Genetically determined variation of respiratory mucins: disease and demography

by Lauren Johnson

A thesis submitted for the Doctor of Philosophy degree at University College London

2010

I, Lauren Johnson confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in my thesis.

Acknowledgements

I would firstly like to thank my supervisor Professor Dallas Swallow, who has given me tireless support throughout my PhD and has more importantly provided me with an incredible start to my research career. I will be eternally greatful to both Dallas and my secondary school science teacher Jakki Ellis, for the role they have played in my scientific education. As a student I really could not have asked for two more inspiring mentors.

I would like to take this opportunity to thank my colleagues at Wolfson House, for their advice and friendship during my time at UCL; Bryony Jones, Laura Horsfall, Andrew Loh, Kate Ingram, Chris Plaster, Adam Powell and Krishna Veeramah. I would also like to thank my collaborators Dr John Holloway, Imran Shah, Dr Karine Rousseau and Dr Nik Maniatis, for their input in my project, and wish to thank Lynne Vinall, Ranji Arasaretnam, Mari-Wyn Burley and Emma Hobson for their laboratory help and data collection. Many thanks to the Melford charitable trust, Annals of Human Genetics and the MRC for providing project funding.

As expected, writing my thesis has not been the easiest or most pleasant time of my life, however my family have made this experience as comfortable for me as possible, and I would especially like to thank my partner's parentsVee and Dave Burroughs who have kindly allowed me to work from their home.

Thank you to my Brothers Luke and Mitchell Johnson. Two of the best men I know, despite all our arguments. Let me not forget my dog Shaquille, who never fails to make me feel better and listens intently to all my problems.

To my partner Lee Burroughs, thank you for that push when my motivation strayed, for your support especially when things got tough and for the much needed laughter. But most of all, thank you for believing in me and for helping me to see just how much I can achieve.

Lastly, I would like to thank my parents, Mark and Pat Johnson. Your support, faith and encouragement have allowed me to find a career I love and that is perhaps one of the greatest gifts. Thank you for teaching me the importance of education and for nurturing whatever I have chosen to do. Your pride in me makes everything worthwhile. I owe so much to you both.

Lauren Johnson
2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aCGH	array comparative genomic hybridisation
ALI	Air Liquid Interface
BALF	Bronchoalveolar Lavage Fluid
bp	base pair
CF	Cystic Fibrosis
CIP	Calf Intestinal Phosphatase
CIP	Confidence Interval
CNP	Copy Number Polymorphism
CNV	Copy Number Variant/Variation
CNVR	Copy Number Variation Region
COPD	Chronic Obstructive Pulmonary Disease
CRE	Creb Response Element
CT	Cytoplasmic Tail
Cys	Cysteine
DGV	Database of Genomic Variants
DPB	Diffuse Panbronchiolitis
DPE	Downstream core Promoter Element
DT	Divergence Threshold
DV	Dependent Variable
ECP	Eosinophil Catonic Protein
EDN	Eosinophil-Derived Neutrotoxin
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EPO	Eosinophil Peroxidase
ETOPD	Exact Test of Population Differentiation
FIP	Familial Interstitial Pneumonia
GalNac	N-acetylgalactosamine
GCD	Goblet Cell Density
GCM	Goblet Cell Metaplasia
GRE	Glucocorticoid Response Element
GT	Glycosyltransferases
GWA	Genome Wide Association
HMMI	Hidden Markov Model
HWE	Hardy Weinberg Equilibrium
IFN-γ	Interferon-γ
Ig	Immunoglobulin
IHGSC	International Human Genome Sequencing Consortium
IL	Interleukin
IL13Rα1	IL13 Receptor α1
IL13Rα2	IL13 Receptor α2
IL1A	Interleukin-1 alpha
IL1B	Interleukin-1 beta
IL1R1	IL1 receptor 1
IL1Ra	Interleukin-1 receptor antagonist
IL1RN	Gene that codes for IL1Ra
IL4Rα	IL4 Receptor α
INR	Initiator Element
IPF	Idiopathic Pulmonary Fibrosis
IRS	Insulin Receptor Substrate
IV	Independent Variable

JAK	Janus Kinase
LD	Linkage Disequilibrium
LDU	Linkage Disequilibrium Unit
MAF	Minor Allele Frequency
MBP	Major Basic Protein
MCMC	Markov chain-Monte Carlo
MLPA	Multiple Ligation Probe Assay
MSA	Multiple Sequence Alignments
mtDNA	mitochondrial DNA
MUC	Mucin
NFκB	Nuclear Factor kappa B
NHBE	Normal Human Bronchial Epithelial cells
OR	Odds Ratio
PCL	Periciliary Liquid Layer
PCR	Polymerase Chain Reaction
PE	Primer Extension
PI3K	Phosphatidylinositol 3 Kinase
PMA	Phorbol 12-myristate 13-acetate
RAO	Recent African Origin
RFLP	Restriction Fragment Length Polymorphism
ROMA	Representational Oligonucleotide Microarray Analysis
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
S	Segregating Sites
SAP	Shrimp Alkaline Phosphatase
SBE	Single Base Extension
SNNPR	Southern Nations Nationalities and Peoples' Regional States
SNP	Single Nucleotide Polymorphism
STAT6	Signal Transducer and Activator of Transcription 6
TACE	Tumour Necrosis-α Converting Enzyme
TFBS	Transcription Factor Binding Site
TGF-α	Transforming Growth Factor alpha
Th1	T helper 1
Th2	T helper 2
TNF	Tumour Necrosis Factor
TSP	Threonine Serine Proline rich
TSS	Transcription Start Site
TYK	Tyrosine kinase
URE	Upstream Regulatory Element
VNTR	Variable Number Tandem Repeat
vWF	Von Willebrand Factor
WHO	World Health Organisation

Abstract

Airway mucus protects and maintains the health of the respiratory tract. Its production is orchestrated by environmental cues, thus inter-individual variation in mucus composition, quantity and rheology is likely to confer differences in disease susceptibility and response, and may also result in environmental specific suitability.

Glycoproteins known as mucins are considered to be the major components of mucus. This project is concerned with the large secreted airway mucins that are encoded by *MUC5AC* and *MUC5B*, with the overall aim being to study their genetic variation in relation to disease and demography.

Using a single base extension genotyping method, this project reports for the first time, significant associations between five dependent allergy related respiratory outcomes, including asthma, and a single nucleotide polymorphism of *MUC5AC* in a European longitudinal cohort. The cause of these associations could not however be refined and therefore further characterisation of the *MUC5AC* gene is essential for understanding the relationship between allergic airways and MUC5AC.

Variants of the *MUC5B* gene have also been explored in relation to asthma. Variation of the *MUC5B* upstream promoter region has been characterised in two asthmatic disease case-control cohorts by Sanger sequencing. Statistically significant associations are reported here between regulatory variants of *MUC5B*, whereby the ‘high’ expressing promoter haplotype is significantly underrepresented in a sample set of severe asthmatic cases as compared to their controls.

To further characterise variation within these genes, the *MUC5B* promoter has also been sequenced in a sample set of eight African populations and the patterns of regulatory diversity have been examined in relation to population differentiation, geographic demarcation and species conservation profiles. We show here for the first time, a statistically significant overrepresentation of the ‘high’ expressing promoter haplotype

in a collection of the Anuak peoples of Ethiopia as compared to four other Ethiopian sample sets of differing Ethnicity.

1 General Introduction

This thesis is about variation within genes encoding highly glycosylated proteins known as mucins, which are expressed at the interface between body and the environment. The main focus of this general introductory section will be on mucin characterisation, assembly and function, with special emphasis on the large secreted respiratory mucins, MUC5AC and MUC5B. The potential role mucins play in inflammatory airway disease will also be explored and mucin gene association studies will be reviewed. Please note that more detailed background information is supplied at the beginning of each chapter.

It is however important to firstly discuss the vast resources available to geneticists and start here by describing the successful completion of the human genome sequence and the current state of knowledge on the frequency of genetic variation.

1.1 The Human Genome

"The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution" Lander 2001- Nature

After a competitive race between the $3 billion government-funded project (international human genome sequencing consortium- IHGSC) and the privately funded Celera consortium project headed by Craig Venter, a rough draft of the human genome was announced in 2000 and published in 2001 (Lander *et al.* 2001; Venter *et al.* 2001). Both drafts were consensus human genome sequences composed of multiple individuals and were 'completed' in 2003. While most of the human genome has largely been read, gaps scatter the sequence and highly repetitive regions such as the centromeres and telomeres remain unsequenced or incomplete. It is very difficult to find official estimates of genome completion, however an unofficial source suggests that 92.3% of the genome has now been sequenced (Genome completion estimates – see web citations on page 206).

The compiled human genome sequence known as the golden path, is publically available in databases such as UCSC Genome Bioinformatics and Ensembl, where the sequence is both extensively annotated with genes and sequence variants, and is supplemented with additional information such as species comparison data. To date the golden path length equates to 3.1 billion nucleotides within which 23,483 protein coding genes are known to exist (Ensembl Assembly and Genebuild – see web citations on page 206).

1.1.1 Variation of the human genome

Genetic variation such as single nucleotide polymorphisms (SNP), copy number variants (CNV) and simple repeat sequences so called microsatellites and minsatellites, are wide-spread throughout the human genome. After publication of the draft sequence, SNP density estimates suggested that one SNP per 1.9Kb or rather 1.42 million SNPs in total, could be found within the human genome (Sachidanandam *et al.* 2001). SNP estimates are ever growing and the actual number of reported SNPs is close to 18 million (Ensembl Assembly and Genebuild – see web citations on page 206), but it should be noted that frequency data are not available for all (some of which may be errors) and a variation can only strictly speaking be classed as a polymorphism if it is present in at least 1% of the tested population.

Copy number variation (CNV) has recently gained considerable interest as a common genetic marker and one comprehensive genome wide CNV study has calculated that as much as 12% of the human genome is affected by CNV (Redon *et al.* 2006).

A considerable fraction of the polymorphism is functional for instance those within the exome that result in amino acid variants or those located within the exon-intron boundaries that cause splice variants. Less is known about the functional significance of most other non-coding variation. However variants within regulatory regions, such as the DNA sequences where transcription factors bind, are likely to alter expression. Due to the degenerate nature of transcription factor binding sites, it is difficult to predict sequence involved in regulation thus making it difficult to establish which non-coding variants are likely to be functional. Experimental procedures are required to determine functional regulatory variants which perhaps explains why the vast majority of genetic disease association studies to date have concentrated on variants that disrupt amino acid sequence rather than the expression levels of the protein.

1.1.2 Blocks of linkage disequilibrium

Genetic markers are said to be in linkage disequilibrium (LD) if they are found together in a test population more often than expected by chance. The extent of LD is variable across the genome and reflects past recombination events. Some sequence regions have undergone more of these past events than other regions and are said to be recombination rich or rather recombination hotspots.

A haplotype is defined as a set of alleles from multiple loci that are inherited together on a single chromosome. It has been suggested that stretches of recombination free haplotype blocks are interspersed with narrow regions known as recombination hotspots where the majority of the recombination events take place (Daly *et al.* 2001; Gabriel *et al.* 2002; Jeffreys *et al.* 2001). The sizes of haplotype blocks vary and one has been shown to extend as long as 804 kb in a European population (Dawson *et al.* 2002).

In general the patterns of LD, or rather the positions of recombination hotspots, appear to be similar for all populations, however the rates of LD decay vary between populations. African populations have been shown to have the highest rate of LD decay which is consistent with an ‘older’ population history and thus supports the out-of-Africa origin of modern day humans (Ke *et al.* 2004). The HapMap project has built a genome-wide map of LD and haplotype blocks for various populations.

1.2 Genetic distance between populations

Extensive genetic differences in allele frequencies are known to exist between geographically distinct populations (Li *et al.* 2008; Novembre *et al.* 2008). These spatial patterns of genetic variation are a result of population specific genetic drift, demographic history, natural selection, and new mutations. In the case of genetic drift, the frequency of a genetic variant can change purely as a result of random sampling within a population. New mutations may arise independently in geographically distinct populations and may remain isolated in the absence of migration and admixture. The demographic history of a population will imprint itself within the genetic diversity of contemporary populations, for example variants may reach a high frequency if a population has undergone a bottleneck event whereby the genetic variants that were available in the ancestral population are reduced to a subset of variants. Lastly, natural selection is likely to play an important role in population differentiation since geographically distinct inhabitants will be subjected to different environmental conditions which will each favour particular heritable phenotypes. If possessing these heritable traits increases an individual's chance of survival and ability to reproduce, then the trait is more likely to be passed onto the next generation and is said to be more successful.

1.3 Genetics and disease

As discussed previously (page 14) variation within the human genome is extensive. Most genetic variants result in no detrimental outcome, however some variants do cause disorders. Monogenic diseases such as Huntingtons chorea and sickle cell anaemia are caused by mutations within single genes, Huntintin (*HTT*) (OMIM entry #143100) and β-globin (*HBB*) (OMIM entry #603903) genes respectively. There are many diseases of complex etiology that are in part a result of genetic variation. Such complex diseases can be caused by variants within multiple genes which confer susceptibility when in combination with each other and/or specific environmental factors. Examples of such complex disorders include, diabetes (OMIM entry #125853), and asthma (OMIM entry #600807). For OMIM reference please see web citations on page 206.

Various study approaches can be used to identify genetic variants that either cause or increase an individual’s susceptibility to disease. Family based linkage studies use disease pedigree genotype data to identify markers or rather areas of the genome that are linked with the disease in all affected indivduals. Population based association studies look to identify genetic variants that are significantly more common in a disease sample set than in a population sample of healthy controls.

Many genetic disease studies seek to identify the causal or susceptibility genetic variants by methods which exploit LD within the genome. Markers in LD with disease causing alleles will pick up signals of association and help to refine the true causal variant. This means that in principle the whole genome can be scanned by typing a small number of variants since information from one genetic marker will be representative for all other markers it is in LD with.

1.3.1 Association studies

Population based association studies can either include case-control or longitudinal cohort sample sets. A case-control association study consists of approximately even numbers of DNA samples from individuals with a particular disorder and samples from healthy controls. The best case-control study involves a matched data set, whereby the disease samples are matched with a corresponding control sample of the same sex and of similar age, ancestry etc. Case-control samples are desirable since they are relatively easy to obtain and allow the researcher to choose the specific disease they are interested in. However disease description is often an issue and can lead to phenotypic heterogeneity. While dataset sizes are also a continuing angst, independent association studies involving the same disease can be accumulated in a meta analysis.

Longitudinal cohorts consist of individuals often recruited from birth and followed for a lengthy period of time. Studies involving longitudinal cohorts benefit from hindsight and thus medical history and lifestyle information are often available. This allows the researcher to combine genetic and non-genetic data in tests of interaction and permits confounder adjustments to be included in tests of association.

Until recently the candidate gene approach was generally used in most genetic studies. This type of study is hypothesis driven and selects genes on the basis of the function of their proteins in health and disease. In general allelic variants that are thought to be functional, such as amino acid altering or regulatory variants, are selected for genotyping. Non functional genetic markers are nevertheless useful since they will highlight regions of association as a consequence of LD. This has been the basis of the SNP tagging approach whereby only a minimal number of SNPs are needed to obtain the maximum information (Chapman *et al.* 2003).

Recent association studies often utilise genetic data that are representative of the whole genome, termed as genome wide association (GWA) studies. With the invention of high throughput technologies such as microarrays, GWA studies have become increasingly accessible in genetic research. By scanning the whole genome, regions significantly associated with disease can be identified with no prior hypotheses. However GWA studies suffer notably from multiple testing since genotyping platforms are now able to test as many as one million variants simultaneously, and this figure is forever growing. This increases the chance of obtaining false positive results and therefore more stringent significance levels are applied which in themselves increase the risk of false negatives. The low power of GWA studies mean only very strong association signals will be correctly identified which is likely to be problematic for conditions involving multiple genes even if their effect size is large.

1.4 Mucus

Mucus is produced at the epithelial surfaces and acts as a primary innate defence mechanism. It forms a protective barrier over the epithelial surface preventing cell desiccation and harm from environmental insults. Mucus is a complex mixture of water, inorganic salts, various immunological proteins and high molecular weight glycoproteins known as mucins.

Epithelial surfaces are in direct contact with the environment, and the epithelial cells produce a plethora of defence proteins. For instance lactoferrin, lysozyme and defensins are produced with the intention to kill bacteria and individually do so in relatively specific manners (Boyton and Openshaw 2002; Fokkens and Scheeren 2000). The mucus itself acts more crudely by entrapping contaminants and thus preventing the offenders from permeating through to the underlying cells.

Mucins are the major component of mucus, and are responsible for its viscoelastic rheology. The biophysical properties of mucus vary depending upon location within the body. For example, the epithelia of the gastrointestinal tract must be protected from gastric secretions and the mucus is viscous and adherent (Allen *et al.* 1993). However the respiratory mucus is required to be 'sticky' and mobile in order to capture and remove inhaled noxious agents via mucociliary transport (Thornton *et al* 2008). Location specific properties of the mucus are the result of the differential expression patterns of a diverse range of mucins with varied characteristics.

1.4.1 Mucins

Mucins share the common feature of a region rich in threonine and serine, which is heavily glycosylated in the mature protein, and is known as the mucin domain. At present 18 different genes have been given the mucin symbol *MUC* (Hugo Gene Nomenclature Commitee – see web citations on page 206) and in most of these genes the sequence that encodes the serine/threonine rich region is tandemly repetitive. Although this region is extensively glycosylated in all mucins, the repeat units and their organisation differ from mucin to mucin. The carbohydrates attached to the mucin domain usually make up more than 70% (Thornton *et al.* 2008) of the total protein mass and they aid in the lubrication and hydration of the epithelium. The carbohydrates have also been shown to directly interact with bacteria and thus prevent these pathogens from entering the epithelial cells (Linden *et al.* 2002; Van de Bovenkamp *et al.* 2003).

1.4.1.1 Mucin glycosylation

Generally the mucin carbohydrate structures are a result of O-glycosylation. During the initial step of O-glycosylation an *N*-acetylgalactosaminyl peptidyltransferase adds a *N*-acetylgalactosamine (GalNac) to serine or threonine residues in the mucin protein backbone producing O-glycans. Elongation of the O-glycans then occurs whereby specific glycosyltransferases (GT) attach either a sialic acid or hexose; galactose (Gal), N-acetylglucosamine (GlcNAc), fucose. Studies of mucin O-glycans have identified four major core structures defined as follows; for core 1 and 2 O-glycans, galactose is transferred to C-3 of the GalNAc which is elongated in the case of core 1. With the core 2 O-glycan branching occurs whereby the next transferase adds to the GalNAc again. In core 3 O-glycans a GlcNAc is added to the GalNAc, this also occurs with core 4 but is followed by further additions to the GalNAc (Rose and Voynow 2006).

The core structures are then elongated by galactose and GlcNAc transferases to form Gal β1, 3/4 GlcNAc units which can be terminated by various substances e.g. fucose, sialic acid, sulphate and even the blood group determinants (Rose and Voynow 2006).

Patterns of mucin domain glycosylation can vary from molecule to molecule in many different ways depending on; the precursor availability, glycosyltransferase expression patterns, the number of tandem repeats, differences in the sequence of tandem repeat units and allelic variation of the transferase as occurs in the ABO Lewis and secretor genes. Thus the mucin glycosylated domain provides a plethora of heterogeneity.

1.4.1.2 Membrane tethered and secreted

There are two main types of mucin, membrane-tethered and secreted. The membrane tethered mucins are comprised of three regions, a short cytoplasmic tail (CT), a single membrane spanning region which passes through the membrane only once and a large extracellular domain mainly composed of the glycosylated tandem repeat region. The transmembrane mucin is not a continuous protein but is cleaved at a site close to the membrane spanning region, producing two subunits which reattach via SDS-liable bonds (Hattrup and Gendler 2008).

The sugar chains cause the transmembrane mucins to extend into the lumen as extended 'bottle-brush' structures which can stretch further than other cell receptors and potentially prevent pathogens from interacting with the epithelial cells (Hattrup and Gendler 2008).

There are two types of secreted mucin, polymeric and non-polymeric. MUC7 is considered non-polymeric because although it is secreted it does not have the capacity to oligomerise. The secreted polymerising mucins have the ability to bond to each other via the disulphide linkages that occur between the cysteine rich regions in the amino and carboxyl terminal regions which flank the central glycosylated region and exhibit extensive homology to human von Willebrand factor (vWF) domains (see section 1.4.2.2.2 for details). The disulphide links are very important for the mucus properties since they are thought to give mucus its gel-like properties (Thornton *et al.* 2008).

1.4.2 Airway mucins

Both secreted and membrane-tethered mucins are expressed within the airways and the relationship between the two mucin types is essential for a healthy respiratory tract. The large secreted oligomeric mucins form a viscous gel which sits on top of the periciliary liquid layer (PCL) where the membrane-tethered mucins reside. The 'sticky' gel layer entraps noxious agents so that they can be removed by ciliary beating within the low viscosity PCL bordering the epithelial cells in a process known as the mucosal ciliary escalator (Curran and Cohn 2009).

1.4.2.1 Membrane-tethered airway mucins

The membrane-tethered mucins MUC1, MUC4 and MUC16 are located within the airway PCL, on the apical surface of the epithelial cells. MUC1 is generally localised to the microvilli, whereas MUC4 and MUC16 tend to be found on the cilia (Hattrup and Gendler 2008).

The membrane-tethered mucins of the airways are not indefinitely attached to the epithelial cells as both MUC1 and MUC4 can be identified within mucus secretions (Hattrup and Gendler 2008). Tumour necrosis-α converting enzyme (TACE/ADAM17) has been shown to cleave MUC1 from uterine epithelia releasing the extracellular domain (Brayman *et al.* 2004), and while this is a plausible explanation for the secreted form of MUC1 in the airways, an alternatively spliced form of MUC1 has also been identified which lacks the CT and transmembrane domains (Hinojosa-Kurtzberg *et al.* 2003). Since 24 different MUC4 transcripts have been identified, alternative splicing is also likely to result in the release of this mucin into the airway lumen (Escande *et al.* 2002).

MUC4 has the ability to interact with the oligomeric mucins in mucus secretions via disulphide linkages since it contains an extracellular vWF cysteine rich D domain. In some diseased airways, mucus has been seen to tether to the epithelium, making clearance difficult, causing the mucus to become stagnant (Thornton *et al.* 2008). It has been suggested that tethering in this instance may be a direct result of interactions between the D domains of MUC4 and the secreted oligomeric airway mucins (Hattrup and Gendler 2008).

MUC1 is known to participate in cell signalling. For example the bacterial protein flagellin has been shown to interact with the extracellular domain of MUC1 which is thought to activate signalling through phosphorylation of the conserved tyrosine residues in the CT. The MUC1 CT has also been localised in the nucleus where it is thought to directly regulate transcription. It has even been suggested that this MUC1 signalling may upregulate the transcription of other mucin genes through NFκB and MAPK pathways (Hattrup and Gendler 2008).

1.4.2.2 The secretory airway mucins

MUC5AC and MUC5B are the predominantly expressed airway mucins (Kirkham *et al.* 2002) accounting for more than 90% of the total mucin found in sputum (Hattrup and Gendler 2008), and it is therefore thought that they give respiratory mucus its characteristic gel-like and ‘sticky’ properties. MUC5B is primarily expressed in the mucous cells of the submucosal glands (Groneberg *et al.* 2002a) while the surface epithelial goblet cells are the principle secretors of MUC5AC (Groneberg *et al.* 2002b; Hovenberg *et al.* 1996). However during airway disease MUC5B can become aberrantly expressed in the goblet cells (Groneberg *et al.* 2002a; Kamio *et al.* 2005).

In the airways MUC2 is expressed at only very low levels and is not easily detectable but may also be aberrantly expressed during disease (Hovenberg *et al.* 1996; Ordonez *et al.* 2001).

The predicted *MUC19* gene has been proposed to code for another large secreted oligomeric mucin (Chen *et al.* 2004). However there is little evidence to suggest that this protein is present in human respiratory mucus (Thornton *et al.* 2008).

1.4.2.2.1 The glycosylated domains of the secretory mucins

Mucin carbohydrate side chains are important in determining mucus biophysical properties since they are essential for mucin expansion when in solution, protease resistance, requisition of pathogens, water holding and ion binding (Thornton *et al.* 2008). Since the glycosylated central domain differs between mucins, alterations in the mucin composition of mucus will change its physical properties.

Within the colon, the major part of MUC2 appears to occur as an insoluble form. (Herrmann *et al.* 1999). Its central region is composed of two repetitive domains separated by approximately 600bp. The first region mainly contains a 16 amino acid repeat motif (table 1.1). The second region is affected by a polymorphism known as variable number tandem repeat (VNTR). The repeat unit of 23 amino acids as seen in table 1.1, has been shown to be present any number of times between 51 to 115 (Toribara *et al.* 1991), and in addition some variation occurs within the tandem repeats. Thus the central glycosylated domain of MUC2 varies greatly between individuals.

Table 1.1 Amino acid repeat motif units within the central regions of MUC2, MUC5AC and MUC5B (Desseyn *et al.* 1997b; Escande *et al.* 2001; Toribara *et al.* 1991).

Mucin	Tandem repeat unit (amino acid)
MUC2	PPTTTPSPPPTSTTTL (region 1) PTTTPITTTTTVTPTPTPTGTQT (region 2)
MUC5AC	TTSTTSAP (consensus) GTTPSPVP (frequently occurs)
MUC5B	ATGSTATPSSTPGTTHTPPVLTTTATTPT

The central domains of MUC5B and MUC5AC are somewhat different from MUC2. While threonine and serine richness is again a prominent feature of their central domains, the glycosylated regions are interrupted by cysteine rich domains termed cys (see figures 1.1 and 1.2). Only one conserved potential O-glycosylation site is present within each of the cys domains meaning that these are virtually free from the constraints inflicted by carbohydrate structures (Desseyn et al. 1997b; Escande et al. 2001). The cys domains are thought to act as hinges, allowing the mucins to be flexible, a property essential for airway mucus whereby it is required to be easily moved via mucosal ciliary transport (Thornton *et al*. 2008). However the very fact that these domains are rich in cysteine implies an additional role with respect to bond interactions. In fact a study of the yeast two-hybrid system has identified interactions between two of these central cys domains and histatin 1 (Iontcheva et al. 2000).

The glycosylated central region of MUC5B is composed of 3570 amino acid residues, coded for by a single exon of 10.7kb. The sequencing by Desseyn et al (Desseyn et al. 1997b) led them to describe the central region as having 19 subdomains (see figure 1.1). Subdomains R01, R02 and R03 have no typical repeat structure but are rich in threonine, serine and proline (TSP). The five subunits RI to RV are also TSP rich, but in addition exhibit a repeating structure composed of various numbers of the 29 amino acid repeat motif shown in table 1.1. The R-ends are very similar to each other and are again TSP rich.

The glycosylated subdomains of MUC5B are interrupted by seven cys subunits (Cys1-7). The cysteine residue positions within these domains are highly conserved when aligned with homologous domains from human MUC5AC, mouse Muc5ac, pig gastric mucin, human MUC2 and the rat Muc2 homologue (Desseyn *et al.* 1997b).

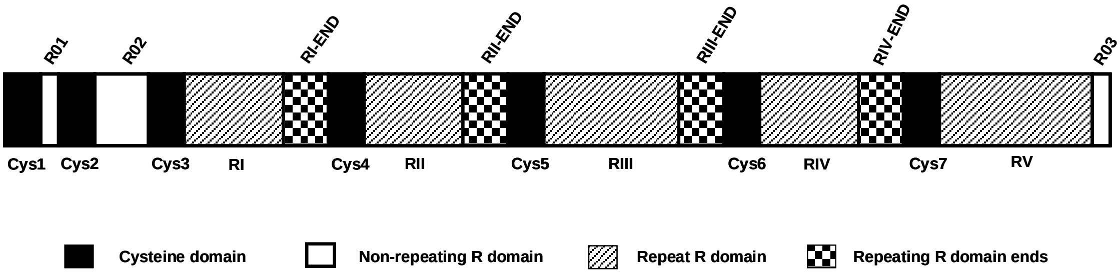


Figure 1.1 MUC5B protein central region and the annotated subdomains. Deduced by Desseyn et al (Desseyn et al. 1997b).

The glycosylated central region of MUC5AC is also coded for by a single exon which is thought to be approximately 10.5kb although it has only been partially sequenced. This central protein region has been deduced to include 17 subdomains (figure 1.2) (Escande et al. 2001). Nine of these subdomains are cysteine rich (Cys1-9) and have a high degree of sequence identity especially with respect to the cysteine residue locations which are completely conserved. Four non-repeating domains termed TSP 1-4 are rich in threonine, serine and proline. TSP1 and TSP3 are almost perfectly identical while

TSP2 and TSP4 are also completely homologous. The remaining four domains are also TSP rich, however they each contain various numbers of the MUC5AC repeat unit of 8 amino acids (table 1.1).

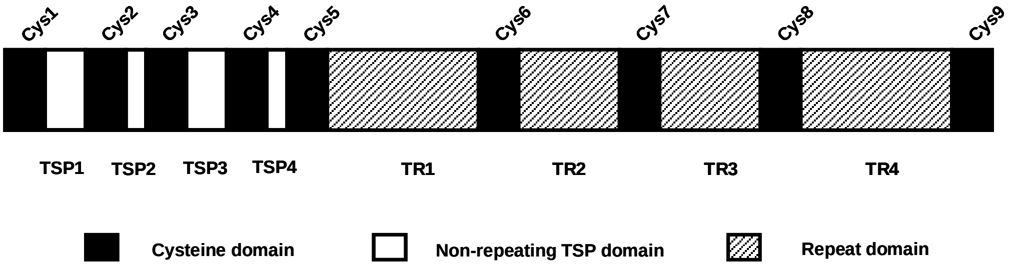


Figure 1.2 MUC5AC protein central region with annotated subdomains (Desseyn *et al.* 1997b; Escande *et al.* 2001).

1.4.2.2.2 Amino and carboxyl terminal regions

The amino and carboxyl terminal regions of the secreted oligomeric mucins are rich in cysteine residues and are therefore thought to be the main sites of oligomerisation via disulphide bonds which are of paramount importance for the gel-like properties of mucus (Desseyn *et al.* 1997b; Thornton *et al.* 2008). Both regions exhibit high sequence homology between the secreted mucins MUC2, MUC5AC and MUC5B, and also share high sequence identity with the human vWF. The most prominent conservation is seen with respect to the numbers and positions of the cysteine residues and also the intron/exon boundaries.

Human vWF is a multimeric glycoprotein found in plasma. It plays an important role in blood clotting, acting to stabilize clotting factor VIII and mediating platelet adhesion. As shown in figure 1.3 the amino and carboxyl terminal regions of the oligomeric mucin proteins have been divided into domains based on those predefined for the human vWF (Buisine et al. 1998; Desseyn et al. 1997a; Desseyn et al. 1998; Escande et al. 2001; Gum, Jr. et al. 1992; Offner et al. 1998).

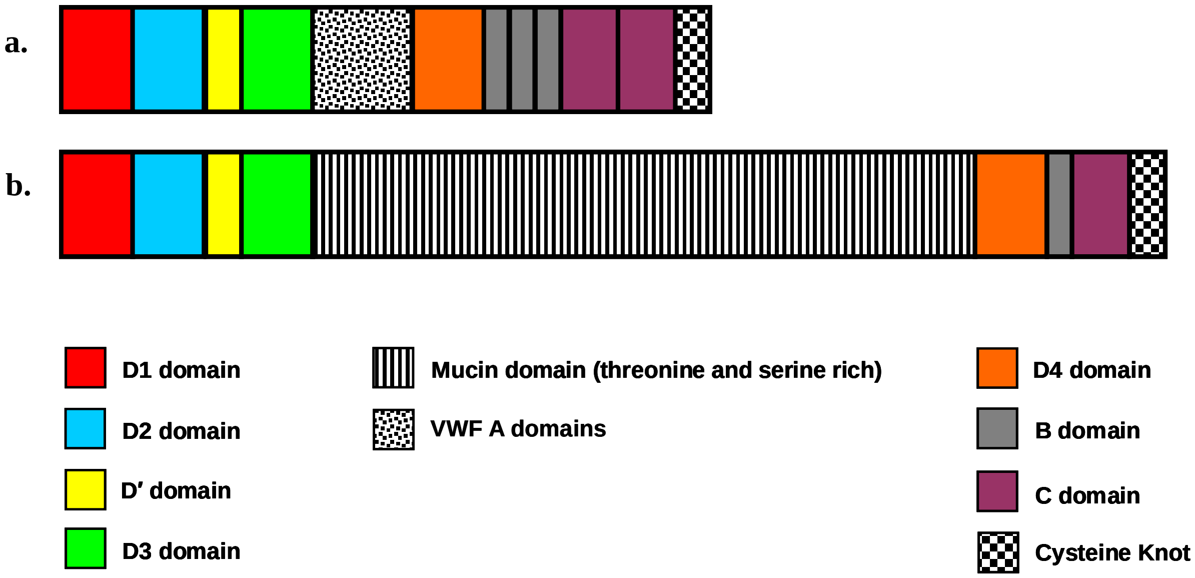


Figure 1.3 Major domains of the secreted oligomeric mucins and human von Willebrand Factor. a. represents human vWF protein divided into its major domains. b. represents the domains of a generic secreted oligomeric mucin (MUC2, MUC5AC and MUC5B) (Perez-Vilar and Hill 1999; Thornton *et al.* 2008; Verweij *et al.* 1987).

1.4.3 Mucin Assembly

It is difficult to study mucin assembly because of the sheer size and complexity of the proteins. Mucus gel can be solubilised by agents such as 6M guanidium chloride, which breaks down non-covalent bonds, and it is therefore thought that the entanglement of mucins gives mucus its gel-like properties (Thornton and Sheehan 2004).

After mucus denaturation, electron microscopy reveals mucin monomers assembled in linear chains held together end to end by disulphide bonds (Thornton and Sheehan 2004). However the procedures used to extract mucins from mucus are very harsh and disruptive. Therefore we cannot be sure that all native state patterns of assembly have been identified.

Protein studies of mucus generally show that samples from healthy and diseased airways largely differ in their mucin composition. However a single study of a mucus plug obtained during autopsy from the airways of a *status asthmaticus* patient, revealed unusual mucin architecture. A low charged glycoform of MUC5B accounted for 96% of the total mucin content within the viscid mucus plug. Electron microscopy of the sample did not reveal the usual linear threads but instead entangled nodes of mucins with emanating linear threads (Sheehan *et al.* 1999). This is the only case of extensive branching and cross-linking identified to date, however it successfully highlights the severity caused by defects in mucin assembly.

A recent study has attempted to characterise the dynamics of MUC5B by allowing intact MUC5B molecules to interact with a variety of surfaces. They have shown that MUC5B forms a structured and hydrated interface that ranges from 40-100nm thick. The regions of the mucin proteins that are not glycosylated, referred to as the naked protein, are responsible for attachment. The carbohydrate structures dominate the interface, and it is this carbohydrate richness that confers the absorptive properties of the interface. Keismer and Sheehan have used microbeads to show that this interface selects what it absorbs based on size. Beads of 60nm are allowed to enter through the interface, however beads of greater than 100nm in size are excluded from the layer (Kesimer and Sheehan 2008).

1.5 Hypersecretory airway diseases

In healthy airways the epithelial layer has adopted a homeostatic mechanism whereby inflammation caused by various environmental insults and pathogens results in the upregulation of mucin expression (Gray *et al.* 2004b; Gray *et al.* 2004a; Koo *et al.* 2002) which consequently allows for greater amounts of mucus to be secreted. The increased airway mucus helps to exclude the noxious agent from the airways by entrapment and removal via the mucociliary escalator and cough. When the inflammatory inducer has been removed the mucosal system reverts back to its balanced state. However during chronic respiratory diseases such as asthma, Cystic fibrosis (CF), diffuse panbronchiolitis (DPB) and chronic obstructive pulmonary disease (COPD), this mechanism persists at a greater intensity and for a longer period of time causing this protective mechanism to actually exacerbate disease symptoms. In fact mucus airway occlusion can often be fatal and a recent study of subjects, who had died in *status* *asthmaticus*, has shown that 95% of the patients had airway narrowing that ranged from 20-100% of the airway lumen due to mucus occlusions (Kuyper *et al.* 2003).

Although enlarged submucosal glands are evident in some respiratory diseases (Bai and Knight 2005; Fahy 2002; Kamio *et al.* 2005; Rogers 2004) the surface epithelial goblet cells are thought to be the principal contributors of excess mucin production during episodes of mucus hypersecretion because mucus occlusions have been found in parts of the airways where goblet cells are the only source of mucus production (Evans and Koo 2009; Turner and Jones 2009). Even in mild and moderate asthmatic airways, the mean volume of stored goblet cell mucin was three times higher in the cases than the controls (Ordonez *et al.* 2001; Turner and Jones 2009).

In healthy humans goblet cells are rarely seen distal to the trachea i.e. the small airways (Curran and Cohn 2009; Williams *et al.* 2006). However during episodes of inflammatory disease, mucous cell numbers within these regions are dramatically increased (Groneberg *et al.* 2002b; Groneberg *et al.* 2002a; Kamio *et al.* 2005). The excess secretory cells are thought to be a result of goblet cell metaplasia in which pre-existing cells adopt an alternative phenotype (Williams *et al.* 2006).

Alteration of mucin composition within the airway mucus has been identified during episodes of inflammatory respiratory disease. In cases of asthma and COPD, MUC5B has been shown to be the primary respiratory mucin within mucus plugs and sputum respectively (Burgel and Nadel 2004; Groneberg *et al.* 2002a), although in healthy airways MUC5AC has been shown to predominate (Kirkham *et al.* 2002).

1.6 Secretory oligomeric mucin gene complex

The genes that code the secreted oligomeric mucins, *MUC6*, *MUC2*, *MUC5AC* and *MUC5B* are located within a 400kb region on chromosome 11p15.5. Pulse field gel electrophoresis was used to determine the order of these genes as *MUC6-MUC2-MUC5AC-MUC5B* (see appendix 4a), in a telomeric to centromeric direction (Pigny *et al.* 1996). The human genome browser annotations for this gene complex are not correct since *MUC5AC* and *MUC5B* are shown to overlap on the golden path sequence, implying that they are the same gene (see appendix 4b). It should also be noted here that due to the difficult nature of the tandemly repeated *MUC5AC* central region, a sequence gap still remains in the golden path for this gene. The gap can be partially supplemented by sequencing data from the literature (Escande *et al.* 2001) however some parts of the central region remain unsequenced.

This secretory mucin gene complex is highly conserved and has been identified in the genomes of the mouse, dog, chimpanzee, rhesus monkey, cow and horse (Thornton *et al.* 2008).

Extensive LD is evident within the *MUC* gene complex since long ranging haplotypes have been shown to extend from *MUC2* to *MUC5B* (Rousseau *et al.* 2007). However, no association between *MUC6* variants and any of the other mucin markers has been identified within the 11p15.5 *MUC* complex, due to a recombination hotspot located between *MUC6* and *MUC2* (Rousseau *et al.* 2007).

1.6.1 Mucin gene associations with respiratory disease

Significant associations have been noted between variation in *MUC7* and *MUC2* VNTRs and asthma phenotypes. The MUC7 repeat domain is composed of units of 23 amino acids which can occur 5, 6 or rarely 8 times (alleles *MUC7*5*, *MUC7*6* and *MUC7*8* respectively). The short allele *MUC7*5* has been shown to be significantly underrepresented in a small asthmatic population (Kirkbride *et al.* 2001). This study has also shown that *MUC7* haplotypes carrying the short length allele are significantly associated with better lung function (higher FEV measures and reduced FEV decline where FEV is defines as forced expiratory volume) (Rousseau *et al.* 2006).

In a small matched asthmatic cohort the distributions of *MUC2* VNTR allele lengths indicate that the larger alleles may protect atopic individuals from developing asthma (Vinall *et al.* 2000). Since the MUC2 protein has only been found at very low levels within the airways it was proposed that the association identified may in fact be due to a linked causal polymorphism within *MUC5AC* or *MUC5B* (Rousseau *et al.* 2007) which code for the dominate airway mucins.

Length variation of the threonine and serine rich central region will affect the extent of glycosylation and therefore VNTR polymorphisms are generally accepted as the most likely functional candidates with respect to disease association. However no length variation has been noted for *MUC5B* and only small differences have been identified for the *MUC5AC* tandem repeat (Escande *et al.* 2001; Vinall *et al.* 2000), thus we might expect any disease association involving these genes to be due to different functional variation. Variation in MUC5AC and MUC5B glycosylation may of course result from sequence variants within the central region, and polymorphisms of this kind have previously been identified for *MUC1* (Fowler *et al.* 2003) and *MUC2* (Toribara *et al.* 1991).

It is also highly plausible that variation within the regulatory regions of *MUC5AC* and *MUC5B* may be associated with hypersecretory respiratory diseases. During chronic airway disease the upregulation of mucin expression in response to insults is known to be exaggerated and prolonged, which could in itself be a direct consequence of regulatory variation. In fact polymorphisms within the *MUC5B* promoter have been shown to be significantly associated with diffuse panbronchiolitis (Kamio *et al.* 2005).

It is clear from the high levels of sequence conservation in the amino and carboxyl regions that these domains are important for mucin function. We also know that the corresponding cysteine rich regions in human vWF are essential for its assembly and that mutations within these regions cause von Willebrand disease due to defects in oligomerisation (Michiels *et al.* 2006). It may therefore be proposed that missense mutations within the amino and carboxyl termini are likely to influence normal function.

Recently a non-synonymous SNP (Ala497Val) located within exon 12 of the *MUC5AC* amino terminal has been shown to be significantly associated with idiopathic interstitial pneumonia. This SNP is located within an amino terminal D domain and it is therefore thought that disease susceptibility may be a result of assembly and oligomerisation defects of the MUC5AC protein (Burch *et al.* 2010).

ENU-induced mutant mice have also highlighted the importance of these regions. Missense mutations within Muc2 have been shown to cause spontaneous ulcerative colitis and chronic diarrhoea in mutant mice *Winnie* and *Eeyore*. A mutation resulting in the substitution of a cysteine to a threonine in the amino terminal D3 domain, and the substitution of a serine to a proline in the carboxyl D4 domain have been identified in *Winnie* and *Eeyore* respectively. Both mice exhibit defects in mucin oligomerisation and secretion, and ER stress in the goblet cells due to protein misfolding has also been noted (Heazlewood *et al.* 2008).

1.7 Project Aims

The overall aim of this thesis is to describe variation within the mucin genes expressed in the respiratory tract, *MUC5AC* and *MUC5B*, and to examine this variation with respect to disease and demography. This thesis is divided into four results sections which aim to examine two specific hypotheses.

We firstlty hypothesise that inter-individual variability of the predominantly expressed respiratory mucins, MUC5AC and MUC5B, will confer differing degrees of suseptibility to respiratory disease.

- **Chapter three** aims to explore this hypothesis by studying the genetic variability of *MUC5AC* with respect to allergy related respiratory disease. This results section examines associations between a single nucleotide polymorphism within the 3′ region of *MUC5AC* and various respiratory outcome data, in a longitudinal birth cohort. Tests of gene-gene interactions are also performed between the same *MUC5AC* variant and a variety of functional inflammatory markers with respect to respiratory outcome.
- **Chapter four** aims to explore this hypothesis by studying both *MUC5AC* and *MUC5B* variability with respect to asthma. This results section describes genetic variation within the *MUC5AC* 3′ region and *MUC5B* regulatory regions in the context of asthmatic case-control association studies. *MUC5AC* 3′ variation has been explored within these sample sets in an attempt to replicate the associations identified in chapter 3. *MUC5B* regulatory variation has been studied in this

results section since haplotype variants have been shown to confer different expression levels of *MUC5B* in two independent studies (Kamio *et al* 2005, Loh *et al* 2010). Taking into account *MUC5B* promoter variant functional information, we hypothesise that the high expressing haploptype will be overrepresented in the asthmatic cases on the basis that asthma is a characteristic hypersecretory disease.

- **Chapter five** aims to identify the *MUC5AC* genetic factor that may be responsible for the disease associations identified in chapter three. This results section introduces a sequencing strategy to identify copy number variation and applies this technique to *MUC5AC* sequence traces in order to explore the possibility that this gene is affected by copy number.

Since mucin expression is affected by environmental cues, we finally hypothesise that the *MUC5B* promoter variants will vary between geographically distinct populations who have experienced different environmental pressures during their demographic histories.

- **Chapter six** aims to explore this hypothesis by characterising genetic variation within the *MUC5B* promoter in African samples of varying geographic location. Promoter variants are examined with respect to population differentiation, diversity measures and species conservation profile.

2 Material and Methods

2.1 DNA Samples

2.1.1 Laboratory volunteer samples

These anonymous samples were donated with consent from laboratory employees and students. The samples are not from a single population or phenotypic group and were used solely for optimisation purposes.

2.1.2 1946 longitudinal birth cohort

The MRC National Survey of Health and Development longitudinal study initially recorded information about all United Kingdom births within the week 3-9th March 1946. A social class stratified cohort of 5362 individuals was followed further. The individuals within this cohort have been studied a total of 21 times, 10 during childhood and 11 in adulthood. The last data collection for which we have collated information was carried out at age 53 years, at which time blood and buccal samples were collected from consenting participants. As a result, DNA is available for 2939 of these individuals (ethical approval reference MREC no 98/2/121). This sample is considered to be representative of a European population since the study began before mass immigration into the United Kingdom.

2.1.3 Matched asthmatic cohort

The matched asthmatic disease cohort is composed of 100 individuals; 50 clinically diagnosed asthmatics and 50 non-asthmatic controls, which were age and sex matched details for which can be found in Vinall *et al* 2000. Both asthmatic and control sample sets are composed of atopic individuals. Atopy was defined by positive skin prick tests which were generally confirmed by elevated IgE levels. All individuals are British with northern European ancestry. DNA was extracted from blood samples collected from St Mary’s Hospital Chest clinic. Ethical approval has been obtained from Parkside Health Authority (EC no 2893). DNA extraction procedures have been described previously (Kirkbride *et al.* 2001; Vinall *et al.* 2000).

2.1.4 Severe asthmatic cohort

The severe asthmatic disease cohort is comprised of 177 individuals; 85 severe asthmatics and 92 ‘hyper’ normal non-asthmatic controls, all are thought to be of northern European ancestry. The asthmatic disease samples have been clinically diagnosed and were defined as severe if they were on step 4 or 5 of the British Thoracic Society’s asthma treatment guidelines. All asthmatic individuals were recruited from Southampton general hospital in collaboration with Dr John Holloway.

The ‘hypernormal’ control sample is comprised of healthy, non-asthmatic individuals believed to be of northern European ancestry. All individuals are recruited blood donors from around the Southampton area and have reported no personal or family history of respiratory disease. No atopy information is available for either sample set.

For both the severe and control sample sets, DNA was extracted from 10ml of blood using the Qiagen Genomic DNA Maxi-prep kit. Ethical approval was obtained from the Southampton and S. W. Hants Joint Ethics Committee.

2.1.5 African samples

Eight African sample sets obtained from The Centre for Genetic Anthropology, have been examined during this project; Ethiopia (n = 380), Mozambique (n = 51), Ghana (n = 57), Malawi (n = 50), Cameroon Lake Chad (n = 65), Cameroon Grassfields-Somie (n = 65), Congo (n = 55), Sudan (n = 30). DNA was extracted from buccal swabs taken only from anonymous males, unrelated at the grandpaternal level. Informed consent was given by all participants. Social and anthropological data are available for each individual which include, participants first and second language, mother and fathers first and second language, birthplace, and self-declared cultural identification.

2.1.6 cDNA samples

The cDNA samples used during this project had been synthesised from RNA extracted from a variety of subclones of the HT29-MTX mucus secreting cells by Wendy Pratt. For the cell culture protocol refer to Lesuffleur *et al* (Lesuffleur *et al.* 1993) and for RNA extraction and cDNA synthesis procedures refer to Wang *et al* (Wang *et al.* 1994).

2.2 Genotyping methods

2.2.1 Six SNP multiplex

The 1946 cohort samples were genotyped for six polymorphisms using a multiplex technique. Three SNPs were genotyped in *MUC5B* (rs2672785, rs2075853 and rs2075859), and one each within the genes *MUC5AC*, *IL4* and *IL13* (rs1132440, rs2070874 and rs1800925 respectively), the relevant details are given in Chapter 3.

2.2.1.1 Multiplex PCR

Optimal polymerase chain reaction (PCR) conditions were defined by altering thermal cycling and primer concentrations, and the final PCR conditions used are as follows. To each 2µl DNA sample (mean concentration of 30ng/µl) the following PCR reagents were added: 1µl of 10X Abgene Buffer IV containing $MgCl_2$ [750mM tris-HCl (pH8.8 at 25°C), 200mM $(NH_4)_2SO_4$, 0.1% v/v Tween 20®, 15mM $MgCl_2$], 0.1µl of 100X dNTPs, 5 p moles of rs1132440, rs2070874, rs1800925, 9 p moles of rs2672785, 3.5 p moles of rs2075853 and 2 p moles of rs2075859 primers. Distilled water was added to make a final reaction volume of 10µl. Table 2.1 shows the *MUC5AC* primer sequences. Please note that primer sequences for all other SNPs can be found in Andrew Loh, PhD thesis (Loh 2007) and Black et al supplementary information (Black *et al.* 2009).

Thermal cycling of the samples was initiated by denaturation at 95°C for 5 minutes, followed by 30 cycles of 30 seconds at 95°C, 30 seconds at annealing temperature 60°C and 30 seconds at 72°C. A final elongation step of 72°C for 5 minutes was added to the end of the thermal programme.

Post optimisation, two PCR products from each plate were run on a 12% 19:1 acrylamide gel and silver stained in order to visualise the multiplex product prior to the next step of PCR product treatment.

4µl of each PCR product was treated, in order to remove remaining primers and residual dNTPs, by adding 1.33 units of Shrimp Alkaline Phosphatase (SAP) (USB Corporation) and 0.8 units of Exonuclease I (NEB, inc) Additions of these reagents were followed by 1 hour incubation at 37°C and enzyme deactivation at 75°C for 15 minutes. The treated PCR products were held at 4°C prior to -20°C storage.

2.2.1.2 Multiplex Single base extension (SBE)

Single base extension (SBE) was the chosen genotyping method for the multiplex. To 1µl of purified PCR product, 1.67µl of SNaPshot™ Multiplex Ready Reaction Mix was added along with 0.66µl of a SBE multiplex primer mix containing equal volumes of all six Primer Extension (PE) primers at the following concentrations: 0.75 µM rs2075853

PE primer, 3.6µM rs2672785 PE primer, 3µM rs2075859 PE primer, 1.8µM rs1132440 *MUC5AC* forward primer, 0.2µM rs1800925 PE IL13 primer, 2.4µM rs2070874 PE IL4 primer (see table 2.2 for details).

The samples were then subjected to 25 thermal cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 30seconds, before being held at 4°C.

In order to remove 5′-phosphate groups from the unincorporated ddNTPs, the SBE products were treated with the following mixture: 0.33 units of calf intestinal phosphatase (CIP), 0.43µl of 10x CIP buffer and 0.23µl of dH_2O. Samples were then incubated at 37°C for 1 hour followed by an enzyme deactivation step of 75°C for 15 minutes, and a final hold of 4°C.

10µl of Hi-Di formamide and 0.25µl of Genescan LIZ 120 size standard was added to 1µl of the SBE product, which were then detected via electrophoresis on the ABI 3730xl Gene Analyzer.

2.2.2 Basic PCR protocol

The basic PCR protocol is used for each of the subsequent PCR reactions and the omitted details regarding annealing temperatures, cycle numbers and primer sequences are shown for each specific PCR reaction in table 2.1.

To each 2µl DNA sample the following PCR reagents were added: 1µl of Abgene 10X Buffer IV containing $MgCl_2$ [750mM tris-HCl (pH8.8 at 25°C), 200mM $(NH_4)_2SO_4$, 0.1% v/v Tween 20®, 15mM $MgCl_2$], 1µl of 10X dNTPs, 2.5 p moles of the forward primer and 2.5 p moles of the reverse primer. Distilled water was added to make a final reaction volume of 10µl. Thermal cycling of the samples was initiated by denaturation at 95°C for 5 minutes, followed by cycling of 30 seconds at 95°C, 30 seconds at the optimal annealing temperature and 1 minute at 72°C. A final elongation step of 72°C for 5 minutes was added to the end of the thermal programme. All PCR products were visualised by agarose gel electrophoresis (1% to 3% gels as appropriate).

2.2.3 DNA sequencing - SNP discovery and genotyping

2.2.3.1 Purification of the PCR product

After amplifying the fragment to be sequenced, the PCR product was purified in order to eliminate excess primers, dNTPs and buffer. PCR products were purified using a method of a home made PCR clean up solution [40% PEG-8000, 1 M NaCl, 2mM Tris-HCl (pH 7.5), 0.2mM EDTA, 3.5mM $MgCl_2$], ethanol wash and centrifugation on the ALC® PK120 (maximum RPM 4000). Subsequent purified PCR products were air dried and re-suspended in distilled water.

2.2.3.2 Sequencing reactions

To 4µl of purified PCR product the following reagents were added: 2.4 p moles of either the forward or reverse primer used in the PCR step was added to a mixture containing 5µl of sequencing buffer, 1µl of ABI BigDye® Terminator mix v1.1 or v3.1 and 0.375µl of DMSO. Each reaction was made up to 15µl volume with distilled water. These sequencing reactions were then subjected to a programme of 98ºC for 10 minutes, followed by 25 cycles of 10 seconds denaturing at 98ºC, 5 seconds annealing at 50ºC and 4 minutes of elongation at 60ºC.

The sequencing reaction products were purified in a two step isopropanol and centrifugation procedure (ALC® PK120 maximum RPM 4000) to remove excess terminator mix, primers and buffer. The resulting purified products were air dried and re-suspended in 10μl of formamide. Electrophoresis of the products on the ABI Gene Analyzer 3730xl was used to detect the sequence results.

Sequencing reactions were mostly conducted with the help of Mari-Wyn Burley at the Centre for Comparative Genomics.

2.2.4 Restriction fragment length polymorphism (RFLP)

2.2.4.1 Variant -614 restriction enzyme digest

PCR steps were in accordance with the described basic PCR protocol and specific details for this assay can be found in table 2.1. To 3μl of the PCR product 2.5 units of BsaI enzyme (NEB, inc) plus 1.5μl of NEB buffer 3 were added and the final reaction volume was made up to 15μl using distilled water and incubated for 16 hours at 50°C.

2.2.4.2 Variant -988 restriction enzyme digest

Since this variant does not fall within a natural enzyme restriction site, an allele specific restriction site was introduced into the PCR product by engineering a site in the reverse primer (see details in table 2.1) in order to perform RFLP as the method of genotyping. The PCR step was in accordance with the basic protocol described in section 2.2.2 and specific details are given in table 2.1. To 3μl of the PCR product 2.5 units of HinfI enzyme (NEB, inc) plus 1.5μl of NEB buffer 2 were added and the final reaction volume was made up to 15μl using distilled water and incubated for 16 hours at 37°C.

2.2.4.2.1 Variant rs1132440 restriction enzyme digest

All repeat typings of the *MUC5AC* variant rs1132440, were performed using RFLP. The PCR step was in accordance with the basic protocol described in section 2.2.2 and specific conditions are shown in table 2.1. To 3µl of the PCR product 2 units of HaeII NEB enzyme, 1.5µl of NEB buffer 4 and 1.5µg of BSA were added and the final reaction volume was made up to 15µl using distilled water and incubated for 16 hours at 37°C.

2.2.5 RT-PCR

For each RT-PCR assay, the following reagents were added to each 2µl cDNA sample: 1µl of Abgene 10X Buffer IV containing $MgCl_2$ [750mM tris-HCl (pH8.8 at 25°C), 200mM $(NH_4)_2SO_4$, 0.1% v/v Tween 20®, 15mM $MgCl_2$], 1µl of 10X dNTPs, 2.5 p moles of the forward primer and 2.5 p moles of the reverse primer (see table 2.1 for primer sequences). Distilled water was added to make a final reaction volume of 10µl. Thermal cycling of the samples was initiated by denaturation at 95°C for 5 minutes, followed by cycling of 30 seconds at 95°C, 30 seconds at the optimal annealing temperature and 30 seconds at 72°C. A final elongation step of 72°C for 5 minutes was added to the end of the thermal programme.

cDNA samples were previously made in the lab by Wendy Pratt and Jo Fowler.

2.2.6 Genotyping of non-mucin genetic markers

Genotyping of genetic markers within the genes *EGFR*, *IL13*, *IL1RN*, *IL1B* and *TNF* were conducted by Lynne Vinall and details of protocols are given in the appendix (appendix 1).

Table 2.1 PCR assay conditions.

PCR assay	Forward primer	Reverse primer	final conc (p moles)	Annealing temp (℃)	thermal cycles
MUC5AC rs1132440	5′ACACCGAGGTGGAAGAGTGC 3′	5′CTGGACAGGGGCACAAGTTC 3′	5.0	60 (multiplex) 59 (RFLP)	30
MUC5B promoter A	5′CCACGGAGCATTCAGGAC 3′	5′CCCTCCCACCACTGCTTAG 3′	5.0	65	35
MUC5B promoter B	5′CACAAGCCACCCAGACTG 3′	5′GAGCCAACACCAGCGTC 3′	5.0	62	35
MUC5B promoter A2	5′GTGCAGTCACAGCCACGA 3′	5′ATACGGTTCAGCCGTGAAAA 3′	5.0	56	35
MUC5B promoter B2	5′TCTCTGCGGGTTTCATGACT 3′	5′ACCTCAGATGCTCCCACCT 3′	5.0	64	38
MUC5B -614 validation	5′CCACGGAGCATTCAGGAC 3′	5′CTCAGTCTGGGTGGCTTGTG 3′	5.0	64	38
MUC5B -988 validation	5′CCACGGAGCATTCAGGAC 3′	5′ACAGCGTCATCTGCAGGA**G**AC 3′	5.0	58	38
MUC5AC 3′ exon 13 to 14	5′ATCAACGGGACCCTGTACC 3′	5′TGCAGATCTGGGTCTCACAG 3′	5.0	63	36
MUC5AC 3′ exon 18 to 19	5′CGACCTGTGCTGTGTACCAT 3′	5′GACACTGGGACGCCTCTCT 3′	5.0	63	36
MUC5B distal promoter to exon 1 (RT-PCR)	5′ACGAGGCCACACCACCCGA 3′	5′CTGCGGCACCACGAGCATG 3′	5.0	65	34
MUC5B distal promoter to exon 2 (RT-PCR)	5′ACGAGGCCACACCACCCGA 3′	5′CCGCCATCCATGGTGTGCC 3′	5.0	65	34
MUC5B exon 1 to exon 2 (RT-PCR)	5′GAGCGCGTGCCGGACGCT 3′	5′CCGCCATCCATGGTGTGCC 3′	5.0	64	30

Details of conditions shown include primers sequences, concentration of primers used, optimal annealing temperatures and number of cycles used during thermal cycling. The red underlined allele within the reverse primer used in the -988 variant validation PCR, represents an engineered site.

Table 2.2 Multiplex Single Base Extension (SBE) reaction details.

Gene	SNP		PCR		Extension primer		Extension product colour
	Identifier	Alleles	Product length (bp)	Annealing temp (℃)	length (bp)	Orientation	
MUC5AC	rs1132440	C/G	243bp	59˚C	25bp	Forward	Black/Blue
Il4	rs2070874	C/T	150bp	59˚C	21bp	Reverse	Blue/Green
Il13	rs1800925	C/T	201bp	59˚C	29bp	Reverse	Blue/Green
MUC5B (exon 2)	rs2672785	A/G	221bp	61˚C	15bp	Reverse	Red/Black
MUC5B (exon 3)	rs2075853	C/T	151bp	61˚C	17bp	Reverse	Blue/Green
MUC5B (exon 9)	rs2075859	C/T	206bp	61˚C	17bp	Forward	Black/Red

2.3 Statistical methods

2.3.1 Deviation from Hardy Weinberg Equilibrium (HWE)

Tests for deviation from Hardy Weinberg Equilibrium (HWE) generate theoretical or expected genotype distributions from the allele frequencies determined from the observed genotype counts. Where p is the frequency of the common allele and 1-p = q is the frequency of the rare allele, the equation representing expected genotype counts is as follows;

$$p^2 + 2pq + q^2 = 1$$

In general a chi-squared test is used to compare the observed and expected genotype distributions. High chi-squared values and associated low p values indicate deviations from HWE. However in cases where expected genotype counts are less than five, chi squared tests are no longer sufficient and may produce spurious results. Thus throughout this project, an exact test is used to calculate the departure from HWE, and is based on that devised by Guo and Thompson (1992) and implemented in Arlequin version 3.1.0.2. This test is said to be analogous to a Fishers exact test however the 2 by 2 contingency table is extended to a triangular contingency table, which is then examined using a Markov-chain random walk algorithm.

Deviations from HWE might be explained by factors such as a non-randomly mating population, population stratification, presence of silent alleles, selection or technical error. In general, tests of deviation from HWE are used to check data for technical errors.

2.3.2 Pearson’s Chi squared test (χ^2)

Pearson’s Chi-squared (χ^2) can be used as a test for independence and goodness of fit. In both cases the observed data must be in the form of counts e.g. genotype numbers in cases and control groups. The chi-squared values generated are evaluated for significance by comparing them with critical values in a chi-squared distribution. If the critical value falls within the 95% confidence interval, the null hypothesis is accepted- the two variables are independent and therefore the distribution of the observed events in each category is not significantly different from that expected by chance.

In order to generate the chi-squared statistic, an expected distribution of the data must firstly be defined which is the number of expected genotypes we would expect to find in each category if the variables were independent. These expected values are generally calculated when all observed data has been entered into a contingency table using the formulae as follows:

$$\text{Expected value} = \frac{\text{row total X column total}}{\text{grand total}}$$

Chi squared is essentially the sum of the squared differences between each observed and expected value divided by the expected value as represented by the formulae below where O represents the observed frequency of a genotype and E represents the expected frequency of a genotype. As the Chi squared value increases the chance of the null hypothesis being rejected also increases.

$$\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i}$$

P values for the appropriate number of degrees of freedom can be looked up in chi squared tables but have been primarily generated by SPSS software in this project.

2.3.3 Fishers Exact Test

As with Pearson’s chi-squared, the Fishers exact test is used as a test for independence. It can be used for analysing small datasets, and is most commonly applied to 2 by 2 contingency tables. The test explores whether there are non-random associations between two categorical variables.

The exact probability of obtaining the observed dataset is calculated. Probabilities are also calculated for all possible contingency tables with data distributions more extreme than the one observed. Two-tailed tests are used when there is no prior hypothesis, meaning that probabilities are calculated for tables more extreme than the observed in both directions. All calculated probability values are summed together to obtain a significance value for the observed data.

2.3.4 Logistic Regression

Logistic regression attempts to predict a dependent variable from a combination of independent variables (IV) or rather predictor factors. The dependent variable (DV) is usually a binary categorical variable for example affected and unaffected disease status. Logistic regression has been used during this project to examine the relationship between genetic variables (IV) and disease outcome (DV).

Logistic regression uses a combination of the predictor variables to predict the probability of a case falling into a particular DV category known as the odds of a category. Any asymmetry of the odds are adjusted for by using log(odds). All logistic regression calculations were performed by Iman Shah.

2.3.5 Permutation test

Permutation tests have been used in this project to test the significance of the genotype distributions and were performed by Dr Nikolas Maniatis. Observed genotype data were randomized one thousand times producing a theoretical distribution of empirical data. These simulated genotype distribution were then ranked. The observed data were then compared to the ranked distributions and a probability value was assigned, indicating the likelihood of obtaining the observed data by chance.

2.3.6 Mann Whitney U and Kruskall Wallis test

The Mann Whitney U and Kruskal Wallis test have been used to compare distributions of continuous data. Both tests are non-parametric meaning that they do not assume normally distributed data and make no assumption that the variances are homogenous. The two tests convert raw data to ranked data and sum the ranks for each population. While the Kruskal Wallis statistic can be used to test any number of groups the Mann Whitney U test can only be used to compare two groups. The Mann Whitney U test is favoured in the case of a two group test since it has greater power.

The Mann Whitney U statistic is calculated by the following formula where n_1 and n_2 are the sample sizes of group 1 and group 2 respectively and R_i refers to the rank of the sample:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$$

The Kruskall Wallis test statistic is calculated by the following formula where H is the test statistic, n is the total number of observations for all groups and R_i refers to the rank of the sample:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{} - 3(n+1)$$

(Formulae - see web citations on page 206 for reference).

2.3.7 Correspondence Analysis

Correspondence analysis is a statistical method used to visualise the relationships between the columns and the rows of a two-way contingency table that contains numeric data for categorical variables. Correspondence analysis is a kind of multi-dimensional scaling since it attempts to capture the maximum extent of correspondence within the contingency table in as few dimensions as possible.

The cross tabulation of frequencies are firstly standardised. This means that the relative frequencies are calculated for each cell by dividing each value by the sum of all entries. The new table of relative frequencies is now called a matrix. The row and column totals within the matrix are known as the row mass and column mass respectively and the mass in this sense can be viewed as a ‘cloud’ of datapoints.

Mass is used to estimate inertia which is the integral of mass multiplied by the squared distance from the centroid (the centre of the cloud). Inertia is a measure of the spread of points and each dimension is given an inertia percentage. The percent of inertia details how much of the interrelationship within the table is captured by each dimension. The first dimension always captures the most and all successive dimensions, orthogonal to the first, capture less.

The main purpose of correspondence analysis is to graphically display the relationships or rather distances between the row and/or column points. The distances are represented by coordinates for each cell within the table. Distance is used to describe the difference between the data distribution patterns across columns for each row and vice versa. The coordinates are plotted onto a two dimensional map and it is usual to plot both row and column points on the same map. However, distances between row and column points may not be interpreted. Conclusions can only be drawn for distances between column points or distances between row points.

2.4 Genetic Analysis software

2.4.1 Arlequin version 3.1.0.2

Arlequin is a population genetics software package (Excoffier *et al.* 2005). It enables the user to perform a variety of intra and inter-population tests on genetic data. During this project Arlequin was used to test for deviation from Hardy Weinberg equilibrium (HWE) by a fishers exact based method which has been discussed previously in section 2.3.1.

Arlequin analysis software has also been used to estimate haplotype frequencies and to calculate the exact test of population differentiation (ETOPD).

The ETOPD calculates pairwise measures between different populations in order to test whether they are significantly different from each other with respect to the distribution of haplotype frequencies. The ETOPD is analogous to a Fishers exact test however the contingency table can be bigger. The contingency table is built with haplotype frequencies and its significance is determined by a Markov-chain procedure as previously mentioned in the HWE section (Raymond and Rousset 1995).

2.4.1.1 Haplotype inference using the Expectation-Maximisation Algorithm

The expectation-maximisation algorithm is a two step iterative method used to infer haplotype frequencies from raw genotype data, which it assumes the genotypes are in Hardy-Weinberg equilibrium. In each step haplotype frequencies are calculated by assigning definite haplotypes to individuals that are unambiguous, i.e. individuals that are entirely homozygous or heterozygous at only one locus, and makes estimates for those that are ambiguous. Since the real number of haplotypes in the ambiguous class is not known, a variable identified as x, is incorporated into the formula, which are used to calculate haplotype frequencies. In this first step the algorithm estimates haplotype frequencies for the ambiguous category assuming random assortment and therefore no linkage disequilibrium between loci. Thus haplotype frequencies are dependent only on the allele frequencies. The overall haplotype frequencies calculated in the first step are then used to estimate x in a second step. This is then used to re-estimate the haplotype frequencies. These two steps are iterated until convergence is established, haplotype frequencies stabilize and likelihood is maximised.

2.4.2 PHASE version 2.1.1 - Bayesian haplotype reconstruction

PHASE analysis software (Stephens *et al.* 2001; Stephens and Donnelly 2003) implements a statistical method to infer haplotypes from genotype data at linked loci by using a Bayesian algorithm. Bayes theorem is a probability theorem that takes two random events, in our case observed genotype and unobserved haplotype data, and produces a posterior probability i.e. the probability of A (haplotypes), given B (genotypes). Thus the posterior probability is calculated from the prior information of observed genotypes.

This haplotype reconstruction method regards all unknown haplotypes as random quantities. It combines prior information with observed genotype information to approximate a posterior distribution using Gibbs sampling which is a type of Markov chain-Monte Carlo (MCMC) algorithm. Individual haplotypes are then estimated from the posterior distribution by looking for the most likely haplotype pairs. The haplotype inference method applied by PHASE differs from other Bayesian methods with regard to its ‘prior knowledge’. It uses an “approximate coalescent prior” to generate expected haplotype patterns, for the given population, based on the evolutionary assumption that a haplotype is more likely to occur if it can be easily generated from a known haplotype i.e. they are closely related.

Essentially the PHASE haplotype inference programme works as follows: After conclusively defining unambiguous haplotypes, it randomly guesses the unknown haplotype pairs. These ambiguous individuals are then chosen at random and the most likely haplotype pair estimations are made with the assumption that all other haplotypes have been resolved correctly. A defined number of iterations are performed between these two steps producing an approximate sample of the posterior distribution, from which individual haplotypes are estimated.

The PHASE case-control permutation test has also been used in this project. A p value is calculated which confirms or rejects the null hypothesis that cases and controls are from the same population and there is no significant difference between them in terms of haplotype frequencies. The PHASE case-control test has also been used in this project to examine haplotypic differences between different ethnic groups.

2.4.3 Regulatory motif prediction - rVISTA version 2.0

rVISTA is a sequence analysis tool implemented via a web server (http://rvista.dcode.org/). It attempts to predict cis-regulatory elements within non-coding DNA sequence by combining database searches and comparative analysis. Homologous sequences from two species are entered in FASTA format. The input sequences are firstly aligned using zPicture and the regions of high conservation are then scanned for sequences that correlate with transcription factor binding sites (TFBS). The TRANSFAC professional V10.2 library is used to search for TFBS. Regulatory elements are notoriously difficult to predict since their DNA sequence is degenerate. However, by utilising conservation information, rVISTA is able to predict biologically functional TFBS with a greater accuracy.

2.4.4 Phylogenetic shadowing - eShadow

eShadow is a web based tool (http://eshadow.dcode.org/) that implements a method known as phylogenetic shadowing, whereby multiple sequences from closely related species are compared and a conservation profile is created from the combination of all mutations. Previous methods of conservation analysis have focused on comparisons between distantly related species, for instance between humans and mice. Although such comparisons have proved invaluable for the identification of critical regions of the genome, they fail to identify functionally important regions between closely related species, for example between humans and other primates. Phylogenetic shadowing is able to accurately predict functionally conserved regions between closely related species since the multiple sequence alignments (MSA) identify an increased number of mutations relative to the base sequence. It can then utilise this cumulative mutation density to detect slowly mutating regions.

eShadow implements phylogenetic shadowing in three steps. Multiple sequence alignments (MSA) of the closely related sequences are firstly generated by the EBI alignment tool ClustalW (http://www.ebi.ac.uk/Tools/clustalw2/index.html). These MSAs are then visualised by eShadow as percentage variation plots, where the x axis refers to the nucleotide position of the base input sequence and the y axis refers to the percentage mismatched nucleotides computed as the number of reference sequence mismatches in a 15bps window divided by the length of the window. It should be noted that the percentage variation is inversely proportional to conservation thus 0% variation indicates complete conservation. Lastly, eShadow statistically evaluates the MSAs to indentify highly conserved regions. It does so by implementing two statistical methods, a Hidden Markov Model (HMMI) and Divergence Threshold (DT). DT utilises a sliding window of predefined length, and calculates the number of matches within the window. If the match number reaches a defined threshold, then the sequence within that window region is identified as being significantly conserved. The HMMI does not use a sliding window but rather examines the MSAs by analysing the overall match and mismatch distributions.

2.4.5 Measures of sequence diversity and tests of neutrality - DnaSP Version 5.10.00

DnaSP is a software package dedicated to analysing polymorphisms from DNA sequences. Input data requires full haplotype sequence data for each individual. During this project DnaSP has been used for measures of sequence diversity (π) and tests of neutrality (Tajima’s D).

2.4.5.1 Sequence diversity - π

Measures of nucleotide diversity (π) represent the probability that two copies of a specific nucleotide will be different if selected randomly from a population or rather a set of sequences (Jobling *et al.* 2004). Values of zero indicate that all sites are monomorphic. The formulae for π is as follows:

$$\pi = n(\Sigma\ x_i x_j \pi_{ij})/(n-1)$$

n represents the number of sequences, x_i and x_j symbolise the frequencies of the $_i$th and $_j$th sequences, and π_{ij} is the proportion of nucleotide differences between the x_i and x_j sequences (Jobling *et al.* 2004). Plots of π have also been created in DnaSP using sliding windows of user defined size. A π value is calculated and plotted for the centre point of each window. The x axis corresponds to the nucleotide position within the input sequence and the y axis refers to the value of π. Peaks represent regions of high sequence diversity and troughs indicate regions of low sequence variation.

2.4.5.2 Test of neutrality - Tajima's D

The expected diversity level of any stretch of sequence is defined as θ (theta). Estimates of θ can be calculated by several different methods which utilise different observations of diversity such as number of segregating sites (total number of polymorphic loci), number of singletons (variants found to occur only once within the population), level of homozygosity, number of alleles or measures of π. Under neutral evolution or rather neutrality, all estimates of θ should be equal.

Tajima's D (Tajima 1989) is a statistical test used to compare estimates of θ based on π and the number of segregating sites (S) and is essentially a measure of the difference between them. If neutral evolutionary forces have acted upon the sequence in question, Tajima's D will be zero since π and S estimates will be equal. However significantly positive values of D may represent balancing selection or population subdivision, and significantly negative values may be indicative of positive selection or population growth (Jobling *et al.* 2004).

2.4.6 Haplotype tree construction - Network version 4.5.0.1.

The Network software has been used in this project to reconstruct phylogenetic trees, which are used to display the relationships between haplotypes. The nodes of the tree refer to a single haplotype and the node circumference is proportional to the frequency of the haplotype, and the branches indicate relationships between the nodes. It should however be noted that phylogenetic trees constructed using recombining autosomal data should be viewed with caution since they do not take into account recombination events.

2.4.7 Linkage Disequilibrium - Ldmax

Loci are said to be in linkage disequilibrium if two alleles from the separate loci are present together on the same chromosome more often than expected by chance.

The simplest measure of LD is a value known as D which calculates levels of LD in relation to allele and haplotype observed frequencies. Let us define two separate loci as A and B. The loci are biallelic and therefore the extent of LD between these loci can be examined. Under random segregation the observed frequency of the AB haplotype, termed as P_{AB}, will be a product of the A and B allele frequencies, termed P_A and P_B respectively. However if the two loci are in LD, the AB haplotype will be observed more often than expected given the allele frequencies. Thus the calculation of D is as follows;

$$D = P_{AB} – P_A P_B$$

If LD exists between the two loci then D values will be significantly different from zero irrespective of a positive or negative sign. It should however be noted that measures of D are no longer used since they are dependent on allele frequency and therefore independent D values can not be compared.

Values of D can however be modified in order to calculate a more robust measure of LD known as D′, which is the measure used throughout this project. D′ is calculated by taking the absolute value of D and dividing by its maximum value given the frequencies of the alleles. Pairwise D′ values range from zero to one. A value of one indicates that no separation of the two loci by recombination has been noted and they are therefore considered to be in complete LD. Nevertheless for low frequency alleles a value of one for D′ can be obtained in the absence of statistically significant association. Values less than one indicate that recombination has taken place between the two loci assuming the phase inference was correct.

Ldmax (University of Michigan. Center for Statistical Genetics) has been used to calculate pairwise measures of LD and does so by firstly estimating haplotype frequencies using the EM algorithm (Excoffier and Slatkin 1995).

2.4.8 Primer design - Primer3

Primer3 is a web based tool (http://frodo.wi.mit.edu/primer3/) used to design primer pairs from an input DNA sequence encompassing the desired region of PCR amplification. Parameters taken into account for primer design such as GC content, can be altered manually. Generally the default parameters were deemed adequate for the design of all primers used in this project.

2.5 List of web resources and tools

Pubmed -	http://www.ncbi.nlm.nih.gov/pubmed/
HapMap -	http://hapmap.ncbi.nlm.nih.gov/
UCSC genome browser -	http://genome.ucsc.edu/
NCBI dbSNP -	http://www.ncbi.nlm.nih.gov/projects/SNP/
rVista 2.0 -	http://rvista.dcode.org/
eShadow -	http://eshadow.dcode.org/
Primer3 version 0.4.0 -	http://frodo.wi.mit.edu/primer3/
ClustalW2 -	http://www.ebi.ac.uk/Tools/clustalw2/index.html
Heavens above -	http://www.heavens-above.com/countries.aspx
Ethnologue -	http://www.ethnologue.com/web.asp
World Health Organisation -	http://www.who.int/en/

Center for Statistical Genetics, University of Michigan (Ldmax)-
http://www.sph.umich.edu/csg/abecasis/gold/index.html
Database of Genomic Variants (DGV)- http://projects.tcag.ca/variation/

2.6 List of Buffers and Solutions

Agarose gel loading buffer: 15% Ficoll in dH_20 plus bromophenol blue and xylene cyanol dyes.

Abgene Buffer IV (10X) containing $MgCl_2$: 750mM Tris-HCl (pH8.8 at 25°C), 200mM $(NH_4)_2SO_4$, 0.1% v/v Tween 20®, 15mM $MgCl_2$.

TBE (5x): 0.44M Tris, 0.44M Boric Acid, 12.5mM EDTA, pH8.2-8.4.

Home made PCR clean up solution: 40% PEG-8000, 1 M NaCl, 2mM Tris-HCl (pH 7.5), 0.2mM EDTA, 3.5mM $MgCl_2$

Home made sequencing buffer: 200mM Tris HCl (pH9), 5mM $MgCl_2$.

NEB buffer 2: 10mM Tris HCl, 10mM $MgCl_2$, 50mM NaCl, 1mM dithiothreitol (pH 7.9 at 25ºC.

NEB buffer 3: 50mM Tris-HCl, 10mM $MgCl_2$, 100mM NaCl, 1mM dithiothreitol (pH 7.9 at 25ºC).

NEB buffer 4: 20mM Tris-acetate, 10mM magnesium acetate, 50mM potassium acetate 1mM dithiothreitol (pH 7.9 at 25ºC)

2.7 Commercial kits

ABI BigDye® Terminator mix v1.1 and v3.1

ABI Prism® SNaPshot™ Multiplex Kit: contents AmpliTaq® DNA polymerase, Fluorescently labelled ddNTPs, Reaction buffer.

2.8 Suppliers

New England Biolabs (NEB)

Applied Biosystems (ABI)

Abgene

Sigma Aldridge (oligos)

USB Corporation

Results Chapters

Chapter 3

MUC5AC variation in the 1946 Cohort

3 *MUC5AC* variation in the 1946 Cohort

Allergy related respiratory disease is primarily orchestrated by Th2 cells and it is the pro-inflammatory cytokines secreted by these cells that coordinate inflammation (see sections 3.1.1.1 and 3.1.3.1). In response to airway inflammation, the body has developed a homeostatic defence system whereby large quantities of airway mucus are secreted in an effort to expel the inflammatory causing agent (see section 3.1.2.1). In chronic airway disease, this mechanism is intensely active for lengthy periods of time and actually exacerbates disease symptoms. The mucus occlusions commonly seen in severe asthmatic airways (Kuyper *et al.* 2003) are testament to the detrimental effects that such a defence system can have when the balance is tipped.

At the cellular level, this mechanism is thought to be controlled by pro-inflammatory cytokines such as interleukin 13 (IL13) and tumour necrosis factor (TNF), which are known to upregulate *MUC5AC* expression (see section 3.1.3.1), and are present at high levels within the asthmatic airways (Bradding *et al.* 1994; Broide *et al.* 1992; Humbert *et al.* 1997; Naseer *et al.* 1997; Ying *et al.* 1997). It is thought that an increase in *MUC5AC* expression in progenitor cells results in their transdifferentiation to goblet cells in a process known as goblet cell metaplasia (GCM) (see section 3.1.2.3). The dramatic increase of goblet cells present in diseased airways results in the characteristic mucous phenotype. It is clear that MUC5AC plays an important role in allergy related disease etiology, we therefore hypothesise that genetic variation affecting the expression of this gene and the properties of the glycoprotein product may alter susceptibility to and severity of asthma and related respiratory disease.

This chapter describes the use of a longitudinal birth cohort for identifying associations between a *MUC5AC* marker and various allergy related and respiratory disease outcomes. Associations are also tested between these same outcomes and variations in genes involved in the inflammatory response with a view to ascertaining possible gene-gene interactions between *MUC5AC* and various inflammatory mediator genes.

The introductory section will discuss known etiology of allergy and asthma as well as MUC5AC and its relationships with various inflammatory mediators.

3.1 Introduction

3.1.1 Allergy

The term ‘allergy’ was coined by Von Pirquet in 1906. Allergy comes from the Greek word *allos* meaning "other" and *ergon* meaning "reaction". Over time allergy has been used interchangeably with the term atopy which generally refers to immunoglobulin (Ig) E mediated disease. In definition, an allergic reaction is an unnecessary immune response to harmless stimuli. Gell and Coombes (1963) divided these inappropriate immune responses into 4 hypersensitivity reaction groups. Categories are based on the time between stimulus trigger and symptom manifestation and the nature of the response. General use of the term allergy often refers to a hypersensitivity 1 reaction. This type of reaction is defined as an immediate response to otherwise harmless proteins known as allergens.

Asthma and hayfever (seasonal or perennial conjunctivorhinitis) are both hypersensitivity 1 allergic disorders. Other allergic disorders categorised in this way include; food anaphylaxis and some intolerances; allergy-related skin disorders such as urticaria and atopic eczema.

The prevalence of allergic disease has been increasing at an alarming rate (Peat and Li 1999) and is primarily thought to be a disorder of the westernized world (Cohn *et al.* 2004). However recent epidemiological studies have highlighted a significant increase in prevalence of allergy in Africa (Obeng *et al.* 2008). This rapid increase in prevalence can only be explained by changing environments.

3.1.1.1 Cellular and Molecular Mechanisms of Allergy: The Importance of Cytokines.

Allergic reactions begin with the processing and presenting of allergens by antigen presenting cells to T helper 2 (Th2) cells in the already sensitized individual. In response, the Th2 cells secrete a defining profile of cytokines, Interleukin (IL) 3, IL4, IL5, IL6, IL9, IL10, IL13 and GM-CSF, which orchestrates the allergic reaction. Infiltration of these Th2 cells into the affected tissue is the “immunological hallmark” of an atopic response (Kay 2000b).

IL13 and IL4 stimulate programmed B cells to produce allergen specific antibody IgE. The allergen then induces cross-linking of the IgE bound to mast cells and subsequently causes these cells to degranulate, releasing preformed inflammatory mediators such as histamine which result in chronic inflammation (Beirman *et al.* 1996).

IL3, IL5 and GM-CSF cytokines stimulate the production and maturation of eosinophil colonies in the bone marrow. These cytokines “prime” the eosinophils by increasing their metabolic activity which is characterised by cytotoxicity enhancement and mediator release (Beirman *et al.* 1996; Kay 2000a).

The circulating eosinophils are attracted to the correct location by the cytokines IL3, IL5 and GM-CSF which are consequently known as chemoattractants. Recruitment to the correct location is initially achieved by adherence of the eosinophils to the endothelial cells near to the site of inflammation. The endothelial cells close to inflammation become more adhesive for the eosinophils due to the local production of cytokines such as IL4 and IL13, which upregulate the production of proteins responsible for this specific adherence (Kay 2000a).

The eosinophils enter into the sites of inflammation by transendothelial migration. This process appears to only occur if the eosinophils have firstly been matured by the cytokines IL3, IL5 and GM-CSF (Moser *et al.* 1992).

The mechanisms by which the eosinophils degranulate remain uncertain. However the mediators released, Major Basic Protein (MBP) Eosinophil catonic protein (ECP) Eosinophil-derived neutrotoxin (EDN) and Eosinophil peroxidase (EPO), are cytotoxic to the bronchial epithelial cells and potent inflammatory substances.

3.1.1.2 Immune system priming

While allergen exposure is universal, individual immune responses to common allergens differ. Non-atopic individuals mount a modest T helper 1 (Th1) immune response to stimuli. They produce Immunoglobulin (Ig) G antibodies and interferon-γ (IFN-γ), a cytokine secreted by Th1 cells. Atopic individuals respond to allergens with an exaggerated Th2/IgE mediated immune response.

The type of immune response we mount against allergens is thought to be determined during early childhood. *In utero* the T helper lymphocytes of the fetus are weakly primed against common allergens. As a result the newborn demonstrates a Th2 dominated immune response. Critical immune developmental stages appear to occur during childhood before the age of 5 years (Kemp and Bjorksten 2003) which define future immune response profiles. Normally, a process known as ‘immune deviation’ should occur in the infant. Where the immune system deviates towards a Th1 dominated allergen response. One possible trigger for this deviation event is thought to be microbial exposure which forms the basis of the hygiene hypothesis, whereby it is believed that the increasing obsession of the Westernised world to prevent child contact with bacteria is thought to be the driving force behind increases in allergy prevalence. In atopic individuals, childhood ‘immune deviation’ has been averted, and their Th2 immunity has been programmed to respond with greater zest and specificity to the allergens (Holt *et al.* 1999; Kay 2000a).

3.1.1.3 Environmental Factors and Allergy Development

Several allergy-protective lifestyle factors and case studies have been used as proof of principle for the ‘hygiene hypothesis’. It has been observed in Europe that children who live on a farm in close proximity to animals, have a dog or have an older brother, achieve a greater degree of protection against atopy (Cookson 1999). Rural living appears to confer allergy protection in Africa as well as Europe, whereby the prevalence of allergic diseases is observably divided between rural and urban Africa, with a far greater incidence of allergy in the cities (Obeng *et al.* 2008).

There are other environmental factors that could explain why some individuals have a better chance of obtaining immune deviation. The dose of the allergen and age at exposure could be deciding factors for an individual’s subsequent immune profile. In fact Scandinavian children born within three months surrounding the birch pollen season are more likely to become allergic to this pollen in life (Cookson 1999).

3.1.1.4 Genetic Variation and Allergy Development

In concert with environmental factors, genetic variation is also likely to play a pivotal role in the development of allergy. It is therefore important to study polymorphisms in genes known to be upregulated or altered in expression during allergic episodes.

Inflammation is the defining symptom of allergy and therefore pro-inflammatory substances will obviously play an important role in allergic disease etiology. In fact, variation in a range of pro-inflammatory cytokine genes has previously been shown to be associated with allergic disease (Albuquerque *et al.* 1998; Apter *et al.* 2008; Black *et al.* 2009; Choi *et al.* 2009b).

It is also important to look at proteins that are upregulated by pro-inflammatory cytokines or are involved in the inflammatory pathway (Tamura *et al.* 2001). Thus *MUC5AC* is a good candidate gene for studies of respiratory allergic diseases such as asthma, since its expression is upregulated in the airways during inflammation. Thus the next part of this introductory section will focus on asthma and the role MUC5AC plays in asthmatic pathology.

3.1.2 Asthma

Asthma is a chronic inflammatory disease of the airways. Periodic bronchoconstriction, inflammation and mucus hypersecretion are all factors which lead to tightening of the airways and reduced airflow. Symptoms include wheezing, breathlessness and coughing. Such intermittent inflammatory episodes are known as asthma attacks. These attacks can be very short and last for just minutes, but some last for hours and even days. Long severe attacks are consistent with a life-threatening condition known as *status asthmaticus*, which can be characterised by mucus plugging and rupturing of the alveoli.

Stimuli such as respiratory viral infection, exercise, emotional stress, air pollutants and ozone are known to trigger asthma attacks. However it has been suggested that 80-90% of cases are allergy related and triggered by a variety of common allergens such as those from the house dust mite, grass pollen and animal danders (Cohn *et al.* 1998; Cookson 2002). When inhaled, the airborne allergens evoke a biphasic allergic reaction in the sensitized asthmatic airways. During the initial acute-phase, allergens induce an inappropriate Th2/IgE mediated immune response, ultimately leading to chronic airway inflammation (Holt *et al.* 1999). In the late-phase of the attack, the Th2 cytokines activate and recruit eosinophils. Mucus plugging and perivascular oedema are also conditions characteristic of the late-phase (Holt *et al.* 1999).

3.1.2.1 Mucus Hypersecretion in Asthmatic Airways

Mucus airway occlusion can often be fatal and a recent study of subjects who had died in *status asthmaticus*, showed that 95% of the patients had airway narrowing that ranged from 20-100% of the airway lumen due to mucus occlusions (Kuyper *et al.* 2003). Mucus plugging is thought to be a marker of asthmatic severity, and it is therefore essential to understand the process of mucus production in both healthy and diseased airways.

3.1.2.2 Cellular Source of Mucus Hypersecretion

MUC5AC and MUC5B are the major secretory mucins expressed in the airways (Kirkham *et al.* 2002). They give the respiratory mucus its characteristic gel-like and sticky properties. MUC5B is primarily expressed in the submucosal glands (Groneberg *et al.* 2002a) while the surface epithelial goblet cells are the principal secretors of MUC5AC (Groneberg *et al.* 2002b; Hovenberg *et al.* 1996). However during airway disease MUC5B can become aberrantly expressed in the goblet cells.

Although enlarged submucosal glands are evident in some respiratory diseases (Bai and Knight 2005; Fahy 2002; Kamio *et al.* 2005; Rogers 2004) the surface epithelium goblet cells are thought to be the principal contributor of excess mucin production during episodes of mucus hypersecretion. This is supported by the fact that the small and medium airways of *status asthmaticus* specimens have been shown to be occluded by mucus (Evans and Koo 2009), and within these regions, goblet cells are the only source of mucin production (Turner and Jones 2009). Even in mild and moderate asthmatic airways, the mean volume of stored goblet cell mucin was three times higher in the cases than the controls (Ordonez *et al.* 2001).

3.1.2.3 The Mucous Phenotype

In healthy human and murine airways, goblet cells are rarely seen distal to the trachea/in small airways (Curran and Cohn 2009; Williams *et al.* 2006). However during episodes of inflammatory disease, mucous cell numbers within these regions are dramatically increased. Abnormally large numbers of goblet cells within the small and medium airways is the defining characteristic of a ‘mucous phenotype’. This condition could be ascribed to either goblet cell ‘hyperplasia’, defined by proliferative events, or goblet cell ‘metaplasia’ in which pre-existing cells adopt an alternative phenotype (Williams *et al.* 2006).

Mouse models of airway inflammation have been extensively used to attempt to solve the conundrum of ‘hyperplasia’ or ‘metaplasia’. Evidence at present appears to favour goblet cell metaplasia (GCM) as the primary cause of the mucous phenotype. Epithelial proliferation is not significantly increased in allergen challenged mice at the point in which excessive numbers of goblet cells first appear, suggesting that no increase in cell number is taking place and mucin production has therefore been initiated in resident cells (Evans *et al.* 2004). A mouse study has also reported that the mucous phenotype following sensitization could not be explained by an increase in epithelial cell numbers since the number of cells per unit surface remained constant. This expansion in mucous cell number could be entirely compensated for by the dramatic decrease in Clara and ciliated cells (Reader *et al.* 2003), implying a process of ‘transdifferentiation’, whereby the Clara and ciliated cells have assumed a goblet-like phenotype. The idea of Clara progenitor cells has also been demonstrated in human smokers. Smoker airways have shown an increase of mucous cell number in locations usually occupied by Clara cells (Lumsden *et al.* 1984).

Both clara and ciliated cells have been implicated as progenitor cells in mouse airways with mucous phenotypes. Direct evidence of airway epithelial cell transdifferentiation has been obtained by ultrastructural analysis of allergen challenged mice, whereby cells exhibit both Clara and goblet characteristics. Co-localization of the ciliated cell marker beta-tubulin and MUC5AC has also been seen in mice infected by the Sendai virus (Turner and Jones 2009).

Cells staining positive for MUC5AC are defined as mucous cells, and therefore increased expression of MUC5AC is the marker of GCM and transdifferentiation (Fahy 2002). MUC5AC has been shown to be the mucin most upregulated in the asthmatic airways of both humans and animals (Kirkham *et al.* 2002; Ordonez *et al.* 2001; Young *et al.* 2007) and it is therefore important to understand why *MUC5AC* becomes aberrantly expressed in progenitor cells.

As mucins give mucus its rheological properties, alterations in the *MUC5AC* expression patterns are likely to change the consistency, viscosity and stickiness of the mucus. Therefore any genetic variation that affects *MUC5AC* expression or protein structure is likely to be functional.

3.1.3 Mediators of GCM and *MUC5AC* Expression

3.1.3.1 Th2 cells mediate GCM

Th2 cells mediate inflammation and are also known to mediate mucus hypersecretion/GCM. A mucous phenotype results when Th2 cells are transferred into the lungs of naïve mice (Cohn *et al.* 1997; Cohn *et al.* 1998). Thus Th2 mediated GCM in humans is likely to have a profound effect in asthmatic airways where a significant enrichment of Th2 cells has been noted in bronchoalveolar lavage fluid (BALF) (Robinson *et al.* 1992).

3.1.3.2 Th2 cytokines regulate *MUC5AC* expression

Various pro-inflammatory cytokines such as, IL1B, IL4 and IL13 have been shown to upregulate *MUC5AC* expression (Gray *et al.* 2004b; Gray *et al.* 2004a; Koo *et al.* 2002) and GCM. A mouse study has suggested that IL-13 may be the only cytokine essential for allergy mediated GCM. Whittaker et al administered IL-13 +/+ or IL-13 -/- Th2 cells to both IL13 wild type and IL13 knock out mice, followed by antigen challenge. Only the mice with a complete IL13 blockade showed no epithelial cell mucus staining in their otherwise inflamed airways. Increased mucus production was evident after antigen challenge in all mice that were either IL13 wildtype or had been administered IL13 positive Th2 cells (Whittaker *et al.* 2002). Thus the elevated IL13 levels seen in human asthmatic lungs (Humbert *et al.* 1997; Naseer *et al.* 1997; Ying *et al.* 1997) are likely to exacerbate mucus production.

3.1.3.3 IL13 signalling

IL13 and IL4 transduce their signals through a common receptor which is a heterodimer composed of the chains IL4 Receptor α (IL4Rα) and IL13 Receptor α1 (IL13Rα1). When cytokine attachment occurs the chain associated signalling intermediates Janus kinase (JAK) 1 and Tyrosine kinase (TYK) 2, are phosphorylated and subsequently activated. Phosphorylation of other tyrosine residues located within the intracellular domain of IL4Rα occurs allowing Insulin receptor substrate (IRS) and Signal transducer and activator of transcription 6 (STAT6) to associate at these sites. IRS and STAT6 themselves become phosphorylated leading to the activation of further downstream signalling pathways. IRS activates the phosphatidylinositol 3 kinase (PI3K) and Ras/MAPK pathways important for cell survival/proliferation and transcriptional regulation respectively. Phosphorylated STAT6 also activates transcription but does so directly by firstly dimerising and then translocating to the nucleus (Hershey 2003; Kasaian and Miller 2008).

The IL13Rα1 60 amino acid intracellular domain does not appear to play a very important role in IL13 signal transduction, however this region in highly conserved with 98% amino acid sequence homology between humans and mice and is therefore likely to be of functional importance (Hershey 2003).

IL13 also interacts with another receptor, IL13 Receptor α2 (IL13Rα2), which is 37% homologous to IL13Rα1. IL13Rα2 binds IL13 with a much greater affinity than IL13Rα1 and it appears that it does not require the aid of another chain for efficient cytokine attachment, although this remains to be proved (Hershey 2003).

IL13Rα2 has no obvious role in signal transduction and has therefore been proposed as a decoy receptor, acting to absorb excess IL13 without inducing changes within the cell. A mouse study has shown that in the absence of functional IL13Rα2, IL13-mediated inflammation, mucus cell metaplasia and mucin gene expression are increased (Zheng *et al.* 2008). This supports the idea that the IL13Rα2 acts as a sink for excess IL13, and in its absence IL13 induces pathways that increase the expression of inflammatory response genes such as mucins.

A recent study has however suggested that IL13Rα2 could mediate signalling through AP1 (Fichtner-Feigl *et al.* 2006). An AP-1 binding site has been noted in the *MUC5AC* promoter region between nucleotides -3576/3570 which is known to be involved in the upregulation of *MUC5AC* transcription during *Haemophilus influenzae* (NTHi) infection (Chen *et al.* 2004).

This AP-1 binding site is also thought to mediate *MUC5AC* transcriptional upregulation in response to cigarette smoke. Gensch et al have shown that the increased mucus production in the lungs of smokers is paralleled by elevated *MUC5AC* mRNA levels (Gensch *et al.* 2004). Using a luciferase reporter gene attached to 3700 nucleotides of the immediate upstream *MUC5AC* sequence, Gensch and colleagues were able to show that *MUC5AC* promoter activity is significantly increased by cigarette smoke. A major smoke response element was subsequently identified at -3700/-3337 nucleotides relative to the transcription start site and mutagenesis of all transcription factor binding sites within this refined region identified AP-1 as the smoke response element (Gensch *et al.* 2004).

Interestingly it has been implicated that AP-1 along with the glucocorticoid response element (GRE) and nuclear factor kappa B (NFκB) binding motifs are the sites at which glucocorticoids/receptor complexes bind to suppress gene expression. As all three motifs have been identified in the *MUC5AC* regulatory region (Gensch *et al.* 2004; Li *et al.* 1998), so inhaled glucocorticoid treatment could directly repress MUC5AC production.

Large intracellular pools of IL13Rα2 have been identified within a small subset of cells including respiratory epithelial cells. The intracellular IL13Rα2 is able to mobilise to the cell surface after treatment with the Th1 defining cytokine IFN-γ (Daines and Hershey 2002).

Cell surface mobilisation has also been noted in the presence of IL4 and IL13 at high concentrations (Yoshikawa *et al.* 2003). This could possibly explain why bronchial epithelial cell cultures grown at air liquid interface (ALI), exhibit increased goblet cell density (GCD) when exposed to low concentrations but not higher concentrations of IL13 (Atherton *et al.* 2003).

3.1.3.4 Pathways involved in IL13 mediated *MUC5AC* upregulation- MAPK

Subject to IL13 binding to its receptor located on the respiratory epithelial cells, the exact mechanisms by which IL13 induces *MUC5AC* gene expression and subsequent GCM remain unclear. However much evidence suggests that the MAPK pathway plays a pivotal role. By inhibiting key components in the MAPK and PI3K pathways (the IRS activated pathways) in human bronchial epithelial ALI cultures during IL13 treatment, GCD in these cultures was significantly reduced (Atherton *et al.* 2003). A similar result was obtained by inhibiting the p38 MAPK pathway in mouse ALI culture cells (Fujisawa *et al.* 2008). Using western blotting Fujisawa et al show that STAT6 is phosphorylated during the first 20 minutes after IL13 treatment and lasts for 48 hours. However p38 MAPK phosphorylation is delayed and does not occur until 36-48 hours post treatment. The upregulation of *MUC5AC* gene expression was also delayed with a significant increase in transcription not occurring until 48 hours post IL13 treatment. The synchronisation of p38 MAPK activation and *MUC5AC* upregulation identifies a potential direct link between the p38 MAPK pathway and *MUC5AC* transcriptional regulation (Fujisawa *et al.* 2008).

It has been suggested that STAT6 may directly mediate IL13-induced *MUC5AC* gene expression. This notion is supported by the fact that STAT6 deficient mice who overexpressed IL13 did not possess a mucous phenotype (Kuperman *et al.* 2002), and a knockdown of STAT6 expression in human ALI culture cells prevented IL13 from inducing GCM (Turner and Jones 2009). However STAT6 does not appear to bind directly to the *MUC5AC* promoter because there is no STAT6 binding motif within the *MUC5AC* immediate promoter (Li *et al.* 1998) and no conserved STAT6 motifs could be located within the whole *MUC2-MUC5AC* intergenic region (Young *et al.* 2007). Why therefore is IL-13 unable to induce GCM in STAT6 knock out mice? In an attempt to answer this problem, Fujisawa and colleagues have suggested that in order for *MUC5AC* gene expression to be upregulated by IL13, STAT6 activation must initially occur which is followed by *de novo* protein synthesis of an unknown protein which results in p38 MAPK activation (Fujisawa *et al.* 2008). This hypothesis is supported by the timing of pathway activation whereby STAT6 is immediately activated by IL13 followed by a delay in p38 MAPK activation and *MUC5AC* transcriptional upregulation.

3.1.3.5 Other pathways involved in MUC5AC regulation- EGFR

Exogenous stimuli such as cigarette smoke, and internal mediators such as the pro-inflammatory cytokines IL13 and TNF (Schmiegel *et al.* 1993; Takeyama *et al.* 1999), have been shown to increase the expression of specific Epidermal growth factor receptor (EGFR) ligands. Thus one may expect significantly elevated EGFR levels in the lungs of smoke-induced COPD sufferers and in allergic asthmatic airways where IL13 and TNF are thought to be in abundance (Bradding *et al.* 1994; Broide *et al.* 1992; Humbert *et al.* 1997; Naseer *et al.* 1997; Ying *et al.* 1997).

EGFR is a membrane glycoprotein that plays a pivotal role in airway epithelial cell repair and differentiation (Curran and Cohn 2009) EGFR and its ligands, Epidermal Growth Factor (EGF) and Transforming Growth Factor alpha (TGF-α), are scarcely found in the adult respiratory epithelium, however they have been shown to be more abundantly present in asthmatic airways (Amishima *et al.* 1998). EGFR activation by its ligands has been shown to stimulate *MUC5AC* gene expression and MUC5AC protein production in human epithelial cell cultures (Takeyama *et al.* 1999). This has been supported in an *in vivo* study in which EGFR and a homologous transmembrane receptor ErbB3, were shown to be significantly upregulated along with MUC5AC in the epithelial cells of smokers compared with non-smokers (O'Donnell *et al.* 2004).

3.1.4 *MUC* Asthma Association

There is previous reported evidence of association between the *MUC2* tandem repeat polymorphism and asthma. In a small asthmatic cohort (n = 100) atopic individuals with longer length alleles appeared to be at lower risk of developing asthma (Vinall *et al.* 2000). Biologically this association appears to be rather unlikely since MUC2 is found at extremely low levels in the airways. *MUC2* gene expression does appear to be upregulated in asthmatic airways although mRNA levels are still 20-fold less than that of *MUC5AC* (Evans and Koo 2009). MUC2 is not easily detectable at protein level (Hovenberg *et al.* 1996; Ordonez *et al.* 2001).

The *MUC5AC* gene is located directly adjacent to *MUC2* on chromosome 11. Linkage disequilibrium has been shown to extend from *MUC2* to *MUC5B* in the 11p15.5 mucin gene complex (Rousseau *et al.* 2007). Thus the association seen between MUC2 and asthma could in fact be a consequence of linkage disequilibrium between the *MUC2* TR and a causative allele in *MUC5AC*.

3.1.5 Inflammatory Mediators – Polymorphisms and asthma association

Several inflammatory response proteins, for example the Th2 cytokines and EGFR, are known to upregulate MUC5AC. Therefore risk polymorphisms within these inflammatory mediators may act in combination with MUC5AC variation to enhance disease risk and/or severity. The final part of this introductory section will therefore explore previously identified genetic associations between asthma and other inflammatory response genes.

3.1.5.1 IL13

Two *IL13* SNPs have been extensively studied with respect to inflammatory disease. IL13 rs1800925 is located within the promoter (-1024C>T) and *IL13* rs20541 is a non-synonymous SNP located within exon 4 (Arg110Gln). Although a high level of LD (Black *et al.* 2009) has been reported between these SNPs, in general the exonic SNP has been shown to exhibit a greater degree of association with inflammatory diseases.

The *IL13* exonic SNP alters an amino acid and therefore an arginine or glutamine can be present at position 110. Computer modelling suggests that this amino acid substitution is likely to be functional as the amino acid at position 110 appears to be important for IL13 tertiary structure. Modelling also shows that the residue at this position directly interacts with the IL13 receptor, and it is thought that a glutamine at this position would enhance binding (Heinzmann *et al.* 2000). Functional studies have confirmed glutamine enhanced binding by showing that having a glutamine at position 110 is significantly more active in inducing STAT6 phosphorylation when compared to the ‘wild type’ IL13 (Vladich *et al.* 2005).

The *IL13* exonic SNP is significantly associated with IgE serum levels and homozygosity of the glutamine allele results in the highest total IgE levels (Graves *et al.* 2000; Leung *et al.* 2001; Maier *et al.* 2006; Wang *et al.* 2003). As previously mentioned, enhanced IgE levels are characteristic of asthmatic airways, it is therefore not surprising that the glutamine allele confers increased risk of asthma (Black *et al.* 2009; Heinzmann *et al.* 2000).

In a Dutch study, significant association was shown between the *IL13* promoter SNP (-1024C>T) and asthma, however the Arg110Gln SNP did not show association (Howard *et al.* 2001). The rare TT promoter homozygote has been acknowledged as the risk genotype for allergic asthma. The potential functionality of this promoter SNP has also been highlighted in a study of T cells. The cells were firstly stimulated to produce high amounts of IL13 and then anti-CD2 was added to inhibit this production. Treated T cells from TT homozygote individuals, significantly resisted this inhibition compared to the CT and CC cells (van der Pouw Kraan TC *et al.* 1999). This indicates that the *IL13* promoter SNP is located within a functionally important site and the risk genotype TT somehow results in a more robust expression pattern.

3.1.5.2 EGFR

A CA microsatellite is located within the CpG rich intron 1 of *EGFR*. Repeat numbers range from 8 to 30, and the transcriptional activity of *EGFR* appears to decline with increasing numbers of repeats. In fact the presence of 21 CA repeats has been demonstrated to reduce *EGFR* transcription by 80% (Gebhardt *et al.* 1999).

The shorter alleles (≤ 16 repeats) have been identified as risk factors in Japanese asthmatics and in particular severe asthmatics (Wang *et al.* 2006).

The EGFR extracellular region is functionally important because this is where binding of the ligands, EGF and TGF-a, occurs. A SNP (rs2227983) at codon position 521 results in an arginine to lysine substitution located within the extracellular subdomain IV. Functional studies have shown that the presence of the lysine allele weakens EGFR growth response to its ligands (Moriai *et al.* 1994).

3.1.5.3 TNF

TNF is a pro-inflammatory cytokine and its gene is located within the class III region of the Major Histocompatibility Complex (MHC) on chromosome 6p21.

The *TNF* promoter SNP rs1800629, more commonly referred to as -308 in the literature, is biallelic with alleles frequently termed *TNF1* (common G) and *TNF2* (rare A). Many studies of various ethnic groups, have noted that the *TNF2* allele confers asthmatic risk and the *TNF1* allele is protective (Chagani *et al.* 1999; Gupta *et al.* 2005; Shin *et al.* 2004; Wang *et al.* 2004; Witte *et al.* 2002; Wu *et al.* 2007b), and this has been confirmed in a meta-analysis (Gao *et al.* 2006). In one study the *TNF11* genotype is shown to be significantly protective of wheeze as well as asthma (Li *et al.* 2006), and the *TNF2* allele has also been identified as a risk allele with respect to chronic bronchitis (Huang *et al.* 1997). This pattern of association is not however universal and in a study of Australian Caucasian children *TNF1* was identified as the risk allele (Albuquerque *et al.* 1998).

In a study of Mexican children *TNF2* was again shown to be significantly associated with asthma, however this association was only apparent in the subgroup of cases that did not have parents who smoked (RR = 2.06 p = 0.0097) (Wu *et al.* 2007b). This finding interestingly highlights the fact that environmental exposures such as cigarette smoke have a profound effect on the outcome of susceptibility polymorphisms.

3.1.5.4 IL1RN and IL1B

The IL1 cytokines, IL1A, IL1B and IL1Ra (Interleukin 1 receptor antagonist) are produced by cells such as monocytes and macrophages and their genes are located in a cluster on chromosome 2q14.2. IL1a and b are pro-inflammatory cytokines, which bind to IL1 receptor 1 (IL1R1) located on target cells and induce signal transduction. IL1Ra, coded for by the *IL1RN* gene, also binds to IL1R1. However IL1Ra is an anti-inflammatory cytokine and on binding to its receptor does not induce a signal and therefore behaves as a natural antagonist of the IL1 pro-inflammatory cytokines, by competitive inhibition. Achieving an optimal ratio of pro and anti- inflammatory cytokines is likely to be of the great importance within the airways and it seems probable that during respiratory disease the delicate balance has been tipped in favour of the pro-inflammatory cytokines. The elevated IL1Ra serum levels seen during asthma attacks (Yoshida *et al.* 1996) are thus likely to be a natural response to inflammation.

An 86 base pair VNTR has been identified within intron 2 of the *IL1RN* gene. In Europeans the repeat commonly occurs four times, however five alleles have been identified ranging in size from 2-6 repeats (see table 3.1 for allele nomenclature). This polymorphism is a good candidate for functional significance as it contains predicted protein binding sites for an α-interferon silencer, a β-interferon silencer and an acute phase response element (Tarlow *et al.* 1993).

The *IL1RN* VNTR has been shown to be significantly associated with asthma. The *IL1RN**2 allele is significantly associated with non-atopic asthma in a Japanese adult population (OR = 5.71 p = 0.0018). This study also identifies a significant association between *IL1RN**2 and lower IL1Ra serum levels in atopic and non-atopic asthmatics (Mao *et al.* 2000). However studies of healthy IL1Ra serum levels have shown the opposite effect, that is the *IL1RN**2 allele is associated with significantly higher levels of IL1Ra (Danis *et al.* 1995; Hurme and Santtila 1998). These discrepancies may be explained by the fact that the *IL1RN**2 allele only appears to confer a plasma level enhancing affect if it is in combination with the rare allele of the *IL1B* -511 promoter SNP (rs16944) (Hurme and Santtila 1998), or by different interactions of trans activators/inhibitors in asthma and non-asthma.

The importance of *IL1RN* and *IL1B* haplotypes has also been noted in a study of smoker lung function. The *IL1RN* VNTR was not shown to be independently associated with a decline in lung function, but when combined as *IL1RN* VNTR/ *IL1B* -511 haplotypes, associations were noted. The *IL1RN**1/*IL1B* -511T was significantly increased in the smokers with rapid decline in lung function, while there was a significant decrease of the *IL1RN**2/*IL1B* -511T haplotype in this group (Joos *et al.* 2001).

The *IL1RN* VNTR and *IL1B* -511 SNP have been shown to be independently associated with asthma in a study of childhood asthma in a Turkish population (Zeyrek *et al.* 2008). However the VNTR genotype A1A1 was identified as the risk genotype whereas the A1A2 genotype was shown to be protective and the same result was seen in an Egyptian population (Settin *et al.* 2008).

Table 3.1 *IL1RN* intron 2 VNTR allele nomenclature and frequencies.

Allele Name	Alternative Name	# of Repeats	PCR Product (bp)	Frequency
*IL1RN**1	A1	4	410	0.736
*IL1RN**2	A2	2	240	0.214
*IL1RN**3	A3	5	500	0.036
*IL1RN**4	A4	3	325	0.007
*IL1RN**5	A5	6	595	0.007

Adapted from (Tarlow *et al.* 1993)

3.2 Hypothesis and Aims

Since inflammatory mediators abundant during allergy related respiratory disease have been shown to alter *MUC5AC* expression (see section 3.1.3), the general aim of this chapter is to investigate the relationship between *MUC5AC* variation and allergy related outcomes in a European longitudinal birth cohort.

We hypothesise that MUC5AC variants will respond differently within allergic airways and therefore propose that functionally significant *MUC5AC* genetic variants might confer altered susceptibility to or severity of allergy and respiratory disease phenotypes.

This chapter focuses on a single *MUC5AC* SNP, typed in the 1946 birth longitudinal cohort and studied in relation to allergy and respiratory outcomes. This SNP was choosen since it has a high minor allele frequency and was one of the few well established SNPs at the start of the study. Gene-gene interactions are also explored to further characterise the relationship between inflammatory markers and *MUC5AC* with respect to these same disease outcomes.

3.3 Results

3.3.1 *MUC5AC* data

In order to investigate the relationship between *MUC5AC* variation and allergic respiratory disease, 2 polymorphisms have been typed within this gene in the 1946 Longitudinal Birth Cohort; the *MUC5AC* VNTR, previously typed by Southern blotting and a synonymous SNP located in exon 19 of the 3′ region (rs1132440). As described in the main introduction of this thesis, it is already known that there is variation in the tandem repeat region of *MUC5AC* which is thought to affect the length of the glycosylated domain (VNTR). However even now in 2010 the human genome sequence is not complete for *MUC5AC*, and due to a paucity of SNP data at the outset of this project, a single SNP (rs1132440) was selected for analysis. *MUC5AC* rs1132440 does not cause an amino acid substitution, thus we do not anticipate functionality. However this SNP was chosen because of its high minor allele frequency and since it is not in complete linkage disequilibrium with the VNTR polymorphism it was predicted to capture more of the overall *MUC5AC* diversity. Genotyping was performed during this project as part of a single base extension (SBE) multiplex assay (see figures 3.1 and 3.2) and genotypes were inferred from observed phenotypes (G, GC and C were interpreted as GG, GC and CC respectively).

3.3.2 Other mucin data within the 11p15.5 gene complex

The cohort has also been typed for other polymorphisms within the 11p15.5 gene complex; *MUC2* VNTR was genotyped by Southern blotting; three exonic *MUC5B* SNPs (rs2672785, rs2075853 and rs2075859) genotyped as part of the SBE multiplex. The *MUC5B* data has been used as part of a previous PhD project from the laboratory (Loh 2007) and will therefore only be discussed briefly.

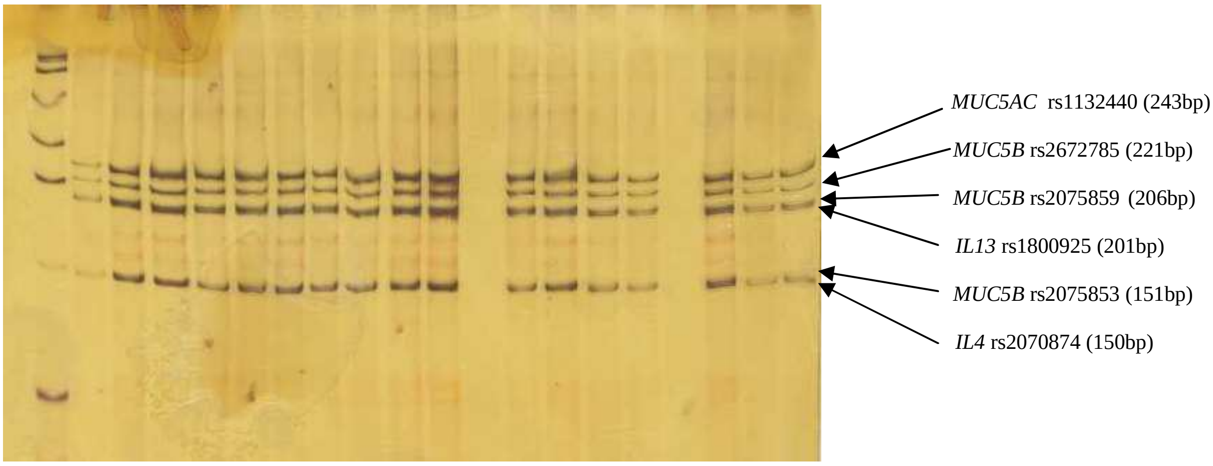

MUC5AC rs1132440 (243bp)
MUC5B rs2672785 (221bp)
MUC5B rs2075859 (206bp)
IL13 rs1800925 (201bp)
MUC5B rs2075853 (151bp)
IL4 rs2070874 (150bp)

Figure 3.1 Silver stained 12% acrylamide gel, used to visualise PCR multiplex. 1.5µl of multiplex PCR product loaded. Corresponding PCR products for each polymorphism are labelled. Each lane corresponds to the multiplexed PCR products of one individual.

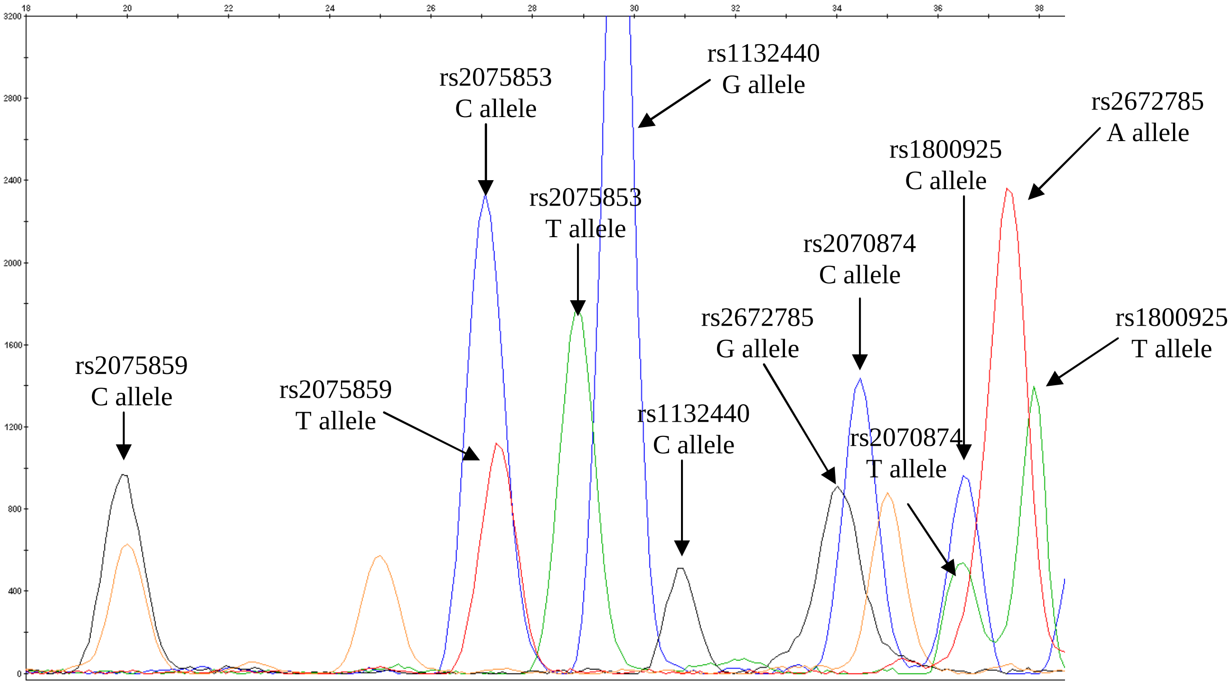


Figure 3.2 Typical chromatogram showing SBE results for the multiplex (genotype data for six SNPs). The sample represented in this figure is heterozygous for all six SNPs.

3.3.2.1 Allele Frequencies: 11p15.5 polymorphisms

A minor allele frequency (MAF) of 0.42 (G) (see table 3.2) was obtained for *MUC5AC* rs1132440 in the Cohort sample. This frequency is similar to European CEPH data submitted to the NCBI SNP database which reports a MAF of 0.39.There was no significant deviation of observed genotype frequencies from those expected under Hardy Weinberg Equilibrium (HWE).

Table 3.2 Minor Allele Frequencies (MAF) for the *MUC5AC* and *MUC5B* polymorphisms typed on the 1946 cohort.

Gene	SNP location	SNP ID	MAF	Total
MUC5AC	exon 19 (3′ region)	rs1132440	0.42 (G)	2910
MUC5B	exon 2 (5′ region)	rs2672785	0.20 (G)	2800
MUC5B	exon 3 (5′ region)	rs2075853	0.07 (T)	2801
MUC5B	exon 9 (5′ region)	rs2075859	0.38 (T)	2797

The *MUC5AC* VNTR has two common alleles (a and b), and several rare length alleles (termed c-k) which are both longer and shorter than the common repeat lengths. To simplify analysis a triallelic model was established for the tandem repeat genotypes: a, b and r with allele frequencies of 0.77, 0.22 and 0.01 respectively, where r includes all rare alleles.

Three *MUC5B* exonic SNPs were genotyped on the cohort as part of the SBE multiplex. SNPs in exons 2 (rs2672785) and 3 (rs2075853) are non-synonymous causing amino acid changes glutamic acid to glycine and arginine to tryptophan respectively. The minor allele frequencies obtained in this cohort for all three SNPs are in good accordance with the HapMap CEPH frequencies.

The *MUC2* VNTR has many alleles ranging from 3.21 to 11.64 Kb in this data set with mean and mode lengths of 7.33 and 7.61 Kb respectively. This was therefore treated as a continuous variable. The allele length data is non-normally distributed thus non-parametric analyses will be performed on this variable.

3.3.2.2 LD within the 11p15.5 complex

Pairwise LD measures were determined for the two *MUC5AC* and three *MUC5B* markers and are presented in the table 3.3 as D′ and chi squared p values. As reported previously adjacent SNPs are in tight LD and for the purpose of this study it should be noted that highly significant LD exists within the *MUC5AC* gene and only a small amount of recombination seems to have occurred between the tandem repeat and rs1132440 SNP (D′ of 0.941).

LD also extends across the two *MUC5* genes as *MUC5AC* rs1132440 and the furthest *MUC5B* SNP in exon 9 rs2075859 are significantly associated (p = 0.009). However, the very small D′ value (0.071) indicates that considerable recombination has occurred between these two markers.

Table 3.3 Pairwise linkage disequilibrium (LD) for *MUC5AC* and *MUC5B* markers.

a

rs1132440	**0.000**			
rs2672785	0.751	**0.000**		
rs2075853	0.047	0.264	**0.000**	
rs2075859	0.803	**0.009**	0.968	**0.000**
	5ACTR	rs1132440	rs2672785	rs2075853

b

rs1132440	**0.941**			
rs2672785	0.014	**0.198**		
rs2075853	0.065	0.062	**0.910**	
rs2075859	0.016	**0.071**	0.001	**0.950**
	5ACTR	rs1132440	rs2672785	rs2075853

LD measures were calculated by LDmax using the 1946 cohort data (n = 2920). Significance of association is shown in figure a, as chi squared p values. Measures of recombination are shown as D′ values in figure b. Where 5ACTR refers to the *MUC5AC* VNTR. SNP rs1132440 is located within *MUC5AC*. SNPs rs2672785 and rs2075853 are located within *MUC5B*.

3.3.2.3 LD between the *MUC5AC* and *MUC2*

This dataset was also used to examine LD between the *MUC2* VNTR and both *MUC5AC* markers. Because of the multiallelic nature of the *MUC2* TR lengths, LD analysis could not be performed with the Ldmax software. Therefore the *MUC2* TR allele distributions were compared between *MUC5AC* rs1132440 genotype groups and between *MUC5AC* VNTR genotype groups (see figure 3.3 not shown for VNTR). A Kruskal-Wallis test was used to compare the medians of the different distributions. Using the binned *MUC2* VNTR data, as seen in figure 3.3, a significant association with *MUC5AC* rs1132440 was identified (p = 0.013), although significance was not reached when the *MUC2* VNTR was treated as raw data (p = 0.084). Statistically significant association was however noted between the *MUC2* TR and the *MUC5AC* TR markers tested as raw data. Thus a general pattern of association can be seen to extend from *MUC2* to *MUC5AC*, which is supported by a previous study from our group (Rousseau *et al.* 2007).

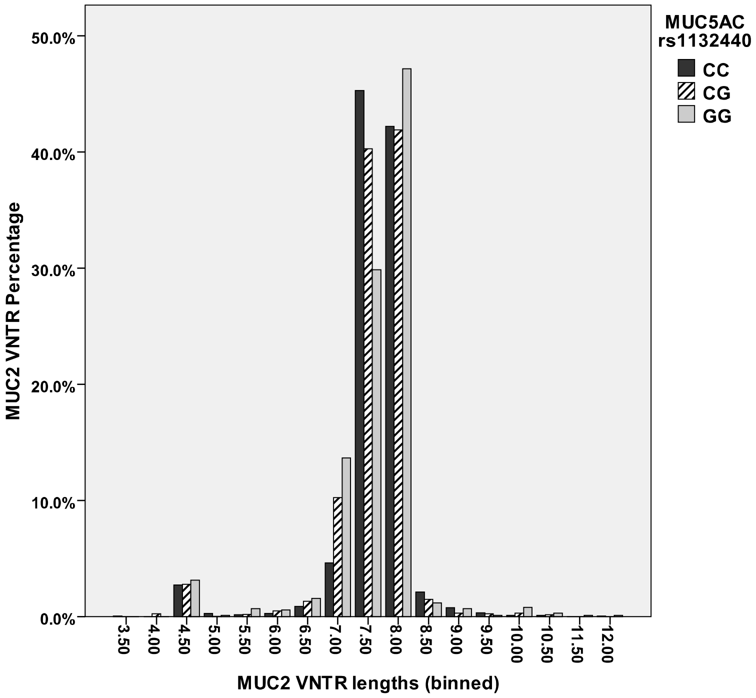


Figure 3.3 Bar chart to show the distribution of *MUC2* TR binned allele lengths with respect to *MUC5AC* rs1132440 genotypes. The medians of the three distributions (one for each *MUC5AC* genotype group) were shown to be significantly different with a p value of 0.013 (Kruskal-Wallis test). This shows that markers *MUC2* TR and *MUC5AC* rs1132440 are significantly associated which is reflective of linkage disequilibrium.

3.3.3 Respiratory Measures and Outcomes

Respiratory outcomes and measures available as 1946 birth cohort data are detailed in table 1 shown in appendix 2. The *MUC5B* SNP data had been previously analysed in the cohort for association with respiratory outcomes. Associations of marginal significance were observed with hayfever, allergy and wheeze (Loh 2007). However in this project the recoded ‘ever’ and ‘never’ respiratory variables, taken from Black et al, were used and none of the *MUC5B* associations remain statistically significant.

The *MUC2* VNTR data set was analysed using a Mann-Whitney test for association with bronchitis, wheeze, hayfever, allergy and asthma outcomes. The data are shown in histogram form in figure 3.4. It is noteworthy that although significance is not reached in any comparisons, there is a trend towards longer *MUC2* TR alleles in the asthma and wheeze affected groups, which is not in agreement with a previous study whereby the longer alleles were shown to confer protection against asthma in an atopic sample set.

Thus all further association analyses detailed in this chapter will concern *MUC5AC* variation, and in particular *MUC5AC* rs1132440. All tested phenotypic variables were non-independent and concerned solely with respiratory disease, function and allergy.

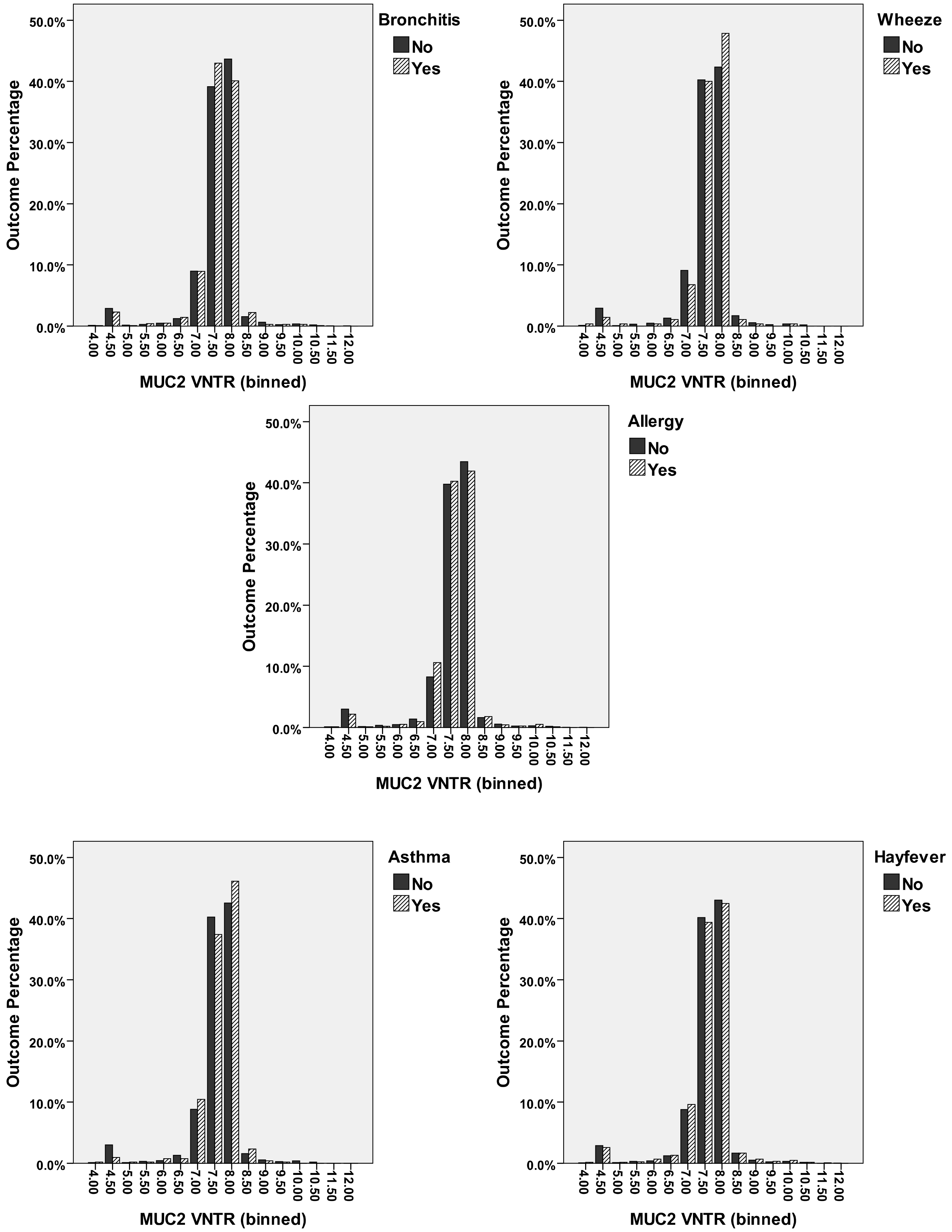


Figure 3.4 Histograms of *MUC2* TR allele lengths in the affected and unaffected groups of the five outcomes, bronchitis (bron89r), wheeze (wzy89c), asthma (evasthm), hayfever (evhay) and allergy (evallerg). The distributions of *MUC2* TR allele lengths in affected and unaffected groups have been analysed using a Mann-Whitney U test of association. Significance was not reached for any outcome.

3.3.3.1 Contingency tables - goodness of fit test

Contingency tables were constructed in order to compare the distribution of genotypes between the; ‘affected’ and ‘unaffected’ groups with respect to disease variables, and ‘yes’ or ‘no’ groups for the confounder variables.

As shown in table 3.4 *MUC5AC* rs1132440 genotypes counts (coded as M5ACX) show statistically significant association with bronchitis (bron89r), wheeze (wzy89c), hayfever (evhay) and allergy (evallerg), and is marginally associated with asthma (evasthm) (Pearsons Chi squared). However *MUC5AC* rs1132440 allele counts were only shown to be significantly associated with bronchitis and wheeze (p = 0.026), with an underrepresentation of the rarer allele (G) in the ‘affected’ groups. No significant associations were identified between rs1132440 and lung functions measures such as FEV.

The *MUC5AC* TR genotype variable (coded as M5ACTRY) only shows marginally significant association with the hayfever outcome. It must be noted that the *MUC5AC* TR data set is smaller than the rs1132440 data set. The discrepancy in n values is due to the fact that the Southern blot method used to genotype the VNTR requires good quality DNA and therefore blood DNA had to be used, whereas the SNP could be genotyped with standard buccal DNA. Fewer participants gave permission for the usage of blood and thus the VNTR sample size is substantially smaller.

The two *MUC5AC* markers (TR and rs1132440) are in LD, and we might therefore expect the significant rs1132440 associations to be reflected by the TR variable, which is not the case (see table 3.4). Therefore analysis of rs1132440 was repeated on the smaller VNTR data set in order to assess whether the association remained significant on the smaller dataset. All outcomes previously shown to be significantly associated with rs1132440 in the full data set were indeed still statistically significant in the reduced set. *MUC5AC* rs1132440 appears therefore to show stronger association and thus by inference greater power to detect the putative real functional variant. Therefore *MUC5AC* rs1132440 will be the only mucin variable considered from this point on.

Table 3.4 Tests of association between *MUC5AC* variables and respiratory outcomes

Outcome	Outcome code	*MUC5AC* rs1132440		*MUC5AC* TR	
		p value	n	p value	n
Bronchitis	Bron89r	**0.016**	2733	0.377	2517
LRTI	Lripy	0.859	2700	0.555	2482
Wheeze	Wzy89c	**0.019**	2641	0.216	2433
Bronchitis	Bronc	0.111	2909	0.909	2673
Wheeze	Wzyc	0.178	2909	0.321	2673
Asthma	Evasthm	**0.061**	2734	0.583	2518
Hayfever	Evhay	**0.003**	2729	**0.044**	2515
Allergy	Evallerg	**0.006**	2720	0.445	2505

Pearson Chi squared p values are shown for association tests between *MUC5AC* rs1132440 genotypes (M5ACX) and the *MUC5AC* TR genotypes (M5ACTRY) and various respiratory outcomes. Significant results are indicated in bold and identify statistically significant differences in genotype distribution between ‘affected’ and ‘unaffected’ groups. For outcome details please refer to table 1 in appendix 2. Where LRTI refers to lower respiratory tract infection.

3.3.3.2 Permutation test of association

Since in genetic studies of this kind there is an issue of multiple testing, another approach was used to test the significance of the genotype distributions. This was a permutation test in which the observed genotype data are randomized one thousand times producing a theoretical distribution of empirical data. The observed data were then compared to the ranked ‘simulated’ genotype distributions and a probability value was assigned, indicating the likelihood of obtaining the observed data by chance.

All outcomes shown to be significantly associated with rs1132440 as defined by the contingency table test remained associated, the empirical chi squared p values being very similar to the chi squared nominal p values (see table 3.5).

Table 3.5 Permutation tests of association between *MUC5AC* rs1132440 genotypes (M5ACX) and respiratory outcomes.

Outcome	Outcome code	Nominal p value	Empirical p value
Bronchitis	Bron89r	**0.016**	**0.020**
Wheeze	Wzy89c	**0.019**	**0.017**
Asthma	Evasthm	0.061	0.066
Hayfever	Evhay	**0.003**	**0.003**
Allergy	Evallerg	**0.006**	**0.007**

Permutation tests were used as an independent means to confirm the significant association results between the M5ACX variable and respiratory outcomes previously identified by chi squared testing of contingency table data.

3.3.3.3 Logistic Regression: Adjusting for Confounders and risk genotypes

Confounders are defined as any factors which could account for, or dilute the association observed between the genetic marker and phenotypic outcome. In this study the potential confounders were chosen on the basis that they were significantly associated with one or more of the outcome variables or considered to be of biological relevance. In addition, region of birth was used to adjust for geographical and population stratification. Table 3.6 shows each of these potential confounders tested against rs1132440, note the significant association between *MUC5AC* rs1132440 and father’s social class. In order to adjust for confounders and verify the robustness of previously identified associations between rs1132440 and respiratory outcomes, a model of multiple logistic regression was applied to these data. Crucially, as seen in table 3.7, all previously identified associations remain significant after adjustments for potential confounders.

Table 3.6 Tests of association between *MUC5AC* rs1132440 genotypes (M5ACX) and potential confounders.

Confounder	Confounder code	p value	n
Cigarette smoker	Cig89cr	0.687	2730
Cigarette smoker	Cig99cr	0.848	2909
Father's social class	Fsc50r	**0.031**	2672
Region of birth	Reg46ar	0.896	2910
Own social class	Scl89r	0.667	2570
Own social class	Scl99r	0.647	2618
Sex	Sexxr	0.341	2910

Pearson chi squared p values are shown for tests of association between the M5ACX variable and various confounder variables. Significant result shown in bold signifies non-independent distributions of genotypes amongst confounder categories.

Table 3.7 Tests of association between *MUC5AC* rs1132440 genotypes (M5ACX) and respiratory outcomes before and after confounder adjustment.

Outcome	Genotypes (CC versus)	Before adjustments		After adjustments		n
		OR (95% CI)	p value	OR (95% CI)	p value	
Bronchitis	GC	1.042 (0.83-1.30)	0.717	1.033 (0.83-1.29)	0.779	2361
	GG	**0.689 (0.50-0.94)**	**0.019**	**0.683 (0.50-0.93)**	**0.017**	
Wheeze	GC	0.938 (0.64-1.37)	0.739	0.898 (0.61-1.32)	0.582	2285
	GG	**0.382 (0.20-0.74)**	**0.005**	**0.372 (0.19-0.73)**	**0.004**	
Hayfever	**GC**	**1.220 (0.99-1.51)**	**0.068**	**1.242 (1.00-1.54)**	**0.049**	2357
	GG	0.930 (0.70-1.23)	0.614	0.949 (0.71-1.26)	0.721	
Allergy	**GC**	**1.277 (1.05-1.56)**	**0.016**	**1.290 (1.06-1.58)**	**0.013**	2351
	GG	0.971 (0.75-1.26)	0.826	0.977 (0.75-1.27)	0.863	
Asthma	GC	0.975 (0.73-1.31)	0.865	0.979 (0.73-1.31)	0.887	2362
	GG	**0.614 (0.40-0.94)**	**0.025**	**0.615 (0.40-0.94)**	**0.026**	

Logistic regression odds ratios (OR) and p values are shown for tests of association between *MUC5AC* rs1132440 genotypes (M5ACX) and respiratory outcomes both before and after adjusting for the possible confounders; smoking status, region of birth, father's social class, own social class, sex. Significant associations are shown in bold and OR 95% confidence intervals are in parentheses.

Logistic regression was used to calculate odds ratios (OR) by comparing each genotypic variable separately with the reference genotype, in this case the common homozygote. An OR significantly less than 1 implies that the variable in question confers protection from the outcome being tested, in comparison to the common homozygote. While an OR significantly greater than 1 suggests an increased risk of having the disease if you carry the genotype in question as compared to the common homozygote. Thus while multiple logistic regression has primarily been used here to adjust for possible confounders, it has also facilitated in identifying protective and risk genotypes.

As seen in table 3.7 the bronchitis (bron89r), wheeze (wzy89c) and asthma (evasthm) rare homozygote genotypes appear to confer protection against the tested outcomes with an OR significantly less than 1 (OR = 0.689, 0.382 and 0.614 respectively, p values ≤ 0.025). This is reflected by the fact that there were significantly less rare alleles in the bronchitis and wheeze 'affected' groups, a pattern mirrored in the asthmatic allele count data, but statistical significance was not reached.

The hayfever (evhay) and allergy (evallerg) logistic regression results (table 3.7) show that the heterozygote genotype confers risk. No rs1132440 allelic association was identified with either of these outcomes and therefore the logistic regression confirms that any association is due to genotype alone, in this case heterozygosity. This result is unexpected as in a model of disease association we would usually expect the risk allele to confer a dominant or additive effect.

It appears from the logistic regression results that the patterns of genotypic association differ between the bronchitis/wheeze/asthma outcomes and the hayfever/allergy outcomes. However as seen in bar charts of the genotype contingency tables (figure 3.5) there is always an overrepresentation of heterozygotes and an underrepresentation of rare homozygotes in the 'affected' groups for all five outcomes.

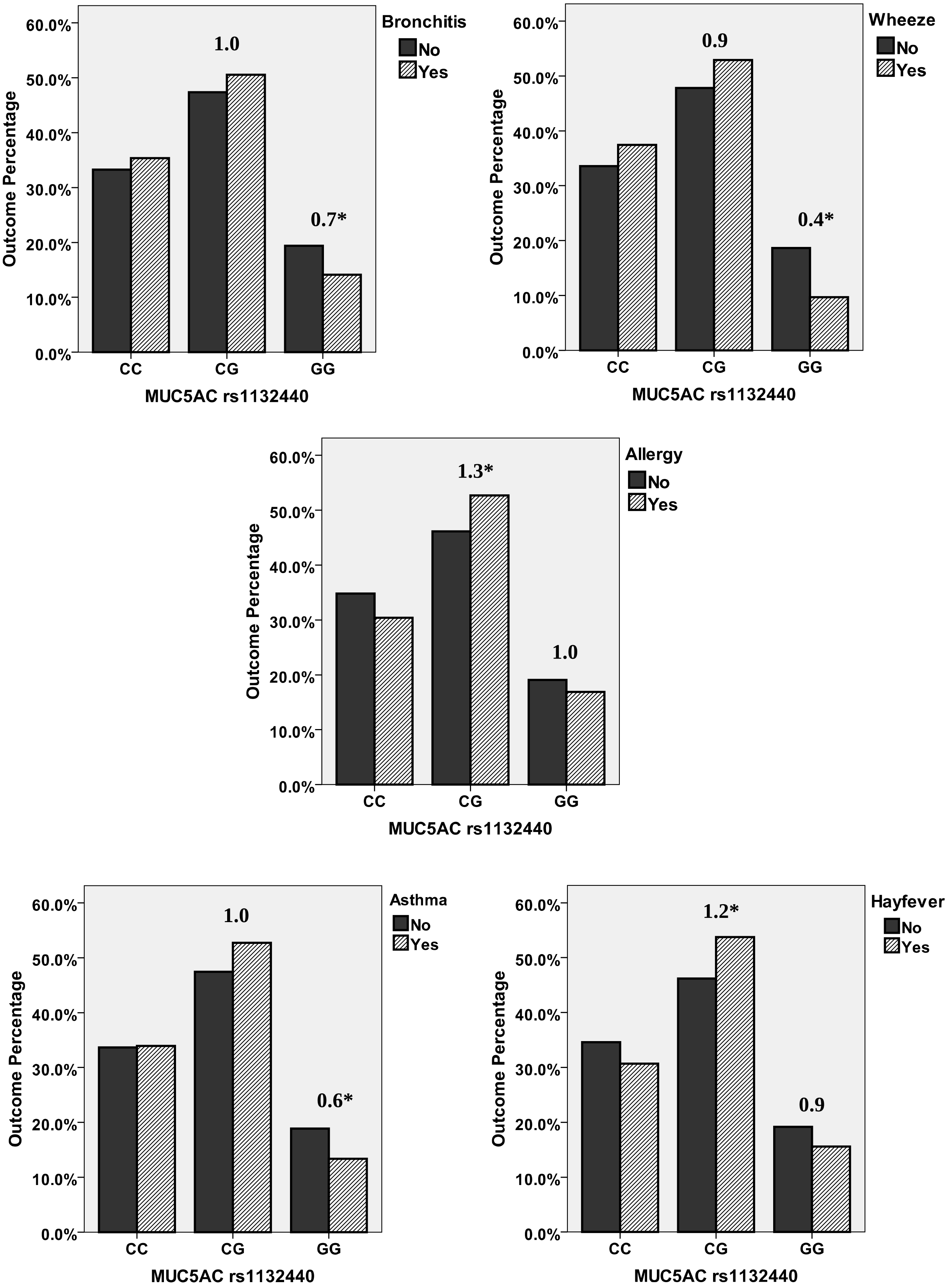


Figure 3.5 Graphical display of *MUC5AC* rs1132440 genotypes (M5ACX) among affected and unaffected groups. The distribution of the rs1132440 genotypes amongst the affected and unaffected groups are shown here as bar charts for outcomes bronchitis (Bron89r), wheeze (Wzy89c), asthma (Evasthm), allergy (Evallerg). Numbers above the CG and GG genotype bars represent the odds ratio for either the CG heterozygote or the GG rare homozygote as compared to the common CC homozygote. * represents significant p values for the corresponding OR.

3.3.3.4 Unexpected heterozygote distribution

Because of the unusual heterozygote association with allergy and hayfever, HWE calculations were performed for the ‘affected’ and ‘unaffected’ groups separately. Although no significant deviation from HWE was observed for the *MUC5AC* rs1132440 cohort data as a whole, deviation was observed for the segregated affected and unaffected allergy (Evallerg) and hayfever (Evhay) samples (table 3.8). For both outcomes the rs1132440 heterozygotes are significantly overrepresented in the affected groups, whereas the unaffected group has a deficit of heterozygotes, and results are shown in detail in table 3.9 a and b.

Table 3.8 HWE p values when considering affected and unaffected groups separately for all five outcomes previously shown to be associated with *MUC5AC* rs1132440 (M5ACX)

Outcome	Outcome code	Affected & Unaffected	Affected	Unaffected
Bronchitis	Bron89r	0.40	0.20	0.10
Wheeze	Whzy89c	0.50	0.07	0.30
Asthma	Evasthm	0.40	0.10	0.15
Hayfever	Evhay	0.40	**0.01**	**0.01**
Allergy	Evallerg	0.50	**0.03**	**0.02**

Bold font signifies significant deviation in genotype distribution from that expected under HWE at the 95% confidence level.

Table 3.9 Comparison of observed and expected *MUC5AC* rs1132440 genotype counts for 'affected' and 'Unaffected' groups with respect to a. Hayfever (Evhay) and b. Allergy (Evallerg).

a.

Genotypes	Hayfever Affected		Hayfever Unaffected	
	Observed	Expected	Observed	Expected
CC	201	216.99	718	690.84
CG	352	320.02	958	1012.31
GG	102	117.99	398	370.84
Total	655	655.00	2074	2073.90

b.

Genotypes	Allergy Affected		Allergy Unaffected	
	Observed	Expected	Observed	Expected
CC	259	274.38	650	625.57
CG	449	418.24	862	910.86
GG	144	159.38	356	331.57
Total	852	852.00	1868	1868.00

In order to exclude the possibility that the difference in the pattern of association observed for the non-independent outcomes was not attributable to the exclusion of different people in the analysis sets because of incomplete outcome data, the regression analyses were repeated on a specific set of individuals who provided data for all five significantly associated outcomes (n = 2263). The results in table 3.10 show that the genotype association patterns remain, although the bronchitis variable (Bron89r) no longer reaches statistical significance (p = 0.065), probably due to the reduced power of the smaller sample size.

Table 3.10 Tests of association between the *MUC5AC* rs1132440 genotype (M5ACX) variable and respiratory outcomes only on samples with full data, before and after confounder adjustment.

Variable	Genotypes (CC versus)	before adjustment		after adjustment	
		OR (95% CI)	p value	OR (95% CI)	p value
Bron89r	GC	1.081 (0.86-1.36)	0.508	1.07 (0.85-1.35)	0.551
	GG	0.742 (0.54-1.02)	0.065	0.737 (0.54-1.01)	0.060
Wzy89c	GC	0.947 (0.65-1.39)	0.781	0.912 (0.62-1.34)	0.638
	GG	**0.386 (0.20-0.75)**	**0.005**	**0.379 (0.19-0.74)**	**0.004**
Evhay	GC	1.233 (0.99-1.54)	0.062	**1.249 (1.00-1.56)**	**0.049**
	GG	0.962 (0.72-1.29)	0.792	0.978 (0.73-1.31)	0.881
Evallerg	**GC**	**1.268 (1.04-1.55)**	**0.021**	**1.282 (1.05-1.57)**	**0.017**
	GG	0.957 (0.73-1.25)	0.747	0.961 (0.74-1.26)	0.772
Evasthm	GC	0.978 (0.73-1.32)	0.881	0.982 (0.73-1.32)	0.903
	GG	**0.642 (0.42-0.99)**	**0.044**	**0.642 (0.42-0.99)**	**0.044**

Logistic regression odds ratios (OR) and p values are shown for tests of association between *MUC5AC* rs1132440 genotypes and respiratory outcomes both before and after adjusting for possible confounders. The same set of individuals with full data has been tested for each outcome, thus all n values = 2263. Significant associations are shown in bold and OR 95% confidence intervals in parentheses.

3.3.4 Inflammatory Mediators, *MUC5AC* and possible gene-gene interactions.

The next step was to consider the role of other genes. As outlined in the introduction, a variety of proteins have been shown to regulate *MUC5AC* expression via allergy mediated pathways, and others have been suggested to play a functional role within the inflamed airways that result from these allergic events.

The particular polymorphic markers typed were selected because they had been significantly associated with allergic respiratory disease in previous studies and/or have been shown to have a direct functional effect (see section 3.1.5). Typing had been done previously by a variety of methods mainly by Lynne Vinall (see appendix 1 for details and acknowledgments). Data were available for seven polymorphic markers within, or in the regulatory regions of, five inflammatory response genes; *EGFR*, *1L13*, *IL1B*, *IL1RN* and *TNF* (see table 3.11 for marker details).

3.3.4.1 Allele and genotype data

The MAFs for the seven inflammatory response markers range from 0.17 to 0.49 (table 3.11) and genotype distributions were in accordance with HWE, with the exception of *IL1B* rs16944 (p ~ 0.01).

It should be noted that the *EGFR* microsatellite and *IL1RN* VNTR are multiallelic. In each case, to simplify analysis, the allelic data were binned into 2 appropriate categories. Derived variable categories were defined by reviewing the literature for allelic functional relevance. With respect to the *EGFR* microsatellite, repeat numbers were defined as either short (S) or long (L). S refers to repeat numbers between 8-18 and repeats of 20 or greater have been denoted as L. The *IL1RN* tandem repeat lengths were categorised as 2 or X; 2 referring to the *IL1RN**2 allele and X includes all other alleles (*IL1RN**1,3,4 and 5). This categorisation was chosen because previous studies have identified *IL1RN**2 as the risk allele.

Table 3.11 Details of inflammatory mediator markers.

Polymorphism	Location	Marker Nomenclature	Alleles	MAF
EGFR Microsatellite	Intron 1	none	L>S §	0.49 (S)
EGFR rs2227983	Exon 13	R497K, R521K	G>A	0.26 (A)
IL13 rs1800925	Promoter	C-1024T, C-1111T C-1112T, C-1055T †	C>T	0.18 (T)
IL13 rs20541	Exon 4	R110Q, R130Q	G>A	0.17 (A)
IL1B rs16944	Promoter	G-511A	G>A	0.34 (A)
IL1RN VNTR	Intron 2	none	X>2‡	0.29 (2)
TNF rs1800629	Promoter	G-308A, G-488A	G>A	0.19 (A)

Note that the marker names refer to physical positions within the gene or its regulatory sequence (for references see introduction). † rs1800925 is a promoter SNP and therefore various notations have been used to describe the believed start of transcription. ‡ X represents alleles *IL1RN**1,3,4 and 5. Please note that the *EGFR* SNP ID rs2227983 was previously rs11543848 in accordance with the NBCI SNP database. § S represents repeat lengths of 8-18 and L refers to lengths of 20-30 repeats. N values range from 2788-2918.

3.3.4.1.1 Allelic association

Prior to investigating gene-gene interactions, all inflammatory mediator gene polymorphisms were tested for association with the outcomes previously shown to be significantly associated with *MUC5AC* rs1132440; bronchitis, wheeze, asthma, hayfever and allergy (bron89r, wzy89c, evasthm, evhay and evallerg respectively).

2 by 2 contingency tables of allele counts were generated for each polymorphism in combination with each of the five outcomes. The chi squared tests performed for each table identified significantly different allelic distributions between the affected and unaffected groups for; the *IL13* promoter SNP (rs1800925) in asthma and allergy (p = 0.038 and 0.024); the *IL13* exonic SNP (rs20541) in asthma and allergy (p = 0.0007 and 0.0104); the *EGFR* SNP (rs2227983) in bronchitis (p = 0.007). Please note that the associations between both *IL13* SNPs and the asthma and allergy outcomes on this data set have been observed previously and are published (Black *et al.* 2009).

For both *IL13* SNPs the rare allele confers risk and is therefore overrepresented in the asthma and allergy affected groups. In contrast, the rare *EGFR* SNP allele is significantly underrepresented in the affected bronchitis group, which implies a protective property, and is shown for the first time.

3.3.4.1.2 Logistic Regression: Genotypic association and confounder adjustment

Logistic regression analysis was performed using genotypes. As seen in table 3.12 all significant associations identified between markers and outcomes with respect to allele counts are also shown to be significantly associated with respect to genotype distribution.

All significant results remain significant after adjusting for confounders in a multiple logistic regression. The confounders used in these adjustments were the same as those used in the *MUC5AC* analysis (table 3.6).

Table 3.12 Tests of association between inflammatory mediator markers and the five respiratory outcomes previously shown to be associated with *MUC5AC* rs1132440.

Variable	*EGFR* (L and S)		*EGFR* (rs11543848)		*IL13* (rs1800925)		*IL13* (rs20541)		*IL1B* (rs16944)		*IL1RN* (TR) 2 and X		*TNF* (rs1800629)	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Bron89r	0.96 (0.77-1.21)	0.99 (0.76-1.29)	0.86 (0.70-1.04)	**0.55*** **(0.35-0.86)**	1.19 (0.97-1.47)	0.85 (0.48-1.49)	1.10 (0.89-1.35)	1.04 (0.60-1.83)	0.98 (0.80-1.19)	0.97 (0.70-1.34)	1.09 (0.90-1.32)	1.02 (0.72-1.44)	0.91 (0.74-1.12)	0.98 (0.59-1.62)
Whzy89c	1.03 (0.69-1.52)	0.86 (0.53-1.40)	0.74 (0.52-1.05)	1.03 (0.55-1.93)	0.98 (0.68-1.41)	0.79 (0.29-2.21)	0.78 (0.54-1.16)	1.08 (0.43-2.73)	1.06 (0.75-1.50)	1.21 (0.71-2.08)	1.13 (0.80-1.60)	1.55 (0.90-2.64)	0.87 (0.61-1.26)	1.65 (0.81-3.37)
Evasthm	1.09 (0.81-1.48)	1.01 (0.71-1.45)	0.89 (0.68-1.16)	1.02 (0.61-1.68)	**1.49*** **(1.14-1.95)**	0.89 (0.40-1.96)	**1.55*** **(1.19-2.03)**	1.73 (0.70-3.34)	1.23 (0.94-1.60)	1.42 (0.94-2.14)	1.17 (0.90-1.53)	1.39 (0.904-2.13)	0.97 (0.74-1.28)	1.51 (0.84-2.71)
Evhay	1.02 (0.83-1.27)	0.96 (0.75-1.23)	1.17 (0.97-1.41)	1.05 (0.73-1.52)	1.17 (0.96-1.42)	1.20 (0.73-1.95)	1.07 (0.88-1.30)	1.48 (0.90-2.43)	1.18 (0.98-1.43)	0.96 (0.70-1.31)	1.05 (0.87-1.26)	1.06 (0.77-1.47)	0.85 (0.70-1.03)	1.25 (0.80-1.97)
Evallerg	1.06 (0.87-1.30)	1.16 (0.92-1.47)	0.96 (0.80-1.14)	0.97 (0.69-1.36)	**1.25*** **(1.05-1.50)**	1.19 (0.75-1.89)	**1.29*** **(1.08-1.55)**	1.21 (0.75-1.97)	1.00 (0.84-1.18)	1.24 (0.94-1.63)	0.94 (0.79-1.11)	1.18 (0.88-1.59)	0.88 (0.74-1.06)	1.34 (0.88-2.04)

ORs from logistic regression analysis represent the level of outcome risk associated with each genotype. The common homozygote has been used as the reference genotype in all cases. 1 refers to common homozygote versus heterozygotes. 2 refers to common homozygotes versus rare homozygotes. Bold font indicates significant ORs and 95% confidence intervals are in parentheses. Adjustments were performed using all confounders previously used for adjustment in M5ACX variable multiple logistic regressions. N values range from 2526-2733.

3.3.4.2 Tests for gene-gene interactions

Since there are potential biological interactions between the inflammatory mediators typed and *MUC5AC* gene expression, we tested for statistical interactions between each of the inflammatory loci and *MUC5AC* rs1132440. It was hypothesised that the functional variants of the inflammatory mediators may interact with MUC5AC to enhance disease association by a combinatorial effect. The possible interactions were tested using the likelihood ratio test with logistic regression.

As can be seen in table 3.13 *EGFR* rs2227983, *IL1RN* VNTR and *TNF* rs1800629 show significant interactions with *MUC5AC* rs1132440 with respect to bronchitis (p = 0.019, 0.009 and 0.046 respectively).

When one displays the data as bar charts (see figure 3.6) we can see that the usual pattern of *MUC5AC* rs1132440 association with bronchitis is only apparent in the sample set of non-carriers for the *IL1RN**2 risk allele. This suggests that the *IL1RN**2 risk allele abolishes any affect MUC5AC may have on the bronchitis outcome. Although the *IL1RN**2 allele has been shown to be significantly associated with asthma (Mao *et al.* 2000), the functional consequences of this allele on the plasma levels of the protein it codes for (IL1Ra) appear to be complicated. The *IL1RN**2 allele has been associated with both elevated and decreased serum levels of IL1Ra (Danis *et al.* 1995; Hurme and Santtila 1998; Mao *et al.* 2000). The Hurme and Santtila study has also suggested that the *IL1RN**2 allele only appears to have a plasma level enhancing affect if in combination with the *IL1B* promoter variant (-511). It is however clear that an increase in IL1Ra serum levels will be beneficial since this protein is an anti-inflammatory cytokine. Thus we might expect any affect of MUC5AC on bronchitis to be more detrimental within airways that contain less of this natural antagonist to inflammation. With this in mind, one might propose the *IL1RN**2 allele to be associated with increased IL1Ra serum levels in this data set, since the MUC5AC affect on bronchitis is only apparent in the *IL1RN**2 non-carriers.

The bar chart in figure 3.7 shows that the patterns of association seen between *MUC5AC* rs1132440 and bronchitis (bron89r) are only apparent when in combination with the *EGFR* rs2227983 genotype that does not contain the rare allele (GG). The rare *EGFR* allele codes for a lysine which has been shown by a previous study to weaken EGFR response to its ligands. Therefore we might expect this weakened response to reduce inflammatory responses and therefore reduce susceptibility to respiratory disease. The *MUC5AC*/*EGFR* interaction pattern appears to support this functional study by showing that the association between *MUC5AC* rs1132440 and bronchitis only exists if the *EGFR* rare allele is not present. The *EGFR* allele that codes for lysine, apparently dampens any effect of *MUC5AC* variants in relation to bronchitis.

A statistically significant result was also obtained for an interaction between *MUC5AC* rs1132440 and the *IL13* promoter SNP rs1800925 with respect to the wheeze outcome (p = 0.013). By plotting the *IL13* promoter risk allele carriers and non-carriers separately (figure 3.8) it can be seen that the *MUC5AC* pattern of association with wheeze (wzy89c) is evident when in combination with both *IL13* rs1800925 risk allele carriers and non-carriers. This is stronger when in combination with the *IL13* rs1800925 risk allele but is not statistically significant.

Table 3.13 Possible interactions between *MUC5AC* rs1132440 and various polymorphisms found in inflammatory response genes.

Outcome	*MUC5AC &* *EGFR* (L&S)	*MUC5AC &* *EGFR* rs2227983	*MUC5AC &* *IL13* rs1800925	*MUC5AC &* *IL13* rs20541	*MUC5AC &* *IL1B* rs16944	*MUC5AC &* *IL1RN*	*MUC5AC &* *TNF* rs1800629
Bronchitis	0.582	**0.019***	0.333	0.273	0.909	**0.009***	**0.046***
Wheeze	0.335	0.244	**0.013***	0.099	0.324	0.983	0.519
Asthma	0.385	0.592	0.781	0.933	0.346	0.637	0.495
Hayfever	0.459	0.476	0.933	0.587	0.052	0.246	0.397
Allergy	0.671	0.562	0.663	0.608	0.340	0.872	0.935

All p values in bold font indicate a statistically significant gene-gene interaction with respect to outcome tested.

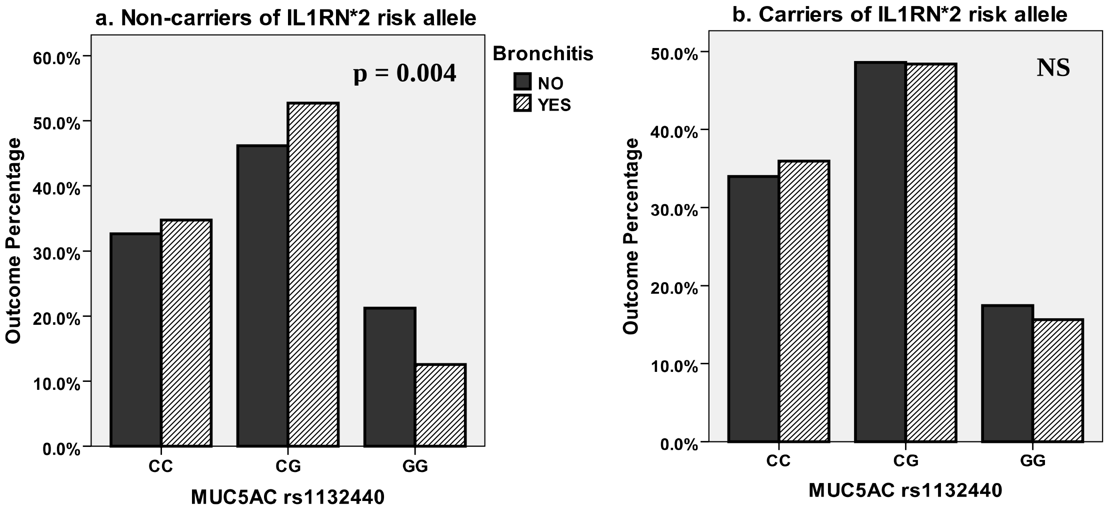


Figure 3.6 Bar charts showing the interaction between *IL1RN* VNTR and *MUC5AC* rs1132440, with respect to Bronchitis (Bron89r). Both show unaffected (NO) and affected (YES) bron89r outcome clusters defined by *MUC5AC* rs1132440 genotypes. Graph a only includes the results from non-carriers of the *IL1RN* risk allele (34, 44, 45, 46, 55). Graph b only includes the results from carriers of the *IL1RN* risk allele (22, 23, 24, 25, 26). Cross tabulations constructed for each data set (with respect to *MUC5AC* rs1132440 genotypes) and tested using Pearson chi squared, show that the distribution of unaffected and affected cases is only significantly different in the non-carriers of the IL1RN*2 allele (p = 0.004). This gene-gene interaction is confirmed in a formal test for interactions (see table 3.13). NS refers to not significant.

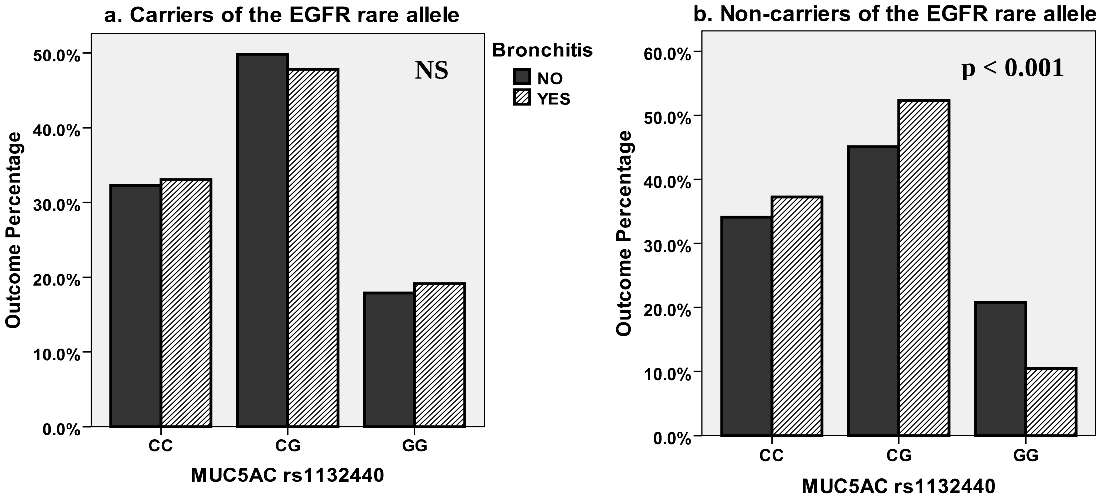


Figure 3.7 Bar charts showing the interaction between *EGFR* rs2227983 and *MUC5AC* rs1132440, with respect to Bronchitis (Bron89r). Both show unaffected (NO) and affected (YES) bron89r outcome clusters defined by *MUC5AC* rs1132440 genotypes. Graph a only includes the results from carriers of the rare *EGFR* allele (AA and AG). Graph b only includes the results from non-carriers of the *EGFR* rare allele (GG). Cross tabulations constructed for each data set (with respect to *MUC5AC* rs1132440 genotypes) and tested using Pearson chi squared, show that the distribution of unaffected and affected cases is only significantly different in the *EGFR* GG homozygotes (p = < 0.001). This gene-gene interaction is confirmed in a formal test for interactions (see table 3.13). NS refers to not significant.

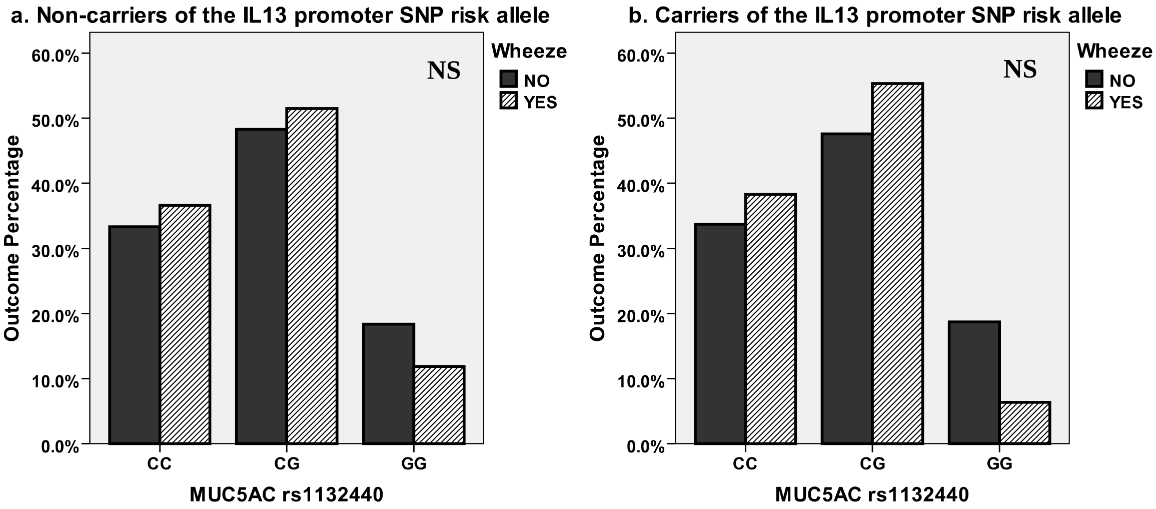


Figure 3.8 Bar charts showing the interaction between *IL13* promoter rs1800925 and *MUC5AC* rs1132440, with respect to wheeze (wzy89c). Both show unaffected (NO) and affected (YES) wzy89c outcome clusters defined by *MUC5AC* rs1132440 genotypes. Graph a only includes the results from non-carriers of the *IL13* rs1800925 risk allele (CC). Graph b only includes results from carriers of the *IL13* rs1800925 risk allele (CT and TT). In neither case were the distribution statistically significant (p = 0.254 p = 0.102 for a and b respectively). NS refers to not significant.

3.4 Discussion

In this study of a longitudinal birth cohort we have indentified statistically significant associations between a *MUC5AC* SNP (rs1132440) and five non-independent respiratory/allergy related outcomes; asthma, allergy, wheeze, hayfever and bronchitis. In order to safeguard against type I errors (false positives) we have utilised two independent goodness-of-fit tests using contingency tables and permutation testing. In both cases the null hypothesis could be rejected with more than 95% certainty showing that the rs1132440 genotypic distributions are significantly different between affected and unaffected groups.

Confounders are also a source of type 1 error because there is the possibility that environmental, population or lifestyle factors are the true cause of the association rather than the genetic variable. Therefore all initial significant associations between rs1132440 and outcome were retested using a model of multiple logistic regression which incorporated confounder adjustment. All associations previously shown to be statistically significant remained so after adjusting for confounders.

The associations identified between *MUC5AC* and the respiratory/allergy related outcomes appear to be robust. Although LD has been shown to span from *MUC2* to *MUC5B*, the *MUC5AC* marker rs1132440 is the only one to show significant association with the five non-independent respiratory and allergy related outcomes. The previously identified *MUC2* TR association with asthma (Vinall *et al.* 2000) was not replicated in this study, however it must be noted that in this case, tests for association only took into account those affected by asthma versus those unaffected. No atopic information was incorporated and therefore the association seen in the small asthmatic cohort could be in direct relation to asthmatic protection in atopic individuals only.

Although *MUC5AC* rs1132440 has been shown to significantly associate with various respiratory and allergy related outcomes, we do not anticipate that this is a causal SNP as it does not alter an amino acid residue. It seems likely that this SNP is in very high

LD with the causal genetic factor. What this factor may be is not obvious because of the unusual excess of heterozygotes, seen in all affected groups but exaggerated with respect to the hayfever and allergy outcomes. One possible explanation for this pattern of heterozygote association might be copy number variation (CNV), which will be explored in chapter 5.

Statistically significant associations have been recently noted between two non-synonymous *MUC5AC* SNPs and the respiratory diseases familial interstitial pneumonia (FIP) and idiopathic pulmonary fibrosis (IPF) (Burch *et al.* 2010). Burch et al hypothesise that these common variants lead to diminished MUC5AC function which leads to an increased susceptibility to FIP and IPF. Both variants result in an amino acid change, Ala497Val and Ala4729Lys. The Ala497Val variant is located within exon 12 of the amino terminal in a vWF-like D domain, and could therefore play an important role in the oligomerisation of MUC5AC. This SNP had not been identified at the outset of this study but is certainly an attractive candidate for future studies involving the 1946 cohort.

The 11p15 chromosome has also been significantly linked to asthma in a genome wide linkage study of Caucasian families (CSGA 1997). While this result has not been replicated in any asthmatic genome wide association studies (GWAS), methods used to analyse GWAS data are notoriously conservative and therefore type II errors are likely to be extensive. The conservative approach taken by the standard analyses is essential because of multiple testing, but it should be recognised that many important genes will be missed as a result of this low sensitivity especially with respect to complex diseases where many genes are thought to be involved in susceptibility and etiology. The extensive data obtained from GWAS are a goldmine of information and it is therefore important that we adopt more sensitive analyses, such as linkage disequilibrium unit (LDU) maps (Andrew *et al.* 2008).

Perhaps one of the most interesting findings reported in this results chapter is the evidence of interactions between *MUC5AC* and various inflammatory functional variants. Although these interactions are not enormously strong, they are consistent with what is known about *MUC5AC* expression during inflammation and its regulation by the inflammatory mediators studied. There is some uncertainty as to whether the *IL13* promoter SNP (rs1800925) is actually functional or whether significant association between this SNP and asthma are in fact due to this marker being in LD with the non-synonymous *IL13* exonic SNP (rs20541) which is considered to be a significant asthmatic risk factor (Black *et al.* 2009). There is however some evidence of interaction between the *IL13* promoter SNP and smoking status with respect to the allergy outcome in this 1946 birth cohort (Black *et al.* 2009). These two interactions involving the *IL13* promoter SNP, be it gene-gene or gene-environment, point to a functional role for this variant which could impact upon *MUC5AC* expression. It is noteworthy that IL13 may be acting in two different ways, firstly to increase IgE levels and subsequently enhance allergy risk, and secondly to increase *MUC5AC* expression.

The newly described interaction between the *EGFR* SNP and *MUC5AC* rs1132440 is of particular interest. The *MUC5AC* association is only apparent if the *EGFR* variation favours a stronger response to its ligands which will subsequently cause enhanced levels of inflammation. It makes biological sense to suggest that bronchitis associated *MUC5AC* variants will have a more detrimental affect on respiratory disease susceptibility and/or severity if present within inflamed airways.

Although the function of the *IL1RN**2 genetic risk allele on the serum levels of its encoded protein appears to be complicated, we do know from this study that the MUC5AC effect on bronchitis is only apparent in the group of *IL1RN**2 non-carriers. If we suggest that the basis of this interaction is biologically similar to *MUC5AC*/*EGFR* interaction, then we might predict that the non-carriers of the *IL1RN**2 allele have increased levels of airway inflammation. Thus by extrapolation, this would suggest that the *IL1RN**2 allele is associated with higher levels of the IL1Ra protein in this sample set, since the carriers of this genetic variant would have reduced levels of inflammation due to an increased abundance of this anti-inflammatory cytokine.

Chapter 4

MUC5AC and *MUC5B* variation in asthmatic case-control cohorts

4 *MUC5AC* and *MUC5B* variation in asthmatic case-control cohorts

In the previous chapter a *MUC5AC* SNP (rs1132440) located within the 3′ region was shown to be significantly associated with asthma in a longitudinal birth cohort. Here *MUC5AC* rs1132440 and other SNPs within close proximity have been typed in two small asthmatic disease cohorts collected in a clinical setting, in an effort to replicate this association and to further characterise *MUC5AC* variation.

Since MUC5B is also a prominent airway mucin like MUC5AC, it is also an attractive candidate gene for respiratory disease association studies. In the second part of this chapter, variation within the *MUC5B* promoter will be described in these same disease cohorts. Studies have shown MUC5B expression to be up-regulated at the mRNA and protein levels during chronic respiratory disease (Burgel *et al.* 2007; Caramori *et al.* 2004; Groneberg *et al.* 2002a; Kamio *et al.* 2005; Kirkham *et al.* 2002). We therefore hypothesise that *MUC5B* regulatory variation, and in this case specifically within the immediate promoter, will result in altered inter-individual susceptibility to and severity of respiratory disease. This hypothesis is supported by a study on the hypersecretory disease diffuse panbronchiolitis (DPB) in which significant association has been identified between *MUC5B* promoter variation and the disease in a Japanese population (Kamio *et al* 2005).

Since the relationship between MUC5AC and asthma has already been explored in detail in chapter 3 this short introductory section will focus on gene regulatory regions and more specifically the *MUC5B* promoter with respect to hypersecretory airway disease.

4.1 Introduction

4.1.1 *MUC5B* expression in hypersecretory airway disease

MUC5B has been shown to be significantly up-regulated at mRNA and protein levels in asthmatic, COPD, CF and DPB airways (Burgel *et al.* 2007; Caramori *et al.* 2004; Groneberg *et al.* 2002a; Kamio *et al.* 2005; Kirkham *et al.* 2002). In sputum obtained from COPD airways, the abundance of MUC5B exceeds that of MUC5AC, a pattern which is reversed in healthy airways (Kirkham *et al.* 2008). In the case of asthma and COPD, MUC5B has been shown to be the primary respiratory mucin within mucus plugs and sputum respectively (Burgel and Nadel 2004; Groneberg *et al.* 2002a). It is therefore important to understand the patterns and mechanisms of *MUC5B* expression.

MUC5AC and MUC5B are the predominant mucins found in the respiratory mucus (Kirkham *et al.* 2002). In healthy airways the MUC5 mucins are expressed in different cells, MUC5AC in surface epithelial goblet cells (Groneberg *et al.* 2002b; Hovenberg *et al.* 1996) and MUC5B in the mucous cells of the submucosal glands (Groneberg *et al.* 2002a). However during hypersecretory respiratory disease *MUC5B* is expressed in the goblet cells as well as the submucosal glands (Groneberg *et al.* 2002a; Kamio *et al.* 2005). This aberrant expression coincides with goblet cell metaplasia (GCM) and submucosal gland hypertrophy but the mechanisms by which ectopic *MUC5B* goblet cell expression occurs remains unknown. We do however know that the pattern of mucin expression seen in diseased airways mirrors that of human tracheobronchial epithelial cells cultured at air liquid interface and is also reminiscent of the patterns of expression in fetal airways at 13 weeks of gestation (Buisine *et al.* 1999).

There is however the possibility that MUC5B protein production occurs at baseline levels in the goblet cells but can only be identified in the cells when expression levels exceed secretion. A mouse mutant has been identified that has a defect in the mucin secretion pathway. Examination of this mouse has identified MUC5B protein in the Clara cells where it was thought previously not to be expressed. The study concludes that this is not aberrant expression but rather MUC5B had not been previously detected because it is secreted as fast as it is made (Zhu *et al.* 2008).

4.1.2 Gene promoters

Since this chapter includes characterisation of *MUC5B* promoter genetic variants, it is appropriate to discuss here the properties and mechanisms of human gene promoters and regulatory regions.

The Promoter of a gene describes a region of DNA containing the cis-acting regulatory motifs needed for the gene to be transcribed. RNA polymerase II is the enzyme that transcribes all protein coding genes. In order to initiate and maintain transcription at a basal level, RNA polymerase II and a variety of general transcription factors bind to DNA motifs upstream of the gene and within close proximity to the transcription start site (TSS).

Basal promoter sequence motifs such as the TATA box and the initiator element (INR) have been shown to be essential for transcription. The TATA box is located 25-30 nucleotides upstream of the TSS. The exact number of gene promoters that contain a TATA box is not known however estimates range from 11% (Bajic *et al.* 2004) to 32% (Suzuki *et al.* 2001). Estimations are difficult to make since many promoters have not been experimentally defined and bioinformatics methods for motif predictions are not altogether reliable since the motif sequences are degenerate. It is however known that the TATA box is sufficient for the initiation of transcription. The initiator motif (INR) is also sufficient for transcriptional activation. It spans across the TSS and can act independently to initiate transcription, or in concert with the downstream core promoter element (DPE) located at +28/+32bp relative to the TSS, providing that the space between the two elements is optimal.

While the basal promoter is essential for gene transcription it confers low transcriptional activity and therefore regulatory regions are needed for adequate expression. Upstream regulatory elements (UREs) and enhancer sequence motifs, located approximately 100-200 nucleotides and up to thousands of nucleotides upstream of the TSS respectively, provide the binding sites for the transcription factors required to stimulate expression levels. UREs and enhancer motifs often overlap (Turner *et al.* 2000).

4.1.3 *MUC5B* Promoters: proximal and distal

The basal promoter of a gene ensures baseline expression, while the URE’s and enhancer elements respond to challenges causing fluctuations in expression levels. For example, a Creb response element (CRE) located in the 5′ upstream region of *MUC5B* at -956 relative to transcription start site, has been shown to respond to the E2 sex hormone causing the upregulation of *MUC5B* expression (Choi *et al.* 2009a). Thus we would expect *MUC5B* expression to be higher in cells expressing E2 or within hormonal rich microenvironments. As opposed to hormonal challenge, diseased airways will be subjected to inflammatory challenge and therefore any *MUC5B* regulatory inflammatory response elements will cause *MUC5B* expression levels to increase during disease. It can also be proposed that genetic variation within the *MUC5B* inflammatory response elements could result in aberrant expression of the gene and may therefore be associated with increased disease susceptibility and/or severity.

Genetic variation within the one kilobase sequence immediately upstream of the *MUC5B* TSS, referred to as the *MUC5B* proximal promoter from now on, has been shown to be significantly associated with diffuse panbronchiolitis (DPB). DPB is a predominantly Asian hypersecretory disease which is symptomatically similar to CF but does not appear to have the same genetic cause. Three *MUC5B* polymorphisms located within the proximal promoter region were shown to be significantly associated with DPB. This variation is likely to be functional since different *MUC5B* proximal promoter haplotypes which include these three associated polymorphisms, have been shown to confer varied expression levels *in vitro* (Kamio *et al.* 2005) and *in vivo* (Loh *et al.*

2010). The low expressing haplotype shows significant negative association with DPB and the high expressing haplotype is overrepresented in the disease group, although statistical significance is not reached in this case.

Much of the literature refers to a single *MUC5B* promoter, which we call here the proximal promoter, however a study by Perrais et al (Perrais *et al.* 2001) has identified a second promoter which is very active in a gastric cancer cell line. This additional promoter is located directly upstream of the proximal promoter and is referred to as the distal promoter. Using a primer extension method, RNA transcripts corresponding to the distal TATA box were identified abundantly in the gastric cancer cell line and were also shown to be present in human trachea indicating that this is an active promoter (Perrais *et al.* 2001).

4.2 Hypothesis and Aims

The first aim of this chapter was to replicate the findings of significant association seen between the *MUC5AC* 3′ SNP and asthma identified in the longitudinal birth cohort (chapter 3). We aim to do this by initially characterising 3′ *MUC5AC* variation by Sanger sequencing in two asthmatic case-control disease cohorts.

Since GCM is characteristic of asthma and MUC5AC is a marker of GCM (see chapter 3) it could be hypothesised that *MUC5AC* variation leads to alter susceptibility to asthma and/or may result in varied asthmatic severity.

The second aim of this chapter is to explore the relationship between *MUC5B* regulatory variants and asthma. *MUC5B* regulatory variants have been shown to be significantly associated with the hypersecretory airway disease DPB (Kamio *et al* 2005) and promoter haplotype variants have been shown to confer different gene expression levels by two independent studies (Kamio et al 2005; Loh *et al* 2010). We therefore hypothesise that the high expressing haplotype will be overrepresented in the asthmatic case sample set since asthma is a characteristic hypersecretory disease.

4.3 Results

The results section will be divided into three parts; analysis of variation within the 3′ end of *MUC5AC*; *MUC5B* proximal promoter variation and an account of the preliminary findings with regard to the putative *MUC5B* distal promoter. Note that all sections report data from asthmatic disease cohorts.

4.3.1 *MUC5AC* genetic variation

In chapter 3 a statistically significant association between *MUC5AC* rs1132440 and various allergy related respiratory outcomes in a longitudinal birth cohort is reported. Here we describe this SNP and others located within close proximity in the 3′ end of *MUC5AC* in two small asthmatic disease cohorts referred to as the matched and severe asthmatic cohorts. The first sample set known as the matched asthmatic cohort (n = 100), is composed of clinically diagnosed atopic asthmatic individuals accompanied by sex and age matched atopic non-asthmatic controls. The second sample set will be referred to as the severe asthmatic cohort (n = 176) and consists of individuals with clinically diagnosed severe asthma (n = 84), and hypernormal controls with no personal or family history of respiratory disease (n = 92). It should be noted that no atopy data is available for the severe cohort. The two cohorts have been treated separately for all analyses in this chapter unless otherwise stated (refer to materials and methods section for sample details).

4.3.1.1 Allelic data and analysis

Sequence was obtained for two *MUC5AC* 3′ end regions: fragments spanning from exon 13 to 14 and exon 19 to 20, the previously tested rs1132440 being located within exon 19. Because of the unusual patterns of association and deviations from HWE seen with rs1132440 in the 1946 cohort (chapter 3), these two neighbouring fragments where chosen in order to investigate whether SNPs within close proximity also exhibited similar patterns of association

In total, 7 polymorphisms were identified within the region spanning exon 13 to 14 and 8 SNPs were noted within the exon 18 to 19 region (see figures 4.1 and 4.2 for annotated sequences) in the two asthmatic cohorts, several of these SNPs have been reported previously. Details and MAFs for all identified SNPs are shown in table 4.1, All MAFs are evidently similar for both the matched and severe cohorts.

Table 4.1 Details of the *MUC5AC* 3′ end SNPs typed in the matched and severe asthmatic cohorts, including minor allele frequencies.

		Matched Cohort		Severe Cohort	
SNP ID	SNP location	MAF	n	MAF	n
novel 116	Intron 13	0.01 (T)	88	0.03 (T)	161
rs2075843	Intron 13	0.20 (A)	89	0.22 (A)	164
10bp VNTR	Intron 13	0.03 (3)	88	0.06 (3)	161
novel 245	Intron 13	0.01 (T)	88	0.01 (T)	161
rs34666042	Intron 13	0.16 (T)	88	0.12 (T)	161
novel 287	Intron 13	0.03 (G)	88	0.05 (G)	161
rs34831688	Exon 14	0.17 (T)	83	0.13 (T)	140
rs35968147	Intron 18	0.19 (C)	88	0.18 (C)	169
rs2075844	Intron 18	0.40 (G)	87	0.41 (G)	170
novel 420	Intron 18	0.01 (T)	88	0.00 (T)	170
novel 422	Intron 18	0.01 (A)	88	0.00 (A)	170
rs28728088	Intron 18	0.19 (T)	88	0.18 (T)	170
novel 514	Intron 18	0.00 (T)	88	0.01 (T)	170
novel 527	Intron 18	0.00 (T)	88	0.003 (T)	170
rs1132440	Exon 19	0.38 (G)	99	0.41 (G)	170

MAF refers to minor allele frequency and SNP location details the intron/exon location of the SNP. Both exonic SNPs are synonymous and thus do not alter an amino acid. N = number of individuals from whom data was obtained. The SNPs reported previously have a rs identifier number. The novel polymorphisms discovered during this project have not been previously reported and therefore have no existing identifiers, thus each has been assigned a number in relation to the location within the PCR fragment.

None of the SNPs in table 4.1 are represented in the HapMap project. The MAF of rs1132440 in the larger of the two cohorts (severe cohort), is in good agreement with that of the 1946 cohort sample set (chapter 3).

All polymorphic markers identified in these sample sets are annotated in figures 4.1 and 4.2 and the novel variants, which have no rs identifiers, have been assigned numbers in accordance with their nucleotide position within the PCR fragment. Three novel SNPs and a 10bp VNTR were identified within intron 13. SNP 116, SNP 245 and the VNTR were confirmed by sequencing in both orientations. SNP 287 was specific to samples typed as heterozygotes for the tandem repeat and it was not therefore possible to confirm this SNP in the forward orientation due to the overlapping peaks generated by the insertion of an extra 10 base pairs.

Four novel SNPs were also identified within intron 18 and confirmed by sequencing in both orientations, however all but one are represented on only one chromosome (singletons). All confirmed novel polymorphisms are at low frequencies (table 4.1) and only two genotypes were identified for each (common homozygote and heterozygote). The intron 13 VNTR is the most frequent of the novel polymorphisms with MAFs of 0.03 and 0.05 (3 repeats) in the matched asthmatic and severe asthmatic cohorts respectively, and is therefore the only novel polymorphism included in the *MUC5AC* 3′ haplotypes subsequently generated in this chapter.

```
GGGCAGGCTGGACAGTGTGCAGC[ATCAACGGGACCCTGTACCAGGTAAGAGCCACGGAG

CTCAGACCCCCTCAGCCATAGGGACGGAGCTTCCCACTGACCCTGAGGCCCAGGTAGACT

TTGGAGCAACTGCCAACTCYGGCCGRGGCCAGGGACTCGAGTCTCTGCAGACACAGCCCA
                   116   rs2075843
CTATCAAGTGTGGCTGAGGCCCGAGGTCGGCCCCAGGTCCCGGAAATATGGACATCTACA
               10bp tandem repeat
CCCTGGCCTGCCTGGCTCCGGGGGGCTCYGGGGGACTTTGCCTCTCYTGGCACCACAGCA
                            245               rs34666042
CAGCCAGGCCKGGATCCCACGGCTCTGTCCTGAGCCGGCTGAGTATGTGGCCCTGCAGAG
          287
TGTGTGGCCTTGTTGGGCACCCCATCCAAGGGGGTGCAGCGTGGGGCTCTGCTCTAGGGA

TGGGGACCCTGGGCTGTGGCCTCTGCACCAAGAGGTGCCACCACGAGTCACCCCAGGGGT

GCAACTCGGCCTGGTAGGAAGCGGCCTGGAGGGGGATGTCTGGGAAGTTGGGGGCAGCAA

GCCAGTGGGGAGGCAGGGGCGGGTCTCCCCAGGGCCCAAGCTCATGAGTGTCTGCTGCCC

TGGCTCTCCCCAGCCCGGCGCCGTGGTCTCCTCGAGCCTGTGCGAAACCTGCAGGTGTGA

GCTGCCGGGTGGCCCCCCATCGGAYGCGTTTGTGGTCAGCTGTGAGACCCAGATCTGCA]
                        rs34831688
ACACACACTGCCCTGTG
```

Figure 4.1 Nucleotide sequence of the *MUC5AC* 3′ terminal region spanning from exon 13 to exon 14. Exons are highlighted in grey. Square brackets define PCR fragment boundaries. Yellow highlight shows positions of novel polymorphisms and the number identifiers for these polymorphisms are defined relative to the start of the PCR fragment. Green highlight represents locations of previously reported SNPs. Note that all polymorphisms annotated in this sequence have been confirmed in this study. Nucleotide abbreviations; K = G/T; R = A/G; Y = C/T.

```
TGGGCTGGTCCTAAACCCTGTGTTCCTCTCCAGAGT[CGACCTGTGCTGTGTACCATAGG

AGCCTGATCATCCAGCAGCAGGGCTGCAGCTCCTCGGAGCCCGTGCGCCTGGCTTACTGC

CGGGGGAACTGTGGGGACAGCTCTTCCATGTACGTGCCTGGGCAGCAGGCAGGGAGACGC

GATTGGCTGTGGGGTGCAGTCAGGGCCCCCAGGGCTCYAGGTGCCAGATAGACGAGGGGC
                                     rs35968147
AGGACCATGAGGGGCCAGRCAAAGGGCTCTGAGGGTGAGGCGGGAAAGGGGTCCTGAGAT
                  rs2075844
GGCAAGGGTGGGGCTGGGGTAACTACATCCCCAGAGCCTGTGTCGGCATCACGCTCTCCT

GTTTACTGAGCTCCGCCAGGAACTTGCCGCAGCCGCCCCGAGTCTCCCTCCCTCCCATCA
                                      422
GCACGGAGCCGGGGTCGGCCCTGGTGGGACTGTTGGYGMCTGGGGAACTGGCAAAGGAGA
                                    420
GCTGGTTGTCAGACACTGGCAGCATGCCTCCAGGAGCAGGGAACACGATGAGGCCGCCCA
          514
GAGCTYGGCAYGGCGCCGGCTTAYGGCAGGAGGCTGGGGTGGCGCAGCAGCTGGTGCTGA
     rs28728088        527
GCAGCCCCTGCCCACAGGTACTCGCTCGAGGGCAACACGGTGGAGCACAGGTGCCAGTGC

TGCCAGGAGCTGCGGACCTCGCTGAGGAATGTGACCCTGCACTGCACCGACGGCTCCAGC

CGGGCCTTCAGCTACACCGAGGTGGAAGAGTGCGGCTGCATGGGCCGGCGSTGCCCTGCG
                                                  rs1132440
CCGGGCGACACCCAGCACTCGGAGGAGGCGGAACCCGAGCCCAGCCAGGAGGCAGAGAGT

GGGAGCTGGGAGAGAGGCGTCCCAGTGTC]CCCCATGCACTGA
```

Figure 4.2 Nucleotide sequence of the *MUC5AC* 3′ terminal region spanning from exon 18 to exon 19. Exons are highlighted in grey. Square brackets define PCR fragment boundaries. Yellow highlight shows positions of novel polymorphisms and the number identifiers for these polymorphisms are defined relative to the start of the PCR fragment. Green highlight represents locations of previously reported SNPs. Note that all polymorphisms annotated in this sequence have been confirmed in this study. Nucleotide abbreviations; K = G/T; M = A/C; R = A/G; S = C/G; Y = C/T.

4.3.1.2 Data analysis

The allelic distribution of all SNPs was compared between the cases and controls in both disease cohorts. 2 by 2 contingency tables were constructed and examined using chi squared tests. No statistically significant difference between cases and controls was noted for any SNP in either cohort.

Because of low expected cell counts, genotype distributions were analysed in two separate ways; by grouping genotypes into common homozygotes versus the heterozygotes and rare homozygotes; by combining the two cohorts in order to increase expected cell counts to sufficient numbers in order to perform chi squared tests. No statistically significant difference in genotype distribution was observed between the cases and controls in either of the analyses performed.

The only significant difference noted between the cases and controls with respect to genotype counts were deviations from HWE for two SNPs rs34666042 and rs34831688 ($p = 0.01$ and 0.02 respectively) in the asthmatic sample set of the severe cohort; no deviation from HWE is noted within the hypernormal controls. Marginal significant deviation ($p = 0.05$) from HWE equilibrium was also noted for rs34666042 in the asthmatic sample set of the matched cohort and once again no significant deviation was seen for any SNPs in the matched controls. However in contrast to the observations in the 1946 cohort (chapter 3), there are less heterozygotes than expected in this disease sample. Genotype distributions for all other SNPs are not shown to deviate from those expected under HWE.

4.3.1.3 Measures of linkage disequilibrium

Table 4.2 D′ measures of pairwise LD for polymorphisms within the *MUC5AC* 3′ end region spanning from exon 13 to exon 19 in asthmatic a. cases and b. controls.

a.

rs34666042	**1.000**					
rs34831688	**1.000**	**1.000**				
rs35968147	**1.000**	**1.000**	**1.000**			
rs2075844	**1.000**	**1.000**	**1.000**	**1.000**		
rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	
rs1132440	**0.914**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088

b.

rs34666042	**1.000**					
rs34831688	**1.000**	**1.000**				
rs35968147	**1.000**	**1.000**	**1.000**			
rs2075844	**1.000**	**1.000**	**1.000**	**1.000**		
rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	
rs1132440	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088

Only polymorphisms with a frequency greater than 10% have been included in these LD analyses. a. Cases include asthmatic samples from both the matched and severe cohorts. b. Controls include control samples from both matched and severe cohorts.

In order to study the patterns of LD across this 3′ end region**,** Ldmax was used to compute pairwise measures of LD for cases and controls separately. The two disease cohorts were combined in order to amalgamate cases and controls; cases include the matched asthmatic samples plus the severe asthmatics, while the controls are represented by the matched atopics plus the hypernormal controls from the severe cohort. As shown in table 4.2 a and b there is almost complete LD across this region as all pairwise associations are statistically significant (with p values of < 0.05 as tested by chi squared) and nearly all have D′ values of 1. The patterns of LD between cases and controls are almost identical except for the D′ values between SNPs rs2075843 and rs1132440, where there appears to have been a small amount of recombination in only the sample set of cases.

4.3.1.4 Haplotypes inference and analysis

Haplotypes were inferred by PHASE with a certainty of 0.99 or greater for the *MUC5AC* 3′ terminal region spanning from exon 13 to exon 19 using the genotype data for 6 of the previously described SNPs and the novel intron 13 VNTR. Data for SNP rs34831688 has been removed from the haplotypic analysis since this marker could not be typed in the individuals who were heterozygotes for the intron 13 TR which could result in uneven dropout due to the complete linkage noted within this region. The inferred *MUC5AC* 3′ region haplotypes are detailed in table 4.3. There are four major haplotypes (HA-HD) which account for all haplotypes identified in the severe asthmatic cohort and 98 per cent of the matched asthmatic haplotypes. The rare haplotypes (HE-HG) appear to be recombinants of the common haplotypes. Please note that the haplotype frequencies obtained from Arlequin (not shown) are almost identical to the PHASE results shown in table 4.3.

The rare novel alleles were assigned to a haplotypic background by firstly identifying a carrier of the novel allele who was also homozygous for a particular haplotype. It was subsequently assumed that the rare allele was only present on this haplotype since it is unlikely that the mutation will have occurred independently more than once. Novel SNPs 287 and 245 are exclusive to the HD haplotype. SNP 287 is always present on the HD haplotype while 245 is present occasionally. Novel alleles 116 and 514 only occur on the most frequent haplotype HA. The other novel alleles cannot be assigned to haplotypic backgrounds since they have only been identified on a single chromosome.

Table 4.3 Details and frequencies of the *MUC5AC* 3′ region haplotypes for all case and controls sample sets.

Haplotype ID	rs2075843	Intron 13 VNTR	rs34666042	rs35968147	rs2075844	rs28728088	rs1132440	Matched asthmatic	Matched atopic	Severe asthmatic	Severe controls
HA	G	2	C	T	A	C	C	0.59	0.62	0.57	0.63
HB	A	2	C	T	G	C	G	0.16	0.21	0.23	0.22
HC	G	2	T	C	G	T	G	0.19	0.13	0.16	0.09
HD	G	3	C	C	G	T	G	0.03	0.02	0.04	0.06
HE	A	2	C	T	G	C	C	0.02	0.00	0.00	0.00
HF	G	2	C	T	G	C	C	0.01	0.00	0.00	0.00
HG	G	2	C	T	A	C	G	0.00	0.01	0.00	0.00

SNP rs34831688 has been removed from haplotype analysis because we were unable to type this variant in the heterozygous intron 13 TR individuals. The shading in this table indicates likely recombination events that may have occurred to produce the rarer haplotypes.

Haplotype frequencies are shown in table 4.3, and they show that the rare recombinant haplotypes have only been identified in the matched asthmatic cohort and are found in both the cases and controls. It is noteworthy that the rare allele of the SNP showing significant deviation from HWE in the cases (rs34666042), is only found on the HC haplotypic background. Haplotype distribution between cases and controls are represented graphically in figure 4.3 for both the severe and matched asthmatic cohorts. The distributions of haplotypes with respect to disease status, are fairly similar for both asthmatic cohorts. Notably the controls have a greater percentage of HA and the cases have a greater percentage of HC.

The differences in haplotype distributions between cases and controls were tested three separate ways, chi squared test, PHASE case-control analysis and the exact test of population differentiation (ETOPD) implemented by Arlequin. A significant difference was noted between the haplotype frequencies of the severe cohort case and control sample sets (p = 0.021) when examined with the ETOPD. However the PHASE case-control analysis and a chi squared test indicated that the difference was not statistically significant. No statistically significant difference in haplotype distributions could be identified in the matched asthmatic cohort.

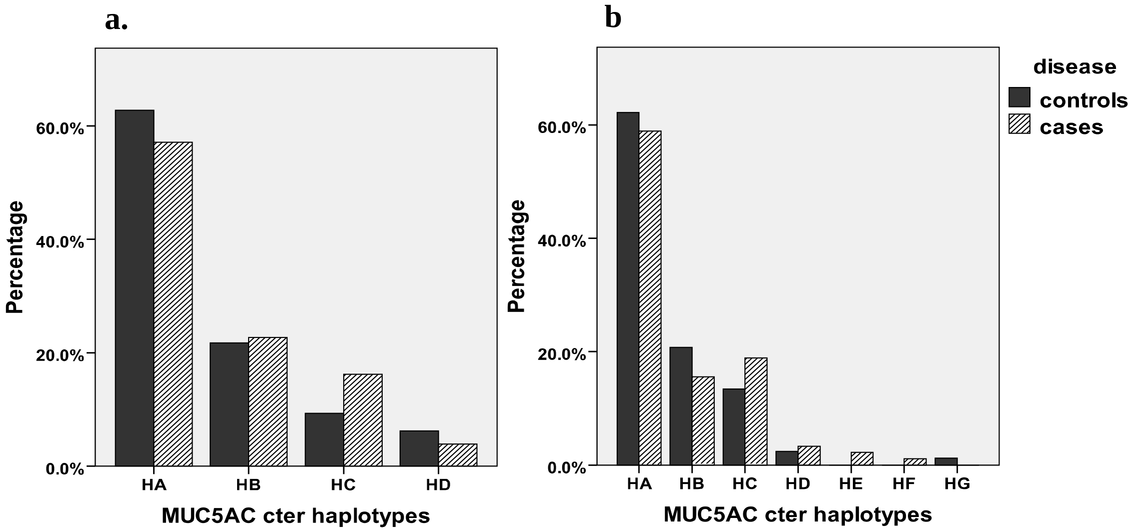


Figure 4.3 Bar chart representing the distribution of *MUC5AC* 3′ region haplotypes between cases and controls of the a. severe asthmatic cohort b. matched asthmatic cohort.

4.3.2 *MUC5B* proximal promoter

As discussed in the introductory section of this chapter, variation in the *MUC5B* immediate promoter region appears to be of functional importance in the hypersecretory disease DPB. Therefore the *MUC5B* proximal promoter (+97 to -1097 with respect to the TSS) was sequenced in our two small asthmatic disease cohorts with a view to examining the corresponding Kamio haplotypes (Kamio *et al.* 2005). We hypothesised that the high expressing H1 haplotype would be overrepresented in the asthmatic populations based on the notion that like DPB, asthma is a hypersecretory disease of the airways.

4.3.2.1 Allelic data

Six polymorphic SNPs were identified (see figure 4.4) and typed by sequencing in the matched asthmatic and severe asthmatic disease cohorts. Minor allele frequencies (MAF) range from 0.05 to 0.24, all shown in table 4.4. Only two of these SNPs are included in the HapMap project, rs885454 and rs7118568, and the MAFs identified for these two European disease cohorts are in good accordance with the HapMap CEPH dataset. It should be noted that the proximal promoter indel (rs17235353) which defines the low expressing H2 haplotype is monomorphic within this European sample, with only the insertion allele being represented.

Table 4.4 Details of the *MUC5B* proximal promoter SNPs typed in the matched and severe asthmatic cohorts, including minor allele frequencies (MAF).

SNP ID	SNP location	Matched Cohort		Severe Cohort	
		MAF	n	MAF	n
rs885455	-919	0.19 (G)	96	0.19 (G)	135
rs885454	-906	0.05 (A)	96	0.06 (A)	136
rs7115457	-237	0.09 (A)	96	0.09 (A)	156
rs7118568	-217	0.09 (G)	96	0.09 (G)	156
rs56235854	-100	0.05 (A)	96	0.05 (A)	154
rs2735738	-78	0.21 (C)	96	0.24 (C)	155

SNP location refers to base position relative to transcription start site and n refers to number of individuals.

In order to compare allelic distributions between cases and controls within each cohort, 2 by 2 contingency tables of allele counts were constructed and tested with chi squared or Fishers exact tests under the null hypothesis that asthmatic and control samples come from the same population and are therefore not significantly different with respect to *MUC5B* promoter allele counts.

Analysis of the severe asthmatic cohort identified statistically significant associations between the two SNPs rs7115457 and rs7118568 and severe asthma with a chi squared p value of 0.015 for both. It should be noted that these SNPs are completely linked and therefore any significance is due to the same association. For both SNPs the rare alleles are underrepresented in the asthmatic disease group, and it should be noted that these rare alleles define the high expressing H1 haplotype.

This association could not be confirmed at the 95% confidence level in the matched asthmatic cohort however rs56235854 showed marginal significance with a p value of 0.054.

```
-1097
CCACGGAGCATTCAGGACGCTGGTGACCAGGGAGCCAGGAGGTGGGAGCATCTGAGGTGCA

GGTCACACGGGCAGGAGGTGTTTGCAAGAGGTATTGCAGCGCGGACGGAGTGTCCTGCAGA
                                                       -919
TGACGCTGTCTGTCCTGTAGATGACGCTCGTCAAGGAGGTTTACCACATAGCCCCRGGAA
        -906                                        rs885455
GCCCACCCRACACCAGCCGGAGGTGCTAGGCTTCTGCGGCTCCCACCTGGGGCAGGCGGAG
        rs885454
GACCCCGGGCAGGTCCAGGACCCCCCGGAGCAGCTGCTTCCTCAACCCTGCCAGGGTTAAT

GAGGAGGCCCCAGAGTGAGGTGGAGGCCAAATGGGACTCAGGGCCGGAGCCTCTGGCCTGG

CTGGATCAGGGCTGGCATTGGACAAGCGCAGCTGACTCCCGATGTGCATGGCCAGGAGACA
          -659/660
CTCTGGGCCTCAGTTTCCCCTTGAATGTGAACCTTGAAACAGATCAGCCCAGAGACCTCCC
          rs17235353
ACGGTCTTCAAGGGGCTCTGGTCAGCTGGGCTGGGGTCTCTGGAAATAGAGCCTCCTCCAG

GGACCCCCACAAGCCACCCAGACTGAGCATCCTGGCCATGTGCATGCCTGAGCTCAGCAGG

AGCCTCCCGGcCTCCCCGTGGGCTAAGCAGTGGTGGGAGGGGAGCTCCAGCCTCGTGGGCC

CTCCCCGGGCCTCGGGGACCCATGGTCAGTGGCTGGGGGTGCTGCCCAGAGGCTGGGATTC

CCTTCCAGCAGGAGCCGCAGTGGGGCTGAGTGTGAGGCAGGCTGGCTGACCACTGTTTCCA

TGGACCCTGCGTCCAAGGCCAGCCCTGCCTTCCAGCGGCTTTGCCATCTAGGACGGGTGCC
      -237                -217
AGGTGGRGTAGGCCCTTCTCTCCCTTSCGATTCTCAGAAGCTGCTGGGGGTGGGGGCGTCC
      rs7115457           rs7118568
TGGGCCTCAGGGCACAGAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCC
                     -100                  -78
CCTCCCCGAGAGCAAACACACRTGGCTGGAGCGGGGAAGAGCAYGGTGCCCTGCGTGGCCT
                     rs56235854            rs2735738
GGCCTGGCTTGGGGCCAAGGCTCCCTGCTACATAAGCTGGGGCCCCCAGGGGAGCAAGCA▌
+1                     PROXIMAL TATA BOX
CCCGGCCCGGCTCCCTCCCTGCCCGTCCCCGTCCCCCCACCCGTGCCAGCCCCCAGGATGG
TSS                                +97
GTGCCCCGAGCGCGTGCCGGACGCTGGTGTTGGCTC
```

Figure 4.4 Annotated sequence corresponding to the *MUC5B* proximal promoter. Grey highlight indicates forward and reverse primers for the fragment A. Pink highlight indicates forward and reverse primers for fragment B. All SNPs typed in the two asthmatics cohorts are highlighted in green, note that all are polymorphic within these sample sets with the exception of the indel rs17235353. Transcription start site is indicated by ▌. The TATA box is highlighted in yellow and the start of translation codon (ATG) is underlined and in bold

4.3.3 Allelic variants and transcription factor binding sites

Figure 4.5 shows predicted transcription factor binding sites (TFBS) within the *MUC5B* proximal promoter most of which have been previously described in the literature (Chen *et al.* 2001; Wu *et al.* 2007a). Only the two Sp1 TFBS have been experimentally determined (Wu *et al.* 2007a). Wu et al have shown that binding of the transcription factor Sp1 to the binding sites located in the *MUC5B* promoter region, is induced by phorbol 12-myristate 13-acetate (PMA). PMA is used to model inflammation and it is therefore proposed that the Sp1 TFBS might play an important role during airway inflammatory disease. It should be noted that all other TFBS reported by Chen et al and Wu et al have only been predicted using the TRANSFAC database and are thus considered to be putative. Three additional predicted binding sites (STAT -919/-912, HNF4 and AP2α) were identified during this project using the web based tool rVista (see materials and methods for details and web address), which predicts cis-regulatory elements by combining TFBS database searches (TRANSFAC professional V10.2 library) with comparative sequence analysis. In this instance the mouse homologous *MUC5B* promoter sequence was used as the comparison sequence. The rodent promoter sequence was chosen as opposed to a primate sequence since the human and mouse are more distantly related and therefore high conservation between these species is likely to be indicative of function. In order to maximise the stringency of the predictions, the matrix similarity was set at 90% and only high-specificity matrices were selected. The rVista tool was also able to predict the MYC (-106/-95) and Sp1 (-124/-114) TFBS that had previously been identified in the literature (Wu *et al.* 2007a).

On examining SNP positions with respect to predicted TFBS, SNPs rs56235854, rs7115457 and rs885455 were shown to fall within MYC, NFκB and STAT binding sites respectively (see figure 4.5).

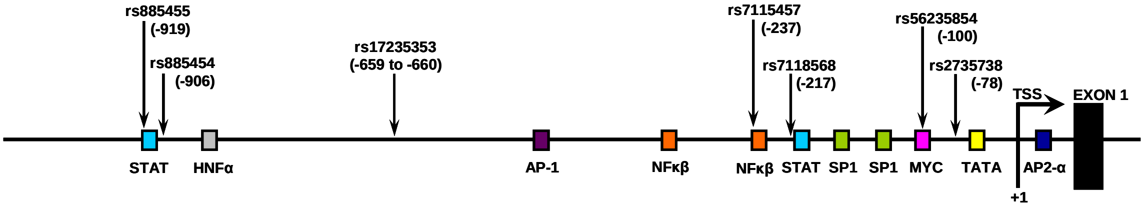


Figure 4.5 Diagrammatic representation of the *MUC5B* proximal promoter with annotated transcription factor binding sites (TFBS) and identified allele variants. The annotated binding sites are located at the following positions relative to the transcription start site; STAT -919/-912, HNFα -866/-861, AP1 -500/-493, NFκB -373/-364, NFκB -238/-229, STAT -213/-205, Sp1 -196/-185, Sp1 -124/-114, MYC -106/-95, TATA -32/-26, AP2α +49/+57. Note that STAT -919/-912, HNF4 and AP2α binding sites were identified during this project with rVista using the human and mouse conservation profile. All other TFBS had been previously identified by either Chen et al or Wu et al (Chen *et al.* 2001; Wu *et al.* 2007a).

4.3.3.1 Genotypic data

Because of the small size of the data sets, the genotypes for all SNPs were grouped into two categories, common homozygotes and carriers of the rare allele (heterozygotes and rare homozygotes). In the severe asthmatic cohort the SNPs shown to significantly associated with asthma at the allelic level, also show statistically significant association with respect to the grouped genotypes. In fact the significance is increased with a chi squared p value of 0.008 for both rs7115457 and rs7118568 and as can be seen in the bar charts in figure 4.6 the severe asthmatic cases show an underrepresentation of the genotypes carrying the rare allele and have a greater proportion of common homozygotes as compared to the hypernormal control sample set. This pattern of association suggests that the rarer alleles confer protection irrespective of genotype.

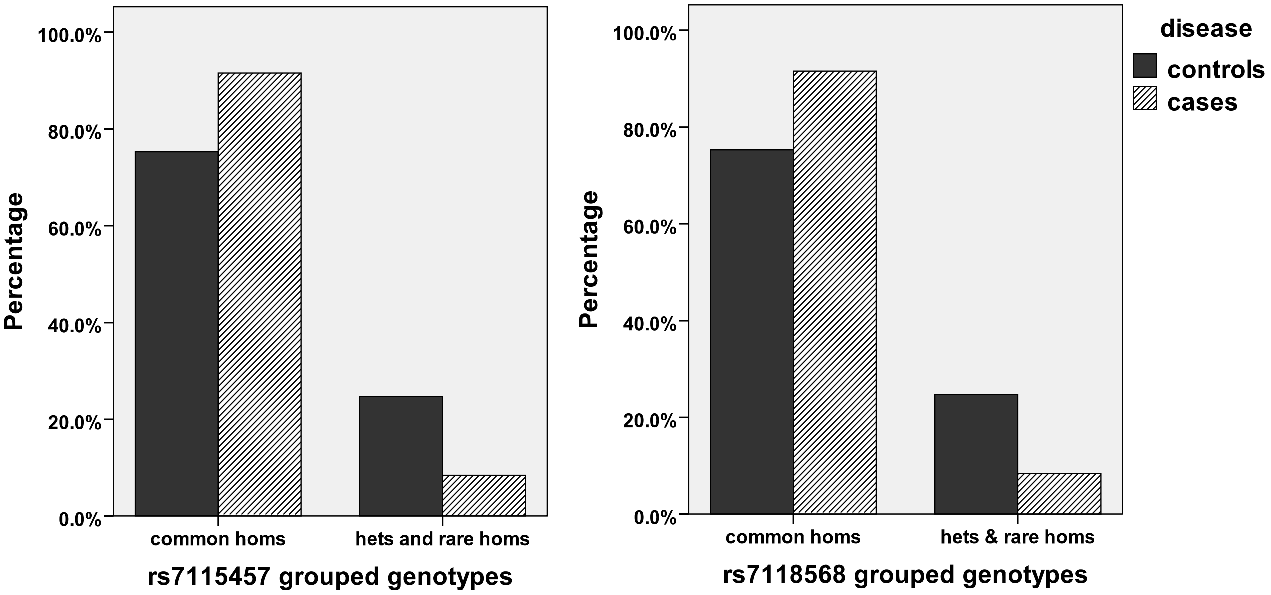


Figure 4.6 Grouped genotype distributions for rs7115457 and rs7118568 between cases and controls in the severe asthmatic cohort. Significant association has been noted for the combined genotypes of both SNPs with a p value of 0.008 for both (2-tail Pearsons chi squared).

In the matched asthmatic cohort the SNP exhibiting borderline significance with respect to allelic distributions rs56235854, is shown to be significantly associated with asthma in this sample set when analysed as grouped genotypes with a chi squared p value of 0.038. It is noteworthy that the pattern of association for the genotype distributions of rs56235854 are very similar to rs7115457 and rs7118568 (figure 4.7) in the severe cohort, in which the common homozygotes are overrepresented in the cases and the heterozygotes/rare homozygotes are underrepresented in the disease sample set.

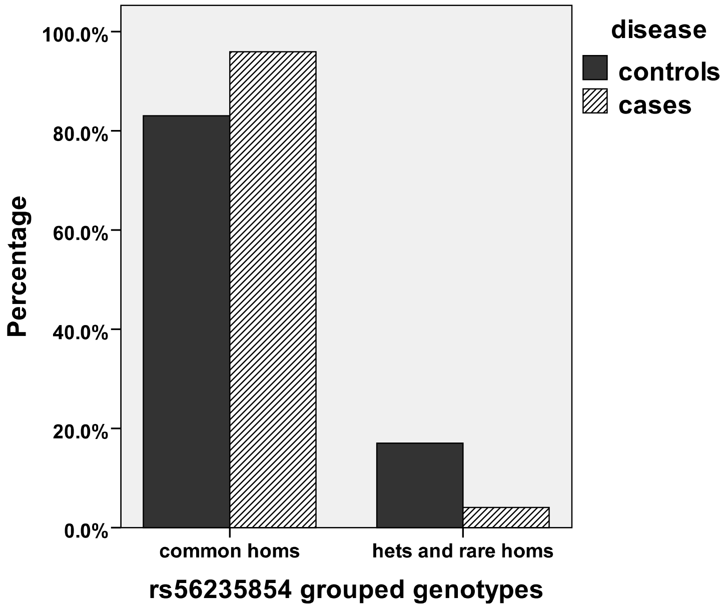


Figure 4.7 Grouped genotype distribution for rs56235854 between cases and controls in the matched asthmatic cohort The grouped gentotype distributions are significantly associated with the asthmatic phenotype with a a p value of 0.038 (Pearsons 2 tail chi squared).

4.3.3.2 Haplotypic data

MUC5B proximal promoter haplotypes for both disease cohorts were constructed using the haplotype inference software PHASE. Haplotypes were assigned identifiers in accordance with those previously identified by Kamio et al. A total of 8 proximal promoter haplotypes were identified in the matched asthmatic cohort and 7 in the severe asthmatic cohort (see table 4.5).

Table 4.5 ***MUC5B*** **proximal promoter haplotype details and frequencies in the matched and severe asthmatic cohorts.**

Haplotype ID	rs885455	rs885454	rs17235353	**rs7115457**	**rs7118568**	**rs56235854**	rs2735738	Matched asthmatic	Matched atopic	Severe asthmatic	Severe controls
H3	A	G	I	**G**	**C**	**G**	T	0.74	0.66	0.67	0.67
H1	G	G	I	**A**	**G**	**G**	C	0.07	0.11	0.05	0.13
H8	A	G	I	**G**	**C**	**G**	C	0.07	0.04	0.09	0.10
H4	G	A	I	**G**	**C**	**G**	C	0.03	0.06	0.07	0.04
H3a	A	G	I	**G**	**C**	**A**	T	0.02	0.09	0.07	0.02
H6	G	G	I	**G**	**C**	**G**	T	0.03	0.02	0.04	0.04
H5	G	G	I	**G**	**C**	**G**	C	0.03	0.01	0.00	0.00
H9	G	A	I	**G**	**C**	**G**	T	0.00	0.01	0.00	0.00
?	A	A	I	**G**	**C**	**G**	T	0.00	0.00	0.009	0.00

Haplotype IDs are in accordance with the Kamio haplotypes (Kamio 2005). H3a is a novel haplotype as it includes the rare allele of the SNP rs56235854 which was not identified in the Japanese population. Note that the rs56235854 rare allele only occurs on the H3 haplotypic background. A severe asthmatic sample SA053 also contains a novel haplotype not previously identified. SNPs in bold have been shown to be significantly associated with asthma in either the severe or matched asthmatic cohorts.

It has previously been shown *in vitro* by reporter construct assays, that H1 haplotype confers the highest expressional activity of the three haplotypes H1, H2 and H3 (H2 not identified in this population), which has been supported by an *in vivo* study (Loh *et al.* 2010). Thus our prior hypothesis was that the H1 haplotype would be overrepresented in the asthmatic samples and therefore all haplotypes were grouped into two categories; H1 or not H1. 2 by 2 contingency tables were constructed in order to compare haplotype counts in asthmatic and control samples and chi squared test were used to test the tables for deviations in haplotype count distributions. Statistical significance was noted in the severe asthmatic cohort with a p value of 0.03 (2 tailed) whereby there is a decrease of H1 haplotypes in the asthmatic sample compared with their controls (see figure 4.8). A similar pattern can be seen in the matched asthmatic cohort where the H1 haplotype frequency is greater in the atopic controls (0.11) than in the asthmatic cases (0.07), though statistical significance is not reached. Thus a significant association between the H1 haplotype and asthma has been noted although the identified pattern of association is opposite from that which we had hypothesised.

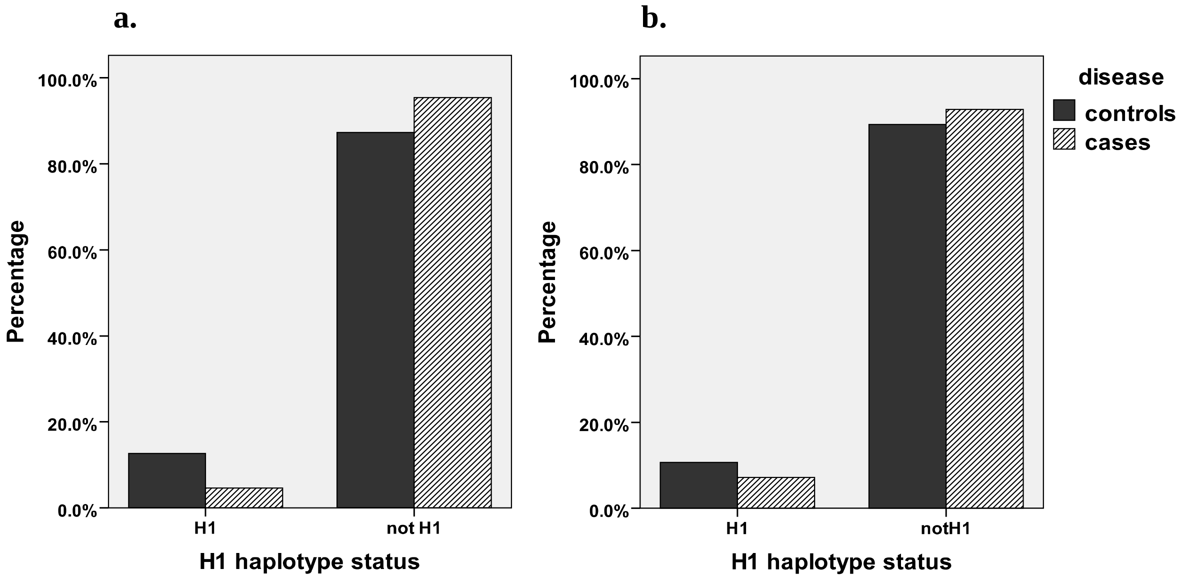


Figure 4.8 Bar charts showing the distribution of H1 and non H1 haplotypes between the cases and controls of the a. severe asthmatic cohort b. matched asthmatic cohort. The distribution depicted in chart a was shown to be statistically significant with a p value of 0.03 (2 tail Pearsons chi squared). The distribution shown in chart b was not shown to be significantly different.

4.3.4 *MUC5B* distal promoter

A preliminary study of the *MUC5B* putative distal promoter (-2268 to -986 relative to the TSS) reported by Perrais et al (Perrais *et al.* 2001) has been conducted during this project.

Within this region 24 SNPs have been reported to the NCBI SNP database, however only 9 of these SNPs are accompanied by supporting polymorphic data within a European population. The aim of this preliminary study was to characterise all sequence variants within the putative distal promoter region in both asthmatic cohorts. However due to limited DNA supply only the matched asthmatic cohort has been included in this preliminary study. RT-PCR was also used to evaluate the validity of this region as an active promoter, since this promoter has been reported only once in the literature and therefore little functional information is available.

4.3.4.1 Allelic data

Within the putative distal promoter region 7 polymorphic SNPs were identified in the matched asthmatic sample set (see appendix 3 for annotated sequence), all of which correspond to previously reported SNPs. Allele frequencies range from 0.006 to 0.49 (table 4.6) and only rs868903 is represented in the HapMap project for which the MAF for the CEPH sample is in good agreement with the disease cohort reported here. All genotypic distributions were in accordance with those expected under HWE.

Table 4.6 Details of the *MUC5B* SNPs located within the putative distal promoter.

SNP ID	SNP Location	MAF
rs11042646	-2019	0.16 (T)
rs55974837	-2002	0.03 (T)
rs35619543	-1996	0.26 (T)
rs12804004	-1947	0.44 (T)
rs868902	-1738	0.45 (A)
rs868903	-1556	0.49 (T)
rs868904	-1437	0.006 (T)

MAF is minor allele frequency, minor alleles are in parentheses, SNP location refers to base position relative to transcription start site. Data for 186 chromosomes were available for rs11042646, rs55974837, rs35619543, rs12804004 and rs868902. Data for 174 chromosomes were available for rs868903 and rs868904.

2 by 2 contingency tables of allele counts were constructed to compare allelic distributions between asthmatic and atopic control samples and tables were tested using a Fishers exact test. No statistically significant associations could be identified between allelic counts and asthma.

4.3.5 Looking for the putative *MUC5B* distal promoter: RT PCR

A previous study has identified active *MUC5B* distal promoter activity in the Kato III gastric carcinoma cell line and has also identified a transcript corresponding to the distal promoter in tracheal RNA using a primer extension method. However no other study has addressed the activity of this distal promoter. In this preliminary study we have used RT-PCR to look for transcripts corresponding to an active *MUC5B* distal promoter in clones of a mucus secreting colon cancer cell line, normal human bronchial epithelial (NHBE) cells grown at air liquid interface (ALI) and fetal lung tissue samples.

In an effort to identify distal promoter activity, three different RT-PCR assays were performed. Two assays were designed to amplify any 5′ UTR that may correspond to a distal promoter transcript, while the third assay was designed as a control for *MUC5B* expression; assay 1 aimed to amplify from the 5′ region of the reported transcript to exon 1; assay 2 from the same region to exon 2; the control assay from exon 1 to exon 2.

Access was available to cDNAs from the following resources; various subclones of the mucus secreting HT29-MTX cell line, at both exponential and stationary phase (n = 8); fetal lung tissues (n = 2); NHBE cells grown at air liquid interface (n = 1). The PCRs were conducted on these cDNAs and *MUC5B* expression was validated in all HT29-MTX subclones, fetal lung tissue and NHBE ALI cells. However, no transcripts corresponding to expected transcripts from the putative *MUC5B* distal promoter (assays one and two) could be identified for any cDNA sample. Figure 4.9 shows a representative experiment and figure 4.10 depicts the *MUC5B* 5′ UTR sequence which has been annotated for all primers and important motifs.

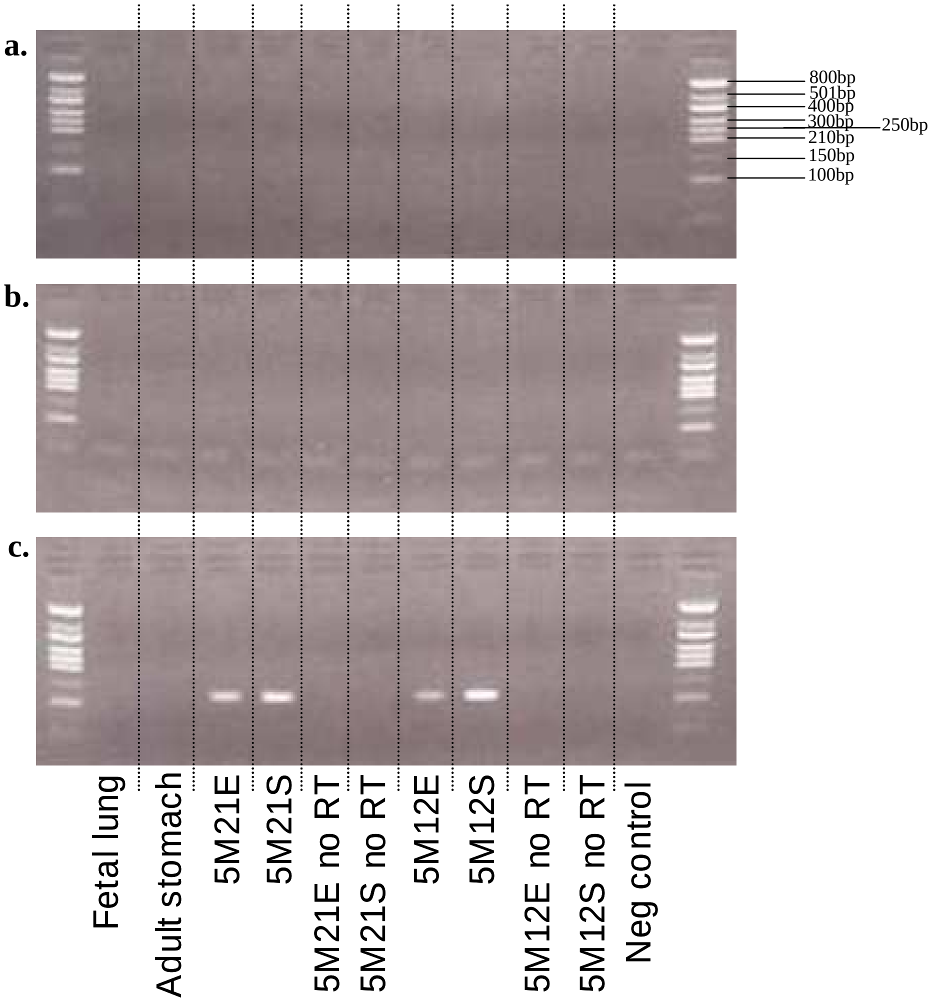


Figure 4.9 ***MUC5B*** **putative distal promoter RT-PCR product visualisation.** a. RT-PCR products aimed to amplify from the 5′ region of the reported transcript to exon 1. b. RT-PCR products aimed to amplify from the 5′ region of the reported transcript to exon 2. c. *MUC5B* control RT-PCR, product visualisation corresponding to a transcript spanning from exon 1 to exon 2. The colonic cell line HT29 was differentiated into the two clones 5M21 and 5M12 by treatment with 10^{-5} M methotraxate. 5M21 has a mucus secreting phenotype. 5M12 has a enterocytic phenotype. E refers to exponential phase. S refers to stationary phase. All RT-PCR products were visualised on 2% agarose gels,

CTGTGACG**TAAATAAA**ACAACAGACCTGGA**C**ACCACCCTAGGGTCCCCATGGGGCCG
distal TATA box **distal TSS**

G**ACGAGGCCACACCACCCGA**CCTGGTGCTTCCTGCCTGGCGTCTGCGCCACGGAGCA
distal FOR primer

TT**CAGGACGCTGGTGACCAGGGA**GCCAGGAGGTGGGAGCATCTGAGGTGCAGGTCAC
NAU 647 (PE REV)- Perrais et al

ACGGGCAGGAGGTGTTTGCAAGAGGTATTGCAGCGGACGGAGTGTCCTGCAGATG

···········891bp of proximal promoter sequence···········

AGCATGGTGCCCTGCGTGGCCTGGCCTGGCTTGGGGCCAAGGCTCCCTGC**TACATAA**
proximal TATA box

GCTGGGGCCCCCAGGGGAGCAAGCA**C**CCGGCCCGGCTCCCTCCCTGCCCGTCCCCGT
proximal TSS

CCCCCCACCCGTGCCAGCCCCCAGG**ATGGGTGCCCCGAGCGCGTGCCGGACGCTGGT**
exon 1 FOR primer

GTTGGCTCTGGCGGCCATGCTCGTGGTGCCGCAGGCAGGTAAGAGCCCCCCACTCCG
exon 1 REV primer

CCCCCTCTCGATGCTGTCTTCACGGCGGGGGTCTCTGCAGGTCGCTTGCCTGGGAGC

···············2,393bp of intron 1 sequence···············

CATTCCCTCTTCCCACA**GAGACCCAGGGCCCTGTGGAGCCGAGCTGGGAGAATGCAG**

GGCACACCATGGATGGCGGTATGTGGCCAGGTTCGGGGGTGGGGGGTTCCTGACCAG
exon 2 REV primer

GCTGGAGGGGCTGGAATTTGGGCTGGGGCAGGCAGACGCCTCTCCAAGCAGCCATGC

Figure 4.10 Annotated *MUC5B* sequence spanning from the putative distal TATA box to exon 2. Boxed sequences shaded in grey represent primers used for the RT-PCR, where FOR is forward and REV is reverse. Bold and underlined sequence represents exons. Proximal and putative distal transcription start sites (TSS) are highlighted in red. Proximal and putative TATA boxes are highlighted in yellow. Sequence highlighted in purple represents the reverse primer elongated by primer extension in order to establish the distal TSS in the Perrais et al study (Perrais *et al* 2001).

4.3.6 *MUC5AC* and *MUC5B* extended haplotypes

Significant disease associations were reported in chapter 3 between a 3′ end *MUC5AC* SNP and in this results chapter we report significant associations between severe asthma and 5′ *MUC5B* promoter SNPs, and some evidence of significant association between severe asthma and *MUC5AC* 3′ haplotypes. A previous study has identified a significant occurrence of an extended haplotype spanning from *MUC2* to *MUC5B* in people of European ancestry (Rousseau *et al.* 2007), and it was therefore postulated that the significant associations reported in this thesis could possibly reflect a single risk haplotype with respect to respiratory disease, thus extended haplotypes spanning from *MUC5AC* to *MUC5B* will be studied in this section with respect to disease status.

Extended haplotypes were inferred from all *MUC5AC* and *MUC5B* markers, except for the indels, the frequencies for which are shown in table 4.7.Three extended haplotypes dominate (E1, E2 and E3), although E1 is very frequent making up 46% and 45% of the haplotypes seen cases and controls respectively. This confirms the findings of the Rousseau et al (Rousseau *et al.* 2007) study though it should be noted that the extended haplotypes examined in this project are smaller since they do not include *MUC2* marker data.

The high expressing *MUC5B* promoter haplotype H1 (shading in blue in table 4.7), occurs on four of the extended haplotypes and it is noteworthy that in each case the H1 carrying haplotype is at a greater frequency in the control sample, the most extreme examples of this being E9 and E17. This finding is consistent with the results reported in section 4.3.3.2 whereby the H1 haplotype is at significantly increased levels in the severe asthmatic control group.

The distribution of the *MUC5AC* haplotypes examined in section 4.3.1.4 showed some evidence of significant difference between cases and controls in the severe asthmatic cohort and from viewing these data graphically (figure 4.3) it appeared that the HA and HC distributions were the cause of this association.

It is evident that both the HA and HC *MUC5AC* haplotypes (orange and red respectively) mainly occur with the H3 *MUC5B* haplotype (shaded pink in table 4.7). Only the extended haplotype E17 includes both the HA and H1 haplotypes, both of which have been shown previously in this chapter to be at increased numbers in the control sample set. Thus the statistically significant *MUC5AC* and *MUC5B* haplotype associations reported with asthma in this chapter are likely to be independent.

Table 4.7 Extended haplotypes ranging from *MUC5AC* to *MUC5B*, in asthmatic cases and controls

	MUC5AC							*MUC5B*							
Haplotype ID	rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088	rs1132440	rs885455	rs885454	rs7115457	rs7118568	rs56235854	rs2735738	Case frequency	Control frequency
E1	G	C	C	T	A	C	C	A	G	G	C	G	T	116	119
E2	G	T	T	C	G	T	G	A	G	G	C	G	T	34	16
E3	A	C	C	T	G	C	G	A	G	G	C	G	T	24	32
E4	A	C	C	T	G	C	G	A	G	G	C	G	C	9	6
E5	A	C	C	T	G	C	G	G	A	G	C	G	C	9	5
E6	G	C	C	T	A	C	C	A	G	G	C	G	C	9	10
E7	G	C	C	T	A	C	C	A	G	G	C	A	T	8	10
E8	G	C	C	T	A	C	C	G	G	G	C	G	T	8	8
E9	A	C	C	T	G	C	G	G	G	A	G	G	C	5	15
E10	G	T	T	C	G	T	G	G	G	A	G	G	C	5	6
E11	A	C	C	T	G	C	G	A	G	G	C	A	T	4	2
E12	G	C	C	T	A	C	C	G	G	G	C	G	C	3	0
E13	G	C	C	C	G	T	G	A	G	G	C	G	T	3	6
E14	G	T	T	C	G	T	G	G	A	G	C	G	C	3	6
E15	A	C	C	T	G	C	G	G	G	G	C	G	C	2	0
E16	A	C	C	T	G	C	C	A	G	G	C	G	T	2	0
E17	G	C	C	T	A	C	C	G	G	A	G	G	C	2	9
E18	G	C	C	C	G	T	G	G	G	A	G	G	C	2	3
E19	G	C	C	T	G	C	C	A	G	G	C	G	T	1	0
E20	G	C	C	T	A	C	C	G	A	G	C	G	C	1	2
E21	G	T	T	C	G	T	G	A	A	G	C	G	T	1	0
E22	G	T	T	C	G	T	G	G	G	G	C	G	T	1	1
E23	G	T	T	C	G	T	G	A	G	G	C	G	C	0	3
E24	G	C	C	T	A	C	C	G	A	G	C	G	T	0	1
E25	G	C	C	T	A	C	G	A	G	G	C	G	T	0	1
E26	A	C	C	C	G	T	G	A	G	G	C	G	C	0	1

All previously described *MUC5AC* and *MUC5B* variants have been included in the haplotype construction except for the *MUC5B* promoter indel since it is monomorphic in this data set. Haplotypes were inferred using PHASE. Blocks shaded pink and blue refer to the *MUC5B* H3 and H1 haplotypes respectively. Blocks shaded orange, green and red refer to the *MUC5AC* haplotypes HA, HB and HC. The case sample set includes asthmatic individuals from both the matched and severe disease cohorts. The control sample set includes the non-asthmatic controls from both the severe and matched disease cohorts. The difference in haplotype distributions between asthmatic cases and controls was not shown to be statistically significant when examined by PHASE case-control (p value = 0.4).

4.3.7 LD across *MUC5AC* and *MUC5B*

It was also of interest to examine whether there was a different pattern of LD between cases and controls. Measures of pairwise LD were calculated using data from only the *MUC5AC* and *MUC5B* SNPs with minor alleles greater than 10 per cent. The pairwise D′ measures shown in table 4.8 were calculated separately for case and control groups, which represent an amalgamation of both the matched and severe asthmatic disease cohorts.

The first analysis was performed using the standard procedure LDmax, and the results are shown in table 4.8a. In general, complete LD is seen across *MUC5AC*. Breakdown of LD between the *MUC5AC* markers rs35968147 and rs34666042 is however seen in both the cases and controls. This apparent breakdown is rather unexpected since the rare alleles of these two SNPs appear together on one of the most frequent haplotypes, HC (see table 4.3).

D′ values less than one are usually interpreted as evidence of historic recombination. However, this assumes correct haplotype inference and for the markers to be in HWE. Since these SNPs have minor allele frequencies greater than 10 per cent and no significant deviation from HWE, this unexpected finding is unlikely to have resulted from a frequency issue and in this case it seems to be a product of the EM algorithm used by LDmax in order to infer phase.

In order to examine these inconsistencies, an unconventional approach was taken by applying the PHASE inferred individual haplotypes to the software HaploXT which is generally used for families. As can be seen in table 4.8b, the breakdown of LD between rs35968147 and rs34666042 is no longer an issue in the HaploXT pairwise results. Thus the pairwise LD measures generated by HaploXT appear to be more reliable. Significant association can be seen between the *MUC5B* markers and several of the *MUC5AC* SNPs, however there is no obvious suggestion that this differs between cases and controls.

Table 4.8 Pairwise LD measures for *MUC5AC* and *MUC5B* markers.

a.

LDmax

Controls

MUC5AC	rs34666042	**1.000**							
	rs34831688	**1.000**	**1.000**						
	rs35968147	0.998	**0.664**	**1.000**					
	rs2075844	**1.000**	**1.000**	**1.000**	**1.000**				
	rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**			
	rs1132440	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**		
MUC5B	rs885455	0.070	0.127	0.046	0.151	0.232	0.094	0.224	
	rs2735738	0.120	0.309	0.271	0.183	**0.320**	0.184	**0.313**	**0.737**
		rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088	rs1132440	rs885455
		MUC5AC							*MUC5B*

Cases

rs34666042	**1.000**							
rs34831688	**1.000**	**1.000**						
rs35968147	**1.000**	**0.722**	**1.000**					
rs2075844	**1.000**	**1.000**	**1.000**	**1.000**				
rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**			
rs1132440	**0.915**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**		
rs885455	0.178	0.001	0.018	0.038	**0.431**	0.058	**0.422**	
rs2735738	**0.243**	0.424	0.338	0.332	**0.376**	0.186	**0.351**	**0.792**
	rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088	rs1132440	rs885455

b.

HaploXT

Controls

MUC5AC	rs34666042	**1.000**							
	rs34831688	**1.000**	**1.000**						
	rs35968147	**1.000**	**1.000**	**1.000**					
	rs2075844	**1.000**	**1.000**	**1.000**	**1.000**				
	rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**			
	rs1132440	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**		
MUC5B	rs885455	**0.267**	0.163	0.163	0.163	**0.375**	0.163	**0.370**	
	rs2735738	**0.357**	**0.297**	**0.297**	**0.297**	**0.475**	**0.297**	**0.471**	**0.715**
		rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088	rs1132440	rs885455
		MUC5AC							*MUC5B*

Cases

rs34666042	**1.000**							
rs34831688	**1.000**	**1.000**						
rs35968147	**1.000**	**1.000**	**1.000**					
rs2075844	**1.000**	**1.000**	**1.000**	**1.000**				
rs28728088	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**			
rs1132440	**0.914**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**		
rs885455	**0.364**	0.008	0.008	0.008	**0.483**	0.008	**0.498**	
rs2735738	**0.549**	0.130	0.130	0.130	**0.675**	0.130	**0.684**	**0.814**
	rs2075843	rs34666042	rs34831688	rs35968147	rs2075844	rs28728088	rs1132440	rs885455

Pairwise D′ measures of LD for cases and controls have been calculated for markers across *MUC5AC* and *MUC5B* using two different methods, LDmax and HaploXT. Pairwise measures of D′ shown to be statistically significant with p values less than 0.05 are in bold. Cases include the asthmatic affected groups from both the severe and matched cohorts. Controls include the non-affected groups from both the severe and matched asthmatic cohorts.

4.4 Discussion

This chapter has explored variation within the genes that code for the respiratory mucins, in relation to asthma. However the most noteworthy finding involves variation within the *MUC5B* promoter, whereby the high expressing H1 haplotype is shown to be significantly underrepresented in the severe asthmatic cases as opposed to their controls, a pattern of association opposite from that hypothesised. It appears to be counter-intuitive that the H1 expressing haplotype should be at a significantly lower frequency within a sample set of hypersecretory disease cases. However it should be remembered that although mucus hypersecretion is likely to exacerbate asthmatic symptoms, the cause of asthma is likely to be allergy related and therefore under some circumstances, higher airway mucus levels may even confer protection against the development of asthma by preventing penetration of insults and allergens into the epithelial cells with a greater efficiency.

The rare alleles of variants rs7115457 and rs7118568, define the H1 haplotype. These two variants were shown to be significantly associated with severe asthma, associations that are likely to be dependent since the SNPs are completely linked. Variant rs7115457 is located within the putative nuclear factor-kappa B (NFκB) binding site located at -238/-229 within the *MUC5B* promoter. NFκB is said to be a prominent transcription factor in chronic airway disease (Barnes 2006). Its active form is located in the nucleus where it binds directly to NFκB promoter motifs. NFκB binding sites can be found on various target inflammatory genes and since the respiratory mucins genes are known to be upregulated by inflammation, the predicted NFκB motifs within the *MUC5B* promoter are likely to be functional.

NFκB is thought to be activated by various stimulants such as, cytokines, viral infections and oxidants such as ozone (Barnes 2006), all of which are known to exacerbate asthmatic symptoms. Variation within the NFκB binding sites of the *MUC5B* promoter may therefore confer susceptibility to asthma or increase asthmatic severity since the variants could alter binding of NFκB to its motif. For instance, during chronic inflammatory airway disease the epithelial surface is immersed in inflammatory cytokines which will in turn lead to the activation of NFκB. In essence this could be the causal event of *MUC5B* aberrant expression in the goblet cells during chronic airway disease. The extent of this aberrant *MUC5B* expression may however by dependent on variation within the NFκB regulatory motif in the promoter.

At the outset of this project it was hypothesised that activation of the putative distal promoter may be the causal event of *MUC5B* aberrant expression in the goblet cells, and therefore preliminary experiments were conducted to characterise and validate the putative *MUC5B* distal promoter proposed by Perrais et al (Perrais *et al.* 2001). However all attempts to isolate transcripts corresponding to an active distal promoter proved unsuccessful and correspondence with the laboratory who had proposed this second promoter proved unsatisfactory. It was subsequently concluded that no evidence supported the activity of a distal promoter and therefore investigations were not furthered.

Although the significant association between the *MUC5AC* SNP rs1132440 identified in chapter 3 could not be confirmed with either asthmatic cohort, the negative results may be a result of reduced power since both disease cohorts are much smaller than the three thousand samples genotyped in chapter 3.

A potentially very interesting finding was made during the analyses in this section with regard to calculating measures of pairwise LD. It appears that some of the breakdowns in LD identified in the LDmax output, may in fact by spurious and a direct result of the EM algorithm used to infer haplotypes prior to LD measure calculation. This issue could pose a great problem since the EM algorithm is also used by Haploview outputs, made publically available by the HapMap project.

Chapter 5

Exploring copy number variation in *MUC5AC*

5 Exploring copy number variation in *MUC5AC*

5.1 Introduction

This chapter considers the possibility that the *MUC5AC* gene is affected by copy number variation (CNV). This introductory section aims to illustrate how wide-spread CNV is thought to be throughout the human genome.

5.1.1 Genome-wide methods for the detection of CNV

Copy number variation or rather a copy number variant, is defined as a DNA segment of 1kb or greater that is detected at different numbers of copies when comparing at least two genomes. A CNV can either be a duplication or deletion but defining which of these categories it belongs to is difficult since the CNV detected is relative to a reference genome (Scherer *et al.* 2007). A CNV becomes a copy number polymorphism (CNP) if the frequency of the variant reaches 1% or greater in a population. Overlapping CNVs are generally merged to produce CNV regions (CNVR). The publically available Database of Genomic Variants (DGV) contains all reported CNVRs which number 8410 at present (February 2010).

Genome-wide microarrays have been very important for the advancement in understanding the extent of copy number diversity within the human genome. A method known as array comparative genomic hybridisation (aCGH) is widely used to detect CNV at the genome-wide level. Test and reference DNA samples are differentially labelled with fluorescence and simultaneously hybridised to the probes present on a microarray. The fluorescence ratio of each probe is measured in order to compare the DNA dosage for each sample at the specific genomic region represented by that probe. Therefore when defining a CNV it can only be said that the test genome has more or less of the DNA segment in comparison to the reference genome, the actual structure of the variant can not be defined by this method.

Various probes have been used for aCGH such as, large-insert clones (40-200kb), and oligonucleotides (25-80bp). Large-insert clone arrays have the best coverage of the genome and the hybridisations are generally the most reliable i.e. low-noise (Carter 2007). However since the DNA inserts are so large they provide low-resolution because it is difficult to refine the CNV boundaries and therefore CNV sizes are likely to be overestimated (Redon *et al.* 2006). Oligonucleotide probes give a much higher resolution, but this is accompanied by high-noise which increases the risk of false positive CNV calls. In order to reduce noise an adapted procedure has been developed known as representational oligonucleotide microarray analysis (ROMA). The test and reference DNA are first digested with a restriction enzyme and then the fragments are amplified by PCR. Small fragments are preferentially amplified and the selected oligonucleotide probes correspond to these amplified fragments (Carter 2007; Sebat *et al.* 2004).

The experimental methods described above detect potential CNV by directly comparing the hybridisation intensities of reference and test DNA samples which are indicative of DNA dosage. However methods have also been developed in which the hybridisation of a single test DNA sample to a microarray is sufficient for CNV detection, such as genotyping array platforms. SNP arrays are primarily used for high-throughput genotyping. Both matched and mismatched probes of 25 bases are present on the array for each allele of a SNP. Thus the intensities at each probe will indicate which alleles are present for that specific SNP in the test sample. As with ROMA, the test DNA is digested and amplified prior to hybridisation. The intensity data can also be used to detect CNV, and special algorithms are used in order to make more reliable calls. Algorithms used such as PennCNV, make calls based on genotyping signal intensities, allele frequencies, distance between neighbouring SNPs, and allelic intensity ratios (Wang *et al.* 2007).

5.1.2 The surprising extent of CNV in the human genome

Although CNV has been associated with disease (Breunis *et al.* 2008; Hollox *et al.* 2008) whole genome microarray studies have identified CNV as a considerable source of genetic diversity in the genomes of healthy individuals (Iafrate *et al.* 2004; Jakobsson *et al.* 2008; Levy *et al.* 2007; Perry *et al.* 2008; Redon *et al.* 2006; Sebat *et al.* 2004). A comprehensive study (Redon *et al.* 2006) examining a HapMap collection of 270 individuals from four populations (Yoriba, Japanese, Han Chinese and European), has suggested that as much as 12% of the human genome is affected by CNV. This study also highlights that many genes are encompassed by the 1447 discrete CNVRs identified.

The extent of CNV in the healthy genome is surprising. However it must be remembered that such alterations in genome architecture will not always alter function. Many CNVs will have a neutral affect on gene expression levels since duplication or deletion events may not even affect a gene. Even in the case of genes, a deletion will not be detrimental if the remaining gene copy is sufficient for normal function and if a gene is duplicated, expression levels will only increase if appropriate regulatory elements are acting on the duplicate and if the new chromatin context allows expression (Henrichsen *et al.* 2009).

5.1.3 Does CNV affect *MUC5AC*?

The aim of this chapter is to explore the possibility that CNV affects *MUC5AC*. The notion that *MUC5AC* falls within, or at least partially within a CNVR was introduced in chapter 1. In chapter 1 genotype data for rs1132440, a SNP in exon 19 of the *MUC5AC* 3′ region, was analysed in the 1946 longitudinal birth cohort and was shown to be significantly associated with five non-independent respiratory outcomes. However the causative genetic component remains unclear since deviations from HWE highlighted an excess of rs1132440 heterozygotes in the hayfever and allergy affected sample sets.

Unexpected numbers of heterozygotes could be indicative of CNV since the variant alleles would no longer be present at only two copies and as can be seen in figure 5.1 various copy number scenarios will result in 'heterozygote' genotyping. It was therefore proposed that copy number affects the *MUC5AC* gene, rs1132440 falls within this CNVR and that *MUC5AC* CNV may be the true causal genetic component with respect to the risk of hayfever and allergy phenotypes.

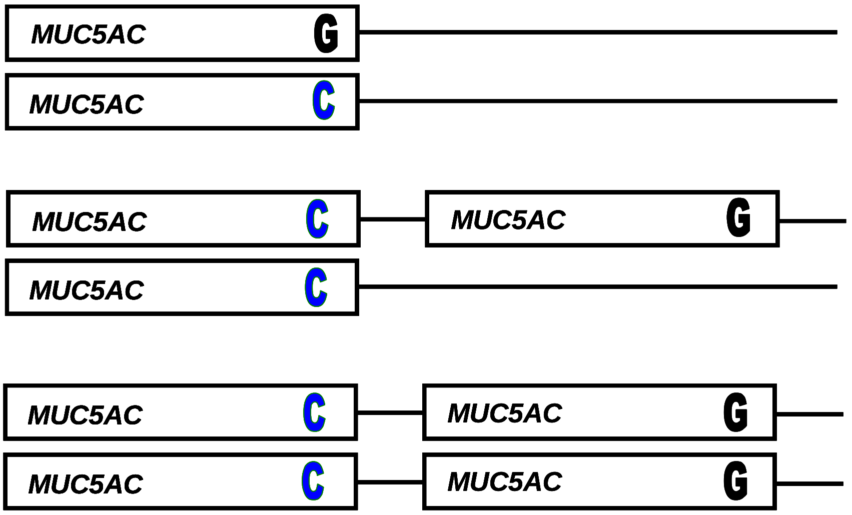


Figure 5.1 True and artificial heterozygotes. Alternative types of possible diplotypes for rs1132440 if located within a CNV region. Three diploid genomes are depicted here with two, three or four copies of the entire or partial MUC5AC gene. For all three scenarios genotyping methods will infer heterozygosity.

At present four genome-wide studies have reported CNVs and CNVRs within or including, the 11p15.5 *MUC* gene complex (see figure 5.2). In one study using a whole genome tile path approach, a CNVR that encompasses the whole 11p15.5 MUC gene complex, has been identified in a single individual (Redon *et al.* 2006). However this approach uses BAC clones as probes and thus the resolution of this method is low and we are unable to refine the CNV call. Another study has also identified a CNVR within the *MUC5AC* and *MUC5B* region using the array comparative genomic hybridisation approach with 60mer oligo probes (Perry *et al.* 2008).

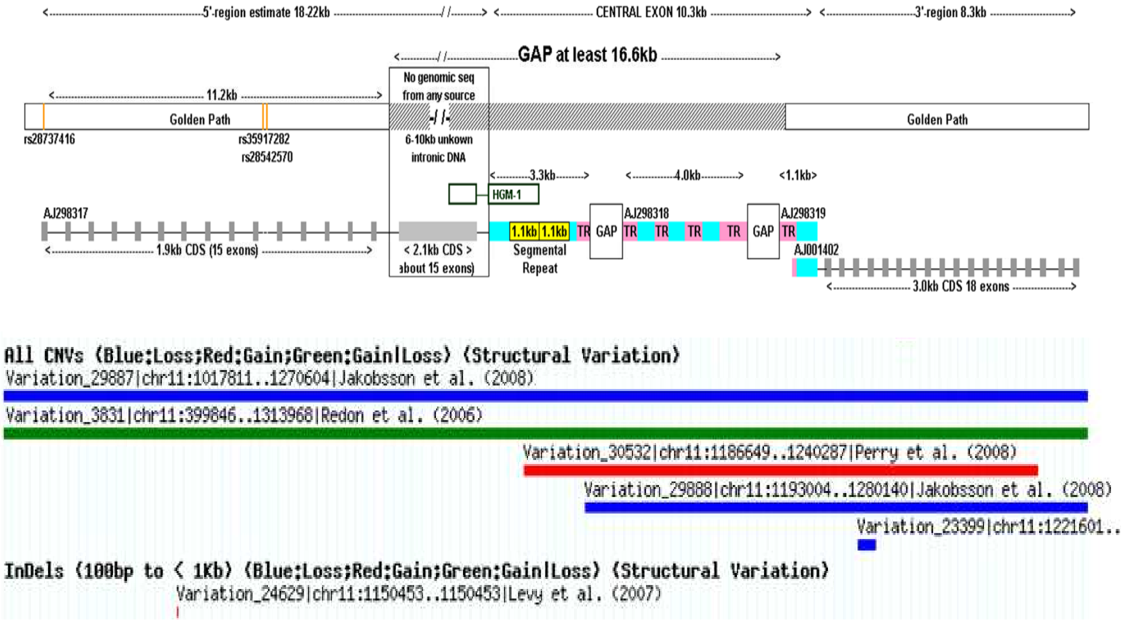

5'-region estimate 18-22kb
CENTRAL EXON 10.3kb
3'-region 8.3kb
GAP at least 16.6kb
11.2kb
No genomic seq from any source
Golden Path
Golden Path
rs28737416
rs35917282
rs28542570
6-10kb unkown intronic DNA
3.3kb
4.0kb
<1.1kb>
HGM-1
AJ298317
AJ298318
AJ298319
1.1kb
1.1kb
TR
GAP
TR
TR
TR
TR
GAP
TR
AJ001402
1.9kb CDS (15 exons)
< 2.1kb CDS > about 15 exons)
Segmental Repeat
3.0kb CDS 18 exons
All CNVs (Blue:Loss;Red:Gain;Green:Gain|Loss) (Structural Variation)
Variation_29887|chr11:1017811..1270604|Jakobsson et al. (2008)
Variation_3831|chr11:399846..1313968|Redon et al. (2006)
Variation_30532|chr11:1186649..1240287|Perry et al. (2008)
Variation_29888|chr11:1193004..1280140|Jakobsson et al. (2008)
Variation_23399|chr11:1221601..
InDels (100bp to < 1Kb) (Blue:Loss;Red:Gain;Green:Gain|Loss) (Structural Variation)
Variation_24629|chr11:1150453..1150453|Levy et al. (2007)

Figure 5.2 Representations of *MUC5AC* with the reported CNVRs. The top half of this figure shows a graphical representation of the MUC5AC gene including the sequence golden path (May 2004 NCBI35/hg17) sequence gap (provided by Ralph Burgess, Undergraduate project student, G.E.E). The coloured blocks aligned beneath the gene diagram are from database of genomic variant (DGV) output, and indicate independently reported CNVRs within the *MUC5AC* region. (http://projects.tcag.ca/cgi-bin/variation/gbrowse/hg18/).

5.1.4 Sequencing strategy for CNV detection

Here we introduce a sequencing strategy for the detection of CNV and report its application with respect to *MUC5AC*. DNA sequencing has been widely used throughout this project as a direct method for SNP discovery and typing. In sequence traces, overlapping peaks present at single loci are inferred as heterozygous genotypes, and it is the examination of these heterozygous peak patterns that forms the basis of this CNV detection approach.

When amplifying regions of sequence that contain true heterozygous loci, equal quantities of fragments containing each allele will be amplified and therefore when sequenced we might expect to see equal peak sizes in the traces at the polymorphic position. Although the entire trace pattern is decidedly reproducible, the size of every peak depends on the local sequence context and terminator dye incorporation rates, therefore peak sizes will vary throughout the trace. Consequently the peaks representing each allele at a heterozygous position will not necessarily be equal. However the relative peak heights are constant from sample to sample in sequence traces from amplified products of the same DNA segment (Dobbs and Gee 2002; Kwok 2010).

5.2 Hypothesis and Aims

This chapter aims to establish a novel method for the determination of copy number variant regions by studying the relative peak patterns at heterozygote loci. We hypothesise that the relative peak heights or rather the peak patterns for heterozygous loci will in fact vary from sample to sample if the polymorphic locus resides within a CNVR due to the unequal amplification of alleles. Thus sequencing and subsequent heterozygote peak examination will be used as a semi-quantitative method and first port of call, to highlight potential CNVR.

5.3 Results

5.3.1 *MUC5AC* fragments sequenced

Two fragments were chosen in the *MUC5AC* 3′ region for subsequent sequencing. The first amplified fragment spans from exon 18 to exon 19 at the 3′ end of the gene. This region was chosen for sequencing since rs1132440 is located within exon 19. The second sequenced fragment spans from exon 13 to exon 14, and was chosen as a second 3′ region fragment since a high number of SNPs have been reported here which increases the chance of obtaining heterozygous peaks. All primers were designed within exons rather than introns because exonic sequences are less likely to accumulate variants if the proposed duplication remains functional.

Sequencing and peak pattern examination of the two 3′ fragments were performed for 176 samples composed of 84 severe asthmatics and 92 healthy controls (refer to materials and methods). Details of SNP discovery and typing data are given in chapter 4 and therefore this results section will only discuss the heterozygote peak examination of sequence traces.

5.3.2 Examining heterozygous peak patterns

Heterozygote peaks were examined at six SNP loci; rs2075843, rs34666042, rs34831688 within the exon 13 to 14 fragment, and rs35968147, rs2075844, rs28728088 within the exon 18 to 19 fragment. Only samples heterozygous for at least one SNP could be included in this analysis and of these only clean sequence traces were considered, reducing the total number of samples examined to 96 (severe asthmatic and healthy controls). Unfortunately the peaks corresponding to rs1132440 heterozygotes could not be examined since this locus was too close to the end of the trace where peaks are low in quality.

For all loci inspected, the heterozygote peak patterns clearly varied from sample to sample. Peak patterns were assessed for their degree of unevenness and were recorded as ‘even ’, ‘slightly uneven’ or ‘very uneven’ and a record made of which allele predominated. Figure 5.4a and b shows examples of heterozygote peaks classified as ‘very uneven’ and figure 5.4c gives an example of ‘even’ peak pattern.

As shown in the figure 5.3 example the observed asymmetric peak patterns were generally similar for all heterozygous loci within the same amplified product, as would be expected if this did indeed represent more copies of one allele than the other.

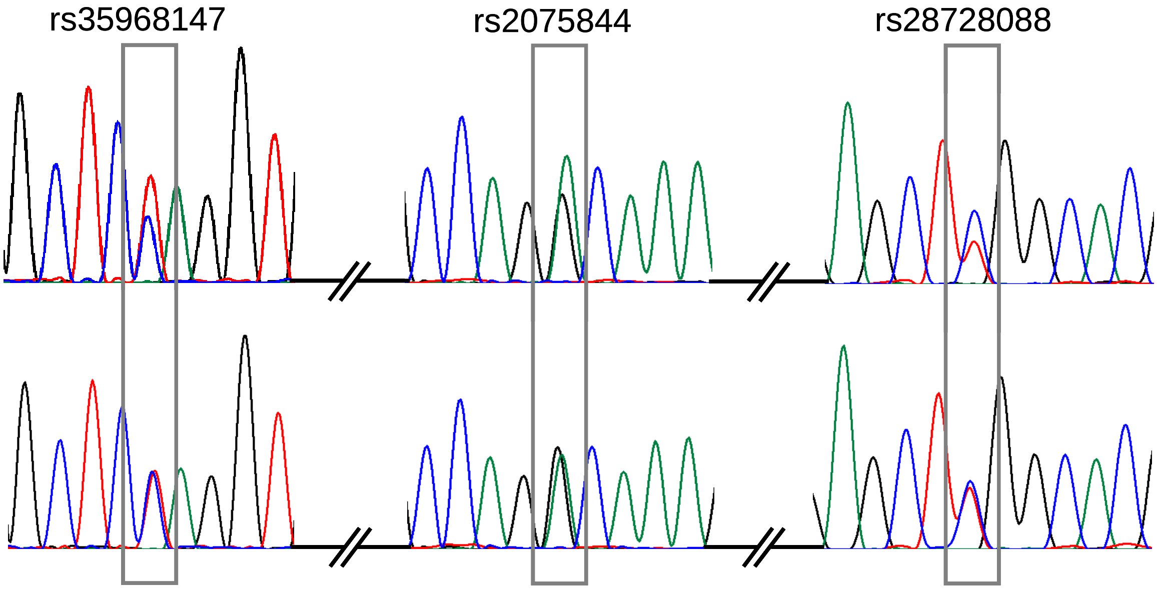


Figure 5.3 Sequence traces for two triple heterozygote individuals. These traces represent three loci within the same fragment spanning from exon 18 to exon 19. Corresponding traces are shown for two individuals, sample H06 in the top row (a) and sample C06 in the bottom row (b), note that both are from the "hypernormal" control sample set. The grey boxes identify SNP loci.

The SNP rs2075844 showed the most exaggerated variation in heterozygote peak patterns and a sample with even peaks and one with very uneven peaks are shown in figure 5.4 as examples. In different individuals the relative peak heights at this locus varied from even (n = 23) to quite uneven (n = 28), and the peak patterns oscillate with respect to which allele peak is the highest; in the case of the 'quite uneven' peaks the A allele peak is higher in 16 samples, and the G allele peak is higher in 12 samples.

If this CNV detection method is viable, one could infer from the peak patterns shown in figure 5.4 that these three individuals have different copy numbers of the sequence region that contains this locus. For instance we might predict that sample AW16 (a) has a duplicate copy of this region containing allele A at the rs2075844 SNP locus and vice versa for sample DS003 (b). Intuitively sample DS033 (c) appears to possess an even number of copies however as explained in the strategy section 1.1.4 all peak heights depend on local sequence context and terminator dye incorporation and therefore even peak patterns may not be indicative of an even number of copies. Thus no direct inference can be made with regard to the numbers of possible copies, and inference of DNA dosage is merely relative.

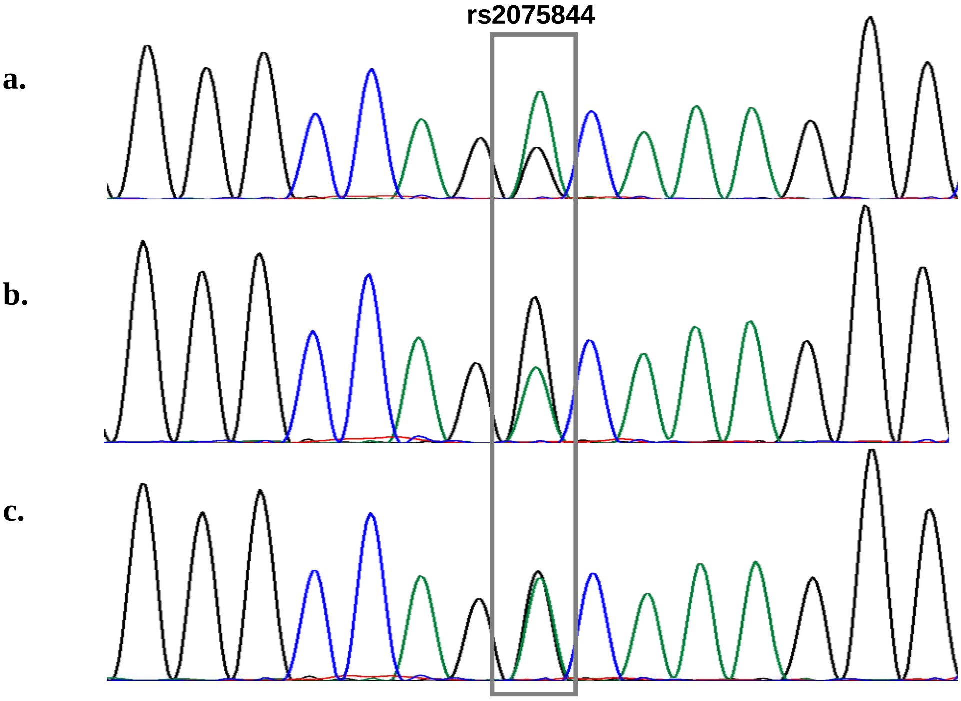


Figure 5.4 Sequence traces of the rs2075844 locus for three heterozygous samples. Samples AW16 (a), DS003 (b), and DS033 (c) are shown here. Note that all three individuals are from the severe asthmatic sample set.

5.3.3 The importance of repeats

In order to inspect the robustness of this CNV detection method, both the amplification and sequencing steps were repeated for the entire severe asthmatic sample set (n = 84), with all reaction conditions kept constant. All genotypes were confirmed and heterozygous peak patterns were compared in only the samples that had very asymmetric or entirely even peak patterns and had clear sequence traces for both the first and second sequence runs. Two SNPs were studied in detail as being ones with the clearest variability. A total of 20 comparisons were made between first and second run sequence traces, of these 8 were discrepant, 6 of which were very discrepant (see table 5.1). These results thus show no evidence of concordance for experiment to experiment results leading to the conclusion that examining peak heights in this region of the *MUC5AC* gene does not give an indication of copy number.

However it should be noted that in discrepant samples in which there was heterozygosity at more than one position, the inconsistencies were true for each polymorphic locus within the same amplified product (see figure 5.5). Thus it might be implied that the observed variance in peak patterns is a product of the PCR amplification step.

Table 5.1 Peak pattern descriptions of rs2075843 and rs2075844 heterozygous samples for two separate sequence runs.

	rs2075843	
Sample	1st run	2nd run
AW22	Very uneven A peak bigger	Very uneven A peak bigger
DS003	Very uneven A peak bigger	Very uneven A peak bigger
DS020	Very uneven G peak bigger	Very uneven G peak bigger
DS033	Very uneven G peak bigger	Very uneven G peak bigger
SA038	Even peaks	Even Peaks
SA137	Very uneven A peak bigger	Even peaks

	rs2075844	
Sample	1st run	2nd run
AW01	Very uneven G bigger	Very uneven A bigger
AW02	Even peaks	Even peaks
AW14	Very uneven G peak bigger	Uneven G peak bigger
AW16	Very uneven A peak bigger	Almost even
AW23	Even peaks	Uneven G bigger
AW27	Very uneven A peak bigger	Very uneven A peak bigger
DS003	Very uneven G peak bigger	Uneven G bigger
DS020	Even peaks	Even peaks
DS030	Even peaks	Even peaks
DS033	Even peaks	Even peaks
SA035	Even peaks	Uneven A bigger
SA043	Even peaks	Very uneven G peak bigger
SA050	Even peaks	Even peaks
SA137	Very uneven G peak bigger	Very uneven G peak bigger

Samples whose peak patterns were very different between sequencing runs have been shaded in blue. Samples whose peak patterns differ but to a lesser degree are shaded in pink.

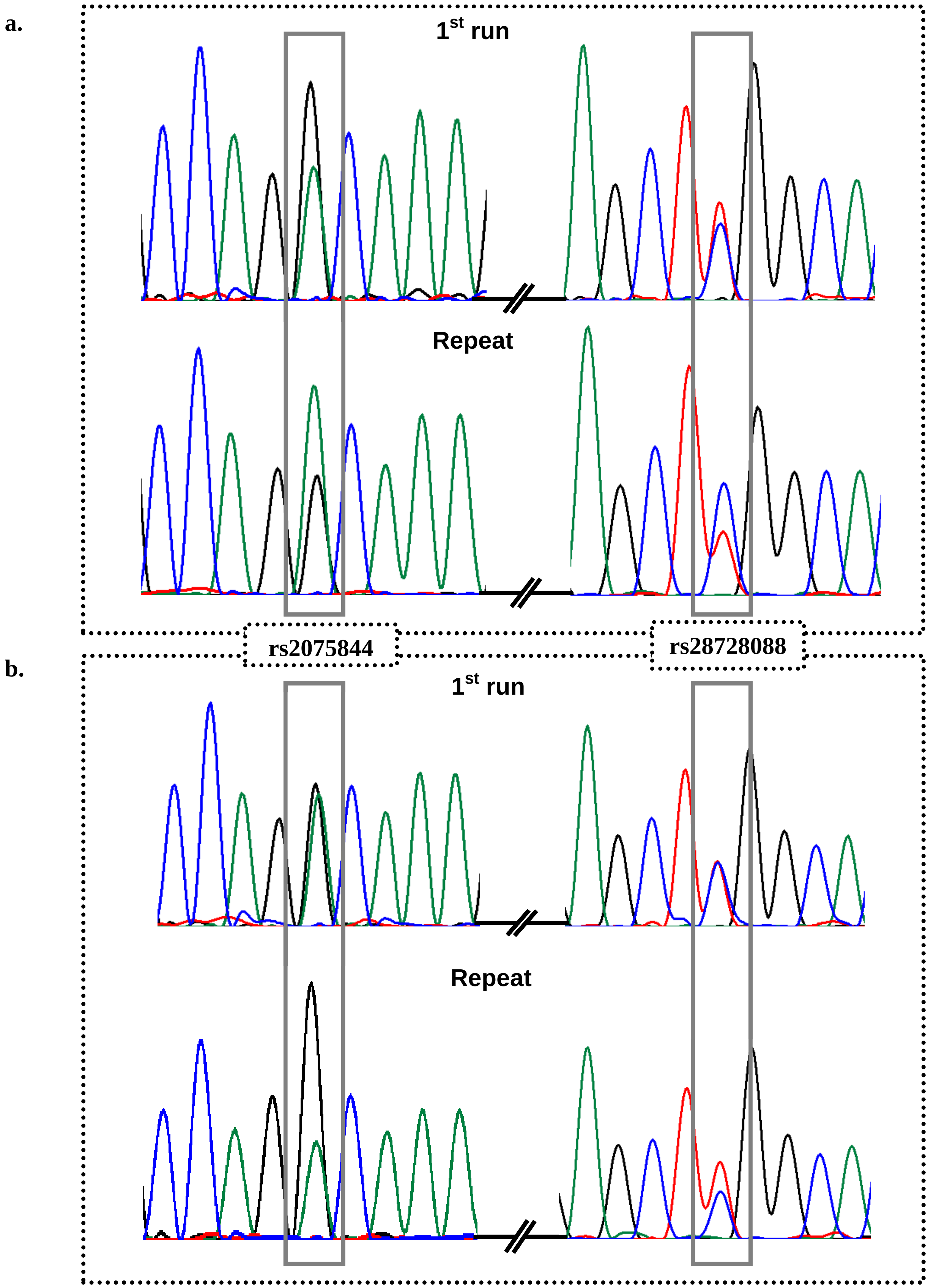


Figure 5.5 Comparison of first run and repeat sequence traces for two samples AW01 (a) and SA043 (b) heterozygous at for rs2075844 and rs28728088.

5.4 Discussion

This results chapter used a sequencing strategy in an attempt to examine whether the *MUC5AC* gene was affected by CNV, however this technique proved to be unreliable since the findings could not be replicated. An independent method of CNV detection, multiple ligation probe assay (MLPA), was used in preliminary experiments as part of a MSc project in the lab. DNA specific probes are ligated to genomic DNA and the ligated DNA/probe complex is then amplified. PCR product quantities are considered to be proportional to the number of copies to which the probe had bound to. *MUC5AC* CNV could not be detected by the MLPA method in an experiment which used 43 samples. However the technique was not well suited to *MUC5AC* since probes could not be designed with the specified levels of GC richness, due to the high GC content of this gene. It should also be noted here that only high quality DNA could be used for MLPA, thus DNA from lymphoblastoid cell lines was used for this experiment, rather than DNA from our disease cohort.

While neither CNV detection method was able to confirm *MUC5AC* CNV, both assays had drawbacks. We hope to follow up this study with other CNV detection techniques in an effort to conclusively confirm or disprove the presence of CNV. At present, blood and sputum samples are being collected from patients with severe asthma. We propose to use the high quality blood DNA for more comprehensive MLPA studies which will include high density probe coverage of the 11p15.5 *MUC* gene complex. Sputum samples will be analysed for different MUC5B isoforms and quantities by our collaborator Karine Rousseau (Manchester Mucin laboratory) to determine whether aberrant mucin isoforms are detected in the same patients.

Unfortunately, *MUC5AC* CNV could neither be confirmed nor disproved by this study, and thus there is no evidence that the deviations from HWE reported in chapter 3, are due to variations in copy number.

The deviations from HWE could be a consequence of chance, however we would be unlikely to see the differences we do between the affecteds and unaffected groups since all samples were randomised and located together on the same plates.

SNPs under primers would result in strand specific PCR or primer extension dropout that might produce spurious results if the SNP is more frequent in either the affected or the unaffected group. Since genetic variants under a primer would result in more homozygous typings, this is unlikely to have been an issue because the greatest deviation in the *MUC5AC* rs1132440 genotype distributions was an increase in the number of heterozygotes in the allergy and hayfever affected groups.

We have no reason to suspect incorrect genotyping of the *MUC5AC* rs1132440 variant in chapter 3 because some samples were also typed by RFLP and all heterozygote examples were confirmed.

If this region of *MUC5AC* is not affected by CNV then we are unsure why the sequencing peak heights are so unstable. From experience and personal correspondence with other researchers, it is unusual for heterozygous peak patterns to shift so dramatically in the opposite directions. The GC richness of the sequence could somehow be causing the heterozygote peak pattern discrepancies. If the degree of allele detection is reliant on the sequence context then this could pose a problem for CNV results in the literature.

Chapter 6

Characterisation of the *MUC5B* promoter in the context of Africa

6 Characterisation of the *MUC5B* promoter in the context of Africa.

As the continent of the origin of modern humans, Africa has the deepest human history and thus harbours much more human genetic diversity than anywhere else in the world. Nevertheless, studies of autosomal genes on Africans remain limited and most SNP discovery projects until very recently have been heavily biased towards non-African populations. Very little mucin research has involved participants of African ancestry and thus the primary aim of this study was to initiate the characterisation of potentially functional variants within the *MUC5B* promoter region in several African populations, with particular emphasis on those of Ethiopian origin.

In the introductory section of this chapter, the history of African populations will be explored in relation to genetic diversity and the recent African origin theory.

6.1 Introduction

6.1.1 The theory of Recent African Origin

Charles Darwin was the first to propose that “our early progenitors lived on the African continent” (Descent of man 1871- Chapter VI - On the Affinities and Genealogy of Man). It is now generally accepted that all non-Africans are descendants of anatomically modern humans originating in Africa. This is known as the recent African origin theory (RAO).

Genetic analysis of contemporary human populations has significantly aided in deciphering the origins of modern humans. Studies involving mitochondrial DNA (mtDNA) and the Y-chromosome have proved to be invaluable for studying the past due to their inability to recombine. Simple phylogenies can be constructed based on sequential mutations. Phylogenetic analysis has identified that the most recent common ancestor of the patrilineal Y chromosome and matrilineal mitochondrial DNA lines, originated in Africa (Cann *et al.* 1987; Underhill *et al.* 2000; Vigilant *et al.* 1991).

The population of modern humans that migrated out of Africa is likely to have been small and thus it is postulated that a bottleneck accompanied the out-of-Africa migratory event. By comparing genetic variation between African and non-African populations, evidence has accumulated in favour of a RAO and subsequent bottleneck. Studies have shown that African populations have greater levels of allelic and haplotypic diversity and more private allelic variants than non-African populations, which only contain a proportion of the African diversity (Campbell and Tishkoff 2008; Tishkoff *et al.* 1996; Tishkoff *et al.* 1998; Tishkoff *et al.* 2000). This extensive African diversity is mirrored by linguistic diversity: approximately 2000 ethnolinguistic groups have been identified within the African continent (Campbell and Tishkoff 2008).

Studies of linkage disequilibrium (LD) have also been used to reflect the age and demographic history of African and non-African populations. Older populations show a greater breakdown in LD, whereas relatively ‘young’ populations or populations, that have undergone a recent bottleneck event, will have increased levels of LD. It has been shown that the extent of LD increases with increasing distance from East Africa (Jakobsson *et al.* 2008), consistent with the notion that East Africans are of the most ancient of all human populations.

6.1.2 The importance of East Africa

While it is generally accepted by the scientific community that anatomically modern humans originated in Africa, the anticipated migratory routes out-of-Africa remain ambiguous. One proposed route suggests that modern humans first left from Ethiopia and migrated to East Asia and Oceania via the Bab-el-Mandeb of the Red sea. Several genetic studies support a migratory route originating from East Africa. It has been shown that a mtDNA haplogroup originally thought to be an ancient East Asian marker, actually originated in East Africa (Quintana-Murci *et al.* 1999) and is at high frequencies in contemporary Ethiopian populations.

The most ancestral Y-chromosome has been identified in Ethiopians, East African Sudanese and the South African Khoisan (Tishkoff and Williams 2002). It is thought that the Khoisan may have once inhabited Eastern Africa (Passarino *et al.* 1998). If this is the case then Y-chromosomal data certainly points to an East African patrilineal origin.

6.1.3 Back migrations into Ethiopia

Back migrations have also played a prominent role in the history of Ethiopia which is in fact thought to have been the first state developed in sub-Saharan Africa (Cavalli-Sforza *et al.* 1994). Past migrations into Ethiopia have caused the Ethiopian gene pool to become a mixture of non-African as well as African genes (Cavalli-Sforza *et al.* 1994). Past major external influences into Ethiopia have come from the Sudanese and Semitic peoples. Levine states that the Sudanese influence occurred in two waves, the first in approximately the third millennium B.C and the second in the first millennium. It is this second wave that is thought to have greatly influenced the Western Ethiopian Anuak ethnic group (Levine 2000), who will discussed later in this introductory section.

Just as the Bab-el-Mandeb is thought to have provided the route out-of-Africa for the first Modern Humans, it is also thought to have been the route by which the peoples of Arabia back migrated into Ethiopia. Levine suggests that a small continuous wave of Arabian influence has occurred in Ethiopia for the past 3 millennia involving Jews, Syrian Christians and Arabian Muslims (Levine 2000).

6.1.4 Ethiopian ethnic groups

Eighty-five different ethnic groups reside in Ethiopia (Ethiopian census 2007 – see web citations on page 206), and although admixture has occurred, it is still possible to identified distinct populations due to linguistic, cultural and geographic disparities. In this project five of these groups have been analysed; the Afar, Amhara, Anuak, Maale and Oromo.

6.1.4.1 Afar

The Afar people are herders (Murdock 1959) and constitute 1.7% of the overall Ethiopian population and 90% of this ethnic group inhabit the Afar region (see figure 6.1) (Ethiopian census 2007 – see web citations on page 206). Many of the Afar live in the Danakil depression (Murdock 1959) which is the lowest point in Africa (CIA world factbook – see web citations on page 206) and is recorded as having the world's highest average temperature (NASA – see web citations on page 206). The relative humidity of the depression is 70% (Leroux 2010), a high value considering that rainfall within this region is very scarce (Leroux 2010).

As can be seen in figure 6.2 Afar is a Cushitic language which branches from the Afro-Asiatic family (Ethnolgue – see web citations on page 206). Cushitic languages are spoken by farmers and pastoralist herders in the majority of Ethiopia and are also spoken in the republic of Sudan (Cavalli-Sforza *et al.* 1994), and in other parts of the horn of Africa.

6.1.4.2 The Amhara

The Amhara ethnic group live mainly in the highlands of the Amhara region (see figure 6.1) where they mainly subsist on agriculture, producing crops like cotton, different cereals and oil plants. The Amhara constitute a high percentage of the Ethiopian population (26.9%) (Ethiopian census 2007 – see web citations on page 206). As can be seen in figure 6.2, Amharic is a Semitic language which branches from the Afro-Asiatic language family (Ethnologue – see web citations on page 206). Semitic languages are spoken in Northern Ethiopia, in all of Arabia and the middle East (Cavalli-Sforza *et al.* 1994).

6.1.4.3 The Anuak

The Anuak mostly live in settlements on the river banks of the Baro river in the Gambella region (13suns - see web citations on page 206) (see figure 6.1) and make up only a small percentage of the overall Ethiopian population, 0.12% (Ethiopian census 2007 – see web citations on page 206). During the rainy season the Anuak subsist on agriculture, producing crops such as cotton, sorghum and various fruit and vegetables. Fishing is prominent during the dry season and hunting is also practised (Murdock 1959; Woube 1998) 13suns – see web citations on page 206). During the rainy season, humidity levels are between 72-78% (Woube 1998). As can be seen in figure 6.2 Anuak is a Nilotic language which comes from the Nilo-Saharan family (Ethnologue – see web citations on page 206).

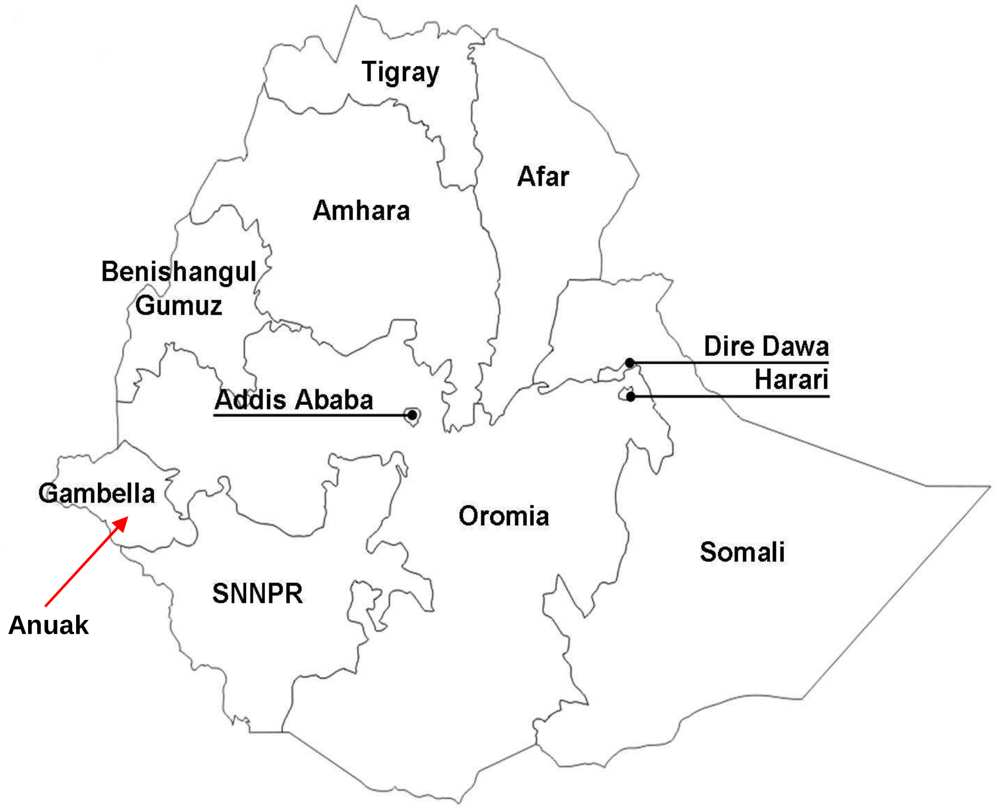


Figure 6.1 Map of Ethiopia's administrative regions. The region/state boundaries displayed on this map are based on those defined by the United Nations emergencies unit for Ethiopia and the United Nations Office for Coordination of Humanitarian Affairs (UN OCHA-Ethiopia). http://www.africa.upenn.edu/eue_web/eue_mnu.htm and http://www.ocha-eth.org/Maps/Maps.htm. Note that SNNPR refers to Southern Nations Nationalities and Peoples' Regional States, and that Addis Ababa and Dire Dawa are chartered cities.

6.1.4.4 The Maale

The Maale (also spelt Male, Marle, Malle) are a minority ethnic group making up only 0.13% of the Ethiopian population, and mostly reside within the Southern Nations Nationalities and Peoples' Regional States (SNNPR) (see figure 6.1) (census 2007). As can be seen in figure 6.2 Maale is an Omotic language which branches from the Afro-Asiatic language family (Ethnologue – see web citations on page 206). This language is spoken in only a small region of Ethiopia (Cavalli-Sforza *et al.* 1994).

6.1.4.5 The Oromo

The Oromo people, previously called the Galla, are pastoralists (Murdock 1959) and mostly inhabit the highlands of the Oromia region (see figure 6.1) and are the largest ethnic group in Ethiopia, making up 34.5% of the population of the country (Ethiopian census 2007 – see web citations on page 206). As can be seen in figure 6.2 Oromo is a Cushitic language which branches from the Afro-Asiatic language family (Ethnologue – see web citations on page 206).

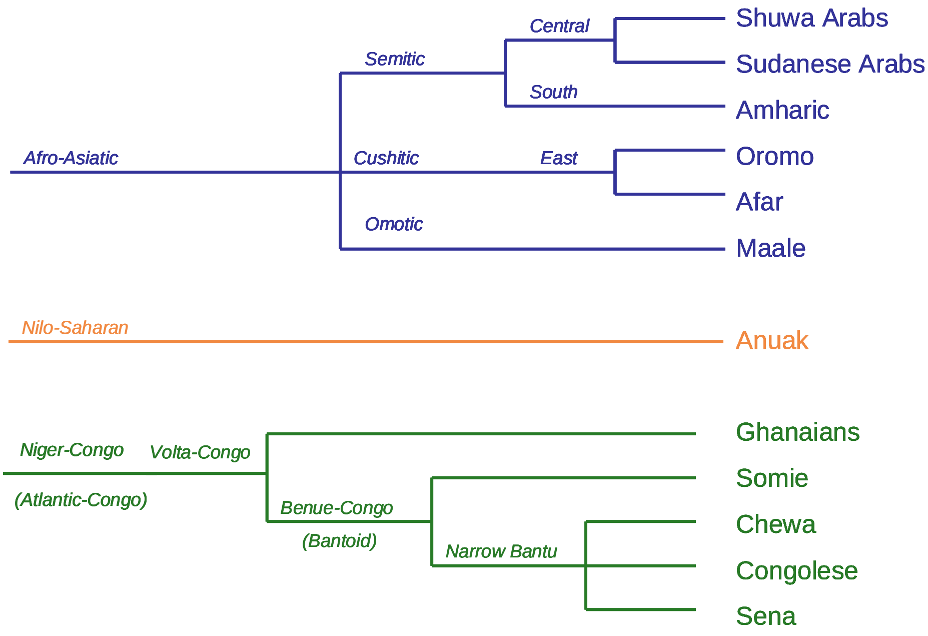


Figure 6.2 African language tree. Language tree of a subset of African language families, extended to show only languages that correspond specifically to this project. Figure kindly provided by Bryony Jones, PhD student, GEE.

6.1.5 Other African populations considered in this project

Seven other African populations have been studied within this project and are defined here by the country/place in which they were collected; Congo-Brazzaville, Ghana, Cameroon Grassfields-Somie, Cameroon Lake Chad, Malawi, Mozambique and Sudan.

The sample collection was made by The Centre for Genetic Anthropology (TCGA), to represent countries influenced by the Bantu expansion. Some 58% are speakers of a Bantoid language.

Although the Malawi sample is entirely composed of an ethnic group called the Chewa (n = 50), the remaining populations are composed of mixed ethnicities. Because of this, the groups have been divided by country. However the Cameroon sample is divided into two geographic locations, Grassfields-Somie and Lake Chad, and it should be noted that these two sample sets are ethnically and linguistically very different. Of the Somie sample 59 out of 63 speak a Bantoid language, whereas no Bantoid derived languages are spoken by the Lake Chad individuals.

6.2 Hypothesis and Aims

The main aim of this study is to characterise *MUC5B* promoter variation in the context of African populations. The initial aim is to identify any novel variants and to investigate their potential to alter function, for instance whether they are located within a TFBS or within a region highly conserved between species. We also intend to explore the overall pattern of upstream *MUC5B* variation with respect to species conservation and measures of selection.

The final aim of this chapter is to study *MUC5B* promoter variation between geographically distinct populations. Since the expression of MUC5B is orchestrated by environmental cues, we hypothesise that the past environments inhabited by populations are likely to have specifically shaped the *MUC5B* promoter variation of the population due to environmental selective pressures. One could predict that frequencies of the *MUC5B* promoter haplotype variants are likely to reflect environmental pressures since they have been shown to confer different expression levels (Kamio *et al* 2005; Loh *et al* 2010).

6.3 Results

In this results section, genetic data will be reported for the *MUC5B* proximal promoter (as described in chapter 4) and analysed in the context of Africa. The *MUC5B* promoter region spanning from -1097 to +97 nucleotides relative to the transcription start site was sequenced in a total of 748 African individuals from eight main geographic populations; Ethiopia, Congo, Ghana, Cameroon Grassfields-Somie, Cameroon Lake Chad, Malawi, Mozambique, and Sudan. The Ethiopian sample was subdivided into five ethnic groups, and studied in detail.

6.3.1 Allelic variants

A total of 15 allelic variants were identified in the African sample as a whole, eight of which had been previously described. Details of these variants, including minor allele frequencies, are shown in table 6.1 and are annotated in the sequence shown in figure 6.4. It should be noted that population specific genotype distributions for all SNPs are in accordance with those predicted under HWE.

14 out of 15 of the variations are present in the Ethiopians; eight and six of these were identified in the Mozambique and Cameroon Lake Chad sample sets respectively and only seven could be identified in the Congo-Brazzaville, Ghana, Cameroon Grassfields-Somie, Malawi and Sudan samples.

All previously reported SNPs (rs885455, rs885454, rs17235353, rs7115457, rs7118568, rs56235854, rs2735738) were present in both African and non-African (European and/or Japanese) populations, except for rs56366237 at position -560, which is African specific.

Seven allelic variants identified in this study have not been previously reported and have therefore been termed ‘novel’ variants. Six out of seven (-988, -946, -614, -420, -221, -89) are private to Ethiopia, and one out of the seven (-194) has been identified only in the Congo, Ghana, Malawi, Mozambique and Sudan samples. The -194 novel appears to be a Niger-Congo specific allele as all carriers of the derived allele (except for one individual) speak languages that belong to the Niger-Congo family. All novel variants are however at low frequencies.

Examination of all novel SNPs in relation to transcription factor binding sites, shows that variant -194 falls within the putative Sp1 binding site located at -196 to -185 (see figure 6.3).

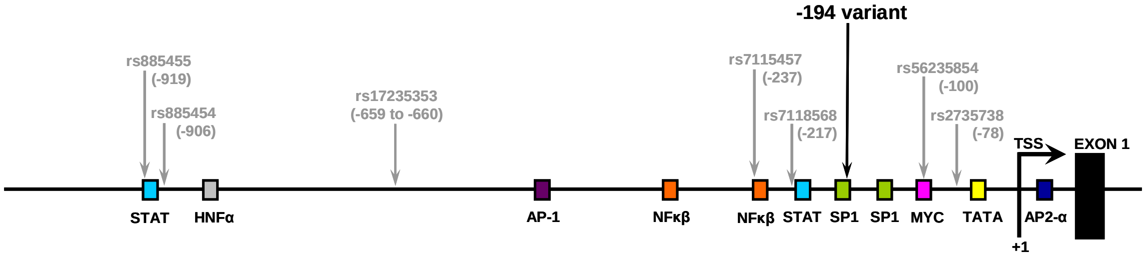


Figure 6.3 Diagrammatic representation of the *MUC5B* proximal promoter with annotated transcription factor binding sites (TFBS) and identified allele variants. The TFBS’s denoted in this figure have either been predicted by rVista during this project or have been previously identified in the literature (Chen *et al.* 2001; Wu *et al.* 2007a). The human *MUC5B* promoter sequence and the homologous mouse Muc5b promoter sequence, were first aligned by zPicture and the generated alignment was then used as an input file for rVista.

Table 6.1 *MUC5B* proximal promoter African allelic variants.

			Position relative to transcription start site and rs identifier														
Country	Ethnic group	Sample number	-988 (G)	-946 (A)	-919 (G)	-906 (A)	-659 to -660 (D)	-614 (T)	-560 (G)	-420 (A)	-237 (A)	-221 (T)	-217 (G)	-194 (T)	-100 (A)	-89 (A)	-78 (C)
			•••	•••	rs885455	rs885454	rs17235353	•••	rs56366237	•••	rs7115457	•••	rs7118568	•••	rs56235854	•••	rs2735738
Ethiopia		380	0.013	0.001	0.238	0.038	0.021	0.011	0.017	0.011	0.125	0.011	0.139	•••	0.030	0.007	0.332
	Afar	76	0.027	•••	0.187	0.067	0.013	0.020	0.007	0.020	0.073	0.007	0.087	•••	0.013	•••	0.293
	Amhara	76	0.020	•••	0.184	0.033	0.020	0.020	0.007	0.020	0.092	0.007	0.099	•••	0.059	•••	0.250
	Anuak	76	•••	•••	0.309	0.007	•••	•••	0.066	•••	0.250	•••	0.296	•••	•••	0.033	0.487
	Malle	76	•••	•••	0.276	0.033	0.026	0.013	•••	0.013	0.105	0.040	0.105	•••	0.026	•••	0.349
	Oromo	76	0.020	0.007	0.237	0.053	0.046	•••	0.007	•••	0.105	•••	0.112	•••	0.053	•••	0.283
Congo-Brazzaville		52	•••	•••	0.423	0.115	•••	•••	0.067	•••	0.212	•••	0.225	0.048	•••	•••	0.373
Ghana		56	•••	•••	0.375	0.116	•••	•••	0.080	•••	0.140	•••	0.184	0.009	•••	•••	0.325
Cameroon Grassfields-Somie		65	•••	•••	0.344	0.109	0.008	•••	0.065	•••	0.177	•••	0.223	•••	•••	•••	0.354
Cameroon Lake Chad		63	•••	•••	0.198	0.079	•••	•••	0.016	•••	0.083	•••	0.117	•••	•••	•••	0.331
Malawi		49	•••	•••	0.439	0.133	•••	•••	0.122	•••	0.117	•••	0.170	0.010	•••	•••	0.298
Mozambique		51	•••	•••	0.370	0.110	0.010	•••	0.153	•••	0.157	•••	0.225	0.020	•••	•••	0.392
Sudan		29	•••	•••	0.328	0.069	•••	•••	0.069	•••	0.190	•••	0.224	0.017	•••	•••	0.414

Table of population specific minor allele frequencies (MAF) for all allelic variants identified in this project. ••• denotes a MAF of zero

```
-1097
CCACGGAGCATTCAGGACGCTGGTGACCAGGGAGCCAGGAGGTGGGAGCATCTGAGGTGCA
                                                -988
GGTCACACGGGCAGGAGGTGTTTGCAAGAGGTATTGCAGCGCGGACGGRGTGTCCTGCAGA
                             -946                       -919
TGACGCTGTCTGTCCTGTAGATGACGCTCRTCAAGGAGGTTTACCACATAGCCCCCRGGAA
        -906                                         rs885455
GCCCACCCRACACCAGCCGGAGGTGCTAGGCTTCTGCGGCTCCCACCTGGGGCAGGCGGAG
        rs885454
GACCCCGGGCAGGTCCAGGACCCCCCGGAGCAGCTGCTTCCTCAACCCTGCCAGGGTTAAT

GAGGAGGCCCCAGAGTGAGGTGGAGGCCAAATGGACTCAGGGCCGGAGCCTCTGGCCTGG

CTGGATCAGGGCTGGCATTGGACAAGCGCAGCTGACTCCCGATGTGCATGGCCAGGAGACA
          -659/660                                      -614
CTCTGGGCCTCAGTTTCCCCTTGAATGTGAACCTTGAAACAGATCAGCCCAGAGACYTCCC
          rs17235353                                 -560
ACGGTCTTCAAGGGGCTCTGGTCAGCTGGGCTGGGGTCTCTGGAAATAGRGCCTCCTCCAG
                                                 rs56366237
GGACCCCCACAAGCCACCCAGACTGAGCATCCTGGCCATGTGCATGCCTGAGCTCAGCAGG

AGCCTCCCGGcCTCCCCGTGGGCTAAGCAGTGGTGGGAGGGGAGCTCCAGCCTCGTGGGCC
      -420
CTCCCCRGGCCTCGGGGACCCATGGTCAGTGGCTGGGGGTGCTGCCCAGAGGCTGGGATTC

CCTTCCAGCAGGAGCCGCAGTGGGGCTGAGTGTGAGGCAGGCTGGCTGACCACTGTTTCCA

TGGACCCTGCGTCCAAGGCCAGCCCTGCCTTCCAGCGGCTTTGCCATCTAGGACGGGTGCC
      -237           -221 -217                   -194
AGGTGGRGTAGGCCCTTCTCTCYCTTSCGATTCTCAGAAGCTGCTGGGGKTGGGGGCGTCC
      rs7115457           rs7118568
TGGGCCTCAGGGCACAGAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCC
                     -100       -89        -78   -72
CCTCCCCGAGAGCAAACACACRTGGCTGGAGCRGGGAAGAGCAYGGTGCYCTGCGTGGCCT
                     rs56235854            rs2735738        ↱
GGCCTGGCTTGGGGCCAAGGCTCCCTGCTACATAAGCTGGGGCCCCCAGGGGAGCAAGCAC
                                                           +1
CCGGCCCGGCTCCCTCCCTGCCCGTCCCCGTCCCCCCACCCGTGCCAGCCCCCAGGATGGG
                                 +97
TGCCCCGAGCGCGTGCCGGACGCTGGTGTTGGCTC
```

Figure 6.4 ***MUC5B*** **proximal promoter sequence annotated with African alleles.** Grey highlight indicates forward and reverse primers for the fragment A. Pink highlight indicates forward and reverse primers for fragment B. All previously identified SNPs highlighted in green. All novel SNPs identified by this project are highlighted in yellow. Transcription start site is highlighted in red and is described as position +1. The TATA box is highlighted in turquoise and the start of translation codon (ATG) is underlined and in bold.

6.3.2 *MUC5B* promoter haplotype

Haplotypes were inferred for all previously reported SNPs with a certainty of 0.9 or greater using PHASE. All novel rare alleles were manually allocated to haplotypic backgrounds by identifying individuals who were both homozygous for a particular haplotype and carriers of the rare novel allele. It was thereafter assumed that the novel rare allele always fell on that particular haplotypic background since it is unlikely that the mutation will have occurred more than once. The previously identified *MUC5B* promoter haplotypes and their identifiers (Kamio *et al.* 2005) were used as the basis for naming the African haplotypes. Additional haplotypes seen in the African populations that had not been seen in the Japanese population (Kamio *et al.* 2005) were given new identifiers.

A total of 23 *MUC5B* promoter haplotypes were identified in the African sample as a whole, of which 20 were found within the Ethiopian population and between 8 and ten could be identified in each of the other African sample sets (see table 6.2). It should be remembered that only 6 *MUC5B* promoter haplotypes were identified in the healthy European controls typed in chapter 2. Table 6.2 describes all haplotypes and includes haplotype frequencies for each African population. Many haplotypes are however rare and therefore the population specific pie charts shown in figure 6.6 represent only the percentages of the three major haplotypes (H3, H1 and H8) plus a category described as ‘other’ which contains all rarer haplotypes.

The phylogenetic tree shown in figure 6.5 was created in order to display the relationships between haplotypes (Network 4.516). On examination of the tree it is apparent that several of the mutations have to have occurred more than once (labelled with stars), and since this is rather unlikely it suggests that several of the haplotypes have occurred by recombination.

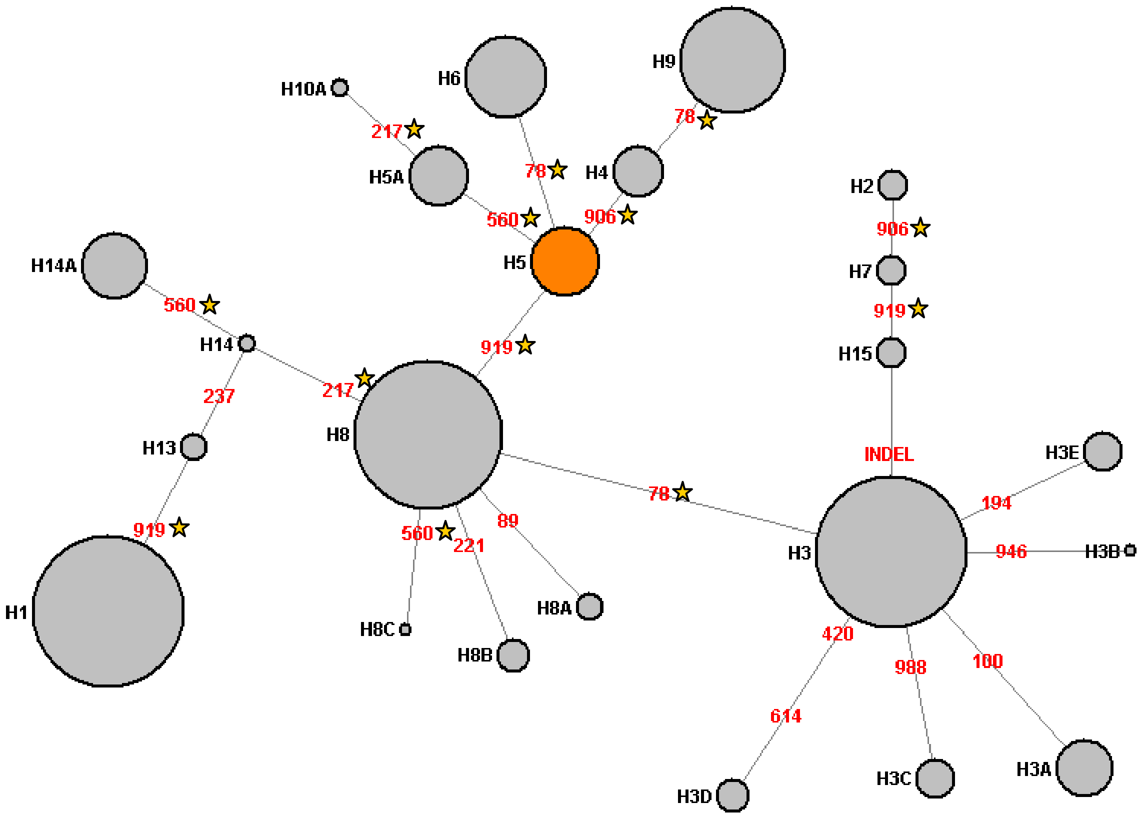


Figure 6.5 ***MUC5B*** **promoter haplotype network.** All African haplotypes have been included in this network. The circumference of each node is relative to the haplotype frequency. The orange node identifies the chimpanzee haplotype. Haplotype identifiers are indicated in black font, while red font signifies the mutation. Mutation identifiers are named with respect to their location relative to the transcription start site. Mutations that are shown to occur more than once on the tree are labelled with stars. Network 4.516 was used to construct the network.

Table 6.2 Sequence details and frequencies of the *MUC5B* proximal promoter haplotypes identified in eight African sample sets.

Haplotype ID	-988 novel	-946 novel	rs885455	rs885454	rs17235353	-614 novel	rs56366237	-420 novel	rs7115457	-221 novel	rs7118568	-194 novel	rs56235854	-89 novel	rs2735738	Ethiopia	Congo-Brazzaville	Ghana	Cameroon Grassfields-Somie	Cameroon Lake Chad	Malawi	Mozambique	Sudan
H5	a	g	G	G	I	c	A	g	G	c	C	g	G	g	C	0.035	0.010	0.010	•••	•••	•••	0.034	0.038
H5a	a	g	G	G	I	c	G	g	G	c	C	g	G	g	C	0.004	0.042	0.030	0.016	•••	0.049	0.068	0.038
H4	a	g	G	A	I	c	A	g	G	c	C	g	G	g	C	0.024	•••	•••	•••	•••	•••	•••	0.019
H10a	a	g	G	G	I	c	G	g	G	c	G	g	G	g	C	•••	•••	•••	0.016	•••	•••	•••	•••
H1	a	g	G	G	I	c	A	g	A	c	G	g	G	g	C	0.124	0.208	0.160	0.186	0.073	0.122	0.148	0.173
H3	a	g	A	G	I	c	A	g	G	c	C	g	G	g	T	0.551	0.448	0.510	0.500	0.564	0.488	0.489	0.538
H3a	a	g	A	G	I	c	A	g	G	c	C	g	A	g	T	0.031	•••	•••	•••	•••	•••	•••	•••
H3b	a	a	A	G	I	c	A	g	G	c	C	g	G	g	T	0.001	•••	•••	•••	•••	•••	•••	•••
H3c	g	g	A	G	I	c	A	g	G	c	C	g	G	g	T	0.013	•••	•••	•••	•••	•••	•••	•••
H3d	a	g	A	G	I	t	A	a	G	c	C	g	G	g	T	0.011	•••	•••	•••	•••	•••	•••	•••
H3e	a	g	A	G	I	c	A	g	G	c	C	t	G	g	T	•••	0.052	0.010	•••	•••	0.012	0.023	0.019
H15	a	g	A	G	D	c	A	g	G	c	C	g	G	g	T	0.005	•••	•••	0.008	•••	•••	0.011	•••
H8	a	g	A	G	I	c	A	g	G	c	C	g	G	g	C	0.114	0.072	0.090	0.113	0.200	0.049	0.045	0.077
H8a	a	g	A	G	I	c	A	g	G	c	C	g	G	a	C	0.007	•••	•••	•••	•••	•••	•••	•••
H8b	a	g	A	G	I	c	A	g	G	t	C	g	G	g	C	0.011	•••	•••	•••	•••	•••	•••	•••
H8c	a	g	A	G	I	c	G	g	G	c	C	g	G	g	C	•••	•••	•••	•••	•••	0.012	•••	•••
H14	a	g	A	G	I	c	A	g	G	c	G	g	G	g	C	0.001	•••	•••	•••	0.009	•••	•••	•••
H14a	a	g	A	G	I	c	G	g	G	c	G	g	G	g	C	0.013	0.010	0.030	0.032	0.018	0.037	0.068	0.038
H13	a	g	A	G	I	c	A	g	A	c	G	g	G	g	C	0.003	0.010	•••	•••	0.018	•••	•••	•••
H7	a	g	G	G	D	c	A	g	G	c	C	g	G	g	T	0.005	•••	•••	•••	•••	•••	•••	•••
H6	a	g	G	G	I	c	A	g	G	c	C	g	G	g	T	0.035	0.021	0.030	0.016	0.027	0.085	•••	0.019
H2	a	g	G	A	D	c	A	g	G	c	C	g	G	g	T	0.008	•••	•••	•••	•••	•••	•••	•••
H9	a	g	G	A	I	c	A	g	G	c	C	g	G	g	T	0.005	0.125	0.130	0.113	0.091	0.146	0.114	0.038

Blue shading represents the ancestral/chimpanzee H5 haplotype. Pink shading represents the most frequent haplotype H3. White cells signify single step mutations. ••• represent frequencies of zero. Alleles in lower case represent those variants discovered for the first time by this project.

The shading in table 6.2 illustrates the possible relationships between the haplotypes. The blue shading signifies the ancestral (chimpanzee) haplotype H5 or haplotypes relating to H5, and pink shading denotes the most frequent haplotype H3 or haplotypes relating to H3. Cells left unfilled represent single step mutations. In table 6.2 the haplotypes shaded with one colour and interspersed with single unfilled cells, can be accounted for by single step mutations on the background of pre-existing haplotypes, whereas haplotypes containing blocks with 2 colours can be accounted for by recombination events. For instance H3a, H3b, H3c, H3e and H15 can all be accounted for by single step mutations occurring on the H3 haplotypic background, and haplotypes H8 and H6 appear to be the direct recombinants between H3 and H5 but it is not possible to accurately describe the recombination blocks.

As can be seen in the pie charts of figure 6.6 H3 has the highest frequency in all African populations as is the case in European samples (Kamio *et al.* 2005)(chapter 4). The percentage of the high expressing haplotype H1 does not appear to vary much between most of the African populations but does seem to be at a lower frequency in the Cameroon Lake Chad sample set. As will be seen in the next section however, there are differences between ethnic groups within the Ethiopian sample. H8 seems to be at an increased frequency in western Africa, where the eastern and south eastern populations, Ethiopia, Malawi and Mozambique have more rare haplotypes, exhibiting greater haplotypic diversity.

The tabulation of haplotype counts amongst the eight African populations shown in figure 6.7b has been used to create the correspondence analysis 2-dimensional map shown in figure 6.7a. correspondence analysis has been used in this instance to represent the interrelationships between single haplotype distributions across populations and the haplotype composition of each population. The correspondence analysis map first and second dimensions capture 59.81% and 15.04% of the inertia respectively. The most noteworthy observation of this correspondence analysis is that the first axis clearly separates Ethiopia from any other population, largely attributable to the existence of more derived haplotypes due to one step mutations from old haplotypes.

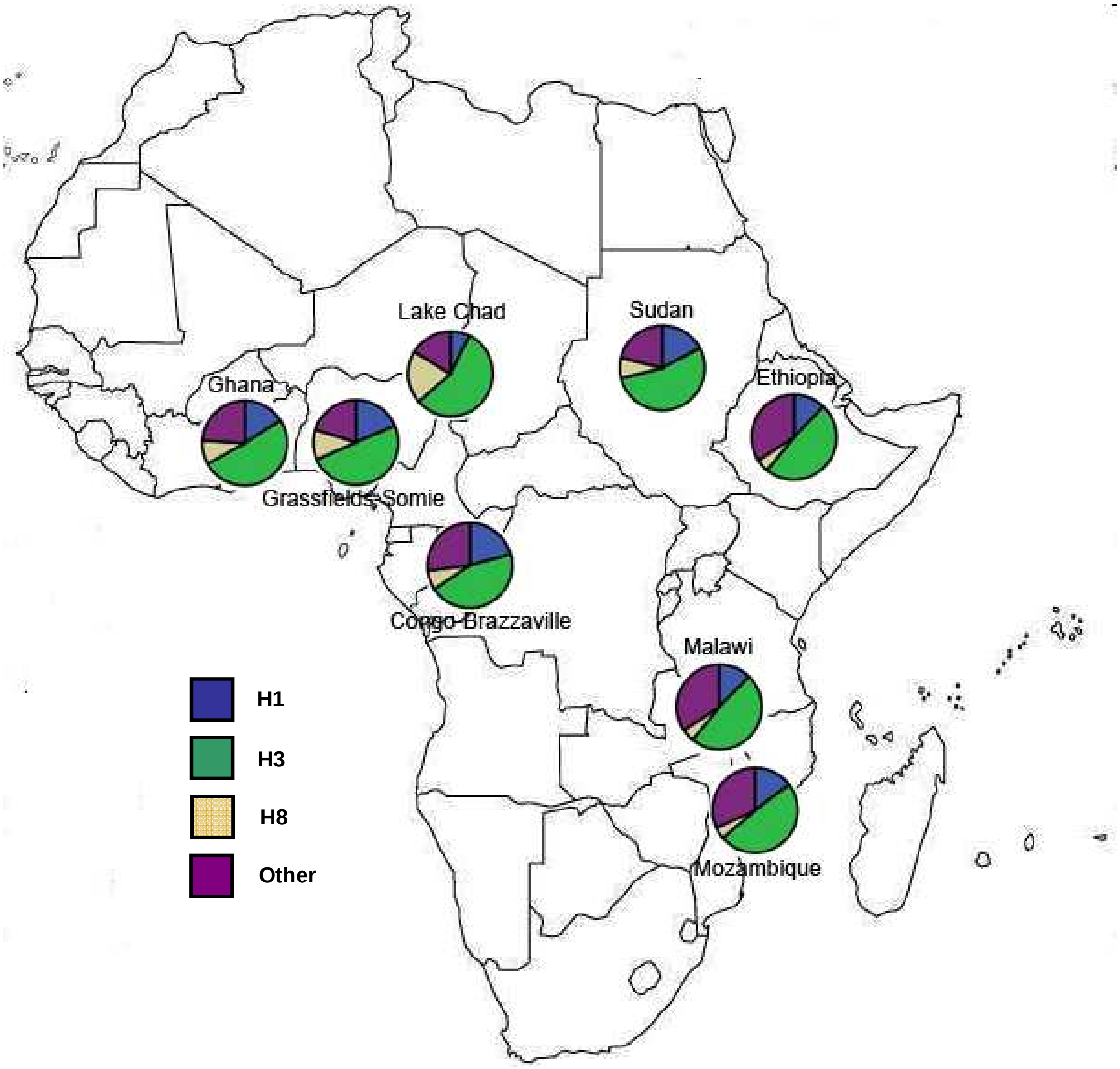
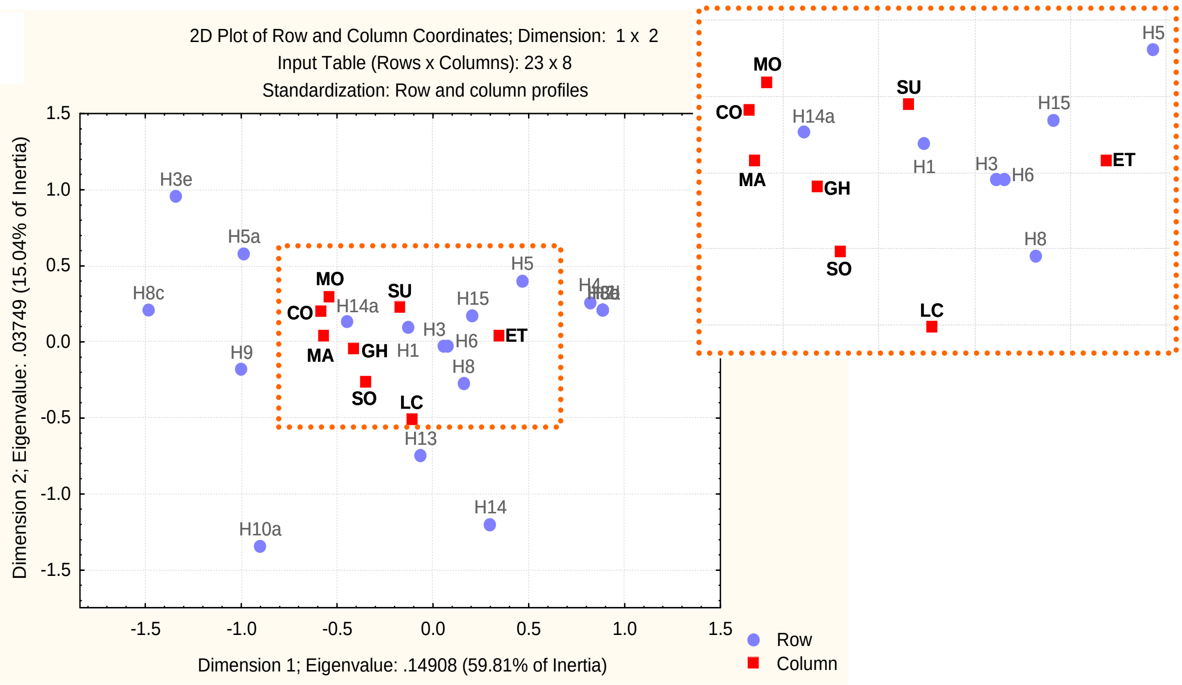


Figure 6.6 Pie charts to depict haplotype percentages for each African population. Each pie chart represents the three major haplotypes H3, H1 and H8, plus a group containing all 'other' haplotypes.

a.



b.

	CO	GH	SO	LC	MA	MO	SU	ET
Haplotype H1	20	16	23	8	10	13	9	93
H2	0	0	0	0	0	0	0	6
H3	43	51	62	62	40	43	28	414
H3a	0	0	0	0	0	0	0	23
H3b	0	0	0	0	0	0	0	1
H3c	0	0	0	0	0	0	0	10
H3d	0	0	0	0	0	0	0	8
H3e	5	1	0	0	1	2	1	0
H4	0	0	0	0	0	0	1	18
H5	1	1	0	0	0	3	2	26
H5a	4	3	2	0	4	6	2	3
H6	2	3	2	3	7	0	1	26
H7	0	0	0	0	0	0	0	4
H8	7	9	14	22	4	4	4	86
H8a	0	0	0	0	0	0	0	5
H8b	0	0	0	0	0	0	0	8
H8c	0	0	0	0	1	0	0	0
H9	12	13	14	10	12	10	2	4
H10a	0	0	2	0	0	0	0	0
H13	1	0	0	2	0	0	0	2
H14	0	0	0	1	0	0	0	1
H14a	1	3	4	2	3	6	2	10
H15	0	0	1	0	0	1	0	4

Figure 6.7 a. Correspondence analysis 2D map representing the interrelationships between the African populations and the *MUC5B* promoter haplotypes shown in the contingency table (b). CO Congo, GH Ghana, SO Grassfields-Somie, LC Lake Chad, MA Malawi, MO Mozambique, SU Sudan, ET Ethiopia. Blue filled circles represent the row coordinates and red filled squares represent the column coordinates. Correspondence analysis performed by Statistica version 9.

6.3.3 Ethiopia in detail

In this section genetic differences between the five Ethiopian ethnic groups will be examined by analysing further the *MUC5B* proximal promoter data described above.

6.3.3.1 Haplotype distributions

In order to explore genetic differences between the five Ethiopian ethnicities, the *MUC5B* promoter haplotype distributions for each ethnic group (see table 6.3) were firstly examined by the exact test of population differentiation implemented by Arlequin (see materials and methods section for details). Statistically significant differences in haplotype frequencies were identified between the Anuak and all other ethnic groups with p values ranging from 0.01 to 0.00001.

The vast difference between the Anuak and all other ethnic groups can be seen in the correspondence analysis shown in figure 6.9. In this two-dimensional map the first and second dimensions capture 56.82% and 20.72% of the total inertia. It is the first axis that clearly separates the Anuak from all other ethnicities.

Because of the previous association identified between asthma and the H1 haplotype, the percentage of H1 haplotypes versus all other haplotypes were compared between the ethnic groups. A statistically significant Pearsons chi squared p value < 0.0001 was obtained and from figure 6.8 it is evident that the Anuak group have a significantly greater percentage of the H1 haplotype than the other ethnic group.

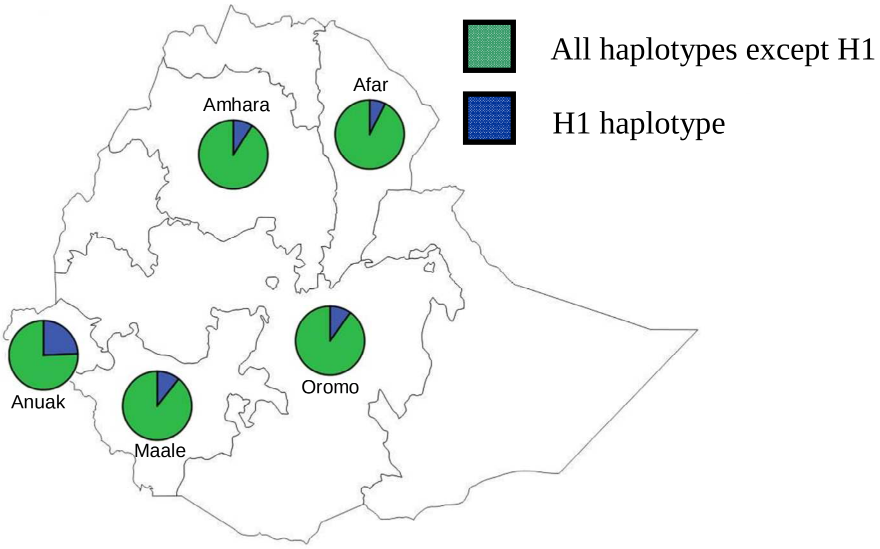


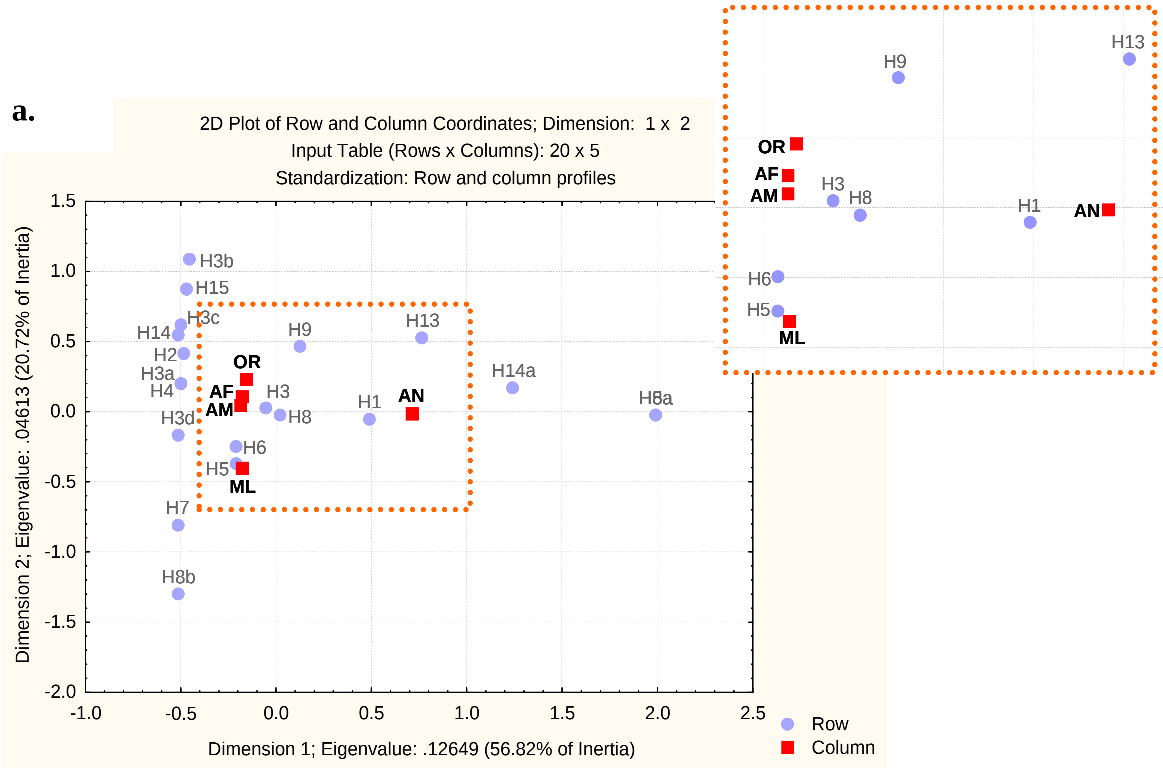
Figure 6.8 Pie charts to show H1 haplotype percentage versus all other haplotypes for five Ethiopian ethnicities.

Table 6.3 Sequence details and frequencies of the *MUC5B* proximal promoter haplotypes identified in the five Ethiopian ethnic groups.

Hap ID	-988 novel	-946 novel	rs885455	rs885454	rs17235353	-614 novel	-560 novel	-420 novel	rs7115457	-221 novel	rs7118568	rs56235854	-89 novel	rs2735738	Afar	Amharic	Anuak	Maale	Oromo	Total Ethiopia
H1	a	g	G	G	I	c	a	g	A	c	G	G	g	C	0.074	0.092	0.243	0.107	0.100	0.124
H2	a	g	G	A	D	c	a	g	G	c	C	G	g	T	0.014	•••	•••	0.007	0.020	0.008
H3	a	g	A	G	I	c	a	g	G	c	C	G	g	T	0.595	0.605	0.487	0.527	0.540	0.551
H3a	a	g	A	G	I	c	a	g	G	c	C	A	g	T	0.014	0.059	•••	0.027	0.053	0.031
H3b	a	a	A	G	I	c	a	g	G	c	C	G	g	T	•••	•••	•••	•••	0.007	0.001
H3c	g	g	A	G	I	c	a	g	G	c	C	G	g	T	0.027	0.020	•••	•••	0.020	0.013
H3d	a	g	A	G	I	t	a	a	G	c	C	G	g	T	0.020	0.020	•••	0.013	•••	0.011
H4	a	g	G	A	I	c	a	g	G	c	C	G	g	C	0.047	0.026	•••	0.020	0.027	0.024
H5	a	g	G	G	I	c	a	g	G	c	C	G	g	C	0.020	0.026	0.020	0.067	0.040	0.035
H5a	a	g	G	G	I	c	g	g	G	c	C	G	g	C	•••	•••	0.020	•••	•••	0.004
H6	a	g	G	G	I	c	a	g	G	c	C	G	g	T	0.027	0.020	0.020	0.060	0.047	0.035
H7	a	g	G	G	D	c	a	g	G	c	C	G	g	T	•••	0.013	•••	0.013	•••	0.005
H8	a	g	A	G	I	c	a	g	G	c	C	G	g	C	0.135	0.092	0.118	0.120	0.107	0.114
H8a	a	g	A	G	I	c	a	g	G	c	C	G	a	C	•••	•••	0.033	•••	•••	0.007
H8b	a	g	A	G	I	c	a	g	G	t	C	G	g	C	0.007	0.007	•••	0.040	•••	0.011
H9	a	g	G	A	I	c	a	g	G	c	C	G	g	T	0.007	0.007	0.007	•••	0.007	0.005
H13	a	g	A	G	I	c	a	g	A	c	G	G	g	C	•••	•••	0.007	•••	0.007	0.003
H14	a	g	A	G	I	c	a	g	G	c	G	G	g	C	0.007	•••	•••	•••	•••	0.001
H14a	a	g	A	G	I	c	g	g	G	c	G	G	g	C	0.007	0.007	0.046	•••	0.007	0.013
H15	a	g	A	G	D	c	a	g	G	c	C	G	g	T	•••	0.007	•••	•••	0.020	0.005

Population specific frequencies are included for each of the five ethnic groups. Number of chromosomes included in these descriptions are as follows; Afar = 148, Amharic = 152, Anuak = 152, Malle = 150 and Oromo = 150. ••• represent frequencies of zero. Haplotypes and SNPs not polymorphic in Ethiopia have not been included. Alleles in lower case represent variants found for the first time by this project.

a.



b.

		AF	AM	AN	ML	OR
Haplotype	H1	11	14	37	16	15
	H2	2	0	0	1	3
	H3	88	92	74	79	81
	H3a	2	9	0	4	8
	H3b	0	0	0	0	1
	H3c	4	3	0	0	3
	H3d	3	3	0	2	0
	H4	7	4	0	3	4
	H5	3	4	3	10	6
	H5a	0	0	3	0	0
	H6	4	3	3	9	7
	H7	0	2	0	2	0
	H8	20	14	18	18	16
	H8a	0	0	5	0	0
	H8b	1	1	0	6	0
	H9	1	1	1	0	1
	H13	0	0	1	0	1
	H14	1	0	0	0	0
	H14a	1	1	7	0	1
	H15	0	1	0	0	3

Figure 6.9 a. Correspondence analysis 2D map representing the interrelationships between the five Ethiopian ethnic groups and the *MUC5B* promoter haplotypes shown in the contingency table (b). AF Afar, AM Amhara, AN Anuak, ML Maale, OR Oromo. Blue filled circles represent the row coordinates and red filled squares represent the column coordinates**.** Correspondence analysis performed by Statistica version 9.

6.3.3.2 H1 status and elevation

There is evidence that the different *MUC5B* promoter haplotypes vary in function (Kamio *et al.* 2005) and *MUC5B* promoter haplotypes have also been shown to be significantly associated with respiratory disease in Europe (chapter 4) and Japan (Kamio *et al.* 2005). It therefore seems possible that the difference in haplotype frequencies between the ethnic groups has resulted from selection as opposed to demography.

Since the respiratory mucins are critical for lung function and elevation is known to affect lung function, we hypothesise that different functional *MUC5B* promoter variants will be favoured at different elevation levels. For instance at higher elevations the air is drier and thus mucin compositions that maximise water retention may be selected for in order to maintain airway hydration. Elevation levels within Ethiopia vary greatly ranging from -125 metres in the Danakil depression to 4,533 metres in Ras Dejen (CIA World factbook – see web citations on page 206) and therefore the Ethiopian data set provided an opportunity to examine *MUC5B* promoter haplotypes with respect to ancestral homeland elevation.

At the outset of this project the idea had been to study the presence or absence of H1 at different elevation levels within the Ethiopian sample as a whole. For each individual, data were available for the location of DNA sample collection and birthplace. In this analysis, birthplace elevation values (obtained from Heavens Above – see web citations on page 206) were used since they are likely to relate more closely to the geographic location of the ancestors of the participants. However, initial observations made this analysis problematic. The overall Ethiopian sample set is stratified with respect to haplotype frequencies due to its composition of five ethnic groups. It was therefore important to determine whether there were significant differences in birthplace elevation levels between the different ethnic groups and as can be seen in figure 6.10 this tends to be the case especially for the Anuak and Maale ethnic groups whereby all individual birthplace elevation levels are highly clustered. This means that any analyses involving birthplace elevation and genetic data, in this Ethiopian sample set, will be confounded by ethnicity.

This problem was circumvented by including only the Amhara and Oromo samples in the analysis. As can be seen from the scatter plots in figure 6.10, birth place elevation levels within these two ethnic groups are varied, making it possible to analyse H1 presence or absence in relation to elevation. In order to maximise sample number, the Amhara and Oromo have been pooled together (n = 124) since no significant difference could be identified between these two populations in the previous haplotypic analysis.

Firstly, each individual was assigned as either a H1 carrier or H1 non-carrier, which simply refers to a typing of at least one H1 haplotype or no H1 haplotype respectively. Initially the 2 by 2 contingency table shown in table 6.4 was constructed to express the H1 status distribution with respect to a high and low birth elevation, defined using the cut off point of 2,400 metres, which is medically recognised as high altitude (Medicine Net – see web citations on page 206). The distribution of these data was not significantly different (Pearsons chi squared 2-sided p value of 0.555).

Table 6.4 Contingency table of the H1 haplotype distribution amongst individuals born at high and low elevation levels.

H1 Status	High	Low	Total
H1 carrier	9	14	23
H1 non-carrier	33	68	101
Total	42	82	124

Data includes a pooled sample of the Amhara and Oromo ethnic groups (n = 124). The high and low categories have been defined by a cut off point of 2,400 metres. H1 carrier status is not significantly different between the high and low elevation sample sets. (Pearsons chi squared 2-sided p value of 0.555).

However significance may have been dampened by the categorisation of birthplace elevation levels and therefore a Mann Whitney U statistical test was used to compare the distributions of raw elevation measures between H1 carrier and H1 non-carrier sample sets (see figure 6.11). Once again no significant association between H1 status and elevation could be identified (p value of 0.824).

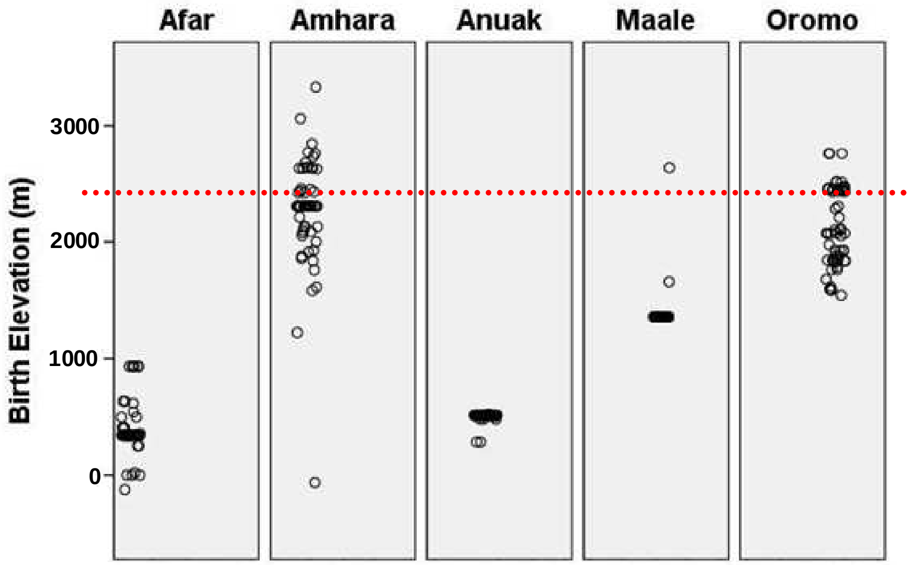


Figure 6.10 Scatter plot of birth elevation levels within each ethnic group. Each data point represents the birth elevation level of a single individual. The red dotted line represents the cut off point of 2,400 metres, used to categorize high and low elevations.

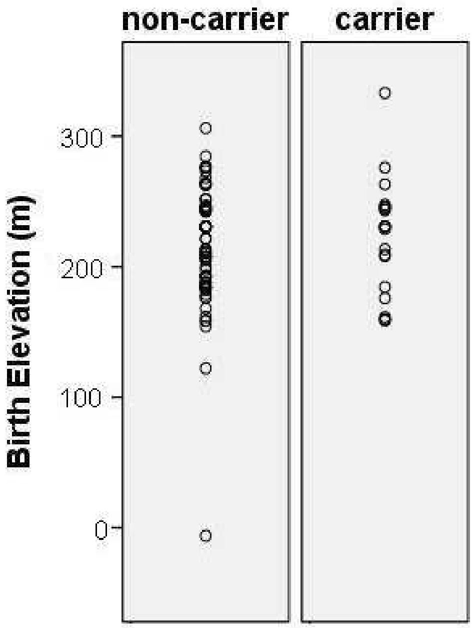


Figure 6.11 Scatter plot of the raw birthplace elevation levels within H1 carrier and H1 non-carrier sample sets. Amhara and Oromo pooled data set. These data were examined with a Mann Whitney U test (p value of 0.824).

6.3.3.3 The high incidence of H1 in the Anuak

The most noteworthy observation of this study is the high incidence of H1 in the Anuak sample. The Anuak have also been noted to be significantly different with respect to other regulatory genetic markers studied in our laboratory, the lactase enhancer and *UGT1A1* promoter TATA box (unpublished - Bryony Jones and Laura Horsfall). This difference may be entirely attributable to the demographic history of the Anuak, whom we know are linguistically different from the other Ethiopian ethnic groups tested within this study since they speak a language descended from the Nilo-Saharan family and all other populations speak languages that belong to the Afro-Asiatic family. However the geography of Ethiopia varies greatly from high mountainous plateaus to inhospitable lowlands and deserts (Levine 2000). The Anuak are the only group in this study that live in a lowland environment on the river banks, and functional genetic variants that may be favourable in such an environment might reach higher frequencies under selective pressures.

Disease prevalence within Ethiopia is also known to vary geographically. For instance, malaria is generally rife in lowland regions such as Gambella where the Anuak originate from, but the highlands (>2,500 metres) are malaria free (WHO). The Gambella region in which the Anuak live has also been reported to have the highest rate of TB and HIV, with prevalence levels only comparable with the three main urban areas, Addis Ababa, Dire Dawa and Harari, plus the Afar region in the case of HIV (WHO Ethiopian TB and HIV reports – see web citations on page 206).

The CIA world fact book (see web citations on page 206) states that meningococcal meningitis is a major respiratory infectious disease that affects Ethiopia, spreading between people via respiratory droplets. There are three clinical manifestations, meningeal syndrome, septicaemia and pneumonia. It is a very serious disease which often results in morbidity or mortality. The bacterium responsible *Neisseria meningitidis* inhabits the nose and throat mucosal membrane. Epidemic levels often accompany the dry seasons or droughts and asymptomatic carriers of the bacteria (5-10% of the population) are thought to play a large role in spreading the disease.

Respiratory inflammation and infection are considered to be risk factors and the highest incidence of this disease occurs throughout the African ‘meningitis belt’ shown in figure 6.12 (WHO meningococcal meningitis fact sheet – see web citations on page 206).

As can be seen in figure 6.12 the ‘meningitis belt’ extends from Senegal to western Ethiopia. Western Ethiopia includes the Gambella region which is inhabited by the Anuak, and it is tempting to speculate that the high frequency of the H1 haplotype within this ethnic group could in fact be in response to the selective pressures of meningococcal meningitis. The higher levels of MUC5B produced by the high expressing H1 haplotype might possibly confer protective properties against meningococcal meningitis and that the Anuak may harbour a greater number of asymptomatic carriers.

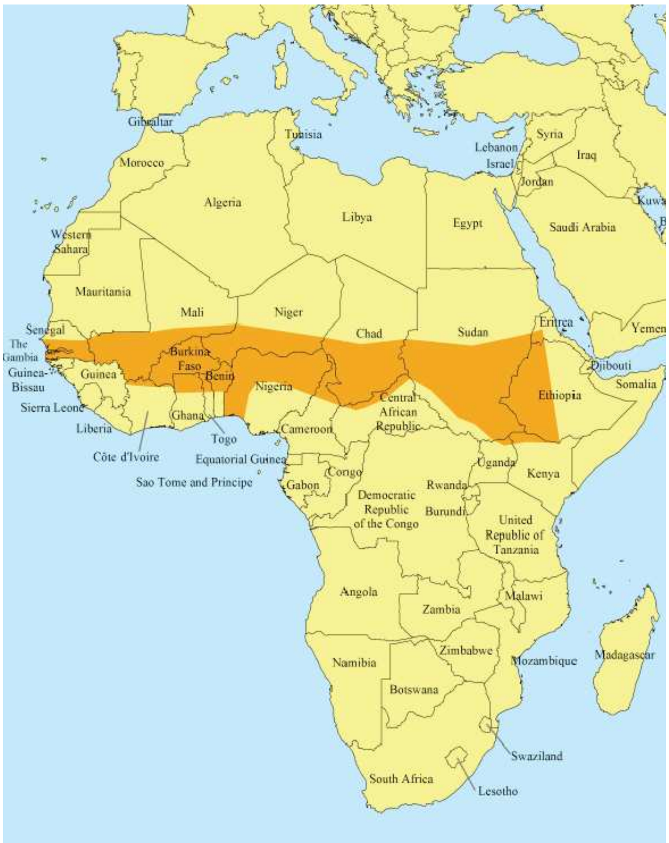


Figure 6.12 African ‘meningitis belt’. Map of Africa displaying the ‘meningitis belt’, the area across sub-Saharan Africa that harbours the highest burden of meningococcal meningitis (http://ehp.niehs.nih.gov/docs/2009/117-1/belt-large.jpg).

6.3.3.4 Nucleotide diversity, tests of neutrality and conservation profiles

In order to investigate the possible effects of selection on the *MUC5B* promoter, nucleotide diversity has been measured and tests of neutrality have been conducted on the Ethiopian data set. These data have also provided an opportunity for examining the properties of the curiously high level of promoter polymorphism with respect to nucleotide sequence conservation profiles.

The expected diversity level of any stretch of sequence is defined as θ (theta) (see materials and methods section 2.4.5). Estimates of θ can be calculated by several different methods which utilise different observations of diversity such as number of segregating sites, number of singletons, level of homozygosity, number of alleles or measures of π (described below). Under neutral evolution or rather neutrality, all estimates of θ should be equal.

A measure of nucleotide diversity (π) (see materials and methods section 2.4.5.1) was calculated for the 1194 nucleotide promoter sequence in each population using analysis software DnaSP (see materials and methods). Measures of π represent the probability that two copies of a specific nucleotide will be the same if selected randomly from a population. Values of zero indicate that all sites are monomorphic. All π values obtained were similar for the five Ethiopian ethnic groups with 11.3×10^{-4}, 11.3×10^{-4}, 16.2×10^{-4}, 12.4×10^{-4} and 12.1×10^{-4} for Afar, Amharic, Anuak, Maale and Oromo populations respectively. Although it should be noted that the Anuak have the largest π value.

Since π was similar for all ethnicities, nucleotide diversity measures were calculated for the Ethiopian population as a whole, using sliding windows of 100 nucleotides to cover the entire input sequence in one nucleotide steps. The π value of each window was plotted at the central point of that window (figure 6.13a). Peaks in the plotted π values represent regions of increased nucleotide diversity, whereas troughs indicate a lack of heterozygosity.

Figure 6.13a shows the π sliding window plot for the Ethiopian population as a whole. The sequence close to the transcription start site exhibits high levels of diversity as does the sequence between approximately -800 to -1000bp relative to the transcription start site.

In order to inspect where these high levels of diversity are with respect to potentially functional regions, a primate conservation profile was constructed for the *MUC5B* proximal promoter (figure 6.13a). Below the diversity plot is a histogram representing sequence conservation between the human, chimpanzee, orangutan and rhesus monkey *MUC5B* proximal promoter sequences, created by phylogenetic shadowing (see materials and methods section). There are two main peaks of conservation. By comparing the two aligned plots in figure 6.13a, it can be seen that the peak furthest from the TSS (approximately -550 to -800), corresponds to a region of little sequence diversity in humans. This is what is expected for regions that are functional, purifying selection having acted upon this region. However the second highly conserved region corresponds to sequence of high diversity in the Ethiopian population, implying that while this region contains functionally important elements, which had been conserved over primate evolution, selective pressures have favoured diversity in this Ethiopian human population.

It should however be noted that the high level of diversity closest to the TSS is interrupted by a steep drop in diversity. This narrow band of low diversity aligns with the highest point of conservation, implying that within this short narrow band of sequence (approximately -159 to -139) variants are not tolerated and therefore this segment of promoter sequence could be considered essential for *MUC5B* expression. This narrow band of sequence lies in between the two putative Sp1 binding sites (-196 to -185 and -124 to -114). The high degree of accumulated diversity either side of this narrow band could indicate that these variants somehow modulate the functional activity of the highly conserved region. As can be seen in figure 6.13b, this narrow band of sequence exhibits a high level of homology (65%) even when the species comparisons are extended to also include the marmoset, horse, dog and mouse. However, a potential TFBS has yet to be identified in this region.

a.

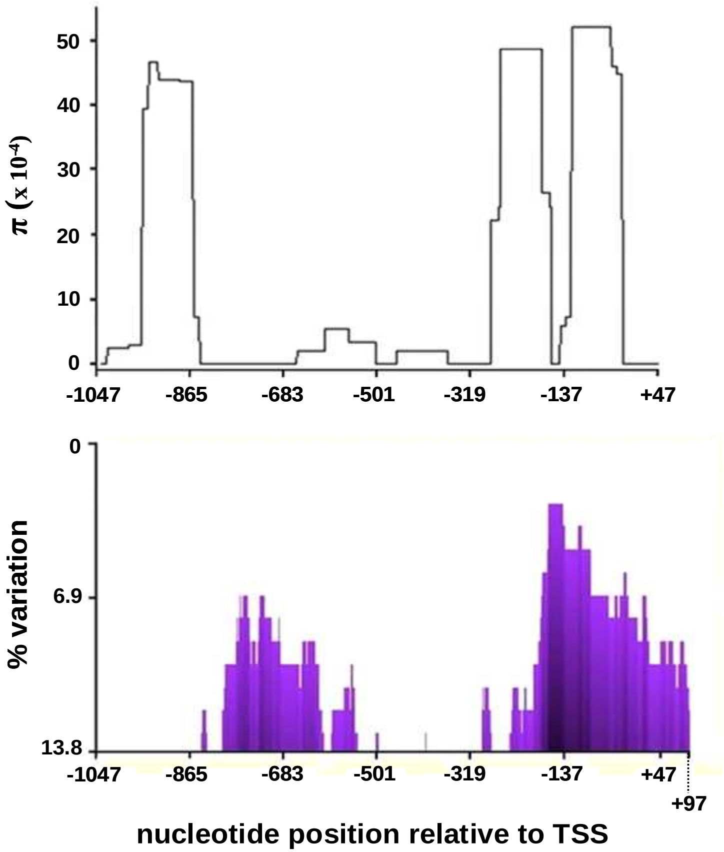


Figure 6.13 Nucleotide diversity and species conservation of the *MUC5B* proximal promoter. a. Aligned plots of nucleotide diversity (π) and sequence conservation profiles for the *MUC5B* proximal promoter. The plot of π values constructed using a sliding window approach (DnaSP), depicts regions of high (peaks) and low (troughs) sequence diversity within the *MUC5B* promoter sequence for the whole Ethiopian sample. The histogram below shows a conservation profile for the human, chimpanzee, orangutan and rhesus monkey *MUC5B* promoter sequences, whereby peaks represent regions of high sequence conservation and troughs represent regions of low sequence conservation (eShadow). The x axis refers to the nucleotide position relative to the transcription start site (TSS). The y axis is a measure of percentage variation, whereby 0 % signifies complete conservation. b. Multiple species alignment (clustalW2) for the sequence region corresponding to the narrow band of high conservation and low sequence diversity at approximately -159 to -139. Grey highlighted sequence corresponds to the putative Sp1 binding sites (-196 to -185 and -124 to -114). Pink highlighted sequence corresponds to the narrow band.

b.

```
                                                                                                    -114
Human       GGGTGGGGGCGTCCTGGGCCTCAG-GGCACAGAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCCCCTCCCCG
Chimp       GGGTGGGGGCTTCCTGGGCCTCAG-GGCACAGAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCCCCTCCCCG
Rhesus      GGGTGGGGGCGTCCTGGGCCTCAG-GGCACAGAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCCCCTCCCCG
Marmoset    GGGCGGGGGCATCTTGG-CCTTAG-AGCACAGAGCTGCAAATCTTTCCTGCTCCAGGCCTCTTCCCTGCCACAGCCCCTCCCCG
Dog         GGGCG-GGGGTCCCGTGAAGCTGG-G-CCCACAGCCGCAAATCCTTCCTTGTCC-GGCGACTTTCCTGCCTCAGCCCCTCCCCA
Horse       GGGGAAGGGGTCCTGCATCTCTGG-GGCCCAGAGCCACAAATCCTTCCTCCTCA-GGGCTCTTTCCTGCCATGGCCCCTCCCCG
Orangutan   GGGTGGGGGCGTCCTGGACCTCAG-GGCACAAAGCTGCAAATCCTTCCTGATCCAGGCCTCTCCCCTGCCACAGCCCCTCCCCG
Mouse       CCACAAGGGGTGGGAGCACTCTGGTGGCACAGTGATGTAAATCCTTCCTTCTCCATGGCCTCTTCCTGCCATGGCCCCTCCCCA
                  ***              *   * **   *    ***** *****  **   *       ******   **********
```

6.3.3.5 Test of neutrality- Tajima’s D

Tajima’s D is a statistical method used to compare estimates of θ based on π with those calculated from the number of segregating sites. Under conditions of neutrality Tajima’s D will be zero. However significantly positive values of D may represent balancing selection and significantly negative values may be indicative of positive selection (Jobling *et al.* 2004).

Values of D were calculated for the Ethiopian group as a whole and for all ethnic groups individually using DnaSP version 5.10.00. No values significantly deviated from those expected under neutrality, though the Anuak was the only population to give a positive result (1.2), all other ethnic groups having negative Tajima’s D values.

Tajima’s D values were also calculated in sliding windows of 100 nucleotides and plotted in the line graphs shown in figure 6.14. It is clear from this figure that the patterns of Tajima’s D values across the *MUC5B* promoter sequence are noticeably different in the Anuak as compared to the other ethnic groups. The most noteworthy observation of the Anuak Tajima’s D plot is the positive double peaks at approximately -300 to -150 and -150 to -30 nucleotides relative to the transcription start site. The steep drop to a Tajima’s D value of zero at approximately -150 corresponds to the narrow band of high conservation and low nucleotide diversity discussed previously (section 6.3.3.4). Either side of this region are two positive Tajima’s D peaks which represent balancing selection or rather a maintenance of multiple alleles. Thus balancing selection may explain the high sequence diversity (high π values) seen either side of the narrow band. It should therefore be noted that while the overall Tajima’s D value for the promoter sequence of the Anuak was not significantly positive, the pattern observed in the sliding window plot is suggestive of balancing selection for these two small regions.

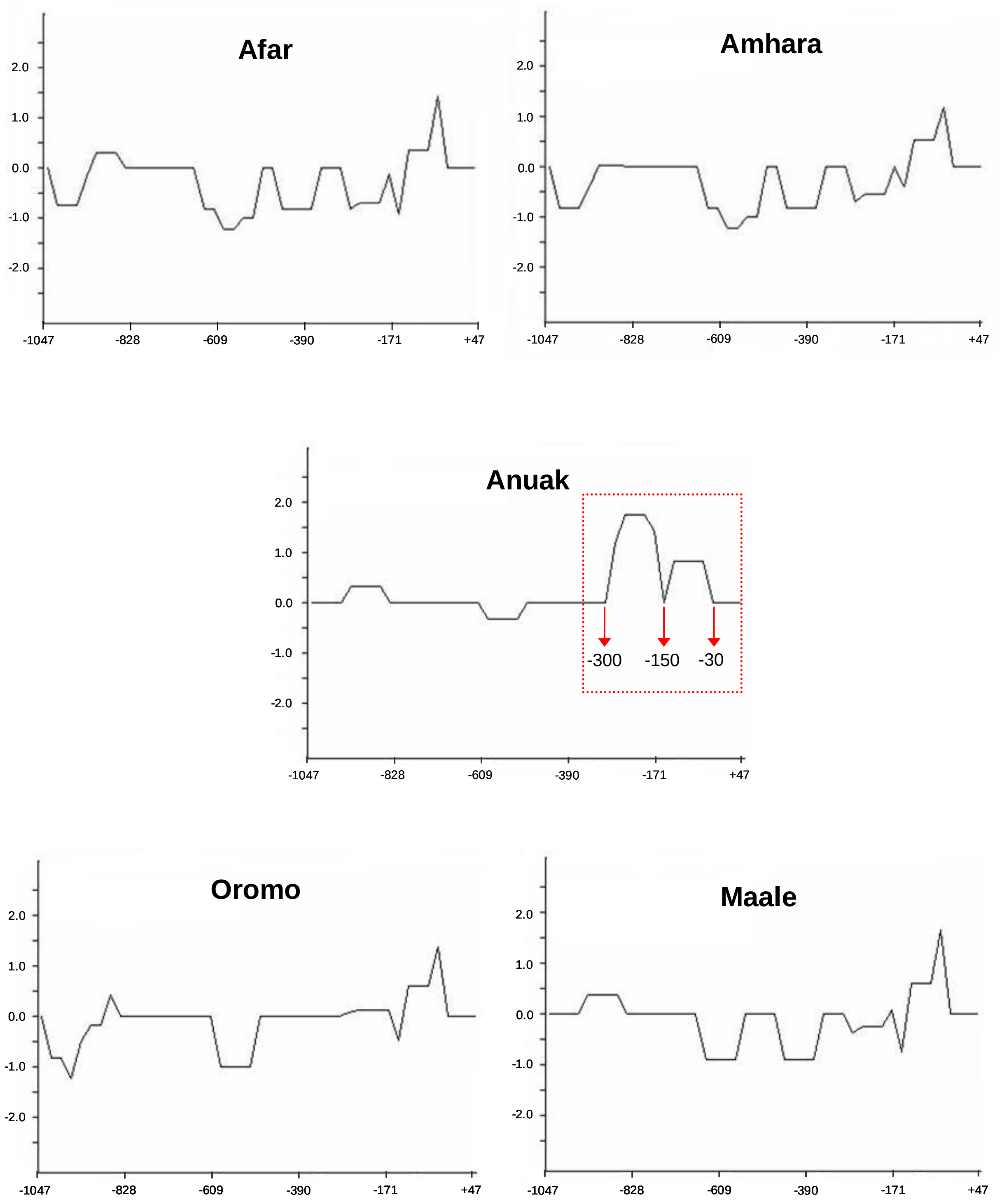


Figure 6.14 Plots of Tajima's D values for the *MUC5B* promoter sequence. Values of D were calculated for sliding windows of 100 nucleotides in length. The value of D for each window was plotted at the windows centre point. Steps between each window was 25 nucleotides. The y axis refers to the D value and the x axis denotes nucleotide position within the sequence relative to the TSS.

6.4 Discussion

This study is the first to characterise *MUC5B* promoter variation within the context of Africa. Approximately 1kb of the sequence directly upstream of the *MUC5B* TSS has been characterised in 748 individuals of African ethnicity, originating from eight main geographic locations. All seven allelic variants previously identified in European (chapter 4) and Japanese (Kamio *et al.* 2005) data sets have also been identified within Africa. However eight additional variants have been discovered within the African populations that were not shown to exist in the non-African samples, which therefore shows that the non-African populations contain only a subset of the African *MUC5B* promoter diversity. These results highlight the importance of genetic studies involving African populations, since African diversity will represent the full spectrum of human genetic variation.

Allelic diversity within the *MUC5B* promoter region appears to be rather extensive. Preliminary characterisation of the *MUC5AC* proximal promoter in the Ethiopian sample set (n = 378) gives a very different picture. In the case of *MUC5AC* we have identified just two SNPs within the 1kb region upstream of the TSS, rs28469016 and rs17859812 with minor allele frequencies of 0.04 and 0.39 respectively. Thus the low level of *MUC5AC* upstream diversity indicates that the extensive promoter variability is not a generalisation for the predominantly expressed respiratory mucins. Although, no functional studies have been performed on the *MUC5AC* promoter and we can not therefore exclude that this arbitrary 1kb upstream region is not the complete promoter of the gene. It is interesting to speculate that the *MUC5B* promoter variation allows variable MUC5B expression because it is an evolutionary adaptation with some haplotypes being favoured in one environment and some in another. Future functional studies will however be required to elucidate whether the novel variants identified during this project have an effect on *MUC5B* expression levels.

Of all the African specific allelic variants, the SNP located at position -194 relative to the TSS, is perhaps the most interesting since it is located within a putative Sp1 binding site. As discussed in section 6.3.3.4, the region containing the two Sp1 binding sites appears to possess the greatest sequence conservation between primates. Sp1 is a zinc finger transcription factor and the two Sp1 motifs located within the *MUC5B* proximal promoter (-196/-185 and -124/-114) have been shown to be functional. In a study using site-directed mutagenesis followed by transfection studies, the Sp1 TFBS at -124/-114 was shown to be essential for basal transcription of *MUC5B* (Wu *et al.* 2007a). The importance of this site is reflected by the complete conservation of the TFBS between human, chimpanzee, orangutan, rhesus monkey, marmoset, dog, horse and mouse sequences. Increased Sp1 binding at the -196/-185 site was shown to responsible for phorbol 12-myristate 13-acetate (PMA) induced *MUC5B* expression (Wu *et al.* 2007a). PMA is a protein kinase C activator used in cell cultures as an inflammatory stimulant, and has been shown to be a potent enhancer of *MUC5B* expression in primary human bronchial cell cultures grown at air liquid interface (ALI). Binding of Sp1 to this site may therefore be important for the upregulation of *MUC5B* by environmental respiratory responses. The -194 variant is likely to be functional since it is located within this environmental response motif and this nucleotide position is conserved between the human, chimpanzee, Orangutan, Rhesus monkey, marmoset, dog and horse genome.

On comparing *MUC5B* promoter nucleotide diversity with a primate conservation profile, an interesting pattern was observed for the region approximately 270 nucleotides upstream of the TSS (see figure 6.13a). Within this region two peaks of high sequence diversity correspond to highly conserved sequence, an unusual pattern since high conservation indicates functionality, implying that diversifying selection or some kind of adaptive selection, has acted upon this region of the human promoter. However the sliding window nucleotide diversity (π) plot shown in figure 6.13a is based only on the Ethiopian sample set and therefore further investigations with regard to possible diversifying selection would benefit from the inclusion of π measurements for various populations around the world.

The sliding window plots created for nucleotide diversity (π) are very effective for studying the patterns of diversity across a length of sequence since the an overall calculation of π for the whole sequence will not be truly representative of sequence segments that confer different functions. For instance, the first 100bp upstream of a genes TSS is of great importance for basal transcription of the gene and therefore sequence variants within this region are more likely to be damaging (Buckland *et al.* 2005). The sliding windows are also very beneficial for calculations such as Tajima's D since they highlight sequence regions that are more susceptible to selective pressures.

The elevated incidence of the high expressing H1 haplotype within the Ethiopian Anuak ethnic group was an interesting but quite unexplained finding. However it is not clear why this divergence between the Anuak and other ethnic groups has occurred and therefore further studies will be needed to determine whether demographic histories or environmental selective pressures have caused the H1 haplotype to reach an elevated frequency in the Anuak. It could be interesting to explore and compare the frequency of this high expressing haplotype in other linguistically similar populations since the Anuak were the only group in this study that speak a Nilo Saharan language. The ultimate goal for the future will be to collect samples of Anuak ethnicity that have been phenotyped for a particular respiratory disease such as TB or phenotyped for *Neisseria meningitidis* carrier status. These sample sets will form a case-control cohort which will be typed for the *MUC5B* proximal promoter haplotypes. Significant associations could then be explored with respect to *MUC5B* promoter variation and respiratory disease susceptibility.

7 General Discussion

The principal aim of this project was to examine genetic variation within *MUC5AC* and *MUC5B* in relation to respiratory disease and demography. Since these genes code for the major components of airway mucus, they are prime candidates for genetic studies on susceptibility and severity of inflammatory respiratory disease.

The first results chapter of this thesis reports significant associations between a synonymous SNP in the 3′ end of *MUC5AC* and five allergy related respiratory outcomes, bronchitis, wheeze, allergy, hayfever and asthma, in a European longitudinal birth cohort. Since these outcomes are not independent because their symptoms are overlapping, it is not surprising that several show significance. It is important to highlight here that the allergy and hayfever outcomes showed the highest level of significance with this *MUC5AC* variant. It should however be noted that this *MUC5AC* exonic SNP does not alter an amino acid and is therefore not considered to be causative, but was chosen at the outset of this project due to a paucity of validated SNPs. The original plan had been to select *MUC5AC* markers likely to be functional or that would give greater coverage as tagging SNPs, to type on the 1946 cohort. However this aim could not be fulfilled during the course of this project since there has been a slow progression of *MUC5AC* research in the literature and a slow accumulation of interesting SNPs in the database. It therefore appeared to be more beneficial to explore the causes of the deviations from HWE that had become evident when the datasets were segregated into affected and unaffected groups with respect to allergy and hayfever, since these results are indicative of CNV which had the potential to be the undiscovered genetic factor causing the significant associations between *MUC5AC* and allergic phenotypes. However the notion of CNV could not be supported experimentally.

In any genetic association study like this, there is a risk of false positive results. This risk is amplified when testing multiple loci simultaneously and the significant associations we see would not stand up to Bonferroni correction. However this issue is of less concern in this project than for example in genomewide studies, since all the research is hypothesis driven, for instance, the multiple *MUC5AC*/inflammatory gene-

gene interaction studies reported in chapter 3, considered only polymorphisms with evidence of functional effect and of genes encoding proteins with likely biological effect on MUC5AC.

As an independent test for the robustness of the associations reported between *MUC5AC* and the various allergy related respiratory outcomes, a permutation test was used which confirmed the significant associations. Since the sample set under investigation was longitudinal, information was also available for various lifestyle and environmental factors such as social class, smoking history, sex and region of birth. These factors are potentially ‘confounders’ since they have the ability to influence the outcomes under study and may cause false positive associations. They were therefore included in the regression models as a method to adjust for these potential confounders.

The mucin genes directly upstream and downstream of *MUC5AC*, *MUC2* and *MUC5B*, were also typed for variants in the longitudinal cohort and tested for association with various respiratory outcomes. However, very little evidence was obtained for any association of *MUC2* or *MUC5B* even though LD across this region was confirmed.

The same *MUC5AC* variant (rs1132440) has also been typed in two small asthmatic cohorts in the second results chapter, but no association could be detected in either case-control disease cohort. This is hardly surprising since the effect size is much smaller than the longitudinal birth cohort which therefore has more power. The reduced power of these asthmatic cohorts may have lead to false negative results especially when considering that associations with asthma had only reached borderline significance in the large longitudinal cohort. It is however curious that some evidence of association was noted between the *MUC5AC* 3′ end haplotypes and severe asthma.

Although the disease cohorts are considerably smaller than the longitudinal cohort, they have the added benefit of being collected in a clinical setting. While the longitudinal cohort provides a wealth of data of relevance to respiratory history and possible confounders, outcomes are self declared, whereas all asthmatic disease cohort participants have been clinically diagnosed which reduces the problem of phenotypic heterogeneity.

Even though there was no significant association of *MUC5B* exonic SNPs and respiratory outcomes in the longitudinal cohort the confirmation that promoter SNPs are associated with altered expression of *MUC5B* (Kamio *et al.* 2005; Loh *et al.* 2010) led us to test these as functional variants in the two small asthmatic disease cohorts. This project reports here for the first time, significant associations between *MUC5B* promoter variants and severe asthma. The *MUC5B* H1 haplotype which appears to direct high expression both *in vivo* (Loh *et al.* 2010) and *in vitro* (Kamio *et al.* 2005), is significantly underrepresented in the severe asthmatic cases as compared to the non-asthmatic control samples, implying that the H1 haplotype confers protection against the development of asthma. This association is somewhat counter-intuitive since one might expect an increase in the incidence of high expressing haplotypes in hypersecretory disease sample sets. However, previous *in vivo* studies on the *MUC5B* promoter have been based only on constitutive expression levels in fetuses (Loh *et al.* 2010) and therefore may not be true for the adult respiratory epithelium particularly after environmental challenge. In order to further understand facultative expression levels of the different *MUC5B* promoter haplotypes, one might perform similar RT-PCR experiments on surgical biopsy tissue from adults with asthma and controls, though such tissue is hard to obtain. It would also be interesting to culture cells grown at air liquid interface, from individuals (asthmatic and not) with the various *MUC5B* haplotypes and compare *MUC5B* expression levels upon exposure to various challenges or rather known asthmatic triggers. It could be suggested that a subtle enhancement of MUC5B production may give the airway mucus greater protective properties and produce a mucosal barrier that is more efficient at excluding inflammatory stimulators, while perhaps over-production clogs the airways

It should be noted that the *MUC5B* rs7115457 variant which determines the H1 haplotype and is significantly associated with severe asthma (chapter 4), is located within a suggested NFκB transcription binding site (Chen *et al.* 2001). If the genetic variant abolishes the NFκB binding site then we might expect the facultative promoter dynamics to become rather complicated when challenged by environmental insults. For instance, by reducing the binding of NFκB to the promoter, other transcription factors may be able to bind to the DNA more efficiently. It should be remembered that the transcription factors driving expression of a gene do not act independently but instead act in concert with each other and environmental cues, in a complexed dynamic process. Thus the *MUC5B* haplotype expression data collected from constitutive *in vivo* fetal studies and *in vitro* reporter construct assays may not be representative of the complex dynamics of the adult *MUC5B* promoter response. Nevertheless the observation of difference in expression in fetuses implies that this SNP can be functional *in vivo*.

In the final results chapter, the distribution of the H1 *MUC5B* promoter haplotype was also shown to be significantly different between Ethiopian ethnic groups. The western Ethiopian Anuak peoples have a significantly greater incidence of H1 compared to four other Ethiopian ethnicities. The Anuak live on lowland river banks in western Ethiopia which is a distinct geographic region from all other groups examined and it is therefore interesting to speculate that environmental specific protective properties of this haplotype have lead to high levels of H1 through positive selection. *Meningococcal meningitis* and tuberculosis endemics are known to be particularly prevalent in the region inhabited by the Anuak and since infection is likely to drive selection it is interesting to speculate that the high incidence of H1 may be a consequence of adaptive protection against respiratory infection.

Characterisation of the *MUC5B* promoter in an African sample set (n = 745) has identified a total of 15 variants within the 1kb sequence upstream of the transcription start site, 7 of which have not been previously identified and are reported here for the first time in this study. This would appear to be a very high level of diversity, and contrast with a preliminary study of the *MUC5AC* promoter in an Ethiopian sample (n = 378), whereby only two SNPs were discovered within the upstream one kilobase sequence. However the functional significance of these novel regulatory SNPs is not known and therefore functional studies such as luciferase assays, electrophoretic mobility shift assays, chromosome immunoprecipitation or *in vivo* heterozygote exonic SNP tagging methods using RNA transcripts, will need to follow.

Perhaps the most interesting novel SNP is the variant at position -194 since it falls within an experimentally determined Sp1 binding site. It is also noteworthy that the high levels of nucleotide diversity within the few hundred bases directly upstream of the TSS, corresponds to high levels of primate species conservation which indicates a region of functional importance. The accumulation and maintenance of variants within this functional region is indicative of balancing selection which might suggest that *MUC5B* promoter diversity has been beneficial. This phenomenon is typical of genes that respond to infectious agents, since different pathogens are prevalent in different geographic regions and thus changes are often in parallel with alterations in environment over time. Also pathogens are continuously evolving and it is therefore advantageous for genetic diversity to exist within the population in order to maximise the chance of survival for at least a small number of individuals.

8 Appendices

Appendix 1- Genotyping of non-mucin genetic markers – conducted by Lynne Vinall

IL1β rs16944

Genotyped using standard PCR conditions with RFLP with BsobI.

PCR primers
Forward 5′ GATTGGCTAGGGTAACAGCACC 3′ (1946-IL1B-C)
Reverse 5′ GGGACAAAGTGGAAGACACACA 3′

IL1RN (tandem repeat)

Standard PCR conditions were used for this assay and visualisation of the PCR product by agarose gel electrophoresis was used to infer allele sizes.

PCR primer
Forward 5′ CTCAGCAACACTCCTA 3′
Reverse 5′ TCCTGGTCTGCAGGTAA 3′

TNF-α rs1800629

A tetra primer arms PCR was used for this assay and visualisation of the PCR product by agarose gel electrophoresis was used to infer SNP genotypes. The four primers, A,B,C and D were added at a ratio of 20:1:1:2 respectively.

Outer PCR primers
Forward – TNFA-C 5′ACCCAAACACAGGCCTCAGGACTCAACA 3′
Reverse – TNFA-D 5′AGTTGGGGACACGCAAGCATGAAGGATA 3′

Inner PCR primers
Forward – TNFA-A 5′TGGAGGCAATAGGTTTTGAGGGGCAGGA 3′
Reverse – TNFA-B 5′ TAGGACCCTGGAGGCTGAACCCCGTACC 3′

EGFR SNP (rs2227983) and EGFR microsatellite

The EGFR SNP and microsatellite/CA repeat were genotyped as a multiplex.

Primers for SNP PCR product
Forward 5′ CAAGGTCATGGAGCACAGG 3′ (CY5)
Reverse 5′ CTGACATTCCGGCAAGAGAC 3′

Primers for the CA repeat
Forward 5′ CTCAAGGTTGGAATTGTGC 3′
Reverse 5′ GCTGTTTGAAGAATTTGAGC 3′ (CY5)

RFLP using AlwNI was used to genotype the SNP. The CA repeat could be genotyped on the basis of PCR product size.

The SNP digest product and the CA repeat PCR product were visualised together on the ALF Express™

Appendix 2

Original Variable	Question asked	Original outcome codes	Recoded Variable	Recoded outcome codes
ALLG89	Have you ever had an allergy (1946-1989)?	0 – No 1 – Yes, once 2 – Yes, recurring	EVALLERG	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
ASTH89	Have you ever had asthma (1946-1989)?	0 – No 1 – Yes, once 2 – Yes, recurring	EVASTHM	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
BRONC89	Have you ever had bronchitis (1989-1999)?	0 – No 1 – Yes, once 2 – Yes, recurring	BRONC89R	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
HAY89	Have you ever had hay fever (1946-1989)	0 – No 1 – Yes, once 2 – Yes, recurring	EVHAY -	0 – Never (original code 0) 1 – Ever (original codes 1 or 2) -
HAYF	In the last ten years (1989-1999), did you have hay fever?	1 – Yes 2 – No		
LRIP	Have you ever had a lower respiratory infection (i.e. bronchitis, broncho pneumonia or pneumonia) in early childhood (0 to 24 mths)?	0 – No attacks of LRI 1 – One attack; no treatment sought 2 – One attack; saw private doctor or was hospital out-patient 3 – One attack; was in-patient at hospital/nursing home 4 – More than one attack; no treatment sought 5 – More than one attack; saw private doctor or was hospital out-patient 6 – More than one attack; was in-patient at hospital/nursing home	LRIPY	0 – Never (original code 0) 1– Ever (original codes 1-6)
WZY89	Has your chest ever sounded wheezy or whistling (1946-1989)?	0 – No 1 – Yes	WZY89C	0 – No or not most days or nights 1 – Yes, most days or nights
	Did you get this most days (or nights)?	0 – No 1 – Yes		
FEVM89	Max FEV1 reading when participant was 43 yrs old.	-	FEVM89C	Outliers removed: 0.01 – 1.00 litres
FEVM99	Max FEV1 reading when participant was 53 yrs	-	FEVM99D	Outliers removed: 0.00 - 0.07,

	old.			0.31 - 1.00, 9.70 – 9.99 litres
FEVM89C – FEVM99D	Difference in Max FEV1 over 10 years (1989-1999).	-	DELTA	-
ALLERGY	In the last ten years (1989-1999), have you had any allergies?	1 – Yes 2 – No	-	-
ASTHMA	In the last ten years (1989-1999), did you have asthma?	1 – Yes 2 – No	-	
BRONC	During the past 3 years (1997-1999), did you have any chest illness, such as bronchitis or pneumonia, which kept you off work or indoors for a week or more?	1 – Yes 2 – No	-	-
WZY	Has your chest ever sounded wheezy or whistling (1946-1999)?	1 – Yes 2 – No	WZYC	0 – No or not most days or nights 1 – Yes, most days or nights
	Did you get this most days (or nights)?	1 – Yes 2 – No		

Table 1 Respiratory measures and outcomes studied in the 1946 birth cohort. Questionnaire data was converted to outcome codes for analysis. Some of these codes were recoded, both are shown in the table. Table adapted from Andrew Loh's thesis (Loh 2007).

Original Variable	Question asked	Original outcome codes	Recoded Variable	Recoded outcome codes
CIG89C	Do you smoke cigarettes? Question was asked in 1966, 1971, 1977, 1982 and 1989.	1 – Currently smoking 2 – Ex-smoker 3 – Never smoked	CIG89CR	Never (code 3) → 0 Ever (codes 1 or 2) → 1
CIG99C	Do you smoke cigarettes? Question asked in 1966, 1971, 1977, 1982, 1989 and 1999.	1 – Currently smoking 2 – Ex-smoker 3 – Never smoked	CIG99CR	Never (code 3) → 0 Ever (codes 1 or 2) → 1
FSC50C	Participant's father's social class at age 4, inferred from father's occupation	1– Professional 2 – Intermediate 3 – Skilled (non-manual) 4 – Skilled (manual) 5 – Partly skilled 6 – Unskilled	FSC50R	Non-manual job (codes 1- 3) → 0 Manual job (codes 4-6) → 1
REG46AR	Participant's region of birth	1 – Scotland 2 – North 3 – Central 4 – London + S.Eastern	-	-
SCL89C	Participant's own social class at age 43, inferred from subject's occupation	1– Professional 2 – Intermediate 3 – Skilled (non-manual) 4 – Skilled (manual) 5 – Partly skilled 6 – Unskilled	SCL89CR	Non-manual job (codes 1- 3) → 0 Manual job (codes 4-6) → 1
SEXX	Sex of participant	1 – Male 2 – Female	SEXXR	Male – 0 Female – 1
SCL99C	Participant's own social class at age 53, inferred from subject's occupation	1– Professional 2 – Intermediate 3 – Skilled (non-manual) 4 – Skilled (manual) 5 – Partly skilled 6 – Unskilled	SCL99CR	Non-manual job (codes 1- 3) → 0 Manual job (codes 4-6) → 1

Table 2 Potential confounder variables studied in the 1946 birth cohort. Questionnaire data was converted to outcome codes for analysis. Some of these codes were recoded, both are shown in the table. Table adapted from Andrew Loh's thesis (Loh 2007)

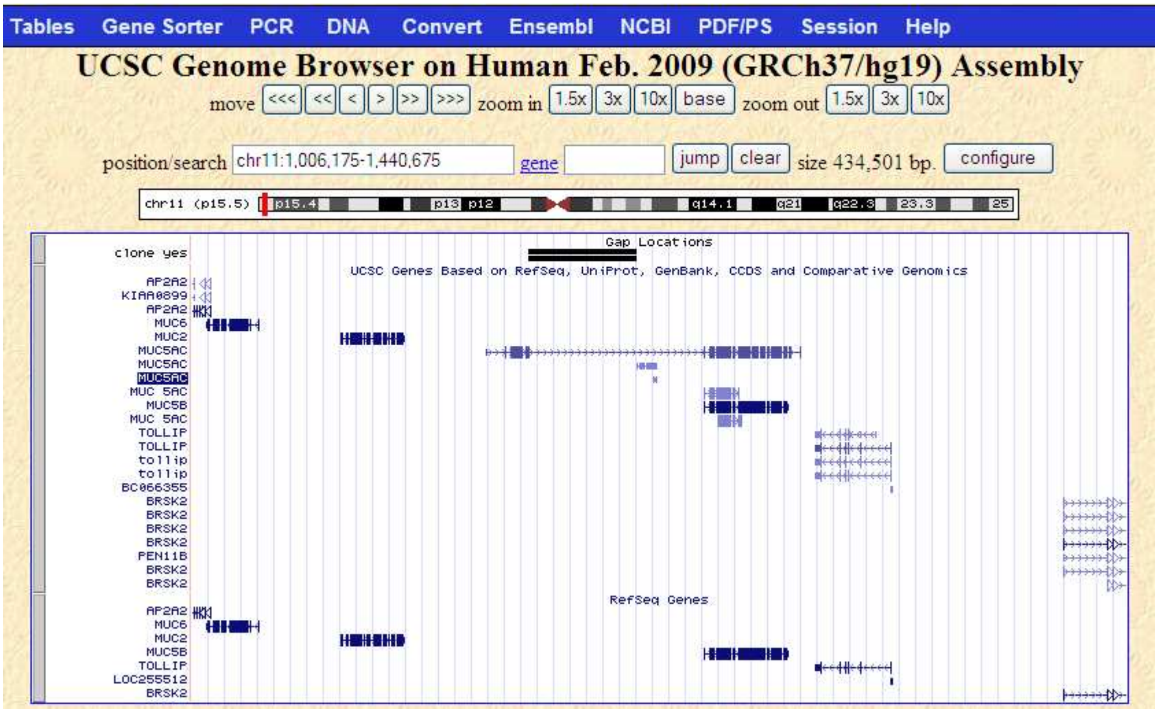
Appendix 3

```
ggggagtctggcccaccctccagaccatcctcaaggcccactggcccaggcatccccgcc

cacccctcccaccgtgccgtgctgcagcgggtctaccggcctggatgtgaaagagagctt

ggagaccccagagacctcggaaccttcagctttggaagtgacgtcggtggggtgggtggg
                                          -2019            -2002
gggggcacaggctctggagtcccggaagtgagcggggagctaYgctgagatctgggagaY
                                     rs11042646       rs55974837
     -1996                                            -1947
ccccctKcccccacccaggtacagggccaggcagaagcccgaggtgtgccctgagKtaaag
     rs35619543                                       rs12804004
aaaccgtcacaaagaacaaagggagaaggcgggttccagcctccaccacagccctcgcgc

tctgaggagccacctgggggcctcagccatgaggggtgacaggtggcaaaacgggccagc

tccgttcacgtcgctgtgcagctgtctccgccctccatctccagaacgttctcacattcc
                       -1738
caagctgaaaccctgtccccatgMaacaccagctcaccatcccctctgccagcccctggc
                       rs868902
gcccaccgtccacactccgtctctgcgggtttcatgactccaggggcagcacacgagtgg

cccctcctgcctttgtcctctgtgtccacctgcctcactctgcacagtgtccccagcttc
                         -1556
ccccatggagcagcctgggccagccYctcctttcacggctgaaccgtattccaccgcac
                         rs868903
ggatcagcctcacgatgctgacccagtcctccgcccagggacacatgggcagcttctgcc
                        -1437
tttgtcagtgatgctgctgtggacWtgggtgtgcaaatgtccctcaggacccgccttcag
                        rs868904
ttcttctggggacagacccagagtggagttgctggtcacccccaccagcagggcacaggg

ctccgggtccccacgtctctgccaacacttcctacttcctgtgtttcttgatccccgcca

tcctattgagcgtgagacaggtcagaagctttgaagatgggctttcgtcttgtcccagaa

atcccacctctaagaatttaacttcagaaagacaaacgcgggggagctggtgcagggccc

gtgacggggactgtgacgtaaataaaacaacagacctggacaccaccctagggtccccat
            DISTAL TATA BOX
```

Putative *MUC5B* distal promoter annotated sequence. The nucleotides highlighted in green have been identified in the matched asthmatic cohort. The putative TATA box has been highlighted in yellow.

Appendix 4

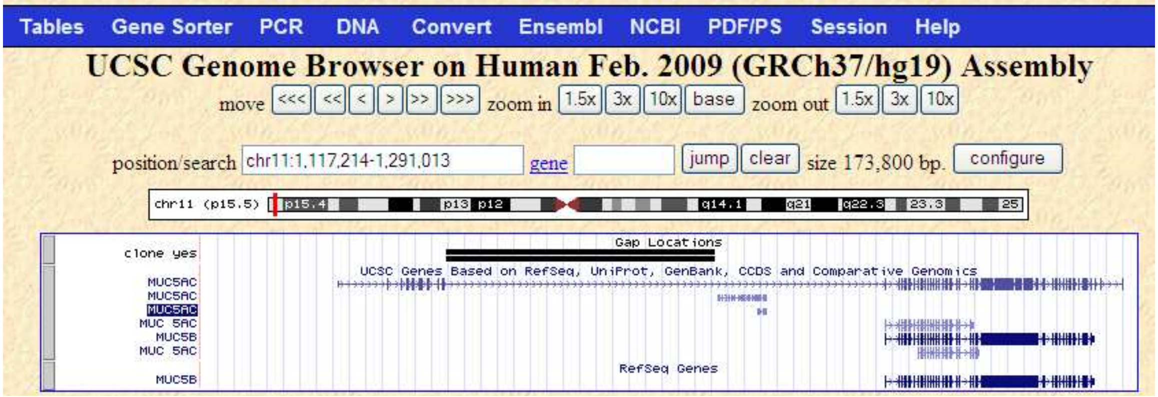
a



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr11:1,006,175-1,440,675 size 434,501 bp.

b



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr11:1,117,214-1,291,013 size 173,800 bp.

Web citations

Ethiopian census 2007 (Summary and statistical report of 2007 population and housing census) - http://www.csa.gov.et/pdf/Cen2007_firstdraft.pdf

Goddard Space Flight NASA - http://www.gsfc.nasa.gov/scienceques2001/20020524.htm

CIA World factbook - https://www.cia.gov/library/publications/the-world-factbook/geos/et.html

Ethnologue - http://www.ethnologue.com/web.asp

13suns - http://www.13suns.com/

Heavens Above (Ethiopian elevation levels) - http://www.heavens-above.com/selecttown.asp?CountryID=ET

Medicine Net - http://www.medterms.com/script/main/art.asp?articlekey=8578

WHO Ethiopian TB report - http://apps.who.int/globalatlas/predefinedReports/TB/PDF_Files/eth.pdf

WHO meningococcal meningitis fact sheet - http://www.who.int/csr/disease/meningococcal/en/index.html

WHO Ethiopian HIV report - http://apps.who.int/globalatlas/predefinedReports/EFS2008/short/EFSCountryProfiles2008_ET.pdf)

Genome completion estimate - http://www.strategicgenomics.com/Genome/index.htm

Ensembl Assembly and Genebuild - http://www.ensembl.org/Homo_sapiens/Info/StatsTable

Hugo Gene Nomenclature Committee (HGNC) - www.genenames.org

Formulae - http://www.statisticssolutions.com/methods-chapter/statistical-tests/mann-whitney-u-test/

OMIM - http://www.ncbi.nlm.nih.gov/omim/

Reference List

1. Albuquerque, R. V., Hayden, C. M., Palmer, L. J., Laing, I. A., Rye, P. J., Gibson, N. A., Burton, P. R., Goldblatt, J., and Lesouef, P. N. (1998). Association of polymorphisms within the tumour necrosis factor (TNF) genes and childhood asthma. *Clin. Exp. Allergy* **28**(5), 578-584.

2. Allen, A., Flemstrom, G., Garner, A., and Kivilaakso, E. (1993). Gastroduodenal mucosal protection. *Physiol Rev.* **73**(4), 823-857.

3. Amishima, M., Munakata, M., Nasuhara, Y., Sato, A., Takahashi, T., Homma, Y., and Kawakami, Y. (1998). Expression of epidermal growth factor and epidermal growth factor receptor immunoreactivity in the asthmatic human airway. *Am. J. Respir. Crit Care Med.* **157**(6 Pt 1), 1907-1912.

4. Andrew, T., Maniatis, N., Carbonaro, F., Liew, S. H., Lau, W., Spector, T. D., and Hammond, C. J. (2008). Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping. *PLoS. Genet.* **4**(10), e1000220.

5. Apter, A. J., Schelleman, H., Walker, A., Addya, K., and Rebbeck, T. (2008). Clinical and genetic risk factors of self-reported penicillin allergy. *J. Allergy Clin. Immunol.* **122**(1), 152-158.

6. Atherton, H. C., Jones, G., and Danahay, H. (2003). IL-13-induced changes in the goblet cell density of human bronchial epithelial cell cultures: MAP kinase and phosphatidylinositol 3-kinase regulation. *Am. J. Physiol Lung Cell Mol. Physiol* **285**(3), L730-L739.

7. Bai, T. R., and Knight, D. A. (2005). Structural changes in the airways in asthma: observations and consequences. *Clin. Sci. (Lond)* **108**(6), 463-477.

8. Bajic, V. B., Choudhary, V., and Hock, C. K. (2004). Content analysis of the core promoter region of human genes. *In Silico. Biol.* **4**(2), 109-125.

9. Beirman, Pearlman, Shapiro, and Busse (1996). Allergy, Asthma, and Immunology from Infancy to Adulthood, Elsevier Health Sciences.

10. Black, S., Teixeira, A. S., Loh, A. X., Vinall, L., Holloway, J. W., Hardy, R., and Swallow, D. M. (2009). Contribution of functional variation in the IL13 gene to allergy, hay fever and asthma in the NSHD longitudinal 1946 birth cohort. *Allergy* **64**(8), 1172-1178.

11. Boyton, R. J., and Openshaw, P. J. (2002). Pulmonary defences to acute respiratory infection. *Br. Med. Bull.* **61**, 1-12.

12. Bradding, P., Roberts, J. A., Britten, K. M., Montefort, S., Djukanovic, R., Mueller, R., Heusser, C. H., Howarth, P. H., and Holgate, S. T. (1994). Interleukin-4, -5, and -6 and tumor necrosis factor-alpha in normal and asthmatic

airways: evidence for the human mast cell as a source of these cytokines. *Am. J. Respir. Cell Mol. Biol.* **10**(5), 471-480.

13. Brayman, M., Thathiah, A., and Carson, D. D. (2004). MUC1: a multifunctional cell surface component of reproductive tissue epithelia. *Reprod. Biol. Endocrinol.* **2**, 4.

14. Breunis, W. B., van, M. E., Bruin, M., Geissler, J., de, B. M., Peters, M., Roos, D., de, H. M., Koene, H. R., and Kuijpers, T. W. (2008). Copy number variation of the activating FCGR2C gene predisposes to idiopathic thrombocytopenic purpura. *Blood* **111**(3), 1029-1038.

15. Broide, D. H., Lotz, M., Cuomo, A. J., Coburn, D. A., Federman, E. C., and Wasserman, S. I. (1992). Cytokines in symptomatic asthma airways. *J. Allergy Clin. Immunol.* **89**(5), 958-967.

16. Buckland, P. R., Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, S. K., and O'Donovan, M. C. (2005). Strong bias in the location of functional promoter polymorphisms. *Hum. Mutat.* **26**(3), 214-223.

17. Buisine, M. P., Desseyn, J. L., Porchet, N., Degand, P., Laine, A., and Aubert, J. P. (1998). Genomic organization of the 3'-region of the human MUC5AC mucin gene: additional evidence for a common ancestral gene for the 11p15.5 mucin gene family. *Biochem. J.* **332 (Pt 3)**, 729-738.

18. Buisine, M. P., Devisme, L., Copin, M. C., Durand-Reville, M., Gosselin, B., Aubert, J. P., and Porchet, N. (1999). Developmental mucin gene expression in the human respiratory tract. *Am. J. Respir. Cell Mol. Biol.* **20**(2), 209-218.

19. Burch, Wise, Garantziotis, Evans, Adler, Speer, Steele, Brown, Loyd, Gudmundsson, Groshong, Dickey, Seibold, Herron, Kervitsky, Talbert, Markin, Zhang, Park, Auerbach, Crews, Slifer, Xu, Potocky, Masinde, Roy, Jancewicz, Schwarz, and Schwartz. Common variants of *MUC5AC* play a role in the development of familial interstitial pneumonia (FIP) and idiopathic pulmonary fibrosis (IPF). 2010.
Ref Type: Unpublished Work

20. Burgel, P. R., Montani, D., Danel, C., Dusser, D. J., and Nadel, J. A. (2007). A morphometric study of mucins and small airway plugging in cystic fibrosis. *Thorax* **62**(2), 153-161.

21. Burgel, P. R., and Nadel, J. A. (2004). Roles of epidermal growth factor receptor activation in epithelial cell repair and mucin production in airway epithelium. *Thorax* **59**(11), 992-996.

22. Campbell, M. C., and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9**, 403-433.

23. Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* **325**(6099), 31-36.

24. Caramori, G., Di, G. C., Carlstedt, I., Casolari, P., Guzzinati, I., Adcock, I. M., Barnes, P. J., Ciaccia, A., Cavallesco, G., Chung, K. F., and Papi, A. (2004). Mucin expression in peripheral airways of patients with chronic obstructive pulmonary disease. *Histopathology* **45**(5), 477-484.

25. Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39**(7 Suppl), S16-S21.

26. Cavalli-Sforza, L. L., Menozzi , P., and Piazza, A. (1994). The history and geography of human genes, Princeton University Press.

27. Chagani, T., Pare, P. D., Zhu, S., Weir, T. D., Bai, T. R., Behbehani, N. A., Fitzgerald, J. M., and Sandford, A. J. (1999). Prevalence of tumor necrosis factor-alpha and angiotensin converting enzyme polymorphisms in mild/moderate and fatal/near-fatal asthma. *Am. J. Respir. Crit Care Med.* **160**(1), 278-282.

28. Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**(1-3), 18-31.

29. Chen, Y., Zhao, Y. H., Di, Y. P., and Wu, R. (2001). Characterization of human mucin 5B gene expression in airway epithelium and the genomic clone of the amino-terminal and 5'-flanking region. *Am. J. Respir. Cell Mol. Biol.* **25**(5), 542-553.

30. Chen, Y., Zhao, Y. H., Kalaslavadi, T. B., Hamati, E., Nehrke, K., Le, A. D., Ann, D. K., and Wu, R. (2004). Genome-wide search and identification of a novel gel-forming mucin MUC19/Muc19 in glandular tissues. *Am. J. Respir. Cell Mol. Biol.* **30**(2), 155-165.

31. Choi, H. J., Chung, Y. S., Kim, H. J., Moon, U. Y., Choi, Y. H., Van, S., I, Baek, S. J., Yoon, H. G., and Yoon, J. H. (2009a). Signal pathway of 17beta-estradiol-induced MUC5B expression in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **40**(2), 168-178.

32. Choi, J. H., Kim, S. H., Cho, B. Y., Lee, S. K., Kim, S. H., Suh, C. H., and Park, H. S. (2009b). Association of TNF-alpha promoter polymorphisms with aspirin-induced urticaria. *J. Clin. Pharm. Ther.* **34**(2), 231-238.

33. Cohn, L., Elias, J. A., and Chupp, G. L. (2004). Asthma: mechanisms of disease persistence and progression. *Annu. Rev. Immunol.* **22**, 789-815.

34. Cohn, L., Homer, R. J., Marinov, A., Rankin, J., and Bottomly, K. (1997). Induction of airway mucus production By T helper 2 (Th2) cells: a critical role for interleukin 4 in cell recruitment but not mucus production. *J. Exp. Med.* **186**(10), 1737-1747.

35. Cohn, L., Tepper, J. S., and Bottomly, K. (1998). IL-4-independent induction of airway hyperresponsiveness by Th2, but not Th1, cells. *J. Immunol.* **161**(8), 3813-3816.

36. Cookson, W. (1999). The alliance of genes and environment in asthma and allergy. *Nature* **402**(6760 Suppl), B5-11.

37. Cookson, W. (2002). Genetics and genomics of asthma and allergic diseases. *Immunol. Rev.* **190**, 195-206.

38. CSGA (1997). A genome-wide search for asthma susceptibility loci in ethnically diverse populations. The Collaborative Study on the Genetics of Asthma (CSGA). *Nat. Genet.* **15**(4), 389-392.

39. Curran, D. R., and Cohn, L. (2009). Advances in Mucous Cell Metaplasia: A Plug for Mucus as a Therapeutic Focus in Chronic Airway Disease. *Am. J. Respir. Cell Mol. Biol.*

40. Daines, M. O., and Hershey, G. K. (2002). A novel mechanism by which interferon-gamma can regulate interleukin (IL)-13 responses. Evidence for intracellular stores of IL-13 receptor alpha -2 and their rapid mobilization by interferon-gamma. *J. Biol. Chem.* **277**(12), 10387-10393.

41. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**(2), 229-232.

42. Danis, V. A., Millington, M., Hyland, V. J., and Grennan, D. (1995). Cytokine production by normal human monocytes: inter-subject variation and relationship to an IL-1 receptor antagonist (IL-1Ra) gene polymorphism. *Clin. Exp. Immunol.* **99**(2), 303-310.

43. Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., and Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**(6897), 544-548.

44. Desseyn, J. L., Aubert, J. P., Van, S., I, Porchet, N., and Laine, A. (1997a). Genomic organization of the 3' region of the human mucin gene MUC5B. *J. Biol. Chem.* **272**(27), 16873-16883.

45. Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P., and Laine, A. (1998). Genomic organization of the human mucin gene MUC5B. cDNA and genomic sequences upstream of the large central exon. *J. Biol. Chem.* **273**(46), 30157-30164.

46. Desseyn, J. L., Guyonnet-Duperat, V., Porchet, N., Aubert, J. P., and Laine, A. (1997b). Human mucin gene MUC5B, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. *J. Biol. Chem.* **272**(6), 3168-3178.

47. Dobbs, and Gee. Application information: Automating heterozygote detection using the human p53 gene as a model. 2002. Beckman Coulter, Inc.
Ref Type: Report

48. Escande, F., Aubert, J. P., Porchet, N., and Buisine, M. P. (2001). Human mucin gene MUC5AC: organization of its 5'-region and central repetitive region. *Biochem. J.* **358**(Pt 3), 763-772.

49. Escande, F., Lemaitre, L., Moniaux, N., Batra, S. K., Aubert, J. P., and Buisine, M. P. (2002). Genomic organization of MUC4 mucin gene. Towards the characterization of splice variants. *Eur. J. Biochem.* **269**(15), 3637-3644.

50. Evans, C. M., and Koo, J. S. (2009). Airway mucus: the good, the bad, the sticky. *Pharmacol. Ther.* **121**(3), 332-348.

51. Evans, C. M., Williams, O. W., Tuvim, M. J., Nigam, R., Mixides, G. P., Blackburn, M. R., DeMayo, F. J., Burns, A. R., Smith, C., Reynolds, S. D., Stripp, B. R., and Dickey, B. F. (2004). Mucin is produced by clara cells in the proximal airways of antigen-challenged mice. *Am. J. Respir. Cell Mol. Biol.* **31**(4), 382-394.

52. Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol. Bioinform. Online.* **1**, 47-50.

53. Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**(5), 921-927.

54. Fahy, J. V. (2002). Goblet cell and mucin gene abnormalities in asthma. *Chest* **122**(6 Suppl), 320S-326S.

55. Fichtner-Feigl, S., Strober, W., Kawakami, K., Puri, R. K., and Kitani, A. (2006). IL-13 signaling through the IL-13alpha2 receptor is involved in induction of TGF-beta1 production and fibrosis. *Nat. Med.* **12**(1), 99-106.

56. Fokkens, W. J., and Scheeren, R. A. (2000). Upper airway defence mechanisms. *Paediatr. Respir. Rev.* **1**(4), 336-341.

57. Fowler, J. C., Teixeira, A. S., Vinall, L. E., and Swallow, D. M. (2003). Hypervariability of the membrane-associated mucin and cancer marker MUC1. *Hum. Genet.* **113**(6), 473-479.

58. Fujisawa, T., Ide, K., Holtzman, M. J., Suda, T., Suzuki, K., Kuroishi, S., Chida, K., and Nakamura, H. (2008). Involvement of the p38 MAPK pathway in IL-13-induced mucous cell metaplasia in mouse tracheal epithelial cells. *Respirology.* **13**(2), 191-202.

59. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and

Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**(5576), 2225-2229.

60. Gao, J., Shan, G., Sun, B., Thompson, P. J., and Gao, X. (2006). Association between polymorphism of tumour necrosis factor alpha-308 gene promoter and asthma: a meta-analysis. *Thorax* **61**(6), 466-471.

61. Gebhardt, F., Zanker, K. S., and Brandt, B. (1999). Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**(19), 13176-13180.

62. Gensch, E., Gallup, M., Sucher, A., Li, D., Gebremichael, A., Lemjabbar, H., Mengistab, A., Dasari, V., Hotchkiss, J., Harkema, J., and Basbaum, C. (2004). Tobacco smoke control of mucin production in lung cells requires oxygen radicals AP-1 and JNK. *J. Biol. Chem.* **279**(37), 39085-39093.

63. Graves, P. E., Kabesch, M., Halonen, M., Holberg, C. J., Baldini, M., Fritzsch, C., Weiland, S. K., Erickson, R. P., von, M. E., and Martinez, F. D. (2000). A cluster of seven tightly linked polymorphisms in the IL-13 gene is associated with total serum IgE levels in three populations of white children. *J. Allergy Clin. Immunol.* **105**(3), 506-513.

64. Gray, T., Coakley, R., Hirsh, A., Thornton, D., Kirkham, S., Koo, J. S., Burch, L., Boucher, R., and Nettesheim, P. (2004a). Regulation of MUC5AC mucin secretion and airway surface liquid metabolism by IL-1beta in human bronchial epithelia. *Am. J. Physiol Lung Cell Mol. Physiol* **286**(2), L320-L330.

65. Gray, T., Nettesheim, P., Loftin, C., Koo, J. S., Bonner, J., Peddada, S., and Langenbach, R. (2004b). Interleukin-1beta-induced mucin production in human airway epithelium is mediated by cyclooxygenase-2, prostaglandin E2 receptors, and cyclic AMP-protein kinase A signaling. *Mol. Pharmacol.* **66**(2), 337-346.

66. Groneberg, D. A., Eynott, P. R., Lim, S., Oates, T., Wu, R., Carlstedt, I., Roberts, P., McCann, B., Nicholson, A. G., Harrison, B. D., and Chung, K. F. (2002a). Expression of respiratory mucins in fatal status asthmaticus and mild asthma. *Histopathology* **40**(4), 367-373.

67. Groneberg, D. A., Eynott, P. R., Oates, T., Lim, S., Wu, R., Carlstedt, I., Nicholson, A. G., and Chung, K. F. (2002b). Expression of MUC5AC and MUC5B mucins in normal and cystic fibrosis lung. *Respir. Med.* **96**(2), 81-86.

68. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E., and Kim, Y. S. (1992). The human MUC2 intestinal mucin has cysteine-rich subdomains located both upstream and downstream of its central repetitive region. *J. Biol. Chem.* **267**(30), 21375-21383.

69. Gupta, V., Sarin, B. C., Changotra, H., and Sehajpal, P. K. (2005). Association of G-308A TNF-alpha polymorphism with bronchial asthma in a North Indian population. *J. Asthma* **42**(10), 839-841.

70. Hattrup, C. L., and Gendler, S. J. (2008). Structure and function of the cell surface (tethered) mucins. *Annu. Rev. Physiol* **70**, 431-457.

71. Heazlewood, C. K., Cook, M. C., Eri, R., Price, G. R., Tauro, S. B., Taupin, D., Thornton, D. J., Png, C. W., Crockford, T. L., Cornall, R. J., Adams, R., Kato, M., Nelms, K. A., Hong, N. A., Florin, T. H., Goodnow, C. C., and McGuckin, M. A. (2008). Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS. Med.* **5**(3), e54.

72. Heinzmann, A., Mao, X. Q., Akaiwa, M., Kreomer, R. T., Gao, P. S., Ohshima, K., Umeshita, R., Abe, Y., Braun, S., Yamashita, T., Roberts, M. H., Sugimoto, R., Arima, K., Arinobu, Y., Yu, B., Kruse, S., Enomoto, T., Dake, Y., Kawai, M., Shimazu, S., Sasaki, S., Adra, C. N., Kitaichi, M., Inoue, H., Yamauchi, K., Tomichi, N., Kurimoto, F., Hamasaki, N., Hopkin, J. M., Izuhara, K., Shirakawa, T., and Deichmann, K. A. (2000). Genetic variants of IL-13 signalling and human asthma and atopy. *Hum. Mol. Genet.* **9**(4), 549-559.

73. Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Hum. Mol. Genet.* **18**(R1), R1-R8.

74. Herrmann, A., Davies, J. R., Lindell, G., Martensson, S., Packer, N. H., Swallow, D. M., and Carlstedt, I. (1999). Studies on the "insoluble" glycoprotein complex from human colon. Identification of reduction-insensitive MUC2 oligomers and C-terminal cleavage. *J. Biol. Chem.* **274**(22), 15828-15836.

75. Hershey, G. K. (2003). IL-13 receptors and signaling pathways: an evolving web. *J. Allergy Clin. Immunol.* **111**(4), 677-690.

76. Hinojosa-Kurtzberg, A. M., Johansson, M. E., Madsen, C. S., Hansson, G. C., and Gendler, S. J. (2003). Novel MUC1 splice variants contribute to mucin overexpression in CFTR-deficient mice. *Am. J. Physiol Gastrointest. Liver Physiol* **284**(5), G853-G862.

77. Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C., Traupe, H., de, J. G., den, H. M., Reis, A., Armour, J. A., and Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* **40**(1), 23-25.

78. Holt, P. G., Macaubas, C., Stumbles, P. A., and Sly, P. D. (1999). The role of allergy in the development of asthma. *Nature* **402**(6760 Suppl), B12-B17.

79. Hovenberg, H. W., Davies, J. R., and Carlstedt, I. (1996). Different mucins are produced by the surface epithelium and the submucosa in human trachea: identification of MUC5AC as a major mucin from the goblet cells. *Biochem. J.* **318 (Pt 1)**, 319-324.

80. Howard, T. D., Whittaker, P. A., Zaiman, A. L., Koppelman, G. H., Xu, J., Hanley, M. T., Meyers, D. A., Postma, D. S., and Bleecker, E. R. (2001). Identification and association of polymorphisms in the interleukin-13 gene with asthma and atopy in a Dutch population. *Am. J. Respir. Cell Mol. Biol.* **25**(3), 377-384.

81. Huang, S. L., Su, C. H., and Chang, S. C. (1997). Tumor necrosis factor-alpha gene polymorphism in chronic bronchitis. *Am. J. Respir. Crit Care Med.* **156**(5), 1436-1439.

82. Humbert, M., Durham, S. R., Kimmitt, P., Powell, N., Assoufi, B., Pfister, R., Menz, G., Kay, A. B., and Corrigan, C. J. (1997). Elevated expression of messenger ribonucleic acid encoding IL-13 in the bronchial mucosa of atopic and nonatopic subjects with asthma. *J. Allergy Clin. Immunol.* **99**(5), 657-665.

83. Hurme, M., and Santtila, S. (1998). IL-1 receptor antagonist (IL-1Ra) plasma levels are co-ordinately regulated by both IL-1Ra and IL-1beta genes. *Eur. J. Immunol.* **28**(8), 2598-2602.

84. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* **36**(9), 949-951.

85. Iontcheva, I., Oppenheim, F. G., Offner, G. D., and Troxler, R. F. (2000). Molecular mapping of statherin- and histatin-binding domains in human salivary mucin MG1 (MUC5B) by the yeast two-hybrid system. *J. Dent. Res.* **79**(2), 732-739.

86. Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de, L. J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**(7181), 998-1003.

87. Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**(2), 217-222.

88. Jobling, Hurles, and Tyler-Smith (2004). Human Evolutionary Genetics, Garland Science.

89. Joos, L., McIntyre, L., Ruan, J., Connett, J. E., Anthonisen, N. R., Weir, T. D., Pare, P. D., and Sandford, A. J. (2001). Association of IL-1beta and IL-1 receptor antagonist haplotypes with rate of decline in lung function in smokers. *Thorax* **56**(11), 863-866.

90. Kamio, K., Matsushita, I., Hijikata, M., Kobashi, Y., Tanaka, G., Nakata, K., Ishida, T., Tokunaga, K., Taguchi, Y., Homma, S., Nakata, K., Azuma, A., Kudoh, S., and Keicho, N. (2005). Promoter analysis and aberrant expression of the MUC5B gene in diffuse panbronchiolitis. *Am. J. Respir. Crit Care Med.* **171**(9), 949-957.

91. Kasaian, M. T., and Miller, D. K. (2008). IL-13 as a therapeutic target for respiratory disease. *Biochem. Pharmacol.* **76**(2), 147-155.

92. Kay (2000a). Allergy and allergic disease: with a view to the future.

93. Kay, A. B. (2000b). Overview of 'allergy and allergic diseases: with a view to the future'. *Br. Med. Bull.* **56**(4), 843-864.

94. Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A. P., Bentley, D., Cardon, L. R., and Deloukas, P. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**(6), 577-588.

95. Kemp, A., and Bjorksten, B. (2003). Immune deviation and the hygiene hypothesis: a review of the epidemiological evidence. *Pediatr. Allergy Immunol.* **14**(2), 74-80.

96. Kesimer, M., and Sheehan, J. K. (2008). Analyzing the functions of large glycoconjugates through the dissipative properties of their absorbed layers using the gel-forming mucin MUC5B as an example. *Glycobiology* **18**(6), 463-472.

97. Kirkbride, H. J., Bolscher, J. G., Nazmi, K., Vinall, L. E., Nash, M. W., Moss, F. M., Mitchell, D. M., and Swallow, D. M. (2001). Genetic polymorphism of MUC7: allele frequencies and association with asthma. *Eur. J. Hum. Genet.* **9**(5), 347-354.

98. Kirkham, S., Kolsum, U., Rousseau, K., Singh, D., Vestbo, J., and Thornton, D. J. (2008). MUC5B is the major mucin in the gel phase of sputum in chronic obstructive pulmonary disease. *Am. J. Respir. Crit Care Med.* **178**(10), 1033-1039.

99. Kirkham, S., Sheehan, J. K., Knight, D., Richardson, P. S., and Thornton, D. J. (2002). Heterogeneity of airways mucus: variations in the amounts and glycoforms of the major oligomeric mucins MUC5AC and MUC5B. *Biochem. J.* **361**(Pt 3), 537-546.

100. Koo, J. S., Kim, Y. D., Jetten, A. M., Belloni, P., and Nettesheim, P. (2002). Overexpression of mucin genes induced by interleukin-1 beta, tumor necrosis factor-alpha, lipopolysaccharide, and neutrophil elastase is inhibited by a retinoic acid receptor alpha antagonist. *Exp. Lung Res.* **28**(4), 315-332.

101. Kuperman, D. A., Huang, X., Koth, L. L., Chang, G. H., Dolganov, G. M., Zhu, Z., Elias, J. A., Sheppard, D., and Erle, D. J. (2002). Direct effects of interleukin-13 on epithelial cells cause airway hyperreactivity and mucus overproduction in asthma. *Nat. Med.* **8**(8), 885-889.

102. Kuyper, L. M., Pare, P. D., Hogg, J. C., Lambert, R. K., Ionescu, D., Woods, R., and Bai, T. R. (2003). Characterization of airway plugging in fatal asthma. *Am. J. Med.* **115**(1), 6-11.

103. Kwok (2010). Methods in Molecular Biology. Single Nucleotide Polymorphisms : Methods and Protocols.

104. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J.,

Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la, B. M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de, J. P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860-921.

105. Leroux (2010). The meteorology and climate of tropical Africa, Springer –Praix press.

106. Lesuffleur, T., Porchet, N., Aubert, J. P., Swallow, D., Gum, J. R., Kim, Y. S., Real, F. X., and Zweibaum, A. (1993). Differential expression of the human mucin genes MUC1 to MUC5 in relation to growth and differentiation of different mucus-secreting HT-29 cell subpopulations. *J. Cell Sci.* **106 (Pt 3)**, 771-783.

107. Leung, T. F., Tang, N. L., Chan, I. H., Li, A. M., Ha, G., and Lam, C. W. (2001). A polymorphism in the coding region of interleukin-13 gene is associated with atopy but not asthma in Chinese children. *Clin. Exp. Allergy* **31**(10), 1515-1521.

108. Levine (2000). Greater Ethiopia: Evolution of a Multi-ethnic Society, Chicago University Press.

109. Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS. Biol.* **5**(10), e254.

110. Li, D., Gallup, M., Fan, N., Szymkowski, D. E., and Basbaum, C. B. (1998). Cloning of the amino-terminal and 5'-flanking region of the human MUC5AC mucin gene and transcriptional up-regulation by bacterial exoproducts. *J. Biol. Chem.* **273**(12), 6812-6820.

111. Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., and Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866), 1100-1104.

112. Li, Y. F., Gauderman, W. J., Avol, E., Dubeau, L., and Gilliland, F. D. (2006). Associations of tumor necrosis factor G-308A with childhood asthma and wheezing. *Am. J. Respir. Crit Care Med.* **173**(9), 970-976.

113. Linden, S., Nordman, H., Hedenbro, J., Hurtig, M., Boren, T., and Carlstedt, I. (2002). Strain- and blood group-dependent binding of Helicobacter pylori to human gastric MUC5AC glycoforms. *Gastroenterology* **123**(6), 1923-1930.

114. Loh , A. Cis-acting polymorphism of *MUC* gene expression. 2007. University College London.
Ref Type: Thesis/Dissertation

115. Loh , A., Johnson , L., Ng , W., and Swallow , D. Cis-acting allelic variation in MUC5B mRNA expression associated with different promoter haplotypes. 2010.
Ref Type: Unpublished Work

116. Lumsden, A. B., McLean, A., and Lamb, D. (1984). Goblet and Clara cells of human distal airways: evidence for smoking induced changes in their numbers. *Thorax* **39**(11), 844-849.

117. Maier, L. M., Howson, J. M., Walker, N., Spickett, G. P., Jones, R. W., Ring, S. M., McArdle, W. L., Lowe, C. E., Bailey, R., Payne, F., Todd, J. A., and Strachan, D. P. (2006). Association of IL13 with total IgE: evidence against an

inverse association of atopy and diabetes. *J. Allergy Clin. Immunol.* **117**(6), 1306-1313.

118. Mao, X. Q., Kawai, M., Yamashita, T., Enomoto, T., Dake, Y., Sasaki, S., Kataoka, Y., Fukuzumi, T., Endo, K., Sano, H., Aoki, T., Kurimoto, F., Adra, C. N., Shirakawa, T., and Hopkin, J. M. (2000). Imbalance production between interleukin-1beta (IL-1beta) and IL-1 receptor antagonist (IL-1Ra) in bronchial asthma. *Biochem. Biophys. Res. Commun.* **276**(2), 607-612.

119. Michiels, J. J., Berneman, Z., Gadisseur, A., van der, P. M., Schroyens, W., van, d., V, and van, V. H. (2006). Classification and characterization of hereditary types 2A, 2B, 2C, 2D, 2E, 2M, 2N, and 2U (unclassifiable) von Willebrand disease. *Clin. Appl. Thromb. Hemost.* **12**(4), 397-420.

120. Moriai, T., Kobrin, M. S., Hope, C., Speck, L., and Korc, M. (1994). A variant epidermal growth factor receptor exhibits altered type alpha transforming growth factor binding and transmembrane signaling. *Proc. Natl. Acad. Sci. U. S. A* **91**(21), 10217-10221.

121. Moser, R., Fehr, J., Olgiati, L., and Bruijnzeel, P. L. (1992). Migration of primed human eosinophils across cytokine-activated endothelial cell monolayers. *Blood* **79**(11), 2937-2945.

122. Murdock , G. P. (1959). Africa - Its peoples and their culture history, McGraw-Hill book company.

123. Naseer, T., Minshall, E. M., Leung, D. Y., Laberge, S., Ernst, P., Martin, R. J., and Hamid, Q. (1997). Expression of IL-12 and IL-13 mRNA in asthma and their modulation in response to steroid therapy. *Am. J. Respir. Crit Care Med.* **155**(3), 845-851.

124. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature* **456**(7218), 98-101.

125. O'Donnell, R. A., Richter, A., Ward, J., Angco, G., Mehta, A., Rousseau, K., Swallow, D. M., Holgate, S. T., Djukanovic, R., Davies, D. E., and Wilson, S. J. (2004). Expression of ErbB receptors and mucins in the airways of long term current smokers. *Thorax* **59**(12), 1032-1040.

126. Obeng, B. B., Hartgers, F., Boakye, D., and Yazdanbakhsh, M. (2008). Out of Africa: what can be learned from the studies of allergic disorders in Africa and Africans? *Curr. Opin. Allergy Clin. Immunol.* **8**(5), 391-397.

127. Offner, G. D., Nunes, D. P., Keates, A. C., Afdhal, N. H., and Troxler, R. F. (1998). The amino-terminal sequence of MUC5B contains conserved multifunctional D domains: implications for tissue-specific mucin functions. *Biochem. Biophys. Res. Commun.* **251**(1), 350-355.

128. Ordonez, C. L., Khashayar, R., Wong, H. H., Ferrando, R., Wu, R., Hyde, D. M., Hotchkiss, J. A., Zhang, Y., Novikov, A., Dolganov, G., and Fahy, J. V.

(2001). Mild and moderate asthma is associated with airway goblet cell hyperplasia and abnormalities in mucin gene expression. *Am. J. Respir. Crit Care Med.* **163**(2), 517-523.

129. Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M., and Santachiara-Benerecetti, A. S. (1998). Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am. J. Hum. Genet.* **62**(2), 420-434.

130. Peat, J. K., and Li, J. (1999). Reversing the trend: reducing the prevalence of asthma. *J. Allergy Clin. Immunol.* **103**(1 Pt 1), 1-10.

131. Perez-Vilar, J., and Hill, R. L. (1999). The structure and assembly of secreted mucins. *J. Biol. Chem.* **274**(45), 31751-31754.

132. Perrais, M., Pigny, P., Buisine, M. P., Porchet, N., Aubert, J. P., and Van Seuningen-Lempire, I. (2001). Aberrant expression of human mucin gene MUC5B in gastric carcinoma and cancer cells. Identification and regulation of a distal promoter. *J. Biol. Chem.* **276**(18), 15386-15396.

133. Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N. A., Park, H. S., Kim, J. I., Seo, J. S., Yakhini, Z., Laderman, S., Bruhn, L., and Lee, C. (2008). The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* **82**(3), 685-695.

134. Pigny, P., Guyonnet-Duperat, V., Hill, A. S., Pratt, W. S., Galiegue-Zouitina, S., d'Hooge, M. C., Laine, A., Van-Seuningen, I., Degand, P., Gum, J. R., Kim, Y. S., Swallow, D. M., Aubert, J. P., and Porchet, N. (1996). Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. *Genomics* **38**(3), 340-352.

135. Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat. Genet.* **23**(4), 437-441.

136. Raymond, M., and Rousset, F. (1995). An exact test for population differentiation. *Evolution* **49**(6), 1280-1283.

137. Reader, J. R., Tepper, J. S., Schelegle, E. S., Aldrich, M. C., Putney, L. F., Pfeiffer, J. W., and Hyde, D. M. (2003). Pathogenesis of mucous cell metaplasia in a murine asthma model. *Am. J. Pathol.* **162**(6), 2069-2078.

138. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C.,

Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature* **444**(7118), 444-454.

139. Robinson, D. S., Hamid, Q., Ying, S., Tsicopoulos, A., Barkans, J., Bentley, A. M., Corrigan, C., Durham, S. R., and Kay, A. B. (1992). Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N. Engl. J. Med.* **326**(5), 298-304.

140. Rogers, D. F. (2004). Airway mucus hypersecretion in asthma: an undervalued pathology? *Curr. Opin. Pharmacol.* **4**(3), 241-250.

141. Rose, M. C., and Voynow, J. A. (2006). Respiratory tract mucin genes and mucin glycoproteins in health and disease. *Physiol Rev.* **86**(1), 245-278.

142. Rousseau, K., Byrne, C., Griesinger, G., Leung, A., Chung, A., Hill, A. S., and Swallow, D. M. (2007). Allelic association and recombination hotspots in the mucin gene (MUC) complex on chromosome 11p15.5. *Ann. Hum. Genet.* **71**(Pt 5), 561-569.

143. Rousseau, K., Vinall, L. E., Butterworth, S. L., Hardy, R. J., Holloway, J., Wadsworth, M. E., and Swallow, D. M. (2006). MUC7 haplotype analysis: results from a longitudinal birth cohort support protective effect of the MUC7*5 allele on respiratory function. *Ann. Hum. Genet.* **70**(Pt 4), 417-427.

144. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**(6822), 928-933.

145. Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39**(7 Suppl), S7-15.

146. Schmiegel, W., Roeder, C., Schmielau, J., Rodeck, U., and Kalthoff, H. (1993). Tumor necrosis factor alpha induces the expression of transforming growth factor alpha and the epidermal growth factor receptor in human pancreatic cancer cells. *Proc. Natl. Acad. Sci. U. S. A* **90**(3), 863-867.

147. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**(5683), 525-528.

148. Settin, A., Zedan, M., Farag, M., Ezz El, R. M., and Osman, E. (2008). Gene polymorphisms of IL-6(-174) G/C and IL-1Ra VNTR in asthmatic children. *Indian J. Pediatr.* **75**(10), 1019-1023.

149. Sheehan, J. K., Howard, M., Richardson, P. S., Longwill, T., and Thornton, D. J. (1999). Physical characterization of a low-charge glycoform of the MUC5B mucin comprising the gel-phase of an asthmatic respiratory mucous plug. *Biochem. J.* **338 (Pt 2)**, 507-513.

150. Shin, H. D., Park, B. L., Kim, L. H., Jung, J. H., Wang, H. J., Kim, Y. J., Park, H. S., Hong, S. J., Choi, B. W., Kim, D. J., and Park, C. S. (2004). Association of tumor necrosis factor polymorphisms with asthma and serum total IgE. *Hum. Mol. Genet.* **13**(4), 397-403.

151. Stephens, M., and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**(5), 1162-1169.

152. Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**(4), 978-989.

153. Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K., and Sugano, S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **11**(5), 677-684.

154. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3), 585-595.

155. Takeyama, K., Dabbagh, K., Lee, H. M., Agusti, C., Lausier, J. A., Ueki, I. F., Grattan, K. M., and Nadel, J. A. (1999). Epidermal growth factor system regulates mucin production in airways. *Proc. Natl. Acad. Sci. U. S. A* **96**(6), 3081-3086.

156. Tamura, K., Arakawa, H., Suzuki, M., Kobayashi, Y., Mochizuki, H., Kato, M., Tokuyama, K., and Morikawa, A. (2001). Novel dinucleotide repeat polymorphism in the first exon of the STAT-6 gene is associated with allergic diseases. *Clin. Exp. Allergy* **31**(10), 1509-1514.

157. Tarlow, J. K., Blakemore, A. I., Lennard, A., Solari, R., Hughes, H. N., Steinkasserer, A., and Duff, G. W. (1993). Polymorphism in human IL-1 receptor antagonist gene intron 2 is caused by variable numbers of an 86-bp tandem repeat. *Hum. Genet.* **91**(4), 403-404.

158. Thornton, D. J., Rousseau, K., and McGuckin, M. A. (2008). Structure and function of the polymeric mucins in airways mucus. *Annu. Rev. Physiol* **70**, 459-486.

159. Thornton, D. J., and Sheehan, J. K. (2004). From mucins to mucus: toward a more coherent understanding of this essential barrier. *Proc. Am. Thorac. Soc.* **1**(1), 54-61.

160. Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., and Krings, M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**(5254), 1380-1387.

161. Tishkoff, S. A., Goldman, A., Calafell, F., Speed, W. C., Deinard, A. S., Bonne-Tamir, B., Kidd, J. R., Pakstis, A. J., Jenkins, T., and Kidd, K. K. (1998). A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am. J. Hum. Genet.* **62**(6), 1389-1402.

162. Tishkoff, S. A., Pakstis, A. J., Stoneking, M., Kidd, J. R., stro-Bisol, G., Sanjantila, A., Lu, R. B., Deinard, A. S., Sirugo, G., Jenkins, T., Kidd, K. K., and Clark, A. G. (2000). Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. *Am. J. Hum. Genet.* **67**(4), 901-925.

163. Tishkoff, S. A., and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3**(8), 611-621.

164. Toribara, N. W., Gum, J. R., Jr., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M., and Kim, Y. S. (1991). MUC-2 human small intestinal mucin gene structure. Repeated arrays and polymorphism. *J. Clin. Invest* **88**(3), 1005-1013.

165. Turner, McLennan, Bates, and White (2000). Molecular biology. The instant notes series., BIOS Scientific Publishers.

166. Turner, J., and Jones, C. E. (2009). Regulation of mucin expression in respiratory diseases. *Biochem. Soc. Trans.* **37**(Pt 4), 877-881.

167. Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L., and Oefner, P. J. (2000). Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**(3), 358-361.

168. Van de Bovenkamp, J. H., Mahdavi, J., Korteland-Van Male, A. M., Buller, H. A., Einerhand, A. W., Boren, T., and Dekker, J. (2003). The MUC5AC glycoprotein is the primary receptor for Helicobacter pylori in the human stomach. *Helicobacter.* **8**(5), 521-532.

169. van der Pouw Kraan TC, van, V. A., Boeije, L. C., van Tuyl, S. A., de Groot, E. R., Stapel, S. O., Bakker, A., Verweij, C. L., Aarden, L. A., and van der Zee, J.

S. (1999). An IL-13 promoter polymorphism associated with increased risk of allergic asthma. *Genes Immun.* **1**(1), 61-65.

170. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., bu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F., V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., and Nodell, M. (2001). The sequence of the human genome. *Science* **291**(5507), 1304-1351.

171. Verweij, C. L., Hart, M., and Pannekoek, H. (1987). Expression of variant von Willebrand factor (vWF) cDNA in heterologous cells: requirement of the pro-polypeptide in vWF multimer formation. *EMBO J.* **6**(10), 2885-2890.

172. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science* **253**(5027), 1503-1507.

173. Vinall, L. E., Fowler, J. C., Jones, A. L., Kirkbride, H. J., de, B. C., Laine, A., Porchet, N., Gum, J. R., Kim, Y. S., Moss, F. M., Mitchell, D. M., and Swallow, D. M. (2000). Polymorphism of human mucin genes in chest disease: possible significance of MUC2. *Am. J. Respir. Cell Mol. Biol.* **23**(5), 678-686.

174. Vladich, F. D., Brazille, S. M., Stern, D., Peck, M. L., Ghittoni, R., and Vercelli, D. (2005). IL-13 R130Q, a common variant associated with allergy and asthma, enhances effector mechanisms essential for human allergic inflammation. *J. Clin. Invest* **115**(3), 747-754.

175. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**(11), 1665-1674.

176. Wang, M., Xing, Z. M., Lu, C., Ma, Y. X., Yu, D. L., Yan, Z., Wang, S. W., and Yu, L. S. (2003). A common IL-13 Arg130Gln single nucleotide polymorphism among Chinese atopy patients with allergic rhinitis. *Hum. Genet.* **113**(5), 387-390.

177. Wang, T. N., Chen, W. Y., Wang, T. H., Chen, C. J., Huang, L. Y., and Ko, Y. C. (2004). Gene-gene synergistic effect on atopic asthma: tumour necrosis factor-alpha-308 and lymphotoxin-alpha-NcoI in Taiwan's children. *Clin. Exp. Allergy* **34**(2), 184-188.

178. Wang, X., Saito, J., Ishida, T., and Munakata, M. (2006). Polymorphism of egfr Intron1 is associated with susceptibility and severity of asthma. *J. Asthma* **43**(9), 711-715.

179. Wang, Y., Harvey, C., Rousset, M., and Swallow, D. M. (1994). Expression of human intestinal mRNA transcripts during development: analysis by a semiquantitative RNA polymerase chain reaction method. *Pediatr. Res.* **36**(4), 514-521.

180. Whittaker, L., Niu, N., Temann, U. A., Stoddard, A., Flavell, R. A., Ray, A., Homer, R. J., and Cohn, L. (2002). Interleukin-13 mediates a fundamental pathway for airway epithelial mucus induced by CD4 T cells and interleukin-9. *Am. J. Respir. Cell Mol. Biol.* **27**(5), 593-602.

181. Williams, O. W., Sharafkhaneh, A., Kim, V., Dickey, B. F., and Evans, C. M. (2006). Airway mucus: From production to secretion. *Am. J. Respir. Cell Mol. Biol.* **34**(5), 527-536.

182. Witte, J. S., Palmer, L. J., O'Connor, R. D., Hopkins, P. J., and Hall, J. M. (2002). Relation between tumour necrosis factor polymorphism TNFalpha-308 and risk of asthma. *Eur. J. Hum. Genet.* **10**(1), 82-85.

183. Woube, M. (1998). Effect of fire on plant communities and soils in the humid tropical savannah of the Gambela region. *Land Degrad. Develop* **9**, 275-282.

184. Wu, D. Y., Wu, R., Chen, Y., Tarasova, N., and Chang, M. M. (2007a). PMA stimulates MUC5B gene expression through an Sp1-based mechanism in airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **37**(5), 589-597.

185. Wu, H., Romieu, I., Sienra-Monge, J. J., del Rio-Navarro, B. E., Anderson, D. M., Dunn, E. W., Steiner, L. L., Lara-Sanchez, I. C., and London, S. J. (2007b). Parental smoking modifies the relation between genetic variation in tumor necrosis factor-alpha (TNF) and childhood asthma. *Environ. Health Perspect.* **115**(4), 616-622.

186. Ying, S., Meng, Q., Barata, L. T., Robinson, D. S., Durham, S. R., and Kay, A. B. (1997). Associations between IL-13 and IL-4 (mRNA and protein), vascular cell adhesion molecule-1 expression, and the infiltration of eosinophils, macrophages, and T cells in allergen-induced late-phase cutaneous reactions in atopic subjects. *J. Immunol.* **158**(10), 5050-5057.

187. Yoshida, S., Hashimoto, S., Nakayama, T., Kobayashi, T., Koizumi, A., and Horie, T. (1996). Elevation of serum soluble tumour necrosis factor (TNF) receptor and IL-1 receptor antagonist levels in bronchial asthma. *Clin. Exp. Immunol.* **106**(1), 73-78.

188. Yoshikawa, M., Nakajima, T., Tsukidate, T., Matsumoto, K., Iida, M., Otori, N., Haruna, S., Moriyama, H., and Saito, H. (2003). TNF-alpha and IL-4 regulate expression of IL-13 receptor alpha2 on human fibroblasts. *Biochem. Biophys. Res. Commun.* **312**(4), 1248-1255.

189. Young, H. W., Williams, O. W., Chandra, D., Bellinghausen, L. K., Perez, G., Suarez, A., Tuvim, M. J., Roy, M. G., Alexander, S. N., Moghaddam, S. J., Adachi, R., Blackburn, M. R., Dickey, B. F., and Evans, C. M. (2007). Central role of Muc5ac expression in mucous metaplasia and its regulation by conserved 5' elements. *Am. J. Respir. Cell Mol. Biol.* **37**(3), 273-290.

190. Zeyrek, D., Demir, E., Alpman, A., Ozkinay, F., Gulen, F., and Tanac, R. (2008). Association of interleukin-1beta and interleukin-1 receptor antagonist gene polymorphisms in Turkish children with atopic asthma. *Allergy Asthma Proc.* **29**(5), 468-474.

191. Zheng, T., Liu, W., Oh, S. Y., Zhu, Z., Hu, B., Homer, R. J., Cohn, L., Grusby, M. J., and Elias, J. A. (2008). IL-13 receptor alpha2 selectively inhibits IL-13-induced responses in the murine lung. *J. Immunol.* **180**(1), 522-529.

192. Zhu, Y., Ehre, C., Abdullah, L. H., Sheehan, J. K., Roy, M., Evans, C. M., Dickey, B. F., and Davis, C. W. (2008). Munc13-2-/- baseline secretion defect reveals source of oligomeric mucins in mouse airways. *J. Physiol* **586**(7), 1977-1992.