THE IDENTIFICATION OF *ATPAF1* AS A NOVEL ASTHMA SUSCEPTIBILITY GENE AND THE CHARACTERIZATION OF FUNCTIONAL REGULATORY VARIANTS.

By

Eric Michael Schauberger

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Genetics

2011

ABSTRACT

THE IDENTIFICATION OF *ATPAF1* AS A NOVEL ASTHMA SUSCEPTIBILITY GENE AND THE CHARACTERIZATION OF FUNCTIONAL REGULATORY VARIANTS.

By

Eric Michael Schauberger

Asthma, the most common chronic disease of childhood, is driven by genetic and environmental determinants. To identify genes that increase the risk of asthma in children, a multiple stage genome-wide association study was conducted in a nested case-control study of a whole-population birth cohort from the Isle of Wight, UK. This study resulted in the identification of a cluster of associated SNPs and SNP haplotypes in the *ATPAF1* gene (ATP synthase mitochondrial F1 complex assembly factor 1) on human chromosome 1p33, with two SNPs achieving significance at a genome-wide level (P=2.26E-5 to 2.2E-8). SNP, haplotype, and/or gene-level associations were confirmed in three of five replication populations.

The *ATPAF1* gene contains 303 reported variants, which were assessed using *in silico* techniques and prioritized through annotated function in public databases, and/or inferred function based on their location in experimentally reported or predicted functional DNA sequences. Twenty-seven variants were prioritized based on the *in silico* results, of which several variants had predicted function as coding, splicing, and/or gene expression regulation. These prioritized variants were targeted in addition to exons, conserved, and regulatory regions for selective resequencing in 40 cohort individuals using Sanger sequencing.

Selective resequencing of 14.6kb resulted in the identification of 35 total variants. This included validation of 9 (of the 27) prioritized variants from the *in silico* screen and 9 new rare variants, including 1 nonsynonymous mutation. Three variants with gene expression regulatory potential were found to be clustered within 600 bp of each other in the promoter/exon 1 of *ATPAF1* in four haplotypes. This region was targeted for analysis using luciferase reporter gene assays in BEAS-2B and COS-7 cell lines. These cell culture assays confirmed promoter functionality and indicated a statistically significant difference in luciferase expression (means ranging between 2-3 fold differences) among the promoter haplotypes.

In conclusion, *ATPAF1* was identified as a childhood asthma susceptibility gene. *In silico* studies coupled with selective resequencing of the *ATPAF1* region provided an efficient method to identify functional variants. DNA variant haplotypes within the *ATPAF1* promoter demonstrated the ability to differentially regulate gene expression. However, the roles of these and other functional variants in the *ATPAF1* promoter and their ability to modulate asthma susceptibility need further study.

DEDICATION

I dedicate this dissertation to my family. Particularly to my understanding, patient, and lovely wife, Katie, who started as a running partner, became a study buddy, best friend, and love of my life. She always believes in me even when I do not believe in myself. There is no doubt that without her continued love and support none of this would have ever have been possible. To my parents, Chuck and Linda, who have always supported me every step of my life and encouraged me to always work hard toward my goals—from the soccer pitch to the lab. To my greyhounds, Salute and Diva, and cats Nutmeg and Basil, who kept me company in the loneliest parts of the middle of the night when much of the writing of this document occurred.

ACKNOWLEDGMENTS

I would like to acknowledge my dissertation committee, Drs. Susan Ewart, Karen Friderici, Marianne Huebner and Andrea Amalfitano for their guidance throughout my dissertation work.

In particular, I would like to acknowledge Dr. Susan Ewart who took me under her wing and taught me what research is really about. She also taught me the art of presentation (through lots of practice!), the art of collaboration (through lots of patience!) and to persevere through failure. Without her mentorship and support this work never would have been completed.

To Dr. Karen Friderici, who adopted me into her lab and mentored me daily. She taught me how to approach problems with unclear answers and coached me through the constant struggle to answer those questions. Without the support and camaraderie of her and her lab, I am unsure how I would have made it through this.

I would also like to thank the large crowd of collaborators involved in my work by mentoring me and/or providing expertise to my project including Drs. Hasan Arshad, Wilfried Karmaus, Marsha Wills-Karp, Carl Langefield, other coauthors, and collaborators from the 1989-90 Isle of Wight birth-cohort too many to mention. I would also like to thank all other past labmates from the Ewart and Friderici labs including Dennis Shubitowski, and Drs. Ravisanker Ramadas, Ellen Wilch, Meghan Drummond, Soumya Korrapati and other members of 6th floor of the MSU Biomedical Physical Science Building for feedback and support.

Finally, I would like to acknowledge the Michigan State University College of Osteopathic Medicine DO/PhD program for its continued support of me and the other future crop of Osteopathic physician scientists.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ATPAF1: ATP synthase mitochondrial F1 complex assembly factor 1

C1ORF223: Chromosome 1 open reading frame 223

CAMP: Childhood Asthma Management Program

CARE: Childhood Asthma Research and Education network

CEU: Utah residents with ancestry from northern and western Europe

FBAT: Family-based association test

FDR: False discovery rate

FEV1: Forced expiratory volume in 1 second

FVC: Forced vital capacity

GALA: Genetics of Asthma in Latino Americans

GEE: Generalized Estimating Equation model

GWAS: Genome-wide association study

IOW: Isle of Wight

KIAA0494: uncharacterized protein LOC9813

LD: Linkage disequilibrium

Luc: Firefly Luciferase

MAF: Minor allele frequency

mRNA: messenger ribonucleic acid

OMIM: Online Mendelian Inheritance In Man

OR: Odds ratio

PC20: Provocation concentration causing a 20% fall in FEV1

PDT: Pedigree disequilibrium test

qPCR: Quantitative Polymerase Chain Reaction

SLEGEN: International Consortium for Systemic Lupus Erythematosus Genetics

SNP: Single nucleotide polymorphism

TDT: Transmission disequilibrium test

TH2: Type-2 T Helper Cell

WTCCC: Wellcome Trust Case Control Consortium

CHAPTER 1

Literature review of asthma, asthma genetics, genome-wide association studies and the 1989-1990 Isle of Wight birth-cohort

ASTHMA

Epidemiology

Asthma is a significant and prevalent disease, afflicting more than 300 million individuals worldwide [1]. Traditionally considered a disease of the developed world--asthma's worldwide prevalence is increasing 50% every decade and becoming epidemic in many countries (Figure 1) [1]. It is responsible for an attributable cost of over $19.7 billion in the United States alone [2]. The most common chronic disorder in childhood, asthma affects more than 9 million children in the United States, with a prevalence rate of 9.4%. It is the number one cause of missed school and the third highest cause of hospitalization among children [2]. The prevalence of asthma in the United Kingdom, already among the highest in the world, increased from 10-15% in 1990 to 15-30% in 2000 [3].

Clinical signs and diagnosis

The clinical signs and chief complaints for asthma are varied between individuals and even in a single individual between asthma attacks. The chief complaints of patients exhibiting asthma often include episodic chest tightness, shortness of breath, and expiratory wheeze (wheezing can be inspiratory in severe cases). Many patients develop a nocturnal cough.

Asthma diagnosis is primarily made through a combination of physical exam, patient and/or parent questionnaire and pulmonary function tests. Physical examination of a person undergoing an asthma attack typically reveals tachycardia, tachypnea with prolonged expirations,

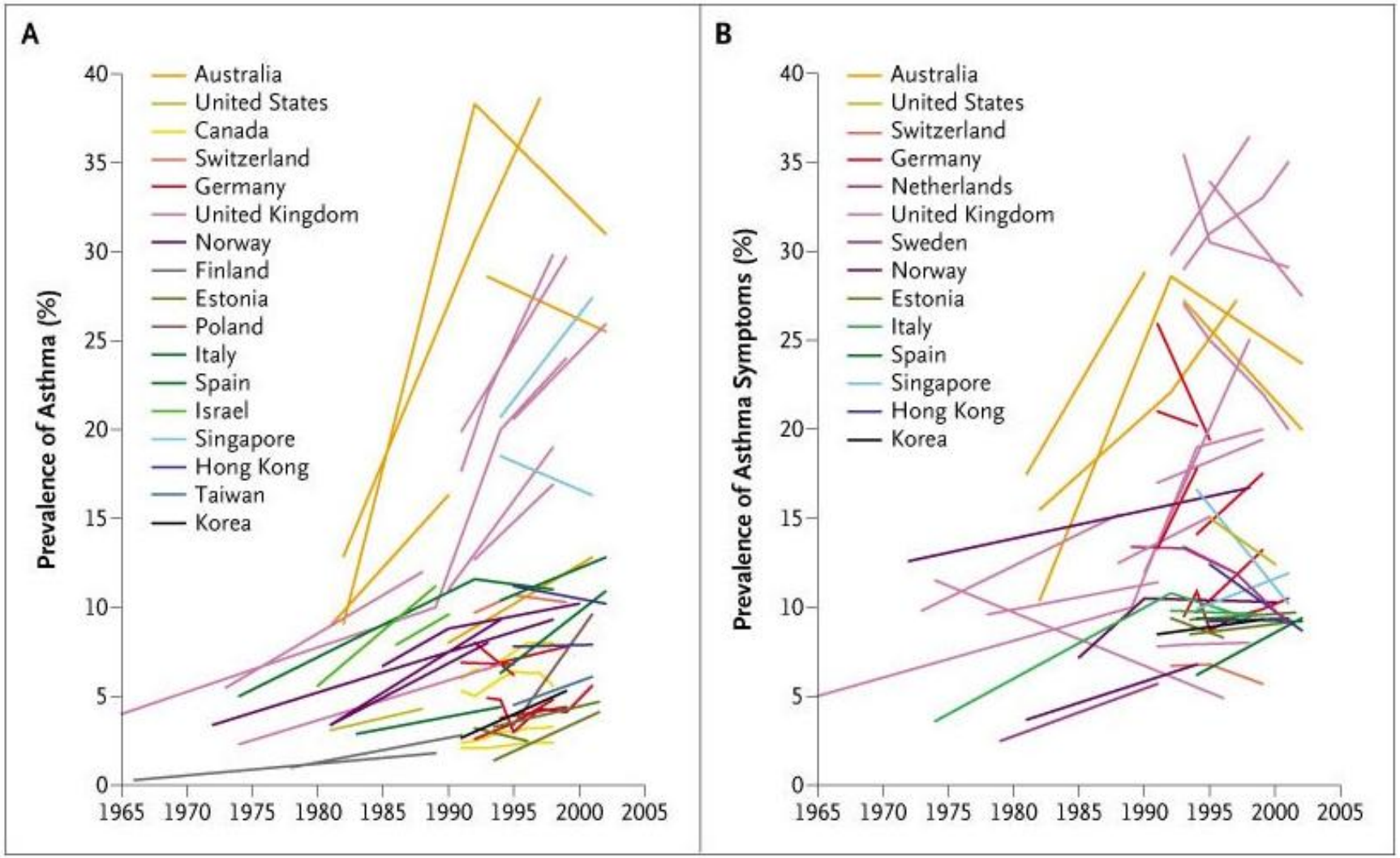

Figure 1. Changes in the prevalence of diagnosed asthma and asthma symptoms over time in children and young adults. Adapted from Eder et al, Figure 1 [3]. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

airtrapping, overinflation of the chest with poor diaphragm mobility, and diffuse high-pitched expiratory wheeze.

In pulmonary function tests, the hallmark of asthma is a reduced forced vital capacity (FVC) and a forced expiratory volume at 1 second (FEV1) / FVC ratio less than 80% of predicted values based on a person's size, age, and sex, but which can be improved with a bronchodilator such as albuterol. A bronchial challenge test, either with methacholine or histamine, is often used as a means to verify asthma. In this test, an asthmatic has a decreased threshold needed to induce a response that will reduce FEV1 and FVC. A detailed examination of sputum can reveal signs of asthma such as Curschmann's spirals (mucus that forms a cast of the small airways) and Charcot-Leyden crystals (eosinophil breakdown products) [5].

The lifecourse of asthma from infancy to adolescence to adulthood is complex. A feature of asthma is its variability from person to person and even from one attack to another in the same person. The occurrence of asthma in early infancy and childhood is difficult to assess. Martinez et al [6] first described a now commonly used classification system to classify wheeze into four categories of phenotypes based on a retrospective study of patient prognosis:

- transient early wheezing - "those with at least one lower respiratory tract illness with wheezing during the first three years of life but no wheezing at six years of age"
- late-onset wheezing- "those with at least one lower respiratory tract illness with wheezing during the first three years of life but no wheezing at six years of age"
- persistent wheezing- "those with at least one lower respiratory tract illness with wheezing during the first three years of life but no wheezing at six years of age"

- no wheezing- “those with no evidence of wheezing at six years of age”

The diagnosis of asthma is usually made by the time children reach school age where it is found at a rate of 10-15% [7]. Of the children with persistent wheeze, studies have shown that the majority will subsequently be diagnosed with asthma. By age 3, many of these children already have abnormal lung function that often persists into adulthood, and by adolescence, many of these early asthmatics will also develop atopy [7]. The morbidity of asthma in children has been shown to vary greatly depending on socioeconomic status [7]. Asthma is often found in association with other related diseases such as hayfever, food allergies and atopic dermatitis [8].

Pathophysiology

Asthma is a chronic obstructive disease of the lung that is characterized by airway hyperresponsiveness, tissue remodeling, and reversible airway obstruction (Figure 2 and 3). Anatomically, asthma primarily involves the lower respiratory system including the bronchi and terminal bronchioles. Etiologic or pathologic classification of the disease is difficult. Asthma can be divided into two classes, atopic and nonatopic, which is based on whether the pathogenesis is due to an allergen-immune system mediated reaction (atopic) or another, non-immune system related cause (nonatopic).

The long term effects of asthma result in histological changes in the bronchi include a thickening of the basement membrane, hypertrophy of submucosal glands, and edema with a mixed inflammatory infiltrate. Most of asthma’s morbidities are reversible but over time the bronchi are narrowed by epithelial hyperplasia and smooth muscle hypertrophy and hyperplasia [9].

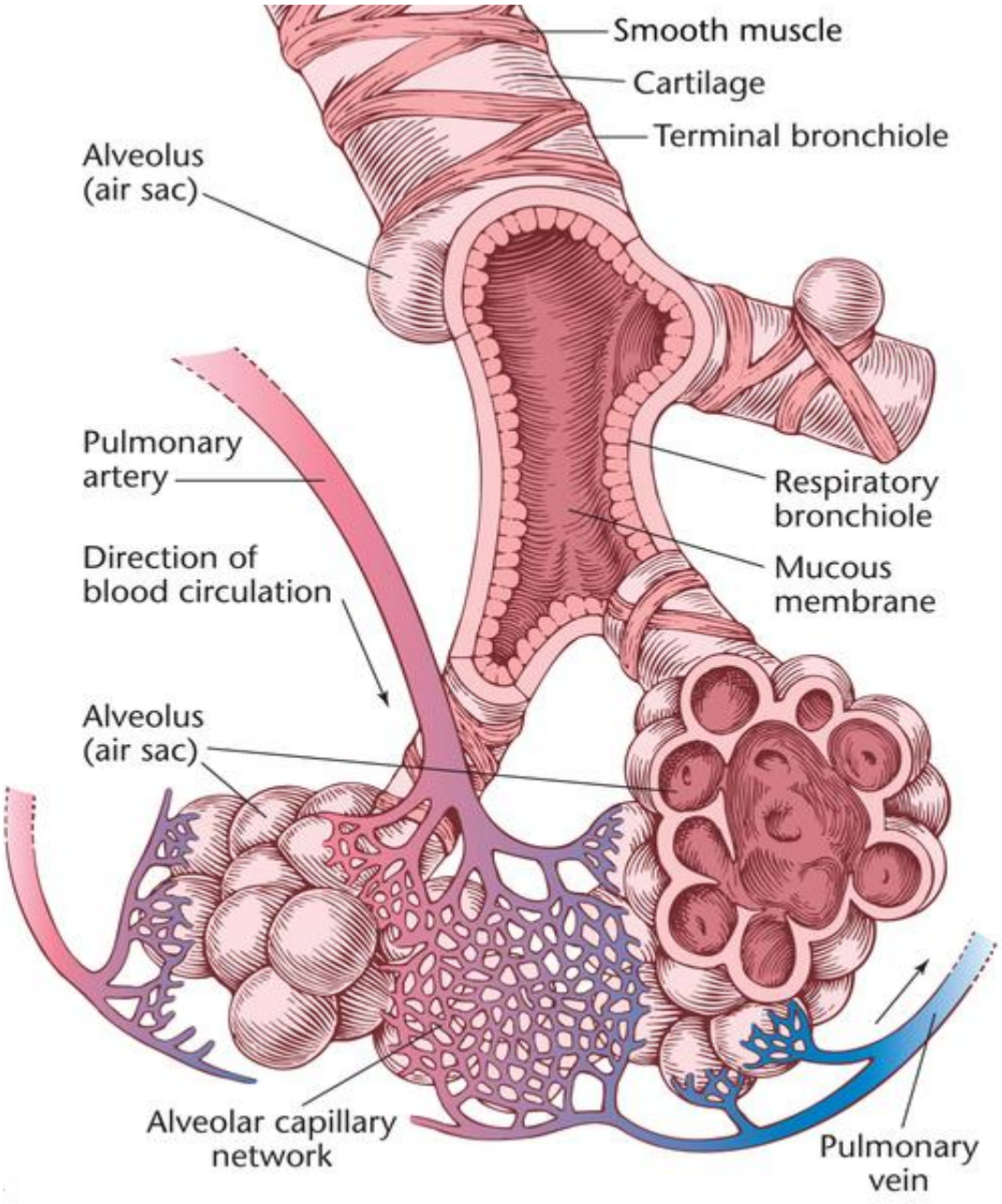

Figure 2: The normal anatomy of lung near the terminal bronchiole. Figure adapted from Figure 2-1 A, Braunwald's Atlas of Internal Medicine (2011) [4] © 2011 Current Medicine Group with kind permission of Springer Science+Business Media.

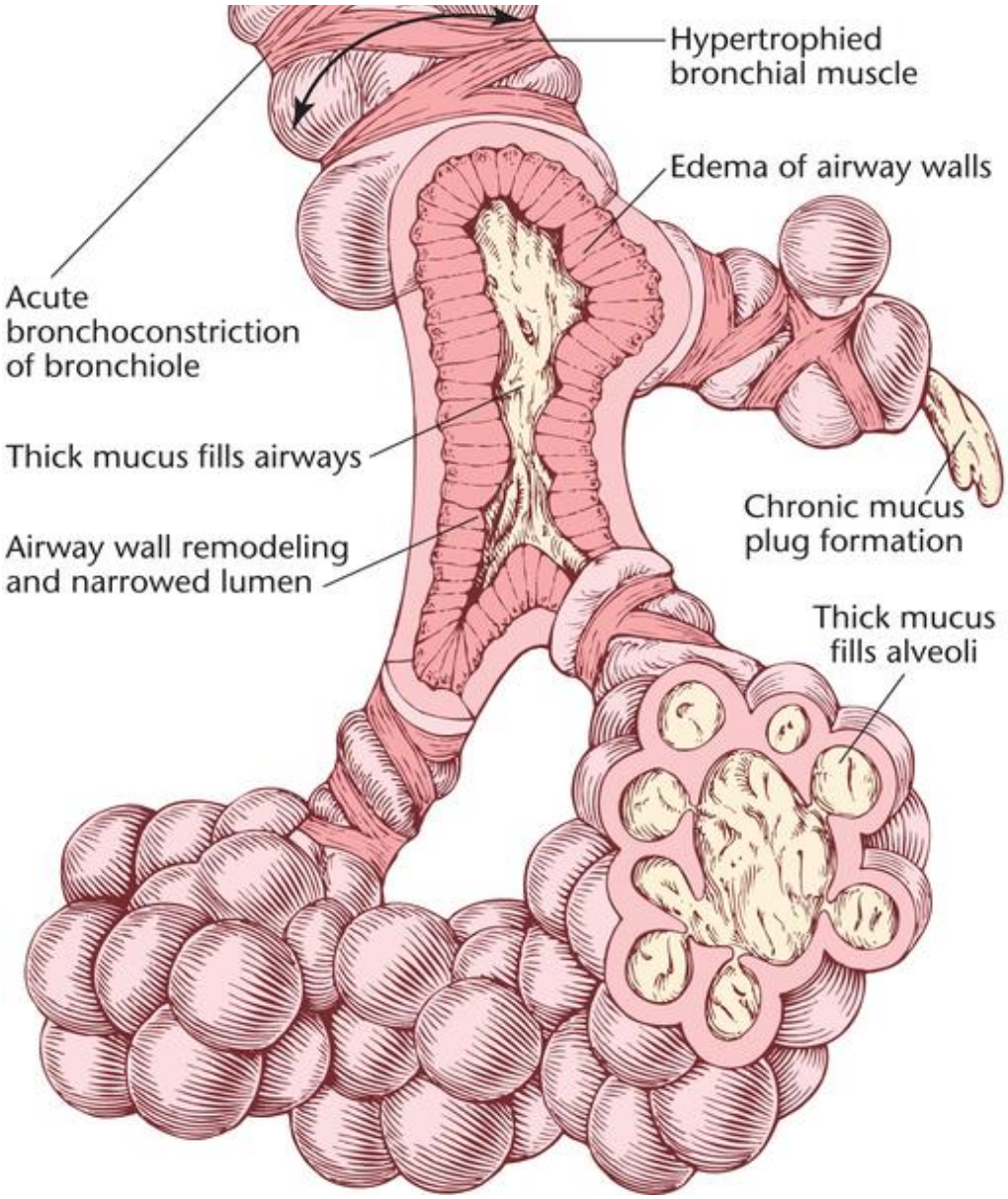

Figure 3: The anatomy of lung near the terminal bronchiole during an acute asthmatic episode demonstrating airflow obstruction in asthma. Smooth muscle along airway walls is constricted and hypertrophied. Along with extensive plugging of the airways by mucus, airway walls are thickened and edematous as a result of inflammatory changes. Eventually, airway walls may be remodeled, leading to fixed obstruction in some patients. Figure adapted from Figure 2-1 B, Braunwald's Atlas of Internal Medicine (2011) [4]

Atopic vs nonatopic asthma types

Atopic (also called allergic or extrinsic) asthma pathogenesis is a Type-1 (immediate) hypersensitivity reaction caused by exposure to allergens. It is responsible for the majority of childhood asthma. Mechanistically, in atopic asthma, an inhaled allergen initially induces sensitization of the immune system by stimulating Type-2 T-Helper cells (TH2) cells to produce interleukin (IL)-4 and IL-5. The IL-4 stimulates antibody producing cells (plasma cells) to class-switch antibody production from immunologlobulin (Ig)G to IgE. IL-5 stimulates production and activation of eosinophils. During a repeat allergen exposure, a complex cascade is initiated when mast cells, which reside predominantly on mucosal surfaces, bind the allergen-IgE antibody complex in a crosslinking manner and release a host of mediators. These mediators, which include histamine, stimulate a constellation of symptoms including bronchoconstriction and mucus production resulting in an “asthma attack”. Several hours later, in the late phase reaction, eosinophils are chemotactically drawn to the airway and activated by eotaxins. The eosinophil releases cytokines, oxygen radicals, and nitric oxide that result in tissue damage and continued bronchoconstriction and results in most of the permanent pathological consequences. The hallmarks of atopic asthma are a positive skin prick test to one or more allergens and the presence of allergen-specific IgE [10].

Nonatopic asthma (nonallergic or intrinsic) is a less common form of asthma that is predominantly seen in adults [10]. It is not a hypersensitivity reaction to a specific allergen and as such afflicted individuals are negative for skin prick tests [10, 11]. Nonatopic asthma is instead found to be triggered by chemical, environmental, or microbial exposures. These triggers include viral infections (rhinovirus, parainfluenza, respiratory syncytial virus), air pollutants (ozone, tobacco smoke), drugs (aspirin), stress, exercise, and cold-air [11]. As would be expected

with no specific allergen trigger, the level of IgE found in the serum is within the normal range although there is some evidence it is elevated when compared to unaffected individuals [11]. The mechanism is not fully understood, but the airway inflammation is similar to that of atopic asthma and is thought to be mediated by a local rather than systemic IgE production [10, 12].

Treatment

At this time, asthma is a disease that cannot be cured and can only be managed. The goal of treatment is to achieve maximum control of symptoms using the fewest medications. The current standard of care is to control mild intermittent and exercise-induced asthma using a short-acting beta-adrenergic agonist as a "rescue inhaler". Mild and moderate persistent asthma is treated using combinations of low dose inhaled corticosteroids and/or leukotriene-receptor antagonists with use of a rescue inhaler as needed. More severe and persistent asthma is often treated using combinations of long-acting beta adrenergic agonists, high doses of inhaled corticosteroids, and leukotriene receptor antagonists. The most severe and persistent cases of asthma, with failure to respond to other therapies, are treated using oral corticosteroids, high doses of inhaled corticosteroids and other immunomodulators [13].

ASTHMA GENETICS

Asthma is a complex genetic disease (OMIM ID: 600807)[14] caused by both genetic and environmental factors. Underlining the role of genetics in this disease, it has been predicted that an individual is three times more likely to have asthma if the mother has asthma and seven times

more likely to be afflicted if both the mother and father have asthma [15]. Several types of genetic studies ranging from twin studies to genome-wide association studies have been utilized to explore the genetics of asthma in human or animal model genomes with varying success. The methods used in asthma genetics have closely paralleled the advancements in the rapidly developing field of genetics as a whole. The identification of the genetics involved in asthma can be divided into methods used before and after the advent of the genome-wide association study (GWAS), which has revolutionized the field of genetics.

As of June 2010, over 600 genes have been reported to be associated with asthma susceptibility (104 genes with ≥5 positive reports) and hundreds more genes have been associated with asthma-related phenotypes [16]. The complexity of dissecting the genetics of asthma is due to a high level of genetic and phenotypic heterogeneity caused partially by inconsistent diagnosing criteria. The genes identified can be divided into four main groups based on proposed function and tissue location, including genes involved/expressed in: 1) innate immunity and immunoregulation, 2) regulation of TH2 cell differentiation and functions, 3) epithelial cells, chemokines and maintenance of the epithelial cell barrier, and 4) lung function including epithelial or smooth muscle expressed genes [17].

Pre-GWAS era of asthma genetics

Twin studies

One frequently used rough measure of genetic influence is the twin study, which measures concordance of a trait between monozygotic and dizygotic twins. The genetic influence has been predicted from twin studies to be 36-79% [18].

Linkage studies

One of the early methods used in genetics to start to identify genes is the genome wide linkage study. The main advantage of this method is that it does not require a prior hypothesis about a potential gene and, thus, can identify genes that do not fit into the current paradigm of asthma. To date there have been genome wide linkage studies for asthma or asthma-related phenotypes completed in 24 different human populations [19]. These studies collectively have resulted in the identification and, in many instances, replication of eight asthma susceptibility genes and many other regions containing linkage with no specific gene identified [19].

An example of a gene identified by a linkage study is *ADAM33*. It is a novel gene found on chromosome 20p13 that had a murine homolog previously described for bronchial hyperresponsiveness (BHR) [20]. The search was narrowed by a subsequent association study and transmission disequilibrium test that showed an over-transmission of a number of polymorphisms in asthmatics. The linkage was strongest in asthmatics that had BHR and lowest in those who had elevations of total serum IgE [21]. The association with *ADAM33,* although not replicated in all association studies, has been one of the most robust asthma susceptibility genes

with replication of association of several SNPs in multiple populations in multiple ethnic backgrounds with asthma[17] and asthma-related phenotypes (allergic rhinitis[22], psoriasis[23], COPD[24], and atopic dermatitis[25]).

The *ADAM33* gene product is a membrane-anchored zinc dependant metalloproteinase that is expressed in lung fibroblasts and bronchial smooth muscle but not in bronchial epithelial cells or immunological cells [17]. It seems most likely involved in the remodeling of airway tissue and not directly impacting the immunological aspect of the disease. It has also been shown to be expressed during branching morphogenesis of lung in both mice and humans, suggesting a role in lung development [17].

Linkage studies have been a powerful method to examine the genetics of a complex trait; however, their usefulness in human genetics seems to have waned as the study of asthma genetics has progressed. Two major drawbacks have likely participated in their decreased use. First, they often suffer a lack of statistical power needed to confirm correlations. Another drawback of linkage studies of complex genetic diseases is that it is often difficult to narrow a region sufficiently enough to identify a single gene, and indeed, regions with linkage may have multiple susceptibility loci [17, 26].

Candidate gene association studies

Another approach to exploring the genes involved in asthma is the candidate gene association study. In this method, the genotype frequencies and/or allele frequencies of variants in a predefined gene or genomic region are compared in unrelated cases (asthmatics) and unrelated

controls. This methodology focus is on genes that have been implicated as having a role in asthma. The gene or genomic region is preselected and relies on prior knowledge of disease mechanisms and pathways [17]. An association study has frequently been done in follow-up to a linkage study once a linked region has been identified.

In comparing and contrasting an association study to a linkage study, for a population of the same size, an association analysis offers more statistical power for the detection of common disease alleles that confer a modest disease risk [17]. As most association studies are completed in a case-control format without families, it is usually easier to recruit a study population. However, family-based populations can be used in association studies through an adaptation called a transmission disequilibrium test (TDT). Association analysis in unrelated individuals requires a higher marker density than does a linkage study due to the regions around each marker having less shared identity by descent or linkage disequilibrium (LD) [17, 27]. The size of a region with a positive association in an association analysis is much narrower than a linkage study and is limited to the size of the LD block containing the signal.

The results of candidate gene association studies for asthma have resulted in over 150 gene associations. Among the proposed genes within the regions found to be associated with asthma are several members that pertain to the immune system including: the cytokine gene cluster on chr5q which contains the genes *IL4* , *IL13*, *IL9*, *IL5*, *CD14* and *ADRB2* (B2 adrenergic receptor); the Major Histocompatibility Complex (MHC) locus (*HLA*, chr6p), *MS4A2* (Fcε receptor (chr11q13), *INFG* (interferon gamma, chr12q), *TCRA* and *TCRB* (T-cell receptors, chr14q) and the *IL4R* (IL-4 receptors, chr16p) [20].

Gene X environmental interactions

Although asthma has a clear genetic component, there are many environmental factors that have been shown to be protective or a risk for asthma [7]. Guerra and Martinez cite the example where both a low and high level of endotoxin exposure is a risk for atopy; whereas a more moderate dose of endotoxin has been shown to be protective for atopy. An extra layer of complexity is the genotype of an allele in the atopy susceptibility gene, *CD14* that alters the amount of exposure to endotoxin that is either moderate risk or protective [28]. A similar study examining the association of the Toll-like receptor 2 (*TLR2*) in farmers' children showed decreasing risk of asthma proportional to endotoxin exposure. This association was not found in children living in the same communities but not living on farms [17].

GWAS era of asthma genetics

Genome-wide association studies (GWAS) are a recent revolutionary approach to dissect complex genetic diseases. This methodology is an innovation that combines the positive attributes of the candidate gene association study with the genome-wide linkage study. Similar to candidate gene association studies, allele and genotype frequencies at SNPs from asthmatics and nonasthmatics are compared to establish association. However, as the name suggest, the scale of these studies is genome-wide (>100,000 SNPs) as opposed to only the candidate regions (<100 SNPs). This new approach to identify susceptibility genes has been enabled by the development of high-throughput SNP microarray technologies. Similar to genome-wide linkage studies, GWAS have the advantage that *a priori* hypotheses are not required, allowing discovery of novel

genes that do not fit into the current paradigm of disease. Recent GWAS have identified previously unsuspected loci for a number of complex genetic diseases [29-35].

Reported associations with asthma

As of June 2010, nineteen GWAS for asthma[35-41] or asthma related traits (pulmonary function [42-45], psoriasis [46-49], YKL-40 level [50], atopic dermatitis[51, 52], plasma eosinophil levels[53]) have been published with 62 SNPs reaching genome-wide significance. The listing of the top reported genes for asthma and several related traits are in Table 1.

Chr17q/ORMDL3

The first and most replicated association to be found by a GWAS for asthma was on Chr17q21 containing the gene *ORMDL3* (ORM1-like 3), a gene of unknown function and ubiquitous expression that encodes a transmembrane protein in the endoplasmic reticulum that alters endoplasmic reticulum mediated Calcium-ion homeostasis. Replication for this region has subsequently been found in additional populations [54-56]. Subsequent findings have shown that the association extends into the neighboring genes, *GSDMB* (gasdermin B), *ZPBP2* (zona pellucida binding protein 2), *IKZF3* (IKAROS family zinc finger 3). Although an expression modulating causative variant was thought to be found, more recent data now supports one or more variants linked to allele specific chromatin remodeling that results in altered cis regulation [57]. A study by Bouzigon et al, showed that variants within this region were found to be associated only with early-onset asthma (onset age 4 years of age of younger) in children with

Table 1 : Top reported results of Asthma GWAS (as of June2010) from GWAS Integrator in www.hugenavigator.net [16]

Gene(s) (sorted by date reported)	Location	Asthma or related	OR/Beta [95% CI] [NR (not reported)]	Top variant P-value	Reference
ORMDL3	17q12	Asthma	1.45[[1.17-1.81]]	9E-11	[35]
CHI3L1	1q32.1	YKL-40 levels	0.3[[NR] ng/ml decrease]	1E-13	[50]
IKZF2	2q34	Plasma eosinophil count	6.3[[4.3-8.3] % increase]	5E-10	[53]
IL1RL1	2q12.1	Plasma eosinophil count	6.4[[4.7-8.1] % increase]	5E-14	[53]
HLA-E	6p21.32	Plasma eosinophil count	4.6[[2.7-6.6] % increase]	3E-6	[53]
WDR36, TSLP	5q22.1	Plasma eosinophil count	6.1[[3.7-8.6] % increase]	1E-6	[53]
SH2B3	12q24.12	Plasma eosinophil count	7.6[[5.9-9.3] % increase]	7E-19	[53]
IL5	5q31.1	Plasma eosinophil count	7.1[[4.9-9.2] % increase]	1E-10	[53]
GATA2	3q21.3	Plasma eosinophil count	9.4[[7.2-11.6] % increase]	9E-17	[53]
CTNNA3	10q21.3	Asthma (toluene diisocyanate-induced)	5[[2.36-10.6]]	6E-6	[38]
Intergenic	9p21.3	Asthma (toluene diisocyanate-induced)	5.29[[2.41-11.61]]	7E-6	[38]
Intergenic	13q12.1	Asthma (toluene diisocyanate-induced)	5.2[[2.47-10.92]]	3E-6	[38]
PDE4D	5q12.1	Asthma	1.18[[1.08-1.30]]	3E-8	[37]
TLE4, CHCHD9	9q21.31	Asthma (childhood onset)	1.64[[1.32-2.04]]	7E-7	[36]
PKDCC	2p21	Atopy	1.92[[1.27-2.86]]	2E-6	[51]
DENND1B, CRB1	1q31.3	Asthma	1.43[[NR] (European ancestry)]	2E-13	[41]
MAVS	20p13	Asthma	NR[NR]	8E-6	[39]
SCG3	15q21.2	Asthma	NR[NR]	2E-6	[39]
RAD50	5q31.1	Asthma	1.64[[1.36-1.97]]	3E-7	[39]

exposure to environmental tobacco smoke [58]. Over the next several years, future studies of this gene should provide interesting knowledge in our understanding of asthma—especially the risk of early exposure to tobacco smoke and gene interaction.

Genetic association study design considerations

In the interpretation and design of a GWAS, the majority of considerations that exist in a typical candidate association study still hold true, but due to their size, frequency of use, and the volume of data that they produce, the design deserves further attention.

Study numbers

An important consideration in study design of GWAS is ensuring having an adequate number of individuals in order to sufficiently power the study. Without sufficient power, the results may not reach significance. Early GWAS were often conducted with several hundred to a thousand individuals and were criticized for lacking power. Recent studies conducted have involved case-control designs with tens of thousands of individuals. The determination of the number of affected and control individuals needed depends on the 1) nature of the trait (dominant, recessive), 2) predicted level of risk (odds ratios (OR)), 3) allele frequency, 4) amount of linkage disequilibrium (LD) exhibited, and 5) level of statistical power required [59].

Study design efficiency

Although technological advancements and the sequencing of the human genome has resulted in large leaps in methods of examining genetic diseases, such as genotyping SNP arrays, the cost of doing these experiments is high. There are several design innovations that have been employed to reduce the cost of these studies including use of a pooled study and a multiple-stage joint analysis study. Typically, a primary genome-screen is used in a case-control population to identify regions of interest which are then followed by one or more replication studies in other populations [60].

A DNA pooling approach is done similar to the standard GWAS, with the exception that the first stage is done using pools of DNA from individuals with the same affection status. Although SNP microarrays were developed for genotyping of individuals (a qualitative measure), they can also be utilized to allelotype DNA pools (a quantitative measure of allele frequencies). The benefit is to substantially decrease the cost of a study (by decreasing the number of microarrays needed) while maintaining the benefits of a GWAS. The second stage involves a focused validation study conducted by genotyping all individuals at the subset of SNPs of interest. The use of pools has the benefit of substantially decreasing genotyping cost but limits the use of the data to the phenotype selected and will only provide the allele frequencies of the pools and not individual genotypes.

A second cost-efficient approach involves a multiple-stage joint analysis methodology. In this method, a higher level of statistical power is maintained by analyzing both the first stage genome-wide screen and the second stage focused study. Utilizing this method, a strategy can be tailored to optimize a study by decreasing genotyping while maintaining statistical power and

decreasing risk of a false positive by optimizing the number of individuals needed and number of SNPs genotyped.[61] For example, in a GWAS involving 500 cases and 500 controls, a disease causing allele with a 25% MAF and imparting a 2 fold susceptibility risk could be detected with >90% statistical power by genotyping only 134 cases in stage 1 (using a high thoroughput genome screen) followed by genotyping the top 0.34% of the SNPs from stage 1 with the other 366 case individuals. This results in reducing genotyping by 25% (which would result in a savings of 70% if stage 2 genotyping costs ten times more per SNP) [61].

Multiple testing concerns

The strength of GWAS to examine association across the whole genome is also ironically its chief problem. Statistically, association studies can require testing more than 100,000 SNPs resulting in many false-positive findings by chance alone using normal significance criteria. Therefore, the significance level must be adjusted to provide a higher level of stringency or the results adjusted. Typically, the goal is to set an appropriate significance level (α=0.05) so that 1 in 20 studies would produce a Type 1 error (false-positive)[62]. This can be done in multiple ways. Using a Bonferroni correction, an experiment, such as a GWAS, with 1E6 independent tests would result in a P-corrected-value= nP-uncorrected = 1E6(0.05) = 5E-8. This is the bar that is now become the standard for "genome-wide significance".

It is important to note that Bonferroni correction is considered conservative as it assumes all tests are independent. This assumption is very inaccurate as the SNPs that are being tested exist in multiple marker haplotypes that are inherited together and thus are not independent. Other forms of multiple testing correction methods including using a false discovery rate (FDR) correction

such as Benjamini-Hochberg, which is less conservative but is usually not used in analysis of genome-wide data and, instead, used in a second stage of multiple-stage design.

Population stratification

It is important to note that a SNP's association with a disease can be an artifact of population substructure or admixture. Population substructure is the result of multiple subgroups that differ in disease prevalence. This can result in false estimates in allele frequency and in false-positive associations. Although stratification can be detected and corrected somewhat, a study designed with well-matched controls (in a case-control study) has been shown to be best prevention [63]. Other solutions include using a family-based study design which, among other benefits, is unaffected by stratification. In the near future, the use of ancestral informative markers to determine the ancestral history and reduce or correct for stratification is likely to reduce risk of spurious associations.

Replication: gene level and variant level

Currently, replication of GWAS results in two or more separate populations is the gold standard for validation of results--although this has been a moving target with increasing stringency over the years demanding more and larger replication populations. Replication of genetic association results (GWAS or otherwise) has proven to be difficult.

Winner's curse

One confounding factor in attempts to replicate GWAS results is the "winner's curse" or the inflated effect seen in the first population an association is detected. This has resulted in many of the first GWAS reports having associations with variants with larger effect than what was found when results were replicated (if the results replicated at all). Theoretically, the maximum magnitude of the effect is inversely proportional to the power of the study to detect an association for a variant with a given minor allele frequency [64]. Therefore, the overestimation of effect is larger for variants with a very low minor allele frequency where statistical power is less.

Loose vs strict replication

An important subject is what classifies as "replication" of an association. To some researchers, replication has a strict requirement of a repeat association and effect with a specific variant(s) [65], while others' definition of replication is far less stringent. The result has been the development of a continuum ranging from strict replication to a loose replication--being an association in the same gene or even LD block but in a different allele or haplotype [26]. Another variable is phenotype replication—that is the replication of findings in the exact same or similar phenotype which are related.

Replication of association findings in multiple populations has become critical. However, it is important to note a lack of replication in one or more populations may simply reflect the heterogenic nature of complex diseases having multiple susceptibility genes, environmental

factors, and population stratification. Even the most replicated genes fail to show association in a substantial number of studies.

Correlation versus causation

Once an association has been established, the next question often is what the functionality of the associated variant. However, as many studies have shown, it is possible that the associated variant is not the causative or even a functional variant but is merely in LD with the causative variant. These causative variants could be another common variant, rare variant, or structural variant that either was not genotyped or lacked sufficient statistical power to be detected by association [64]. Moreover, it is also possible that a group of variants have functional significance as a group or haplotype. Due to the nature of LD often spanning multiple genes, the causative variant may be in a different gene than the associated variant(s). Although coding variants (especially those resulting in missense mutations) are often the first examined if there is an association, there are several noteworthy examples in which associations with a coding variant were later found to be caused by a functional non-coding variant [64].

Moving from association of a variant(s) to identifying the true causative variant(s) often involves a multiple stage approach which can include 1) utilization of available *in silico* resources to identify variants in regulatory control regions or splicing control variants, 2) resequencing of the genomic region containing the variant to validate all possible variants or discover novel variants, 3) functional studies (including *in vitro* and *in vivo* models) to characterize variants to evaluate their possible causative role in the phenotype being studied.

Summary of asthma susceptibility regions and genes

As a whole, the field of asthma genetics has progressed significantly and has resulted in a deeper understanding of the pathogenesis of asthma. Combining the results of linkage studies with candidate and genome-wide association studies has resulted in the identification of several susceptibility genes with strong support. The genes with the most consistent positive associations and the proposed mechanism or involvement in asthma pathogenesis are listed in Table 2. Due to the somewhat recent advent of the GWAS, this table is biased towards linkage studies and the subsequent candidate association studies that have been conducted to confirm the linkage studies. Also, the number of articles with a positive association is used as a surrogate for the actual number of positive associations.

These results also underscore that although GWAS are a powerful methodology, there is still room for candidate gene association studies. This is illustrated by the fact that many candidate genes with the most support have failed to replicate in GWAS. It is important to note, however, that even the most replicated disease genes have not been replicated in every study. Among the compiled results, however, are a few that were also found in genome wide linkage studies including the major histocompatibility complex (MHC), interferon gamma (INF-G) , the cellular receptor for IgE and the cytokine gene cluster.

ISLE OF WIGHT 1989-1990 BIRTH COHORT

Study design

The primary population used in this dissertation consists of an unselected whole population birth cohort from the Isle of Wight, UK [66]. Between 1 January 1989 and 28 February 1990, 1,536

Table 2: Top replicated asthma susceptibility genes from linkage and association studies (including GWAS)[16]

Gene Symbol	Number of Articles: All	MetaAnalysis	GWAS	G xE interaction	Predicted function or pathway relevant to asthma (Adapted from Vercelli [17])
ADRB2	104	6	0	43	Bronchial smooth-muscle relaxation
IL13	60	1	1	10	TH2 effector functions
TNF	55	4	0	12	Inflammation
IL4	54	1	0	5	TH2 differentiation and IgE induction
CD14	49	3	0	9	Innate immunity — microbial recognition
IL4R	48	2	0	3	α-chain of the IL-4 and IL-13 receptors
GSTP1	44	2	0	22	Environmental and oxidative stress — detoxification
GSTM1	42	2	0	20	Environmental and oxidative stress — detoxification
ADAM33	38	2	0	3	Cell–cell and cell–matrix interactions
TGFB1	33	2	1	4	Immunoregulation, cell proliferation
GSTT1	30	3	0	11	Environmental and oxidative stress — detoxification
MS4A2	28	1	0	3	Subunit of the high affinity IgE receptor
LTC4S	28	2	0	11	Cysteinyl leukotriene biosynthesis — inflammation
HLA-DRB1	26	0	0	1	Antigen presentation
HLA-DQB1	21	0	1	2	Antigen presentation
FLG	21	2	0	3	Epithelial barrier integrity
CYSLTR1	21	0	0	5	Cysteinyl leukotriene receptor — inflammation
TLR4	21	1	0	3	T-cell-response inhibition and immunoregulation
CCL5	20	2	0	1	Monocyte, T-cell and eosinophil chemoattractant
ALOX5	20	0	0	7	Synthesis of leukotrienes
IL10	20	0	0	2	Immunoregulation
IFNG	17	0	0	1	Immunoregulation
NPSR1	17	0	0	1	Regulation of cell growth and neural mechanisms
NOS3	17	0	0	3	Nitric oxide synthesis — cell–cell communication
LTA	16	1	0	3	Inflammation
CTLA4	16	0	0	2	T-cell-response inhibition and immunoregulation

children were born on the island and 1,456 children (≈95%) were enrolled in the study. Data related to asthma and allergies were collected at birth and at ages 1 (N=1,167), 2 (N=1,174), 4 (N=1,218), 10 (N=1,373) and 18 years (N=1,313) using detailed questionnaires. At age 10, the International Study of Asthma and Allergy in Childhood (ISAAC)[67] questionnaire was used to asses respiratory, dermatological and nasal symptoms related to asthma and allergy. Details on extent of disease morbidity and asthma treatment were also collected. A summary of collected data and basic characteristics is present in Table 3. Pulmonary function tests were conducted to collect baseline lung function. At age 4 and 10, skin prick test to 14 common food and aeroallergens was performed to establish atopy status. Blood samples were collected for DNA analysis and total serum IgE and specific antigen IgE analyses [66]. It is important to note that unlike many other asthma studies, individuals in this birth cohort were not selected based on allergy or asthma status, which can creates bias.

Geography

The Isle of Wight is an island (23 miles x 13 miles) immediately off the Southern coast of England with a population of approximately 130,000 people (Figure 4) [68].

Demographics relevant to asthma

The island consists of a predominantly semi-rural environment of small towns and villages with a large tourism industry. There is a low prevalence of farming and air pollution on the island, thereby decreasing the influence of agricultural and industrial environmental factors on asthma.

Table 3: Summary of the Isle of Wight birth cohort

Birth (January 1989 – February 1990): Family history of allergy, household pets, parental smoking, socioeconomic status, gestational and parturient factors, cord blood IgE, maternal IgE	n = 1,456
Age 1 year: Asthma and other atopic manifestations (physician diagnosis and symptoms), household pets, parental smoking, method of feeding, chest infections	n = 1,167 (80.2%)
Age 2 years: Asthma and other atopic manifestations, household pets, parental smoking, method of feeding, chest infections	n = 1,174 (80.6%)
Age 4 years: Asthma and other atopic manifestations, household pets, parental smoking, method of feeding, chest infections, skin prick test	n = 1,218 (83.7%)
Age 10 years: Asthma and other atopic manifestations (additional: ISAAC questionnaire), household pets, parental smoking, skin prick test (n = 1,036), serum IgE and blood sampling for gene analysis (n = 945), baseline spirometry (n = 981) and methacholine bronchial challenge (n = 784)	n = 1,373 (94.3%)
Some basic characteristics:	
Maternal smoking during pregnancy	24.7%
Recurrent chest infection at age 1 year	7.4%
Owned dog at recruitment (birth)	28.6%
Eczema at age 1-year	9.6%
Wheezing at age 1-year	11.4%
Wheezing at age 4-years	11.4%
Wheezing at age 10-years	21.4%
Skin prick test positive at age 4-years	16.7%
Skin prick test positive at age 10-years	26.9%

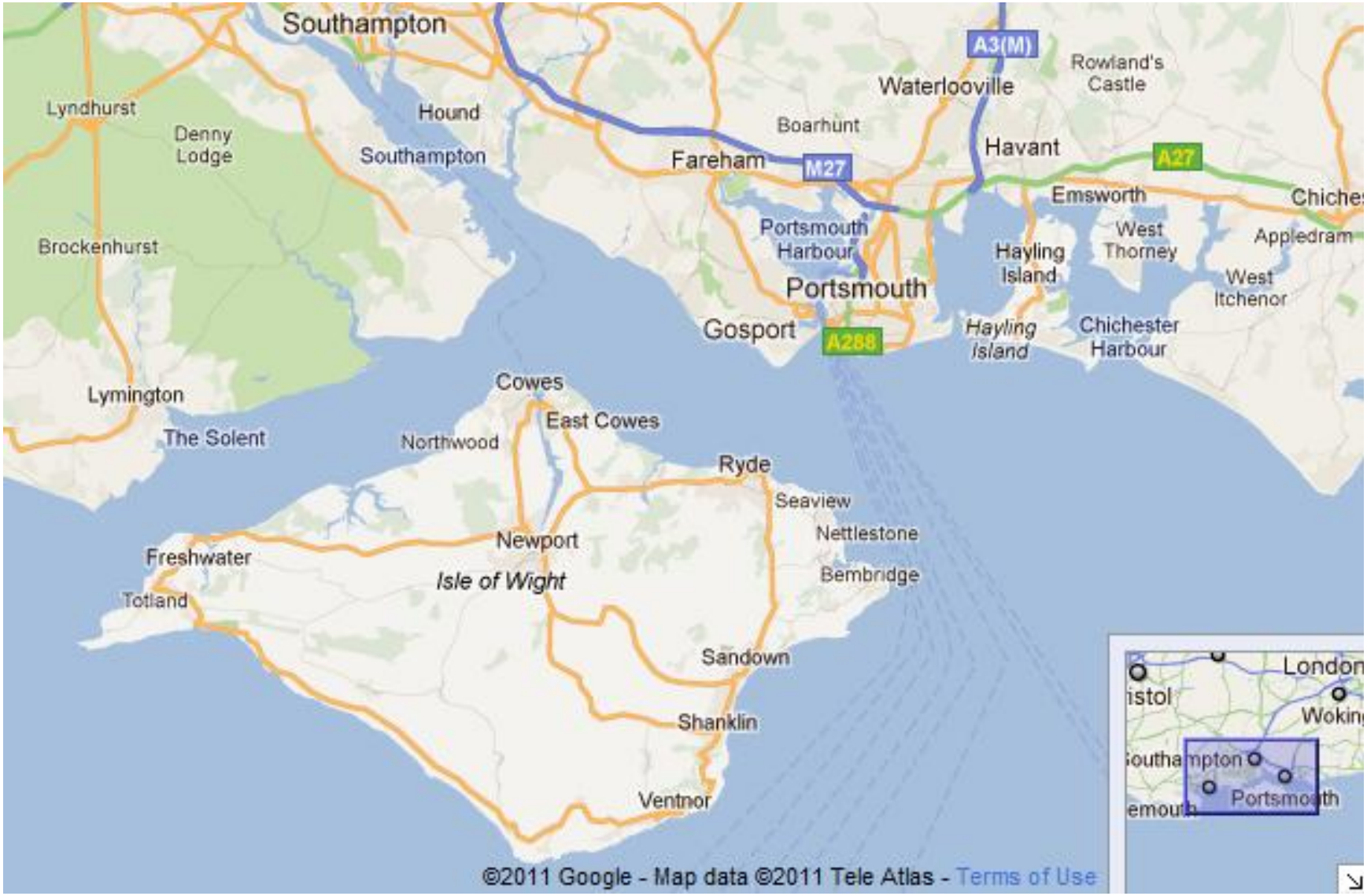

Figure 4: Location of the Isle of Wight, U.K. (23 miles x 13 miles) immediately off the southern coast of England. © 2011 Google and © 2011 Tele Atlas

The population is not geographically or genetically isolated and is approximately 99% Caucasian ethnicity.

Key accomplishments/findings of the Isle of Wight birth cohort study

Prevalence

The key benefit of a longitudinal study, like the Isle of Wight birth cohort, is the ability to examine data at multiple ages. This is important in diseases like asthma which are often transient as children grow. The prevalence of asthma and asthma related traits were recorded at multiple periods. The cumulative prevalence of wheeze at age 4 was reported to be 30% and at age 10 to be 40% with the prevalence of asthma being approximately 25% at age 10 [66]. As a comparison, the prevalence of asthma in children (age 6-14) in the U.K. (20-25%), is among the highest of all European countries [69]. At age 10, an unusually high level of nonatopic wheeze was found [66].

Natural history of asthma

The Isle of Wight birth cohort longitudinal data support the different wheezing phenotypes documented by Martinez et al [6] and described above in Clinical signs and Diagnosis section. Of the children in the study who were reported to wheeze anytime by age 10, 81% had transient early wheezing which is demonstrated by the improvement of the majority of children (63%) between the ages of 4 and 10. The other 37% of children continue to wheeze and developed into typical childhood asthma (persistent wheezers). These persistent wheezers were symptomatic at

age 10, were usually atopic, had reduced lung volumes (including FEV1 and FVC), BHR, and had the most severe symptoms and required oral and inhaled corticosteroids [70].

Identified risk factors

Environment

Several key findings regarding the risk or protection of environmental factors have been discovered or corroborated in the Isle of Wight birth cohort. One key example includes exploring the role of breastfeeding and prenatal tobacco smoke exposure. Karmaus et al found that breastfeeding decreased the risk of recurrent lower respiratory infections and the risk of asthma associated with recurrent lower respiratory infections. In addition, breastfeeding was found to modify the effect of prenatal smoke exposure [71] and enhance lung volume [72] in children.

Genetic risk factors

Results of the Isle of Wight birth cohort support that asthma has an inheritable component. Data support that positive family history is a significant risk factor for asthma and wheeze (as well as for eczema and allergy) [70, 73]. Three susceptibility genes have previously been implicated in the Isle of Wight birth cohort for asthma or related phenotypes. The gene *GATA3,* whose protein for a transcription factor involved in the development and activation of TH2 cells, was identified as increasing susceptibility risk for atopic eczema [74] and rhinitis [75]. DNA variants in *IL13* (interleukin-13) and *IL1RN* (interleukin-1 receptor antagonist), although not found to be associated with asthma alone, were found to be associated with increased risk of developing asthma in children with *in utero* exposure to tobacco smoke [76, 77]. These findings highlight the utility of a birth-cohort over a traditional case-control study, which most likely would not

have had such detailed longitudinal data needed to detect this gene-environment interaction. Additional DNA variants in *IL13* have subsequently been found to have an interaction with child birth order and allergic sensitization [78]. Gene-gene interactions between *GATA3* and *IL13* variants were also found to impact the risk of rhinitis in cohort children [75].

REFERENCES

REFERENCES

1. Braman, S.S., *The Global Burden of Asthma.* Chest, 2006. **130**(1_suppl): p. 4S-12.

2. American Lung Association. *Asthma*. 2007 [cited.

3. Eder, W., M.J. Ege, and E. von Mutius, *The asthma epidemic.* N Engl J Med, 2006. **355**(21): p. 2226-35.

4. Braunwald, E., *Atlas of internal medicine*. 1999, Philadelphia, PA: Current Medicine. 1 v. (various pagings).

5. Kumar, V., et al., *Robbins and Cotran pathologic basis of disease*. 7th ed. 2005, Philadelphia: Elsevier Saunders. xv, 1525 p.

6. Martinez, F.D., et al., *Asthma and wheezing in the first six years of life. The Group Health Medical Associates.* N Engl J Med, 1995. **332**(3): p. 133-8.

7. Subbarao, P., P.J. Mandhane, and M.R. Sears, *Asthma: epidemiology, etiology and risk factors.* CMAJ, 2009. **181**(9): p. E181-90.

8. Stedman, T.L., *Stedman's medical dictionary*. 2006, Lippincott Williams & Wilkins: Philadelphia. p. 1 electronic text.

9. Holgate, S.T., *Pathogenesis of asthma.* Clin Exp Allergy, 2008. **38**(6): p. 872-97.

10. Barnes, P.J., *Immunology of asthma and chronic obstructive pulmonary disease.* Nat Rev Immunol, 2008. **8**(3): p. 183-92.

11. Virchow, J., *Intrinsic Asthma*, in *Asthma and rhinitis*, W.W. Busse and S.T. Holgate, Editors. 2000, Blackwell Science: Oxford ; Malden, MA, USA. p. 1355-1375.

12. Takhar, P., et al., *Class switch recombination to IgE in the bronchial mucosa of atopic and nonatopic patients with asthma.* J Allergy Clin Immunol, 2007. **119**(1): p. 213-8.

13. National Heart, L., and Blood Institute, National Asthma Education and Prevention Program,, *Expert panel report 3: Guidelines for the diagnosis and management of asthma.* 2010.

14. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.

15. Litonjua, Augusto A., et al., *Parental History and the Risk for Childhood Asthma . Does Mother Confer More Risk than Father?* Am. J. Respir. Crit. Care Med., 1998. **158**(1): p. 176-181.

16. Yu, W., et al., *Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations.* Bioinformatics. **26**(1): p. 145-146.

17. Vercelli, D., *Discovering susceptibility genes for asthma and allergy.* Nature Reviews Immunology, 2008. **8**(3): p. 169-182.

18. Zhang, J., P.D. Pare, and A.J. Sandford, *Recent advances in asthma genetics.* Respir Res, 2008. **9**: p. 4.

19. Bouzigon, E., et al., *Meta-analysis of 20 genome-wide linkage studies evidenced new regions linked to asthma and atopy.* Eur J Hum Genet, 2010. **18**(6): p. 700-6.

20. Wills-Karp, M. and S.L. Ewart, *Time to draw breath: asthma-susceptibility genes are identified.* Nat Rev Genet, 2004. **5**(5): p. 376-87.

21. Cookson, W. and M. Moffatt, *Making sense of asthma genes.* N Engl J Med, 2004. **351**(17): p. 1794-6.

22. Cheng, L., et al., *Polymorphisms in ADAM33 are associated with allergic rhinitis due to Japanese cedar pollen.* Clin Exp Allergy, 2004. **34**(8): p. 1192-201.

23. Siroux, V., et al., *Replication of association between ADAM33 polymorphisms and psoriasis.* PLoS One, 2008. **3**(6): p. e2448.

24. Wang, X., et al., *Association of ADAM33 gene polymorphisms with COPD in a northeastern Chinese population.* BMC Medical Genetics, 2009. **10**(1): p. 132.

25. Matsusue, A., et al., *ADAM33 genetic polymorphisms and risk of atopic dermatitis among Japanese children.* Clinical Biochemistry, 2009. **42**(6): p. 477-483.

26. Holloway, J.W. and G.H. Koppelman, *Identifying novel genes contributing to asthma pathogenesis.* Current Opinion in Allergy and Clinical Immunology, 2007. **7**(1): p. 69-74.

27. Jorgenson, E. and J.S. Witte, *A gene-centric approach to genome-wide association studies.* Nat Rev Genet, 2006. **7**(11): p. 885-91.

28. Guerra, S. and F.D. Martinez, *Asthma genetics: from linear to multifactorial approaches.* Annu Rev Med, 2008. **59**: p. 327-41.

29. Samani, N.J., et al., *Genomewide association analysis of coronary artery disease.* New England Journal of Medicine, 2007. **357**(5): p. 443-453.

30. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-U7.

31. Hakonarson, H., et al., *A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene.* Nature, 2007. **448**(7153): p. 591-U7.

32. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-1345.

33. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-1341.

34. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-1336.

35. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-U5.

36. Hancock, D.B., et al., *Genome-Wide Association Study Implicates Chromosome 9q21.31 as a Susceptibility Locus for Asthma in Mexican Children.* PLoS Genet, 2009. **5**(8): p. e1000623.

37. Himes, B.E., et al., *Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene.* Am J Hum Genet, 2009. **84**(5): p. 581-93.

38. Kim, S.H., et al., *Alpha-T-catenin (CTNNA3) gene was identified as a risk variant for toluene diisocyanate-induced asthma by genome-wide association analysis.* Clinical & Experimental Allergy, 2009. **39**(2): p. 203-212.

39. Li, X., et al., *Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions.* Journal of Allergy and Clinical Immunology, 2010. **125**(2): p. 328-335.e11.

40. Mathias, R.A., et al., *A genome-wide association study on African-ancestry populations for asthma.* Journal of Allergy and Clinical Immunology, 2009. **125**(2): p. 336-346.e4.

41. Sleiman, P.M.A., et al., *Variants of DENND1B Associated with Asthma in Children.* N Engl J Med, 2009. **362**(1): p. 36-44.

42. Hancock, D.B., et al., *Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function.* Nat Genet, 2009. **42**(1): p. 45-52.

43. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nat Genet, 2009. **42**(1): p. 36-44.

44. Wilk, J., et al., *Framingham Heart Study genome-wide association: results for pulmonary function measures.* BMC Medical Genetics, 2007. **8**(Suppl 1): p. S8.

45. Wilk, J.B., et al., *A Genome-Wide Association Study of Pulmonary Function Measures in the Framingham Heart Study.* PLoS Genet, 2009. **5**(3): p. e1000429.

46. Capon, F., et al., *Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene.* Hum. Mol. Genet., 2008. **17**(13): p. 1938-1945.

47. Liu, Y., et al., *A Genome-Wide Association Study of Psoriasis and Psoriatic Arthritis Identifies New Disease Loci.* PLoS Genet, 2008. **4**(3): p. e1000041.

48. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-[kappa]B pathways.* Nat Genet, 2009. **41**(2): p. 199-204.

49. Zhang, X.-J., et al., *Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21.* Nat Genet, 2009. **41**(2): p. 205-210.

50. Ober, C., et al., *Effect of Variation in CHI3L1 on Serum YKL-40 Level, Risk of Asthma, and Lung Function.* N Engl J Med, 2008. **358**(16): p. 1682-1691.

51. Castro-Giner, F., et al., *A pooling-based genome-wide analysis identifies new potential candidate genes for atopy in the European Community Respiratory Health Survey (ECRHS).* BMC Medical Genetics, 2009. **10**(1): p. 128.

52. Esparza-Gordillo, J., et al., *A common variant on chromosome 11q13 is associated with atopic dermatitis.* Nat Genet, 2009. **41**(5): p. 596-601.

53. Gudbjartsson, D.F., et al., *Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.* Nat Genet, 2009. **41**(3): p. 342-347.

54. Galanter, J., et al., *ORMDL3 gene is associated with asthma in three ethnically diverse populations.* American Journal of Respiratory and Critical Care Medicine, 2008. **177**(11): p. 1194-1200.

55. Madore, A.M., et al., *Replication of an association between 17q21 SNPs and asthma in a French-Canadian familial collection.* Human Genetics, 2008. **123**(1): p. 93-95.

56. Tavendale, R., et al., *A polymorphism controlling ORMDL3 expression is associated with asthma that is poorly controlled by current medications.* Journal of Allergy and Clinical Immunology, 2008. **121**(4): p. 860-863.

57. Verlaan, D.J., et al., *Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease.* Am J Hum Genet, 2009. **85**(3): p. 377-93.

58. Bouzigon, E., et al., *Effect of 17q21 variants and smoking exposure in early-onset asthma.* N Engl J Med, 2008. **359**(19): p. 1985-94.

59. Weiner, M.P., S.B. Gabriel, and J.C. Stephens, *Genetic variation : a laboratory manual*. 2007, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. xxi, 472 p.

60. Butcher, L.M., et al., *SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children.* Human Molecular Genetics, 2005. **14**(10): p. 1315-1325.

61. Skol, A.D., et al., *Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.* Nat Genet, 2006. **38**(2): p. 209-13.

62. Rice, T.K., N.J. Schork, and D.C. Rao, *Methods for handling multiple testing.* Adv Genet, 2008. **60**: p. 293-308.

63. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nature Reviews Genetics, 2005. **6**(2): p. 95-108.

64. Ioannidis, J.P., G. Thomas, and M.J. Daly, *Validating, augmenting and refining genome-wide association signals.* Nat Rev Genet, 2009. **10**(5): p. 318-29.

65. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature Reviews Genetics, 2008. **9**(5): p. 356-369.

66. Kurukulaaratchy, R.J., et al., *Characterization of wheezing phenotypes in the first 10 years of life.* Clinical and Experimental Allergy, 2003. **33**(5): p. 573-578.

67. *Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee.* Lancet, 1998. **351**(9111): p. 1225-32.

68. Isle of Wight Council, *Mid-year Population Statistics*. 2007: Newport, Isle of Wight, UK.

69. World Health Organization. Prevalence of asthma and allergies in children 2007 [cited; Available from: http://www.euro.who.int/Document/EHI/ENHIS_Factsheet_3_1.pdf.

70. Kurukulaaratchy, R.J., S. Matthews, and S.H. Arshad, *Does environment mediate earlier onset of the persistent childhood asthma phenotype?* Pediatrics, 2004. **113**(2): p. 345-50.

71. Karmaus, W., et al., *Long-term effects of breastfeeding, maternal smoking during pregnancy, and recurrent lower respiratory tract infections on asthma in children.* J Asthma, 2008. **45**(8): p. 688-95.

72. Ogbuanu, I.U., et al., *Effect of breastfeeding duration on lung function at age 10 years: a prospective birth cohort study.* Thorax, 2009. **64**(1): p. 62-6.

73. Arshad, S.H., M. Stevens, and D.W. Hide, *The effect of genetic and environmental factors on the prevalence of allergic disorders at the age of two years.* Clinical & Experimental Allergy, 1993. **23**(6): p. 504-511.

74. Arshad, S.H., et al., *Polymorphisms in the interleukin 13 and GATA binding protein 3 genes and the development of eczema during childhood.* Br J Dermatol, 2008. **158**(6): p. 1315-22.

75. Huebner, M., et al., *Patterns of GATA3 and IL13 gene polymorphisms associated with childhood rhinitis and atopy in a birth cohort.* Journal of Allergy and Clinical Immunology, 2008. **121**(2): p. 408-414.

76. Sadeghnejad, A., et al., *IL13 gene polymorphisms modify the effect of exposure to tobacco smoke on persistent wheeze and asthma in childhood, a longitudinal study.* Respiratory research, 2008. **9**: p. 10.

77. Ramadas, R.A., et al., *Interleukin-1R antagonist gene and pre-natal smoke exposure are associated with childhood asthma.* Eur Respir J, 2007. **29**(3): p. 502-508.

78. Ogbuanu, I., et al., *Birth order modifies the effect of IL13 gene polymorphisms on serum IgE at age 10 and skin prick test at ages 4, 10 and 18: a prospective birth cohort study.* Allergy, Asthma & Clinical Immunology. **6**(1): p. 6.

CHAPTER 2

Identification of genes associated with childhood asthma

This chapter is a modified version of an accepted manuscript. It has been modified to include more thorough details that were not able to be included in the submission due to space limits. Please refer to publication for updated version.

Schauberger EM, Ewart SL, Arshad SH, Huebner M, Karmaus W, Holloway JW, Friderici KH, Ziegler JT, Zhang H, Rose-Zerilli MJ, Barton SJ, Holgate ST, Kilpatrick JR, Harley JB, Lajoie-Kadoch S, Harley IT, Hamid Q, Kurukulaaratchy RJ, Seibold MA, Avila PC, Rodriguez-Cintrón W, Rodriguez-Santana JR, Hu D, Gignoux C, Romieu I, London SJ, Burchard EG, Langefeld CD, Wills-Karp M. Identification of ATPAF1 as a novel candidate gene for asthma in children. Journal Allergy Clinical Immunology. 2011 Jun 20. [Epub with no published date provided at time of dissertation]. PMID: 21696813

ABSTRACT

Background: Asthma is a common disease of children with a complex genetic origin. Understanding the genetic basis of asthma susceptibility will allow disease prediction and risk stratification.

Objective: We sought to identify asthma susceptibility genes in children.

Methods: A nested case-control genetic association study of Caucasian children from a birth cohort was conducted. Single nucleotide polymorphisms (SNPs, N=116,024) were genotyped in pools of DNA samples from cohort children with physician-diagnosed asthma (N=112) and normal controls (N=165). A genomic region containing the *ATPAF1* gene was significantly associated with asthma. Additional SNPs within this region were genotyped in individual samples from the same children and in five independent study populations consisting of Caucasian, Puerto Rican, Mexican, or various ancestries. SNPs were also genotyped, and if missing imputed, in two control populations. *ATPAF1* expression was measured in bronchial biopsies from asthmatics and controls.

Results: Asthma was associated with a cluster of SNPs and SNP haplotypes containing the *ATPAF1* gene with two SNPs achieving significance at a genome-wide level (p=2.26E-5 to 2.2E-8). SNP and gene-level associations were confirmed in three populations, but not in the Mexican populations. Haplotype association was confirmed in the Caucasian population. ATPAF1 mRNA expression was significantly (p=0.01) higher in bronchial biopsies from asthmatics than controls.

Conclusion: Genetic variation in the *ATPAF1* gene predisposes children of different ancestry to asthma.

INTRODUCTION

Asthma is a debilitating chronic inflammatory disease of the conducting airways whose symptoms often manifest early in childhood. Many affected children struggle with this disease throughout their lives. Asthma is remarkably common, with the prevalence in children exceeding 30% in some parts of the world [1, 2]. Indeed, it is a disease of concern world-wide with particularly high incidence in Northwestern Europe, the USA, and in some populations of Hispanic ancestry [2-4]. Gene variations, in tandem with environmental factors, are believed to be the primary drivers behind asthma development and symptom exacerbations. Therefore, we undertook a study to determine susceptibility genes for asthma in a birth cohort of children from the U.K. with the rationale that understanding genetic factors will allow us to predict disease risk. Despite numerous studies, few genes have emerged as underlying asthma in the majority of populations examined [5-7] thus, the hunt for asthma susceptibility genes continues.

The resources available for genetic association studies have expanded tremendously in recent years on several fronts. First, technology for genotyping has advanced such that hundreds of thousands of genotypes can be generated as a matter of course and simultaneously the methods for analyzing these massive datasets have been devised and refined. In addition, there are a growing number of populations thoroughly characterized for asthma and related phenotypes that have subsequently been extensively genotyped. Consequently, these populations provide a powerful means for assessing the genetic contributions to asthma. The Isle of Wight birth-cohort is one such population that was established over 20 years ago for the purpose of investigating asthma during childhood. The replication populations that were examined for this report are similarly well-established, providing access to children with asthma and their families. The more

recent genetic data generated in these populations can provide a new dimension to our understanding of asthma in children.

METHODS

Study design

The objective of our study was to identify asthma susceptibility genes in our birth cohort of children, which has been extensively characterized for asthma and allergies [8]. We used an efficient and sequential strategy to optimize the search for asthma susceptibility genes (Figure 5). Specifically, we first genotyped 116,204 SNPs in pooled DNA samples (N=5 pools) from Isle of Wight cohort children with physician-diagnosed asthma (N=112 children) and pooled DNA samples (N=5 pools) from cohort children without evidence of asthma or wheeze at any time from birth through 10 years of age (N=165 children). A genomic region, containing the *ATPAF1* gene was significantly associated with asthma (Figure 6 and 7). Additional SNPs within this region, were genotyped in individual samples from the same children, who were of European ancestry, and in five independent replicate study populations consisting of European (one affected sib pair family study), Puerto Rican (one case-parent trios study), and Mexican (two case-parent trios studies) ancestry, as well as in the childhood studies from the National Heart, Lung, and Blood Institute SNP Health Association Resource (SHARe) Asthma Resource Project (SHARP) consisting of case-parent trios of European American, African American, Hispanic, or other ancestry [9]. SNPs were also genotyped or imputed in consortia controls from the International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN) [10] and controls from the Wellcome Trust Case Control Consortium (WTCCC)[11].

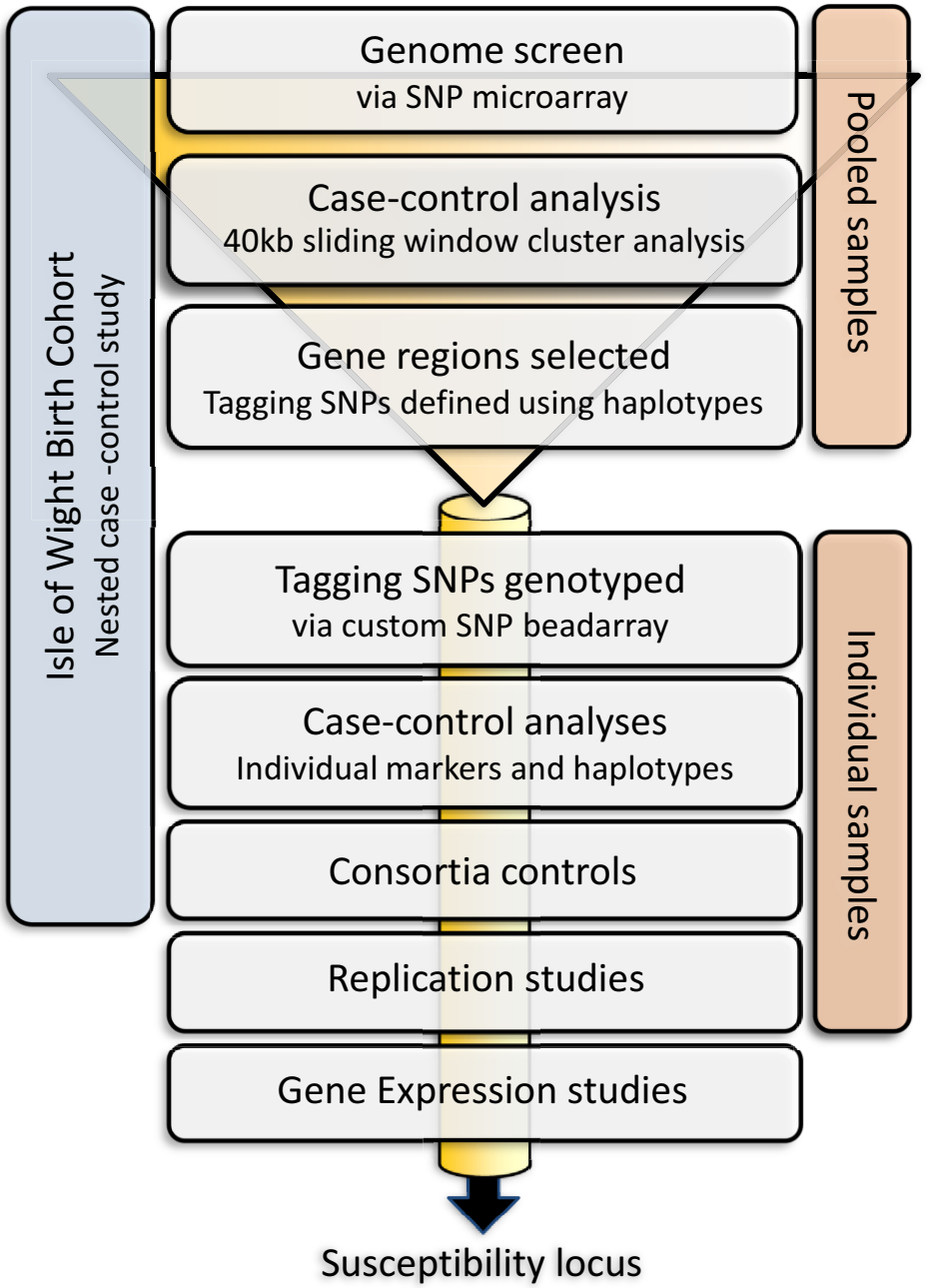

Figure 5: The study design incorporated a sequential strategy to identify asthma associations in the primary population, pursuit of validation in five replication populations, and evidence of functional relevance in asthmatics.

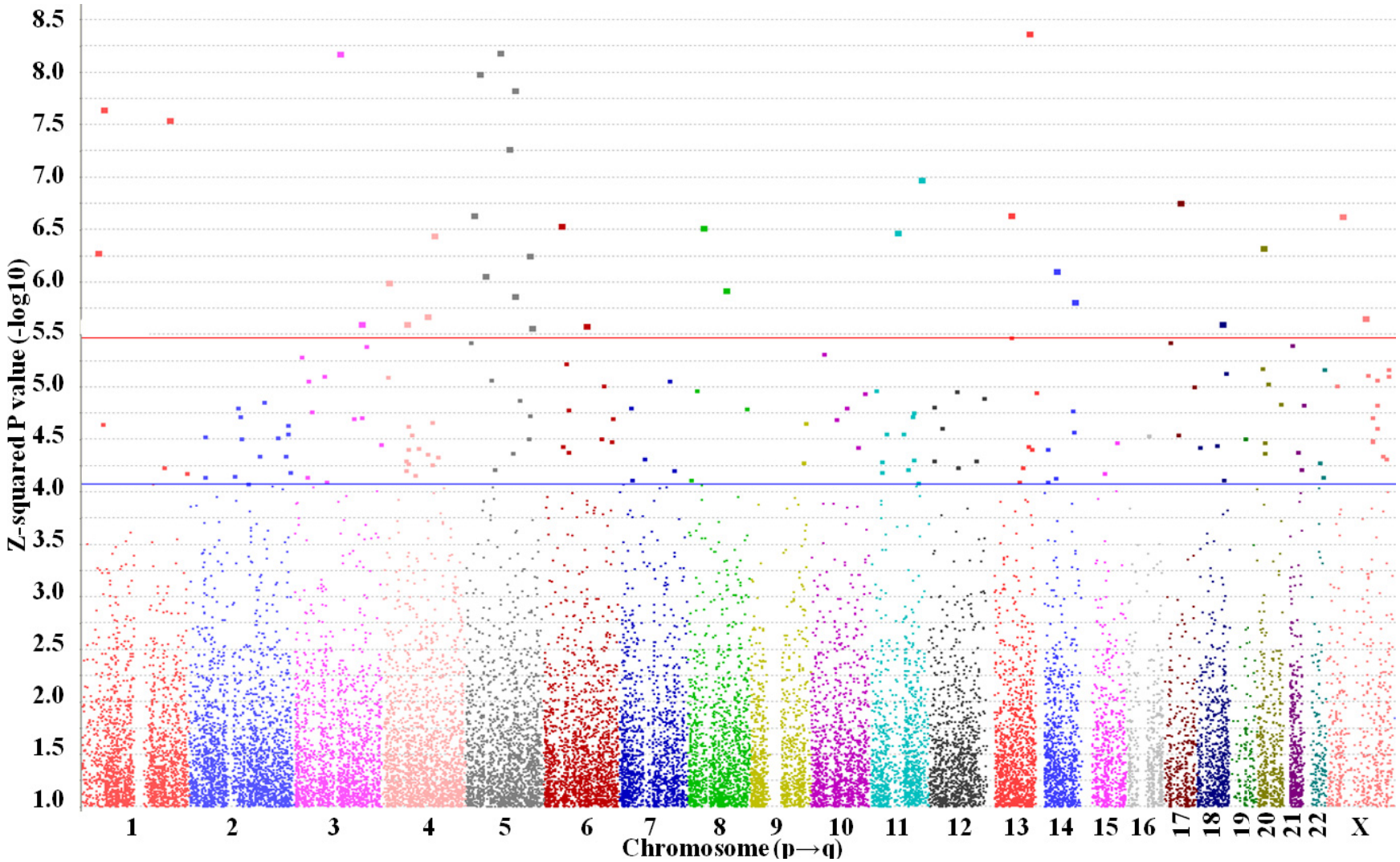

Figure 6: Genome-wide association results of SNP microarray for asthma vs control. The horizontal axis is genomic location from Chromosomes 1 through 22 in the orientation of p to q (left to right). Vertical axis is the Z2 P values displayed in a -logarithmic scale. The 1% and 5% False Discovery Rate (FDR) are indicated by red and blue lines respectively.

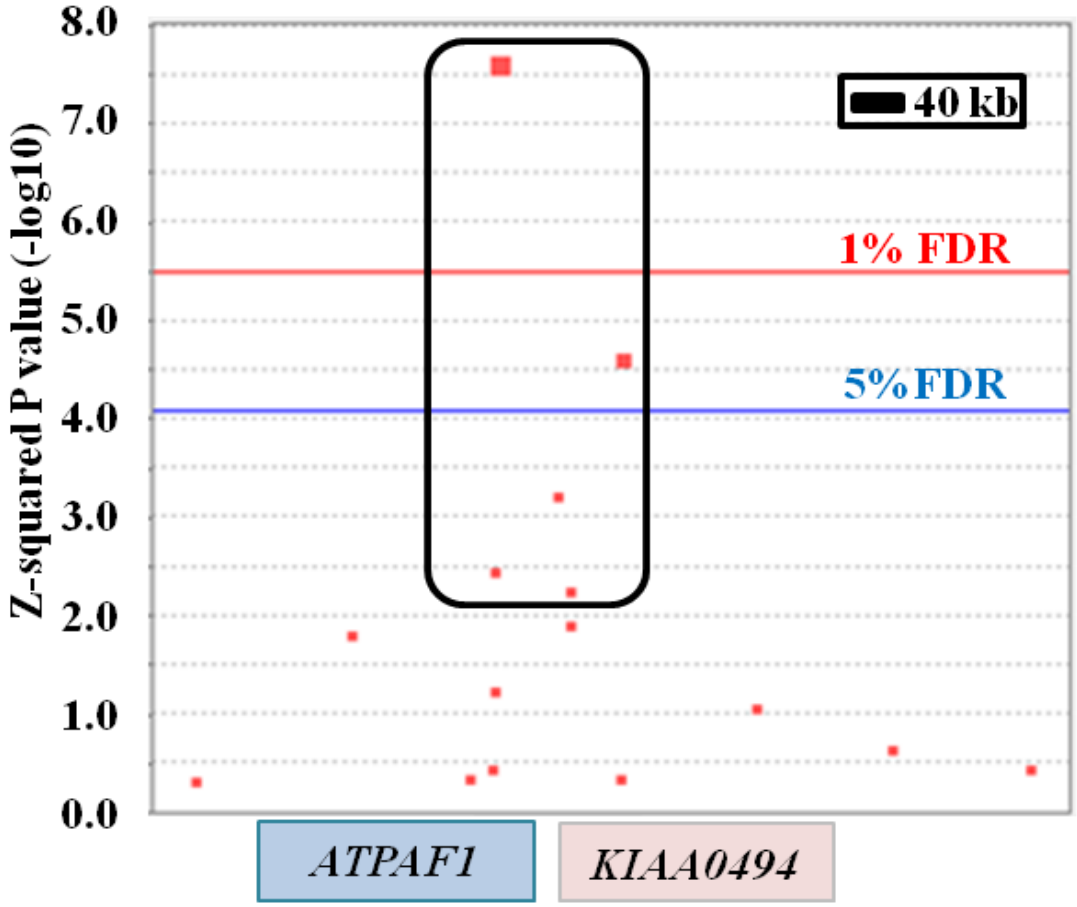

Figure 7: Region on Chromosome 1p33 with a sustained significance. Z2 P value range = 2.2E-8 to 0.0124 across a cluster of 5 SNPs (circled). The 1% and 5% False Discovery Rate (FDR) are indicated. The location of the gene, *C1ORF223*, is not illustrated but is located in between the *ATPAF1* and *KIAA0494* genes.

Subjects

Isle of Wight birth-cohort

The primary study population consisted of children (N=1,456) born and enrolled between January 1, 1989 and February 28, 1990 in the Isle of Wight, U.K. whole population birth-cohort study[8]. A nested case-control study was conducted on a subset of the cohort children (N=277). Children were phenotyped for asthma at ages 1, 2, 4 and 10 years, with asthma diagnosis at age 10 (N=112) based on a minimum criteria of physician-diagnosed asthma plus wheeze in the previous 12 months, using a validated questionnaire [12] and controls (N=165) randomly selected from among cohort children who had never been given a diagnosis of asthma and, in addition, had never wheezed in their life.

Consortia controls

Additional consortium controls were acquired from 1) the WTCCC (N=3,004), which derive from European Caucasian control subjects[11] from the 1958 British Birth Cohort and from blood donors recruited by the three national UK Blood Services, and 2) the SLEGEN (N=3,471) containing females from UK and USA origins [10]. The asthma status of the consortia controls is unknown.

Replication populations

The five additional populations of asthmatic children examined for replication purposes have been described elsewhere [13-18] and their key characteristics are summarized in Table 4.

Wessex

The Wessex population (University of Southampton, 1997-98 Family-based study) consisted of families with at least two biologic siblings with a current physician's diagnosis of asthma and who were taking asthma medication on a regular basis [18]. Phenotyping was based on validated health survey questionnaires completed by each family member. During 1997-98, 1,481 individuals in 341 families were recruited. Selection criteria included families with at least two children in each family with asthma. This resulted in families consisting of children aged 5 to 21 years with an average size of 4.3 children per family and a range of 2 to 11 children. There were a total of 836 cases (physician diagnosed with current use of asthma medication) and 624 controls (negative for asthma). The children's age, ethnicity and socioeconomic background were similar to the Isle of Wight population and they were recruited from the same geographical area. Similar to the Isle of Wight cohort, information on several asthma-related phenotypes were collected. However, unlike the Isle of Wight cohort which is a case-control study, the Wessex study is a family-based study design [18].

Table 4: Characteristics of the Isle of Wight primary population, consortia controls, and replication populations

	Isle of Wight 1989-1990 Birth Cohort (IOW)	International Consortium for SLE Genetics (SLEGEN)	Wellcome Trust Case Control Consortium (WTCCC)	Wessex family study	Mexico Childhood Asthma Study	Genetics of Asthma in Latino Americans (GALA)		Childhood Asthma Management Program (CAMP)	Childhood Asthma Research and Education (CARE)
Location	UK	UK & US	UK	UK	MX	PR & US	MX & US	US	US
Ancestry	Caucasian	Caucasian	Caucasian	Caucasian	Mexican	Puerto Rican	Mexican	Caucasian, African American, Hispanic, other	Caucasian, African American, Hispanic, other
Population structure	Nested case-control	Controls	Controls	Family-based	Case-parent trios	Case-parent trios	Case-parent trios	Case-parent trios	Case-parent trios
Numbers	112 cases; 165 controls	3,471 controls	3,004 controls	341 familes; 1,481 individuals	492 trios	395 trios	298 trios	442 trios #	131 trios †
Age of cases	10 yr	N/A	N/A	5-21 yr	5-17 yr	8-40 yr $	8-40 yr $	5-12 yr	24-35 mo & 6-14 yr
Asthma criteria	PD asthma + symptoms	N/A	N/A	PD asthma + medications	PD asthma	PD asthma + symptoms	PD asthma + symptoms	Asthma symptoms or medications	Asthma symptoms

PD, Physician diagnosed; N/A, not available; UK, United Kingdom; $ Asthma onset during childhood

Number of CAMP-affected offspring trios per ancestra group: Caucasian N= 334, African American N=42, Hispanic N=30, other

†Number of CARE-affected offspring trios per ancestral group: Caucasian N=95, African American N=10, Hispanic N=16, other

GALA

A Puerto Rican and one Mexican population were from the Genetics of Asthma in Latino Americans (GALA) study [14] in which Puerto Rican subjects (N=395 trios) were recruited in New York, NY and Puerto Rico and Mexican subjects (N=298 trios) were recruited in San Francisco, CA and Mexico City, Mexico. These populations consisted of asthma cases (diagnosed during childhood) and both parents, with cases defined as having physician-diagnosed asthma and two or more symptoms (wheeze, cough, shortness of breath) in the past two years.

Mexico Childhood Asthma Study

The second Mexican population, the Mexico Childhood Asthma Study [15], was also a case-parent trio design recruited from Mexico City, Mexico in which cases (N=600) were children (age 5-17) with a diagnosis of asthma by a pediatric allergist based on clinical symptoms and response to treatment.

SNP Health Association Resource Asthma Resource Project studies

The fifth replication population derived from the publically available data from the SNP Health Association Resource Asthma Resource Project studies [9, 19], (study accession: phs000166.v1.p1) which contains families from the Childhood Asthma Research and Education (CARE, N=1,492 [20]) network and Childhood Asthma Management Program (CAMP,

N=1,738) [9, 13, 17]). The CAMP study contained children (age 5 to 12 years) with cases defined as chronic asthmatic (medication use and either asthma symptoms at least twice per week or at least two usages per week of an inhaled bronchodilator). The CARE study contained children (24-35 months or 6-14 years of age) with cases having a positive asthma predicted index based on at least three exacerbations of wheezing during the previous twelve months.

Ethics approval was granted by local research ethics committees for the Isle of Wight population and written (parental) consent was obtained. Ethics approval for each of the other populations has been reported previously [13-18] and permission for access to the consortia and SHARP data were obtained [9].

Gene expression studies were conducted on bronchial biopsy samples collected from asthmatic (severe asthma (N=8) with FEV1<80% predicted and on oral or inhaled steroids; or mild asthma (N=5) FEV1>80% predicted, on bronchodilators only) and control (normal FEV1 and PC20, nonatopic, N=4) individuals from Montreal, Quebec, Canada [21, 22]. A more detailed description of this population is below.

Genetic association analyses

The initial analyses were performed on pools consisting of equimolar amounts of DNA from asthmatic (5 pools, N=22-23 subjects/pool) and control (5 pools, N=30-37 subjects/pool) children from the Isle of Wight population. The pools were allelotyped with Affymetrix GeneChip Human Mapping 100K arrays (Santa Clara, CA, USA; 1 array per DNA pool, total 10 microarrays) to determine the collective allele frequencies of each pool (see Supplementary

Methods for detailed description). Hybridization intensity comparisons of the case and control pools were used to identify significant allele frequency differences for each SNP [23]. The data were normalized to controls and ranked based on Z2 p values of case-vs-control frequencies. An allelic test was used in the analysis of pooled SNP microarray data as individual genotypes were not available. After quality control procedures, the 100K SNP microarrays yielded a set of 98,921 SNPs sufficient for analysis (Figure 6). A whole genome false discovery rate cut-off of $\alpha = 0.05$ was determined to have a Z2 P value=2.27E-5. To prioritize genes of interest, a 40 kb sliding window identified clusters of significant SNPs. Parameters of cluster size, SNP numbers, and level of significance were considered with priority given to small clusters of high significance. This analysis yielded 60 clusters throughout the genome with a Z2<0.005. The cluster containing the individual SNP with the highest rank among all SNPs on the microarray (rs2289447, Z2 P value = 2E-8, rank 6th) was located on chromosome 1p33 which contains the genes *ATPAF1* (ATP synthase mitochondrial F1 complex assembly factor 1) and *KIAA0494* and *C1ORF223* (both uncharacterized) (Figure 7, Table 5, Table 6). This region had sustained significance (Z2 P value range = 0.0124 to 2.2E-8) across a cluster of 6 SNPs (rs2289447, rs1150068, rs1048380, rs2275380, rs1150064 and rs1440486) and was therefore, selected for further study.

To investigate the genes on chromosome 1p33 identified by the sliding-window analysis, this region was targeted for genotyping in individuals from the nested study. Nine SNPs identified in the gene discovery analysis plus an additional four tagging SNPs were selected using Caucasian (CEU Population, Release 21a/Phase II Jan 07, B35 data set) Hapmap data [24]. Tagger (http://www.broad.mit.edu/mpg/tagger/)[25] was used to Tag SNP selection (SNPs with MAF of ≥10% and pairwise R2 threshold of 100%) within and 10kb surrounding (5' and 3') the region

Table 5: Genetic associations for asthma on chromosome 1p33-p32.31 in the Isle of Wight and replication populations

				Primary Population (IOW)			
				IOW pooled GWAS	**Individual IOW samples**	**IOW + SLEGEN**	**IOW + WTCCC**
			Genotyping platform	Affymetrix 100K GeneChip array	Custom Illumina Goldengate	Illumina HumanHap300	Affymetrix 500K GeneChip array
			Model	Allelic	Additive	Additive	Additive
			Software	LatteThunder	SNPGWA	SNPGWA	SNPGWA
			Statistical test	Z2 p-value$	Cochran-Armitage p-value	Cochran-Armitage p-value	Cochran-Armitage p-value
SNP	Chr1 Position	IOW Minor allele	IOW MAF				
rs1258000	47094267	G	0.296	-	0.0282	0.0024 (0.0024)	(0.2090)
rs2289447	47118168	T	0.255	2.20E-08	0.0156	(0.0001)	(0.0070)
rs620431	47118189	A	0.283	-	0.0091	(0.0004)	(0.0104)
rs1150068	47118918	C	0.269	0.0034	0.0065	(0.0002)	(0.0068)
rs654509	47127171	G#	0.017	-	-	-	-
rs601060	47139580	G#	0.241	-	-	-	-
rs1048380	47142538	T	0.265	0.0006	0.0044	(0.0001)	(0.0067)
rs2275380	47147728	G	0.509	0.0124	0.0612	0.0171 (0.01708)	0.0587 (0.0573)
rs1150064	47148044	T	0.269	0.0053	0.0084	(8.79E-05)	(0.0066)
rs6665021	47152520	G#	0.058	-	-	-	-
rs4660956	47158486	T#	0.242	-	-	-	-
rs1440487	47167075	T	0.233	0.4309	0.7177	0.2167 (0.1878)	0.7623 (0.7025)
rs1440486	47167238	A	0.265	2.26E-05	0.0095	(0.0002)	(0.0105)
rs10749863	47174328	G#	0.246	-	-	-	-
rs2218189	47176818	G	0.269	-	0.0060	0.0001 (0.0001)	(0.0077)
rs6670495	47187908	A	0.250	-	0.0060	(7.93E-06)	(0.0009)
rs6662321	47220059	T#	0.042	-	-	-	-

P-values from imputed data indicated in parentheses

#Minor allele/MAF from HapMap CEU data when IOW data not available

$False discovery rate cut-off for α=0.05 is Z2 p-value=2.27E-05

Table 5 (cont'd)

	Replication Populations				
	Wessex	**GALA**		**Mexico Childhood Asthma Study**	**CAMP+CARE**
Genotyping platform	Kbiosciences Kaspar Competitive PCR	ABI Taqman/ Affymetrix 6.0		Illumina 550	Affymetrix 6.0
Model	Allelic	Dominant		Allelic	Allelic
Software	PLINK	FBAT		PLINK	PDTphase
Statistical test	TDT p-value	TDT p-value		TDT p-value	PDT p-value
SNP		Puerto Rican	Mexican		
rs1258000	0.1113	0.0230	0.0750	0.4503	-
rs2289447	0.1216	-	-	-	-
rs620431	0.2476	0.1680	0.2210	-	-
rs1150068	0.2286	-	-	-	-
rs654509	-	N/A	N/A	-	0.0047
rs601060	-	0.0253	N/A	-	0.0588
rs1048380	0.0926	-	-	-	-
rs2275380	0.0312	0.4270	0.5910	0.5525	0.2386
rs1150064	0.1576	-	-	-	-
rs6665021	-	N/A	N/A	-	0.0196
rs4660956	-	0.0431	N/A	-	0.0782
rs1440487	0.2399	0.3400	0.1060	0.3173	0.0620
rs1440486	0.0926	-	-	-	-
rs10749863	-	0.0304	N/A		0.0495
rs2218189	0.1576	-	-	0.6152	-
rs6670495	0.2222	0.6240	0.3800	-	-
rs6662321	-	-	-	-	0.0460

P-values from imputed data indicated in parentheses

$False discovery rate cut-off for α=0.05 is Z2 p-value=2.27E-5

N/A= result not available due to too few rare homozygus individuals needed for dominant model test

Table 6. Allele frequencies of pooled DNA samples from Isle of Wight children for SNPs within and flanking *ATPAF1, C1ORF223,* and *KIAA0494* genes

SNP	Gene Symbol	Function	Chr 1 Position	Asthma					Control				
				Pool 1 (n=)22	Pool 2 22	Pool 3 22	Pool 4 23	Pool 5 23	Pool 1 30	Pool 2 30	Pool 3 30	Pool 4 37	Pool 5 37
rs10489769	NSUN4	intron	46579290	0.642	0.799	0.720	0.702	0.784	0.727	0.706	0.752	0.721	0.570
rs10489770	NSUN4	intron	46580184	0.826	0.757	0.749	0.739	0.793	0.717	0.769	0.766	0.794	0.866
rs952947			46775476	0.437	0.224	0.205	0.182	0.218	0.300	0.231	0.134	0.230	0.329
rs6429606	MKNK1	intron	46835487	0.223	0.342	0.141	0.328	0.126	0.216	0.296	0.267	0.356	0.246
* rs2181412			46869234	0	0	0	0	0	0	0	0	0	0
rs1273237	ATPAF1	intron	46881838	0.045	0.042	0.060	0.060	0.070	0.059	0.051	0.054	0.053	0.043
* rs1933932	ATPAF1	intron	46881915	0	0	0	0	0	0	0	0	0	0
rs631840	ATPAF1	intron	46890531	0.920	0.981	0.896	0.938	0.919	0.967	0.872	0.892	0.962	1
rs620913	ATPAF1	intron	46890654	0.708	0.570	0.665	0.575	0.580	0.534	0.631	0.506	0.479	0.558
rs2289447	ATPAF1	intron	46890755	0.969	0.831	0.920	0.957	0.730	0.741	0.578	0.570	0.761	0.674
rs1150068	ATPAF1	intron	46891505	0.805	0.708	0.706	0.795	0.739	0.676	0.577	0.687	0.587	0.648
rs1048380	KIAA0494	3' UTR	46915125	0.731	0.689	0.825	0.825	0.733	0.597	0.695	0.698	0.647	0.489
rs2275380	KIAA0494	intron	46920315	0.609	0.397	0.379	0.333	0.414	0.474	0.567	0.447	0.513	0.632
rs1150064	KIAA0494	intron	46920631	0.229	0.278	0.280	0.223	0.308	0.351	0.395	0.394	0.359	0.382
rs1440487	KIAA0494	intron	46939662	0.556	0.868	0.777	0.742	0.753	0.759	0.642	0.899	0.655	0.725
rs1440486	KIAA0494	intron	46939825	0.098	0.198	0.180	0.152	0.248	0.222	0.492	0.396	0.168	0.451
rs720413			46992410	0.485	0.519	0.415	0.492	0.440	0.336	0.540	0.346	0.383	0.424
* rs10493124			47002322	0	0	0	0	0	0	0	0	0	0
rs2405335	CYP4B1	intron	47044772	0.2694	0.2939	0.2428	0.1290	0.4212	0.2575	0.2504	0.2810	0.2048	0.2174
rs10493125			47098257	0.2501	0.0897	0.1155	0.1670	0.1197	0.0618	0.1615	0.1982	0.2231	0.1577
rs1002378	CYP4Z1	intron	47317899	0.4859	0.4587	0.4463	0.5505	0.3939	0.4959	0.4471	0.4640	0.4226	0.4388
rs2405340	CYP4Z1	intron	47322106	0.5046	0.5179	0.4573	0.6875	0.5523	0.2817	0.3415	0.5722	0.5372	0.5261
rs1343294	CYP4A22	intron	47377339	0.8217	0.7462	0.8115	0.8748	0.5377	0.6459	0.7217	0.7635	0.8552	0.8008

*SNPs monomorphic in HapMap CEU population; n = number of DNA samples per pool

containing the *ATPAF1*, *C1ORF223,* and *KIAA0494* genes. SNPs were genotyped in individual cohort children (N=277) using the Illumina GoldenGate Custom Panel (San Diego, CA) bead array (detailed description in Supplementary Methods).

Haploview[26] was used to examine Hardy-Weinberg equilibrium, minor allele frequencies, and linkage disequilibrium (LD). Cochran-Armitage trend tests implemented with SNPGWA software (v.4.0)[10] were used to determine asthma association with individual polymorphisms in Isle of Wight subjects and in Isle of Wight and consortia data combined. The three genes were found to be within a single LD block (Figure 8A and B). Thus, the targeted SNP approach validated the pooled DNA genome screen methodology. Odds ratios for SNPs were calculated using SNP.assoc (v.1.4)[27] implemented in R (v.2.5.1) [28]. SNPs that were non-informative or outside the LD block were excluded and the remaining 11 SNPs were used to establish haplotypes (Table 7, Figure 8A). Haplotypes with population frequency >5% (haplotypes I, II, III) were examined. Chi-square p-values determined with PLINK software (v.1.0) [29] were used to detect asthma associations with haplotypes. Significant association was identified if p-value was <0.05. Regression models were used to compute odds ratios for the haplotypes for the case-control data using SNPassoc (v.1.4)[27] and haplo.stats (v.1.3.8)[30] programs implemented in R (v.2.5.1),[28] incorporating haplotype ambiguity.

Consortia control data from SLEGEN and WTCCC were independently combined with Isle of Wight control data. Several SNPs in common with the Isle of Wight SNPs had been directly genotyped in the SLEGEN and WTCCC control populations. In addition, a high degree of LD in this region allowed imputation to infer additional genotypes using BimBam software[31] and CEU Hapmap build 36 [24]. Hardy Weinberg equilibrium was calculated for both WTCCC and

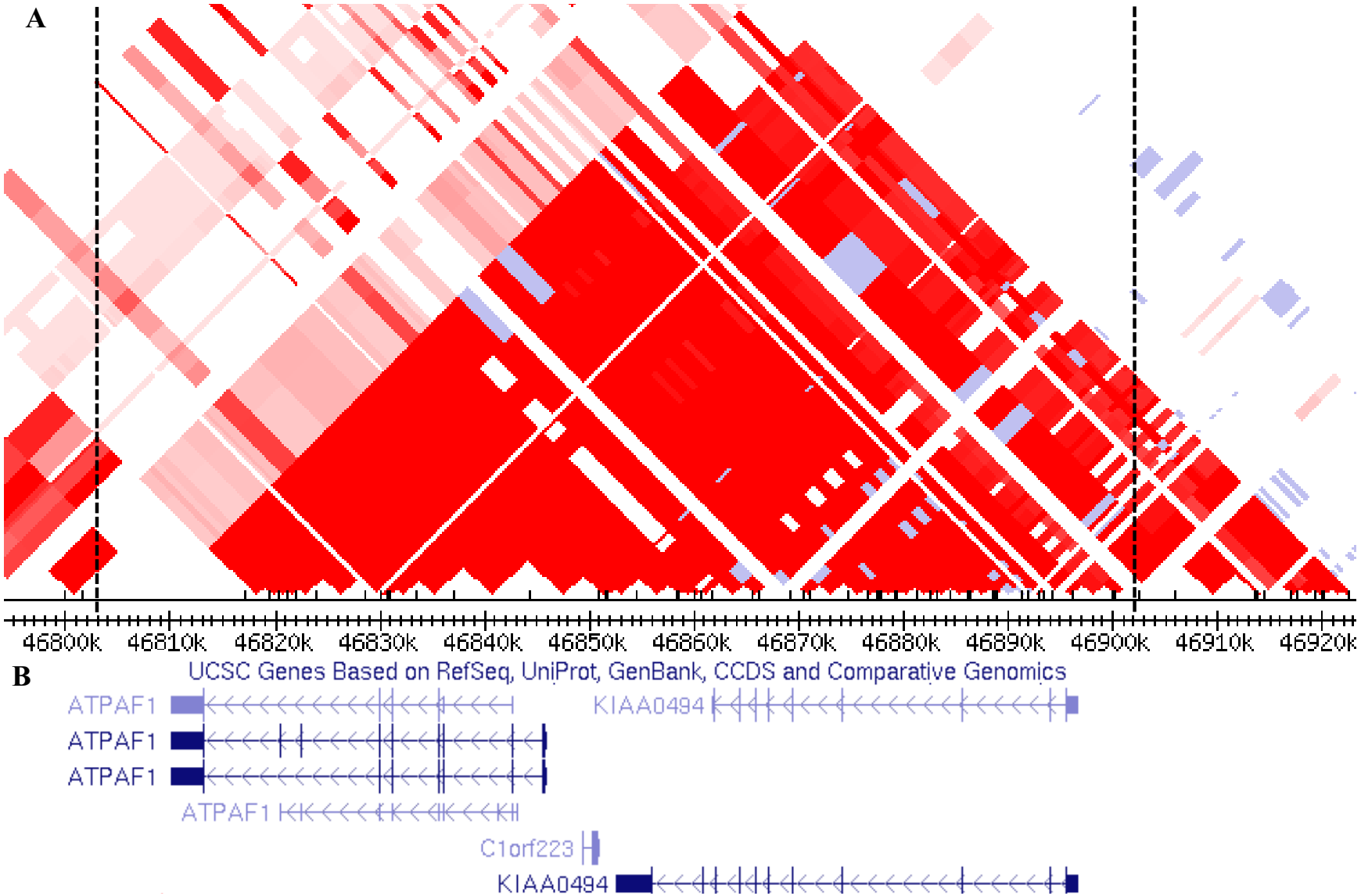

Figure 8: *ATPAF1*, *C1ORF223, KIAA0494* genes HapMap with the (A) LOD linkage disequilibrium plot with block pattern based on confidence intervals (Gabriel et al) [26] (B), and the UCSC gene track [24, 45].

Table 7. Haplotype associations for asthma in children of European ancestry

													Isle of Wight Case-Control Study $				Wessex Family-Based Study #				
Haplotype	rs1258000	rs2289447	rs620431	rs1150068	rs1048380	rs2275380	rs1150064	rs1440487	rs1440486	rs2218189	rs6670495		Haplotype Frequency Controls	Haplotype Frequency Cases	χ^2	χ^2 p value	Haplotype Frequency Parents	Transmitted Haplotypes	Untransmitted Haplotypes	χ^2	χ^2 p value
I	A	C	C	T	C	A	A	C	G	A	T		0.4796	0.5736	4.4450	0.0350	0.4807	255.9	204	5.8520	0.0156
II	A	C	C	T	C	G	A	T	G	A	T		0.2243	0.2158	0.0524	0.8190	0.2210	154.1	185	2.8100	0.0937
III	G	T	A	C	T	G	T	C	A	G	A		0.2452	0.1436	7.9600	0.0048	0.2140	147	168	1.4000	0.2367

$ Isle of Wight data based on individual data from SNP beadarray; asthma cases (n=109), control (n=163)

Wessex population of 341 families (n=1,481 total individuals)

SLEGEN. SNPGWA [10] software with an additive model using a Cochran-Armitage trend test was used to determine statistical significance ($p<0.05$).

SNPs in 1p33-p32.31, including the tag SNPs associated with asthma in the Isle of Wight population, were examined in the replication populations. Individuals in the Wessex population were genotyped for 13 SNPs in the LD block using KASPar competitive allele-specific PCR (Kbiosciences, Herts, U.K.) Asthma associations with individual polymorphisms in the Wessex families were determined by transmission disequilibrium tests (TDT)[29] for genotyped data. Haplotype associations were determined by TDT Chi-square p-values [29]. Odds ratios for the haplotypes were determined by the same method as in the Isle of Wight population. The GALA Puerto Rican and Mexican populations were analyzed using family-based association test (FBAT, v.2.03) with an additive model to assess associations between the asthma and individual variants, as well as haplotypes [32]. The Mexico Childhood Asthma Study population was analyzed for an allelic association using a TDT implemented in PLINK [29]. Genotype data from families in the CAMP [13] and CARE [20] studies were available in the 1p33-p32.31 region with 52 SNPs in the *ATPAF1-C1ORF223-KIAA0494* LD block. Single SNP as well as two SNP and three SNP sliding windows across the region were analyzed by Chi-square using PDTphase [33]. Hardy-Weinberg equilibrium had been examined in all populations, Mendelian errors checked within the families, and erroneous genotypes set as missing data.

The initial association analysis done in the Isle of Wight population was performed on a genome-wide level and correction for multiple testing was performed. The genome-wide false discovery rate cut off of $\alpha=0.05$ was Z2 p-value of 2.27E-5 (Figures 5 and 6) The follow-up studies in the Isle of Wight and replication populations were performed on a limited number of SNPs that were

located exclusively within a single LD block, thus, multiple testing corrections for these studies were not applied.

Expression of ATPAF1 in bronchial biopsies

Bronchial biopsy collection

Endoscopic biopsies were obtained using 3.5 mm cup forceps from the lower lobe of the right lung and stored in the MCI/Meakins-Christie Tissue Bank (McGill University, Montreal, Quebec, CAN): Normals (4), mild asthmatics (6) and severe asthmatics (9). Patients with severe asthma were selected from the severe asthma program at McGill University for which the inclusion and exclusion criteria are published [21, 22]. Of the severe asthmatics, their FEV1 were less than 80% predicted and they were all on inhaled or oral corticosteroids. The average age was 41 and all were atopic and were nonsmokers. The mild asthmatics had an average age of 40, an FEV1 of more than 80% and all of them except for one were atopic, they were on bronchodilators but not on inhaled or oral corticosteroids and did not suffer from any serious respiratory illness. The normal group have an average age of 43 are non-atopic with normal FEV1 and PC20 and none were smokers.

Quantitative PCR

Expression of the asthma-associated gene, *ATPAF1*, along with *S9* ribosomal protein gene were measured by the StepOnePlus PCR system (Applied Biosystems, Foster City, CA, USA) in the

bronchial biopsies. Total RNA was extracted from biopsied tissues using the RNeasy micro kit extraction columns (Qiagen, Valencia, CA, USA). Messenger RNA was reverse transcribed with oligo(dT) (Amersham Pharmacia Biotech, Pittsburgh, PA, USA) and Superscript II (Invitrogen, Carlsbad, CA, USA) in the presence of RNAguard (Amersham Pharmacia Biotech, Pittsburgh, PA, USA). *ATPAF1* levels were normalized using ribosomal protein *S9* gene expression. Primers spanned at least one intron. Primer sequences were: ATPAF1.sense 5'-AAGTGGAGTTCAGTACCTGTCCA-3'; ATPAF1.antisense 5'-GGCTCAGTCCTGTCCAACA-3'; S9.sense 5'-TGCTGACGCTTGATGAGAAG-3'; S9.antisense 5'-CGCAGAGAGAAGTCGATGTG-3'. Expression data were analyzed with Kruskal-Wallis followed by Dunn's multiple comparison tests using GraphPad Prism (v.4.0, GraphPad Software, San Diego, CA, USA).

RESULTS

SNP associations

In the gene discovery stage of the study conducted in the Isle of Wight cohort, the ~100K SNP microarrays yielded a set of 98,921 SNPs of sufficient quality for analysis (Figure 6). The 40 kb sliding window approach used to identify clusters of SNPs with allele frequency differences between asthmatics and controls yielded 60 clusters (Z2 p-value<0.005) throughout the genome. The cluster containing the individual SNP with the highest rank (rs2289447, Z2 p-value=2E-8, ranked 6th among all SNPs on the microarray) was located on chromosome 1p33-p32.31 in the *ATPAF1* and neighbouring *C1ORF223* and *KIAA0494* genes (Figures 6, 7, Table 5). This region had sustained significance (Z2 p-value range=0.0124 to 2.2E-8) across a cluster of 6 SNPs

(rs2289447, rs1150068, rs1048380, rs2275380, rs1150064, and rs1440486) and was therefore selected for further study.

To investigate the genes on chromosome 1p33-p32.31 identified in the pooled samples by the cluster analysis, this region was targeted for focused genotyping in individual samples; the same subjects were used as in the pools. Specifically, SNPs identified in the gene discovery analysis plus additional tagSNPs identified in *ATPAF1, C1ORF223,* and *KIAA0494* and within 10 kb of the 5' and 3' flanking regions of these genes were genotyped in the 277 individual cohort children (Table 5). These genes were found to be within a single LD block (Figure 8). Ten of eleven informative SNPs in this LD block were significantly associated with asthma in the Isle of Wight population (Table 5) and the minor alleles were found to be protective (ORs=0.547 to 0.699; Table 8).

Control data from the Isle of Wight cohort and the consortia populations were combined and compared to the Isle of Wight case data. A comparison of SNPs that were directly genotyped and imputed demonstrates a similar minor allele frequency (Table 8). Results of analysis from the variants that were directly genotyped were found to be similar to those of imputed genotypes (for example, rs2218189 had a p=0.00010 from direct genotyping and p=0.00011 from genotype imputation). The use of consortia controls in our association analysis supported the original associations in the Isle of Wight cohort (nine SNPs) and added support for an additional SNP (rs2275380). It was found that an additive model best supported association in the custom SNP beadarray and the consortia data.

Individuals in the Wessex, U.K. replication population were genotyped for eleven SNPs in the 1p33-p32.31 LD SNPs block. The same LD pattern was identified as in the original association

Table 8. Alleles and frequencies in Isle of Wight, consortia, and replication populations

	CEU Hapmap		Isle of Wight						SLEGEN	WTCCC	Wessex		GALA			Mexico Childhood Asthma		CAMP + CARE	
														Puerto Rican	Mexican				
SNP	Minor Allele	MAF	Minor Allele	Control MAF	Asthma MAF	OR	Lower CI	Upper CI	Control MAF	Control MAF	Minor Allele	MAF	Minor Alelle	MAF	MAF	Minor Allele	MAF	Minor Allele	Case MAF
rs1258000	G	0.27	G	0.30	0.22	0.52	0.33	0.82	0.30	0.30	G	0.29	G	0.37	0.44	G	0.43		
rs2289447	T	0.23	T	0.25	0.17	0.47	0.29	0.78	0.25	0.25	T	0.24							
rs620431	A	0.26	A	0.28	0.18	0.49	0.31	0.78	0.28	0.28	A	0.26	A	0.45	0.48				
rs1150068	C	0.25	C	0.27	0.17	0.45	0.27	0.73	0.27	0.27	C	0.26							
rs654509	G	0.02																A	0.01
rs601060	G	0.24											G	0.16	0.11			G	0.22
rs1048380	T	0.23	T	0.27	0.17	0.44	0.27	0.73	0.27	0.27	T	0.25							
rs2275380	G	0.49	G	0.51	0.42	0.78	0.55	1.11	0.51	0.49	A	0.49	A	0.36	0.41	A	0.43	A	0.43
rs1150064	T	0.25	T	0.27	0.17	0.46	0.28	0.75	0.27	0.27	T	0.26							
rs6665021	G	0.06																G	0.00
rs4660956	T	0.24											T	0.14	0.10			T	0.20
rs1440487	T	0.22	T	0.23	0.24	1.32	0.88	1.98	0.23	0.24	T	0.25	T	0.16	0.11	T	0.08	T	0.22
rs1440486	A	0.23	A	0.27	0.17	0.46	0.28	0.75	0.27	0.27	A	0.25							
rs1074986	G	0.25											A	0.16	0.11			G	0.22
rs2218189	G	0.23	G	0.27	0.17	0.46	0.28	0.75	0.27	0.27	G	0.26				G	0.43		
rs6670495	A	0.26	A	0.25	0.15	0.49	0.29	0.81	0.25	0.25	A	0.24	A	0.46	0.41				
rs6662321	T	0.04																T	0.10

study (Figure 9). Transmission disequilibrium tests in this replication population confirmed the association of rs2275380 with asthma diagnosis (Table 5). Several other SNPs approached significance. In addition, a very high level of LD was found in this region, which was very similar to both the Isle of Wight and the Caucasian Hapmap data (Figure 8A, 9A and B) (which also validates the use of Hapmap in selecting tagging SNPs).

The GALA Puerto Rican and Mexican populations were specifically genotyped for five SNPs in the 1p33-p32.31 region selected from the Isle of Wight study for purposes of replication, in addition several SNPs had been genotyped in the LD block as part of the GALA SNP array study. The Mexico Childhood Asthma Study population had been genotyped for fifteen SNPs in this region, four of which were identical to SNPs genotyped in the Isle of Wight cohort, seven of which were in complete LD (r2=1) and two in high LD (r2=0.87) with SNPs genotyped in the Isle of Wight cohort, and one was a novel SNP. SNPs rs1258000, rs601060, rs4660956, and rs10749863 were associated ($p<0.05$) with asthma in the Puerto Rican population (Table 5). No SNPs reached significance in either Mexican population (Table 5). Minor allele frequencies differed between the Caucasian and Hispanic populations, although they were similar between the Isle of Wight, Wessex, and HapMap CEU populations, as well as between the Puerto Rican and the two Mexican populations (Table 8).

Data from the CAMP and CARE family studies analyzed by PDT analysis revealed asthma associations ($p<0.05$) with four single SNPs (rs654509, rs6665021, rs10749863, rs6662321; $p=0.0046$ to 0.049, Table 5, Table 9).

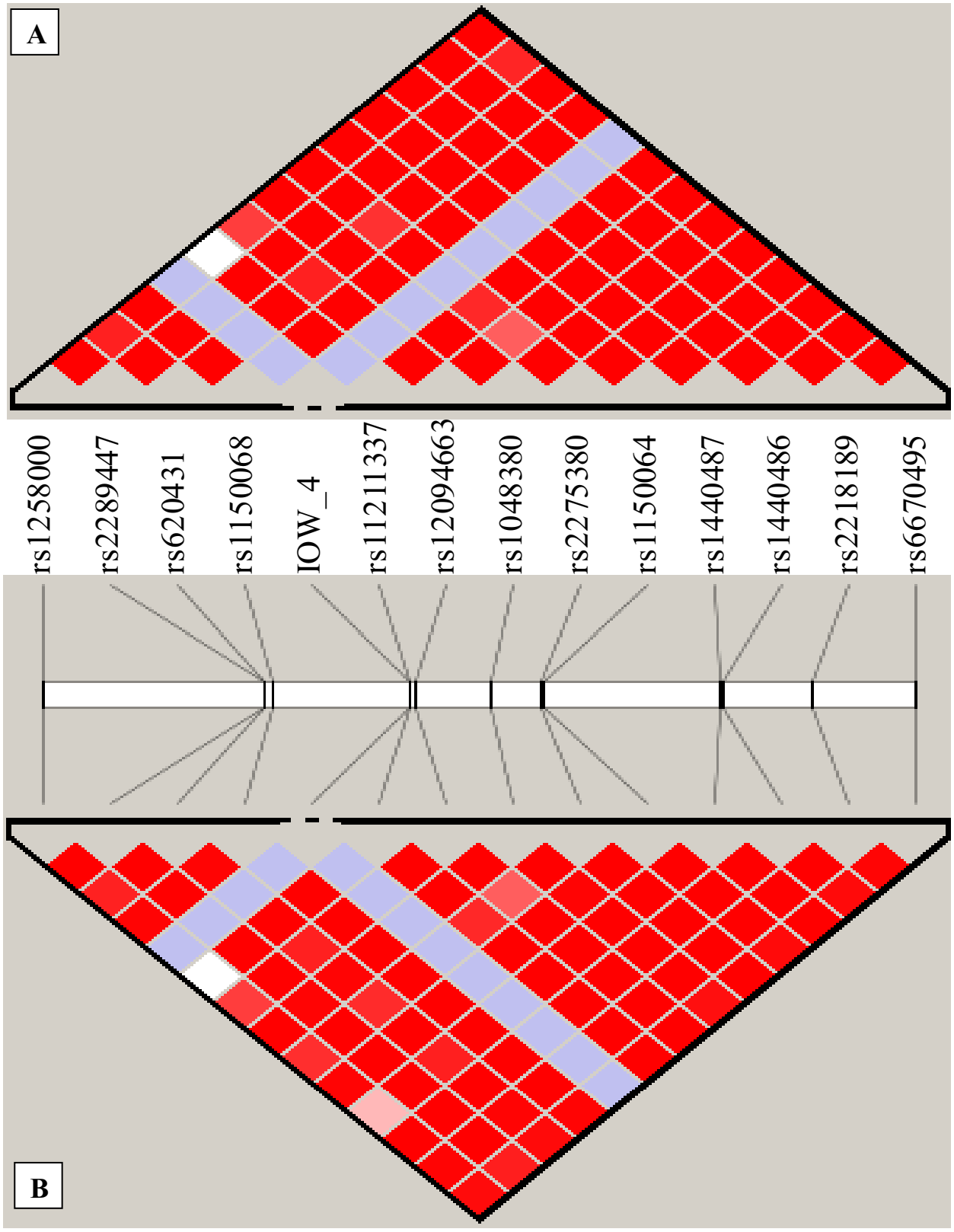

Figure 9: The D'/LOD LD plots from individuals from the Isle of Wight case-control study (A) and Wessex family-based study (B) generated with Haploview [26].

Table 9: Results of sliding-window haplotype association by PDT in CAMP and CARE families

1-marker		2-marker			3-marker			
SNP	p-value	SNP1	SNP2	p-value	SNP1	SNP2	SNP3	p-value
rs7354865	0.9632	rs682000	rs654509	0.0006	rs7354865	rs682000	rs654509	0.0213
rs682000	0.3173	rs720413	rs6662321	0.0222	rs682000	rs654509	rs12094663	0.0249
rs654509	0.0047							
rs12094663	0.9259							
rs6665021	0.0196							
rs10749863	0.0495							
rs720413	0.8999							
rs6662321	0.0460							

Haplotype associations

A haplotype analysis was conducted in the Isle of Wight cohort and other replication populations (except the Mexico Childhood Asthma study). SNPs that were non-informative or outside the LD block were excluded and the remaining eleven SNPs were used to establish haplotypes (Figure 9, Table 7). While common haplotypes were found in the Isle of Wight and Wessex populations, there were some differences in haplotype structure between Hispanic and Caucasian populations. Haplotypes with population frequency greater than 5% (haplotypes I, II, III in Caucasians) were examined. In the Isle of Wight cohort haplotypes I and III were found to confer asthma risk (Chi-square p=0.035) and protection (Chi-square p=0.0048), respectively (Table 7). Odds ratios for the haplotypes indicated more than two-fold decreased risk of asthma associated with haplotype III (OR 0.45, 95% CI 0.26–0.78, p=0.0042) as compared to haplotype I as a reference. Transmission disequilibrium tests in the Wessex population confirmed the association of haplotype I with increased risk of asthma diagnosis (Chi-square p=0.0156). In the Puerto Rican families a rare haplotype (0.02 frequency) was found to be associated with risk for asthma (Chi-square p=0.038). This rare haplotype was present in the Isle of Wight and Wessex populations but did not confer asthma risk. Four two- and three-marker haplotype associations with asthma were seen in CAMP and CARE families (p=0.0006 to 0.024; Table 9).

ATPAF1 relevance to asthma

The expression of ATPAF1 in bronchial biopsy samples obtained from subjects with severe asthma was markedly (50-fold) elevated as compared to controls ($p<0.01$, Figure 10).

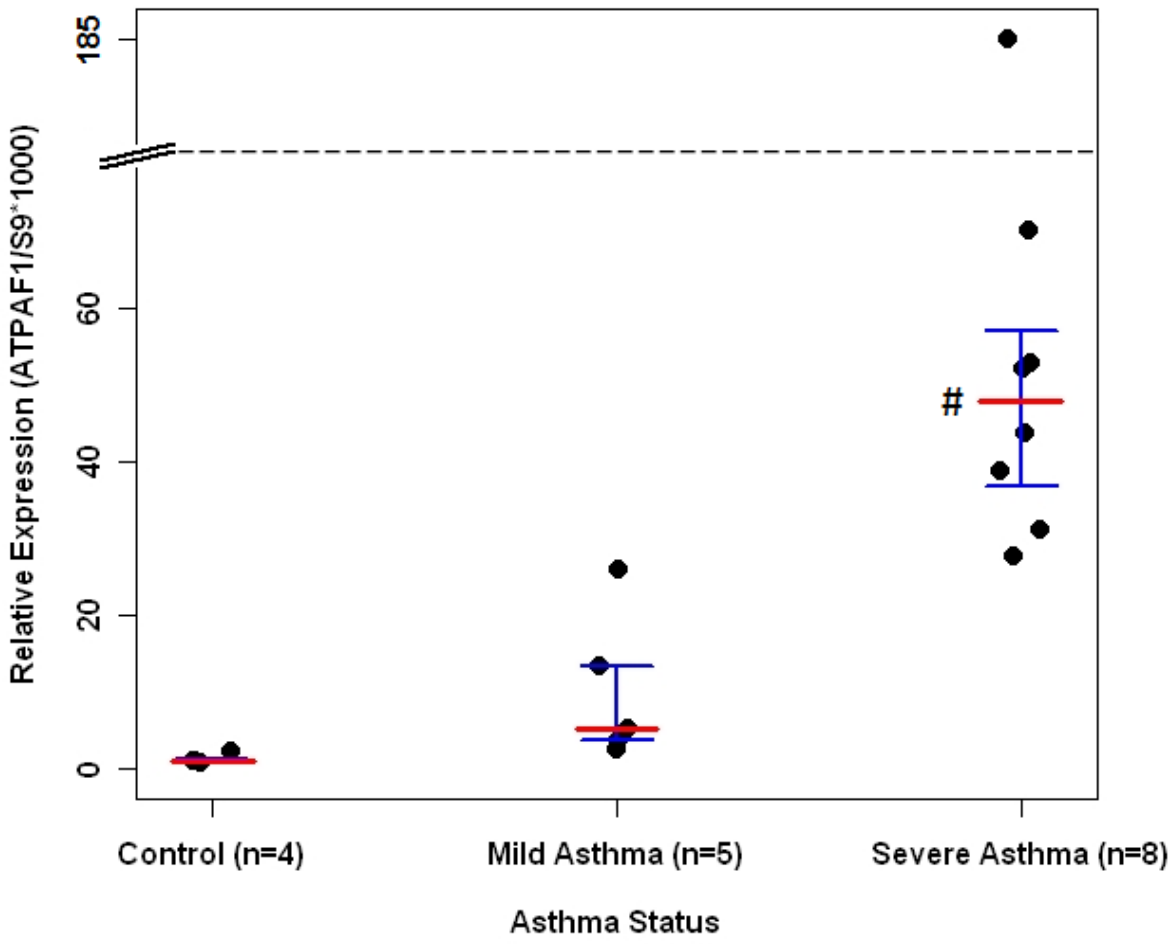

Figure 10: Gene expression study of *ATPAF1* from biopsied bronchial tissue [21, 22] A # indicates statistical significance for severe asthma versus no asthma ($p < 0.01$). Median expression value indicated by red horizontal line; and first and third quantiles indicated by brackets.

DISCUSSION

A genomic region on chromosome 1q33-q32.31 met genome-wide significance for asthma in the Isle of Wight cohort. Subsequent detailed examination using a combination of targeted genotyping and imputation in the primary population and consortia controls confirmed the association with an LD block containing *ATPAF1, C1ORF223,* and *KIAA0494* genes. Replication studies pursued in five independent populations revealed two instances of strict replication with specific SNPs examined in the Isle of Wight children. Further gene-level association with additional SNPs genotyped using other platforms was found in three of five replication populations. Thus, while not all significant findings in the primary population were replicated, major trends in association were identified across the LD block in all but the two Mexican populations. Data demonstrating differential upregulation of *ATPAF1* expression in asthmatics as compared to control subjects lends further support for a role for this gene, which is novel to asthma.

Past evidence of linkage of Chromosome 1p to asthma and related phenotypes

While no previous asthma association study has been as finely mapped on 1p3 as the current study, our work supports previous studies that identified asthma linkages to the 1p31 to 1p36 region [34-42]. Indeed, the chromosome region 1p has been consistently implicated in genome screens for asthma in populations of different races ethnicities (Table 10) [34-36, 38]. The Collaborative Study on the Genetics of Asthma (CSGA) suggested chromosome region 1p3 as a locus of interest for asthma susceptibility in Caucasian (1p31-p32) and Hispanic (1p33-p32) populations [42]. Subsequently, in the French Epidemiological Study on the Genetics and

Table 10: Previous reports of linkage of asthma and asthma related phenotypes to Chromosome 1p

Study/Population	Peak Linkage	Peak Location (cM)	Peak LOD Score	P-value	Phenotype	Reference
French EGEA (Epidemiological Study on the genetics and Environment of Asthma)						
Caucasian	1p31	94	2.5	0.0006	asthma	[38]
	1p33	75.7	1.4	0.005	asthma	[35]
	1p36	4.2	1.52	0.004	$\%FEV_1$	[36]
CSGA (Collaborative Study on the Genetics of Asthma						
Caucasian	1p31-p32	97	0.54		asthma	[37]
	1p31-p32	97	1.29	0.034	tobacco smoke exposed asthma	[37]
	1p33	76	0.6		asthma	[41]
Hispanic	1p33	76	2.92	0.0002	asthma	[41]
	1p33	76	1.4		asthma	[40]
	1p33	76	1.29	0.007	asthma	[34]
German	1p21	146.7		0.0135	IgE	[43]
	1p21	146.7		0.0156	RAST	[43]
	1p33	118.1		0.0045	IgE	[43]
	1p36	33.2		0.0094	absolute number eosinophils	[43]
	1pter	237.2		0.0098	IgE	[43]
	1pter	151.2		0.0023	peak exp flow restriction	[43]
Danish	1p34	50	1.22		RAST	[39]
	1p36	6	2.02		asthma	[39]

Environment of Asthma (EGEA) study, chromosome 1p31 was found to be associated with asthma [38]. Several asthma-related phenotypes (total serum IgE, absolute number of eosinophils) were linked to the 1p33-p36 region in a German study [43] and a Danish study [39]. Thus, our finding of association of asthma with a 115 kb LD block at 1p33-p32.31 is supportive of these earlier studies.

Background information on the genes ATPAF1, C1ORF223, and KIAA0494

The childhood asthma susceptibility locus that was identified on human chromosome 1p33 contains three genes: *ATPAF1, C1ORF223,* and *KIAA0494* (Figure 7 and 10) [44] The association was contained within an ≈120kb (chr1:46,862,453-46,979,069 (UCSC Genome Browser, Feb 2009 build[45]) linkage disequilibrium (LD) block (CEU Hapmap data[24]). These three genes together occupy approximately 92 kb of the LD block with *ATPAF1* spanning 35 kb (46,870,998-46,906,686bp) in nine exons, *C1ORF223* 2kb (46,910,087-46,911,843) in two exons and *KIAA0494* 44 kb (46,922,422-46,957,323) in eleven exons[45]. *C1ORF223* is between the two larger genes and is in opposite transcription orientation. Due to the overlap of coding and regulatory sequence among the three genes in the associated LD block, additional studies on each gene are warranted. So while we cannot exclude *C1ORF223* or *KIAA0494* from having a role in asthma, we chose to prioritize further study of *ATPAF1* because of its potential role in asthma based on the evidence of differential expression in a bronchial biopsy.

ATPAF1

The gene *ATPAF1* (ATP synthase mitochondrial F1 complex assembly factor 1, also known as *ATP11*, OMIM ID:608917)[46] is a nuclear gene encoding ATPAF1 (ATP11p). A soluble protein, ATPAF1, is exported to the mitochondria where it is required for proper assembly of the F1 subunit of the F1F0ATP-synthase. ATP synthase is a multi-subunit enzyme consisting of a F1 and F0 subunits which catalyzes the synthesis of ATP from ADP and P. ATPAF1 functions as a chaperone protein by binding to unassembled β subunits of the F1 subunit of the ATP synthase and preventing the F1 α and β subunits from aggregating in the matrix of the mitochondria [47, 48]. The F1 subunit is critical for ATP generation as it contains nucleotide binding (ADP/ATP) and catalytic sites. As ATPAF1 is required for correct assembly of this complex,[48] it is not surprisingly preserved in all eukaryotic lineages capable of ATP synthesis via oxidative phosphorylation [48, 49].

ATPAF1 is moderately expressed in all tissues (Figure 11)[45, 50]. *In situ* hybridization in mouse embryos demonstrates global expression during development [51]. Although it seems intuitive that due to ATPAF1's necessary function in assembly of the ATP synthase, that *ATPAF1* would have increased expression in cells with a high energy demand (heart, muscle, brain, brown adipose tissue), qRT-PCR analysis shows that expression is consistent across all examined tissues (airway tissue was not examined).

A female infant with a missense mutation in *ATPAF2,* a gene with similar function to *ATPAF1*, was described in a case report. She presented with several dysmorphic features, severe brain atrophy, an enlarged liver, hypertonia, limb flexion contractures, and inability to suck with

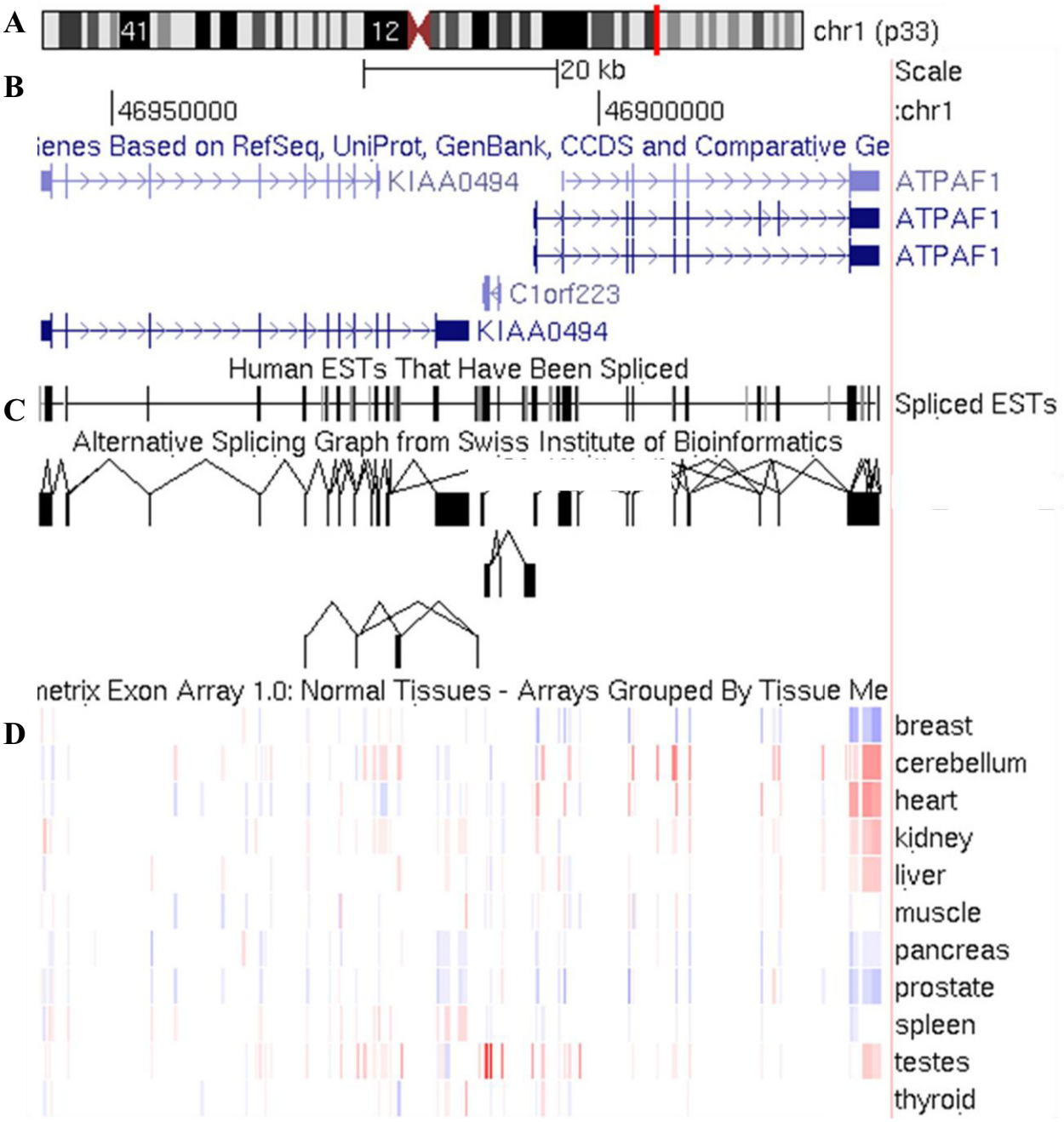

Figure 11: Location schematic of *ATPAF1*, *C1ORF223*, *KIAA0494* on Chr 1p33 (A) from UCSC Genome browser[45]. The gene track illustrating their location (B) and the complex splicing pattern that exists between *ATPAF1* and *KIAA0494* is illustrated (C). The relative expression of exons by tissue is demonstrated (D). In D, red indicates increased expression while blue color reflects a decreased expression.

eventual death at 14 months. Pathology revealed she had marked decrease in ATP synthase activity in liver and muscle [52].

KIAA0494

KIAA0494 encodes an uncharacterized calcium-binding protein that is expressed in all tissues (Figure 11), with suggestion of a high level of expression in lymphocytes [45]. It has a conserved domain similar to a family of ATPases that aids in structural maintenance of chromosomes (SMC) during cell division and chromosome partitioning. KIAA0494 is predicted to contain a transmembrane domain and two calcium-binding helix-loop-helix motifs ("EF-hands") [44].

C1ORF223

The gene *C1ORF223* encodes an uncharacterized integral protein. It has a very high level of expression found in the testes but is also expressed at much lower levels elsewhere (Figure 11)[45]. *C1ORF223* is in opposite transcription orientation to *ATPAF1* and *KIAA0494*. This results in 294 bp of the transcript being antisense to an alternative ATPAF1-KIAA0494 fused transcript, thus suggesting the possibility of regulation of alternative expression [53].

Alternative splicing results in many transcript isoforms

Alternative transcript isoforms of both *ATPAF1* and *KIAA0494* have been widely reported. This locus is considered "complex" with over 19 alternatively spliced variants due to 7 possible alternative promoters and 10 validated alternative polyadenylation sites (Figure 11). In multiple instances, an ATPAF1-KIAA0494 fused transcript has been reported in lung and B-cell germinal centers. This and other transcript isoforms have been reported to result in proteins which localize to the cytoplasm or plasma membrane while still maintaining the majority of the ATPAF1 functional domain (yet remain uncharacterized thus far) [53, 54]. Of unknown consequence, *KIAA0494* has been reported to contain an unusual TG 3' splice site in intron 1 [55]. Transcription of *C1ORF223* alone produces 3 alternatively spliced transcripts [53]. In relevance to asthma, asthma candidate genes with unique transcript isoforms specific to airway tissues have been reported in the past [56].

Functional significance of associated SNPs

Functional significance is predicted for several of the SNPs associated with asthma in this study [24, 45]. Specifically, sequence encompassing rs1258000 has resemblance to alignment patterns typical of regulatory elements per analysis by ESPERR Regulatory Potential (7xReg) software implemented in the UCSC Genome Browser [45, 57]. Similarly, rs620431 has high regulatory potential by 7xReg and also lies 60 bp downstream of the exon 6/intron 6 boundary making it a potential splicing modulating element for the alternatively spliced exon 7. In addition, rs6665021 significant in the CAMP and CARE datasets is coding in *KIAA0494*. The most direct evidence of functional relevance of the *ATPAF1* gene in asthma comes from its differential expression in

bronchial tissue between asthmatics and controls (Figure 10). *ATPAF1* was highly expressed in bronchial biopsies from those with severe asthma. Not only does this suggest a mechanism by which the gene may modify asthma risk, but it is also consistent with the findings of Chen et al [58] in which they report that genes that are differentially expressed have a greater likelihood of containing variants that cause disease.

Study Design

We utilized an initial strategy involving pooled DNA samples because it was practical, efficient, and cost effective [59, 60]. We were able to effectively rank the SNPs, and thereby genome regions, that were most closely associated with asthma. Association of the top-ranked region in chromosome 1p33-p32.31 was then confirmed by targeted genotyping of DNA samples from individual children. Preliminary evidence is supportive of the clinical relevance of the *ATPAF1* gene in asthma, by virtue of its differential expression between asthmatics and controls. This efficient approach could be applied to study genetic association for a number of diseases to increase cost effectiveness and may be particularly useful for diseases with multiple genes of small effect.

We opted to use a nested case-control design, augmented with consortia controls. The use of consortia controls in smaller scale case-control studies has been utilized before and has the benefit of mitigating the concern of lack of statistical power while remaining efficient [10, 17]. Although we were technically unable to genetically match the Isle of Wight controls with consortia controls, both consortia were Caucasian and had nearly identical minor allele frequencies for SNPs in the region examined (Table 8).

The risk of reporting statistical significance merely by chance is a major concern in most association studies due to the increased number of tests. However, this is unlikely in the present study as several SNPs retained significance at the genome-wide level after correction for multiple testing, the outcomes were consistent across individual SNP and haplotype analyses, and the common haplotypes showed association with asthma in the replication population of the same race. In addition, data showing functional relevance of *ATPAF1* further reinforce the validity of our findings.

For replication, we chose populations that had previously been studied for asthma genetics. The Wessex children's age, ethnicity, and socioeconomic background were similar to the Isle of Wight population and they were recruited from the same geographical area. However, the study design collected affected sib pairs and parents, therefore, TDT analysis was performed. In this population significant association with asthma was found with rs1440486 and several other SNPs approached significance. In addition, a common eleven SNP haplotype conferred risk in both Caucasian populations examined.

In the Puerto Rican population significant association was found with rs1258000, rs601060, rs4660956, and rs10749863. In contrast, no SNPs were significant in the Mexican population from the GALA study or the Mexico Childhood Asthma Study. Thus, this region may not convey asthma susceptibility to children of Mexican descent. Differences in strength of association of individual susceptibility genes between Puerto Rican and Mexican populations have previously been noted [61]. In addition, large differences in asthma prevalence and mortality rates between Hispanics of Puerto Rican, Cuban, and Mexican heritage have been reported [3, 4]. Variability in replication between cohorts has been a feature of studies of asthma genetics [7, 62, 63]. Other possible explanations for lack of replication in the Mexican ancestry

populations include type II error due to lack of power and differences in environmental exposures between cohorts.

Further support for our findings come from the publicly available data from the CAMP and CARE datasets within the SHARP studies. The analysis found association with three SNPs not tested in the other populations (rs654509, rs6665021, rs6662321) and replicated results in one SNP (rs10749863) with several other SNPs trending towards association. The association with haplotypes in these populations also support the replication of the 1p33 as an asthma susceptibility region.

The replication study results support both strict and loose replication of the chromosome 1p33 region as an asthma susceptibility region. In addition, the statistical genetic model found to be the strongest in each population varied between allelic and genotypic dominant and additive (Table 5). These results raise the likelihood that there might be more than one functional and causative variant [6].

In conclusion, our sequential strategy, as well as the use of well-phenotyped populations, led to identification of an association between *ATPAF1* region variants and asthma. Studies to further understand the mechanistic role of this gene in asthma are being pursued.

SUPPLEMENTARY METHODS

Genotyping/Allelotyping methods for the Isle of Wight birth cohort. Genotyping methods used in other populations is described in Table 5.

Pooled affymetrix GeneChip Mapping Array genotyping. DNA samples assessed by NanoDrop spectrophotometry (Wilmington, DE, USA) to have A260/A280 range of 1.65-2.0 and A260/A230 range of 1.0-2.2 qualified for inclusion in a pool. DNA samples were separated on 0.8% agarose gels to confirm lack of degradation or RNA contamination. Equimolar amounts of DNA from individuals were combined for a total of 250 ng/pool. DNA pools were digested with XbaI or HindIII enzyme, adapter ligated, and PCR amplified. Then samples were separated on 4% agarose gels to ensure DNA fragmentation in the 100-300 bp range. PCR yields (>1,200 ng/μl accepted) were compared between microarray chips to ensure uniformity and PCR products were separated on 2% agarose gels to ensure the proper range of product was amplified. GeneChip Genotyping software (v.4.0, Affymetrix, Inc.)[64] was used for relative quality control assessment, detection rates (>98.5%), and allele distributions (<10% difference between pools). Hybridization intensity comparisons of the case and control pools were used to identify significant allele frequency differences for each SNP (Table 6).

Illumina SNP beadarray genotyping. Individual sample genotyping by custom Illumina GoldenGate assays included a ThermoElectron KingFisher96 automated magnetic bead wash. An autoclustering algorithm was used on all SNPs. Clusters of SNPs were manually inspected when they had low call rate (<98.5%), low clustering score (<0.6), or significant departure from Hardy-Weinberg equilibrium (P<0.05). A total of 38 of the 96

clusters were edited to adjust the autoclustering algorithm. The genotype success rate for each SNP was ≥98.9% and the overall call rate was 99.7%. Five samples failed completely and were not reported here. Two samples assayed in duplicate as technical replicates had complete concordance. Eight samples from a different study included here had a 99.8% call rate.

ACKNOWLEDGEMENTS

The authors thank Dennis Shubitowski and David Hutchings for technical assistance and Beth Cobb for administrative support. Mrs. Sharon Matthews helped with Isle of Wight phenotype data collection at age 10 years. Andrea Mogas, Susan Foley and Alejandro Vazquez extracted RNA and made cDNA from the biopsy tissues. The authors acknowledge the NIH GWAS Data Repository (dbGAP) and the investigators involved in the National Heart, Lung, and Blood Institute SNP Health Association Resource (SHARe) Asthma Resource Project (SHARP) for use of their genotype and phenotype data. The authors thank Hao Wu and Grace Chiu for this analysis of the Mexico City Childhood Asthma Study data.

Funding support: This study was funded by the National Institutes of Health, grants R01 AI061471, R01 HL67736, P01 HL076383, T32 GM063483, Asthma UK (364) and the Asthma, Allergy and Inflammation Research Charity. The Wessex Family Cohort was originally recruited in collaboration with Genome Therapeutics Corporation and Schering-Plough. Richard and Edith Strauss Foundation of Canada and Dr. Ron Olivenstein supported the severe asthma program and collection of bronchial biopsies. The GALA studies were supported by HL078885, HL088133, AI077439, ES015794, Robert Wood Johnson Foundation Amos Medical Faculty Development Program, Flight Attendant Medical Research Institute (FAMRI). The Mexico Childhood Asthma Study was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (Z01 ES49019). Subject enrollment was also supported in part by the National Council of Science and Technology (grant 26206-M), Mexico. Dr. Romieu was supported in part by the National Center for Environmental Health at the Centers for Disease Control. The CAMP study was supported by contracts with the National Heart, Lung, and Blood Institute (NO1-HR-16044, NO1-HR-16045, NO1-HR-16046, NO1-HR-

16047, NO1-HR-16048, NO1-HR-16049, NO1-HR-16050, NO1-HR-16051, and NO1-HR-16052) and by General Clinical Research Center grants from the National Center for Research Resources (M01RR00051, M01RR0099718-24, M01RR02719-14, and RR00036). The CARE study was supported by grants (HL071742-01, HL004519-04, 5U10HL064287, 5U10HL064288, 5U10HL064295, 5U10HL064307, 5U10HL064305, and 5U10HL064313) from the National Heart, Lung, and Blood Institute. This study was carried out in part in the General Clinical Research Centers at Washington University School of Medicine (M01 RR00036) sponsored by the National Institutes of Health and the National Jewish Medical and Research Center (M01 RR00051).

REFERENCES

REFERENCES

1. Eder, W., M.J. Ege, and E. von Mutius, *The asthma epidemic.* N Engl J Med, 2006. **355**(21): p. 2226-35.

2. Pearce, N., et al., *Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC).* Thorax, 2007. **62**(9): p. 757-765.

3. Carter-Pokras, O.D. and P.J. Gergen, *Reported asthma among Puerto Rican, Mexican-American, and Cuban children, 1982 through 1984.* Am J Public Health, 1993. **83**(4): p. 580-582.

4. Homa, D.M., D.M. Mannino, and M. Lara, *Asthma mortality in U.S. Hispanics of Mexican, Puerto Rican, and Cuban heritage, 1990-1995.* Am J Respir Crit Care Med, 2000. **161**(2 Pt 1): p. 504-9.

5. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-U5.

6. Vercelli, D., *Discovering susceptibility genes for asthma and allergy.* Nature Reviews Immunology, 2008. **8**(3): p. 169-182.

7. Holloway, J.W. and G.H. Koppelman, *Identifying novel genes contributing to asthma pathogenesis.* Current Opinion in Allergy and Clinical Immunology, 2007. **7**(1): p. 69-74.

8. Kurukulaaratchy, R.J., et al., *Characterization of wheezing phenotypes in the first 10 years of life.* Clinical and Experimental Allergy, 2003. **33**(5): p. 573-578.

9. *The database of Genotypes and Phenotypes (dbGaP).* Available from: http://www.ncbi.nlm.nih.gov/gap.

10. Harley, J.B., et al., *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.* Nat Genet, 2008. **40**(2): p. 204-10.

11. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

12. *Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee.* Lancet, 1998. **351**(9111): p. 1225-32.

13. *Long-term effects of budesonide or nedocromil in children with asthma. The Childhood Asthma Management Program Research Group.* N Engl J Med, 2000. **343**(15): p. 1054-63.

14. Burchard, E.G., et al., *Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma.* Am J Respir Crit Care Med, 2004. **169**(3): p. 386-92.

15. David, G.L., et al., *Nicotinamide adenine dinucleotide (phosphate) reduced:quinone oxidoreductase and glutathione S-transferase M1 polymorphisms and childhood asthma.* Am J Respir Crit Care Med, 2003. **168**(10): p. 1199-204.

16. Hancock, D.B., et al., *Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children.* PLoS Genet, 2009. **5**(8): p. e1000623.

17. Himes, B.E., et al., *Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene.* Am J Hum Genet, 2009. **84**(5): p. 581-93.

18. Van Eerdewegh, P., et al., *Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness.* Nature, 2002. **418**(6896): p. 426-430.

19. Guilbert, T.W., et al., *The Prevention of Early Asthma in Kids study: design, rationale and methods for the Childhood Asthma Research and Education network.* Control Clin Trials, 2004. **25**(3): p. 286-310.

20. Guilbert, T.W., et al., *Long-term inhaled corticosteroids in preschool children at high risk for asthma.* N Engl J Med, 2006. **354**(19): p. 1985-97.

21. Foley, S.C., et al., *Increased expression of ADAM33 and ADAM8 with disease progression in asthma.* Journal of Allergy and Clinical Immunology, 2007. **119**(4): p. 863-871.

22. Shannon, J., et al., *Differences in airway cytokine profile in severe asthma compared to moderate asthma.* Chest, 2007: p. chest.07-1881.

23. Yang, H.C., et al., *New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification.* Genetics, 2005. **169**(1): p. 399-410.

24. Gibbs, R.A., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.

25. de Bakker, P.I.W., et al., *Efficiency and power in genetic association studies.* Nature Genetics, 2005. **37**(11): p. 1217-1223.

26. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-265.

27. Gonzalez, J.R., et al., *SNPassoc: an R package to perform whole genome association studies.* Bioinformatics, 2007. **23**(5): p. 644-645.

28. R Development Core Team, *R: A Language and Environment for Statistical Computing*2007, Vienna, Austria.

29. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

30. Lake, S.L., et al., *Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous.* Human Heredity, 2003. **55**(1): p. 56-65.

31. Scheet, P. and M. Stephens, *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.* Am J Hum Genet, 2006. **78**(4): p. 629-44.

32. Laird, N.M., S. Horvath, and X. Xu, *Implementing a unified approach to family-based tests of association.* Genet Epidemiol, 2000. **19 Suppl 1**: p. S36-42.

33. Dudbridge, F., *Pedigree disequilibrium tests for multilocus haplotypes.* Genet Epidemiol, 2003. **25**(2): p. 115-21.

34. Blumenthal, M.N., et al., *A genome-wide search for allergic response (atopy) genes in three ethnic groups: Collaborative Study on the Genetics of Asthma.* Human Genetics, 2004. **114**(2): p. 157-164.

35. Bouzigon, E., et al., *Clustering patterns of LOD scores for asthma-related phenotypes revealed by a genome-wide screen in 295 French EGEA families.* Human Molecular Genetics, 2004. **13**(24): p. 3103-3113.

36. Bouzigon, E., et al., *Scores of asthma and asthma severity reveal new regions of linkage in EGEA study families.* European Respiratory Journal, 2007. **30**(2): p. 253-259.

37. Colilla, S., et al., *Evidence for gene-environment interactions in a linkage study of asthma and smoking exposure.* Journal of Allergy and Clinical Immunology, 2003. **111**(4): p. 840-846.

38. Dizier, M.H., et al., *Genome screen for asthma and related phenotypes in the French EGEA study.* American Journal of Respiratory and Critical Care Medicine, 2000. **162**(5): p. 1812-1818.

39. Haagerup, A., et al., *Asthma and atopy - a total genome scan for susceptibility genes.* Allergy, 2002. **57**(8): p. 680-686.

40. Mathias, R.A., et al., *Genome-wide linkage analyses of total serum IgE using variance components analysis in asthmatic families.* Genetic Epidemiology, 2001. **20**(3): p. 340-355.

41. Xu, J.F., et al., *Genomewide screen and identification of gene-gene interactions for asthma-susceptibility loci in three US populations: Collaborative study on the genetics of asthma.* American Journal of Human Genetics, 2001. **68**(6): p. 1437-1446.

42. *A genome-wide search for asthma susceptibility loci in ethnically diverse populations.* Nat Genet, 1997. **15**(4): p. 389-392.

43. Wjst, M., et al., *A genome-wide search for linkage to asthma.* Genomics, 1999. **58**(1): p. 1-8.

44. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Research, 2007. **35**: p. D26-D31.

45. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

46. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.

47. Sheluho, D. and S.H. Ackerman, *An accessible hydrophobic surface is a key element of the molecular chaperone action of Atp11p.* Journal of Biological Chemistry, 2001. **276**(43): p. 39945-39949.

48. Wang, Z.G., P.S. White, and S.H. Ackerman, *Atp11p and Atp12p are assembly factors for the F-1-ATPase in human mitochondria.* Journal of Biological Chemistry, 2001. **276**(33): p. 30773-30778.

49. Pickova, A., M. Potocky, and J. Houstek, *Assembly factors of F1F0-ATP synthase across genomes.* Proteins-Structure Function and Bioinformatics, 2005. **59**(3): p. 393-402.

50. Pickova, A., et al., *Differential expression of ATPAF1 and ATPAF2 genes encoding F(1)-ATPase assembly proteins in mouse tissues.* FEBS Lett, 2003. **551**(1-3): p. 42-6.

51. Visel, A., C. Thaller, and G. Eichele, *GenePaint.org: an atlas of gene expression patterns in the mouse embryo.* Nucleic Acids Res, 2004. **32**(Database issue): p. D552-6.

52. De Meirleir, L., et al., *Respiratory chain complex V deficiency due to a mutation in the assembly gene ATP12.* Journal of Medical Genetics, 2004. **41**(2): p. 120-124.

53. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.

54. Sprenger, J., et al., *LOCATE: a mammalian protein subcellular localization database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D230-3.

55. Szafranski, K., et al., *Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns.* Genome Biol, 2007. **8**(8): p. R154.

56. Laitinen, T., et al., *Characterization of a common susceptibility locus for asthma-related traits.* Science, 2004. **304**(5668): p. 300-4.

57. King, D.C., et al., *Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences.* Genome Res, 2005. **15**(8): p. 1051-60.

58. Chen, R., et al., *FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease.* Genome Biol, 2008. **9**(12): p. R170.

59. Butcher, L.M., et al., *SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children.* Human Molecular Genetics, 2005. **14**(10): p. 1315-1325.

60. Docherty, S.J., et al., *Applicability of DNA pools on 500 KSNP microarrays for cost-effective initial screens in genomewide association studies.* Bmc Genomics, 2007. **8**: p. 7.

61. Galanter, J., et al., *ORMDL3 gene is associated with asthma in three ethnically diverse populations.* Am J Respir Crit Care Med, 2008. **177**(11): p. 1194-200.

62. Holloway, J.W., I.A. Yang, and S.T. Holgate, *Genetics of allergic disease.* J Allergy Clin Immunol. **125**(2 Suppl 2): p. S81-94.

63. Rogers, A.J., et al., *Assessing the reproducibility of asthma candidate gene associations, using genome-wide data.* Am J Respir Crit Care Med, 2009. **179**(12): p. 1084-90.

64. Affymetrix. *GeneChip® Genotyping Analysis Software (GTYPE)*. 2007; Available from: http://www.affymetrix.com/products/software/specific/gtype.affx.

CHAPTER 3

Determination of variants in *ATPAF1* that modulate susceptibility to asthma

INTRODUCTION

ATPAF1 (ATP synthase mitochondrial F_1 complex assembly factor 1) was recently identified as a childhood asthma susceptibility gene [1]. This gene encodes an assembly factor for the F_1 complex of mitochondrial ATP synthase. Its function is to bind the β subunit which is necessary to prevent improper polymerization of β subunits into nonproductive homooligomers [2]. In addition to asthma susceptibility gene support from association studies, gene expression data from bronchial biopsies show a 50-fold increase in asthmatics as compared to nonasthmatic/control individuals [1].

Despite the evidence for a role for *ATPAF1* in asthma, the mechanism of its involvement in disease pathogenesis is unknown. As is common in many other genes implicated in asthma or other complex genetic diseases, the causative DNA variant(s) has not been identified. For example, the role of *ORMDL3*, the first asthma gene identified through a genome wide association study (GWAS), in asthma pathogenesis has yet to be determined. The ORMDL3 protein has been localized in the endoplasmic reticulum (ER) of cells where it has a hypothesized function of managing calcium ion concentration [3]. Yet there is still debate about 1) whether the ER is the correct location and/or mechanism of ORMDL3's function in relevance to asthma pathogenesis, 2) what are the causative variant(s) and how are they involved in asthma pathogenesis [3]. To assist in modern genetics struggle to identify and validate causative variants in regions targeted thru GWAS studies, *in silico* tools have been developed to explore the function(s) of DNA variants [4-7]. These databases, though extremely helpful, are reliant on programmers' continuous updating to add new sources of data. The *in silico* approach presented here to screen variants for function utilized these automated tools and augmented these sources with data from various tracks of the UCSC Genome browser [8]. This method benefits from the

combining of complementary data, for example, sequence based transcription factor binding site prediction with Chromatin immunoprecipitation (ChIP) data for transcription factors.

In order to discern the function of variants in and around the *ATPAF1* gene and their potential as causative variants for asthma, a detailed and systematic approach was required. Specifically, we utilized a step-wise approach to: 1) identify variants and/or gene regions with a high probability of having functional relevance, 2) resequence targeted regions to discover new variants and validate those already cataloged in our population, and 3) prioritize the variants for functional studies.

METHODS

Summary of experimental design

Variants within or near the *ATPAF1* gene (≈ 7,500 bp in 5' direction and ≈ 500 bp 5' direction [limit until adjacent gene]) were identified in public databases (dbSNP [9], Ensembl [10], and UCSC Genome Browser [8]) and screened for either known or predicted function based on 1) annotated function of variants in public databases (dbSNP [9], Ensembl [10], and UCSC Genome Browser [8]), and 2) inferred function of variants based on the context of their location in experimentally derived or *in silico* predicted functional DNA sequences or proteins. The variants were prioritized based on the evidence and those with the highest rankings were resequenced to validate their existence in asthmatic and nonasthmatic individuals. Regions with a predicted high-level of functional DNA sequences (exons, promoter) were also identified and resequenced for purposes of variant discovery. Novel variants identified by resequencing were

screened for predicted function and a final list of validated variants was constructed. Clusters of two or more variants with predicted function and/or in regions with predicted function were prioritized for future studies.

In silico analysis of ATPAF1 variants

Categories and sources of functional predictive evidence

Data from UCSC genome browser [8] (available as "tracks"), and the SNPlogic[5], and SNP-Nexus [7] programs were mined in order to specify ATPAF1 functional elements. The locations of these functional elements were intercepted with validated and unvalidated variants from the NCBI dbSNP [9] and Ensembl [10] databases, and UCSC Genome Browser [8] using Galaxy[11]. The data were divided into categories by function (Table 11) using the algorithm described in Figure 12. Locations of repetitive sequences were also collected but not used for the functional screen as the repetitive sequence does not preclude all functionality (for example, palindrome sequences) [12, 13]. Variant association data were not used for functional screening of variants but were used in selection of variants examined in functional studies.

Scoring scheme to predict variant function

The functional potential of each variant was inferred through the quantity and quality of supportive evidence from the above data sources. Variants were classified into one or more of the possible categories: 1) nonsynonymous-coding variants, 2) regulatory variants, and/or 3)

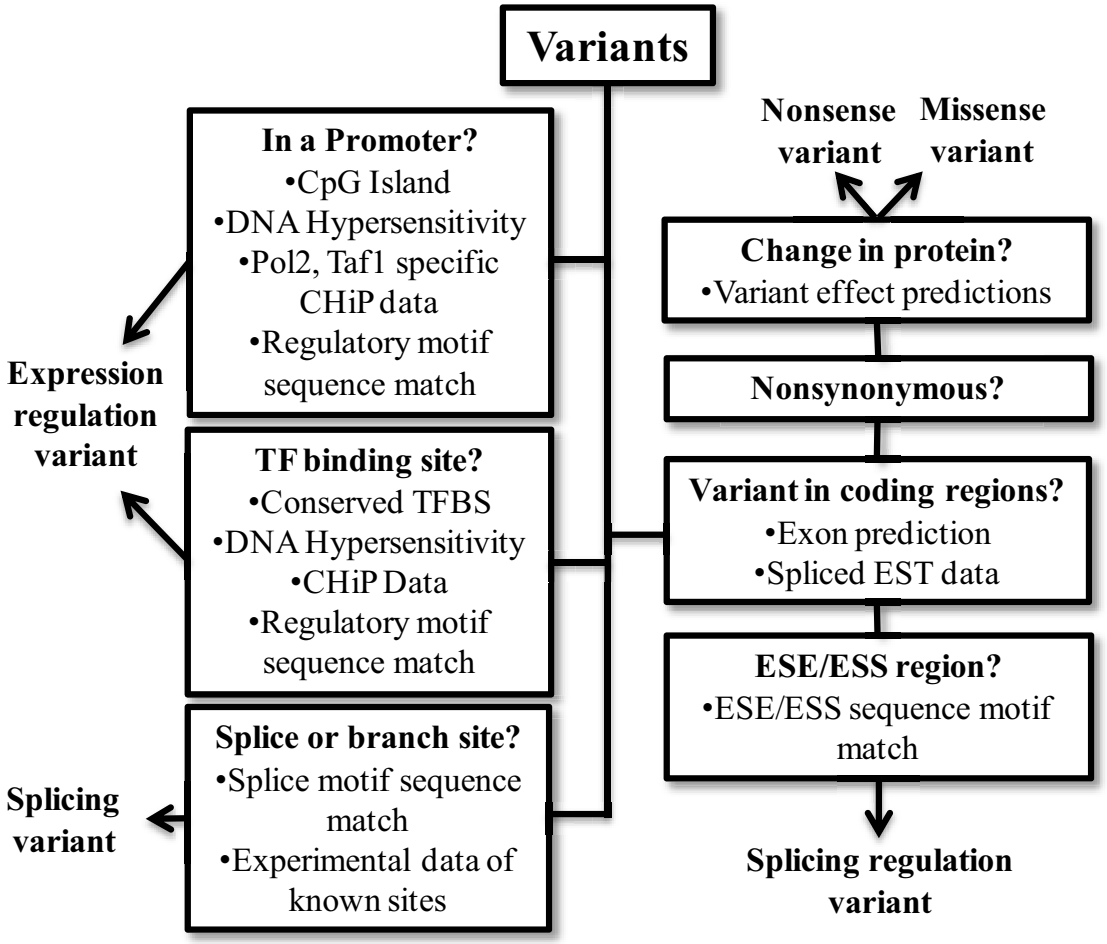

Figure 12: Decision tree for *in silico* variant function prediction. See Tables 11 and 12 for references and scoring scheme.

Table 11: Resources used to define functional categories.

Functional category	Resources	Comments
Coding regions	RefSeq EST data[14], Aceview[15], ENCODE Gencode[16], Affymetrix exon array probe locations, SIB Alternative Transcripts[17]	Nonsynonymous variants were further examined for impact on protein using POLYPHEN[18], SIFT[19], and SNPs3D[20]
Regulatory Regions	**Promoter sites:** ORegAnno[21], FirstEF[22]), CpG repeats islands[23] **Transcription Factor Binding Sites:** Conserved TFBS[8], whole-genome chromatin immunoprecipitation (Yale WG[24], ENCODE CHIP[25]) and DNA hypersensitivity (DNAse, FAIRE techniques)[26]	The creation of new variant TFBS was predicted using DELTA-MATCH[27], and Pupasuite [28]
Splice sites and splicing control elements	Pupasuite, RESCUE-ESE[29], and ESE Finder[30], Location relative to Exon-Intron boundary.	All possible exons locations were determined using ENCODE Gencode [16]
Conserved Regions	28-way PhastCons conservation track [8]	

splicing modulating variants. For each source, a score was estimated based on the quantity and statistical significance (if available for given source) of the supporting data (Table 12). The scores for the individual data sources were summed within each functional category in order to create a categorical score.

In order to take into account the conservation data in variant selection, the level of conservation encompassing the variant was used to determine a score (Table 12). The conservation score was added the Regulatory score for each variant to create a "Regulatory plus Conservation score". A "Splicing Control plus Conservation score" was calculated for the subset of variants having a Splicing Control score >1. To prevent the conservation from skewing the selection of splice control variants to those with a high conservation, conservation scores were only added to variants with some splicing control evidence. A "Regulatory plus Splicing plus Conservation score" was calculated for all variants in order to allow for variants that have data supporting multiple functional mechanisms. Due to the amount of data, the conservation data skewing was less profound.

Both "Regulatory plus Conservation" and "Splicing Control plus Conservation" categories had score thresholds set to allow the prioritization of the top variants in each category. However, differences in the types and abundance of data available for Regulatory and Splicing categories resulted in differences in approach to setting a threshold.

The Regulatory category had multiple sources of predictive and experimental evidence--often with quantitative scores related to P-values. The threshold score (Regulatory plus Conservation) was set to a threshold requirement of >3 points, which was chosen to require that the variants

Table 12: Scoring scheme used in the prioritization of variants for sequencing

Function	Scoring scheme
Regulatory	CpG islands, promoter or first exon (FirstEF), Stat1 or Pol2 TFBS (Yale whole-genome ChIP, Chip Mapper, Pupasuite, ENCODE DNA Hypersensitivity sites and TFBS ChIP, EIO/JCVI Nuclease accessibility sites): 1pt/each Sources with quantitative data were assigned a score based on level of significance calculated by source: Deltamatch TFBS affinity prediction (Dif z-score=0.25-0.49 (1pt); 0.50-0.99 (2pts); 1(3pts)), Conserved TFBS (z-score= 1.64-2.32 (2pts); >2.33 (3pts)). 7x Regulatory Potential (RP score= 0.10-0.19 (1pt) , 0.20-0.29 (2pts) , 0.30-0.40 (3pt), etc.).
Splicing Control	Variants were predicted to be in exon splicing control elements through the use of Pupasuite and ESE Finder/RESCUE ESE: 2 pts/each. The variants' locations from an intron/exon border based on all known transcripts were also calculated: Variants <100bp:1pt. Intron splicing elements were predicted based on proximity to exon/intron border (within 100 bp) and conservation data.
Conservation	Conservation data from 28-way alignment were used to assign a conservation score as follows. An alignment score of 0.70-0.79 (1pt); 0.80-0.89 (2pts); 0.90-0.99 (3pts); 1 (4pts)

would need support from several experimental and predictive evidences in order to increase specificity.

This threshold approach differs from the Splicing Control category threshold, which was reliant on fewer sources of quantitative data. The threshold score (Splicing control plus Conservation) was set (>1 point) to include 1) all variants with any predictive Exon Splicing Enhancer (ESE) evidence and 2) variants within 100 bp of splice junctions with conservation support.

Nonsynonymous variants were determined using data from the above mentioned sources. A subset of the data is presented to illustrate locations of possible coding regions. The putative effects of the coding variants were determined using the POLYPHEN, SIFT, and SNPs3D prediction engines. These data resulted in qualitative determination rather than a quantitative score.

Resequencing of ATPAF1

Individual Selection

Forty individuals, 20 asthmatics and 20 without asthma, were selected for sequencing from the Isle of Wight birth-cohort study population nested case-control subset used in the previous study that identified the association between asthma and *ATPAF1* (Table 11, Appendix). Forty individuals were chosen based on estimates that targeted sequencing of ≥30 individuals would result in a >95% statistical power to detect variants with >5% MAF [31]. In order to simplify sequencing, haplotype data from genotyping were used to remove individuals with heterozygous haplotypes from selection consideration, with the exception of individuals with rare haplotypes

that occurred only in the heterozygous state [1]. A random number generator was used to select individuals for consideration. After selection, haplotype frequencies for the 40 individuals were compared to the larger nested case-control from the previous association study. Several individuals in each group were changed in order to keep the haplotype frequencies consistent with those found in the association study (Table 12, Appendix).

Primer Design and sequencing scheme

The location of the prioritized variants from the *in silico* screen were combined with the locations of functional regions to create genome intervals that would be targeted for sequencing. These functional regions included 1) *ATPAF1* coding regions (including those from coding prediction and reported EST data), 2) regions highly conserved between placental mammals, and 3) expression regulation regions determined by motif matching of known regulatory sequence from 7 species of placental mammals. Functional categories and regions overlapped and were not mutually exclusive. The locations of the prioritized variants and putative functional regions were combined into 22 intervals. Subsequently, some intervals were merged to make sequencing more practical. Thirty-five amplicons, some of which were overlapping, were designed to sequence these intervals (Figure 13). The range of the length of amplicon was 104 to 695 bp and the average length was 487 bp. Primer sequences are located in Table 13.

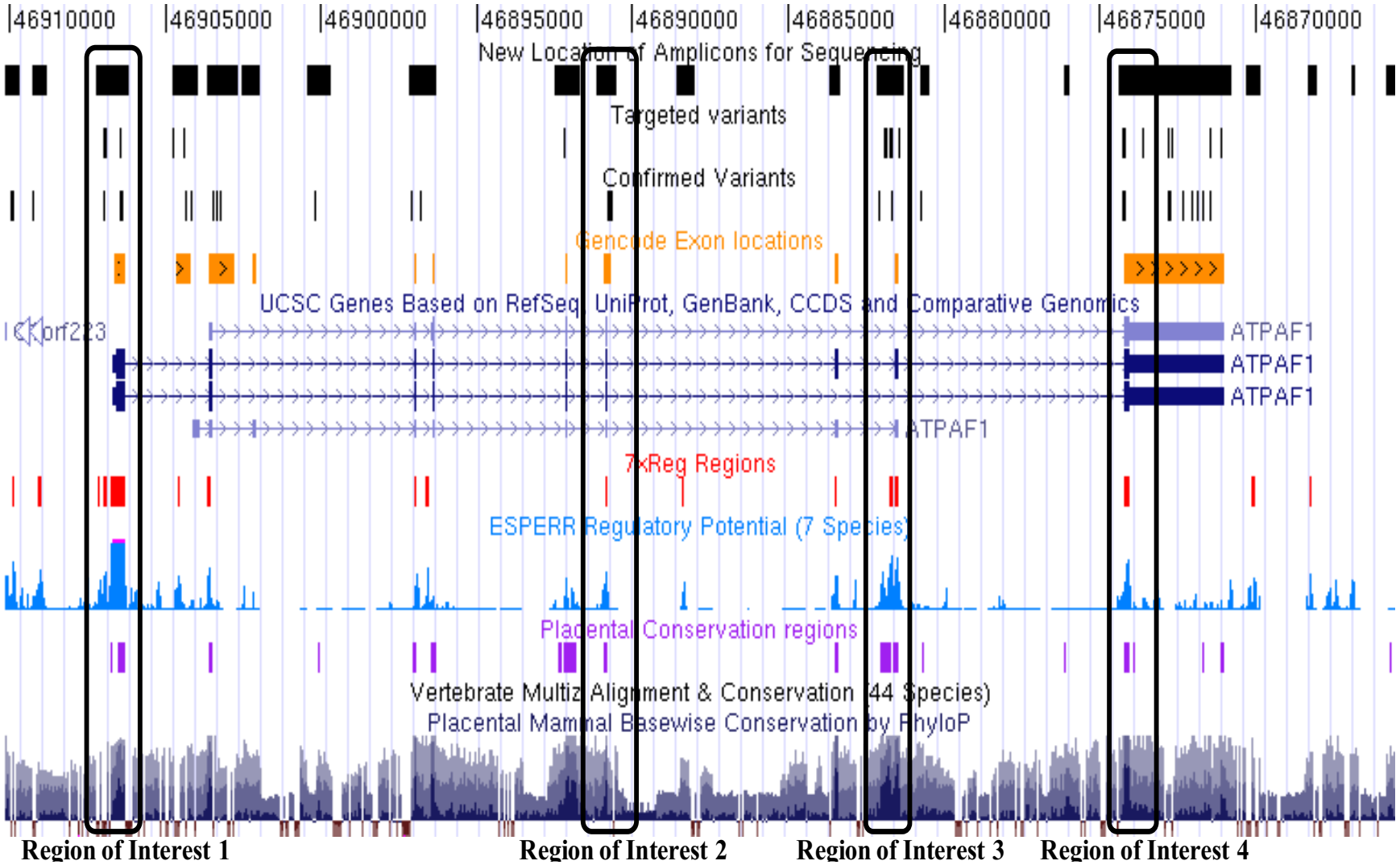

Figure 13: Gene track of Chr1p33 from UCSC Genome browser [8] indicating the location of variants and functional support data. From top to bottom: locations of sequenced regions, locations of variants, Gencode showing coding regions, regulatory potential and conservation tracks showing the targeted regions. The defined Regions of Interest with variants with predicted function are indicated.

Table 13: Primers used for Sanger sequencing (listed in order from 5' to 3' of gene)

Primer Name	Primer Direction	Sequence (5'-3')
ATPAF1_1	pF	ACTCACCCACCCATCTGCCC
	pR	CCACTTGGTTCCCTTGAGCAGGTTT
ATPAF1_2	pF	TGATCCCACCATCAGGGCCA
	pR	AAAGCCAGGGGTCTGGGTGAG
ATPAF1_3	pF	GGAAAAGGAGGCTTGGGAGGCA
	pR	GCTCTCGCCGCCCTGTGTAT
ATPAF1_4	pF	GGCATACACAGGGCGGCGAGAGCG
	pR	GCCTCGCGGAGCCTGGCTACATCATC
ATPAF1_5.1	pF	GAGCTCCTTGAGGTAGGGAGTCTGA
	pR	CAGGCTGGGGTGCTGTCCC
ATPAF1_5.2	pF	GGAGCCGGCCCACTACAGAG
	pR	TCATCTTGCTGGATCACCTCTTCAGAC
ATPAF1_6.1	pF	GGCCCAGTTGGTTTACCCTTTGGT
	pR	TGACCCCTGCTCTACATCAAAGTCCC
ATPAF1_6.2	pF	GGTGGTCAAGGCAAGGCACAC
	pR	TGGCCCCAGAGCCCATCCTC
ATPAF1_7	pF	TTAGCGCCGTCTCTGGCATA
	pR	TCACTTAAAATTCAGGATACCCAAGGC
ATPAF1_8.1	pF	TTTATGAGGCCGGGCGTGG
	pR	TGTGTTAGCCGTTAGCCTTCCTC
ATPAF1_8.2	pF	AGCGGAGATCACACCACTGCACTCCA
	pR	AGAAATGGGTGGAAGCACCTCATGACTACA
ATPAF1_9	pF	CACTGCGCCCAGCCAACAAT
	pR	TGCGGCACTTTGCAGCTGGT
ATPAF1_10	pF	GCCGCAGGACTCATTGGTAAGG
	pR	CCTTTCATTGTGGCTAGAAAGAGGTGTC
ATPAF1_11.1	pF	GGTCTGCCTGACTGTGAACCCT
	pR	GGAGCTTCCAATCTCTGGTGCTCA
ATPAF1_11.2	pF	ACCCACATGTCACAGACACCCCTCTCTT
	pR	AGGCTCCAAGTCAGCTCAGGTATAGGTCACT
ATPAF1_12	pF	CTCCCTCAAACTGACTTGGTGACTTGAT
	pR	AGCCAGAGGCTTACTGGGTCAAAC
ATPAF1_13.1	pF	GCTGCCAGAAGGCCTAAGGGAAA
	pR	CCTGGGCAACGTAAGTGAGATCTTG
ATPAF1_13.2	pF	TGGCCTCAAATGACCCTCCCACCT
	pR	AGCTGAGCATGGGGGTGCATGTTTGT
ATPAF1_14	pF	GCAGATCAAGACTGCGGTGAGC
	pR	AGCCTCAGCAAGGGTTTCTAGTCC
ATPAF1_15.1	pF	CCACCTGTGCTATGTAGGGGCTTT
	pR	GTGAGTTGGAGGACTGGTATGTGGTG
ATPAF1_15.2	pF	CTTGATTTGGTAGCTGCTTGTGCC
	pR	TCTTGAGCAAGACAACTTCACTGTCA

Table 13 (cont'd)

ATPAF1_16	pF	TGCTCATAGTTAATGATGCCCACTGTGTT
	pR	TCCTCACTGTGTCCTAGGTTTGTGGGT
ATPAF1_17	pF	TCCTGGCTGTGACCTTGGCTG
	pR	GCAGGCACTAATCTCTGAGACTGCAT
ATPAF1_18.1	pF	TTGGTACACAGAGAAGGGGCACCTAACCA
	pR	TGAGTTCCTGAGGAGGGGGTGGGC
ATPAF1_18.2	pF	GCCCTGTGCAAGAAAGGACCTCAT
	pR	GCTTTTGTTCAGCTGACTGGTGC
ATPAF1_18.3	pF	GCCCACCCCCTCCTCAGGAACT
	pR	ACCCCAAGCCAGCTACTTACACTAAAGGC
ATPAF1_18.4	pF	TGTTCGAGGTTATAACAAGAGACAGCCA
	pR	GCCCATAGCCCTGACACAATGGA
ATPAF1_18.5	pF	TGGATGAAAGCAAATGTTCCCAGATTGGAT
	pR	GCCTGGGCAAGACAGAGCAAG
ATPAF1_18.6	pF	CCCAGTATTGTTGTCAGCTGCCT
	pR	GACCAGAAGCAGGTCCAGAATTTACAT
ATPAF1_18.7	pF	TGGGCCTGTCCCTAAGCTCTTT
	pR	ACCATGCTAGCACCAGGGGA
ATPAF1_18.8	pF	TCCTCAGTCCCCAACTGCCAAG
	pR	TCAAAGTGTGGCTGGCTAATTACCTACAT
ATPAF1_18.9	pF	CTTTAAGATGCAGTAGCATCACCTAAAGTG
	pR	CTGTGAGCAGCTGTTAAGTTGTACTG
ATPAF1_19	pF	ATGGAGTGGAGGCAATAACTTAATGGTCAG
	pR	TGACTGTAGCTTGGGAGGCCAGC
ATPAF1_20	pF	AGCCCTTACATGGATTATCTCAATCCTACCAA
	pR	AGGTGTGTGCCACCACATCCAGC
ATPAF1_21	pF	GGCCTTACAGACTATCCAGTCACCAG
	pR	GGTGCAGGTCAGTTAGCAGGTAGAGA

PCR and Sanger sequencing

DNA from each subject was whole-genome amplified using the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Piscataway, NJ) according to manufacturer protocol to ensure an abundant supply of genomic DNA. Genomic-amplified DNA was PCR amplified using GoTaq Polymerase (Promega, Madison, WI) using a protocol with a low concentration of primers (reaction concentration= 0.1 μM primers) and dNTPs (reaction concentration= 0.1 mM dNTPs) to enable sequencing of PCR product. PCR products were separated by TAE gel electrophoresis to verify amplification of a single product. Primer sets found to require a higher concentration of primers and/or dNTPs to amplify properly were treated with a combination of Exonuclease 1 (USB, Cleveland, OH) and Antarctic phosphatase (New England Biolab, Ipswich, MA). The PCR additive Betaine (reaction concentration=1.0M, Sigma-Aldrich, St. Louis, MO) was used to amplify the primerset in the GC rich region of the ATPAF1 promoter. PCR amplification was followed by Sanger sequencing at Michigan State University Research Technology Support Facility core facility. Each amplicon was sequenced in the forward and reverse directions using the amplification primers (Table 13).

Analysis

The sequence data were analyzed using Lasergene software (v.8.0, DNASTAR, Madison, WI). Sequences were aligned to each other and the reference sequence from UCSC Genome browser [8]. The software indicated possible location of variants that were then examined and manually verified.

PLINK was used to establish the most likely haplotypes from the sequence data [32]. These data were also integrated with SNP genotype data from the previous asthma association study in which the *ATPAF1* gene was implicated. This combined dataset allowed mapping of variant alleles onto asthma risk haplotypes. This information was used during the post-validation prioritization to select variants inferred to modulate asthma risk.

Post-sequencing functional candidate selection

After the analysis of the sequence data to validate/discover new variants, the data from the functional screen were used to prioritize the variants. Association data from the Isle of Wight cohort [1] were incorporated to improve selection of functional candidates with possible relevance to asthma.

Imputation Analysis

An imputation case-control association analysis was used to predict association between asthma and the variants in the *ATPAF1* gene. This step was not part of the functional screen but was available for use in the postsequencing functional candidate selection.

The data were derived directly from genotypes determined in the Isle of Wight cohort or through imputation analysis implemented in PLINK [32] utilizing Hapmap data (CEU Population, R23a)[33]. The imputation was based on 14 previously genotyped SNPs. Every SNP imputation was based on a minimum of 2 SNPs (average = 2.12). A SNP information content metric (INFO)

of greater than 0.8 was used as a quality control requirement. Analysis was available for a total of 94 variants, of which 14 were genotyped and the remaining 80 imputed. A P-value of <0.05 was considered significant.

RESULTS

In silico functional screen and variant selection

Three hundred three variants in or near the *ATPAF1* gene (≈ 7,500 bp in 5' direction and ≈ 500 bp 5' direction [limit until adjacent gene]) were identified and screened for function using *in silico* techniques. This functional screen resulted in the identification and prioritization of 27 functional variant candidates in the *ATPAF1* gene for resequencing (Table 14).

Nonsynonymous variants

The analysis of coding variants revealed that one variant, rs11211337, results in a missense mutation. However, this change was predicted to have a benign effect on the *ATPAF1* protein [18]. An additional four coding variants were targeted but are predicted to be synonymous changes.

Table 14: Functional variant candidates targeted for resequencing by *in silico* screen

Variant (SNP unless noted)	Location (Chr 1)	Validated†	Location/ Function	Promoter/Regulatory Element Evidence	Splicing Control Element Evidence
rs11579762	46859089	X	intergenic	Large change in predicted binding affinity of several TFs	
rs6676940	46859747		intergenic	Change in predicted affinity to TF, TFBS (ChIP)	
rs6669274	46860703	X	intergenic	Multiple DNA Hypersensitivity sites (HS), Regulatory motiff	
rs58854755	46860725		intergenic	Multiple DnaseHS, Regulatory motiff	
rs56238166	46871098		3'UTR	Regulatory motiff, High conservation	High conservation
rs11541308	46871410		3'UTR		Predicted loss of ESE
rs1258068	46872636	X	3'UTR		Predicted loss of ESE
rs1139758	46872746	X	3'UTR		Predicted loss of ESE
rs724309	46873602	X	3'UTR		Predicted loss of ESE
rs41298535	46874152		coding-synon	Regulatory motiff, High conservation	Predicted ESE, In or near exon/intron boundary
rs4660952	46874185		coding-synon	Regulatory motiff, High conservation	Predicted loss of ESE
rs61783140	46874206		coding-synon	Regulatory motiff, High conservation	High conservation
rs57221043	46881403	X	intron		Proximal to splice site
rs58805974	46881416		intron		Proximal to splice site
rs35508143 $	46881660		intron	Conserved TFBS, TFBS (ChIP), Reg. motiff, DNaseHS, High conservation	Proximal to splice site, High conservation
rs1745377	46881676	X	intron	TFBS (ChIP), Regulatory motiff, DNaseHS, High conservation	
rs58289597	46881826		intron	TFBS (ChIP), High conservation	

$Indel variant, †Validation status from dbSNP129, #Validation on Affymetrix 100k SNP array used in Chapter 2

Table 14 (cont'd)

Variant (SNP unless noted)	Location (Chr 1)	Validated†	Location/ Function	Promoter/Regulatory Element Evidence	Splicing Control Element Evidence
rs60517812	46881879		intron	High conservation, Near conserved TFBS	
rs59407595	46881903	X	intron	Conserved TFBS, Regulatory motiff, High conservation	
rs620913 #	46890654	X	intron	DNaseHS	Proximal to splice site
rs11589863	46891639		intron	High conservation, Change in predicted affinity of TF	
ENSSNP226730	46904372		intron	DNaseHS	
rs636871	46904710		intron		Proximal to splice site
rs11211337	46906398	X	missense	Promoter, CpG Island, TFBS (CHiP), DnaseHS	Predicted loss ESE, Proximal splice site, Conservation
rs56218488	46906405		coding-synon	Promoter, CpG Island, TFBS (CHiP), DnaseHS	Proximal to splice site, Conservation
rs61364505	46906896		near-gene-5'	Promoter, CpG Island, TFBS (CHiP), DnaseHS	
rs12094663	46906908	X	near-gene-5'	Promoter, CpG Island, TFBS (CHiP), DnaseHS	

$Indel variant, †Validation status from dbSNP129, #Validation on Affymetrix 100k SNP array used in Chapter 2

Regulatory variants

Combining the results of the Regulatory category plus the Conservation data resulted in the identification of 98 variants. Twenty-three variants, representing the approximately the top 25% of variants in this category, met the threshold score of >3 points. This threshold was set to require that the majority of variants would need support from several experimental and predictive evidences in order to increase specificity. In order to focus on variants most likely to be present in our population, the rare variants were removed from consideration. This resulted in three variants being removed from consideration and the 20 variants described in Table 14.

Splicing Control

There were 26 variants identified for predicted Splicing Control. These variants were then combined with the conservation data to increase the likelihood of predicting functional variants. Fifteen variants had both Splicing Control evidence plus Conservation that met the threshold (>1 point) for promotion to validation in resequencing (Table 14). This threshold resulted in the selection of 1) all variants with any predictive exon splicing enhancer (ESE) evidence and 2) variants within 100 bp of splice junctions with conservation support. Two variants (rs723333 and rs41298535) were found to be within 3 bp of exon/intron splice boundaries and may directly affect the affinity of splicing machinery.

Multiple categories

A total of eight variants met the threshold in both regulatory and splicing control categories (Table 14). Other variants that had functional evidence in multiple categories yet may have narrowly missed meeting threshold in any single category were examined by adding all scores together. The addition of strong conservation data to Regulatory plus Splicing Control did not yield any additional variants as variants with strong evidence of function were already targeted.

Summary

A total of 34 variants met threshold for targeted resequencing for validation for any one or several categories of predicted function (nonsynonymous, regulatory, splicing control). However, cohort genotype data (combined with Hapmap CEU LD data) suggests seven of these candidates have a very low MAF (≤0.008) in both asthmatic and control individuals; therefore, these variants were excluded from further consideration. Of the remaining 27 variants (=34-7 rare variants), 11 had prior validation data and 16 were unknown (Table 14). Of the variants with prior validation, 10 had strong data supporting their existence in Caucasian populations. These 27 variants were targeted for validation by resequencing in cohort children. Eight of these variants met the threshold in both regulatory and splicing control categories.

Resequencing

A total of 14.6 kb of DNA within and flanking the *ATPAF1* gene was sequenced in 40 individuals. This resulted in the identification of 34 variants in the Isle of Wight subset of individuals. This includes nine novel variants and validation of 25 previously reported variants, which is fewer than the 77 variants reported to be in the sequenced region based on dbSNP build 129 data + 1000Genome Project [9] (Figure 13, Table 15). Of the 27 variants targeted by the *in silico* screen before resequencing, only nine were validated in sequenced individuals. Thirteen rare variants (MAF <0.05) were found with fifteen of their eighteen alleles found in asthmatics with one asthmatic individual possessing five rare alleles (Table 16). Only one other individual had more than one rare variant. Nine novel variants were discovered, of which all were rare. One nonsynonymous-missense mutation was discovered.

The genetic structure of the subset of individuals sequenced was found to approximate the larger Isle of Wight cohort. The approximate MAFs and LD found in the sequenced sample (n=40) were representative of the entire nested-control when compared to data from genotyping all individuals (N=277; Figure 14). The accuracy of MAF and LD in the cohort is necessary for establishing the relative risk conferred by a specific allele so it will be applicable to the larger population.

After resequencing, the validated and novel variants were functionally screened using the same *in silico* technique used before resequencing. It is important to note that many variants had overlapping functions and fit into several categories. Based on functional predictions it was predicted that: five variants may have regulatory function (three with strong support from

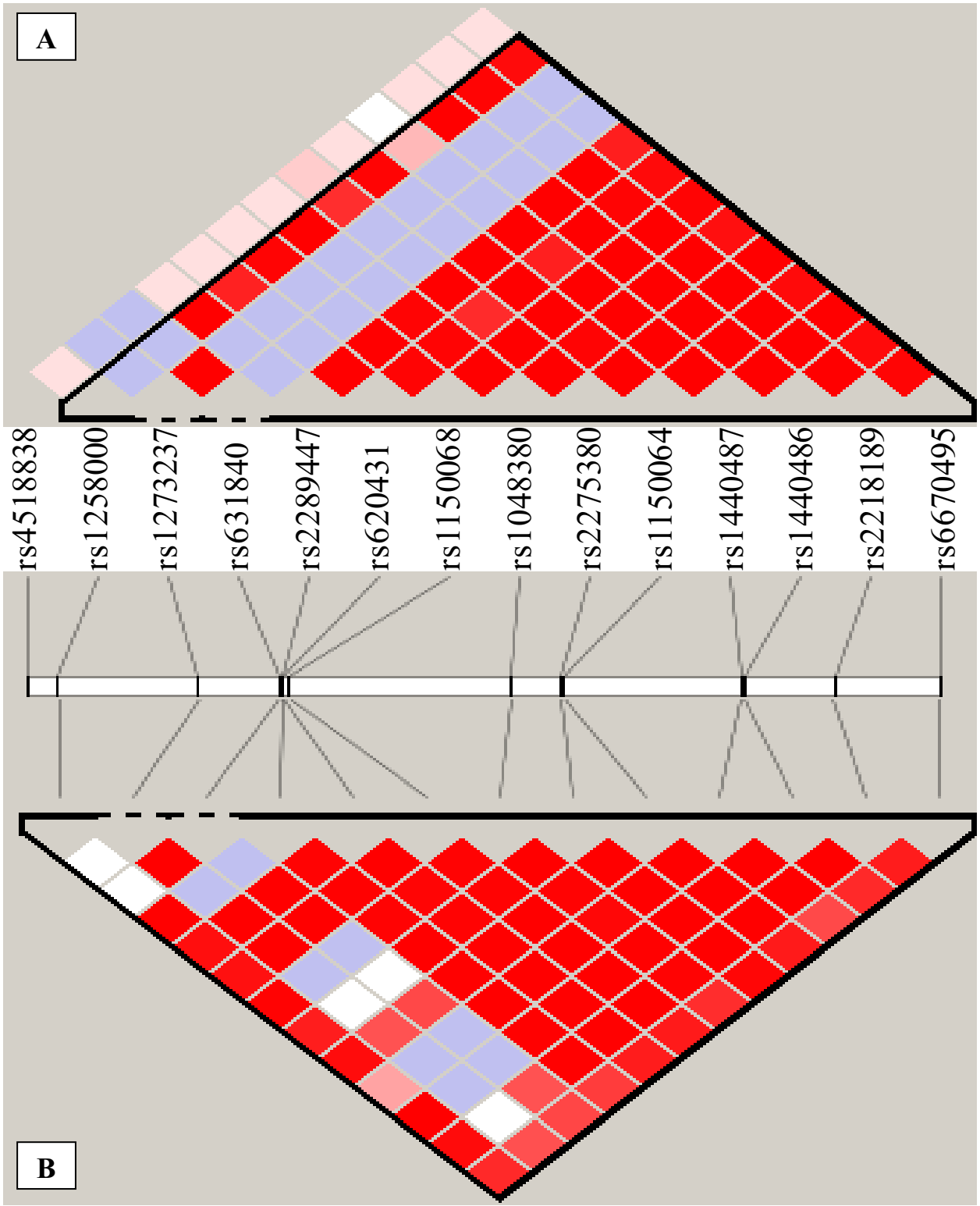

Figure 14: Haplotype LD (D'/LOD) plot for the (A) subset of individuals (N=40) examined by targeted resequencing including genotype from past association study and (B) all individuals (N=277) from the association study.

Table 15: Validated and discovered variants from resequencing in the Isle of Wight subset in and near the *ATPAF1* gene.

	Variant (all SNPs)	Location (Chr 1)	Isle of Wight MAF	Location/ Function	Conserved region	Promoter/Regulatory Element Evidence	Splicing Control Element Evidence
	rs11541308	46871410	0.064	3'UTR			Predicted loss of ESE
	rs11403101	46871686	0.09	3'UTR			
	rs4660950	46871822	0.488	3'UTR		TFBS (predicted change in affinity)	
	IOW_9	46872023	0.026	3'UTR			
	rs723334	46872308	0.231	3'UTR			
	rs1139759	46872686	0.487	3'UTR			
	rs1139758	46872746	0.224	3'UTR			Predicted loss of ESE
ROI4	rs41298535	46874152	0.013	coding-synon	X	Regulatory motiff	Predicted gain ESE, <100bp from splice site
ROI4	rs4660952	46874185	0.013	coding-synon	X	Regulatory motiff	Predicted loss of ESE, <50bp from splice site
ROI4	rs61783140	46874206	0.013	coding-synon	X	Regulatory motiff	Predicted loss of ESE, <50bp from splice site
	IOW_8	46880719	0.038	intron			
ROI3	IOW_7	46881628	0.013	intron	X	Regulatory motiff, DNAHS(FAIRE)	<100bp from splice site
ROI3	rs1745377	46881676	0.244	intron	X	TFBS (ChIP), Regulatory motiff, DNaseHS	
	rs1933933	46882060	0.212	intron			
ROI2	rs620913	46890654	0.211	intron		DNAHS	<50bp from splice site
ROI2	rs2289446	46890683	0.237	intron			<100bp from splice site
ROI2	rs2289447	46890755	0.184	intron			<100bp from splice site
ROI2	rs620431	46890776	0.197	intron		Regulatory motiff	<100bp from splice site
	rs34308534	46896795	0.225	intron			
	rs611468	46897066	0.212	intron			<100bp from splice site

Table 15 (cont'd)

	Variant (all SNPs)	Location (Chr 1)	Isle of Wight MAF	Location/ Function	Conserved region	Promoter/Regulatory Element Evidence	Splicing Control Element Evidence
	rs656263	46900156	0.257	intron			
	IOW_6	46903176	0.012	intron			
	rs2218190	46903181	0.475	intron			
	rs2218191	46903215	0.262	intron			
	IOW_5	46903301	0.028	intron			
	rs594118	46903460	0.227	intron			<50bp from splice site
	rs72898751	46904115	0.013	intron			
	rs12402844	46904321	0.051	intron			
ROI1	IOW_4	46906346	0.013	missense	X	CpG Island, Promoter, Regulatory motiff, Pol2 site, DNAHS (DNAse)	Predicted loss of ESE, <50bp from splice site
ROI1	rs11211337	46906398	0.275	missense	X	CpG Island, Promoter, Regulatory Motiff, CpG Island, TFBS (CHiP), DNAHS	Predicted loss of ESE, <50bp from splice site
ROI1	rs12094663	46906908	0.238	near-gene-5'	X	Promoter, Regulatory motiff, CpG Island, TFBS (CHiP), DNAHS	Predicted loss of ESE, <100bp from splice site
	IOW_3	46909189	0.013	near-gene-5'		Regulatory motiff	
	IOW_2	46909837	0.013	near-gene-5'		Regulatory motiff	
	IOW_1	46909937	0.026	near-gene-5'		Regulatory motiff	

Table 16: Rare variant findings with individual keyed by asthma status.

Rare Variant	Individual (study number)			Key	
IOW_9	425	500		Atopic Asthma	
rs41298535	549			NonAtopic Asthma	
rs4660952	549			Nonasthma	
rs61783140	549				
IOW_8	40	237			
IOW_7	338				
IOW_6	549				
IOW_5	99	441			
rs72898751	549				
IOW_4	32				
IOW_3	902				
IOW_2	37				
IOW_1	12	237			

multiple sources); seven variants may have function in effecting splicing (three with strong support); and six variants may have a role in regulatory and/or splicing modulating control (three with strong support; Table 15, Figure 15). Two variants, including one novel missense mutation, are nonsynonymous and also have strong support in regulatory and/or splice modulating role. Fourteen variants do not currently have a predicted function.

Post-sequencing functional candidate selection

Variants with the strong functional evidence were found often clustered together (Table 15, Figure 13). Four such regions were assigned numerically as Region of Interest (ROI) 1 through 4. These regions contained a total of twelve variants (ten previously known variants and two novel variants). Five of the twelve variants were predicted to have multiple functional consequences, including two that were nonsynonymous/missense mutations, three had splice modulating, and three had gene regulatory putative functions (Figure 16). An example of a region with strong functional support, ROI1 contains three variants (IOW_4, rs11211337, rs12094663) located within 600 bp of each other near the *ATPAF1* promoter and Exon 1. Two of these SNPs are missense mutations and all three of them are predicted to control both gene regulation and splice modulation.

Also pertinent to the selection of candidates for causative variants was the strength of the association of the variants with asthma. The imputation analysis utilizing Caucasian Hapmap data combined with Isle of Wight genotype data showed protective associations (OR=0.6 [no confidence intervals generated], P-value range=0.0057-0.0135) with several validated variants

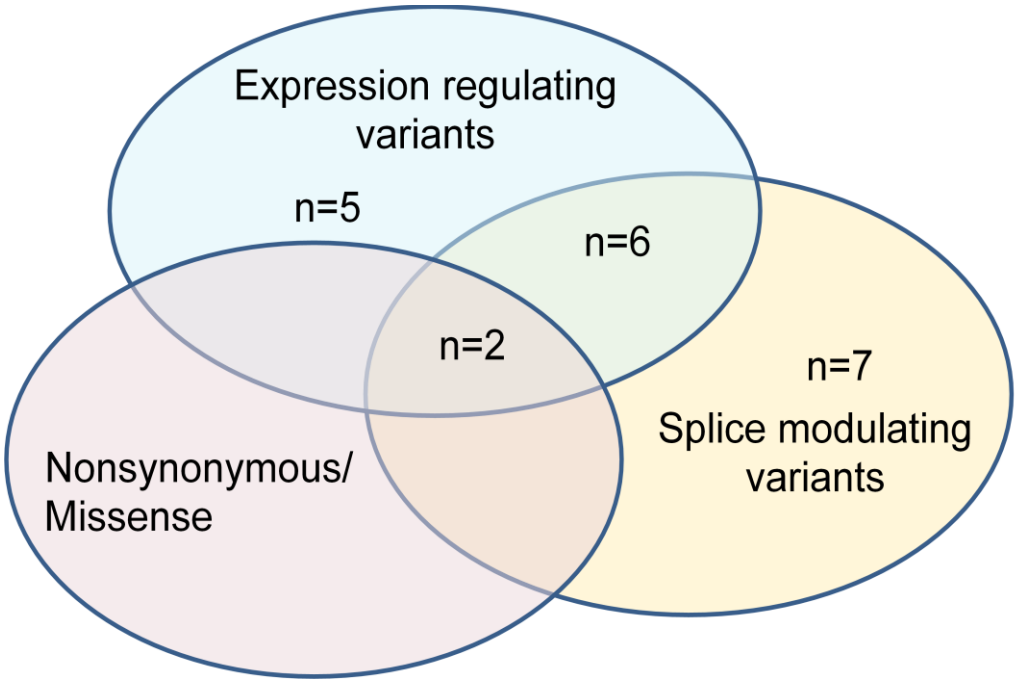

Figure 15: Venn diagram of predicted function for all validated variants (N=34) found by resequencing (20 variants predicted to have function (above),14 variants were not predicted to have function (not shown))

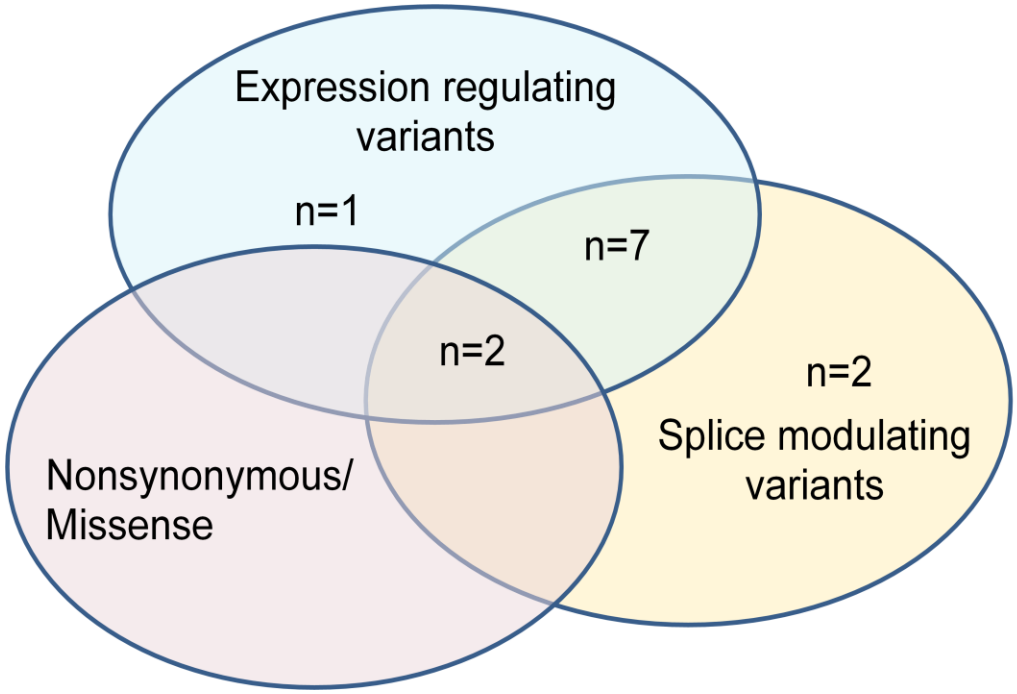

Figure 16: Venn diagram of variants selected as functional candidates (N=12) in the four regions of interest.

including rs723334, rs1745377 [ROI3], rs2289447 [ROI2], rs620431 [ROI2], rs611468, and rs12094663 [ROI1].

DISCUSSION

We utilized an *in silico* functional screen combined with targeted resequencing to narrow the search for functional variants in the region containing the asthma susceptibility gene, *ATPAF1*, from 303 down to 12 in four functional regions of interest. The screen resulted in the targeting of 27 variants for validation by resequencing with several variants classified into more than one functional category (regulatory, splice controlling, missense). In addition to the targeting of these variants, DNA regions in the *ATPAF1* gene and flanking 10 kb with predicted functional role (exons, conserved, and regulatory regions) were examined by targeted resequencing in 20 asthmatic and 20 nonasthmatic individuals. In the resequencing, a total of 34 variants were found with nine of the 27 variants prioritized in the screen validated and nine novel rare variants discovered in sequenced individuals. The following were identified: two missense mutations with predicted function in regulation and splice control, seven variants with regulation and splice control, one variant with regulation function only, and two variants with splice control only (Table 15, Figure 15).

Variants were re-evaluated for functional candidate selection. Twelve variants (ten variants that were previously known, two novel variants) were prioritized for functional study in four regions of interest. Nine of the twelve variants were predicted to have possible multiple functional consequences including two that were nonsynonymous/missense mutations, two had splice modulating, and one variant had gene regulatory putative functions (Table 15, Figure 16). The

use of an *in silico* functional screen has been shown to be a powerful method of narrowing the search for disease variants from candidate gene association studies [34].

Functional variants were predicted based on their location in or near *cis*-acting functional DNA elements using many diverse sources of functional element predictions combined with a variety of experimentally derived data. For example, although conservation is a powerful predictor of function, it is nonspecific. Here, conservation was combined with regulatory sequence similarity, ChIP, and DNase HS, or ESE motif mapping to provide a robust method of selecting putative regulatory variants. It is important to note that not all sources of data were immediately relevant to asthma (for example, ChIP data from non asthma relevant cell lines) but this does not preclude functionality. It is also recognized that the effects of an allele of a putative functional variant may only be seen in specific cells or tissues.

A scoring system was used to rank the variants for selection for sequencing. There are no previously published methods for combining the various available datasets relevant to functional prediction. Our results are robust with the majority of the top scoring variants selected using several schemes. In the analysis presented here, a subset of *in silico* data was used for simplification; however, an analysis using the complete data-set yields a similar set of functional variant candidates.

Targeted resequencing has become a popular follow-up study to GWAS. In the present study, we used targeted resequencing to validate predicted functional variants while also hunting for novel variants in functional regions of the *ATPAF1* gene. The value of the strategy employed here of combining an *in silico* functional screen of variants with resequencing of functional regions allowed us to leverage knowledge of variants. Resequencing of predicted functional regions

alone would have ignored data of variants already in databases. However, our strategy also allowed us to detect novel variants. The data support that in the Isle of Wight cohort, the variants in *ATPAF1* that have, until now, gone undiscovered are the rare variants. All nine novel variants discovered had a MAF <5% while the other 25 variants that were validated (21 common, 4 rare variants) in the resequencing had an average MAF of 20%.

It was anticipated that few if any nonsense mutations would be identified in the *ATPAF1* gene. ATPAF1's function as a chaperone protein for the F1-subunit of mitochondrial ATP synthase is required for assembly of this complex [35]. Null mutants in yeast are non-viable [36]; therefore, it is predicted that nonsense mutations would have been evolutionarily been selected against in humans.

The sequencing of both asthmatic and control individuals was chosen to allow for identification of protective variants that would be missed by sequencing asthmatics only. It is also clear that there are many variants not in the dbSNP database so findings in cases need to be compared to controls. The study found that 15 of 18 rare alleles found in the resequencing (in 13 SNPs) were in asthmatics (Table 16, Figure 17). The functional data revealed that many of these rare variants were also the most promising variants functionally. Region of interest 1, near the *ATPAF1* promoter, contains IOW_4, a rare variant found in an asthmatic that is predicted to be a missense mutation but also could affect splicing and gene regulation. Region of interest 3, near the alternatively spliced exon 8, contains the novel variant IOW_7, which is found in an asthmatic and implicated in splice modulation and gene regulation. All three variants in Region of interest 4 (in exon 9, all previously known) were found clustered within ≈50 bp of each other in a single asthmatic individual and are predicted to have splicing and regulatory properties. These findings could reflect an increased risk that is consistent with the multiple rare variant hypothesis in lieu

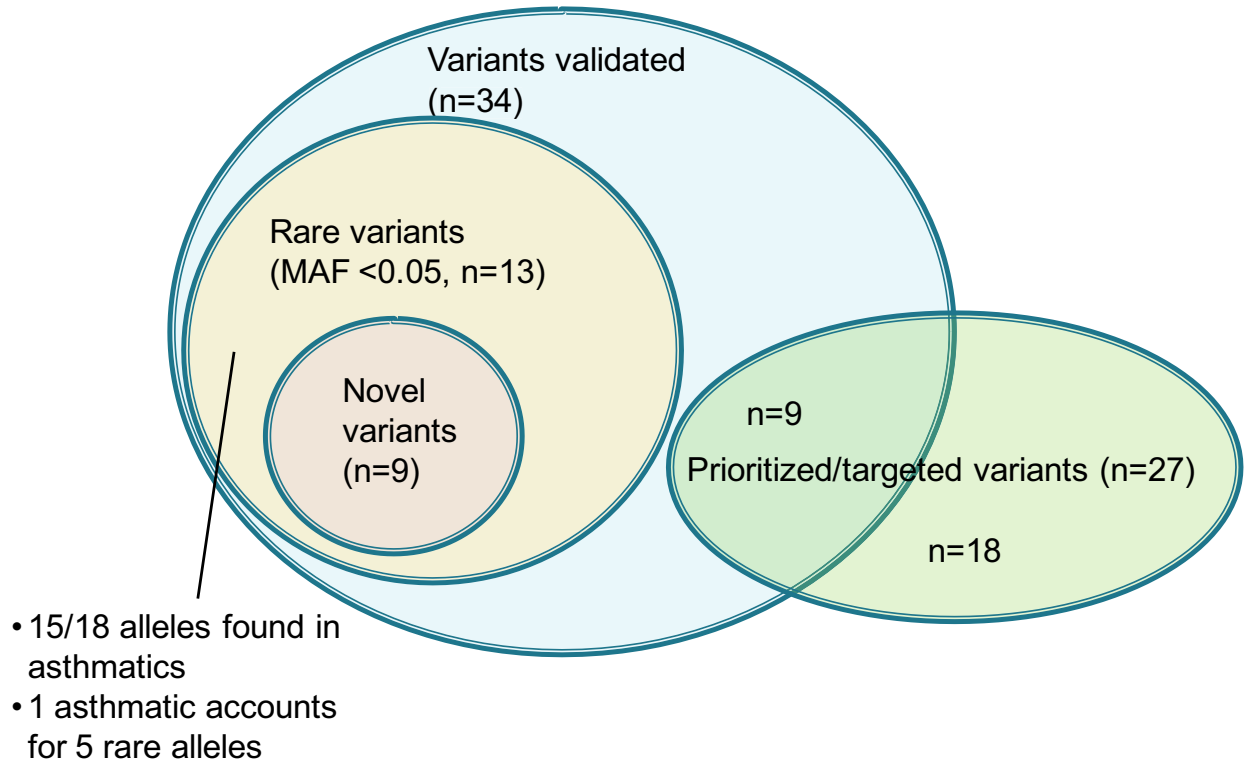

Figure 17: Venn diagram of variant validation and variant allele frequency from resequencing results. A total of 14.7 kb of DNA was sequenced in 40 individuals from a subset of study individuals from the Isle of Wight birth cohort. In the region sequenced there were a reported 77 variants (dbSNP 129 [1] +1000 Genome project at the time of sequencing). Numbers in parenthesis refer to the number of variants for a category while numbers without parenthesis refer to the number of variants in the overlapping region.

of the Common Disease Common Variant (CDCV) hypothesis. This is supported by finding by Dickson et al [37], who demonstrated in a simulation study that associations that were detected for common variants could come from the compound effect of several rare variants in what they called "synthetic associations". Furthermore, the occurrence of multiple functional variants among unrelated cases and controls provides both epidemiological and biological support for the causal role of the gene or pathway in a disease [38].

The DNA used in this study for resequencing was derived by whole genome amplification, which increases the risk of introducing errors into the template used for sequencing. However, this procedure utilizes the Phi29 DNA polymerase which has a reported error rate of 1 in 10^7 bp (roughly 100-fold less than Taq polymerase) [39]. Thus, the risk of an error is negligible. Therefore, it is most likely that the rare variants discovered in the resequencing are real and perhaps recent changes.

The importance of the work presented here is the identification of potentially functional DNA variant candidates in the *ATPAF1* asthma susceptibility gene. These variants will be explored in future studies. The *in silico* approach presented here to functionally screen variants represents a blueprint for a method that could be applied to association results of other genetic studies of any complex genetic disease.

REFERENCES

REFERENCES

1. Schauberger, E.M., et al., *Identification of ATPAF1 as a novel candidate gene for asthma in children.* J Allergy Clin Immunol, 2011.

2. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Research, 2007. **35**: p. D26-D31.

3. Vercelli, D., *Genetics and biology of asthma 2010: La' ci darem la mano.* Journal of Allergy and Clinical Immunology, 2010. **125**(2): p. 347-348.

4. Yuan, H.Y., et al., *FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W635-41.

5. Pico, A.R., et al., *SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system.* Nucleic Acids Res, 2009. **37**(Database issue): p. D803-9.

6. Xu, Z. and J.A. Taylor, *SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies.* Nucl. Acids Res., 2009: p. gkp290.

7. Chelala, C., A. Khan, and N.R. Lemoine, *SNPnexus: A web database for functional annotation of newly discovered and public domain Single Nucleotide Polymorphisms.* Bioinformatics, 2008.

8. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

9. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucl. Acids Res., 2001. **29**(1): p. 308-311.

10. Flicek, P., et al., *Ensembl 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D707-14.

11. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.

12. Nowak, R., *Mining treasures from 'junk DNA'.* Science, 1994. **263**(5147): p. 608-10.

13. Kaushik, M. and S. Kukreti, *Structural polymorphism exhibited by a quasipalindrome present in the locus control region (LCR) of the human beta-globin gene cluster.* Nucleic Acids Res, 2006. **34**(12): p. 3511-22.

14. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.

15. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.

16. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.

17. Jongeneel, C.V., et al., *An atlas of human gene expression from massively parallel signature sequencing (MPSS).* Genome Res, 2005. **15**(7): p. 1007-14.

18. PolyPhen. 2008; Available from: http://genetics.bwh.harvard.edu/pph/.

19. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

20. Yue, P., E. Melamud, and J. Moult, *SNPs3D: candidate gene and SNP selection for association studies.* BMC Bioinformatics, 2006. **7**: p. 166.

21. Griffith, O.L., et al., *ORegAnno: an open-access community-driven resource for regulatory annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D107-13.

22. Davuluri, R.V., *Application of FirstEF to find promoters and first exons in the human genome.* Curr Protoc Bioinformatics, 2003. **Chapter 4**: p. Unit4 7.

23. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.

24. Zhang, Z.D., et al., *Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions.* Genome Res, 2007. **17**(6): p. 787-97.

25. Boguski, M.S., *ENCODE and ChIP-chip in the genome era.* Genomics, 2004. **83**(3): p. 347-8.

26. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.

27. Deltamatch. *http://www.deltamatch.org*.

28. Conde, L., et al., *PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W621-5.

29. Fairbrother, W.G., et al., *RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W187-90.

30. Cartegni, L., et al., *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res, 2003. **31**(13): p. 3568-71.

31. Johnson, G.C.L., et al., *Haplotype tagging for the identification of common disease genes.* Nature Genetics, 2001. **29**(2): p. 233-237.

32. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

33. Gibbs, R.A., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.

34. Barnes, M.R., *Bioinformatics for geneticists : a bioinformatics primer for the analysis of genetic data*. 2nd ed2007, Chichester, England ; Hoboken, NJ: Wiley. xxii, 554 p.

35. Wang, Z.G., P.S. White, and S.H. Ackerman, *Atp11p and Atp12p are assembly factors for the F-1-ATPase in human mitochondria.* Journal of Biological Chemistry, 2001. **276**(33): p. 30773-30778.

36. Ackerman, S.H., J. Martin, and A. Tzagoloff, *Characterization of ATP11 and detection of the encoded protein in mitochondria of Saccharomyces cerevisiae.* J Biol Chem, 1992. **267**(11): p. 7386-94.

37. Dickson, S.P., et al., *Rare Variants Create Synthetic Genome-Wide Associations.* PLoS Biol. **8**(1): p. e1000294.

38. McClellan, J. and M.C. King, *Genetic heterogeneity in human disease.* Cell, 2010. **141**(2): p. 210-7.

39. Esteban, J.A., M. Salas, and L. Blanco, *Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization.* J Biol Chem, 1993. **268**(4): p. 2719-26.

CHAPTER 4

The characterization of *ATPAF1* expression and putative functional variants

Disclaimer: Dr. Marianne Huebner performed the GEE model statistical analysis. Methods for the GEE model analysis were adapted from her descriptions.

INTRODUCTION

Asthma is a complex genetic disease of the conducting airways of the respiratory tract involving TH2 cells and cytokines produced by these cells. Genetic studies of asthma have revealed associations with more than 100 genes, several of which have been identified using whole genome linkage or association studies [1]. A key benefit of whole genome studies is the identification of genes not previously known to play a role in asthma. However, without this *a priori* knowledge, it frequently is difficult to assign a role in disease pathogenesis to an associated gene, much less an associated DNA variant. If a variant does not lead to a nonsynonymous change, other plausible mechanisms for a variant's biological role include modulation of gene expression or alternative splicing. These variants often require significant follow-up studies to uncover the underlying biological mechanism which will, hopefully, lead to understanding a possible pathological mechanism.

Recently, we identified and replicated the association of the chr1p33 region, containing the genes *ATPAF1*, *KIAA0494*, and *C1ORF223,* as playing a role in asthma susceptibility [2]. Although variants in this susceptibility locus could affect the function of any of these three genes, *ATPAF1* was chosen as the best candidate after a gene expression study in bronchial biopsy samples from severe asthmatics showed a 50-fold increase in ATPAF1 mRNA as compared to controls [2].

We previously conducted an *in silico* screen followed by selective resequencing in and surrounding the *ATPAF1* gene. This resulted in the identification of three putatively functional

SNPs (rs11211337, rs12094663, and novel variant IOW_4 (238G>T)) in the promoter region of *ATPAF1*. These SNPs are present in four haplotypes in a Caucasian birth cohort from the Isle of Wight, U.K., which is the population in which we identified association between *ATPAF1* and asthma. The variants are predicted to be involved in the regulation of *ATPAF1* transcription based on (1) predicted promoter locations, (2) experimental data validating binding sites for transcription factors in various cell lines, (3) experimental data for DNAse I hypersensitivity, (4) predicted regions of gene transcription regulation, (5) location in regions that are conserved across multiple species, and (6) the variants' location in CpG islands (Figure 18) [2].

The location and gene structure of *ATPAF1* (chr1:46870998-46906686) and the three SNPs are depicted in Figure 18. The details of the SNPs is as follows: rs12094663 (chr1:46906908; MAF=23.75% in Isle of Wight cohort, 25.9% in Hapmap CEU population)[3, 4] is located 220bp upstream of the *ATPAF1* gene. SNP rs11211337 (chr1:46906398; MAF=27.5% in the Isle of Wight cohort)[2, 3] is a missense mutation that results in a serine to glycine amino acid change, which is predicted to be a benign change)[5]. IOW_4 (chr1:46906346; 238G>T) is a novel rare variant (MAF=1.25%) discovered in the Isle of Wight cohort. It is a missense mutation that results in an arginine (basic amino acid) to leucine (neutral nonpolar) substitution that is predicted to have a deleterious ("potentially damaging" according to Polyphen [5]) effect on the protein. These variants have been mapped onto larger haplotypes already determined to confer risk or protection for asthma [2]. The variants are predicted to be involved in the regulation of *ATPAF1* transcription. We hypothesize that these three regulatory variants modulate expression of *ATPAF1*. Therefore, we designed the following studies to further characterize these variants and their effect on *ATPAF1* gene function in airway-specific cells in order to understand the role of *ATPAF1* in asthma.

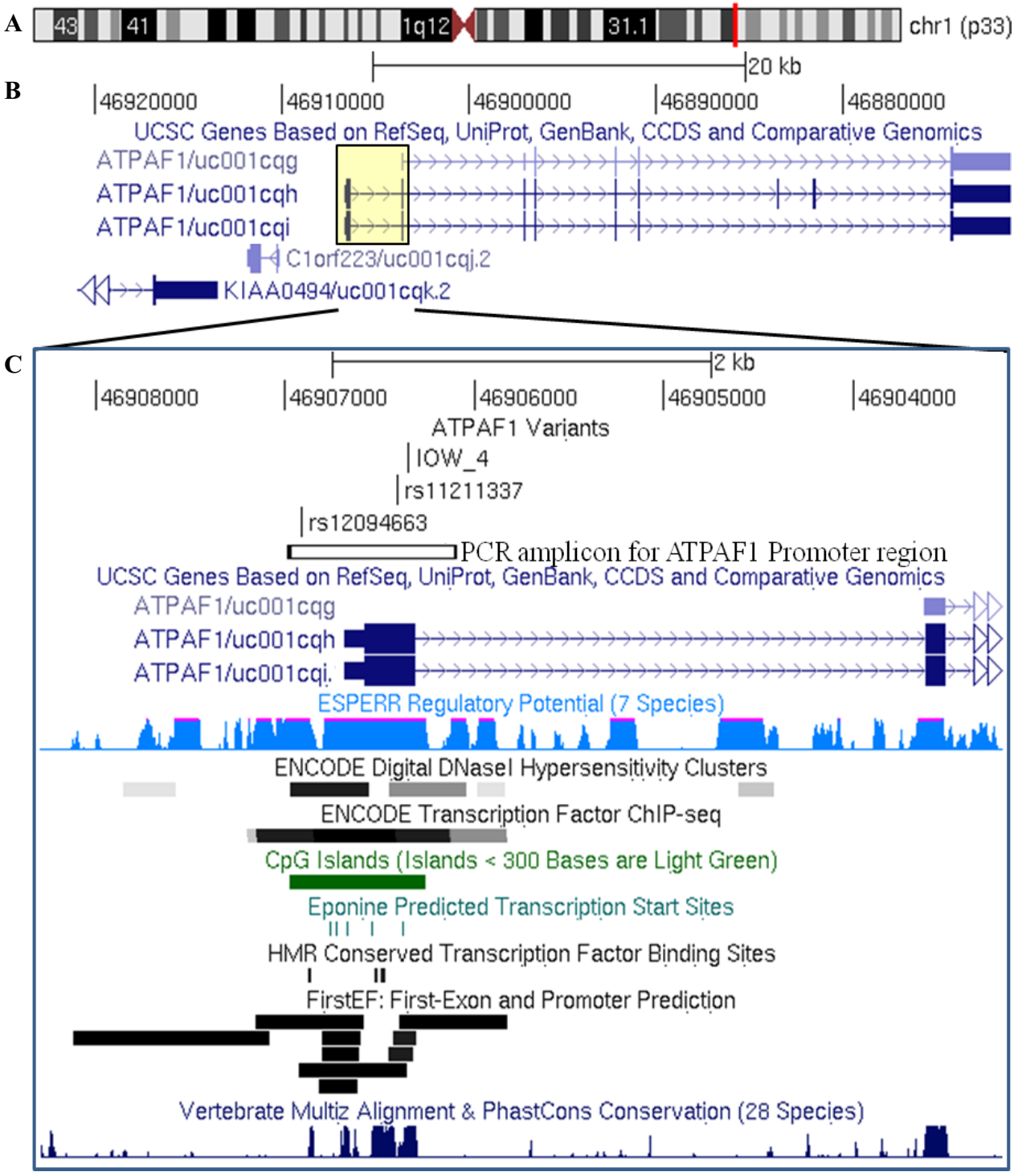

Figure 18: UCSC Genome Browser view of *ATPAF1* promoter region. (A) Chromosome ideogram showing the location of 1p33 and the *ATPAF1* gene (indicated by red dash). (B) *ATPAF1* gene structure and transcript isoforms. (C) Magnified region of *ATPAF1* promoter with location of the three promoter variants. Tracks located below (in order): illustration of the

amplicon that was inserted into luciferase reporter vectors, the exons of the *ATPAF1*. The lower half of the figure contains tracks relevant to promoter/gene expression regulatory activity including locations of CpG islands, transcription start sites, conserved transcription factor binding sites, locations of predicted promoter and exon locations, experimental data validating binding sites for transcription factors in various cell lines (shade indicates strength of signal), experimental data for DNAseI hypersensitivity, predicted regions of gene transcription regulation, and conservation regions [3].

METHODS

Experimental design

African green monkey fibroblasts-like cells (COS-7) were used for their high transfection efficiency. Human bronchial epithelial cells (BEAS-2B) and human type 2 alveolar epithelial cells (A549) were selected for study because of their relevance to asthma. A protocol of IL-4 stimulation of these cells was developed and utilized in an attempt to mimic an asthma-like state, recognizing the limitations of simulating asthma in cell culture. To document that IL-4 stimulation was effective, expression of biomarkers of the TH2 response was measured in cultured cells at several time points from 0 to 48 hours post-stimulation. The expression of *ATPAF1* and neighboring *KIAA0494* and *C1ORF223* genes were also examined.

To explore the gene regulatory functionality of the ATPAF1 SNPs rs11211337, rs12094663, and IOW_4, DNA fragments containing four different haplotypes of these SNPs were cloned upstream of the luciferase reporter gene in an expression vector for use in *in vitro* cell culture experiments. Cell cultures were transfected with the expression vectors and baseline values of luciferase expression of each haplotype were collected, after which the experiments were repeated in the presence of IL-4 stimulation. BEAS-2B, A549, and COS-7 cell lines were used in these studies, although the transfections with A549 cells were unsuccessful on multiple attempts and, therefore, are not presented here.

Expression of ATPAF1 and biomarkers of IL4 stimulation in asthma relevant cell lines

Cell Culture

Human bronchial epithelial cells (BEAS-2B [CRL-9609], American Type Culture Collection, Manassas, VA), human type 2 alveolar epithelial cells (A549 [CCL-185], American Type Culture Collection), and African green monkey fibroblasts-like cells (COS-7 [CRL-1651] American Type Culture Collection) were used for cell culture experiments. BEAS-2B cells were grown on plates precoated with 0.01 mg/mL fibronectin (BD Biosciences, Bedford, MA), 0.03 mg/mL Collagen (Collagen 1 Rat-tail, Invitrogen, Carlsbad, CA), 0.01 mg/mL BSA Factor V (Invitrogen) in BEBM media with BEGM SingleQuots added (Lonza, Walkersville, MD). A549 and COS-7 cells were grown in F-12 media and DMEM media, respectively, supplemented with 10% fetal calf serum (Invitrogen).

Timecourse experiment

A timecourse study of gene expression was conducted in each cell line under baseline conditions and in the presence of IL-4 (10 ng/mL; Recombinant Human IL-4, R&D Systems, Minneapolis, MN) or vehicle (PBS) for controls [6, 7]. The IL-4 stock solution was prepared by dilution from a stock solution with PBS with 0.1% BSA. Cells were seeded in triplicate (biologic replicates) in six-well plates 24 hours before the start of the timecourse at a density of $1x10^4$ cells per well. At time point 0 hour cells were fed with appropriate media (F-12 for A549, DMEM for COS-7, BEBM with no serum for BEAS-2B) containing 10% FBS (Invitrogen) and either IL-4 or vehicle added. Cells were collected using TRIZOL reagent (Invitrogen) at 0, 4, 16, 24, and 48 hours for

A549 and COS-7 cells and at 0, 24, and 48 hours for BEAS-2B cells. Total RNA was purified by TRIZOL protocol, DNAse treated according to manufacturer protocol (Qiagen), and a sample of RNA was separated by gel electrophoresis on a TAE gel to determine quality.

Quantitative reverse transcriptase PCR (qRT-PCR)

Reverse transcriptase (Superscript III RNase H$^-$, Invitrogen) and random primers (Applied Biosystems, Foster City, CA) were used to create cDNA from 1-2 μg of total RNA. Gene expression was examined using a Step One Plus quantitative PCR system (Applied Biosystems) using Power Sybr Green PCR Mastermix (Applied Biosystems) and primers (Table 17) designed using Primer3 software implemented in Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Primers were designed to span exon-exon junctions of genes and/or to have multiple hundred base pairs of intronic sequence separating the genomic DNA from the spliced cDNA version. Primers were synthesized (IDT DNA, Coralville, IA) and tested using cDNA, non-reverse transcribed RNA, and no-template controls prior to their use in quantitative PCR.

ATPAF1 expression was examined with primers designed to capture expression of all isoforms and the two main isoforms (*ATPAF1*/uc001cqh, containing all nine exons; *ATPAF1*/uc001cqi, containing all exons minus exons seven and eight)(Figure 18)[3]. The expression of the adjacent genes *KIAA0494* and *C1ORF223* was also measured. Eotaxin 2 [*CCL24*], eotaxin 3 [*CCL26*], TGF-β2 [*TGFB2*], and PPAR-γ [*PPARG*]) were examined to validate the intended effect of IL-4 on cell lines. The endogenous control reference gene was DNA-directed RNA polymerase II subunit A (*POLR2A*).

Table 17: Primers used for SYBR Green qPCR

Primer name	Gene Name/Description	Primer Direction	Sequence (5'-3')
ATPAF1_All	ATP synthase mitochondrial F1 complex assembly factor 1 (All transcripts)	pF	AAGGACAAGACTCTCAGTTCAATCT
		pR	CTGCAGGAATAACTGCGTAGA
ATPAF1_cqh	ATPAF1 cqh isoform (full length, containing all nine exons)	pF	TTATTCCTGCAGAAAAGTTTGATT
		pR	CCTCGGGTTGTCCTACAAA
ATPAF1_cqi	ATPAF1 cqi isoform (all exons minus exons seven and eight)	pF	GCAGAAGAAATAAAACAGATTTGG
		pR	AGCAACATTTGTTGGACAGG
C1ORF223	Uncharacterized protein C1ORF223	pF	CCTGGGATGTGGGTCTCAT
		pR	CTCTTCCCGCCTCAAGTTCT
CCL26	Eotaxin-3	pF	TGGACCTGGGTGCGAAGCTATGA
		pR	TGGATGGGTACAGACTTTCTTGCCTCT
KIAA0494	Uncharacterized protein KIAA0494	pF	TTCAGCATGATGGGTGATAGA
		pR	CTTCTTTCTTTATGCTTTGGATTTT
POLR2A	DNA-directed RNA polymerase II subunit A	pF	ATCGGAAGCACATGACTGAC
		pR	TTCCTCCTCTTGCATCTTGTT
PPARG	Peroxisome proliferator-activated receptor gamma (PPAR-γ)	pF	TGCCTTGCAGTGGGGATGTCT
		pR	GCGGACTCTGGATTCAGCTGGT
TGFB2	Transforming growth factor, beta 2 (TGF-β_2)	pF	GGAGCGACGAAGAGTACTACGCC
		pR	AGAAAGTGGGCGGGATGGCA

qRT-PCR Analysis

DNA amplification was conducted by PCR using an initial 10 minute 95°C denature and 40 cycles of a two-step cycle consisting of a 15 second, 95°C denature and a 1 minute, 60°C annealing/extension. Template-free controls were included in each plate for each gene. Quantification was performed by the standard curve method with a standard curve determined for each gene on each plate. DNA melting curves for PCR products were examined to ensure specific amplification. All standard curves had a PCR efficiency ranging between 80 to 110% (mean= 95.39%, with the exception of *PPARG* which was 70.8%[BEAS-2B] and 72%[A549]).

Technical replicates (n=2) of each biological replicate (n=3) for each cell type, timepoint, or treatment type were used to decrease error. Each wells' data were normalized to the *POLR2A* gene. Technical replicates for each sample were averaged in order to determine the average Ct for each biological replicate. The biological replicate data were averaged and each gene's data calibrated to the 4 hr PBS (A549 cells) or 24 hr PBS (BEAS-2B cells), as these time points contained the lowest levels of expressed genes.

Statistical analyses and charts for expression were made in GraphPad Prism. Timecourse data were analyzed using a one-way ANOVA with a Newman-Keuls Multiple Comparison Test. A two-tailed, unpaired t test was used for pairwise comparisons. A $P<0.05$ was considered statistically significant.

Examination of ATPAF1 promoter haplotypes

Cloning techniques

Four constructs were prepared for the cloning, one for each of the four promoter haplotypes identified in the Isle of Wight study population. Sequence data for the SNPs rs11211337, rs12094663, and IOW_4 were utilized to select individuals containing the target haplotypes. Two individuals from the 1989-90 Isle of Wight birth cohort study[8] who together carried the four haplotyes of interest (individuals #10 nonasthmatic female, haplotypes A/B; #32 asthmatic male, haplotypes C/D) were selected for use as genomic DNA template. The haplotypes were used as cloned inserts into the pGL3-Basic vector and were constructed from an 896 bp segment containing the haplotypes and flanking regions (Figure 18). Primers were designed using Primer3 software [9] to amplify 896 bp of genomic DNA located 300 bp 5' of the 5'UTR to 227 bp 3' of Exon 1 (chr1:46,906,092-46,906,987 bp). Restriction enzyme digestion sites (BglII and HinDIII) were designed near the 5' end of each primer to enable ligation into the pGL3-Basic vector's reported multiple cloning site. The primers with indicated restriction sites in brackets are as follows: pF-5'-tgtg-[agatct(BglII)]-accttcgatgctgaaggac-3', pR-5'-tgtg-[aagctt(HinDIII)]-cagacaaacgtggaacgag-3'. High fidelity polymerase PFU Turbo (Stratagene, La Jolla, CA) was used in the PCR.

The amplified DNA fragments (haplotype inserts) were cloned into the pGL3-Basic reporter vector (Promega, Madison, WI) upstream of the promoterless firefly luciferase reporter gene using standard cloning techniques [10]. BglII and HinDIII (New England Biolabs, Ipswich, MA) were used to digest the insert and T4 DNA ligase (New England Biolabs) was used to ligate. The vector was transformed into competent subcloning efficiency DH5α *E. coli* (Invitrogen) and

plated on Amp+ LB plates; multiple colonies were inoculated into Amp+ LB cultures. A Wizard Plus miniprep DNA isolation kit (Promega) was used to prepare plasmid DNA. The plasmid constructs were verified using restriction digests followed by agarose gel electrophoresis to ensure proper insertion. The plasmid constructs were further verified and screened for their insert haplotypes by Sanger sequencing using primers from Promega (GL2 and RV3), which sequence across the insert and multiple cloning site. Sequence data were analyzed using Lasergene software (DNASTAR, Madison, WI) by comparison to the expected plasmid plus insert sequence. No variants other than the three target SNPs were found in the insert sequences.

Frozen stocks of DH5α *E. coli* containing the confirmed constructs with the four haplotypes were used to inoculate fresh LB plates, which were used to inoculate starter cultures and finally 100 mL cultures for use in an endotoxin-free DNA isolation (EndoFree Plasmid Maxi Kit; Qiagen, Valencia, CA). The pRL-TK control vector (Promega) containing the *Renilla* luciferase gene with a constitutively active thymidine kinase promoter was prepared as above and verified with restriction digest.

A Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE) was used to determine DNA concentration of the purified stock DNA plasmid preps. DNA dilutions were made with final DNA concentration at 20 ng/μL for pGL3 vectors and 5 ng/μL for pRL-TK vector. The Nanodrop spectrophotometer, which is reported to detect within 2 ng for this range of DNA [11], followed by an agarose gel electrophoresis was used to validate that concentrations were equal.

COS-7 and BEAS-2B cell transfection

For each haplotype-pGL3 construct, the vector was cotransfected using FuGene 6 (Roche Diagnostics, Mannheim Germany) with pRL-TK control vector in the cell line, BEAS-2B and COS-7 cells. For each cell line, the quantity of total Fugene and the ratios of total vector-DNA: Fugene, and pGL3: pRL-TK were optimized in order to minimize reporter cross-talk. The positive control used was the pGL3-promoter vector containing a SV-40 promoter in front of the firefly luciferase. The promoterless pGL3-Basic vector minus an insert was used as a negative control to compare baseline reporter gene activity without a promoter.

Cells were grown according to protocol in cell culture treated 96 half-area well plates (Product #3885, Corning, Lowell, MA). The 96 well plates were seeded at 6E3 cells per well with 40 μL of media (DMEM + 10% FBS for COS-7, BEBM + singlequots added for BEAS-2B). When cells reached approximately 75% confluence, they were cotransfected with the optimized amount of pGL3 and pRL-TK control vector. Multiple transfection (FuGene + DNA) mixes were created for each haplotype-pGL3 vector. The optimized tranfection mixture consisted of a total plasmid DNA mass of 25 ng of vectors with a 12:1 pGL3: pRL-TK plasmid DNA ratio resulting in 23 ng of pGL3 vector + 1.9 ng of pRL-TK vector per well. A Fugene:DNA ratio of 6:1 was determined to be optimum with 0.15 μL Fugene per well. Multiple DNA mastermixes for each transfection were created to assess accuracy of pipetting. Each DNA mastermix was used to transfect multiple wells with each well defined as a mastermix replicate. Each mastermix replicate had 3 wells that served as biological replicates. The N provided in Figures 22 to 25 are the number of mastermix replicates for each cell stimulation type and/or haplotype. In addition, in the experiments where IL-4 stimulation was compared to PBS, the same transfection reagent-DNA

mixes were used (for example, mastermix 1 was used to transfect cells with IL-4 was later added to half the wells).

A single mix of FuGene was prepared in serum free media (DMEM for COS-7, BEBM for BEAS-2B) and aliquoted to each DNA mix. After 15 minutes incubation, 2.5 uL of the transfection mix (FuGene+DNA) was added to each well and the plates were briefly spun in a centrifuge. At least 3 wells on each plate were left untransfected to serve as background for firefly luciferase and *Renilla* luciferase. For the experiments involving the IL-4 stimulation, 96 full-well plates were used and all quantities and volumes were doubled (except 100 μL of media). The use of full-well plates resulted in a higher level of luciferase activity than in half-well plates for all samples. Immediately before transfection, cells in half of the wells on the plate were fed with media containing 10 ng/mL of IL-4 in PBS containing 0.1% BSA, the other half were fed with PBS + 0.1% BSA only.

Cell lysis/data collection

The cells were incubated for 48 hours with the cell lysis/luciferase data collected according to the Dual-Glo Luciferase Assay System (Promega) protocol. An equal volume (40 μL) of Dual-Glo buffer was added directly to each well to lyse and provide the luciferase substrate. The plates were rocked and incubated at room temperature for 10 minutes before the level of luciferase for each well was quantified using a Veritas (Promega) 96 well plate luminometer. Plates were read at least 3 times at both a 0.5 and a 1.0 second integration times. An equal volume of Stop and Go buffer (included in DualGlo) was added to each well, the plates were rocked and incubated for 10 minutes and plates read in the same manner as above. In the

experiments involving IL-4 stimulation, the media for the cells was completely removed and 50 μL of DualGlo reagent was added. An equal volume of Stop and Glo reagent was added.

Analysis

Data quality control

The data for each well were averaged for the given integration time. The average for the luciferase firefly and *Renilla* firefly background was subtracted from each well's luciferase and *Renilla* luciferase luminescence, respectively. Each well's firefly luciferase activity level was normalized to its *Renilla* firefly luminescence. The resulting ratios for each biologic replicate group (from the same transfection mix) were averaged resulting in the transfection replicate average. Outliers for each group were identified by residual plot and removed from analysis. The mean and standard error for each cell type, stimulation condition and/or haplotype was determined from the transfection replicates.

Statistical testing

Data points for every mastermix replicate were charted in a grouped scatter plot using GraphPad Prism 5 (GraphPad Software, La Jolla, CA). The mean and standard error of the mean were calculated for each group. Differences in firefly luciferase/*Renilla* luciferase activities among haplotypes, between treatments, and any interactions were identified using a 2-way mixed model ANOVA for repeated measures and a Bonferroni Post-test to compare replicates. In addition,

generalized estimating equation (GEE) models [12] were also applied to account for correlations and to estimate average responses of observations sharing the same covariates, such as plate effect. The GEE model with an exchangeable correlation structure was assumed and the identity link function for a normal random variable was specified. The GEE model was fitted with the R library geepack (R programming language v. 2.10.1, open source), which yielded estimates for the haplotype, treatment and plate effects. Differences were estimated between each haplotype group using Haplotype A as referent and between treatment (IL-4) and control (PBS). Statistical significance was accepted at $P<0.05$. Imputation of the promoter haplotypes within the larger context of haplotypes found in the association study [2] were determined using PHASE implemented in PLINK software [13].

RESULTS

Time-course study of ATPAF1 gene expression

The results of *ATPAF1* gene expression in bronchial epithelial tissue in severe asthmatics in Schauberger et al [2] suggested that ATPAF1's mechanistic involvement in the pathogenesis of asthma may be related to expression differences. Therefore, expression of *ATPAF1* and neighboring genes in the chromosome 1p33 region were examined using cell culture techniques involving human bronchial epithelial cells and type 2 alveolar epithelial cells under stimulation to IL-4. In addition, gene expression data for markers of IL-4 stimulation (Eotaxin-3 [*CCL26*], *PPARG*, *TGFB2)* were also examined to confirm stimulation.

A549 cell line

A549 cell line is a Type 2 alveolar epithelial cells commonly used in respiratory research and relevant to asthma. Expression of markers of IL-4 stimulation and of genes in chr1p33 was examined in a timecourse study. The results of the timecourse support the response to IL-4 stimulation (Figure 19, Table 18). The relative expression of Eotaxin 3 (*CCL26*) shows a statistically significant increase in expression in IL-4 stimulated cells throughout the timecourse. This response increased from 4 hours (15.7-fold [fold change different than PBS control at same timepoint], P=0.0035) and stayed constant at 8 hours (14.6-fold, P=0.0002), before increasing at 24 hours (69.4-fold, P=0.0069) and 48 hours (204.3-fold, P=0.0004).

PPARG shows a distinct trend of IL-4 upregulation that began at 4 hours (1.7-fold), was present at 8 hours (1.4-fold), and became statistically different at 24 hours (2.1-fold, P=0.0399) and 48 hours (1.7-fold, P=0.0156).

TGFB2 expression in A549 cells showed a downregulation pattern when cells were stimulated with IL-4. Timepoints at 4 hours (0.483, P=0.0045) and 8 hours (0.547, P=0.0315) were significant while the later timepoints were not significant.

The effect of IL-4 stimulation on *ATPAF1* and adjacent genes was minimal (Figure 20, Table 18). Primer sets designed to capture the expression of different *ATPAF1* transcript isoforms revealed that the overall expression pattern of *ATPAF1* (all transcript isoforms) was consistent over time and IL-4 independent. Specific expression of the full-length *ATPAF1* transcript (*ATPAF1*/uc001cqh, containing all nine exons) revealed no significant changes or trend. The expression of the *ATPAF1* alternative transcript (*ATPAF1*/uc001cqi, containing all exons minus exons 7 and 8) and *C1ORF223* gene were found to be very low and difficult to measure

Table 18: qPCR gene expression data for A549 and BEAS-2B cells

	Markers of IL-4 stimulation									**Gene transcripts in chr1p33 region**								
	TGFB2			*PPARG*			*CCL26*			*ATPAF1* (Full length transcript)			*ATPAF1* (All transcripts)			*KIAA0494*		
	Mean	SEM	Fold Δ	Mean	SEM	Fold Δ	Mean	SEM	Fold Δ	Mean	SEM	Fold Δ	Mean	SEM	Fold Δ	Mean	SEM	Fold Δ
A549																		
0hr	1.12	0.07	-	1.26	0.01	-	N/A	N/A	-	0.59	0.54	-	0.76	0.07	-	0.73	0.07	-
4hr PBS	1	0.07	-	1	0.12	-	1	0.08	-	1	0.2	-	1	0.09	-	1	0.17	-
4hr IL-4	0.48	0.06	0.483	1.67	0.37	1.673	15.7	2.36	15.66	1.07	0.09	1.07	0.66	0.06	0.665	0.4	0.06	0.405
8hr PBS	0.84	0.06	-	1.44	0.05	-	1.14	0.18	-	0.57	0.22	-	0.88	0.19	-	0.85	0.1	-
8hr IL-4	0.46	0.1	0.547	1.95	0.35	1.352	16.8	1.16	14.66	0.63	0.34	1.104	0.79	0.11	0.901	1.04	0.24	1.227
24hr PBS	1.12	0.05	-	1.15	0.05	-	1.37	0.15	-	0.99	0.26	-	0.78	0.07	-	0.87	0.09	-
24hr IL-4	1.01	0.06	0.905	2.45	0.43	2.131	95.2	18.4	69.4	0.89	0.05	0.898	1.14	0.22	1.461	1.22	0.24	1.404
48hr PBS	1.38	0.07	-	1.65	0.15	-	0.46	0.05	-	0.78	0.26	-	0.76	0.1	-	0.9	0.02	-
48hr IL-4	1.14	0.14	0.828	2.85	0.26	1.729	94.3	8.83	204.3	1.01	0.08	1.292	1.18	0.11	1.553	1.03	0.05	1.149
BEAS-2B																		
24hr PBS	1	0.16	-	1	0.02	-	1	0.17	-	1	0.19	-	1	0.1	-	1	0.17	-
24hr IL-4	1.65	0.32	1.653	2.27	0.05	2.267	252	20.5	251.8	1.15	0.09	1.145	1.35	0.14	1.354	1.07	0.15	1.074
48hr PBS	1.17	0.21	-	0.74	0.01	-	1.18	0.87	-	3.65	1.84	-	1.34	0.08	-	1.32	0.31	-
48hr IL-4	2.39	0.52	2.053	1.84	0.03	2.482	426	37.7	361.2	2.71	0.76	0.743	1.7	0.24	1.27	0.95	0.16	0.723

Note: Fold Δ relative to PBS control at same time point

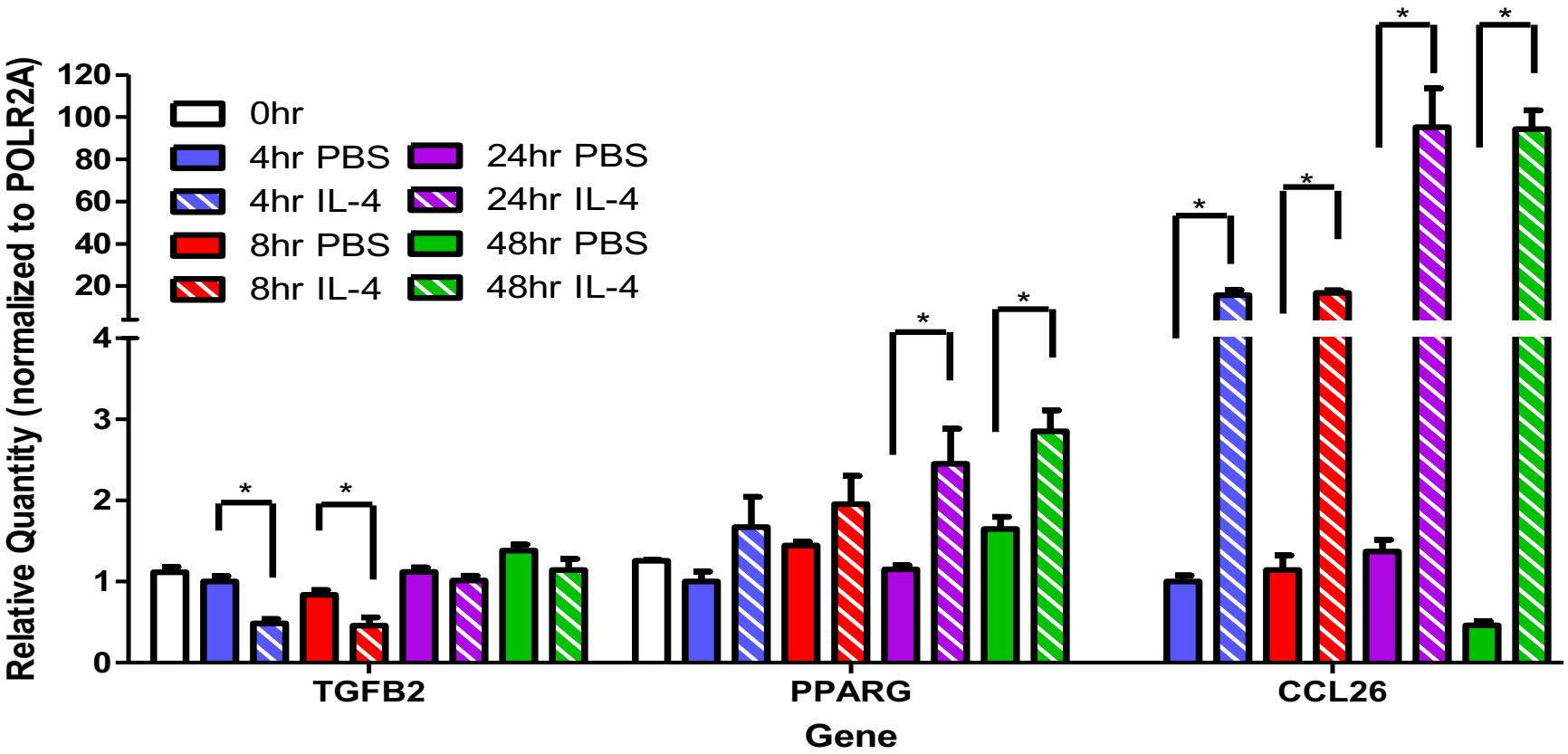

Figure 19: Relative expression of markers of IL-4 stimulation (or PBS) in a timecourse study in A549 cells. The qPCR data have been normalized to POLR2A reference gene and calibrated to the lowest quantity in each gene. Each bar represents the Mean±SEM, N=3. The expression of two other genes in the Chr1p33 locus on *C1ORF223* and *ATPAF1* (alternative transcript missing exons 7 and 8) was found to be very low in A549cells and is not presented. Data for 0hr timepoint for CCL26 is missing *P<0.05 by one-way ANOVA with a Newman-Keuls Multiple Comparison Test. Significant differences between PBS and IL-4 stimulation within timepoints are indicated.

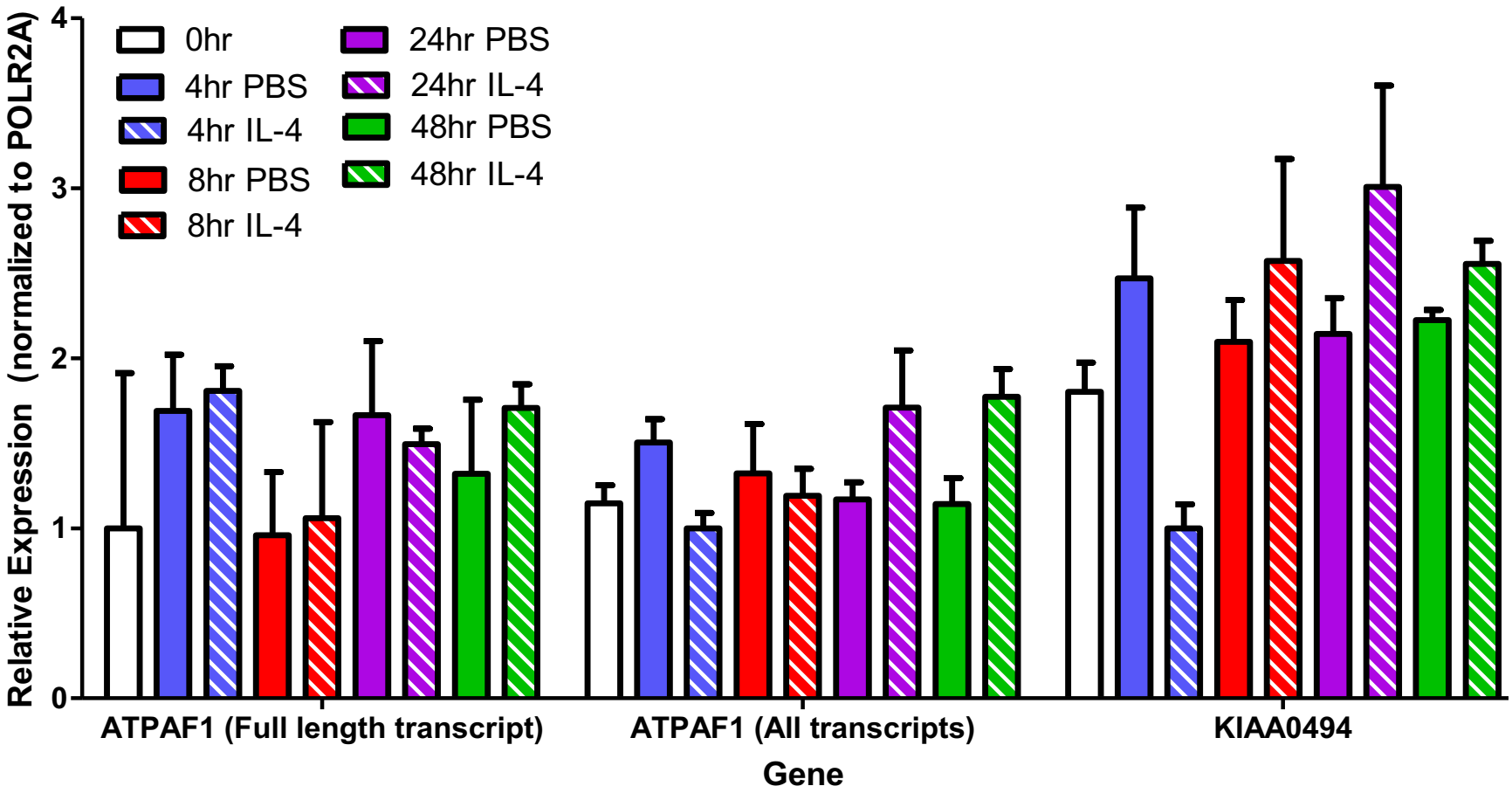

Figure 20: Relative expression of genes in the asthma susceptibility locus on Chr1p33 in an IL-4 stimulation (or PBS) timecourse study in A549 cells. The qPCR data have been normalized to *POLR2A* reference gene and calibrated to the 4hr PBS (=1). Each bar represents the Mean±SEM, N=3. The expression of two other genes in the Chr1p33 locus on *C1ORF223* and *ATPAF1* (alternative transcript missing exons 7 and 8) was found to be very low in A549cells and is not presented.*P<0.05 by one-way ANOVA with a Newman-Keuls Multiple Comparison Test.

consistently by qPCR and are not reported here. The expression of *KIAA0494* was found to be consistent.

BEAS-2B cell line

A second cell line, BEAS-2B, human bronchial epithelial cell line was used to compare the result of the timecourse study done in A549 cell line. The results of the timecourse of BEAS-2B cell line support the response to IL-4 stimulation (Figure 21, Table 18). Two markers of IL-4 stimulation show upregulation of expression activity. The relative expression of Eotaxin 3 (*CCL26*) shows a statistically significant 251.8-fold increase in IL-4 stimulated cells at 24 hours (P= 0.0003) and a 361-fold increase at 48 hours (P=0.0004). The expression of *PPARG* showed a trend toward higher expression at 24 hours (2.3-fold, P= 0.0653) and a statistically significant increase in expression at 48 hours (2.5-fold, P=0.0222) in IL-4 stimulated cells. *TGFB2* exhibited a trend of increase gene expression at both 24 hours (1.7-fold, P=0.1411) and 48 hours (2.1-fold, P=0.0943) that was not statistically significant.

The effect of IL-4 stimulation on *ATPAF1* and the adjacent genes in BEAS-2B was minimal (Figure 22, Table 18). No statistically significant changes were found in these data. The expression of the *ATPAF1* alternative transcript (*ATPAF1*/uc001cqi, containing all exons minus exons seven and eight) and *C1ORF223* gene were found to be very low and difficult to measure consistently by qPCR and are not reported here.

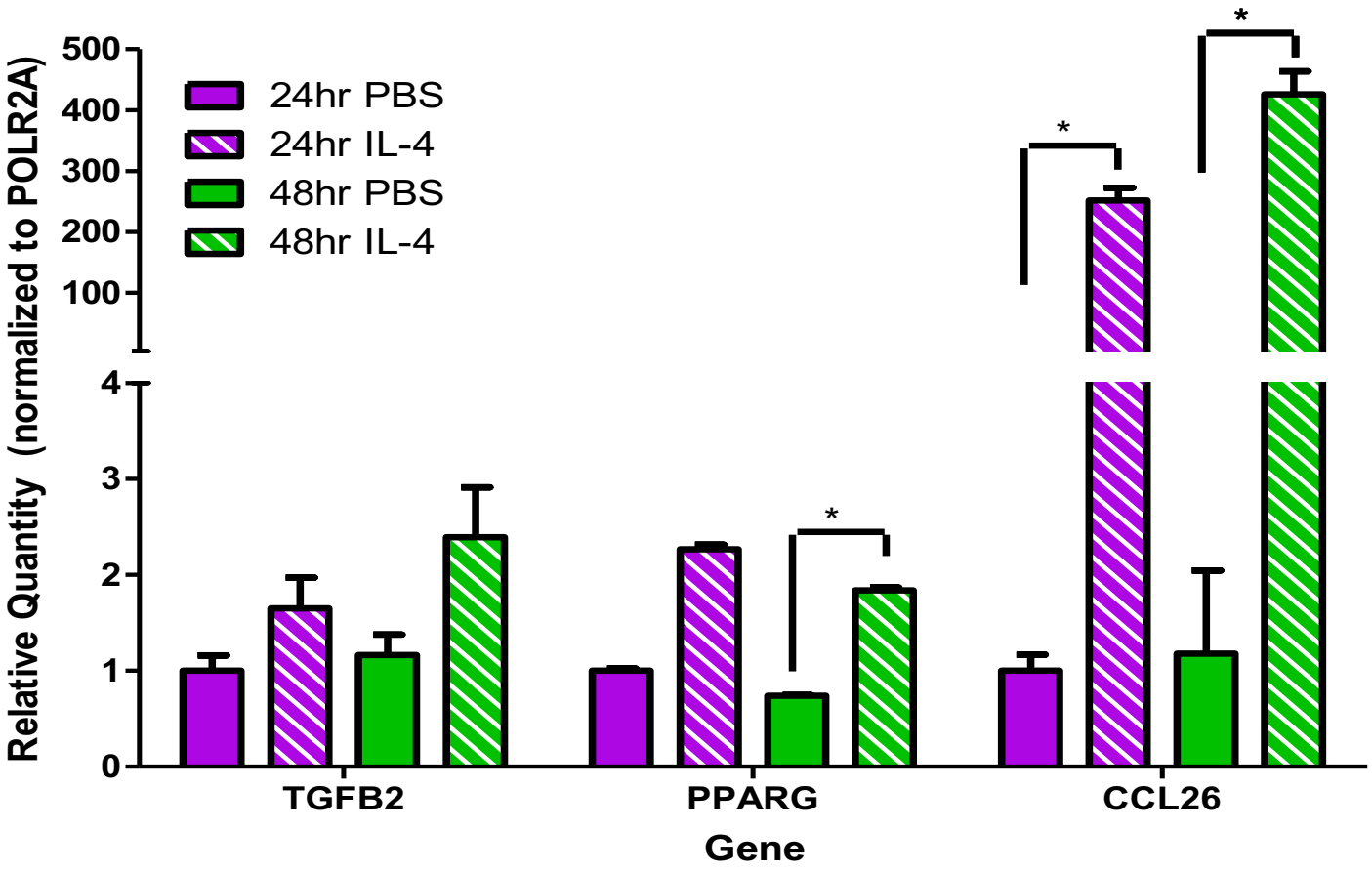

Figure 21: Relative expression of gene markers of IL-4 stimulation (or PBS) in a timecourse study in BEAS-2B cells. The qPCR data have been normalized to POLR2A reference gene and calibrated to the lowest quantity in each gene. Each bar represents the Mean±SEM, N=3. *P<0.05 by one-way ANOVA with a Newman-Keuls Multiple Comparison Test. Significant differences between PBS and IL-4 stimulation within timepoints are indicated.

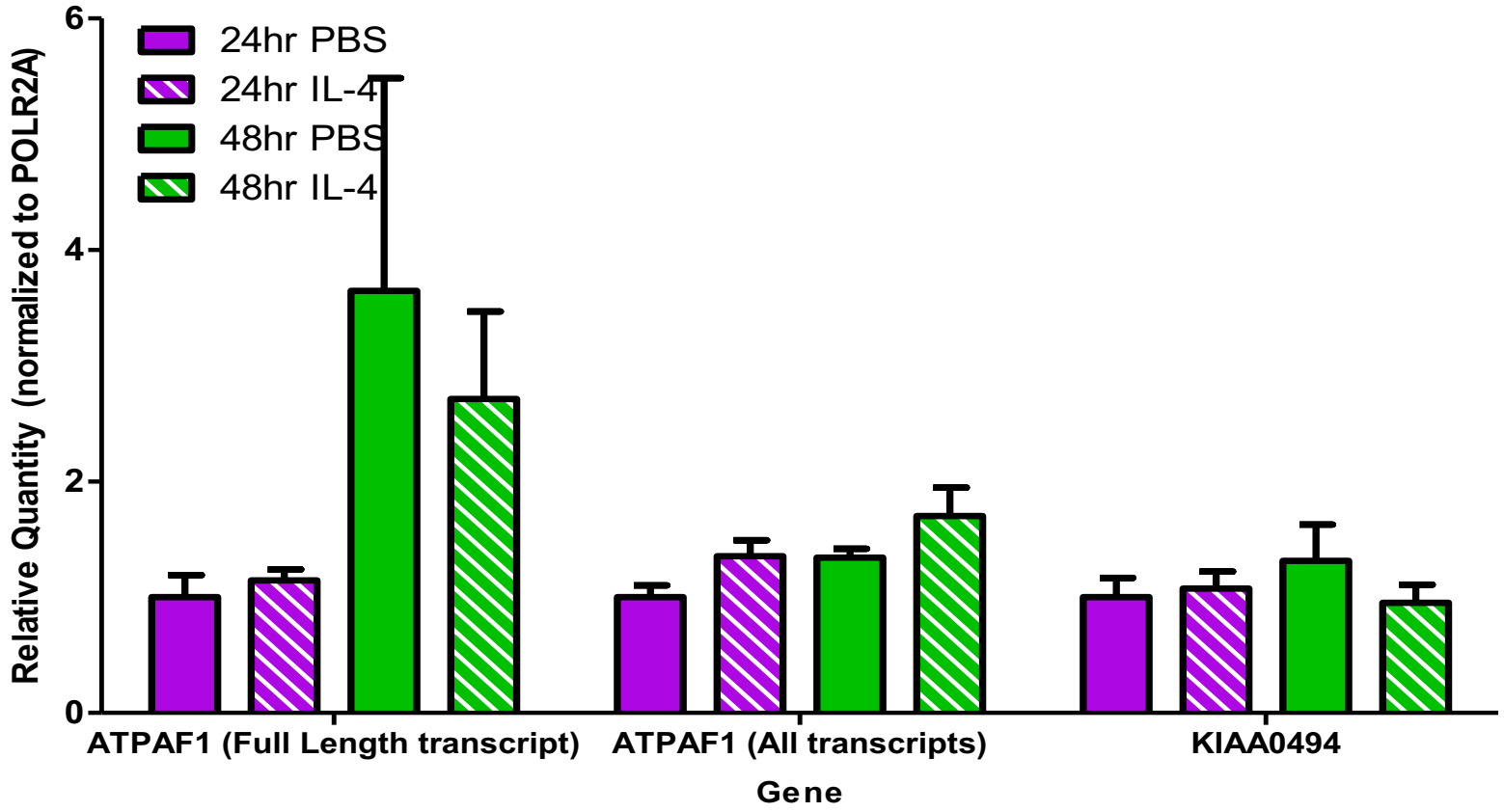

Figure 22: Relative expression of genes in the asthma susceptibility locus on Chr1p33 in an IL-4 stimulation (or PBS) timecourse study in BEAS-2B cells. The qPCR data have been normalized to *POLR2A* reference gene and calibrated to the lowest quantity in each gene. Each bar represents the Mean±SEM, N=3. The expression of two other genes in the Chr1p33 locus on *C1ORF223* and *ATPAF1* (alternative transcript missing exons 7 and 8) was found to be very low in BEAS-2B cells and is not presented.*P<0.05 by one-way ANOVA with a Newman-Keuls Multiple Comparison Test.

Luciferase activity of ATPAF1 promoter haplotypes

Having demonstrated that BEAS-2B cells respond appropriately to stimulations pertinent to asthma, we then decided to examine the effects of promoter haplotype specific expression using a luciferase reporter gene system in cells under IL-4 stimulation and nonstimulated cells. COS-7 cells were also utilized because they are a cell line with a high level of transfection efficiency.

Nonstimulated cells

Results are reported for BEAS-2B and COS-7 cells transfected with one of the four *ATPAF1* promoter haplotypes. Under nonstimulated conditions, BEAS-2B and COS-7 cells show consistent and marked upregulation of the firefly luciferase reporter (relative light units normalized to *Renilla* luciferase) as compared to the pGL3-Basic vector with no insert, indicating promoter activity in all haplotypes (Figure 23).

The relative strengths of the luciferase activity in COS-7 cells were 9.562±0.6880 (Haplotype A, mean±SEM), 6.304±0.6424 (Haplotype B), 7.981±0.7466 (Haplotype C), and 11.10±0.4864 (Haplotype D) (Figure 23). The negative control (pGL3-Basic, no insert) had a relative activity of 0.05054±0.01137 and the positive control; pGL3-Pro had a relative activity of 127±25.88. Haplotype B significantly differed from haplotypes A and D, while haplotype D differed from haplotypes B and C.

Although COS-7 cells show differential expression of ATPAF1 promoter haplotypes, they are a kidney cell line not immediately relevant to asthma. Therefore, further investigations were

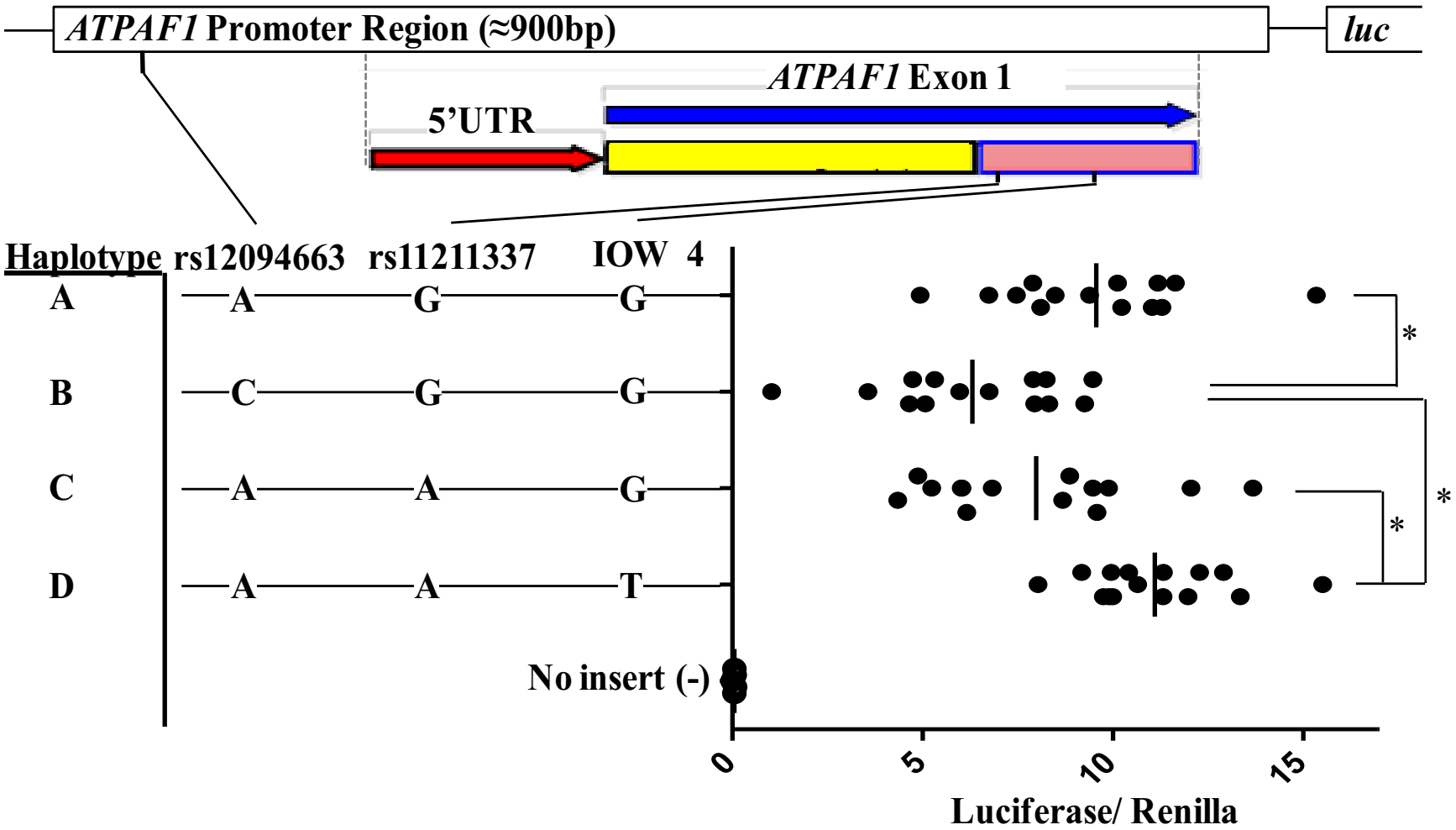

Figure 23: Variable expression of luciferase reporter gene in *ATPAF1* promoter haplotypes in COS-7 cells. *ATPAF1* Promoter-pGL3 construct design (top) showing the location of the ATPAF1-Promoter region insert (≈900bp) relative to the firefly luciferase gene. The insert contains the *ATPAF1* 5'UTR, Exon 1 of *ATPAF1* with the locations of the mitochondrial signal peptide (yellow) and the mature protein (after signal peptide cleavage; pink), flanking 5' noncoding, and 3' *ATPAF1* intron 1. The locations of SNPs rs12094663,

rs11211337, and IOW_4 are indicated. The four different *ATPAF1* Promoter Haplotypes (left), naming convention (A-D), and genotypes at SNPs in haplotypes are indicated. Shown in the lower right panel is luciferase expression under baseline conditions in COS-7 cells from two pooled experiments. N=14 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. *P<0.05 by a Kruskal-Wallis test followed by a Dunn's multiple comparison test.

completed in human bronchial epithelial (BEAS-2B) cell line, a commonly used cell line in respiratory research [14, 15].

The results of the BEAS-2B cells (Figure 24) showed a similar pattern of luciferase activity, but at a lower level. The relative strength of luciferase activity were 2.808±0.2031 (Haplotype A), 2.432±0.1697 (Haplotype B), 3.057±0.2682 (Haplotype C), 4.628±0.3854 (Haplotype D). The positive and negative controls had a luciferase relative activity of 379±20.09 and 0.0733±0.02603, respectively. Haplotype D differed significantly from haplotypes A and B.

IL-4 stimulated cells

The results of the ATPAF1 promoter haplotypes suggested differential reporter expression based on the three-SNP haplotypes. Based on the results showing that IL-4 stimulation can induce several genes relevant to asthma, we further investigated promoter haplotype gene expression when stimulated with IL-4 in both COS-7 and BEAS-2B cell lines.

<u>COS-7 cell line</u>

The COS-7 cells relative luciferase activity is as follows for PBS and IL-4 (Figure 25): Haplotype A (PBS, 45.474±8.901; IL-4, 44.795±4.700), Haplotype B (PBS, 30.620±2.451; IL-4, 31.756±1.477), Haplotype C (PBS, 34.334±2.005; IL-4, 36.514±2.702), Haplotype D (PBS, 66.947±6.694; IL-4, 65.049±2.552), Negative control (PBS, 0.584±0.090; IL-4, 0.440±0.116), and Positive control (PBS, 3705.615±528.922; IL-4, 3048.5398±380.956).

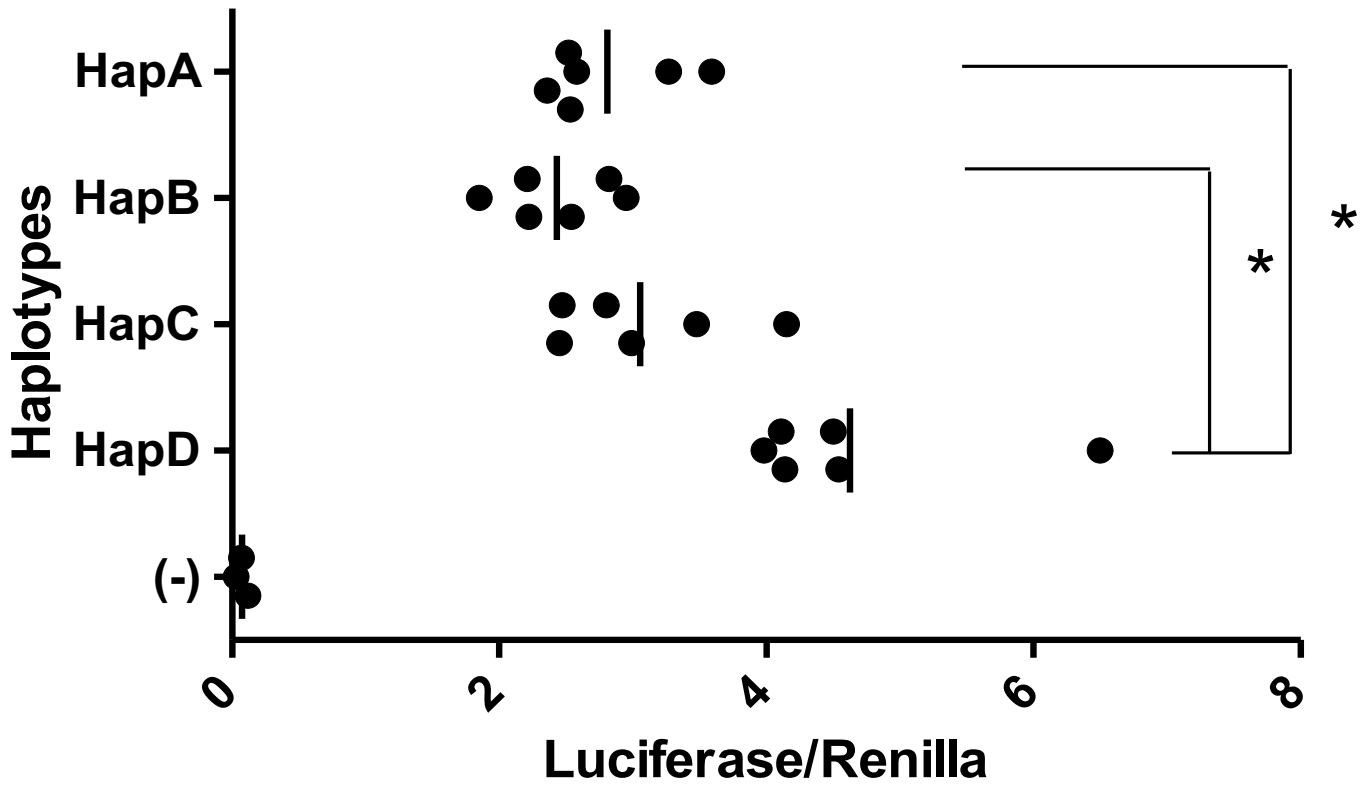

Figure 24: Comparison of relative strength of *ATPAF1* promoter haplotypes using a luciferase reporter assay in BEAS-2B cells under nonstimulation conditions. N=6 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. The mean is indicated with a horizontal line. *P<0.05 by a Kruskal-Wallis test followed by a Dunn's multiple comparison test.

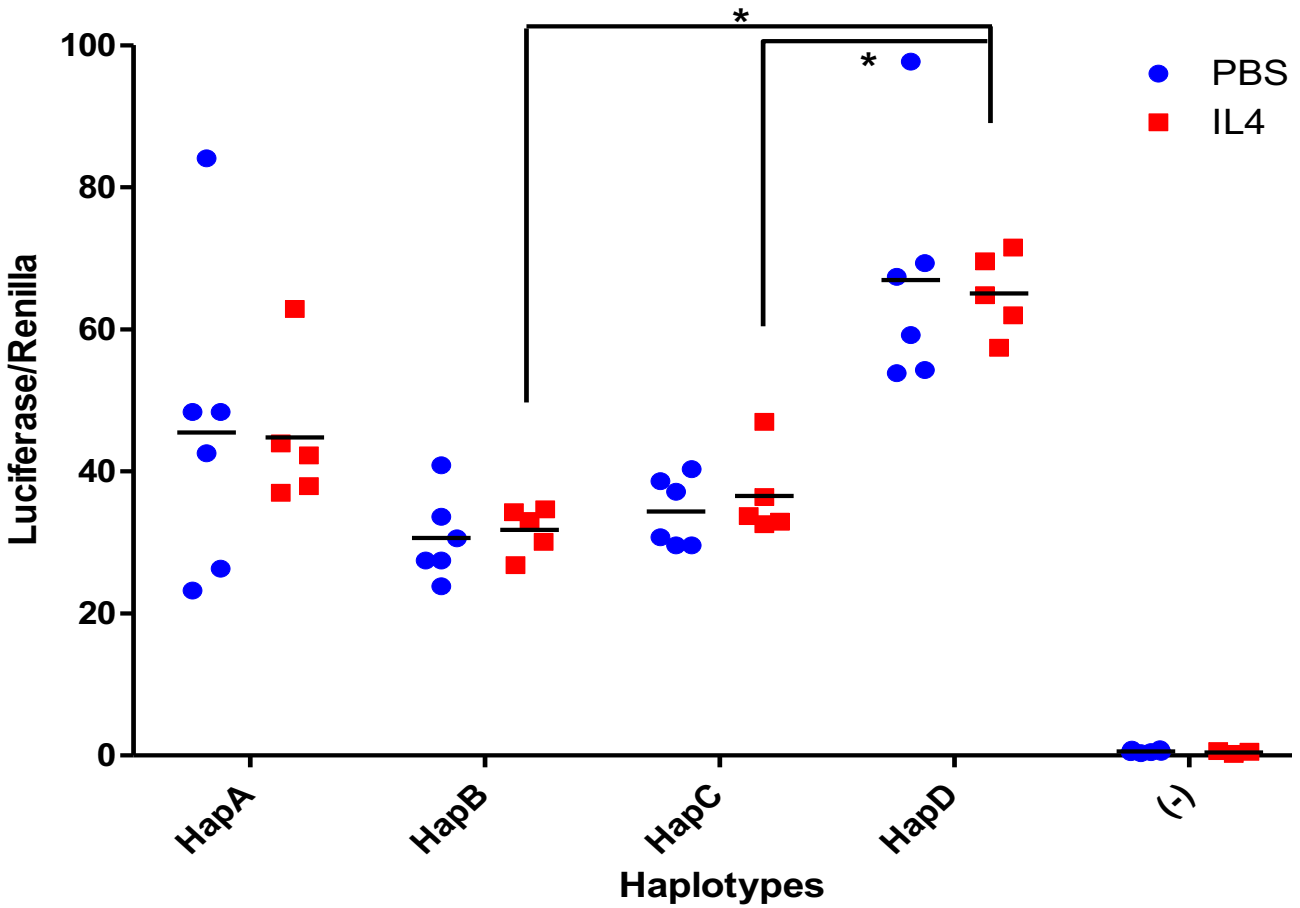

Figure 25: Comparison of relative strength of *ATPAF1* promoter haplotypes using a luciferase reporter assay in COS-7 cells under IL-4 stimulation. N=6 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. The mean is indicated with a horizontal line. A two-way mixed model ANOVA for repeated measures showed significance for IL-4 treatment (P=0.0092) and haplotype (P<0.0001). *P<0.05 by a Kruskal-Wallis test followed by a Dunn's multiple comparison test. No individual pairwise comparison of IL-4 to PBS control in a haplotype was found significant.

A two-way mixed model ANOVA for repeated measures was used to examine the effect of both haplotype and IL-4 stimulation on luciferase activity. Treatment was found to not affect the results whereas haplotype was found to be significant ($P<0.0001$). No interaction between haplotype and treatment was found. A GEE model was also applied to confirm findings and to take into account plate effect. In reference to Haplotype A, Haplotypes B ($P=9.4x10^{-8}$) and C ($P=0.000025$) were found to be significantly less. Treatment with IL-4 resulted in a statistically significant ($P=0.0001$) decrease in expression (on average) as compared to the PBS control (estimated -8.5 units [95%C.I., -14.42 to -2.75]). Plate effect was found to be significant and was taken into account in the GEE model. All IL-4 stimulated data points were on the same plate with half of the PBS controls (the other PBS controls were on a second plate). A Kruskal-Wallis test followed by a Dunn's multiple comparison tests for pairwise comparison was used to examine differences in expression based on haplotype. This analysis revealed that Haplotype B vs D, and C vs D were significantly different ($P<0.05$). No individual pairwise comparison of IL-4 to PBS control in a haplotype was found to be significant.

BEAS-2B cell line

We sought to validate the results of haplotype specific expression of luciferase activity in COS-7 cells in an asthma relevant cell line. BEAS-2B cells were again used for this purpose. As in the COS-7 cells, the cell line was stimulated with IL-4. The BEAS-2B relative luciferase activity was as follows for PBS and IL-4 (Figure 26): Haplotype A (PBS, 8.357±0.876; IL-4, 6.766±0.546), Haplotype B (PBS, 6.571±0.176; IL-4, 5.288±0.192), Haplotype C (PBS, 9.688±0.586; IL-4, 7.948±0.514), Haplotype D (PBS, 15.810±0.632; IL-4, 12.362±1.774),

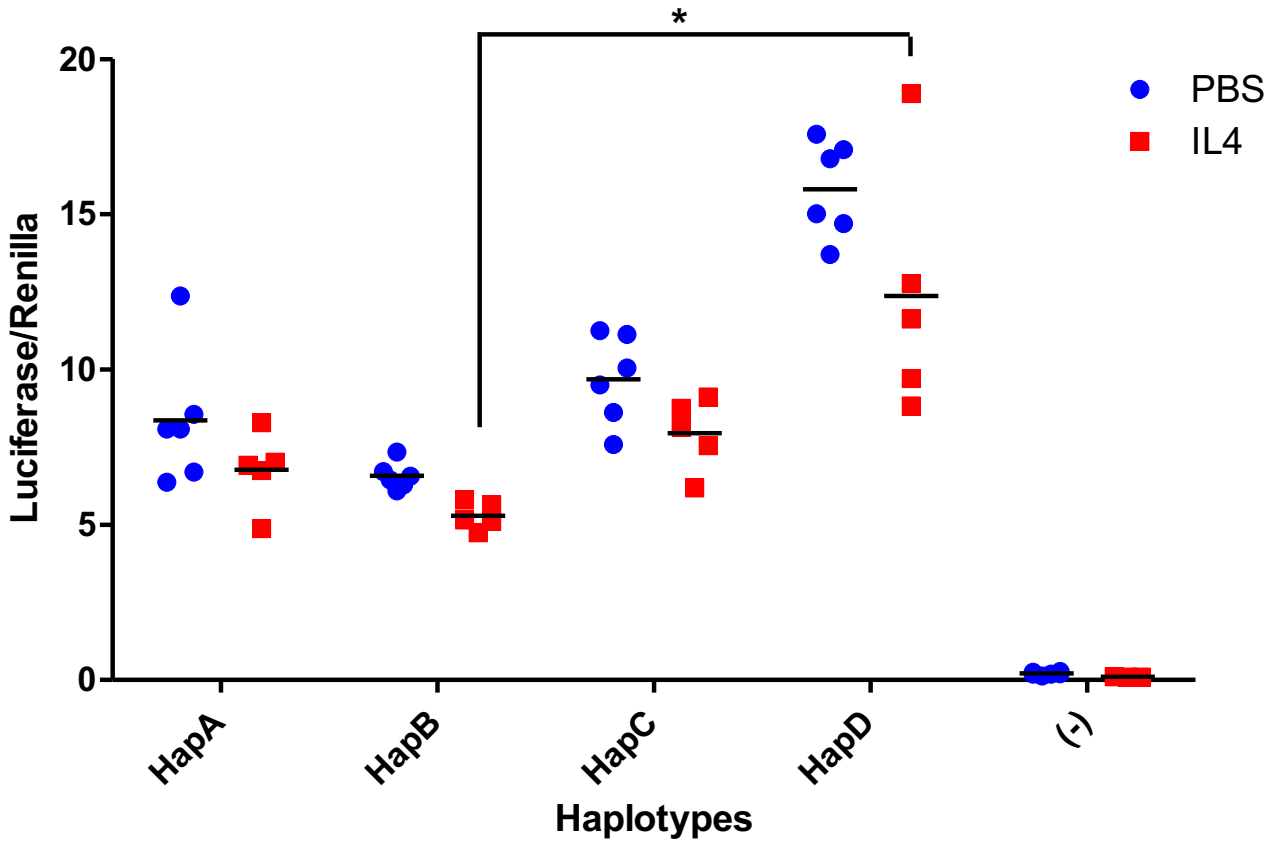

Figure 26: Comparison of relative strength of *ATPAF1* promoter haplotypes using a luciferase reporter assay in BEAS-2B cells under IL-4 stimulation. N=6 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. The mean is indicated with a horizontal line. A two-way mixed model ANOVA for repeated measures showed significant decrease in expression for IL-4 treatment (P=0.0092) and haplotype (P<0.0001). *P<0.01 by a Kruskal-Wallis test followed by a Dunn's multiple comparison test. No individual pairwise comparison of IL-4 to PBS control in a haplotype was found to be significant.

Negative control (PBS, 0.207±0.021; IL-4, 0.098±0.004), and positive control (PBS, 49.446±8.957; IL-4, 577.634±90.344).

A two-way mixed model ANOVA for repeated measures was used to examine the effect of both haplotype and IL-4 stimulation on luciferase activity. Treatment (P=0.0092) and haplotype (P<0.0001) were both found to significantly affect the results on a global scale, however, no interaction between treatment and haplotype was found. The GEE model found that Haplotype A was significantly different from the every other haplotype (Haplotype B, P=0.0017; Haplotype C, P=0.0011; Haplotype D, P=3.1E-21). Treatment with IL-4 was found to result in a statistically significant (P =0.0005) decrease in expression as compared to the PBS control (estimated average -2.01 units [95%C.I., -3.15 to -0.88). The effect of conducting the experiment on two plates was negligible (P=0.9778). A Kruskal-Wallis test followed by a Dunn's multiple comparison tests for pairwise comparison revealed that Haplotype B vs D were statistically significant different (P<0.05). No individual pairwise comparison of IL-4 to PBS control in a haplotype was found to be significant.

DISCUSSION

The present study was undertaken to determine the functional relevance of variants in the *ATPAF1* gene in asthma. Haplotypes containing three SNPs with putative function in the promoter of *ATPAF1* were examined for ability to modulate gene expression using a firefly luciferase reporter gene. There was a significant difference in the expression of the reporter in four *ATPAF1* promoter haplotypes. This trend was consistent across cell lines (COS-7 and BEAS-2B) with the highest expression seen in Haplotype D and A followed by C and B.

Interleukin-4 stimulation resulted in a statistically significant decrease in luciferase reporter activity across all haplotypes in BEAS-2B and COS-7 cells. A parallel study was conducted to explore the effect of IL-4 stimulation on the endogenous expression of the *ATPAF1* gene. Timecourse experiments were conducted in BEAS-2B and A549 cell lines under IL-4 stimulation. No change in *ATPAF1* expression was found despite evidence that cells were indeed stimulated (Figures 18 and 20).

Haplotype A is the common haplotype (frequency = 0.473) and was previously found to confer asthma risk [2]. Haplotypes A and D have a significant increase in luciferase expression when compared to Haplotypes B and C. Haplotypes B and D are the most dissimilar by alleles and by luciferase expression. Haplotype B consistently has the lowest luciferase activity and it is the only haplotype with the rare allele of rs12094663. Several transcription factors are predicted to bind at the locus containing rs12094663, including YY1 (The Yin Yang-1) transcription factor whose site is completely abolished by the minor allele [16].

The Haplotypes A and C differ at SNP rs11211337, but have a similar pattern of luciferase expression. Based on this evidence it appears that rs11211337 by itself does not modulate reporter expression.

Haplotypes C and D differ in only IOW_4. Although IOW_4 is a private mutation found in a single individual, its effect appears to be important in increasing expression from levels seen in Haplotype C to those in Haplotype D. Interestingly, the reporter expression with Haplotype A compared to Haplotype D is relatively less in the BEAS-2B cell line than in the COS-7 cell line, which could reflect a different complement of transcription factors.

Biologically, making sense of how the luciferase expression results may reflect risk for asthma is somewhat difficult. Haplotype A was a part of a larger 10-SNP haplotype that was found to confer a risk for asthma (frequency: 57.36% in asthmatics versus 47.96% in controls, Figure 27). This is consistent with elevated *ATPAF1* gene expression in bronchial biopsies of asthmatics [2]. Haplotypes C and D were found to be subhaplotypes of the previously reported Haplotype II [2]. Haplotype II was found to confer protection in a replication population of Caucasian subjects from Wessex, U.K. but was neutral in the Isle of Wight subjects [2]. Haplotype B is representative of the third most common haplotype (and is a subhaplotype of Haplotype III), which was found to confer protection (24.52% in asthmatics vs 14.36% in controls). Although Haplotype D is a rare subhaplotype (only different at IOW_4), the studies presented here support the effect of a single variant to modulate expression. Haplotype D was found in an asthmatic individual (#32) who otherwise has the statistically neutral haplotype (Haplotype II). It is possible that many other rare variants and subhaplotypes may exist that lead to a similar result.

The luciferase vector construct utilized here was designed to explore the modulation of the *ATPAF1* promoter by SNPs functioning as a haplotype (such as modifying a transcription factor binding site). However, there are several design considerations worth noting. First, the *ATPAF1* promoter haplotypes described here contain three SNPs, two of which are within the first exon and in the coding region of *ATPAF1*. A traditional luciferase vector is designed to explore a promoter upstream of a gene's coding region. This does not preclude the functionality of the variants in impacting *ATPAF1* expression, but does pose difficulty in designing a suitable luciferase expression vector. Second, the *ATPAF1* gene contains a mitochondrial leader sequence, which is used to direct the native ATPAF1 protein to the mitochondria. This leader sequence could impact luciferase production and/or activity, so the vector was designed to

Previous Haplotype Designation	Promoter Haplotype	Variants														IOW Haplotype Frequency
		rs6670495	rs2218189	rs1440486	rs1440487	rs1150064	rs2275380	rs1048380	rs12094663	rs11211337	IOW_4	rs1150068	rs620431	rs2289447	rs1258000	
I	A	T	A	G	C	A	A	C	A	G	G	T	C	C	A	0.473
II	C	T	A	G	T	A	G	C	A	A	G	T	C	C	A	0.211
III	B	A	G	A	C	T	G	T	C	G	G	C	A	T	G	0.199
		T	A	G	C	A	A	C	C	G	G	T	C	C	A	0.024
		T	G	A	C	T	G	T	C	G	G	C	A	T	G	0.021
		T	A	G	T	A	G	C	A	A	G	T	C	C	G	0.018
		T	A	G	C	A	A	C	A	A	G	T	C	C	G	0.014
II	D	T	A	G	T	A	G	C	A	A	T	T	C	C	A	0.011
									Promoter Variants							

Figure 27: *ATPAF1* promoter haplotypes integrated into haplotypes from an association study conducted in the Isle of Wight birth cohort [2]. The haplotype frequencies were based on the combined data set from the resequencing and genotyping. Haplotypes were determined using PHASE implemented in PLINK software [13].

prevent the luciferase production activity from being compromised by the mitochondria leader while also capturing the extent of the promoter activity (which extends through intron 1 of *ATPAF1*).

There are several ramifications of the vector design possibly impacting the transcription and translation of the luciferase reporter gene that need to be considered. First, although it is likely that an ATPAF1-luciferase fused transcript is generated, the possibility of an ATPAF1-luciferase fused protein is not likely due to the presence of multiple stop codons in the several hundred base pairs of intron between the *ATPAF1* exon and luciferase. The preceding *ATPAF1* exon contains a splice donor site, whereas the luciferase does not contain a typical splice acceptor site, thus making normal intron splicing unlikely. However, sequence analysis revealed several atypical splice acceptor sites near the start of the luciferase gene (based on sequence similarity with known human acceptor sites). This includes a potential splice acceptor site with strong sequence similarity to known acceptors 5 bp upstream of the luciferase AUG start site. In addition, there are several other splice acceptor sites in the first 20 bp of the luciferase gene, which could allow splicing of a ATPAF1-luciferase fused protein that may still have luciferase functionality [18]. In the engineering of luciferase vectors fused gene-luciferase gene constructs are occasionally designed intentionally. The resulting luciferase fused protein is often still functional in luciferase activity.

Even if a fused protein is not produced, there are mechanisms which would allow for independent translation of luciferase. In eukaryotes, it is thought that translation usually initiates at the first AUG downstream from the 5' cap, which in the vector construct utilized here, would be at the start of *ATPAF1* and not at the luciferase [17]. Although most eukaryotic proteins are translated by the binding of the ribosome on a 5'cap of an mRNA transcript, alternative

mechanisms to initializing translation do exist through internal ribosomal entry sites (IRES), which allow for the independent translation of the luciferase [19]. In the DNA sequence between the 3' end of the first exon of *ATPAF1* and the 5'end of luciferase (≈160 bp) there is very strong sequence similarity with several viral and cellular IRES sequences [20]). The luciferase gene does possess a Kozak sequence immediately preceding the translation start site, which would increase ribosome binding and translation [17]. If the luciferase translated independently of ATPAF1, this would remove the concern that a fused ATPAF1-luciferase protein with a mitochondrial leader sequence could disrupt proper functioning of the luciferase.

In summary, although there are potential vector design complications, limitations, and concerns regarding the transcription and translation of the luciferase reporter gene, the results appear to be robust and reproducible in two separate cell lines with reporter gene activity significantly above the negative control pGL3 vector without insert in both the IL-4 and PBS controls. The results show that there is differential expression of the luciferase reporter which is dependent on the ATPAF1 promoter haplotype. Further studies, including the use of qPCR with primers specific to the luciferase vector need to be completed in order to better understand the transcription and translation specifics.

A limitation of the interpretation of the study is the inability to accurately compare the results of the experimental luciferase reporter expression to a positive control of activity in the both COS-7 and BEAS-2B cell lines (in data not presented). Both of these cell lines were immortalized using SV-40 and constitutively expresses T-antigen. The T-antigen is known to substantially increase firefly luciferase reporter gene expression in the positive control vector, pGL3-Pro, which contains SV-40 promoter elements upstream of the gene reporter [17]. The result was that the luciferase activity of the positive control was on a scale of orders of magnitude different from the

vectors containing inserts, making direct comparisons between them difficult. Despite this, pGL3-Pro clearly serves as a positive control. Furthermore, appropriate comparisons can be made between the vectors containing inserts and the negative control. Another concern is the plate effect that was seen in COS-7 cells but not in the BEAS-2B cells. The two-way ANOVA did not take this into account whereas the GEE model did. The experiment was designed so that all IL-4 stimulated cells would be on a single plate, whereas the PBS controls were split between the plates. Thus, comparing haplotypes within IL-4 treatment is immune to this concern. This, however, likely is the source of the disagreement between GEE and ANOVA on effect of treatment in the COS-7 cells.

The effects of upregulation of key transcription factors involved in asthma on the *ATPAF1* promoter haplotypes were examined by stimulation of cells with IL-4. The effect of IL-4 stimulation shows an overall global decrease in luciferase expression in all but the positive control (SV-40 containing pGL3-Pro) in BEAS-2B cells whereas no effect was seen in COS-7. The mechanism that leads to suppression of reporter expression is unknown but may be influenced by changing availability of transcription factors or state of the cell. Interleukin-4 stimulation has been shown to upregulate expression and/or activates numerous genes involved in the asthma immunological cascade including transcription factors [6, 7, 21]. It is possible that the cells have either trans-effects from many genes being upregulated at the same time resulting in decreased availability of transcription factors to express or simply different combinations of expressed transcription factors.

In the study here, the expression of *CCL26*, *TGFB2*, *PPARG* were used as markers of IL-4 stimulation. The gene expression results show that IL-4 causes a marked increase in *CCL26* and less of an effect on *PPARG* expression. Other gene expression studies performed examining

eotaxin-2 (*CCL24*) (not presented here) mirror the response seen in *CCL26*. The *CCL26* and *CCL24* results are consistent with data from Atasoy et al. showing IL-4 stimulation resulting in an increase in eotaxin expression due to increased mRNA stability. The *TGFB2* results are inconsistent with what was expected based on previous reports from other authors [6]; however, the raw expression data show that *TGFB2* is already highly expressed (Ct range from 22 to 27) in both A549 and BEAS-2B cell lines and, therefore, it may not be possible to further upregulate expression. Of importance, the reference gene used in the gene expression studies to normalize data, DNA-directed RNA polymerase II subunit A (*POLR2A*), was specifically chosen as it has been shown to be "minimally influenced by stimulation" and thus unlikely to affect results [21].

The timecourse experiments showed that there were no significant changes in expression of *ATPAF1* or the neighboring genes in the LD block, *KIAA0494* and *C1ORF223* in response to IL-4 over time in A549 or BEAS-2B cell lines. Gene expression data specific for ATPAF1 transcript isoforms showed no detectable change in isoform species. Based on the data from the qPCR designed specifically to detect the *ATPAF1* transcript isoform missing exons seven and eight, we conclude this isoform is found in very low quantities and is not upregulated by IL-4 stimulation. The expression of *C1ORF223* was also found to be in very low quantities. Microarray data show this gene is primarily expressed in the testis; therefore, this result was not unexpected [3, 22]. Although the IL-4 stimulation results do not mirror the large increase in expression of *ATPAF1* seen previously in bronchial biopsies of severe asthmatics [2], this can simply be explained by the acknowledgment that asthma is a very complex disease and the cell culture model using IL-4 stimulation may not be sufficient to simulate the milieu of cytokines and long term effects of severe asthma.

In conclusion, variants from the promoter region of the asthma susceptibility gene, *ATPAF1*, were found to modulate levels of luciferase expression in a reporter gene system, thus supporting their role in gene regulation. This expression of luciferase reporter was significantly affected by cell stimulation with IL-4 while parallel studies examining the role of IL-4 stimulation on endogenous gene expression of *ATPAF1* found IL-4 stimulation to have no effect.

Finally, the results presented supply functional support for the *ATPAF1* promoter region variants (rs12094663, rs11211337, IOW_4) involvement in modulating *ATPAF1* expression. In addition, these results provide mechanistic evidence for both the involvement of these variants and further support of *ATPAF1*'s role in asthma. Future studies are necessary to validate and further explore the role of these and possibly other *ATPAF1* promoter region variants in gene expression regulation and asthma more directly.

REFERENCES

REFERENCES

1. Yu, W., et al., *Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations.* Bioinformatics. **26**(1): p. 145-146.

2. Schauberger, E.M., et al., *Identification of ATPAF1 as a novel candidate gene for asthma in children.* J Allergy Clin Immunol, 2011.

3. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

4. Gibbs, R.A., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.

5. PolyPhen. 2008; Available from: http://genetics.bwh.harvard.edu/pph/.

6. Kikuchi, T., et al., *Differentiation-dependent responsiveness of bronchial epithelial cells to IL-4/13 stimulation.* Am J Physiol Lung Cell Mol Physiol, 2004. **287**(1): p. L119-26.

7. van Wetering, S., et al., *Epithelial differentiation is a determinant in the production of eotaxin-2 and -3 by bronchial epithelial cells in response to IL-4 and IL-13.* Mol Immunol, 2007. **44**(5): p. 803-11.

8. Kurukulaaratchy, R.J., et al., *Characterization of wheezing phenotypes in the first 10 years of life.* Clinical and Experimental Allergy, 2003. **33**(5): p. 573-578.

9. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

10. Sambrook, J., E.F. Fritsch, and T. Maniatis, *Molecular cloning : a laboratory manual, Chapter 1 (Plasmid Vectors)*. 2nd ed. Vol. 1. 1989, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

11. Thermo Scientific. *NanoDrop 1000 Spectrophotometer V3.7 User's Manual*. 2008 6-17-2010]; Available from: http://www.nanodrop.com/Library/nd-1000-v3.7-users-manual-8.5x11.pdf.

12. Liang, K.-Y. and S.L. Zeger, *Longitudinal data analysis using generalized linear models.* Biometrika, 1986. **73**(1): p. 13-22.

13. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

14. Yamamoto, S., et al., *Upregulation of interleukin-4 receptor by interferon-gamma: enhanced interleukin-4-induced eotaxin-3 production in airway epithelium.* Am J Respir Cell Mol Biol, 2004. **31**(4): p. 456-62.

15. Matsukura, S., et al., *Interleukin-13 upregulates eotaxin expression in airway epithelial cells by a STAT6-dependent mechanism.* Am J Respir Cell Mol Biol, 2001. **24**(6): p. 755-61.

16. Messeguer, X., et al., *PROMO: detection of known transcription regulatory elements using species-tailored searches.* Bioinformatics, 2002. **18**(2): p. 333-4.

17. Promega Corporation, *pGL3 Luciferase Reporter Vector Manual*, 2002: Madison, WI.

18. Desmet, F.O., et al., *Human Splicing Finder: an online bioinformatics tool to predict splicing signals.* Nucleic Acids Research, 2009. **37**(9).

19. Vagner, S., B. Galy, and S. Pyronnet, *Irresistible IRES - Attracting the translation machinery to internal ribosome entry sites.* Embo Reports, 2001. **2**(10): p. 893-898.

20. Mokrejs, M., et al., *IRESite-a tool for the examination of viral and cellular internal ribosome entry sites.* Nucleic Acids Research, 2010. **38**: p. D131-D136.

21. Radonic, A., et al., *Guideline to reference gene selection for quantitative real-time PCR.* Biochem Biophys Res Commun, 2004. **313**(4): p. 856-62.

22. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.

ADDENDUM HISTAMINE DATA

INTRODUCTION

As a comparison to the luciferase gene expression experiments performed in the COS-7 and BEAS-2B cell lines detailed above using IL-4, the effects of histamine on gene expression of luciferase was examined. Histamine is known to play various roles in allergic airway inflammation and is specifically produced by degranulating mast cells that begins a immunologic cascade resulting in many of the symptoms of asthma and allergy [1]. This data is presented here rather than in the main text because the data lacks qPCR data supporting that histamine responded to stimulation as intended; however, there are data in BEAS-2B and other cell lines that have been published data supporting stimulation through the dose and timeframe of stimulatory effect used here [2].

METHODS

The methodology used for the cell culture is as above except the only difference is the use of histamine in place of IL-4. Immediately before transfection, cells in half of the wells on the plate were fed with media containing 10 uM of Histamine PBS containing 0.1% BSA, the other half were fed with PBS + 0.1% BSA only. All other luciferase data collection and analysis methods utilized are as above.

RESULTS

In COS-7 cells, the two-stage ANOVA shows that treatment (P=0.0456) and haplotype statistically affect the luciferase expression (Figure 28). There is no interaction between haplotype and the method of stimulation. The GEE analysis shows that the effect of histamine is statistically significant (P=0.00013) increase over PBS (estimated 29.2 units more [95% CI, 14.25 to 44.26])

In BEAS-2B cells, the two-stage ANOVA analysis supported that treatment (P<0.0001) and haplotype (P<0.0001) are extremely significant for effecting luciferase expression (Figure 29). The GEE analysis shows that the effect of histamine is a statistically significant (P=0.0465) increase over PBS (estimated 1.2 units more [low=0.019, high=2.42]). No interaction was seen between haplotype and method of stimulation.

DISCUSSION

The results of the histamine stimulation of the cell lines support that histamine has a statistically significant effect on luciferase reporter gene expression in both COS-7 and BEAS-2B cells. The results of effect on histamine stimulation as compared PBS control were more profound in the BEAS-2B cell line as compared to COS-7 cell line.

The data also support differential expression of the luciferase reporter gene based on promoter haplotype. This is consistent with results seen with differential expression of *ATPAF1* promoter haplotype presented in Chapter 4 of the dissertation text. Although not statistically analyzed, the

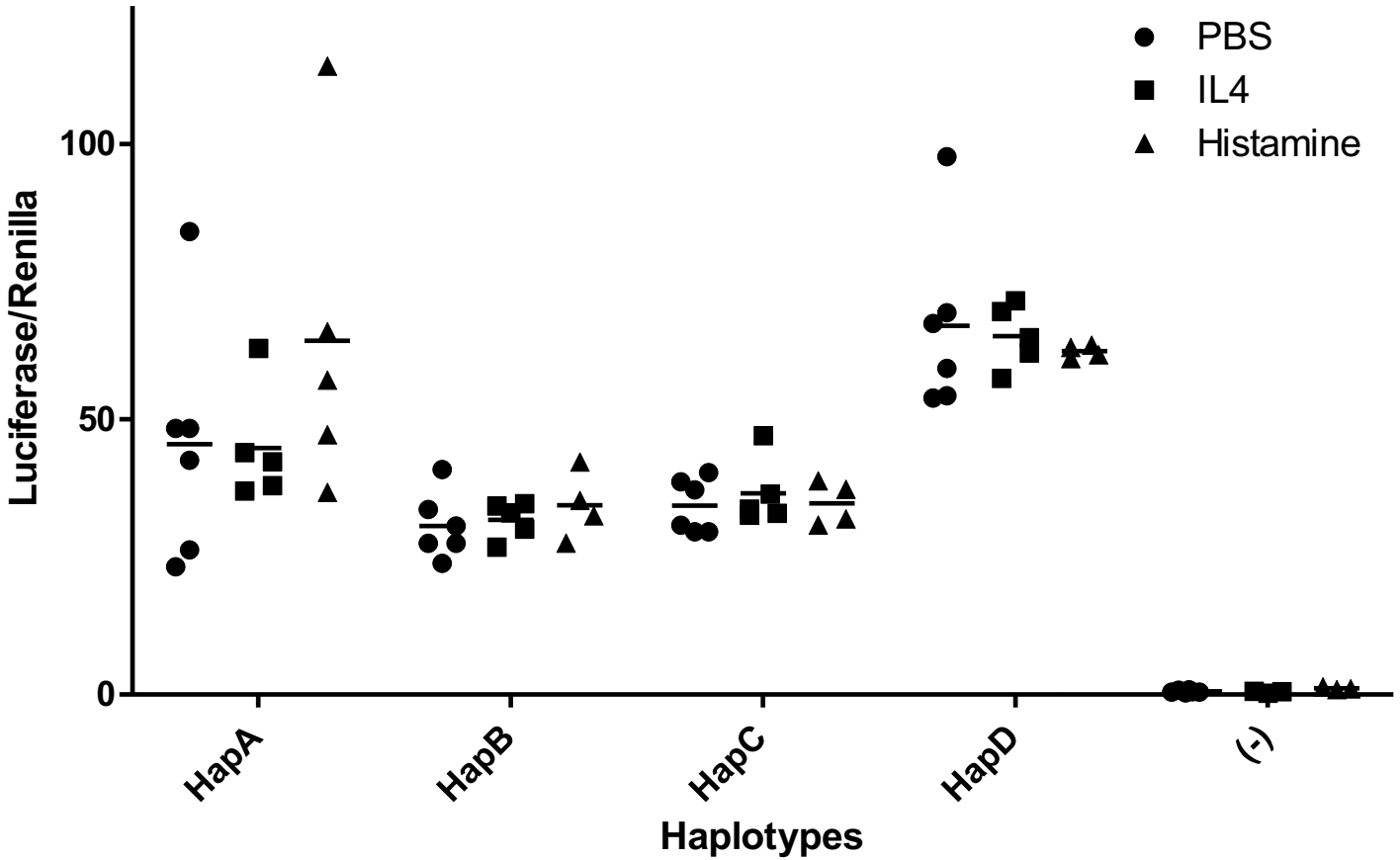

Figure 28: Comparison of relative strength of *ATPAF1* promoter haplotypes using a luciferase reporter assay in COS-7 cells under IL-4 or histamine stimulation. N=5 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. The mean is indicated with a horizontal line.

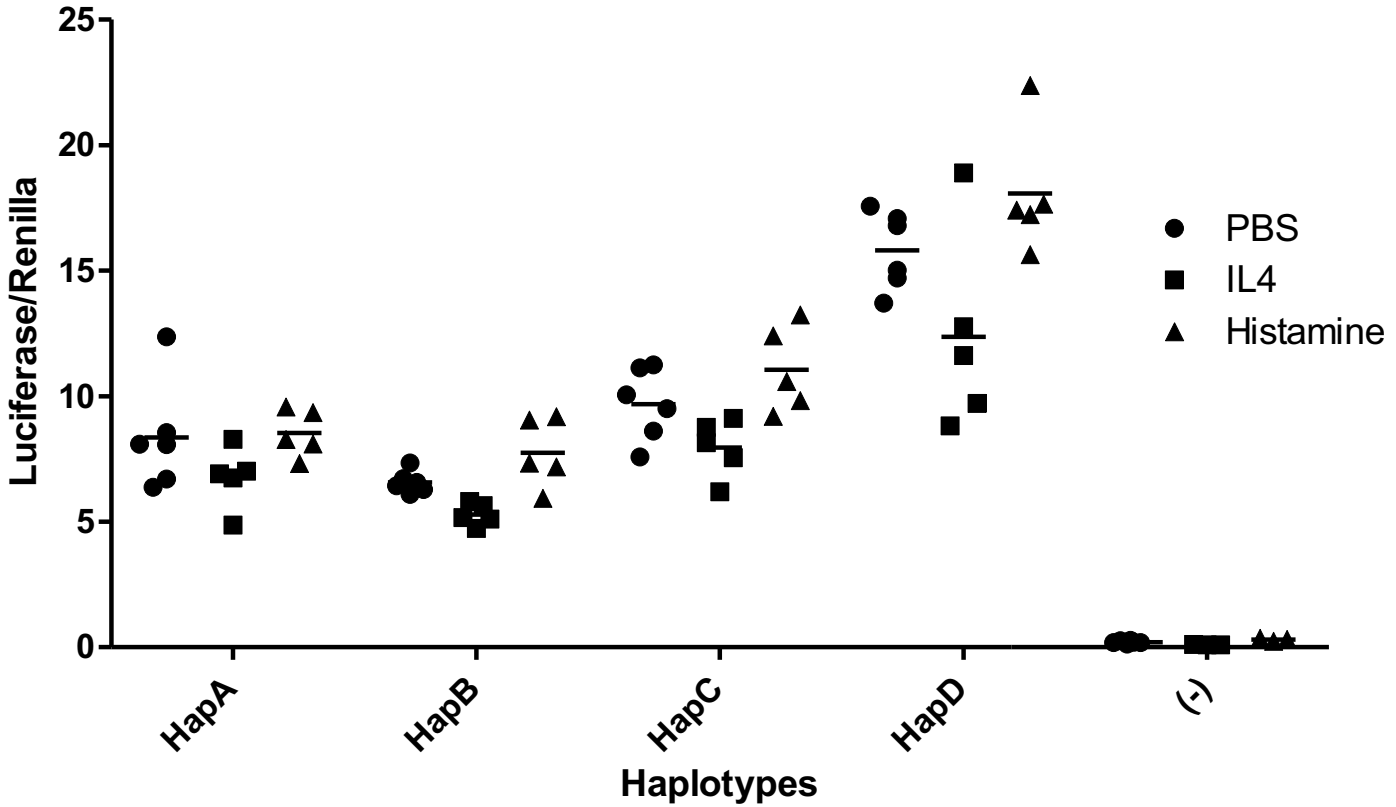

Figure 29: Comparison of relative strength of *ATPAF1* promoter haplotypes using a luciferase reporter assay in BEAS-2B cells under IL-4 or histamine stimulation. N=5 with each dot indicating the average of a mastermix replicate with each being the average of three biological replicates. The mean is indicated with a horizontal line.

results of the IL-4 stimulation are visually different in BEAS-2B and, specifically, in haplotypes C and D.

CONCLUSIONS and FUTURE DIRECTIONS

Although one can only surmise that the histamine stimulation was successful based off of the significant difference in luciferase expression (and all other known variables being equal), these data provide further evidence supporting the functional role of rs12094663, rs11211337, IOW_4 to modulate *ATPAF1* gene expression. These data also support that histamine has a significant effect on reporter gene expression and may play a role in the mechanism in which ATPAF1 is involved in asthma pathogenesis. These results need further experimental data and additional support—specifically, the quantitative PCR data supporting histamine stimulation.

REFERENCES

REFERENCES

1. Barnes, P.J., *Immunology of asthma and chronic obstructive pulmonary disease.* Nat Rev Immunol, 2008. **8**(3): p. 183-92.

2. Takizawa, H., et al., *HISTAMINE ACTIVATES BRONCHIAL EPITHELIAL-CELLS TO RELEASE INFLAMMATORY CYTOKINES IN-VITRO.* International Archives of Allergy and Immunology, 1995. **108**(3): p. 260-267.

CHAPTER 5

Summary, Discussion and Future directions

In less than 5 years, the use of GWAS to identify genes for complex diseases has changed dramatically. For example, the first GWAS (starting around 2005) utilized technology of SNP arrays that could genotype around 100,000 SNPs. Modern GWAS now use arrays that can genotype 1.8 million markers (SNPs plus copy number variant markers). This represents much of what is the rapid advancement in modern genetics. Over the course of this dissertation work, the science and techniques involved have continuously evolved (almost monthly).

The intricacies of the use of GWAS presents many difficulties such as determining the number of case and control individuals needed for sufficient statistical power to detect association in variants with a low MAF and low predicted effect for a disease. Although these considerations should not be minimized, as a future physician scientist interested in translational research, it is my opinion that these concerns seem to pale in comparison to deciding how to proceed after an associated variant and/or gene has been implicated. My point is exhibited by the sheer number of published GWAS (in asthma or other complex genetic diseases) with identified variants but which lack any follow-up work to identify causative variant or even function of the gene. GWAS has the extraordinary benefit of uncovering new genes in pathways not currently known to be involved with disease. However, it appears the great struggle of how to proceed from correlation to causation then begins. Outlining a method of moving from a variant association to some functional significance is the goal of this dissertation. In the studies I have presented, I have detailed a progressive method of identifying functional variants in an asthma susceptibility gene utilizing a multiple-stage genome-wide association study followed by an *in silico* screen, targeted resequencing and functional studies of identified variants.

The GWAS that is detailed in this dissertation was conducted as a case-control from the Isle of Wight birth cohort [1]. This study resulted in the identification of a genomic region on

chromosome 1p33-p32.31 that met genome-wide significance for asthma [2]. Subsequent detailed examination using a combination of targeted genotyping and imputation in the primary population and consortia controls confirmed the association with an LD block containing *ATPAF1, C1ORF223,* and *KIAA0494* genes. Replication studies pursued in five independent populations revealed two instances of strict replication with specific SNPs examined in the Isle of Wight children. Further gene-level association with additional SNPs genotyped using other platforms was found in three of five replication populations. Thus, while not all significant findings in the primary population were replicated, major trends in association were identified across the LD block in all but the two Mexican populations. Data demonstrating differential upregulation of *ATPAF1* expression in asthmatics as compared to control subjects lent further support for a role for *ATPAF1*, a novel asthma susceptibility gene.

An *in silico* functional screen was coupled to targeted resequencing of 40 individuals to narrow the search for functional variants in *ATPAF1*. This resulted in ≈650 reported variants screened for possible role in regulation and/or splice modulation with 27 variants targeted for resequencing for validation. In addition to the targeting of these variants, DNA regions in the *ATPAF1* gene with predicted functional role (exons, conserved, and regulatory regions) were examined by targeted resequencing. Nine of the 27 variants prioritized in the screen were validated and nine novel rare variants were discovered. After the resequencing, variants were evaluated for functional candidate selection. Eleven variants (nine variants that were previously known, two novel variants) were prioritized for functional study in four regions of interest. These variants included two coding-nonsynonymous, seven splice modulating, and eight gene regulatory putative functions.

A cluster of three variants near the promoter of *ATPAF1* (Region of Interest 1) with putative gene regulatory function were examined for the ability to modulate gene expression using a firefly luciferase reporter construct. It was found that there is a significant difference in the strength of the *ATPAF1* promoter haplotypes which is evident by the relative differences in luciferase activity between haplotypes. This trend is consistent across two cell lines (COS-7 and BEAS-2B). In a parallel study, interleukin-4 treatment was used to stimulate asthma relevant genes in the two cell lines. This resulted in statistically significant decrease in luciferase activity across all haplotypes in BEAS-2B but not COS-7 cells. This data provides experimental functional evidence supporting the results of *in silico* study, but more importantly, the ability of the *ATPAF1* promoter variants to modulating the expression of *ATPAF1*. This further corroborates that these variants (and likely others in the promoter region and elsewhere in ATPAF1) may work through gene expression to impact asthma and/or allergy susceptibility.

There are several future directions for this project worth pursuing including follow-up functional studies and expanding the *in silico* and resequencing study. First, I recommend that the additional experiments be conducted to clarify the concerns raised in Chapter 4 regarding the transcription and translation of the firefly luciferase reporter gene. This is outlined in detail below. No formal methods section will be written; instead the methods will be interwoven.

The first recommended study is to investigate if the *ATPAF1* and firefly luciferase gene (*Luc*) are spliced after mRNA transcription prior to translation. Although only the first exon of *ATPAF1* (and several hundred base pairs of flanking intron) was cloned into pGL3 luciferase vector [3], the *Luc* gene (it is engineered as one continuous gene without introns) does not have a splice acceptor site at it 5', making the possibility of a fused transcript less likely; however, the possibility of splicing cannot be ruled out completely and must be examined. Examining the

splicing also will provide necessary information regarding the length of the mRNA transcript between the *ATPAF1* and *Luc* genes.

The second experiments pertain to examining the expression of the *ATPAF1-luc* transcript using qPCR, which then can be compared to data from Chapter 4 to examine whether the relative expression of *Luc* is consistent with luciferase luminescence (and protein translation).

The cell culture of BEAS-2B and COS-7 cells should be cultured with the same methods and under the same conditions described in Chapter 4 methods. For the purposes of detecting splicing of exon 1 of ATPAF1 with luciferase, a full timecourse experiment is not necessary, instead IL-4 stimulated (and nonstimulated) cells should be harvested for RNA and purified using Trizol method (also described in Chapter 4).

Polymerase-chain reaction can be used to examine splicing. Several general points I'd like to point out when considering designing or interpreting data in this section or the qPCR section below. I write these after having to rethink about this again, and again—hopefully this helps:

1. In order to remove amplification of native ATPAF1, one of the two primers MUST be placed on unique sequence with all or some on the pGL3 (either *Luc* or plasmid DNA) sequence.

2. Exon 1 of *ATPAF1* and *Luc* gene are expressed/transcribed together (just as the normal *ATPAF1* gene would be except *Luc* is where exon 2 would normally be located (minus several hundred bp of intron).

3. If splicing occurs between *ATPAF1* and *Luc*, then the *ATPAF1* intron 1-2 will be spliced out (= PCR with primers on intron will fail). If splicing does NOT occur, then the intron

will remain in mRNA and a primer will anneal (= the PCR will amplify). If splicing occurs, the overall length of the amplicon (with at least one primer on a unique, non-endogenous, sequence) would be ideal, whereas if no splicing occurs, the amplicon would be too long for qPCR.

4. Alternatively, primers can be designed so one of primers span *ATPAF1* exon 1- *Luc* junction. If splicing occurs, these primers will be specific for ATPAF1-luciferase splicing. However, if the reaction fails, there is no way to determine if there were technical reasons for the failure or the product (a fused splice product) is not present.

5. pGL3-luc plasmid DNA is in the cells and can be amplified by PCR if DNAse is not used to remove the plasmid DNA prior to RT of mRNA. If splicing occurs, then primer sets with primer on exon junctions will further remove this possibility.

6. All primers must be BLASTed against all human mRNA sequences and should be specific to the mRNA if possible. DNAse can be used to remove the native or plasmid DNA copies.

As mentioned above, the splicing, primers can be designed to anneal right on the probable splice site, or near the ends. The length of the amplicon product can be used to determine whether splicing has occurred. Primers have been designed (Table 19, Figure 30) using Primer BLAST [4] for both primers that anneal on the *ATPAF1* exon and *Luc* respectively (ATPAF1-Luc 2pF and 2pR) and on the exon-exon junction/splice site (ATPAF1-Luc 1pF and 1pR). The advantage of using primer on the exon junction/splicing site include less risk of plasmid DNA contamination and a discernable positive or negative results, but no positive controls are available to test the primers, whereas primers off of the exon junction has the advantage of

Table 19: Primer sequences for exploration of ATPAF1-Luc transcription/translation

Primer	**Sequence (5' to 3')**	**Comment**
ATPAF1-Luc 1pF	CGACCGCTACCGCGACAAGAT	Anneals to *ATPAF1* exon 1
ATPAF1-Luc 1pR	TGGCGCCGGGCCTTTCTTTAT	Anneals to *Luc*. Product 76bp amplicon if splicing occurs, 334bp amplicon product if no splicing occurs.
ATPAF1-Luc 1pF (alternate)	CAAAGCCCTGGCTCGTTCCACG	Anneals to intron between *ATPAF1-Luc*. Used with 1pR, the amplicon product is 105bp.
ATPAF1-Luc 2pF	CTGCGCAGGATGGAAGACGCC	Anneals to *ATPAF1* exon 1-luc exon-exon (if splicing occurs)
ATPAF1-Luc 2pR	AGCCTTATGCAGTTGCTCTCCAGC	Anneals to *Luc*. Product 96bp if splicing occurs, No product if splicing does not occur.

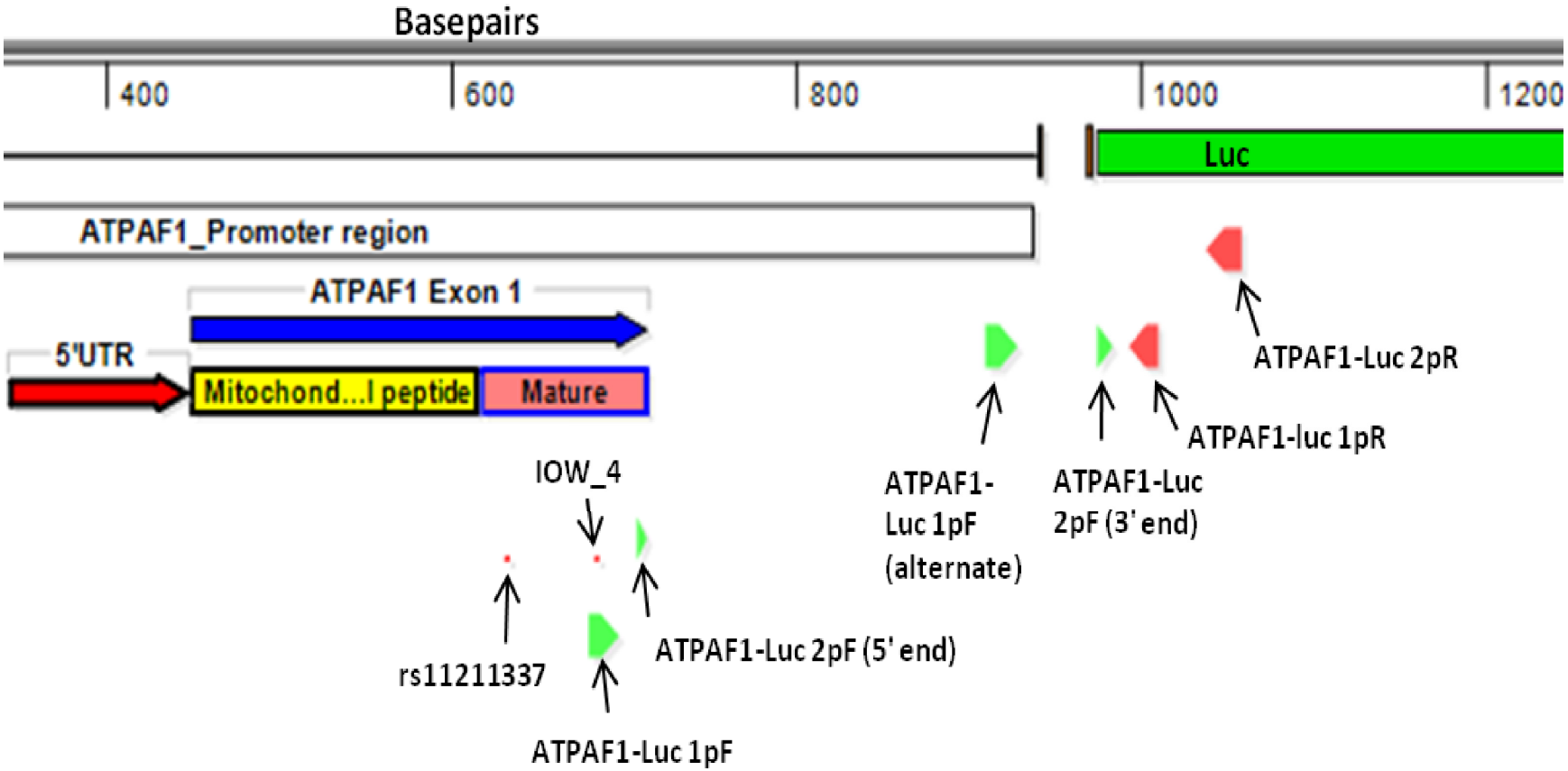

Figure 30: Schematic of the pGL3-ATPAF1 vector at the *ATPAF1* Exon 1 and Luciferase junction. The *ATPAF1* 5'UTR, Exon 1 of *ATPAF1* with the locations of the mitochondrial signal peptide (yellow) and the mature protein (after signal peptide cleavage; pink), luciferase gene (Luc; green) are indicated. Single nucleotide polymorphisms and primer annealing locations for primers in Table 18 are illustrated. Note that SNP rs12094663 is outside (5') of this view. Primers are numerically paired.

always giving you a PCR product and a different product length can be used to determine if splicing has occurred. In primer set ATPAF1 1pF and 1pR, if splicing occurs the product would be 76 bp and if no splicing has occurred the product size would be 334 bp. Other primers would need to be designed if splicing occurs at an alternative splice acceptor site located within the luciferase gene.

Next, pending the results of the splicing experiment, a qPCR study could be used to examine the relative quantities of mRNA transcripts produced by ATPAF1 Region of Interest 1 haplotype. If a spliced *ATPAF1-luc* fused transcript does indeed exist, then primers need to be designed differently than if there is no splicing—specifically, a qPCR amplicon needs to be between 100 and 150 bp and a spliced product provides the opportunity to design primers specific to the transcript. The primers ATPAF1-Luc 1pF and 1pR would result in a product that would be too long for qPCR if splicing does NOT occur. An alternative primer (ATPAF1-luc 1pF alternate) primer was designed with a shorter amplicon that can be used for qPCR if no splicing has occurred (the intron would remain). The qPCR methods are similar to those described in Chapter 4. Primer sequences for this experiment have been designed in Table 19.

Although the experiment described above does not directly examine translation of luciferase, qPCR can be used to examine the relative expression of the *ATPAF1-luc* transcript based by haplotype and compare the data to luciferase results reported in Chapter 5. If results are inconsistent, then it is possible that the transcription/translation coupling that is a premise for this experiment is not true and additional experiments and/or alternative methods would need to be explored.

There are several other followup studies worth examining that could examine the role of these variants on transcription factor binding could involve an electrophoretic mobility shift assay in order to validate the transcription factor binding site. It might also be possible to explore the promoter of *ATPAF1* by narrowing the region of activity by performing a series of deletions and site-specific mutagenesis to isolate variants. Furthermore, within this region, the two variants, rs11211337 and IOW_4 have been explored in relevance to their regulatory role as part of the *ATPAF1* promoter but should also be explored as missense mutations for their functional consequence on protein structure and function and cellular function.

The other three regions implicated in the postsequencing functional prioritization (Chapter 3) should be explored for their possible role in alternative splicing and gene regulation. The variants that are reported here should be genotyped in the remaining members of the nested case-control or other replicate populations to confirm their association with asthma. Allelic-specific differential expression of the *ATPAF1* gene should be examined for these variants in asthmatic individuals in an array of tissues—this would be fruitful in narrowing the search to the function and tissue these variants may have a functional role in.

With the continued advancement of technology and *in silico* tools, there are several future additional paths with this project. There will, undoubtedly, be continued advancement of new *in silico* tools and increased availability of expanded variant databases. The data should be remined in the future to explore possible functionality missed in the current study. The resequencing could be expanded in scope including utilizing a larger set of cohort individuals through next generation sequencing and broadening the target range to include both *KIAA0494* and *C1ORF223*.

On a personal note, my PhD training has been one exhilarating experience. I believe my training has been as much in science of genetics as it has been in the art of collaboration. Collaboration is a powerful tool—one that can move painfully slow, but a tool that I believe will continue to allow me to do great things and stay involved in research as I continue into the tedious parts of my clinical training.

REFERENCES

REFERENCES

1. Kurukulaaratchy, R.J., et al., *Characterization of wheezing phenotypes in the first 10 years of life.* Clinical and Experimental Allergy, 2003. **33**(5): p. 573-578.

2. Schauberger, E.M., et al., *Identification of ATPAF1 as a novel candidate gene for asthma in children.* J Allergy Clin Immunol, 2011.

3. Promega Corporation, *pGL3 Luciferase Reporter Vector Manual*, 2002: Madison, WI.

4. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

APPENDIX

APPENDIX

Supplemental resequencing data

Sequence of *ATPAF1* Regions of Interest

The four regions of interest containing variants with strong evidence of function in the ATPAF1 gene are below. The sequence encompassing each variant with the variant and predicted function listed. The exons are indicated by italics/red after a vertical line (|). Nonfunctional variants are listed (no caps). Splicing=Splicing control; Reg=Regulatory potential; missense=missense mutation. Rare variants (MAF<0.05) are indicated.

Region of Interest 1 (ROI1) contains the SNPs rs12094663, rs11211337, IOW_4. This region is in the 5'UTR/Exon 1 and contains variants with splicing control, regulatory potential, and missense mutations.

gcactacagcgagtgccaggggcaagcaccgcgaggac(**A/C;rs12094663,splicing,reg)**tggagatccgctggctttaga
gaacggacttggcgcgcagaggctgctggtcccagagggaggaagggcgttgaatttgggcagcctccctggaggaggcgggccgtg
gcatacacagggcggcgagagcgcgcagtgccggggccagagcggcgcgccccctcctccccgccccgcctcgccgctatcccgcg
cccggaaccttgggcgcgcggaacctc|*ggtgccggagcgcgggcgcggcgcactcggagcgggatggagcgggcgcgcgggcc*
gcgggccgaggccgacgcgcctgaggaggaggaggaggcgggggcggccatggctgctgtggtggtggcggctgcgggtggcgc
gggaccggcggtcctgcaggtggccggtctctaccggggcctgtgcgcggtgcgcagccgcgccctgggcctggggctcgtgtcacc
*cgcgcagctgcgcgtcttcccagtgcgccccggctcgggccggcccgaggggggcgccgac(**A/G ;rs11211337,***
missense,splicing,reg)**gcagcggggtcggggccgaggccgagctccaggccaacccttctacgacc**(G/T; IOW_4,
***rare, missense, splicing, reg)**ctaccgcgacaagatccagctgctgcgcag*|gtgcgggcctgggttggctg

Region of Interest 2 (ROI2) contains rs620431, rs2289446. This region is near Exon 6 and contains variants with regulatory potential, however these variants are near an exon so could also be examined for splicing though no functional data supports this.

ctgtctttctgtcataagagtgatttgaaaatgatgaaacttaaatgactgttttagttgtattataattctagaattcagtcttggtgaattacatgtg

atagtggccttttgtgttttactatgtgtaggttaactttattcttttccaacctacag*gcagaaaagtttgatttgatctggaaccgggctcagtc*

*ctgtccaaca*gtaagtttctggtgatcacgtaagggttgcttatcaatgttgccaaacaagtattttccagcc(**C/A ;rs620431,**

reg)atcacgtcatattccatac(t/c;rs2289447)gtctcacatgttatagtgctctgtttacttaataa

agatgctcacatgttgattgtactaat aagttatagt(**G/T ;rs2289446; reg)**ctctgtttacttaataaagatgctcaca

(c/t;rs620913)gttgattgtactaataagtgacattttcagccttatgaaatcatgtgg

Region of Interest 3 (ROI3): contains rs1745377 and IOW_7. This region is near Exon 8 and contains variants with predicted regulatory potential and splicing control.

accacataccagtcctccaactcac(**T/C rs1745377, reg)**agccacccaaggggctgttaagagccagaatttcctct

agactgtct(**A/G ;IOW_7, rare, splicing, reg)**ttgctttgaaattgcagtttgttgtgatggtaccttattgagcctt

gataagtgttccaccttgcag|*acccgaggggaagctgcagccagccagctgattttatatcactatcctgaacttaaggaagaaaagg*

gcatagtgctgatgactgcagaaatggattccacatttctg|gtaagaattaacttgcaaaactctttggatttacctcatggcactgaaggtat

tactcttctgccaaagctcatgagagagaaaatttcaaagagctacatataccttgtttcttacttattttaaagttgacagtgaagttgtcttgctc

aagatcttccaatttgaagttagtacaga

Region of Interest 4 (ROI4) contains rs61783140, rs4660952, rs41298535. This region is near Exon 9 and contains variants with predicted regulatory potential and splicing control.

tttaccttacag|*aatgttgctgaggcacagtgcat(**C/A ;rs61783140, rare, splicing)**gccaaccaagt tcagctctt(**C/A*

; rs4660952, rare, splicing)** tacgctactgatcggaaagagacctacgggtt(**A/G ;rs41298535, rare, splicing,

***regulatory)**gtggagacctttaacctcagaccaaatgagttcaaatatatgtct*

gtcatcgctgaattggagcaaagcggacttggagcagaactgaaatgtgcccagaaccaaaataagacttagaactgtacaggttgg

cccttcacctagttgactcagccctcgatagtctagagcccacccccctcctcaggaactcaagagctcagcatttataatgagcagttggt

aatgagttgccctatgtgcttgtcgcaa

Table 20: Isle of Wight individuals used for resequencing.

Asthma status	Individual Sample	Haplotypes (using convention in Chapter 2)	Haplotypes		Probability of correct assignment
Control	10	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	67	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	70	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	75	4/1	GCCTCGATGAT	ACCTCAACGAT	0.7809
	302	5/3	GTACTGTCAGT	GTACTGTCAGA	0.5
	392	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	441	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	464	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	472	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	500	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	621	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	669	3/1	GTACTGTCAGA	ACCTCAACGAT	1
	749	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	902	Minor*	GCACTGTCAGA	ACATCAACGAT	1
	941	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	1141	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	1196	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	1354	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	1421	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	1459	2/2	ACCTCGATGAT	ACCTCGATGAT	1
Asthma	12	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	16	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	32	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	37	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	40	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	99	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	211	1/2	ACCTCAACGAT	ACCTCGATGAT	1
	237	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	338	5/1	GTACTGTCAGT	ACCTCAACGAT	1
	408	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	425	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	477	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	549	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	560	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	562	2/2	ACCTCGATGAT	ACCTCGATGAT	1
	727	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	818	1/1	ACCTCAACGAT	ACCTCAACGAT	1
	850	1/1	ACCTCAACGAT	ACCTCAACGAT	1

Table 20 (cont'd)

	960	3/3	GTACTGTCAGA	GTACTGTCAGA	1
	1521	4/1	GCCTCGATGAT	ACCTCAACGAT	0.7809

*Both Haplotypes are <1% in the Isle of Wight cohort

Table 21: Haplotype frequency of subset of Isle of Wight individuals used in resequencing compared to frequency of individuals genotyped for association study (Chapter 2).

	Controls			**Asthmatics**		
Haplotype #	Frequency	%	Overall Population %	Frequency	%	Overall Population %
1	18	0.45	0.48	23	0.58	0.57
2	10	0.25	0.22	9	0.23	0.22
3	8	0.2	0.25	6	0.15	0.14
4	1	0.03		1	0.03	
5	1	0.03		1	0.03	
Unclassified	2			0		
Total # chromosomes	40			40		