CIS-ACTING POLYMORPHISM OF *MUC* GENE EXPRESSION

BY

ANDREW XIONG WEN LOH

SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY COLLEGE LONDON

OCTOBER 2007

THE GALTON LABORATORY
DEPARTMENT OF BIOLOGY
UNIVERSITY COLLEGE LONDON
WOLFSON HOUSE
4 STEPHENSON WAY
LONDON NW1 2HE

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Mucins are high molecular weight glycoproteins that are the principal protein components of mucus and are found on epithelial surfaces throughout the body, where they play a protective role. Changes in mucin expression have been demonstrated in various diseases, such as asthma and COPD; and in the respiratory system, overproduction of mucus can cause blockage of the airways. Experimental studies show co-ordinated upregulation of both mRNA and mucin protein in response to inflammation.

Recent work suggests that allelic differences in gene expression are a common occurrence in the genome. The major research question addressed in this thesis is whether the *MUC* genes exhibit this phenomenon and if such allelic differences in expression might contribute to an individual's susceptibility to disease. Two mucins, *MUC4* (membrane-bound) and *MUC5B* (secreted) were examined, and their relative allelic transcript levels studied by a single-base extension (SBE) method and related to haplotypes of the two genes.

The first aim for *MUC4* was to explore the possible influence of tandem repeat length on mRNA expression. However it was discovered that *MUC4* alternative splicing is much more extensive than previously reported, with hints that this might be genetically determined, but making the original question unanswerable.

The remainder of the thesis foccussed on *MUC5B.* For the expression studies, a total of 42 heterozygous samples were tested by SBE, of which 10 showed strong evidence of an allelic difference in *MUC5B* expression. These constitutive differences were small (greatest difference about two-fold). However, a highly significant association between *MUC5B* promoter haplotype heterozygosity and the magnitude of the detected allelic differences in *MUC5B* expression was demonstrated. A high expressing haplotype was identified (H1) and carriers of marker alleles for this haplotype were tested in a longitudinal birth cohort (n=2807) to examine this variation in relation to respiratory outcomes (eg. asthma), as well as to measures of lung function.

Significant correlations were detected between *MUC5B* and allergy and wheeze history, although the significance of these correlations is unclear and could reflect differences in MUC5B properties rather than expression, or might have arisen from multiple testing.

ACKNOWLEDGEMENTS

I wish to begin by thanking my primary supervisor Prof Dallas Swallow for her support and guidance these past few years. Her great enthusiasm for research, vast experience in the field of mucins and astute scientific judgement have been influential throughout my doctoral work. I am particularly grateful that whilst Dallas was always interested in my research, she also allowed me the freedom I needed to explore things on my own. It has truly been an honour and joy to work with her.

I also wish to express my gratitude to my second supervisor Prof Sue Povey for her advice and concern at many points during my time here at the Galton Laboratory. I have benefited greatly from Sue's wisdom and I am grateful for the times she put my mind at ease when things seemingly were not going so well.

I am greatly indebted to several past members of the Galton Lab: Mrs Lynne Vinall, Dr Ana Teixeira and in particular, Dr Karine Rousseau. They served as my informal mentors and between them taught me almost everything I know about experimental work in the lab. My thanks also go to Dr Wendy Ng who helped with the various RNA extractions and with the Southern blot experiments, thereby saving me valuable time.

Mr Imran Shah, Ms Lauren Johnson, Ms Ranji Arasaretnam, Mr Christopher Plaster and Mr Krishna Veeramah assisted me in various technical aspects of my PhD work. Imran performed the various logistic regression and statistical analyses on the 1946 birth cohort data; Lauren assisted with the genotyping of the 1946 cohort; Ranji helped with the gDNA extractions; Chris assisted with running the ABI 3100 sequencer and Krishna wrote several programs, which greatly reduced the time required to format my data before it could be used by the various computer programs in my research. To each of you, many thanks. I am also grateful to Mr Jed Digby who provided the fetal samples for this project.

I have had the privilege of sharing an office with Ms Catherine Ingram over the past year or so. Kate was a wonderful friend to talk to, to confide in and I have very much enjoyed sharing an office with her.

I also wish to thank the Overseas Research Students Award Scheme (ORSAS) for giving me a research grant to do my doctoral work.

Lastly, I wish to thank my Father, Dr James Loh, who has been both my strongest critic and strongest supporter. Dad, thank you for looking after me all these years, for listening patiently whenever I complained about problems with my research and encouraging me to soldier on, and for spending extended periods of time with me in London during the past two years, cooking and cleaning when you could have stayed back home and let someone cook and clean for you instead.

I wish to dedicate this thesis to the memory of my Mother, Mei Loh, who passed away after a long and difficult struggle with cancer during the course of my PhD. Words are insufficient to describe my gratitude for all that she has done for me. Mom, I miss you and love you very much - this one is for you.

Andrew X W Loh
Oct 2007, London

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

This thesis is concerned with the study of *cis*-acting allelic variation in mucin gene expression. This first chapter describes the main molecular and genetic features of *MUC4* and *MUC5B*, the two mucin genes investigated in this thesis, provides a brief overview of mammalian gene regulation and concludes with a review of our current understanding of allelic variation in mRNA expression in humans.

1.1 MUCINS

1.1.1 Mucus and its Functions

Mucus is a viscous, slimy, gel-like material that covers various epithelial surfaces found throughout the human body, thereby forming an important interface between the individual and the external environment (reviewed by (Rose 1992).

Mucus performs a number of critical functions: in the respiratory system for example, the large airways are covered with mucus, which is continually driven towards the surface by ciliated cells (Sleigh 1983). This mucus layer provides a perpetual aqueous environment for the cells and also helps to trap foreign material and bacteria passing through the air passages (Lillehoj and Kim 2002). Here, mucus is thought to function primarily as a protective barrier against infection and desiccation.

In the stomach, mucus forms a buffer between the gastric lumen and mucosal surface, thus producing a pH gradient from highly acidic in the lumen to near neutral at the mucosal surface (Allen et al. 1993; Turnberg and Ross 1984). This protects the stomach lining from the corrosive effects of stomach acid. In addition, the mucus layer forms a physical barrier to the pepsin found in digestive fluid, preventing proteolytic digestion of the underlying epithelium (Allen et al. 1986; Kaunitz 1999). The mucus layer is continually renewed, thereby countering peptic erosion of the gel

and preserving its viscoelasticity and protective properties (Slomiany and Slomiany 1991).

Elsewhere in the digestive tract, mucus acts as a lubricant by facilitating the passage of food and faecal material during the process of digestion, thus protecting the underlying cells from mechanical damage (Allen 1981).

1.1.2 Mucus and Disease

Changes in the quality and quantity of mucus have been observed in various inflammatory diseases. For example, mucus hypersecretion is often observed in asthmatic patients and plays a central role in the pathogenesis of severe airway obstruction and asphyxiation in fatal asthma attacks (Rogers 2004; Sidebotham and Roche 2003). Conversely, a deficiency in mucus production can lead to problems such as dry eye syndrome (Pflugfelder, Solomon, and Stern 2000). In intestinal mucosal diseases, there is evidence that the mucus layer is thinner in ulcerative colitis compared to controls; the opposite is true in Crohn's disease versus controls (Pullan et al. 1994).

1.1.3 Mucins and the *MUC* Genes

Mucus is a complex mixture, consisting of water, ions, proteins, lipids and glycoproteins. Mucins are the main glycoproteins components found in mucus (reviewed by (Boat and Cheng 1980; Strous and Dekker 1992)). Mucins are responsible for the viscosity of mucus and are therefore critical to its proper function (Gum, Jr. 1992; Vinall et al. 1998). At least 18 distinct genes encoding mucins have been identified so far (HUGO Committee 7 A.D.). These can be classified under two main categories. These are the genes that encode secreted mucins and the genes that encode membrane-bound mucins, as shown in Table 1-1 below (Moniaux et al. 2001).

Secreted	Membrane-bound
MUC2	*MUC1*
MUC5AC	*MUC3A*
MUC5B	*MUC3B*
MUC6	*MUC4*
MUC7	*MUC12*
MUC19	*MUC13*
	MUC15
	MUC16
	MUC17
	MUC18
	MUC20

Table 1-1 – Classification of genes that encode mucins. *MUC* genes in bold text are gel-forming mucins. *MUC* genes in blue text form a gene complex found on Chromosome 11p15.5 (Pigny et al. 1996). The complete structure and sequence of *MUC8* is unknown, so that this gene cannot be classified yet.

Most of the secreted mucins form gels by forming large disulphide cross-linked oligomers (Carlstedt, Lindgren, and Sheehan 1983; Carlstedt and Sheehan 1984; Hovenberg et al. 1996), whereas membrane-bound mucins are monomeric and are anchored to the cell surface by a hydrophobic membrane-spanning domain (reviewed by (Fowler, Vinall, and Swallow 2001).

The *MUC* genes are not all members of a single gene family although they do have a number of features in common. One of the more characteristic features of the *MUC* genes is the presence of a large central tandem repeat (TR) region, which exhibits a large variation in the number (VNTR) as well as sequence of the tandem repeat units (Fowler, Vinall, and Swallow 2001; Swallow et al. 1987). This VNTR is also reflected in mucin mRNAs and is a direct consequence of the variation observed at the DNA level (Debailleul et al. 1998). The TR regions typically contain a high percentage of serine and threonine residues. As a result, the TR regions carry most of the O-glycosylation found in mucins (Van den et al. 1998) . Hence, changes in the TR length are likely to affect the properties of the mature mucins through variations in size as well as extent of glycosylation. There is increasing evidence that these or other differences in the *MUC* genes influence an individual's susceptibility to various diseases such as cancer and other inflammatory diseases (Kirkbride et al. 2001; Kyo et al. 1999; Silva et al. 2001; Vinall et al. 2000).

The focus of this project is on MUC4 and MUC5B. These two mucins are therefore discussed in detail in the following sections.

1.1.4 MUC4

MUC4 encodes a large mucin with a C terminal membrane anchor and is expressed in numerous tissues such as the lungs, colon, cervix and salivary glands (Gipson et al. 1997; Liu et al. 2002; Nguyen et al. 1996; Ogata et al. 1992; Troxler et al. 1997). *MUC4* was not detected in normal pancreas, gall bladder, biliary epithelial cells and the liver (Balague et al. 1994; Vandenhaute et al. 1997). In the colon, *MUC4* is strongly expressed in columnar and goblet cells at the crypt base, but tends to diminish with increasing cell maturity such that expression at the surface epithelium becomes sporadic (Winterford et al. 1999).

MUC4 is located on chromosome 3q29 and its entire genomic organisation has been successfully elucidated (Moniaux et al. 1999). Figure 1-1 shows the deduced cDNA structure of *MUC4*. The characteristic mucin TR domain is located in exon 2 of this gene and observed tandem repeat lengths vary between 7-19 kb, brought about by a large variation in the number of 48 bp repeat units (Nollet et al. 1998). As a result, the MUC4 glycoprotein can potentially extend up to 2.12 μm above the cell surface, making it one of the largest transmembrane mucins discovered so far (Moniaux et al. 1999).

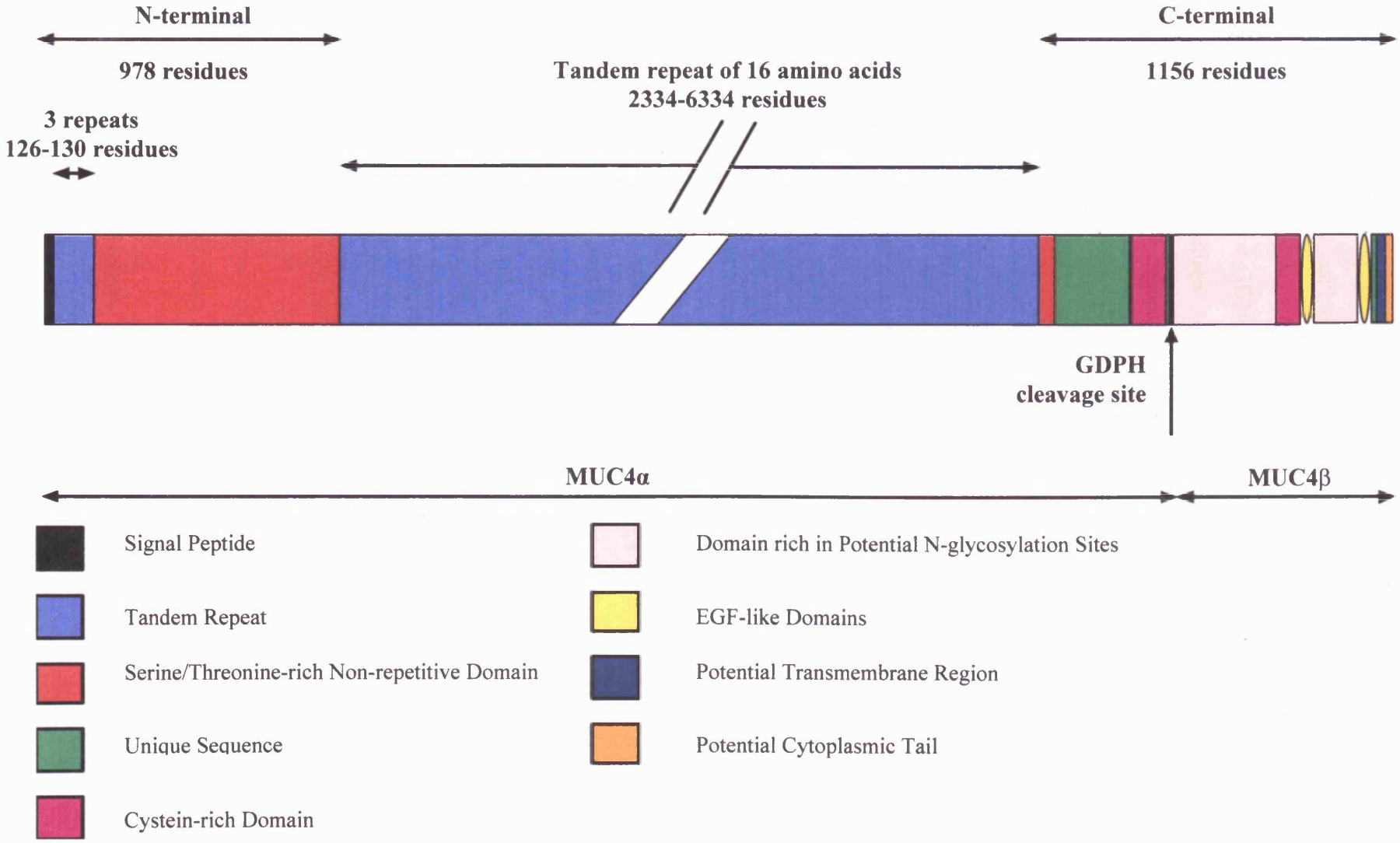


Figure 1-1 – Schematic representation of the deduced structure of MUC4

The human *MUC4* gene is considered to be the orthologue of the sialomucin complex (SMC) or rat Muc4 (Carraway et al. 2000; Moniaux et al. 1999). SMC/Muc4 is a heterodimer, which is composed of two subunits: an O-glycosylated mucin subunit ASGP-1, which is tightly bound to ASGP-2, a N-glycosylated transmembrane subunit (Rossi et al. 1996). The two subunits are translated by a single unique cDNA. MUC4-like SMC/Muc4 precursor contains a GDPH cleavage site, like that which gives rise to the two subunits found in SMC (Moniaux et al. 1999). This therefore suggests that human MUC4 could also be cleaved into two subunits, now known as MUC4α and MUC4β (Moniaux et al. 1999). Therefore, MUC4 could potentially exist in both soluble (MUC4α) and membrane-bound (MUC4β +/- MUC4α) forms.

The hydropathy profile[1] of MUC4β revealed a hydrophobic region of about 24 amino acid residues (Moniaux et al. 1999) which is thought to constitute a membrane-spanning region and hence the classification of MUC4 as a membrane-bound mucin. In addition, this subunit contains three extracellular Epidermal Growth Factor (EGF) like domains (Choudhury et al. 2000). In other proteins, these domains can interact with members of the ErbB receptor tyrosine kinase family and cause activation, thereby playing a role in cell signalling (Holbro and Hynes 2004). However, this association between EGF-like domains and ErbB receptor tyrosine kinases has only been demonstrated between the EGF1 domain in rat ASGP-2 and ErbB2 so far (Carraway, III et al. 1999). This binding has been shown to induce tyrosine auto-phosphorylation of ErbB2, which in turn initiates downstream signalling pathways such as the multiple mitogen activated protein kinase pathways (MAPK) and the phosphtidyl 3–kinase pathway (PI3K) (Carraway et al. 2003). Since these pathways are involved in cell transformation, this suggests that ErbB2 plays a role in this process. Indeed, overexpression of ErbB2 has been found in many cancerous conditions such as breast (Slamon et al. 1987; Slamon et al. 1989), colorectal (Kapitanovic et al. 1997), ovarian (Meden and Kuhn 1997) and non-small-cell lung (Yu et al. 1997).

---

[1] Amino acids are hydrophobic, hydrophilic or neutral (Lodish et al. 2000). Each amino acid is assigned a hydropathy score, which reflects the extent of its hydrophobicity. For a given amino acid sequence, these scores can be used in a hydropathy profile to predict how the protein would interact with aqueous environments and therefore its resultant folding pattern (Kyte and Doolittle 1982; von Heijne 1992)

The significance of ErbB2 (also known as HER2) overexpression has been particularly well studied in breast cancer. ErbB2 overexpression can upregulate MMP-9 and MMP-2 protease activities and increase the invasiveness of breast cancer cells (Tan, Yao, and Yu 1997). In addition, Erbb2 overexpression may lead to a stronger angiogenic response and an enhanced resistance to apoptosis (Kumar and Yarmand-Bagheri 2001; Yu et al. 1998). Together, these can contribute to an increased metastatic potential in breast cancer cells. Indeed, ErbB2 overexpressing breast tumours have been shown to be more resistant to various treatments, including chemotherapy and hormone therapy (Kim et al. 2002a; Leitzel et al. 1995). As such, several therapies have been developed to target the expression or function of the ErbB2 receptor (Badache and Goncalves 2006).

Although it is not known if the EGF-like domains in human MUC4 can interact in a similar fashion to rat ASGP-2, the expression of MUC4 has been found to be very similar to that of ErbB2. Thus, MUC4 cell signalling may play a crucial role in cell proliferation and therefore be important for tumour progression.

Extensive alternative splicing of *MUC4* has been detected. This will be discussed in further detail in section 3.1.2.

1.1.5 MUC5B

MUC5B encodes a large secreted mucin (Van Klinken et al. 1998). *MUC5B* is expressed primarily in the bronchial glands, saliva (where it is sometimes known as MG1), gall bladder and endocervix (Audie et al. 1993; Audie et al. 1995; Campion et al. 1995; Gipson et al. 1999; Keates et al. 1997; Nielsen et al. 1997; Troxler et al. 1997; van Klinken et al. 1998). Like *MUC4*, the genomic organisation of *MUC5B* is known (Desseyn et al. 1998). The deduced protein structure inferred from cDNA is shown in Figure 1-2.

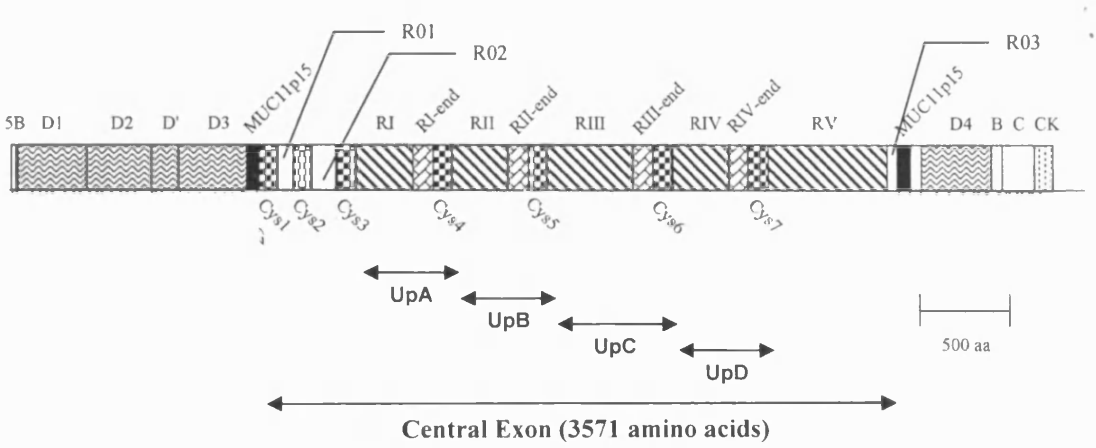


Figure 1-2 – Schematic Representation of the Deduced Structure of MUC5B protein. Adapted from (Desseyn et al. 1998). Cys: cysteine-rich subdomain; R: tandem repeat subdomain; Up: super-repeat unit.

The gene is comprised of 48 exons and the gene possesses an unusually large 10713 bps central exon, containing the tandem repeat domain. The central exon consists of 19 subdomains (Desseyn et al. 1997). There are seven cysteine-rich subdomains (Cys1 to Cys7), eight tandem repeat subdomains (R01 to RV) and four R-end subdomains (RI-end to RV-end). The tandem repeat subdomains contain imperfectly conserved repeats of 87 bps, whilst the R-end subdomains show striking sequence similarities to each other. The 19 subdomains are arranged in a particular pattern, giving rise to four super-repeat units (UpA to UpD). Each of these super-repeat units consists of a tandem repeat subdomain, followed by an R-end subdomain and lastly a cysteine-rich subdomain. The central region is flanked by MUC11p15-like domains, which are similar to those also present in other 11p15.5 *MUC* genes. In contrast to the other *MUC* genes, *MUC5B* shows no evidence of variation in TR length (Vinall et al. 1998).

The amino and carboxyl termini of MUC5B contain domains (D1, D2, D', D3, D4, B and C) that show extensive homology to the similar domains found in human pre pro-von Willebrand factor (Desseyn et al. 1998; Offner et al. 1998) as well as in other 11p15.5 mucins such as *MUC2* and *MU5AC* (see Figure 1-3).

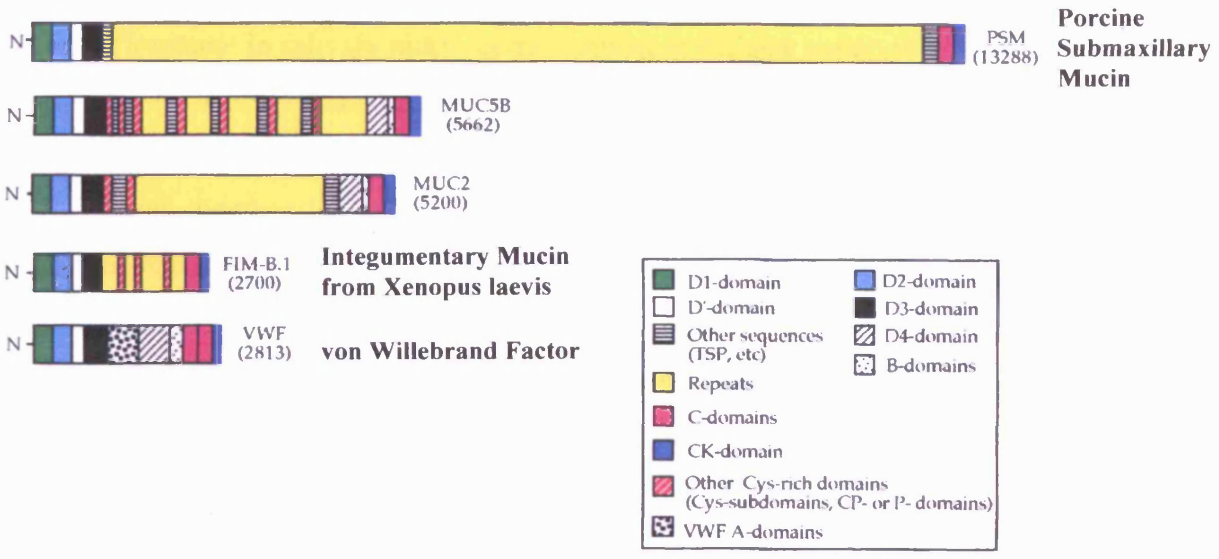


Figure 1-3 – Comparison between von Willebrand Factor and four secreted mucins. Adapted from (Perez-Vilar and Hill 1999)

Human von Willebrand factor (vWF) is a multimeric glycoprotein found in plasma, where it plays a vital role in blood clotting, by stabilising clotting factor VIII and by mediating platelet adhesion (de Groot 2002; Perutelli, Biglino, and Mori 1997). After protein translocation, the signal peptide is cleaved off, leaving behind pro-vWF. Pro-vWF subunits then form dimers in the endoplasmic reticulum and are subsequently transported to the Golgi, where multimerisation takes place (Sadler 1998). The multimerisation of vWF occurs via disulphide linkages between the D domains (Dong et al. 1994; Verweij, Hart, and Pannekoek 1987; Voorberg et al. 1990). The process is thought to involve the sequence CGLCG, which is similar to the sequence of the disulphide isomerase active site (Mayadas and Wagner 1992). As a result, the process is self-catalysed. The CGLCG motifs are found in the D1, D2 and D3 domains of MUC5B and *MUC5AC*. However, the motifs are present only in D1 and D3 of MUC2. In addition, sequence alignments between the three *MUC* genes show the greatest sequence identity in the D3 domains (Offner et al. 1998), suggesting that D3 is the most likely region for the formation of disulphide linkages.

The mature vWF is finally produced when the propeptide is cleaved off via the D′ domain (Sadler 1998). Since MUC5B also contains a D′ domain, it has been proposed that a similar D′ cleavage event occurs in the production of the mature MUC5B protein. Indeed, by using antibodies specific to D1 and D2, Wickstrom et al successfully identified

a protein fragment in salivary mucin corresponding to the size expected if D′ cleavage had occurred in MUC5B (Wickstrom and Carlstedt 2001).

Taken together, the current data suggests that mucin and vWF assembly might be conserved and that the D3 domains may play a role in MUC5B polymerisation, similar to that seen in vWF. Indeed, work by Perez-Vilar et al on porcine submaxillary mucin (the porcine homolog of human MUC5B) demonstrated that PSM could potentially form disulphide-linked dimers via its N-terminal D domains (Perez-Vilar et al. 1998).

Salivary MUC5B has been shown to bind bacteria in various studies, suggesting a possible role in oral disease. Veerman et al demonstrated that MUC5B binds *Hemophilus parainfluenzae* but not other bacteria species such as *Streptococcus* and *Staphylococcus* (Veerman et al. 1995). Periodate acid treatment, partial deglycosylation, or addition of monosaccharides did not affect MUC5B binding to *H. parainfluenzae*. This indicated that the binding was not affected by the carbohydrate side chains but was likely instead to occur via a non-glycosylated domain.

In another study, Bosch et al (Bosch et al. 2000) examined the effects of acute stress on the salivary levels of the carbohydrate structure sulfo-Lewis (sulfo-Le), which is found on MUC5B and has been identified as an adhesion molecule for *Helicobacter pylori* (Namavar et al. 1998; Veerman et al. 1997). They successfully demonstrated a direct correlation between stress-mediated biochemical changes and enhanced saliva-mediated adherence of *H. pylori.*

1.1.6 Evidence of Changes in *MUC* Gene Expression in relation to Disease

A large number studies describe altered expression of mucins in disease, in particular in cancer. Mucins have therefore often been considered as cancer markers. Table 1-2 summarises the studies on MUC4 and Table 1-3 the studies on MUC5B. In some studies, upregulation has been shown at the RNA level while in others increased staining with antibodies could also reflect epitope alterations and subcellular localisation.

In addition to those studies listed, I have myself been involved in a study on biliary tract cancer, where I assisted with the western blot analysis and which is currently being prepared as a full manuscript. This is shown in abstract form in Appendix I. Our study showed evidence for upregulation of *MUC4* at the mRNA and protein levels in biliary tract cancer. In the case of MUC4, the best antibodies recognise the TR domain and are therefore the epitopes that are at risk of being impeded by glycosylation.

Differences in epitope specificity may account for some for the discrepancies that appear to be in the literature. However some discrepancies are also seen for the RNA studies, see for example (Nguyen et al. 1996) and (Lopez-Ferrer et al. 2001b) in Table 1-2.

Disease	Type of Evidence	Observation and Conclusions	Reference
Colon cancer	mRNA (northern blots)	*MUC4* mRNA at comparable or higher levels compared to normal.	(Ogata et al. 1992)
	Protein (immunohistochemistry)	In 50% of the hyperplastic polyps, MUC4 was reduced but in the remaining cases was similar to normal. Loss of MUC4 expression was observed in all serrated adenomas.	(Biemer-Huttmann et al. 1999)
Pancreatic cancer	mRNA (northern blots)	Enhanced expression of *MUC4* in pancreatic cancer compared to normal.	(Balague et al. 1994)
	mRNA (northern blots)	High levels of *MUC4* mRNA expression in pancreatic cancer.	(Hollingsworth et al. 1994)
	mRNA (*in-situ* hybridisation)	*MUC4* undetectable in normal pancreas but detected in pancreatic cancer.	(Balague et al. 1995)
	mRNA (slot blots)	*MUC4* expressed in pancreatic cancer but not in chronic pancreatitis and normal pancreas samples.	(Andrianifahanana et al. 2001)
	mRNA (*In-situ* hybridisation) and Protein (immunohistochemistry)	Highly increased expression was observed in late pancreatic intraepithelial neoplasia (PanIN-3, pre-cursor lesions) and infiltrating ductal adenocarcinoma of the pancreas compared to normal tissue.	(Park et al. 2003)
	Downregulation of MUC4 using cell transfection experiments and anti-sense RNA. Results confirmed with western blots and immunofluorescence analyses.	Anti-sense transfected cells exhibited diminished growth, diminished clonogenic ability and 3-fold decrease in motility compared to controls. Demonstrates direct association of MUC4 mucin with the metastatic pancreatic cancer phenotype.	(Singh et al. 2004)
Invasive ductal carcinoma of the pancreas (IDC)	Protein (immunohistochemistry)	MUC4 expression significantly correlated with poor prognosis.	(Saitou et al. 2005)
Intraductal papillary mucinous neoplasms (IPMNs) of the pancreas	Protein (immunohistochemistry)	Expression of MUC4 found in cancerous lesions of IPMNs, but undetectable in normal and hyperplastic lesions.	(Kanno et al. 2006)
Gastric cancer	mRNA (northern blots)	Increased levels of *MUC4* mRNA in gastric cancer compared to normal.	(Ho et al. 1995)

Disease	Type of Evidence	Observation and Conclusions	Reference
Lung cancer	mRNA (northern blots)	Raised levels of *MUC1*, *MUC3* and *MUC4* mRNA in cancerous tissue. Lung squamous-cell, adenosquamous, and large-cell carcinomas were characterized by increased levels of *MUC4* alone.	(Nguyen et al. 1996)
	mRNA (semi-quantitative RT PCR)and protein (immunohistochemistry)	Reduced expression of MUC4, MUC5AC, and MUC8 was observed in non-small cell carcinomas vs. normal tissue.	(Lopez-Ferrer et al. 2001b)
	mRNA (northern blots)	Intensity of expression of *MUC1* and *MUC4* was always superior in cancer vs. normal tissue.	(Seregni et al. 1996)
	mRNA (northern blots)	44% of the cell lines and 72% of the tumour samples tested showed high levels of *MUC4* mRNA expression	(Hanaoka et al. 2001)
	Protein (Immunohistochemistry)	High MUC4 expression in small-sized lung adenocarcinomas correlates significantly with poor prognosis.	(Tsutsumida et al. 2007)
	mRNA (RT PCR, i*n-situ* hybridisation) and protein (immunohistochemistry)	MUC4 expressed at protein and mRNA levels only in lung adenocarcinomas and not in malignant or benign mesothelial cells. Potential as specific marker for lung cancer.	(Llinares et al. 2004)
Biliary tract cancer	mRNA (*in-situ* hybridisation)	Increased expression of MUC4 and MU5AC of mRNA found in cholangiocarcinomas and the biliary epithelium, especially for dysplastic cells of stone-containing intrahepatic bile ducts compared with normal controls.	(Lee and Liu 2001a)
Extrahepatic bile duct carcinoma	Protein (immunohistochemistry)	High MUC4 expression associated with poor prognosis.	(Tamada et al. 2006)
Salivary gland mucoepidermoid carcinoma	Protein (immunohistochemistry)	Significant trend towards reduction in MUC4 antigen expression in high-grade tumours compared with low-grade and intermediate-grade tumours.	(Weed et al. 2004)
	Protein (immunohistochemistry)	MUC4 expression associated with lower grade tumours and better prognosis.	(Alos et al. 2005) and (Handra-Luca et al. 2005)

Disease	Type of Evidence	Observation and Conclusions	Reference
Ovarian cancer	mRNA (northern blots)	Significant decrease in expression of *MUC3* and *MUC4* with increasing cancer stage. Trend toward improved patient survival occurred with increased expression of MUC4.	(Giuntoli et al. 1998)
	Protein (Immunohistochemistry)	MUC4 expression detected in ovarian cancer samples but not or weakly detected in normal controls.	(Chauhan et al. 2006)
Prostate cancer	Protein from tissue (immunohistochemistry) mRNA from cell lines (semi-quantitative RT PCR)	Expression of MUC4 significantly down regulated in prostate cancer tissues and cancer cell lines vs. normal/benign tissue and normal prostate epithelial cell lines.	(Singh et al. 2006)
Barrett's oesophagus (pre-cancerous)	mRNA (*in-situ* hybridisation)	Upregulation of *MUC4* and *MUC1* expression in severely dysplastic and neoplastic oesophageal tissue. Possible use as markers of early progression to oesophageal cancer.	(Arul et al. 2000)
	mRNA (semi-quantitative RT PCR)	*MUC4* levels were significantly higher in high-grade intraepithelial neoplasia and adenocarcinoma, than in Barrett's oesophagus, which was in turn lower than in normal squamous epithelium.	(Bax et al. 2004)
Cervical Dysplasia (pre-cancerous)	mRNA (*in-situ* hybridisation) and protein (immunohistochemistry, western blots)	MUC4 was strongly detected in dysplastic cervical epithelia but was absent in normal stratified cervical epithelium.	(Lopez-Ferrer et al. 2001a)
Crohn's disease	mRNA (dot-blots)	Expression levels of *MUC3*, *MUC4*, and *MUC5B* significantly lower in both healthy and involved ileal mucosa of patients with Crohn's disease compared to controls.	(Buisine et al. 1999a)
Cap polyposis	mRNA (*in-situ* hybridisation)	*MUC4* was overexpressed in cap polyposis compared to normal.	(Buisine et al. 1998a)
Chronic sinusitis	mRNA (semi-quantitative RT PCR)	Upregulation of *MUC4*, *MUC5AC*, *MUC5B*, *MUC7*, and *MUC8* compared to normal.	(Jung et al. 2000)
Black-pigment gallstones	mRNA (*In-situ* hybridisation)	Presence of *MUC2*, *MUC4*, and increased expression of *MUC1*, *MUC3*, *MUC5B* and *MUC6* in black stone-containing gallbladders vs. gall bladders with no stones.	(Lee and Liu 2001b)

Table 1-2 – Survey of literature reporting evidence of a change in *MUC4* expression in disease vs. normal states.

Disease	Type of Evidence	Observation and Conclusions	Reference
Asthma	Protein (immunohistochemistry)	In mild asthmatics, large amounts of MUC5B, but not MUC5AC, positive extracellular mucus was found in the airway lumen as plugs. Similar MUC5B positive mucus plugs were not detected or observed in normal controls.	(Groneberg et al. 2002)
	Protein (western blots)	Significant increase in MUC5B and MUC5AC in sputum from asthmatics vs. normal controls.	(Kirkham et al. 2002)
Chronic obstructive pulmonary disease (COPD)	Protein (immunohistochemistry)	COPD specifically associated with increased expression of MUC5B in bronchiolar lumen and increased expression of MUC5AC in bronchiolar epithelium vs. normal controls.	(Caramori et al. 2004)
Chronic rhinosinusitis (CRS)	mRNA (semi-quantitative RT PCR)	Expression levels of *MUC5B* and *MUC5AC* in CRS significantly increased compared to in normal sinus mucosa.	(Kim et al. 2004)
	Protein (ELISA)	MUC5B secretion was significantly upregulated in CRS compared with normal controls.	(Viswanathan et al. 2006)
	Protein (immunohistochemistry)	MUC5AC and MUC5B expression was significantly increased in the sinus mucosa of CRS patients vs. normal sinus mucosa.	(Ding and Zheng 2007)
Cystic fibrosis (CF)	Protein (immunohistochemistry)	In epithelium of patients with cystic fibrosis, MUC5B and MUC5AC-staining was significantly increased compared to normal controls. Staining for MUC5B and MUC5AC also increased in airway lumens.	(Burgel et al. 2007)
Nasal polyps (NP)	mRNA (*in-situ* hybridisation) and protein (immunohistochemistry)	*MUC5B* expression raised in bilateral NP, NP from CF patients and antrochoanal (AC) polyps vs. normal nasal mucosa.	(Martinez-Anton et al. 2006)
Chronic otitis media (COM) and mucoid otitis media (MOM)	mRNA (northern blots, *in-situ* hybridisation) and protein (western blots, ELISA)	*MUC5B* and *MUC4* were upregulated 4.2- and 6-fold, respectively, in middle ears with COM or MOM. Upregulation of mucin genes was accompanied by increase of MUC5B and MUC4 producing cells in middle ear mucosa.	(Lin et al. 2001)
Mucoid otitis media (MOM)	mRNA (northern blots, *in-situ* hybridisation) and protein (western blots, ELISA)	5-fold and 6-fold increase in MUC5B and MUC4 expression respectively in middle ears with MOM vs. normal controls.	(Lin et al. 2003)
Otitis media with effusions (OME)	mRNA (semi-quantitative RT PCR)	*MUC5B* and *MUC5AC* expression increased 2-3 folds in OME vs. normal controls.	(Elsheikh and Mahfouz 2006)

Disease	Type of Evidence	Observation and Conclusions	Reference
Lung cancer	mRNA (slot blots)	*MUC5B* was overexpressed in cancerous vs. normal tissue. Overexpression of *MUC5B* correlated with increased risk of post-operative relapse.	(Yu et al. 1996)
Endometrial cancer	mRNA (slot blots)	Endometrial tumours showed increased expression of *MUC1*, *MUC5B* and *MUC8* over normal tissues.	(Hebbar, Damera, and Sachdev 2005)
Breast cancer	Protein (immunohistochemistry)	MUC5B was overexpressed in breast cancer but was generally undetected in normal breast epithelium.	(Sonora et al. 2006)
Gallstones	mRNA (*in-situ* hybridisation)	Stronger and more extensive expression of *MUC5B, MUC3* and *MUC6* in stone-containing gallbladders vs. normal gallbladder.	(Lee and Liu 2002)

Table 1-3 - Survey of literature reporting evidence of a change in *MUC5B* expression in disease vs. normal states.

1.2 CONTROL OF GENE EXPRESSION

Control of gene expression is clearly central to the generation of tissue and developmental specificity as well as to disease related changes such those as seen in the previous section. This section describes the mechanisms that are responsible for the regulation of gene expression.

1.2.1 Restriction of Gene Expression

1.2.1.1 Spatial Restriction

The so-called housekeeping genes encode products that are required for key cellular processes such as translation and metabolism and are found in all cells of the body (Watson JD et al. 1965). Other genes exhibit spatial restriction and show differential expression depending on their location. This leads to phenomena such as tissue-specific and intracellular-specific gene expression (Durkin et al. 1997; Hughes 1997; Lauffart et al. 2006; Saito-Hisaminato et al. 2002).

The *MUC* genes exhibit an extensive range of cellular and tissue-specific expression. *MUC1* is perhaps the most widely expressed mucin. It is found in tissues as

diverse as the stomach, ocular surface and the reproductive tract (Gipson et al. 1997; Ho et al. 1993; Inatomi et al. 1995; Watanabe 2002). Other mucins have a more limited range of expression. *MUC7* for example, is mainly found on epithelial surfaces of the respiratory system (Biesbrock, Bobek, and Levine 1997).

1.2.1.2 Temporal Restriction

Temporal restriction of genes is also a common occurrence: this includes developmental stage-specific, cell differentiation stage-specific and cell cycle stage-specific expression (Cox and Hirsh 1985; Dutcher and Hartwell 1983; Gober et al. 1991; Kok et al. 1993; Porter et al. 2005).

The most studied form of temporal restriction in the *MUC* genes is that of developmental-stage specific expression. Various studies have looked at the expression of mucins in different organ systems during development (Buisine et al. 1998b; Buisine et al. 1999b; Buisine et al. 2000; Reid, Gould, and Harris 1997; Reid and Harris 1998). They compared the expression of various *MUC* genes in tissue originating from fetuses of different gestational ages as well as tissue from adults. Certain of the *MUC* genes, such as *MUC5AC* in the trachea (Reid, Gould, and Harris 1997) and *MUC4* in the colon (Reid and Harris 1998) were expressed in fetuses but were subsequently absent in the same tissue taken from adults. In other cases, such as *MUC4* in the trachea, there was the opposite pattern, with a gradual increase in expression level that eventually reaches a maximal level in adults (Buisine et al. 1999b).

1.2.2 Levels of Gene Regulation

Regulation of mammalian gene expression occurs at two broad levels. These are:

- **Transcriptional regulation**, which can be further subdivided into the following:
    - **Epigenetic mechanisms and long range control of gene expression by chromatin structure**, which are methods of gene regulation not attributed directly to DNA sequence. These include processes such as

histone acetylation and methylation of genes (Robertson 2002; Sharma et al. 2005; Sproul, Gilbert, and Bickmore 2005). These mechanisms affect chromatin structure, physically controlling access of the transcription machinery to genes.

- **Control of transcription initiation**, which involves the binding of transcription factors to DNA regulatory sequences (Maston, Evans, and Green 2006; Villard 2004). The vast majority of gene regulation is believed to occur at this level (Mitchell and Tjian 1989).

- **Post-transcriptional regulation**, which includes processes such as alternative splicing, mRNA transport and stability, and translation efficiency (Audic and Hartley 2004; Bevilacqua et al. 2003; Blobe et al. 1993; Day and Tuite 1998; Ren and Stiles 1994; Shim and Karin 2002).

The focus of this project is on variation in transcription regulation of the *MUC* genes. Hence, the other levels of gene regulation will not be discussed in detail.

1.2.3 An Overview of the Control of Transcription Regulation – *Cis* vs. *Trans*-acting

Eukaryotic transcription is initiated by the assembly of a transcription preinitiation complex (PIC), consisting of RNA polymerase II and other general transcription factors, such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (see Figure 1-4) (Conaway and Conaway 1993; Tang et al. 1996). The complex directs RNA polymerase II to the transcription site start and is responsible for the basal transcription (low-level expression) of a gene.

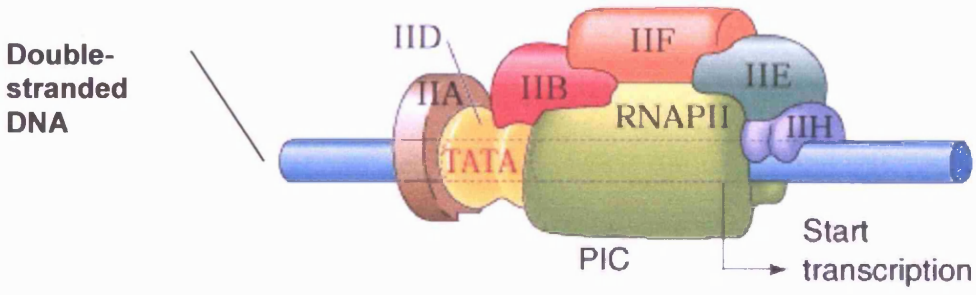


Figure 1-4 – The Transcription Preinitiation Complex (PIC). Adapted from (Zawel and Reinberg 1992)

The basal expression levels of a gene can be increased or decreased by the binding of sequence-specific transcription factors (activators or repressors, respectively) to various regulatory sequence elements, such as proximal promoter elements, enhancers and silencers (Villard 2004). Thus, transcriptional regulation is based on controlling the binding of *trans*-acting transcription factors to *cis*-acting regulatory sequences:

Transcription factors are usually transcribed from a distantly located gene(s), after which they migrate to their site of action. Since these factors act on distantly located targets, they are said to be *trans*-acting. Indeed, these factors can potentially act on multiple regulatory sequences situated at a great distance from their site of origin as well as on both homologues of a target gene(s) (Griffiths et al. 1999; Strachan and Read 2000)

Conversely, the regulatory elements they bind to tend to be located within the vicinity of the gene being regulated. These sequences are physically on the same DNA molecule as the gene being regulated and only influence the expression of that gene and indeed that homologue. Hence, these regulatory elements are said to be *cis*-acting (Griffiths et al. 1999; Strachan and Read 2000).

1.2.4 Classes of *Cis*-acting Regulatory Elements involved in Transcriptional Regulation

The various classes of *cis*-acting regulatory elements are briefly discussed in this section.

1.2.4.1 Core Promoter Elements

Core promoters are sequence elements on which the PIC assembles and are therefore found in the immediate vicinity of the transcription start site (Smale and Kadonaga 2003). Figure 1-5 shows some of the core promoter elements found in humans.

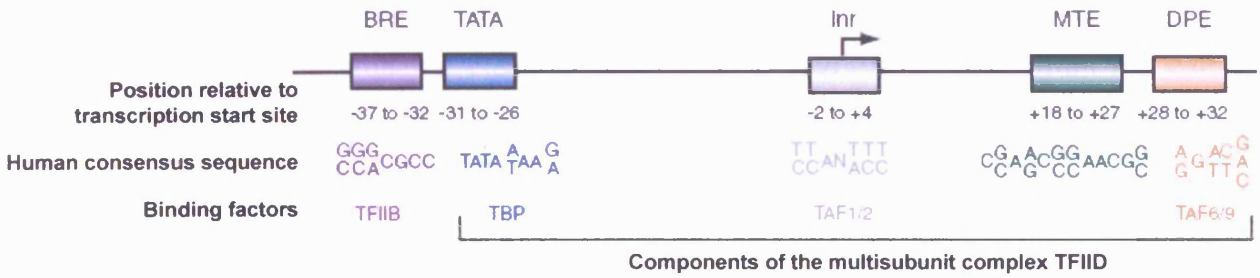


Figure 1-5 – Core Promoter Elements. Adapted from (Maston, Evans, and Green 2006)

- **BRE (TFIIB recognition elements)** are recognised by TFIIB (see Figure 1-5). In comparison, the other core promoter elements described to date are TFIID interaction sites (Smale and Kadonaga 2003). The BRE was originally identified as an element that had a positive effect on transcription *in vitro*, using purified basal factors (Lagrange et al. 1998). It has since been shown *in vitro* using crude nuclear extracts and *in vivo* to repress basal transcription, thereby increasing the amplitude of transcriptional stimulation in the presence of an activator (Evans, Fairley, and Roberts 2001).
- **TATA (TATA Boxes)** are short and highly conserved sequences comprised solely of Thymines and Adenines. These are located approximately 26-31 bps upstream of the start site and are bound by the TATA-box-binding protein (TBP) subunit of TFIID (Thomas and Chiang 2006).
- **Inr (Initiators)** are found in some eukaryotic genes as an alternative to TATA boxes (Smale and Baltimore 1989; Weis and Reinberg 1992). In comparison to TATA boxes, these are much less conserved.
- **MTE (Motif Ten Elements)** are a recently discovered class of core promoter element. The MTE can promote transcription by RNA polymerase II in conjunction with the Inr but independently of the TATA-box or DPE and can therefore compensate for the loss of either of the two elements (Lim et al. 2004).
- **DPE (Downstream Promoter Elements)** are, as the name suggests, located 28-32 bps *downstream* of the transcription start site (Burke and Kadonaga 1997). These are thought to be the downstream counterparts of TATA boxes and appear to be conserved from drosophila to humans. This observation suggests an important role in basal transcription.

1.2.4.2 Proximal Promoter Elements

Proximal promoter elements tend to be located 100-200 bps from the start site (Anthony J.F.Griffiths et al. 1999; Griffiths et al. 1999). Notable sequence elements falling under this classification include:

- **GC Boxes**, which have the consensus sequence GGGCGG and are usually found within 100 bps of the start site. These bind the Sp1 transcription factor and modulate the basal transcription of the core promoter (Li et al. 2004).
- **AP1 (Activator Protein 1) Sites** have the consensus sequence TGAC and bind the transcription factor Activator Protein 1. AP1 is a heterodimer of c-Jun and c-Fos both of which are also transcription factors.AP1 has been shown to play an important role in signalling, cell proliferation and apoptosis (Karin, Liu, and Zandi 1997).

1.2.4.3 Enhancers

Enhancers are sequence elements that *enhance* the basal transcription level of genes (Khoury and Gruss 1983; Serfling, Jasin, and Schaffner 1985). They differ in terms of their location from proximal promoter elements that also up-regulate gene transcription (i.e. activator-binding sites). Whereas proximal promoter elements are always upstream of the transcription initiation site, enhancers can theoretically be found at a great distance from the genes they affect. They can be located upstream of, downstream of or even within the genes themselves (Lettice et al. 2003). Unlike core promoter elements, enhancers can exert their influence irrespective of their orientation (Blackwood and Kadonaga 1998).

1.2.4.4 Silencers

Silencers function in the opposite way to enhancers – they decrease the basal transcription level of genes, thereby *silencing* their expression (Brand et al. 1985). They are thus repressor-binding sites. They share similar characteristics in terms of location and orientation, although some position-dependent silencers, known as Negative Regulatory Elements (NRE) have been discovered (Ogbourne and Antalis 1998).

1.2.4.5 Response Elements

Response elements are an interesting class of *cis*-acting regulatory element. As their name implies, these elements allow the level of transcription of a gene to be modulated in *response* to specific external stimuli. They tend to be situated within 1 kb upstream of the start site (Kobayashi et al. 1997; Strachan and Read 2000). Some examples of response elements include those that activate transcription in response to specific hormones such as glucocorticoids and growth hormones and intracellular secondary messengers such as cyclic AMP (Bergad, Towle, and Berry 1999; Hayashi et al. 2004; Verma, Blass, and Davidson 1997).

1.2.5 Transcriptional Regulation of *MUC4* and *MUC5B* Expression

This section summarises current knowledge on the promoters of *MUC4* and *MUC5B* and reviews the various studies so far on the regulation of *MUC4* and *MUC5B* expression using experimental systems such as ALI cultures. Schematic diagrams of the promoters, with the relative positions of putative *cis*-acting regulatory sequences (described in section 1.2.4), are shown on the following pages.

1.2.5.1 *MUC4* Promoter

A schematic diagram of the *MUC4* promoter is shown in Figure 1-6. This diagram was constructed using information from the work of Perrais and colleagues (Perrais et al. 2001b).

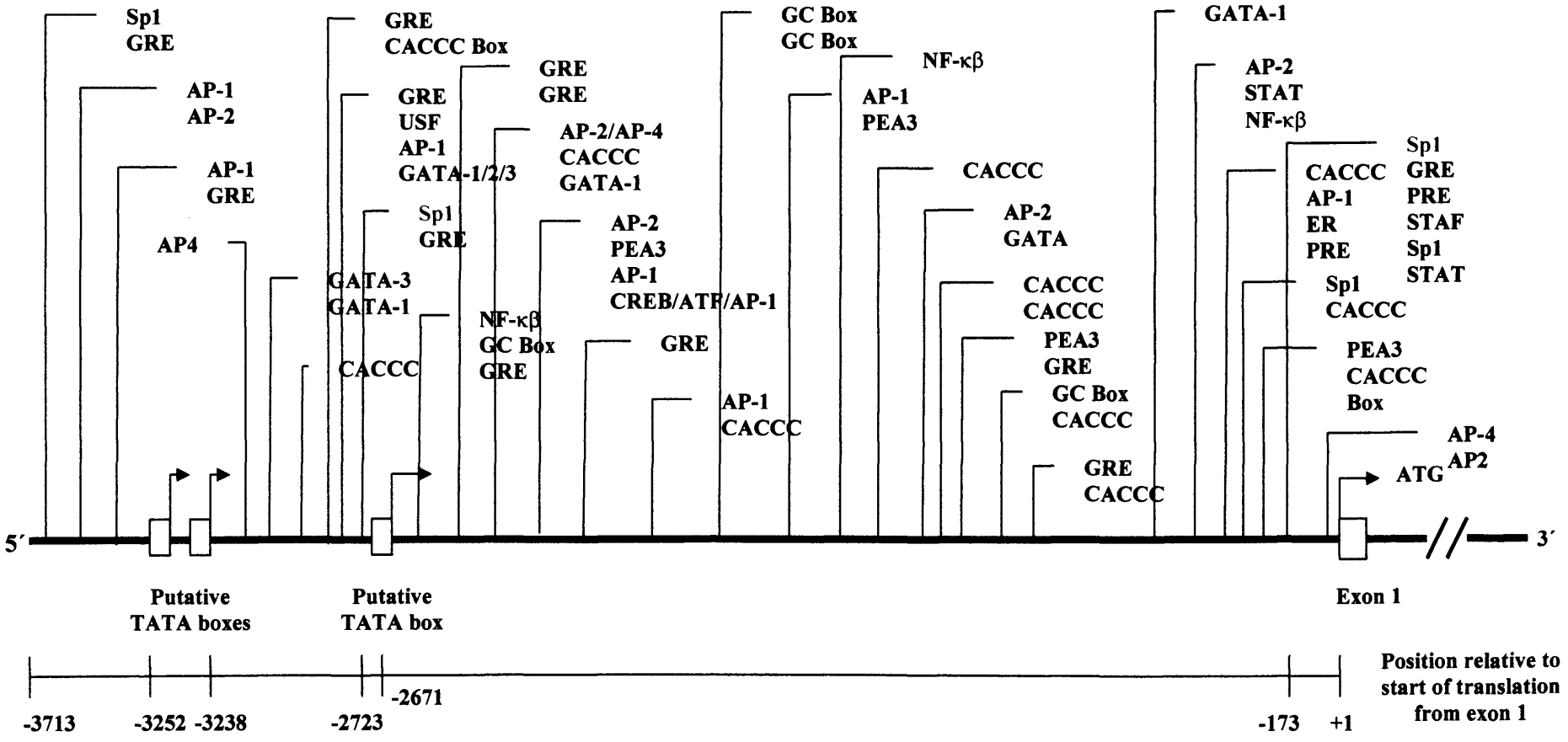


Figure 1-6 - Schematic diagram of *MUC4* promoter. Positions of putative transcription factor binding sites (*cis* regulatory elements, determined computationally) relative to start of translation from exon 1. Sites in red have been validated experimentally to bind their associated transcription factors (Perrais et al. 2001b) Diagram not to scale.

1.2.5.2 *MUC5B* Promoter

A schematic diagram of the *MUC5B* promoter is shown in Figure 1-7. This diagram was constructed using information primarily from the work of Van Seuningen and colleagues (Chen et al. 2001; Perrais et al. 2001a; Van, I et al. 2000).

A notable feature of the *MUC5B* promoter region is the presence of a putative distal TATA box, in addition to the proximal TATA box first described (Van, I et al. 2000). This suggested that *MUC5B* transcription could be initiated from two different locations. In order to address this, Perrais and colleagues performed primer extension experiments using a reverse primer located 131 bps downstream of the distal TATA box (Perrais et al. 2001a). They successfully obtained a specific extension product of 109 bps, corresponding to a transcription start site located 23 bps downstream of the distal TATA box, implying that transcription could indeed be initiated from the distal TATA box.

To complement this finding, RT PCR experiments on total RNA from KATO III gastric cancer cells were carried out, using a pair of primers that encompassed the distal TATA box, the proximal promoter and the proximal transcription start site. PCR products of the expected sizes were obtained, showing that the entire region was transcribed and that both distal and proximal TATA boxes were active.

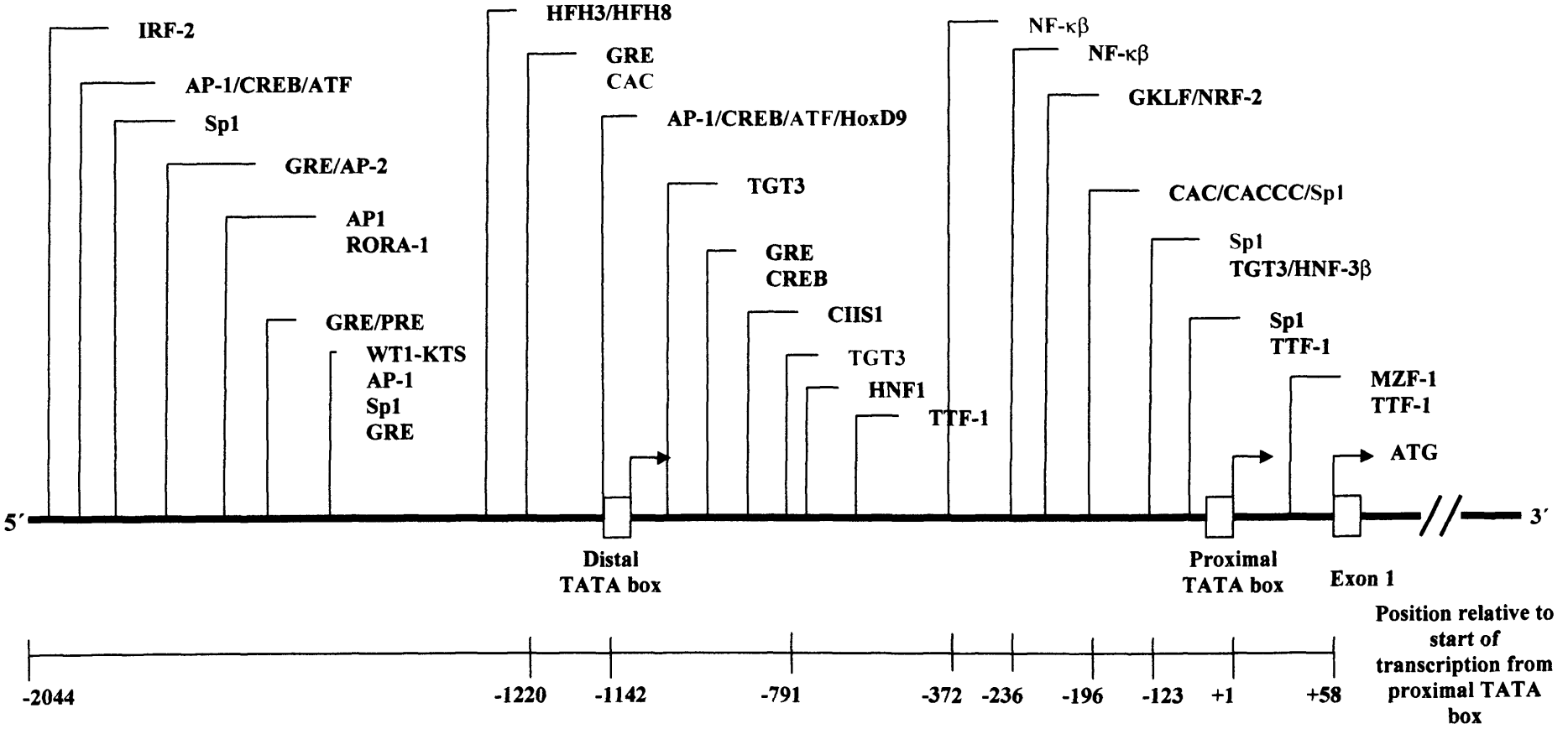


Figure 1-7 - Schematic diagram of *MUC5B* promoter. Positions of putative transcription factor binding sites (*cis*-acting regulatory elements, determined computationally) relative to start of transcription from proximal TATA box are shown. Sites in red have been validated experimentally to bind their associated transcription factors (Perrais et al. 2001a; Van, I et al. 2000)). Diagram not to scale.

1.2.5.3 Evidence of changes in *MUC4* and *MUC5B* Expression in Response to Inflammatory Mediators

Numerous studies have investigated the transcriptional regulation of *MUC4* and *MUC5B* expression, often by using Normal Bronchial Epithelial Cells (NHBE) grown at the Air Liquid Interface (ALI) as a model system. These have been used widely for the study of respiratory epithelium because cells grown in this manner have been shown to retain many morphological and functional properties of *in vivo* airway epithelial cells (Gray et al. 1996) and have been shown to express mucins (Bernacki et al. 1999). Such cells can be treated various inflammatory mediators and the change in mucin expression studied.

Examples of experimental evidence of changes in *MUC4* and *MUC5B* expression in response to inflammatory mediators are summarised in Table 1-4 and Table 1-5 respectively.

Inflammatory Mediator Tested	Rationale	Experimental system used	Conclusions from Study	Reference
Estrogen Dexamethasone Progesterone	Steroid hormones.	mRNA levels quantified by slot blotting using radiolabelled oligonucleotide probes and NHBE cells.	*MUC4* expression levels raised 3-4 fold by estrogen, 1.5-2 fold by dexamethasone. No effect with progesterone but interferes with upregulation by estrogen.	(Gollub et al. 1995)
Tumour necrosis factor-alpha (TNF-α) Interferon gamma (INF-γ)	Proinflammatory cytokines	*MUC4* Transfection assays using CAPAN-1 and 2 pancreatic cell lines.	Mild effect when TNF-α or INF-γ used alone. Synergistic effect (10-12 fold activation) when used together or when INF-γ used with transforming growth alpha (TGF-α) in CAPAN-2 cells.	(Perrais et al. 2001b)
Neutrophil Elastase	Inflammatory protease. In COPD, airway epithelium permanently exposed to neutrophil elastase.	Steady state mRNA levels in NHBE cells measured by RT PCR. MUC4 protein levels measured by western blotting.	Neutrophil elastase increased *MUC4* expression in time and dose dependent manner, and raised MUC4 glycoprotein levels.	(Fischer et al. 2003)
Bile Acids	Abnormal gastro-oesophageal reflux and bile acids have been linked to the presence of Barrett's oesophageal premalignant lesion associated with an increase in mucin-producing goblet cells and *MUC4* gene overexpression.	Luciferase reporter assays containing *MUC4* promoter sequences.	Taurocholic, taurodeoxycholic, taurochenodeoxycholic and glycocholic bile acids and sodium glycocholate strongly activate *MUC4* expression.	(Mariette et al. 2004)

Inflammatory Mediator Tested	Rationale	Experimental system used	Conclusions from Study	Reference
Interleukin-4 (IL4)	Inflammatory cytokines	Real-time PCR using mRNA from NCI-H650 cell lines. MUC4 glycoprotein levels quantified by western blotting.	IL4 and IL9 upregulate *MUC4* mRNA and protein levels in a time and concentration dependent manner.	(Damera, Xia, and Sachdev 2006)
Interleukin-9 (IL9)				(Damera et al. 2006)
IL-1 beta Lipopolysaccharide (LPS)	Previously shown to induce *MUC2* and *MUC5AC* expression (Kim et al. 2002b).	Level of *MUC4* mRNA measured by RT PCR and the amount of the MUC4 mucin glycoprotein estimated by ELISA using cultured nasal polyps obtained from chronic rhinosinusitis patients.	IL-1 beta and LPS upregulate *MUC4* expression at mRNA and protein levels.	(Bai, Song, and Kim 2007)

Table 1-4 - Examples of studies on changes in *MUC4* expression in response to inflammatory mediators

Inflammatory Mediator Tested	Rationale	Experimental system used	Conclusions from Study	Reference
Bacterial Lipopolysaccharides (LPS)	Bacterial inflammation in mucosa is accompanied by morphological and proliferative changes in goblet cells and mucin hypersecretion	RT PCR and ELISA using total RNA and secretions from HT29-MTX cells, a human colon carcinoma derived, mucin-secreting goblet cell line	LPS (100 ng/ml) increased *MUC5B* mRNA expression 2.1 fold and stimulated a 31% increase in MUC5B secretion.	(Smirnova et al. 2003)
Acrolein	Acrolein is an aldehyde in tobacco smoke that could cause/exacerbate respiratory diseases such as bronchitis and asthma, where there is mucus hypersecretion.	Steady state mRNA levels in NCI-H292 human lung cancer cells measured by RT PCR	*MUC5B* expression unaffected by exposure, suggesting *MUC5B* is constitutively expressed in cells tested.	(Borchers, Carty, and Leikauf 1999)
prostaglandin E2 TNF-alpha	Eicosanoid inflammatory mediators associated with acrolein exposure			
Interleukin 4 (IL-4)	IL-4 was shown to dramatically upregulate 15-LO expression and production of 15-LO metabolites (Jayawickreme et al. 1999), which is in turn thought to stimulate mucin secretion (Marom et al. 1983)	RT PCR using total RNA extracted from NHBE cells in ALI culture	IL-4 unexpectedly reduced *MUC5B* expression. Results from Marom and colleagues may have differed because bronchial explants in their studies contained not only a variety of mesenchymal cells but also inflammatory cells, which might when exposed to 15-LO metabolites stimulate the release of mucin.	(Jayawickreme et al. 1999)

Inflammatory Mediator Tested	Rationale	Experimental system used	Conclusions from Study	Reference
Human neutrophil defensins (human neutrophil peptides 1-3 [HNP1-3])	Known to induce airway epithelial cell proliferation and therefore might play important role in wound repair	*MUC5B*mRNA levels in NCI-H292 human lung cancer cells measured by RT PCR	HNP1-3 increased *MUC5B* mRNA expression, suggesting a role for defensins in mucous cell differentiation	(Aarbiou et al. 2004)
phorbol 12-myristate 13-acetate (PMA)	Known inflammatory mediator (protein kinase C activator), which initiates signalling cascades that lead to a variety of cell events, including gene expression, cell differentiation, cell proliferation, and respiratory secretions.	Steady state mRNA levels in NCI-H292 human lung cancer cells measured by RT PCR	*MUC5B* expression unaffected by exposure, suggesting *MUC5B* is constitutively expressed in cells tested.	(Borchers, Carty, and Leikauf 1999)
		ALI cultures using NHBE cells, an immortalized NHBE cell line HBE1, and human lung cancer cell line A549. Real-time RT PCR and ELISA performed.	PMA induced *MUC5B* mRNA in all primary NHBE, HBE1, and A549 cells in a time and dose dependent manner. MUC5B secretions also increased.	(Yuan-Chen et al. 2007)

Table 1-5 - Examples of studies on changes in *MUC5B* expression in response to inflammatory mediators. Discrepancy in results between studies on PMA and its effect on *MUC5B* expression might be due to different cell lines used.

1.3 ALLELIC VARIATION IN GENE EXPRESSION

Allele-specific differences (allelic variation) in levels of gene expression are known to exist in humans. The following sections discuss this phenomenon in detail:

1.3.1 Causes of Allelic Variation in mRNA Expression

1.3.1.1 Genomic Imprinting and X-chromosome Inactivation

Classical Mendelian inheritance is obeyed when genes located on paternally and maternally inherited chromosomes are equally expressed. This is now understood to be an overly simplistic view of what happens in reality and that exceptions exist.

For example, genomic imprinting results in monoallelic expression of a gene, depending on which parent that allele was inherited from (Brannan and Bartolomei 1999). Examples of imprinted genes include insulin-like growth factor-2, IGF2, which exhibits parental imprinting, and PEG1, which is paternally imprinted in fetuses, whereas both alleles are expressed in adults (Ekstrom 1994; Kobayashi et al. 1997; Reik et al. 2000; Riesewijk et al. 1997).

Another example of monoallelic expression is that of X-chromosome Inactivation, which results in the random inactivation of one of the two X chromosomes in the cells of a female (Lyon 1999; Migeon 1994). This provides a mechanism for controlling gene dosage and compensates for the fact that males have a single X chromosome compared to females (Lyon 1989).

1.3.1.2 Regulatory Polymorphisms

DNA polymorphisms are defined as DNA sequence variations that occur at a frequency of at least 1% in a particular population. The most frequent type of polymorphism is the Single Nucleotide Polymorphism or Simple Nucleotide

Polymorphism (SNP), which is so called because it involves a single-base change (Human Genome Project Information 2006). These are thought to occur very often throughout the human genome, at a typical rate of one SNP per 1000 bases (Zhao et al. 2003). Other polymorphisms include various types of repeat number variations (VNTR) such as mini and microsatellites, as well as insertions and deletions (indels).

Broadly speaking, polymorphisms fall under two main categories: coding and regulatory. Coding polymorphisms occur in the coding regions of genes and are therefore limited to the exons. These can cause non-synonymous amino acid changes, frame shifts and premature stop codons, and have been the focus of most genetic studies concerning association with function (Cargill et al. 1999). This is because of the possibility of such changes in amino acid sequence affecting the structure and consequently, the function of a protein of interest.

On the other hand, regulatory polymorphisms are located within the various non-coding *cis* elements described in section 1.2.4. These variations can affect the complex interactions between *trans*-acting factors and *cis*-acting regulatory sequences, thus affecting transcriptional regulation. This could in turn influence the level of gene expression.

Regulatory polymorphisms thus give rise to an added degree of sophistication to the expression of genes, allowing control of the quantity of encoded proteins produced by an individual, in addition to the quality, as determined by differences in amino acid sequence, caused by coding polymorphisms. Thus, differences in gene expression levels could potentially account for a major part of the variation found between individuals and indeed in between species (Rodriguez-Trelles, Tarrio, and Ayala 2003).

1.3.2 Methods for Detecting Allelic Variation in mRNA Expression due to *Cis*-acting Regulatory Polymorphisms

1.3.2.1 *In Vitro* Methods

In vitro approaches most commonly involve reporter assays (Alam and Cook 1990). These monitor the transcriptional activity of a synthetic reporter construct using a quantifiable reporter protein (such as luciferase) and usually involve the putative promoter and upstream regions of the gene studied. Such assays can be used to compare the relative transcriptional activities of allele-specific constructs, revealing the presence of allelic variation (Coleman et al. 2002).

An alternative *in vitro* approach for detecting allelic variation utilises Electrophoretic Mobility Shift Assays (EMSA) (Kako et al. 1998). Each allele in an individual heterozygous for the gene of interest is used as a probe. These are tagged with radioactive or fluorescent labels to allow detection. The probes are then incubated with nuclear extracts from relevant cell types and then analysed on polyacrylamide gels. Non-specific binding is controlled for by introducing a non-specific competitor DNA at increasing concentrations. Protein-bound DNA will migrate more slowly through the gel and show up as a higher molecular weight band (called a 'shift') compared to unbound DNA. The intensity of this band indicates the stability of the binding. Thus, the demonstration of differential transcription binding between the two alleles indicates the presence of allelic variation (Mottagui-Tabar et al. 2005).

1.3.2.2 *In Vivo* Methods

The general strategy used for quantifying relative allelic expression *in* vivo is depicted in Figure 1-8. In this example, the G allele results in strong binding of a *trans*-acting regulatory protein such as a transcription factor to the *cis*-acting DNA regulatory element. This results in high expression. Conversely, the A allele results in weaker binding of the regulatory protein and results in lower expression.

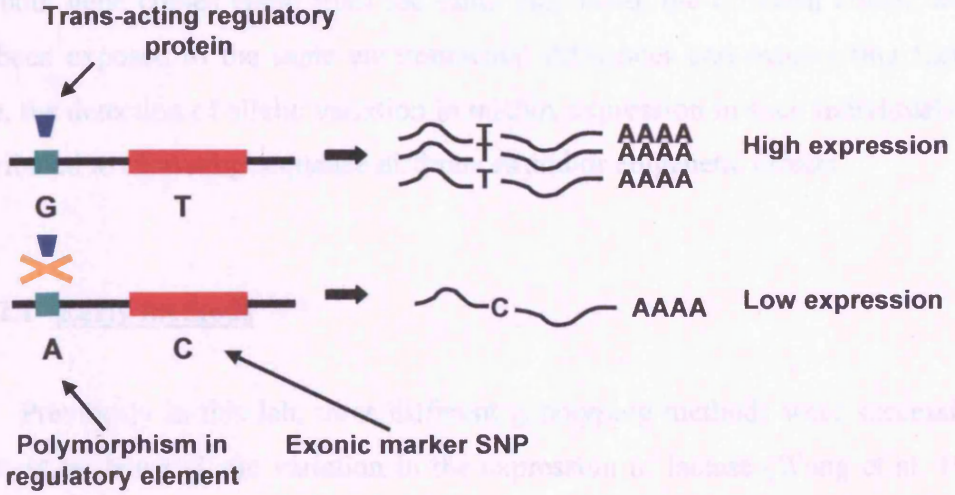


Figure 1-8 - Strategy for detecting presence/extent of allelic variation in mRNA expression

Individuals heterozygous for the gene being studied are used. A marker SNP is used to distinguish between the two alleles. For most of the following methods, the marker SNP is exonic, because it needs to be transcribed in order to be detected in the processed mRNA.

These methods usually begin with obtaining total RNA from an individual heterozygous for the gene of interest. RT PCR is performed using primers encompassing a region containing the marker SNP. The RT PCR product contains allele-specific PCR fragments, representative of the different relative amounts of allelic transcripts in the starting total RNA. By determining the relative abundance of each allelic transcript, the presence and extent of allelic variation in mRNA expression can be determined.

The key to these methods is to replicate the procedure in gDNA as well as in cDNA, from individuals heterozygous for the gene of interest. The gDNA results act as a reference since the results obtained represent a 1:1 ratio of each allele (important: assuming there is no gene duplication). This in principle corrects for any bias in representation of one allele over another.

The comparison of different alleles within the same individual provides a number of benefits: firstly, both gene copies originate from the same tissue source, which controls for tissue and developmental-stage-specific differences; secondly,

since both gene copies come from the same individual, the different alleles would have been exposed to the same environmental influences and *trans*-acting factors. Hence, the detection of allelic variation in mRNA expression in such individuals can be attributed to *cis*-acting sequence differences and/or epigenetic effects.

1.3.2.2.1 Early methods

Previously in this lab, three different genotyping methods were successfully employed to detect allelic variation in the expression of lactase (Wang et al. 1994; Wang et al. 1995). All the methods generate visible allele-specific bands. A significant difference in relative intensity of the allele-specific bands in cDNA compared to that observed in the genomic DNA indicates that allelic variation is present.

1.3.2.2.1.1 Restriction Fragment Length Polymorphisms (RFLP) and Single-strand Conformation Polymorphism (SSCP)

PCR is performed to obtain a product containing an exonic SNP in gDNA. An enzyme digestion is carried out and the restriction fragments are analysed by gel electrophoresis to identify heterozygous individuals. The process is then repeated using cDNA from these heterozygous individuals.

In Figure 1-9, individuals 4 and 5 show a substantially lower expression of the – allele compared to the + allele in the cDNA. In contrast, individual 6 shows an equal expression of both alleles and therefore a lack of allelic variation.

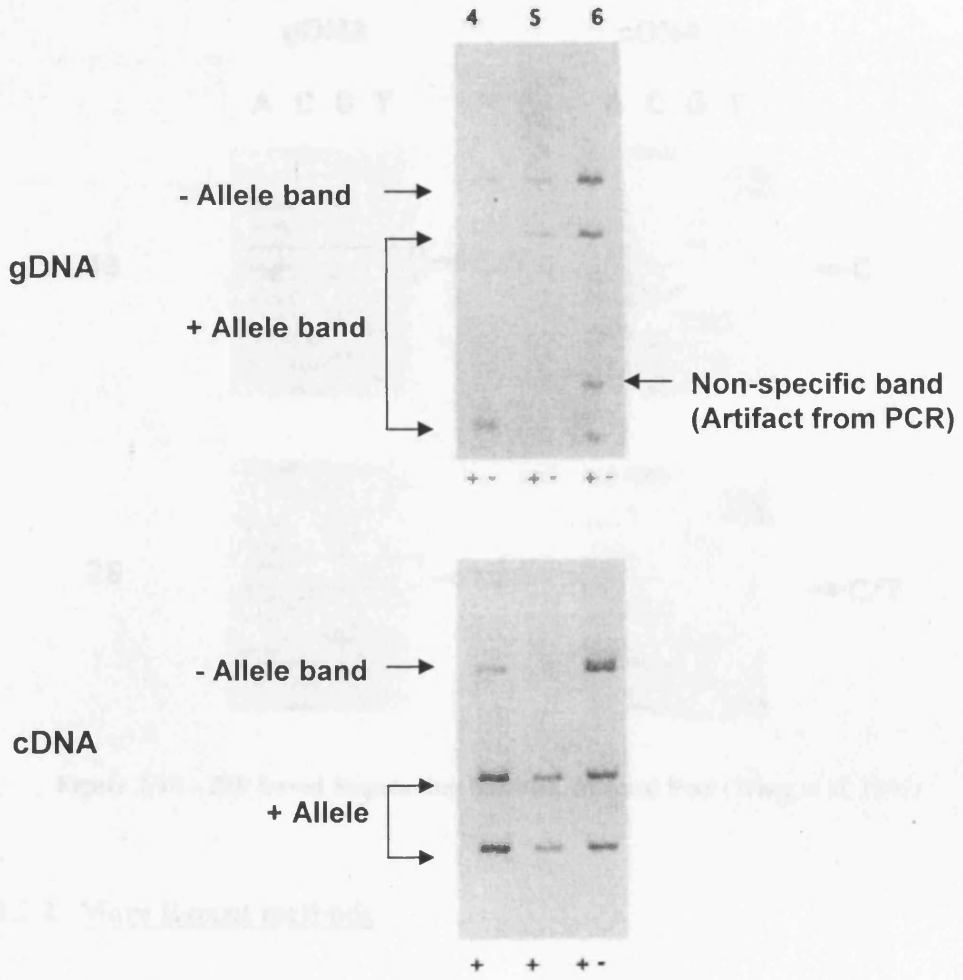


Figure 1-9 – RFLP Method. Adapted from (Wang et al. 1995)

The same approach was also taken for SSCP.

1.3.2.2.1.2 33P-based Sequencing

gDNA and cDNA from individual heterozygous for the gene of interest (lactase) were sequenced by a 33P-based method.

In Figure 1-10 below, individuals 43 and 26 are shown with their respective gDNA and cDNA autoradiographs. Both individuals are heterozygous (CT) at locus 666, as indicated by the blue boxes. The C and T bands are of the same relative intensities. A similar result is seen in the cDNA results for individual 26. In comparison, the cDNA result for individual 43 shows much higher expression of the C allele than T allele, indicating an allelic difference in mRNA expression.

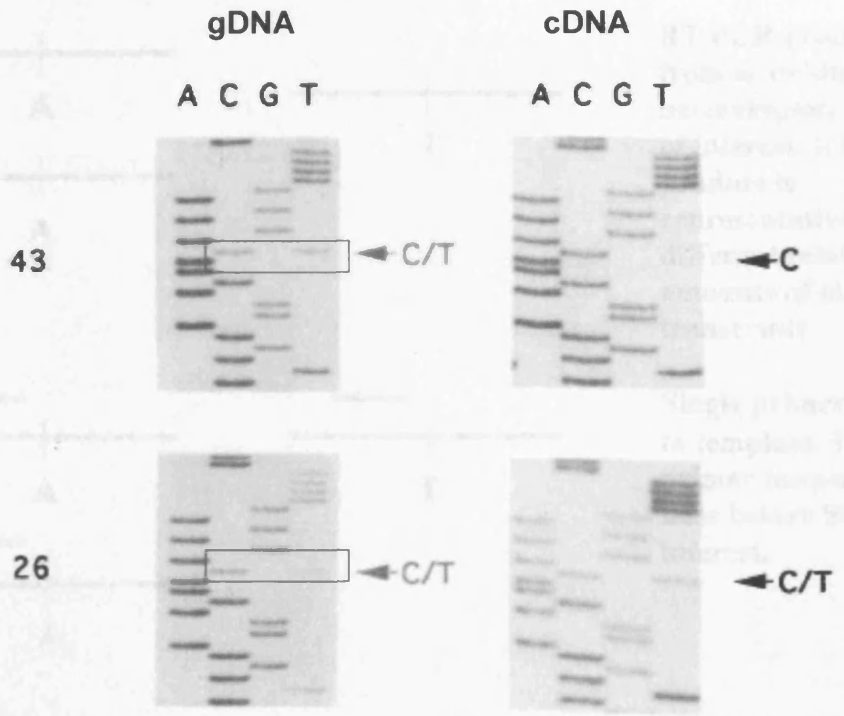


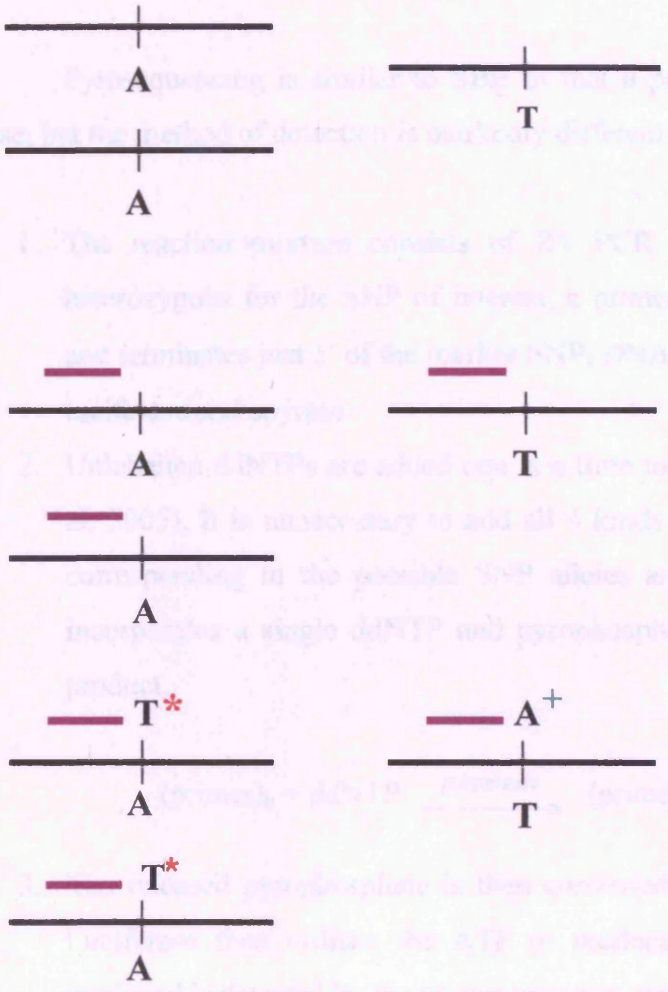
Figure 1-10 – 32P-based Sequencing Method. Adapted from (Wang et al. 1995)

1.3.2.2.2 More Recent methods

1.3.2.2.2.1 Single-base Extension (SBE)

Single-base extension, also known as primer extension or minisequencing, has been used in various allelic variation studies (Bray et al. 2003; Yan et al. 2002).

Figure 1-11 shows how the procedure works.



RT PCR product from individual heterozygous for SNP of interest. RT PCR product is representative of the different relative amounts of allelic transcripts

Single primer anneals to template. 3′ end of primer terminates one base before SNP of interest.

Primer is extended by a single nucleotide using ddNTPs. ddNTPs are labelled with different fluorescent dyes.

Relative abundance of each allelic transcript determined from strength of fluorescent signals.

T allele expression > A allele expression

Figure 1-11 – SBE Method

1.3.2.2.2.2 Pyrosequencing

Pyrosequencing is similar to SBE in that a primer is extended by a single base, but the method of detection is markedly different (Ahmadian et al. 2000).

1. The reaction mixture consists of RT PCR products from an individual heterozygous for the SNP of interest, a primer that anneals to the sequence and terminates just 5′ of the marker SNP, DNA polymerase, ATP sulfurylase, luciferase and apyrase.
2. Unlabelled ddNTPs are added one at a time to the reaction mixture (Zhou et al. 2005). It is unnecessary to add all 4 kinds of ddNTPs; only the ddNTPs corresponding to the possible SNP alleles are required. DNA polymerase incorporates a single ddNTP and pyrophosphate (PPi) is released as a by-product.

$$(primer)_n + ddNTP \xrightarrow{polymerase} (primer)_{n+1} + PPi$$

3. The released pyrophosphate is then converted to ATP by ATP sulfurylase. Luciferase then utilises the ATP to produce light. The amount of light produced is detected by the pyrosequencing machine.

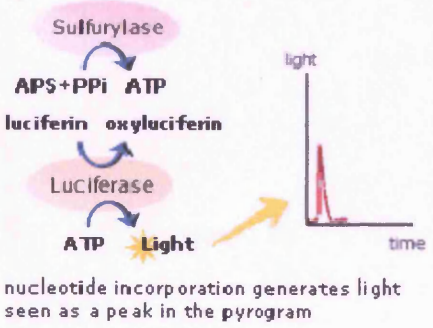


Figure 1-12 – Pyrosequencing Method

4. Excess ATP is removed by apyrase.

$$ATP \xrightarrow{Apyrase} ADP + AMP + phosphate$$

5. The next ddNTP is now added and the detection procedure is repeated. The relative amounts of each allelic transcript can then be inferred by comparing the amount of light produced when each ddNTP was added.

1.3.2.2.2.3 SBE with MALDI-TOF MS

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) in conjunction with SBE has been successfully developed into a high throughput technique for studying allelic variation by Ding et al (Ding and Cantor 2003).

Figure 1-13 shows how this method works. A SBE reaction is carried out, similar to that described in section 1.3.2.2.2.1. However, the crucial difference is that the reaction contains three different ddNTPs, with the remaining base as a dNTP. In addition, none of the nucleotides are dye-labelled. The dNTP chosen must be a base complementary to one of the two SNP alleles (dGTP in Figure 1-13). This modified SBE reaction results in either a single-base or a double-base extension. The different sized allele-specific products can then be resolved by MALDI-TOF MS and quantified.

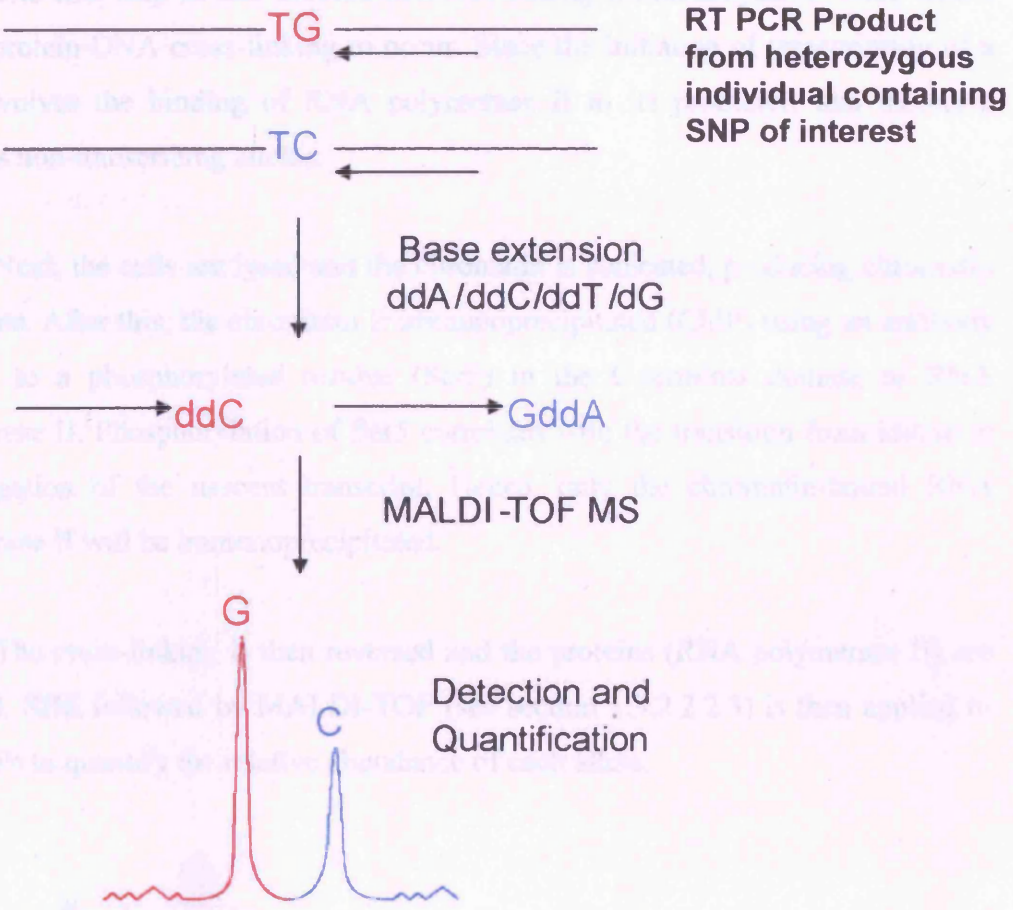


Figure 1-13 – MALDI-TOF MS Method. Adapted from (Ding and Cantor 2003)

1.3.2.2.2.4 HaploChIP

The main limitation of the above methods is the requirement for the marker SNP to be transcribed. It is not always possible to find such a SNP within a gene of interest and this somewhat limits the usefulness of the aforementioned techniques.

The Haplotype-specific chromatin immunoprecipitation (HaploChIP) method overcomes this limitation (Knight et al. 2003). A marker SNP is still required but that SNP in theory can be located anywhere in the gene. The method discriminates between the two alleles based on differential initiation of transcription from genomic DNA.

The first step in this method involves adding formaldehyde to cells, which causes protein-DNA cross-linking to occur. Since the initiation of transcription of a gene involves the binding of RNA polymerase II to its promoter, this therefore excludes non-transcribing alleles.

Next, the cells are lysed and the chromatin is sonicated, producing chromatin fragments. After this, the chromatin is immunoprecipitated (ChIP) using an antibody specific to a phosphorylated residue (Ser5) in the C-terminal domain of RNA polymerase II. Phosphorylation of Ser5 correlates with the transition from initiation to elongation of the nascent transcript. Hence, only the chromatin-bound RNA polymerase II will be immunoprecipitated.

The cross-linking is then reversed and the proteins (RNA polymerase II) are digested. SBE followed by MALDI-TOF (see section 1.3.2.2.2.3) is then applied to the ChIPs to quantify the relative abundance of each allele.

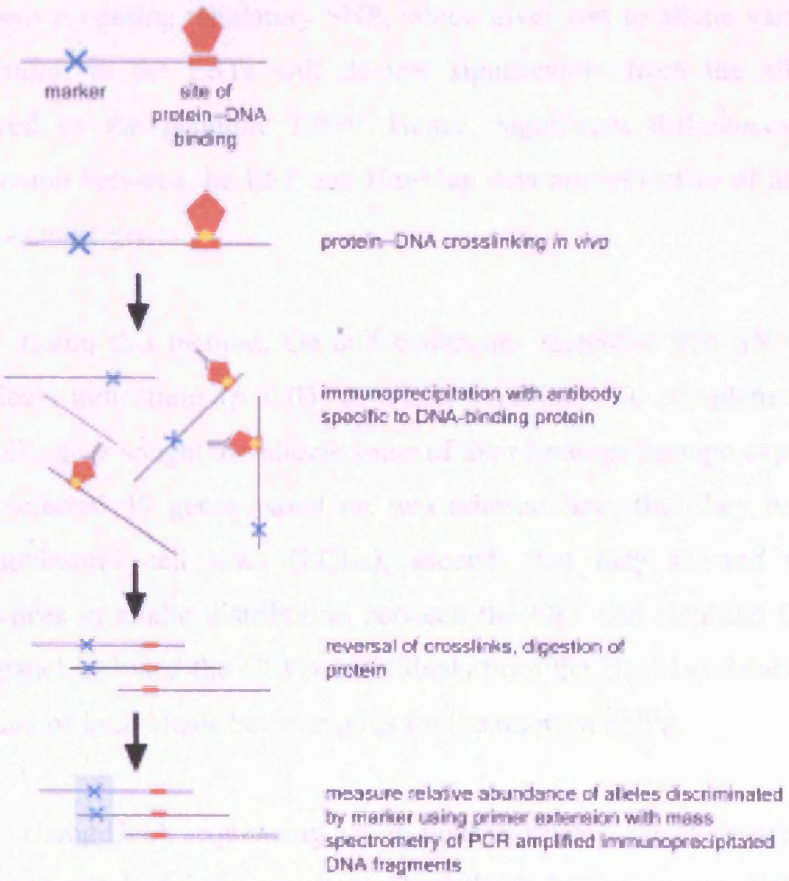


Figure 1-14 – HaploChIP Method (Knight et al. 2003)

1.3.2.3 *In Silico* Methods

The prediction of genes showing allelic variation in mRNA expression has been performed using computers. An example of this *in silico* approach is the work by Ge and colleagues:

They employed a novel approach of using the dbEST and HapMap databases to detect allelic variation (Ge et al. 2005). First, the HapMap database was used to determine the expected allele frequencies for a set of SNPs, which were common in the four distinct populations available from the database. Next, they compared these allele frequencies to the allele counts obtained from dbEST for the set of SNPs. The rationale behind their strategy was as follows:

If a marker SNP in observed ESTs is in linkage disequilibrium with an unknown *cis*-acting regulatory SNP, which gives rise to allelic variation, the allelic distribution in the ESTs will deviate significantly from the allelic distribution observed in the genomic DNA. Hence, significant differences in SNP allele distribution between the EST and HapMap data are indicative of allelic variation in the associated genes.

Using this method, Ge and colleagues identified 976 SNPs which showed significant indications ($p<0.05$) of allelic variation. To complement their data, Ge and colleagues sought to validate some of their findings through experimental means. They selected 40 genes based on two criteria: first, that they were expressed in lymphoblastoid cell lines (LCLs); second, that they showed more significant differences in allelic distribution between the EST and HapMap CEPH data. The LCL panel included the CEPH individuals from the HapMap database, allowing the selection of individuals heterozygous for the relevant SNPs.

Using DNA sequencing and in-house software, the allele predicted to be over expressed matched their experimental results in 242 cases, the alleles were found to be equally expressed or undefined in 348 cases and the opposite allele was

overexpressed in 105 cases. Overall, despite differences between the experimentally validated and the predicted findings, the *in silico* method Ge and colleagues employed resulted in the production of a list of genes enriched for allelic variation.

1.3.3 Evidence Suggesting that *MUC* Gene Expression may be Influenced by Regulatory Polymorphisms

This section presents the arguments to suggest that it would be worthwhile to explore the possibility that *MUC* gene expression may be affected by regulatory polymorphisms and hence exhibit allelic variation in mRNA expression.

1.3.3.1 Heritable Differences in Gene Expression Levels are Common in the Human Genome

Firstly, a number of studies in the last few years have indicated that genetically determined differences in gene expression are widespread. One such study used cDNA microarrays to study genes expressed in lymphoblastoid cells, which were derived from 35 unrelated individuals from the Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigrees (Cheung et al. 2003).

The hybridisation signal intensities of each individual were compared against a reference sample, pooled from 10 individuals for 813 genes. Of the 813 genes, 40 genes that showed the greatest variation in expression levels between the 35 individuals were selected. These 40 genes included several that were already known to be highly variable amongst individuals, such as *HLA-DRB1* and drug metabolising genes like *CYP3A4* and *DHFR*, supporting the validity of their results.

Cheung and colleagues sought to investigate the genetic basis of the variation in gene expression observed by selecting five genes from the 40 highly variable genes and examining their transcript levels in three groups of people – 49 unrelated individuals, 35 offspring from 5 CEPH families and 10 pairs of monozygotic twins. Results from quantitative real-time PCR experiments indicated that for all five genes,

the expression levels were least variable in monozygotic twins and most variable amongst the unrelated individuals.

Although subsequent F tests showed that the difference in variance for the unrelated individuals versus the monozygotic twins was only marginally significant, the results suggest that inter-individual differences in gene expression are widespread throughout the genome and that these differences are due at least in part to underlying genetic differences.

A study by Morley and colleagues supported this conclusion.(Morley et al. 2004). They also used microarrays to measure the baseline expressions levels of genes in immortalised B cells from 94 CEPH grandparents. From the 8500 or so genes on the microarrays, 3554 genes which showed greater variance in expression levels between individuals than between replicate measurements (within individuals) were chosen for further study.

Genotypes for SNPs were obtained from the SNP consortium (Matise et al. 2003) and used to carry out genome-wide linkage analysis for the 3554 expression phenotypes in 14 CEPH families. This allowed the authors to isolate expression phenotypes with 'regulators', which were gene regions that were found to be significantly linked to the expression levels when a lod score of ~3.4 was used. 984 expression phenotypes with 'regulators' were identified, suggesting that a large proportion of human quantitative traits were subject to gene regulation and are influenced by underlying genetic determinants.

1.3.3.2 Allelic Variation in Non-imprinted Autosomal Genes - A Heritable and Common Occurrence in the Human Genome

The evidence that much of this variation in gene expression is allele-specific first came from the study of individual genes:

A recent study of allelic variation in mRNA expression was by Yan and colleagues (Yan et al. 2002). They investigated 13 genes for allelic variation in

mRNA expression in 96 CEPH families using mRNA from lymphoblastoid cell lines and the SBE method. They detected significant differences in relative expression levels for six of the 13 genes examined. The extent of the differences varied amongst the heterozygotes, ranging from a 1.3-4.3 fold difference depending on the gene. Next, they re-examined these six genes in nine families where a member of each family had previously exhibited allelic variation. In three of the families, they showed that altered allelic expression levels were consistently inherited together with a single haplotype, providing convincing proof that the allelic variation was caused by *cis*-acting inherited polymorphisms.

Since half the genes examined in this study exhibited allelic variation in mRNA expression, the authors proposed that this phenomenon might be a relatively common occurrence in humans and this led to a follow-up genome-wide survey. Affymetrix HuSNP chips were used to analyse 1063 SNPs, each of which were located in transcribed regions (Lo et al. 2003). A computational method was developed to extract the fluorescent intensity of the allele-specific probes and to quantify the relative levels of expression of each allele present. cDNA from seven fetal tissues sources such as kidney and liver was used.

There were significant differences in allelic expression for 326 of the 602 (54%) genes studied. In about 28% of the genes, this difference was greater than four-fold. Genes known to be imprinted and genes subjected to X-chromosome inactivation showed monoallelic expression, just as expected.

In order to substantiate these results, allele-specific real-time quantitative PCR was performed on a sample set of seven genes; two genes were known to be imprinted, one gene was shown to have equally expressed alleles, according to the HuSNP experiments, whilst the remaining four genes showed allelic variation in gene expression. With one exception, the results were concordant with the HuSNP findings.

A study by Bray and colleagues examined 15 genes that are expressed in the brain using mRNA derived from post-mortem brain (Bray et al. 2003). Their purpose was to investigate the extent to which Yan and colleagues' findings could be

extrapolated to native human tissue. Polymorphisms located within the 15 genes were selected and used to identify an average of 19 heterozygote individuals per polymorphism. These were then used to compare the relative expression levels of the different alleles by SBE. They found that seven of the 15 genes showed significant differences in allelic expression in one or more individuals - approximately half of the genes, as was also reported by Yan and colleagues in their study (Yan et al. 2002).

A final piece of evidence supporting the notion that allelic variation is widespread throughout the genome was an extensive literature survey of *in vitro* allelic variation studies up to the year 2002. This survey revealed that 107 genes found throughout the human genome exhibited allelic variation (Rockman and Wray 2002). From reporter assay experiments, 63% of these 107 genes showed allelic differences of two-fold or greater. However, since most of the genes tested had been pre-selected, based on prior evidence of probability or functionality, this high proportion is unlikely to be a true reflection of the entire genome.

1.4 HYPOTHESIS AND AIMS OF THE PROJECT

1.4.1 Hypothesis

Mucins have been shown to be aberrantly expressed in a wide variety of inflammatory diseases and types of cancer (see section 1.1.6). There is also research, which shows that mucin expression levels are altered at both mRNA and protein levels in response to inflammatory mediators (see section 1.2.5.3). Finally, there is evidence, particularly in the case of *MUC5B* (see section 5.1.1), to suggest that these changes in expression may be due at least in part to inherited regulatory polymorphisms affecting transcriptional regulation of the *MUC* genes (see section 1.3.3). Hence, allelic variation in *MUC* mRNA expression could be important in the study of complex diseases that involve inflammation and mucin alterations. Thus:

We hypothesise that cis-acting regulatory polymorphism is likely to cause allelic variation in mRNA expression of the MUC genes. This may affect the quantity

of particular mucins produced by an individual and thus the properties of the mucus and also the cellular response to inflammation. This in turn is likely to influence susceptibility to disease and disease severity.

1.4.2 Aims of Project

The main aims of this project were as follows:

- To establish an effective method for determining the presence and extent of allelic variation in mRNA expression of the *MUC* genes.

  Two genes were selected as examples: *MUC4,* which is membrane-associated and *MUC5B,* which is secreted and gel-forming.

- To determine the influence of template length (due to VNTR) on the level of *MUC4 mRNA* expression.
- To investigate the presence and degree of constitutive allelic variation in *MU5B* and *MUC4* mRNA expression.
- To test for association between *MUC5B* genotypes/high expressing haplotypes and respiratory outcomes in a longitudinal birth cohort.

CHAPTER 2

MATERIALS AND METHODS

2 MATERIALS AND METHODS

2.1 SAMPLES

2.1.1 Human Fetal Samples

The human fetal samples were taken from of 117 aborted fetuses from which one or more of the following tissue types were collected: lung, trachea, salivary glands, stomach, gall bladder, liver and intestine. The cohort comprises of tissue originating from both male and female fetuses (although the sex was not determined in a large majority of individuals), with menstrual ages ranging from 11 to 20.4 weeks. Information regarding the method of abortion was also recorded.

The samples were collected over a 7-year period, from 1997 to 2004, by the former MRC Tissue Bank at Hammersmith hospital, London. Fully informed consent for all the samples was obtained from the respective mothers and ethical approval was obtained (Ethical approval reference: UCL/UCLH 04/0011).

Each sample was labelled with an identifier consisting of a unique number representing the individual, followed by a single letter: G (gall bladder), I (intestine, both large and small), L (lung), LV (liver), S (stomach), SG (salivary gland) and T (trachea), representing the tissue of origin.

2.1.2 MRC National Survey of Health and Development (NSHD) - 1946 Longitudinal Birth Cohort)

Details of this cohort are described in chapter 6 (see section 6.1).Blood and buccal samples were collected (Ethical approval reference MREC no 98/2/121) from the cohort at age 53 and used to extract genomic DNA (gDNA), giving a total of 2939 members represented. Plated buccal gDNA was available for this project (~30 ng per well) and epidemiological information was available though NSHD.

2.2 LIST OF BUFFERS AND SOLUTIONS

- 5 X TBE

  0.44 M Tris, 0.44 M Boric acid, 12.5 mM EDTA, pH 8.2-8.4

- Sample loading buffer for agarose gel electrophoresis

  0.25 % Bromophenol blue, 0.25 % Xylene Cyanol FF, 15 % Ficoll.type-400

- *Back Extraction* buffer

  4 M Guanidine Thiocyanate, 50 mM sodium citrate, 1 M Tris

- *Depurinating* solution

  0.25 M HCl

- *Denaturing* solution

  1.5 M NaCl, 0.5 M NaOH

- *Neutralising* solution

  1.5 M NaCl, 0.5 M Tris HCl, 0.001 M EDTA, pH 7.2

- *Denhardts* solution (100 X)

  2 % (w/v) Ficoll type-400, 2 % (w/v) Bovine Serum Albumin (BSA), 2% (w/v) Polyvinylpyrrolidone (PVP), filtersterilised with a 0.2 μM membrane

- *Pre-hybridisation* solution

  60 ml 20 X SSC, 120 ml dH_2O, 10 ml Denhardts (100 X), 10 % (w/v) SDS

- 20 X SSC (Sodium Chloride, Sodium Citrate)

  3 M NaCl (Sodium Chloride), 0.3 M $Na_3C_6H_5O_7.2H_2O$ (Sodium Citrate), pH 7.0

- 'Home-made' *Microclean* solution
  1 M NaCl, 2 mM Tris-HCl, 0.2 mM EDTA, 40 % PEG-8000 (polyethylene glycol), 3.5 mM $MgCl_2$

- 'Home-made' *Better Buffer* solution
  200mM Tris-HCl (pH 9.0), 5 mM $MgCl_2$

- *ABgene PCR Buffer IV* (10 X concentration)
  1.25 ml of 750 mM Tris-HCl (pH 8.8 @ 25 °C), 200 mM $(NH_4)_2SO_4$, 0.1 % (v/v) Tween 20®, 15 mM $MgCl_2$

- *NEBuffer 2*
  10 mM Tris-Hcl, 10 mM Mg Cl_2, 50 mM NaCl, 1 mM dithiothreitol (pH 7.9 @ 25 °C)

- *NEBuffer 3*
  50 mM Tris-Hcl, 10 mM Mg Cl_2, 100 mM NaCl, 1 mM dithiothreitol (pH 7.9 @ 25 °C)

- *NEBuffer 4*
  20 mM Tris-acetate, 10 mM magnesium acetate, 50 mM potassium acetate, 1 mM dithiothreitol (pH 7.9 @ 25 °C)

2.3 RNA/DNA EXTRACTION FROM FETAL TISSUE

2.3.1 Tissue Homogenisation

All fetal tissue samples were stored at -70 °C. All equipment (pestles, mortars, and forceps) were cleaned with Presept solution and baked at 180 °C for 6 hours in order to denature any nucleases. The RNA and DNA extraction protocols are optimised for samples ranging from 10-60 mg in weight. Hence, a mallet was used to break off a suitably sized piece of tissue whenever a sample was estimated to exceed the 60 mg limit. Otherwise, the entire sample was used.

First, the piece of tissue was transferred using forceps into a cyrotube. The cyrotube was then dropped into a container filled with liquid nitrogen to keep the tissue frozen. Liquid nitrogen was poured into a mortar and a pestle was placed inside in order to pre-chill both the mortar and pestle. A separate mortar and pestle was used for each sample. Meanwhile, the cryotube was weighed using a balance and its weight taken as soon as the reading stabilised. The sample was then emptied into the mortar containing liquid nitrogen. The empty cryotube was weighed and the difference between the readings gave an estimate of the weight of the tissue.

The sample was homogenised with the pestle and mortar, using liquid nitrogen.

2.3.2 RNA Extraction

The RNA extraction method employed was the *RNA Bee* (Biogenesis Ltd) extraction solution, which is based on the phenol-chloroform extraction method developed by Chomczynski (Chomczynski 1993).

After the sample was completely homogenised, 20 µl of *RNA Bee* per mg of tissue was added directly to the mortar. The pestle was used to mix the *RNA Bee* with the sample and the resultant semi-frozen mixture left to thaw.

Once the mixture had thawed completely, it was transferred to a clean 1.5 ml microcentrifuge tube using a pipette. The pipette was also used to estimate the total volume of mixture recovered (some of the mixture was left behind on the mortar). Only filtered pipette tips were used from this point forth. Next, 1 µl of chloroform for every 5 µl of mixture was added to the tube. The tube was then shaken vigorously for 30 sec, before it was placed on ice and left to stand for five min.

Next, the tube was centrifuged at 12000 rpm (MSE Micro Centaur centrifuge) for 15 min at 4 °C. The aqueous phase containing RNA was then transferred into a clean 1.5 ml microcentrifuge tube. Great care was taken not to disturb the interface and risk contaminating the aqueous phase with DNA whilst doing so. Thus, in

practice, not all of the aqueous phase was recovered, in order to leave a safety buffer between the interface and aqueous phase. Next, 500 µl of isopropanol was added in order to precipitate the RNA. The sample was left to stand for 10 min at room temperature before it was centrifuged at 12000 rpm (MSE Micro Centaur centrifuge) for 15 min at 4 °C. The tube containing the organic phase was stored at -70 °C for gDNA extraction later on.

The supernatant was removed and care was taken not to lose the RNA pellet. 75 % ethanol equal in volume to that of the starting mixture was added, making sure that the pellet was dislodged from the tube wall. The sample was then centrifuged at 7000 rpm (MSE Micro Centaur centrifuge) for 5 min. Finally, the supernatant was removed and the pellet was re-suspended in 50 µl of RNase-free dH_2O via gentle pipetting.

Optical density of the total RNA was measured and its concentration estimated from the A260/A280 ratios. The concentration was adjusted to 200 $ng\mu l^{-1}$ and the sample was stored at -70 °C.

1 µl of total RNA was checked on 1 % agarose gels and deemed as good quality if the 18s and 28s ribosomal RNA bands were visible.

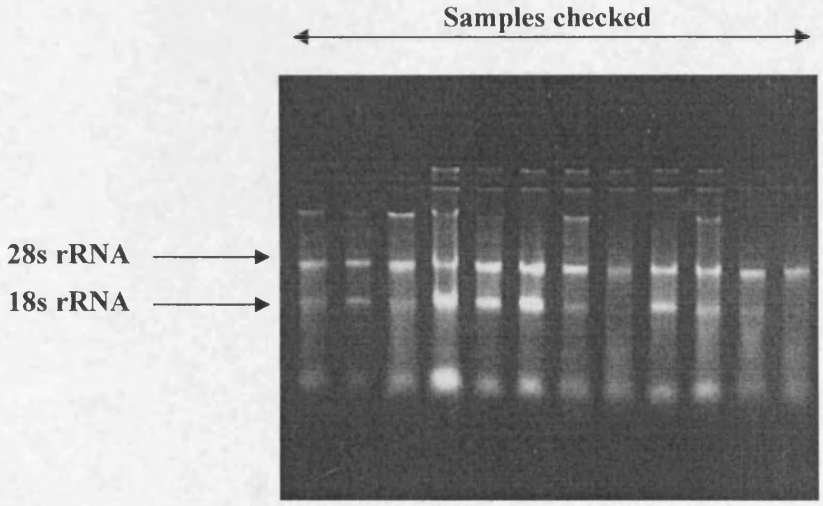


Figure 2-1 - RNA samples run on 1 % agarose gel to assess their quality

2.3.3 gDNA Extraction following RNA Extraction

The organic phase left behind during the RNA extraction procedure (see section 2.3.2) contains gDNA and proteins. The following procedure was used to extract the gDNA from the mixture:

First, the microcentrifuge tube containing the mixture was allowed to defrost completely. Then the tube was inverted a few times to mix its contents and centrifuged at 12000 rpm (MSE Micro Centaur centrifuge) for 10 min at 4 °C to re-establish the various separation phases. After this, the remaining aqueous phase was removed, thereby reducing any RNA contamination of the gDNA.

0.5 X the volume of RNA Bee used (see section 2.2.2) of *back extraction buffer* was added. Next, the tube was shaken vigorously for 30 sec and then stored for 10 min at room temperature. Then, the tube was centrifuged at 12000 rpm (MSE Micro Centaur centrifuge) for 15 min at 4 °C.

The supernatant was discarded and 1 ml of 75% ethanol was added to the pellet. The mixture was stored at room temperature for 10 min and mixed periodically by inversion. Next, the sample was centrifuged at 2000 rpm (MSE Micro Centaur centrifuge) for 5 min at 4 °C. The ethanol was removed and the pellet was allowed to briefly dry for 3 min at room temperature. The pellet was dissolved in 50 µl of 8mM NaOH solution by gentle pipetting. Finally, 8.6 µl of 0.1M HEPES buffer was added to bring the pH of the gDNA down to 8.4, a pH more suitable for PCR reactions.

2.3.4 gDNA Extraction for Southern Blots

Genomic DNA for Southern blots needs to be of high quality, so for this purpose the gDNA was extracted directly from fetal tissue, rather than as a by-product from the RNA extraction, as described in the previous section.

An aliquot of *Cell Lysis Solution* (Puregene DNA Purification Kit) was placed into the water bath set at 53 °C and allowed to warm up. This prevents the SDS from precipitating out of the *Cell Lysis Solution* at low temperatures. Meanwhile, a piece of fetal tissue was homogenised under liquid nitrogen as described in section 2.2.1 and was left at room temperature for 10 min. Next, 600 µl of the pre-warmed *Cell Lysis Solution* was added to the mortar and mixed thoroughly with the ground tissue before the lysate was transferred to a clean 1.5 ml microcentrifuge tube using a pipette.

600 µl of *Cell Lysis Solution* is suitable for tissues weighing 10-20 mg. If the piece of tissue greatly exceeded 20 mg, an additional 600 µl of *Cell Lysis Solution* was added and the total volume of the lysate was divided into two microcentrifuge tubes. The tubes were henceforth treated like two separate samples for the rest of the extraction.

3 µl of Proteinase K (20 mg/ml) solution was then added to the tube and its contents were mixed by inverting the tube 25 times. The mixture was left to incubate overnight in the 55 °C water bath.

The following morning, the sample was cooled to room temperature and 200 µl of *Protein Precipitation Solution* (Puregene DNA Purification kit) was added to the lysate. The tube was vortexed at high speed for 20 sec, thoroughly mixing its contents before it was centrifuged at 7200 g (Fisher Scientific Force 7 centrifuge) for 5 min. A protein pellet should be visible at his stage. If not, the tube was re-vortexed, placed on ice for 5 min and the centrifugation step repeated.

The supernatant was transferred into a clean 1.5 ml microcentrifuge tube containing 600 µl of 100% isopropanol (2-propanol) and its contents mixed by inverting gently 50 times. The tube was then centrifuged at 7200 g (Fisher Scientific Force 7 centrifuge) for 2 min. The supernatant was poured off and the tube was drained on clean absorbent paper. 600 µl of 70% ethanol was added and the tube was inverted several times to wash the DNA pellet. Following this, the tube was centrifuged at 7200 g (Fisher Scientific Force 7 centrifuge) for 2 min. The ethanol

was discarded, ensuring that the DNA pellet was not lost. The tube was inverted, drained on a clean piece of absorbent paper and allowed to air dry for 10-15 min.

Lastly, the pellet was re-suspended in 100 µl of *DNA Hydration Solution* (Puregene DNA Purification Kit) before the tube was left on a rocker at 4 °C overnight. The gDNA was stored at 4 °C.

1 µl of gDNA was checked on 1% agarose gels. If no band was visible, (sometimes because the DNA pellet was not completely dissolved) the sample was incubated at 37 °C for 2 hours and rechecked.

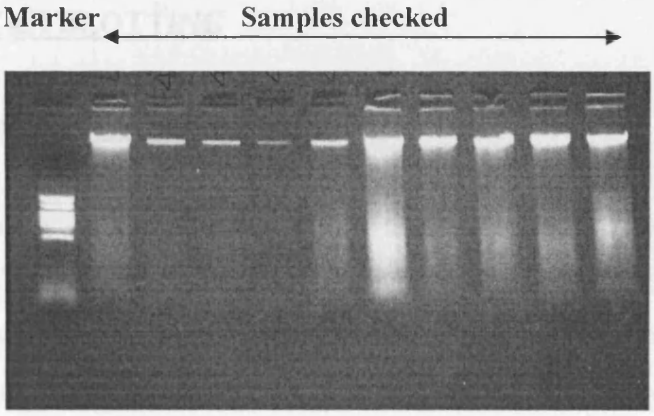


Figure 2-2 - Extracted gDNA run on a 1% agarose gel

2.3.5 Reverse Transcription (Single-stranded cDNA Synthesis)

The extracted RNA was converted to cDNA using the following method:

For each sample, 2 µl (~400 ng) of total RNA was put into two 0.2 ml microcentrifuge tubes (Eppendorf), one tube as a + RT and the second as a – RT control. A master mix comprising of 8 µl of dH_2O, 2 µl of 10 X dNTPs and 0.5 µl (0.25 µg) of random hexamers was added to each tube. The tubes were heated at 65 °C for 5 min in a thermocycler (MJ).

Next, a master mix containing 4 µl of 5 X buffer (provided with Moloney Murine Leukemia Virus (MMLV), Invitrogen), 2 µl of 0.1M DTT (Invitrogen) and

0.5 µl of RNaseOut RNase inhibitor (Invitrogen) was added to both tubes. 1 µl (200 units) of MMLV (Invitrogen) was added to the + RT tube whilst 1 µl of dH_2O was added to the – RT tube.

Reverse transcription was performed in a thermocycler using the following conditions: 37 °C for 2 min, then 25 °C for 10 min, followed by 37 °C for 50 min and finally 70 °C for 15 min. Lastly, the cDNA was diluted 1:10 in dH_2O and stored as smaller aliquots at -20 °C.

2.4 DETERMINATION OF TANDEM REPEAT LENGTHS USING SOUTHERN BLOTTING

2.4.1 Digestion of DNA Samples

In order to determine the *MUC4* tandem repeat (TR) lengths for an individual, gDNA from that individual (obtained by the method described in section 2.3.4) was digested using *Pvu II*. This enzyme has recognition sites relatively close to both 5′ and 3′ ends of the *MUC4* TR (as well as elsewhere throughout the genome), but crucially does not cut within the TR itself. Hence, digestion with this enzyme will result in gDNA fragments containing the entire *MUC4* TR, for both alleles, which can be subsequently detected and sized.

For each sample, 12 µl of gDNA was digested with a master mix composed of the following: 1.7 µl (17 units) of *Pvu II* (New England Biolabs), 1.7µl of NEBuffer 2 (New England Biolabs), 0.6 µl (0.1M) of Spermidine (Promega, Southampton, UK) and 1 µl of dH_2O. The samples were incubated at 37 °C overnight.

2.4.2 Gel Electrophoresis

5 µl of ficoll loading buffer was added to each sample, before they were loaded onto a 0.8% agarose gel (2.4 g of agarose in 300 ml of 1 X TBE and 3 µl of

EtBr (5 $mg.ml^{-1}$). In addition to the samples, 15 µl of *Raoul Molecular Weight Marker* was run in the first and last lanes of the gel. Two digested genomic controls of known molecular size were run as well. The first control is a genomic sample prepared from the cell line Caco-2 whilst the second is a mixture of two genomic samples (180/210). Electrophoresis was performed in 2 litres of 1 X TBE at 50V, 50mA for 24 hours. After this, the gel was visualised under UV light with a ruler lined up next to the *Raoul* marker bands and a photo was taken.

2.4.3 Southern Blotting

The gel was subjected to Southern blotting using a *VacuGene XL Vacuum Blotting Unit* (Amersham Pharmacia Biotech). The gDNA fragments were vacuum blotted onto a *Hybond N+ nylon transfer membrane* (Amersham Biosciences) using a pressure of 75 millibars. The following solutions were used as transfer buffers in this specific order: first, the *depurinating* solution for 30 min, next the *denaturing* solution for 30 min, then the *neutralising* solution for 30 min and finally 20 X SSC for 2 hours.

The membrane was placed between two pieces of 3mm filter paper (Whatman) and dried by baking in an oven at 80 °C for two hrs before it was stored in a cool dark place.

2.4.4 Radioactive Detection of DNA Fragments

After the Southern blotting was complete, the membranes were ready to be probed in the following fashion:

2.4.4.1 Membrane Pre-hybridisation

200 ml of pre-hybridisation solution was brought up to 65 °C, poured into a large clean plastic box and the membranes were added one at a time into the box. The box was swirled after each membrane was added to ensure that a layer of

solution was between each membrane. Then the box was placed in the water bath at 65 °C and incubated with agitation for four hrs.

2.4.4.2 Labelling of Probe

The probe was radiolabelled as follows: 6 µl (25-50 ng) of a *MUC4* TR specific cDNA probe (JER64) (Porchet et al. 1991), 54 µl of dH_2O and 10 µl of random primers from the *Megaprime DNA Labelling System* (Amersham Biosciences) were combined in a 1.5 ml microcentrifuge tube. The mixture was denatured at 95 °C for 5 min and snap cooled on ice. 20 µl of labelling buffer (Megaprime DNA Labelling System), 6 µl of *Redivue* 5'[α^{32}P]dCTP (370 MBq/ml, GE Healthcare Life Sciences) and 4 µl of Klenow (Megaprime DNA Labelling System) was added. The mixture was then incubated at 37 °C for 30 min.

Following this, the labelled probe was purified using a *Nick column* (Amersham Biosciences) to remove the unincorporated nucleotides. The reaction mixture was applied to the top of the column and eluted with 400 µl of 3 X SSC. This 'void' volume was collected in a 15 ml centrifuge tube and kept aside. The labelled probe was then recovered by adding another 400 µl of 3 X SSC to the column. The unincorporated radionucleotides are left behind, within the column itself.

The efficiency of the radionucleotide incorporation was estimated by measurements with a Geiger counter, using the following formula:

$$\text{Incorporation rate (\%)} = \frac{\text{counts per minute (cpm) of the labelled probe}}{\text{cpm of the labelled probe + cpm of the unincorporated radionucleotides (column)}} \text{ X } 100$$

Equation 2-1 - Estimation of radionucleotide incorporation rate

The probe was used only if the incorporation rate was above 50%. Otherwise, the labelling procedure was repeated. This procedure labels enough probe to test up to five membranes.

2.4.4.3 Probing of Membrane

50 µl of Herring Sperm DNA (10 mg/ml, Promega) was then added to the probe to block non-specific hybridisation of the probe to the membrane and the mixture was boiled for 5 min. The box containing the membranes and prehybridisation solution was retrieved from the water bath and the membranes were removed. The probe was then introduced into the prehybridisation solution and the membranes were returned one at a time to the solution. Lastly, the box was placed back in the water bath and incubated with agitation at 65 °C overnight.

2.4.4.4 Washing of Membranes

The next morning, the prehybridisation solution was discarded and the membranes were washed to remove any unbound probe. Three wash solutions were used consecutively; all contained 0.1% (w/v) SDS but had decreasing amounts of SSC (2%, 0.2% and 0.1% SSC) such that the solutions were of increasing stringency. (2% solution for 30 min, and the duration of the washes for the 0.2% and 0.1% solutions was dependent on Geiger counter readings performed after each wash, the goal being to wash off as much unbound probe as possible, without washing off too much of the bound probe in the process.) All washes were carried out at 65 °C in a water bath.

2.4.4.5 Probe Detection

The membranes were wrapped in cling film and placed in a *Fuji FG-8 film cassette* (Fuji film) together with a piece of *Super HR-E30 Fuji medical X-ray film* (Fuji film). The X-ray films were exposed to the membranes overnight in a -70 °C freezer and the autoradiographs developed the following day using a *Compact X2*

automatic X-ray film processor (Xograph Imaging Systems Ltd). Depending on the results obtained, longer or shorter exposure times were used for subsequent pieces of film, to obtain the best results possible.

2.4.4.6 Tandem Repeat Size Determination

Once the best possible autoradiograph was obtained for a particular membrane, it was scanned using the *Grab-IT 2.59* (Ultra-Violet Products) computer program. The digital image of the autoradiograph was then processed using the *GelWorks ID Intermediate v4.01* (Ultra-Violet Products) software package. This software package uses the bands of known size from the Raoul markers to plot a curve of size (kb) against distance migrated (cm). This curve is then used to determine the size of the other bands on the autoradiograph. There are two main advantages of using this system, rather than sizing the bands by hand. The first is that the process is faster and more accurate. The second is that the program can take into consideration distortions in the gel run (such as if the DNA fragments did not migrate in a straight line during electrophoresis) by using the Caco-2 and 182/210 samples as internal controls.

2.5 POLYMERASE CHAIN REACTION (PCR)

Two different protocols were used, depending on whether normal Taq polymerase or *Thermo-Start* Taq was used. Following PCR, the products were subjected to electrophoresis on 2% agarose gels and visualised under UV light to ensure that the PCR products were of the expected sizes.

2.5.1 Taq polymerase PCR

A 10 µl reaction was carried out, using 2 µl of gDNA (~30 ng), 1 µl of each primer (5 µM), 1 µl of 10 X Buffer IV (containing $MgCl_2$, ABgene), 1 µl of 10 X dNTPs (200 µM), 0.05 µl of Taq polymerase (0.25 units, ABgene) and 3.95 µl of dH_2O.

A standard PCR cycling program was used, although the annealing temperature as well as number of cycles was adjusted depending on the particular PCR reaction, as shown in Table 2-1.

95 °C initial denaturation for 5 min, followed by 30 sec of denaturation at 95 °C, 30 sec at the annealing temperature (dependant on the primers used) and 30 sec of extension at 72 °C. These last three steps were repeated up to 35 times (depending on the PCR) and were followed by a final 5 min extension at 72 °C.

Purpose	Primer Name	Sequence of Primer (5′ to 3 ′)	Product Size (bp)	Annealing Temp (°C)	No. of Cycles
MUC4 rs2259292 genotyping	*MUC4 Ex4 For DNA*	TCGGGGAACACAGGACAG	318	63	32
	MUC4 Ex4 Rev DNA	CGTCTCGAAGCAGGAGAGAG			
MUC4 rs2259102 genotyping	*rs2259102 Forward Exonic*	ATGGTGAACACAGCCTGCTA	132	58	30
	rs2259102 Reverse Exonic	TCCACTGGGCAGGATAGG			
MUC4 rs2550240 genotyping	*MUC4X7F*	AACACCTACCAAGCCATCCTC	263	65	31
	MUC4I7R	GGAGTCAGCGTGAGTCAGAAG			
MUC4 rs2291652 genotyping	*rs2291652 For DNA/cDNA*	CGACGTGGTCTTCCAGCCGAT	282	61	32
	rs2291652 rev DNA	TGCTGGCTGCTCCTGCTCTC			
MUC4 rs3205933 genotyping	*MUC4exon24-F*	TTCATCGGTTGCTTTCTGTG	475	58	30
	MUC4exon24-R	GGCAGGTTCCTTGTCATCTC			
MUC4 rs2291652 gDNA allelic ratios	*MUC4 cDNA Forward*	CGGTGGTGGAGGCGTTCT	278	61.4	32
	rs2291652 For DNA/cDNA	CGACGTGGTCTTCCAGCCGAT			
MUC5B rs2075859 RFLP genotyping	*rs2075859 For DNA*	CAGCGCGCGCAGCACTG	204	62	32
	rs2075859 Rev DNA	CCTCCCAGCCAGTGCCCTTC			

Purpose	Primer Name	Sequence of Primer (5′ to 3 ′)	Product Size (bp)	Annealing Temp (°C)	No. of Cycles
MU5AC rs1132440 multiplex genotyping	*MUC5AC rs1132440 F*	ACACCGAGGTGGAAGAGTGC	243	61	32
	MUC5AC rs1132440 R	CTGGACAGGGGCACAAGTTC			
IL4 rs2070874 multiplex genotyping	*IL4 rs2070874 F*	CTCATTTTCCCTCGGTTTCA	150		
	IL4 rs2070874 R	GAAGCAGTTGGGAGGTGAGA			
IL13 rs1800925 multiplex genotyping	*IL13 rs1800925 F*	TGACATCAACACCCAACAGG	201		
	IL13 rs1800925 R	CACCAGTCTCTGCAGGATCA			
MUC5B rs2672785 multiplex genotyping + gDNA allelic ratios	*rs2672785 Multiplex For gDNA*	CATTCCCTCTTCCCACAGAG	222		
	rs2672785 Rev DNA	GGGTGCTGGGTATTGTCCT			
MUC5B rs2075853 multiplex genotyping + RFLP genotyping + gDNA allelic ratios	*rs2075853 For DNA*	GGGAGGGTGTCTCTGCTTC	151		
	rs2075853 Rev DNA	GGCAGCATCTGCTTACGG			
MUC5B rs2075859 multiplex genotyping + gDNA allelic ratios	*rs2075859 Multiplex For gDNA*	ACATGCAGCACCAGGAGTGT	207		
	rs2075859 Multiplex Rev gDNA	GCATGAGGGGCTCTGCTT			
MUC5B loci 6, 7, 8 sequencing	*5BpromF1*	CCACGGAGCATTCAGGAC	576	61	32
	5BpromR2	CTCAGTCTGGGTGGCTTGTG			
MUC5B loci 9, 10, 11 sequencing	*5BpromF2*	CACAAGCCACCCAGACTG	585	61	32
	5BpromR1	GAGCCAACACCAGCGTC			

Table 2-1 - Details of primers and conditions for PCR reactions using Taq

2.5.2 *Thermo-Start* Taq Polymerase

For some PCRs and all RT PCRs, to reduce the spurious annealing of the primers, *Thermo-Start* Taq was used:

A 10 µl reaction was carried out, using 3 µl of gDNA (~30 ng) or cDNA, 1 µl of each primer (5 µM), 1 µl of *Thermo-Start* buffer (ABgene), 1 µl of $MgCl_2$ (25mM), 1 µl of 10 X dNTPs (200 µM), 0.05 µl of *Thermo-Start* Taq polymerase (0.25 units, ABgene) and 1.95 µl of dH_2O.

The following cycling program was used: (As was the case with normal Taq polymerase PCR, the number of cycles and annealing temperature was dependent on the particular PCR. These details are shown in Table 2-2.)

95 °C initial denaturation for 15 min, followed by 30 sec of denaturation at 95 °C, 30 sec at the annealing temperature (dependant on the primers used) and 30 sec of extension at 72 °C. These last three steps were repeated up to 35 times (depending on the PCR) and were followed by a final 5 min extension at 72 °C.

The initial denaturing step is 10 minutes longer than that for the normal Taq PCR (see section 2.5.1). *Thermostart* Taq is a chemically modified form of the normal Taq polymerase and an extended heating time is needed to activate the enzyme. The principle behind this is that the Taq will not be able to function whilst the PCR reactions are being set up and before the reaction mixture has reached 95 °C, thereby minimising the initial formation of spurious PCR products that could be amplified in the cycling steps.

Purpose	Primer Name	Sequence of Primer (5′ to 3 ′)	Product Size (bp)	Annealing Temp (°C)	No. of Cycles
MUC5B rs2672785 Tetra ARMS genotyping	*rs2672785 FI DNA*	CCCTGTGGAGCCGAGCTGTGA	296	62	32
	rs2672785 RO DNA	GTCAGACGCATGGCTGCTTGGAG			
	rs2672785 RI DNA	CATGGTGTGCCCTGCATGCC	149		
	rs2672785 FO DNA	TTTCCATAGCATTGAGCAGGAGCCG			
MUC5B rs2672785 + rs2075853 cDNA allelic ratios	*MUC5B E1-3 For cDNA*	GACGCTGGTGTTGGCTCTG	175	59.2	35
	rs2075853 Rev cDNA	GCTCAGGCTGGGGAAGACA			
MU5B rs2075859 cDNA allelic ratios	*MUC5B E8-E9 boundary*	CTCTGCCCCCGGACCT	165	60.1	35
	rs2075859 Rev cDNA	AGCCAGAGTGCGTGATGTC			
MUC4 E3-E4 cDNA	*MUC4 E4 For Out cDNA*	GACAGACGGTGGGAGACGCA	264	62	35
	MUC4 E4 Rev Out cDNA	TAGAGGGAATCACGGAGAGAGGAGC			
MUC4 E5-E7 cDNA	*MUC4X5F*	GACGATGCTGACTTCTCCAC	277	61	35
	MUC4 Ex6 Rev cDNA	ATCCCACCGCTCTGGTAG			
MUC4 E7-E9 cDNA	*MUC4X7F*	AACACCTACCAAGCCATCCTC	251	60	35
	MUC4X9R	CGGTGTAGCCTGTAGAACTGC			
MUC4 E23-E24 cDNA	*MUC4 cDNA Forward*	CGGTGGTGGAGGCGTTCT	168	57	35
	MUC4 E24 Rev cDNA	CCAGGTCGTAGCCCTTGTAGC			
MUC4 E24-25 cDNA	*MUC4 E24 For CDNA*	CTTCAGATGCGATGGCTACA	207	58	35
	MUC4 E25 Rev cDNA	GCGTCGAGTTTCATGCTCA			
Ribosomal RNA controls	*Ribo1*	CCTGGTTCTGGGGACCTGGAC	144	65	26
	Ribo2b	CAGGTCCAGGGGTCTTGGTCC			

Table 2-2 – Details of primers and conditions for PCR reactions using *Thermo-Start* Taq

For RT PCR, the primers either spanned intron-exon boundaries or were located in exons with large introns in between so that only cDNA and not contaminating gDNA produced an amplicon. If this happened, products amplified from contaminating gDNA could be easily identified from their large size.

As an added precaution, whenever RT PCR was carried out on a sample for the first time, the PCR was also performed for its – RT control and checked to ensure that a product was not obtained.

2.5.3 Standard Procedure for Agarose Gel Electrophoresis

A standard size marker was run with all agarose gels. The size markers were made in the lab using PCR products of known sizes, as shown in Figure 2-3.

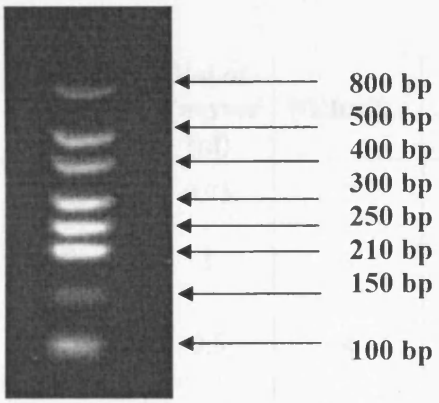


Figure 2-3 - PCR Size Marker

1 to 3 % (w/v) agarose gels were typically used with 50 ml 1 X TBE as the gel and running buffer. 1 µl of EtBr (5 $mgml^{-1}$) was added to the gels and to the running buffer. Gel electrophoresis was carried out at 60 mA / 60 V for 30-45 mins before the gels were visualised using a UV transilluminator (Ultra Violet Products, Cambridge).

2.6 GENOTYPING

Throughout this project, the polymorphisms were typed as inferred genotypes: i.e. phenotypes C, CT and T were assumed to be genotypes CC, CT and TT.

2.6.1 Restriction Enzyme Digestion

The *MUC4* SNPs and the *MUC5B* rs2075853 and rs2075859 SNPs were genotyped by restriction fragment length polymorphism (RFLP). Table 2-3 shows the enzyme required to genotype each SNP, the type of NEB enzyme buffer, the volume of enzyme, enzyme buffer and dH_2O required, and the expected sizes of the allele-specific bands. 3 µl of PCR product was digested in all the assays and 0.15 µl Bovine Serum Albumin (BSA) was added to each sample for rs2075859 genotyping.

Gene	SNP	Enzyme	Vol of Enzyme (µl)	NEBuffer	Vol of Buffer (µl)	Vol of dH_2O (µl)	Alleles and Band sizes (bp)
MUC4	rs2259292	Sfo I	0.75	2	1	5.25	A = 318 G = 177, 141
MUC4	rs2259102	Mbo I	1	3	1.5	9.5	C = 132 G = 79, 53
MUC4	rs2550240	ScrF I	0.5	4	1	5.5	C = 105, 104, 32, 22 G = 209, 32, 22
MUC4	rs2291652	Mbo I	1	4	1.2	6.8	C = 261, 21 T = 282
MUC4	rs3205933	AlwN I	0.5	4	1.2	7.3	A = 475 G = 292, 183
MUC5B	rs2075853	Msp I	1	2	1.5	9.5	C = 88, 66 T = 151
MUC5B	rs2075859	Bts I	1	4	1.5	9.35	C = 187, 17 T = 204

Table 2-3 - List of enzymes and conditions for restriction enzyme digests

The digest reactions were then left overnight at 37 °C in an incubator. All 15 µl of each sample was subjected to electrophoresis the next day on a 3 to 4% agarose gel (depending on the particular assay) and visualised using a UV transilluminator.

2.6.2 Tetra-ARMS Allele Specific PCR (Ye, Humphries, and Green 1992)

The *MUC5B* rs2672785 SNP was genotyped using this method, which uses allele-specific primers. This method uses four primers in a PCR reaction. These are the Forward Outer (FO), Reverse Outer (RO), Forward Inner (FI) and Reverse Inner (RI) primers. FO and RO are non allele-specific primers that flank the region surrounding the SNP to be genotyped. FI and RI are allele-specific primers (for allele 1 and allele 2, respectively) whose 3′ ends terminate at the SNP of interest. In addition, the inner primers have a mismatch mutation at the third base from their 3′ ends. This mutation helps to make the primers more allele-specific.

Depending on the alleles present in an individual, different PCR products will be generated. FO and RO will always generate a product, because these primers are non allele-specific. FI and RO will only generate a product if allele 1 is present, whereas FO and RI will generate a product only when allele 2 is present. Hence, a homozygote for allele 1 will have the FO-RO and the FI-RO products, a homozygote for allele 2 will have the FO-RO and FO-RI products, whilst a heterozygote will have all three possible PCR products. These products are differentiated by gel electrophoresis. It is therefore vital to design the primers such that the products are sufficiently different in size from one another.

A 10 µl PCR reaction was carried out and consisted of the following: 2 µl of gDNA (~30 ng), 1 µl of each primer (5 µM), 1 µl of *Thermo-start* buffer (ABgene), 1 µl of $MgCl_2$ (25mM), 1 µl of 10 X dNTPs (200 µM), 0.05 µl of *Thermo-Start* Taq polymerase (0.25 units, ABgene) and 0.95 µl of dH_2O.

The PCR conditions described in section 2.5.2 were used. All 10 µl of the PCR products were subject to gel electrophoresis and genotyped under UV light.

2.6.3 Sequencing

2.6.3.1 Purifying PCR Products

15 µl of PCR product was prepared for each sample using PCR plates (ABgene). Next, 5 µl of the product was run on an agarose gel to check that the amplification was successful. The remaining 10 µl of PCR product was purified by adding 20 µl of 'home-made' *Microclean* and 10 µl of dH_2O. The plates were covered and the samples were thoroughly mixed by running the plates against the edge of a lab bench. The plate was centrifuged at 145 g (MSE Mistral 2000 centrifuge) for 45 min. Next, the plate lids were removed, the plates inverted, placed on pieces of paper towel and centrifuged for 1 min at 20 g, in order to remove the supernatant.

After this, 150 µl of 70% EtOH was added to each sample. The plates were covered and centrifuged for 25 min at 145 g. As before, the plate lids were removed, the plates were inverted, placed on paper towel and centrifuged for 1 min at 20 g. Lastly the samples were allowed to air-dry for 15 min at room temperature before the DNA pellets were resuspended in 10 µl of dH_2O.

2.6.3.2 Sequencing Reaction

3 µl of purified PCR sample was subjected to electrophoresis on an agarose gel and checked under UV light. Depending on the intensity of the product, 3 to 7 µl of PCR product was subsequently used. The following was added as a master mix to each sample: 5 µl of 'home-made' *Better Buffer*, 1 µl of *Big Dye v1.1 Terminator Mix* (Applied Biosystems), 0.48 µl of primer (5 µM) and enough dH_2O to make up the volume of reaction to 15 µl.

The sequencing reaction was carried out as follows: 96 °C for 1 min, 96 °C for 10 sec, 50 °C for 5 sec, and 60 ° C for 4 mins. The last three steps were repeated for 25 cycles.

2.6.3.3 Sequencing Reaction Clean-up

Following the sequencing reactions, 80 µl of 80% isopropanol was added to each sample. The plates were covered and run across the edge of the lab bench to mix the samples thoroughly. The samples were left at room temperature for 10 min. After this, the plates were centrifuged at 2500 rpm for 45 min (MSE MISTRAL 2000 centrifuge). The supernatant was removed in an identical fashion to that used in the PCR purification step (see section 2.6.3.1) Then 150 µl of 70% ethanol was added to each sample and the plates were centrifuged at 2500 rpm for 15 min. The supernatant was removed in the same way as before. Finally, the sample pellets were allowed to air dry at room temperature for 10 min.

2.6.3.4 Electrophoresis and Detection of Sequencing Products

The samples were resuspended in 10 µl of Hi-Di formamide and run on an *ABI PRISM 3730XL Genetic Analyser* (Applied Biosystems) using a 50 cm array and POP-6 polymer (Applied Biosystems). Hi-Di formamide was used in place of standard de-ionised formamide because the low conductivity of Hi-Di formamide aids in the efficient transfer of samples onto the sequencer's capillaries. A standard Big Dye v1.1 specific run module was used. The results were analysed using *Sequencing Analysis v5.1.1* (Applied Biosystems)

2.6.4 Multiplex Single Base Extension (SBE)

SBE was used to genotype all three *MUC5B* SNPs for the work described in chapter 6, as well as 3 other markers that formed part of another project. The method uses the *SNaPshot Multiplex kit* (Applied Biosystems) and reduces genotyping cost and time by multiplexing at both the PCR and SBE stages.

2.6.4.1 Multiplex PCR

Six different sets of primers were multiplexed together in the following fashion:

A 10.05 µl PCR reaction was carried out, comprised of the following: 2 µl of gDNA (~30 ng), 0.5 µl (18 µM) of *rs2672785 Multiplex For gDNA ver 2* primer, 0.5 µl (18 µM) of *rs2672785 Rev DNA* primer, 0.5 µl (7 µM) of *rs2075853 For DNA* primer, 0.5µl (7 µM) of *rs2075853 Rev DNA* primer, 0.5 µl (4 µM) of *rs2075859 Multiplex For gDNA* primer, 0.5 µl (4 µM) of *rs2075859 Multiplex Rev gDNA* primer, 0.5 µl (10 µM) of *MUC5AC rs1132440 F* primer, 0.5 µl (10 µM) of *MUC5AC rs1132440 R* primer, 0.5 µl (10 µM) of *IL13 rs1800925 F* primer, 0.5 µl (10 µM) of *IL13 rs1800925 R* primer, 0.5 µl (10 µM) of *IL14 rs2070874 F* primer, 0.5 µl (10 µM) of *IL14 rs2070874 R* primer, 1 µl of Buffer IV (ABgene), 1 µl of 10X dNTPs (200 µM) and 0.05 µl of Taq polymerase (0.25 units, ABgene).

The following PCR conditions were used: 95 °C initial denaturation for 5 min, followed by 30 sec of denaturation at 95 °C, 30 sec of annealing at 61 °C and 30 sec of extension at 72 °C. These last three steps were repeated for 31 cycles and were followed by a final 5 min extension at 72 °C.

2.6.4.2 PCR Product Purification

Excess primers and dNTPs from the multiplex PCR reactions had to be removed in order to prevent them from interfering with the subsequent SBE reactions. Shrimp Alkaline Phosphatase (SAP) was used to remove the 5'-phosphates from dNTPs, rendering them useless and Exonuclease I (Exo I) functions as a 3' → 5' single strand exonuclease, thereby removing any excess primers left over from the multiplex PCR. SAP was used rather than any other sort of phosphatase because it inactivates fully when heated at 75 °C.

1.33 units of SAP and 0.8 units of Exo I (made up as a master mix) were added to 4 µl of PCR product. The PCR products were placed in a thermocycler and

the following program was run: 37 °C for 60 min, then 75 °C for 15 min and finally a 4 °C hold. If the SBE reaction was not carried out immediately, the samples were stored at - 20 °C.

2.6.4.3 SBE

For each sample, 1.67 µl of *SNaPshot Multiplex Ready Reaction Mix* (Applied Biosystems) and 0.66 µl of primer mix consisting of the following: 0.75 µM of *rs2075853 PE* primer, 3.6 µM of *rs2672785 PE* primer, 3 µM of *rs2075859 PE* primer, 1.8 µM of *PE MUC5AC FORWARD rs1132440* primer, 0.2 µM of *PE IL13 rs1800925* primer and 2.4 µM of *PE IL4 rs2070874* primer, was added as a master mix to 1 µl of purified PCR product.

The samples were placed in a thermocycler and the following program was run: 96 °C for 10 sec, followed 50 °C for 5 sec, then 60 °C for 30 sec. These three steps were repeated for 25 cycles before a final 4 °C hold.

2.6.4.4 Post-extension Treatment

The post-extension treatment removes the 5′-phosphoryl groups from the unincorporated ddNTPs and alters their migration, preventing them from co-migrating with the fragments of interest during the electrophoretic run and interfering with the interpretation of results. The post-extension treatment was performed by adding 1 µl of a master mix consisting of 0.33 µl of calf intestinal phosphatase (CIP) (1 unit/µl), 0.43 µl of 10X CIP buffer and 0.23 µl of dH_2O to each sample. Next, the samples were placed in a thermocycler and the following program was run: 37 °C for 60 min, followed by 75 °C for 15 min and a 4 °C hold.

2.6.4.5 Electrophoresis and Detection of SBE Products

2 µl of each sample was combined with 10 µl of Hi-Di formamide and 0.25 µl of *Genescan LIZ 120* size standard. The size standard acts as an internal control. This circumvents the problem of slight variations in the migration rates of identically sized products between different capillaries and allows samples run in different capillaries to be sized accurately for comparison with each other.

The samples were denatured at 95 °C for 5 min before they were snap-cooled on ice. The samples were then run on an *ABI PRISM 3730XL Genetic Analyser* (Applied Biosystems) using the built-in run module for *SNaPshot* products. *Genemapper version 4.0* (Applied Biosystems) was used to analysis the data generated from the run. *Bins* for each marker were assigned. These are a set of instructions that tell the software what size a particular SBE primer should be, the colours of the peaks expected, and the alleles those colours represent. Once this was completed, the genotypes could be called automatically by the software, based on the size and colours of the peaks.

Table 2-4 shows details of the SBE primers, as well as the expected coloured peaks and the SNP alleles they represent.

Primer Name	Sequence (5′ to 3′)	Orientation	Alleles and Peak colours
rs2291652 PE	GGTCTTCCAGCCCAT	Sense	Black = C Red = T
rs2672785 PE	GTGTGGCCTGCATTC	Anti-sense	Red = A Black = G
rs2075853 PE	AACAAAGCTCACGCGCC	Anti-sense	Blue = C Green = G
rs2075859 PE	CAGCGCGCGCAGCACTG	Sense	Black = C Red = T
PE MUC5AC FORWARD rs1132440	AAAAAAAAAAGCTCGGCGCAGGGCA	Sense	Black = C Blue = G
PE IL4 rs2070874	GATATGCAGTGAGAATGTGAG	Anti-sense	Blue = C Green = T
PE IL13 rs1800925	TTTTTTTTGCCTTTTCCTGCTCTTCCCTC	Anti-sense	Blue = C Green = T

Table 2-4 - Details of SBE primers

2.7 TESTING FOR ALLELIC VARIATION IN mRNA EXPRESSION USING SBE

For the purpose of detecting and determining the extent of allelic variation, SBE was used. The primary difference between the SBE method used here and that described in section 2.6.4 is that the PCRs and SBEs are all carried out as singleplexes.

2.7.1 PCR and PCR Product Cleanup

The relevant PCRs were performed on individuals known to be heterozygous for the relevant SNPs, using the conditions and procedures described in section 2.5.2.

After the PCRs were complete, 3 µl of each sample was subject to gel electrophoresis on a 2 % agarose gel. The PCR products were checked to ensure that bands of the expected size were clearly visible. The PCR products were purified using the procedure described in section 2.6.4.2.

2.7.2 SBE and Post-extension Treatment

Each reaction was set up in the following manner: 1 µl of cleaned up PCR sample, 0.66 µl of the relevant extension primer (all at 1µM) and 1.67 µl of *SNaPshot Multiplex Ready Reaction Mix* (Applied Biosystems).

The samples were then placed in a thermocycler and the program described in section 2.6.4.3 was used. After the SBE was complete, the post-extension treatment described in section 2.6.4.4 was performed.

2.7.3 Electrophoresis, Detection of SBE Products

16 µl of Hi-Di formamide was added to each sample. The samples were then denatured at 95 °C for 5 min before they were snap-cooled on ice. The SBE products

were run on an *ABI PRISM 3100 Genetic Analyser* (Applied Biosystems) using a 50 cm array and *POP-6* polymer (Applied Biosystems). Dye set E was used in conjunction with a GeneScan run module, which encodes the parameters shown in the table below:

Run Temperature	60 °C
Cap Fill Volume	184 Steps
Current Tolerance	100 μAmps
Run Current	100 μAmps
Voltage Tolerance	0.6 kV
Pre-run Voltage	15 kV
Pre-run Time	180 sec
Injection Voltage	3 kV
Injection Time	22 sec
Run Voltage	15 kV
Number of Steps	10
Voltage Step Interval	60 sec
Data Delay Time	600 sec
Run Time	3000 sec

Table 2-5 – Parameters for SBE Electrophoresis Run on *ABI 3100*

The results were then analysed using *GeneScan Analysis Software version 3.7* (Applied Biosystems).

The different alleles are distinguishable by the colour of their peaks. The peak heights are representative of the abundance of a particular allele-specific product. Hence, the relative abundance of one allele-specific product compared to the other. was calculated as follows:

$$\text{Allelic Ratio of Sample X for SNP A} = \frac{\text{Peak Height of Allele 1 of SNP A for Sample X}}{\text{Peak Height of Allele 2 of SNP A for Sample X}}$$

Equation 2-2 - Method for calculating allelic ratios

2.8 STATISTICAL ANALYSIS

2.8.1 χ^2 (Chi-Square) Test

Pearson's χ^2 test, commonly referred to as *the* χ^2 test, is a statistical hypothesis test in which the test statistic is evaluated with reference to the χ^2 distribution. The formula for calculating the χ^2 test statistic is as follows:

$$\chi^2_{n-1} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Equation 2-3 - Formula for calculating χ^2 test statistic. O_i = an observed frequency; E_i = an expected (theoretical) frequency, asserted by the null hypothesis; n = the number of possible outcomes of each event.

In this thesis, the χ^2 test is used as a test for independence in two main ways: Firstly, to evaluate the deviation from the expected Hardy Weinberg distribution of a population and secondly, to evaluate the statistical significance of correlations from cross tabulations and logistic regression analyses.

2.8.2 Fisher's Exact Test

Fisher's exact test is a statistical significance test used in the analysis of categorical data where sample sizes are small. The test is used to examine the significance of the association between two variables in a 2 x 2 contingency table. Although a χ^2 Test could be used in such a situation, Fisher's exact test has two main advantages over the χ^2 Test. Firstly, it can used when any of the cell values in the contingency table is below 10 and secondly, when the data is very unevenly distributed within the contingency table (for example, a value of 1000 in one cell and a value of 5 in another). The sampling distribution of the test statistic only approximates a χ^2 distribution in such cases, making the χ^2 test unsuitable.

2.8.3 Student's t Test

There are a number of uses of the t test, including determining whether the mean of a normally distributed population has a value specified in a null hypothesis. However, in this thesis, the t test was used to test if the means of two groups were significantly different from each other (see section 5.7). There are different forms of the t-test depending on whether two groups of equal size are compared and if the data is paired. For the data analysed in this thesis, two unpaired, unequal sized groups were compared and the calculation used is as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \text{ where } s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Equation 2-4 – Formula for calculating t test statistic. 1: first group; 2: second group; s: standard deviation; n: number of samples; $\overline{X}$: mean of group

2.8.4 Bayesian Algorithm

The Bayesian algorithm is used by the *Phase* program (Scheet and Stephens 2006; Stephens, Smith, and Donnelly 2001) to infer the haplotypes and their probabilities in an individual. As the name suggests, this algorithm is based on Bayes theorem, which relies on conditional probabilities.

The algorithm begins by making an initial guess of the haplotypes, based on the allele frequencies. It then resolves all the unambiguous haplotypes (such as those in an individual homozygous at all the loci tested). Next, it randomly selects an individual from amongst the ambiguous samples and makes an estimate of the probability of the haplotypes in this individual, under the assumption that all the other haplotypes are correct. This process is repeated continuously (as a Markov Chain Monte Carlo method) until stable probabilities are obtained.

The method relies on the assumption that an unresolved haplotype should most likely resemble a resolved haplotype, on the evolutionary basis that haplotypes are derived from other haplotypes.

2.8.5 Exact Test of Population Differentiation Based on Haplotype Frequencies (Goudet et al. 1996)

This is a test of non-random distribution of haplotypes into population samples under the hypothesis of panmixia. It is similar to the Fisher's exact test where a 2 x 2 contingency table is constructed, except here a k x r table is constructed instead, where k is the number of haplotypes and r the number of populations.

In this thesis, this test is used to determine if the haplotype distribution is significantly different between two different populations, although the test can accommodate more than two populations.

2.8.6 Cross Tabulations

Cross tabulations (or crosstabs for short) are essentially contingency tables and show the distribution of two or more variables simultaneously. They are often used for a number of reasons, including ease of use and because they can be used with any level of data: nominal, ordinal, interval, or ratio.

In this thesis, cross tabulations in conjunction with χ^2 tests are mainly used to test for significant associations between *MUC5B* genotypes/haplotypes and various respiratory outcomes (see section 6.4)

2.8.7 Analysis of Variance (ANOVA)

In general, the purpose of analysis of variance (ANOVA) is to test for significant differences between the means of two or more groups. In this thesis, ANOVA is used primarily to test for significant correlations between *MUC5B* genotypes/haplotypes and measures of lung function (continuous data in the form of FEV_1 readings, see section 6.4). It was used in preference to the Student t test since

more than two groups were compared; this becomes computationally more difficult for t tests and greatly increases the chance of type I errors due to multiple pairs of comparisons.

2.8.8 Likelihood Ratio Test

A likelihood-ratio test is a statistical test in which a ratio is calculated between the maximum probabilities of a result under two different hypotheses. The general form of the test statistic (S) is as follows:

$$S = -2\ln\left(\frac{L_{H0}}{L_H}\right)$$

Equation 2-5 – Formula for calculating likelihood ratio test statistic. L_{H0}: maximum probability of an observed result under the null hypothesis; L_H: maximum probability of an observed result under the alternative hypothesis

This test statistic approximates to a χ^2 distribution, which can be used test for a statistically significant difference in likelihood between the two hypotheses. In this project, the likelihood ratio test is used primarily to compute a P value for the adjusted logistic regressions (see section 6.5)

2.8.9 Logistic Regression

Logistic regression is a variation of ordinary regression, which is used when the dependent variable is a dichotomous variable (such as 'yes' or 'no' for a disease).

In particular, logistic regression allows one to calculate the odds ratios for a variety of individual risk factors that may be contributing to the occurrence of a disease. The odds ratios represent the risk of disease for those with the exposure being studied (*MUC5B* in this thesis), relative to the risk of disease among those without the exposure — adjusted to remove the effects of all other confounding variables.

For the work described in this thesis, logistic regression was used in conjunction with the likelihood ratio test to compare an unadjusted model vs. an adjusted model (adjusted for confounding variables) and generate an overall P value for assessing statistical significance of *MUC5B* with a respiratory outcome (see section 6.5).

2.9 COMPUTER SOFTWARE

- *LDmax* (Abecasis and Cookson, 2000)
  http://www.sph.umich.edu/csg/abecasis/GOLD/
  Part of GOLD software package. Uses the Slatkin and Excoffier (1995) algorithm to analyse genotype data.

- *Phase v2.1* (Scheet and Stephens 2006; Stephens, Smith, and Donnelly 2001)
  http://stephenslab.uchicago.edu/software.html
  Uses Bayesian algorithm to infer haplotype distributions.

- *Arlequin v3.11* (Excoffier, L. G. Laval, and S. Schneider (2005)
  http://cmpg.unibe.ch/software/arlequin3
  Integrated software package for population genetics data analysis.

- *GeneScan Analysis Software version 3.7* (Applied Biosystems)
  https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=601382
  Software package for general fragment analysis. (Discontinued)

- *Genemapper Software v4.0* (Applied Biosystems)
  https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=600798
  Genotyping software package.

- *Sequencing Analysis Software v5.1.1* (Applied Biosystems)
  https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=601780
  Software for base calling and sequence analysis.

- *Grab-IT 2.59* (Ultra-Violet Products)
  Software for capturing images of gels/autoradiographs. (Discontinued)

- *GelWorks ID Intermediate v4.01* (Ultra-Violet Products)
  Software for analysing images of gels/autoradiographs. (Discontinued)

CHAPTER 3

STUDY OF ALLELIC VARIATION IN *MUC4* EXPRESSION

3 STUDY OF ALLELIC VARIATION IN *MUC4* EXPRESSION

3.1 INTRODUCTION

This chapter is concerned with the study of *cis*-acting allelic variation in *MUC4* mRNA expression. *MUC4* was chosen for two reasons: it is an example of a membrane-associated mucin and because it is clearly overexpressed in disease, for example in pancreatic cancer (Andrianifahanana et al. 2001; Balague et al. 1994; Kanno et al. 2006) and biliary tract cancer, as found in a parallel study that I was involved in (see Appendix I). This particular property of *MUC4* makes it an attractive diagnostic marker for pancreatic cancer (Andrianifahanana et al. 2001). The upregulation of *MUC4* in these conditions appears to occur at the mRNA level and is most probably a change in gene regulation in these diseases. As such, it was of interest to question whether there is any evidence of inter-individual differences in this response.

In order to test for the presence/extent of constitutive allelic variation in *MUC4* mRNA expression, the SBE method was chosen (see section 1.3.2.2.2.1). This was for a number of reasons, including the relatively low cost and successful use by other researchers to detect allelic variation in mRNA expression (Cheung et al. 2003; Yan et al. 2002)

Two potential confounders/limitations to this study needed to be addressed first; these are tandem repeat length variation and alternative splicing in *MUC4*.

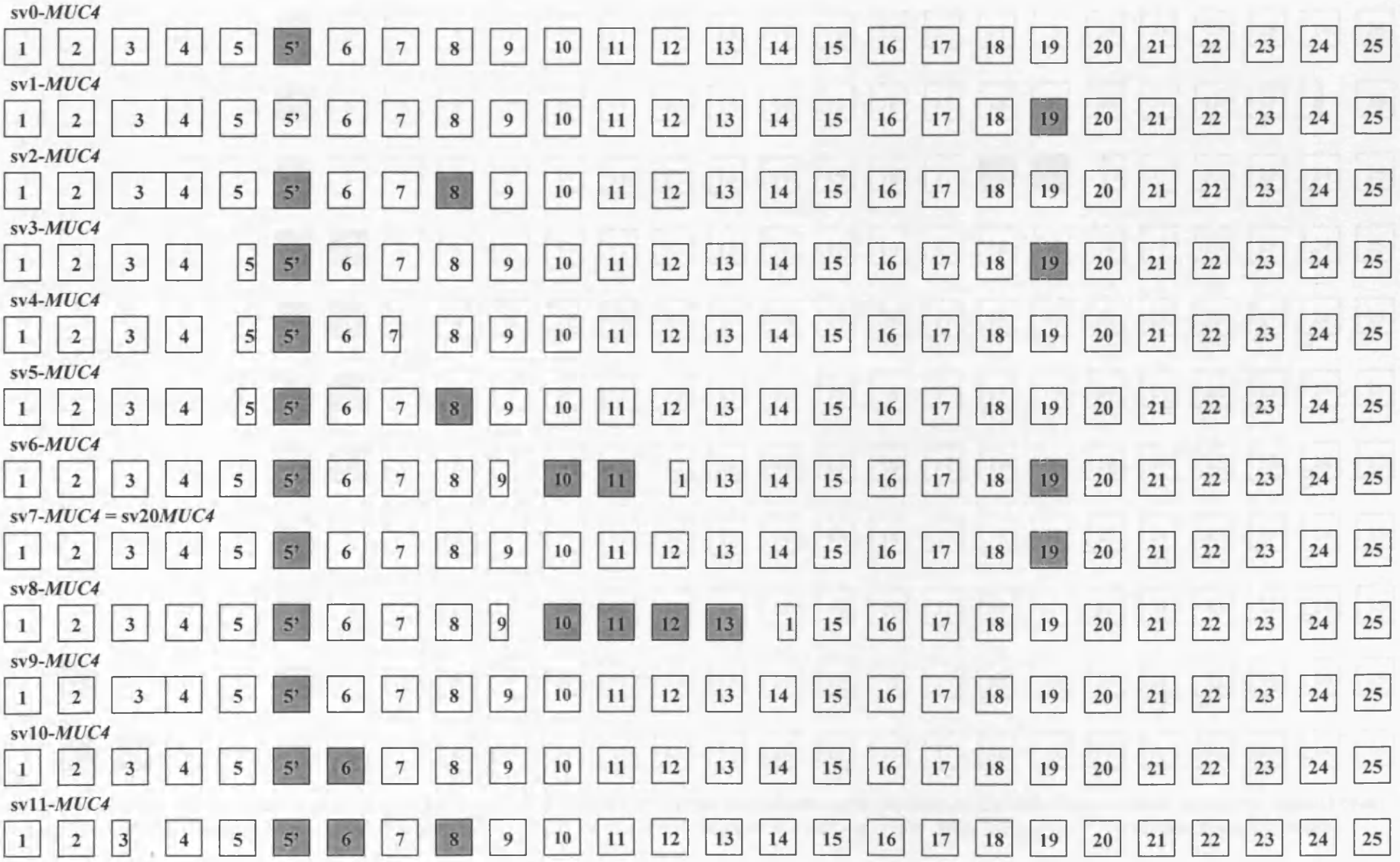
3.1.1 Tandem Repeat Length Variation in *MUC4*

MUC4 TR length varies highly between individuals and this could influence allelic variation in *MUC4* gene expression; the idea being that alleles with longer TR lengths would take more time to transcribe and hence longer TR length transcripts would be less abundant than short TR length transcripts. Thus, one of the major aims

of this project is to investigate the influence of TR length on allelic variation in *MUC4* expression.

3.1.2 Alternative Splicing in *MUC4*

It has been known for some time that *MUC4* undergoes extensive alternative splicing. 24 distinct transcripts of *MUC4* have been reported (Escande et al. 2002; Moniaux et al. 2000). These are generated by different combinations of alternative splicing, such as the alternative use of exons and the use of cryptic donor/acceptor splice sites as shown in Figure 3-1. The splice variants are named sv0-MUC4 to sv21-MUC4, MUC4/X and MUC4/Y. sv0-MUC4 is the predominant spliceoform and encompasses the full MUC4 cDNA sequence. MUC4/X and MUC4/Y are unique in that they have a complete deletion of exon 2 and hence if at all translated, will result in proteins totally lacking the O-glycosylated TR region (Moniaux et al. 2000). Other splice variants result from different combinations of splice events and these transcripts are shorter than sv0-MUC4.


sv0-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv1-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv2-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv3-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv4-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv5-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv6-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 1 13 14 15 16 17 18 19 20 21 22 23 24 25
sv7-MUC4 = sv20MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv8-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 1 15 16 17 18 19 20 21 22 23 24 25
sv9-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv10-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
sv11-MUC4
1 2 3 4 5 5' 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

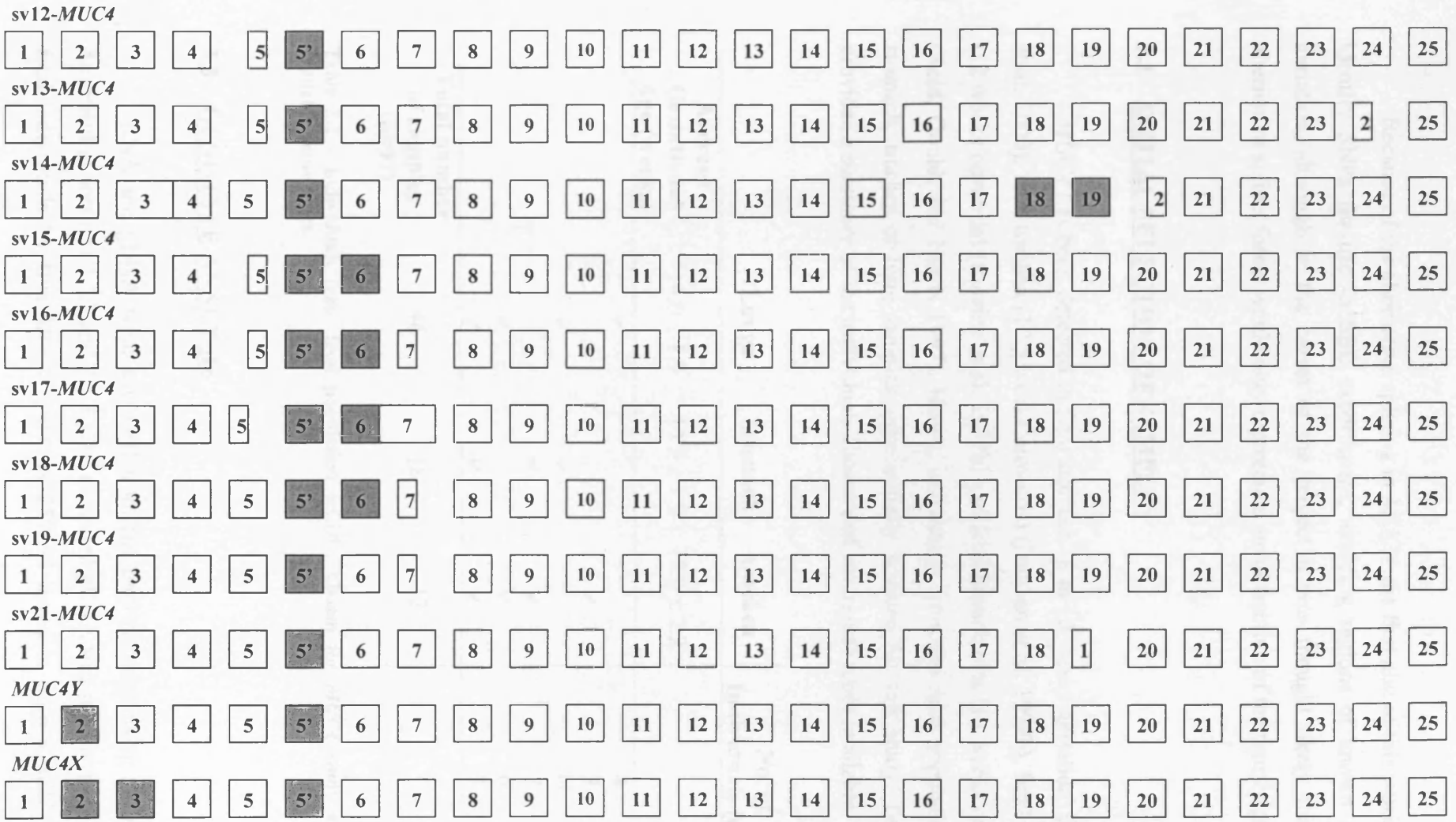


Figure 3-1 – Known *MUC4* splice variants. Clear and shaded boxes denote constitutive and deleted exons respectively. Extended boxes indicate insertions; reduced boxes deletions. The side of the box extended or deleted indicates the involvement of a cryptic acceptor site (left) or cryptic donor site (right). 5' is a cassette exon. Not to scale.

Because of the alternative splicing in *MUC4, t*he first aim of this work was to identify SNPs for use in SBE experiments, avoiding regions of known splicing variation, although at the outset of the project it was thought likely that these alternative spliceoforms would only represent a small fraction of the transcripts.

3.2 INITIAL SELECTION OF SAMPLES

MUC4 has been detected in fetal stomach (8 to 18 weeks gestation) (Buisine et al. 2000), fetal trachea (10.5 weeks onwards) (Buisine et al. 1999b), fetal bronchi (12 weeks onwards) (Buisine et al. 1999b) and fetal bronchioles (13 weeks onwards) (Reid, Gould, and Harris 1997). Hence, individuals from the fetal population with stomach, trachea or lung samples were initially selected for this study. Table 3-1 provides a summary of the individuals chosen and the tissues types available.

	Lungs	Stomach	trachea	No. of Individuals (n=51)
Average Gestational Age (weeks)	14.0 ± 1.9	13.9 ± 1.9	15.0 ± 2.2	
	✔			4
			✔	3
	✔	✔		34
		✔	✔	2
	✔		✔	6
	✔	✔	✔	2
Total number of samples (n=97)	46	38	13	

Table 3-1 - Individuals from fetal population initially chosen for *MUC4* study and their available tissue types

3.3 GENOTYPE ANALYSIS

RNA and gDNA were extracted from the fetal samples using the methods described in sections 2.3.2 and 2.3.3. The extracted gDNA was used to genotype the fetal individuals for five *MUC4* SNPs: rs2259292 found in exon 4, rs2259102 in

exon 6, rs2550240 in exon 7, rs2291652 in exon 23 and rs3205933 in exon 24. These SNPS were chosen because they had been successfully validated either in the HapMap databases or in previous work, could be used as markers for allelic variation expression analysis since they were exonic (see section 1.3.2.2) and because they were located in exons which were usually not spliced out (see Figure 3-1). Representative genotyping results for the five SNPs are shown in Figure 3-2, Figure 3-3, Figure 3-4, Figure 3-5 and Figure 3-6.

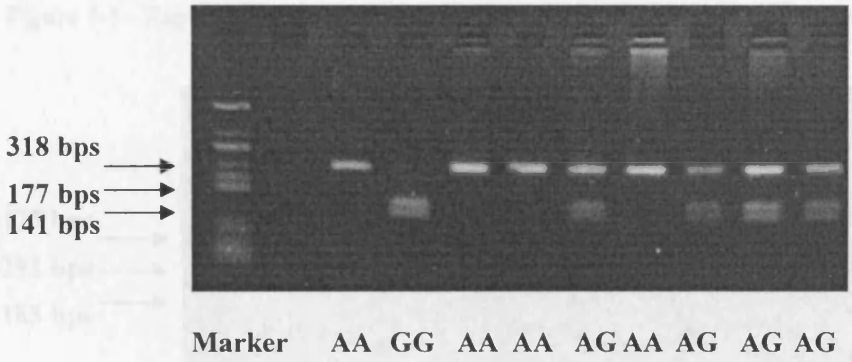


Figure 3-2 - Representative genotyping results for rs2259292 (exon 4). Digested using Sfo I.

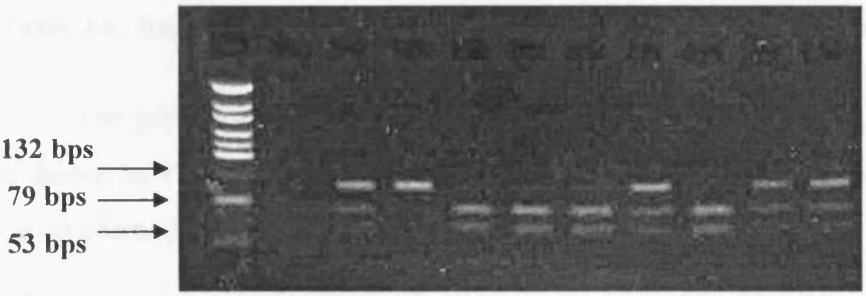


Figure 3-3 - Representative genotyping results for rs2259102 (exon 6). Digested using Mbo I.

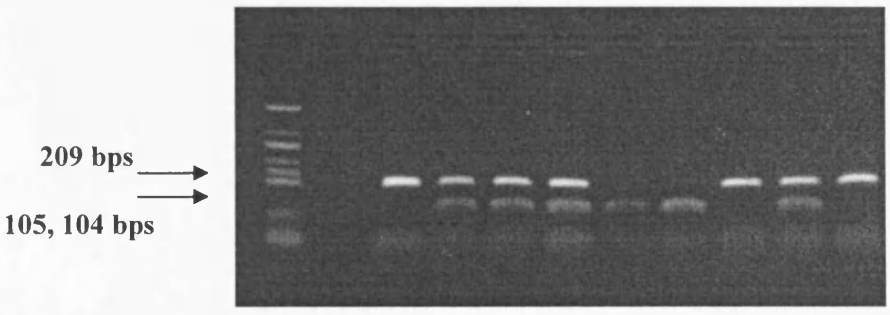


Figure 3-4 - Representative genotyping result for rs2550240 (exon 7). Digested using ScrF I.

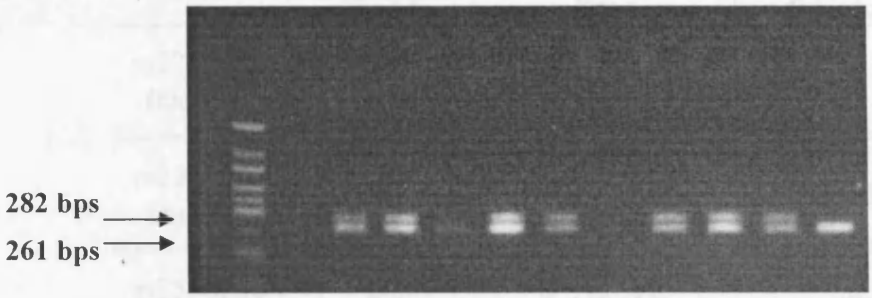


Figure 3-5 - Representative genotyping result for rs2291652 (exon 23). Digested using Mbo I.

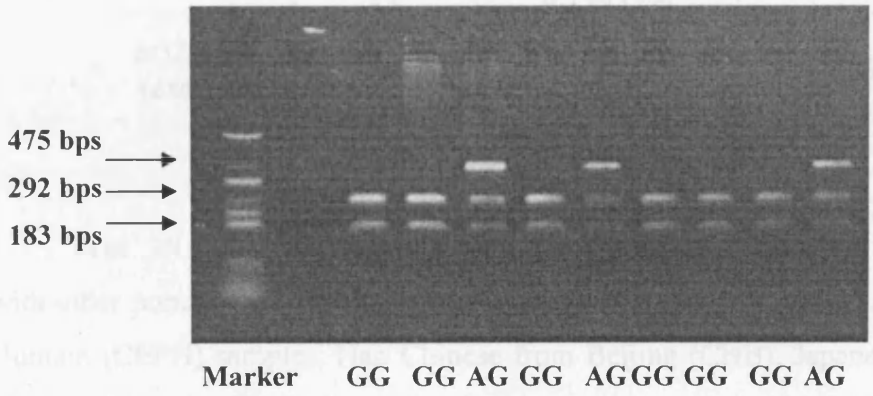


Figure 3-6 - Representative genotyping result for rs3205933 (exon 24). Digested using AlwN I.

The genotype frequencies were determined for each *MUC4* SNP and these are shown in Table 3-2. χ^2 tests were performed for each SNP to test for deviation from expected Hardy-Weinberg frequencies.

SNP	Genotypes	Fetal Genotype Freq	HW χ^2 P Value
rs2259292 (exon 4)	AA AG GG	0.300 (15) 0.600 (30) 0.100 (5)	0.210
rs2259102 (exon 6)	CC CG GG	0.020 (1) 0.314 (16) 0.667 (34)	0.852
rs2550240 (exon 7)	CC CG GG	0.340 (17) 0.460 (23) 0.200 (10)	0.909
rs2291652 (exon 23)	CC CT TT	0.373 (19) 0.431 (22) 0.196 (10)	0.736
rs3205933 (exon 24)	AA AG GG	0.000 (0) 0.143 (7) 0.857 (42)	0.865

Table 3-2 – *MUC4* genotype frequencies in fetal population

The *MUC4* SNPs allele frequencies in the fetal population were compared with other populations in the HapMap database: Centre d'Etude du Polymorphisme Humain (CEPH) samples, Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT) and Yoruba (YOR) from Nigeria. The results are shown in Table 3-3:

SNP	Allele	Allele Frequencies				
		Fetal	CEPH	CHB	JPT	YOR
rs2259292 (exon 4)	G	0.400	-	-	-	0.356
rs2259102 (exon 6)	C	0.176	0.108	0.344	0.261	0.000
rs2550240 (exon 7)	G	0.441	0.533	0.444	0.318	0.875
rs2291652 (exon 23)	T	0.412	0.542	0.578	0.689	0.025
rs3205933 (exon 24)	A	0.071	-	-	-	-

Table 3-3 - *MUC4* SNP allele frequencies in fetal population compared to other HapMap populations. Allele shown for each SNP is the minor allele in fetal population. - : Data not available

The allele frequencies in the fetal population are the most similar, for the three SNPs for which there are data, to those from the HapMap CEPHs. The HapMap CEPH samples were collected in the 1980s in the US from individuals with western and northern European ancestry. Unfortunately, information regarding the

ethnicity of the fetal samples was not available, but it is likely that the majority of the individuals from the fetal population are of northern/western European ancestry, although the sample set, collected from the London area, undoubtedly includes representatives of non-European groups.

There is however no significant deviation from Hardy-Weinberg equilibrium, suggesting no very serious problem with population admixture.

3.4 DETERMINATION OF *MUC4* TANDEM REPEAT LENGTH

In order to conserve fetal lung, stomach and trachea samples for RNA extraction, genomic DNA was extracted (using the method described in 2.3.4) from intestine or liver samples whenever these were available.

MUC4 tandem repeat allele lengths were determined by Southern blotting. Figure 3-7 shows a photo of one of the gels taken before blotting (see section 2.4.2) whilst Figure 3-8 shows its matching autoradiograph. The distance from the wells of each Raoul marker band was determined from the photo and their positions were marked on the autoradiographs. The *MUC4* TR bands were then sized using the method described in section 2.4.4.6.

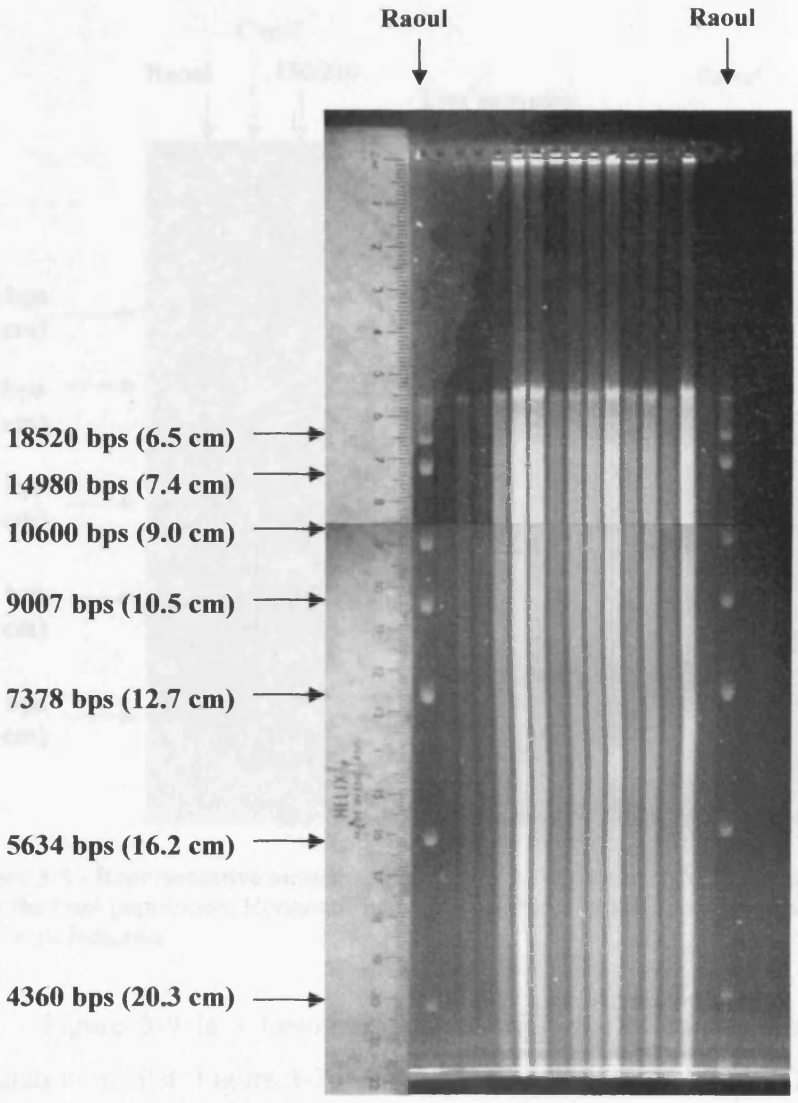


Figure 3-7 – Photo of gel of *Pvu II* digested gDNA. Size of Raoul marker bands and distance from wells indicated.

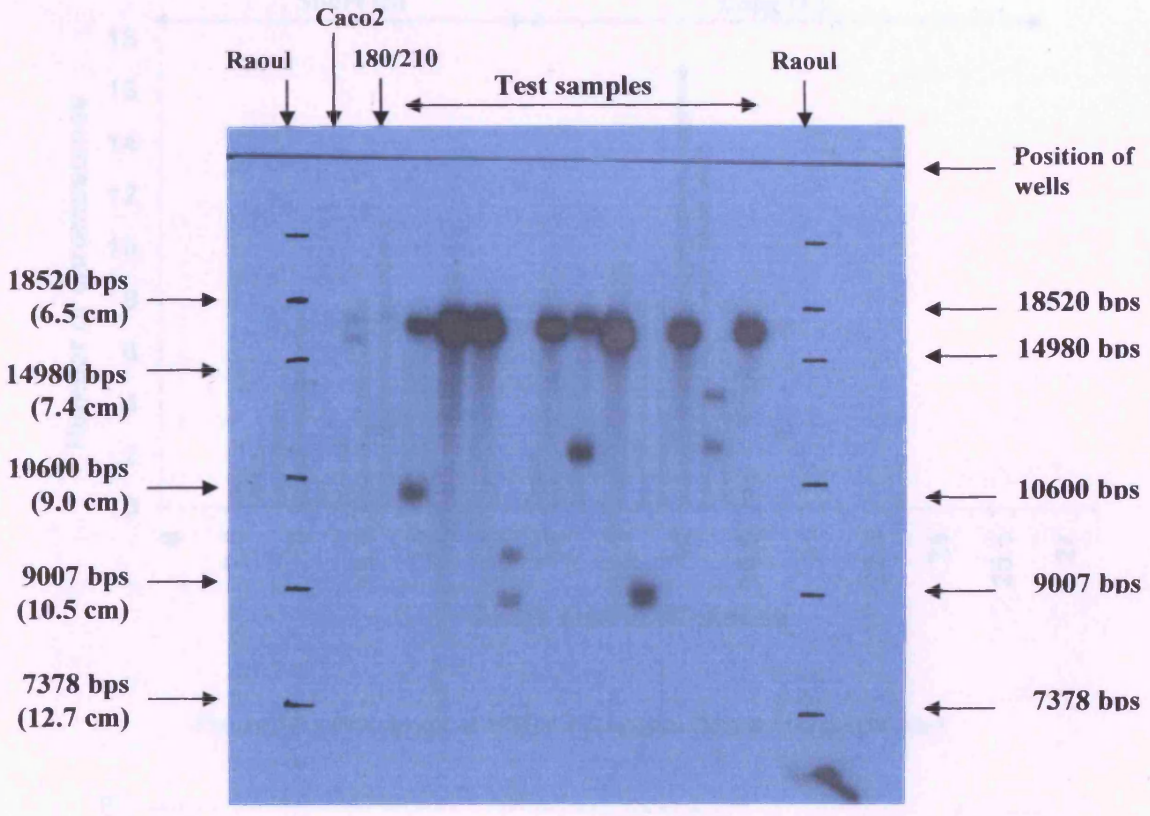


Figure 3-8 - Representative autoradiograph used to determine *MUC4* TR lengths of individuals from the fetal population. Horizontal bars show positions of Raoul marker bands. Sizes and distance from wells indicated.

Figure 3-9 is a histogram of the *MUC4* TR lengths detected in the fetal population, whilst Figure 3-10 shows the distribution of *MUC4* TR lengths in a UK population studied previously by Vinall and colleagues (Vinall, Pratt, and Swallow 2000).

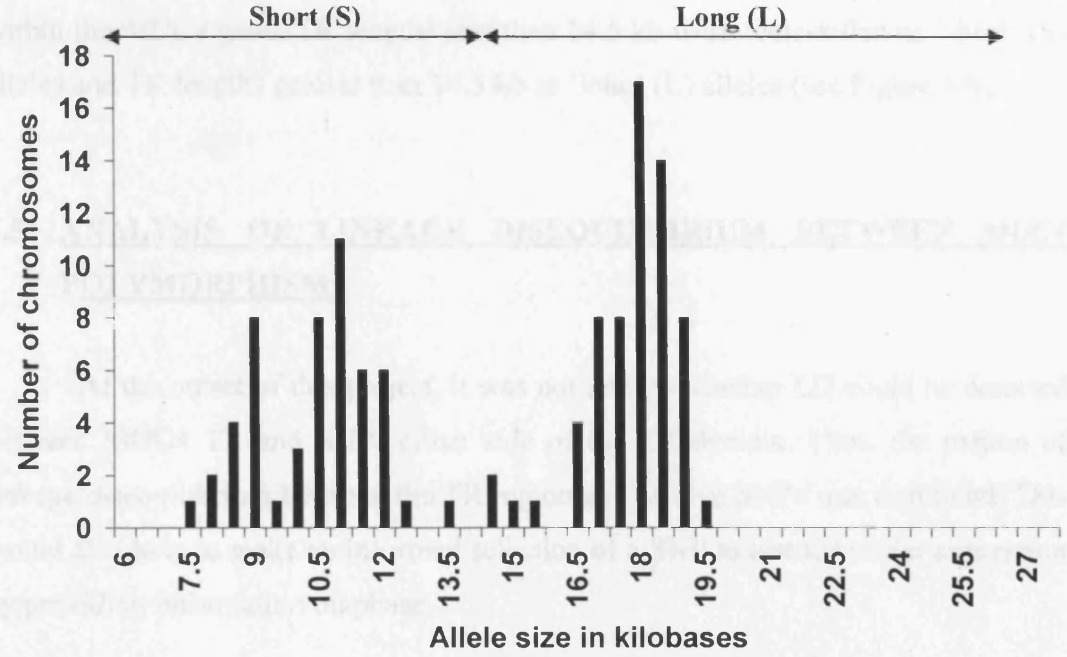


Figure 3-9 - Histogram of *MUC4* TR lengths (kb) in fetal population

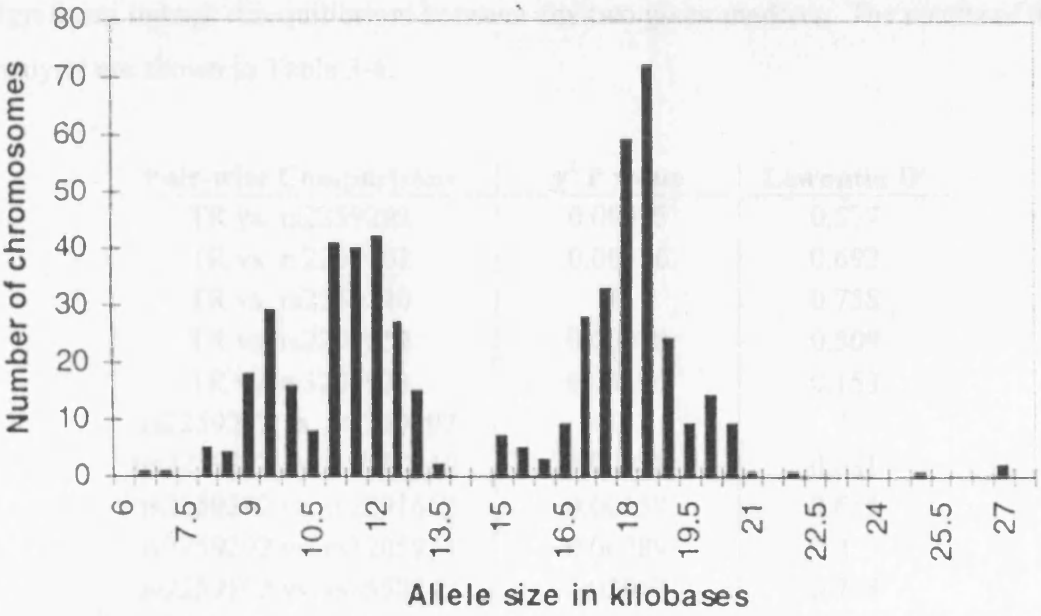


Figure 3-10 - Histogram of *MUC4* TR lengths (kb) in UK population (Vinall, Pratt, and Swallow 2000)

As can be seen, in both cases, the TR lengths appear to fit a tri-modal distribution with three main groups of TR lengths, the first ranging from 7.5 to 9.5 kb, the second ranging from 10 to 14 kb and the third most distinct group from 14.5 to 19.5 kb. Thus, in order to better analyse TR lengths in relation to SNP markers

within the *MUC4* gene, TR lengths less than 14.5 kb were reclassified as 'short' (S) alleles and TR lengths greater than 14.5 kb as 'long' (L) alleles (see Figure 3-9).

3.5 ANALYSIS OF LINKAGE DISEQUILIBRIUM BETWEEN *MUC4* POLYMORPHISMS

At the outset of this project, it was not known whether LD could be detected between MUC4 TR and SNPs either side of the TR domain. Thus, the pattern of linkage disequilibrium between the TR region and all five SNPs was examined. This would also help to make an informed selection of a SNP to test for allelic expression by providing information on phase.

The *MUC4* SNP genotypes and TR LS length alleles were used to calculate D' values between each marker and a χ^2 test P value, which indicates whether there is significant linkage disequilibrium between any two given markers. The results of the analysis are shown in Table 3-4.

Pair-wise Comparisons	χ^2 P value	Lewontin D'
TR vs. rs2259292	0.00005	0.577
TR vs. rs2259102	0.00956	0.692
TR vs. rs2550240	0	0.758
TR vs. rs2291652	0.00088	0.509
TR vs. rs3205933	0.73496	0.153
rs2259292 vs. rs2259102	0	1
rs2259292 vs. rs2550240	0.00058	0.681
rs2259292 vs. rs2291652	0.00159	0.655
rs2259292 vs. rs3205933	0.06089	1
rs2259102 vs. rs2550240	0.02017	0.768
rs2259102 vs. rs2291652	0.13935	0.512
rs2259102 vs. rs3205933	0.6144	0.647
rs2550240 vs. rs2291652	0	0.620
rs2550240 vs. rs3205933	0.01688	1
rs2291652 vs. rs3205933	0.31702	0.495

Table 3-4 - Measures of linkage disequilibrium between *MUC4* TR and SNP markers.
Significant LD (P value <0.05) between markers shown in bold.

The degree of linkage disequilibrium (LD) between each of the *MUC4* markers is depicted in a graphical form in Figure 3-11. The diagram is to scale and provides an indication of the physical distance between each marker.

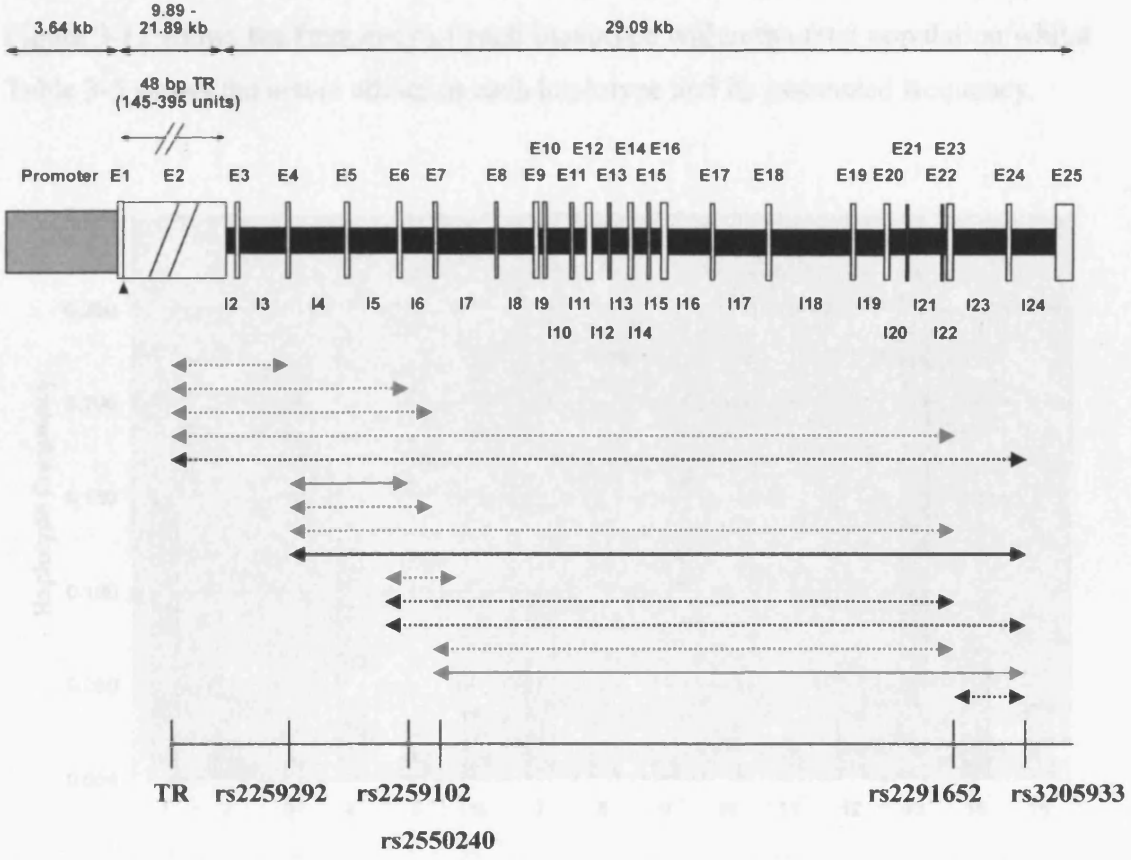


Figure 3-11 – *MUC4* scale gene diagram showing LD between *MUC4* TR and SNPs. Red lines indicate significant χ P values (<0.05). Dotted lines indicate D′ values less than 1.

There seems to be a striking degree of LD within the *MUC4* gene. The most prominent associations appear to be those between the TR domain and the five SNP markers. With the exception of rs3205933, which is some 30 kb (possibly much further depending on the actual number of TR units) away, all the other SNPS exhibit a statistically significant association with the TR domain. The breakdown of LD between the TR domain and rs3205933 is probably due to recombination between rs2291652 and rs3205933, as evident from the lack of association between these two SNPs. There also appear to be significant associations between many pairs of SNPs, although the D′ values suggest that the strength of these associations varies quite a bit.

3.6 HAPLOTYPE ANALYSIS

Given that rs2259292, rs2259102, rs2550240 and rs2291652 were significantly associated with the TR domain, genotype data from these four SNPs and the TR domain were used to construct haplotypes using the *Phase v2.1* program. Figure 3-12 shows the frequency of each haplotype within the fetal population whilst Table 3-5 shows the actual alleles in each haplotype and its associated frequency.

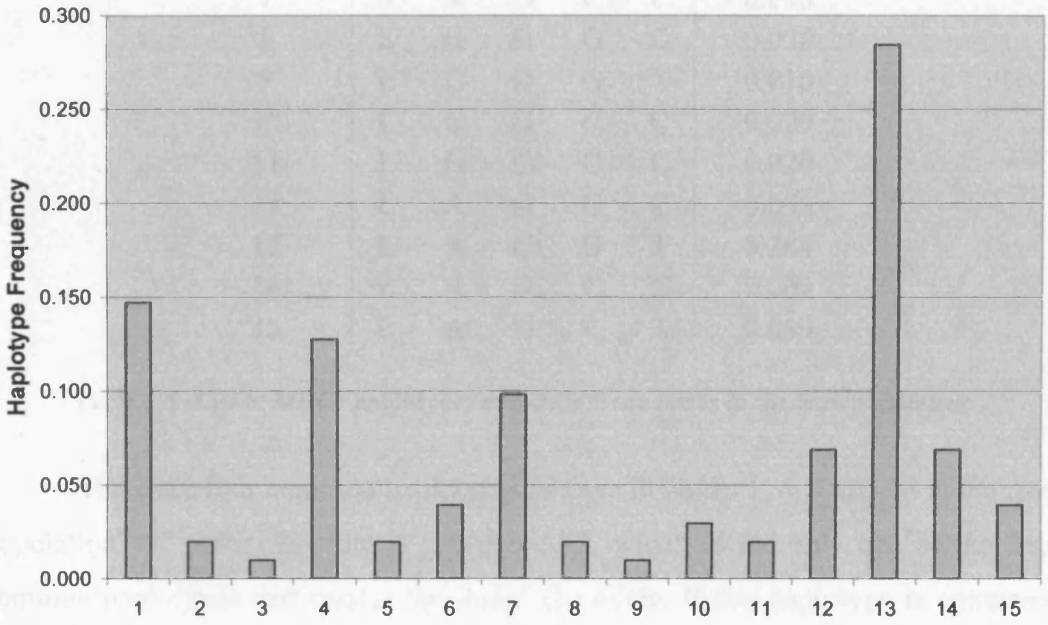


Figure 3-12 - *MUC4* haplotypes and their frequencies in the fetal population determined by *Phase* program

Haplotype Identifier	TR	rs2259292	rs2259102	rs2550240	rs2291652	Haplotype Freq
1	**S**	**G**	**G**	**C**	**C**	**0.147**
2	S	G	G	C	T	0.020
3	S	G	C	G	C	0.010
4	**S**	**G**	**C**	**C**	**C**	**0.127**
5	S	G	C	C	T	0.020
6	S	A	G	G	T	0.039
7	**S**	**A**	**G**	**C**	**C**	**0.098**
8	L	G	G	G	C	0.020
9	L	G	G	G	T	0.010
10	L	G	G	C	C	0.029
11	L	G	C	C	C	0.020
12	L	A	G	G	C	0.069
13	**L**	**A**	**G**	**G**	**T**	**0.284**
14	L	A	G	C	C	0.069
15	L	A	G	C	T	0.039

Table 3-5 -List of *MUC4* haplotypes and their frequencies in the fetal population

There are four common haplotypes (shown in bold): 1, 4, 7 and 13 in the fetal population. Of particular note is haplotype 13, which is the only one of the four common haplotypes that carries the 'long' (L) allele. If this haplotype is compared with the other three common haplotypes, it seems that the rs2550240 and rs2291652 SNPs could function effectively as tags for the TR domain (G associated with L for rs2550240 and T with L for rs2291652).

3.7 RT PCR *MUC4* EXON EXPRESSION ANALYSIS

The next task was to amplify the regions in *MUC4* cDNA containing the SNPs selected - regions that had been selected in order to avoid alternative splicing. Early results showed that products were not obtained in all cases. Thus, there was an indication of alternative splicing in these regions of *MUC4* as well. In order to gain a better understanding of the extent of alternative splicing of these regions in fetal tissue, all samples (not just heterozygotes) were amplified for all five SNP regions.

RT PCR using ribosomal RNA specific primers were used as positive controls for cDNA quality, since rRNA should be constitutively expressed. RT PCR was conducted using a certain number of rounds of amplification in order to keep the reaction as far as possible in the linear range (35 rounds for the test PCRs and 26 rounds for the rRNA controls). Figure 3-13 is a representative result from a selection of lung samples and shows significant differences in relative expression of the various exons in different individuals.

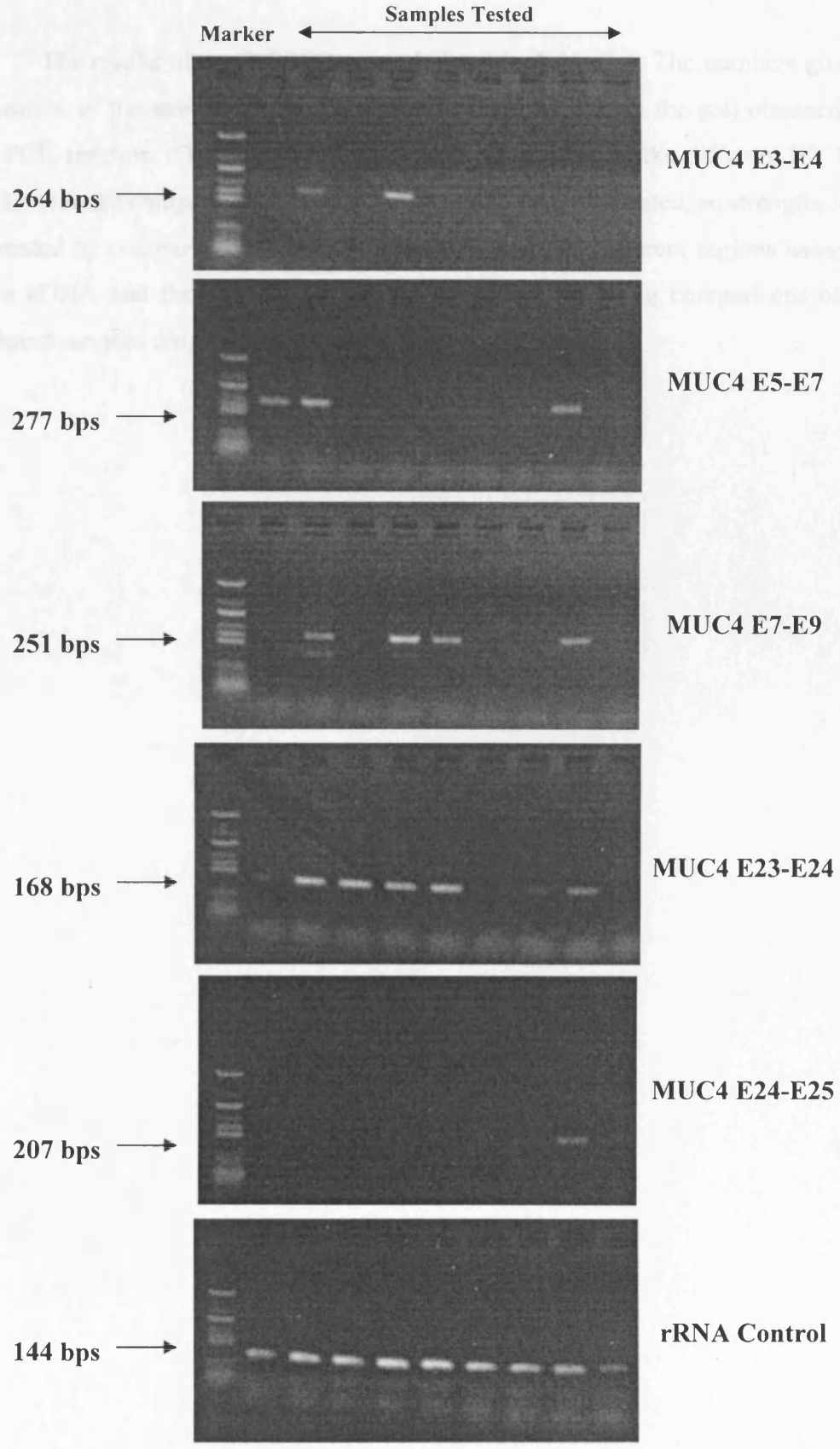


Figure 3-13 - Representative RT PCR results showing variability in expression of *MUC4* exons in a series of lung samples. 10 µl of RT PCR product electrophorised for each sample.

The results of the RT PCRs are tabulated in Table 3-6. The numbers give an indication of the strength of the PCR product (estimated from the gel) obtained per RT PCR reaction. The differences in relative efficiencies of the different RT PCR reactions make comparisons across different reactions complicated, so strengths were estimated by comparing the relative product strength for different regions using the same cDNA and the differences between samples by making comparisons of the different samples amplified in the same PCR experiments.

		Lungs					Stomach					Trachea				
	Target SNP RT PCR	rs2259292 Ex3-4	rs2259102 Ex5-7	rs2550240 Ex7-8	rs2291652 Ex23-24	Rs3205933 Ex 23-25	rs2259292 Ex3-4	rs2259102 Ex5-7	rs2550240 Ex7-8	rs2291652 Ex23-24	rs3205933 Ex23-25	rs2259292 Ex3-4	rs2259102 Ex 5-7	rs2550240 Ex7-8	rs2291652 Ex23-24	rs3205933 Ex 23-25
SAMPLES	4	0	2	0	2	1	2	3	4	3	4	-	-	-	-	-
	8	-	-	-	-	-	-	-	-	-	-	3	4	4	4	4
	22	0	0	3	2	3	-	1	2	3	3	-	-	-	-	-
	23	-	0	0	3	3	-	0	0	4	4	-	-	-	-	-
	25	0	0	0	2	3	3	0	4	4	4	-	-	-	-	-
	27	3	3	4	3	4	4	4	4	4	4	-	-	-	-	-
	28	0	0	0	0	0	4	4	0	4	0	-	-	-	-	-
	30	-	-	-	-	-	0	4	4	4	3	-	4	4	4	4
	31	0	-	0	1	2	-	-	-	-	-	3	3	4	4	3
	32	-	-	-	-	-	-	-	-	-	-	4	2	4	4	4
	36	0	1	0	4	4	0	0	0	4	0	-	-	-	-	-
	39	0	0	0	0	0	0	4	0	2	3	-	-	-	-	-
	40	-	4	4	4	4	4	4	0	3	4	-	-	-	-	-
	49	0	0	1	2	1	0	0	0	2	2	-	-	-	-	-
	50	0	0	2	2	2	0	1	1	3	3	-	-	-	-	-
	51	0	0	1	0	2	0	1	4	3	4	-	-	-	-	-
	54	-	3	4	3	0	2	4	4	3	4	-	-	-	-	-
	57	0	4	0	0	0	4	4	4	4	4	-	-	-	-	-
	58	0	2	0	0	3	1	2	4	3	4	-	-	-	-	-
	60	1	3	4	3	4	1	4	4	2	3	-	-	-	-	-
	63	-	3	4	3	1	-	-	-	-	-	2	3	4	3	3
	64	0	1	3	3	3	-	4	4	3	3	-	-	-	-	-
	66	0	1	1	3	3	-	2	4	4	4	-	-	-	-	-
	67	0	0	2	1	3	2	4	4	4	4	-	-	-	-	-
	68	-	0	0	2	1	0	0	-	2	-	-	-	-	-	-
	69	0	0	0	0	-	1	4	-	3	-	-	-	-	-	-
	71	0	-	2	2	2	2	-	1	2	1	-	-	-	-	-
	72	0	0	0	1	1	4	4	-	4	-	-	-	-	-	-
	73	-	0	2	4	3	0	2	1	3	4	-	-	-	-	-
	74	0	0	0	2	1	0	1	0	3	4	-	-	-	-	-
	82	-	-	0	3	4	3	4	4	4	4	-	-	-	-	-
	83	0	4	0	0	0	0	4	4	4	4	-	-	-	-	-
	85	-	0	0	0	0	0	1	4	3	4	-	-	-	-	-
	86	1	0	3	2	3	-	-	-	-	-	-	2	4	4	4
	87	-	-	-	-	-	-	-	-	-	-	4	4	4	4	4
	88	0	-	1	3	2	-	-	-	-	-	4	4	4	4	4
	89	-	-	-	-	-	-	-	-	-	-	0	3	4	3	1
	90	-	0	1	-	2	-	-	-	4	4	-	-	-	-	-
	91	1	2	2	3	3	0	0	2	3	3	-	-	-	-	-
	92	0	0	1	2	1	0	0	0	0	0	-	-	-	-	-
	94	0	0	1	4	2	-	-	-	-	-	4	4	4	4	4
	95	-	0	0	0	0	0	0	0	0	0	-	-	-	-	-
	97	0	0	1	2	3	-	-	-	-	-	-	-	-	-	-
	98	0	-	0	2	3	1	2	3	3	3	4	3	4	4	4
	100	0	0	2	2	-	-	-	2	3	-	-	-	-	-	-
	103	-	-	0	2	0	-	-	-	-	-	3	0	3	3	3
	104	0	0	0	3	3	3	4	4	4	4	-	-	-	-	-
	110	-	0	0	0	1	0	3	3	3	4	-	-	-	-	-
	112	0	0	0	1	1	-	-	-	-	-	-	-	-	-	-
	115	-	0	0	0	0	-	-	-	-	-	-	-	-	-	-
	116	-	0	0	2	1	-	-	-	-	-	-	-	-	-	-

Table 3-6 - RT PCR *MUC4* exon expression analysis. 0: no visible product; 1: trace; 2: clearly visible product; 3: strong product; 4: very strong product; -: sample not available or not tested

As can be seen, there is a great deal of inter-sample variability in *MUC4* exonic expression. In general, the trachea samples tended to produce strong products across all five RT PCR reactions. The expression in stomach was more varied, with some samples producing strong products for some exons and weak or undetectable products in others. The lung samples showed large inter-individual variability and had the weakest expression of *MUC4* overall, with many of the samples failing to produce any detectable products for the majority of the RT PCR reactions. In addition, there seemed to be relatively lower expression of the 5′ exons, with the 3′ exons being expressed more frequently and often more strongly, although it is hard to exclude this being due to relative differences in PCR efficiency of the different reactions. There was no correlation between gestational age and the RT PCR products obtained.

However, the lack of expression of some exons, but strong expression of others in some of the individuals provides clear evidence of additional previously unreported alternative splicing in *MUC4*, which is different in different samples and therefore suggests inter-individual variability in this.

3.8 TESTING FOR ALLELIC VARIATION IN *MUC4* mRNA EXPRESSION BY SBE METHOD

Using the information from the work described in the preceding sections of this chapter, rs2291652 was chosen as best marker SNP for testing for allelic variation in *MUC4* gene expression. This SNP was chosen for the following reasons: it appeared as though RT PCR templates would be obtainable for all samples; it has a high heterozygosity, increasing the number of informative individuals available for testing by SBE (see section 3.3); it is in significant LD with the TR domain (see section 3.5) and can function as a tag for the 'long' TR allele (see section 3.6).

The initial strategy was to compare rs2291652 heterozygotes with two long TR alleles or two short alleles against rs2291652 heterozygotes with both L and S TR alleles.

If the TR length were having a substantial effect, we would expect to see a greater difference in allelic expression in the TR LS individuals compared to the TR homozygotes.

Given that the T allele of rs2291652 is associated with the 'long' TR allele, the hypothesis at this stage was that if TR length were exerting a strong influence on *MUC4* mRNA allelic variation, the SBE experiments would show that the C allele (associated with 'short' TR alleles) was usually expressed at a higher level than the T allele.

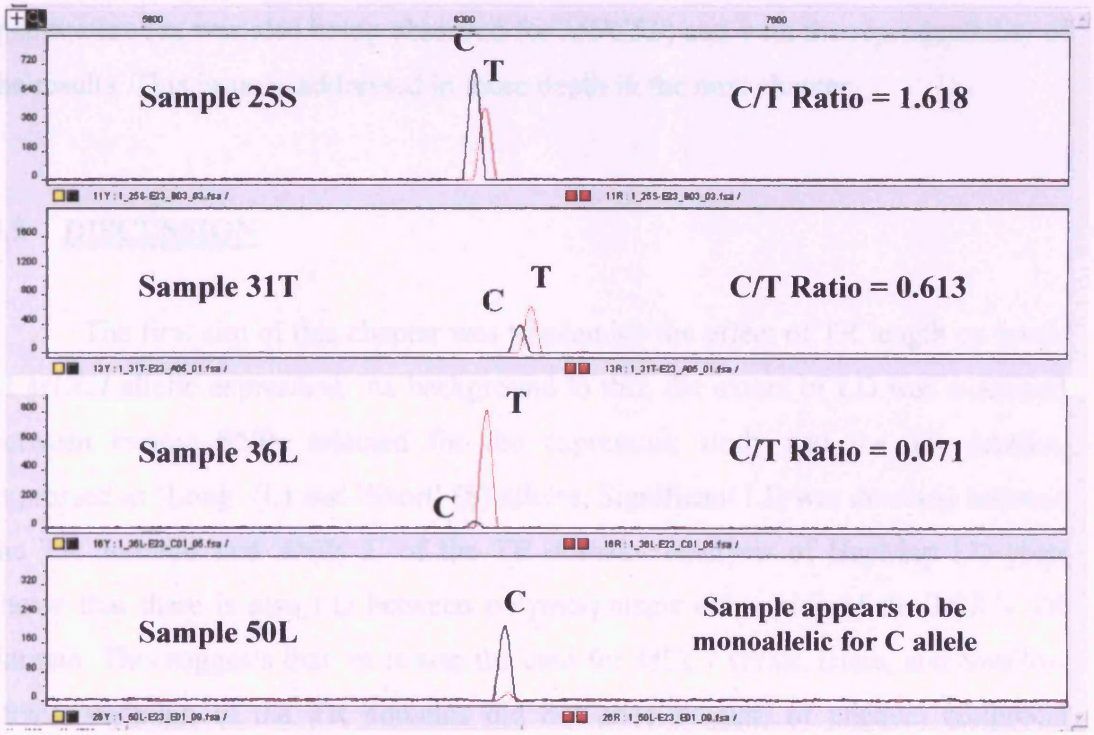


Figure 3-14- Representative preliminary SBE results for rs2291652. Four different cDNA samples from heterozygous individuals shown.

Preliminary SBE results showed an imbalance of allelic expression with the C allele more highly expressed in some samples but the opposite effect observed in a few others (examples shown in Figure 3-14). In a few of the samples, there was almost mono-allelic expression. The allelic differences observed were much greater than that expected for allelic variation due to TR length differences or even regulatory polymorphism.

On completion of the general semi-quantitative RT PCR, which was done whilst this was in progress, it became clear that in a few cases, no PCR product was obtainable for this fragment (Ex 23-24, see Table 3-6) even though other regions of *MUC4* were expressed.

Both these observations therefore suggested that there is genetic variation in alternative splicing of *MUC4* that is influencing allelic variation in *MUC4* mRNA expression.

It was also clear from the results however, that there were issues with the quantitation (as was also being observed for *MUC5B*) and with the reproducibility of the results. This issue is addressed in more depth in the next chapter.

3.9 DISCUSSION

The first aim of this chapter was to examine the effect of TR length on levels of *MUC4* allelic expression. As background to this, the extent of LD was examined between exonic SNPs selected for the expression study and the TR domain, expressed as 'Long' (L) and 'Short' (S) alleles. Significant LD was detected between the TR domains and SNPs 3′ of the TR domain. Analysis of HapMap LD plots shows that there is also LD between polymorphisms either side of the *MUC4* TR domain. This suggests that, as is also the case for *MUC1* (Pratt, Islam, and Swallow 1996), variation of the TR domains did not arise because of unequal reciprocal recombination at meiosis, but rather by non-reciprocal events.

The work in this chapter highlights the challenges faced when investigating allelic variation in gene expression, in particular the difficulties posed by alternative splicing.

MUC4 had been reported to exhibit a high degree of alternative splicing; this was determined using RT PCR on cDNA derived from pancreatic cancer cell lines (Choudhury et al. 2000; Moniaux et al. 2000) as well as from epithelial tissue

obtained from stomach, trachea, salivary glands and other organs (Moniaux et al. 2000). Spliceoforms were identified by using different primers spanning the length of the gene and by comparing the products obtained to the genomic sequence. However, it was not clear that these alternative spliceoforms were major components of the *MUC4* transcriptome.

Here we found evidence of even more extensive *MUC4* alternative splicing, obtained by RT PCR using primers specifically chosen to avoid regions of known splicing. There were slight differences between the different tissues examined, but more note-worthy was the large inter-sample variability in *MUC4* exon expression, suggesting the possibility of inter-individual differences. The 3′ most end of the *MUC4* transcript was less affected by splicing than the rest of the regions examined and was therefore selected for the mRNA allelic variation studies. Nevertheless, there was some variability in expression and the idea that there was *cis*-acting variation in splicing was strongly suggested by these experiments.

Although the SBE results were encouragingly interesting, the technical difficulties encountered, plus the fact that there were four potential genetic sources of variation: *cis*-acting regulatory polymorphism; TR; possible tissue-specificity and alternative splicing, led us to conclude that untangling the relative roles of these different effects would be very difficult, and that this might be better done with different sample sets and by using a different approach. For example, alternative splicing in *MUC4* could be further studied by using microarrays and allelic variation in mRNA expression by a method that does not rely on RNA, such as HaploChIP (see section 1.3.2.2.2.4). In view of the extreme complexity of *MUC4*, the rest of the work described in this thesis is focussed on *MUC5B*.

As to whether TR length influences *MUC4* expression, it seems that this question is unanswerable at present in view of the extensive alternative splicing observed. It is relevant to note that in another parallel study in the group, the effect of TR length on *MUC1* expression was addressed and there was no apparent correlation between the two (Ng 2007).

The significance of the large number of different *MUC4* splice variants is subject to speculation. Many of the *MUC4* splice events reported previously result in the introduction of stop codons and thus if translated, many of the different spliceoforms would produce an identical apomucin. This limits the total number of different putative MUC4 proteins. It is possible that the potential diversity of transcripts could help to modulate the properties of MUC4. Alternatively, the various spliceoforms might serve to reduce the level of expression of the main sv0-MUC4 isoform.

As for *MUC4*, a number of splice variants have been reported for *MUC1* (reviewed in (Imbert et al. 2006) and some are directly influenced by a single exonic SNP 8 nucleotides from the start of exon 2 (Ng 2007). In addition, some variation in the levels of these transcripts was observed. It could well be that there are such variations in *MUC4*. At present our limited observations were from fetal tissue and may not apply to adults, but these suggest that this is something to be watchful for in disease studies.

Whatever the situation, it seems hardly surprising that there have been difficulties in getting good antibody reagents and that there are discrepancies in expression studies for *MUC4*. This issue needs to be fully understood if *MUC4* is to be a good disease marker.

CHAPTER 4

DEVELOPMENT OF SINGLE-BASE EXTENSION METHOD FOR DETECTING ALLELIC DIFFERENCES IN mRNA EXPRESSION

4 DEVELOPMENT OF SINGLE-BASE EXTENSION METHOD FOR DETECTING ALLELIC DIFFERENCES IN mRNA EXPRESSION

4.1 INTRODUCTION

This chapter is concerned with the development of the single-base extension (SBE) method to test for allelic variation in mRNA expression of the MUC genes.

Preliminary experiments showed that there was an allelic bias in the representation of the different alleles in the gDNA of heterozygous individuals, specifically, an over representation of A over G in marker SNP rs2672785, T over C in rs2075853 and C over T in rs2075859. Representative SBE traces are shown in Figure 4-1, Figure 4-2 and Figure 4-3.

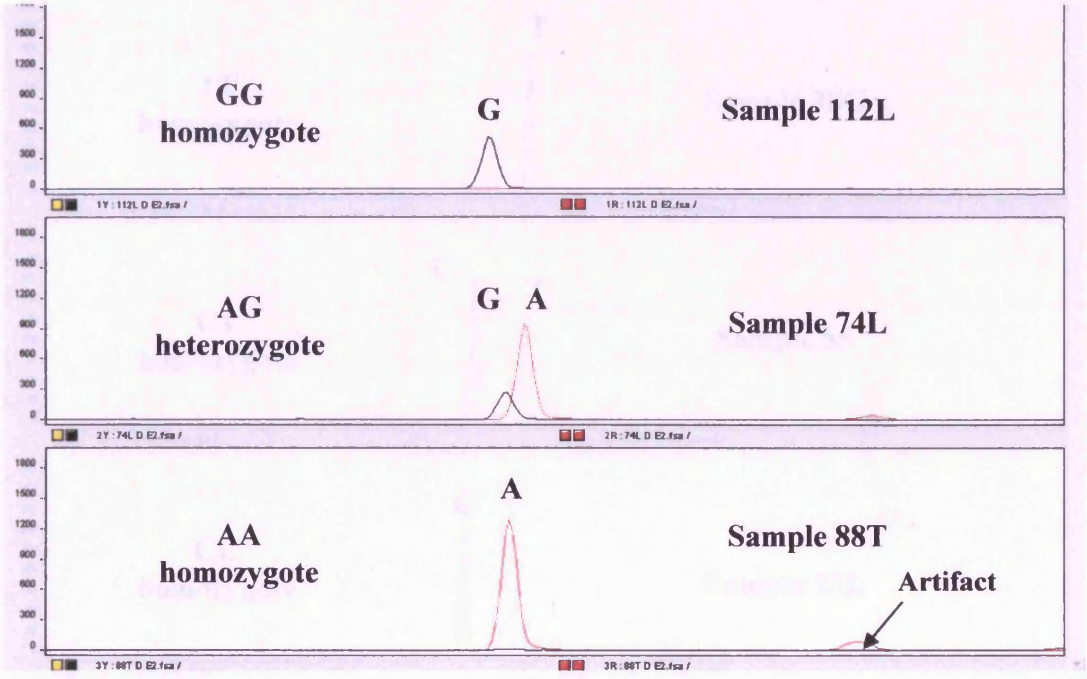


Figure 4-1 - Representative SBE traces from gDNA for rs2672785. One example of each genotype shown, confirming that SBE was detecting the alleles accurately. Samples run on *ABI 3100.*

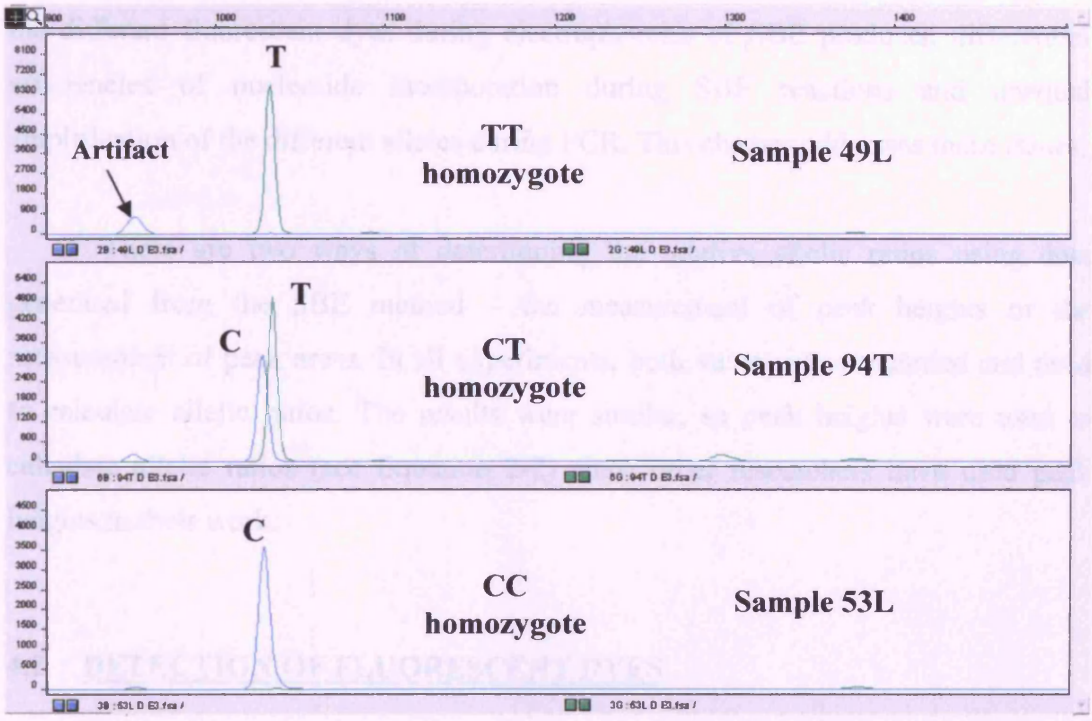


Figure 4-2 -Representative SBE traces from gDNA for rs2075853. One example of each genotype shown, confirming that SBE was detecting the alleles accurately. Samples run on *ABI 3100.*

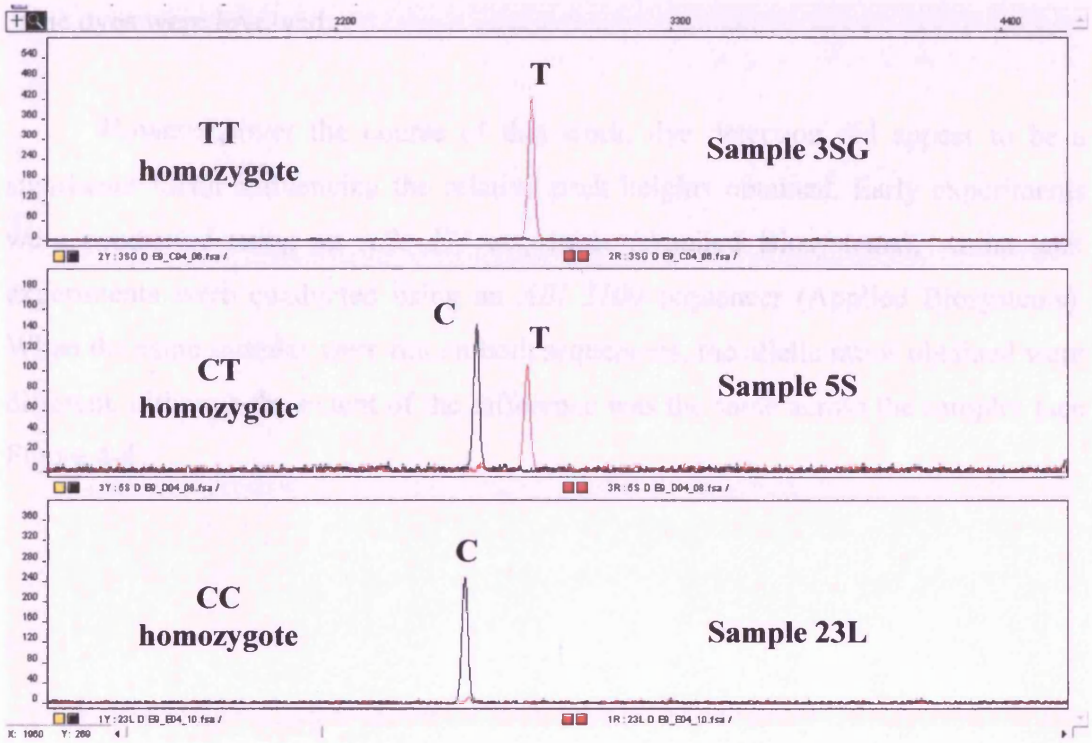


Figure 4-3 - Representative SBE traces from gDNA for rs2075859. One example of each genotype shown, confirming that SBE was detecting the alleles accurately. Samples run on *ABI 3100.*

Previously, three technical sources of unequal representation of allele-specific products had been highlighted (Moskvina et al. 2005). These are unequal detection of

the different fluorescent dyes during electrophoresis of SBE products, differential efficiencies of nucleotide incorporation during SBE reactions and unequal amplification of the different alleles during PCR. This chapter addresses these issues.

There are two ways of determining the relative allelic ratios using data generated from the SBE method – the measurement of peak heights or the measurement of peak areas. In all experiments, both values were recorded and used to calculate allelic ratios. The results were similar, so peak heights were used to calculate allelic ratios (see Equation 2-2) since other researchers have used peak heights in their work.

4.2 DETECTION OF FLUORESCENT DYES

At first sight, it appeared that dye detection was not a major issue because different relative signal strengths were obtained for different SNPs, even when the same dyes were involved.

However, over the course of this work, dye detection did appear to be a significant factor influencing the relative peak heights obtained. Early experiments were conducted using an *ABI 377* sequencer (Applied Biosystems), whilst later experiments were conducted using an *ABI 3100* sequencer (Applied Biosystems). When the same samples were run on both sequencers, the allelic ratios obtained were different, although the extent of the difference was the same across the samples (see Figure 4-4).

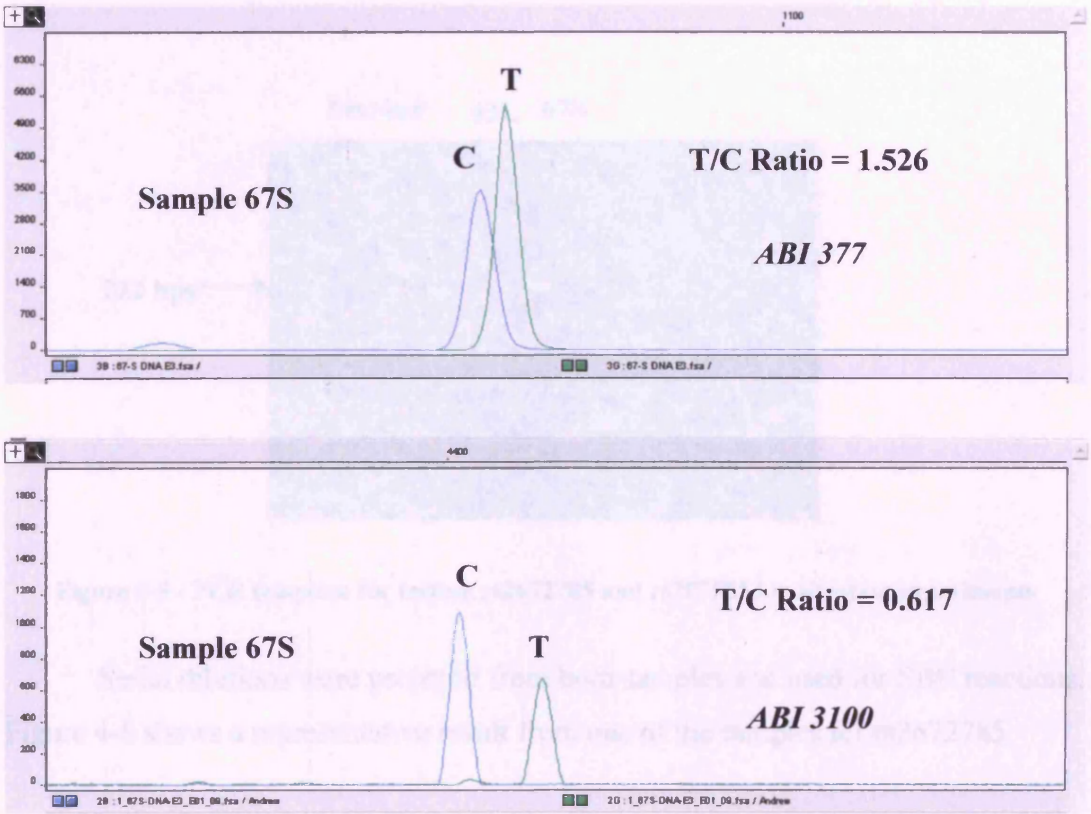


Figure 4-4 – Representative comparison of allelic ratios detected between *ABI 377* and *ABI 3100* sequencers. Same sample run on both machines. Separation of dyes and relative peak heights different between sequencers for SNP marker rs2075853.

It was thus clear that the use of the machines was not interchangeable and for the remainder of this project (Chapter 5), only the *ABI 3100* was used for detecting allele-specific differences in mRNA expression.

4.3 NUCLEOTIDE INCORPORATION

Two factors that might affect nucleotide incorporation were investigated. In each case, three *MUC5B* SNPs were tested using two compound heterozygotes.

4.3.1 Template Concentration

Figure 4-5 shows the electrophoresis of 3 µl of PCR product using gDNA from two *MUC5B* compound heterozygotes with different levels of *MUC5B* expression. These two rather different samples were used to explore the effect of a greater range of PCR product concentrations on the allelic ratios measured.

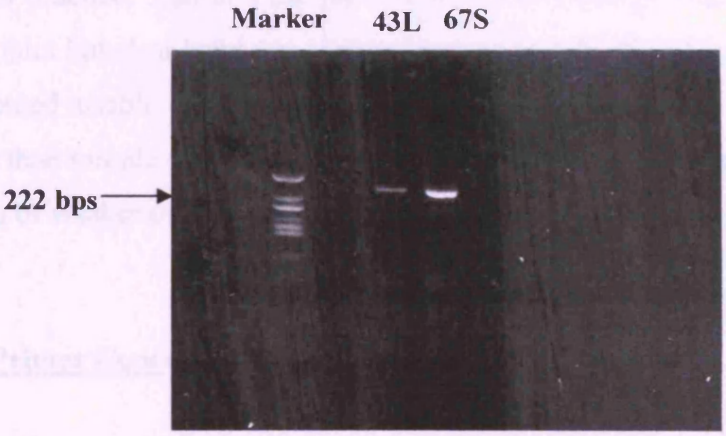


Figure 4-5 - PCR template for testing rs2672785 and rs2075853 in titration experiments

Serial dilutions were prepared from both samples and used for SBE reactions. Figure 4-6 shows a representative result from one of the samples for rs2672785.

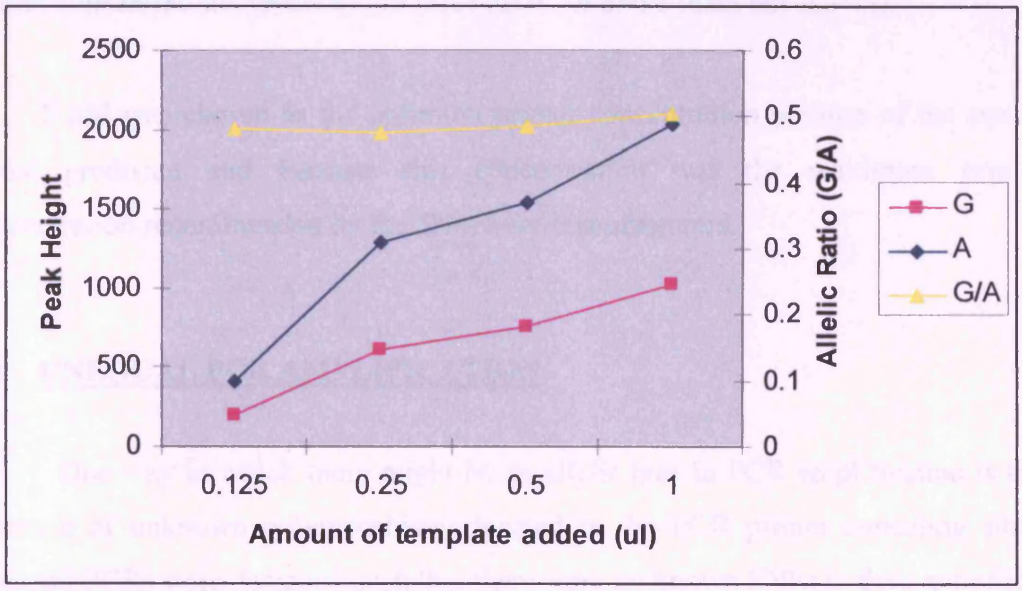


Figure 4-6 - Graph of peak heights/allelic ratio against amount of template used for rs2672785

The experiments showed that for both samples and for all three SNPs that the strength of the signal (peak height) is increased with the amount of template used, suggesting that SBE is quantitative. Importantly, the amount of template added does not significantly alter the allelic ratios obtained. These results suggested that careful control of the amount of template used for the SBE reactions would not be necessary.

In practice, 3 µl of PCR product from every sample was run on an agarose gel. If a faint but clear band was visible (such as sample 43 in Figure 4-5), the sample was deemed usable for SBE reactions. Some PCR products from gDNA were stronger than sample 67 in Figure 4-5 but all PCR products derived from cDNA were as strong or weaker than sample 67.

4.3.2 Primer Concentration Titration

A similar experiment was conducted to that described in section 4.3.1. This time the amount of template was fixed at 1 µl, but the concentration of primer used varied from 0.25 µM to 2 µM. Increasing the primer concentration resulted in higher peak heights just as was observed when template concentration was raised. Again, the allelic ratios measured did not change significantly throughout the range of primer concentrations tested for all three *MUC5B* SNPs (data not shown).

1 µM was chosen as the optimum primer concentration because of the strong signal produced and because this concentration was the maximum primer concentration recommended by the *SNaPshot* manufacturers.

4.4 UNEQUAL PCR AMPLIFICATION

One way in which there might be an allelic bias in PCR amplification is the presence of unknown polymorphisms located in the PCR primer annealing sites. Thus, all PCRs were designed such that there were no known SNPs in their annealing sites and that the SNP being tested was well spaced from the PCR primers.

4.5 THE POSSIBILITY OF SEGMENTAL GENE DUPLICATIONS OF THE SBE PRIMER ANNEALING SITE

Allelic imbalance in gDNA could also be due to the existence of variable number of copies of target sequence due for example to segmental duplications as

has been recently reported to be widespread in the genome (Conrad and Antonarakis 2007; Freeman et al. 2006).

An important assumption made when using SBE for detecting allelic variation is that the SBE primer anneals only once for every allele. If segmental duplications exist, the PCR product used as template for the SBE reactions may contain more than one SBE primer-annealing site. For example:

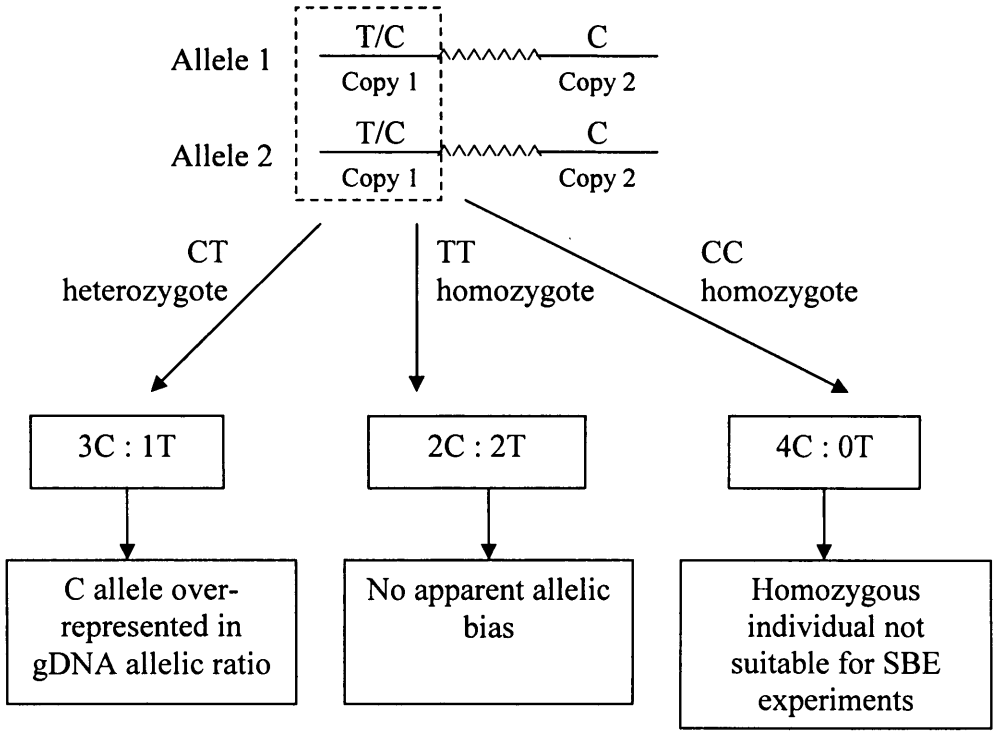


Figure 4-7 – The effect of a segmental duplication on the measured allelic ratios from a SBE reaction.

In this example, copy 1 of the gene contains a SNP, whereas that SNP is non-polymorphic in copy 2. In the case of a true heterozygote, the C allele will be over-represented, since copy 2 will always produce two extra C allele primer-annealing sites. This will result in an allelic ratio overly biased for the C allele. In the case of a TT homozygote, the 1:1 ratio of each allele is fortuitously preserved. Whilst the allelic ratios obtained would certainly represent the results expected when an equal number of copies of each allele were present, it would be inappropriate to use that individual's cDNA since he/she is really a homozygote.

If the segmental duplications are polymorphic (i.e. there is copy number variation (CNV)), some individuals will have only one copy of the gene (and the allele it carries) whilst others might have multiple copies. In such a case, one would observe significant inter-individual differences in the allelic ratios and would be alerted to the possibility of segmental duplications complicating the investigation.

To test whether the detected allelic imbalances reflects the existence of multiple copies of the gene or CNV, we had to use a template for which we knew no duplications were present. Thus, two long oligonucleotides were designed for every SNP tested. Each oligonucleotide was 29-30 bps long and contained the SBE primer-annealing site, but had a different SNP allele. Table 4-1 shows the sequences of the long oligonucleotide used and the SNP alleles they represent.

SNP ID	Sequence 5′ to 3′	Allele Represented
rs2672785	AGCTGGG***G***GAATGCAGGGCACACCATGGAT	G
rs2672785	AGCTGGG***A***GAATGCAGGGCACACCATGGAT	A
rs2075853	CCCACC***C***GGCGCGTGAGCTTTGTTCCACCC	C
rs2075853	CCCACC***T***GGCGCGTGAGCTTTGTTCCACCC	T
rs2075859	TCCTC***G***CAGAGCTGCGCGCGCTGGGGGTT	C
rs2075859	TCCTC***A***CAGAGCTGCGCGCGCTGGGGGTT	T

Table 4-1 – Sequences of long oligonucleotides and the SNP alleles they represent. Position of SNP in sequence shown in bold italics.

These were made up to equimolar concentrations (0.5 μM) using the manufacturers' directions. Equal volumes of each allele-specific oligonucleotide could then be mixed together to artificially create a heterozygote with a perfect 1:1 representation of each allele that could be used for SBE reactions.

Figure 4-8 shows a typical result when different proportions of the two alleles were mixed in a titration experiment. The titration experiments demonstrate that there is a linear relationship between the allelic ratio determined and the proportion of the oligonucleotides in the mixture tested. This meant that SBE was quantitative and could be used to measure allelic ratios.

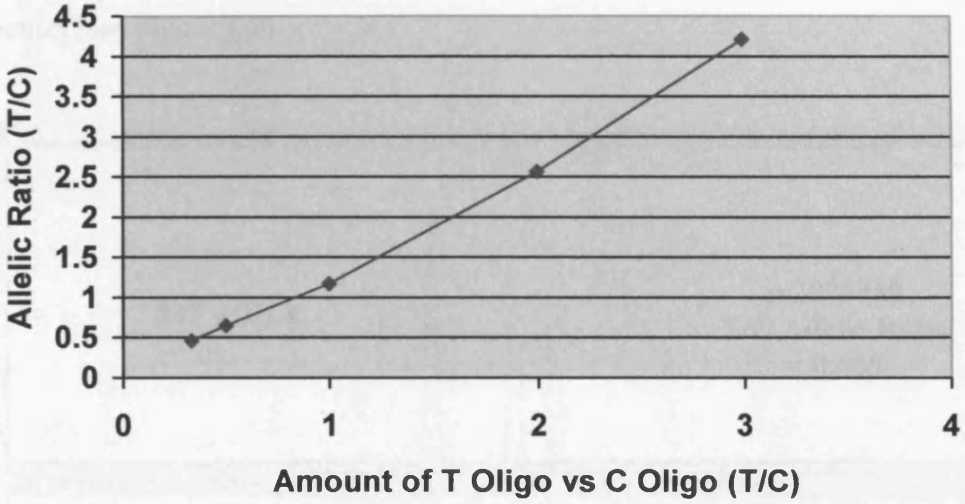


Figure 4-8 – Long oligonucleotide allelic ratios (T/C) for rs2075853. Oligonucelotide mixtures run on *ABI 377*

Crucially, the 1:1 mixtures gave results that were similar to those obtained from heterozygous individuals for all three *MUC5B* markers (see Table 4-2).

Allelic Ratios	rs2672785 (G/A)	rs2075853 (T/C)	rs2075859 (T/C)
Averaged gDNA	0.699 ± 0.021	0.634 ± 0.059	0.669 ± 0.050
Averaged Oligonucleotide Mixtures	0.789 ± 0.023	0.620 ± 0.038	0.579 ± 0.042

Table 4-2 - Allelic ratios from gDNA compared with allelic ratios from oligonucleotide mixtures. All samples were run on *ABI 3100.* Averaged allelic ratio results for oligonucleotide mixtures obtained from three replicates for each SNP. gDNA from a small number of individuals were initially tested for comparison with the oligonucleotide mixture allelic ratios. The averaged gDNA allelic ratios shown here however were derived from the complete sample set described in Chapter 5, in order to give a more complete representation.

These results suggest that imbalance in detection is not due to segmental duplications but rather difference in labelled nucleotide incorporation. It therefore follows that in the absence of large inter-individual differences in gDNA allelic ratios, segmental duplications are unlikely to be an issue.

4.6 CALCULATING ALLELIC RATIOS IN cDNA

Because of the imbalance in allelic detection observed in gDNA from heterozygotes, there is therefore a need to adjust the calculated allelic ratios from

cDNA, in order to take into account the degree of allelic bias in a particular SBE reaction (see Figure 4-9).

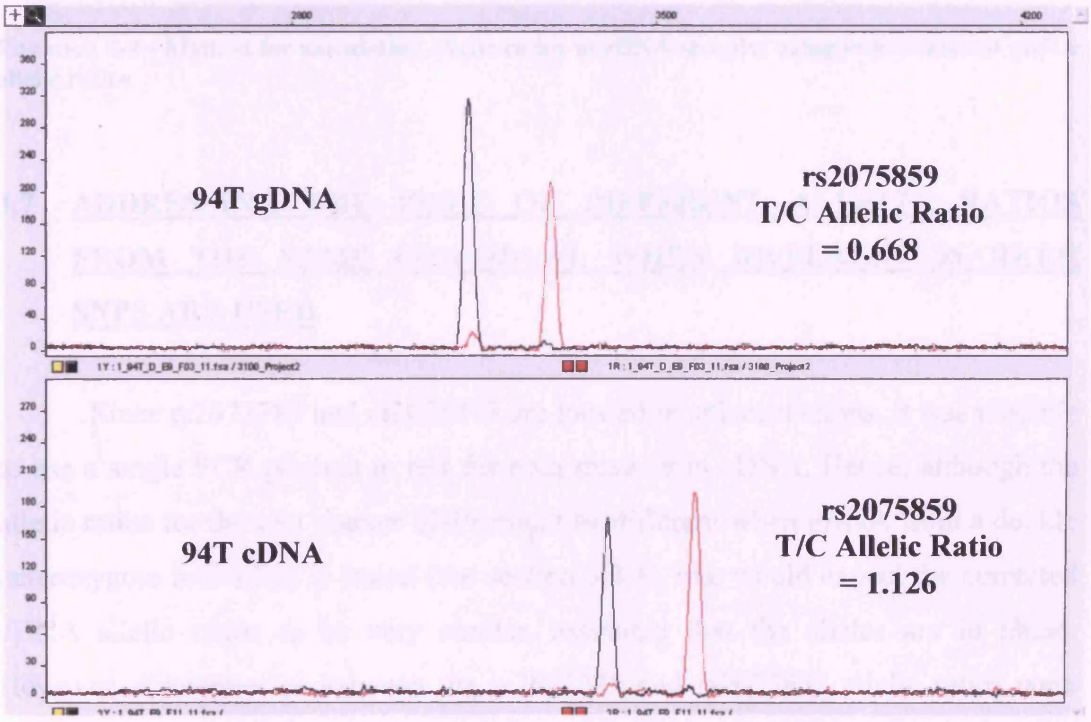


Figure 4-9 - Example showing difference in allelic ratios for gDNA vs. cDNA for a single individual.

For the work described in chapter 5, normalisation of the cDNA allelic ratios for each SNP was conducted using an overall gDNA allelic ratio obtained from all the samples tested for that SNP (see Equation 4-1). Allelic ratios for the cDNA samples were also normalised using individualised allelic ratios derived from their corresponding gDNA samples (see Equation 4-2). Very little difference was observed between the two correction methods and the rationale for using these two methods will be discussed later in Chapter 5.

$$\text{Corrected cDNA Allelic Ratio of Sample X for SNP A} = \frac{\text{cDNA Allelic Ratio of Sample X for SNP A}}{\textit{Overall gDNA Allelic Ratio for SNP A}}$$

Equation 4-1 - Method for calculating allelic ratios in cDNA samples using overall gDNA allelic ratio

$$\text{Corrected cDNA Allelic Ratio of Sample X for SNP A} = \frac{\text{cDNA Allelic Ratio of Sample X for SNP A}}{\textit{Mean gDNA Allelic Ratio of Sample X for SNP A}}$$

Equation 4-2 - Method for calculating allelic ratios in cDNA samples using individualised gDNA allelic ratios

4.7 ADDRESSING THE ISSUE OF DIFFERENT ALLELIC RATIOS FROM THE SAME INDIVIDUAL WHEN DIFFERENT MARKER SNPS ARE USED

Since rs2672785 and rs2075853 are located in adjacent exons, it was possible to use a single PCR product to test for both markers in cDNA. Hence, although the allelic ratios for the two marker SNPs might be different when gDNA from a double heterozygote individual is tested (see section 3.2.3), one would expect the corrected cDNA allelic ratios to be very similar, assuming that the alleles are in phase. However, discrepancies between the rs2672785 and rs2075853 allelic ratios were noted when five compound heterozygotes were tested for both SNPs.

Nonetheless, when the averaged corrected cDNA allelic ratio values (derived from three replicates using same cDNA but PCR and SBE done separately) for rs2075853 are plotted against those for rs2672785, there is an apparent correlation between the results. Greater signal differences were detected for rs2672785 than for rs2075853.

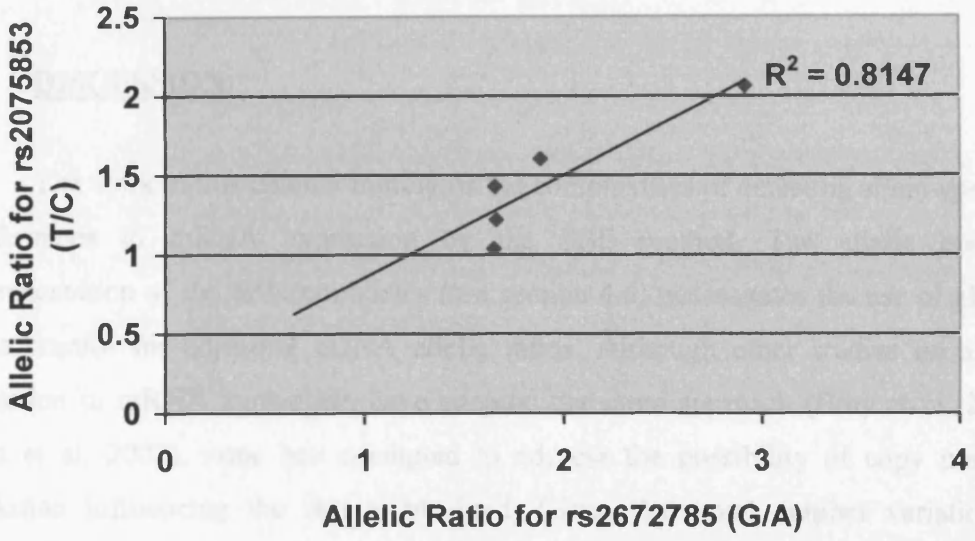


Figure 4-10 - Graph of Allelic Ratios for rs2075853 against Allelic Ratios for rs2672785. Each point on graph represents the averaged allelic ratios for a different individual.

The explanation for this is not obvious. It was possible that this was related to the sequence context of the two different SNPs in terms of GC% content and thus local melting temperatures and primary or secondary structures formed differed, although it is necessary to hypothesise a differential effect on the two alleles at one or both loci.

In order to address the issue of melting temperatures and secondary structures, two sample treatments were attempted. The first was a 5 min denaturation included before the cycling steps in the SBE reaction and the second was an extension of the 96 °C denaturation steps from 10 sec to 20 sec, in the hope that these would help ensure that the DNA strands were fully unfolded and separated before the primer-annealing step. The additional denaturation times did not appear to affect the results in any significant way and the discrepancies in allelic ratios between the two SNP markers remained.

In view of the fact that we were also testing a third SNP, it was decided to continue collecting the data and to combine results from more than one SNP. To do this, it was necessary to obtain full haplotype information on the samples, in order to determine phase and to link the exonic SNPs to the published promoter haplotypes, as described in full detail in the next chapter.

4.8 DISCUSSION

The work in this chapter highlights the complexities of detecting allele-specific differences in mRNA expression by the SBE method. The allelic bias in representation of the different alleles (see section 4.6) necessitates the use of gDNA as a control for adjusting cDNA allelic ratios. Although other studies on allelic variation in mRNA expression have adopted the same approach (Bray et al. 2003; Yan et al. 2002), none has attempted to address the possibility of copy number variation influencing the results obtained. Given that copy number variation is increasingly considered widespread throughout the genome, this was an important consideration. Here I show that the allelic ratios obtained from the oligonucleotide mixes correspond well with the allelic ratios obtained from gDNA.

Although other studies have used different marker SNPs within the same gene and have noted good correlation between the markers, this is the first study that uses two marker SNPs (rs2672785 and rs2075853) found on the same PCR product. Hence, the absolute values of the allelic ratios for cDNA were expected to be similar, although the direction of the allelic bias might be different depending on the phase. However, there were clear differences between the allelic ratios from the different marker SNPs. In rs2075853, the difference in allelic imbalance appeared to be suppressed in comparison with rs2672785. Nonetheless, there was an apparent correlation between the markers and the higher expressing alleles for both SNPs appeared to be in phase for all the five samples tested.

It is difficult to see what might be the real cause of the discrepancies between the markers, although it is almost certainly technical. It seems unlikely that it is due to non-linearity of the reaction because titration experiments failed to show that changes in primer and template concentrations significantly influence allelic ratios obtained. It is most probable that the problem is sequence-specific and reflects differences in conformation between PCR product derived from cDNA and gDNA. It is possible that these differences could have been corrected by using additives such

as DMSO, but it seemed likely that such additives would interfere with the SBE reactions.

It is likely that all of the allelic variation detection techniques that rely on allele-specific primers are susceptible to this type of problem. The results from such experiments need to be carefully considered in light of this, because the extent of the allelic variation observed is highly dependent on the marker chosen.

CHAPTER 5

STUDY OF ALLELIC VARIATION IN *MUC5B* EXPRESSION

5 STUDY OF ALLELIC VARIATION IN *MUC5B* EXPRESSION

5.1 INTRODUCTION

The chapter is concerned with the study of allelic variation in *MUC5B* mRNA expression. Unlike *MUC4*, *MUC5B* shows no evidence of a variable TR length or of extensive alternative splicing. Hence, it should be in principle easier to detect and account for allele-specific differences in mRNA expression. In addition, *in vitro* experiments carried out by Kamio and colleagues had suggested the presence of allelic differences in *MUC5B* mRNA expression. The background of their work is described here:

5.1.1 *MUC5B* and Diffuse Panbronchiolitis (DPB) – Initial Research Suggesting the Presence of Allelic Variation in *MUC5B* Expression

Diffuse Panbronchiolitis (DPB) is an inflammatory airway disease characterised by a thickening of the wall of the respiratory bronchiole, coupled with infiltration of lymphocytes, plasma cells and foamy histiocytes around respiratory bronchioles (Kudoh and Keicho 2003). The disease primarily affects East Asians. Patients suffer from chronic coughs, produce large amounts of sputum and are prone to chronic bacterial infections (Homma et al. 1983). At advanced stages, patients become superinfected with *Pseudomonas aeruginosa* and prognosis is often bleak (Kudoh and Keicho 2003).

Since mucus hypersecretion is observed in DPB patients and because mucins are the principal proteins found in mucus, Kamio and colleagues hypothesised that transcriptional regulation of the mucins might be altered in DPB (Kamio et al. 2005). They explored this possibility by analysing polymorphisms located in the promoter regions of *MUC2*, *MUC4*, *MUC5AC*, *MUC5B* and *MUC7*, all of which are known to be expressed in the respiratory tract (Copin et al. 2000).

Their test population was 92 unrelated Japanese patients diagnosed with DPB and their control population was 128 anonymous individuals selected from the general Japanese population. A case-control association study was then conducted using 22 different polymorphisms, spread across the five *MUC* genes chosen. Three (rs885454, rs17235353 and rs7115457) of the 22 polymorphisms, all found in *MUC5B*, were significantly associated with DPB (P values 0.0052, 0.0001 and 0.0488 respectively). Armed with this information, Kamio and colleagues analysed *MUC5B* haplotype frequencies in cases vs. controls.

To do so, they used an SSCP-based molecular haplotyping method to determine the haplotypes (comprised of 6 *MUC5B* promoter polymorphisms, including the three polymorphisms significantly associated with DPB) present in the Japanese test subjects (Kamio et al. 2004). Figure 5-1 shows the positions of these polymorphisms on the *MUC5B* promoter.

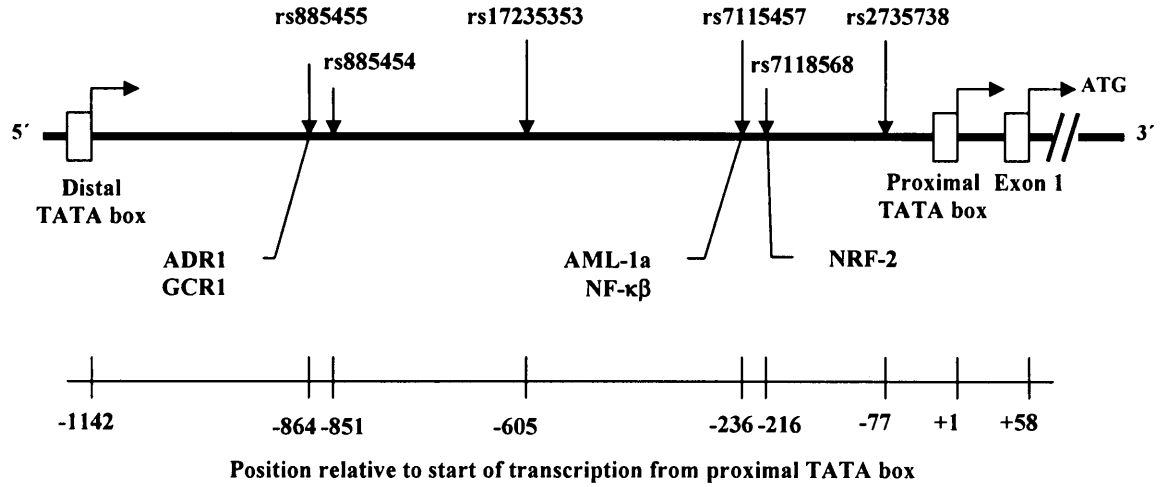


Figure 5-1 - *MUC5B* promoter polymorphisms identified by Kamio and colleagues. Relative positions of polymorphisms shown. Putative transcription factor binding sites detected using TRANSFAC shown (see section 5.8). Diagram not to scale.

Thirteen promoter haplotypes were detected, shown in Table 5-1, of which three major haplotypes (H1, H2 and H3) collectively represented approximately 80% of all the alleles. Kamio and colleagues found that the haplotype distribution was significantly different between cases and controls. In particular, the frequency of H2

was significantly decreased in the patient population. H1 was increased in patients but the difference did not reach statistical significance (P value = 0.063).

Promoter Haplotype Identifier	rs885455	rs885454	rs17235353	rs7115457	rs7118568	rs2735738	Haplotype Freq	
							Controls	DPB Patients
H1	G	G	I	A	G	C	0.379	0.467
H2	G	A	D	G	C	T	0.238	0.098
H3	A	G	I	G	C	T	0.191	0.212
H4	G	A	I	G	C	C	0.094	0.098
H5	G	G	I	G	C	C	0.035	0.038
H6	G	G	I	G	C	T	0.023	0.054
H7	G	G	D	G	C	T	0.016	0.005
H8	A	G	I	G	C	C	0.008	0.000
H9	G	A	I	G	C	T	0.008	0.160
H10	G	G	I	G	G	C	0.004	0.000
H11	G	G	I	A	G	T	0.004	0.000
H12	A	A	I	G	C	C	0.000	0.005
H13	A	G	I	A	G	C	0.000	0.005

Table 5-1 - *MUC5B* promoter haplotypes detected by Kamio et al. For rs17235353, I: CA insertion, D: CA deletion

Motivated by these findings, Kamio and colleagues investigated whether the different *MUC5B* haplotypes affected *MUC5B* transcription levels by utilising reporter assays. DNA sequences representing H1, H2 and H3 were transfected into *MUC5B* expressing NCI-H292 cells and relative luciferase activity in the steady state was measured. They found that H2 resulted in the lowest transcriptional activity, whilst H3 was intermediate and H1 the highest.

Taken together, Kamio and colleagues' findings suggest that polymorphisms in the *MUC5B* promoter region might affect the expression levels of *MUC5B* (due to allele-specific differences in expression) and influence the pathogenesis of DPB.

The work in this chapter investigates the presence and extent of allelic variation in the constitutive expression of *MUC5B* in fetal tissue and attempts to address the question of whether the *MUC5B* promoter haplotypes are differentially expressed *in vivo,* using the SBE method.

5.2 INITIAL SELECTION OF SAMPLES

Individuals from the fetal population were initially selected for this study based on the range of tissues available. *MUC5B* has been detected in fetal trachea (13 weeks onwards) (Reid, Gould, and Harris 1997) and fetal gall bladder (18 weeks onwards but not tested in less mature tissue) (Buisine et al. 1999b; Reid, Gould, and Harris 1997). *MUC5B* is also expressed in bronchi (13 weeks onwards) (Buisine et al. 1999b), bronchioles (23 weeks onwards) (Reid, Gould, and Harris 1997) and strongly expressed in adult salivary glands (Alos et al. 2005; Veerman et al. 2003), suggesting the possibility of detecting *MUC5B* in fetal lung and salivary glands. Thus, only individuals for whom trachea, gall bladder, salivary glands or lung tissue were available were chosen. RNA and gDNA were extracted from these tissues. Table 5-2 below gives a summary of the individuals chosen and the tissue types available from them.

	Trachea	**Gall Bladder**	**Salivary Glands**	**Lung**	**No. of Individuals (n=92)**
Average Gestational Age (weeks)	15.2 ± 1.6	15.8 ± 0.9	19.7 ± 1.1	14.2 ± 1.8	
	✔				13
		✔			12
			✔		2
				✔	53
	✔	✔			1
		✔		✔	4
	✔			✔	6
	✔	✔		✔	1
Total Number of Samples (n=105)	21	18	2	64	

Table 5-2 – Individuals from fetal population initially chosen for *MUC5B* study and their available tissue types

5.3 GENOTYPE ANALYSIS

The extracted gDNA was used to genotype these individuals for three *MUC5B* SNPs, rs2672785 found in exon 2, rs2075853 in exon 3 and rs2075859 in

exon 9. These three SNPS were chosen because they had been successfully validated in both the HapMap and dbSNP databases and because they were located in exons, allowing them to function as marker SNPs for the SBE experiments. Representative genotyping results for the three SNPs are shown in Figure 5-2, Figure 5-3 and Figure 5-4.

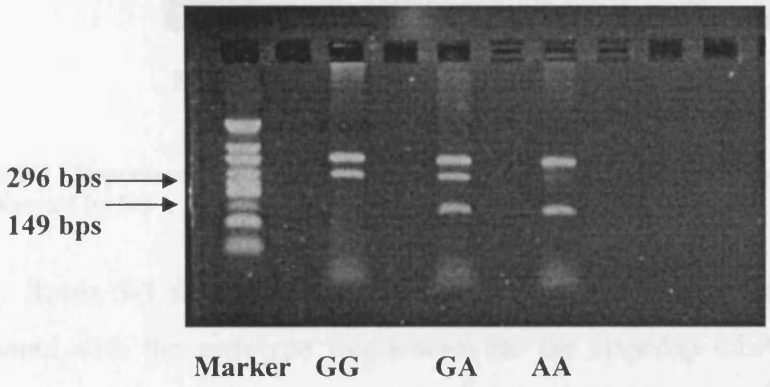


Figure 5-2 – Representative *MUC5B* rs2672785 (exon 2) genotyping results. Typed by Tetra-ARMS PCR

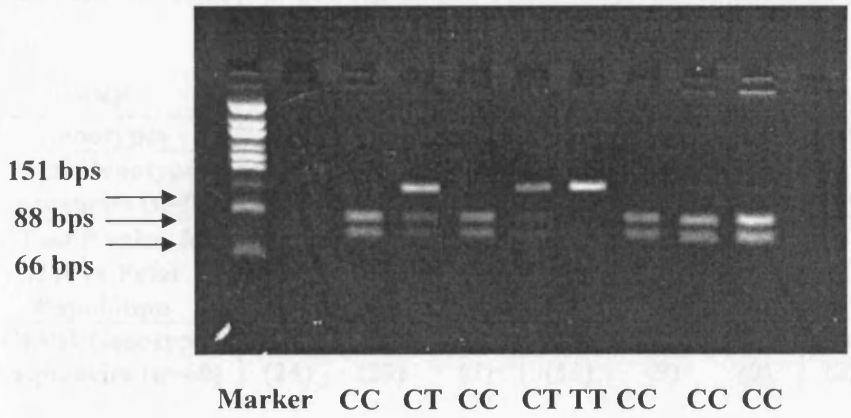


Figure 5-3 - Representative *MUC5B* rs2075853 (exon 3) genotyping results. Digested by Msp I.

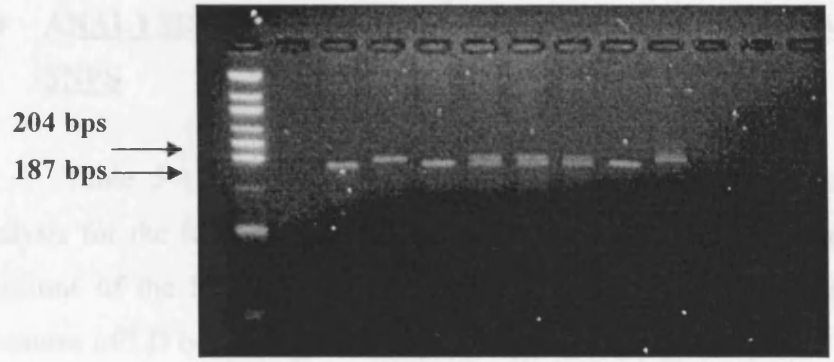


Figure 5-4 - Representative *MUC5B* rs2075859 (exon 9) genotyping results. Engineered restriction site. Digested by Bts I.

Table 5-3 shows the *MUC5B* genotype frequencies for the fetal population compared with the genotype frequencies for the HapMap CEPHs sample set and gives an indication of the number of informative individuals available for use in SBE experiments. Chi Square tests were performed for each SNP to test for deviation from expected Hardy-Weinberg frequencies.

SNP	rs2672785			rs2075853			rs2075859		
Genotypes	AA	AG	GG	CC	CT	TT	CC	CT	TT
Fetal Genotype Frequencies (n=92)	0.511 (47)	0.402 (37)	0.087 (8)	0.750 (69)	0.217 (20)	0.033 (3)	0.326 (30)	0.457 (42)	0.217 (20)
χ^2 Test P value for HW in Fetal Population		0.983			0.607			0.766	
CEPH Genotype Frequencies (n=60)	0.567 (34)	0.383 (23)	0.050 (3)	0.850 (51)	0.150 (9)	0.000 (0)	0.417 (25)	0.450 (27)	0.133 (8)

Table 5-3 – *MUC5B* exonic SNP genotype frequencies in fetal population compared with HapMap CEPH population

The genotype frequencies between the two groups are very similar for all three markers, as was also noted for the *MUC4* SNPs (see section 3.3). The P values indicate that there is no significant deviation from Hardy-Weinberg equilibrium and suggests no serious genotyping errors or problems with population admixture.

5.4 ANALYSIS OF LINKAGE DISEQUILIBRIUM BETWEEN MUC5B SNPS

Table 5-4 shows the results of the *MUC5B* linkage disequilibrium (LD) analysis for the fetal and HapMap CEPH populations, whilst Figure 5-5 shows the positions of the SNPs on a scale diagram of the *MUC5B* gene structure and the measures of LD between the SNPs.

Pair-wise Comparisons	rs2672785 vs. rs2075853		rs2672785 vs. rs2075859		rs2075853 vs. rs2075859	
	Chi Sq P value	Lewontin D′	Chi Sq P value	Lewontin D′	Chi Sq P value	Lewontin D′
Fetal Population	<0.00001	1	0.53145	0.089	0.00009	0.800
HapMap CEPHs	<0.00001	1	0.58915	0.156	0.01715	0.758

Table 5-4 - Measures of LD between *MUC5B* exonic SNPs in fetal and HapMap CEPH populations

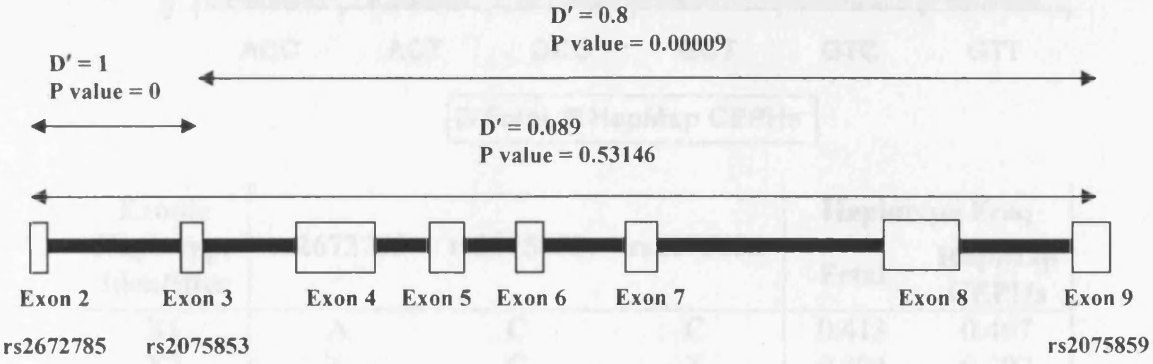


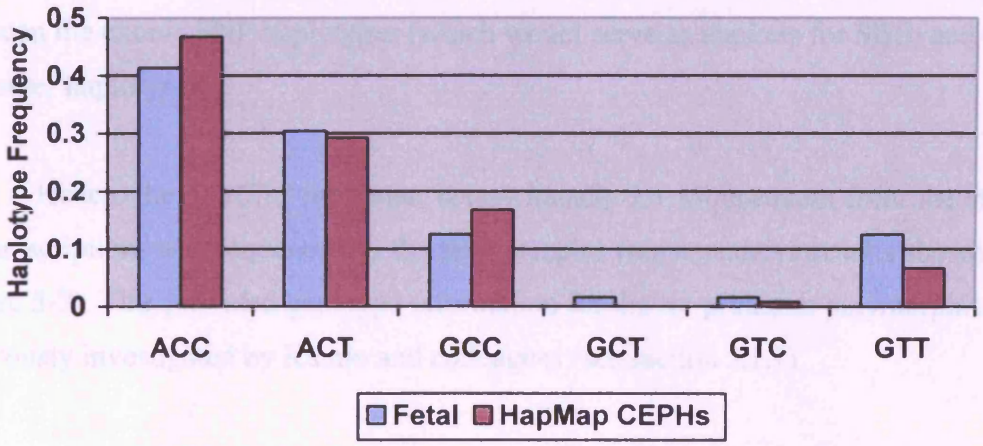
Figure 5-5 – Scale diagram showing positions of *MUC5B* exonic SNPs and associated measures of LD in fetal population

The LD analysis shows that there is significant LD between rs2672785 and rs2075853 and between rs2075853 and rs2075859 but a lack of significant association between rs2672785 and rs2075859. The pattern of LD is very similar in both fetal and HapMap populations.

5.5 HAPLOTYPE ANALYSIS

5.5.1 Haplotype Reconstruction using *MUC5B* Exonic SNPs

The exonic SNP genotypes were used to reconstruct haplotypes using the *Phase* program. Figure 5-6 shows the various deduced haplotypes and their associated frequency within the fetal and HapMap CEPH populations. To make the analysis comparable since the CEPH genotyping data originates from family trios, all 30 offspring samples were excluded from the haplotype reconstruction and haplotypes were inferred without using family information.



Exonic Haplotype Identifier	rs2672785	rs2075853	rs2075853	Haplotype Freq	
				Fetal	HapMap CEPHs
X1	A	C	C	0.413	0.467
X2	A	C	T	0.304	0.292
X3	G	C	C	0.125	0.167
X4	G	C	T	0.016	0.000
X5	G	T	C	0.016	0.008
X6	G	T	T	0.125	0.067

Figure 5-6 - *MUC5B* Exonic Haplotype Frequencies in fetal and HapMap CEPH populations

There appear to be six haplotypes occurring within the fetal population: two common haplotypes (ACC and ACT), two less common (GCC and GTT) and two rare haplotypes (GCT and GTC). These have been numbered X1 to X6. The haplotype frequencies appear to be quite similar between the fetal and HapMap CEPH groups. The most striking difference between the two groups is probably the nearly twice as frequent X6 haplotype in the fetal vs. CEPH populations, which

could be due to slight admixture within the fetal group. However, the haplotype distributions between the two populations are not significantly different (exact test of population differentiation using *Arlequin* program, P value = 0.71582).

5.5.2 Haplotype Reconstruction using *MUC5B* Exonic SNPs and Promoter Polymorphisms from Kamio Study

A major aim of this project was to determine whether the *MUC5B* promoter haplotypes are associated with expression *in vivo* by using the SBE method (see section 5.1.1). In order to achieve this, it was first necessary to establish the link between the exonic SNP haplotypes (which would serve as markers for SBE) and the promoter haplotypes.

Hence, the *MUC5B* promoter, approximately 1.1 kb upstream from the start of transcription, was sequenced in the fetal samples (representative results shown in Figure 5-7). This provided genotype information for the six promoter polymorphisms previously investigated by Kamio and colleagues (see section 5.1.1).

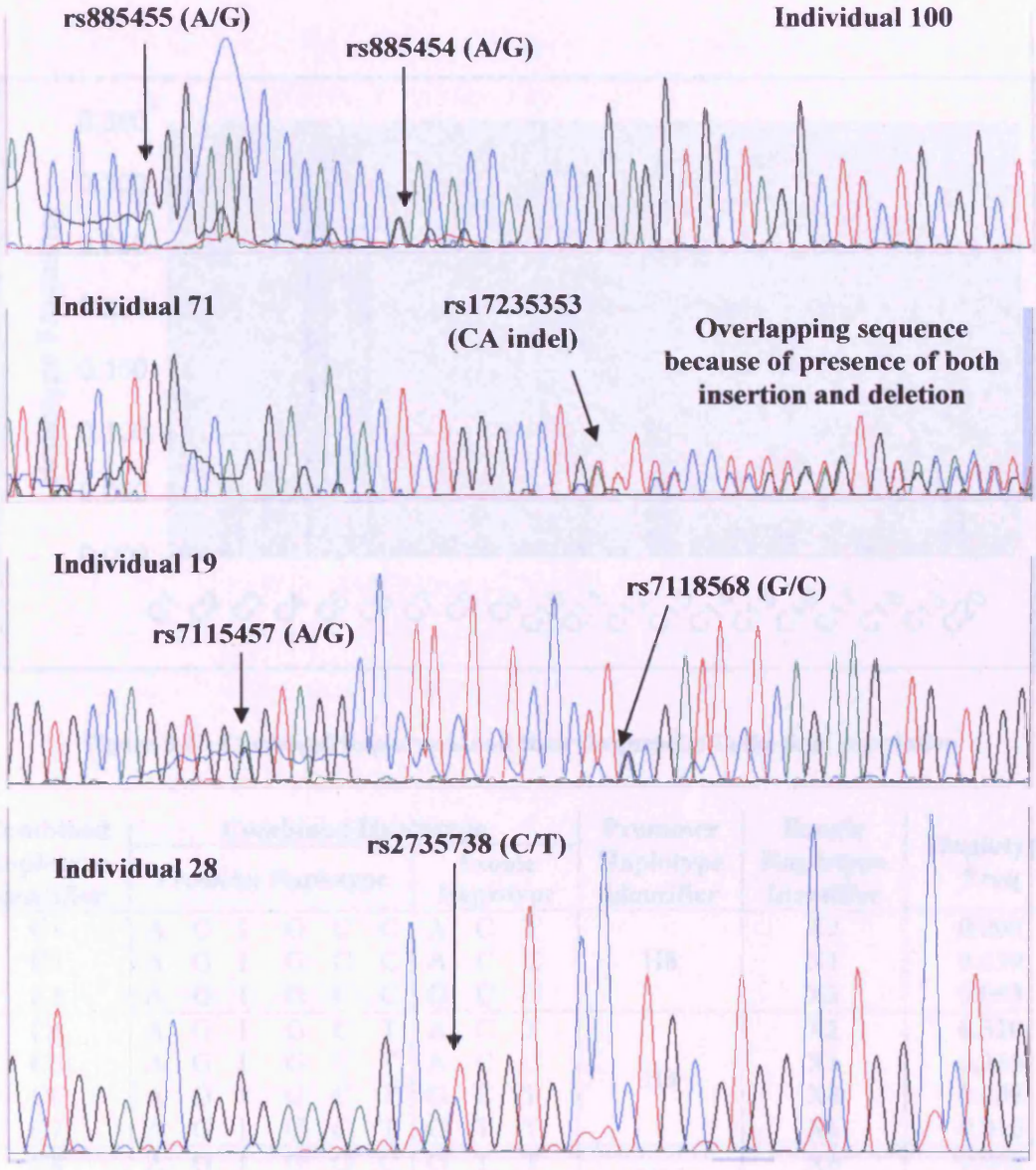


Figure 5-7 – Representative electropherograms of the *MUC5B* promoter sequence. Sequences from individuals heterozygous for the various promoter polymorphisms shown. Sequences analysed using *Sequencing Analysis v5.1.1.*

These data were used in conjunction with the exonic SNP genotypes as input for the *Phase* program, in order to generate 'combined' haplotypes containing all nine *MUC5B* polymorphisms.

Figure 5-8 shows the frequency of each combined haplotype within the fetal population. Table 5-5 shows the actual alleles in each combined haplotype, its associated frequency, as well as the combination of promoter and exonic haplotypes (see Figure 5-6 and Table 5-1) that make up each combined haplotype.

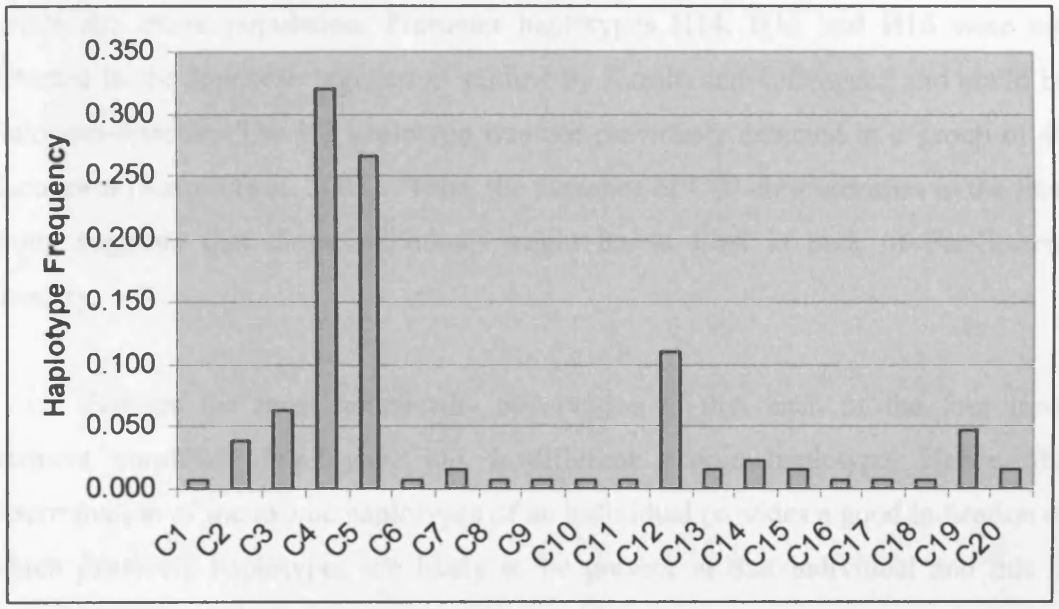


Figure 5-8 - Combined haplotypes and their frequencies in the fetal population

Combined Haplotype Identifier	Combined Haplotype		Promoter Haplotype Identifier	Exonic Haplotype Identifier	Haplotype Freq
	Promoter Haplotype	Exonic Haplotype			
C1	A G I G C C	A C T	H8	X2	0.008
C2	A G I G C C	A C C		X1	0.039
C3	A G I G C C	G C C		**X3**	**0.063**
C4	A G I G C T	A C T	H3	**X2**	**0.320**
C5	A G I G C T	A C C		**X1**	**0.266**
C6	A G I G C T	G C T		X4	0.008
C7	A G I G C T	G T T		X6	0.016
C8	A G I G G C	G T T	H14	X6	0.008
C9	A G I G G C	G T C		X5	0.008
C10	G G I A C C	G C C	H15	X3	0.008
C11	G G I A C C	G T T		X6	0.008
C12	G G I A G C	G T T	H1	**X6**	**0.109**
C13	G G I A G C	G T C		X5	0.016
C14	G G I G C C	G C C	H5	X3	0.023
C15	G G I G C T	A C T	H6	X2	0.016
C16	G G I G C T	A C C		X1	0.008
C17	G A I A G C	G T T	H16	X6	0.008
C18	G A I G C C	G C T	H4	X4	0.008
C19	G A I G C C	G C C		X3	0.047
C20	G A D G C T	A C C	H2	X1	0.016

Table 5-5 - Combined MUC5B haplotypes and their frequencies in the fetal population (n = 92). Haplotypes occurring at frequency greater than 0.05 shown in bold.

Inspection of these data reveals the presence of four combined haplotypes at a frequency greater than 0.05 within the fetal population. These are C3, C4, C5 and

C12 shown in bold in Table 5-5. Some of the remaining haplotypes occur only once within the entire population. Promoter haplotypes H14, H15 and H16 were not detected in the Japanese population studied by Kamio and colleagues and could be European-specific. The H2 haplotype was not previously detected in a group of 40 Europeans (Kamio et al. 2005). Thus, the presence of C20 chromosomes in the fetal group suggests that these individuals might be, at least in part, of Far-Eastern ancestry.

Perhaps the most noteworthy observation is that each of the four most frequent combined haplotypes has a different exonic haplotype. Hence, the determination of the *exonic* haplotypes of an individual provides a good indication of which *promoter* haplotypes are likely to be present in that individual and this is exploited in the analysis described in Chapter 6.

5.6 TESTING FOR ALLELIC VARIATION IN *MUC5B* MRNA EXPRESSION BY SBE METHOD

5.6.1 Final Selection of Samples for Study

A large number of samples from the initial 92 individuals chosen for this study (see section 5.3) could not be tested for the following two reasons:

- Some individuals were non-informative (i.e. homozygous) at all three marker SNPs.
- *MUC5B* expression is generally much lower in fetal tissue compared to adult tissue. Thus, the cDNA from some samples failed to produce a sufficiently strong PCR product when used in RT PCR reactions. This was particularly an issue for cDNA derived from lungs, most probably because the piece of lung tissue used to extract the mRNA contained insufficient bronchial tissue. cDNA derived from trachea, gall bladder and salivary glands tended to produce strong PCR products.

The final sample set consists of 42 samples from 38 different individuals.

5.6.2 SBE Results

Figure 5-9, Figure 5-10 and Figure 5-11 are representative of the SBE results obtained.

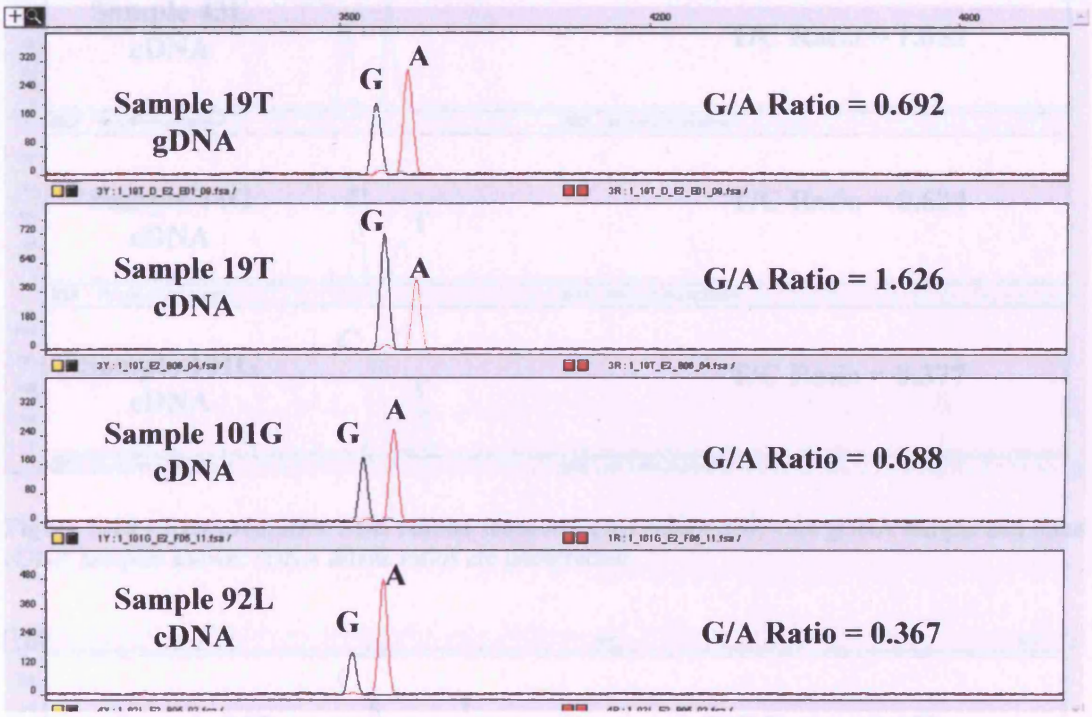


Figure 5-9 - Representative SBE results from *MUC5B* rs2672785. One gDNA sample and three cDNA samples shown. cDNA allelic ratios are uncorrected.

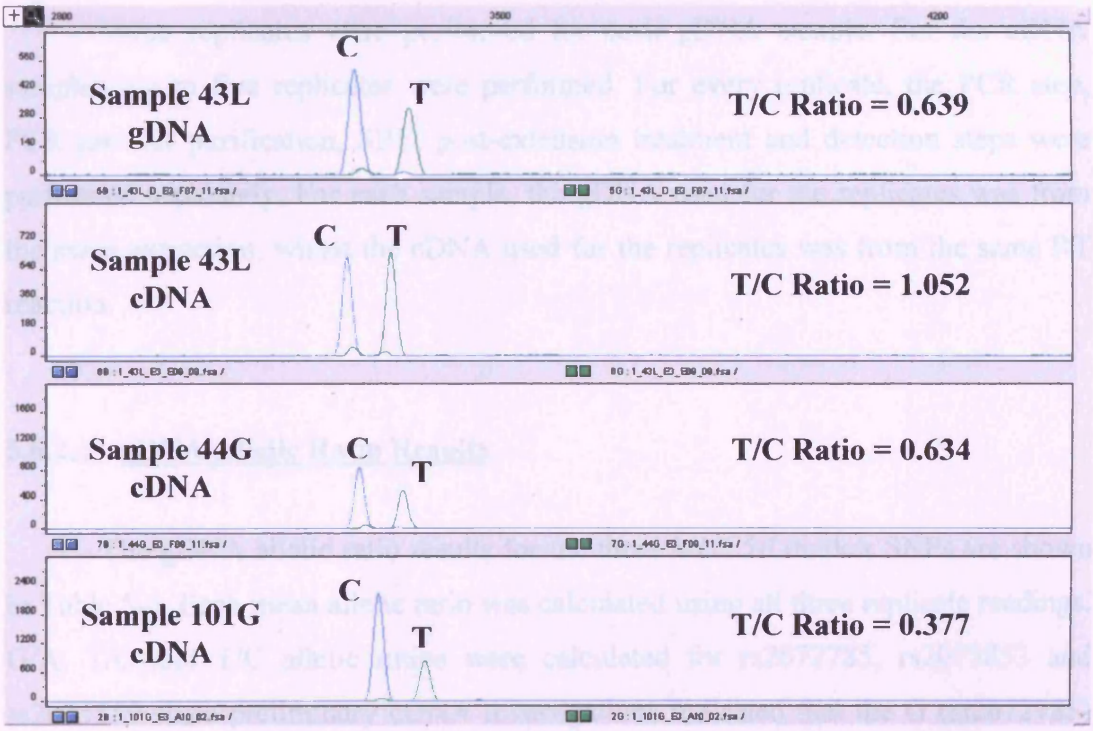


Figure 5-10 - Representative SBE results from *MUC5B* rs2075853. One gDNA sample and three cDNA samples shown. cDNA allelic ratios are uncorrected.

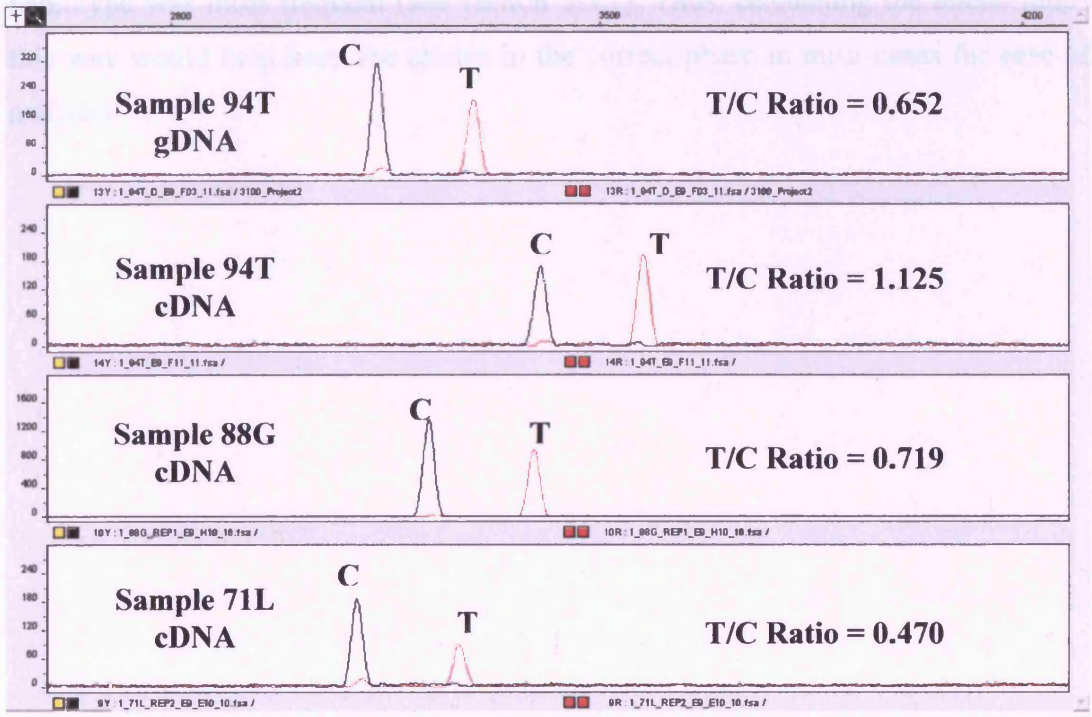


Figure 5-11 - Representative SBE results from *MUC5B* rs2075859. One gDNA sample and three cDNA samples shown. cDNA allelic ratios are uncorrected.

Three replicates were performed for each gDNA sample. For the cDNA samples, up to five replicates were performed. For every replicate, the PCR step, PCR product purification, SBE, post-extension treatment and detection steps were performed separately. For each sample, the gDNA used for the replicates was from the same extraction, whilst the cDNA used for the replicates was from the same RT reaction.

5.6.2.1 gDNA Allelic Ratio Results

The gDNA allelic ratio results for the three *MUC5B* marker SNPs are shown in Table 5-6. Each mean allelic ratio was calculated using all three replicate readings. G/A, T/C and T/C allelic ratios were calculated for rs2672785, rs2075853 and rs2075859 since preliminary cDNA investigations indicated that the G (rs2672785) and T (rs2075853) alleles were usually expressed at a higher level than their counterpart alleles (see section 4.1) and haplotype analysis indicated that the ACC haplotype was most frequent (see section 5.5.1). Thus, calculating the allelic ratios this way would help keep the alleles in the correct phase in most cases for ease of analysis.

	rs2672785 (G/A allelic ratio)		rs2075853 (T/C allelic ratio)		rs2075859 (T/C allelic ratio)	
	Mean	**StDev**	**Mean**	**StDev**	**Mean**	**StDev**
3SG	0.712	0.108	0.543	0.051	-	-
4SG	-	-	-	-	0.757	0.068
6T	-	-	-	-	0.638	0.003
9G	0.689	0.052	-	-	-	-
10T	-	-	-	-	0.663	0.040
11G	-	-	0.662	0.056	-	-
12G	-	-	-	-	0.655	0.056
14G	-	-	-	-	0.634	0.009
15T	-	-	-	-	0.654	0.041
16T	0.718	0.074	-	-	0.615	0.042
17T	0.690	0.028	-	-	-	-
19T	0.689	0.042	0.761	0.096	-	-
26L	0.708	0.035	-	-	0.742	0.010
28L	0.752	0.026	0.609	0.057	0.619	0.010
29L	0.690	0.050	-	-	0.702	0.016
30T	0.667	0.078	-	-	-	-
34L	0.672	0.080	0.668	0.033	-	-
43L	0.706	0.058	0.665	0.025	-	-
44G	0.708	0.060	0.600	0.052	0.633	0.047
45G	0.715	0.016	-	-	-	-
53L	-	-	-	-	0.736	0.067
57L	-	-	-	-	0.606	0.065
63L	-	-	-	-	0.689	0.017
66L	0.709	0.037	-	-	-	-
67S	0.690	0.028	0.562	0.070	0.617	0.032
71L	-	-	-	-	0.674	0.037
73L	-	-	-	-	0.643	0.025
86T	-	-	-	-	0.757	0.023
88T	-	-	-	-	0.661	0.009
92L	0.690	0.071	-	-	-	-
94T	0.723	0.064	0.639	0.012	0.648	0.017
97L	0.703	0.069	0.631	0.017	0.606	0.027
98T	-	-	-	-	0.702	0.082
100L	0.668	0.045	-	-	0.649	0.035
101G	0.676	0.033	0.635	0.029	-	-
103T	-	-	-	-	0.785	0.016
112L	-	-	-	-	0.656	0.023
116L	-	-	-	-	0.654	0.032
Overall	**0.699**	**0.021**	**0.634**	**0.059**	**0.669**	**0.050**

Table 5-6 - *MUC5B* gDNA allelic ratio results. The overall mean value represents the averaged mean allelic ratio of all the samples for a particular SNP. The overall stdev represents the inter-individual variation in mean allelic ratios for a particular SNP.

The overall allelic ratios were calculated by taking an average of all the replicate readings available for each marker SNP, whilst the overall standard deviations reflect the inter-individual variation in mean allelic ratios.

5.6.2.2 Corrected cDNA Allelic Ratio Results

For cDNA, extreme outlier results were excluded from the mean allelic ratios, which were derived from at least three replicates. The corrected cDNA allelic ratios for the three *MUC5B* SNPs are shown in Table 5-7. For the data shown, the allelic ratios of the samples were corrected using the overall gDNA allelic ratios for each SNP (see Equation 4-1 and Table 5-6), although they were also corrected using the mean gDNA allelic ratios from their corresponding gDNA samples (data not shown, see Equation 4-2 and Table 5-6). There was very little difference between the allelic ratios when corrected using either method.

	rs2672785 (G/A allelic ratio)			rs2075853 (T/C allelic ratio)			rs2075859 (T/C allelic ratio)		
	Mean	**StDev**	**95% CI**	**Mean**	**StDev**	**95% CI**	**Mean**	**StDev**	**95% CI**
3SG	2.451	0.183	0.455	1.260	0.149	0.371	-	-	-
4SG	-	-	-	-	-	-	0.887	0.191	0.475
6T	-	-	-	-	-	-	1.124	0.092	0.229
9G	1.574	0.031	0.076	-	-	-	-	-	-
10T	-	-	-	-	-	-	1.040	0.018	0.044
11G	-	-	-	1.215	0.038	0.094	-	-	-
12G	-	-	-	-	-	-	0.610	0.230	0.571
14G	-	-	-	-	-	-	1.140	0.200	0.497
15G	-	-	-	-	-	-	0.885	0.231	0.574
15T	-	-	-	-	-	-	1.062	0.086	0.214
16T	1.141	0.077	0.192	-	-	-	*0.926*	0.103	0.256
17T	1.193	0.125	0.310	-	-	-	-	-	-
19T	2.082	0.269	0.667	1.061	0.206	0.511	-	-	-
26L	1.623	1.050	2.607	-	-	-	*0.553*	0.105	0.260
28L	1.846	0.234	0.582	1.027	0.056	0.139	*0.917*	0.326	0.810
29L	2.140	0.334	0.830	-	-	-	*0.969*	0.111	0.276
30T	1.439	0.219	0.543	-	-	-	-	-	-
34L	2.244	0.267	0.663	1.216	0.045	0.111	-	-	-
43L	2.918	0.203	0.504	1.559	0.092	0.229	-	-	-
44G	1.839	0.208	0.517	0.991	0.058	0.144	1.580	0.115	0.285
45G	1.234	0.108	0.268	-	-	-	-	-	-
53L	-	-	-	-	-	-	1.938	0.480	1.191
57L	-	-	-	-	-	-	1.270	0.246	0.610
63L	-	-	-	-	-	-	1.088	0.218	0.541
66L	2.472	0.693	1.721	-	-	-	-	-	-
67L	2.296	0.536	0.852	1.284	0.095	0.235	*0.918*	0.328	0.522
71L	-	-	-	-	-	-	0.661	0.099	0.245
73L	-	-	-	-	-	-	1.006	0.374	0.929
86T	-	-	-	-	-	-	1.016	0.179	0.445
88G	-	-	-	-	-	-	1.008	0.066	0.163
88T	-	-	-	-	-	-	1.087	0.061	0.152
92L	0.420	0.202	0.502	-	-	-	-	-	-
94T	1.618	0.072	0.179	0.912	0.055	0.136	1.634	0.134	0.333
97L	1.115	0.071	0.176	0.725	0.008	0.021	1.032	0.163	0.405
97G	1.499	0.079	0.196	0.547	0.216	0.535	0.747	0.182	0.453
98T	-	-	-	-	-	-	1.182	0.147	0.365
100L	1.650	0.381	0.945	-	-	-	1.091	0.413	1.026
101G	1.004	0.046	0.113	0.590	0.021	0.051	-	-	-
103T	-	-	-	-	-	-	1.465	0.092	0.227
112L	-	-	-	-	-	-	1.679	0.286	0.709
116G	-	-	-	-	-	-	1.230	0.115	0.286
116L	-	-	-	-	-	-	0.518	0.292	0.726
Overall	**1.705**	**0.594**	-	**1.032**	**0.302**	-	**1.075**	**0.330**	-

Table 5-7 – *MUC5B* corrected cDNA allelic ratio results. Allelic ratios corrected using overall gDNA allelic ratios (see Equation 4-1). The overall mean value represents the averaged mean allelic ratio of all the samples for a particular SNP. The overall stdev represents the inter-individual variation in mean allelic ratios for a particular SNP. 95% Confidence intervals of the mean = standard error of mean * 4.3 (t-distribution, α =0.05, 2 degrees of freedom). Allelic ratios in italics were inverted when used to calculate the averaged corrected cDNA allelic ratios (see section 5.7).

Overall, approximately 50% of the samples showed mean corrected cDNA allelic ratios of 20% or more above or below the expected ratio of 1. This effect was seen in all tissues.

Figure 5-12 shows the standard deviations for various aspects of the SBE experiments for all three *MUC5B* SNPs.

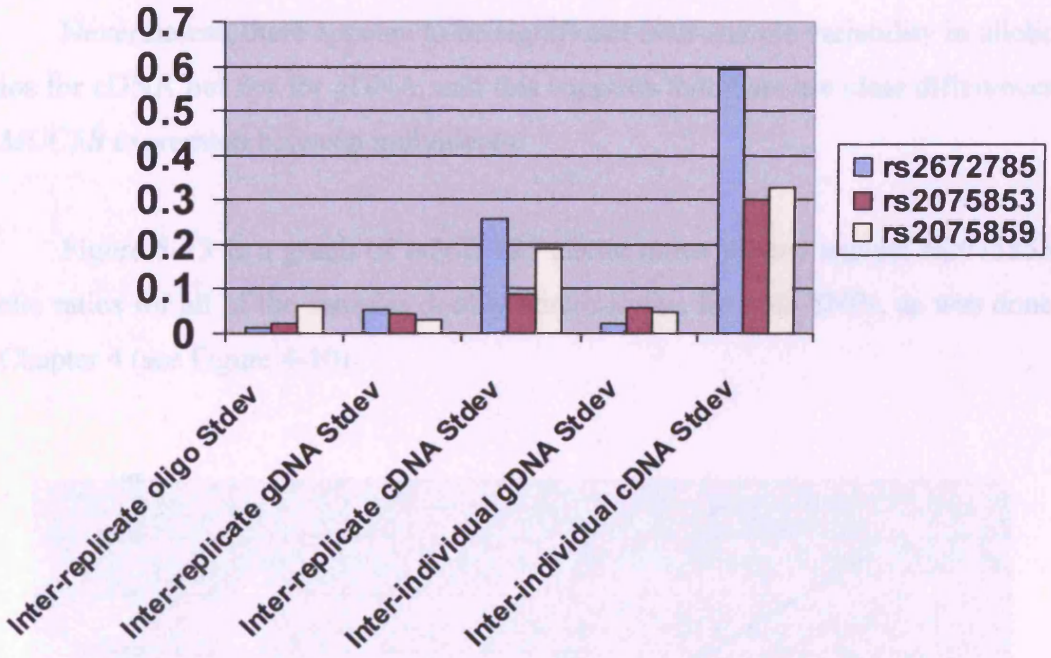


Figure 5-12 - Graph of standard deviations for SBE experiments. Inter-replicate oligonucleotide Stdev calculated for three replicates per SNP. Inter-replicate gDNA and cDNA Stdev values are averaged Stdev across all the samples tested. Inter-individual gDNA and cDNA Stdev represent the variation in mean allelic ratios across all the samples tested.

Several observations can be made from these results. First, there is a very small standard deviation in allelic ratios between the oligonucleotide mixture replicates, implying that the SBE reaction and detection of SBE products are highly reproducible (since the oligonucleotide mixtures do not require a PCR step prior to the SBE reaction). The reproducibility of the SBE reaction was confirmed in experiments in which the SBE reactions were repeated using the same PCR templates and the reproducibility was shown to be good (data not shown).

Second, the low standard deviation in allelic ratios for gDNA suggests that the reproducibility of the readings is good, whereas it appears to be noticeably less so for cDNA. Given that the cDNA used is from the same RT, it follows that the experimental variability must arise from the subsequent steps in the entire process. Since we have already established that the SBE reaction and detection steps are highly reproducible, this suggests that the major source of experimental variability is at the PCR step.

Nevertheless, there appears to be significant inter-sample variability in allelic ratios for cDNA but not for gDNA, and this suggests that there are clear differences in *MUC5B* expression between individuals.

Figure 5-13 is a graph of rs2672785 allelic ratios plotted against rs2075853 allelic ratios for all of the samples doubly heterozygous for both SNPs, as was done in Chapter 4 (see Figure 4-10).

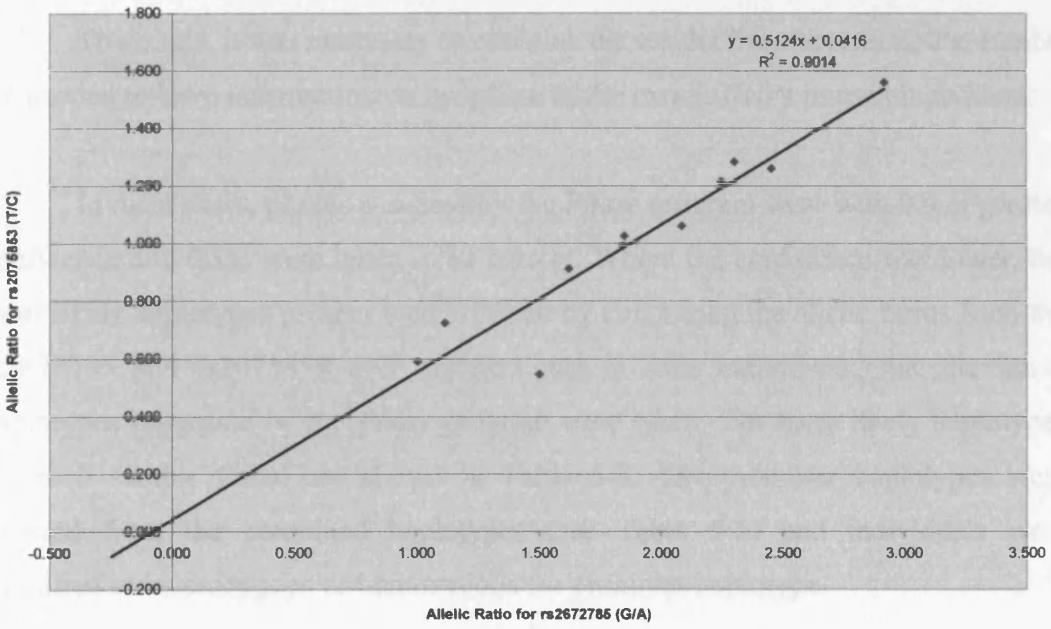


Figure 5-13 - Graph of rs2075853 (T/C) allelic ratios against rs2672785 (G/A) allelic ratios. Each point on graph represents the averaged allelic ratios for a different individual.

This new graph plotted using a larger data set confirms that there is in fact a very good correlation between the allelic ratios from both SNPs. In addition, this graph also implies that the G allele (rs2782785) and T allele (rs2075853) are in phase

for all of the samples. Again, however, the allelic ratios for rs2672785 are inflated in comparison to rs2075853, implying that despite our best efforts with the technique, the allelic ratios obtained using different SNPs are not directly interchangeable. No correlation was found between the other SNP markers (data not shown), probably because of variation in phase, which was to be expected from the breakdown of LD (see section 5.4).

5.7 RELATING SBE RESULTS TO HAPLOTYPES

It was now possible to address the question of whether the *MUC5B* promoter haplotypes are differentially expressed *in vivo*. If the promoter haplotypes are differentially expressed *in vivo*, we hypothesised that the samples heterozygous for promoter haplotype would be more likely to exhibit a significant difference in *MUC5B* allelic expression than samples that are homozygous for promoter haplotype.

To do this, it was necessary to combine the results from several SNPs. Hence, we needed to have information on the phase of the exonic SNPs in each individual:

In most cases, phases assigned by the *Phase* program were with 0.9 or greater confidence and these were taken to be correct. Where the confidence was lower, the most likely haplotypes present were inferred by comparing the allelic ratios from the rs2672785 and rs2075859 SNP markers and in four individuals, the alternative haplotypes suggested by the *Phase* program were taken. The most likely haplotypes for each sample tested are shown in Table 5-8. The promoter haplotypes were inferred from the combined haplotypes (see Table 5-5) and individuals were classified as heterozygous or homozygous for promoter haplotype.

The allelic ratios from all three SNPs were then be averaged and are shown in Table 5-8 as 'Averaged Corrected cDNA Allelic Ratios'. These averaged allelic ratios represent the difference in *MUC5B* expression between the two *MUC5B haplotypes*.

As an initial analysis, the averaged corrected cDNA allelic ratios from individuals homozygous for promoter haplotype were compared to those from individuals heterozygous for promoter haplotype (see Figure 5-14).

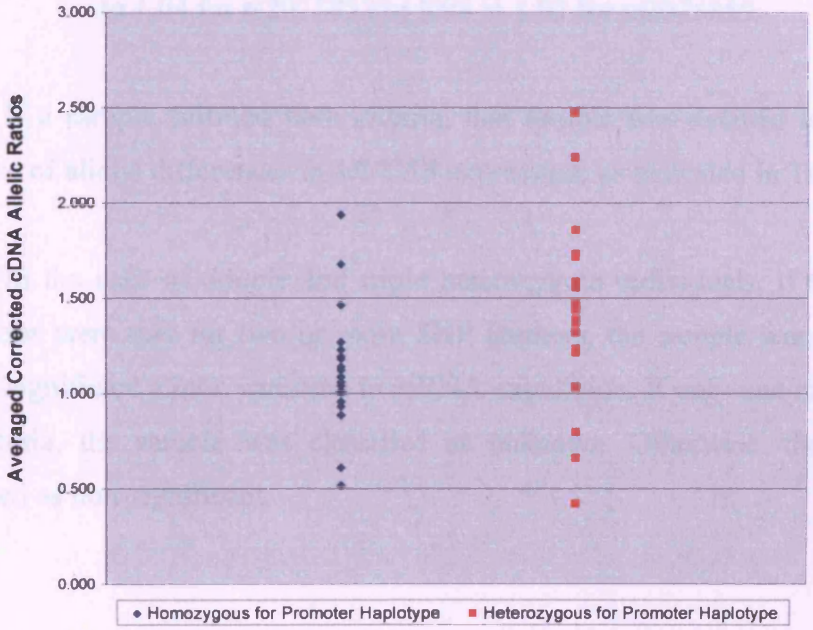


Figure 5-14 - Comparison of averaged corrected cDNA allelic ratios in individuals homozygous for promoter haplotype vs. individuals heterozygous for promoter haplotype

As can be seen, the averaged corrected cDNA allelic ratios for homozygous individuals tends to cluster around 1.0 (no allelic differences in expression) whilst the allelic ratios for heterozygous individuals tend to vary more widely. Indeed, a t-test indicates that the distribution of allelic ratios is significantly different between the two groups (P Value = 0.012). This suggests that there is a significant association between heterozygosity vs. homozygosity for promoter haplotype and the extent of the allelic differences detected.

In order to test the robustness of this observation, the data was also analysed in another way. Here the SNP markers were not averaged and were considered separately as follows:

- The allelic ratio was beyond the 0.8 to 1.2 range. This range was shown to discriminate from a control gene where there was no difference in allelic expression (Yan et al. 2002).

- The 95% confidence interval for the cDNA mean allelic ratio did not overlap with the 95% confidence interval derived from the gDNA overall mean allelic ratio. In practice, samples were not significant if their confidence intervals extend into 0.99 to 1.01 for rs2672785, 0.96 to 1.04 for rs2075853 or 0.98 to 1.02 for rs2075859.

If a sample fulfilled both criteria, that sample was deemed to show strong evidence of allelic differences in *MUC5B* expression, as indicated in Table 5-8.

In the case of double and triple heterozygote individuals, if the above two conditions were met for two or more SNP markers, the sample was considered to exhibit significant allelic variation in mRNA expression. If only one marker fulfilled the criteria, the sample was classified as unknown. Otherwise, the sample was classified as non-significant.

Samples	Promoter Haplotypes	Exonic Haplotypes	Averaged Corrected cDNA Allelic Ratios	Strong Evidence of Allelic Difference in *MUC5B* expression?
3SG	H1/H8	X6/X2	1.855	?
4SG	H3/H3	X2/X1	0.887	No
6T	H3/H3	X2/X1	1.124	No
9G	**H4/H8**	**X3/X1**	**1.574**	**Yes**
10T	H6/H6	X2/X1	1.040	No
11G	**H1/H8**	**X5/X3**	**1.215**	**Yes**
12G	H3/H3	X2/X1	0.610	No
14G	H3/H3	X2/X1	1.140	No
15G	H3/H3	X2/X1	0.885	No
15T	H3/H3	X2/X1	1.062	No
16T	H4/H3	X3/X2	1.111	No
17T	H8/H8	X3/X1	1.193	No
19T	H1/H3	X6/X2	1.572	?
26L	H5/H3	X3/X2	1.715	?
28L	*H1/H3*	*X5/X2*	1.321	No
29L	H8/H3	X3/X2	1.586	?
30T	H8/H3	X3/X1	1.439	No
34L	**H1/H3**	**X6/X2**	**1.730**	**Yes**
43L	**H1/H3**	**X6/X2**	**2.239**	**Yes**
44G	**H1/H3**	**X6/X1**	**1.470**	**Yes**
45G	H5/H3	X3/X1	1.234	No
53L	H3/H3	X2/X1	1.938	No
57L	H3/H3	X2/X1	1.270	No
63L	H3/H3	X2/X1	1.088	No
66L	H8/H3	X3/X1	2.472	No
67L	***H1/H8***	***X5/X2***	1.556	**Yes**
71L	**H3/H2**	**X2/X1**	0.661	**Yes**
73L	H3/H3	X2/X1	1.006	No
86T	H3/H3	X2/X1	1.016	No
88G	H3/H3	X2/X1	1.008	No
88T	H3/H3	X2/X1	1.087	No
92L	**H4/H3**	**X3/X1**	**0.420**	**Yes**
94T	**H1/H3**	**X6/X1**	**1.388**	**Yes**
97G	H3/H3	X6/X1	0.957	No
97L	H3/H3	X6/X1	0.931	No
98T	H3/H3	X2/X1	1.182	No
100L	H4/H3	X4/X1	1.371	?
101G	*H15/H3*	*X6/X2*	0.797	?
103T	**H3/H3**	**X2/X1**	**1.465**	**Yes**
112L	H1/H1	X6/X5	1.679	No
116G	H3/H3	X2/X1	1.230	No
116L	H3/H3	X2/X1	0.518	No

Table 5-8 - Promoter, exonic haplotypes and averaged corrected cDNA allelic ratios for each individual tested. Allelic ratios represent difference in expression of first haplotype vs. second haplotype in exonic and promoter haplotype columns. Yes: Sample shows strong evidence of significant allelic differences in *MUC5B* expression (also in bold). No: Sample does not show strong evidence of significant allelic differences in *MUC5B* expression. ?: Unknown. Haplotypes shown in italics are those where the alternative haplotypes suggested by the *Phase* program were used.

This information, together with promoter haplotype heterozygosity is tabulated in Table 5-9.

	Heterozygous for Promoter Haplotype	Homozygous for Promoter Haplotype
Allelic Difference in *MUC5B* Expression	9	1
No Allelic Difference in *MUC5B* Expression	5	21

Table 5-9 - 2 by 2 table showing number of samples showing strong evidence of allelic variation in *MUC5B* expression and whether they are heterozygous or homozygous for promoter haplotype

A Fisher's Exact Test (P value <0.0002) showed that there was a highly significant difference in distribution between promoter haplotype heterozygosity and significance of allelic variation in mRNA expression, and that samples which were heterozygous for promoter haplotype were much more likely to demonstrate strong evidence of differences in allelic expression.

Hence, these results strongly supported the original hypothesis that the *MUC5B* promoter haplotypes are differentially expressed *in vitro* as well as *in vivo*.

5.7.1 Inferring Relative Transcriptional Activities of Promoter Haplotypes

The data was also examined in another way. Using all the data, the averaged corrected cDNA allelic ratios from the four most common diplotypes (H3H3, H1H3, H1H8 and H3H8) were compared.

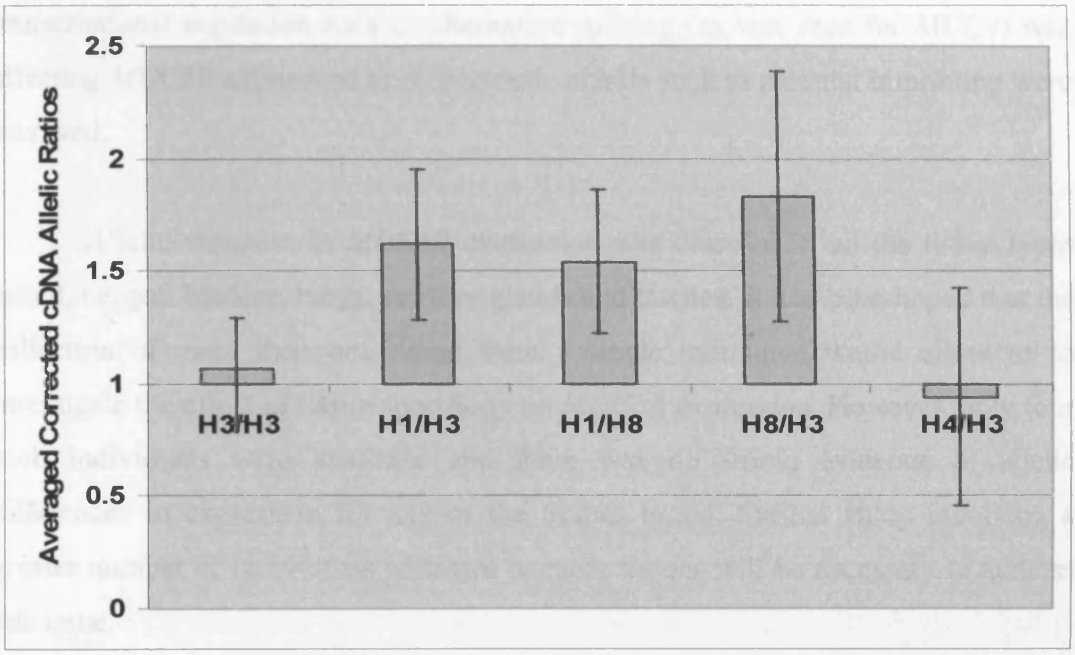


Figure 5-15 – Comparison of averaged corrected cDNA allelic ratios between common diplotypes. Allelic ratios shown derived from taking an average from all the individuals of a particular diplotype. Error bars are one standard deviation from the mean.

Figure 5-15 clearly shows that H1 is more highly expressed than H3 and H8, whilst H8 is more highly expressed than H3. However, H4 is either lower than, equal to, or higher than H3. Thus, it is difficult to fit all possible haplotypes into a scheme to rank their relative expression levels. This was the case even when only using the more stringently selected samples showing the strongest evidence of allelic differences in *MUC5B* expression (see Table 5-8). This is probably at least in part due to H3 for which there is clearly also variability in allelic expression in the homozygotes.

5.8 DISCUSSION

The work in this chapter demonstrates that *MUC5B* exhibits allelic variation in mRNA expression in fetal tissue. Large inter-sample differences were detected, although the extent of the difference in allelic expression was generally small; the greatest difference detected between the expression of two alleles in a single individual was about two-fold. In none of the samples was there evidence of anything approaching mono-allelic expression, which might have been found if post-

transcriptional regulation such as alternative splicing (as was seen for *MUC4*) was affecting *MUC5B* expression or if epigenetic effects such as parental imprinting were involved.

Allelic variation in *MUC5B* expression was detected in all the tissue types tested, i.e. gall bladder, lungs, salivary glands and trachea. It had been hoped that the collection of more than one tissue from a single individual would allow us to investigate the effect of tissue-specificity on *MUC5B* expression. However, only four such individuals were available and there was no strong evidence of allelic differences in expression for any of the tissues tested. Further study involving a greater number of individuals with two or more tissues will be necessary to address this issue.

Slight inter-individual variations were observed in gDNA allelic ratios. Single factor ANOVA analyses indicated that these differences are statistically significant for rs2075853 and rs2075859 (P values <0.005 and <0.000005 respectively), although not for rs2672785 (P value = 0.977).

The nature of these inter-individual differences is unclear. These could be due to an unknown polymorphism(s) present within the SBE primer-annealing site or within the gDNA PCR primer-annealing sites. Either might have led to altered primer binding efficiencies in some individuals, which could have ultimately affected the allelic ratios. With the benefit of hindsight, it would have been better to first sequence the primer-annealing sites to check for unknown polymorphisms, but this could not be done retrospectively for this project due to time-constraints.

In an attempt to allow for polymorphisms under the primers, cDNA allelic ratios were corrected in two ways: first, by using an overall gDNA allelic ratio for each SNP, derived from all the individuals tested; second by using individualised gDNA allelic ratios. The corrected cDNA allelic ratios were then compared between the two correction methods. The rationale behind this was that if there were unknown polymorphisms under the primer-annealing sites, these are highly unlikely to occur in every single individual. There was very little difference in the results when

corrected using either method, suggesting that this variation was not an important confounder.

The inter-replicate standard deviations for cDNA were noticeably higher than for gDNA. The small inter-replicate standard deviations for the oligonucleotide mixes suggested that the main source of variability between replicates occurred at the PCR stage. The PCR step is common to both gDNA and cDNA samples and is performed in a similar fashion, albeit with different primers and thermocycling conditions. It follows that the difference in reproducibility between gDNA and cDNA is most likely due to the nature of the samples themselves:

The expression of *MUC5B* in fetal tissue is known to be lower compared to in adults (Buisine et al. 1999b; Reid, Gould, and Harris 1997). Thus, it is possible that the lack of starting template could affect the final relative proportions of each allelic transcript in the PCR product, leading to variations in allelic ratios between different PCR replicates. This could have resulted in the greater standard deviation in allelic ratios for cDNA compared to gDNA. Nevertheless, this is puzzling in view of the fact that many investigators routinely use 40 or more cycles of amplification for quantitative work, whilst here only 35 cycles were used.

It would be useful to repeat the SBE experiments using cDNA obtained from a different reverse transcription reaction. Whilst the reproducibility of the cDNA allelic ratios might still be an issue, this would nonetheless allow us to take an average of the results from two separate RTs.

In this chapter, a significant correlation between heterozygosity for promoter haplotype and differential *MUC5B* allelic expression was demonstrated. By using the data obtained from SBE experiments, an attempt was made to infer the relative expression levels of the promoter haplotypes. This was met with some difficulty due to the lack of comparisons between some of the haplotypes as well as the fact that the inference was based on a small number of samples. As such, it was not possible to elucidate the relative expression levels of every promoter haplotype present in the samples tested. Nevertheless, it was quite clear that haplotype H1 is expressed more highly than H3 and H8. The association with promoter haplotype and expression

levels is an argument against parental imprinting as the cause of the observed allelic differences in *MUC5B* expression.

Reporter assays had previously demonstrated that promoter haplotype H1 was the higher expressing than H3, which was in turn expressed at a higher level than H2 (Kamio et al. 2005). Here, we confirm part of this study that H1 is higher expressing than H3 *in vivo*.

The SBE results (on the basis of one sample) suggest that H2 is higher expressing than H3, unlike what was previously shown (Kamio et al. 2005). However, here we provide strong suggestive evidence of variation in H3 expression levels, which could reflect the involvement of more distant regulatory elements not examined in the reporter assays.

The H1 haplotype differs at four nucleotide positions from the H3 haplotype and at three from the H8 haplotype (see Table 5-1). The sequences surrounding and including these four differences were analysed on the TRANSFAC database (http://motif.genome.jp/MOTIF.html) using the default cut-off score of 85.

One of the nucleotide differences is at rs7115457, which was previously reported to reside in a putative NF-κβ transcription factor-binding site (Chen et al. 2001) (see Figure 5-1). This binding site was not detected when the default cut-off score was used but was detected when the score was reduced. The A allele on the H1 haplotype background results in the disruption of this binding site. This is somewhat counterintuitive since one might have expected the higher expressing H1 haplotype to be associated with the more active NF-κβ transcription factor-binding site, leading to higher expression at least under conditions of inflammation.

Another nucleotide difference is at rs7118568, which resides in a putative NRF-2 (nuclear factor, erythroid 2 related factor) binding site, detected using the default cut-off score. The G allele on the H1 haplotype background again results in the disruption of this binding site. Interestingly, NRF-2 has been shown to play an important role in protection against oxidative stress and influences inflammatory processes and is thought to be important in diseases such as emphysema and cancer

(Cho, Reddy, and Kleeberger 2006; Ishii et al. 2005; Yu and Kensler 2005). If the binding site were to be genuine, it is plausible that *MUC5B* expression might be upregulated by NRF-2 in response to oxidative stress.

CHAPTER 6

STUDY OF *MUC5B* EXPRESSION IN RELATION TO RESPIRATORY OUTCOMES IN A LONGITUDINAL BIRTH COHORT

6 STUDY OF *MUC5B* EXPRESSION IN RELATION TO RESPIRATORY OUTCOMES IN A LONGITUDINAL BIRTH COHORT

6.1 INTRODUCTION

6.1.1 The 1946 Birth Cohort

The MRC National Survey of Health and Development is a sample of all births in England, Wales and Scotland in a week (3–9th March) in 1946. The 1946 birth cohort was originally composed of 16,695 individuals before it was stratified by social class and reduced to 5362 members. The cohort study was started at the time of the immediate post-war baby boom in Britain, before the subsequent mass immigration into Britain. Hence, the vast majority of the participants are of Northern European ancestry.

This population was studied at birth, a further 10 times up to age 15 years, and then 11 more times in adulthood, most recently at 53 years. Data including birth weight, height, adult respiratory outcomes (for example diagnosed with asthma, wheezing), FEV_1 (forced expiratory volume in 1 second) was collected.

At age 53, 3035 of the 3673 (83%) with whom contact was attempted provided information. The rest of the original sample had died, previously refused information or were living abroad. Importantly, the study sample at 53 years was shown to be representative of the national population of a similar age when compared with Census data (Wadsworth et al. 2003).

Acute challenges to the respiratory tract, such as infectious pathogens, toxins and allergens activate lung inflammatory/immune response mediators. Some of these inflammatory mediators can function as secretagogues to activate secretion of mucins from surface goblet cells and/or glandular secretory cells, whilst others upregulate *MUC* gene expression, resulting in sustained mucin hypersecretion (see section

1.2.5.3). Since mucins have been shown to bind bacteria and play an important role in protection of the epithelial cell layer, these raised levels of mucin production are part of the normal response to inflammation, and typically revert to baseline levels once the inflammatory mediators have been removed, presumably in response to anti-inflammatory mechanisms.

However, in several chronic inflammatory airway diseases such as asthma, chronic bronchitis and cystic fibrosis, patients typically manifest goblet cell hyperplasia and glandular hyperplasia/hypertrophy (Fahy 2002), which results in permanently elevated baseline levels of mucin production. As a result, mucus hypersecretion and mucus airway obstruction are prominent features of such respiratory diseases.

Besides an increase in expression levels, the composition and properties of mucins can be altered in disease states. For example, an increase in the low-charge glycoform of MUC5B was demonstrated in airway secretions in asthmatics vs. controls (Sheehan et al. 1999a), whilst MUC5B glycoproteins of an extreme size were discovered in the mucus plugs from a patient who died of acute asthma (Sheehan et al. 1995a).

From the work in the previous chapter, it was shown that promoter haplotype H1 was a high expressing haplotype and that the exonic haplotype X6 was a good marker for this haplotype. In chronic inflammatory airway diseases such as asthma and bronchitis, the expression of *MUC5B* is raised compared to normal (Kirkham et al. 2002). Our hypothesis therefore was that carriers of the X6 *MUC5B* exonic haplotype would be over-represented in groups of people with these diseases. In addition to this characterised variation, other variations within the *MUC5B* gene, in which there may be an association with the *MUC5B* SNPs tested, may alter bacterial binding or affect MUC5B properties in other ways.

This chapter examines the possibility of significant associations between *MUC5B* genotypes/exonic haplotypes, respiratory outcomes and measures of lung function in the 1946 birth cohort. The focus is on asthma and measures of lung function, although various other associated variables are also examined.

6.2 GENOTYPE & HAPLOTYPE ANALYSIS

For this study, genotyping was conducted as part of a multiplex analysis, which consisted of other markers of interest to the group. This assay was optimised by adjusting the relative amounts of each SBE primer in the reaction mix. A representative result from the SBE multiplex used to genotype the cohort for the *MUC5B* exonic SNPs and three other loci is shown in Figure 6-1.

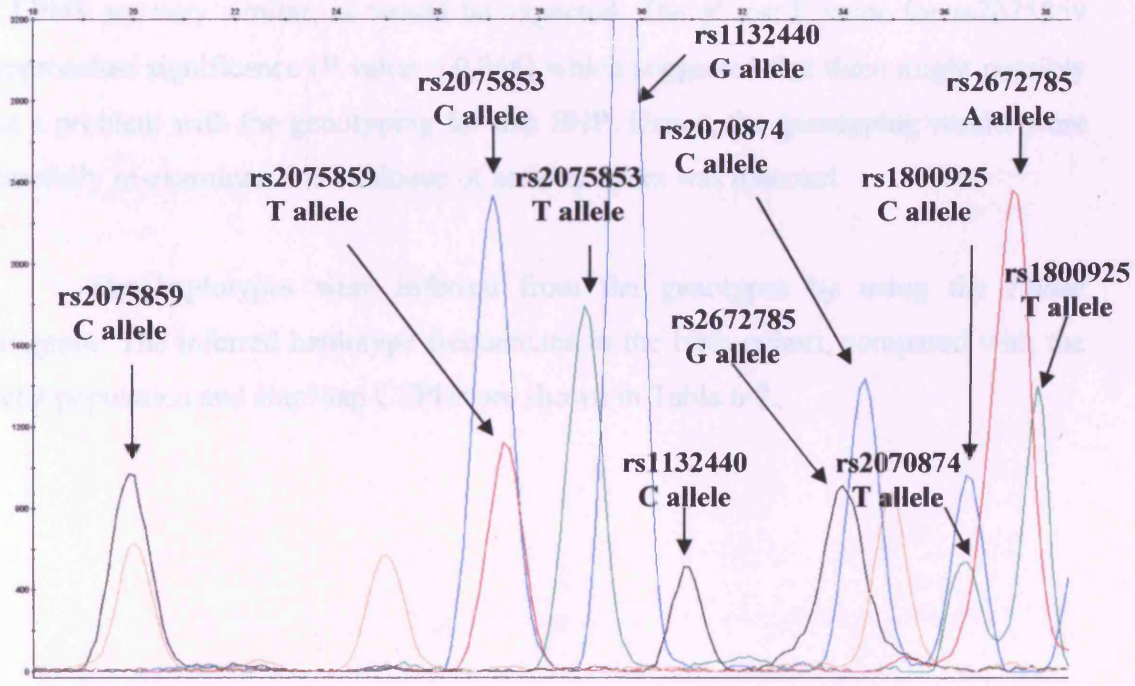


Figure 6-1 - Representative genotyping result from SBE multiplex. Individual heterozygous at all six loci shown (Sample N188). Orange peaks are LIZ-120 size standards. All samples run on *ABI 3730XL* sequencer and genotyped automatically using *Genemapper 4.0*.

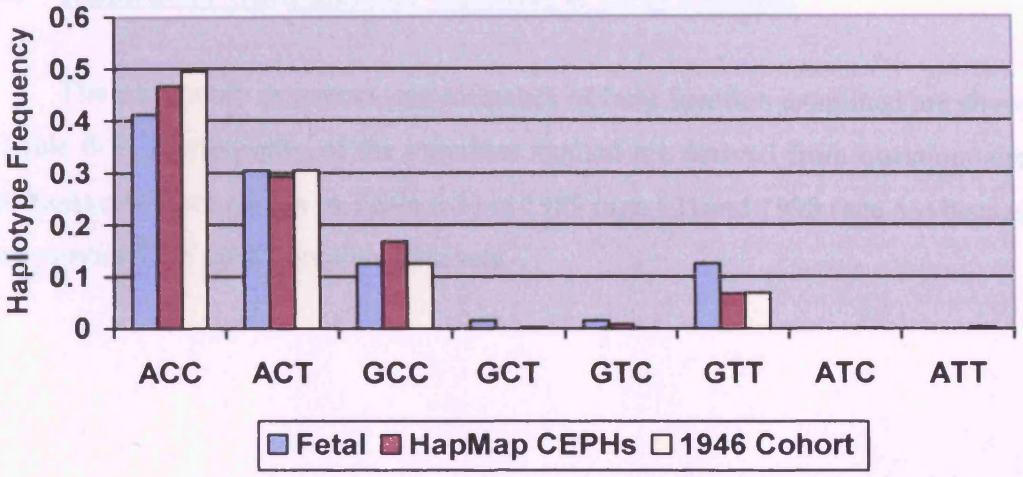
Table 6-1 shows the *MUC5B* genotype frequencies for the 1946 population again compared with the genotype frequencies for the HapMap CEPHs sample set. χ^2 tests were performed for each SNP to test for deviation from expected Hardy-Weinberg frequencies.

SNP	rs2672785			rs2075853			rs2075859		
Genotypes	AA	AG	GG	CC	CT	TT	CC	CT	TT
1946 Cohort Genotype Frequencies (n=2807)	0.643 (1804)	0.322 (905)	0.035 (98)	0.858 (2409)	0.137 (384)	0.005 (14)	0.399 (1120)	0.449 (1260)	0.152 (427)
χ^2 Test P value for HW in 1946 cohort		0.491			0.953			0.066	
CEPH Genotype Frequencies (n=60)	0.567 (34)	0.383 (23)	0.050 (3)	0.850 (51)	0.150 (9)	0.000 (0)	0.417 (25)	0.450 (27)	0.133 (8)

Table 6-1- *MUC5B* exonic SNP genotype frequencies in 1946 cohort vs. HapMap CEPHs

As can be seen, the genotype frequencies between the cohort and HapMap CEPHS are very similar, as would be expected. The χ^2 test P value for rs2075859 approached significance (P value = 0.066) which suggested that there might possibly be a problem with the genotyping for this SNP. Hence, the genotyping results were carefully re-examined. No evidence of serious errors was detected.

The haplotypes were inferred from the genotypes by using the *Phase* program. The inferred haplotype frequencies in the birth cohort, compared with the fetal population and HapMap CEPHs are shown in Table 6-2.



Exonic Haplotype Identifier	rs2672785	rs2075853	rs2075859	Haplotype Freq		
				1946 cohort	Fetal	HapMap CEPHs (n=2807)
X1	A	C	C	0.495	0.413	0.467
X2	A	C	T	0.304	0.304	0.292
X3	G	C	C	0.125	0.125	0.167
X4	G	C	T	0.002	0.016	0.000
X5	G	T	C	0.001	0.016	0.008
X6	G	T	T	0.068	0.125	0.067
X7	A	T	C	0.001	0.000	0.000
X8	A	T	T	0.004	0.000	0.000

Table 6-2 - *MUC5B* exonic SNP haplotype frequencies in 1946 cohort, fetal and HapMap CEPH populations

Haplotype inference showed that the haplotype frequencies between the three population groups are highly similar. Two new exonic haplotypes were identified in the 1946 cohort: X7 and X8, although both are very rare in the population.

6.3 VARIOUS VARIABLES TESTED AND RATIONALE FOR TESTING THEM

This section describes the variables tested for correlation with the *MUC5B* exonic genotypes and X6 haplotype carrier status.

6.3.1 Respiratory Outcomes and Measures of Lung Function

The respiratory outcomes and measures of lung function examined are shown in Table 6-3. The majority of the variables studied are derived from questionnaires (questions asked are shown in Table 6-3) in 1989 (age 43) and 1999 (age 53) because these provided the most complete data sets.

Original Variable	Explanation/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	ALLG89R	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
ALLERGY	In the last ten years (1989-1999), have you had any allergies?	1 – Yes 2 – No	-	-
ASTH89	Have you ever had asthma (1946-1989)?	0 – No 1 – Yes, once 2 – Yes, recurring	ASTH89R	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
ASTHMA	In the last ten years (1989-1999), did you have asthma?	1 – Yes 2 – No	-	
BRONC89	Have you ever had bronchitis (1946-1989)?	0 – No 1 – Yes, once 2 – Yes, recurring	BRONC89R	0 – Never (original code 0) 1 – Ever (original codes 1 or 2)
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	-	-
HAY89	Have you ever had hay fever (1946-1989)	0 – No 1 – Yes, once 2 – Yes, recurring	HAY89R -	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		

Original Variable	Explanation/Question asked	Original outcome codes	Recoded variable	Recoded outcome codes
LRIP	Cohort member's mother was asked if cohort member 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		
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		
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 old.	-	FEVM99D	Outliers removed: 0.00 - 0.07, 0.31 - 1.00, 9.70 – 9.99 litres
FEVM89C – FEVM99D	Difference in Max FEV1 over 10 years (1989-1999).	-	DELTA	-

Table 6-3 – Respiratory outcomes and measures of lung function examined in 1946 cohort. Answers from questionnaires were converted into outcome codes in order to be analysed. Some of the original variables were recoded for the analysis performed in this project and the recoded outcome codes are shown above.

As can be seen, the questions asked were not identical on the different occasions. Although subsidiary questions were asked concerning whether or not the doctor or health worker had confirmed the diagnosis, previous experience showed that this served to reduce the data set slightly and only altered responses marginally (i.e. most answering 'yes' to these questions had a doctor diagnosis).

The rationale for choosing to examine these particular variables is described in the following sections:

6.3.1.1 Asthma History (ASTH89R and ASTHMA)

Asthma is a respiratory disease in which there is constriction of the airways and an over-secretion of mucus in the respiratory tract. In the most severe cases of asthma, the excess mucus forms gelatinous plugs, which occlude the airways (Rogers 2004; Sidebotham and Roche 2003). Asthmatics also suffer from decreased lung function, which in a large proportion of patients is due to mucus hypersecretion (Rogers 2004). Since mucins are the main glycoproteins found in mucus secretions, mucus hypersecretion is by extension associated with an increase in mucin production (Morcillo and Cortijo 2006), and relative changes in mucin gene expression will alter the composition of the secreted mucins.

Research has shown that the airway immune response plays an important part in the clinical manifestation of this disease (reviewed in (Ying et al. 2006b)). Atopy (a tendency to develop allergic reactions and is characterised by elevated levels of allergy-specific IgE antibodies) is commonly associated with asthma, although some asthmatics are non-atopic (Humbert et al. 1999)

Allergens are processed and presented by antigen-presenting cells such as dendritic cells and macrophages. Depending on the cytokines acting on the precursor cells at the time of antigen presentation, Th1 or Th2 cells are produced. The two T helper cell types secrete different cytokine patterns, which influence other immune cells. This distinction between the two T helper cell types is critical because only the

type 2 pattern of the Th2 cells (IL4, IL5, IL6, ILL9, IL10 and IL13) is associated with asthma (Nakamura and Hoshino 2005; Robinson et al. 1992).

It is widely accepted that atopy and also susceptibility to asthma is heritable and a number of twin studies have shown an increased concordance between monozygotic vs. dizygotic twins (Duffy et al. 1990; Nieminen, Kaprio, and Koskenvuo 1991). In addition, other studies have demonstrated that asthma and its associated phenotypes are more prevalent amongst relatives of asthmatics than in relatives of non-asthmatics, lending further support for the notion that there is a genetic component in this disease (Sibbald, Horn, and Gregg 1980; Sibbald and Turner-Warwick 1979; Ying et al. 2006a).

Various genes have been shown to be associated with asthma. These include ADAM33, (a disintegrin and metalloprotease (ADAM)33) (Holgate et al. 2006; Holloway et al. 2004), *GPRA* (G-protein-coupled receptor for asthma susceptibility), *IL4* (interleukin 4) (reviewed in (Holloway, Beghe, and Holgate 1999; Holloway and Koppelman 2007)) and genes associated with atopy such as IL13.

Two genetic studies on *MUC* genes have been carried out by our lab to investigate the possibility of an association with susceptibility to asthma. The first of which established an association between *MUC7* and asthma (Vinall et al. 2000), with a particular allele (*MUC7*5*) found at significantly lower frequency in asthmatics (Kirkbride et al. 2001) vs normal. The second identified an association between longer *MUC2* TR alleles and a cohort of atopic, non-asthmatic patients than in asthmatic patients. In addition, significant linkage between 11p15 (where the *MUC5B* gene resides) and asthma in Caucasians has been discovered (A genome-wide search for asthma susceptibility loci in ethnically diverse populations. The Collaborative Study on the Genetics of Asthma (CSGA) 1997). Taken together, this suggests that the expression levels and properties of MUC5B could be important:

Together with MUC5AC, MUC5B is a major mucin species found in airway secretions from asthma patients and its expression has been shown to be increased in asthma patients compared to normal controls (Groneberg et al. 2002; Kirkham et al.

2002). In addition, the properties of MUC5B have been shown to be altered in asthma (Sheehan et al. 1995b; Sheehan et al. 1999b)

Thus, the *MUC5B* genotypes/haplotypes an individual possesses could play a prominent role in the etiology of asthma and this respiratory outcome was therefore examined in this project.

6.3.1.2 Bronchitis History (BRONC89R and BRONC)

Bronchitis is an inflammation of the bronchial tubes, accompanied by coughing symptoms, mucus hypersecretion and is often secondary to an upper respiratory tract viral infection. Chronic bronchitis is a more persistent condition that is strongly associated with smoking, although not all smokers get chronic bronchitis (Urrutia et al. 2005). The BRONC89R and BRONC variables would include participants suffering from chronic bronchitis as well as some for whom bronchitis was due to a one off-secondary infection.

6.3.1.3 Wheeze (WZY89C and WZYC)

The wheeze variable is a measure of breathing difficulties, which results from a range of respiratory dysfunction including asthma, chronic obstructive pulmonary disease and chronic bronchitis. This variable is a simple concept for the participant to answer and was previously shown to correlate well with FEV_1 and be associated with *MUC7* variation (Rousseau et al. 2006).

6.3.1.4 Hay Fever History (HAY89R and HAYF)

Hay fever or seasonal allergic rhinitis is characterised by sneeze attacks and itchy runny noses in sufferers, triggered by pollen during early spring to late summer in the UK. The main distinction with this variable and the aforementioned ones is that although there is mucus hypersecretion, airway obstruction is much less of a feature of hay fever.

6.3.1.5 Allergy History (ALLG89R and ALLERGY)

The allergy variable is a broad measure that would include various allergic conditions such as allergic rhinitis and skin allergies. Unfortunately, the way the data was collected for this variable (cohort members were asked 'Have you ever had any allergies?') means that the answers given were largely dependent on the subject's understanding of what constitutes an allergy. Thus, for example, whilst hay fever is an allergic disease, a cohort member might have answered 'yes' to whether he/she ever had hay fever but 'no' to whether he/she ever had an allergy. Indeed, examination of the data showed that whilst there was a significant correlation between positive responses for allergy and asthma/hay fever, there were discrepancies.

6.3.1.6 Lower Respiratory Tract Infections in Infancy (LRIPY)

This variable represents the presence or absence of lower respiratory tract infections in members of the cohort during early childhood (0-24 mths), determined from interviewing the mothers of the cohort members at age 2. Since mucins are known to bind bacteria and are believed to play a protective role in defence against infection, it was possible that variation in *MUC5B* expression might be associated with susceptibility to early childhood infections Hence, there was interest in investigating the relationship between this variable and *MUC5B*.

6.3.1.7 Measures of Lung Function - Fixed Expiratory Volume in one second (FEV_1) (FEVM89C, FEVM99D and DELTA)

Lung function was measured using a Micro Medical Micro Plus turbine spirometer (Stirling, Scotland) at ages 43 and 53 years by nurses during home visits. Two readings of FEV_1 were taken and the maximum reading (in litres) was used to derive the FEVM89C and FEVM99D measures. DELTA represents the decline (because lung function decreases with age) in FEV_1 over 10 years, calculated using

the FEVM89C and FEM99D values. Reduced FEV_1 results from poor lung development and also from congestion and constriction of the airways (Vestbo, Prescott, and Lange 1996).

6.3.2 Potential Confounding Variables

A confounding variable is a variable that is correlated with an outcome of interest. Such variables need to be taken into consideration when testing for an association between respiratory outcomes and *MUC5B*, since these may result in a spurious correlation between the two.

The variables shown in Table 6-4 were identified as potential confounders and were tested for correlation with *MUC5B*, so that any that showed a positive correlation could be included in an adjusted model.

Original Variable	Explanation/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
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
SEXX	Sex of participant	1 – Male 2 – Female	SEXXR	Male – 0 Female – 1

Table 6-4 - Potentially Confounding Variables. Answers from questionnaires were converted into outcome codes in order to be analysed. Some of the original variables were recoded for the analysis performed in this project and the recoded outcome codes are shown above.

6.4 RESULTS OF TESTS FOR ASSOCIATION WITH VARIABLES

The results of the tests for association with *MUC5B* exonic SNPs/X6 haplotype carrier status and respiratory outcomes, using ANOVA, cross-tabulations and χ^2 tests are shown in Table 6-5. The results of the tests for association with *MUC5B* and potential confounding variables are shown in Table 6-6. All analysis was carried out using *SPSS*.

Variables Tested	Type of Test Used	
ALG89R	Crosstab	

Variables Tested	Type of Test Used	Explanation	rs2672785		rs2075853		rs2075859		*MUC5B* Haplotype	
			No. Tested	P Value	No. Tested	P Value	No. Tested	P Value	No. Tested	P Value
CIG89CR	Crosstab	Cigarette smoking history (1966-1989)	2626	**0.038**	2627	0.604	2623	**0.026**	2626	0.591
CIG99CR	Crosstab	Cigarette smoking history (1966-1999)	2799	0.121	2800	0.847	2796	**0.032**	2799	0.789
FSC50R	Crosstab	Father's social class at age 4	2572	0.740	2573	0.700	2569	0.661	2572	0.407
SCL89R	Crosstab	Social class at age 43	2472	0.269	2473	0.723	2469	0.858	2890	0.569
SCL99R	Crosstab	Social class at age 53	2517	0.723	2518	0.857	2514	0.310	2517	0.796
REG46AR	Crosstab	Region of birth	2800	0.925	2801	0.267	2797	0.626	2800	0.185
SEXX	Crosstab	Sex	2800	0.599	2801	0.812	2797	0.155	2800	0.500

Table 6-6 – Results of tests for correlation with potential confounding variables. Significant (P value <0.05) correlations are shown in bold.

As can be seen from the results, there are significant associations between *MUC5B* and a number of respiratory outcomes (ALLERGY, HAY89R, HAYF and WZY89C, see Table 6-5) as well as a significant correlation between cigarette smoking and *MUC5B* (see Table 6-6), although it is not clear why this should be. Cigarette smoking history is also marginally correlated with ALLERGY in this cohort (P value = 0.082, analysis done by A.Teixeira) and is therefore a confounding variable that must be adjusted for.

All analyses were adjusted for region of birth (associated with HAY89R (P value = 0.009) and ALLERGY (P value = 0.010), analysis done by A.Teixeira) even though it did not appear to be associated with *MUC5B,* because hidden population stratification is a well-known genetic confounder.

In the case of the measures of lung function (FEVM89C, FEVM99D and DELTA), FEV_1 values are usually expressed in a clinical setting as a percentage as expected for age, height and sex. In this project, since raw FEV_1 readings were used, these had to be adjusted for height and sex. An adjustment for age was unnecessary since the population tested is a birth cohort.

6.5 ADJUSTMENT FOR CONFOUNDING VARIABLES

The *MUC5B* alleles that exhibited evidence of a significant association with respiratory outcomes in the preliminary analyses (i.e. only bolded values; see Table 6-5,) were reanalysed by logistic regression using *STATA 10*, to correct for cigarette smoking history and region of birth.

Variables Tested	Type of Test Used	Explanation	rs2672785		rs2075853		rs2075859		*MUC5B* Haplotype	
			No. Tested	P Value	No. Tested	P Value	No. Tested	P Value	No. Tested	P Value
ALLERGY	Crosstab	Allergy history (1989-1999)	2798	0.543	2799	**0.043**	2795	**0.039**	2798	**0.021**
	Adjusted Log Regression	Allergy history (1989-1999) adjusted for cigarette smoking history and region of birth	-	-	2799	**0.033**	2795	**0.039**	2798	**0.017**
HAY89R	Crosstab	Hay fever history (1946-1989)	2628	0.489	2629	**0.030**	2625	0.090	2628	**0.049**
	Adjusted Log Regression	Hay fever history (1946-1989) adjusted for cigarette smoking history and region of birth	-	-	2623	0.060	-	-	2622	0.054
HAYF	Crosstab	Hay fever history (1989-1999)	2798	0.311	2799	**0.050**	2795	0.635	2798	0.749
	Adjusted Log Regression	Hay fever history (1989-1999) adjusted for cigarette smoking history and region of birth	-	-	2799	0.104	-	-	-	-
WZY89C	Crosstab	Wheeze history (1946-1989)	2541	**0.022**	2542	0.317	2538	0.398	2541	0.054
	Adjusted Log Regression	Wheeze history (1946-1989) adjusted for cigarette smoking history and region of birth	2535	**0.029**	-	-	-	-	2535	0.062
FEVM89C	ANOVA	Max FEV_1 at age 43	2577	0.325	2578	0.837	2574	0.473	2577	0.583
	Adjusted Log Regression	Max FEV_1 at age 43 adjusted for height and sex	2563	0.927	2564	0.878	2560	0.634	2563	0.989
FEVM99D	ANOVA	Max FEV_1 at age 53	2693	0.799	2694	0.721	2690	0.131	2577	0.815
	Adjusted Log Regression	Max FEV_1 at age 53 adjusted for height and sex	2684	0.699	2685	0.892	2681	0.095	2684	0.477
DELTA	ANOVA	Change in Max FEV_1 over 10 years (1989-1999)	2492	0.803	2493	0.811	2489	0.082	2492	0.589
	Adjusted Log Regression	Change in Max FEV_1 over 10 years (1989-1999) adjusted for height and sex	2483	0.738	2484	0.793	2480	0.057	2483	0.535

Table 6-7 - Comparison of results before and after adjusting for confounding variables. Significant (P value <0.05) correlations are shown in bold. Original results from cross tabulations compared with adjusted results from logistic regressions.

After adjusting for confounding variables, the associations previously detected between *MUC5B* and hay fever (HAY89R and HAYF) are now no longer statistically significant. This suggests that it is most likely that cigarette smoking history was confounding the results and that there is no significant link between *MUC5B* and hay fever.

The measures of lung function were still not significantly associated with *MUC5B* after adjustments.

In contrast, the associations between *MUC5B* and allergy history (ALLERGY) and between *MUC5B* and wheeze history (WZY89C) have remained significant, suggesting that there may be a genuine correlation between these outcomes and *MUC5B*.

6.6 TEST FOR INTERACTION BETWEEN *MUC5B* AND *IL13*

Allergies are associated with a production of elevated levels of IgE antibody. In individuals suffering from allergies, IgE antibodies bind to receptors on mast cells and basophil surfaces, releasing mediators that initiate an immunological cascade and the accompanying inflammatory symptoms. IL13 is a cytokine produced by activated Th2 cells and plays an essential role in IgE synthesis. Thus, IL13 is important in the development of allergies.

A SNP in the IL13 gene (IL13 R110Q) involves an amino-acid substitution and has been shown to be associated with an increase in IgE (Graves et al. 2000; Howard et al. 2001; Leung et al. 2001). This polymorphism is thought to be functionally important in ligand binding and IL13 110Q was shown to be significantly more active in IgE switching than IL13 110R in a cellular model (Vladich et al. 2005). Indeed, work in our group showed that this polymorphism was correlated significantly with asthma and the allergy variables in this cohort (see Table 6-8).

Respiratory Outcome	P Value	
	Log Regression	Crosstabs
ASTH89R	0.002	<0.0001
ASTHMA	Not done	0.042
ALLG89R	0.042	0.030
ALLERGY	Not done	0.007

Table 6-8 - Results of tests for association with IL13 R100Q genotypes and asthma and allergy history variables. Analysis done by A.Teixiera. P values for logistic regression and cross tabulations shown. Significant association not observed for wheeze history.

Table 6-9 shows the logistic regression result for ALLERGY, using IL13 Q carriers vs. non-carriers.

Respiratory Outcome	Odds Ratio and 95% CI	P Value
ALLERGY (Allergy history 1989-1999)	1.320 (1.071-1.629)	0.009

Table 6-9 – Result of logistic regression analysis between IL13 Q carriers and ALLERGY. Analysis done by A.Teixeira.

Published research suggests that IL13 induces *MUC5AC* expression (Liu et al. 2004; Yasuo et al. 2006). Thus, it is conceivable that IL13 might affect *MUC5B* expression levels in a similar fashion. Therefore, a test for interaction[2] between *MUC5B* and IL13 R110Q, affecting allergy history (ALLERGY), which was shown to be significantly correlated with *MUC5B* (see Table 6-7) was performed. The results of the tests for interaction, using logistic regression are shown in Table 6-10.

MUC5B marker	P Value
rs2075853	0.4469
rs2075859	**0.0358**
MUC5B Haplotype	0.7270

Table 6-10 – P values obtained for tests for interaction between *MUC5B* and *IL13E* influencing allergy history outcomes. Statistically significant (P value <0.05) results are shown in bold. All results adjusted for cigarette smoking history, region of birth, sex and social class.

The results show that the rare allele of IL13 together with the common allele of rs2075859 confers a greater risk of allergy. Curiously, the results indicate that there is a significant interaction between rs2075859 and IL13 R110Q but not

[2] Two variables are said to be interacting when together they exert a multiplicative effect on an outcome of interest

between rs2075853 and *MUC5B* haplotype, suggesting that the *IL13* only interacts with certain *MUC5B* genotypes to influence allergy.

6.7 DISCUSSION

The work in this chapter has shown that there is a significant correlation between certain *MUC5B* genotypes/haplotypes, wheeze history (WZY89C) and allergy history (ALLERGY). In some cases, there is an association with a respiratory outcome for rs2672785 or rs2075853 but never both together. This is unexpected since both SNPs are in strong LD with each other (see section 5.4).

People counts for the statistically significant cross-tabulations are shown in Appendix II. For wheeze, there are more G allele (rs2672785) carrying individuals with wheeze than without wheeze, which was expected since the G allele is on the high expressing X6 haplotype background. Curiously, the X6 haplotype is under-represented in individuals with allergies rather than the other way around. It is not obvious why this is the case, although it is plausible that the high expressing X6 haplotype is protective against allergic responses, acting by clearing allergens more efficiently from the nasal and respiratory tracts.

The significant correlations observed between *MUC5B* genotypes and cigarette smoking history are quite unexpected, as this implies that there is a difference in distribution of *MUC5B* alleles between smokers (both current and ex) and non-smokers. It is difficult to propose a rationale explanation for this and it is most likely that the positive correlation is simply due to chance.

There are a number of possible reasons for the significant correlations observed in this study:

If the associations are genuine, it is possible that the significant associations detected reflect differences in MUC5B properties, rather than differences in mRNA expression levels, since both rs2672785 (G: glycine, A: glutamine) and rs2075853

(T: tryptophan, C: arginine) result in non-synonymous, non-conserved amino acid changes, which could alter the physical properties of MUC5B in airway secretions.

Alternatively, the SNPs tested might be in LD with other functional substitutions in *MUC5B* and the positive correlations observed could be a reflection of this.

The significant associations seen could also have resulted from Type I errors (false positives). These could be due to problems such as population stratification and systematic genotyping errors. Population stratification is unlikely to be a problem since the 1946 cohort pre-dates the mass immigration into the UK beginning in the 1950s. The possibility of systematic genotyping errors has been addressed at the beginning of this study by checking for significant deviations from expected Hardy-Weinberg frequencies (see section 6.2)

However, given the number of variables tested, there is a risk that the significant correlations observed have arisen due to multiple testing. A Bonferroni correction can be applied to help control for this (Bland and Altman 1995). Since some of the variables, such as HAYF and ALLERGY, are likely to be closely related to each other (i.e. not independent), a standard Bonferroni correction is likely to be overly conservative (thereby increasing Type II errors). Hence, other statistical methods such as *false discovery rate* and *permutation testing* might be more appropriate for such variables (Benjamini et al. 2001; Doerge and Churchill 1996).

In order to better address the possibility of Type I errors in this study, it would be useful to test other cohorts for *MUC5B*, such as the 1958 longitudinal cohort (Power 1992) and a small asthma cohort we have in our lab, which has yet to be tested.

It would be also be interesting to test the 1946 cohort for other markers in the 11p15.5 *MUC* gene complex, as is currently being done for a SNP in *MUC5AC* by another lab member.

CHAPTER 7

FINAL DISCUSSION AND CONCLUSIONS

7 FINAL DISCUSSION AND CONCLUSIONS

There is mounting evidence to suggest that *cis*-acting regulatory variants are widespread throughout the genome and that allelic variation in mRNA expression might be a common occurrence. Here, a principal aim of this project was to address the issue of whether *MUC4* and *MUC5B* exhibited allelic variation in mRNA expression. In order to achieve this aim, a single-base extension (SBE) based method was developed. It was hoped that using this technique would enable us to test for allelic variation in *MUC4* expression and to determine if TR length had any influence on levels of expression. Unfortunately, we had not anticipated the extent to which *MUC4* was alternatively spliced and the various complications caused by this led us to abandon this aspect of the project.

However, background data showed that *MUC4* TR length was associated with SNPs downstream of the TR domain, and by inference from HapMap, also upstream. The RT PCR data showed evidence of more extensive alternative splicing than previously reported. Interestingly, there seem to be tissue and inter-individual differences in this and preliminary experiments suggested that this might be *cis*-acting. This clearly needs further investigation.

The rest of this project focussed on *MUC5B.* Using the SBE method, clear allelic differences in *MUC5B* expression were demonstrated. These differences were relatively small, which was probably why the various technical issues affected the results so much.

Fetal tissue was used as a source of mRNA for testing the constitutive expression of *MUC5B in vivo*. The initial plan was to compare the cDNA allelic ratios from different tissues originating from the same individual in order to test for tissue-specificity in *MUC5B* allelic variation. Unfortunately, difficulties in obtaining sufficient *MUC5B* cDNA template from a sizeable proportion of the samples meant that this comparison could only be done for a small handful of individuals. The results were therefore inconclusive.

In retrospect, the decision to collect multiple fetal tissue types was a double-edged sword. Whilst doing so allowed the opportunity to test for tissue-specificity in a wide range of different tissue types, it reduced the total number of individuals of a single tissue type available for testing. As a result, in order to increase the total number of individuals tested, allelic ratios from different tissue sources had to be used. Thus, it is quite likely that if tissue-specific effects were present, this might have masked the true extent of allelic variation in *MUC5B* mRNA expression.

Although it might have been better to concentrate on just one tissue type, there is still the possibility of *intra*-tissue specificity to contend with, as demonstrated by Wilkins and colleagues for bone morphogenic protein 5 (*BMP5*) (Wilkins et al. 2007). This is likely to be more crucial for the more complex tissues, such as stomach and lungs, where a different subset of cell types could conceivably be sampled each time. The best way to approach this problem is unclear.

Nevertheless, a highly significant association between promoter haplotype heterozygosity and the magnitude of the allelic differences in *MUC5B* expression *in vivo* was demonstrated. This finding supported a previous *in vitro* study in which differential expression of *MUC5B* promoter haplotypes was shown. Importantly, this finding also provided evidence for the allelic differences observed being due to *cis-acting* regulatory variants rather than an epigenetic effect such as parental imprinting.

An attempt was made to rank the relative expression levels of the different promoter haplotypes. Whilst this was possible for a few of the haplotypes, this could not be accomplished for every haplotype due to the small numbers of samples and the variability in allelic ratios even for individuals homozygous for promoter haplotype. Nevertheless, the H1 promoter haplotype was clearly established as a high expressing *MUC5B* haplotype.

Hence, carrier status for H1 and *MUC5B* exonic SNP genotypes were examined in a longitudinal birth cohort in relation to various respiratory outcomes. Significant correlations between wheeze and allergy history and *MUC5B* were

observed. If correct, rather than being a consequence of multiple testing, these might reflect differences in MUC5B properties rather than in mRNA expression levels, possibly due to other polymorphisms associated with the ones tested here. Further work on the cohort by testing for association with other polymorphisms in the 11p15.5 *MUC* gene complex might help to shed further light on this.

It would have been interesting to test markers in the cohort that could distinguish between the H3 and H8 promoter haplotypes, since H8 was shown to be higher expressing than H3. Unfortunately, since the cohort work was done concurrently with the SBE work, the information regarding the relative expression levels of H3 vs. H8 was not available in time.

In conclusion, the work in this thesis has clearly demonstrated that there is constitutive allelic variation in *MUC5B* mRNA expression and that this variation is most likely due to *cis*-acting regulatory polymorphisms in the promoter. It would be of great interest to test for this phenomenon in other *MUC* genes.

It would now be interesting to test for facultative allelic differences in *MUC5B* expression. It is possible that the true significance and extent of the allelic differences in expression are manifested only in the context of inflammation and that much greater allelic differences in expression might be observed. Respiratory cells grown at the air liquid interface (ALI culture), which can differentiate into mucus producing and ciliated cells, would be an ideal cellular model to study this (Gray et al. 1996). The ALI culture cells could be tested using the SBE method, after challenging the cells with inflammatory mediators known to upregulate *MUC5B* expression, such as human neutrophil defensins and PMA (Aarbiou et al. 2004; Yuan-Chen et al. 2007). The advantage of this kind of approach is that cells from the same individual could be tested before and after challenging with inflammatory mediators, thereby controlling for many experimental variables.

It would also be interesting to see if high expressing alleles/haplotypes respond more (or perhaps less) to certain challenges. For example, by using this approach, it should be possible to test whether NF-κβ and NRF-2 play a role in controlling *MUC5B* expression and whether this differs in H1 and H3 haplotypes.

One could for example treat ALI culture cells with bacterial LPS, which is known to upregulate NF-κβ and then examine the relative allelic expression in heterozygotes for promoter haplotype. Since there is suggestive evidence of more distant regulators such as enhancers and/or silencers, it would be important to test this in several different individuals, as well as to test individuals homozygous for promoter haplotype.

Ultimately, it will be possible to test whether the differences in allelic expression also correlate with total expression levels and with the relative amounts of MUC5B protein produced by these alleles. This would be an ultimate demonstration of allelic differences in mRNA expression.

APPENDICES

Appendix I – MUC4 Biliary Tract Cancer Study

I contributed to this project by:

- Supervising and assisting with the western blotting, including experimental design.
- Making the first biliary RNA sample, showing that it was possible to extract RNA from bile.

Are MUC4 and/or MUC5AC Useful Tumour Markers for Biliary Tract Cancer (BTC)?

Wolf-rudiger Matull, Andrew Loh, Andreola Fausto, Zwelal Adiguzel, Maesha Deheragoda, Uzma Qureshi, Dallas M. Swallow, Stephen P. Pereira

Introduction: Alterations in epithelial mucin expression are associated with carcinogenesis. MUC4 contains an EGF-like domain that can induce cell proliferation and differentiation via the ERBB2 receptor (1). In pancreatic disease, MUC4 is a diagnostic and prognostic marker for malignancy (2) (3), while MUC5AC has been proposed as a serum marker for BTC (4). Methods: We investigated MUC4 and MUC5AC expression in: (i) 79 archive Biliary tissues (69) BTC, 10 benign) by immunohistochemistry (IHC), and (ii) bile and serum specimens from 72 patients with biliary obstruction (39 BTC, 8 other malignancies, 9 primary sclerosing cholangitis (PSC), 16 benign) by western blot and quantitative real-time RT PCR (qPCR). Two monoclonal MUC4 antibodies (against transmembrane and secretory unit epitopes) and a mono- and polyclonal MUC5AC antibody were used. qPCR was based on the Taqman® method. Mucin gene primers were custom-synthesised. MUC4 and MUC5AC mRNA expression was normalised against the housekeeping gene GAPDH. Results: In archive tissues, MUC4 and MUC5AC proteins were detected exclusively in 37% and 10% of BTC samples, respectively, vs. in none of the benign ($p=0.02$ and $p+0.4$) respectively). In bile, MUC4 protein was detected in 27% of BTC and 29% of PSC cases, but not in benign conditions

(p=0.04 for BTC vs. non-BTC). qPCR revealed a 2.3-fold increased expression of *MUC4* mRNA in the bile of patients with BTC compared with benign disease. No such difference was seen for MUC5AC. In serum, MUC4 was detected in 5% of patients overall, with no significant difference between BTC and non-BTC patients. MUC5ACwas found exclusively in BTC and PSC sera (44% and 13%, respectively; p<0.001 for BTC vs. non-BTC). Conclusion: Biliary MUC4 and serum MUC5AC are highly specific tumour associated mucins in BTC, which may be of potential value in the early detection of BTC. References: (1) Nature Rev 2004; 4:45-60. (2) Br J Cancer 2004;91:1633-8. (3) J Clin Pathol 2005;58:845-52. (4) Cancer Lett 2003;195:93-9.

BTC prediction markers

(Western Blotting)	Bile MUC4	Serum MUC5AC	Bile MUC4 or serum MUC5AC
Sensitivity	27% (9/34)	44% (17/39)	58% (22/38)
Specificity	93% (26/28)	96% (26/27)	87% (22/23)
PPV	82% (9/11)	94% (17/18)	88% (22/25)
NPV	51% (26/51)	54% (26/48)	56% (20/36)

Appendix II – People Counts for Statistically Significant Cross Tabulations in 1946 Birth Cohort

Variable	Marker	P Value	Genotype	Never		Ever		Total
				No. Tested	% of Total	No. Tested	% of Total	
HAYF	*MUC5B* Haplotype	0.749	Non-carrier	1984	86.75%	446	87.28%	2430
			X6 Carrier	303	13.25%	65	12.72%	368
			TOTAL	*2287*	*100.00%*	*511*	*100.00%*	*2798*
	rs2075853	0.050	CC	1962	85.75%	439	85.91%	2401
			CT	318	13.90%	66	12.92%	384
			TT	8	0.35%	6	1.17%	14
			TOTAL	*2288*	*100.00%*	*511*	*100.00%*	*2799*
HAY89	rs2075853	0.030	CC	1902	86.57%	359	83.10%	2261
			CT	287	13.06%	68	15.74%	355
			TT	8	0.36%	5	1.16%	13
			TOTAL	*2197*	*100.00%*	*432*	*100.00%*	*2629*
	MUC5B Haplotype	0.049	Non-carrier	1922	87.52%	363	84.03%	2285
			X6 Carrier	274	12.48%	69	15.97%	343
			TOTAL	*2196*	*100.00%*	*432*	*100.00%*	*2628*
WZY89	rs2672785	0.022	AA	1555	64.98%	84	56.76%	1639
			AG	762	31.84%	54	36.49%	816
			GG	76	3.18%	10	6.76%	86
			TOTAL	*2393*	*100.00%*	*148*	*100.00%*	*2541*
ALLERGY	rs2075853	0.043	CC	1896	84.95%	505	89.07%	2401
			CT	324	14.52%	60	10.58%	384
			TT	12	0.54%	2	0.35%	14
			TOTAL	*2232*	*100.00%*	*567*	*100.00%*	*2799*
	rs2075859	0.039	CC	879	39.43%	236	41.70%	1115
			CT	990	44.41%	263	46.47%	1253
			TT	360	16.15%	67	11.84%	427
			TOTAL	*2229*	*100.00%*	*566*	*100.00%*	*2795*
	MUC5B Haplotype	0.021	Non-carrier	1921	86.10%	509	89.77%	2430
			X6 Carrier	310	13.90%	58	10.23%	368
			TOTAL	*2231*	*100.00%*	*567*	*100.00%*	*2798*

Variable	Marker	P Value	Genotype	Never		Ever		Total
				No. Tested	% of Total	No. Tested	% of Total	
CIG89CR	rs2672785	0.038	AA	511	65.94%	1185	64.02%	1696
			AG	229	29.55%	612	33.06%	841
			GG	35	4.52%	54	2.92%	89
			TOTAL	*775*	*100.00%*	*1851*	*100.00%*	*2626*
	rs2075859	0.026	CC	332	42.84%	717	38.80%	1049
			CT	314	40.52%	855	46.27%	1169
			TT	129	16.65%	276	14.94%	405
			TOTAL	*775*	*100.00%*	*1848*	*100.00%*	*2623*
CIG99CR	rs2075859	0.032	CC	340	42.24%	776	38.98%	1116
			CT	330	40.99%	923	46.36%	1253
			TT	135	16.77%	292	14.67%	427
			TOTAL	*805*	*100.00%*	*1991*	*100.00%*	*2796*

REFERENCES

1. (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.

2. Aarbiou J, Verhoosel RM, Van Wetering S, De Boer WI, Van Krieken JH, Litvinov SV, Rabe KF, and Hiemstra PS (2004) Neutrophil defensins enhance lung epithelial wound closure and mucin gene expression in vitro. *Am.J.Respir.Cell Mol.Biol.* 30 (2):193-201.

3. Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyren P, Uhlen M, and Lundeberg J (2000) Single-nucleotide polymorphism analysis by pyrosequencing. *Anal.Biochem.* 280 (1):103-110.

4. Alam J and Cook JL (1990) Reporter genes: application to the study of mammalian gene transcription. *Anal.Biochem.* 188 (2):245-254.

5. Allen A (1981) Structure and function of gastrointestinal mucus. In: Physiology of the Gastrointestinal Tract. In Johnson L.R (ed).Raven Press, NY.

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

7. Allen A, Hutton DA, Leonard AJ, Pearson JP, and Sellers LA (1986) The role of mucus in the protection of the gastroduodenal mucosa. *Scand.J.Gastroenterol.Suppl* 125:71-78.

8. Alos L, Lujan B, Castillo M, Nadal A, Carreras M, Caballero M, de Bolos C, and Cardesa A (2005) Expression of membrane-bound mucins (MUC1 and MUC4) and secreted mucins (MUC2, MUC5AC, MUC5B, MUC6 and MUC7) in mucoepidermoid carcinomas of salivary glands. *Am.J.Surg.Pathol.* 29 (6):806-813.

9. Andrianifahanana M, Moniaux N, Schmied BM, Ringel J, Friess H, Hollingsworth MA, Buchler MW, Aubert JP, and Batra SK (2001) Mucin (MUC) gene expression in human pancreatic adenocarcinoma and chronic pancreatitis: a potential role of MUC4 as a tumor marker of diagnostic significance. *Clin.Cancer Res.* 7 (12):4033-4040.

10. Anthony J.F.Griffiths et al (1999) Regulation of Gene Transcription. Modern Genetic Analysis, 1 ed.W.H. Freeman and Company.

11. Arul GS, Moorghen M, Myerscough N, Alderson DA, Spicer RD, and Corfield AP (2000) Mucin gene expression in Barrett's oesophagus: an in situ hybridisation and immunohistochemical study. *Gut* 47 (6):753-761.

12. Audic Y and Hartley RS (2004) Post-transcriptional regulation in cancer. *Biol.Cell* 96 (7):479-498.

13. Audie JP, Janin A, Porchet N, Copin MC, Gosselin B, and Aubert JP (1993) Expression of human mucin genes in respiratory, digestive, and reproductive tracts ascertained by in situ hybridization. *J.Histochem.Cytochem.* 41 (10):1479-1485.

14. Audie JP, Tetaert D, Pigny P, Buisine MP, Janin A, Aubert JP, Porchet N, and Boersma A (1995) Mucin gene expression in the human endocervix. *Hum.Reprod.* 10 (1):98-102.

15. Badache A and Goncalves A (2006) The ErbB2 signaling network as a target for breast cancer therapy. *J.Mammary.Gland.Biol.Neoplasia.* 11 (1):13-25.

16. Bai CH, Song SY, and Kim YD (2007) Effect of Glucocorticoid on the MUC4 Gene in Nasal Polyps. *Laryngoscope.*

17. Balague C, Audie JP, Porchet N, and Real FX (1995) In situ hybridization shows distinct patterns of mucin gene expression in normal, benign, and malignant pancreas tissues. *Gastroenterology* 109 (3):953-964.

18. Balague C, Gambus G, Carrato C, Porchet N, Aubert JP, Kim YS, and Real FX (1994) Altered expression of MUC2, MUC4, and MUC5 mucin genes in pancreas tissues and cancer cell lines. *Gastroenterology* 106 (4):1054-1061.

19. Bax DA, Haringsma J, Einerhand AW, van Dekken H, Blok P, Siersema PD, Kuipers EJ, and Kusters JG (2004) MUC4 is increased in high grade intraepithelial neoplasia in Barrett's oesophagus and is associated with a proapoptotic Bax to Bcl-2 ratio. *J.Clin.Pathol.* 57 (12):1267-1272.

20. Benjamini Y, Drai D, Elmer G, Kafkafi N, and Golani I (2001) Controlling the false discovery rate in behavior genetics research. *Behav.Brain Res.* 125 (1-2):279-284.

21. Bergad PL, Towle HC, and Berry SA (1999) Definition of a high affinity growth hormone DNA response element. *Mol.Cell Endocrinol.* 150 (1-2):151-159.

22. Bernacki SH, Nelson AL, Abdullah L, Sheehan JK, Harris A, Davis CW, and Randell SH (1999) Mucin gene expression during differentiation of human airway epithelia in vitro. Muc4 and muc5b are strongly induced. *Am.J.Respir.Cell Mol.Biol.* 20 (4):595-604.

23. Bevilacqua A, Ceriani MC, Capaccioli S, and Nicolin A (2003) Post-transcriptional regulation of gene expression by degradation of messenger RNAs. *J.Cell Physiol* 195 (3):356-372.

24. Biemer-Huttmann AE, Walsh MD, McGuckin MA, Ajioka Y, Watanabe H, Leggett BA, and Jass JR (1999) Immunohistochemical staining patterns of MUC1, MUC2, MUC4, and MUC5AC mucins in hyperplastic polyps,

serrated adenomas, and traditional adenomas of the colorectum. *J.Histochem.Cytochem.* 47 (8):1039-1048.

25. Biesbrock AR, Bobek LA, and Levine MJ (1997) MUC7 gene expression and genetic polymorphism. *Glycoconj.J.* 14 (4):415-422.

26. Blackwood EM and Kadonaga JT (1998) Going the distance: a current view of enhancer action. *Science* 281 (5373):60-63.

27. Bland JM and Altman DG (1995) Multiple significance tests: the Bonferroni method. *BMJ* 310 (6973):170.

28. Blobe GC, Khan WA, Halpern AE, Obeid LM, and Hannun YA (1993) Selective regulation of expression of protein kinase C beta isoenzymes occurs via alternative splicing. *J.Biol.Chem.* 268 (14):10627-10635.

29. Boat TF and Cheng PW (1980) Biochemistry of airway mucus secretions. *Fed.Proc.* 39 (13):3067-3074.

30. Borchers MT, Carty MP, and Leikauf GD (1999) Regulation of human airway mucins by acrolein and inflammatory mediators. *Am.J.Physiol* 276 (4 Pt 1):L549-L555.

31. Bosch JA, de Geus EJ, Ligtenberg TJ, Nazmi K, Veerman EC, Hoogstraten J, and Amerongen AV (2000) Salivary MUC5B-mediated adherence (ex vivo) of Helicobacter pylori during acute stress. *Psychosom.Med.* 62 (1):40-49.

32. Brand AH, Breeden L, Abraham J, Sternglanz R, and Nasmyth K (1985) Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* 41 (1):41-48.

33. Brannan CI and Bartolomei MS (1999) Mechanisms of genomic imprinting. *Curr.Opin.Genet.Dev.* 9 (2):164-170.

34. Bray NJ, Buckland PR, Owen MJ, and O'Donovan MC (2003) Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum.Genet.* 113 (2):149-153.

35. Buisine MP, Colombel JF, Lecomte-Houcke M, Gower P, Aubert JP, Porchet N, and Janin A (1998a) Abnormal mucus in cap polyposis. *Gut* 42 (1):135-138.

36. Buisine MP, Desreumaux P, Debailleul V, Gambiez L, Geboes K, Ectors N, Delescaut MP, Degand P, Aubert JP, Colombel JF, and Porchet N (1999a) Abnormalities in mucin gene expression in Crohn's disease. *Inflamm.Bowel.Dis.* 5 (1):24-32.

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

38. Buisine MP, Devisme L, Maunoury V, Deschodt E, Gosselin B, Copin MC, Aubert JP, and Porchet N (2000) Developmental mucin gene expression in the gastroduodenal tract and accessory digestive glands. I. Stomach. A relationship to gastric carcinoma. *J.Histochem.Cytochem.* 48 (12):1657-1666.

39. Buisine MP, Devisme L, Savidge TC, Gespach C, Gosselin B, Porchet N, and Aubert JP (1998b) Mucin gene expression in human embryonic and fetal intestine. *Gut* 43 (4):519-524.

40. Burgel PR, Montani D, Danel C, Dusser DJ, and Nadel JA (2007) A morphometric study of mucins and small airway plugging in cystic fibrosis. *Thorax* 62 (2):153-161.

41. Burke TW and Kadonaga JT (1997) The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.* 11 (22):3020-3031.

42. Campion JP, Porchet N, Aubert JP, L'Helgoualc'h A, and Clement B (1995) UW-preservation of cultured human gallbladder epithelial cells: phenotypic alterations and differential mucin gene expression in the presence of bile. *Hepatology* 21 (1):223-231.

43. Caramori G, Di Gregorio C, Carlstedt I, Casolari P, Guzzinati I, Adcock IM, Barnes PJ, Ciaccia A, Cavallesco G, Chung KF, and Papi A (2004) Mucin expression in peripheral airways of patients with chronic obstructive pulmonary disease. *Histopathology* 45 (5):477-484.

44. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, and Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat.Genet.* 22 (3):231-238.

45. Carlstedt I, Lindgren H, and Sheehan JK (1983) The macromolecular structure of human cervical-mucus glycoproteins. Studies on fragments obtained after reduction of disulphide bridges and after subsequent trypsin digestion. *Biochem.J.* 213 (2):427-435.

46. Carlstedt I and Sheehan JK (1984) Macromolecular properties and polymeric structure of mucus glycoproteins. *Ciba Found.Symp.* 109:157-172.

47. Carraway KL, Price-Schiavi SA, Komatsu M, Idris N, Perez A, Li P, Jepson S, Zhu X, Carvajal ME, and Carraway CA (2000) Multiple facets of sialomucin complex/MUC4, a membrane mucin and erbb2 ligand, in tumors and tissues (Y2K update). *Front Biosci.* 5:D95-D107.

48. Carraway KL, Ramsauer VP, Haq B, and Carothers Carraway CA (2003) Cell signaling through membrane mucins. *Bioessays* 25 (1):66-71.

49. Carraway KL, III, Rossi EA, Komatsu M, Price-Schiavi SA, Huang D, Guy PM, Carvajal ME, Fregien N, Carraway CA, and Carraway KL (1999) An

intramembrane modulator of the ErbB2 receptor tyrosine kinase that potentiates neuregulin signaling. *J.Biol.Chem.* 274 (9):5263-5266.

50. Chauhan SC, Singh AP, Ruiz F, Johansson SL, Jain M, Smith LM, Moniaux N, and Batra SK (2006) Aberrant expression of MUC4 in ovarian carcinoma: diagnostic significance alone and in combination with MUC1 and MUC16 (CA125). *Mod.Pathol.* 19 (10):1386-1394.

51. Chen Y, Zhao YH, Di YP, 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.

52. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, and Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat.Genet.* 33 (3):422-425.

53. Cho HY, Reddy SP, and Kleeberger SR (2006) Nrf2 defends the lung from oxidative stress. *Antioxid.Redox.Signal.* 8 (1-2):76-87.

54. Chomczynski P (1993) A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. *Biotechniques* 15 (3):532-537.

55. Choudhury A, Moniaux N, Winpenny JP, Hollingsworth MA, Aubert JP, and Batra SK (2000) Human MUC4 mucin cDNA and its variants in pancreatic carcinoma. *J Biochem.(Tokyo)* 128 (2):233-243.

56. Coleman SL, Hoogendoorn B, Guy C, Smith SK, O'Donovan MC, and Buckland PR (2002) Streamlined approach to functional analysis of promoter-region polymorphisms. *Biotechniques* 33 (2):412, 414, 416.

57. Conaway RC and Conaway JW (1993) General initiation factors for RNA polymerase II. *Annu.Rev.Biochem.* 62:161-190.

58. Conrad B and Antonarakis SE (2007) Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. *Annu.Rev.Genomics Hum.Genet.*

59. Copin MC, Devisme L, Buisine MP, Marquette CH, Wurtz A, Aubert JP, Gosselin B, and Porchet N (2000) From normal respiratory mucosa to epidermoid carcinoma: expression of human mucin genes. *Int.J.Cancer* 86 (2):162-168.

60. Cox GN and Hirsh D (1985) Stage-specific patterns of collagen gene expression during development of Caenorhabditis elegans. *Mol.Cell Biol.* 5 (2):363-372.

61. Damera G, Xia B, Ancha HR, and Sachdev GP (2006) IL-9 modulated MUC4 gene and glycoprotein expression in airway epithelial cells. *Biosci.Rep.* 26 (1):55-67.

62. Damera G, Xia B, and Sachdev GP (2006) IL-4 induced MUC4 enhancement in respiratory epithelial cells in vitro is mediated through JAK-3 selective signaling. *Respir.Res.* 7:39.

63. Day DA and Tuite MF (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J.Endocrinol.* 157 (3):361-371.

64. de Groot PG (2002) The role of von Willebrand factor in platelet function. *Semin.Thromb.Hemost.* 28 (2):133-138.

65. Debailleul V, Laine A, Huet G, Mathon P, d'Hooghe MC, Aubert JP, and Porchet N (1998) Human mucin genes MUC2, MUC3, MUC4, MUC5AC, MUC5B, and MUC6 express stable and extremely large mRNAs and exhibit a variable length polymorphism. An improved method to analyze large mRNAs. *J Biol.Chem.* 273 (2):881-890.

66. Desseyn JL, Buisine MP, Porchet N, Aubert JP, 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.

67. Desseyn JL, Guyonnet-Duperat V, Porchet N, Aubert JP, and Laine A (1997) 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.

68. Ding C and Cantor CR (2003) A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc Natl.Acad.Sci.U.S.A* 100 (6):3059-3064.

69. Ding GQ and Zheng CQ (2007) The expression of MUC5AC and MUC5B mucin genes in the mucosa of chronic rhinosinusitis and nasal polyposis. *Am.J.Rhinol.* 21 (3):359-366.

70. Doerge RW and Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* 142 (1):285-294.

71. Dong Z, Thoma RS, Crimmins DL, McCourt DW, Tuley EA, and Sadler JE (1994) Disulfide bonds required to assemble functional von Willebrand factor multimers. *J.Biol.Chem.* 269 (9):6753-6758.

72. Duffy DL, Martin NG, Battistutta D, Hopper JL, and Mathews JD (1990) Genetics of asthma and hay fever in Australian twins. *Am.Rev.Respir.Dis.* 142 (6 Pt 1):1351-1358.

73. Durkin ME, Loechel F, Mattei MG, Gilpin BJ, Albrechtsen R, and Wewer UM (1997) Tissue-specific expression of the human laminin alpha5-chain, and mapping of the gene to human chromosome 20q13.2-13.3 and to distal mouse chromosome 2 near the locus for the ragged (Ra) mutation. *FEBS Lett.* 411 (2-3):296-300.

74. Dutcher SK and Hartwell LH (1983) Test for temporal or spatial restrictions in gene product function during the cell division cycle. *Mol.Cell Biol.* 3 (7):1255-1265.

75. Ekstrom TJ (1994) Parental imprinting and the IGF2 gene. *Horm.Res.* 42 (4-5):176-181.

76. Elsheikh MN and Mahfouz ME (2006) Up-regulation of MUC5AC and MUC5B mucin genes in nasopharyngeal respiratory mucosa and selective up-regulation of MUC5B in middle ear in pediatric otitis media with effusion. *Laryngoscope* 116 (3):365-369.

77. Escande F, Lemaitre L, Moniaux N, Batra SK, Aubert JP, and Buisine MP (2002) Genomic organization of MUC4 mucin gene. Towards the characterization of splice variants. *Eur.J Biochem.* 269 (15):3637-3644.

78. Evans R, Fairley JA, and Roberts SG (2001) Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev.* 15 (22):2945-2949.

79. Fahy JV (2002) Goblet cell and mucin gene abnormalities in asthma. *Chest* 122 (6 Suppl):320S-326S.

80. Fischer BM, Cuellar JG, Diehl ML, deFreytas AM, Zhang J, Carraway KL, and Voynow JA (2003) Neutrophil elastase increases MUC4 expression in normal human bronchial epithelial cells. *Am.J.Physiol Lung Cell Mol.Physiol* 284 (4):L671-L679.

81. Fowler J, Vinall L, and Swallow D (2001) Polymorphism of the human muc genes. *Front Biosci.* 6:D1207-D1215.

82. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, and Lee C (2006) Copy number variation: new insights in genome diversity. *Genome Res.* 16 (8):949-961.

83. Ge B, Gurd S, Gaudin T, Dore C, Lepage P, Harmsen E, Hudson TJ, and Pastinen T (2005) Survey of allelic expression using EST mining. *Genome Res.* 15 (11):1584-1591.

84. Gipson IK, Ho SB, Spurr-Michaud SJ, Tisdale AS, Zhan Q, Torlakovic E, Pudney J, Anderson DJ, Toribara NW, and Hill JA, III (1997) Mucin genes expressed by human female reproductive tract epithelia. *Biol.Reprod.* 56 (4):999-1011.

85. Gipson IK, Spurr-Michaud S, Moccia R, Zhan Q, Toribara N, Ho SB, Gargiulo AR, and Hill JA, III (1999) MUC4 and MUC5B transcripts are the prevalent mucin messenger ribonucleic acids of the human endocervix. *Biol.Reprod.* 60 (1):58-64.

86. Giuntoli RL, Rodriguez GC, Whitaker RS, Dodge R, and Voynow JA (1998) Mucin gene expression in ovarian cancers. *Cancer Res.* 58 (23):5546-5550.

87. Gober JW, Xu H, Dingwall AK, and Shapiro L (1991) Identification of cis and trans-elements involved in the timed control of a Caulobacter flagellar gene. *J.Mol.Biol.* 217 (2):247-257.

88. Gollub EG, Waksman H, Goswami S, and Marom Z (1995) Mucin genes are regulated by estrogen and dexamethasone. *Biochem.Biophys.Res.Commun.* 217 (3):1006-1014.

89. Goudet J, Raymond M, de Meeus T, and Rousset F (1996) Testing differentiation in diploid populations. *Genetics* 144 (4):1933-1940.

90. Graves PE, Kabesch M, Halonen M, Holberg CJ, Baldini M, Fritzsch C, Weiland SK, Erickson RP, von Mutius E, and Martinez FD (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.

91. Gray TE, Guzman K, Davis CW, Abdullah LH, and Nettesheim P (1996) Mucociliary differentiation of serially passaged normal human tracheobronchial epithelial cells. *Am.J.Respir.Cell Mol.Biol.* 14 (1):104-112.

92. Griffiths A et al (1999) Regulation of Gene Transcription. Modern Genetic Analysis, 1 ed.W.H. Freeman and Company.

93. Groneberg DA, Eynott PR, Lim S, Oates T, Wu R, Carlstedt I, Roberts P, McCann B, Nicholson AG, Harrison BD, and Chung KF (2002) Expression of respiratory mucins in fatal status asthmaticus and mild asthma. *Histopathology* 40 (4):367-373.

94. Gum JR, Jr. (1992) Mucin genes and the proteins they encode: structure, diversity, and regulation. *Am J Respir.Cell Mol.Biol.* 7 (6):557-564.

95. Hanaoka J, Kontani K, Sawai S, Ichinose M, Tezuka N, Inoue S, Fujino S, and Ohkubo I (2001) Analysis of MUC4 mucin expression in lung carcinoma cells and its immunogenicity. *Cancer* 92 (8):2148-2157.

96. Handra-Luca A, Lamas G, Bertrand JC, and Fouret P (2005) MUC1, MUC2, MUC4, and MUC5AC expression in salivary gland mucoepidermoid carcinoma: diagnostic and prognostic implications. *Am.J.Surg.Pathol.* 29 (7):881-889.

97. Hayashi R, Wada H, Ito K, and Adcock IM (2004) Effects of glucocorticoids on gene transcription. *Eur.J.Pharmacol.* 500 (1-3):51-62.

98. Hebbar V, Damera G, and Sachdev GP (2005) Differential expression of MUC genes in endometrial and cervical tissues and tumors. *BMC.Cancer* 5:124.

99. Ho SB, Niehans GA, Lyftogt C, Yan PS, Cherwitz DL, Gum ET, Dahiya R, and Kim YS (1993) Heterogeneity of mucin gene expression in normal and neoplastic tissues. *Cancer Res.* 53 (3):641-651.

100. Ho SB, Shekels LL, Toribara NW, Kim YS, Lyftogt C, Cherwitz DL, and Niehans GA (1995) Mucin gene expression in normal, preneoplastic, and neoplastic human gastric epithelium. *Cancer Res.* 55 (12):2681-2690.

101. Holbro T and Hynes NE (2004) ErbB receptors: directing key signaling networks throughout life. *Annu.Rev.Pharmacol.Toxicol.* 44:195-217.

102. Holgate ST, Yang Y, Haitchi HM, Powell RM, Holloway JW, Yoshisue H, Pang YY, Cakebread J, and Davies DE (2006) The genetics of asthma: ADAM33 as an example of a susceptibility gene. *Proc.Am.Thorac.Soc.* 3 (5):440-443.

103. Hollingsworth MA, Strawhecker JM, Caffrey TC, and Mack DR (1994) Expression of MUC1, MUC2, MUC3 and MUC4 mucin mRNAs in human pancreatic and intestinal tumor cell lines. *Int.J.Cancer* 57 (2):198-203.

104. Holloway JW, Beghe B, and Holgate ST (1999) The genetic basis of atopic asthma. *Clin.Exp.Allergy* 29 (8):1023-1032.

105. Holloway JW, Davies DE, Powell R, Haitchi HM, Keith TP, and Holgate ST (2004) The discovery and role of ADAM33, a new candidate gene for asthma. *Expert.Rev.Mol.Med.* 6 (17):1-12.

106. Holloway JW and Koppelman GH (2007) Identifying novel genes contributing to asthma pathogenesis. *Curr.Opin.Allergy Clin.Immunol.* 7 (1):69-74.

107. Homma H, Yamanaka A, Tanimoto S, Tamura M, Chijimatsu Y, Kira S, and Izumi T (1983) Diffuse panbronchiolitis. A disease of the transitional zone of the lung. *Chest* 83 (1):63-69.

108. Hovenberg HW, Davies JR, Herrmann A, Linden CJ, and Carlstedt I (1996) MUC5AC, but not MUC2, is a prominent mucin in respiratory secretions. *Glycoconj.J.* 13 (5):839-847.

109. Howard TD, Whittaker PA, Zaiman AL, Koppelman GH, Xu J, Hanley MT, Meyers DA, Postma DS, and Bleecker ER (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.

110. Hughes SE (1997) Differential expression of the fibroblast growth factor receptor (FGFR) multigene family in normal human adult tissues. *J.Histochem.Cytochem.* 45 (7):1005-1019.

111. HUGO Committee Human Gene Nomenclature Database http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl 1-12-0007

112. Human Genome Project Information SNP Fact Sheet http://www.ncbi.nlm.nih.gov/About/primer/snps.html 24-2-2006

113. Humbert M, Menz G, Ying S, Corrigan CJ, Robinson DS, Durham SR, and Kay AB (1999) The immunopathology of extrinsic (atopic) and intrinsic

(non-atopic) asthma: more similarities than differences. *Immunol.Today* 20 (11):528-533.

114. Imbert Y, Darling DS, Jumblatt MM, Foulks GN, Couzin EG, Steele PS, and Young WW, Jr. (2006) MUC1 splice variants in human ocular surface tissues: possible differences between dry eye patients and normal controls. *Exp.Eye Res.* 83 (3):493-501.

115. Inatomi T, Spurr-Michaud S, Tisdale AS, and Gipson IK (1995) Human corneal and conjunctival epithelia express MUC1 mucin. *Invest Ophthalmol.Vis.Sci.* 36 (9):1818-1827.

116. Ishii Y, Itoh K, Morishima Y, Kimura T, Kiwamoto T, Iizuka T, Hegab AE, Hosoya T, Nomura A, Sakamoto T, Yamamoto M, and Sekizawa K (2005) Transcription factor Nrf2 plays a pivotal role in protection against elastase-induced pulmonary inflammation and emphysema. *J.Immunol.* 175 (10):6968-6975.

117. Jayawickreme SP, Gray T, Nettesheim P, and Eling T (1999) Regulation of 15-lipoxygenase expression and mucus secretion by IL-4 in human bronchial epithelial cells. *Am.J.Physiol* 276 (4 Pt 1):L596-L603.

118. Jung HH, Lee JH, Kim YT, Lee SD, and Park JH (2000) Expression of mucin genes in chronic ethmoiditis. *Am.J.Rhinol.* 14 (3):163-170.

119. Kako K, Wakamatsu H, Hamada T, Banasik M, Ohata K, Niki-Kuroiwa T, Suzuki S, Takeuchi J, and Ishida N (1998) Examination of DNA-binding activity of neuronal transcription factors by electrophoretical mobility shift assay. *Brain Res.Brain Res.Protoc.* 2 (4):243-249.

120. Kamio K, Matsushita I, Hijikata M, Tanaka G, Nakata K, Ishida T, Tokunaga K, Kobashi Y, Taguchi Y, Homma S, Nakata K, Azuma A, Kudoh S, and Keicho N (2005) Promoter Analysis and Aberrant Expression of MUC5B gene in Diffuse Panbronchiolitis. *Am J Respir.Crit Care Med.*

121. Kamio K, Matsushita I, Tanaka G, Ohashi J, Hijikata M, Nakata K, Tokunaga K, Azuma A, Kudoh S, and Keicho N (2004) Direct determination of MUC5B promoter haplotypes based on the method of single-strand conformation polymorphism and their statistical estimation. *Genomics* 84 (3):613-622.

122. Kanno A, Satoh K, Kimura K, Hirota M, Umino J, Masamune A, Satoh A, Asakura T, Egawa S, Sunamura M, Endoh M, and Shimosegawa T (2006) The expression of MUC4 and MUC5AC is related to the biologic malignancy of intraductal papillary mucinous neoplasms of the pancreas. *Pancreas* 33 (4):391-396.

123. Kapitanovic S, Radosevic S, Kapitanovic M, Andelinovic S, Ferencic Z, Tavassoli M, Primorac D, Sonicki Z, Spaventi S, Pavelic K, and Spaventi R (1997) The expression of p185(HER-2/neu) correlates with the stage of disease and survival in colorectal cancer. *Gastroenterology* 112 (4):1103-1113.

124. Karin M, Liu Z, and Zandi E (1997) AP-1 function and regulation. *Curr.Opin.Cell Biol.* 9 (2):240-246.

125. Kaunitz JD (1999) Barrier function of gastric mucus. *Keio J.Med.* 48 (2):63-68.

126. Keates AC, Nunes DP, Afdhal NH, Troxler RF, and Offner GD (1997) Molecular cloning of a major human gall bladder mucin: complete C-terminal sequence and genomic organization of MUC5B. *Biochem.J.* 324 (Pt 1):295-303.

127. Khoury G and Gruss P (1983) Enhancer elements. *Cell* 33 (2):313-314.

128. Kim DH, Chu HS, Lee JY, Hwang SJ, Lee SH, and Lee HM (2004) Up-regulation of MUC5AC and MUC5B mucin genes in chronic rhinosinusitis. *Arch.Otolaryngol.Head Neck Surg.* 130 (6):747-752.

129. Kim R, Tanabe K, Uchida Y, Osaki A, and Toge T (2002a) The role of HER-2 oncoprotein in drug-sensitivity in breast cancer (review). *Oncol.Rep.* 9 (1):3-9.

130. Kim YD, Kwon EJ, Park DW, Song SY, Yoon SK, and Baek SH (2002b) Interleukin-1beta induces MUC2 and MUC5AC synthesis through cyclooxygenase-2 in NCI-H292 cells. *Mol.Pharmacol.* 62 (5):1112-1118.

131. Kirkbride HJ, Bolscher JG, Nazmi K, Vinall LE, Nash MW, Moss FM, Mitchell DM, and Swallow DM (2001) Genetic polymorphism of MUC7: allele frequencies and association with asthma. *Eur.J Hum Genet* 9 (5):347-354.

132. Kirkham S, Sheehan JK, Knight D, Richardson PS, and Thornton DJ (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.

133. Knight JC, Keating BJ, Rockett KA, and Kwiatkowski DP (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat.Genet.* 33 (4):469-475.

134. Kobayashi S, Kohda T, Miyoshi N, Kuroiwa Y, Aisaka K, Tsutsumi O, Kaneko-Ishino T, and Ishino F (1997) Human PEG1/MEST, an imprinted gene on chromosome 7. *Hum.Mol.Genet.* 6 (5):781-786.

135. Kok RG, Christoffels VM, Vosman B, and Hellingwerf KJ (1993) Growth-phase-dependent expression of the lipolytic system of Acinetobacter calcoaceticus BD413: cloning of a gene encoding one of the esterases. *J.Gen.Microbiol.* 139 (10):2329-2342.

136. Kudoh S and Keicho N (2003) Diffuse panbronchiolitis. *Semin.Respir.Crit Care Med.* 24 (5):607-618.

137. Kumar R and Yarmand-Bagheri R (2001) The role of HER2 in angiogenesis. *Semin.Oncol.* 28 (5 Suppl 16):27-32.

138. Kyo K, Parkes M, Takei Y, Nishimori H, Vyas P, Satsangi J, Simmons J, Nagawa H, Baba S, Jewell D, Muto T, Lathrop GM, and Nakamura Y (1999) Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin gene, MUC3. *Hum Mol.Genet* 8 (2):307-311.

139. Kyte J and Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J.Mol.Biol.* 157 (1):105-132.

140. Lagrange T, Kapanidis AN, Tang H, Reinberg D, and Ebright RH (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* 12 (1):34-44.

141. Lauffart B, Dimatteo A, Vaughan MM, Cincotta MA, Black JD, and Still IH (2006) Temporal and spatial expression of TACC1 in the mouse and human. *Dev.Dyn.* 235 (6):1638-1647.

142. Lee KT and Liu TS (2001a) Altered mucin gene expression in stone-containing intrahepatic bile ducts and cholangiocarcinomas. *Dig.Dis.Sci.* 46 (10):2166-2172.

143. Lee KT and Liu TS (2001b) Mucin gene expression in gallbladder epithelium with black pigment stone ascertained by in situ hybridization. *Kaohsiung.J.Med.Sci.* 17 (10):517-523.

144. Lee KT and Liu TS (2002) Mucin gene expression in gallbladder epithelium. *J.Formos.Med.Assoc.* 101 (11):762-768.

145. Leitzel K, Teramoto Y, Konrad K, Chinchilli VM, Volas G, Grossberg H, Harvey H, Demers L, and Lipton A (1995) Elevated serum c-erbB-2 antigen levels and decreased response to hormone therapy of breast cancer. *J.Clin.Oncol.* 13 (5):1129-1135.

146. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, and de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum.Mol.Genet.* 12 (14):1725-1735.

147. Leung TF, Tang NL, Chan IH, Li AM, Ha G, and Lam CW (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.

148. Li L, He S, Sun JM, and Davie JR (2004) Gene regulation by Sp1 and Sp3. *Biochem.Cell Biol.* 82 (4):460-471.

149. Lillehoj ER and Kim KC (2002) Airway mucus: its components and function. *Arch.Pharm.Res.* 25 (6):770-780.

150. Lim CY, Santoso B, Boulay T, Dong E, Ohler U, and Kadonaga JT (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* 18 (13):1606-1617.

151. Lin J, Tsuboi Y, Rimell F, Liu G, Toyama K, Kawano H, Paparella MM, and Ho SB (2003) Expression of mucins in mucoid otitis media. *J.Assoc.Res.Otolaryngol.* 4 (3):384-393.

152. Lin J, Tsuprun V, Kawano H, Paparella MM, Zhang Z, Anway R, and Ho SB (2001) Characterization of mucins in human middle ear and Eustachian tube. *Am.J.Physiol Lung Cell Mol.Physiol* 280 (6):L1157-L1167.

153. Liu B, Lague JR, Nunes DP, Toselli P, Oppenheim FG, Soares RV, Troxler RF, and Offner GD (2002) Expression of membrane-associated mucins MUC1 and MUC4 in major human salivary glands. *J Histochem.Cytochem.* 50 (6):811-820.

154. Liu J, Zhang Z, Xu Y, Xing L, and Zhang H (2004) Effects of glucocorticoid on IL-13-induced Muc5ac expression in airways of mice. *J.Huazhong.Univ Sci.Technolog.Med.Sci.* 24 (6):575-577.

155. Llinares K, Escande F, Aubert S, Buisine MP, de Bolos C, Batra SK, Gosselin B, Aubert JP, Porchet N, and Copin MC (2004) Diagnostic value of MUC4 immunostaining in distinguishing epithelial mesothelioma and lung adenocarcinoma. *Mod.Pathol.* 17 (2):150-157.

156. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, and Lee MP (2003) Allelic variation in gene expression is common in the human genome. *Genome Res* 13 (8):1855-1862.

157. Lodish et al (2000) Protein Structure and Function. Molecular Cell Biology, 4th ed.W.H. Freeman and Company.

158. Lopez-Ferrer A, Alameda F, Barranco C, Garrido M, and de Bolos C (2001a) MUC4 expression is increased in dysplastic cervical disorders. *Hum.Pathol.* 32 (11):1197-1202.

159. Lopez-Ferrer A, Curull V, Barranco C, Garrido M, Lloreta J, Real FX, and de Bolos C (2001b) Mucins as differentiation markers in bronchial epithelium. Squamous cell carcinoma and adenocarcinoma display similar expression patterns. *Am.J.Respir.Cell Mol.Biol.* 24 (1):22-29.

160. Lyon MF (1989) X-chromosome inactivation as a system of gene dosage compensation to regulate gene expression. *Prog.Nucleic Acid Res.Mol.Biol.* 36:119-130.

161. Lyon MF (1999) X-chromosome inactivation. *Curr.Biol.* 9 (7):R235-R237.

162. Mariette C, Perrais M, Leteurtre E, Jonckheere N, Hemon B, Pigny P, Batra S, Aubert JP, Triboulet JP, and Van S, I (2004) Transcriptional regulation of human mucin MUC4 by bile acids in oesophageal cancer cells is promoter-

dependent and involves activation of the phosphatidylinositol 3-kinase signalling pathway. *Biochem.J.* 377 (Pt 3):701-708.

163. Marom Z, Shelhamer JH, Sun F, and Kaliner M (1983) Human airway monohydroxyeicosatetraenoic acid generation and mucus release. *J.Clin.Invest* 72 (1):122-127.

164. Martinez-Anton A, Debolos C, Garrido M, Roca-Ferrer J, Barranco C, Alobid I, Xaubet A, Picado C, and Mullol J (2006) Mucin genes have different expression patterns in healthy and diseased upper airway mucosa. *Clin.Exp.Allergy* 36 (4):448-457.

165. Maston GA, Evans SK, and Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. *Annu.Rev.Genomics Hum.Genet.*

166. Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, Ehm M, Glanowski S, He C, Heil J, Markianos K, McMullen I, Pericak-Vance MA, Silbergleit A, Stein L, Wagner M, Wilson AF, Winick JD, Winn-Deen ES, Yamashiro CT, Cann HM, Lai E, and Holden AL (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am.J.Hum.Genet.* 73 (2):271-284.

167. Mayadas TN and Wagner DD (1992) Vicinal cysteines in the prosequence play a role in von Willebrand factor multimer assembly. *Proc.Natl.Acad.Sci.U.S.A* 89 (8):3531-3535.

168. Meden H and Kuhn W (1997) Overexpression of the oncogene c-erbB-2 (HER2/neu) in ovarian cancer: a new prognostic factor. *Eur.J Obstet.Gynecol.Reprod.Biol.* 71 (2):173-179.

169. Migeon BR (1994) X-chromosome inactivation: molecular mechanisms and genetic consequences. *Trends Genet.* 10 (7):230-235.

170. Mitchell PJ and Tjian R (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* 245 (4916):371-378.

171. Moniaux N, Escande F, Batra SK, Porchet N, Laine A, and Aubert JP (2000) Alternative splicing generates a family of putative secreted and membrane-associated MUC4 mucins. *Eur.J Biochem.* 267 (14):4536-4544.

172. Moniaux N, Escande F, Porchet N, Aubert JP, and Batra SK (2001) Structural organization and classification of the human mucin genes. *Front Biosci.* 6:D1192-D1206.

173. Moniaux N, Nollet S, Porchet N, Degand P, Laine A, and Aubert JP (1999) Complete sequence of the human mucin MUC4: a putative cell membrane-associated mucin. *Biochem.J* 338 (Pt 2):325-333.

174. Morcillo EJ and Cortijo J (2006) Mucus and MUC in asthma. *Curr.Opin.Pulm.Med.* 12 (1):1-6.

175. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, and Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430 (7001):743-747.

176. Moskvina V, Norton N, Williams N, Holmans P, Owen M, and O'donovan M (2005) Streamlined analysis of pooled genotype data in SNP-based association studies. *Genet Epidemiol.* 28 (3):273-282.

177. Mottagui-Tabar S, Faghihi MA, Mizuno Y, Engstrom PG, Lenhard B, Wasserman WW, and Wahlestedt C (2005) Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC.Genomics* 6 (1):18.

178. Nakamura Y and Hoshino M (2005) TH2 cytokines and associated transcription factors as therapeutic targets in asthma. *Curr.Drug Targets.Inflamm.Allergy* 4 (2):267-270.

179. Namavar F, Sparrius M, Veerman EC, Appelmelk BJ, and Vandenbroucke-Grauls CM (1998) Neutrophil-activating protein mediates adhesion of Helicobacter pylori to sulfated carbohydrates on high-molecular-weight salivary mucin. *Infect.Immun.* 66 (2):444-447.

180. Ng W (2007) Genetic Polymorphism in the Control of Alternative Splicing, Green College, University of Oxford.

181. Nguyen PL, Niehans GA, Cherwitz DL, Kim YS, and Ho SB (1996) Membrane-bound (MUC1) and secretory (MUC2, MUC3, and MUC4) mucin gene expression in human lung cancer. *Tumour.Biol.* 17 (3):176-192.

182. Nielsen PA, Bennett EP, Wandall HH, Therkildsen MH, Hannibal J, and Clausen H (1997) Identification of a major human high molecular weight salivary mucin (MG1) as tracheobronchial mucin MUC5B. *Glycobiology* 7 (3):413-419.

183. Nieminen MM, Kaprio J, and Koskenvuo M (1991) A population-based study of bronchial asthma in adult twin pairs. *Chest* 100 (1):70-75.

184. Nollet S, Moniaux N, Maury J, Petitprez D, Degand P, Laine A, Porchet N, and Aubert JP (1998) Human mucin gene MUC4: organization of its 5'-region and polymorphism of its central tandem repeat array. *Biochem.J.* 332 (Pt 3):739-748.

185. Offner GD, Nunes DP, Keates AC, Afdhal NH, and Troxler RF (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.

186. Ogata S, Uehara H, Chen A, and Itzkowitz SH (1992) Mucin gene expression in colonic tissues and cell lines. *Cancer Res.* 52 (21):5971-5978.

187. Ogbourne S and Antalis TM (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem.J.* 331 (Pt 1):1-14.

188. Park HU, Kim JW, Kim GE, Bae HI, Crawley SC, Yang SC, Gum JR, Jr., Batra SK, Rousseau K, Swallow DM, Sleisenger MH, and Kim YS (2003) Aberrant expression of MUC3 and MUC4 membrane-associated mucins and sialyl Le(x) antigen in pancreatic intraepithelial neoplasia. *Pancreas* 26 (3):e48-e54.

189. Perez-Vilar J, Eckhardt AE, DeLuca A, and Hill RL (1998) Porcine submaxillary mucin forms disulfide-linked multimers through its amino-terminal D-domains. *J.Biol.Chem.* 273 (23):14442-14449.

190. Perez-Vilar J and Hill RL (1999) The structure and assembly of secreted mucins. *J.Biol.Chem.* 274 (45):31751-31754.

191. Perrais M, Pigny P, Buisine MP, Porchet N, Aubert JP, and Seuningen-Lempire I (2001a) 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.

192. Perrais M, Pigny P, Ducourouble MP, Petitprez D, Porchet N, Aubert JP, and Van S, I (2001b) Characterization of human mucin gene MUC4 promoter: importance of growth factors and proinflammatory cytokines for its regulation in pancreatic cancer cells. *J.Biol.Chem.* 276 (33):30923-30933.

193. Perutelli P, Biglino P, and Mori PG (1997) von Willebrand factor: biological function and molecular defects. *Pediatr.Hematol.Oncol.* 14 (6):499-512.

194. Pflugfelder SC, Solomon A, and Stern ME (2000) The diagnosis and management of dry eye: a twenty-five-year review. *Cornea* 19 (5):644-649.

195. Pigny P, Guyonnet-Duperat V, Hill AS, Pratt WS, Galiegue-Zouitina S, d'Hooge MC, Laine A, Van Seuningen I, Degand P, Gum JR, Kim YS, Swallow DM, Aubert JP, and Porchet N (1996) Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. *Genomics* 38 (3):340-352.

196. Porchet N, Nguyen VC, Dufosse J, Audie JP, Guyonnet-Duperat V, Gross MS, Denis C, Degand P, Bernheim A, and Aubert JP (1991) Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem.Biophys.Res.Commun.* 175 (2):414-422.

197. Porter K, Komiyama NH, Vitalis T, Kind PC, and Grant SG (2005) Differential expression of two NMDA receptor interacting proteins, PSD-95 and SynGAP during mouse development. *Eur.J.Neurosci.* 21 (2):351-362.

198. Power C (1992) A review of child health in the 1958 birth cohort: National Child Development Study. *Paediatr.Perinat.Epidemiol.* 6 (1):81-110.

199. Pratt WS, Islam I, and Swallow DM (1996) Two additional polymorphisms within the hypervariable MUC1 gene: association of alleles either side of the VNTR region. *Ann.Hum.Genet.* 60 (Pt 1):21-28.

200. Pullan RD, Thomas GA, Rhodes M, Newcombe RG, Williams GT, Allen A, and Rhodes J (1994) Thickness of adherent mucus gel on colonic mucosa in humans and its relevance to colitis. *Gut* 35 (3):353-359.

201. Reid CJ, Gould S, and Harris A (1997) Developmental expression of mucin genes in the human respiratory tract. *Am.J.Respir.Cell Mol.Biol.* 17 (5):592-598.

202. Reid CJ and Harris A (1998) Developmental expression of mucin genes in the human gastrointestinal system. *Gut* 42 (2):220-226.

203. Reik W, Constancia M, Dean W, Davies K, Bowden L, Murrell A, Feil R, Walter J, and Kelsey G (2000) Igf2 imprinting in development and disease. *Int.J.Dev.Biol.* 44 (1):145-150.

204. Ren H and Stiles GL (1994) Posttranscriptional mRNA processing as a mechanism for regulation of human A1 adenosine receptor expression. *Proc.Natl.Acad.Sci.U.S.A* 91 (11):4864-4866.

205. Riesewijk AM, Hu L, Schulz U, Tariverdian G, Hoglund P, Kere J, Ropers HH, and Kalscheuer VM (1997) Monoallelic expression of human PEG1/MEST is paralleled by parent-specific methylation in fetuses. *Genomics* 42 (2):236-244.

206. Robertson KD (2002) DNA methylation and chromatin - unraveling the tangled web. *Oncogene* 21 (35):5361-5379.

207. Robinson DS, Hamid Q, Ying S, Tsicopoulos A, Barkans J, Bentley AM, Corrigan C, Durham SR, and Kay AB (1992) Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N.Engl.J.Med.* 326 (5):298-304.

208. Rockman MV and Wray GA (2002) Abundant raw material for cis-regulatory evolution in humans. *Mol.Biol.Evol.* 19 (11):1991-2004.

209. Rodriguez-Trelles F, Tarrio R, and Ayala FJ (2003) Evolution of cis-regulatory regions versus codifying regions. *Int.J.Dev.Biol.* 47 (7-8):665-673.

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

211. Rose MC (1992) Mucins: structure, function, and role in pulmonary diseases. *Am.J.Physiol* 263 (4 Pt 1):L413-L429.

212. Rossi EA, McNeer RR, Price-Schiavi SA, Van den Brande JM, Komatsu M, Thompson JF, Carraway CA, Fregien NL, and Carraway KL (1996) Sialomucin complex, a heterodimeric glycoprotein complex. Expression as a soluble, secretable form in lactating mammary gland and colon. *J Biol.Chem.* 271 (52):33476-33485.

213. Rousseau K, Vinall LE, Butterworth SL, Hardy RJ, Holloway J, Wadsworth ME, and Swallow DM (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.

214. Sadler JE (1998) Biochemistry and genetics of von Willebrand factor. *Annu.Rev.Biochem.* 67:395-424.

215. Saito-Hisaminato A, Katagiri T, Kakiuchi S, Nakamura T, Tsunoda T, and Nakamura Y (2002) Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res.* 9 (2):35-45.

216. Saitou M, Goto M, Horinouchi M, Tamada S, Nagata K, Hamada T, Osako M, Takao S, Batra SK, Aikou T, Imai K, and Yonezawa S (2005) MUC4 expression is a novel prognostic factor in patients with invasive ductal carcinoma of the pancreas. *J.Clin.Pathol.* 58 (8):845-852.

217. Scheet P and Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am.J.Hum.Genet.* 78 (4):629-644.

218. Seregni E, Botti C, Lombardo C, Cantoni A, Bogni A, Cataldo I, and Bombardieri E (1996) Pattern of mucin gene expression in normal and neoplastic lung tissues. *Anticancer Res.* 16 (4B):2209-2213.

219. Serfling E, Jasin M, and Schaffner W (1985) Enhancers and Eukaryotic Gene-Transcription. *Trends in Genetics* 1 (8):224-230.

220. Sharma RP, Grayson DR, Guidotti A, and Costa E (2005) Chromatin, DNA methylation and neuron gene regulation--the purpose of the package. *J.Psychiatry Neurosci.* 30 (4):257-263.

221. Sheehan JK, Howard M, Richardson PS, Longwill T, and Thornton DJ (1999b) 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.

222. Sheehan JK, Howard M, Richardson PS, Longwill T, and Thornton DJ (1999a) 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.

223. Sheehan JK, Richardson PS, Fung DC, Howard M, and Thornton DJ (1995b) Analysis of respiratory mucus glycoproteins in asthma: a detailed study from a patient who died in status asthmaticus. *Am.J.Respir.Cell Mol.Biol.* 13 (6):748-756.

224. Sheehan JK, Richardson PS, Fung DC, Howard M, and Thornton DJ (1995a) Analysis of respiratory mucus glycoproteins in asthma: a detailed study from a patient who died in status asthmaticus. *Am.J.Respir.Cell Mol.Biol.* 13 (6):748-756.

225. Shim J and Karin M (2002) The control of mRNA stability in response to extracellular stimuli. *Mol.Cells* 14 (3):323-331.

226. Sibbald B, Horn ME, and Gregg I (1980) A family study of the genetic basis of asthma and wheezy bronchitis. *Arch.Dis.Child* 55 (5):354-357.

227. Sibbald B and Turner-Warwick M (1979) Factors influencing the prevalence of asthma among first degree relatives of extrinsic and intrinsic asthmatics. *Thorax* 34 (3):332-337.

228. Sidebotham HJ and Roche WR (2003) Asthma deaths; persistent and preventable mortality. *Histopathology* 43 (2):105-117.

229. Silva F, Carvalho F, Peixoto A, Seixas M, Almeida R, Carneiro F, Mesquita P, Figueiredo C, Nogueira C, Swallow DM, Amorim A, and David L (2001) MUC1 gene polymorphism in the gastric carcinogenesis pathway. *Eur.J Hum Genet* 9 (7):548-552.

230. Singh AP, Chauhan SC, Bafna S, Johansson SL, Smith LM, Moniaux N, Lin MF, and Batra SK (2006) Aberrant expression of transmembrane mucins, MUC1 and MUC4, in human prostate carcinomas. *Prostate* 66 (4):421-429.

231. Singh AP, Moniaux N, Chauhan SC, Meza JL, and Batra SK (2004) Inhibition of MUC4 expression suppresses pancreatic tumor cell growth and metastasis. *Cancer Res.* 64 (2):622-630.

232. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, and McGuire WL (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235 (4785):177-182.

233. Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, Keith DE, Levin WJ, Stuart SG, Udove J, Ullrich A, and . (1989) Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* 244 (4905):707-712.

234. Sleigh MA (1983) Ciliary function in transport of mucus. *Eur.J.Respir.Dis.Suppl* 128 (Pt 1):287-292.

235. Slomiany BL and Slomiany A (1991) Role of mucus in gastric mucosal protection. *J.Physiol Pharmacol.* 42 (2):147-161.

236. Smale ST and Baltimore D (1989) The "initiator" as a transcription control element. *Cell* 57 (1):103-113.

237. Smale ST and Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu.Rev.Biochem.* 72:449-479.

238. Smirnova MG, Guo L, Birchall JP, and Pearson JP (2003) LPS up-regulates mucin and cytokine mRNA expression and stimulates mucin and cytokine secretion in goblet cells. *Cell Immunol.* 221 (1):42-49.

239. Sonora C, Mazal D, Berois N, Buisine MP, Ubillos L, Varangot M, Barrios E, Carzoglio J, Aubert JP, and Osinaga E (2006) Immunohistochemical analysis of MUC5B apomucin expression in breast cancer and non-malignant breast tissues. *J.Histochem.Cytochem.* 54 (3):289-299.

240. Sproul D, Gilbert N, and Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. *Nat.Rev.Genet.* 6 (10):775-781.

241. Stephens M, Smith NJ, and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am.J.Hum.Genet.* 68 (4):978-989.

242. Strachan T, Read A (2000) Control of Gene Expression. Human Molecular Genetics, 2 ed.BIOS Scientific Publishers.

243. Strous GJ and Dekker J (1992) Mucin-type glycoproteins. *Crit Rev.Biochem.Mol.Biol.* 27 (1-2):57-92.

244. Swallow DM, Gendler S, Griffiths B, Corney G, Taylor-Papadimitriou J, and Bramwell ME (1987) The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature* 328 (6125):82-84.

245. Tamada S, Shibahara H, Higashi M, Goto M, Batra SK, Imai K, and Yonezawa S (2006) MUC4 is a novel prognostic factor of extrahepatic bile duct carcinoma. *Clin.Cancer Res.* 12 (14 Pt 1):4257-4264.

246. Tan M, Yao J, and Yu D (1997) Overexpression of the c-erbB-2 gene enhanced intrinsic metastasis potential in human breast cancer cells without increasing their transformation abilities. *Cancer Res.* 57 (6):1199-1205.

247. Tang H, Sun X, Reinberg D, and Ebright RH (1996) Protein-protein interactions in eukaryotic transcription initiation: structure of the preinitiation complex. *Proc.Natl.Acad.Sci.U.S.A* 93 (3):1119-1124.

248. Thomas MC and Chiang CM (2006) The general transcription machinery and general cofactors. *Crit Rev.Biochem.Mol.Biol.* 41 (3):105-178.

249. Troxler RF, Iontcheva I, Oppenheim FG, Nunes DP, and Offner GD (1997) Molecular characterization of a major high molecular weight mucin from human sublingual gland. *Glycobiology* 7 (7):965-973.

250. Tsutsumida H, Goto M, Kitajima S, Kubota I, Hirotsu Y, Wakimoto J, Batra SK, Imai K, and Yonezawa S (2007) MUC4 expression correlates with poor prognosis in small-sized lung adenocarcinoma. *Lung Cancer* 55 (2):195-203.

251. Turnberg LA and Ross IN (1984) Studies of the pH gradient across gastric mucus. *Scand.J.Gastroenterol.Suppl* 92:48-50.

252. Urrutia I, Capelastegui A, Quintana JM, Muniozguren N, Basagana X, and Sunyer J (2005) Smoking habit, respiratory symptoms and lung function in young adults. *Eur.J.Public Health* 15 (2):160-165.

253. Van den SP, Rudd PM, Dwek RA, and Opdenakker G (1998) Concepts and principles of O-linked glycosylation. *Crit Rev.Biochem.Mol.Biol.* 33 (3):151-208.

254. van Klinken BJ, Dekker J, van Gool SA, van Marle J, Buller HA, and Einerhand AW (1998) MUC5B is the prominent mucin in human gallbladder and is also expressed in a subset of colonic goblet cells. *Am.J.Physiol* 274 (5 Pt 1):G871-G878.

255. Van Klinken BJ, Einerhand AW, Buller HA, and Dekker J (1998) Strategic biochemical analysis of mucins. *Anal.Biochem.* 265 (1):103-116.

256. Van S, I, Perrais M, Pigny P, Porchet N, and Aubert JP (2000) Sequence of the 5'-flanking region and promoter activity of the human mucin gene MUC5B in different phenotypes of colon cancer cells. *Biochem.J.* 348 Pt 3:675-686.

257. Vandenhaute B, Buisine MP, Debailleul V, Clement B, Moniaux N, Dieu MC, Degand P, Porchet N, and Aubert JP (1997) Mucin gene expression in biliary epithelial cells. *J Hepatol.* 27 (6):1057-1066.

258. Veerman EC, Bank CM, Namavar F, Appelmelk BJ, Bolscher JG, and Nieuw Amerongen AV (1997) Sulfated glycans on oral mucin as receptors for Helicobacter pylori. *Glycobiology* 7 (6):737-743.

259. Veerman EC, Ligtenberg AJ, Schenkels LC, Walgreen-Weterings E, and Nieuw Amerongen AV (1995) Binding of human high-molecular-weight salivary mucins (MG1) to Hemophilus parainfluenzae. *J.Dent.Res.* 74 (1):351-357.

260. Veerman EC, van den Keijbus PA, Nazmi K, Vos W, van der Wal JE, Bloemena E, Bolscher JG, and Amerongen AV (2003) Distinct localization of MUC5B glycoforms in the human salivary glands. *Glycobiology* 13 (5):363-366.

261. Verma M, Blass C, and Davidson EA (1997) Upregulation of the tracheobronchial mucin gene involves cyclic AMP response elements. *Indian J.Biochem.Biophys.* 34 (1-2):118-123.

262. Verweij CL, 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.

263. Vestbo J, Prescott E, and Lange P (1996) Association of chronic mucus hypersecretion with FEV1 decline and chronic obstructive pulmonary disease morbidity. Copenhagen City Heart Study Group. *Am.J.Respir.Crit Care Med.* 153 (5):1530-1535.

264. Villard J (2004) Transcription regulation and human diseases. *Swiss.Med.Wkly.* 134 (39-40):571-579.

265. Vinall LE, Fowler JC, Jones AL, Kirkbride HJ, de Bolos C, Laine A, Porchet N, Gum JR, Kim YS, Moss FM, Mitchell DM, and Swallow DM (2000) Polymorphism of human mucin genes in chest disease: possible significance of MUC2. *Am J Respir.Cell Mol.Biol.* 23 (5):678-686.

266. Vinall LE, Hill AS, Pigny P, Pratt WS, Toribara N, Gum JR, Kim YS, Porchet N, Aubert JP, and Swallow DM (1998) Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Hum.Genet.* 102 (3):357-366.

267. Vinall LE, Pratt WS, and Swallow DM (2000) Detection of mucin gene polymorphism. *Methods Mol.Biol.* 125:337-350.

268. Viswanathan H, Brownlee IA, Pearson JP, and Carrie S (2006) MUC5B secretion is up-regulated in sinusitis compared with controls. *Am.J.Rhinol.* 20 (5):554-557.

269. Vladich FD, Brazille SM, Stern D, Peck ML, 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.

270. von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J.Mol.Biol.* 225 (2):487-494.

271. Voorberg J, Fontijn R, van Mourik JA, and Pannekoek H (1990) Domains involved in multimer assembly of von willebrand factor (vWF): multimerization is independent of dimerization. *EMBO J.* 9 (3):797-803.

272. Wadsworth MEJ, Butterworth SL, Hardy RJ, Kuh DJ, Richards M, Langenberg C, Hilder WS, and Connor M (2003) The life course prospective design: an example of benefits and problems associated with study longevity. *Social Science & Medicine* 57 (11):2193-2205.

273. Wang Y, Harvey C, Rousset M, and Swallow DM (1994) Expression of human intestinal mRNA transcripts during development: analysis by a semiquantitative RNA polymerase chain reaction method. *Pediatr.Res.* 36 (4):514-521.

274. Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M, Rossi M, Auricchio S, and Swallow DM (1995) The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum.Mol.Genet.* 4 (4):657-662.

275. Watanabe H (2002) Significance of mucin on the ocular surface. *Cornea* 21 (2 Suppl 1):S17-S22.

276. Watson JD et al (1965) The functioning of higher eucaryotic genes. Molecular Biology of the Gene, 1 ed.Benjamin Cummings.

277. Weed DT, Gomez-Fernandez C, Pacheco J, Ruiz J, Hamilton-Nelson K, Arnold DJ, Civantos FJ, Zhang J, Yasin M, Goodwin WJ, and Carraway KL (2004) MUC4 and ERBB2 expression in major and minor salivary gland mucoepidermoid carcinoma. *Head Neck* 26 (4):353-364.

278. Weis L and Reinberg D (1992) Transcription by RNA polymerase II: initiator-directed formation of transcription-competent complexes. *FASEB J.* 6 (14):3300-3309.

279. Wickstrom C and Carlstedt I (2001) N-terminal cleavage of the salivary MUC5B mucin. Analogy with the Van Willebrand propolypeptide? *J.Biol.Chem.* 276 (50):47116-47121.

280. Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, and Loughlin J (2007) Extreme context-specificity in differential allelic expression. *Hum.Mol.Genet.*

281. Winterford CM, Walsh MD, Leggett BA, and Jass JR (1999) Ultrastructural localization of epithelial mucin core proteins in colorectal tissues. *J.Histochem.Cytochem.* 47 (8):1063-1074.

282. Yan H, Yuan W, Velculescu VE, Vogelstein B, and Kinzler KW (2002) Allelic variation in human gene expression. *Science* 297 (5584):1143.

283. Yasuo M, Fujimoto K, Tanabe T, Yaegashi H, Tsushima K, Takasuna K, Koike T, Yamaya M, and Nikaido T (2006) Relationship between calcium-activated chloride channel 1 and MUC5AC in goblet cell hyperplasia induced by interleukin-13 in human bronchial epithelial cells. *Respiration* 73 (3):347-359.

284. Ye S, Humphries S, and Green F (1992) Allele specific amplification by tetra-primer PCR. *Nucleic Acids Res.* 20 (5):1152.

285. Ying S, Zhang G, Gu S, and Zhao J (2006a) How much do we know about atopic asthma: where are we now? *Cell Mol.Immunol.* 3 (5):321-332.

286. Ying S, Zhang G, Gu S, and Zhao J (2006b) How much do we know about atopic asthma: where are we now? *Cell Mol.Immunol.* 3 (5):321-332.

287. Yu CJ, Shun CT, Yang PC, Lee YC, Shew JY, Kuo SH, and Luh KT (1997) Sialomucin expression is associated with erbB-2 oncoprotein overexpression, early recurrence, and cancer death in non-small-cell lung cancer. *Am J Respir.Crit Care Med.* 155 (4):1419-1427.

288. Yu CJ, Yang PC, Shun CT, Lee YC, Kuo SH, and Luh KT (1996) Overexpression of MUC5 genes is associated with early post-operative metastasis in non-small-cell lung cancer. *Int.J.Cancer* 69 (6):457-465.

289. Yu D, Jing T, Liu B, Yao J, Tan M, McDonnell TJ, and Hung MC (1998) Overexpression of ErbB2 blocks Taxol-induced apoptosis by upregulation of p21Cip1, which inhibits p34Cdc2 kinase. *Mol.Cell* 2 (5):581-591.

290. Yu X and Kensler T (2005) Nrf2 as a target for cancer chemoprevention. *Mutat.Res.* 591 (1-2):93-102.

291. Yuan-Chen WD, Wu R, Reddy SP, Lee YC, and Chang MM (2007) Distinctive epidermal growth factor receptor/extracellular regulated kinase-independent and -dependent signaling pathways in the induction of airway

mucin 5B and mucin 5AC expression by phorbol 12-myristate 13-acetate. *Am.J.Pathol.* 170 (1):20-32.

292. Zawel L and Reinberg D (1992) Advances in RNA polymerase II transcription. *Curr.Opin.Cell Biol.* 4 (3):488-495.

293. Zhao Z, Fu YX, Hewett-Emmett D, and Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* 312:207-213.

294. Zhou GH, Gotou M, Kajiyama T, and Kambara H (2005) Multiplex SNP typing by bioluminometric assay coupled with terminator incorporation (BATI). *Nucl.Acids Res.* 33 (15):e133.