algorithm produces a series of classification trees, grown from bootstrapped training samples with replacement; each participant's outcome status is predicted from the aggregate of all trees in which they were not part of the training sample. We used 'balanced sampling', by specifying by the *sampsize* parameter, to draw the same number of observations from the minority class and majority class for each bootstrap sample, so that each individual classification tree was grown from a balanced sub-sample[58]. We also altered the parameters for the number of predictors to test per node (*mtry* = 10% of all predictors available) and the number of

trees to grow in the forest (*ntree* = 7500) due to sparsity in the data and to produce stable VIMs, respectively.

We effectively implemented the RF algorithm recursively as a means of reducing the size of the data[62]. For the recursive RF feature selection we (1) ran the RF algorithm on all available predictors via the randomForest package in R[63], (2) extracted out-of-bag (OOB) misclassification rates and the variable importance measures, (3) sorted the predictors by their variable importance measures (VIMs), (4) excluded half of the predictors with the smallest VIMs, and (5) repeated the sequence until the stop criteria was met. Our stop criterion was a leveling-off of the asthma-specific misclassification rate; this would indicate that the predictors contributing noise to the RF classifications had been effectively excluded and the remaining predictors offered some ability to distinguish between asthmatics and non-asthmatics. Predictors that remained once the stop criteria were met were then annotated with relevant genetic information and analyzed for significant associations with asthma in the stage-2 validation sample.

Stage 2 (n_{S2}=279) – Validation Tests in Independent Sample

The methylation levels for the selected CpGs were then converted to M-values as previously described, then tested for crude association with asthma status via logistic regression. The p-value distribution was then investigated via histogram, to determine whether small p-values from the validation tests were likely due to random chance. We generated q-values to estimate the false discovery rate (FDR)[64] using the qvalue package in R[65]. Sites from the stage-2 analyses that were within a 10% FDR were then selected as our positive findings. We then validated that the direction (positive vs. negative) of the association, between methylation levels and asthma, was the same by plotting the

parameter estimates for the selected sites for both stage-1 and stage-2 logistic regression models. Any sites that did not show the same direction of association were excluded from our positive findings.

Stage 3 (N=370) – Characterization of Findings

Post-hoc analyses were then conducted on the positive findings from stage-2 analyses, in the full sample. We assessed whether the odds of higher average methylation was greater among asthmatics compared to non-asthmatics using logistic regression. Potential confounders (sex and cell-type proportions from blood samples) were determined *a priori*. Due to logistical constraints, complete blood counts (CBC) were not possible for our participants, thus we predicted cell type proportions of CD8T cells, CD4T cells, natural killer cells, B-cells, monocytes, eosinophils and other granulocytes[66] using constrained projection[67,68] via the minfi package in R[69]. To determine whether the observed associations with asthma were driven by any confounding factors, we produced three logistic regression models for each CpG site with asthma status as the dependent variable and M-values as the independent variable, while including: (1) no additional covariates for Model 1, (2) confounders associated with asthma status via t-tests as adjustment covariates for Model 2, and (3) all potential confounders identified *a priori* as adjustment covariates for Model 3. CpG sites that retained independent associations with asthma, after adjustment for confounders, were considered as our top findings.

As a post-hoc analysis, we investigated whether our top findings between DNA-M and asthma were also associated with some continuous measures of airway obstruction, reversible airway obstruction, and airway hyper-reactivity commonly used in the evaluation and management of asthma: FEV_1/FVC, BDR, and FeNO, respectively. To

conduct these analyses, we implemented Model 1 (crude) and Model 3 (adjusted) linear regression models for each of the above dependent variables. Also, because the underlying biological mechanisms leading to airway obstruction may differ between those with allergic- versus non-allergic-asthma, we conducted a sensitivity analysis to see whether the observed associations differed between those with and without atopy. For these analyses we conducted Model 1 (crude) and Model 3 (adjusted) regressions (logistic for asthma status, and linear for FEV_1/FVC, BDR, and FeNO), stratified by atopy status. Normal values of FEV_1, FVC are dependent on age, sex, and height, thus all statistical analyses utilizing these, or combinations of these, measures included sex and height as adjustment covariates (all participants were matched on age by study design).

Last, to understand the functionality of the selected CpG sites, pathway analysis was performed on the genes annotated to CpG sites that were internally validated in stage-2. Where a CpG site was annotated to more than one gene, all annotated genes were included in the list. Three CpGs were not annotated to any gene, so were not included in pathway analysis. The resulting list of 22 asthma-associated genes and their parameter estimate values was submitted to Ingenuity Pathway Analysis software (IPA, Qiagen), using default analysis parameters. The networks tool was used to find gene expression networks containing multiple differentially methylated genes.

2.3 Results:

All participants were 18 years old at the time of epigenetic screening for DNA-M and ascertainment of physician diagnosed asthma status. Within our sample, 13.9% (n=51) of participants were asthmatic, meaning they had received an asthma diagnosis from a physician at any point in their life and had asthmatic symptoms and/or took an

asthma medication within the previous 12 months. The majority of our sample was female (66.2%), and there was no difference in the proportion of males and females between asthmatics and non-asthmatics. Asthmatics had substantially higher prevalence of concurrent atopy, determined by SPTs, (66.0% vs 29.5%; $\chi 2$ P-value < 0.0001) (Table 2.1). We also compared a number of lung function measures, markers of airway reactivity, and proportions of circulating cell-types, between asthmatics and non-asthmatics. On average, there was no difference in FVC (means: 4.50 vs 4.48; T-test P-value = 0.93), but asthmatics had lower FEV1 (means: 3.74 vs 3.94; T-test P-value = 0.079), lower FEV1/FVC Ratio (means: 0.83 vs. 0.88; T-test P-value = 0.0006) greater BDR (means: 7.74 vs. 4.17; T-test P-value = 0.0004) and greater FeNO (medians: 21.0 vs 14.0; T-test P-value = 0.0005).

Because methylation measurements were obtained from peripheral blood samples, and persons with asthma likely have different circulating blood compositions than those without asthma, we also compared the proportions of cell types between asthmatics and non-asthmatics as estimated using the Houseman method [67,68]. On average, those with asthma had significantly greater proportions of circulating B-Cells (0.053 vs 0.045; T-test *P*-value = 0.028) and eosinophils (0.043 vs 0.021; T-test *P*-value = 0.0007) when compared to those with asthma, but only minor differences for all other cell-types.

To conduct the genome-scale DNA-M screening, this study first implemented a supervised feature selection on DNA-M sites from the Illumina 450K microarray (stage-1), then validated associations between DNA-M and asthma status for the set of selected features (stage-2), and lastly tested the validated sites for their associations with asthma, and with measures of airway-obstruction and hyper-reactivity, while adjusting for cellular

heterogeneity and sex (stage-3). To implement this design, the full data set (N=370) was randomly divided into two independent sets: a stage-1 sample (n_{S1}=91) used for feature selection and a stage-2 sample used for validation testing (n_{S2}=279). Prior to running the stage-1 analyses, we compared the stage-1 and stage-2 samples to ensure that randomization was effective for important variables: prevalence of asthma, prevalence of female sex, average values for airway obstruction and airway hyper-reactivity, as well as average cell-type proportions (Table 2.2). Only BDR was significantly different between stage-1 and stage-2 samples (4.99 vs. 3.58; T-test *P*-value = 0.006).

Stage 1 Results (n_{S1}=91) – Recursive RF feature Selection

The RF feature selection was implemented on the stage-1 sample, with a starting set of 248,336 CpG sites. This was a four step process which utilized RF to sort the predictors by their VIMs, then dropped least informative predictors in recursive iterations; each iteration reduced the set of predictors by 50%, until all uninformative predictors had been excluded. The overall forest's ability to predict asthma in the discovery sample was tracked across all iterations. We observed a leveling-off of the misclassification rates at the 12th iteration, corresponding to zero misclassification of asthmatics and 1% misclassification of non-asthmatics in the stage-1 sample (Figure 2.1). This iteration included 121 CpG sites as possible predictors, which were annotated with relevant genetic information and included in the supplemental materials. The forest grown in the 12th iteration of stage-1 was then used to predict asthma status of stage-2 participants, with very poor accuracy (Sensitivity=20.5% & Specificity=92.9%).

The low sensitivity was likely due to the recursive RF data reduction over-fitting to the stage-1 sample, resulting in the inclusion of sites that were spuriously associated

with asthma and over-fitting the cut-points of truly associated sites used to distinguish asthmatics from non-asthmatics in the stage-1 sample. To determine which of the selected sites were likely to be truly informative for asthma, the selected sites were then tested for their crude associations with asthma status via logistic regression in the stage-2 sample.

Stage 2 Results (n_{S2}=279) – Validation Tests in Independent Sample

The p-values for the crude associations exhibited a strong left peak with positive skew (Figure 2.2), indicating that the small p-values were unlikely to be purely chance findings. Thus q-values were generated for the 121 crude tests, 24 of which were associated with asthma (*P*-value range: 0.000045 – 0.027) at a FDR of 10% (Table 2.3). For all of these 24 sites, the direction of the association was the same for stage-1 and stage-2 logistic regression models (Figure 2.3). The majority of these sites were negatively associated with probability of asthma: for 21 (87.5%) of the validated sites, asthmatics were more likely to have lower average methylation levels than non-asthmatics; asthma was associated with higher average methylation for only 3 validated sites (12.5%). Given that these 24 sites were selected from stage-1 and validated in stage-2, for both direction of association and statistically significant FDR q-values, they were selected as candidate CpGs for follow-up analyses to better understand their associations and potential roles in asthma.

Stage 3 (N=370) – Characterization of Findings

The remaining analyses utilized methylation data from all participants (N=370), to provide the most accurate estimates of the relationships between DNA-methylation at these candidate CpGs and asthma (Table 2.4). We first produced crude logistic regression

models (Model 1) for each of the 24 sites with asthma status as the dependent variable and M-values as the independent variable. As expected, all sites had strong crude associations with asthma (negative OR range: 0.09 to 0.34; positive OR range: 7.99 to 9.51) and 95% CIs that excluded the null for all.

Then we added the proportions of eosinophils and B-cells, which were present in greater proportions among asthmatics (T-test *P*-values = 0.0007 and 0.0396, respectively), as covariates to the regression models (Model 2) for each site. We found that the majority of the ORs attenuated towards the null and 14 (58.3%) of the confidence intervals crossed the null, indicating that these sites were not independently associated with asthma, and may instead represent markers of eosinophils or B-cells. However, 10 sites retained independent associations with asthma after adjustment (negative OR range: 0.13 to 0.35; positive OR range: 8.96 to 10.56).

Last, we added in overall cellular heterogeneity (proportions of CD8T cells, CD4T cells, B-cells, monocytes, natural killer cells, eosinophils, and other granulocytes) and sex as covariates (Model 3). After adjusting for sex and proportions of all predicted cell-types, only one more site lost statistical significance: 15 (62.5%) of the 95% confidence intervals crossed the null. Given the potential dependence of associations between DNA-M and asthma on overall cellular heterogeneity[67,70] and sex[71], as well as the consistency of our results between Model 2 and Model 3, we considered the 9 sites that retained independent associations with asthma in Model 3 as our top positive findings. Lower average methylation levels were independently associated with increased probability of asthma at cg25578728 in *CHD7* (Model 3 OR = 0.109 (0.034, 0.333)), cg16658191 in *HK1* (Model 3 OR = 0.129 (0.032, 0.481)), cg00100703 in *UNC45B*

(Model 3 OR = 0.177 (0.059,0.502)), cg07948085 [intergenic] (Model 3 OR = 0.280 (0.095, 0.784)), cg04359558 in *LITAF* (Model 3 OR = 0.352 (0.172, 0.694)), and cg20417424 in *ST6GALNAC5* (Model 3 OR = 0.353 (0.118, 0.998)). Whereas higher average methylation levels were independently associated with increased probability of asthma at cg19974715 [intergenic] (Model 3 OR = 8.785 (2.536, 31.879)), cg01046943 in *NUP210* (Model 3 OR = 9.626 (2.229, 43.439)) and cg14727512 in *DGCR14* (Model 3 OR = 10.464 (3.021, 38.637)). One site, cg19232164 in *REXO2*, retained independent associations with asthma in Model 2 (OR = 0.254 (0.065,0.936)) but not in Model 3 (OR = 0.281 (0.073, 1.073)).

To investigate which cell-types were the strongest confounders of our observed associations we generated two correlation matrices of: (1) pair-wise correlations of predicted cell proportions and the 9 sites with independent associations for Matrix 1 (Figure 2.4), as well as (2) pair-wise correlations of predicted cell proportions and the 15 sites that were confounded by cellular heterogeneity for Matrix 2 (Figure 2.5). Interestingly, although all 9 sites in Matrix 1 showed strong associations with asthma independent of cellular heterogeneity, methylation levels at some of these sites were still moderately or even strongly correlated with proportions of eosinophils. For instance, cg16658191 was moderate-to-strongly correlated with proportions of eosinophils (r= -0.67) while cg07948085 (r= -0.54), cg20417424 (r= -0.46), and cg00100703 (r= -0.44) were moderately correlated with proportions of eosinophils. All other sites in Matrix 1 only exhibited weak correlations ($|r| < 0.40$) with any cell proportions. Although proportions of B-cells differed between those with and without asthma, methylation levels at the 15 CpGs in Matrix 2 only exhibited very weak correlations with B-cell

proportions but all were strongly negatively correlated with eosinophil proportions. CpG sites in Matrix 2 exhibited very weak correlations with any of the other cell-type proportions, indicating that eosinophils were likely driving the majority of confounding from cellular heterogeneity.

Because the underlying biological mechanisms leading to asthmatic symptoms may differ between those who have allergic hypersensitivity and those who do not, we conducted a sensitivity analysis to determine whether the associations between DNA-M and asthma differed between those with and without atopy among our top 9 findings. Persons with atopy were far more likely to be asthmatic than persons without atopy (26.2% vs. 7.1%). Also circulating cell-mixture showed larger differences between persons with and without asthma, among those with atopy: asthmatics had lower estimated proportions of natural killer cells (0.06 vs 0.08; T-test *P*-value = 0.0096) and higher estimated proportions of both B-cells (0.06 vs 0.05; T-test *P*-value = 0.0211) and eosinophils (0.05 vs 0.03; T-test *P*-value = 0.0027). However, among those without atopy, there were no significant differences for any estimated cell-type proportions between those with and without asthma. We also compared whether measures of airway obstruction (FEV1/FVC Ratio), reversible airway obstruction (BDR), and airway hyper-reactivity (FeNO) were differed between those with and without atopy. As expected, higher log-FeNO was associated with asthma, but only among those with atopy (T-test *P*-value = 0.006), whereas FEV_1/FVC ratio and BDR were associated with asthma among participants with (T-test *P*-values: 0.0153 and 0.0045, respectively) and without (T-test *P*-values: 0.0094 and 0.043, respectively) atopy.

Crude logistic regression models, stratified by atopy status (Table 2.5), revealed that all sites were strongly associated with asthma among persons with atopy, except for cg01046943 in *NUP210*; while 6 of the 9 sites were associated with asthma among those without atopy. After adjusting for sex and cellular heterogeneity cg16658191 in *HK1* (OR_{Atopy} = 0.09 (0.01,0.77)), cg04359558 in *LITAF* (OR_{Atopy} = 0.29 (0.10,0.81)), and cg14727512 in *DGCR14* (OR_{Atopy} = 22.99 (3.43,194.38)) were only associated with asthma among persons with atopy; cg00100703 in *UNC45B* ($OR_{No\text{-}Atopy}$ = 0.03 (0.01,0.19)), cg07948085 [intergenic] ($OR_{No\text{-}Atopy}$ = 0.15 (0.02,0.89)), and cg01046943 in *NUP210* ($OR_{No\text{-}Atopy}$ = 35.26 (3.32,444,37)) were only associated with asthma among persons without atopy; both cg25578728 in *CHD7* (OR_{Atopy} = 0.15 (0.02,0.82); and $OR_{No\text{-}Atopy}$ = 0.06 (0.01,0.35)) and cg19974715 [intergenic] (OR_{Atopy} = 26.32 (3.57,251.98); and $OR_{No\text{-}Atopy}$ = 8.32 (1.22,59.89)) were associated with asthma, irrespective of atopy status. Finally, cg20417424 in *ST6GALNAC5* was not associated with asthma within either atopy-strata after adjustment for sex and cellular heterogeneity, though the CI only barely crossed the null among those with atopy (OR_{Atopy} = 0.20 (0.03,1.06)). These findings provide evidence that there may be some shared epigenetic regulation (or dys-regulation) involved in asthmatic processes, but also that there are likely some epigenetic alterations that are unique to allergic and non-allergic asthma, respectively.

We next investigated whether the top 9 sites were associated with various measures of overall airway obstruction (FEV1/FVC ratio), reversible airway obstruction (BDR), and airway hyper-reactivity (FeNO), stratified by atopy status (Table 2.6). Because the established normal values for FEV1 and FVC are dependent on age, height and sex, we included height and sex in all statistical models for FEV1/FVC ratio and

BDR. Higher average methylation levels at cg00100703 in *UNC45B* were associated with increased FEV1/FVC ratio (beta = 0.03 P-Value = 0.032), among those without atopy, and with decreased average log-FeNO (beta = -0.21 P-Value = 0.024) among those with atopy. Higher average methylation levels at cg07948085 [intergenic] were associated with lower FeNO (beta = -0.20 P-Value =0.028), but only among those with atopy. Also, higher methylation levels at cg14727512 in *DGCR14* were associated with higher log-FeNO (beta = 0.32 P-Value = 0.005) among those with atopy. Higher methylation levels at cg16658191 in *HK1* were associated with lower BDR (beta = -4.91 P-Value = 0.036), but only among those with atopy. Higher methylation levels at cg19974715 [intergenic] were associated with increased BDR (beta = 4.89 P-Value = 0.016), only among those with atopy. Lastly, higher methylation levels at cg20417424 in *ST6GALNAC5* were associated with increased FEV1/FVC ratio , both for those with (beta = 0.05 P-Value = 0.022) and without (beta = 0.03 P-Value = 0.022) atopy as well as decreased BDR (beta = -2.41 P-Value = 0.017) only among those without atopy.

The functional analyses for the 22 genes submitted to IPA revealed five statistically significant canonical pathways related to trehalose degradation II (*P*-value = 3.41E-03), GDP-glucose biosynthesis (*P*-value = 6.80E-03), glucose and glucose-1-phosphate degradation (*P*-value = 7.93E-03), UDP-N-acetyl-D-galactosamine biosynthesis II (*P*-value = 1.02E-02), and CDP-diacylglycerol biosynthesis I (*P*-value = 1.80E-02). Interestingly, the top network associated with our gene list, *Cell Morphology, Tissue Morphology, Cellular Development*, (Figure 2.6) included 7 of the 22 genes from our list: *HK1, LITAF, ALCAM, KCNH2, SIK2, ST6GALNAC5*, and *DGCR14.*

2.4 Discussion:

Asthma was prevalent in 13.8% of our epigenetic sub-sample; the majority of which had atopy (66%) and on average had normal FEV1/FVC (mean = 83%), though lower than non-asthmatics, high BDR (mean = 7.74 L), and high FeNO (median = 21.00 log(ppb)). Using a unique feature-selection technique and two-stage design, we identified 24 CpG sites within a 10% FDR associated with asthma-status at age 18; nine of which had associations with asthma, atopy-specific asthma, airway obstruction, reversible airway obstruction, and/or airway hyper-reactivity, independent of cellular heterogeneity and sex. Gene annotations for these nine sites (Table 2.7) revealed that all of our top findings were novel, as none of the genes encompassing these CpG sites had been previously implicated in studies of asthma; though many of them are involved in biological processes that may be important in asthma pathogenesis and persistence. Many of the top sites were within genes involved in a network that influences cell and tissue morphology. The present findings require replication in other cohorts with larger samples.

Three of these CpG sites were most strongly associated with asthma among persons with atopy: cg16658191 within the body (or 1st exon) of hexokinase 1 (*HK1*), cg14727512 within the 3'UTR of DiGeorge Syndrome Critical Region Gene 14 (*DGCR14*) and within the 1st exon of testis-specific serine kinase 2 (*TSSK2*), and cg04359558 within the body of lipopolysaccharide-induced TNF-α factor (*LITAF*). Among persons with atopy, methylation levels at cg16658191 were also associated with BDR, a measure of reversible airway obstruction, and methylation levels at cg1427512 were associated with FeNO, a marker of airway inflammation. *HK1* encodes for a

hexokinase-1 which is integral in glucose metabolism, provides resistance to TNF-induced apoptotic signals[72], and mutations within this gene have been associated with hemolytic anemia in mice[73]. Eosinophils of asthmatics not on corticosteroids[74] and activated T-cells of atopic asthmatics[75] have been shown to be resistant to apoptotic signals resulting in prolonged pro-inflammatory activity, though many aspects of the molecular mechanisms involved in this resistance are still unclear[75]. *DGCR14* is thought to encode for a component of a spliceosome. Deletions at or near this gene have been implicated in the etiology of DiGeorge syndrome, a heterogenous developmental disorder that often is characterized by T-cell deficiency and/or craniofacial malformations[76]. Also of note, increased expression of *DGCR14* has been shown to up-regulate *IL17α*, thus promoting Th17 cell differentiation[77], and Th17 cells have been implicated for their roles in non-atopic and steroid-resistant asthma[78]. *LITAF*, which has primarily been studied in mouse models, encodes for a protein that is expressed in response to lipopolysaccharide (LPS) challenge, and forms a complex with STAT6B which then up-regulates the transcription of monocyte chemoattractant protein-1 (MCP-1)[79] and tumor necrosis factor alpha (TNF-alpha)[80]. Both MCP-1 and TNF-alpha are pro-inflammatory cytokines, and thus over-expression could lead to prolonged inflammatory states[79,80].

Another three sites were most strongly associated with asthma among those without atopy: cg07948085 [intergenic], cg01046943 within the body of nucleoporin 210kDa (*NUP210*), and cg00100703 within the 3'UTR of unc-45 homolog B (*UNC45B*). Interestingly, despite these sites being most strongly associated with asthma among persons without atopy, cg07948085 and cg00100703 were associated with FeNO among persons with atopy; whereas cg00100703 was associated with FEV1/FVC ratio among

persons without atopy. *NUP210* encodes for a major component of the nuclear pore complex which is essential for transporting molecules between the nucleus and the cytoplasm; this particular component is involved in muscle cell differentiation and is integral for maintaining nuclear envelope/ER homeostasis[81]. *NUP21* was also found to be up-regulated in multiple tumor cell lines[82] and may have some anti-apoptotic effects[81]. The functions of the protein encoded by *UNC45B* are largely unknown, although it is thought to have a role in myoblast fusion and sarcomere organization[83]. However, *UNC45B* does lie within chromosomal region 17q12, which is within the asthma susceptibility region 17q12-21[71,84–86]. SNPs within 17q12-21 the genes *ZPBP2*, *ORMDL3*, *GSDMA*, and *GSDMB* have previously been associated with childhood asthma and with FeNO levels[85,87]. Berlivet et al (2012) recently found that promoter methylation within these genes was highly correlated with their expression levels, providing evidence for epigenetic-genetic interactions that may alter asthma risk[87].

One site, cg20417424 within the TSS1500 of (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 5 (*ST6GALNAC5*), was associated with overall asthma, but not with either atopy-specific strata. Yet, methylation levels at this site were associated with FEV1/FVC ratio, among persons with and without atopy, and with BDR among persons without atopy. The protein encoded by *ST6GALNAC5* is involved in modifying cell-surface proteins, influencing how they participate in cell-cell and cell-extracellular matrix interactions[88].

The final two sites, cg25578728 within the body of chromodomain helicase DNA binding protein 7 (*CHD7*) and cg19974715 [intergenic], were strongly associated with asthma both among those with and without atopy. And, cg19974715 was also associated

with BDR among persons with atopy. *CHD7* encodes for a protein that is involved in chromatin remodeling, and thus may have widespread affects on gene-expression. Recent work with mouse models has shown that functional *CHD7* is necessary for proper craniofacial development and formation of the airways during development and in humans may result in CHARGE syndrome[89].

Although none of the above genes, which contained our 9 identified CpG sites (within 100kb), have been previously implicated in studies of asthma, the roles that these genes play in inflammation, cell surface protein modifications, development of the respiratory tract, and peripheral airway restriction are all important mechanisms in asthma pathogenesis. Thus these may represent some under-studied candidate genes and CpG sites that could play important roles in asthma etiology. However, the above findings must be interpreted within the limitations of this study.

Due to the cross-sectional design, measurements of DNA-M, asthma, spirometry, and FeNO were all obtained at the same point in time; thus, temporality between DNA-M and our different outcomes cannot be established. It is possible that these findings are a product of reverse causation, in which the observed variations in DNA-M were actually caused by asthma or asthma treatments. Follow-up studies with longitudinally collected DNA-M levels and asthma assessments could provide more insight into which of these variations might contribute to the development of asthma or the persistence of asthma, and which may come about as a result of asthma. Also, some may consider the use of questionnaire-based asthma determinations a limitation of this study. However, the questions used in our diagnosis came from the validated ISAAC questionnaire[5]. Also, despite the many different presentations of asthma, multiple studies have been successful

at replicating findings using general asthma-status as an outcome, which may support that there are some common molecular mechanisms driving the various asthmatic phenotypes. Also, we found that eight of our nine sites were associated with asthma stratified by atopy status and six of the nine sites were associated with airway obstruction (FEV1/FVC), reversible airway obstruction (BDR), and/or airway hyper-reactivity (FeNO). Lastly, although there were multiple functionally interesting genes identified in this study, we only evaluated variations in DNA-M at the identified CpG sites. Thus it is unclear whether the variations in DNA-M at these sites would result in any regulatory effect on the expression of any of these genes. However, current paradigms indicate that methylation within promoter regions typically leads to transcriptional silencing while methylation within the gene-body is more frequently associated with increased expression and alternative splicing[32]. Thus interesting follow-up work could involve investigations into whether variations in DNA-M at these sites are associated with corresponding variations in gene-expression, and if those variations in gene-expression are related to the frequency and severity of asthmatic symptoms, FEV1/FVC, BDR, or FeNO.

In conclusion, we found that DNA-M variations at nine novel CpG sites were associated with prevalent asthma at age 18. For some of these sites, the association was strongest among persons with atopy (cg16658191 in *HK1*, cg04359558 in *LITAF*, and cg14727512 in *DGCR14*), some among persons without atopy (cg00100703 in *UNC45B*, cg07948085 [intergenic], and cg01046943 in *NUP210*), while others were associated with asthma regardless of their atopy status (cg25578728, cg01046943, and cg19974715). We also found that cg00100703 in *UNC45B* was associated with

FEV1/FVC and FeNO; cg07948085 [intergenic] was associated with FeNO; cg14727512 in *DGCR14* was associated with FeNO; cg16658191 in *HK1* was associated with BDR, cg19974715 [intergenic] was associated with BDR; and cg20417424 in *ST6GALNAC5* was associated with FEV1/FVC.

Table 2.1: Comparison of lung function and cell-type distributions between those with and without asthma in the overall epigenetic sub-sample.

	Asthmatics (n=51)		**Non-Asthmatics (n=319)**		**Asthma Comparison**	
Categorical Factors	**n**	**Percent**	**n**	**Percent**	**χ^2 Statistic**	**P-Value**
Female Sex	35	68.6	210	65.8	0.054	0.816
Atopy	33	66.0	93	29.5	23.811	< 0.001
Continuous Factors	**n**	**Mean (SD)**	**n**	**Mean (SD)**	**T-Statistic**	**P-Value**
FVC	49	4.50 (0.874)	307	4.48 (0.857)	-0.09	0.9307
FEV_1	49	3.74 (0.750)	307	3.94 (0.719)	1.78	0.0794
FEV_1/FVC Ratio	49	0.83 (0.088)	307	0.88 (0.064)	3.63	0.0006
BDR	48	7.74 (6.243)	294	4.17 (4.802)	-3.78	0.0004
FeNO[†]	47	21.00 (35.37)	300	14.0 (17.67)	-3.70	0.0005
CD8T	51	0.075 (0.081)	319	0.072 (0.048)	-0.27	0.7907
CD4T	51	0.133 (0.048)	319	0.123 (0.045)	-1.36	0.1755
Natural Killer Cells	51	0.073 (0.058)	319	0.087 (0.061)	1.58	0.1193
B Cells	51	0.053 (0.022)	319	0.045 (0.025)	-2.21	0.0306
Monocytes	51	0.080 (0.026)	319	0.082 (0.024)	0.57	0.5676
Eosinophils	51	0.043 (0.043)	319	0.021 (0.023)	-3.60	0.0007
Other Granulocytes	51	0.536 (0.107)	319	0.563 (0.100)	1.74	0.0927

FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; BDR, bronchodilator reversibility; FeNO, fractional-exhaled nitric oxide

[†] Due to heavy positive skew of FeNO, medians were presented in the mean column, and log(FeNO) was used for computing the T-test.

Table 2.2: Comparison of stage 1 and stage 2 samples for asthma variables and cell proportions.

	Stage 1 Sample (n_{s1}=91)		**Stage 2 Sample (n_{s2}=279)**		**Sample Comparison**	
Asthma Prevalence	**n**	**%**	**n**	**%**	**Chi-Statistic**	**P-Value**
Non-Asthmatic	79	86.81	240	86.02	0.0002	0.987
Asthmatic	12	13.19	39	13.98		
Lung Function	**n**	**Mean (SD)**	**n**	**Mean (SD)**	**T-Statistic**	**P-Value**
FVC	88	4.497 (0.801)	268	4.482 (0.877)	0.152	0.879
FEV_1	88	3.917 (0.688)	268	3.909 (0.739)	0.095	0.925
FEV_1/FVC Ratio	88	0.875 (0.075)	268	0.875 (0.068)	0.004	0.997
BDR	83	5.933 (5.085)	259	4.269 (5.141)	2.773	0.006
FeNO[†]	83	14.00 (17.46)	264	15.00 (23.50)	-0.493	0.623
Cellular Heterogeneity	**n**	**Mean (SD)**	**n**	**Mean (SD)**	**T-Statistic**	**P-Value**
CD8T	91	0.074 (0.055)	279	0.071 (0.053)	0.458	0.647
CD4T	91	0.128 (0.040)	279	0.123 (0.047)	1.006	0.316
Natural Killer	91	0.085 (0.052)	279	0.085 (0.064)	0.072	0.942
B Cells	91	0.049 (0.026)	279	0.045 (0.024)	1.093	0.276
Monocytes	91	0.080 (0.024)	279	0.082 (0.024)	-0.559	0.577
Eosinophils	91	0.021 (0.024)	279	0.025 (0.029)	-1.242	0.216
Other Granulocytes	91	0.555 (0.090)	279	0.561 (0.105)	-0.537	0.592

FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; BDR, bronchodilator reversibility; FeNO, fractional-exhaled nitric oxide

[†] Due to heavy positive skew of FeNO, medians were presented in the mean column, and log(FeNO) was used for computing the T-test.

Table 2.3: Stage 2 (n_{s2}=279) – Crude logistic regression results for the sites within a 10% FDR.

CpG ID	β_1	P-Val.	Q-Val.	Gene	Gene Region
cg16658191	-1.90	0.00004	0.003	*HK1*	Body;1stExon
cg25578728	-1.83	0.00036	0.010	*CHD7*	Body
cg00100703	-1.68	0.00049	0.010	*UNC45B*	3'UTR
cg07948085	-1.49	0.00054	0.010		
cg09241885	-1.02	0.00102	0.015	*C20orf118*	TSS200
cg11310939	-1.46	0.00120	0.015	*MARCH3*	5'UTR
cg01069468	-1.69	0.00231	0.025	*SYNGAP1*	Body
cg04359558	-1.03	0.00304	0.029	*LITAF*	Body
cg06866208	-1.44	0.00520	0.041		
cg13753183	-1.66	0.00529	0.041	*APTX*	Body;1stExon;5'UTR
cg06648780	-1.29	0.00728	0.051	*ALCAM*	Body
cg24491618	-1.34	0.00961	0.062	*KCNH2*	Body
cg09597192	-1.23	0.01154	0.066	*AGPAT1*	5'UTR
cg15344640	-1.39	0.01234	0.066	*LMAN2*	Body
cg26252077	-0.93	0.01360	0.066	*NFIA*	Body
cg24368962	-1.17	0.01433	0.066	*SIK2*	Body
cg09278187	-1.24	0.01453	0.066	*FOXJ3*	3'UTR
cg20417424	-1.19	0.01610	0.068	*ST6GALNAC5*	TSS1500
cg10704177	-0.70	0.01690	0.068	*INADL*	5'UTR
cg14727512	1.53	0.01809	0.070	*TSSK2;DGCR14*	1stExon;3'UTR
cg08940169	-0.85	0.01908	0.070	*ZFPM1*	Body
cg01046943	1.75	0.02345	0.081	*NUP210*	Body
cg19974715	1.51	0.02415	0.081		
cg19232164	-1.32	0.02681	0.086	*REXO2*	3'UTR

β_1 = parameter estimate for the effect of DNA-M on asthma status
P-Val = p-value corresponding to the significance of β_1
Q-Val = FDR-adjusted p-values, adjusted for 121 regression models

Table 2.4: Pooled Sample (n=370) – Crude and adjusted ORs for M-values predicting asthma status.

	Model 1		**Model 2**		**Model 3**	
CpG ID	**Crude OR**	**Crude CI**	**Adj.* OR**	**Adj. CI**	**Adj.** * OR**	**Adj. CI**
cg16658191	**0.098**	**(0.04, 0.23)**	**0.136**	**(0.04, 0.49)**	**0.129**	**(0.03,0.48)**
cg19232164	**0.098**	**(0.03, 0.28)**	**0.254**	**(0.06, 0.94)**	0.281	(0.07,1.04)
cg13753183	**0.109**	**(0.04, 0.31)**	0.364	(0.08, 1.62)	0.347	(0.07,1.58)
cg25578728	**0.116**	**(0.05, 0.28)**	**0.163**	**(0.06, 0.44)**	**0.109**	**(0.03,0.33)**
cg01069468	**0.118**	**(0.04, 0.31)**	0.409	(0.09, 1.78)	0.486	(0.10,2.23)
cg00100703	**0.120**	**(0.05, 0.28)**	**0.190**	**(0.07, 0.52)**	**0.177**	**(0.06,0.50)**
cg06866208	**0.137**	**(0.05, 0.34)**	0.449	(0.11, 1.89)	0.532	(0.12,2.33)
cg24491618	**0.143**	**(0.06, 0.35)**	0.500	(0.12, 2.06)	0.550	(0.13,2.31)
cg15344640	**0.148**	**(0.05, 0.39)**	0.973	(0.19, 4.66)	1.095	(0.20,5.74)
cg11310939	**0.155**	**(0.07, 0.34)**	0.378	(0.10, 1.35)	0.434	(0.11,1.57)
cg06648780	**0.157**	**(0.07, 0.36)**	0.347	(0.09, 1.19)	0.365	(0.09,1.33)
cg07948085	**0.157**	**(0.07, 0.34)**	**0.308**	**(0.11, 0.82)**	**0.280**	**(0.09,0.78)**
cg09597192	**0.158**	**(0.07, 0.37)**	0.702	(0.16, 3.12)	1.010	(0.21,4.83)
cg20417424	**0.170**	**(0.07, 0.40)**	**0.359**	**(0.12, 0.99)**	**0.353**	**(0.12,0.99)**
cg24368962	**0.174**	**(0.07, 0.39)**	0.650	(0.16, 2.69)	0.489	(0.11,2.14)
cg09278187	**0.176**	**(0.07, 0.44)**	0.574	(0.16, 1.96)	0.674	(0.19,2.35)
cg04359558	**0.233**	**(0.12, 0.42)**	**0.358**	**(0.18, 0.70)**	**0.352**	**(0.17,0.69)**
cg08940169	**0.265**	**(0.14, 0.50)**	0.990	(0.29, 3.40)	1.158	(0.31,4.30)
cg26252077	**0.277**	**(0.14, 0.53)**	0.698	(0.25, 1.87)	0.710	(0.25,1.93)
cg09241885	**0.286**	**(0.16, 0.49)**	0.513	(0.22, 1.17)	0.447	(0.18,1.08)
cg10704177	**0.340**	**(0.20, 0.57)**	1.140	(0.39, 3.34)	1.608	(0.51,5.08)
cg14727512	**7.997**	**(2.58, 26.09)**	**10.562**	**(3.12, 37.97)**	**10.464**	**(3.02,38.64)**
cg01046943	**9.464**	**(2.36, 39.32)**	**9.639**	**(2.25, 42.35)**	**9.626**	**(2.23,43.44)**
cg19974715	**9.516**	**(2.99, 31.59)**	**8.968**	**(2.63, 31.92)**	**8.785**	**(2.54,31.88)**

OR, odds ratio; CI, 95% confidence interval; Adj., adjusted

* Odds ratios adjusted for potential confounders strongly associated with asthma status: proportions of B-cells and eosinophils.

** Odds ratios adjusted for sex and overall cellular heterogeneity (proportions of CD8T cells, CD4T cells, B-cells, monocytes, natural killer cells, eosinophils, and other granulocytes). Sites whose 95% CIs didn't cross the null are emphasized with bold text.

Table 2.5: Crude and adjusted ORs for M-values predicting asthma status stratified by atopy status, among the top nine sites.

Atopy (n= 126; 33 with asthma)				
	Crude OR	**Crude CI**	**Adj** OR**	**Adj CI**
cg16658191	**0.10**	**(0.03,0.32)**	**0.09**	**(0.01,0.77)**
cg25578728*	**0.13**	**(0.03,0.43)**	**0.15**	**(0.02,0.82)**
cg00100703	**0.19**	**(0.06,0.58)**	0.36	(0.07,1.61)
cg07948085	**0.21**	**(0.07,0.56)**	0.27	(0.05,1.19)
cg20417424	**0.12**	**(0.03,0.41)**	0.20	(0.03,1.06)
cg04359558	**0.18**	**(0.07,0.41)**	**0.29**	**(0.10,0.81)**
cg14727512	**14.11**	**(2.71,86.41)**	**22.99**	**(3.43,194.38)**
cg01046943	5.61	(0.74,45.72)	3.85	(0.39,40.55)
cg19974715*	**20.76**	**(3.70,140.35)**	**26.32**	**(3.57,251.98)**
No Atopy (n=239; 17 with asthma)				
	Crude OR	**Crude CI**	**Adj** OR**	**Adj CI**
cg16658191	0.23	(0.04,1.17)	0.18	(0.02,1.41)
cg25578728*	**0.15**	**(0.04,0.62)**	**0.06**	**(0.01,0.35)**
cg00100703	**0.09**	**(0.02,0.36)**	**0.03**	**(0.01,0.19)**
cg07948085	**0.22**	**(0.05,0.85)**	**0.15**	**(0.02,0.89)**
cg20417424	0.61	(0.14,2.53)	0.59	(0.11,2.88)
cg04359558	0.43	(0.15,1.15)	0.40	(0.13,1.16)
cg14727512	**7.14**	**(1.18,46.99)**	5.45	(0.83,39.75)
cg01046943	**38.23**	**(4.20,390.06)**	**35.26**	**(3.32,444.37)**
cg19974715*	**7.06**	**(1.14,44.18)**	**8.32**	**(1.22,59.89)**

OR, odds ratio; CI, 95% confidence interval; Adj., adjusted

* Sites that were significantly associated with asthma in both strata, after adjusting for sex and cellular heterogeneity.

** Odds ratios adjusted for sex and overall cellular heterogeneity (proportions of CD8T cells, CD4T cells, B-cells, monocytes, natural killer cells, eosinophils, and other granulocytes). Sites whose 95% CIs didn't cross the null are emphasized with bold text.

Table 2.6: Parameter estimates from linear regressions for measures of airway obstruction (FEV1/FVC), hyper-reactivity (BDR) and airway inflammation (FeNO) for the top 9 sites, stratified by atopy status.

		FEV1/FVC				**BDR**				**FeNO**			
CpG ID	**Strata**	**Crude**	**P-Val**	**Adj***	**P-Val**	**Crude**	**P-Val**	**Adj***	**P-Val**	**Crude**	**P-Val**	**Adj****	**P-Val**
cg00100703	Atopy	**0.04**	**0.031**	0.01	0.627	-1.97	0.166	0.58	0.739	**-0.37**	**<0.001**	**-0.21**	**0.024**
	No Atopy	**0.03**	**0.003**	**0.03**	**0.032**	-0.98	0.251	-0.25	0.790	-0.04	0.255	-0.02	0.650
cg01046943	Atopy	-0.02	0.555	-0.03	0.486	2.43	0.365	1.13	0.680	0.21	0.203	0.17	0.253
	No Atopy	-0.03	0.191	-0.03	0.190	1.83	0.223	2.10	0.166	0.02	0.745	0.03	0.608
cg04359558	Atopy	0.01	0.317	-0.01	0.631	-1.14	0.219	0.32	0.772	**-0.18**	**0.002**	-0.04	0.484
	No Atopy	0.01	0.153	0.01	0.558	0.62	0.328	1.10	0.103	0.00	0.958	0.01	0.645
cg07948085	Atopy	**0.03**	**0.043**	0.02	0.480	**-3.13**	**0.007**	-2.88	0.087	**-0.38**	**<0.001**	**-0.20**	**0.028**
	No Atopy	0.02	0.179	0.01	0.649	0.87	0.322	1.54	0.138	-0.01	0.902	0.00	0.971
cg14727512	Atopy	-0.02	0.448	-0.03	0.262	1.62	0.436	1.20	0.574	0.25	0.060	**0.32**	**0.005**
	No Atopy	-0.01	0.547	-0.01	0.539	0.31	0.795	0.57	0.638	-0.08	0.146	-0.03	0.580
cg16658191	Atopy	**0.04**	**0.021**	0.05	0.135	**-3.83**	**0.006**	**-4.91**	**0.036**	**-0.39**	**<0.001**	-0.13	0.336
	No Atopy	0.02	0.101	0.01	0.558	-0.49	0.654	-0.44	0.740	-0.05	0.366	0.01	0.859
cg19974715	Atopy	-0.04	0.215	-0.03	0.346	**5.73**	**0.005**	**4.89**	**0.016**	0.08	0.545	0.03	0.812
	No Atopy	-0.02	0.281	-0.01	0.365	-0.57	0.661	-0.93	0.479	0.04	0.499	0.00	0.955
cg20417424	Atopy	**0.07**	**<0.001**	**0.05**	**0.022**	**-3.70**	**0.012**	-1.63	0.361	**-0.29**	**0.001**	0.01	0.920
	No Atopy	**0.03**	**0.006**	**0.03**	**0.022**	**-2.18**	**0.020**	**-2.41**	**0.017**	-0.06	0.171	0.01	0.877
cg25578728	Atopy	0.03	0.089	0.00	0.870	-1.91	0.176	0.78	0.665	-0.13	0.127	0.12	0.229
	No Atopy	0.01	0.305	0.00	0.787	-1.23	0.185	-0.68	0.512	-0.07	0.100	-0.03	0.530

FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; BDR, bronchodilator reversibility; FeNO, fractional-exhaled nitric oxide

***** Parameter estimates adjusted for height, sex and overall cellular heterogeneity (proportions of CD8T cells, CD4T cells, B-cells, monocytes, natural killer cells, eosinophils, and other granulocytes). Sites with P-values < 0.05 are emphasized with bold text.

****** Parameter estimates adjusted for sex and overall cellular heterogeneity (proportions of CD8T cells, CD4T cells, B-cells, monocytes, natural killer cells, eosinophils, and other granulocytes). Sites with P-values < 0.05 are emphasized with bold text.

Table 2.7: Annotations and biological functions of genes associated with the nine CpG sites associated with asthma, independent of cellular heterogeneity and sex.

CpG ID	Ch	Gene Region	Gene ID	Gene Name	Gene Description
cg00100703	17	3'UTR	*UNC45B*	unc-45 homolog B	Encodes a protein that is involved in myoblast and sarcomere organization.
cg01046943	3	Body	*NUP210*	Nucleoporin 210kDa	Encodes a protein that is a major component in the nuclear pore complex, which controls the flow of molecules between the nucleus and cytoplasm.
cg04359558	16	Body	*LITAF*	Lipopolysaccharide-Induced TNF-α Factor	Encodes a protein which promotes TNF-α factor expression.
cg07948085	10	-	-	-	N/A
cg14727512	22	3'UTR	*DGCR14*	DiGeorge Syndrome Critical Region Gene 14	Encodes a component of a spliceosome; deletions in this region associated with DiGeorge syndrome.
cg16658191	10	Body; 1stExon	*HK1*	Hexokinase 1	Encodes a hexokinase involved in glucose metabolism; associated with anemia.
cg19974715	17	-	-	-	N/A
cg20417424	1	TSS1500	*ST6GALNAC5*	(alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 5	Encodes a protein that modifies cell-surface proteins that are important in cell-cell or cell-extracellular matrix interactions.
cg25578728	8	Body	*CHD7*	Chromodomain Helicase DNA Binding Protein 7	Encodes a protein that is involved in regulating gene-expression via chromatin remodeling and loss-of-function results in CHARGE syndrome.

Ch, Chromosome; TSS1500, transcription start site within 1500 base pairs; UTR, untranslated region

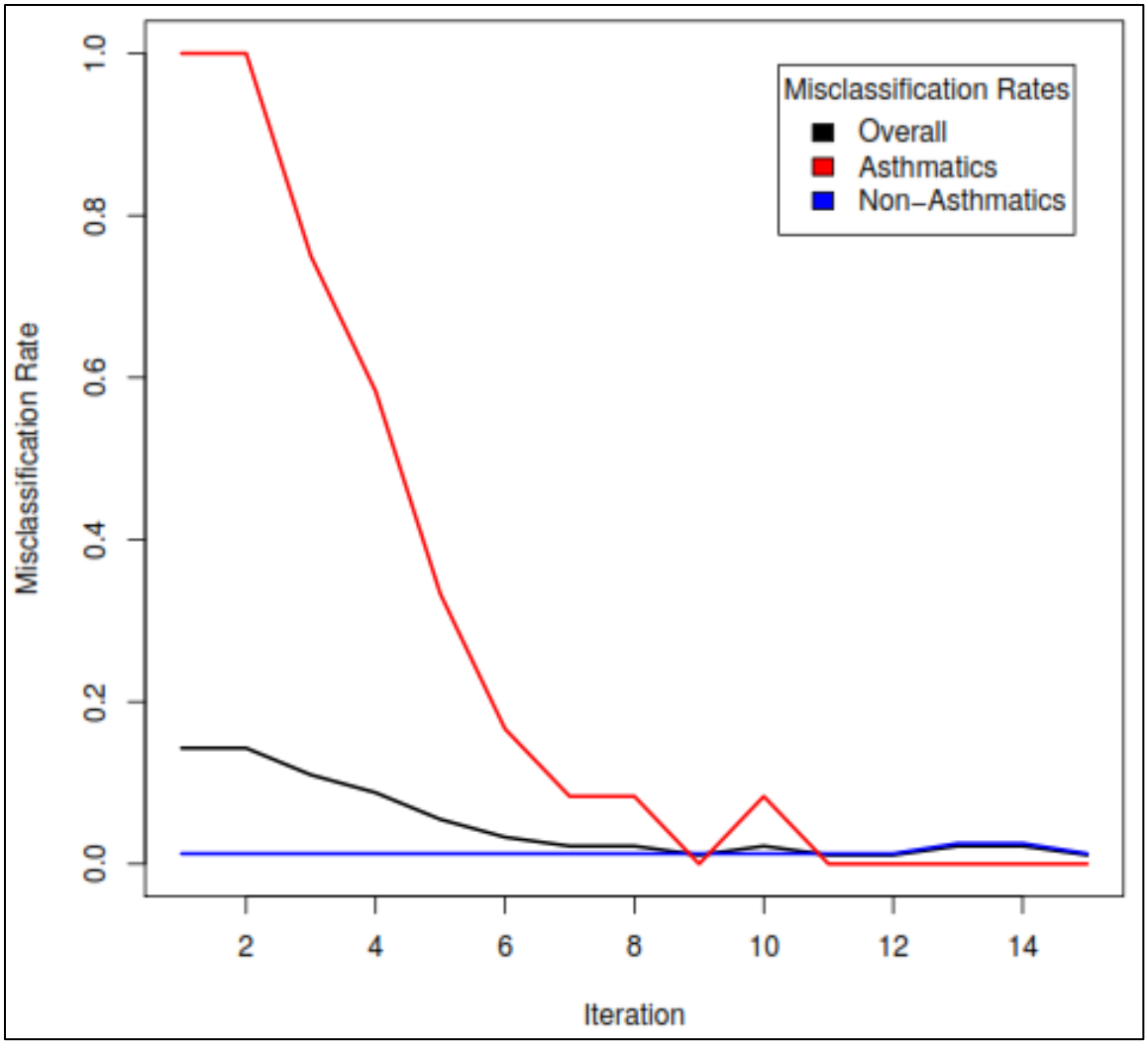

Figure 2.1: Tracking of the misclassification rates (y-axis) across iterations (x-axis) of the recursive RF feature selection.

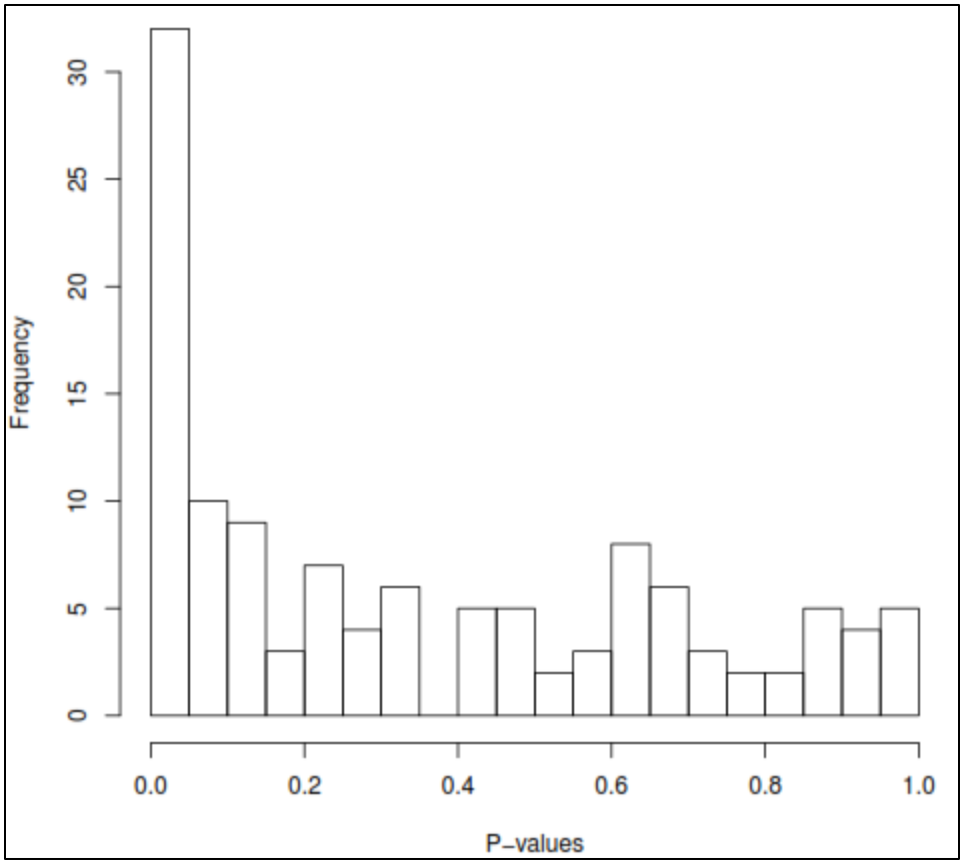

Figure 2.2: Histogram of P-values from 121 regressions in the stage 2 analyses.

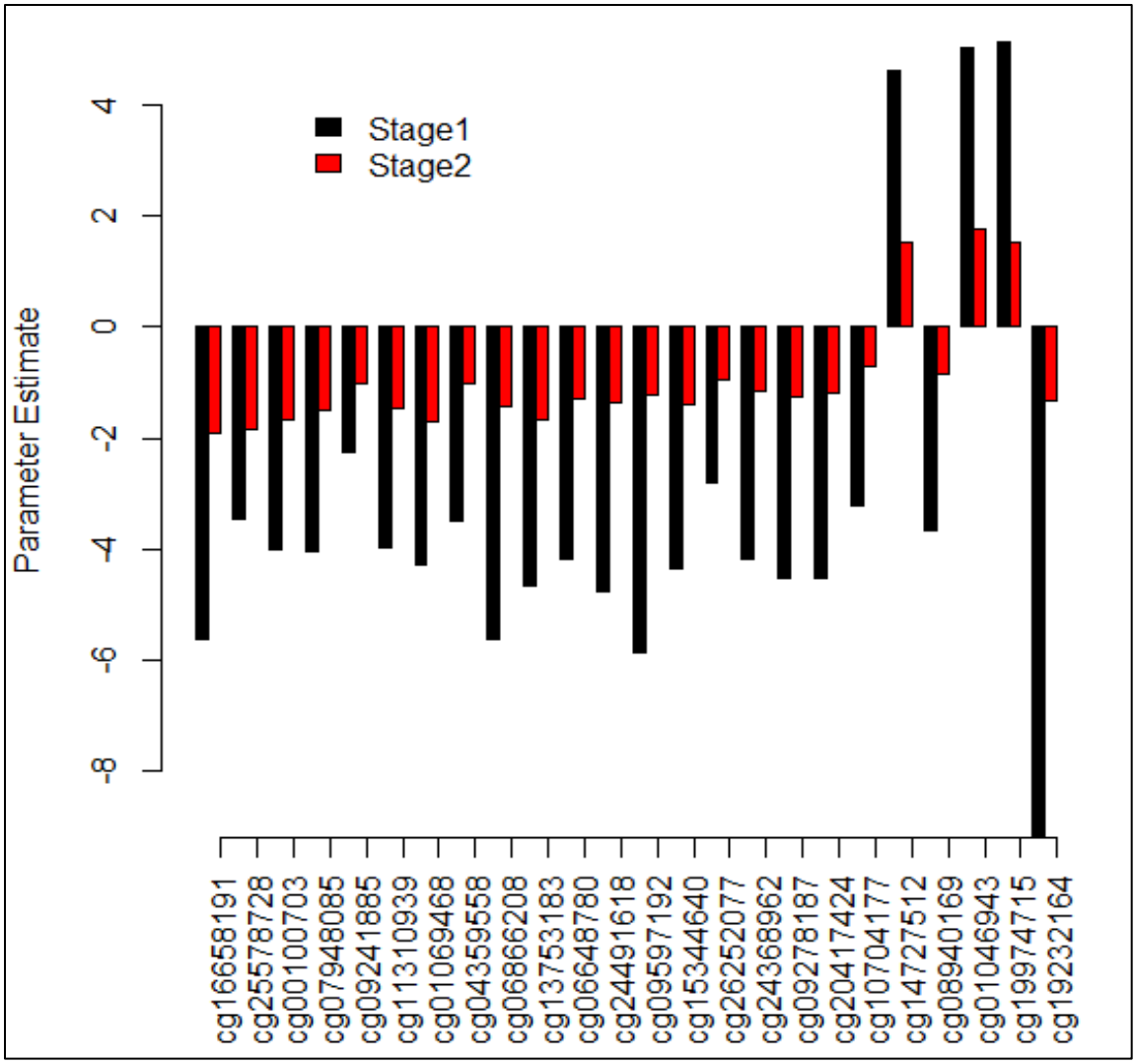

Figure 2.3: Comparison of parameter estimates from stage 1 and stage 2 logistic regression models, among the 24 sites within a 10% FDR in stage 2 analyses.

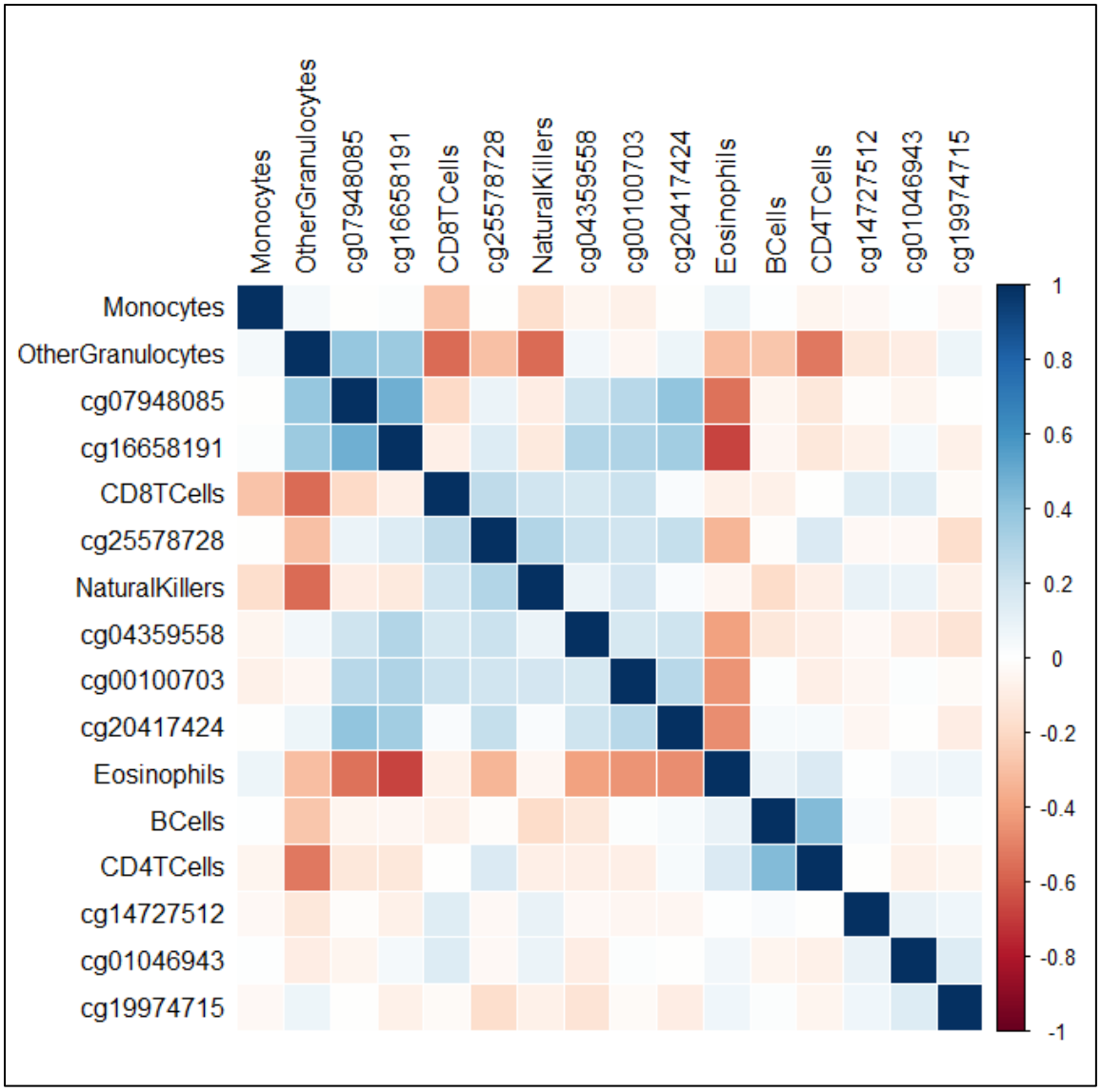

Figure 2.4: Correlation matrix of 9 CpGs associated with asthma, independent of cellular heterogeneity, and cell predicted cell proportions.

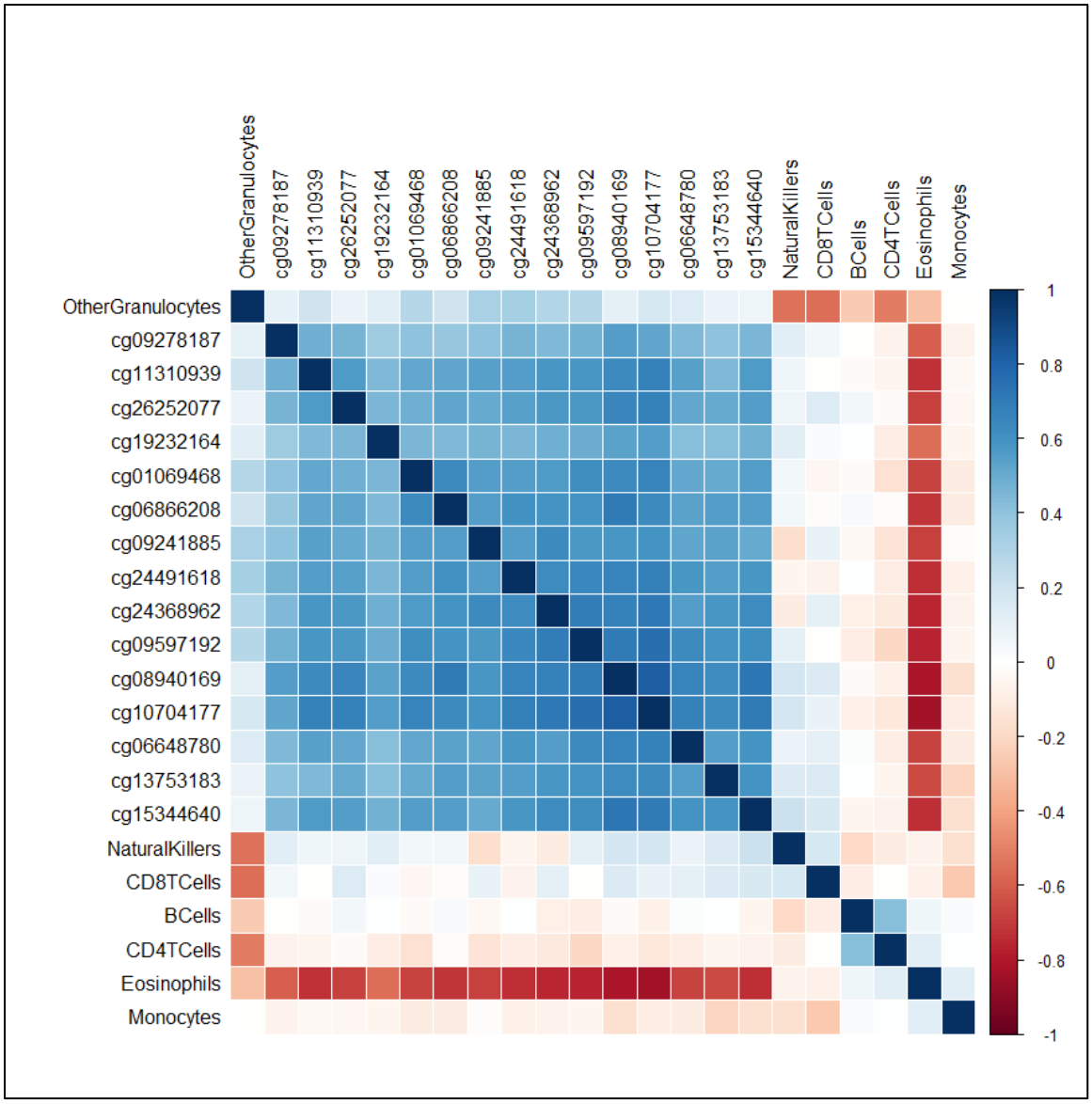

Figure 2.5: Correlation matrix of 15 CpGs associated with asthma, but confounded by cellular heterogeneity.

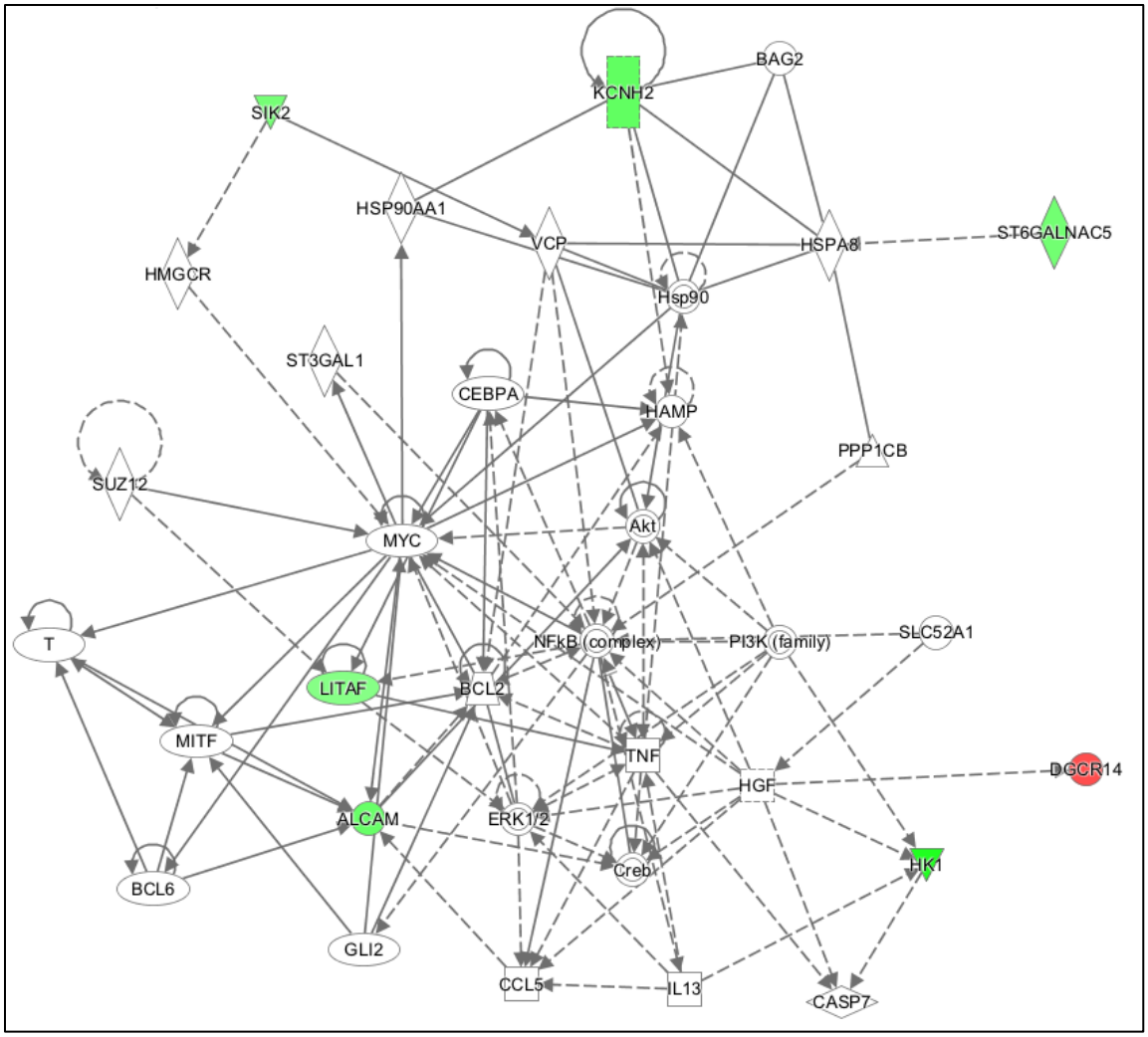

Figure 2.6: Cell and tissue morphology network, including seven genes with higher (red) or lower (green) methylation levels associated with asthma status.

CHAPTER 3

EXPLORATORY STUDY OF GENETIC AND EPIGENETIC VARIATIONS ASSOCIATED WITH YOUNG ADULT WHEEZE CLUSTERS

3.1 Introduction:

Asthma and wheeze illness, defined by restricted airflow, encompass a wide array of symptoms and physiological characteristics including allergic hypersensitivity, impaired lung function and responsiveness to bronchodilators[1]. The frequency and severity of these symptoms are somewhat dependent on the time of onset and gender, and the combination of all these factors influence the management and prognosis of disease. Recent studies have made use of clustering to classify persons with asthma and/or wheeze based on clinical and demographic features such as race, gender, age of onset of illness, atopic status, obesity, comorbidities, severity of disease, and lung function[4,90–92]. Previous work within the Isle of Wight (IOW) birth cohort , an unselected population birth cohort in the United Kingdom, identified six distinct clusters of young adult wheeze, characterized by 13 important components of wheezing illness in 18 year olds[4]. Each cluster represented a group of individuals with recent wheeze or asthma symptoms, which were relatively homogeneous in gender, time of onset, allergic sensitivity, lung function, and levels of treatment.

It has long been recognized that T-cell polarization towards a Th2 phenotype, which drive pro-inflammatory responses, was an important cellular component of allergy and allergic asthma[15,93] and $CD4^{+}$T cells tend to commit to type 2 helper (Th2) T

cells in response to an allergen among persons with atopy[8,94]. A set of cytokines, referred to collectively as the Th2 path, are largely responsible for Th2 commitment, have been associated with asthma[33,38,95]. Although early work primarily indicated Th2-polarization was important in allergic asthma, recent work has shown that Th2 cytokines are involved in the pathogenesis of both allergic and non-allergic asthma[3]. The transcription factor GATA-binding protein 3 (GATA3) is one of the most important regulator of Th2 cytokines[3], and has been found to be up-regulated in both atopic and non-atopic asthma[96]. Interleukin (IL)4 also drives naïve T cells towards a Th2 phenotype, while IL13 and IL4 both promote IgE production which is involved in allergy and allergic asthma[3]. The similar functions of IL4 and IL13 are likely related to the fact that both have affinity for the IL4 receptor (IL4Rα)[97]. Also, signal transducer and activator of transcription (STAT)-6 and IL4R are upregulated among both atopic and non-atopic asthma, although both appear to be more highly expressed in atopic asthma[96]. *STAT6* expression is induced by IL13 and IL4 activation of IL4Rα, and has consistently been associated with high IgE and childhood asthma[97]. IL3, IL4 and STAT6 may also be involved the development of airway airway and hypertrophy of smooth muscle[98]. Recent research has found that *IL13*, but not *IL4*, may also be expressed from non-Th2 cells but still confer type-2-like immune response[99].

Genetic variations in *IL4*[95,100], *IL4R*[100], *IL13*[95], *STAT6*[101], and *GATA3*[102] genes have previously been associated with asthma and allergic disease. More recently, DNA-M variations within these genes (*IL4*, *IL4R*, *IL13*, *STAT6*, and *GATA3*)[1,35,103,104] have been associated T-cell polarization and/or risk of asthma. We have also shown that the effects of DNA-M on risk of allergic disease can be dependent on genetic variations in

SNPs[35,36,103], and that risk of asthma and asthma transitions were associated with interactions between CpG sites and SNPs in the Th2 pathway[104]. However, it is still largely unknown whether such variations within this pathway lead to an overall underlying asthma trait, or to specific clinical or physiological characteristics of asthma.

We hypothesized that persons with clinically similar asthma and wheezing illness would share similar underlying genetic and epigenetic variations within the Th2 path. To test this hypothesis, we compared DNA-M levels and genotypes among selected CpG sites and SNPs within *GATA3*, *STAT6*, *IL4*, *IL4R*, and *IL13* among those without wheeze and those within young adult wheeze clusters. Furthermore, we aimed to identify whether a subset of genetic and epigenetic variations that could most effectively distinguish between young adult wheeze clusters.

3.2 Methods:

The Isle of Wight birth cohort

The Isle of Wight (IOW) birth cohort was established to study the natural history of asthma and allergies in children born between January 1, 1989 and February 28, 1990 in Isle of Wight, UK. The study was approved by the local research ethics committee (now named the National Research Ethics Service, NRES Committee South Central – Southampton B, 06/Q1701/34) and written informed consent was provided by the infants' parents. Details about the birth cohort have been described in detail elsewhere[46,47]. After exclusion of adoptions, prenatal deaths and refusals, 1,456 children were enrolled, and followed-up at 1 (n=1,167; 80.2%), 2 (n=1,174; 80.6%), 4 (n=1,218; 83.7%), 10 (n=1,373; 94.3%), and 18 (n=1,313; 90.2%) years of age. At each follow-up, participants were administered detailed questionnaires and evaluated for manifestations of allergic

disease. Questionnaires included the International Study of Asthma and Allergies in Childhood (ISAAC)[5] as well as study-specific questions about allergic disease and relevant risk factors . Most of those who attended the follow-up visit in-person were also assessed for spirometry, BDR, BHR, FeNO, and skin prick tests (SPTs). At the 18-year follow-up, a random subset of participants was selected to take part in epigenetic screening (n=368).

Dependent Variables

The primary dependent variable for young adult wheeze was a nominal variable with five levels; the details of the clustering used to create this variable are described in elsewhere[4]. Participants with asthma or wheezing illness at age 18 (n=309) were clustered on 13 variables via k-means clustering. Six distinct clusters were identified. Because cluster six (C6) was under-sampled during the random selection of the epigenetic sub-sample (described below), we selected clusters one through five for analysis in this study.

DNA Methylation

For epigenetic screening, blood samples were collected at the 18 year follow-up. DNA was extracted from whole blood using a standard salting out procedure [51]. DNA concentration was determined by the PicoGreen dsDNA quantitation kit (Molecular Probes, Inc., OR, USA). One microgram of DNA was bisulfite-treated for cytosine to thymine conversion using the EZ 96-DNA methylation kit (Zymo Research, CA, USA), following the manufacturer's standard protocol. Genome-wide DNA methylation was assessed using the Illumina Infinium HumanMethylation450K BeadChip (Illumina, Inc., CA, USA), which interrogates >484,000 CpG sites, regions of DNA where a cytosine

nucleotides are followed by a guanine nucleotide, associated with approximately 24,000 genes. The BeadChips were scanned using a BeadStation, and the methylation levels (β value, described below) were calculated for each queried CpG locus using the Methylation Module of BeadStudio software. Arrays were processed using a standard protocol as described elsewhere[52], with multiple identical control samples assigned to each bisulphite conversion batch to assess assay variability and samples were randomly distributed on microarrays to control against batch effects.

Data Cleaning

The program for data cleaning was written in R (R Development Core Team, 2012). Quality control (QC) measures were employed to improve the reliability of data prior to analysis. In our study, the detection *P*-value reported by BeadStudio (Illumina software to process raw intensities) was used as a QC measure of probe performance. Probes whose detection *P*-values > 0.01 in >10% of the samples were removed[53]. The methylation data were then preprocessed and technical variations removed via peak-correction using the Bioconductor IMA (Illumina methylation analyser) package[55]. Dropping control probes and probes with poor detection *P*-values. The arrays were processed in three different batches; batch number was recorded as a categorical variable which was used in ComBat to adjust for inter-array variation[54]. Methylation levels for each queried CpG were calculated as beta (β) values. These represent the proportions of methylated probes for each specific CpG site and can be interpreted as percent methylation.

SNP Assessments

We selected SNPs potentially related to asthma within five genes in the Th2 pathway (*IL4*, *IL4R*, *IL13*, *GATA3*, and *STAT6*). Genotyping was conducted on DNA extracted from blood or saliva samples for 1,211 cohort subjects using GoldenGate Genotyping Assays (Illumina, Inc., CA, USA), the details of which are reported elsewhere[36]. In total, DNA methylation at 107 CpG sites and 42 SNPs were available: 72 CpGs and 17 SNPs in *GATA3*, 12 CpGs and 13 SNPs in *IL4R*, 9 CpGs and 7 SNPs in *IL13*, 5 CpGs and 4 SNPs in *IL4*, and 9 CpGs and 1 SNP in *STAT6*.

Statistical Analyses

All statistical analyses were carried out in R. Due to the relatively small sample size and large number of potential predictors, we used unsupervised data-reduction techniques to reduce the number CpG sites and SNPs for analyses. For CpG sites, we first explored potential outliers; observations that appeared to be erroneous or substantially different from the distribution were recoded as missing. This resulted in 13 observations within 11 CpG sites being recoded to missing. Then we selected the most variable sites (standard deviation > 0.02), resulting in 21 CpG sites for analyses. For SNPs, we identified linkage disequilibrium (LD) blocks using Haploview[105], and only included those SNPs that were not included in a block as well as one SNP from each block[35,103] as a representative for that block. When applicable, we preferentially included SNPs from LD blocks that had been previously studied in relation to asthma, lung-function or allergic disease.

We then explored whether methylation levels within *GATA3* (10 CpGs), *IL4R* (3 CpGs), *STAT6* (1 CpG), *IL13* (5 CpGs), and *IL4* (2 CpGs) varied by wheeze-clusters via

ANOVA models. Prior to implementing the ANOVA models, we converted the methylation β values into M-values via $\log_2(\beta / (1-\beta))$ which more closely approximate a normal distribution[57]. We also explored whether genetic variation within *GATA3* (8 SNPs), *IL4R* (9 SNPs), *STAT6* (1 SNPs), *IL13* (4 SNPs), and *IL4* (2 SNPs) differed by wheeze-cluster via χ^2 tests. Then, to see whether there was consistency between wheeze cluster associations and wheeze vs non-wheeze associations, we collapsed clusters of wheeze (C1-C5) into a single group of persons with wheeze, and compared them to persons without wheeze at the same 24 SNPs and 21 CpG sites, using χ^2 tests and t-tests respectively.

Lastly, among the SNPs and CpG sites that showed differential variations among the wheeze clusters, we produced a conditional inference tree[106] to explore whether a set of genetic and epigenetic loci, rather than individual loci, may be more effective at distinguishing between the different wheeze clusters. Many tree-based methods exist; we chose to use conditional inference trees (CTREE) over classification and regression trees (CART) or recursive partitioning and regression trees (RPART) because: 1) their variable selection process is based on tests of statistical significance rather than information, and 2) their variable selection process is not biased towards selecting continuous predictors or predictors with more possible cut-points[107]. Nonetheless, as with other tree-based methods, conditional inference trees are biased towards correctly classifying the majority class in a highly imbalanced dataset. To minimize this bias, we excluded persons without wheeze at age 18 from our analysis.

3.3 Results:

The five clusters represented unique combinations of physiologic, clinical, and symptomatic characteristics (Table 3.1). Cluster 1 (C1) represented a childhood-onset wheeze group with the best lung-function metrics. Cluster 2 (C2) was a childhood-onset wheeze group with intermediate lung function metrics and intermediate proportions allergic diseases. Cluster 3 (C3) was a mostly female, childhood-onset wheeze group with high prevalence of allergic disease (atopy, rhinitis, and eczema), the most impaired lung-function, and high BDR and FeNO. Cluster 4 (C4) was a mostly female, latest-onset wheeze group with the low prevalence of atopy, and low BDR, IgE and FeNO. Cluster 5 (C5) was a mostly female, late-childhood onset wheeze group with normal lung-function and intermediate allergic disease prevalence.

Due to the small sample sizes within each cluster, we reduced the number of tests performed on SNPs by identifying blocks of SNPs in LD analyses. Then we only performed analyses on one SNP per block as well as SNPs outside of LD-blocks (Figure 3.1, Figure 3.2, and Figure 3.3), resulting in 24 SNPs for analyses. In univariate results (Table 3.2), only one SNP (rs1058240) showed significant variation across the six groups (*P*-value = 0.0075), although this finding was not robust under Bonferroni adjustment for 24 tests ($\alpha = 0.002$). For these analyses, GG and AG were analyzed together as one class since there were only 9 participants with GG genotype, all of which were among those without wheeze. The clusters ‘C2’ and ‘C3’ had the highest proportions of GG/AG (52.6% and 71.4% respectively). Cluster ‘C5’, had the lowest proportion of GG/AG (9.1%). Clusters ‘C1’ and ‘C4’ had proportions of GG/AG (35.7% and 37.5%% respectively) similar to those without any wheeze at age 18 (32.7%).

We also reduced the number of CpG sites for analyses by only selecting the most variable CpGs (those with a standard deviation ≥ 0.02), resulting in 21 CpG sites for analyses. In univariate results (Table 3.3), seven CpG sites showed significant variation across the six groups: cg25368824 in the TSS200 of *IL4* (*P*-value = 0.00005), cg12377972 in the 5'UTR of *IL4* (*P*-value = 0.049), cg06584121 in the TSS200 of *IL13* (*P*-value = 0.0032), cg131993853 also in the TSS200 of *IL13* (*P*-value = 0.0031), as well as three CpG sites with the body of *GATA3* cg12181459 (*P*-value = 0.024), cg25630514 (*P*-value = 0.023), and cg00463367 (*P*-value = 0.00029). Only cg25368824 and cg00463367 survived the Bonferroni adjustment for 21 tests ($\alpha = 0.0024$).

Among persons without wheeze at age 18 as a reference group, cluster 'C4' had the most unique epigenetic pattern. We found that cluster 'C4' was consistently different from those free of wheeze; it had significantly higher methylation at cg25368824 (*P*-value = 0.0063), and significantly lower methylation levels at cg06584121 (*P*-value = 0.0010), cg06967316 (0.0015), cg1218159 (*P*-value = 0.0015), cg25630514 (*P*-value = 0.0013) and cg00463367 (*P*-value = 0.00004).

Then we tested whether a set of the nominally significant (*P*-values < 0.05) CpGs and/or SNPs could distinguish between the five clusters of participants with wheeze using conditional inferences trees (Figure 3.4). We found that two predictors could significantly distinguish some, but not all, clusters. Among persons with wheeze, those with methylation levels ≤ 0.579 at cg25630514 (*P*-value = 0.019) were most likely to belong to cluster 4 (47%), whereas those with methylation > 0.579 at cg25630514 were most likely to belong to cluster 5 (51%) if they also had genotype AA at rs1058240 and most

likely to belong to cluster 2 (46%) if they had genotype GG or AG at rs1058240 (P-value = 0.021).

Last, we investigated whether the same CpG sites and SNPs identified in the above analyses would have been identified if our outcome variable were a dichotomous measure of person with wheeze (combined C1-C5, n=75) vs. persons without wheeze at age 18 (n=293). Only one CpG site was found to have differential methylation between person with and without wheeze, whereas 3 SNPs were found to have differential genetic variation. Average methylation at cg12377972 within *IL4* was lower among persons with wheeze compared to those without wheeze (mean and standard deviation = 0.906 (0.02) vs 0.913 (0.02), respectively; T-test *P*-value = 0.031). Three other SNPs, all within *IL4R*, did show nominal variation between those with and without wheeze. At rs1110470, persons with wheeze were less likely to have the minor allele compared to those without wheeze, AA (16.4% vs 23.7%, respectively) or AG (44.8% vs 52.6%, respectively) (χ^2 *P*-value = 0.04). At rs3024604, persons with wheeze were more likely than those without to have the minor allele GG or AG (23.2% vs 21.1%, respectively; χ^2 *P*-value = 0.03). Lastly, at rs8832, persons with wheeze were more likely than person without wheeze (χ^2 *P*-value = 0.04) to be homozygous for the minor allele AA (32.4% vs 19.1%, respectively), similar for heterozygous AG (43.7% vs 47.7%, respectively), and less likely to be homozygous for the major allele GG (23.9% vs 33.2%, respectively).

3.4 Discussion:

We found that there were epigenetic variations at three CpGs (cg12181459, cg25630514, and cg00463367) in the body of *GATA3*, one CpG (cg25368824) within 200 base pairs (bp) of the transcription start site (TSS200) of *IL4*, one CpG (cg12377972)

within the 5' untranslated region (5'UTR) of *IL4*, and two CpG (cg06584121 and cg06967316) within the TSS200 of *IL13*, as well as genetic variation at one SNP (rs1058240) within the 3'UTR of *GATA3* that were dependent on the type of wheeze as determined by clustering on 13 clinical characteristics. Only cg25368824 and cg00463367 maintained statistical significance after Bonferroni-adjustment for multiple testing.

We also observed greater epigenetic variation when comparing the clusters to each other (7 CpG sites and 1 SNP were nominally significant) and greater genetic variation when comparing those with and without wheeze overall (3 SNPs and 1 CpG site were nominally significant). Additionally, only cg12377972 was statistically significant in both analyses, likely because those without wheeze had the highest methylation levels, while all of the wheeze clusters had varying levels of methylation but all lower than that of the non-wheeze group. For all other nominally significant sites, some wheeze clusters had higher average methylation when compared to non-wheeze while others had lower average methylation. These findings support the idea that asthma and wheeze are heterogeneous diseases with different symptomatic profiles which may result from various unique pathophysiology[2].

The cluster (C4) characterized by being mostly female, late-onset wheeze with low prevalence of allergic diseases and low levels for BDR, FeNO and serum IgE, was most different from the other clusters across the epigenetic loci. We did not have expression of these genes, but could assume that epigenetic regulation of expression at the nominally significant CpG sites follows the common paradigm: DNA-M is negatively correlated with expression in promotors and positively correlated with expression in gene

bodies[32]. Then we would expect persons in this cluster to have lower expression of *IL4*, higher expression of *IL13*, and lower expression of *GATA3*. *IL4*, *IL13*, and *GATA3* all promote Th2 differentiation[108,109], but differ somewhat in their activities. GATA3 is often referred to as the master regulator of Th2 differentiation, as it is a transcription factor that up-regulates the production of other pro-Th2 cytokines[110]. Though IL4 and IL13 are structurally similar and share the same receptor (IL4R), IL4 appears to be more important in the polarization of Th2 phenotype, while *IL13* plays its role in the effector phase[109]. There is also new research showing that while IL13 and IL4 are both produced in Th2 cells, but only IL13 is produced from type-2 innate lymphoid cells (ILC2)[110], a cell type that has only recently been implicated and asthma and allergic disease[99]. Despite relatively few studies of the activity of ILC2s in humans, they have been suggested to play a role in both Th2-like asthma and severe asthma[99].

This has important implications for generalizing findings or for conducting validation studies of wheeze-illness in independent samples. As suggested by others[18,26,27], the ability to replicate a finding likely depends on whether the asthmatics being studied have similar distributions of many characteristics (e.g., time of onset, measures of lung function, bronchial reactivity, amount of treatment being taken, and prevalence of allergic sensitization).

Interestingly, our main finding among SNPs (rs1058240) has received some attention in previous studies of allergic disease. Within the IOW this SNP was associated with rhinitis and allergic rhinitis[111] while others observed that increasing frequency of the minor allele was associated with decreased allergic sensitization and that presence of the major allele (G) was associated with susceptibility for asthma and atopic asthma (Slager

2010). More recently, Yang et al. showed that rs1058240 is at the binding site for three different microRNAs (miRNA), and that genetic variations at this SNP were associated with differential *GATA3* expression, where AA variants had significantly higher expression when compared with either AG or GG variants[112]. This could be an example of genetic (SNP) and epigenetic (miRNA) interaction influencing disease. Interesting follow-up work could investigate the combined influence of miRNA expression, GATA3 expression, and polymorphisms at rs1058240, on various wheeze phenotypes.

These findings should be interpreted within the limitations of this study. This work was exploratory, and the observed associations are limited in their generalizability but should new hypotheses for further investigation. One limitation was the use of k-means clustering on 13 characteristics to define the outcome variable. This would be difficult to reproduce within an independent cohort due to variations in such characteristics across populations. Second, the clusters within our epigenetic sample had small sample sizes (minimum cluster sample size = 7), and thus may not be very representative of persons with similar characteristics. Future studies in larger samples would provide stronger evidence on whether variations in Th2-path genes are indeed variable by sets of multiple physiologic, clinical and symptomatic characteristics. Because this analysis involved a sub-sample of the original data, we compared average levels and proportions of the clinical characteristics among persons with wheeze in our epigenetic sub-sample (n=75) to those in the full cohort (n= 279). Overall, the sub-sample was no different from the full sample for proportions of females, or prevalence of rhinitis, eczema, or atopy (χ^2 P-values > 0.05), as well as for average measures of FEV1, FVC, FEV1/FVC, FEF2575, BDR, FeNO, total IgE, BTS Step, or age of onset (T-test P-values

> 0.05), so it is unlikely that selection of the epigenetic sub-sample biased the observed associations. Despite these limitations, we believe follow-up investigations of these findings are warranted given that phenotypic-heterogeneity of asthma, and the likely corresponding heterogeneity of underlying mechanisms, has been long recognized but often studied with individual characteristics of wheeze-illness rather than overall wheeze phenotype based on multiple characteristics. These findings suggest that using a complex outcome variable, such as young adult wheeze clusters, may lead to new findings but it is still unclear whether those findings can be generalizable and offer clinical utility.

In summary, we found that DNA-M varied based on different wheeze-illness phenotypes at cg25630514 in the body of *GATA3* and at cg25368824 in the TSS200 of *IL4*, suggesting that epigenetic regulation of these cytokines differs among persons with wheeze that have different physiologic, clinical and symptomatic characteristics. We also found that genetic variation at rs1058240 in the 3'UTR of *GATA3*, which is at the binding site for several miRNAs, differed among persons with different wheeze phenotypes. Given that this study was exploratory, these associations should be investigated in independent study populations.

Table 3.1: Prevalence and average values of physiologic, clinical, and symptomatic characteristics of the five wheeze clusters (n=75).

	C1 (n=14)	C2 (n=20)	C3 (n=7)	C4 (n=9)	C5 (n=25)
Categorical Factors					
	n (%)	**n (%)**	**n (%)**	**n (%)**	**n (%)**
Female	7 (50.0)	12 (60.0)	6 (85.7)	8 (88.9)	18 (72.0)
Rhinitis	12 (85.7)	10 (50.0)	7 (100.0)	5 (55.6)	12 (48.0)
Eczema	3 (21.4)	3 (15.0)	2 (33.3.0)	1 (11.1)	3 (12.0)
Atopy	12 (85.7)	9 (45.0)	6 (100.0)	2 (25.0)	12 (48.0)
Continuous Factors					
	Mean (SD)	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**
FEV_1	4.57 (0.85)	3.58 (0.77)	3.33 (0.82)	3.12 (0.35)	3.96 (0.75)
FVC	5.14 (1.10)	4.74 (1.10)	4.50 (0.84)	3.88 (0.35)	4.29 (0.62)
FEV_1/FVC	0.90 (0.10)	0.77 (0.11)	0.74 (0.08)	0.81 (0.12)	0.92 (0.11)
$FEF_{25:75}$	4.93 (0.62)	3.47 (0.61)	2.5 (0.84)	3.62 (0.52)	5.00 (1.35)
BDR	4.57 (1.96)	6.96 (3.63)	12.45 (3.79)	2.76 (3.78)	2.62 (1.73)
FeNO	64.29 (34.22)	26.22 (28.93)	64.33 (22.61)	13.38 (5.63)	26.17 (35.71)
Serum IgE	810.2 (780.3)	217.5 (342.5)	439.0 (237.6)	84.8 (31.4)	257.1 (535.8)
BTS Step	2.29 (0.95)	1.86 (0.69)	2 (0.71)	1.33 (0.58)	1.83 (0.75)
Onset of Wheeze	6.36 (5.53)	4.53 (4.65)	5 (3.85)	14.56 (2.35)	10.22 (5.84)

FEV_1: Forced expiratory volume (L) in one second (s).
FVC: Forced vital capacity (L).
FEV1/FVC: Ratio of FEV_1 to FVC.
$FEF_{25\text{-}75\%}$: Forced expiratory flow 25-75% (L/s).
BDR: Bronchodilator reversibility; percent change in FEV_1 post-salbutamol (600mg) challenge.
FeNO: Log_{10} of fractional exhaled nitric oxide (ppb).
Serum IgE: Log_{10} of total IgE.
BTS Step: British Thoracic Society step of therapy (0-4), higher values reflect higher levels of treatment.
Onset of Wheeze: Age (in years) when wheeze first appeared.

Table 3.2: Proportions of genetic variants among selected Th2 SNPs within young adult wheeze clusters.

SNP Annotations		**Comparing genetic variation across 6 groups**							
***GATA3* SNPs**	**Genotype**	**Total n (%)**	**Wheeze-free n (%)**	**C1 n (%)**	**C2 n (%)**	**C3 n (%)**	**C4 n (%)**	**C5 n (%)**	**P-Val.**
rs2229359	AAorAG	47 (13.5%)	41 (14.6%)	0 (0%)	1 (5.6%)	2 (28.6%)	1 (12.5%)	1 (4.8%)	0.28
	GG	302 (86.5%)	239 (85.4%)	14 (100%)	17 (94.4%)	5 (71.4%)	7 (87.5%)	20 (95.2%)	
rs3802604	AA	144 (42.1%)	111 (40.7%)	4 (28.6%)	10 (52.6%)	1 (14.3%)	6 (75%)	12 (60%)	0.16
	AG	149 (43.6%)	121 (44.3%)	8 (57.1%)	7 (36.8%)	4 (57.1%)	1 (12.5%)	8 (40%)	
	GG	49 (14.3%)	41 (15%)	2 (14.3%)	2 (10.5%)	2 (28.6%)	1 (12.5%)	0 (0%)	
rs568727	AA	34 (10.5%)	30 (11.6%)	1 (7.7%)	1 (5.6%)	1 (16.7%)	0 (0%)	0 (0%)	0.46
	AC	142 (43.7%)	111 (43%)	8 (61.5%)	8 (44.4%)	4 (66.7%)	2 (28.6%)	9 (40.9%)	
	CC	149 (45.8%)	117 (45.3%)	4 (30.8%)	9 (50%)	1 (16.7%)	5 (71.4%)	13 (59.1%)	
rs3802600	AAorAT	115 (33.1%)	97 (35.1%)	6 (42.9%)	2 (10.5%)	3 (42.9%)	2 (25%)	5 (22.7%)	0.19
	TT	232 (66.9%)	179 (64.9%)	8 (57.1%)	17 (89.5%)	4 (57.1%)	6 (75%)	17 (77.3%)	
rs422628	AA	196 (56.3%)	156 (56.3%)	8 (57.1%)	9 (47.4%)	3 (42.9%)	5 (62.5%)	15 (68.2%)	0.95
	AG	134 (38.5%)	105 (37.9%)	6 (42.9%)	9 (47.4%)	4 (57.1%)	3 (37.5%)	7 (31.8%)	
	GG	18 (5.2%)	16 (5.8%)	0 (0%)	1 (5.3%)	0 (0%)	0 (0%)	0 (0%)	
rs1058240	**GGorAG**	**114 (33.3%)**	**89 (32.7%)**	**5 (35.7%)**	**10 (52.6%)**	**5 (71.4%)**	**3 (37.5%)**	**2 (9.1%)**	**0.0075**
	AA	**228 (66.7%)**	**183 (67.3%)**	**9 (64.3%)**	**9 (47.4%)**	**2 (28.6%)**	**5 (62.5%)**	**20 (90.9%)**	
rs434645	AAorAG	98 (28.6%)	73 (26.8%)	5 (35.7%)	9 (47.4%)	4 (57.1%)	2 (25%)	4 (18.2%)	0.15
	GG	245 (71.4%)	199 (73.2%)	9 (64.3%)	10 (52.6%)	3 (42.9%)	6 (75%)	18 (81.8%)	
rs12412241	AA	33 (9.6%)	30 (10.9%)	0 (0%)	1 (5.3%)	0 (0%)	1 (12.5%)	1 (4.8%)	0.53
	AG	138 (40%)	106 (38.5%)	4 (28.6%)	9 (47.4%)	5 (71.4%)	5 (62.5%)	8 (38.1%)	
	GG	174 (50.4%)	139 (50.5%)	10 (71.4%)	9 (47.4%)	2 (28.6%)	2 (25%)	12 (57.1%)	

Table 3.2: Proportions of genetic variants among selected Th2 SNPs within young adult wheeze clusters *(cont'd)*.

SNP Annotations		**Comparing genetic variation across 6 groups**							
IL4R **SNPs**	**Genotype**	**Total n (%)**	**Wheeze-free n (%)**	**C1 n (%)**	**C2 n (%)**	**C3 n (%)**	**C4 n (%)**	**C5 n (%)**	**P-Val.**
rs2057768	AA	23 (6.6%)	15 (5.4%)	1 (7.1%)	4 (21.1%)	0 (0%)	0 (0%)	3 (13.6%)	0.19
	AG	151 (43.1%)	122 (43.7%)	7 (50%)	8 (42.1%)	3 (42.9%)	1 (12.5%)	10 (45.5%)	
	GG	176 (50.3%)	142 (50.9%)	6 (42.9%)	7 (36.8%)	4 (57.1%)	7 (87.5%)	9 (40.9%)	
rs1110470	AA	75 (22.3%)	64 (23.7%)	2 (14.3%)	4 (21.1%)	1 (14.3%)	2 (25%)	2 (11.1%)	0.19
	AG	172 (51%)	142 (52.6%)	7 (50%)	8 (42.1%)	4 (57.1%)	5 (62.5%)	5 (27.8%)	
	GG	90 (26.7%)	64 (23.7%)	5 (35.7%)	7 (36.8%)	2 (28.6%)	1 (12.5%)	11 (61.1%)	
rs3024604	GGorAG	49 (14.3%)	33 (12.1%)	3 (21.4%)	4 (21.1%)	2 (28.6%)	2 (25%)	4 (20%)	0.21
	AA	293 (85.7%)	240 (87.9%)	11 (78.6%)	15 (78.9%)	5 (71.4%)	6 (75%)	16 (80%)	
rs3024622	CC	38 (11.6%)	28 (10.6%)	1 (7.7%)	5 (29.4%)	0 (0%)	0 (0%)	4 (21.1%)	0.28
	CG	140 (42.8%)	114 (43.3%)	8 (61.5%)	7 (41.2%)	3 (50%)	3 (37.5%)	5 (26.3%)	
	GG	149 (45.6%)	121 (46%)	4 (30.8%)	5 (29.4%)	3 (50%)	5 (62.5%)	10 (52.6%)	
rs4787423	GGorAG	78 (22.3%)	61 (21.9%)	3 (21.4%)	4 (21.1%)	3 (42.9%)	1 (12.5%)	5 (22.7%)	0.85
	AA	271 (77.7%)	217 (78.1%)	11 (78.6%)	15 (78.9%)	4 (57.1%)	7 (87.5%)	17 (77.3%)	
rs1805011	CCorCA	74 (21.4%)	60 (21.7%)	3 (21.4%)	4 (22.2%)	2 (28.6%)	1 (12.5%)	4 (18.2%)	0.99
	AA	272 (78.6%)	216 (78.3%)	11 (78.6%)	14 (77.8%)	5 (71.4%)	7 (87.5%)	18 (81.8%)	
rs8832	AA	76 (21.8%)	53 (19.1%)	5 (35.7%)	9 (47.4%)	1 (14.3%)	1 (12.5%)	7 (31.8%)	0.16
	AG	163 (46.8%)	132 (47.7%)	7 (50%)	5 (26.3%)	5 (71.4%)	4 (50%)	10 (45.5%)	
	GG	109 (31.3%)	92 (33.2%)	2 (14.3%)	5 (26.3%)	1 (14.3%)	3 (37.5%)	5 (22.7%)	
rs12102586	AAorAG	59 (16.9%)	44 (15.8%)	5 (35.7%)	3 (15.8%)	1 (14.3%)	2 (25%)	4 (18.2%)	0.45
	GG	290 (83.1%)	234 (84.2%)	9 (64.3%)	16 (84.2%)	6 (85.7%)	6 (75%)	18 (81.8%)	
rs16976728	AA	64 (18.4%)	51 (18.3%)	2 (14.3%)	5 (27.8%)	0 (0%)	1 (12.5%)	5 (22.7%)	0.69
	AG	152 (43.7%)	118 (42.4%)	7 (50%)	7 (38.9%)	6 (85.7%)	3 (37.5%)	11 (50%)	
	GG	132 (37.9%)	109 (39.2%)	5 (35.7%)	6 (33.3%)	1 (14.3%)	4 (50%)	6 (27.3%)	

Table 3.2: Proportions of genetic variants among selected Th2 SNPs within young adult wheeze clusters *(cont'd)*.

SNP Annotations		**Comparing genetic variation across 6 groups**							
STAT6 **SNPs**	**Genotype**	**Total n (%)**	**Wheeze-free n (%)**	**C1 n (%)**	**C2 n (%)**	**C3 n (%)**	**C4 n (%)**	**C5 n (%)**	**P-Val.**
rs1059513	GGorAG	66 (19%)	52 (18.6%)	4 (28.6%)	5 (26.3%)	0 (0%)	1 (14.3%)	3 (14.3%)	0.65
	AA	282 (81%)	227 (81.4%)	10 (71.4%)	14 (73.7%)	7 (100%)	6 (85.7%)	18 (85.7%)	
IL4R **SNPs**	**Genotype**	**Total n (%)**	**Wheeze-free n (%)**	**C1 n (%)**	**C2 n (%)**	**C3 n (%)**	**C4 n (%)**	**C5 n (%)**	**P-Val.**
rs1881457	CCorAC	119 (34.2%)	96 (34.7%)	5 (35.7%)	8 (42.1%)	2 (28.6%)	2 (25%)	6 (27.3%)	0.94
	AA	229 (65.8%)	181 (65.3%)	9 (64.3%)	11 (57.9%)	5 (71.4%)	6 (75%)	16 (72.7%)	
rs1800925	AAorAG	121 (34.9%)	97 (34.9%)	6 (42.9%)	9 (47.4%)	2 (28.6%)	1 (12.5%)	6 (30%)	0.60
	GG	226 (65.1%)	181 (65.1%)	8 (57.1%)	10 (52.6%)	5 (71.4%)	7 (87.5%)	14 (70%)	
rs20541	AAorAG	119 (34.1%)	99 (35.5%)	3 (21.4%)	6 (31.6%)	2 (28.6%)	3 (37.5%)	6 (28.6%)	0.92
	GG	230 (65.9%)	180 (64.5%)	11 (78.6%)	13 (68.4%)	5 (71.4%)	5 (62.5%)	15 (71.4%)	
rs2243204	AAorAG	63 (18.1%)	50 (17.9%)	2 (14.3%)	6 (31.6%)	1 (14.3%)	1 (12.5%)	3 (14.3%)	0.75
	GG	286 (81.9%)	229 (82.1%)	12 (85.7%)	13 (68.4%)	6 (85.7%)	7 (87.5%)	18 (85.7%)	
IL4R **SNPs**	**Genotype**	**Total n (%)**	**Wheeze-free n (%)**	**C1 n (%)**	**C2 n (%)**	**C3 n (%)**	**C4 n (%)**	**C5 n (%)**	**P-Val.**
rs2070874	AAorAG	84 (24.3%)	73 (26.4%)	3 (21.4%)	2 (10.5%)	1 (14.3%)	2 (28.6%)	3 (14.3%)	0.57
	GG	262 (75.7%)	204 (73.6%)	11 (78.6%)	17 (89.5%)	6 (85.7%)	5 (71.4%)	18 (85.7%)	
rs2243263	GGorCG	64 (18.4%)	54 (19.5%)	2 (14.3%)	3 (15.8%)	2 (28.6%)	1 (12.5%)	2 (9.1%)	0.81
	CC	284 (81.6%)	223 (80.5%)	12 (85.7%)	16 (84.2%)	5 (71.4%)	7 (87.5%)	20 (90.9%)	

SNP, single nucleotide polymorphism; C1, cluster 1; C2, cluster 2; C3, cluster 3; C4, cluster 4; C5, cluster 5; P-Val., p-value from a Fisher's Exact test.

Table 3.3: Average DNA-M levels among selected Th2 CpGs within young adult wheeze clusters.

CpG Annotations				**Mean (sd) within each wheeze cluster**[¥]					
Gene	**CGs**	**CpG Location**	**P-val.**[€]	**Wheeze-free (n=293)**	**C1 (n=14)**	**C2 (n=20)**	**C3 (n=7)**	**C4 (n=9)**	**C5 (n=25)**
IL4	**cg25368824**	**TSS200**	**0.00005**	**0.86 (0.02)**	**0.84 (0.05)**	**0.85 (0.04)**	**0.82 (0.04)**	**0.88 (0.02)**	**0.87 (0.03)**
	cg12377972	**5'UTR; 1stExon**	**0.049**	**0.91 (0.02)**	**0.90 (0.03)**	**0.90 (0.02)**	**0.89 (0.03)**	**0.91 (0.01)**	**0.91 (0.02)**
IL4R	cg26937798	5'UTR	0.15	0.08 (0.02)	0.08 (0.02)	0.07 (0.03)	0.07 (0.02)	0.05 (0.01)	0.07 (0.02)
	cg16649560	5'UTR	0.076	0.19 (0.05)	0.17 (0.03)	0.17 (0.05)	0.18 (0.03)	0.24 (0.07)	0.18 (0.04)
	cg01165142	Body	0.42	0.69 (0.05)	0.69 (0.05)	0.70 (0.05)	0.71 (0.05)	0.72 (0.09)	0.69 (0.04)
IL13	cg04303330	TSS1500	0.82	0.29 (0.04)	0.29 (0.04)	0.29 (0.05)	0.29 (0.03)	0.31 (0.04)	0.30 (0.04)
	cg13566430	TSS1500	0.63	0.20 (0.03)	0.19 (0.02)	0.19 (0.02)	0.18 (0.03)	0.20 (0.02)	0.20 (0.03)
	cg06584121	**TSS200**	**0.0033**	**0.89 (0.02)**	**0.89 (0.01)**	**0.89 (0.02)**	**0.88 (0.02)**	**0.86 (0.02)**	**0.90 (0.01)**
	cg06967316	**TSS200**	**0.0032**	**0.84 (0.02)**	**0.83 (0.02)**	**0.83 (0.03)**	**0.84 (0.02)**	**0.81 (0.05)**	**0.85 (0.02)**
	cg11798521	3'UTR	0.52	0.83 (0.02)	0.83 (0.02)	0.83 (0.02)	0.84 (0.01)	0.82 (0.02)	0.83 (0.02)
GATA3	cg17124583	Body	0.11	0.04 (0.02)	0.05 (0.02)	0.06 (0.02)	0.04 (0.02)	0.04 (0.01)	0.04 (0.02)
	cg12181459	**Body**	**0.023**	**0.39 (0.06)**	**0.41 (0.07)**	**0.40 (0.05)**	**0.40 (0.05)**	**0.32 (0.07)**	**0.39 (0.07)**
	cg11430077	Body	0.16	0.11 (0.04)	0.12 (0.04)	0.12 (0.05)	0.12 (0.04)	0.08 (0.02)	0.11 (0.03)
	cg22770911	Body	0.79	0.52 (0.05)	0.54 (0.03)	0.52 (0.05)	0.51 (0.04)	0.53 (0.04)	0.51 (0.04)
	cg10163955	Body	0.13	0.74 (0.05)	0.75 (0.04)	0.75 (0.05)	0.71 (0.05)	0.76 (0.08)	0.74 (0.04)
	cg04492228	Body	0.33	0.19 (0.04)	0.19 (0.04)	0.19 (0.04)	0.18 (0.04	0.22 (0.08)	0.18 (0.03)
	cg17489908	Body	0.37	0.25 (0.04)	0.24 (0.04)	0.25 (0.05)	0.25 (0.04)	0.28 (0.06)	0.24 (0.04)
	cg22892607	Body	0.073	0.06 (0.02)	0.07 (0.02)	0.07 (0.02)	0.08 (0.04)	0.06 (0.01)	0.06 (0.02)
	cg25630514	**Body**	**0.023**	**0.67 (0.06)**	**0.69 (0.07)**	**0.66 (0.08)**	**0.67 (0.03)**	**0.58 (0.10)**	**0.66 (0.05)**
	cg00463367	**Body**	**0.00029**	**0.19 (0.05)**	**0.22 (0.05)**	**0.20 (0.06)**	**0.21 (0.05)**	**0.13 (0.06)**	**0.18 (0.03)**
STAT6	cg07926491	3'UTR	0.91	0.94 (0.01)	0.94 (0.02)	0.93 (0.02)	0.93 (0.02)	0.94 (0.02)	0.93 (0.02)

[€] ANOVA models used methylation M-values to test for significance of DNA-M variations.

[¥] The mean and sd (standard deviation) were calculated using methylation β-values.

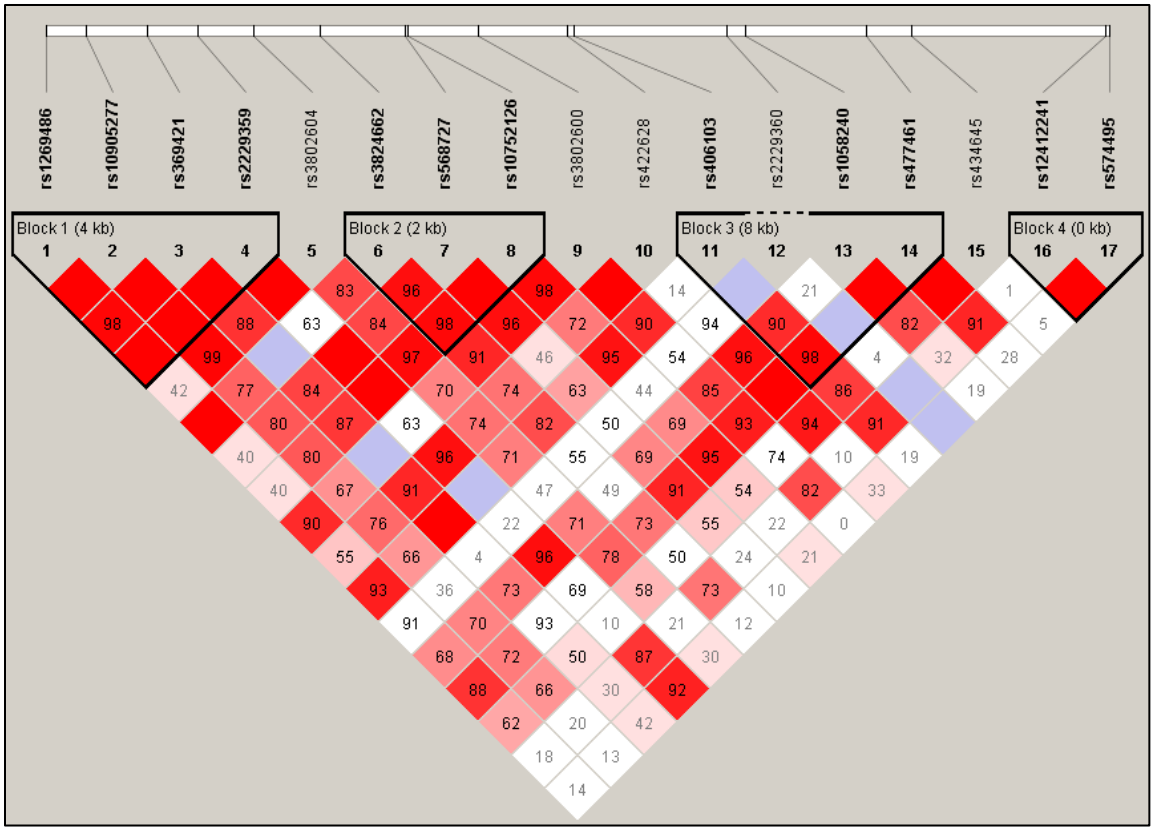

Figure 3.1: LD Plot for *GATA3*

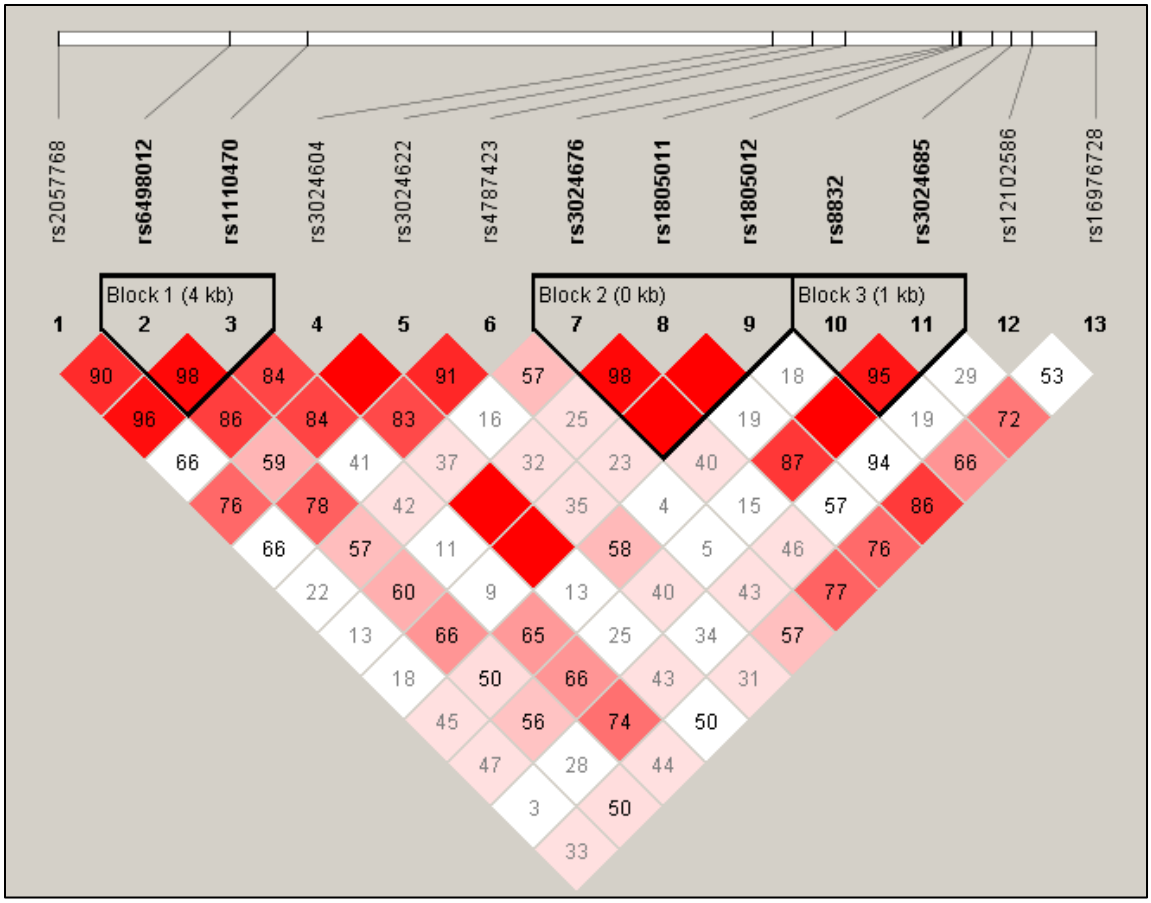

Figure 3.2: LD Plot for *IL4R*

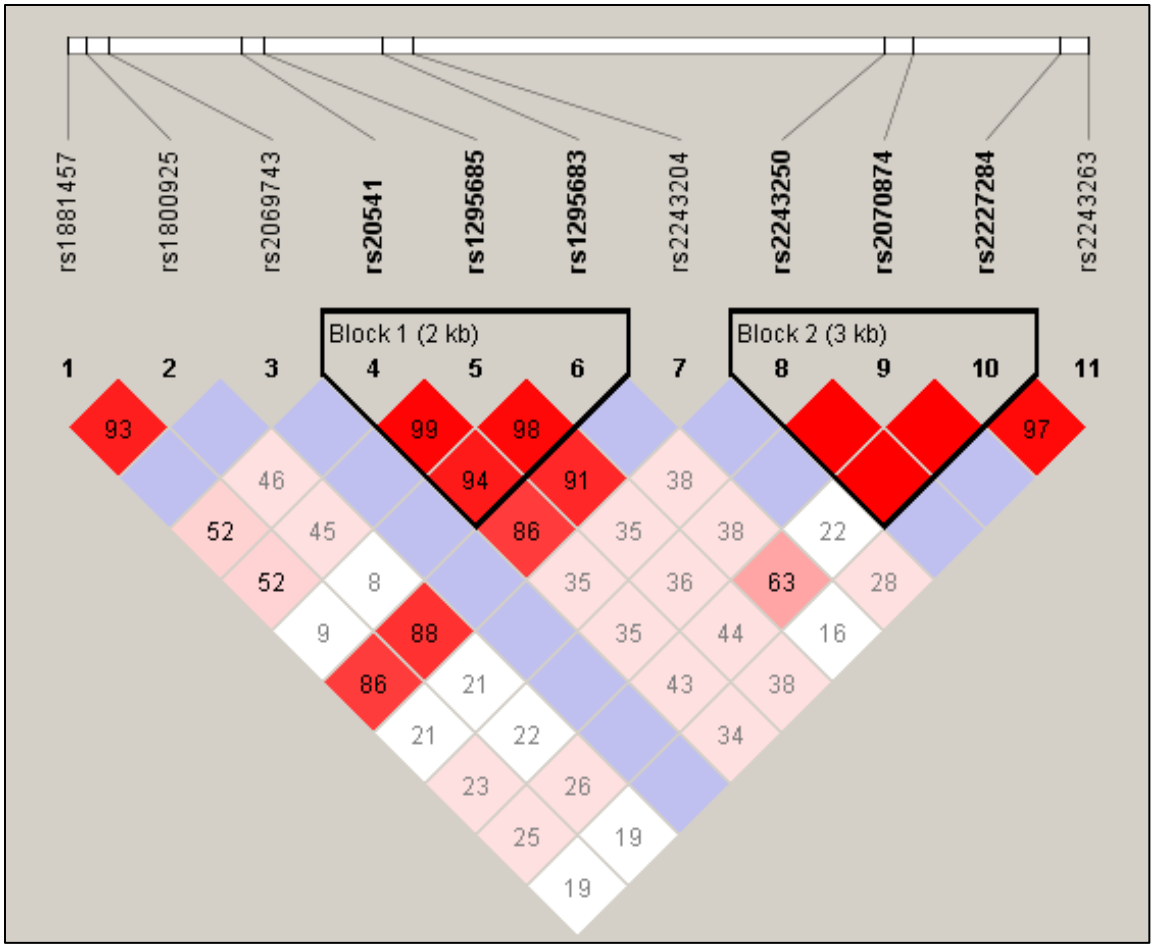

Figure 3.3: LD Plot for *IL4 and IL13.*

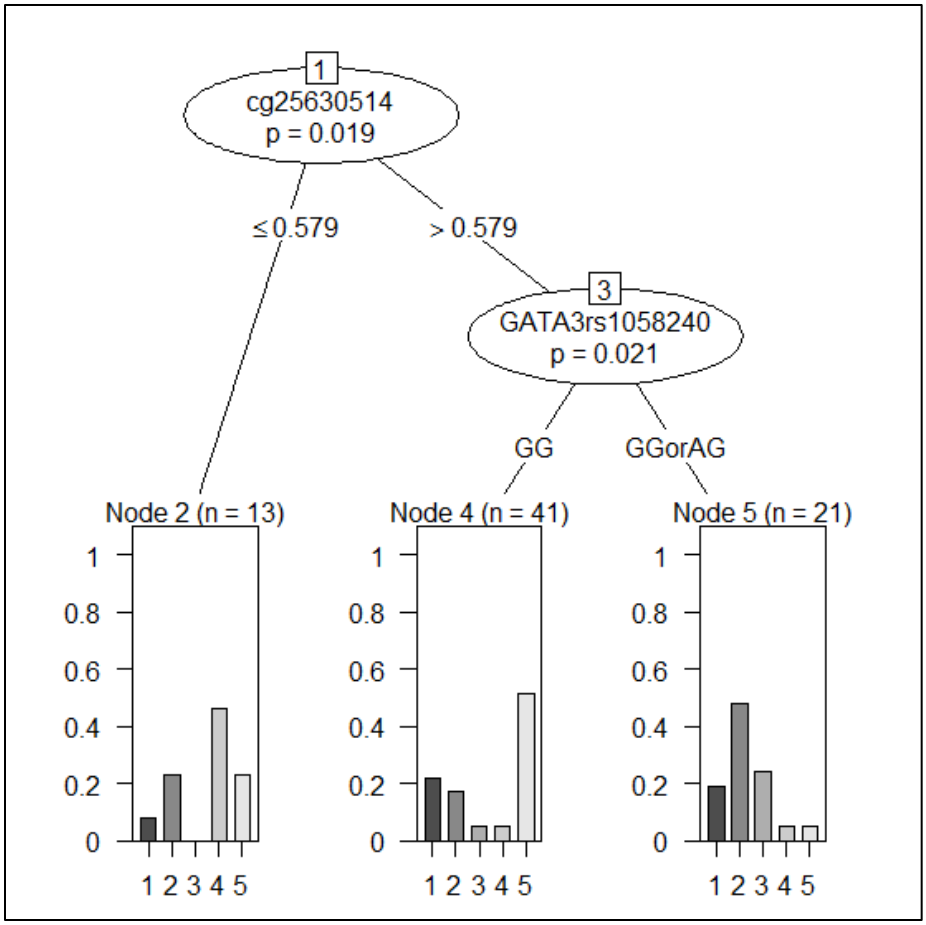

Figure 3.4: Conditional inference tree for classifying wheeze clusters with CpGs and SNPs that were nominally significant via ANOVA and Fisher's Exact Tests.

CHAPTER 4

CORD BLOOD EXPRESSION LEVELS OF THE NOVEL GENES, HK1 AND LITAF, PREDICT WHEEZE WITHIN FIRST OF LIFE

4.1 Introduction:

Wheeze, defined as a high-pitched whistle associated with narrowing of the airways, is common among infants[113] and it is often the defining symptom of asthma at any age. However, many of the mechanisms which make infants particularly susceptible to wheeze, including a more compliant chest wall as well as differences in smooth muscle tone and tracheal cartilage composition, tend to improve as they grow older[113]. Thus, most infant wheeze resolves within the first year after birth, with the majority of these infants not being affected by asthma or recurrent wheeze later in life[114]. However, given that wheezing is an indicator of narrow or obstructed airways at any age, there are undoubtedly similarities between infant wheeze, particularly that which occurs without colds, and later-life wheezing-illnesses such as asthma[115]. In fact, childhood wheeze unrelated to cough or cold is an important criteria in the asthma predictive index (API), as well as later modifications of the API, which use parental and early-life characteristics to predict childhood diagnosis of asthma[116]. Thus, it is important to not only identify which phenotypic presentations of infant wheeze are predictive of later life asthma, but also to identify molecular mechanisms that are involved in both infant and later-life wheeze illnesses, as these may help us to distinguish which infants with wheeze are more likely to be affected by wheezing illnesses beyond infancy.

It is recognized that persons affected by asthma and wheeze-illness often have markers of immune dysregulation at birth, suggesting that prenatal factors likely contribute to the differences in immune-programming that predispose certain individuals to the development of asthma[117]. The intrauterine environment likely influences the development of the fetal immune system[118] as well as modeling of the fetal airways given that pre-term and low birthweight babies are more likely to experience wheeze and/or develop asthma[119]. Previous studies have shown that cytokine levels in cord blood are associated with risk of wheeze or allergy in infancy and early childhood[120]. For instance, infants with lower levels of production of interleukin (IL)-13[121], IL-10 and interferon (IFN)-γ[122] at birth were more likely to develop atopic diseases during infancy or childhood [121] while higher levels of IL-4 and IFN-γ have been associated with decreased risk[123] and higher levels of IL-8 with increased risk[120] of asthma and wheeze in early childhood. Most of these cytokines are involved in influencing T-helper (Th)-1 or Th-2 cell differentiation. However, it is likely that the expression of other genes involved in asthma and wheeze could be identified in cord blood as predictors of infant wheeze.

Recently, we identified nine novel CpG sites that were differentially methylated in association with asthma at age 18 years in an epigenome-wide association study, seven of which were within novel genes (*unpublished data*): (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 5 (*ST6GALNAC5*), DiGeorge Syndrome Critical Region Gene 14 (*DGCR14*), Nucleoporin 210kDa (*NUP210*), Chromodomain Helicase DNA Binding Protein 7 (*CHD7*), Lipopolysaccharide (LPS) Induced TNF-α Factor (*LITAF*), unc-45 homolog B (*UNC45B*), and Hexokinase 1 (*HK1*). However, it was unclear whether the expression of

these genes was related to asthma and/or wheeze, and whether DNA-M at the identified CpG sites were functionally related to the expression of those genes. For the present study, we hypothesized that cord blood expression levels of these genes may predispose infants to wheeze within the first year of life, which is an early life risk factor for later being diagnosed with asthma. We also hypothesized that the expression of these genes would be correlated with DNA-M levels at the previously identified CpG sites: cg20417424 within 1500 base pairs (bp) of the transcription start sites (TSS1500) of *ST6GALNAC5*, cg14727512 within the three prime untranslated-region (3'UTR) of *DGCR14*, cg01046943 within the body of *NUP210*, cg25578728 within the body of *CHD7*, cg04359558 within the body of *LITAF*, cg00100703 within the 3'UTR of *UNC45B*, and cg16658191 within the 1st exon of *HK1*. Therefore, in the present study we investigated whether the expression levels of these seven genes were associated with wheeze within the first 12 months of life in the IOW Third Generation cohort.

4.2 Methods:

The Isle of Wight birth cohort

The Isle of Wight (IOW) birth cohort was established to study the natural history of asthma and allergies in children born between January 1, 1989 and February 28, 1990 in Isle of Wight, UK. Ethical approval was obtained from the Isle of Wight Local Research Ethics Committee (now named the National Research Ethics Service, NRES Committee South Central—Southampton B) for the 18 years follow-up (06/Q1701/34) and NRES Committee South Central—Hampshire B (09/H0504/129) for the third generation study. Details about the birth cohort have been described in detail elsewhere[46,47]. The offspring of the original IOW birth cohort were recruited into the

Third Generation study, and were followed up at 3 months, 6 months and 12 months after birth. Genome-wide gene-expression and DNA-M levels were obtained from cord blood samples at birth. Information on wheeze and maternal smoking was obtained via questionnaires that were administered to parents of the infants at scheduled follow-up visits. Data on sex, birthweight, gestational age, birth date, and type of delivery were abstracted from medical records. Infants included in this study had information about wheeze during at least one of the infant follow-up visits, and were non-missing for gene-expression and DNA-M for all of the selected mRNA and CpG probes (n=80).

Dependent Variables

The primary dependent variable for this study, *wheeze*, was a dichotomous variable defined as wheeze when the infant did not have a cough or cold, during at least one of the 3 infant follow-up visits. This was measured by an affirmative response to "Has your child had wheezing or whistling in the chest?" as well as an affirmative response to "Does your child wheeze in between cold or chest infection?", reported at any of the 3 month, 6 month, or 12 month post-birth follow-up visits. Three alternate definitions of wheeze were utilized for sensitivity analyses. *Any-wheeze* was defined as wheeze (whether or not associated with a cough or cold) reported at any of the infant follow-up visits. *Wheeze-frequency* was defined as the number of follow-up periods during which an infant experienced wheeze (when the infant did not have a cough or cold). *Any-wheeze-frequency* was defined as the number of follow-up periods during which an infant experienced any-wheeze (whether or not associated with a cough or cold).

Gene Expression Arrays

At birth, cord blood samples were collected with PAXgene RNA kits. RNA integrity was verified with the Agilent 2100 Bioanalyzer system. Genome-wide mRNA expression was assessed via one color (Cy3) experiments with the Agilent (Agilent Technologies, Santa Clara, CA) SurePrint G3 Human Gene Expression 8x60k v2 microarray kits. Array content was sourced from RefSeq, Ensembl, UniGene, and GenBank databases and provides full coverage of the human transcriptome in 50,599 biological features (including replicate probes and control probes). The oligos were 60mer in length and each transcript was tagged at least once and some had multiple tagging oligos for genes with documented splice variants. Data QC indices and analyses were performed with Agilent GeneSpring software.

For this study we selected a set of seven candidate probes measuring the expression of genes (Accession IDs): *NUP210* (NM_024923), *HK1* (NM_033500), *ST6GALNAC5* (NM_030965), *LITAF* (NM_004862), *CHD7* (NM_017780), *UNC45B* (NM_173167), and *DGCR14* (NM_022719).

DNA Methylation

Blood samples for epigenetic screening were collected at birth from the cord blood; DNA was extracted using a salting out procedure described elsewhere[51]. DNA concentration was determined by the PicoGreen dsDNA quantitation kit (Molecular Probes, Inc., OR, USA). One microgram of DNA was bisulfite-treated for cytosine to thymine conversion using the EZ 96-DNA methylation kit (Zymo Research, CA, USA), following the manufacturer's standard protocol. Genome-wide DNA methylation was assessed using the Illumina Infinium HumanMethylation450K BeadChip (Illumina, Inc.,

CA, USA), which interrogates >484,000 CpG sites, regions of DNA where a cytosine nucleotides are followed by a guanine nucleotide, associated with approximately 24,000 genes. The BeadChips were scanned using a BeadStation, and the methylation levels (β value, described below) were calculated for each queried CpG locus using the Methylation Module of BeadStudio software. Arrays were processed using a standard protocol as described elsewhere[52], with multiple identical control samples assigned to each bisulphite conversion batch to assess assay variability and samples were randomly distributed on microarrays to control against batch effects.

Data Cleaning

The program for data cleaning was written in R (R Development Core Team, 2012). Quality control (QC) measures were employed to improve the reliability of data prior to analysis. In our study, the detection *P*-value reported by BeadStudio (Illumina software to process raw intensities) was used as a QC measure of probe performance. Probes whose detection *P*-values > 0.01 in >10% of the samples were removed[53]. The methylation data were then preprocessed and technical variations removed via peak-correction using the Bioconductor IMA (Illumina methylation analyser) package[55]. The arrays were processed in six different batches; batch number was recorded as a categorical variable which was used in ComBat to adjust for inter-array variation[54]. Methylation levels for each queried CpG were calculated as beta (β) values. These represent the proportions of methylated probes for each specific CpG site and can be interpreted as percent methylation. For this study we selected a set of candidate probes we identified in a genome-wide DNA-M study for physician diagnosed asthma, that were within the body or regulatory region of the genes listed above (*unpublished data*):

cg00100703, cg01046943, cg04359558, cg14727512, cg16658191, cg20417424, and cg25578728.

Statistical Analyses

All statistical analyses were carried out in R, with statistical significance determined at an α of 0.05. Testing for differences in sex, type of delivery, maternal smoking and season of birth, between infants with and without wheeze, were done via χ^2 tests. Testing for differences in birth weight, gestational age and predicted cell-type proportions, between infants with and without wheeze, were done via T-tests. Robust regressions, which minimize the effects of outliers on the parameter estimates and standard errors, were fit to test for associations between mRNA transcripts and other variables. Robust logistic regression models were used to test the associations between expression levels of the seven mRNA transcripts and infant wheeze. Robust poisson regression models were used to test the associations between the seven mRNA transcripts and infant wheeze frequency; only infants that were non-missing for all three infant follow-up visits were included in these models. Robust linear regression models were used to test the associations between expression of each mRNA transcript and DNA-M at that gene's selected CpG site.

Signed weighted correlation networks of gene-expression between the seven mRNA transcripts were produced with the WGCNA package in R[124,125]. A soft threshold power of three and minimum-module size of two were used to identify whether the seven genes belonged to different modules. Eigengenes were produced to summarize the overall expression of each module. Pairwise spearman correlations were produced to describe how well each gene was represented by its corresponding eigengene. To observe whether

a particular module was associated with infant wheeze, we produced multivariable logistic regression models with infant wheeze as the outcome, module eigengenes as the predictors, and sex, season of birth, and predicted cell-proportions as adjustment covariates.

4.3 Results:

Infants in the IOW third generation study (n=80) consisted of similar proportions of males (47.5%) and females (52.5%), typically born in spring months of April, May or June (31.2%), primarily with normal delivery (69.2%), of normal average gestational age (mean = 39.4 weeks std = 1.6), and normal average birth weight (mean = 3482.4g std = 531.3) (Table 4.1). In our sample, 76.3% of infants never experienced wheeze during the first year of life, while 23.7% experienced wheeze during at least one infant follow-up period. We then compared the distribution of *a priori* identified confounders between infants that experienced wheeze to those that did not. There were no significant differences in sex, season of birth, or predicted cell-type proportions from cord blood samples. We also compared whether characteristics of the pregnancy or the infants at-birth that have been implicated as predictors of infant wheeze, such as cesarean delivery[126], maternal smoking at pregnancy[119,127], gestational age (pre-term birth)[119], and birthweight[128] differed between infants with and without wheeze. We observed no statistically significant differences in type of delivery, maternal smoking during pregnancy, gestational age, or birthweight.

Among these infants, we tested whether gene-expression was associated with infant wheeze for seven candidate genes, within which DNA-M levels were previously identified as markers of young adult prevalent asthma (*unpublished data*). Robust logistic

regression models (Table 4.2) revealed that increased expression levels of *HK1* in cord blood were associated with increased risk of infant wheeze (β_1 = 1.114, *P*-value = 0.027). Although sex, season of birth, and predicted cell-proportions were not significantly associated with wheeze (Table 4.1), we adjusted for these in our models as they were identified *a priori* as potential confounders. Even after adjustment for sex, season of birth, and cell-type proportions, greater expression of *HK1* (β_1 = 1.80, *P*-value = 0.030) significantly increased the risk of infant wheeze, whereas lower expression of *LITAF* (β_1 = -1.563, *P*-value = 0.047) was associated with increased risk of infant wheeze. Also, lower expression of *NUP210* was associated with infant wheeze after adjusting for sex and month of birth (β_1 = -2.106, *P*-value = 0.037), but adjusting for cord blood cell-mixture eliminated this association. Cord blood expression of *ST6GALNAC5*, *CHD7*, *UNC45B*, and *DGCR14* were not associated with infant wheeze (P-values > 0.05).

Our primary analyses included infants that were evaluated for wheeze during at least one of the three infant follow-up visits. However, those that attended all three follow-ups had greater opportunity to report wheeze. To evaluate the potential bias of missing follow-up visits, we conducted the same analyses as above, but only included those (n=61) that had answered the questionnaire for all three follow-up visits (45 without infant wheeze and 16 with infant wheeze). After excluding those with any missing assessments for wheeze, high *HK1* expression was still significantly associated with increased risk of wheeze (*P*-value = 0.049) and low *LITAF* expression was still significantly associated with increased risk of wheeze (*P*-value = 0.014) (Table 4.3). These findings suggest that missing one or two infant follow-ups did not substantially bias the observed associations.

Infant wheeze can also be transient, and recurrent wheeze, whether or not associated with a cough or cold, has been shown to increase the risk of being diagnosed with asthma later in life [116]. Thus we also tested for associations between the expression of *HK1* and *LITAF* with wheeze-frequency (45 never experienced wheeze, 9 experienced wheeze during one follow-up period, and 7 experienced wheeze during two-or-more follow-ups) and any-wheeze-frequency (27 never experienced any-wheeze, 11 experienced any-wheeze during one follow-up period, 17 experienced any-wheeze during two follow-u periods, and 6 experienced any-wheeze during all three follow-up periods). However, lower levels of expression of *LITAF* were significantly associated with increased wheeze-frequency (*P*-value = 0.010) and any-wheeze frequency (P-value = 0.006), while higher levels of expression of *HK1* were marginally associated with increased wheeze-frequency (*P*-value = 0.071) and significantly associated with any-wheeze-frequency (*P*-value = 0.007). We produced box plots to visualize the distribution of expression by any-wheeze-frequency. A dose-response relationship between *HK1* expression and any-wheeze frequency was not obvious (Figure 4.1), whereas *LITAF* expression does have an apparent inverse dose-response relationship with any-wheeze-frequency (Figure 4.2).

We then investigated whether there were linear associations between gene-expression of the seven selected genes and DNA-M levels at the selected CpG sites. Only one CpG probe was significantly associated with that gene's expression levels: higher methylation levels at cg16658191, which is within the first exon (or body), were associated with decreased expression of *HK1* (β_1 = -5.92 *P*-value = 0.049; adjusted for sex and month of birth). Also of note, higher DNA-M levels at cg14727512, which is

within the first exon (or 3'UTR), were marginally associated with increased expression of *DGCR14* (β_1 = 6.21 *P*-value = 0.089; adjusted for sex and month of birth). The linear associations for the five other CpG-gene pairs were not even marginally significant (*P*-values > 0.10) indicating that these DNA-M sites were not involved in the up- or down-regulation of the expression of the genes they are within, at least not in cord blood.

We then investigated the expression patterns among the seven selected genes; pairwise correlations indicated that some of these genes may not be independently expressed in cord blood (Table 4.4). Nine of the 21 pairwise correlations had *P*-values < 0.05, though most correlations were low-to-moderate (ρ < 0.50). Only *CHD7* and *NUP210* had at least a moderate-to-strong correlation (ρ = 0.50 *P*-value < 0.0001). Given the correlation structure among these seven genes, we explored whether a set (or sets) of correlated genes could be represented by eigengenes, and whether those eigengenes were predictive of infant wheeze. We identified two modules of genes with correlated gene-expression via weighted gene co-expression network analysis (WGCNA). Six of the seven genes were moderately-to-strongly correlated (0.40 rho < 1.0 with *P*-values < 0.0001) with the eigengene of either module (Table 4.5), while *HK1* did not fit within either module. Eigengene 1 was primarily representative of the expression levels of *ST6GALNAC5* (ρ = 0.82) and *UNC45B* (ρ = 0.74). Eigengene 2 was representative of *NUP210* (ρ = 0.80), *CHD7* (ρ = 0.73), *DGCR14* (ρ = 0.54), and *LITAF* (ρ = 0.49). The eigengene values for these two modules plus the expression of *HK1*, the combination of which summarizes the overall variations across all seven genes, were then tested in association with infant wheeze in a multiple logistic regression model. After adjusting for sex, season of birth, and cell mixture, eigengene 2 was significantly associated with

infant wheeze (β_{E2} = -7.19 *P*-value = 0.039) while HK1 was marginally associated (β_{HK1} = -6.76 *P*-value = 0.064) and eignengene 1 was not associated (*P*-value = 0.28) with infant wheeze. Also of note, in a multivariate poisson regression model that was adjusted for the same covariates, eigengene 2 was associated with wheeze frequency (β_{E2} = -5.88 *P*-value = 0.0045), while *HK1* and eigengene 1 were not (*P*-values = 0.70 and 0.14 respectively).

So, taking all results together, the expression of *HK1* was the strongest individual predictor of infant wheeze, after adjusting for sex, season of birth, and cell-mixture. However, when considering the expression of all seven genes, the overall expression of *DGCR14*, *NUP210*, *CHD7*, and *LITAF*, as represented by eigengene 2, was the strongest predictor of infant wheeze and wheeze frequency.

4.4 Discussion:

To the best of our knowledge, this was the first study to investigate whether cord blood expression levels of *ST6GALNAC5*, *DGCR14*, *NUP210*, *CHD7*, *LITAF*, *UNC45B*, or *HK1* were associated with infant wheeze. We found that high levels of HK1 and low levels of LITAF in cord blood were related to risk of wheeze and any-wheeze-frequency in the first year of life, independent of sex, season of birth, and cellular heterogeneity within the cord blood sample. We also found that low levels of NUP210 were associated with increased risk of infant wheeze, though this association appeared to be driven by differences in cellular heterogeneity and not a significant predictor after adjusting for cell-type proportions. Cord blood expression levels of the four other genes tested (*UNC45B*, *ST6GALNAC5*, *DGCR14*, and *CHD7*) were not significant predictors of infant wheeze. Also of note, other potential predictors of infant wheeze that are established at or

before birth (type of delivery, maternal smoking during pregnancy, gestational age, and birthweight) were not associated with infant wheeze in our sample, suggesting that *HK1* and *LITAF* may be unique predictors of infant wheeze without cough or cold. These findings provide insight into previously unstudied molecular risk factors of infant wheeze. Also, since infant wheeze is an early life risk factor for later being diagnosed with asthma, and DNA-M variations at these sites have been associated with asthma at age 18, these findings may provide a potential link between infant wheeze and young adult asthma.

Although the expression of *LITAF* and *HK1* were identified as novel predictors of infant wheeze in the present study, previous work has characterized some of their biological functions which are suggestive for how they may influence wheeze. Genetic variations within HK1 have been associated with retinitis pigmentosa[129] and congenital hyperinsulinism[130]. Also, increased expression of *HK1* has been observed in tumor cells [131] and HIV-infected macrophages[132] which helps to stabilize the mitochondrial membrane allowing those cells to avoid apoptosis[72], but has never previously been observed in association with infant wheeze. Yet, its functional roles make a biologically plausible link to wheeze-illness. The lifespan of pro-inflammatory cells, such as neutrophils, can be regulated via apoptotic signals and delayed response to such signals has been associated with inflammatory disease[133]; for instance, neutrophils of asthmatics appear to be resistant to CCL2-induced apoptosis[134].

The functions of LITAF have primarily been studied in mechanistic animal models, and it has been shown to form a complex with STAT6B, then translocate into the nucleus resulting in the upregulation of multiple inflammatory cytokines[79]. LITAF and/or

STAT6B also promote the expression of VEGF is known to promote angiogenesis in the airways as well as the proliferation and differentiation of bronchial endothelial cells[135,136] and T-cells[137]. TNF-α is an important pro-inflammatory cytokine, and SNPs within TNF-α have been associated with asthma and wheeze, which may mediate the risk of maternal smoking during pregnancy for wheeze or asthma[138]. Finally, LITAF has been suggested to play a role in p53-dependent apoptosis[139,140] and treatment with *Cipangopaludina chinensis*-derived LITAF (CcLITAF) induced apoptosis in a human lung cancer cell line[141]. Overall, LITAF can promote apoptosis and is an up-stream regulator of both VEGF[79] and TNF-α[139], which influence multiple immune processes. Thus, the dysregulation of the expression of LITAF may play critical roles in inflammation, immune-cell-differentiation, angiogenesis and apoptosis, providing plausible links through which LITAF may make important, yet under-studied molecular contributions to wheezing among infants.

We also found that lower levels of DNA-M at cg16658191 within the 1st exon of *HK1* were associated with increased expression of *HK1*. This provides evidence that methylation at this site may serve as a transcriptional repressor of this gene, though this should be validated in an experimental design. Given that DNA-M can be stable over time[28], *HK1* epigenetic-programming at or before birth could prospectively influence the risk of wheeze over the first year of life and possibly later in life. Previously we observed that low levels of DNA-M at cg16658191 were associated with higher probability of having prevalent asthma at age 18 (*unpublished data*), and in the present study we have observed that low levels of DNA-M at the same CpG site were associated higher expression of *HK1*, and higher expression of *HK1* was predictive of infant wheeze.

We understand that our findings should be considered within the limitations of this study. First, the candidate genes were selected due to previously identified DNA-M associations in these genes with asthma at age 18. Although variations in DNA-M have been shown to have functional relationships to gene-expression, this relationship can be complex and sometimes inconsistent. Many factors other than DNA-M can increase or decrease the expression of a particular gene: chromatin remodeling factors, histone modifications, micro RNA activity, and transcription factor binding. Thus, it is possible that the previously identified DNA-M associations would not result in changes in gene-expression that could then be associated with infant wheeze. Also, infant wheeze is a common condition that often resolves in infancy or early childhood[114] and is likely quite distinct from young adult wheeze and asthma. Thus, gene-activity related infant wheeze may not be the same as gene-activity related to asthma at age 18. These differences may explain the lack of association between expression of *DGCR14*, *UNC45B*, *CHD7*, and *ST6GALNAC5* infant wheeze. We may not have accounted for all confounders, which could have resulted in some residual confounding. Lastly, due to the small sample size in the present study, these findings should be confirmed in an independent cohort with a large sample.

Despite these limitations, this study was strong in many ways. First, its prospective nature eliminates the possibility of reverse causation for the observed associations. Second, we were able to adjust for cellular heterogeneity of the cord blood samples, despite the inability to collect blood cell differentials due to logistical constraints. Instead we utilized genome-wide DNA-M data to estimate the relative proportions of major leukocyte components (B-cells, NK cells, CD4+T cells, CD8+T

cells, monocytes, eosinophils, and other granulocytes)[67]. The inclusion of cell-proportions in our statistical models allowed us to identify whether differences in gene-expression were merely acting as markers of different cellular compositions, as appears to be the case with *NUP210* in our sample. However, we understand that this method of predicting cell-type proportions also has its limitations. This method is currently the gold-standard for predicting cell-type proportions from DNA-M and has been utilized in cord blood samples previously[142], though the algorithm's accuracy has only been validated in adult blood samples as of yet[68]. The adjusted analyses rely on the assumption that the algorithm performs as well with cord blood samples as it has with adult blood samples. To ensure that our top findings were not driven by any bias that may have been introduced by this limitation, we produced the adjusted regression models with and without cell-proportions as adjustment covariates (Table 2; Columns 3 and 2, respectively). Our two main findings (*HK1* and *LITAF*) were not affected by these adjustments; however *NUP210* was no longer a significant predictor of infant wheeze. Thus we emphasized the more conservative models, those adjusted for of estimated cell-type proportions, for our top findings.

In summary, we found that higher expression of HK1 and lower expression of LITAF in cord blood were predictive of infant wheeze, and that the expression of HK1 may be regulated by DNA-M at cg16658191. Of greatest interest are the parallels that can be drawn between our findings and the evidence in the literature suggesting that up-regulation of *HK1*[72,132] and down-regulation of *LITAF*[140,141] may both have anti-apoptotic effects; combined with the evidence that anti-apoptotic effects in pro-inflammatory cells results in prolonged inflammation[133] and have been associated with asthma[134]. Thus our

observations of higher HK1 and lower LITAF in cord blood as predictors of infant wheeze may represent an anti-apoptotic predisposition for certain infants to experience wheeze within the first year of life.

Table 4.1: Distribution of risk factors for infant wheeze and asthma, stratified by infant wheeze occurring apart from a cough or cold.

Variables	**Overall (n=80)**	**No Wheeze w/o cough/cold (n=61)**	**Wheeze w/o cough/cold (n=19)**	***P*-value**
Infant Sex	n (%)	n (%)	n (%)	χ^2 Test
Male	38 (47.5)	28 (45.9)	10 (52.6)	0.80
Female	42 (52.5)	33 (54.1)	9 (47.4)	
Season of Birth	n (%)	n (%)	n (%)	χ^2 Test
Jan/Feb/Mar	19 (23.8)	13 (21.3)	6 (31.6)	0.19
Apr/May/Jun	25 (31.2)	22 (36.1)	3 (15.8)	
Jul/Aug/Sep	19 (23.8)	12 (19.7)	7 (36.8)	
Oct/Nov/Dec	17 (21.2)	14 (23.0)	3 (15.8)	
Cell-Mixture	mean (sd)	mean (sd)	mean (sd)	T-test
$CD8^+T$	0.057 (0.031)	0.059 (0.028)	0.049 (0.039)	0.27
$CD4^+T$	0.134 (0.044)	0.132 (0.038)	0.142 (0.059)	0.46
Natural Killers	0.059 (0.046)	0.053 (0.038)	0.076 (0.064)	0.16
B-Cells	0.109 (0.034)	0.106 (0.029)	0.116 (0.046)	0.37
Monocytes	0.108 (0.022)	0.108 (0.022)	0.106 (0.022)	0.66
Eosinophils	0.049 (0.035)	0.046 (0.028)	0.061 (0.050)	0.23
Other Granulocytes	0.506 (0.095)	0.516 (0.076)	0.473 (0.138)	0.21
Delivery Type	n (%)	n (%)	n (%)	χ^2 Test
Normal	54 (69.2)	44 (73.3)	10 (55.6)	0.34
Cesarean	16 (20.5)	11 (18.3)	5 (27.8)	
Other	8 (10.3)	5 (8.3)	3 (16.7)	
Smoke During Pregnancy	n (%)	n (%)	n (%)	χ^2 Test
No	50 (64.1)	40 (67.8)	10 (52.6)	0.36
Yes	28 (35.9)	19 (32.2)	9 (47.4)	
Gestational Age	mean (sd)	mean (sd)	mean (sd)	T-test
Weeks of Gestation	39.4 (1.6)	39.5 (1.5)	39.2 (1.9)	0.57
Birth Weight	mean (sd)	mean (sd)	mean (sd)	T-test
Birth Weight (grams)	3482.9 (531.3)	3481.2 (522.2)	3488.9 (576.3)	0.96

Table 4.2: Logistic regression results for cord blood gene-expression predicting wheeze within the first year of life.

	Crude		**Adjusted***		**Adjusted****	
Gene	β_1	***P*-value**	β_1	***P*-value**	β_1	***P*-value**
NUP210	-1.024	0.219	**-2.106**	**0.037**	-0.585	0.596
HK1	**1.114**	**0.027**	**1.319**	**0.020**	**1.800**	**0.030**
ST6GALNAC5	0.207	0.625	0.380	0.492	0.372	0.587
LITAF	-0.741	0.108	**-1.222**	**0.032**	**-1.563**	**0.047**
CHD7	-0.729	0.274	-1.146	0.129	-0.876	0.319
UNC45B	-0.471	0.312	-0.841	0.134	-0.765	0.236
DGCR14	-0.959	0.183	-1.346	0.086	-1.594	0.103

* Adjusted for sex and month of birth.

** Adjusted for sex, month of birth and cell mixture (predicted proportions of $CD4^+$T cells, monocytes, natural killer cells, B cells, eosinophils, and other granulocytes).

Table 4.3: Results from sensitivity analyses (excluding those with missing follow-up) and frequency of wheeze within the first year of life.

	Wheeze[¥]		**Wheeze**[¥,**,*]		**Wheeze-Frequency**[¥,*]		**Any-Wheeze-Frequency**[¥¥,*]	
Gene	β_1	***P*-value**	β_1	***P*-value**	β_1	***P*-value**	β_1	***P*-value**
HK1	1.800	0.030	1.619	0.049	1.368	0.071	0.737	0.007
LITAF	-1.563	0.047	-2.693	0.014	-1.489	0.010	-0.818	0.006

[¥] Wheeze unrelated to a cough or cold.

[¥¥] Wheeze whether or not related to a cough or cold.

[*] Only included infants that evaluated for wheeze at all three infant follow-up visits (n=61).

[**] Because of the small number of cases and large number of adjustment covariates, this model could not be fit with robust logistic regression and instead a standard logistic GLM was used.

Table 4.4: Spearman pairwise-correlation matrix of gene-expression levels in cord blood (n=80).

Rho (*P*-value)	*ST6GAL-NAC5*	*HK1*	*LITAF*	*UNC45B*	*DGCR14*	*NUP210*	*CHD7*
ST6GAL-NAC5	-	0.04 (0.74)	-0.09 (0.43)	**0.26 (0.02)**	-0.18 (0.11)	-0.18 (0.10)	**-0.39 (<0.01)**
HK1		-	-0.09 (0.45)	-0.19 (0.08)	**-0.26 (0.02)**	-0.09 (0.44)	**-0.22 (0.04)**
LITAF			-	**0.27 (0.02)**	0.17 (0.14)	**0.23 (0.04)**	0.17 (0.15)
UNC45B				-	-0.01 (0.95)	**0.29 (<0.01)**	0.21 (0.06)
DGCR14					-	**0.27 (0.01)**	0.22 (0.06)
NUP210						-	**0.50 (<0.01)**
CHD7							-

Table 4.5: Pairwise correlations between gene-expression levels and eigengene values.

Rho (*P*-value)	**Eigengene 1**	**Eigengene 2**	***HK1***
ST6GALNAC5	**0.82 (< 0.0001)**	-0.29 (0.007)	0.04 (0.74)
DGCR14	-0.13 (0.24)	**0.54 (< 0.0001)**	-0.26 (0.02)
NUP210	0.04 (0.74)	**0.80 (< 0.0001)**	-0.09 (0.43)
CHD7	-0.14 (0.20)	**0.73 (< 0.0001)**	-0.22 (0.05)
LITAF	0.07 (0.52)	**0.49 (< 0.0001)**	-0.09 (0.45)
UNC45B	**0.74 (< 0.0001)**	0.34 (0.003)	-0.19 (0.08)
HK1	-0.09 (0.41)	-0.24 (0.03)	**1 (< 0.0001)**

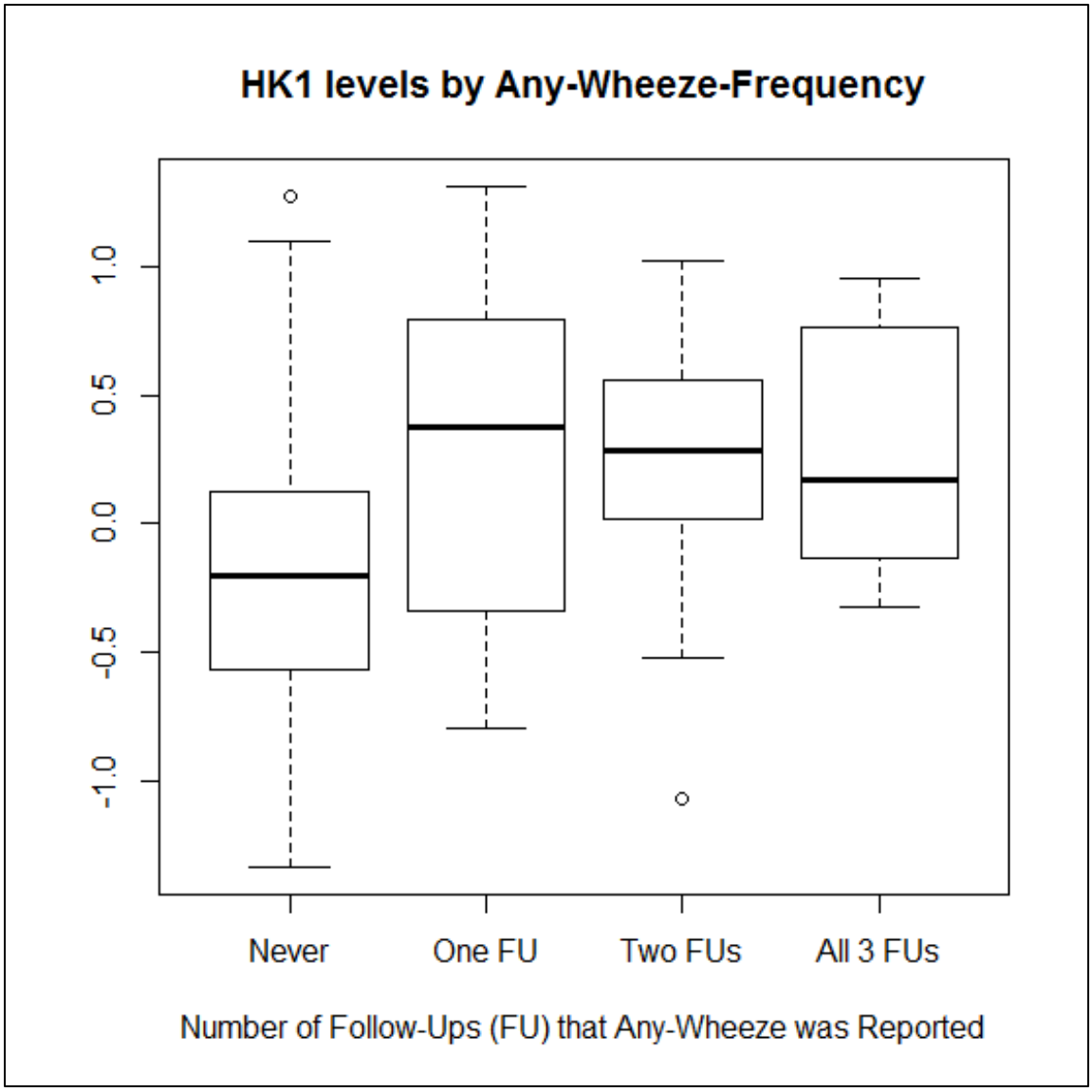

Figure 4.1: Distribution of *HK1* expression levels by any-wheeze-frequency.

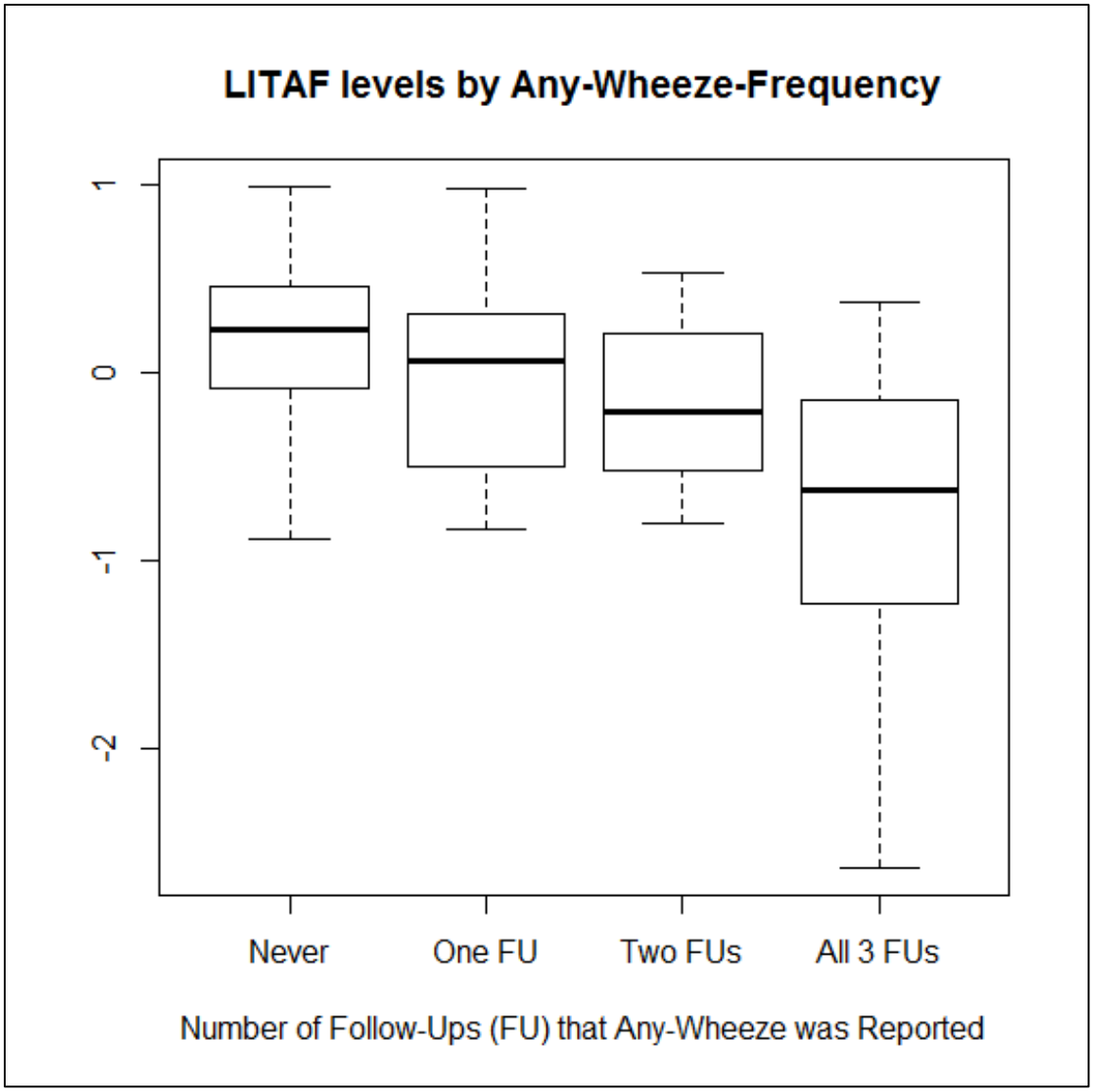

Figure 4.2: Distribution of *LITAF* expression levels by any-wheeze-frequency.

CHAPTER 5

CONCLUSIONS AND FINAL REMARKS

Overall, this dissertation achieved its objectives by exploring genetic and epigenetic variations in the Th2-path associated with complex wheeze phenotypes (Aim 2), identified novel epigenetic loci whose DNA-M levels were associated with physician diagnosed asthma (Aim 1), and related the expression of the genes encompassing the novel epigenetic loci to infant wheeze (Aim 3).

Summary of Findings and Innovations (Aim 1 and Aim 3):

From our genome-wide DNA-M association study, we found that nine epigenetic loci were differentially methylated in association with prevalent asthma at age 18: cg25578728 in the body of *CHD7*, cg16658191 in the 1st exon of *HK1*, cg00100703 in the 3'UTR of *UNC45B*, cg07948085 [intergenic], cg04359558 in the body of *LITAF*, cg20417424 in the TSS1500 of *ST6GALNAC5*, cg19974715 [intergenic], cg01046943 in the body of *NUP210* and cg14727512 in the 3'UTR of *DGCR14*. Interestingly, even though these were strongly associated with physician diagnosed asthma, they appear to drive different underlying aspects of asthma. For instance, DNA-M in *HK1*, *LITAF*, *DGCR14* were more strongly associated with atopic asthma, DNA-M in *NUP210* and *UNC45B* were more strongly associated with non-atopic asthma, whereas DNA-M within *CHD7* was associated with both atopic and non-atopic asthma. We also observed varying degrees of association between the nine sites with lung function (FEV1/FVC), BDR, and

FeNO. These findings suggest that DNA-M at these nine sites influence different physiologic mechanisms which may influence asthma. For GWAS, replication has been a gold-standard for eliminating false-positives, overfitting, and other biases[30]; thus, a replication study in an independent cohort is currently underway.

We then hypothesized that if the seven genes associated with our Aim 1 findings represented true underlying mechanisms that contribute to asthma, some of these genes could be differentially expressed in cord blood in association with infant wheeze. We found that the expression levels of both *HK1* and *LITAF* were associated with risk of experiencing wheeze within the first year of life. Also, *HK1* expression shared an inverse linear relationship with DNA-M levels at cg16658191 whereas expression of *LITAF* was not linearly associated with DNA-M levels at cg04359558.

We did not have data on gene-expression in the young adults, so we compared our findings in Aim 1 with Aim 3 via currently accepted paradigms for how DNA-M relates to gene-expression. DNA-M in the promoter region or the 1st exon is consistently associated with transcriptional repression, yet DNA-M within the gene-body is most frequently associated with promotion of transcription [32]. Assuming these paradigms hold true, then low DNA-M at cg04359558 (body) is a marker of low expression of LITAF and low DNA-M at cg16658191 (1st exon) is a marker of high expression of HK1, both of which were associated with increased odds of prevalent asthma at age 18. The relationships between expected expression levels and young adult asthma were consistent with the relationships between cord-blood expression levels and infant wheeze.

Taken together, these analyses culminated in the discovery of novel genes (*HK1* and *LITAF*) that were under differential epigenetic regulation in young adults with

asthma, and whose expression levels in cord blood were predictive of infant wheeze. Up-regulation of *HK1* and down-regulation of *LITAF* in cord blood were associated with increased risk of wheeze within the first year of life. Despite wheeze-predictive expression patterns for *HK1* and *LITAF* being present at birth, the establishment of DNA-M within *HK1* and *LITAF* appear to occur prenatally and after birth, respectively. These findings provide a direct link between our observations in Aim 1 and Aim 3, that increased *HK1* and decreased *LITAF* expression increases the risk of wheeze and/or asthma, apparently independent of age.

Not only did the findings for *HK1* and *LITAF* from Aim 1 and Aim 3 complement each other, but they were consistent with plausible biological links to wheeze and asthma. Previous research as shown that up-regulation of *HK1*[72,132] and down-regulation of *LITAF*[140,141] can induce delayed apoptosis of some cells and it is well recognized that apoptotic-resistant pro-inflammatory cells produce prolonged inflammation[133]. Thus high levels of HK1 and low levels of LITAF in cord blood may produce, or be representative of, cells that are more resistant to apoptotic signaling, putting those infants at greater risk for wheeze illness within the first year of life. Given that these genes appear to be involved in wheeze and asthma in both infants and young adults, future studies should investigate whether cord blood expression patterns of *HK1* and *LITAF* can improve upon current techniques for predicting which infants with wheeze will later be diagnosed with asthma, such as the modified asthma predictive index (mAPI)[116]. Improvement of such algorithms could allow for better prospective management of wheeze in infants and children. The discovery of these genes also provides new targets to study for potential therapeutic or preventive techniques.

Another innovation from Aim 1 of this dissertation was the use of recursive random forest (RF) as a data-reduction technique for an epigenome-wide dataset. High dimensional feature selection techniques are necessary for analyzing such large datasets as they compress these data into more manageable subsets[143]. RF has been utilized in multiple GWAS, but only infrequently utilized recursively[58,144–146] and to our knowledge only implemented with large-scale epigenetic data in a few previous investigations of atopy[62] and eczema[147] in the IOW birth cohort. Also, despite the recommendation of Leo Brieman, the original author of RF[60] to test multiple values for *mtry* and *ntree* then to pick the 'best' one[63], many previous studies with RF simply utilized the default values. In our analyses, we found that altering the default values of *ntree*, *mtry*, and *sampsize*, were critical to stable selection of predictors from data-reduction process, emphasizing the need to carefully evaluate such parameters when implementing the RF algorithm, at least when used recursively for variable selection.

Summary of Findings and Innovations (Aim 2):

We explored whether persons with more homogenous wheeze-illness would share similar genetic and/or epigenetic variations within asthma candidate genes (*IL4*, *IL4R*, *IL13*, *STAT6*, and *GATA3*) from the Th2-path. From this study, we found that DNA-M at cg25630514 in the body of *GATA3* and at cg25368824 in the TSS200 of *IL4* significantly varied based on different clinical characteristics among persons with wheeze, suggesting that epigenetic regulation of these cytokines differs among persons with wheeze or asthma that have different clinical characteristics.

To our knowledge, no one has investigated whether combinations of genetic and epigenetic features within the Th2 pathway, one of the most heavily studied sets of genes

in asthma research, can distinguish between complex manifestations of asthma, such as those represented by the wheeze-illness clusters identified in the IOW birth cohort[4]. Multiple recent studies have used clustering techniques to identify groups of asthmatics that share similar physiologic or symptomatic characteristics[90–92], though none of those studies have attempted to relate such clusters to genetic or epigenetic variations. Thus, the primary innovation from Aim 2 was that epigenetic variations, and likely other molecular mechanisms, differed in their association with complex wheeze phenotypes.

Most interesting was cluster C4, characterized by being mostly female, late-onset wheeze with low prevalence of allergic diseases and low levels for BDR, FeNO and serum IgE, which differed substantially from the other clusters across many epigenetic loci. Assuming that epigenetic regulation of expression at the nominally significant CpG sites follows the same paradigm discussed above[32], then we would expect persons in this cluster (C4) to have lower expression of *IL4*, higher expression of *IL13*, and lower expression of *GATA3*, which may be an indicator of *IL13* expression from non-Th2 cells[99], though further human studies are needed to clarify the differential expression and activity of *IL4* and *IL13*. Overall, these findings suggest that the underlying epigenetic profiles for wheeze-illnesses differ based on a multitude of characteristics, and emphasize the need to thoroughly characterize multiple symptoms and physiologic components of asthma to effectively compare asthma findings across different studies and different populations. However, given the exploratory nature and small sample size of this study, further studies are necessary to confirm our observed findings.

Final Remarks:

Advancements in epigenetic epidemiology are crucial to furthering our understanding of asthma etiology. With this dissertation we identified novel epigenetic and genetic loci that may clarify some of the underlying mechanisms involved in asthma etiology and observed differential directions of association in Th2 genes based on young adult wheeze clusters. Of particular note were the findings that high levels of *HK1* expression and low levels of *LITAF* expression increased the risk of wheeze/asthma in infants and young adults, and that DNA-M may regulate, or act as markers of, the activity of these genes. Based on current knowledge of *HK1* and *LITAF*, the observed expression and/or DNA-M patterns may increase resistance to apoptotic signaling among pro-inflammatory cells, though further functional research is needed to clarify such mechanisms. These genes offer promise as potential markers of asthma and wheeze, and could be investigated in future studies as targets for therapy or for inclusion into algorithms, such as the asthma predictive index, for predicting later-life-asthma from early-childhood characteristics. We also found that complex wheeze clusters differed in their epigenetic patterns within the Th2 path, suggesting that future studies of asthma carefully evaluate what symptomatic and physiologic traits are represented within their asthma cases when attempting to generalize findings or to conduct validation studies.

REFERENCES

1. Kim HY, DeKruyff RH, Umetsu DT. The many paths to asthma: phenotype shaped by innate and adaptive immunity. *Nat Immunol*. 2010;11(7):577-584. doi:10.1038/ni.1892.

2. Lötvall J, Akdis C a., Bacharier LB, et al. Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol*. 2011;127(2):355-360. doi:10.1016/j.jaci.2010.11.037.

3. Barnes PJ. Intrinsic asthma: Not so different from allergic asthma but driven by superantigens? *Clin Exp Allergy*. 2009;39(8):1145-1151. doi:10.1111/j.1365-2222.2009.03298.x.

4. Kurukulaaratchy RJ, Zhang H, Raza a., et al. The diversity of young adult wheeze: a cluster analysis in a longitudinal birth cohort. *Clin Exp Allergy*. 2014;44(5):724-735. doi:10.1111/cea.12306.

5. Asher MI, Keil U, Anderson HR, et al. International study of asthma and allergies in childhood (ISAAC): rationale and methods. *Eur Respir J*. 1995;8(3):483-491. doi:10.1183/09031936.95.08030483.

6. Subbarao P, Mandhane PJ, Sears MR. Asthma: Epidemiology, etiology and risk factors. *Cmaj*. 2009;181(9). doi:10.1503/cmaj.080612.

7. *National Surveillance of Asthma : United States ,*.; 2010.

8. Kay A, Rosen F. Allergy and Allergic Diseases. *N Engl J Med*. 2001;344(1):30-37.

9. Prescott S, Saffery R. The role of epigenetic dysregulation in the epidemic of allergic disease. *Clin Epigenetics*. 2011;2(2):223-232. doi:10.1007/s13148-011-0028-4.

10. *The Global Asthma Report 2014*.; 2014.

11. Beran D, Zar HJ, Perrin C, Menezes AM, Burney P. Burden of asthma and chronic obstructive pulmonary disease and access to essential medicines in low-income and middle-income countries. *Lancet Respir Med*. 2015;3(2):159-170. doi:10.1016/S2213-2600(15)00004-1.

12. Simpson CR, Sheikh A. Trends in the epidemiology of asthma in England: a national study of 333,294 patients. *J R Soc Med*. 2010;103(3):98-106. doi:10.1258/jrsm.2009.090348.

13. ALA. *Trends in Asthma Morbidity and Mortality*.; 2012.

14. Yang I V., Schwartz D a. Epigenetic mechanisms and the development of asthma. *J Allergy Clin Immunol*. 2012;130(6):1243-1255. doi:10.1016/j.jaci.2012.07.052.

15. Durham AL, Wiegman C, Adcock IM. Epigenetics of asthma. *Biochim Biophys Acta*. 2011;1810(11):1103-1109. doi:10.1016/j.bbagen.2011.03.006.

16. Palmer LJ, Burton PR, James a L, Musk a W, Cookson WO. Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families. *Eur J Hum Genet*. 2000;8(11):853-860. doi:10.1038/sj.ejhg.5200551.

17. Thomsen SF, Duffy DL, Kyvik KO, Backer V. Genetic influence on the age at onset of asthma: A twin study. *J Allergy Clin Immunol*. 2010;126(3):626-630. doi:10.1016/j.jaci.2010.06.017.

18. Berenguer AG, Rosa A, Brehm A. Asthma-snapshot or motion picture? *Front Genet*. 2013;4(April):73. doi:10.3389/fgene.2013.00073.

19. Svanes C, Zock JP, Antó J, et al. Do asthma and allergy influence subsequent pet keeping? An analysis of childhood and adulthood. *J Allergy Clin Immunol*. 2006;118(3):691-698. doi:10.1016/j.jaci.2006.06.017.

20. Bertelsen RJ, Carlsen KCL, Granum B, et al. Do allergic families avoid keeping furry pets? *Indoor Air*. 2010;20(3):187-195. doi:10.1111/j.1600-0668.2009.00640.x.

21. Von Mutius E, Vercelli D. Farm living: effects on childhood asthma and allergy. *Nat Rev Immunol*. 2010;10(12):861-868. doi:10.1038/nri2871.

22. Reponen T, Lockey J, Bernstein DI, et al. Infant origins of childhood asthma associated with specific molds. *J Allergy Clin Immunol*. 2012;130(3):639-644.e5. doi:10.1016/j.jaci.2012.05.030.

23. Murk W, Risnes KR, Bracken MB. Prenatal or early-life exposure to antibiotics and risk of childhood asthma: a systematic review. *Pediatrics*. 2011;127(6):1125-1138. doi:10.1542/peds.2010-2092.

24. Beasley R, Clayton T, Crane J, et al. Association between paracetamol use in infancy and childhood, and risk of asthma, rhinoconjunctivitis, and eczema in

children aged 6-7 years: analysis from Phase Three of the ISAAC programme. *Lancet*. 2008;372(9643):1039-1048. doi:10.1016/S0140-6736(08)61445-2.

25. Cheelo M, Lodge CJ, Dharmage SC, et al. Paracetamol exposure in pregnancy and early childhood and development of childhood asthma: a systematic review and meta-analysis. *Arch Dis Child*. 2014;100(1):81-89. doi:10.1136/archdischild-2012-303043.

26. Stranger BE, Stahl E a, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*. 2011;187(2):367-383. doi:10.1534/genetics.110.120907.

27. Buchanan A V, Weiss KM, Fullerton SM. Dissecting complex disease: the quest for the Philosopher's Stone? *Int J Epidemiol*. 2006;35(3):562-571. doi:10.1093/ije/dyl001.

28. Talens RP, Boomsma DI, Tobi EW, et al. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J*. 2010;24(9):3135-3144. doi:10.1096/fj.09-150490.

29. Sandoval J, Heyn H a., Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. 2011;6(6):692-702. doi:10.4161/epi.6.6.16196.

30. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet*. 2013;14(8):585-594. doi:10.1038/nrg3405.

31. Adams RLP. DNA methylation: The effect of minor bases. *Biochem J*. 1990;265:309-320.

32. Jones P a. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13(7):484-492. doi:10.1038/nrg3230.

33. Bégin P, Nadeau KC. Epigenetic regulation of asthma and allergic disease. *Allergy Asthma Clin Immunol*. 2014;10(1):27. doi:10.1186/1710-1492-10-27.

34. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. *Where Genotype Is Not Predictive of Phenotype: Towards an Understanding of the Molecular Basis of Reduced Penetrance in Human Inherited Disease.*; 2013. doi:10.1007/s00439-013-1331-2.

35. Soto-Ramírez N, Arshad SH, Holloway JW, et al. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics*. 2013;5(1):1. doi:10.1186/1868-7083-5-1.

36. Ziyab AH, Karmaus W, Holloway JW, Zhang H, Ewart S, Arshad SH. DNA methylation of the filaggrin gene adds to the risk of eczema associated with loss-of-function variants. *J Eur Acad Dermatol Venereol*. 2014;27(3):e420-e423. doi:10.1111/jdv.12000.DNA.

37. Baccarelli A, Rusconi F, Bollati V, et al. Nasal cell DNA methylation, inflammation, lung function and wheezing in children with asthma. *Epigenomics*. 2012;4(1):91-100. doi:10.2217/epi.11.106.Nasal.

38. Ho S-M. Environmental epigenetics of asthma: an update. *J Allergy Clin Immunol*. 2010;126(3):453-465. doi:10.1016/j.jaci.2010.07.030.

39. Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, Sun F. HapBlock: Haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics*. 2005;21(1):131-134. doi:10.1093/bioinformatics/bth482.

40. Lovinsky-Desir S, Miller RL. Epigenetics, asthma, and allergic diseases: a review of the latest advancements. *Curr Allergy Asthma Rep*. 2012;12(3):211-220. doi:10.1007/s11882-012-0257-4.

41. Breton C V., Byun HM, Wang X, Salam MT, Siegmund K, Gilliland FD. DNA methylation in the arginase-nitric oxide synthase pathway is associated with exhaled nitric oxide in children with asthma. *Am J Respir Crit Care Med*. 2011;184(2):191-197. doi:10.1164/rccm.201012-2029OC.

42. Gaffin JM, Raby B a, Petty CR, et al. β-2 adrenergic receptor gene methylation is associated with decreased asthma severity in Inner-City School Children. *Clin Exp Allergy*. 2013:681-689. doi:10.1111/cea.12219.

43. Yang I V., Tomfohr J, Singh J, et al. The clinical and environmental determinants of airway transcriptional profiles in allergic asthma. *Am J Respir Crit Care Med*. 2012;185(6):620-627. doi:10.1164/rccm.201108-1503OC.

44. Seumois G, Chavez L, Gerasimova A, et al. Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nat Immunol*. 2014;15(8). doi:10.1038/ni.2937.

45. Qiu W, Baccarelli A, Carey VJ, et al. Variable DNA methylation is associated with chronic obstructive pulmonary disease and lung function. *Am J Respir Crit Care Med*. 2012;185(4):373-381. doi:10.1164/rccm.201108-1382OC.

46. Raza A, Kurukulaaratchy RJ, Grundy JD, et al. What does adolescent undiagnosed wheeze represent? Findings from the Isle of Wight Cohort. *Eur Respir J*. 2012;40(3):580-588. doi:10.1183/09031936.00085111.

47. Arshad SH, Hide DW. Arshad (1992) Effects of environmental factors on the development of allergic disorders in infancy.pdf. *J Allergy Clin Immunol*. 1992;90(2):235-241.

48. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J*. 2005;26(2):319-338. doi:10.1183/09031936.05.00034805.

49. Crapo R, Casaburi R, Coates A, et al. American Thoracic Society Guidelines for Methacholine and Exercise Challenge Testing — 1999. *Am J Respir Crit Care Med*. 2000;161(10):309-329.

50. Dreborg S. Dreborg (1989) The skin prick test in the diagnosis of atopic allergy.pdf. *J Am Acad Dermatol*. 1989;21(4):820-821.

51. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 1988;16(3):1215.

52. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295. doi:10.1016/j.ygeno.2011.07.007.

53. Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, et al. Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. *PLoS One*. 2010;5(3):e9749. doi:10.1371/journal.pone.0009749.

54. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037.

55. Wang D, Yan L, Hu Q, et al. IMA : An R package for high-throughput analysis of Illumina ' s 450K Infinium methylation data. *Bioinformatics*. 2012:1-3.

56. Chen YA, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8(2):203-209. doi:10.4161/epi.23470.

57. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587.

58. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative random forest for gene selection from microarray data. *BMC Bioinformatics*. 2013;14:261. doi:10.1186/1471-2105-14-261.

59. Goldstein B a, Polley EC, Briggs FBS. Random forests for genetic association studies. *Stat Appl Genet Mol Biol*. 2011;10(1):32. doi:10.2202/1544-6115.1691.

60. Breiman L. Random Forests. *Mach Learn*. 2001;(45):5-32.

61. Goldstein B a, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genet*. 2010;11(1):49. doi:10.1186/1471-2156-11-49.

62. Everson TM, Lyons G, Zhang H, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med*. (*in press)*

63. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(December):18-22.

64. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003;100(16):9440-9445. doi:10.1073/pnas.1530509100.

65. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J R Stat Soc Ser B Stat Methodol*. 2004;66(1):187-205. doi:10.1111/j.1467-9868.2004.00439.x.

66. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLoS One*. 2012;7(7). doi:10.1371/journal.pone.0041361.

67. Houseman EA, Accomando WP, Koestler DC, et al. Open Access DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(86).

68. Koestler DC, Christensen BC, Karagas MR, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: A validation analysis. *Epigenetics*. 2013;8(8):816-826. doi:10.4161/epi.25430.

69. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049.

70. Liu Y, Li H, Xiao T, Lu Q. Epigenetics in immune-mediated pulmonary diseases. *Clin Rev Allergy Immunol*. 2013;45(3):314-330. doi:10.1007/s12016-013-8398-3.

71. Naumova AK, Al Tuwaijri A, Morin A, et al. Sex- and age-dependent DNA methylation at the 17q12-q21 locus associated with childhood asthma. *Hum Genet*. 2013;132(7):811-822. doi:10.1007/s00439-013-1298-z.

72. Schindler A, Foley E. Hexokinase 1 blocks apoptotic signals at the mitochondria. *Cell Signal*. 2013;25(12):2685-2692. doi:10.1016/j.cellsig.2013.08.035.

73. Peters LL, Lane PW, Andersen SG, Gwynn B, Barker JE, Beutler E. Downeast anemia (dea), a new mouse model of severe nonspherocytic hemolytic anemia caused by hexokinase (HK(1)) deficiency. *Blood Cells Mol Dis*. 2001;27(5):850-860. doi:10.1006/bcmd.2001.0454.

74. Kankaanranta H, Lindsay M a., Giembycz M a., Zhang X, Moilanen E, Barnes PJ. Delayed eosinophil apoptosis in asthma. *J Allergy Clin Immunol*. 2000;106(1 I):77-83. doi:10.1067/mai.2000.107038.

75. Potapinska O, Demkow U. T lymphocyte apoptosis in asthma. *Eur J Med Res*. 2009;14:192-195.

76. Davies EG. Immunodeficiency in DiGeorge Syndrome and Options for Treating Cases with Complete Athymia. *Front Immunol*. 2013;4(October):322. doi:10.3389/fimmu.2013.00322.

77. Takada I. DGCR14 Induces *Il17a* Gene Expression through the RORγ/BAZ1B/RSKS2 Complex. *Mol Cell Biol*. 2015;35(2):344-355. doi:10.1128/MCB.00926-14.

78. Cosmi L, Liotta F, Maggi E, Romagnani S, Annunziato F. Th17 cells: new players in asthma pathogenesis. *Allergy*. 2011;66(8):989-998. doi:10.1111/j.1398-9995.2011.02576.x.

79. Tang X, Yang Y, Yuan H, You J, Burkatovskaya M, Amar S. Novel transcriptional regulation of VEGF in inflammatory processes. *J Cell Mol Med*. 2013;17(3):386-397. doi:10.1111/jcmm.12020.

80. Merrill JC, You J, Constable C, Leeman SE, Amar S. Whole-body deletion of LPS-induced TNF- factor (LITAF) markedly improves experimental endotoxic shock and inflammatory arthritis. *Proc Natl Acad Sci*. 2011;108(52):21247-21252. doi:10.1073/pnas.1111492108.

81. Gomez-Cavazos JS, Hetzer MW. The nucleoporin gp210/Nup210 controls muscle differentiation by regulating nuclear envelope/ER homeostasis. *J Cell Biol*. 2015;208(6):671-681. doi:10.1083/jcb.201410047.

82. De las Heras JI, Batrakou DG, Schirmer EC. Cancer biology and the nuclear envelope: A convoluted relationship. *Semin Cancer Biol*. 2013;23(2):125-137. doi:10.1016/j.semcancer.2012.01.008.

83. Price MG, Landsverk ML, Barral JM, Epstein HF. Two mammalian UNC-45 isoforms are related to distinct cytoskeletal and muscle-specific functions. *J Cell Sci*. 2002;115(Pt 21):4013-4023. doi:10.1242/jcs.00108.

84. Van Der Valk RJP, Duijts L, Kerkhof M, et al. Interaction of a 17q12 variant with both fetal and infant smoke exposure in the development of childhood asthma-like symptoms. *Allergy Eur J Allergy Clin Immunol*. 2012;67(6):767-774. doi:10.1111/j.1398-9995.2012.02819.x.

85. Van Der Valk RJP, Duijts L, Timpson NJ, et al. Fraction of exhaled nitric oxide values in childhood are associated with 17q11.2-q12 and 17q12-q21 variants. *J Allergy Clin Immunol*. 2014;134(1):46-55. doi:10.1016/j.jaci.2013.08.053.

86. Bisgaard H, Bønnelykke K, Sleiman PM a, et al. Chromosome 17q21 gene variants are associated with asthma and exacerbations but not atopy in early childhood. *Am J Respir Crit Care Med*. 2009;179(3):179-185. doi:10.1164/rccm.200809-1436OC.

87. Berlivet S, Moussette S, Ouimet M, et al. Interaction between genetic and epigenetic variation defines gene expression patterns at the asthma-associated locus 17q12-q21 in lymphoblastoid cell lines. *Hum Genet*. 2012;131(7):1161-1171. doi:10.1007/s00439-012-1142-x.

88. Tsuchida A, Okajima T, Furukawa K, et al. Synthesis of disialyl Lewis a (Lea) structure in colon cancer cell lines by a sialyltransferase, ST6GalNAc VI, responsible for the synthesis of ??-series gangliosides. *J Biol Chem*. 2003;278(25):22787-22794. doi:10.1074/jbc.M211034200.

89. Sperry ED, Hurd E a., Durham M a., Reamer EN, Stein AB, Martin DM. The chromatin remodeling protein CHD7, mutated in CHARGE syndrome, is necessary for proper craniofacial and tracheal development. *Dev Dyn*. 2014;243(9):1055-1066. doi:10.1002/dvdy.24156.

90. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med*. 2008;178(3):218-224. doi:10.1164/rccm.200711-1754OC.

91. Moore WC, Meyers D a., Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med*. 2010;181(4):315-323. doi:10.1164/rccm.200906-0896OC.

92. Schatz M, Hsu JWY, Zeiger RS, et al. Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. *J Allergy Clin Immunol*. 2014;133(6):1549-1556. doi:10.1016/j.jaci.2013.10.006.

93. Robinsons D, Hamid Q, Ying S, et al. Predominant Th2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N Engl J Med*. 1992.

94. Couto Alves A, Bruhn S, Ramasamy A, et al. Dysregulation of complement system and CD4+ T cell activation pathways implicated in allergic response. *PLoS One*. 2013;8(10):e74821. doi:10.1371/journal.pone.0074821.

95. Kabesch M, Schedel M, Carr D, et al. IL-4/IL-13 pathway genetics strongly influence serum IgE levels and childhood asthma. *J Allergy Clin Immunol*. 2006;117(2):269-274. doi:10.1016/j.jaci.2005.10.024.

96. Christodoulopoulos P, Cameron L, Nakamura Y, et al. T H2 cytokine-associated transcription factors in atopic and nonatopic asthma: Evidence for differential signal transducer and activator of transcription 6 expression. *J Allergy Clin Immunol*. 2001;107(4):586-591. doi:10.1067/mai.2001.114883.

97. Godava M, Vrtel R, Vodicka R. STAT6 - Polymorphisms, haplotypes and epistasis in relation to atopy and asthma. *Biomed Pap*. 2013;157(2):172-180. doi:10.5507/bp.2013.043.

98. Walford HH, Doherty T a. STAT6 and lung inflammation. *Jak-Stat*. 2013;2(4):e25301. doi:10.4161/jkst.25301.

99. Kabata H, Moro K, Koyasu S, Asano K. Group 2 innate lymphoid cells and asthma. *Allergol Int*. 2015;64(3):227-234. doi:10.1016/j.alit.2015.03.004.

100. Beghe B, Barton S, Rorke S, et al. Polymorphisms in the interleukin-4 and interleukin-4 receptor alpha chain genes confer susceptibility to asthma and atopy in a Caucasian population. *Clin <html_ent glyph="@amp;" ascii="&"/> Exp Allergy*. 2003;33(8):1111-1117. doi:10.1046/j.1365-2222.2003.01731.x.

101. Gao PS, Heller NM, Walker W, et al. Variation in dinucleotide (GT) repeat sequence in the first exon of the STAT6 gene is associated with atopic asthma and differentially regulates the promoter activity in vitro. *J Med Genet*. 2004;41(7):535-539. doi:10.1136/jmg.2003.015842.

102. Pykäläinen M, Kinos R, Valkonen S, et al. Association analysis of common variants of STAT6, GATA3, and STAT4 to asthma and high serum IgE phenotypes. *J Allergy Clin Immunol*. 2005;115(1):80-87. doi:10.1016/j.jaci.2004.10.006.

103. Guthikonda K, Zhang H, Nolan VG, et al. Oral contraceptives modify the effect of GATA3 polymorphisms on the risk of asthma at the age of 18 years via DNA methylation. *Clin Epigenetics*. 2014;6(1):17. doi:10.1186/1868-7083-6-17.

104. Zhang H, Tong X, Holloway JW, et al. The interplay of DNA methylation over time with Th2 pathway genetic variants on asthma risk and temporal asthma transition. *Clin Epigenetics*. 2014;6(1):8. doi:10.1186/1868-7083-6-8.

105. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263-265. doi:10.1093/bioinformatics/bth457.

106. Hothorn T, Hornik K, Zeileis a. ctree: Conditional Inference Trees. *CranAtR-ProjectOrg*.

107. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9:307. doi:10.1186/1471-2105-9-307.

108. Zhou M, Ouyang W. The function role of GATA-3 in Th1 and Th2 differentiation. *Immunol Res*. 2003;28(1):25-37. doi:10.1385/IR:28:1:25.

109. Wills-Karp M. Interleukin-13 in asthma pathogenesis. *Immunol Rev*. 2004;202:175-190. doi:10.1111/j.0105-2896.2004.00215.x.

110. Zhu J. T helper 2 (Th2) cell differentiation, type 2 innate lymphoid cell (ILC2) development and regulation of interleukin-4 (IL-4) and IL-13 production. *Cytokine*. 2015;2:32-34. doi:10.1016/j.cyto.2015.05.010.

111. Huebner M, Kim D, Ewart S, Karmaus W, Sadeghnejad A, Arshad S. Patterns of GATA3 and IL13 gene polymorphisms associated with childhood rhinitis and atopy in a birth cohort. *J Allergy Clin Immunol*. 2008;121(2):408-414. doi:10.1016/j.biotechadv.2011.08.021.Secreted.

112. Yang F, Chen F, Gu J, Zhang W, Luo J, Guan X. Genetic variant rs1058240 at the microRNA-binding site in the GATA3 gene may regulate its mRNA expression. *Biomed Rep*. 2014;2(3):404-407. doi:10.3892/br.2014.254.

113. El-Gamal YM, El-Sayed SS. Wheezing in infancy. *World Allergy Organ J*. 2011;4(5):85-90. doi:10.1097/WOX.0b013e318216b41f.

114. Panettieri R a., Covar R, Grant E, Hillyer E V., Bacharier L. Natural history of asthma: Persistence versus progression-does the beginning predict the end? *J Allergy Clin Immunol*. 2008;121(3):607-613. doi:10.1016/j.jaci.2008.01.006.

115. Aalderen WM Van. Childhood Asthma : Diagnosis and Treatment. 2012;2012.

116. Chang TS, Lemanske RF, Guilbert TW, et al. Evaluation of the modified asthma predictive index in high-risk preschool children. *J Allergy Clin Immunol Pract*. 2013;1(2):152-156. doi:10.1016/j.jaip.2012.10.008.

117. Martino D, Prescott S. Epigenetics and prenatal influences on asthma and allergic airways disease. *Chest*. 2011;139(3):640-647. doi:10.1378/chest.10-1800.

118. Tadaki H, Arakawa H, Sugiyama M, et al. Association of cord blood cytokine levels with wheezy infants in the first year of life. *Pediatr Allergy Immunol*. 2009;20(3):227-233. doi:10.1111/j.1399-3038.2008.00783.x.

119. Duijts L. Fetal and infant origins of asthma. *Eur J Epidemiol*. 2012;27(1):5-14. doi:10.1007/s10654-012-9657-y.

120. Tadaki H, Arakawa H, Sugiyama M, et al. Association of cord blood cytokine levels with wheezy infants in the first year of life. *Pediatr Allergy Immunol*. 2009;20(3):227-233. doi:10.1111/j.1399-3038.2008.00783.x.

121. Williams TJ, Jones C a, Miles E a, Warner JO, Warner J a. Fetal and neonatal IL-13 production during pregnancy and at birth and subsequent development of atopic symptoms. *J Allergy Clin Immunol*. 2000;105(5):951-959. doi:10.1067/mai.2000.106211.

122. Neaville W a., Tisler C, Bhattacharya A, et al. Developmental cytokine response profiles and the clinical and immunologic expression of atopy during the first year of life. *J Allergy Clin Immunol*. 2003;112(4):740-746. doi:10.1016/S0091-6749(03)01868-2.

123. Macaubas C, De Klerk NH, Holt BJ, et al. Association between antenatal cytokine production and the development of atopy and asthma at age 6 years. *Lancet*. 2003;362(9391):1192-1197. doi:10.1016/S0140-6736(03)14542-4.

124. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. *J Stat Softw*. 2012;46(11).

125. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559. doi:10.1186/1471-2105-9-559.

126. Magnus MC, Håberg SE, Stigum H, et al. Delivery by cesarean section and early childhood respiratory symptoms and disorders: The Norwegian Mother and Child Cohort Study. *Am J Epidemiol*. 2011;174(11):1275-1285. doi:10.1093/aje/kwr242.

127. Henderson a J, Sherriff a, Northstone K, Kukla L, Hruba D. Pre- and postnatal parental smoking and wheeze in infancy: cross cultural differences. Avon Study of Parents and Children (ALSPAC) Study Team, European Longitudinal Study of Pregnancy and Childhood (ELSPAC) Co-ordinating Centre. *Eur Respir J Off J Eur Soc Clin Respir Physiol*. 2001;18(2):323-329.

128. Carolina S, Wandalsen G, Mallol J, Sole D. Wheezing and low birthweight. *Pediatr Allergy Immunol*. 2014;26:80-82. doi:10.1111/pai.12314.

129. Sullivan LS, Koboldt DC, Bowne SJ, et al. A Dominant Mutation in Hexokinase 1 (HK1) Causes Retinitis Pigmentosa. *Invest Ophthalmol Vis Sci*. 2014;55(11):7147-7158. doi:10.1167/iovs.14-15419.

130. Pinney SE, Ganapathy K, Bradfield J, et al. Dominant Form of Congenital Hyperinsulinism Maps to HK1 Region on 10q. *Horm Res Paediatr*. 2013;80(1):18-27. doi:10.1159/000351943.

131. Shoshan-Barmatz V, Ben-Hail D, Admoni L, Krelin Y, Tripathi SS. The mitochondrial voltage-dependent anion channel 1 in tumor cells. *Biochim Biophys Acta - Biomembr*. 2014. doi:10.1016/j.bbamem.2014.10.040.

132. Sen S, Kaminiski R, Deshmane S, et al. Role of Hexokinase-1 in the survival of HIV-1-infected macrophages. *Cell Cycle*. 2015;14(7):980-989. doi:10.1080/15384101.2015.1006971.

133. Luo HR, Loison F. Constitutive neutrophil apoptosis: Mechanisms and regulation. *Am J Hematol*. 2008;83(4):288-295. doi:10.1002/ajh.21078.

134. Yang EJ, Choi E, Ko J, Kim DH, Lee JS, Kim IS. Differential effect of CCL2 on constitutive neutrophil apoptosis between normal and asthmatic subjects. *J Cell Physiol*. 2012;227(6):2567-2577. doi:10.1002/jcp.22995.

135. Psarras S, Volonaki E, Skevaki CL, et al. Vascular endothelial growth factor-mediated induction of angiogenesis by human rhinoviruses. *J Allergy Clin Immunol*. 2006;117(2):291-297. doi:10.1016/j.jaci.2005.11.005.

136. Leigh R, Oyelusi W, Wiehler S, et al. Human rhinovirus infection enhances airway epithelial cell production of growth factors involved in airway remodeling. *J Allergy Clin Immunol*. 2008;121(5):1238-1245.e4. doi:10.1016/j.jaci.2008.01.067.

137. Mor F, Quintana FJ, Cohen IR. Angiogenesis-inflammation cross-talk: vascular endothelial growth factor is secreted by activated T cells and induces Th1 polarization. *J Immunol*. 2004;172(7):4618-4623. doi:10.4049/jimmunol.172.7.4618.

138. Panasevich S, Lindgren C, Kere J, et al. Interaction between early maternal smoking and variants in TNF and GSTP1 in childhood wheezing. *Clin Exp Allergy*. 2010;40(3):458-467. doi:10.1111/j.1365-2222.2010.03452.x.

139. Myokai F, Takashiba S, Lebo R, Amar S. A novel lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha gene expression: molecular cloning, sequencing, characterization, and chromosomal assignment. *Proc Natl Acad Sci U S A*. 1999;96(8):4518-4523. doi:10.1073/pnas.96.8.4518.

140. Hong YH, Lillehoj HS, Hyen Lee S, Woon Park D, Lillehoj EP. Molecular cloning and characterization of chicken lipopolysaccharide-induced TNF-?? factor (LITAF). *Dev Comp Immunol*. 2006;30(10):919-929. doi:10.1016/j.dci.2005.12.007.

141. Yang S, Li P, Mi Z. LPS-induced TNF?? factor (LITAF) in the snail Cipangopaludina chinensis: Gene cloning and its apoptotic effect on NCI-H446 cells. *Fish Shellfish Immunol*. 2012;32(2):268-272. doi:10.1016/j.fsi.2011.11.011.

142. Koestler DC, Avissar-Whiting M, Andres Houseman E, Karagas MR, Marsit CJ. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. *Environ Health Perspect*. 2013;121(8):971-977. doi:10.1289/ehp.1205925.

143. Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*. 2012;13(1):59. doi:10.1186/1471-2105-13-59.

144. Archer KJ, Kimes R V. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal*. 2008;52(4):2249-2260. doi:10.1016/j.csda.2007.08.015.

145. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;7:3. doi:10.1186/1471-2105-7-3.

146. Moorthy K, Mohamad MS. Random forest for gene selection and microarray data classification. *Bioinformation*. 2011;7(3):142-146. doi:10.6026/97320630007142.

147. Quraishi BM, Zhang H, Everson TM, et al. Identifying CpG sites associated with eczema via random forest screening of epigenome-scale DNA methylation. *Clin Epigenetics*. 2015;7(1):68. doi:10.1186/s13148-015-0108-y.