MUC1 polymorphism in relation to susceptibility to *Helicobacter pylori* gastritis

by

Ana Sofia Santos Teixeira

Submitted for the degree of Doctor of Philosophy

University College London

September 2005

The Galton Laboratory
Wolfson House
4 Stephenson Way
London NW1 2HE

Esta tese é dedicada aos meus pais, Maria Teresa e Ramiro,
por todo o carinho e apoio incondicional que me deram durante estes quatro anos,
mesmo estando a milhares de kilómetros de distância.
Muito, muito obrigada... por tudo.

Abstract

The gene *MUC1* encodes a transmembrane mucin glycoprotein that is expressed on the apical surface of most epithelia and is aberrantly expressed in cancer. MUC1 contains an extended domain of tandemly repeated (TR) amino-acid sequence, which acts as the backbone for a large amount of O-linked glycosylation, and which varies in length and sequence in different alleles. Previous studies on *MUC1* tandem repeat variation in patients with gastritis (Vinall *et al*, 2002) and gastric cancer (Carvalho *et al*, 1997) showed an overrepresentation of short TR alleles in the patient groups when compared with normal controls. The major aim of this thesis is to pursue this observation further.

MUC1 allele and three locus haplotype frequencies were compared in 3 populations of different ancestry, from UK, Nigeria and Portugal, which show dramatic differences in gastric disease incidence; in patients with gastric disease, patients with other gastrointestinal disease; as well as associated controls. There were differences between the Nigerians and unselected European control groups, but there was no significant difference between the groups collected in London and Porto. Analysis of the gastric disease groups showed an over-representation of a particular *MUC1* haplotype. A search was made, by sequencing and using a bioinformatics approach, for additional polymorphic markers within and surrounding the *MUC1* gene, that might act as convenient markers for future disease association studies. Patterns of Linkage Disequilibrium were established across a 600Kb genomic region containing *MUC1* using information in the HapMap resource and this information was used to assist in the selection of the single nucleotide polymorphisms (SNPs) within a 70Kb region to test on disease groups.

During the course of this thesis work, a second UK cohort of patients and controls was collected and characterised and a replication study attempted. DNA samples from a total of 154 Northern Europeans classified as *H. pylori* gastritis (n=33), former *H. pylori* gastritis (n=44), No *H. pylori* gastritis (n=18) and normal (n=59). Examination of the *MUC1* polymorphisms failed to show over-representation, in the *H.* pylori gastritis group, of the same haplotype found in the first gastritis cohort. The extended 70Kb haplotypes showed the expected association in the first cohort but no significant differences in the second cohort. However, in the population overall, it was noteworthy that there is a very high frequency haplogroup containing long tandem repeat arrays and this was somewhat lower in frequency in both *H. pylori* gastritis groups.

Acknowledgments

Firstly, I would like to thank the GABBA PhD program (Graduate Program in Areas of Basic and Applied Biology), and particularly the two coordinators of the 2000 class, Professor Fátima Carneiro and Professor José Manuel Castro Lopes, as well as Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Tecnology) for sponsoring this PhD studentship.

Also, I would like to express my deep gratitude to my supervisor, Professor Dallas Swallow, for all the enormous support and enthusiasm, even when I had lost faith in my experiments. Thank you very much for your kindness, for being always available and also for the excellent research I had the opportunity to do in the group.

I would like to thank Professor Sue Povey for employing me in the last year of my PhD, providing a very enjoyable and relevant way of earning some money while writing my thesis.

My thanks go also to the people in lab in Porto, Professor Leonor David and Dr Luis Filipe Silva, for always being available and for providing the samples from the Portuguese patients. I would also like to acknowledge everyone in the lab in Porto for always being so nice in my short visits.

There are also several people I would like to thank for their expertise and work done for my thesis: Jo Fowler, for helping in my first steps with the very specialised techniques applied to *MUC1*; Dr Adil Elamin for the enormous effort in sample collection; Dr Steve Pereira for supervising the clinical aspects; Professor Marco Novelli for supervising the histological evaluation and personally looking at all the slides with us; members of the Histopathology Department for processing the sections; MariWyn Burley for help with the sequencing and Mrs Ranji Arasaretnam for all her technical help, including the DNA extraction from many of the blood samples used in this thesis.

Also very important were the people I had the privilege to meet in these four years, and that will always be part of my life. So, firstly I would like to thank to the members of Wolfson House, past and present (in alphabetical order) Edward Almond, Catherine Bridge, Luis and Lupe Carvajal, Lewis Dartnell, Ian Evans, Jan Engelstaedter, Kate Ingram, Abi Jones, Andrew Loh, Charlotte Mulcare (a dear friend that shared the office with me during the writing-up!) Wendy Ng, Krishna Veeramah (thank you very much for helping me with the colour printing!), Danny White, and finally Lorenzo Zanette, to whom I would like to thank particularly for putting up with my bad mood so many times and try to cheer me up when I needed!

Also, I would like to thank my past and present flatmates, Richard Benninger, Arabella Bramley, Ruth Harding, Lorenzo Milazzo and Horácio Montenegro. It was a pleasure to share the flat with you, and thank you for all the encouragement to resist!

Very important too are the ones that were far and still gave a lot of support. I would like to thank my family, specially my parents Maria Teresa and Ramiro, and my brother Pedro for the constant support and encouragement to persevere and finish the thesis (and go back to Portugal...). Also to all my friends in Portugal, and specially (again in alphabetical order!) Ana Basto, Helena Carvalho, Sofia Magalhães and Marta Vilela for being as good friends as ever and never forget!

Finally, a special big thanks to Karine Rousseau and Lynne Vinall for during these four years being not only good colleagues and teaching me so much, but specially for being such good friends! Thank you very very much!

Table of contents

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

List of Tables

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 6:

Abbreviations

AMPS- ammonium persulphate

bp- base pair(s)

BSA- Bovine Serum Albumin

CagA- cytotoxin-associated gene A

cDNA- complementary DNA

CT- Cytoplasmic tail

DMSO- Dimethyl Sulfoxide

DNA- Deoxyribonucleic Acid

EC- Extracellular

EDTA- Ethylenediaminetetraacetic acid

GI- gastrointestinal

HCl- Hydrochloric Acid

HWE- Hardy-Weinberg Equilibrium

IARC- International Agency for Research on Cancer

IM- Intestinal Metaplasia

Kb- thousand base pairs

LD- Linkage Disequilibrium

Lewis-b blood group antigen- Le^b

LREC- Local Research Ethics Committee

M_r- protein molecular weight

MRC- Medical Research Council

MREC- Multicenter Research Ethics Committee

min- minute(s)

mRNA- messenger RNA

MVR- Minisatellite Variant Repeat

NCBI- National Center for Biotechnology Information

NIH- National Institutes of Health (USA)

nt- nucleotide(s)

NeuAc- Neuraminic acid

ON- overnight

PAI- Pathogenicity Island

RM- Recurrent miscarriage

RNA- Ribonucleic Acid

rpm- rotations per minute

S.D.- standard deviation

SDS- Sodium Dodecyl Sulphate

SN- supernatant

SNP(s)- Single Nucleotide Polymorphism(s)

TEMED- Tetramethylethylene diamine

TM- Transmembrane

TR-Tandem repeat

TSS(s)- Transcription Start Site(s)

UTR- Untranslated Region

UV- Ultra Violet

VacA- vacuolating cytotoxin

CHAPTER 1

INTRODUCTION

This thesis is concerned with a study of the variation of membrane associated mucin MUC1 in relation to gastric disease. This introduction gives a brief background to mucins, a more detailed account of MUC1, as well as describing gastric disease, *Helicobacter pylori*, gastric disease risk.

1.1 Mucins

Mucins are heavily glycosylated glycoproteins in which carbohydrate accounts for more than 50% of the molecular mass (Shimizu and Yamauchi, 1982). Glycosylation is O-linked through the hydroxyls on serines and threonines (Gendler, 2001;Shimizu and Yamauchi, 1982). They are highly heterogeneous molecules which are major components of mucus and the mucous layer on the surface of epithelia. They are encoded by multiple genes which have been given the symbol *MUC*, though they are not necessarily closely related. The major secreted gel forming mucins are encoded by a gene family located on chromosome 11, and there is a second family of genes that encodes mucins with a membrane anchor, which is located on chromosome 7 (*MUC3A*, *MUC3B*, *MUC11/12* and *MUC20*). However most of the other genes (including *MUC1*) do not show close sequence similarity to any other gene in the genome. Mucins do however have some features in common. They contain large domains rich in serine and threonine which form the backbone for the O-glycosylation and these domains usually contain repetitive sequence. Variation in the length and sequence of these repeat arrays has come to be the hallmark of mammalian mucins (Fowler et al., 2001;Gross et al., 1992;Nguyen et al., 1990;Pigny et al., 1995;Swallow et al., 1987a;Vinall et al., 1998).

Mucins play a major role in protection of epithelial surfaces. The secreted mucins are vital to the clearance of particulate matter from the airways; they act as lubricants; they protect against extreme acidity in the stomach, and they interact with bacteria. Many bacterial/carbohydrate interactions have been described, and the bacteria which colonise the mucous layer use mucus as food (Bry et al., 1996;Roussel et al., 1988). They also play many roles at the epithelial surface, as will be evident from the more detailed description of MUC1 below. The membrane mucins for example play a role in signalling and the cytoplasmic domains interact with other proteins.

1.2 MUC1

1.2.1 MUC1 from a historical perspective

MUC1 was the first gene to be described that encodes one of this class of glycoproteins involved in the protection and lubrication of human mucosal surfaces. For that reason, a very large amount of literature is available in relation to this particular gene and protein, yet there are a lot of unanswered questions.

In 1982, Cartron *et al* (Cartron et al., 1982), in an attempt to describe the human urine components that carry particular blood group antigens, described a glycoprotein of apparent M_r 340,000 that exhibited the red cell Sd^a antigen. It had a typical composition of a mucin, with low contents of cysteine and methionine and high content of serine, threonine, alanine and proline. In addition, the carbohydrate component did not show detectable amounts of mannose or fucose, ruling out the possibility of this being the Tamm-Horsfall (T-H) glycoprotein, a known glycoprotein that was a candidate at the beginning of the study.
The *MUC1* polymorphism was described for the first time in 1983 as a family of glycoconjugates present not only in urine, but also in kidney and lung (Karlsson et al., 1983). These glycoconjugates were called PUM, 'peanut-reactive urinary mucin', given that they were first found in human urine samples and were able to bind peanut agglutinin, which was being used as a way of detecting glycoconjugates. The PUM molecules had different mobilities in polyacrylamide gels in urine samples from different individuals. Four codominant alleles were found, and the authors predicted that this was a genetically determined polymorphism encoded by an autosomal gene locus.
In parallel with this, several different authors were characterising monoclonal antibodies that, in the context of their studies on cancer, might be useful diagnostically. These antibodies were raised against several tissues and body fluids, including breast cancer cell lines, human milk fat globule membrane (HMFG) and cervical carcinoma cell lines. It was quickly realised that most of these antibodies could recognise a milk glycoprotein or a urinary protein (Ashall et al.,

1982;Bramwell et al., 1983;Bramwell et al., 1985;Burchell et al., 1983;Hilkens et al., 1984;Kufe et al., 1984;Price et al., 1986;Taylor-Papadimitriou et al., 1981).

The recognition that this protein was encoded by the same gene as PUM came from Swallow and collaborators, when testing several of the monoclonal antibodies available (Swallow et al., 1986). In addition to the genetically determined variation, the authors found at least 3 distinct components of the different allelic forms of the glycoprotein in the urine, which were preferentially detected with different agglutinins. This was a reflection of the sialic acid content of the different constituents, which altered the affinity to the lectins. Also, one antibody was shown to bind preferentially to more sialylated products (Ashall et al., 1982), whereas the other antibodies bound both sialylated and desialylated forms, yet showing differences in the binding affinities (Bramwell et al., 1983;Bramwell et al., 1985;Burchell et al., 1983;Taylor-Papadimitriou et al., 1981).

The PUM glycoprotein was also found to be present in other body fluids, such as milk, as well as in several epithelia and in particular tumours of epithelial origin (Burchell et al., 1983;Girling et al., 1989;Griffiths et al., 1988b;Griffiths et al., 1988a;Karlsson et al., 1983).

The discovery that all the antigens recognised by the various antibodies were in fact all variants of the same protein, and that at least some of the antibodies recognised a protein epitope, allowed Gendler *et al* (Gendler et al., 1987;Gendler et al., 1990) to use an antibody to obtain cDNAs from an expression library and consequently to obtain information on the DNA sequence and deduced amino acid composition.

Using the same cDNA clone and the DNA sequence information, the PUM variability was established to be a genomic polymorphism (Swallow et al., 1987a;Gendler et al., 1987) which could be detected using restriction fragment length polymorphism (RFLP) assay. Using the restriction endonuclease EcoRI, the authors found high variability between the DNA samples studied, with alleles ranging from 7 to 12 Kb inherited in the family in a Mendelian fashion. This genomic polymorphism also showed correlation with the mobility (apparent size) differences observed previously in the protein (Karlsson et al., 1983). Swallow *et al* also used the *HinfI* restriction endonuclease that cuts smaller fragments, from 3 to 7 kb, displaying an even larger number of alleles, but with the same relative differences in size. This led them to suggest that the *PUM* locus had a tandem repeat area that showed a different size in different individuals due to variation in the number of repeats. In the same year, this gene locus was assigned to the long arm of chromosome 1, more precisely in the region 1q21 (Middleton-Price et al., 1988;Swallow et al., 1987b).

1.2.2 Constitution of MUC1: gene and protein

Several independent laboratories attempted to clone this gene, with several authors reporting partial or entire cDNA sequences (Gendler et al., 1987;Gendler et al., 1990;Hareuveni et al., 1990;Lan et al., 1990;Siddiqui et al., 1988;Wreschner et al., 1990). The comprehensive *MUC1* gene structure and most of its gene sequence was shown for the first time by Lancaster and colleagues (Lancaster et al., 1990). *MUC1* is composed of 7 exons, spanning from 5 and 11 kb, depending on the tandem repeat size (between 20 and 125 repeats).

1.2.2.1 *MUC1* promoter and regulation of transcription

All the functional elements of the promoter of *MUC1* so far described are localised in the small intergenic region that extends for less than 3kb upstream from the beginning of the gene (Sequence accession number: X69118). At position -2663 from *MUC1* transcription start site (TSS) is the polyadenylation signal (CATAAA) for the adjacent gene upstream, Thrombospondin 3 (*THBS3*).

In 1999, a group of researchers made an extensive description of potential binding sites and respective transcription factors found in a fragment of about 800bp immediately upstream *MUC1* using computational methods (Zaretsky et al., 1999). Cis-elements of several types were identified in this sequence, and included ubiquitous cis-transcription elements, as well as ones that are present in particular cell types, such as mammary epithelial cells, haematopoietic cells, immunospecific cells, hepatocytes, muscle cells and even elements specific for viral promoters. From this wide variety, only a few elements have been studied and reported as functional in transcription of *MUC1*.

Experimentally, two DNAse I hypersensitive sites were identified in the promoter region. The one at –750bp from the beginning of the gene does not seem to correlate with *MUC1* expression and is not tissue specific, whereas the other, at –250bp, is present only in *MUC1* expressing tissues and correlates with the transcriptional activity of the gene (Shiraga et al., 2002). Also, a 743bp fragment in the 3`-region of the promoter was shown to induce full expression of a reporter gene (Kovarik et al., 1993).

The *MUC1* promoter has a TATA box 25 nucleotides upstream the transcription start site (TSS). In *MUC1*, both gene and promoter region are very GC rich. That high level of GC in the promoter accounts for the numerous GC boxes found, some of which act as cis-elements involved in the transcription of *MUC1*. Two of these are known to be functional Sp1 binding sites ("consensus" binding sequence: GGGCGG), one located at –99/-90 from the TSS and the other much farther, at – 576/-567 (Kovarik et al., 1993;Kovarik et al., 1996;Morris and Taylor-Papadimitriou, 2001). The –576/-567 cis-element binds the Sp1 transcription factor forming a complex that is functional in activating the transcription of *MUC1* in breast cancer cells. The other GC box at position –99/-90 also binds Sp1, activating transcription in *MUC1* expressing cells, but the way in which this element is involved is rather more complex. Because of the AG nucleotides 5` to this Sp1 site, this sequence is also a site for attachment of another DNA binding protein, SpA, which by competing with Sp1 for binding, acts as a repressor of *MUC1* transcription. Interestingly, when inactivating mutations are introduced in the sequence for both Sp1 and SpA, and consequently the binding of both is abolished, the non-expressing cells show a dramatic increase in expression of the reporter gene (Kovarik et al., 1996).

In this region of the promoter, there are also three purine/pyrimidine mirror repeat elements (M-PMR), M-PMR1 (-641/-615), M-PMR2 (-253/-237) and M-PMR3(-133/-102). However, until now, no functional role was detected in M-PMR1 and M-PMR2, despite the co-localisation of the later with one of the DNAse I hypersensitive sites described previously. Conversely, The M-PMR3 element, localised just before the SpA binding site (-101/-90), seems to act as a repressor of transcription, since deletion of this sequence results in increased expression of the reporter gene in all cell types. When active, most probably when it is in the single stranded state, it intervenes in the binding occurring at the –101/-90 sequence, favouring SpA binding to the sequence as opposed to Sp1.

Another cis-element in the *MUC1* promoter that seems to have a role in tissue-specific transcription regulation is the sequence fragment positioned at –84/-72. It contains two copies of the consensus E-box sequence (CANNTG) that together constitute the E-MUC1 enhancer element. Mutation of this sequence, which abolishes the binding of the transcription factor, leads to increased expression in non-expressing cells and reduced expression in epithelial cells. Also, when this site is mutated simultaneously with the proximal Spl site, the non-expressing cells show a 12-fold increase in expression (Kovarik et al., 1996), indicating the presence of a synergistic interaction between these two cis-elements in the regulation of tissue-specific transcription. In this case, the induced transcription may be due to activation of the distal Spl site, together with other elements, possibly another GC box at position –57/-44.

After Kovarik and collaborators established this Spl-based transcription mechanism (Kovarik et al., 1996), another group of investigators aimed to investigate whether this mechanism which has been shown to be functional by reporter gene assays was also functional with the *MUC1* gene itself (Zaretsky et al., 1999). Using mouse DA3 cells (mouse mammary epithelial cells) transiently transfected with human *MUC1* and the 743bp of promoter sequence, the authors demonstrated that human MUC1 protein was being produced in these cells, using a monoclonal antibody (mAb) specific for the MUC1 protein isoforms containing the tandem repeat. In the T47D breast carcinoma cell lines Zaretsky and collaborators found using RNase-protection assay three new transcription start sites (TSSs) together with the major start site, all within a 70bp fragment from the first site. These cells have TGFβ-receptors, and the fragment analysed contains two TGFβ elements, the authors reported that treating these cells with TGFβ1 lead to activation of two further TSSs (Zaretsky et al., 1999). Indeed, deletion constructs of the 743bp *MUC1* promoter region in a reporter gene, demonstrated that according to the cell line used, this fragment has promoter function in the 3` and 5`-parts suggesting a dual promoter, and the possibility that *MUC1* transcription may occur from different transcription start sites. This could mean that these different TSSs found, and perhaps others still to be found, can be used alternatively according to tissue-type and transcription factors involved.

1.2.2.2 Gene

Exon 1, which is 126bp long, contains a 72bp of 5̀-untranslated region, followed by the ATG initiation codon. The ATG area in MUC1- CCACCATGA is almost homologous to the Kozak consensus sequence- ACCATGG, the published sequence proved to be responsible for effective translation of DNA sequences in eukaryotic cells (Kozak, 1987a;Kozak, 1987b).

Exon 2 is the largest and the most variable exon of the gene. Not only does it comprise the tandem repeat domain, that accounts for the majority of the protein backbone, but can it vary in length by 27 nucleotides at the 5` end, due to a splicing event. The 27 nucleotides in the elongated sequence are acquired by translating sequence that originally belonged to intron 1. This alternative splice event is thought, on the basis of association in 15 samples, to be due to a single nucleotide polymorphism (SNP), A to G transition, occurring 8 nucleotides after the beginning of exon 2 (Ligtenberg et al., 1991). According to the authors, the nucleotide present in that position, A or G, determines the spliceoform. This conclusion is supported by the recent findings of Wendy Ng in the group.

The major source of variation in exon 2 is the tandem repeat domain, which is composed of a variable number of tandem repeats (VNTR) of 60 nucleotides each. The number of repeats is very variable, ranging from about 20 to 125 repeat units, and displaying a bimodal allele length distribution, with the two most common alleles having approximately 40 and 80 repeat units. The repeat units also vary in sequence as will be discussed in detail later (Fowler et al., 2003).

The region that follows, between intron 2 and exon 6, is composed of small exons intercalated with small introns, all no longer than 150 nucleotides long. Within this region, no sequence variability is known to date, though some reports describe some MUC1 protein isoforms that splice in this area (Baruch et al., 1999;Wreschner et al., 1990). This region, though not very big, is also extremely important for the mature form of the protein. Part of exon 5 sequence codes for the transmembrane domain of the protein, while exon 6 and part of exon 7 code for the cytoplasmic tail.

Between exon 6 and 7, lies a very big intron, about 1017 nucleotides long, that also contains a CA microsatellite that can have from 11 to 14 repeats (Pratt et al., 1996). Exon 7, 378bp long, contains the final 72 translated nucleotides, followed by a TAG stop codon and a 303bp of 3̀-untranslated region (UTR) with a single AATAAA polyadenylation signal.

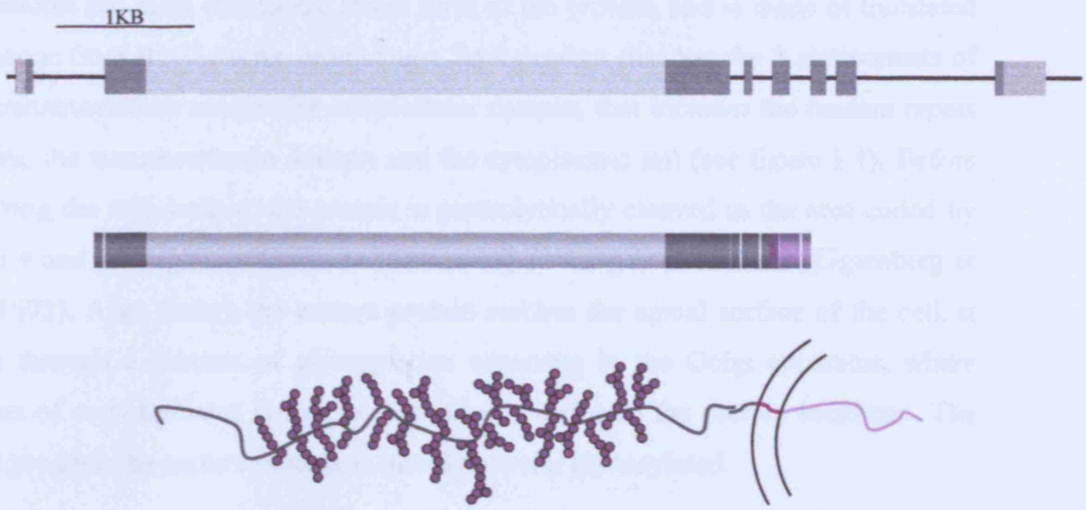


Figure 1.1: Schematic representation of *MUC1* gene (A), mRNA (B) and mature glycosylated protein (C). (A) The gene has the 7 exons represented at scale, with the two UTR in diagonal stripes boxes and the TR domain in pale grey in exon 2. (B) *MUC1* mRNA with the TR domain in grey, TM domain coloured in dark lilac and the CT in paler lilac. The same colours are used for the TM and CT domains of the protein (C), and the glycosylation in the EC domain is shown as lilac circles.

1.2.2.3 MUC1 Protein

MUC1 protein is fairly ubiquitously expressed in epithelial cells. It is present at high levels in the gastro-intestinal tract, (particularly the stomach, but much less in the colon), respiratory tract and female reproductive tract. Being a transmembrane protein, MUC1 usually localises in the apical surfaces of the epithelial cells, in the interface between the organism and the extracellular environment.

The MUC1 protein sequence was deduced from several cDNA clones that were made available in the early 90's. The first cDNA clone and respective protein sequence was published by Gendler *et al* (Gendler et al., 1990) showing a 474 amino acids sequence with only one repeat from the tandem repeat array represented. Sequences showing high degree of identity with this sequence were published at about the same time (Lan et al., 1990;Ligtenberg et al., 1990;Wreschner et al., 1990), confirming the sequence of the non TR parts of the molecule. This original sequence constitutes the most commonly found form of the protein, and is made of translated sequence from the 7 exons, producing a final product that has the 3 components of the transmembrane mucin: the extracellular domain, that includes the tandem repeat region, the transmembrane domain and the cytoplasmic tail (see figure 1.1). Before reaching the cell surface, the protein is proteolytically cleaved in the area coded by exon 4 and undergoes subsequent reannealing forming heterodimers (Ligtenberg et al., 1992). Also, before the mature protein reaches the apical surface of the cell, it goes through a process of glycosylation occurring in the Golgi apparatus, where chains of carbohydrates O-link to threonine or serine in the protein backbone. The final product, the mature protein, is usually heavily glycosylated.

1.2.2.3.1 Protein sequence present in the databases

Despite the knowledge on the protein sequence of MUC1, no actual sequence was obtained from protein sequencing. Instead, all sequences available and published come from translation of mRNA fragments obtained. Nevertheless, a consensus sequence for MUC1 protein is available at http://us.expasy.org/sprot/ (Protein accession number: P15941). According to the information available in the web-resources, MUC1 protein is made up of 1255 amino acids. This sequence contains 42 repeat units.

1.2.3 MUC1 processing

1.2.3.1 Glycosylation and Biosynthesis

As mentioned at the beginning of the chapter, mucins are glycoproteins with a very high molecular weight. This high molecular weight is partly due to the heavy glycosylation associated with the protein backbone. The glycosylation of proteins occurs by attachment of the monosaccharide chains to specific amino acid residues. According to the type of link two different types of glycosylation can occur: N- type or O- type, the latter being the most common in mucins.

O-glycosylation adds carbohydrate side chains to the hydroxyl groups of the amino acids serine and threonine and initiates through the addition of galactosamine. Indeed because of the TR array, which constitutes more than half of the protein backbone, according to allele length, serine and threonine residues that account for some 25% or more of the amino acid content of the protein.

N-glycans in general constitute a small portion of the total carbohydrate chains present in mucins. N-glycosylation initiates on asparagine amino acid residues, requiring the amino acid motif Asn-X-Ser/Thr to occur. The inner core is constituted by two GlcNAc residues followed by 3 mannose residues. This inner core can subsequently elongate to make diverse forms.

The processing of the MUC1 mucins starts shortly after the initiation of translation.

According to the results of Hilkens *et al* (Hilkens and Buijs, 1988), this oligomannose glycan addition (N-glycosylation) occurs 1 minute after the biosynthesis of the protein, within the endoplasmic reticulum (ER). O-glycosylation occurs in the Golgi apparatus, after the proteolytic cleavage, and mucins are processed through the cis, midi and trans Golgi, acquiring in the process more complex and branched glycosylation (Figure 1.2) (Hanisch et al., 1996;Hilkens and Buijs, 1988;Litvinov and Hilkens, 1993). Figure 1.3 shows the 5 potential glycosylation sites per repeat marked circles. The glycosylation of each repeat and of the entire repeat array is tissue specific and is dramatically changed in neoplasias (Burchell et al., 2001;Hanisch et al., 1996;Hanisch and Muller, 2000;Lloyd et al., 1996;Reis et al., 1998).

The N-glycans are likely to be in the extracellular domain, C-terminal to the tandem repeat array, where 5 potential N-glycosylation sites (Asp-X-Ser/Thr) can be observed (Ligtenberg et al., 1990).

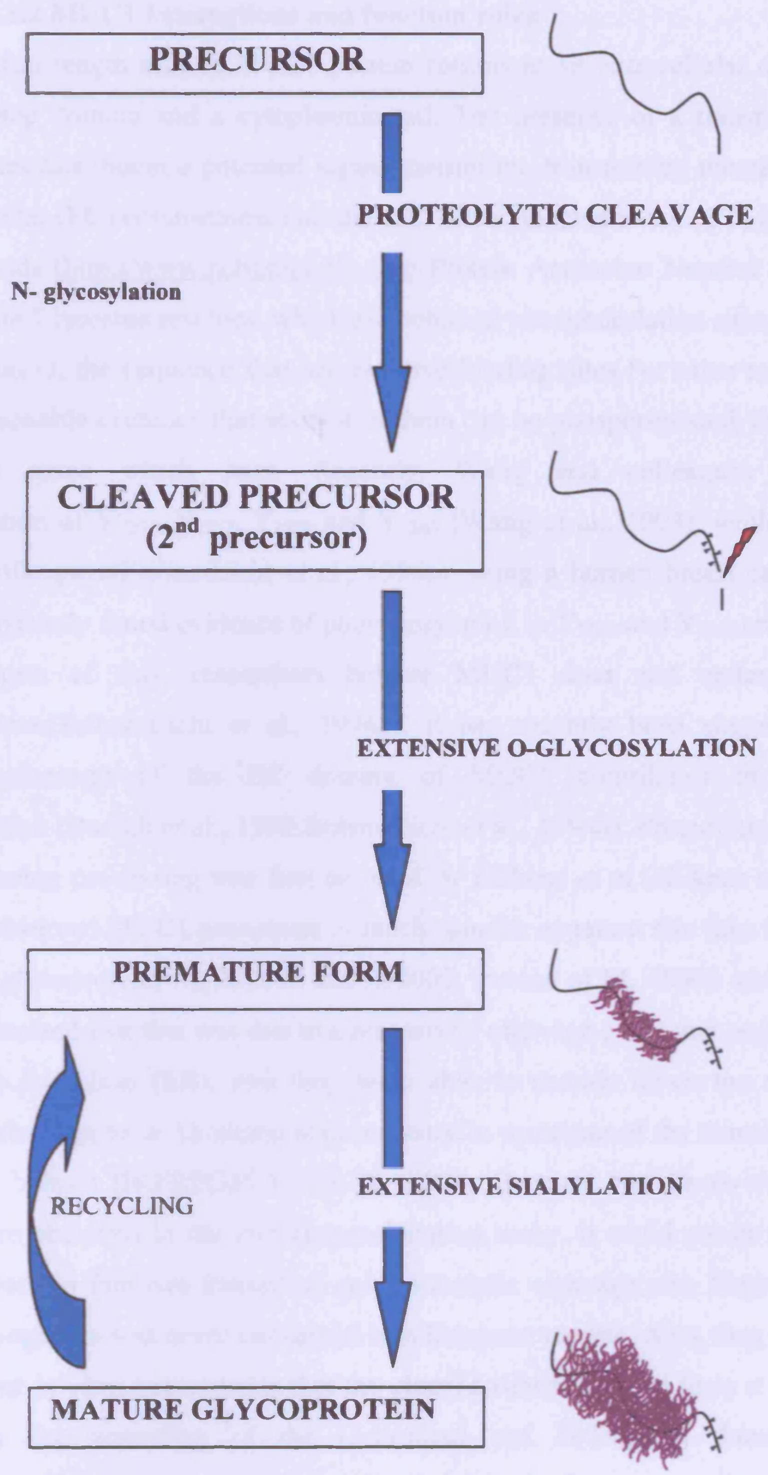


Figure 1.2: Schematic representation of *MUC1* biosynthesis from the initial precursor protein backbone to the fully glycosylated mature protein.

1.2.3.2 MUC1 Interactions and function roles

The full length mature MUC1 protein comprises an extracellular domain, a transmembrane domain and a cytoplasmic tail. The presence of a transmembrane domain makes this mucin a potential signal transmitter, transmitting messages from the extracellular (EC) environment into the cell. The cytoplasmic tail (CT) comprises 74 amino acids (http://www.ncbi.nlm.nih.gov/ Protein Accession Number P15941), which include 7 tyrosine residues, which are potential phosphorylation sites, and also contains areas in the sequence that are putative binding sites for other molecules. There is reasonable evidence that about 4 of them can be phosphorylated, though not all authors agree which ones. Recently, Wang and colleagues reported phosphorylation of Y_{1203}, Y_{1212}, Y_{1229} and Y_{1243} (Wang et al., 2003), while Zrihan-Licht and colleagues(Zrihan-Licht et al., 1994a), using a human breast cancer cell line, had previously found evidence of phosphorylation in Y_{1229} and Y_{1243} and Y_{1218}.

In spite of this, researchers believe MUC1 does not undergo auto-phosphorylation(Zrihan-Licht et al., 1994a). It has recently been suggested that proteolytic cleavage of the EC domain of MUC1 contributes to the CT phosphorylation (Baruch et al., 1999;Zrihan-Licht et al., 1994a). Proteolytic cleavage of MUC1 during processing was first reported by Hilkens *et al* (Hilkens and Buijs, 1988) who observed MUC1 precursors of much smaller apparent size than the newly synthesised glycoprotein, M_r 250000 and 350000, instead of M_r 45000 and 650000. They hypothesised that this was due to a proteolytic cleavage event that occurs in the endoplasmic reticulum (ER), and they were able to narrow down the region of proteolytic cleavage to an 18 amino acid sequence in upstream of the transmembrane anchor (see below) (IKFRPGSVVVQLTLAFRE). Because two bands very close together were observed in the immunoprecipitation assay, it could not be ruled out that there were in fact two instead of one proteolytic cleavage site. However, this second cleavage site was never confirmed in subsequent reports. Also, they observed in *in vivo* and *in vitro* experiments that the cleaved subunits could form a complex, possibly by the annealing of the C-terminal and N-terminal domains, *i.e.* heterodimers. The cleavage site was finally narrowed to a precise position: $FRPG_{1097}/S_{1098}VVV$, 65 amino acids upstream the transmembrane domain (Parry et al., 2001). Parry *et al* could not observe a second proteolytic cleavage site in their experiments. Using a cell line transfected with a construct of MUC1 lacking the tandem repeat domain, they observed that the proteolytic cleavage also occurred at the same position, showing that the TR domain was not essential for proteolysis. It was then established that this mechanism of cleavage occurs on the fully glycosylated MUC1, possibily still in the endoplasmic reticulum.

This proteolytic cleavage site is located in a region of MUC1 EC domain downstream of the TR in a domain known as the SEA module (Bork and Patthy, 1995;Wreschner et al., 2002). SEA modules usually occur in membrane docked proteins that undergo extensive post-translation modifications, in particular O-linked glycosylation. Also, the preservation of the structural fold of the SEA sequence is crucial. These sequence domains contain some highly conserved motifs across proteins and across species, in particular the GVSSS cleavage site.

Two other small sequence fragments are highly important in this module: one in the sequence Y_{1065}YQEL, the other in the sequence L_{1058}ED, 39 and 32 amino acids upstream the cleavage site, respectively. Without the presence of at least these 3 pieces of sequence, the proteolytic cleavage does not occur.

Yet, once the proteolytic cleavage occurs inside the cell, the two subunits reassociate subsequently and form heterodimers in the cell membrane. The authors speculate that these "off" and "on" reactions, cleavage and reannealing, lead to conformational changes in the membrane associated unit, recruiting tyrosine kinases and serine/threonine kinases that will phosphorylate the CT domain, this way initiating a signalling cascade.

Evidently other factors must be involved in the signalling process, in particular bacterial adhesins with affinity to the MUC1 ligands. It has been shown in a hamster tranfection model that *Pseudomonas aeruginosa* flagellin protein acts as an adhesin with the ability to regulate phosphorylation of serine and tyrosine amino acid residues of the Muc1 cytoplasmic tail (Lillehoj et al., 2001;Lillehoj et al., 2002;Lillehoj et al., 2004). These phosphorylation/dephosphorylation events lead to a signalling cascade Grb2-Sos-Ras-MEK1/2-ERK1/2, initiating an inflammatory response. Interestingly, there is also evidence that MUC1 interacts with the bacterium *Helicobacter pylori* in the human gastric mucosa via the BabA adhesin (Linden et al., 2004). *H. pylori* binds to MUC1 under neutral pH conditions. Also, under the same conditions, the BabA adhesin from the bacteria binds other 'mucin-like' components that do not react with anti-MUC1 antibody, but react with H-type-1 ligand structure. The authors suggest that the BabA adhesion to MUC1 could for example lead to detachment of the EC domain of the protein from the cell surface and subsequently induce a signalling pathway over the epithelial barrier enhancing the inflammatory response.

Another MUC1 interaction that seems to play an important role is the one that takes place with β-catenin.

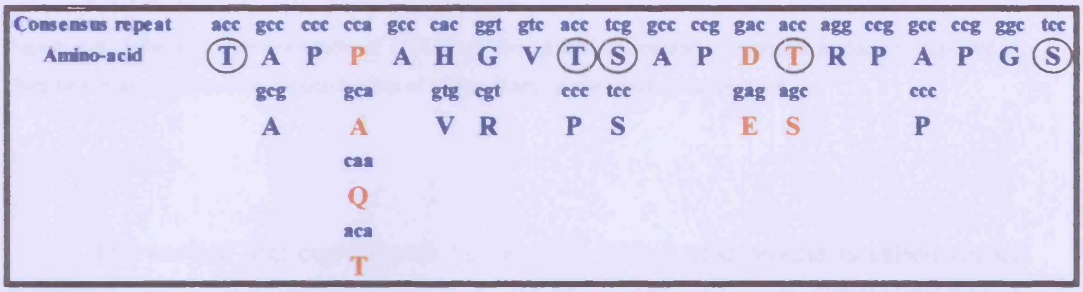
This molecule has different functions according to its intracellular location; it has an important role in cell-cell adhesion, linking to the cell cytoskeleton the cytoplasmic tail of the adhesion molecule E-cadherin, but also as a transcription factor if present in the nucleus. The MUC1 CT has a peptide domain that interacts with β-catenin (SAGNGGSSLS) (Yamamoto et al., 1997). This interaction depends on various factors, including the amount of MUC1 present in the cell, the presence of E-cadherin and the phosphorylation state of MUC1 CT (Carraway et al., 2003;Li et al., 1998;Ren et al., 2002;Wen et al., 2003;Yamamoto et al., 1997). Different kinases that phophorylate different CT amino acids may have either an enhancing or inhibitory effect on the MUC1- β-catenin interaction (Li et al., 1998;Ren et al., 2002). As a consequence of this complex formation, E-cadherin is no longer able to bind β-catenin, compromising in this way the cell-cell adhesion mechanism. In experiments in a cancer cell line, where MUC1 is not polarised, it has been shown that MUC1 can inhibit E-cadherin adhesion properties in a length dependent manner (Wesseling et al., 1996). Long MUC1 molecules inhibit adhesion by simple steric hindrance, while small MUC1 molecules only inhibit when sialylated which causes charge repulsion between cells.

If this MUC1 CT- β-catenin interaction occurs in the nucleus, this coupling will help to stabilise β-catenin and prevent its degradation by the APC protein. β-catenin accumulation in the nucleus will therefore result in activation of its role as transcription factor (Wen et al., 2003). MUC1 can in this way be seen as a modulator of nuclear activity of β-catenin.

1.2.4 MUC1 hypervariability:

In addition to variability of tandem repeat number and in contrast to the early view that most of the tandem repeats were identical in sequence it is now known that there is considerable variability from repeat unit to repeat unit and that these nucleotide changes also differ across the array in different alleles (Figure 1.3) (Engelmann et al., 2001;Fowler et al., 2003;Muller et al., 1999;Siddiqui et al., 1988;Wreschner et al., 1990). Curiously a high proportion of the nucleotide changes lead to amino-acid changes.

The only well characterised single nucleotide polymorphism is the g/a SNP (G3506A, with accession number rs4072037) at the beginning of exon 2 that is mentioned above. As well as its relevance to splicing studies to show that this SNP is tightly associated with a microsatellite polymorphism in intron 6 of the gene provide the important and initially unexpected evidence that the TR variability cannot have been generated by unequal crossing over at meiosis, but is more likely to have been generated by non reciprocal exchanges (Pratt et al., 1996).



Consensus repeat	acc	gcc	ccc	cca	gcc	cac	ggt	gtc	acc	tcg	gcc	ccg	gac	acc	agg	ccg	gcc	ccg	ggc	tcc
Amino-acid	T	A	P	P	A	H	G	V	T	S	A	P	D	T	R	P	A	P	G	S
		gcg		gca		gtg	cgt		ccc	tcc			gag	agc			ccc			
		A		A		V	R		P	S			E	S			P			
				caa																
				Q																
				aca																
				T																

Figure 1.3: Nucleotide and amino acid 'consensus' sequences of the MUC1 TR domain and reported changes. The amino acids enclosed in circles represent the 5 potential glycosylation sites of the TR; the nucleotide sequence changes reported so far are show below the consensus sequence, as well as the amino acids encoded by those changes. The highlighted aminoacids (orange) are the change that were studied in our lab, in particular the DT to ES change.

1.2.5 Splice variants of MUC1

Despite the hypervariability of the tandem repeat array of *MUC1* gene, to date not many polymorphisms have been identified and confirmed in other regions of the gene. However MUC1 does, as indicated undergo alternative splicing, and some of this is apparently under genetic control.

Soon after the first MUC1 protein sequence was available, several variants of the original found in a variety of cell lines started to be reported, all generated by alternative splicing events.

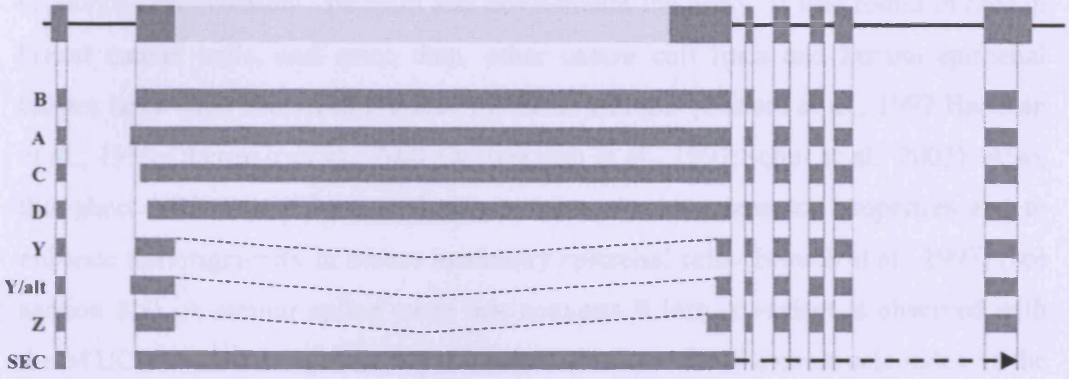


Figure 1.4: Schematic representation of *MUC1* and the several spliceoforms reported to date, as reported by Obermair *et al.* A more detailed description of each spliceoform is given in section 1.6.

The variant that corresponds to the 1255 amino acid protein is called variant B. Three other isoforms that are very similar to this have also been described. Variant A, is very similar to variant B, except for 9 extra amino acids included in the protein. These amino acids are coded by 27 nucleotides in the 3` end of exon 2. This alternative splice event is believed to be due to an A to G transition in nucleotide 3506 (DNA accession number M61170), 8 nucleotides from the start of exon 2 of variant B (Ligtenberg et al., 1991) in which the G allele is associated with the insertion. Variants C and D, though submitted directly to Genbank in 1992 by Buluwela and collaborators (Genbank accession numbers: Z17324 and Z17325), were only formally reported in 2001 by Obermair and collaborators in cervical carcinoma cell lines (Obermair et al., 2001). The two variants, C and D, correspond to deletions of 3 and 12 amino acids, respectively, from the start of exon 2 in comparison with variant B. Thus there must be 4 different splice acceptor sites in this area of the gene. The two main variants A and B have different signal peptides and after cleavage by signal peptidase have different amino-terminal ends. Variant C is predicted to end up like B, while D loses a cleavage site (see Chapter 7, figure 7.1). All four forms are thought to be the same with respect to the rest of the protein.

Other splice variants lacking large exonic regions were also observed by several authors. Zrihan-Licht and collaborators were the first to report a small MUC1 protein with an invariable size of 1.2Kb (Zrihan-Licht et al., 1994b). This isoform, called MUC1/Y, is devoid of the tandem repeat array, as well as the exon 2 sequences immediately upstream and downstream the array. It was found in human breast cancer cells, and since then, other cancer cell lines and human epithelial tissues have been shown to produce the same isoform (Baruch et al., 1997;Hartman et al., 1999;Obermair et al., 2001;Oosterkamp et al., 1997;Schut et al., 2003). Also, this short MUC1 protein was shown to have cytokine receptor properties and to enhance tumorigenicity in mouse mammary epithelial cells (Baruch et al., 1997) (see section 1.5). A similar splice event that converts B into A variant is observed with the MUC1/Y isoform generating the MUC1/Yalt, with 27 extra nucleotides at the beginning of exon 2.

Two different groups described another isoform similar to MUC1/Y simultaneously and named it differently: MUC1/X or MUC1/Z (Baruch et al., 1997;Oosterkamp et al., 1997). This isoform is similar to MUC1/Y, but it uses a different splice acceptor site within exon 2 that is 54 nucleotides upstream the MUC1/Y acceptor site. This leads to the formation of an identical protein with only 18 extra amino acids in the extracellular domain of the protein. However, despite this high degree of similarity, it is not yet known whether this form has the same properties for the MUC1/Y (see below). Because the MUC1/Z name is the one found in subsequent publications, this is the name given to this spliceoform in Figure 1.2.

MUC1 protein can also apparently exist as a soluble form. The spliceoform MUC1/SEC was reported for the first time by Wreschner *et al*, when characterising the MUC1 cDNAs in human breast cancer tissue (Wreschner et al., 1990). This isoform contains the translated sequence corresponding to exon 1 and 2, and a unique sequence of 11 extra amino acids coded by the first 33 nucleotides of intron 2 (VSIGLSFPMLP), where it encounters a STOP codon. This mRNA extends the length of the rest of the genomic DNA, without splicing at the exon/intron boundaries, creating an mRNA with an extremely long 3`UTR. This molecule does not have a transmembrane domain and a cytoplasmic tail it is thought to be released as a soluble form, without attaching to the cell membrane at any stage. Nonetheless, it has been postulated to play an important role in the mechanism of cell interaction. In *in vitro* experiments MUC1/SEC has been shown to be able to bind MUC1/Y isoform with high affinity, and to induce phosphorylation in the cytoplasmic tail of MUC1/Y (Baruch et al., 1999). However, how this leads to recruitment of kinases inside the cell to induce phosphorylation is still not understood.

1.2.6 MUC1 homologies from other mammals:

The *MUC1* gene and respective protein are known to be present in a variety of species, at least in mammals, where several species have the gene in their genome and are also able to produce the corresponding protein.

Spicer and collaborators constructed alignments between the MUC1 cDNA and protein of cow, rabbit, golden hamster, guinea pig, mouse, gibbon and human (Spicer et al., 1995). cDNA from the various species was obtained by RT-PCR. Clones from cDNA libraries were isolated to obtain the tandem repeat domain, which is particularly difficult to amplify.

The extracellular domain (EC) of MUC1 was generally poorly conserved across species. The MUC1 tandem repeat (TR) domain in particular seems to have diverged widely, which is perhaps not surprising in view of the high mutation rate of repetitive sequences. While human, gibbon and bovine *MUC1* have a variable number of tandem repeats (VNTR), rabbit *MUC1* seem to have a non polymorphic tandem repeat array, with a fixed number of repeats in all individuals, a characteristic previously observed in the mouse (Spicer et al., 1991). Also, the tandem repeat array shows a high degree of diversity in nucleotide sequence and respective amino acid sequence, from repeat to repeat. Though gibbon TR seems to be very similar to the human consensus TR sequence, the other species show several modifications in the nucleotide sequence. Nonetheless, MUC1 tandem repeat does maintain some characteristics between species: it has an average length of 20 amino acids per repeat and shows a high content of Serine/Threonine and Proline amino acids which are potential glycosylation sites in the protein.

The EC proximal region of the TM domain is also poorly conserved, except for some very specific regions. The YYQEL sequence (amino acids 1065 to 1069, sequence accession number: P15941) in present in all species, and is though to be involved in the apical sorting of MUC1. This corroborates work by Pemberton *et al* who found that the extracellular domain of MUC1, excluding the tandem repeat domain, is involved in the apical sorting of the molecule, together with the CQC sequence present in the C-terminal side of the transmembrane domain (Pemberton et al., 1996). The KDEL amino acid sequence is also present in all the species compared, and is similar to a short sequence known to be involved in the retention and return of resident proteins in the ER (Pelham, 1996). The cytoplasmic tail and the transmembrane domain of the MUC1 protein represented the most highly conserved regions (Spicer et al., 1995). In the CT, 6 out of 7 tyrosine (Y) residues as well as the T-X-R sequence are conserved across all species, indicative of the importance of the these potential phosphorylation sites in the protein.

1.3 Gastric disease

1.3.1 Gastrointestinal tract- stomach

The gastrointestinal tract is in direct contact with the external environment. It is colonised with many different sorts of bacteria. These range from "friendly" bacteria, that acquired the ability to live in a symbiotic relationship with us, helping to break down certain substances that the gut is not able to do on its own, to ones that acquired the skill to survive in the most inhospitable environment possible, and still produce damage.

After mastication in the mouth, the stomach is the first stopping point for foods and other ingested materials. It is a major digestive organ, which is particularly exposed to external factors.

1.3.1.1 Stomach: anatomy, histology and physiology:

The stomach is a dilated part of the digestive tract that is connected with the oesophagus in its upper region, and with the duodenum at the other side. It is divided into four main regions: fundus, body, antrum and pylorus. The fundus is the region in the upper side of the stomach that makes contact with the oesophagus through the cardia; it is here that the digestive gases are collected. The body is the major part; it is responsible for the secretion of hydrochloric acid and pepsinogen. The antrum extends from the body to the pyloric region, and together they constitute the area where most of the mucous glands can be found. Together with the lower body, the antrum is also the region responsible for the strong peristaltic movements during digestion.

Despite the differences between the regions of the stomach, the basic histology is very similar. The stomach mucosa is made-up of gastric glands that open into the stomach lumen through the gastric pits. The gastric pit is mainly made of mucus-secretor columnar cells, and the pit narrows into the deeper glands of the stomach. Usually, in the body region the gastric pits tend to be short, and the gland long. On the other hand, in the antrum/pylorus region, deeper pits are seen with shorter glands.

Each gland comprises a variety of cells playing different roles, according to their secretions. On the base of the gland are numerous Chief cells (or Zymogenic cells) that secrete pepsinogen, an inactive precursor of the proteolytic enzyme pepsin, that is able to initiate the break down of proteins during digestion; the enzyme is produced in an inactive state to prevent proteolytic damage of the gastric mucosa. In the middle region of the gland, the neck, several cells coexist. The most numerous cells in the neck are the mucous neck cells that secrete mucus lining the surface epithelium. The parietal cells (or oxyntic cells) secrete hydrochloric acid and intrinsic factor; the hydrochloric acid maintains a very acidic pH in the stomach, that activates pepsinogen into pepsin and helps creating a hostile environment to possible colonisers; intrinsic factor is a glycoprotein important in the intestinal absorption of vitamin B12, which in it turn is an important enzyme cofactor crucial in the development of erythrocytes. The neck also hosts a number of endocrine cells that produce several hormones involved in in the digestive process. Of the endocrine cells, the G cells are noted for the production of gastrin, a hormone that is involved in digestion by stimulating the production of hydrochloric acid, pepsinogen and intrinsic factor, and by increasing stomach motility by inducing the distension of the antrum. Finally, a large number of stem cells, situated mostly in the neck and isthmus (upper part of the gland) serve to replace the lost cells and renovate the epithelium.

1.3.2 Mucins in the normal stomach

As mentioned before, many of the cells in the stomach epithelium produce mucus, a layer that lines the epithelium, lubricating and protecting it from external agents. The mucus is composed mostly of water, minerals and high molecular weight glycoproteins, in particular mucins.

The mucins typically expressed in the stomach are MUC1,MUC5AC and MUC6. In the normal stomach, MUC1 is typically expressed at the apical surface of the cells lining the gastric mucosa (Vinall et al., 2002). MUC5AC and MUC6 are the secreted gel-forming mucins that actively contribute to the viscous and elastic nature of the mucous. While MUC5AC is mainly secreted by the surface foveolar cells, MUC6 is secreted by the neck mucus cells deeper in the glands. Despite this difference in the localisation of the secreting cells, these two mucins are expressed together throughout the gastric epithelium and glands, and are the major constituents of the gastric mucus. On rare occasions, some MUC2, MUC3 and MUC4 may occur in the normal stomach, exhibiting a patchy and weak expression. These three mucins, together with some MUC5B, have been also observed to be expressed in the stomach at early developmental stages, and subsequently downregulated at a later stage during fetal life (Buisine et al., 2000).

1.3.3 Gastric disease and environmental factors involved

Gastric disease, and in particular gastric cancer, is a major burden to - human healthcare across the world. According to the cancer global estimates for 2002 from the International Agency for Research on Cancer (IARC) gastric cancer is the fourth most prevalent cancer in the world, after lung, breast and colorectal cancers. However, stomach cancer is the second most common cause of death from cancer, after lung cancer (Parkin et al., 2005). In Europe specifically, stomach cancer ranks fifth in the most common cancers and third in death rates (Boyle and Ferlay, 2005). This high mortality rate associated with gastric cancer is probably due to the fact that it tends to be diagnosed at a late stage.

Despite some controversy on the definition of types of gastric cancer, it is widely accepted that it can be divided in two major types: the 'intestinal type', that accounts for about 90% of the human gastric adenocarcinomas, and 'diffuse type'

(LAUREN, 1965). While the diffuse gastric cancer does not have a defined aetiology, some of the causal agents for the 'intestinal type' are widely known.

This thesis focuses on the study of precancerous lesions that occur in the stomach prior to intestinal type cancer, and agents that may intervene in the process. *Helicobacter pylori*, the bacterium that colonise the human stomach, has been accepted as a major causal agent for gastric disease, and ultimately cancer. In 1994, the IARC classified *H. pylori* as a human type I carcinogen (1994a;1994b). This bacterium, and the role played in gastric carcinogenesis will be address in more detail later in this chapter. Other established agents that may be causal of intestinal type carcinoma consist mainly of lifestyle factors, in particular diet, alcohol intake and smoking habits. There is suggestive evidence that certain foods, such as fried food, foods with high salt concentration (such as dried salted fish, cured and smoked meat) and fibrous vegetables all contribute to the chemical or mechanical damage of the gastric mucosa by breaking the mucous barrier. Also, nitrogen rich nutrients such as nitrites are converted in N-nitroso compounds, a potential mutagen and carcinogen (Correa et al., 1975). On the other hand, consumption of fresh fruit and vegetables, and particularly vitamin C are effective in reducing gastric cancer risk. In contrast, the role of alcohol intake and smoking remains unclear. Several studies focused on these two agents showed a positive association of smoking and/or specific alcohol drinking with gastric cancer. The association were usually not very strong, but significant, nonetheless. Specifically, an increased risk of gastric cancer was found in smokers that were infected with *Helicobacter pylori* (Tredaniel et al., 1997;Zaridze et al., 2000).

In summary, despite gastric disease being undoubtedly a multifactorial disease, *Helicobacter pylori* does play a major role in the gastric cancer and the process of gastric carcinogenesis, in the development of intestinal type gastric cancer, rather than diffuse type (Parsonnet et al., 1991).

1.3.4 Gastric cancer precursor lesions

The intestinal type gastric cancer is the product of a long multiple sequential changes pathway, that usually takes several decades to reach neoplasia (Correa et al., 1975;Correa, 2005). That sequential mechanism was has proposed originally in 1975 by Correa and collaborators, with the multistep proposed model shown in figure 1.5.

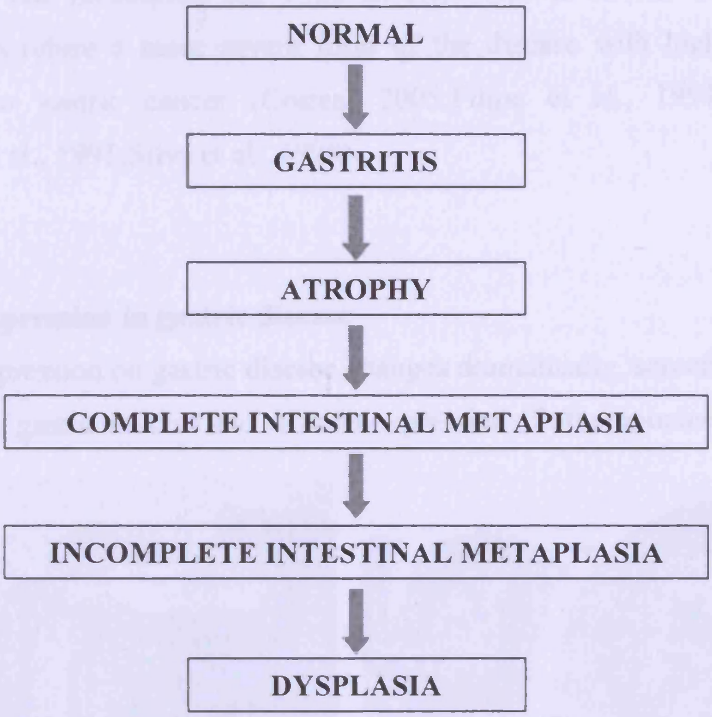

NORMAL
GASTRITIS
ATROPHY
COMPLETE INTESTINAL METAPLASIA
INCOMPLETE INTESTINAL METAPLASIA
DYSPLASIA

Figure 1.5: Multistep process of gastric cancer precursor lesions. (Adapted from Correa, 2005)

Usually in the presence of *Helicobacter pylori*, the stomach develops gastritis, as an attempt to fight the infection. However, the fact that most bacteria do not remain attached to the epithelial cells, leads to chronic gastritis with large number of mononuclear cells in the inflammatory infiltrate, and sometimes accompanied of acute inflammation (active chronic gastritis) marked by the presence of neutrophils (Dixon et al., 1996). In some cases, the chronic gastritis develops not only in the antrum, where it is usually observed, but also in the body region of the stomach. This kind of gastritis tends to evolve to a stage characterised by the loss of gastric glands- atrophy, that frequently evolves to intestinal metaplasia (IM) (Dixon et al., 1996).

The intestinal metaplasia can be divided into 2 major groups: Complete IM, also know as Type I IM, characterised by a histological morphology typical of small intestine and presence of sialomucins, and Incomplete IM, that can be Type II or Type III, exhibiting a typical colon morphology, producing neutral and sialomucins in Type II and sulphomucins in Type III, but exhibiting a mix of gastric and colonic characteristics. The Incomplete IM Type III phenotype is considered to be more unstable and therefore a more severe form of the disease with higher chance of developing into gastric cancer (Correa, 2005;Filipe et al., 1994;Reis et al., 1999;Rokkas et al., 1991;Silva et al., 1990).

1.3.5 Mucin expression in gastric disease

Mucin expression on gastric disease changes dramatically, sometimes with loss of expression of gastric mucins and *de novo* expression of other mucins.

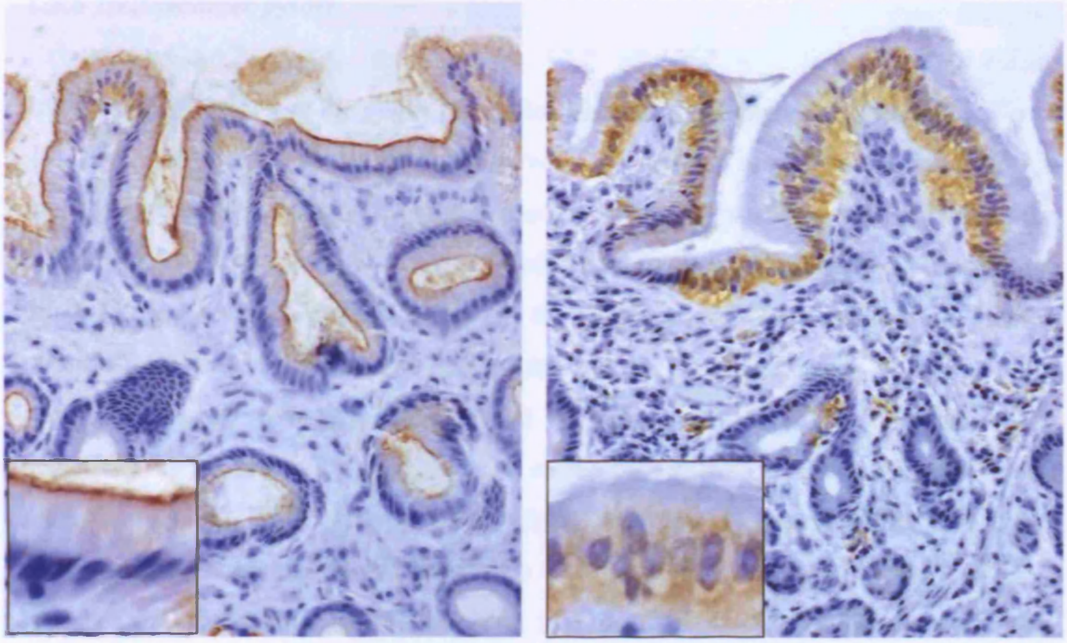


Figure 1.6: MUC1 immunohistochemistry in normal antral type mucosa, with a typical apical staining (left hand side), and antral type mucosa with *H. pylori* gastritis (right hand side), where the gastric pits lose completely the apical staining and acquire a marked perinuclear staining. (Adapted from Vinall *et al*, 2002)

In gastric chronic and acute inflammation, there is a general decrease in the expression of the 3 main mucins, MUC1, MUC5AC and MUC6 (Ho et al., 1995). It is also reported loss of the typical MUC1 apical staining and a more prominent perinuclear staining is observed (Figure 1.6) (Vinall et al., 2002).

As the process of gastric carcinogenesis progresses, the pattern of mucin expression c hanges. In intestinal metaplasia, t he gastric mucins not only decrease expression, but in some cases lose completely the gastric mucins and replaces them by intestinal mucins (Ho et al., 1995). In Complete IM, the mucosa loses almost completely the gastric mucin expression and expresses *de novo* MUC2 and other small intestine mucins; in the Incomplete IM, both type II and III seem to have a similar mucins expression pattern, with a mix of gastric and colonic mucins, in particular MUC2 (Ho et al., 1995;Reis et al., 1999;Silva et al., 2002;Teixeira et al., 2002).

1.3.6 *Helicobacter pylori*

1.3.6.1 Brief history of *H. pylori* and colonisation of the human mucosa

For a long time, investigators observed the presence of gram-negative bacilli in the stomach, in some cases detected in association with gastric disease. However, the presence of these bacteria was disregarded for more than 40 years, and observations of those organisms were attributed to the presence of other well know human colonisers, such as *Pseudomonas*, contaminating the gastric samples.

Just over 20 year ago, Marshall and Warren recognised for the first time these organisms as a novel species of bacteria, which they called "pyloric campylobacter", later known as *Helicobacter pylori* (Marshall and Warren, 1984). The two investigators characterised the bacterium, describing it as a gram-negative bacillus type bacteria that could acquire either an S-shape or curved rod appearance, usually exhibiting four sheathed flagella at one end. It was better cultured in a microaerophilic environment at 37°C. In the same report, in a study performed with a sample of 100 individuals, the authors found that bacterial colonisation was in close association with chronic gastritis, and was common in those individuals with peptic ulceration of the stomach and duodenum (Marshall and Warren, 1984). Furthermore, in the absence of inflammation the bacteria were rare, suggesting a possible role in the development of gastric disease.

Since then, this bacterium has been exhaustively studied, mainly due to its unique characteristics; the bacteria are able to colonise the Hydrochloric Acid (HCl) rich environment of the human stomach for decades. This amazing ability to colonise such inhospitable environment is probably the result from an extremely high mutation rate, leading to an enormous sequence diversity that is about 50-fold greater than the one observed in humans (Falush et al., 2003). Because of this high mutation rate, only partial linkage disequilibrium (LD) is observed in the bacteria (Falush et al., 2003;Spratt, 2003).

1.3.6.2 *H. pylori* main features

Helicobacter pylori is a gram-negative bacillus type bacteria with unique mechanisms that enable it to survive in the human stomach, an extremely inhospitable environment, as well as virulence factors able to induce disease in humans. The main barriers to overcome in the stomach are the extremely acidic environment together with the frequent peristaltic movements. *H. pylori* has evolved a mechanism to alter pH, by production of urease.

The *Helicobacter pylori* urease is a high molecular weight protein with a very high affinity for the substrate, urea, and produced in large amounts by the organism, due to the crucial role it plays in bacterial survival in the host (Mobley et al., 1988).

Despite the highly acidic pH of the human stomach, *Helicobacter pylori* is only able to survive under neutral pH. For that purpose, the bacteria created an adaptation mechanism to control the pH in the cytoplasm and extracellular microenvironment around itself. Urease is therefore an essential enzyme for the successful colonisation of the human stomach by *Helicobacter pylori.* The urease gene cluster comprises seven genes that must be all expressed for the efficiency of the mechanism. The genes can be divided into two classes: the structural genes and the accessory genes. *ureA* and *ureB* encode the urease structural subunits, producing proteins of M_r 26,657 and M_r 60,473, respectively (Clayton et al., 1989;Clayton et al., 1990;Labigne et al., 1991). However, although there are only two genes contributing to the enzyme structure, their expression is not enough for the activity of the enzyme. For that, the bacterium needs to express the five accessory genes, *ureB. ureE*, *ureF*, *ureG* and *ureH*, that encode accessory proteins involved in assembly and insertion of Nickel in the apoenzyme, essential for the occurrence of catalytic activity. Finally, the *ureI* gene, also an accessory gene, codes for an inner membrane protein with six transmembrane segments with both the N- and C-termini located in the periplasm. This protein acts as a H^+- gated urea channel to transfer urea into the cytoplasm (Weeks et al., 2000).

The bacterium regulates the pH by activating the urease enzyme in the presence of urea, hydrolysing it into ammonia and carbon dioxide, and in this way lowering the pH of the bacteria microenvironment. To avoid extreme alkalinisation of the environment, that can also be deleterious to the bacteria, the organism developed a mechanism of preventing continuous urea uptake, thought to be through mRNA decay of the *ureI* mRNA which then breaks the pathway for further urea hydrolysis (Akada et al., 2000).

Each bacterium exhibits 3 to 6 unipolar sheathed flagella. Each flagellum is encoded by 2 distinct genes, *flaA* and *flaB*, that share poor homology between them and are present in distinct areas in the *H. pylori* genome. Each gene encodes one of the two structural subunits that constitute the flagella; a subunit of M_r 53,000, FlaA that is the major constituent and FlaB, the M_r 54,000 subunit that is the minor constituent (Kostrzynska et al., 1991;Leying et al., 1992;Suerbaum et al., 1993). Together they encode a protein named flagellin, involved in the bacterial motility. The FlaB subunit, expressed in low amounts, assembles to the hook, and forms the basal portion of the filament, and the FlaA subunit, expressed at high levels, constitutes the distal portion of the flagellum. The flagellum is sheathed in a membrane-like structure containing lipopolysaccharides (Geis et al., 1993;Kostrzynska et al., 1991). To become active, the flagellin is previously glycosylated, and contains mainly neuraminic Acid (NeuAc) (Josenhans et al., 2002).

1.3.6.2.1 Virulence factors

Given the numerous *Helicobacter pylori* strains known, reflected in the DNA sequence diversity observed for this organism, it was difficult to make a good correlation between strains present in the patients stomach and disease outcome. Xiang and collaborators looked at the expression of vacuolating cytotoxin (VacA) and the cytotoxin-associated gene A (CagA antigen), two known virulence markers, in several strains of *H. pylori* and divided them into two major types of strains (Xiang et al., 1995). Type I bacteria have the *cagA* gene and produce the CagA protein and the vacuolating cytotoxin, and are considered to produce the most virulent phenotype in the patient's stomach. Type II bacteria do not have the *cagA* gene and do not express the CagA protein or the vacuolating cytotoxin, inducing a less virulent phenotype in the patients. Also, the investigators observed presence of other intermediate phenotypes that either produced or had the gene for one of the two virulence factors, identifying them as subgroups in the type I bacteria.

The *cag* pathogenicity island (PAI) was at first thought to be only one gene, named *cagA* expressing a protein with a M_r 180,000 molecular weight. This antigen was found to be highly immunogenic, and present in patients' sera infected with *H. pylori*. The protein expression was found to be associated with the expression of the vacuolating cytotoxin, a virulence factor, and therefore considered a virulence factor itself (Covacci et al., 1993). Later on, Censini *et al* found that type I strains contain not only the *cagA* gene, but also flanking it a 40kb insertion of foreign DNA that was called the *cag* region (Censini et al., 1996). The *cag* region is situated within the glutamate racemase gene, and is made up of 2 regions separated by a sequence also present in *cag* negative strains. Within these two sections of sequence lay the *cagA* gene, one of the 27 genes that compose this region (Lamarque and Peek, 2003). The genes encoded in these region present strong homologies to constituents of type IV secretion systems. This observation is consistent with the fact that these proteins could be a mechanism of export the VacA cytotoxin into the host epithelial cells (Censini et al., 1996;Lamarque and Peek, 2003).

The vacuolating properties of the vacuolating cytotoxin was first described by Leunk and collaborators, when in 1988 they observed intracellular vacuolisation of cultured cells *in vitro* in presence of *Helicobacter pylori* (Leunk et al., 1988). The factor producing this effect was then determined to be a M_r 87,000 protein, partially homologous to ion channel and transport proteins (Cover and Blaser, 1992).

1.3.7 Genes and gastric disease

It is clear from the studies described above that inflammatory disease of the stomach and gastric cancer are provoked by *H. pylori* infection and there are also other environmental risk factors, such as smoking and alcohol. However it is clear that not all people get gastric cancer if infected and also that there is great variability in susceptibility to infection and inflammation. Such observations implicate inter-individual genetic differences. There is increasing literature to suggest variation in the genes encoding proinflammatory cytokines/receptors such as IL-1β IL-1RN and TNF-α (El Omar et al., 2000;El Omar et al., 2001;El Omar, 2001;Machado et al., 2001;Machado et al., 2003) alters disease risk. The key role of MUC1 at the gastric surface led to the consideration of this gene as a candidate intervening in the development of gastric disease as will be described in more detail in Chapter 3, 4 and 6. In this thesis variation in *MUC1* is examined in case control studies.

In undertaking association studies it is important to appreciate that the genome is not comprised of sets of totally independent genes. Physical linkage resulting from co-location of loci in the same chromosomal region can lead to genetic linkage in families, and closely linked loci can remain associated over many generations, leading to so-called Linkage Disequilibrium. Methods of determining LD are described in the Methods section (Chapter 2). It is now known that blocks of Linkage Disequilibrium exist in the genome, as well as regions known as recombination hotspots (Jeffreys et al., 2004;Jeffreys et al., 2005;Stumpf and Goldstein, 2003). The existence of LD can aid association studies since fewer markers are required (Goldstein and Weale, 2001;Goldstein et al., 2003;Weale et al., 2003), but on the other hand in large regions of association make it harder to find causal elements. LD in a region reduces the numbers of haplotypes (allele combinations) seen in a population. While haplotypes are most reliably determined by study of several generation families, many algorithms are now available for inferring haplotypes from population data. These are also described in the Chapter 2.

1.4 Aims

The aims of this thesis are to explore *MUC1* gene diversity in the context of interpopulation differences and disease susceptibility, in particular gastric disease. Two independent studies previously reported an over-representation of short *MUC1* tandem repeat alleles in patients with gastric cancer (Carvalho et al., 1997) and with *Helicobacter pylori* gastritis (Vinall et al., 2002) and when compared with controls. The overall aim of this thesis was to pursue this observation further. The specific aims were:

- To characterise intra tandem repeat variability a the first small cohort of gastritis patients and controls available in the lab

- To examine three locus haplotype distribution in this and other patient and control cohorts

- To characterise the distribution of *MUC1* polymorphisms and three locus haplotypes in human populations of different ancestries

- To determine the extent of linkage disequilibrium around the *MUC1* gene

- To search for new polymorphisms within the immediate vicinity of *MUC1*

- To characterise suitable markers in the *MUC1* gene and in the region around it that would be suitable for association studies and further haplotype characterisation.

- To collect and characterise a new cohort of gastritis patients and controls, for a replication study.

CHAPTER 2

MATERIALS AND METHODS

2.1 Human Samples

Fully informed consent was obtained for all samples collected and in the UK LREC or MREC ethical approval obtained.

- First gastritis cohort: "The relationship between genetically determined variations in the intestinal mucins and intestinal injuries"
  Ethical committee: UCL/UCLH 95/3037
- Second gastritis cohort: " Genetic hypervariability of the *MUC1* membrane associated mucin in relation to gastritis and gastric cancer.
  Ethical committee: UCL/UCLH 01/0237
- MRC National Survey of Heath and Development: "Ageing and health in middle life in a MRC National Survey of Heath and Development
  Ethical committee: MREC/98/2/121

2.1.1 Northern European origin samples

2.1.1.1 MRC National Survey of Health and Development

The 1946 cohort was initially made up of 15000 individuals, all born in England, Scotland and Wales in the same week in March 1946, but was stratified by social class and reduced to 5362 individuals. Lifestyle and other information have been collected at intervals since then. In 1999, blood and buccal samples were collected from consenting individuals, and about 3000 samples from the original 5000 were obtained.

Samples were obtained from across Great Britain (England, Scotland and Wales), including 349 samples collected in the London area. These samples collected in London were the ones used in this project, as a matched geographical control group for the two gastritis cohorts, also collected in London.

2.1.1.2 Gastrointestinal Clinic – University College London Hospitals samples

2.1.1.2.1 The first Gastritis cohort

These samples were collected in collaboration with Dr. Martin Sarner. Patients undergoing gastric endoscopy or colonoscopy were approached by the researcher and invited to take part in the study. A questionnaire was completed and all patients signed a consent form agreeing to take part in the study. Each individual provided a blood sample, for DNA extraction, as well as two biopsies for research collected during the endoscopy procedure together with those required for routine histopathology. Patients also gave information about smoking habits and ethnic origin. Sample collection took place between April 1996 and September 1997, from a total of 181 patients. Some of these patients, for whom DNA was available, had undergone upper gastrointestinal tract endoscopy, and an additional biopsy was collected from these to perform a Campylobacter-like organism test (CLO-test) (Marshall et al., 1987) and determine *H. pylori* infection status. Disease status was determined by histological examination conducted in collaboration with Dr. Marco Novelli, using as a scoring reference the updated Sydney system (Dixon et al., 1996). *Helicobacter pylori* infection status was determined by three different methods: CLO-test, performed using one of the biopsies from the patient, routine histology and Immunohistochemistry, performed with paraffin wax embedded biopsy sections and a monoclonal antibody specific for *H. pylori*.

The individuals were divided into subgroups, according to disease and *H. pylori* infection status.

The individuals from whom colonic biopsies were obtained together with one from whom a duodenal sample was obtained were included in the patient control group.

2.1.1.2.2 Second gastritis survey

This series of samples were collected in collaboration with Dr Steve Pereira and the work of patient recruitment and patient data categorisation was shared between myself and Dr Adil Elamin with the assistance of Prof. Marco Novelli for the histological evaluation. All patients had been referred for upper GI endoscopy (n=239) and similar information was collected to that obtained for the first cohort. Histological slides were assessed by at least two investigators.

2.1.1.2.3 Oesophageal disease group

Samples from individuals with oesophageal disease (n=103) were collected at the same hospital in collaboration with Dr. Laurence Lovat. From each patient, a blood sample was collected as well as biopsies, for further disease diagnosis. Dr. Laurence Lovat and colleagues (including histopathologists) made the disease categorisation and recorded ethnic origin of the samples. Patients were subdivided into five major groups according to disease status: squamous cell carcinoma, adenocarcinoma, mixed carcinoma (squamous cell carcinoma and adenocarcinoma), Barrett's oesophagus (a pre-cancerous lesion) and oesophagitis, an inflammatory disease. However, in this thesis, the oesophageal disease group of patients is used in total as a control group for the gastritis study together with the colon disease patients.

2.1.2 Portuguese origin samples:

2.1.2.1 Control group

The Portuguese control group consists of 155 individuals, who were donors recruited from the blood bank of S. João Hospital in Porto, Portugal. The collection was made during two periods of time: during one month in 1993 and one month in 1995. Informed consent was obtained from each individual, and a 10 ml blood sample was taken. Samples were obtained in collaboration with Prof. Leonor David from the Mucin group in the Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Portugal. No gastric disease information was available from this control group. DNA was extracted using a Salt-chloroform extraction method (Mullenbach et al., 1989) and Southern blot data for *MUC1* tandem repeat was collected by Dr Filipa Carvalho in Porto.

2.1.2.2 Viana samples- *Helicobacter pylori* gastritis group

The samples were part of a prospective study on the development of gastric disease. The individuals were selected from the workers of a Naval Shipyard in Viana do Castelo, North Portugal. This particular region was chosen due to the very high incidence of gastric cancer in the area. Endoscopies were performed after informed consent of individuals presenting symptoms of dyspepsia. A blood sample was collected from each individual, as well as several gastric biopsies. Histological diagnosis was made in S. João Hospital, in the Pathology department, where the slides were assessed by several pathologists.

Only patients with lesions in the stomach were included in the study due to the very low number of individuals showing normal biopsies (n=3). Apart from these 3 patients all others had evidence of gastritis with or without intestinal metaplasia. No individuals had stomach ulcer, dysplasia or gastric carcinoma.

2.1.2.3 Gastric cancer group

Samples were collected from individuals suffering from gastric cancer and undergoing surgery. All patients gave informed consent and participated voluntarily in the study. Samples were obtained during surgery from carcinoma tissue and adjacent non-neoplastic gastric mucosa. The latter was used to extract DNA, which was done using a Salt-chloroform extraction method (Mullenbach et al., 1989).

2.1.3 Nigerian origin samples:

In collaboration with Dr. Rosemary Ekong, a sample of 94 individuals of Nigerian ancestry was obtained. All individuals were asked to give a buccal sample by rubbing the inner cheek with cotton buds. Later, these buccal samples were used for DNA extraction (Freeman et al., 1997). Although the samples were anonymised, family relationships were recorded. Five individuals were thus withdrawn from the study for population c omparison since they were related to other members of the cohort.

The individuals belong to several ethnic groups: 85 individuals (90% of the total sample group) belong to the Ibibio tribe (one of them belonging to the Okobo tribe, closely related), 6 are Yoruba/Isekiri and 3 individuals are Igbo/Ikwere

2.2 Materials

2.2.1 Buffers and Solutions

Agarose gel loading buffer:

0.25% Bromophenol blue, 0.25% Xylene Cyanol FF, 15% Ficoll (Type 400; Pharmacia) in H_2O

Formamide loading buffer:

1:3 parts of Blue Dextran (50mg/ml) and Deionised Formamide.

5x TBE:

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

20x SSC:

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

100x Denhardts:

2% (w/v) Ficoll type 400, 2% (w/v) Bovine Serum Albumin (BSA), 2% (w/v) Polyvinylpyrrolidone (PVP), filtered sterilised with a 0.2μm membrane.

Jeffreys' buffer:

11.1x concentrated, adapted from Jeffreys *et al* (Jeffreys et al., 1990)

0.499M Tris-HCL (pH8.8), 0.112M Ammonium Sulphate ($(NH_4)_2SO_4$), 0.050M Magnesium Chloride ($MgCl_2$), 0.074M 2-mercaptoethanol, 48.8μM EDTA, 11.1mM of each dNTPs (Amersham Pharmacia biotech), 1.25μg/μl Bovine Serum Albumin (BSA) (DNAse free, Amersham Pharmacia biotech)

PCR clean-up solution:

1M of NaCl, 2mM Tris-HCl, 0.2mM EDTA, 40% PEG-8000 (polyethylene glycol), 3.5mM $MgCl_2$. The solution is sterilized, aliquoted and kept in the fridge.

Sequencing buffer:

200mM Tri-HCl pH 9, 5mM $MgCl_2$. The solution is filtered, sterilized and aliquoted and freezed.

Depurinating solution:

0.25M hydrochloric acid (HCl)

Denaturing solution:

1.5M Sodium Chloride (NaCl) 0.5M Sodium Hydroxide (NaOH)

Neutralising solution:

1.5M NaCl, 0.5M Tris HCl, 0.001M EDTA, pH7.2

Deprobing solution:

0.1% SDS solution, prepared by adding 990ml of distilled water to 10ml of 10% filtered SDS.

Commercial buffers:

- New England Biolabs:

NEBuffer 1: (1x)

10mM Bis Tris Propane-HCl, 10mM $MgCl_2$, 1mM dithiothreitol (pH 7.0 at 25°C)

NEBuffer 2: (1x)

10mM Tris-HCl, 10mM $MgCl_2$, 50mM NaCl, 1mM dithiothreitol (pH 7.9 at 25°C)

NEBuffer 3: (1x)

50mM Tris-HCl, 10mM $MgCl_2$, 100mM NaCl, 1 mM dithiothreitol (pH 7.9 at 25°C)

NEBuffer 4: (1x)

20mM Tris-acetate, 10mM magnesium acetate, 50mM potassium acetate, 1mM dithiothreitol (pH 7.9 at 25°C)

- Boehringer Mannheim GmbH, now part of Roche Applied Science

suRE/Cut Buffer B: (1x)

10mM Tris-HCl, 5mM $MgCl_2$, 100mM NaCl, 1mM 2-mercaptoethanol (pH 8.0 at 37°C)

- GIBCO BRL, now part of Invitrogen

REact 3:

50mM Tris-HCl, 10mM $MgCl_2$, 100mM NaCl, (pH 8.0)

REact 4:

20mM Tris-HCl, 5mM $MgCl_2$, 50mM KCl, (pH 7.4)

2.2.2 Equipment

Thermal cyclers:

- Helena Biosciences Phoenix thermal cycler- Helena Biosciences, Sunderland, UK
- GeneAmp- PCR System 9700- PE Applied Biosystems, Warrington, Cheshire UK
- MJ Research PTC-200 Peltier Thermal Cycler- Genetic Research Instrumentation (GRI), Essex, UK
- MJ Research PTC-225 Tetrad PCR Systems- Genetic Research Instrumentation (GRI), Essex, UK

Sequencers:

- ALFexpress DNA sequencer- Amersham Pharmacia Biotech, Bucks, UK
- ABI PRISM 377 DNA Sequencer- ABI, Applied Biosystems
- ABI PRISM 3700 DNA Analyser- ABI, Applied Biosystems

Electrophoresis equipment:

- CONSORT Microcomputer electrophoresis power supply- Flowgen Bioscience Limited, Nottingham, UK
- Power supply ECPS 3000/150- Amersham Pharmacia Biotech, Bucks, UK

- Mini Gel Electrophoresis tank 10 x 7 cm- Flowgen Bioscience Limited, Nottingham, UK
- Electro-Fast Stretch gel systems, 96 wells plus 12 wells for marker- ABgene House, Epsom, UK.
- Horizon 20.25 Horizontal Gel Electrophoresis Apparatus- Gibco BRL, Life Technologies, Inc, Paisley, UK.
- Horizon 11.14 Horizontal Gel Electrophoresis Apparatus- Gibco BRL, Life Technologies, Inc, Paisley, UK.

Others:

- Dark Reader DR45M- Clare Chemical Research, Inc. GRI, Essex, UK
- UV Transilluminator- Ultra Violet Products (UVP), Cambridge, UK
- PCR plates- Thermo-Fast 96 Semi-Skirted- ABgene House, Epsom, UK
- Adhesive PCR Film- ABgene House, Epsom, UK
- Fuji Medical X-Ray Film Super HR-E30- Fuji Photo Film Ltd, London, UK
- Fuji cassette FG-8- Fuji Photo Film Ltd, London, UK
- pH meter- Mettler-Toledo Gmbh, Beaumont Leys Leicester, UK

2.3 Methods

2.3.1 DNA Extraction

2.3.1.1 DNA extraction from whole blood

Blood was collected in tubes containing EDTA and stored at -70°C before DNA extraction. DNA was extracted from 3ml of blood using a PUREGENE kit (Gentra Systems, supplied by Flowgen Bioscience, Wilford, Nottingham, UK) and reconstituted in 250µl of DNA hydration solution (Tris [hydroxymethyl] aminomethane and EDTA (ethylenediaminetetracetic acid), Puregene). Aliquots were stored at 4°C and -70°C.

2.3.1.2 DNA collection and extraction from buccal cells

One of the sets of samples from the MRC 1946 Longitudinal Cohort, (used for the PCR based tests) and samples from Nigeria were both collected using buccal cells as a source for DNA. Each individual was asked to rub the inner cheek with 10 cotton buds, rubbing a different part of the inside of the mouth with each bud for about 20 seconds. The cotton swabs were placed in a previously prepared 15 ml tube containing 2.5ml of a solution of STE buffer (100 mM NaCl, 10mM Tris HCl pH8.0, 10mM EDTA pH8.0), 0.2mg/ml of Proteinase K and 0.5% of SDS (Freeman et al., 1997;Meulenbelt et al., 1995).

To extract the DNA the tubes were incubated in a waterbath at 65°C for 2 hours. The contents of the tubes were inverted and transferred into 50ml conical tubes, and spun at 647g for 10 minutes in a MSE Mistral 300E bench model centrifuge at room temperature. The swabs were drained and discarded, and the liquid poured into a 15ml centrifuge tube. 300μl of Yeast Reagent 3 (YR3) (Autogen Bioclear, Wiltshire, UK) was added to the tube contents and after vigorously mixing for 1 minute the tubes were spun for 25 min in a MSE-Europa 24M centrifuge at 7000rpm. The supernatant was transferred into a clean 15ml tube and another 300μl of YR3 were added plus vigorously mixed. The centrifugation procedure was repeated. The supernatant was then transferred into a new 15ml tube, and 1.8ml of isopropanol were added. After mixing, the contents were spun down at 7000rpm in the MSE-Europa 24M centrifuge for 25 minutes. The supernatant was discarded and the pellets were washed with 70% ethanol (EtOH) and centrifuged at the same speed for 10 minutes using the same centrifuge. The EtOH was discarded and the pellets were air dried for 15 minutes. Finally, 400μl of DNA hydration buffer (TE) was added and left to ressuspend overnight at 4°C. The DNA was kept at 4°C or in some cases an extra aliquot stored at -70°C.

2.3.2 PCR general method

Polymerase Chain Reaction (PCR) was performed in reaction volumes ranging from 10μl to 25μl. The reaction mix was prepared as followed, unless stated otherwise. The final concentration of the reagents in the mix was: 1:10 dilution of 10x buffer IV (containing $MgCl_2$) was added to 0.2μM of each dNTP, 0.5μM of each primer and 0.025U/μl of *Taq* DNA polymerase. All products were supplied by ABgene (Surrey, UK). Some PCR products required addition of 5% (v/v) glycerol or 2% Dimethyl Sulfoxide (DMSO) (Finnzymes, supplied by GRI, Essex, UK) to increase PCR stringency and obtain specific PCR products.

Products were amplified using one of the thermal cyclers described a bove (Equipment section) and subjected to electrophoresis in a 2% agarose gel (unless stated otherwise), together with a PCR ladder.

2.3.3 Marker PCR

All the markers used in the agarose gels and in the ALF express were made in the lab. One antisense primer was used (Marker AS), combined with several sense primers to give final products of specific sizes: 100, 150, 210, 250, 300, 400, 501, 600, 700, 800, 900, 1000, 1200 and 1400 base pairs. For the ALF markers an antisense Cy5 labelled primer was used. The PCR was carried out as described in the general PCR methods. All the primers used for these PCRs are listed in Table 2.1. Table 2.2 has a summary of the proportions of each reagent used in the PCR.

Primer name	Primer sequence from 5' to 3'	Orientation
Marker AS	CTCTCAGGTCTGGTGTCATCC	antisense
100	CTGTTCTCTTAAGCTTACTACTGACAG	sense
150	ACATGATTCAGTCTAATCAATGGAT	sense
210	GAGGTTGAGTCATACCAAATAGTG	sense
250	TGTCTAACCACTCACTGTCATGA	sense
300	GCCTTGTCCTATGGAAGGC	sense
400	CTTCCATCCTCCAAGGCC	sense
501	TGCCTGTTGCTGTCCCTG	sense
600	CAGCTATGGGAGAGGAAGCA	sense
700	CCCAGCAGGCCCAGAA	sense
800	AAGCCTCACATATCTCCGTTT	sense
900	ACTTCTGGAAAATGACTGAGGC	sense
1000	GCCCAGTTTTGTGGCATCT	sense
1200	CCTGTCTTGGAGTCCCCAA	sense
1400	GATGTTTTCTCTGAGTCTCCCTT	sense

Table 2.1: List of primers for the marker PCRs. For each primer, the sequence is indicated from 5` to 3`, as well as the primer orientation.

Reagents	Amount (µl)
Buffer 4	1
10 x dNTP's	1
Marker AS	1
Sense primer	1
Taq	0.05
dH2O	4.95
DNA	1

Table 2.2: List of reagents and proportions used for each PCR for the marker PCR. The reagents and concentrations are the same indicated in the PCR general method.

2.3.4 *MUC1* Tandem Repeat Size determination

2.3.4.1 DNA digestion

To determine the allele sizes of the *MUC1* tandem repeat (TR), each genomic DNA sample was digested with the restriction enzyme *Hinf I*, which cuts the *MUC1* product as shown in figure 2.5. For each digestion, 12 µl of genomic DNA (200-400ng/µl) were digested with a mix with a final mix concentration of 1x NEB Buffer 2 (New England Biolabs), 17 units of restriction enzyme *Hinf I* (New England Biolabs, Herts, UK), and 0.1M o f Spermidine (Promega, Southampton, UK). The mixture was incubated overnight (ON) at 37ºC.

2.3.4.2 Southern Blot

The digests were loaded into a 0.8% agarose (Sigma-Aldrich, Dorset, UK) 24cm (300ml) gel, and subjected to 2 V/cm for 24 hours. 15µl of Raoul Molecular Weight Marker (100µg/ml) (Qbiogene, Cambridge, UK) was loaded into two wells and two digested genomic DNA controls containing allelic bands of known size (the first prepared from the cell line Caco-2 and the second a mixture of two other DNA samples) were also used to allow accurate sizing of the bands.

The gel was subjected to 'Southern Blotting', using a V acuGene XL V acuum Blotting Unit (Amersham Pharmacia Biotech). The DNA bands were transferred onto a positively charged nylon transfer membrane Hybond-N+ membrane (0.222mx0.222m) (Amersham Biosciences) as follows:

The membrane was placed in the Vacuum Blot unit on a gel supporting screen and the gel placed on the top, applying 75 millibars of pressure. The blot was made by covering the top of the gel with depurinating solution for 30 min, followed by denaturing solution for 30 min, and neutralising solution for 30 min. The DNA transfer was made with 20x SSC for 2 hours. The membrane was wrapped in 3mm filter paper and baked at 80°C for at least 2 hours.

2.3.4.3 Probing the membrane

Pre-hybridisation

The membranes were firstly pre-hybridised in a solution containing 5x Denhardts (2% Ficoll 400, 2% Bovine Serum Albumin (BSA), 2% Polyvinylpyrrolidone (PVP)), 0.5% (w/v) SDS and 6x SSC. The membranes were incubated at 65°C, with agitation, for at least 4 hours.

Probe labelling

The MUC1 cDNA probe (pum24p), produced in our lab (Yonezawa et al., 1991), was labelled with ^{32}P using the Megaprime DNA Labelling System (Amersham Biosciences, Bucks, UK). The cDNA probe (30-60ng) was mixed with water and the random primer from the kit and denatured for 5 min at 95°C and put in ice. The kit reaction buffer was the added to this mix. The probe was labelled with $\alpha^{32}P$ dCTP (ICN Biomedicals, Basingstoke, UK) using 2.22 MBq of label and 4 μl of Klenow enzyme (1U/μl), incubating at 37°C for at least 30 min. After the incubation, the labelled probe was purified, by separating the different fractions in a 'Nick Column' containing DNA Grade Sephadex G-50 (Pharmacia Biotech, Amersham Biosciences, Bucks, UK). The unlabelled probe was eluted by adding 400 μl of 3x SSC, and the labelled probe was then recovered by adding a further 400 μl of 3x SSC. The incorporation rate was measured with a Geiger counter in counts per minute (cpm), using the following formula:

$$\text{Incorporation rate (\%)} = \frac{\text{cpm of the labelled probe}}{\text{cpm of the labelled probe} + \text{cpm remaining in the column}} \times 100$$

The probe was only used if an incorporation rate of over 50% was obtained, and stored at room temperature.

Membrane probing

Once the membranes were pre-hybridised the membranes were removed from the pre-hybridisation solution. 50µl of Sonicated Herring Sperm DNA (10mg/ml) (Promega, Southampton, UK) were added to the labelled probe, to block non-specific hybridisation of the probe to the membrane. The probe was boiled for 5 min to denature and added to the pre-hybridisation solution. The membranes were returned to the solution one by one, and incubated at 65°C with agitation overnight.

Membrane washing

The membranes were washed with 3 solutions containing decreasing amounts of SSC, to increase the stringency. All the solutions had 0.1% (w/v) of SDS, and decreasing amounts of SSC: 2% SSC, 0.2% SSC and 0.1% SSC. The incubations were carried out at 65°C, with agitation. The time with each washing solution was dependent on the number of counts read after each wash, monitored with the Geiger counter.

Membrane development

The membranes were wrapped in cling film and placed overnight within a Fuji cassette FG-8 with Super HR-E30 Fuji medical X-ray film (Fuji film). The film was then developed using a Compact X2 Automatic X-ray Film Processor (Xograph Imaging Systems Ltd, Glos, UK). If necessary, longer or shorter exposures would be done, to obtain a better result.

2.3.4.4 Allele size determination

Once the autoradiographs are developed, the alleles are sized by scanning the autorad using Grab-IT2.59 (for Windows 95) software, taking particular care in lining up the top of the filter in the image. After scan, the analysis was done using the computer software GelWorks ID Intermediate v4.01 (both from Ultra-Violet Products, Cambridge, UK). With GelWorks, the size of each band is determined in relation to the size of the bands of the Raoul marker and the other controls, which bands have known sizes (Figure 2.3).

Because of the bimodal distribution of the alleles, and in order to simplify the analysis, the alleles are divided into two categories: large alleles (L), with bands bigger than 5 kb, and short alleles, comprising all the bands smaller than 5kb. This cut-point was defined by the dip between the two peaks, 5Kb, in the bimodal distribution (Vinall et al., 2002) (Figure 2.3).

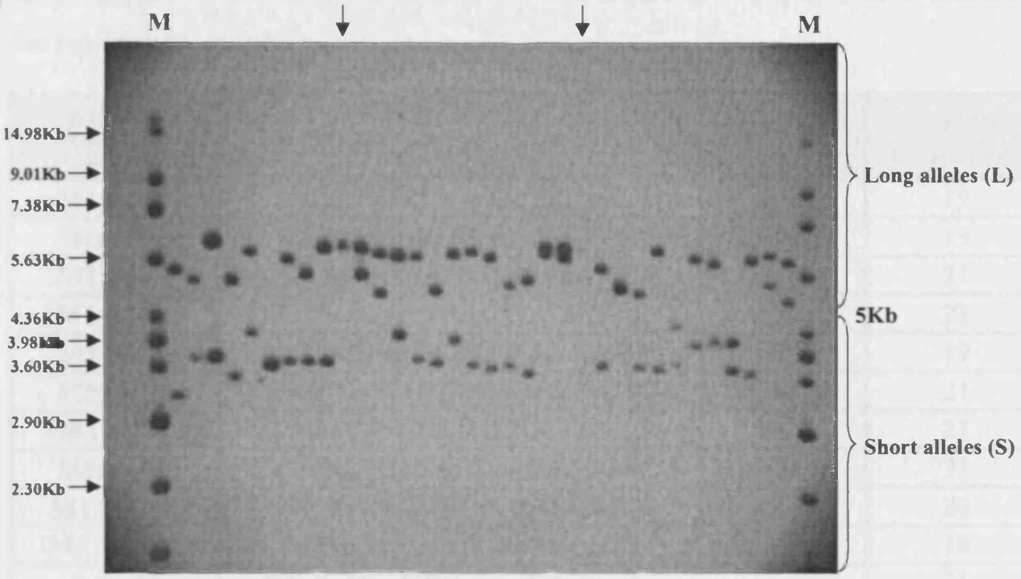


Figure 2.3: Southern Blot obtained with samples digested with the *Hinf I* restriction enzyme. The autoradiograph shows the Raoul molecular weight marker in each side of the gel and the bands size on the left side (M), and the size of the bands in Kb. The two internal controls, positioned below the arrows facing downwards, are the *Hinf I* restriction products of a cell line, Caco-2, and a mixture of 2 DNA samples.

2.3.5 Long PCR across tandem repeat

Because most individuals are heterozygous for *MUC1* TR size, and in order to obtain haplotypic MVR maps, the first step for the MVR analysis was a long PCR across the entire TR array, using specific primers for the regions flanking the array. For this purpose DNA of good quality was necessary, because if the DNA is nicked it is not possible to obtain PCRs across the entire TR region, and therefore this does not produce good MVR starting material.

The long PCR was made in a 25μl reaction, using 2 μl (80-160ng/μl) of DNA diluted 1:5 from the stock. Together with the DNA, the final reaction mix contained 1x 'Jeffreys buffer' (composition in the buffers and solution section), 0.9 units of DyNazyme EXT DNA polymerase (Finnzymes, obtained through GRI Ltd, Essex, UK), 2% DMSO, 0.5 μM of forward primer Exon2S (Table 2.4) and 0.5 μM of reverse primer MUC1E2AS (Table 2.4). The reaction is performed in a Phoenix

Thermocycler (Helena Biosciences), with 1 step of 1m30s at 96ºC for initial denaturation, followed by 22 cycles of 40s at 96ºC for denaturing, 30s at 60ºC for annealing and 4 min at 68ºC for extension. One final step of 5 min at 68ºC concluded the reaction.

Primer name	Primer sequence from 5' to 3'	Orientation	Number of nucleotides
M1proF1	TGATCTGATGCGCTCCAAT	sense	19
M1proR1	GAGAGGACGTTGGAACGTG	antisense	19
M1proF2	CGCATCTGCCTCTTAAGTACA	sense	21
M1proR2	GCTGGATAATGAGTGGACTAGG	antisense	22
M1proF3	ACAGGGAGCGGTTAGAAGG	sense	19
M1proR3	CAGCCTCTGTCTGCAATTCTT	antisense	21
M1proF4	CAGTCTCCTTTCTTCCTGCTG	sense	21
M1proR4	TACGCTGCTGGTCATACTCAC	antisense	21
M1 3utrF1	GGAGATGTGAGGAGGAGGTG	sense	20
M1 3utrR1	CTGTGCTGGGTGTGGTAGG	antisense	19
Exon2S	AAGGAGACTTCGGCTACCCAG	sense	21
Exon2AS	CTGGGACCGAGGTGACATGG	antisense	20
MUC1E2AS	TGTGCACCAGAGTAGAAGCTGA	antisense	22
1946- AGS	CCTAAACCCGCAACAGTTGTTAC	sense	23
Cy5- Exon2AS	*CTGGGACCGAGGTGACATCC	antisense	20
1946- CAS	AGAGAGTTTAGTTTTCTTGCTCC	sense	23
Cy5- CAA	*TCTTGGCTCTAATCAGCCC	antisense	19

Table 2.4: List of primers for the *MUC1* gene, including the promoter region and the region immediately downstream of the gene. For each primer, the sequence is indicated from 5` to 3`, as well as the primer orientation and nucleotide length. The labelling in the Cy5 primers in marked with an asterisk (*).

2.3.5.1 Allele separation

To separate the two alleles in the samples heterozygous for the TR size, the PCR products were loaded onto a 1% agarose (Sigma-Aldrich) 14cm (100ml) gel, and run at 2.1 V/cm for 19 hours. To improve the visualisation of the bands, the gel was post-stained for 30 min in the dark with SYBR Gold nucleic acid gel stain (Molecular Probes, Leiden, The Netherlands) diluted 10000x in TBE pH7-8.5.

Taking as a reference the *Hinf I* fragment sizes previously determined for the same individuals, and subtracting 1.02kb from that value (corresponding to the size of the region between the cut sites and the beginning and end of the TR array) (see Figure 2.5), the expected sizes of the TR bands in the gel were calculated. The bands in the gel were visualised and excised using a Dark Reader transilluminator (Clare Chemical Research), to allow DNA visualisation without nicking the DNA. Because some of the bands were not visible, despite using SYBR Gold to stain the DNA, the flanking lanes in the gel were loaded with 1 Kb DNA ladder (Gibco BRL, Invitrogen, Paisley, UK), to help determine the position in the gel from where the bands should be extracted. The gel slices after extraction were placed in an eppendorf, together with 50 µl of water with 10ng/µl of Herring Sperm DNA (Promega), crushed with a tip and frozen and thawed 3 times, to release the DNA from the gel slice.

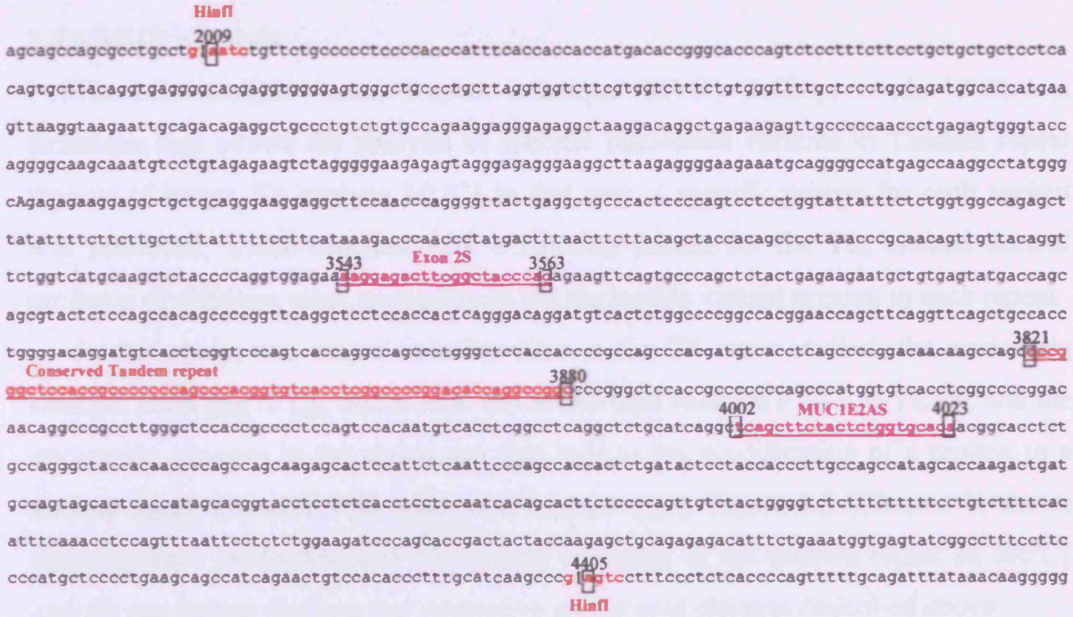

HinfI
2009
agcagccagcgcctgcctgaatctgttctgcccccctccccacccatttcaccaccaccatgacaccgggcacccagtctcctttcttcctgctgctgctcctca
cagtgcttacaggtgaggggcacgaggtggggagtgggctgccctgcttaggtggtcttcgtggtctttctgtgggttttgctccctggcagatggcaccatgaa
gttaaggtaagaattgcagacagaggctgccctgtctgtgccagaaggagggagaggctaaggacaggctgagaagagttgcccccaaccctgagagtgggtacc
aggggcaagcaaatgtcctgtagagaagtctaggggaagagagtagggagagggaaggcttaagaggggaagaaatgcaggggccatgagccaaggcctatggg
cAgagagaaggaggctgctgcagggaaggaggcttccaacccaggggttactgaggctgcccactccccagtcctcctggtattatttctctggtggccagagct
tatatttcttcttgctcttatttttccttcataaagacccaaccctatgactttaacttcttacagctaccacagcccctaaacccgcaacagttgttacaggt
3543
Exon 2S
3563
tctggtcatgcaagctctaccccaggtggagaaaggagacttggctacccagagaagttcagtgcccagctctactgagaagaatgctgtgagtatgaccagc
agcgtactctccagccacagccccggttcaggctcctccaccactcagggacaggatgtcactctggccccggccacggaaccagcttcaggttcagctgccacc
3821
tggggacaggatgtcacctcggtcccagtcaccaggccagccctgggctccaccaccccgccagcccacgatgtcacctcagccccggacaacaagccagcgccg
Conserved Tandem repeat
3880
ggctccaccgccccccagcccacggtgtcacctcggccccggacaccaggccggccccgggctccaccgcccccccagcccatggtgtcacctcggccccggac
4002
MUC1E2AS
4023
aacaggcccgccttgggctccaccgcccctccagtccacaatgtcacctcggcctcaggctctgcatcaggctcagcttctactctggtgcagacggcacctct
gccagggctaccacaaccccagccagcaagagcactccattctcaattcccagccaccactctgatactcctaccacccttgccagccatagcaccaagactgat
gccagtagcactcaccatagcacggtacctcctctcacctcctccaatcacagcacttctccccagttgtctactggggtctctttctttttcctgtcttttcac
atttcaaacctccagtttaattcctctctggaagatcccagcaccgactactaccaagagctgcagagagacatttctgaaatggtgagtatcggcctttccttc
4405
cccatgctcccctgaagcagccatcagaactgtccacaccctttgcatcaagcccgagtcctttccctctcaccccagttttgcagatttataaacaaggggg
HinfI

Figure 2.5: Nucleotide sequence of the *MUC1* sequence surrounding the TR region (extracted from sequence with accession number M61170), with the nucleotide numbers displayed on the top of framed nucleotides. The conserved tandem repeat nucleotide sequence is typed in red and underlined and the *Hinf I* sites are also marked in red. The two primers used in the 'Long PCR across TR' are shown in pink.

2.3.5.2 Gel slice check PCR

To make sure that the gel slice was removed from the correct part of the gel and therefore contained the right PCR product, and to determine the relative amount of PCR product present in the gel slice, a semi-quantitative PCR was performed. After spinning the gel slice in an Eppendorf tube for 2 min at 13000rpm, 1µl of gel slice PCR product was added to a final concentration of mix of 1x buffer IV (with $MgCl_2$)

(ABgene), together with 0.2µM of each dNTP (ABgene), 0.5µM of forward primer Exon2S, 0.5µM of reverse primer Exon2AS, 10% glycerol (v/v) (BDH Biomedical, Poole, Dorset, UK), and 0.25U of *Taq* DNA polymerase (ABgene), in a final reaction volume of 10µl. The reaction was performed in a Phoenix thermocycler (Helena Biosciences) with 27 cycles, each with 40s of denaturation at 96°C, 30s of annealing time at 66°C and 30s of extension time at 68°C.

The final PCR product was 205bp long, and run in a 2% agarose gel at 7.5 V/cm for 45 min. The intensity of the bands, due to the relatively low number of cycles, was assumed to be proportional to the amount of PCR product in the gel slice.

2.3.6 MVR analysis

The Minisatellite Variant Repeat technique (MVR) (Jeffreys et al., 1990) is a technique that allows the analysis of specific nucleotide variants in Tandem repeat regions of genes. To analyse *MUC1* in this way, a specific primer for each variant was produced, which together with a flanking primer for the TR, created several products of different sizes, according to the nucleotide variant present in each repeat.

In this study, two major variation sites in the TR were studied: the nucleotide changes from gag to cac that lead to the amino acid changes PDTR to PESR, and the nucleotide changes in the codon cca that lead to the modification of a proline to a alanine (gca), glutamic acid (caa), or threonine (aca). Figure 1.3 (Chapter 1) shows the consensus nucleotide and amino acid sequence of the tandem repeat of *MUC1* and the nucleotide changes and respective amino acid changes described above.

The method used to characterise the map of variants in the tandem repeats was an adaptation of the Minisatellite Variant Repeat (MVR) technique developed by Jeffreys *et al* (Jeffreys et al., 1990;Jeffreys et al., 1991), and is described in detail in Fowler *et al* (Fowler et al., 2003), also available in appendix 3 of this thesis.

2.3.7 Analysis of the G3506A SNP in exon 2 and the CA microsatellite in intron 6 of the *MUC1* gene:

To determine the genotypes for the two markers flanking the TR region, a multiplex PCR was used, using primers to amplify the 2 regions flanking the polymorphisms: a SNP in exon 2 that occurs before the TR (G3506A) and a microsatellite in intron 6 composed of CA dinucleotides with 11 to 14 repeats (Figure 2.6). The reaction mix was prepared by adding 1µl of DNA (diluted 1:20-ie approximately 15ng) to a final concentration of mix with 1:10 dilution of 10x Buffer IV (with 1.5mM of $MgCl_2$), 0.2µM of each dNTP, 0.2µM of each primer, 1946-AGS, Cy5-Exon2AS, 1946-CAS and Cy5-CAA, 2.5% of glycerol (v/v), 0.025U of *Taq* DNA polymerase and water up to a volume of 12µl. The PCR was prepared in 96 well semi-skirted well (ABgene) and run in a GeneAmp- PCR System 9700 thermal cycler (PE, Applied Biosystems) with the following conditions: pre-denaturation at 95°C for 5 min, followed by 32 cycles of 96°C for 30 sec, 62°C for 1 min and 72°C for 1 min and a final extension of 72°C for 5 min.

Both reverse primers were Cy5 labelled, a fluorescent marker recognised by the ALF express machine.

2.3.7.1 Restriction enzyme digestion

After amplifying the PCR products as described above, 5µl of the PCR product were digested with 1Unit (U) of the restriction enzyme *AlwN I* with 1:10 dilution of NEB buffer 4 at 37°C overnight. This enzyme cuts the G3506A SNP PCR product whenever an adenine is present at the SNP position. The restriction site is abolished when the guanine is present.

The final product was composed of 2 fragments of 265bp and 173-181bp (depending on the number of repeats), corresponding to the G3506A SNP and CA microsatellite PCR products, respectively. If the nucleotide adenine is present in SNP position, the 265bp fragment is divided in 2 fragments of 28bp and 237bp.

The samples were stored frozen away from light, to avoid degradation of the Cy5 dye. To prepare the samples for running in the ALF genotyper machine, 2µl of the PCR digest product were mixed with 2µl of Loading Dye (Amersham Pharmacia

Biotech), 2μl each of Cy5 labelled of 100 and 300 base pair marker, both prepared in our lab.

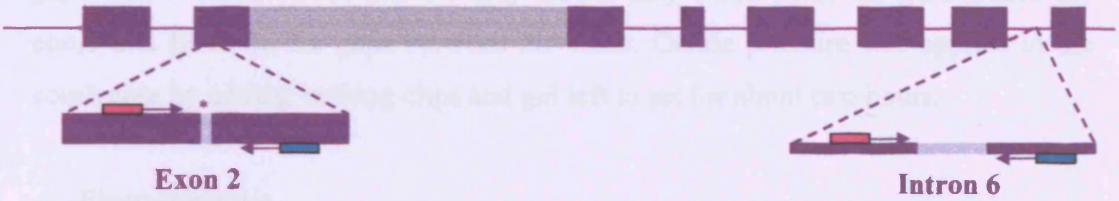


Figure 2.6: Schematic representation of the PCRs for the two *MUC1* gene polymorphisms flanking the TR (grey box). The g/a SNP is present in exon 2 (•), just before the TR. The CA microsatellite is positioned in intron 6, spanning from 11 to 14 repeats.

2.3.7.2 Detection of the products

The samples were genotyped using an ALF express DNA Sequencer (Pharmacia Biotech), designed to detect fluorescently labelled DNA molecules separated by electrophoresis on vertical polyacrylamide gels. Cy5 labelled DNA is detected by a laser positioned near the bottom of the gel.

Preparation of the gel

Before preparing the gel mixture, the plates, spacers and comb were thoroughly cleaned, first with tap water and with a non-scratch brush, and then with Isopropanol (AnalaR, BDH, Poole, Dorset, UK). To help with the formation of neat wells, Bind-Silane (γ-methacryloxy-propyl-trimethoxysilane) (Amersham Pharmacia Biotech) was applied to the top 2cm of each of the plates and the excess removed by wiping with a tissue containing a small amount of isopropanol. Once all the pieces were dry, the spacers were placed on the sides of the bottom glass plate and covered with the top glass plate and the two clipped together, making sure that the spacers did not move. Finally the comb was inserted between the top of the two plates and held in place with bulldog clips.

The gel was prepared by pouring into a 50ml cylinder, 40ml of SequaGel-6, a 19:1 acrylamide/bis-acrylamide solution, and 10ml of SequaGel Complete buffer Reagent, containing TBE and TEMED (both solution are from National

Diagnostics). The mixture was then transferred to a squeezy bottle and 400µl of 10% Ammonium Persulfate (AMPS) (Bio-Rad, Bio-Rad House, Hemel Hempstead, Hertfordshire, UK) was added and mixed gently. The mixture was poured into the gap at the bottom of the plates using the squeezy bottle until the gel reached the comb and filled all the gaps between the wells. Gentle pressure was applied to the comb area by adding bulldog clips and gel left to set for about two hours.

Electrophoresis

Once set, the gel was inserted into the ALFexpress machine, filling the top and bottom buffer tanks with 1x TBE. The comb was removed and the wells rinsed with TBE using a syringe. Using the AM v3.0 software (Amersham Pharmacia Biotech) the run settings were adjusted to 1500V, 38mA, 25W, for 400 minutes with a sampling interval of 1 second. A pre run of approximately 10 minutes was used, to prewarm the gel to 45°C and to stabilise the laser value, ideally with a value above 600. During the pre-run, the samples were denatured for 5 min at 95°C, and were kept in ice during the loading of the gel to avoid reannealing.

After the pre-run, the program was paused and the wells were rinsed again with TBE. The samples were then loaded into the wells. A positive control containing a mix of 2 samples (with different genotypes) and the 100+300bp markers was loaded in the first well, to allow comparison of the genotypes with the control. After loading the samples the run was resumed.

2.3.7.3 Analysis of the data

Once the run was completed, the data was analysed using the 'Fragment Manager' software (Amersham Pharmacia Biotech). The 100 and 300 base pair markers allow the sequences to be aligned and therefore the alleles to be accurately sized, despite the CA microsatellite alleles differing by only 2bp. Figure 2.7 demonstrates an example of the output of an ALF run.

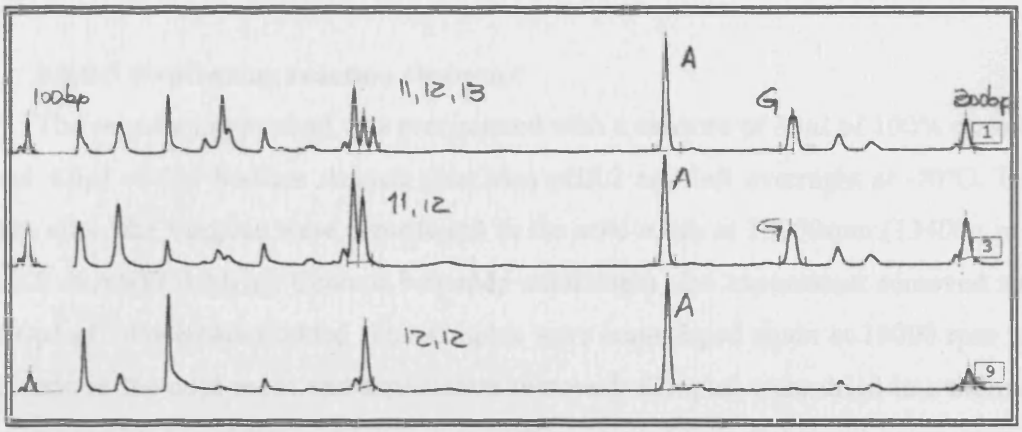


Figure 2.7: Representation of the output file of an ALF express run using Fragment Manager software. The first lane shows a control sample, that results from the combination of 2 samples with different genotypes. The second and third lane show two samples, one heterozygous for both markers (11,12 AG) and the other homozygous for both markers (12,12 AA).

2.3.8 Sequencing

2.3.8.1 PCR-product clean-up

10μl of PCR reaction product was mixed in a 1:1:2 ratio of PCR product: distilled water: lab-made PCR clean-up solution (composition in the 'Buffers and Solutions' section) and left for 10 minutes at room temperature. The mixture was subsequently spun at 13000 rpm (13400g) for 15 minutes (MSE (SANYO) Micro Centaur benchtop centrifuge), the supernatant (SN) discarded with a pipette and 100μl of 100% ethanol added. The tube was spun again at 13000 rpm for 5 minutes and supernatant removed with a pipette. After repeating the ethanol precipitation procedure, the tubes were left to dry at 60°C for about 20 minutes. The pellets were resuspended in 10μl of distilled water and left overnight at 4°C.

2.3.8.2 Sequencing reaction

5μl of the purified PCR product was added to 2.4pmol of primer, 1μl of Big Dye Terminator mix v3.1 (PE Applied Biosystems), 1μl of lab-made sequencing buffer (composition in the 'Buffers and Solutions' section) and water, to a total volume of

15µl (per reaction). The sequencing reaction was performed in a GeneAmp- PCR System 9700 thermal cycler, at 96°C for 45 seconds, 50°C for 30 seconds and 60°C for 4 minutes for 25 cycles.

2.3.8.3 Sequencing reaction clean-up:

The sequencing product was precipitated with a mixture of 30µl of 100% ethanol and 4.5µl of 3M Sodium Acetate (NaOAc) pH5.2 and left overnight at -70°C. The day after, the samples were centrifuged in the cold room at 13000rpm (13400g in a MSE (SANYO) Micro Centaur benchtop centrifuge), the supernatant removed and 100µl of 70% ethanol added. The samples were centrifuged again at 13000 rpm for 15 min in the cold room and supernatant removed. Samples were dried in a thermal cycler for 20 min at 60°C, and stored at -20°C protected from light until sequencing.

Just before sequencing, the samples were resuspended in a sequencing loading buffer containing deionised formamide and blue dextran, in a 3:1 ratio. The samples were kept in the dark wrapped with aluminium foil at 4°C.

2.3.8.4 Sequencing gel:

Before preparing the gel, the sequencing plates were washed with 1% Alconox (Alconox powdered precision cleaner, PC International Limited, Abington, Cambridge, UK) and rinsed well, first with tap water and then with Millipore water. The spacers and comb were carefully cleaned with Millipore water using damp tissue. Once dry, the plates and spacers were assembled.

The gel was prepared in a beaker by mixing 9g of Urea with 2.85ml of 40% Accugel 29:1 acrylamide: bisacrylamide (National Diagnostics, Hessle, East Riding of Yorkshire, UK), 0.25g Amberlite (PE Applied Biosystems) and 12.5ml of Millipore water. The mixture was heated until the urea crystals were dissolved completely and then filtered through Whatman 1 cellulose filter paper (7 cm) in a vacuum system. The recovered product was transferred to a 25ml cylinder and 2.5ml of TBE was added and topped up with Millipore water to 25ml. 17.5µl of TEMED and 125µl of 10% Ammonium Persulphate (AMPS) were pipetted into the mixture, which was decanted back to the beaker and stirred quickly and gently. The mixture was quickly poured into the assembled plates with a syringe and the square side of the comb inserted straight afterwards and a weight put on the top of the plates near the comb area. The gel was left to set for at least two hours.

2.3.8.5 Electrophoresis:

Electrophoresis was either performed in an ABI PRISM 377 DNA Sequencer (Applied Biosystems) using the Data collection software version 2.5 or on a capillary system as indicated below.

Before the run, the comb was removed from the polymerised gel and the area of the gel formed by the comb cleaned carefully with distilled water to remove any unset gel. The comb was then cleaned and replaced in the comb area with the 36 well 'shark's-tooth' side towards the gel and sunk about 2 mm into it. The area of the plates through which the laser passes was also thoroughly cleaned with ethanol to make sure no dust or dirt interfered with the laser readings. The plate was placed into the machine and the buffer tanks filled with 1x TBE.

A plate check was performed using the software to check the level of background fluorescence and a pre-run was performed at 1Kv and 35 minutes while the sequencer reached 51°C. Meanwhile, the samples were denatured at 95°C. Once the plates had reached 51°C, the samples were loaded into the gel in every other well and then run into the gel for 2 minutes. The intervening wells were then loaded. Once all the samples were loaded the 7 hour run was started at 1.68Kv, 50mA at 51°C, with a laser detection of 1200 scans per hour.

- Capillary System

Many of the samples were analysed in an ABI PRIMS 3700 DNA Analyser (Applied Biosystems). The BigDye Terminator v 1.1 Cycle Sequencing Kit (Applied Biosystems) was used for the sequencing reaction.

2.3.8.6 Sequences analysis:

The sequences obtained were analysed using Sequence Analysis v 3.3 and Sequence Navigator v 1.0.1, software programs from ABI (Applied Biosystems). Sequences were aligned using Sequencer Navigator and exported to Sequence analysis to analyse the sequences or using Chromas v2.0 and ChromasPro v1.3 software (free trial version, http://www.technelysium.com.au).

2.3.9 Other SNPs from database- markers flanking the gene

Single Nucleotide polymorphisms outside the *MUC1*gene were found using the NCBI SNP database (NCBI dbSNP, http://www.ncbi.nlm.nih.gov/) and HapMap (http://www.hapmap.org/). The strategies used are described in detail in Chapter 5. 8 new SNPs were selected from the databases and assays designed to type the polymorphisms using restriction enzyme digestion.

The SNPs selected and respective rs accession numbers are:

- THBS3 P1- rs2075571
- THBS3 P2- rs2066981
- PCR5 Taqαl- rs2070803
- PCR5 BsmF I- rs4971101
- 497- rs4971100
- MTX1p- rs1045253
- KCP2- rs4971088
- EFNA1- rs9297

2.3.9.1 PCR reaction

The PCRs for the SNPs selected were performed as previously described in PCR general methods, but with varying reaction volumes. All PCRs were done in 96 well plates, using the GeneAmp- PCR System 9700 or the MJ Research PTC-200 Peltier Thermal Cyclers. Table 2.8 shows a summary of the primer sequences used for each PCR, the reaction volume, annealing temperature, extension time, number of cycles, extra reagents and product size.

4μl of each PCR product was loaded onto a 96 well 'stretch' gel (Electro-fast Stretch gel systems, ABgene) with 3μl of loading buffer. Each row of the gel also contained a molecular weight marker to allow size comparison of the products (lab-made). The gel was run at 100V and 50mA and visualised under U.V. light.

2.3.9.2 Restriction enzyme digestion

Once the size of the PCR products was confirmed, the samples were digested with a restriction endonuclease following the company guidelines. Table 2.9 shows a summary of the restriction assays. The restriction enzyme digestions at 60°C or 65°C were carried out in a thermal cycler for 16 hours. An incubator maintained at 37°C was used for the rest of the digestions, and samples were left also for 16 hours.

	Primer name	**Primer sequence 5'-3'**	**Reaction volume**	**Annealing temperature**	**Extension time**	**Number of cycles**	**Glycerol 50%(v/v)**	**Product size**
PCR5	**M1 3'utrF1**	GGAGATGTGAGGAGGAGGTG	25µl	64°C	1 min	30	1:20	711 bp
	M1 3'utrR1	CTGTGCTGGGTGTGGTAGG						
THBS3 P1	**THBS3AltF1**	GGCTGACCAAGAGATACTGGTC	10µl	60°C	1 min	30	-----	427 bp
	THBS3rev1	GGTGAGAGCTTCTGAACATCC						
THBS3 P2	**THBS3for2**	GAGGTCAGAGCCCAGAAGG	20µl	65°C	1 min	30	-----	291 bp
	THBS3rev2	CAACTGCCTTGTGCTCCTG						
497	**4971100For**	CAACTCCCTATTTAACACCCTC	12µl	57°C	30 sec	30	-----	223 bp
	4971100Rev	AGAGAAAGGATGTTGGGAAG						
KCP2	**KCP2For**	TTTTTACCAGCTCTGCCCGA	12µl	57°C	30 sec	30	-----	217 bp
	KCP2Rev	CAGGGAGCAAGGCCTTCA						
EFNA1	**EFNA1For**	GGCTGAGAGCCAGTACAAATA	12µl	57°C	30 sec	30	-----	205 bp
	EFNA1Rev	CTTTAAGGCCAGGTGTGGTA						

Table 2.8: Summary of the PCR conditions used for the non *MUC1* PCRs. In the table the PCR name and respective primers are shown, as well as PCR reaction volume, PCR cycling conditions, extra reagents and final PCR product size.

	Product size	Restriction enzyme	Enzyme units	Buffer	Extra reagents	Amount of PCR product	Total volume	Incubation temperature	Nucleotide change	Fragments generated
PCR5	711 bp	Taq$^{\alpha}$ I (1)	0.5 U	Buffer B	-----	2µl	10µl	65°C	C/t	464bp/247bp
PCR5	711 bp	Bsmf I (2)	1 U	NEB4	1% BSA	2µl	10µl	65°C	C/t	463bp/248bp/ 164bp/84bp
THBS3 P1	427 bp	HinP1 I (2)	1 U	NEB2	-----	3µl	15µl	37°C	a/G	261bp/164bp/ 145bp/116bp
THBS3 P2	291 bp	BstE II (2)	5 U	NEB3	-----	3µl	15µl	60°C	c/T	216bp/75bp
497	223 bp	Dde I (3)	1 U	REact3	-----	2µl	10µl	37°C	a/G	178bp/45bp
KCP2	217 bp	EcoR V (2)	5 U	NEB3	1% BSA	2µl	10µl	37°C	a/T	197bp/20bp
EFNA1	205 bp	Acc I (2)	1 U	NEB4	-----	2µl	10µl	37°C	A/g	185bp/20bp

Table 2.9: Summary of the restriction enzyme digestion conditions used for the non *MUC1* PCRs. In the table the PCR name is shown together with the endonucleases, buffers, extra reagents, assay volume and incubation temperature. The polymorphism is also shown with the nucleotide that creates the restriction site marked in capital letter in bold. The last column shows the sizes of the possible fragments generated. (1) Boehringer Mannheim GmbH, now part of Roche Applied Science; (2) New England BioLabs; (3) GIBCO BRL, now part of Invitrogen.

2.4 Statistical analysis

2.4.1 Hardy Weinberg Equilibrium:

The Hardy-Weinberg Equilibrium (HWE) derives from a mathematical model that predicts that gene pool frequencies are intrinsically stable throughout generations, despite the constant occurrence of evolutionary pressure. A deviation from HWE can occur under several conditions: non-random mating, inbreeding, selection and genetic drift (that usually occurs due to small population size). In a specific research project, it may deviate from equilibrium due to sampling errors, population stratification, technical problems (genotyping errors) and small sample sizes.

A population is said to be in Hardy-Weinberg Equilibrium for a particular locus when the genotypic frequencies observed in a given population are not significantly different from the expected frequencies.

$$\boxed{p^2 + q^2 + 2pq = 1}$$

The equation presented in the box above represents the Hardy-Weinberg principle for a bi-allelic locus, where p^2 represents the frequency of homozygous for one of the alleles, q^2 represents the frequency of homozygous for the other allele and 2pq represents the frequency of heterozygous, in a given population. Significance of deviation from HWE is usually determined using a chi-square test, but in this thesis a kind of exact test implemented in Arlequin was used.

2.4.2 Tests of independence (Non-parametric tests):

2.4.2.1 Chi-square (χ^2) test:

χ^2 tests for goodness of fit, ie tests how well the observed frequency distribution fits the expected frequency distribution.

$$\boxed{\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}}$$

The χ^2 significance is then determined in the Pearson χ^2 contingency table at 95% confidence interval with (r-1)(c-1) degrees of freedom, where r is the number of rows and c the number of columns of the distribution.

In the context of HWE it uses the observed genotype frequencies to calculate the allele frequencies of a certain locus and in a way calculate the expected frequencies for the genotype distribution.

2.4.2.2 Fisher's exact test:

The Fisher's exact test calculates an exact probability value for the relationship between 2 variables with 2 variants each, in a 2 by 2 contingency table.

		Locus 2		
	alleles	1	2	Totals
Locus 1	1	A	B	A + B
	2	C	D	C + D
	Totals	A + C	B + D	A + B + C + D = N

The p value is calculated:

$$\boxed{p=\frac{(A+B)!\ (C+D)!\ (A+C)!\ (B+C)!}{N!\ A!\ B!\ CD!}}$$

The exact probability is calculated for the observed table and all possible tables that deviate more from the expected than the one actually observed, and these are summed, as well as those tables that deviate in the opposite direction (for a two tailed test). The table is said to be significantly deviated from expectation when $p \leq 0.05$.

2.4.2.3 Mann-Whitney U test:

Mann Whitney tests whether two independent samples come from identical populations or whether these populations have unequal distributions. It is started by calculation of the number of observation in each group, where:

- n1 is the number of observation in the larger group
- n2 is the number of observations in the smaller group

The two sample groups are ranked together in an increasing order of magnitude as a single series. After the ranking, the two sample groups are separated, each observation maintaining the previously given rank. For each group, the sum of the ranks is made and a value assigned for each group:

- R1 is the sum of the ranks of population 1 with n1 number of observations
- R2 is the sum of the ranks of population 2 with n2 number of observations

$$\boxed{U = n1n2 \frac{n1(n1+1)}{2} - R1}$$

or

$$\boxed{U = n1n2 \frac{n2(n2+1)}{2} - R2}$$

Calculation of U statistics:

If U exceeds the critical value for U at the specified significance level, then the two groups are said to have distributions that are significantly different.

2.4.3 Haplotype prediction:

2.4.3.1 Expectation-maximisation algorithm (EM):

The Expectation-maximisation algorithm (EM) is an iterative process that estimates haplotype frequencies through successive steps (Excoffier and Slatkin, 1995). It assumes that the observed genotypes are in HWE.

Based on the observed genotype frequencies:

		Marker B			
	Genotypes	BB	Bb	bb	
Marker A	AA	AA/BB	AA/Bb	AA/bb	AA
	Aa	Aa/BB	Aa/Bb	Aa/bb	Aa
	aa	aa/BB	aa/Bb	aa/bb	aa
		BB	Bb	bb	N

it computes the haplotypes as follows:

		Marker B			
	Haplotypes	BB	Bb	bb	
Marker A	AA	AB-AB	AB-Ab	Ab-Ab	
	Aa	AB-aB	AB-ab/ Ab-aB	Ab-ab	
	aa	aB-aB	aB-ab	ab-ab	

All haplotype frequencies are easily calculated, except for the double heterozygous AaBb that can be either AB-ab or Ab-aB, where the phase is ambiguous. In the EM algorithm the first step is to estimate the initial haplotype frequencies of the double heterozygotes based on the allele frequencies (expectation step). The initial haplotype frequencies estimated are then used to estimate the proportion of AB-ab in the AB-ab/Ab-aB heterozygotes. Once this value is estimated it is used to re-estimate the AB-ab proportion in the heterozygotes and compute new haplotype frequencies (maximisation step). The iterative process repeats until the haplotype distribution stabilises.

The EM algorithm makes no assumption about linkage Disequilibrium between loci.

2.4.3.2 Bayesian algorithm (based on Bayes theorem):

The Bayesian algorithm to infer haplotype distribution is described in Stephens *et al* (Stephens et al., 2001). It is an inference method that adjusts the estimates of a population parameter in light of prior knowledge.

It uses a variant of the Markov chain-Monte Carlo (MCMC) algorithm. It estimates all unambiguous haplotypes and then repeatedly chooses an ambiguous individual and estimates the probability under the assumption that the other haplotypes are correct. This process is repeated until the haplotype probabilities become stationary. It assumes the most simple relationship between the haplotypes.

2.4.4 Linkage Disequilibrium measurement:

Linkage Disequilibrium is the non-random association of alleles at linked loci. There are several coefficients used to measure LD:

2.4.4.1 D and D` coefficients:

D measures the observed frequency of co-occurrence of an allele A_1 of locus A and an allele B_1 of locus B on the same chromosome (P_{11}) and the expected frequency of co-occurrence under Linkage Disequilibrium

$$D= P_{11}-p_1q_1$$

where p_1 and q_1 are the allele frequencies of A_1 and B_1, respectively. If D differs significantly from zero, the loci are in LD.

Because D is dependent on allele frequencies in the population,

its maximum value is $D_{max}= \min (p_1q_2, p_2q_1)$

its minimum value is $D_{min}= \max (p_1q_1, p_2q_2)$

and D can be scaled to

$$\boxed{D`=\frac{D}{D_{max}}}$$

2.4.4.2 χ^2 and associated p value:

Determined as shown above for tests of independence.

2.4.5 Exact test of population differentiation (Goudet et al., 1996;Raymond and Rousset, 1995)

The Exact test of population differentiation produces a contingency table similar to Fisher's exact test, but with a variable number of cell (columns and rows). All potential contingency tables are generated using a Markov chain, and calculated the probability of observing each table. For haplotypic data, the table is built using haplotype frequencies (Raymond and Rousset, 1995), and for genotypic data with unknown gametic phase, the table is built based on the genotype frequencies (Goudet et al., 1996).

2.4.6 Software packages

Alrequin: http://lgb.unige.ch/arlequin/, version 2.0

PHASE2.0 and Phamily:
http://www.stat.washington.edu/stephens/phase.html
http://archimedes.well.ox.ac.uk/cgi-bin/pise/phamily1
Websites no longer available.

GOLD: http://www.sph.umich.edu/csg/abecasis/GOLD/index.html

LDmax: part of GOLD software package- uses the EM to estimate haplotype frequencies (Excoffier and Slatkin, 1995).

CHAPTER 3

MUC1 AND GASTRIC DISEASE IN UK
-1996/1997 COHORT STUDY-

Introduction

As described in detail in Chapter 1, there is evidence that mucins, as part of the gastric mucosa/external environment interface, play a role in the competence of *Helicobacter pylori* to establish in the stomach. Several groups reported interactions between the MUC5AC mucin and the bacteria and there was also involvement of another non-MUC5AC mucin, which was almost certainly MUC1 (Linden et al., 2004). The blood group-related antigen Le^b is also a ligand for *H. pylori*, and this is potentially expressed on the carbohydrate chains of MUC1 (Boren et al., 1993). Furthermore, dramatic changes were observed in the pattern of expression of *MUC1* in *H. pylori* gastritis (Vinall et al., 2002).

Studies in our lab and other labs reported a correlation between *MUC1* tandem repeat (TR) size and presence of gastric disease, namely *H. pylori* gastritis and gastric cancer (Carvalho et al., 1997;Silva et al., 2001;Silva et al., 2003;Vinall et al., 2002) in different populations. All reports include an over-representation of small alleles in the disease groups when compared with the control groups. It was not however clear whether this was due to the length of the repeat domain itself, the internal sequence of the TR array or perhaps due to linked polymorphisms elsewhere in the gene which affect the properties of MUC1 or level of expression.

At the outset of this thesis work, data was available for 3 polymorphic markers in the *MUC1* gene for the published gastritis group and controls (Vinall et al., 2002) and various other disease and control groups (Joanna Fowler and Lynne Vinall, unpublished data), though some of this was incomplete, or unchecked, and not fully analysed; I contributed to this work by completing the data collection and furthering the analysis by studying Linkage Disequilibrium (LD) and haplotype distribution.

3.1 Samples collected

3.1.1 First small gastritis cohort

During the period between 1996 and 1997, 181 individuals needing gastrointestinal endoscopy examination in the Middlesex hospital, London, were recruited for the first gastrointestinal study. All individuals filled in a questionnaire. Blood samples as well as biopsies (either colon or stomach) for histology were collected from each individual.

Gastric samples were collected from 59 individuals of Northern European extraction who had had upper gastrointestinal tract endoscopy examination and from whom blood samples were also available. DNA was extracted from the blood samples, and biopsies were analysed and scored for possible lesions in the gastrointestinal tract, specifically in the stomach antrum, and presence of *Helicobacter pylori*. Only Northern European individuals were included, to avoid the problem of population stratification, which might mask significant results or produce false positives.

7 of the individuals had evidence of gastritis but no *H. pylori* infection and were excluded from the comparisons because of the small number of cases, and lack of knowledge as to whether or not those individuals had *H. pylori* eradication treatment before.

3.1.2 Oesophageal cancer and other GI disease

51 colon samples from Northern European individuals who had had a colonoscopy examination and were diagnosed according to presence or absence of colon disease were used as part of a control population c ollected in the same GI clinic, Middlesex Hospital. The other group included in this category was a cohort of individuals with oesophageal disease (n=101), collected by Dr Laurence Lovat. These 2 groups were pooled into the same category, named Middlesex hospital- other diseases, and for which no gastritis information and *Helicobacter pylori* infection status is available. Each of the groups is characterised in the Material and Methods section. From this point onwards, this group will be denominated 'Mdx- others'.

3.1.3 Non-hospital controls

Finally, a fourth group of individuals included in this analysis as a control population belongs to the 1946 cohort, described in more detail in the Materials and Methods section. This was composed of 349 DNA samples collected from individuals living in the London area, from whom no information on gastric disease or *Helicobacter pylori* infection status is available. These are identified below as the 'London area' group, representing a random London population.

3.1.4 Characterisation of the samples from the first small gastritis cohort

The table 3.1 summarises the number of females and males in each of the two subgroups, normal and *H. pylori* gastritis, independently of the category. The table shows that there are more males in the affected group and more females in the controls GI clinic controls.

	Males		Females	
	Number of individuals	**% of individuals**	**Number of individuals**	**% of individuals**
H. pylori **gastritis**	14	56	9	33
Normal	11	44	18	67
Total cohort	25	-	27	-

Table 3.1: Distribution of males and females in each of the disease groups and the total cohort in terms of number of individuals and percentages. Though there are more females overall, 51.9%, the majority of *H. pylori* gastritis population is male (60.9%). In the normal GI clinic controls the majority of the individuals are females (62.1%).

In relation to the mean age of the individuals in these two subgroups, there is not a significant difference between sexes, with the mean age of 53 for the *H. pylori* gastritis patients (ranging from 25 to 81 years old) and 46 for the normal controls (ranging from 17 to 75). The slightly older age of the gastritis patients may reflect the age-related occurrence of the disease where prolonged colonisation of the bacteria, together with other factors, triggers a series of mechanisms that over time will lead to gastritis, and symptoms severe enough to require a visit to the doctor.

3.2 *MUC1* genotype and haplotype distribution for the 3 polymorphic loci within the gene

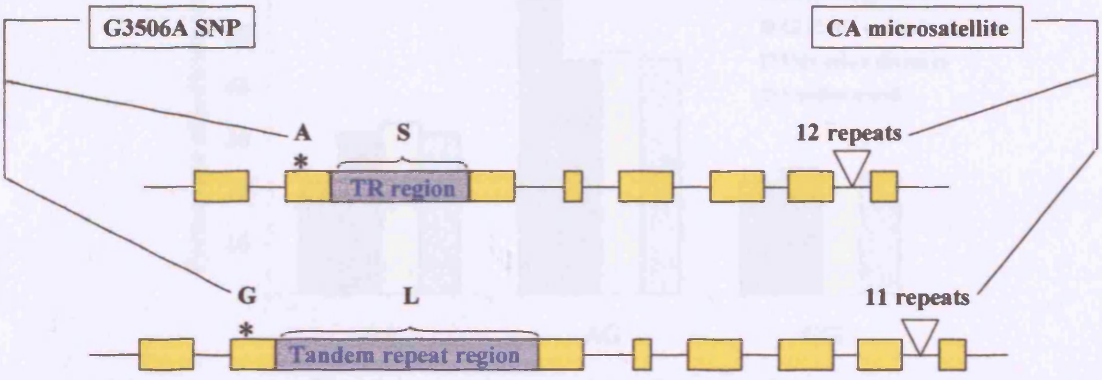


Figure 3.2: Schematic representation of *MUC1* gene. Yellow boxes represent the 7 gene exons. In exon 2 the grey boxes are shown with different lengths to illustrate the variable number of tandem repeats (VNTR) polymorphism, and the g/a SNP polymorphism is also shown in exon 2 marked with an asterisk (*). In intron 6, the CA microsatellite is marked with open arrowheads and two of the possible number of repeats, 11 and 12.

3.2.1 *MUC1* genotypes and allele distribution

Figure 3.2 shows a schematic representation of the *MUC1* gene and polymorphisms within it. One large domain composed of repeats arrayed in tandem makes up most of exon 2 and is flanked by a single nucleotide polymorphism (SNP) and a microsatellite. The TR array has an extremely variable number of repeats, which in the European population exhibits a bimodal distribution (see Figure 4.2 in Chapter 4).

Figure 3.3 (A, B and C) shows the genotype distribution of the three *MUC1* markers: G3506A SNP, tandem repeat polymorphism (TR) and CA microsatellite, respectively, in the *Helicobacter pylori* gastritis group and the three control groups: GI clinic controls, Middlesex Hospital- other diseases and London area. The corresponding allele frequencies of the groups are show in Table 3.4 A, B and C respectively.

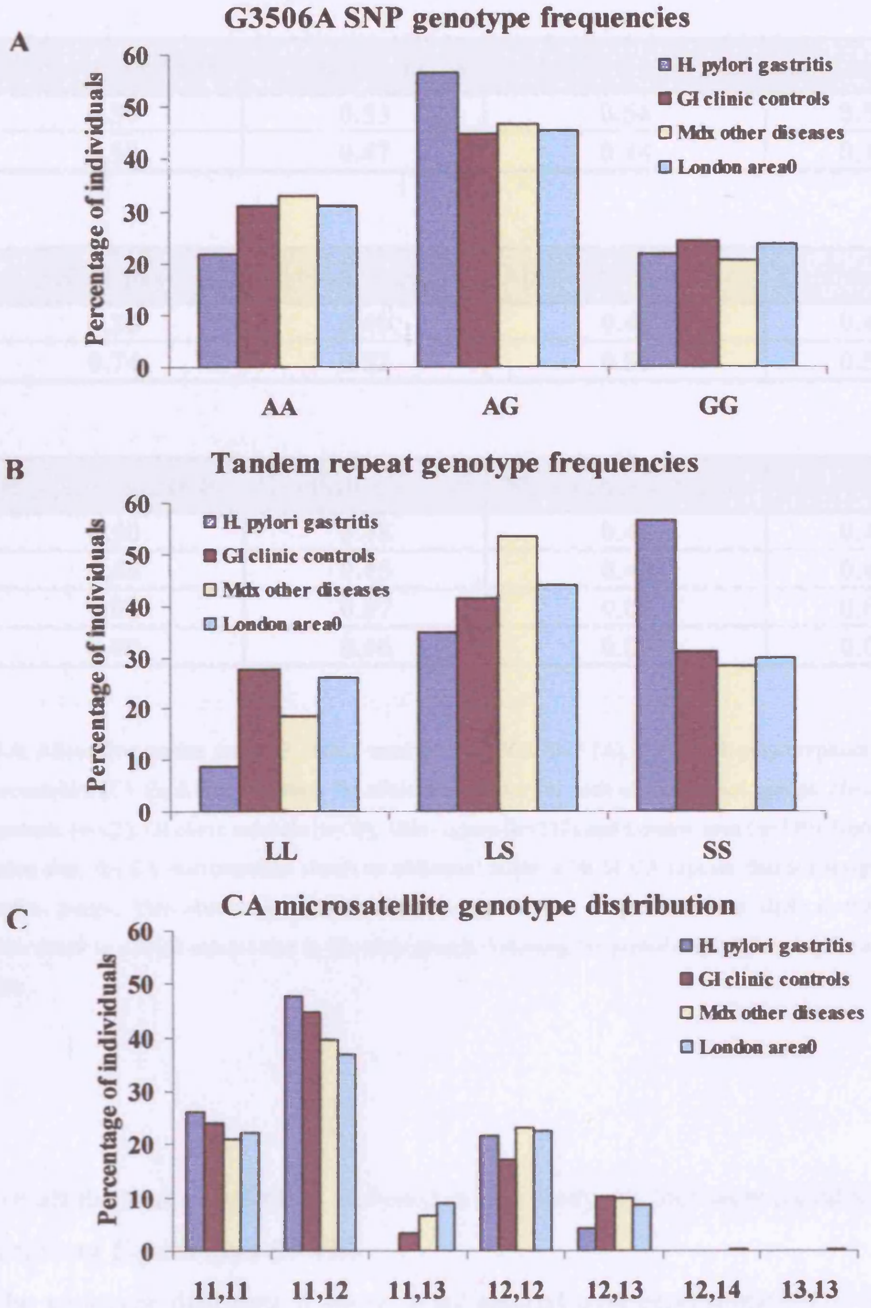


Figure 3.3: Genotype frequencies for 3 *MUC1* markers: G3506A SNP (A), TR length polymorphism (B) and CA microsatellite (C). Each graph displays the genotype frequencies for each of the groups studied: *Helicobacter pylori* gastritis (n=23), GI clinic controls (n=29), Mdx- others (n=152) and London area (n=349).

A

	H. pylori gastritis	GI clinic controls	Mdx other diseases	London area0
A	0.50	0.53	0.56	0.54
G	0.50	0.47	0.44	0.46

B

	H. pylori gastritis	GI clinic controls	Mdx other diseases	London area0
L	0.26	0.48	0.45	0.48
S	0.74	0.52	0.55	0.52

C

	H. pylori gastritis	GI clinic controls	Mdx other diseases	London area0
11	0.50	0.48	0.44	0.45
12	0.48	0.45	0.48	0.46
13	0.02	0.07	0.08	0.09
14	0.00	0.00	0.00	0.00

Table 3.4: Allele frequencies for the 3 *MUC1* markers: G3506A SNP (A), TR length polymorphism (B) and CA microsatellite (C). Each table displays the allele frequencies for each of the studied groups: *Helicobacter pylori* gastritis (n=23), GI clinic controls (n=29), Mdx- others (n=152) and London area (n=349). Note that for the London area, the CA microsatellite shows an additional allele, with 14 CA repeats, that is not represented in the other groups. This observation can be explained by the low frequency of the allele in the general population allied to a small sample size in the other groups, lowering the probability of showing the allele just by chance.

For all the four populations included in this study, all loci were found to be in Hardy-Weinberg Equilibrium (HWE).

The genotype distribution shows a substantial over-representation of the SS genotype in the *Helicobacter pylori* gastritis group (Figure 3.3(B)) not only as compared with the GI clinic controls (Vinall et al., 2002) but also with all other controls. Contingency tables and Fisher's exact test calculations show that the tandem repeat allele distribution of the *Helicobacter pylori* gastritis group is not only significantly different from the GI clinic controls, but also from Mdx-other disease and the London area, with 2-tailed p-values of 0.026, 0.016 and 0.0036, respectively (see Table 3.4(B)).

On the other hand, other markers are more similar between the groups, except for a small and non-significant increase in AG heterozygous for the SNP polymorphism in the gastritis groups, when compared with the others.

The lack of difference between the flanking markers in the various groups in comparison with the TR repeat length suggested that an examination of the haplotypic distribution would be worthwhile.

3.2.2 *MUC1* haplotypes

To follow these observations further, *MUC1* haplotypes for the 3 polymorphisms were inferred. Two different software packages were used; Arlequin, that uses Expectation Maximisation EM algorithm to infer haplotype frequencies, and PHASE, that uses a Bayesian method (involving conditional probabilities) to infer the haplotypes of individual people with associated probabilities (Stephens et al., 2001).

The haplotypes inferred using the two methods are displayed in figure 3.5 (A) and (B). Observing the two bar charts, it is easily seen that the two methods generate similar haplotypes and the respective frequencies are also similar.

Using the Arlequin statistical package which gives a measure of Linkage Disequilibrium (LD), it was also found that for most groups, the 3 markers are in LD with each other generating two major haplotypes, A S 12 and G L 11, which together account for nearly 80% of the chromosomes in most groups (see Figure 3.2). Interestingly, in the gastritis population the strong LD observed in the other groups is broken. Though the g/a SNP remains in LD with the Tandem Repeat domain and the CA microsatellite, the LD between the TR and the CA microsatellite is disrupted (see table 3.6).

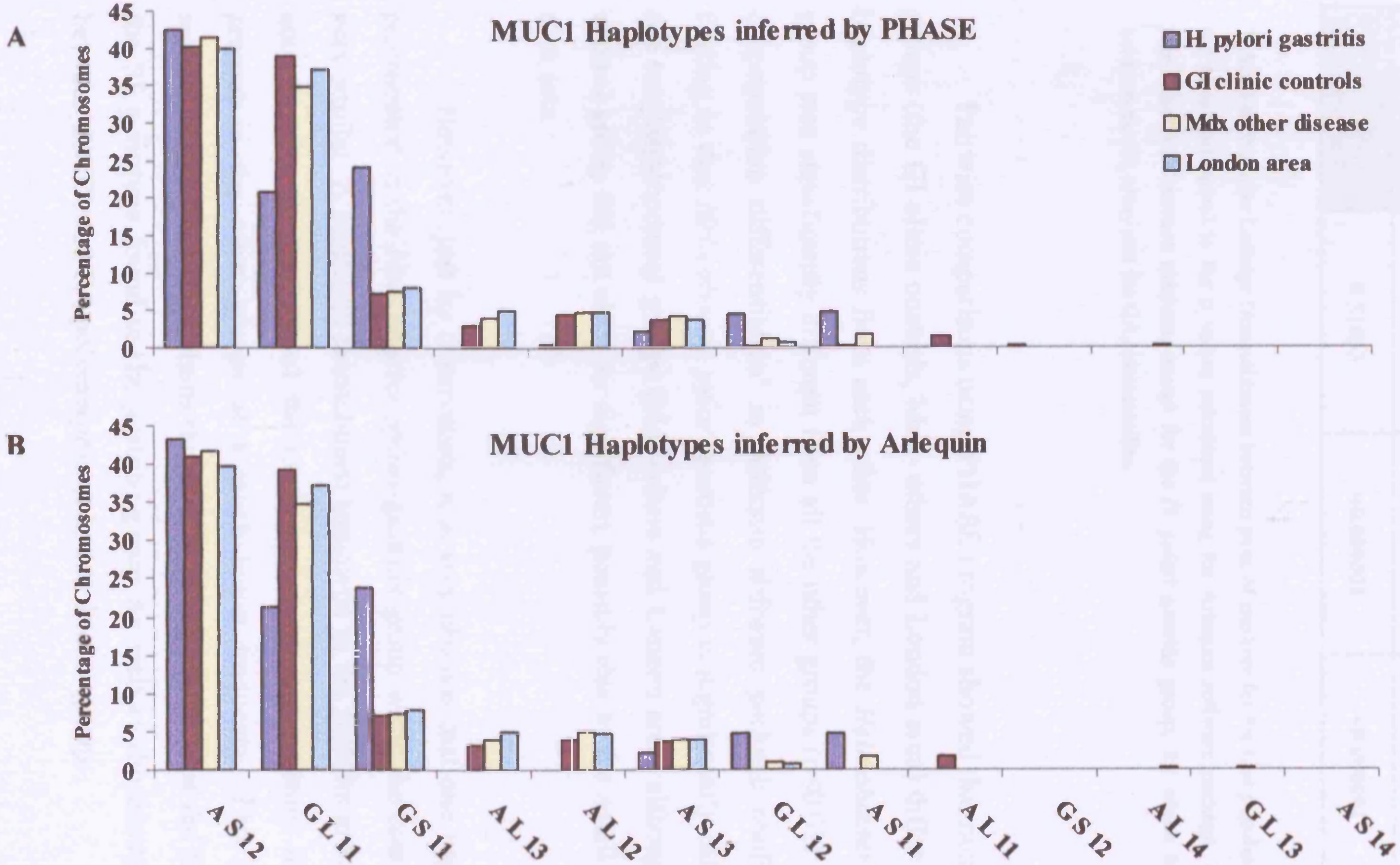


Figure 3.5: *MUC1* haplotype distribution in 4 different groups: *Helicobacter pylori* gastritis, GI clinic controls, Mdx- others and London area. The haplotypes were inferred using 2 different statistical methods: (A) a Bayesian method and (B) a maximum-likelihood method, the EM algorithm.

	Helicobacter pylori gastritis	GI clinic controls	Mdx- other diseases	London area
G3506A SNP and TR domain	0.00485	0.0001	<0.000001	<0.000001
G3506A SNP and CA microsatellite	<0.000001	<0.000001	<0.000001	<0.000001
TR domain and CA microsatellite	0.51653	<0.000001	<0.000001	<0.000001

Table 3.6: Pairwise Linkage Disequilibrium between pairs of markers for the four populations. The values on the table correspond to the p values calculated using the Arlequin software package. All groups show a significant LD between markers, except for the *H. pylori* gastritis group, for which the LD is disrupted between the TR array and the CA microsatellite.

Pairwise comparisons using PHASE program showed that none of the control groups (the GI clinic controls, Mdx- others and London area) differ significantly in haplotype distributions from each other. However, the *Helicobacter pylori* gastritis group was significantly different from all the other groups (p<0.02). The 'exact test of population differentiation' in Arlequin software package confirms this basic finding in that *Helicobacter pylori* gastritis group is significantly different from the the two larger control groups (Mdx-others and London area) although the GI clinic control group did not show as significant, possibly due to the small size of the two data sets.

However, just by observation, it is very obvious that one haplotype is over-represented in the *Helicobacter pylori* gastritis group while the control groups look very similar. A particular recombinant haplotype in the gastritis group- G S 11- that accounts for almost 25% of the total number of chromosomes in this group, is present in the other groups at a much lower frequency. This high frequency recombinant haplotype explains the much higher frequency of the S allele and also the SS genotype found in the gastritis group, as well as the disruption of linkage between the TR and CA polymorphisms in this disease group.

3.3 MVR analysis

Previous work from this lab, and elsewhere, has described considerable variability between repeat sequence variation in the *MUC1* tandem repeats (Engelmann et al., 2001;Fowler et al., 2003). The main variation studied by us was the change in sequence which alters the more common peptide motif in the repeat unit from PDTR to PESR. The results of this study, which I contributed to by data completion, haplotype assignment and data analysis are published (Fowler et al., 2003) and a copy of the paper which includes relevant methods is appended to this thesis. Alleles were separated by PCR amplification across the TR array and electrophoresis to separate the bands, then used separately for minisatellite variant repeat analysis (MVR). Complete maps were obtained from 119 alleles as well as 30 partial maps. The MVR maps of 90 short alleles and 29 long alleles were compared, even though the majority of the long alleles were incomplete maps. Because of this gap in the MVR data for the long alleles, no statistical comparison was done. However, the analyses suggested that there were more PESR repeats in the long alleles and that the block structure was a little different and less regular (Fowler et al., 2003). The observation of over-representation of one haplotype GS11 in the gastritis cohort, which was subsequently also seen in the patients with gastric cancer (see Chapter 4) led us to consider whether these alleles (GS11) might differ from the standard AS12 alleles and perhaps resemble long alleles in terms of block structure and also have a greater PESR content. To make comparison of block structure was not possible with the limited number of these variant alleles, but a simple comparison of the proportion of PESR content was attempted.

To do this, only Northern European chromosomes were tested, and the *MUC1* allele maps were classified into GS and GA haplotype by a combination of methods. In some cases haplotype assignment was unambiguous (ie where at least one of the two polymorphism was homozygous). In some cases it was tested directly by allele specific PCR across the TR region and in a few cases phase was inferred using haplotype inference software. It was possible to assign haplotype with reasonable certainty for a total of 80 NE chromosomes (70 AS and 10 GS).

The number of each type of repeat was counted for each allele and expressed as a proportion of the total repeat number. Comparison of the two groups of alleles shows that although there were indeed slightly more PESR repeats this was not statistically significant. It should be emphasised that only 38 alleles from patients with gastritis were fully mapped. Attempts to continue this work and to analyse other inter-repeat variation were unfortunately not successful.

Haplotype	Number of chromosomes	Average of PESR variant	Standard Deviation
SA	70	0.2050	0.0348
SG	10	0.2283	0.0514

Table 3.7: Summary of the number of small TR MVR alleles from NE individuals. The average amount of the PESR variant is slightly higher in the SG chromosomes than in the SA chromosomes but this difference was not statistically significant p=0.1939 p- 2tail, p=0.969 p- 1 tail.

3.4 Discussion

The 4 populations studied were all collected in London, three of the groups in the same GI clinic, so that all groups would be comparable. The two GI clinic control groups and the 'London area' group showed genotypic and allelic frequencies similar to each other, but statistically different from the *Helicobacter pylori* gastritis group for the TR distribution. In the *H. pylori* gastritis group there was a significant increase of S alleles and of the proportion of individuals carrying the SS genotype.

In the haplotype prediction, PHASE software program generated a larger number of haplotypes than the Arlequin program. This is due to the fact that PHASE estimates the haplotype frequencies without taking into account the number of chromosomes in each population and thus predicts some rare haplotypes that would occur in less than 1 chromosome. However, the two software packages produced extremely similar haplotype distributions. Again, the 3 control groups showed very similar haplotype distributions, and distinct from the gastritis group. The latter showed an over-representation of a particular recombinant allele, G S 11, that was represented at small proportions of the chromosomes in the other groups.

In preliminary inspection of the MVR patterns, it seemed that the G S 11 haplotype showed a distinct pattern, with larger blocks of PESR variants interspersed between the PDTR blocks rather like the large TR alleles which generally share the same flanking markers. However, the analyses of the proportion of PESR variant in SA and SG chromosomes showed that the small difference between the two was not statistically significant. This may be in part due to the very small number of SG chromosomes with available MVR maps.

Since a different haplotype distribution was observed in the gastritis patients, it was of interested to establish whether there are inter-population differences in the distribution of these haplotypes and whether a change in frequency was observed in other patients with gastric disease. This is considered in Chapters 4 and 6.

CHAPTER 4

MUC1 HAPLOTYPE DISTRIBUTION IN POPULATIONS OF DIFFERENT ANCESTRY

PORTUGUESE POPULATION- STUDIES ON PATIENTS

Introduction

As seen in a previous chapter, in Northern Europeans *MUC1* usually exhibits a bimodal distribution of TR lengths and very high Linkage Disequilibrium between the S and L TR lengths and the 2 flanking polymorphisms in the gene. However, little was known about other populations, and whether the bimodality and LD were still maintained. In the first part of this chapter, 3 populations of different ancestries are compared, not only in terms of allele length and genotypic frequencies, but also haplotype distribution.

For this study, samples were available from an unselected population of Nigerian origin. No previous analysis of the TR domain of MUC genes has been reported for Africans, partly because of the lack of availablility of suitable DNA samples. Normally this means using DNA from blood, but because of HIV and AIDS blood is rarely sent from Africa to the UK these days. For this study an attempt was made to use buccal DNA. DNA was obtained from 10 buccal swabs from each donor as described in section 2. In some cases it was found necessary to reprecipitate and after this procedure DNA of adequate quality was obtained in a total of 60/86 samples from unrelated individuals (70% of cases) to obtain MUC1 TR data by Southern blot analysis (see Figure 4.1).

TR data and DNA were also available from unselected Portuguese blood donors.

In addition, data and samples were available from Portuguese patient groups and they are considered in the second half of this chapter.

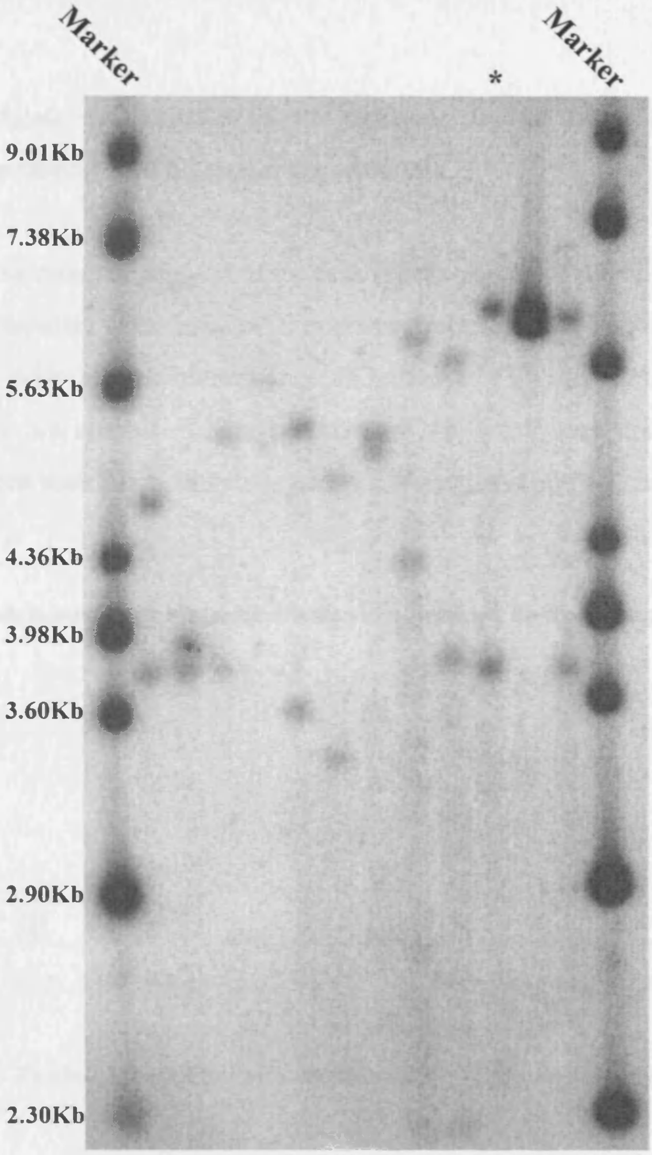


Figure 4.1: Southern Blot obtained using Nigerian samples digested with the *Hinf I* restriction enzyme. The autoradiograph shows the Raoul molecular weight marker in each side of the gel and the bands size on the left side. The * symbol marks the lane position of the internal positive control, digested DNA from the cell line Caco-2.

4.1 *MUC1* genotype and haplotype distribution in populations of different ancestry

4.1.1 Allele distribution of the tandem repeat array in the Northern European population and Nigerian population

In this section the lengths of tandem repeat alleles of the Nigerian population group were compared with those of Europeans from the UK. The data are displayed in histograms with alleles binned in 0.5Kb classes (Figure 4.2). The Portuguese population was not compared here because the TR sizing was done in another lab, using a different restriction enzyme, and a different method to size the alleles (see section 4.1.2.1).

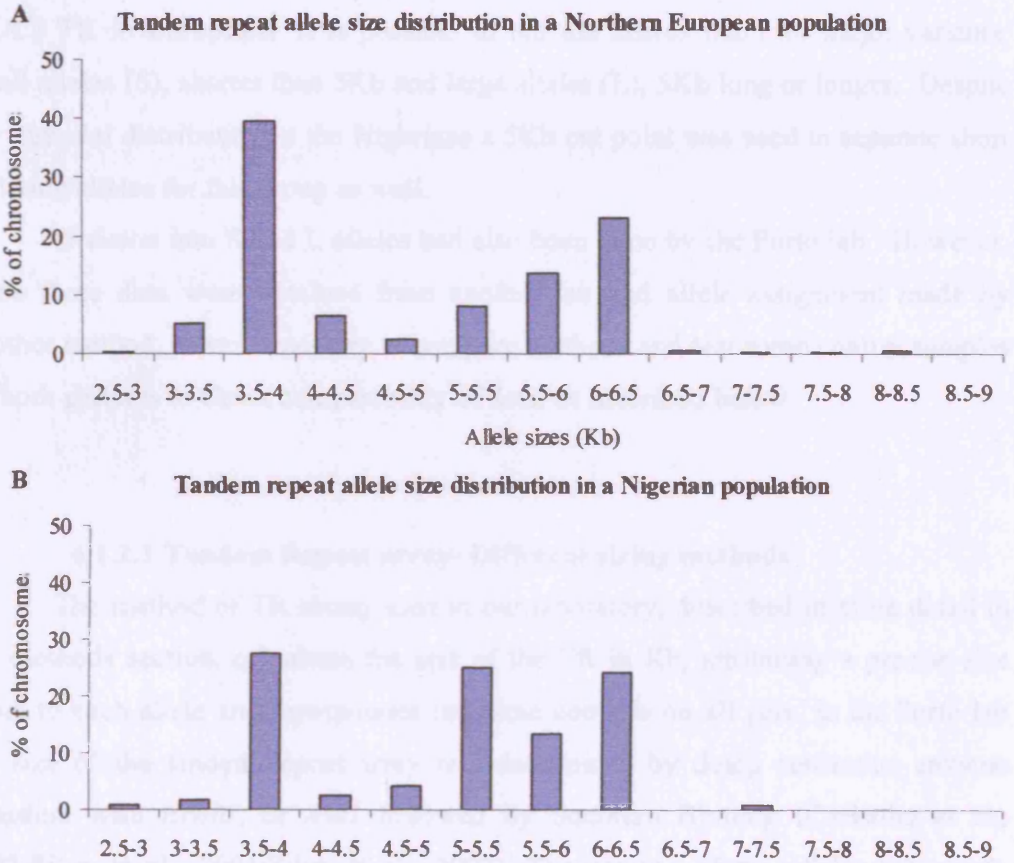


Figure 4.2: Allele size distribution in a population of Northern European extraction, collected in London (A), and Nigerian population (B). The NE group shows a bimodal distribution, though the second peak show a very abrupt decrease. The Nigerian group shows a trimodal distribution, never observed in a NE population, with an increase of large alleles near the cut-point area (5-5.5Kb).

The histogram in figure 4.2 shows the distribution of the TR array size in the Nigerians compared with UK residents. As previously reported, the Northern European population shows a bimodal distribution of the alleles with the dip between the peaks at about 5Kb length (Vinall et al., 2002). The distribution in the Nigerians in contrast was trimodal with 2 peaks above 5kb. Despite these differences, a Mann-Whitney test reveals no statistically significant difference between the two populations (P (2-tailed)= 0.1978).

4.1.2 Comparison of three populations of different ancestry: Northern European, Nigerian and Portuguese

As seen in the previous chapter, because of the bimodal distribution of the *MUC1* TR in Europeans, it is possible to bin the alleles into two major variants: small alleles (S), shorter than 5Kb and large alleles (L), 5Kb long or longer. Despite the trimodal distribution in the Nigerians a 5Kb cut point was used to separate short and long alleles for this group as well.

Division into S and L alleles had also been done by the Porto lab. However, since these data were obtained from another lab and allele assignment made by another method, it was necessary to compare methods and test some control samples on both systems to check comparability of data, as described below.

4.1.2.1 Tandem Repeat array- Different sizing methods

The method of TR sizing used in our laboratory, described in more detail in the methods section, calculates the size of the TR in Kb, attributing a precise size value to each allele and incorporates the same controls on all gels. In the Porto lab the size of the tandem repeat array was determined by doing restriction enzyme digestion with *EcoRI*, or *Alu1* followed by Southern Blotting (Carvalho et al., 1997;Silva et al., 2001;Silva et al., 2003). The scoring of the alleles was made analysing the autoradiograms visually. The bands were scored into 15 discrete allele classes according to the band position in relation to the molecular marker λ*HindIII* and other samples. The bands ranged in size from 8.9Kb to 12.6Kb, and the alleles were assigned a number corresponding to the size, with allele 1 being the largest, and allele 15 the smaller in molecular weight. Though this sizing method is very different from the one used in our lab, where we use the actual size of the bands and treat it as a continuous variable, the cut-point between large (L) and small (S) alleles seems to coincide, and therefore the L and S terminology used by the Portuguese groups seems to be the same (see Figures 4.3 and 4.4).

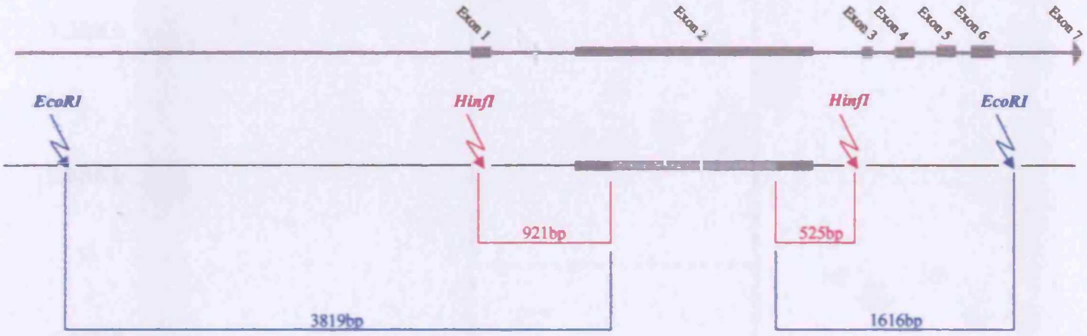


Figure 4.3: Schematic representation of the *MUC1* gene (long arrow at the top) and the relative positions of *HinfI* and *EcoRI* cut sites to exon2 and Tandem Repeat region. The distance in base pair number from the enzyme cut sites to the TR domain is shown in the open boxes (blue for the *EcoRI* and purple for the *HinfI*). The *Alu I* cut points are not marked here since this enzyme was not used in any of the analyses for this study.

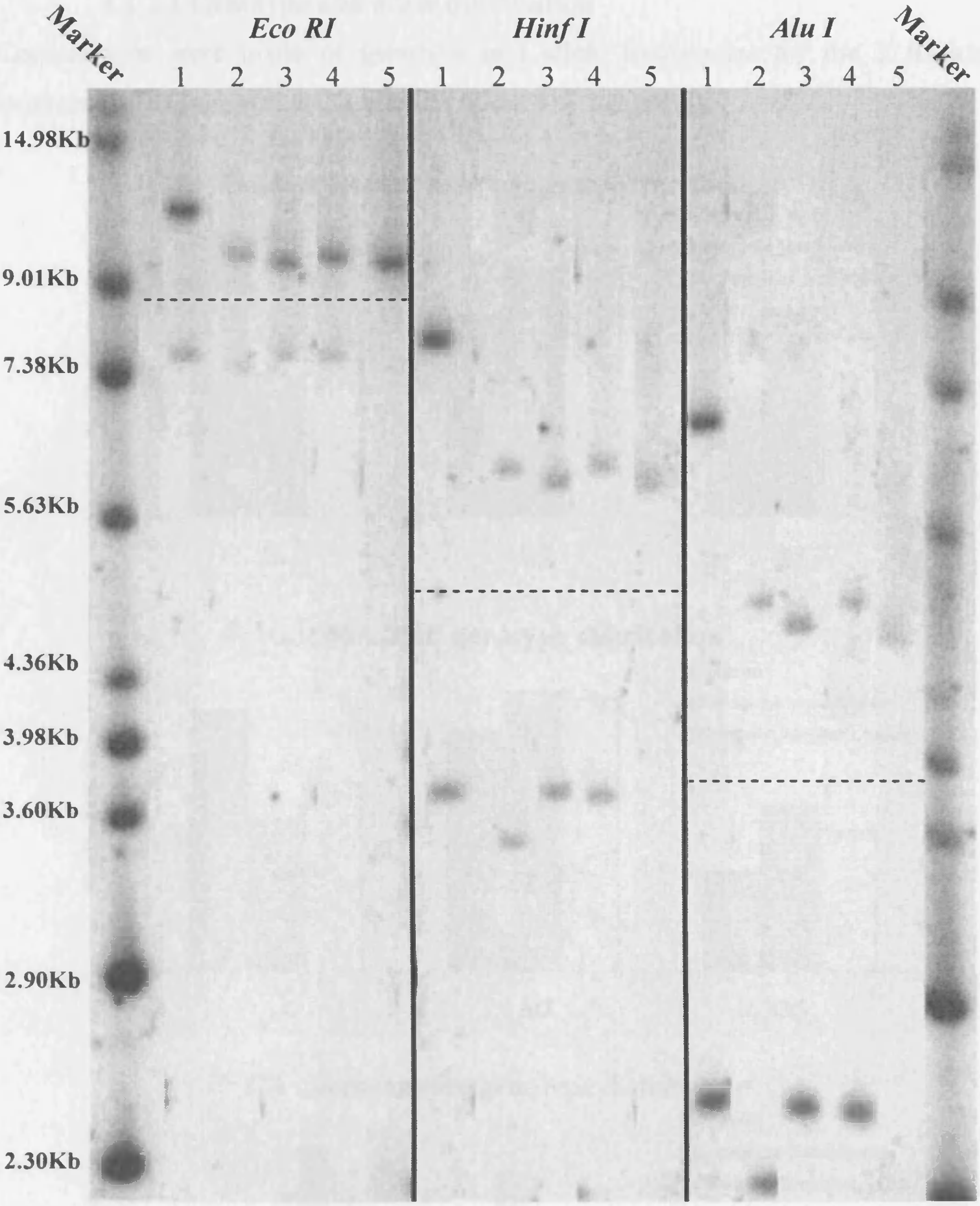


Figure 4.4: Southern Blot obtained using 5 different samples digested with the 3 restriction enzymes used in the lab in London, *Hinf I*, and the lab in Portugal, *Eco RI* and *Alu I*. The autoradiograph shows the Raoul molecular weight marker in each side of the gel and the bands size on the left side. The dashed lines correspond to the calculated cut points for each enzyme. It is easily observed that the three enzymes have a consistent cut pattern for each sample. This figure shows the results are reproducible using different restriction enzymes.

4.1.2.2 Genotype and allele distribution

Comparisons were made of genotype and allele frequencies for the 2 flanking markers and the simplified TR alleles (figure 4.5, Table 4.6).

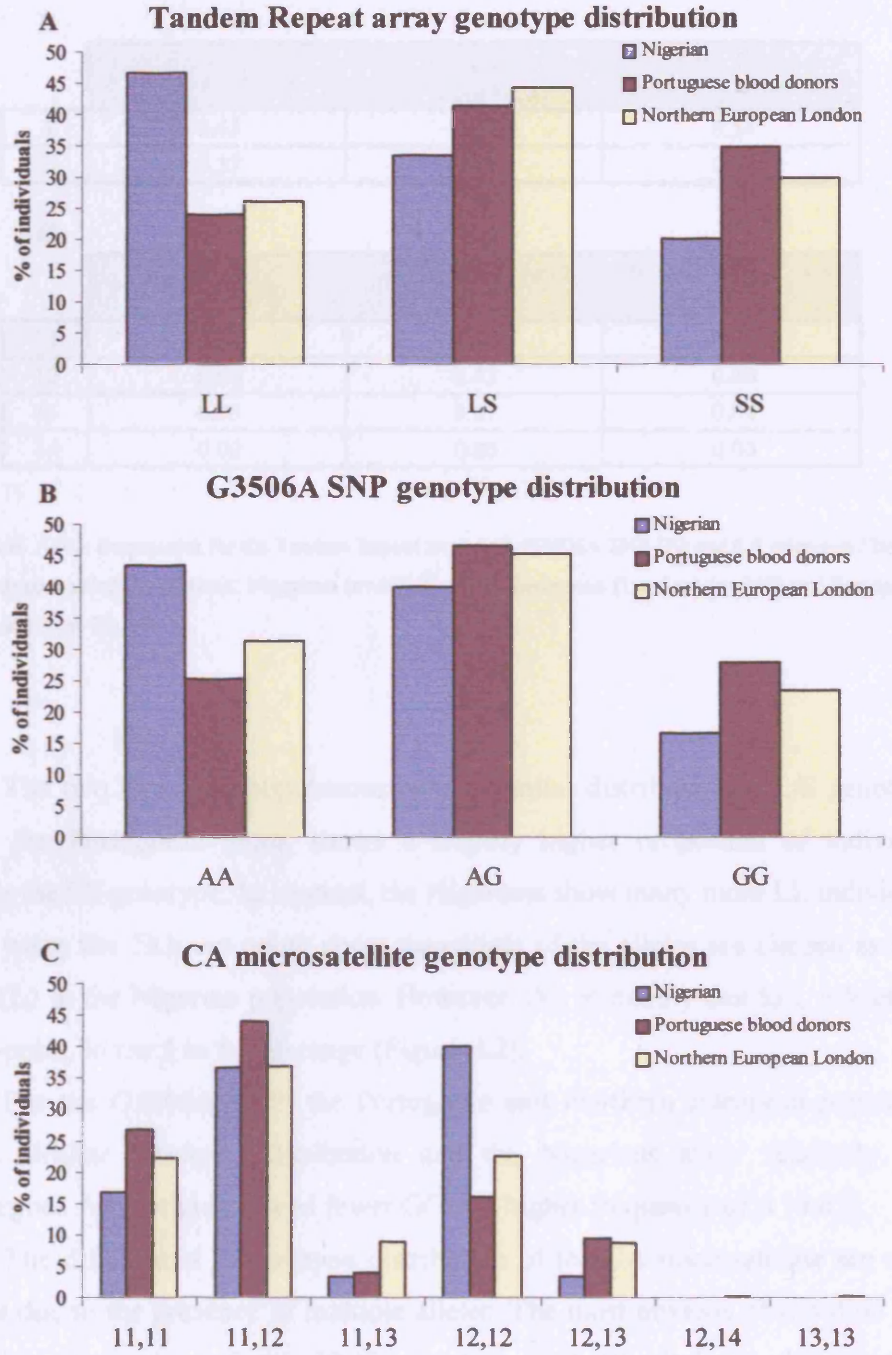


Figure 4.5: Genotype frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the three studied populations: Nigerians (n=60), Northern Europeans (London) (n=349) and Portuguese blood donors (n=75).

A

	Nigerian	Portuguese blood donors	Northern European London
L	0.63	0.45	0.48
S	0.37	0.55	0.52

B

	Nigerian	Portuguese blood donors	Northern European London
A	0.63	0.49	0.54
G	0.37	0.51	0.46

C

	Nigerian	Portuguese blood donors	Northern European London
11	0.37	0.51	0.45
12	0.60	0.43	0.46
13	0.03	0.07	0.09
14	0.00	0.00	0.00

Table 4.6: Allele frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the three studied populations: Nigerians (n=60), Northern Europeans (London) (n=349) and Portuguese blood donors (n=75).

The two European populations have a similar distribution of L/S genotypes, though the Portuguese group shows a slightly higher proportion of individuals carrying the SS genotype. In contrast, the Nigerians show many more LL individuals. In fact using the 5Kb cut point, about two-thirds of the alleles are classed as Large alleles (L) in the Nigerian population. However, this is mainly due to L alleles near the cut-point, in the 5 to 5.5Kb range (Figure 4.2).

For the G3506A SNP, the Portuguese and Northern European populations have a similar genotype distribution and the Nigerians show relatively more homozygous AA individuals and fewer GG and higher frequency of A (0.63).

The differences in genotype distribution of the CA microsatellite are not so obvious due to the presence of multiple alleles. The most obvious observation is the very high proportion of the 12,12 genotype and 12 allele in the Nigerians, accompanied by a less obvious decrease in the frequency of the 11,11 genotype. 63% of the alleles in the CA microsatellite with 12/13 repeats. The other two populations show no significant differences in the genotype/allele distribution, despite the presence of a minor allele with 14 repeats in the Northern European population that is not present in the others. These differences in allele frequencies between the Nigerian population and the two European populations are statistically significant. A Fisher's exact test using allele counts revealed 2-tailed p values of 0.0031 between the Nigerians and Portuguese blood donors and a p value of 0.0022 between Nigerians and Northern Europeans. The two European populations were not significantly different from each other (p value= 0.4714).

4.1.2.3 Haplotype distribution for the 3 populations:

Haplotype distributions were inferred for the 3 populations using: Arlequin and PHASE, as used previously in Chapter 3.

Goodness of fit to Hardy-Weinberg Equilibrium (HWE) was also tested using the Arlequin software package. The G3506A SNP and CA microsatellite showed no deviation from HWE for the three populations tested. However, while the TR array L/S genotype distribution is in HWE in the Portuguese blood donors, it shows a significant deviation in the Northern Europeans and Nigerians, with p values of 0.03132 and 0.02705, respectively. This deviation from equilibrium is in both cases due to a lack of LS heterozygotes.

Figure 4.7 shows the haplotype distribution inferred using both methods. From direct observation it is obvious that Arlequin (Figure 4.7.A) and PHASE (Figure 4.7.B) produce very similar haplotype distributions, giving consistency to the predictions.

For all three populations, the two major haplotypes are the same, A S 12 and G L 11 (See Figure 4.7). However, the A L 12 haplotype, which is relatively rare in the two populations of European extraction, is much more frequent in the Nigerians, accounting for more than 25% of the chromosomes in this group. Nonetheless, the high LD between markers previously observed in the Northern Europeans and responsible for the appearance of two major haplotypes, A S 12 and G L 11, is not disrupted in the Nigerians, with the LD p values calculated by the Arlequin software package lower than 0.0001.

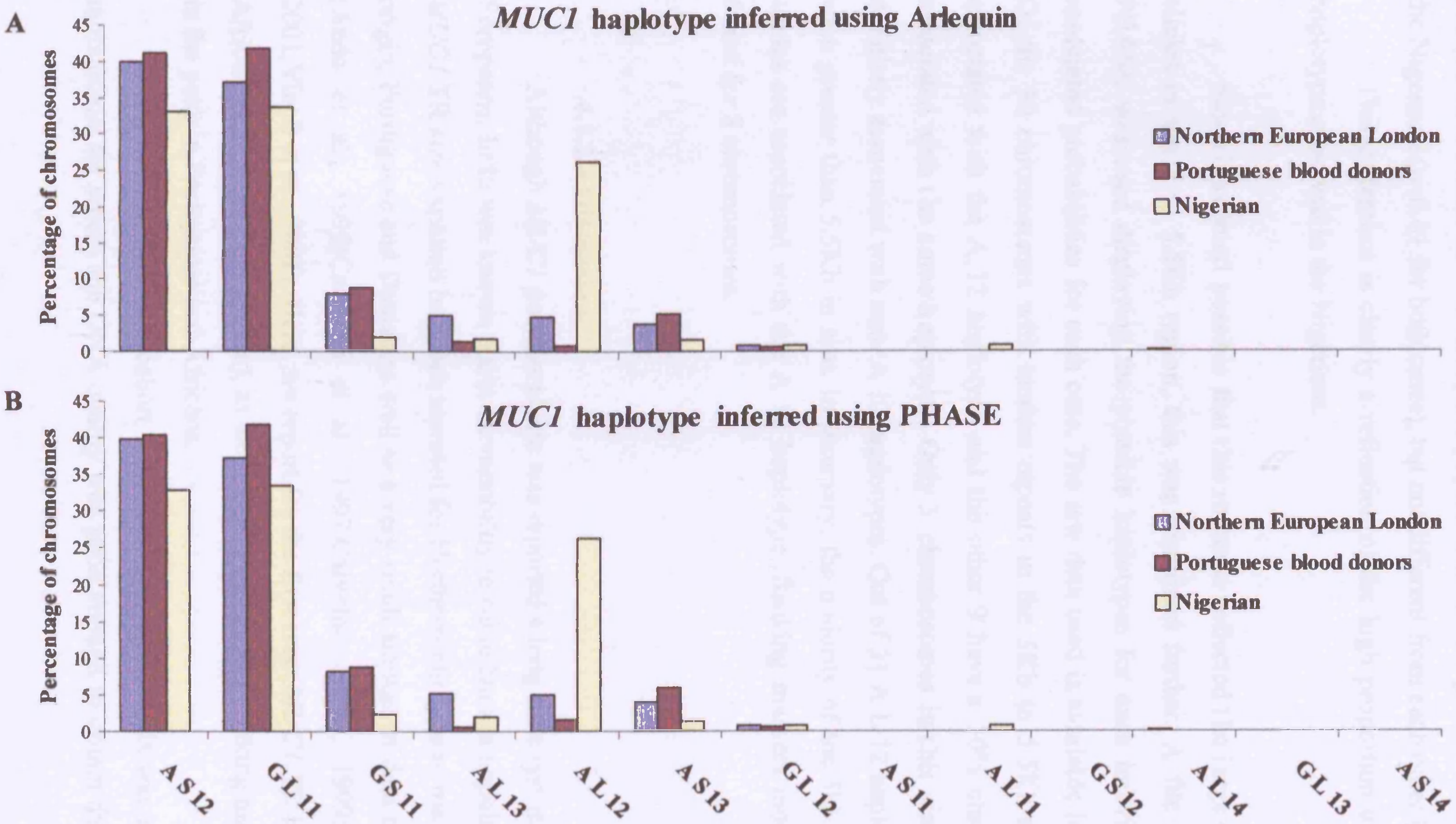


Figure 4.7: *MUC1* haplotype distribution for the three studied populations: Nigerians (n=60), Northern Europeans- London (n=349) and Portuguese blood donors (n=75). The haplotypes were inferred using 2 different statistical methods: (A) a maximum-likelihood method, the EM algorithm, and (B) a Bayesian method.

Comparison using the 'Exact test of sample differentiation' (performed by Arlequin) shows that the Nigerian population is significantly different from the Northern European London cohort and the Portuguese blood donors' cohort ($p<0.00001$ for both pairs of populations). Comparisons of pairs of populations using PHASE also found the two European populations to be significantly different from the Nigerians ($p=0.01$ for both cases), but not different from each other ($p=0.07$).

This difference is clearly a reflection of the high proportion of the A L 12 haplotype observed in the Nigerians.

Since it seemed possible that this increase reflected the large number of L alleles in the 5 to 5.5Kb region, this was examined further. A file generated by PHASE was used displaying the possible haplotypes for each individual and the associated probabilities for each case. The raw data used is available in appendix 1. Of the 30 chromosomes with tandem repeats in the 5Kb to 5.5Kb range, 18 are associated with the A 12 haplotype and the other 9 have a 50% chance of being associated with the same haplotype. Only 3 chromosomes in this size range were definitely associated with non- A 12 haplotypes. Out of 31 A L 12 haplotypes only 4 were greater than 5.5Kb in size. In summary, the majority of the 5kb to 5.5Kb TR alleles are associated with the A 12 haplotype, flanking markers more commonly found for S chromosomes.

4.1.2.4 Discussion

Although *MUC1* polymorphism was reported a long time ago in the Northern Europeans, little was known about its variability in other human populations. So far, *MUC1* TR size variation has been reported for Northern Europeans, mainly of British origin, Portuguese and Danish as well as a very small amount of data from Japanese (Ando et al., 1998;Carvalho et al., 1997;Carvalho et al., 1999;Silva et al., 2001;Vinall et al., 2002). Here, we report for the first time *MUC1* TR information in Africans (a Nigerian population), as well as data on the two flanking markers present in the gene in Portuguese and Africans.

The original Nigerian cohort of 86 unrelated individuals was reduced to 60 individuals, for whom the DNA quality was good enough to obtain data for the TR allele sizes. The results obtained from those that were included were however clear (Figure 4.1).

The Nigerian population revealed a significantly different distribution of the TR array alleles, which seemed to be trimodal, due to a large number of chromosomes with TR in the 5-5.5Kb size range. This was accompanied by a decrease in the proportion of small alleles in the 3-3.5Kb class, though the overall range in size of TR alleles seemed similar to that in Europeans.

The slight deviation from HW Equilibrium of the TR genotypes in Nigerians and Northern Europeans was of possible concern. However, it may not be of real significance because of the large number of comparisons made; it may also reflect an artefact of binning allele sizes. The presence of very small alleles, missed in our system is also a possible explanation, though rather unlikely. The loss of a few large alleles in the Nigerians is conceivable because the samples collected were from buccal DNA and the DNA quality is not as good as blood, and results for the TR using Southern Blotting are more difficult to obtain. However, nearly 80% of the samples showed 2 different length alleles making nether of these likely. The same situation applies to the NE population, where 87% of the individuals showed 2 different alleles.

A noteworthy feature of the Nigerian data, however, was that the flanking markers were, as in Europe, highly associated, with no evidence of greater haplotypic diversity than in Europeans, Most studies indicate that there is more diversity and regions of LD tend to be shorter in Africans (Gabriel et al., 2002;Hollox et al., 2001;Tishkoff et al., 2000) but here is still limited information available as to inter group differences. . Using the European cut point for the length of the TR alleles, one haplotype class that is rare in Europeans was frequent. This haplotype is different from the one found frequently in the UK gastritis study. It could be attributable to a particular TR allele (length 5-5.5Kb) being at a particularly high frequency in the Nigerian population we tested. It should be noted that most of these samples came from a single tribal group Ibibio (n=85 from the original group). If suitable DNA becomes available it will be of interest to study these TR variations in different groups and in different parts of Africa. In Nigeria, the *Helicobacter pylori* infection is usually associated with moderate to severe chronic gastritis (Oluwasola and

Ogunbiyi, 2004). Despite the high incidence of *H.* pylori infection in the Nigerian population (80 to 85%), the prevalence of atrophy or intestinal metaplasia is quite low (Oluwasola and Ogunbiyi, 2003;Oluwasola and Ogunbiyi, 2004). This may explain the very low incidences of gastric cancer in Nigeria, about 1 case per 100.000 individuals, according to the 2002 statistics from Globocan (http://www-dep.iarc.fr/globocan/database.htm).

4.2 Portuguese Population- Studies on patients

Introduction

The Portuguese population has a very high incidence of gastric disease, including gastric cancer. The gastric cancer rates in Portugal are the highest in Europe, and one of the highest in the world. According to Globocan 2002 database (http://www-dep.iarc.fr/globocan/database.htm), the number of new cases per year per 100.000 individuals (Incidence Crude Rate) in Portugal is of 43.4 in males and 25.9 in females. In the face of these facts, it has been interesting to study the Portuguese population to assess whether genetic predisposition could play an important role, and the Portuguese lab considered *MUC1* as one of the possible candidates (Carvalho et al., 1997).

The samples in this study were obtained in collaboration with Prof Leonor David and Dr Raquel Seruca, from the Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Portugal. Three different groups of samples were studied: a cohort of blood donors, from S. João Hospital, Porto, from whom no gastric disease information is available, a cohort of individuals with gastritis and dyspepsia, collected in a naval shipyard in Viana do Castelo, North Portugal, and a cohort of gastric cancer samples collected also in S. João Hospital. The sample collection and DNA extraction was done in Portugal, and *MUC1* Tandem Repeat size information was also produced in Portugal by Dr Filipa Carvalho and Dr Luis Filipe Silva.

The aim of this part of the work was to determine whether a disease associated haplotype could also be detected in this population.

4.2.1 Blood donors and gastric cancer

In the first analysis the blood donors and gastric cancer samples will be analysed together since they were collected in the same hospital, and therefore more likely to be representative of 2 subgroups of the same population. The 'Viana' samples were collected in a different region of Portugal, and the collection procedure was different and will be referred to later in this chapter.

A previous study of the *MUC1* Tandem Repeat (TR) polymorphism involving the two populations in this particular study reported a significant difference on the distribution of the TR alleles and genotypes, showing a significant higher prevalence of small alleles (S) and individuals with SS genotypes in the gastric cancer group in relation to the blood donors control group (Carvalho et al., 1997).

In our study, only a subset of each of the above-mentioned populations was used, rather than the entire group of individuals in the report, because most of the DNA samples were not available. They were either degraded or used up. As a result, data was obtained for a total of 75 samples from the blood donors group of the original 166 and 43 samples from the gastric cancer cohort of the original 66.

The TR information that was already available for the individuals was used and analysed in conjunction with the two *MUC1* flanking markers. Allelic frequencies and genotypic frequencies are produced and the haplotype distribution inferred.

4.2.1.1 Genotype and allele distribution

Figure 4.8 shows the genotype distribution for each of the 3 *MUC1* markers for blood donors and gastric cancer patients. Figure 4.8.A shows the dramatic difference in Tandem Repeat genotype frequencies between the two groups. While the blood donors show a relatively balanced distribution of LL and SS genotypes and a higher proportion of LS heterozygotes, the gastric cancer group has a very similar percentage of heterozygotes LS, but the fraction of SS homozygotes is markedly increased and the LL homozygotes amount declines dramatically, to about 7% of the total number of gastric cancer patients.

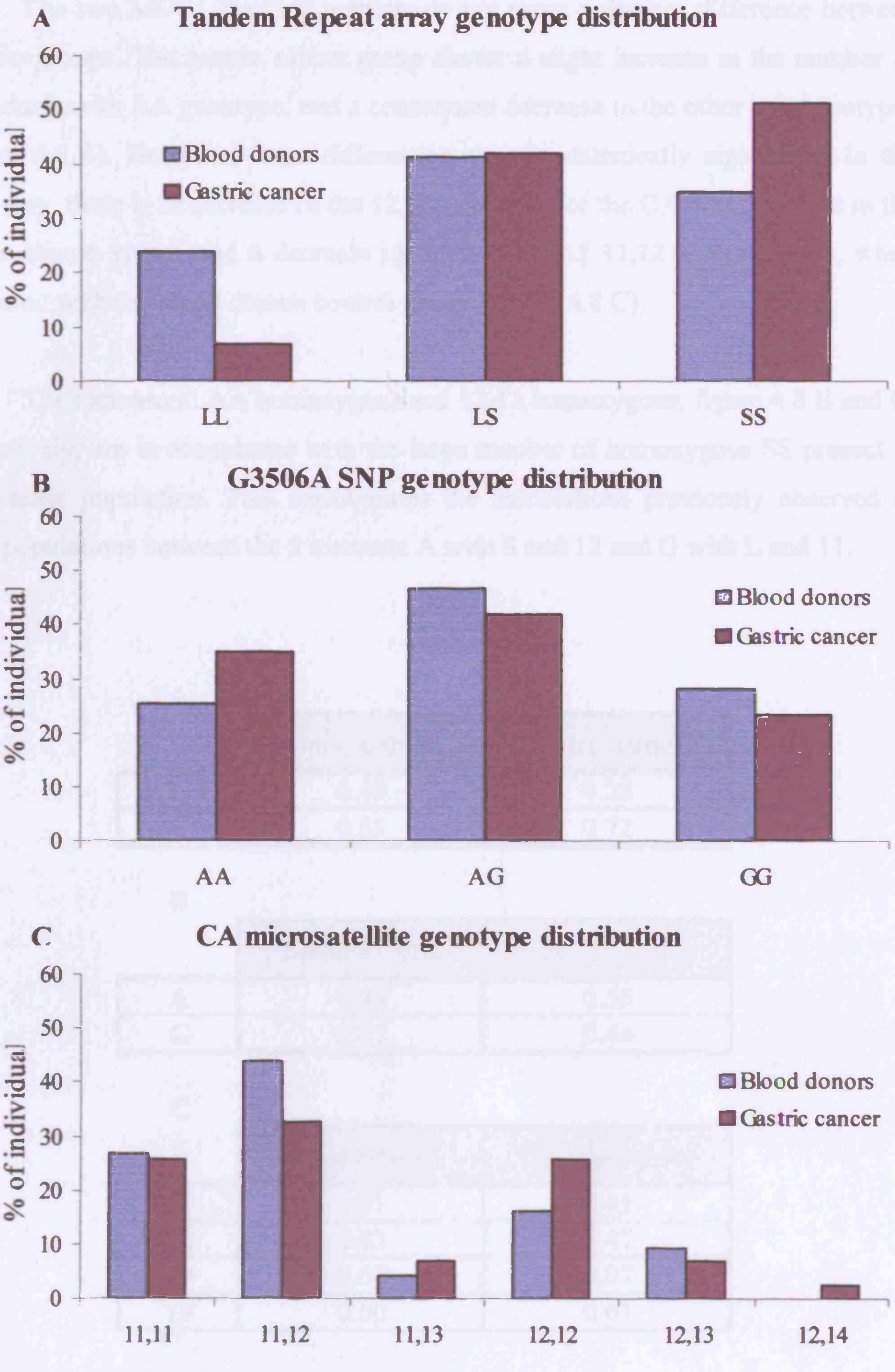


Figure 4.8: Genotype frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the two Portuguese populations: Blood donors (n=75) and Gastric cancer patients (n=43).

The two *MUC1* flanking markers do not show a distinct difference between the two groups. The gastric cancer group shows a slight increase in the number of individuals with AA genotype, and a consequent decrease in the other two genotypes (Figure 4.8.B). However, these differences are not statistically significant. In the same way, there is an increase of the 12,12 genotype for the CA microsatellite in the gastric cancer group, and a decrease in the fraction of 11,12 heterozygotes, when compared with the blood donors control group (Figure 4.8.C).

The increase in AA homozygous and 12,12 homozygous, figure 4.8 B and C, respectively, are in accordance with the large number of homozygous SS present in the disease population. This corroborates the associations previously observed in other populations between the 3 markers: A with S and 12 and G with L and 11.

A

	Blood donors	Gastric cancer
L	0.45	0.28
S	0.55	0.72

B

	Blood donors	Gastric cancer
A	0.49	0.56
G	0.51	0.44

C

	Blood donors	Gastric cancer
11	0.51	0.45
12	0.43	0.47
13	0.07	0.07
14	0.00	0.01

Table 4.9: Allele frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the two Portuguese populations: Blood donors (n=75) and Gastric cancer patients (n=43).

Table 4.9 shows the allele frequencies for the blood donors group and the cohort of gastric cancer patients. As is easily observed, the percentage of S alleles is much higher in the gastric cancer group than in the control group of blood donors (Table 4.9.A). Despite the reduction in the number of individuals in each group compared with the ones reported by Carvalho *et al*, the difference in L and S allele distribution for the TR polymorphisms is still statistically significant. A Fisher's exact test revealed a 2-tailed p-value of 0.0124. However, the two other polymorphisms tested in our lab seem to have similar allele distribution, with no significant difference between the two population groups (Tables 4.9.B and 4.9.C).

4.2.1.2 Haplotype distribution

Using the Arlequin software package and PHASE program, the haplotype distribution for the two populations was inferred. Once more, both programs reveal very similar haplotype distributions, as shown in figure 4.10 (A and B). It is interesting to observe a substantial increase in the particular haplotype G S 11 in the gastric cancer group, where it accounts for about 16% of the chromosomes. This haplotype is the same one shown to be over-represented in the *Helicobacter pylori* gastritis group in Northern Europeans (see chapter 3). However, the difference was not statistically significant, although comparison of two locus haplotypes (AG and SL) almost reached significance (p=0.068).

The 3 loci were in HWE in both populations, and were in tight Linkage Disequilibrium. Unlike the gastritis group in the Northern Europeans, the *MUC1* loci in the Portuguese gastric cancer patients group remain in LD, even though there is an increase in the 'recombinant' GS11 haplotype.

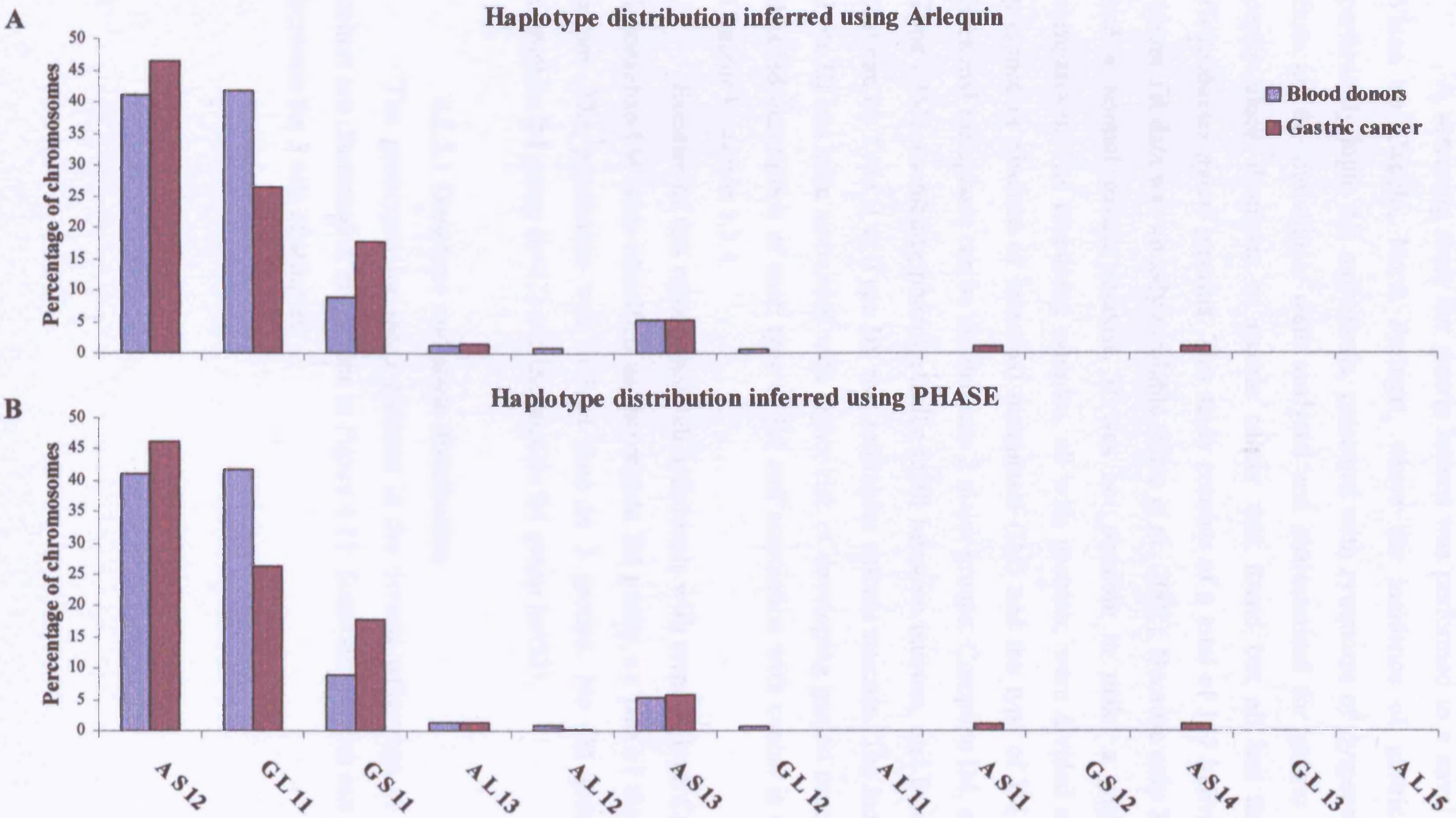


Figure 4.10: *MUC1* haplotype distribution for the two Portuguese populations: Blood donors (n=75) and Gastric cancer patients (n=43). The haplotypes were inferred using 2 different statistical methods: (A) a maximum-likelihood method, the EM algorithm, and (B) a Bayesian method.

4.2.2 Naval Shipyard in 'Viana do Castelo' (North Portugal)- A patients "enriched" study

A screening study for gastric lesions was performed in a naval shipyard in Viana do Castelo, North Portugal, where the incidence of gastric disease is particularly high. All individuals, presented with symptoms of dyspepsia. Biopsies from all the individuals were analysed and characterised for gastric lesions. No peptic ulcer, dysplasia or gastric cancer was found but all but three showed *Helicobacter pylori* gastritis. This study consists of a total of 157 individuals, from whom TR data was already available (Silva et al., 2001). Because only 3 individuals had a normal gastric mucosa, it was not possible to make a gastritis/normal comparison. The remaining samples, all with gastritis, were divided according to presence or absence of intestinal metaplasia (IM) and the type of IM found. The intestinal metaplasia can be divided into 2 major groups: Complete IM, also know as Type I IM, resembling phenotypically small intestine mucosa, and Incomplete IM, that can be Type II or Type III, that resembles colonic mucosa. The incomplete IM Type III has been associated with higher risk of developing gastric cancer. A more detailed description of each type of IM and association with cancer is described in Chapter 1, section 1.3.4.

Because of this relative risk, all individuals with areas of both Complete and Incomplete IM were classified as Incomplete IM group, as part of the higher risk group. The population was divided then in 3 groups: No IM group (n=113), Complete IM group (n=12) and Incomplete IM group (n=32).

4.2.2.1 Genotype and allele distribution

The genotypes for the 3 markers in the diverse subgroups of the 'Viana' cohort are illustrated in the graphs in Figure 4.11. Some differences can be observed between the 3 sets of samples.

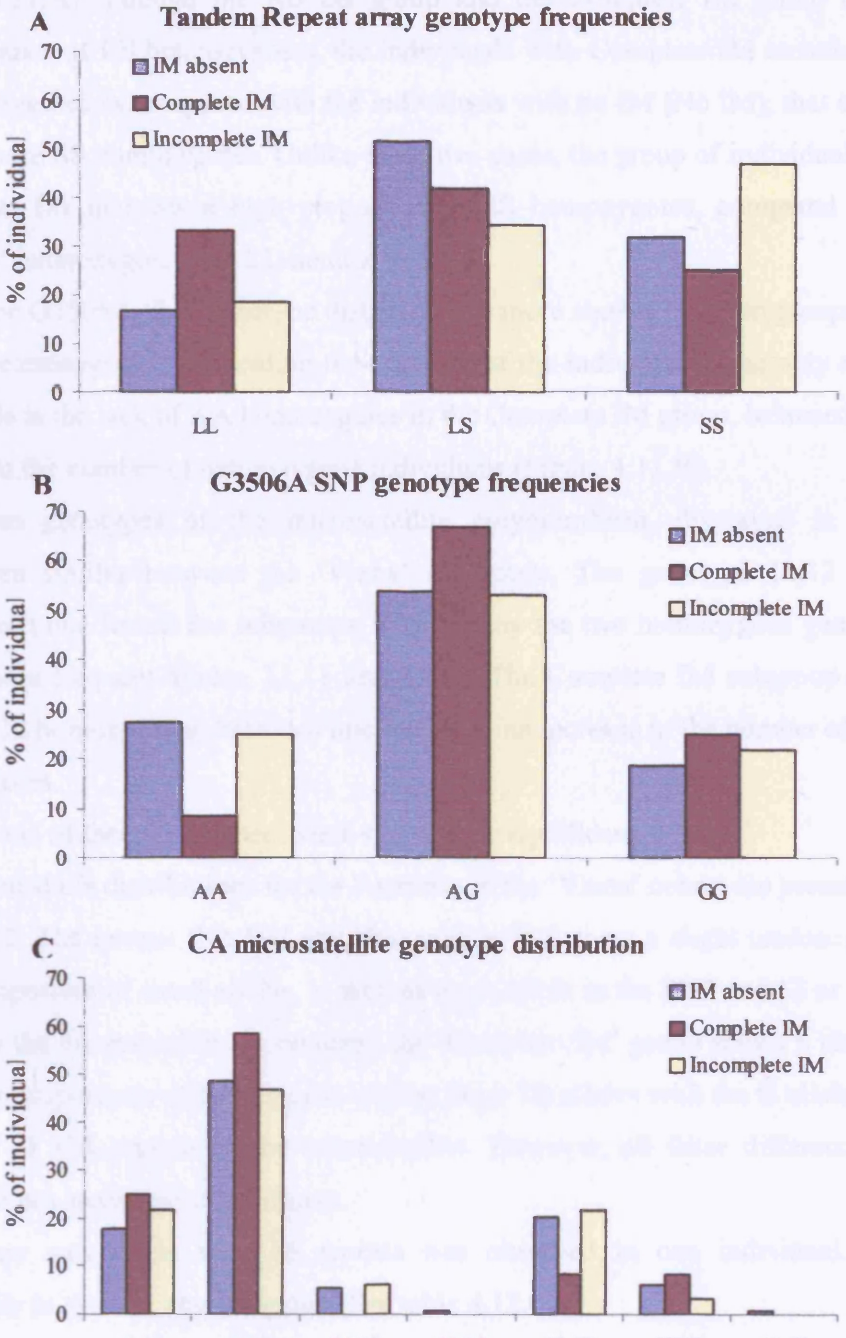


Figure 4.11: Genotype frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the three 'Viana' subgroups: No IM (n=113), Complete IM (n=12) and Incomplete IM (n=32).

The tandem repeat genotype frequency is somewhat different in each c ase (Figure 4.11.A). Though the No IM group and the Complete IM group have a predominance of LS heterozygotes, the individuals with Complete IM include more LL homozygotes as compared with the individuals with no IM (No IM), that include slightly more SS homozygotes. Unlike these two cases, the group of individuals with Incomplete IM includes a high proportion of SS homozygotes, compared to the number of heterozygotes and LL homozygotes.

The G3506A SNP genotype distribution is more similar between groups, with the AG heterozygotes representing the majority of the individuals. The only remark to be made is the lack of AA homozygotes in the Complete IM group, balanced by an increase in the number of heterozygous individuals (Figure 4.11.B).

The genotypes of the microsatellite polymorphism, displayed in figure 4.11.C, are similar between the 'Viana' subgroups. The genotype 11,12 is the predominant one for all the subgroups, followed by the two homozygous genotypes for the more frequent alleles, 11,11 and 12,12. The Complete IM subgroup shows fewer 12,12 homozygotes, balanced once more by an increase in the number of 11,12 heterozygotes.

None of these differences were statistically significant.

The allele distributions for the 3 subsets in the 'Viana' cohort are presented in Table 4.12. The groups 'No IM' and 'Incomplete IM' show a slight tendency for a higher proportion of small alleles, as well as the A allele in the SNP and 12 or 13 CA repeats in the microsatellite. In contrast, the 'Complete IM' group shows a tendency for higher proportions of the opposite alleles: large TR alleles with the G allele in the SNP and 11 CA repeats in the microsatellite. However, all these differences are small, and not statistically significant.

One new allele with 15 repeats was observed in one individual, more specifically in the 'IM absent' group (See table 4.12.C).

A

	No Intestinal Metaplasia	Complete Intestinal Metaplasia	Incomplete Intestinal Metaplasia
L	0.42	0.54	0.36
S	0.58	0.46	0.64

B

	No Intestinal Metaplasia	Complete Intestinal Metaplasia	Incomplete Intestinal Metaplasia
A	0.54	0.42	0.52
G	0.46	0.58	0.48

C

	No Intestinal Metaplasia	Complete Intestinal Metaplasia	Incomplete Intestinal Metaplasia
11	0.45	0.54	0.48
12	0.48	0.42	0.47
13	0.07	0.04	0.05
14	0.00	0.00	0.00
15	0.00	0.00	0.00

Table 4.12: Allele frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the three 'Viana' subgroups: No IM (n=113), Complete IM (n-12) and Incomplete IM (n=32).

4.2.2.2 Haplotype distribution

Arlequin and PHASE inferred haplotypes for the 'Viana' subgroups are shown in figure 4.13, A and B, respectively. Despite the high degree of similarity between the results obtained with the two methods, PHASE predicts a broader range of haplotypes, including several with low frequencies. Doing a chromosome count, most of these low frequency haplotypes represent chromosomes with a predicted frequency of less than 1 chromosome.

Using the Arlequin software package, the goodness of fit to HWE was assessed as well as the LD between pairs of loci. All loci were found to be in HWE for the three subgroups.

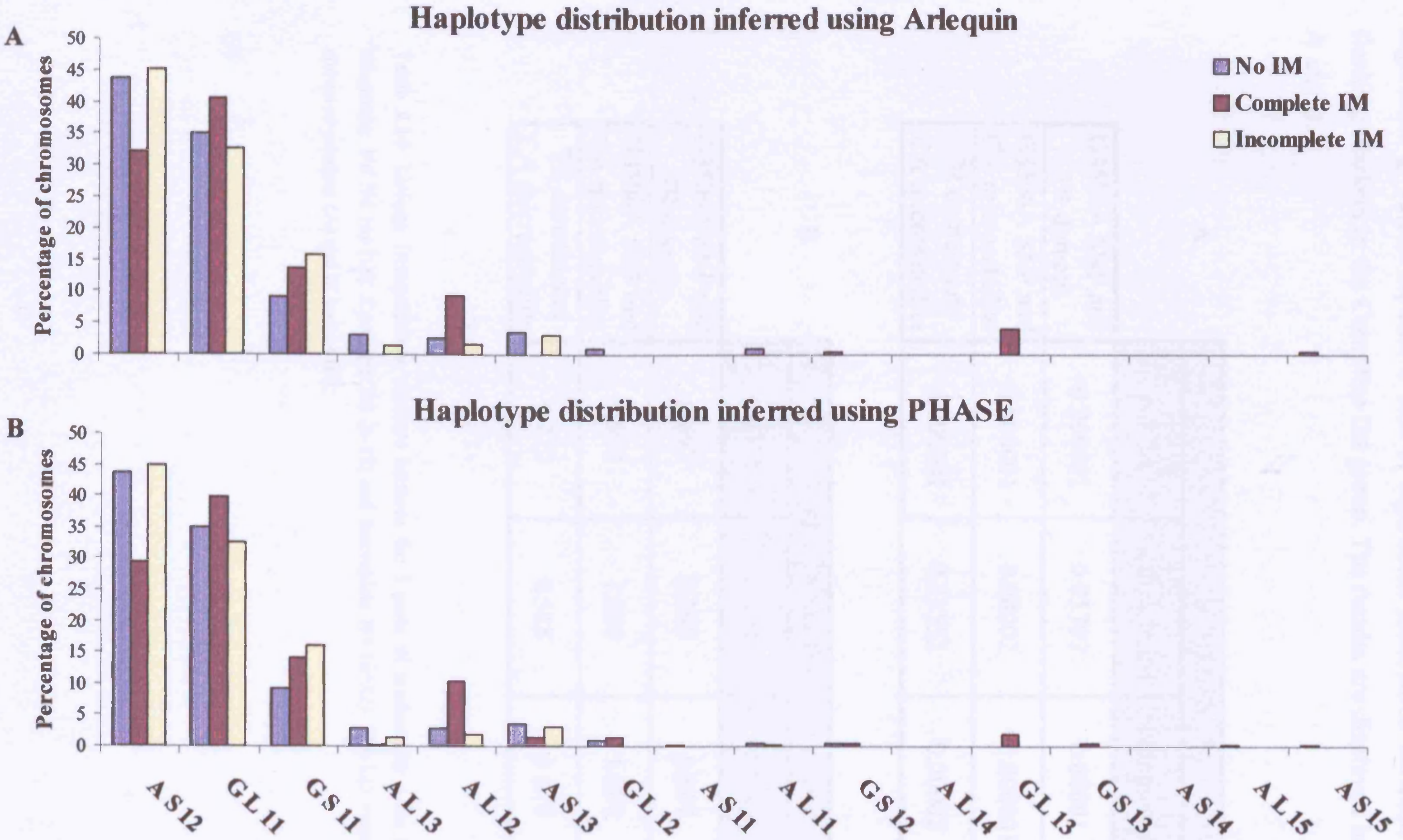


Figure 4.13: *MUC1* haplotype distribution for the three 'Viana' subgroups: No IM (n=113), Complete IM (n=12) and Incomplete IM (n=32). The haplotypes were inferred using 2 different statistical methods: (A) a maximum-likelihood method, the EM algorithm, and (B) a Bayesian method.

In the No IM group and the Incomplete IM group all loci were in LD with each other, but for the Complete IM group, the LD was disrupted between the Tandem Repeat and the two flanking markers. This analysis was also confirmed using the LDmax program, part of GOLD interface software, which also uses the EM algorithm. χ^2 values indicate a lack of significant association between LS and the flanking markers in the Complete IM group. The results are displayed in tables 4.14 A and B.

A	VIANA SUBSETS		
	No IM	Complete IM	Incomplete IM
G3506A SNP and TR domain	<0.000001	0.05707	0.00001
G3506A SNP and CA microsatellite	<0.000001	0.00002	<0.000001
TR domain and CA microsatellite	<0.000001	0.13552	0.00002

B	VIANA SUBSETS		
	No IM	Complete IM	Incomplete IM
G3506A SNP and TR domain	0.697	0.565	0.807
G3506A SNP and CA microsatellite	0.954	1.000	1.000
TR domain and CA microsatellite	0.712	0.528	0.818

Table 4.14: Linkage Disequilibrium measure between the 3 pairs of markers for r the three 'Viana' subgroups: No IM (n=113), Complete IM (n-12) and Incomplete IM (n=32). The LD measurements are shown as p values (A) and D` values (B).

The A S 12 and G L 11 haplotypes account for most of the chromosomes, in all three groups. However, while the groups with no IM and Incomplete IM have more A S 12 chromosomes than G L 11, the Complete IM group is the opposite. It is also interesting to observe that the haplotype over-represented in the *Helicobacter pylori* gastritis group in Northern Europeans, G S 11, is present in relatively high proportions in the three groups, showing an increase in frequency from the non-IM group (~7%) to the two groups with intestinal metaplasia. The highest frequency is found in the Incomplete IM group (16%), the group with highest cancer risk. Nevertheless, the differences in haplotype distribution were not statistically significant.

4.3 Discussion

The Portuguese population is of interest as the only European country with a high rate of gastric cancer. In fact, the incidence of gastric cancer in the Portuguese is one of the highest in the world, being only surpassed by Japan, North Korea, South Korea and some Eastern European counties (Globocan 2002 database- http://www-dep.iarc.fr/globocan/database.htm). One factor that plays a vital role in the high incidence of gastric disease is the very high incidence of infection by *Helicobacter pylori*. It is estimated that the incidence of *H. pylori* in Portuguese individuals 50 years old or older is as high as 80% (Lunet and Barros, 2003); in younger generations the infection rate declines, to about 50% of infection rate in individuals in their twenties (Prof. Fátima Carneiro, personal communication). Despite the high incidence of *H. pylori*, this factor alone does not explain such a high incidence of gastric disease, and in particular gastric cancer. It is therefore reasonable to assume that in combination with the bacteria infection, environmental factors as well as genetic factors are key aspects in the disease aetiology.

Carvalho *et al* investigated the potential association of gastric cancer and the epithelial mucin *MUC1*, looking at the length of the tandem repeat of the gene (Carvalho et al., 1997). Their observation of a significant over-representation of small *MUC1* tandem repeats in the gastric cancer patients when compared with blood donors controls led to the studies in London. The aim of this work was to determine whether the same haplotype that was over-represented in the UK gastritis patients was also frequent in the Portuguese patients with gastric disease. A consistent increase of this haplotype was indeed found in the gastric cancer group as compared with the blood donors. Even though not statistically significant, it was noteworthy that this haplotype, G S 11, is the same one found to be over-represented in the Northern European *Helicobacter pylori* gastritis population. It should also be taken into account that the blood donors' group, treated here as the normal controls, were not assessed for gastric disease status and because of the very high incidence of gastric cancer and precursor lesions in the Portuguese population, it is reasonable to assume that some of those individuals will suffer from gastric disease, and therefore the effects of the genetic background for disease association may be underestimated. A similar increase was also found in the 'Viana' subsets, from the individuals with no IM to the ones with IM, and within these, from the individuals with Complete IM to Incomplete IM though also lacked statistical significance. It seemed that this could be the very small sample size for the two IM subgroups, with 12 individuals in the Complete IM subgroup and 32 in the Incomplete IM. A breakdown of LD was observed in the Complete IM group; however, that was not observed in any other Portuguese disease group, and could again be an artefact of such a small sample size. Bigger sample sizes would be required to access the question of whether or not the effect observed in Northern Europe reflects in the Portuguese population. Another point that could affect the analysis is the male: female ratio, that in the 'Viana' group is very high. Because the samples were collected in a Naval Shipyard, most of the workers were males. In fact, the proportion male: female in the samples is as high as 8.8:1 and males have a higher gastric cancer incidence than females and there may be sex difference in the relative role of *MUC1*. It is possible that other genes have a relatively larger effect in men than women.

In conclusion, the clear increase of the recombinant haplotype G S 11 frequency with disease severity is noteworthy. It is possible that the study had insufficient power to show significance and also that other genes have a bigger effect in the Portuguese population.

CHAPTER 5

IDENTIFICATION OF NEW POLYMORPHISMS IN THE *MUC1* GENE AND ITS VICINITY

Introduction

The evidence developed so far suggested that a particular *MUC1* haplotype was associated with increased susceptibility to gastric disease. The aims of the work in this chapter were to determine how this haplotype fitted in with the general pattern of Linkage Disequilibrium in the vicinity of *MUC1* and to search for other potential associated markers that might be suitable and transposable for PCR for more extensive epidemiological studies.

This chapter describes the strategies used for searching polymorphisms, in particular single nucleotide polymorphisms (SNPs), both in the gene and in the vicinity of the gene. Both sequencing and bioinformatics approach were used.

5.1 *MUC1* gene resequencing

Complete sequencing of the 5` end of *MUC1* gene, the promoter region, exon 1, intron 1 and the beginning of exon 2 was carried out with the help of Mari Wyn Burley. A schematic representation of the region and primers used is shown in figure 5.1, indicating the exons location and position of the primers. The region to be sequenced was divided into 3 PCR products, around 700bp each.

The aim was to sequence 8 samples, though a total of 12 individuals were used (Table 5.2). The samples were chosen according to their genotypes for the *MUC1* TR and the two flanking markers, in order to obtain as diverse as possible set of genotypes and haplotypes, to maximise the chances of picking up new polymorphisms associated with disease haplotypes.

A comprehensive list of the primers used and sequencing techniques is described in the materials and methods, section sequencing. The sequences were analysed using the ChromasPro software (Technelysium Pty Ltd).

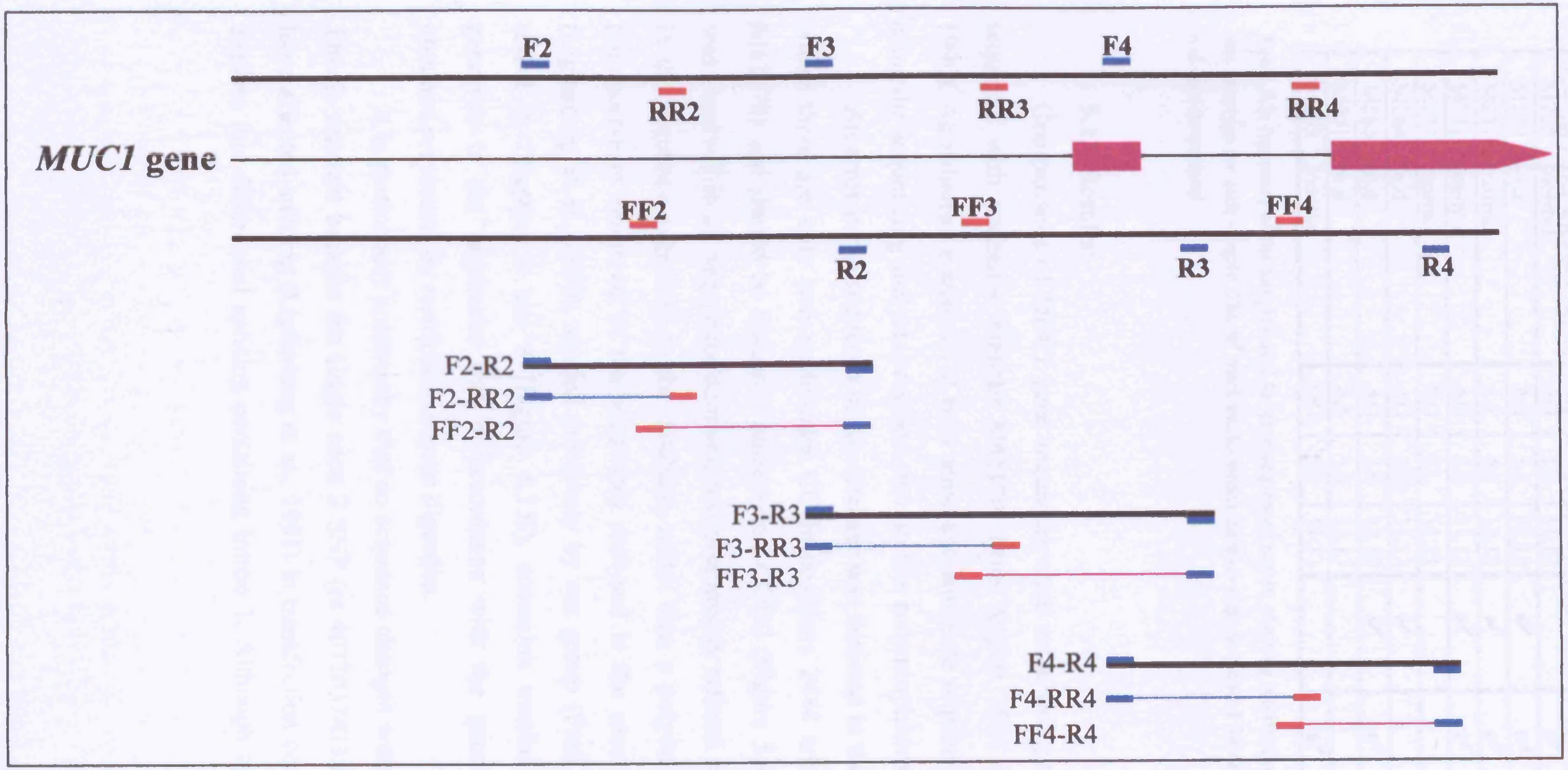


Figure 5.1: Sequencing strategy for the 5` end of *MUC1* gene. The genomic region includes promoter region, exon 1, intron 1 and the beginning of exon 2 (exons represented in purple boxes and promoter region and intron represented as a line). The forward and reverse primers for the initial PCRs are represented as blue lines and named F or R (for forward and reverse, respectively) followed by a number (2-4). The internal primers are represented as red lines and named FF or RR (for forward and reverse, respectively) followed by a number (2-4). These internal primers were used to amplify and sequence regions that were not easily read using only the flanking primers.

Sample	Disease status	Genotypes			Region 2	Region 3	Region 4
M22	*H. pylori* gastritis	GG	SS	11,11			✔
M44	*H. pylori* gastritis	GG	SS	11,11	✔	✔	✔
M189	normal	AG	SS	11,12			✔
M205	normal	AG	LL	11,12		✔	✔
M230	n.d	AG	LS	12,12	✔	✔	✔
M231	normal	AA	SS	12,12	✔	✔	✔
M232	normal	AG	LS	11,12	✔	✔	✔
M233	normal	GG	LL	11,11			✔
M264	n.d	AG	SS	11,12	✔		✔
M266	n.d	AG	LS	12,12	✔	✔	✔
M271	n.d	AA	LS	12,13			✔
M293	normal	GG	SS	11,11		✔	✔

Table 5.2: Summary of the samples used for each sequenced region, showing information about disease status and genotype for each sample. The ✔ mark marks which samples were sequenced for each region.
n.d.- not determined

5.1.1 Results:

Comparisons of *MUC1* gene sequences were made using as a reference the sequence with accession number M61170, from August 2001 (Lancaster et al., 1990). An exhaustive analysis of both sense and antisense sequences was performed using the sequencing analysis programs, but no new polymorphisms were found.

An error in the sequence in the database was detected in the promoter region where there are only two nucleotides CC in positions 2444 and 2445 (sequence M61170) and should be instead 4 nucleotides, CCCC (Figure 5.3.A). This change was observed in all sequenced samples, and presumably reflects a sequencing error in the sequence submitted to the database, rather than a polymorphism. The only polymorphism observed in the sequences analysed is the exon 2 G3506A SNP (Ligtenberg et al., 1990), studied previously by our group (Pratt et al., 1996) and tested in Chapters 3 and 4 (Figure 5.3.B), accession number rs4072037. The genotype in the sequences was in accordance with the genotypic information obtained previously by restriction enzyme digestion.

It is particularly noteworthy that no sequence changes were seen in intron 1. This is relevant because the single exon 2 SNP (rs 4072037/G3506A) was said to have affected splicing (Ligtenberg et al., 1991) in transfection constructs studies to explore this differential splicing containing intron 1. Although the Hilkens' group had claimed there was no difference in this region, this was a particularly difficult sequence and it was reassuring to obtain good sequence runs to confirm this.

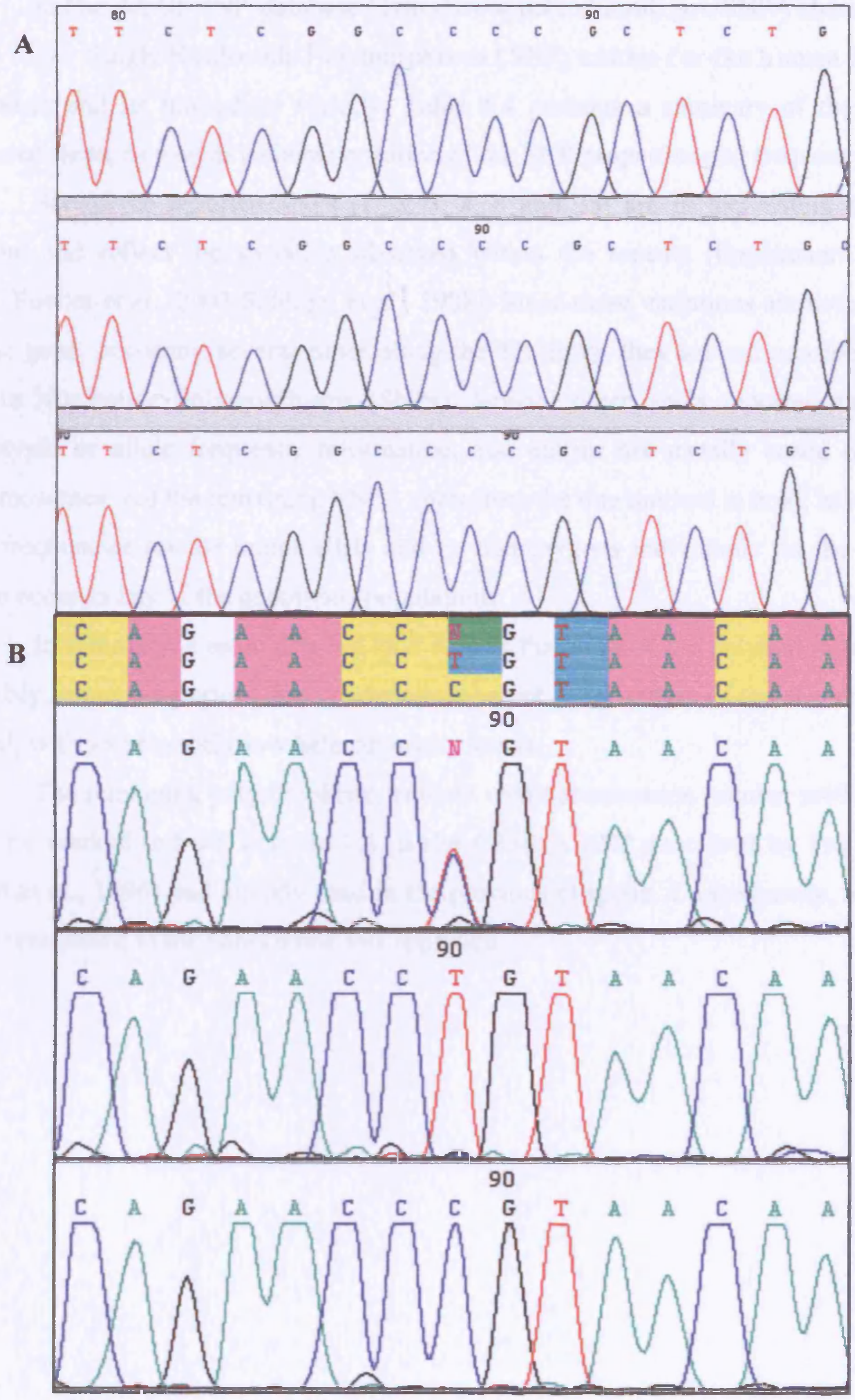


Figure 5.3: Sequencing chromatograms obtained using ChromasPro software. Figure A illustrates 3 different samples showing 4 C nucleotides instead of the 2 Cs found in the database sequence M61170 used as reference. Figure B shows an example of polymorphism in exon 2 detected in the sequence reaction in the antisense strand for 3 different samples and the 3 possible genotypes: CT, TT and CC, that correspond to the GA, AA and GG genotypes in the sense strand.

5.2 Search of NCBI SNP database for *MUC1* SNPs

In the NCBI SNP database (http://www.ncbi.nlm.nih.gov/SNP/) there are a total of 30 Single Nucleotide Polymorphisms (SNP) entries for the human *MUC1* sequence and its immediate vicinity. Table 5.4 presents a summary of the SNPs reported there, as well as a characterisation of the SNP properties and frequency.

Six of the reported SNPs (1, 2, 3, 4, 5 and 25) are in fact within the TR region, and reflect the variation observed within the repeats (Engelmann et al., 2001;Fowler et al., 2003;Siddiqui et al., 1988). Since those variations are not unique in the gene, occurring several times along the TR array, they are not considered as Single Nucleotide Polymorphisms (SNPs). Several other SNPs reported have no genotypic or allele frequency information, and entries are usually based on two chromosomes. All the remaining SNPs, apart from the one marked in bold, have very low frequencies for the minor allele and no homozygous individuals for the minor allele occur in any of the genotyped populations.

In summary, a more detailed look reveals that most of the polymorphisms are possibly sequencing errors due to artefacts inherent to the nature of the sequence, or if real, with an extremely low heterozygosity levels.

The remaining polymorphism, entered with the accession number rs4072037, the one marked in bold in table 5.4, is the G3506A SNP described by Pratt *et al* (Pratt et al., 1996) and already used in the previous chapters. Consequently, no new SNP was added to the panel using this approach.

	Acession number	Polymorphism	Validation status	Location	Number of tested chromosomes	Number of tested populations	Minor allele frequency	Amino-acid change
1	rs12743084	C/G	not validated	TR domain	2	0	n.d	
2	rs12742111	G/T	not validated	TR domain	2	0	n.d	
3	rs12737964	G/T	not validated	TR domain	2	0	n.d	
4	rs12737963	G/T	not validated	TR domain	2	0	n.d	
5	rs12728535	G/T	not validated	TR domain	2	0	n.d	
6	rs12566541	G/T	not validated	intron 6	2	0	n.d	
7	rs12411216	A/C	not validated	promoter region	2	0	n.d	
8	rs11807598	G/T	not validated	promoter region	3	0	n.d	
9	rs11590286	C/T	not validated	after 3' UTR	2	0	n.d	
10	rs11465211	A/G	validated	after 3' UTR	178	1	0.006	
11	rs11465210	A/G	validated	after 3' UTR	178	1	0.006	
12	rs11465209	C/T	not validated	exon 7	180	1	0.006	
13	rs11465208	-/C	validated	intron 6	172	1	0.012	
14	rs11465207	A/G	validated	exon 5	170	1	0.012	Asn>Ser
15	rs11465206	C/T	not validated	intron 2	166	1	0.006	
16	rs11465205	A/G	not validated	exon 2	154	1	0.006	Ser> Gly
17	rs11465204	A/G	validated	promoter region	158	1	0.013	
18	rs11465203	C/T	validated	promoter region	180	1	0.011	
19	rs11465202	C/G	not validated	promoter region	178/90/88/120	4	0.006/0/0/0.025	
20	rs11465201	A/C	not validated	promoter region	178	1	0.006	
21	rs7365893	G/T	not validated	promoter/ intergenic region	15	0	n.d	
22	rs7364559	C/G	not validated	promoter/ intergenic region	11	0	n.d	
23	rs6664861	C/G	not validated	intergenic MUC1/THBS3	2	0	n.d	
24	rs6427222	A/G	not validated	intergenic MUC1/THBS4	44/42/38	3	0/0.095/0	
25	rs4971063	C/G	not validated	TR domain	2	0	n.d	
26	**rs4072037**	A/G	validated	exon 2	114/90/88/120/136	5	0.421/0.200/0.0159/0.492/0.353	
27	rs1611773	C/T	validated	3' UTR	176/178/184	3	0.006/0.006/0.07	
28	rs1611772	C/T	validated	3' UTR	174/184/180	3	0.006/0.02/0.006	
29	rs1611771	A/G	not validated	intron 5	180/184/170	3	0.006/0/0.006	
30	rs1611770	A/G	validated	exon 4	176/170/118/90/88	5	0.023/0.018/0/0/0	Met> Val

Table 5.4: Summary of the SNPs reported in the NCBI dbSNP. Each entry is identified with the accession number from the database, location in the gene, number of chromosomes genotyped, allele frequencies and amino-acid change. The G3506A SNP, used in the previous chapters, is marked in bold in entry number 26.

5.3 Characterisation of the 100kb genomic region housing *MUC1*:

At the onset of this project *MUC1* was not present on the human genome sequence despite the fact that the full gene structure had been known since 1990 (Gendler et al., 1990;Lan et al., 1990;Lancaster et al., 1990). Although the gaps in the sequence were subsequently filled it was important to carefully check the genomic region and understand the disposition of the coding exons and flanking sequences. The 100kb region, shown in detail in figure 5.5, is very rich in genes, containing a total 8 genes and 2 pseudogenes. These are summarised below.

EPGL1:

EPGL1, also known as *LERK-1* or *EFNA1* (*EPGL1* being the accepted gene name according to HUGO nomenclature) is encoded by 5 exons. It encodes a protein with MW 22,000 called Ephrin-A1, made up of 187 amino acids and with one potential N-glycosylation site. Ephrin-A1 belongs to a subfamily of receptor protein-tyrosine kinases and was first described by Dixit *et al* (Dixit et al., 1990) as a novel TNF-induced gene in endothelial cells. It is induced by proinflammatory agents and was show to play a role in angiogenesis (Holzman et al., 1990).

Ephrins can be divided in two subclasses. EFNA1 belongs to the ephrin-A class, a group of ligands that are membrane anchored by a glycosylphosphotidylinositol (GPI) linkage. It has affinity to several receptor tyrosine kinases from the EphA receptor family.

LOC55974:

The *LOC55974* locus was initially found in a cDNA library (RefSeq in NCBI database: NM_018845). The gene and intron/exon boundaries were electronically predicted and characterised as having 6 exons spanning a region of 3Kb. The gene was placed between *EFNA1* and *DPM3*. It encodes a MW 19,000 protein with 167 amino acids, and is predicted to be a membrane protein, called "stromal cell protein". The protein has conserved domains homologous to pfam03083, a saliva protein family in drosophila. However, the function of the protein encoded by *LOC55974* remains unknown.

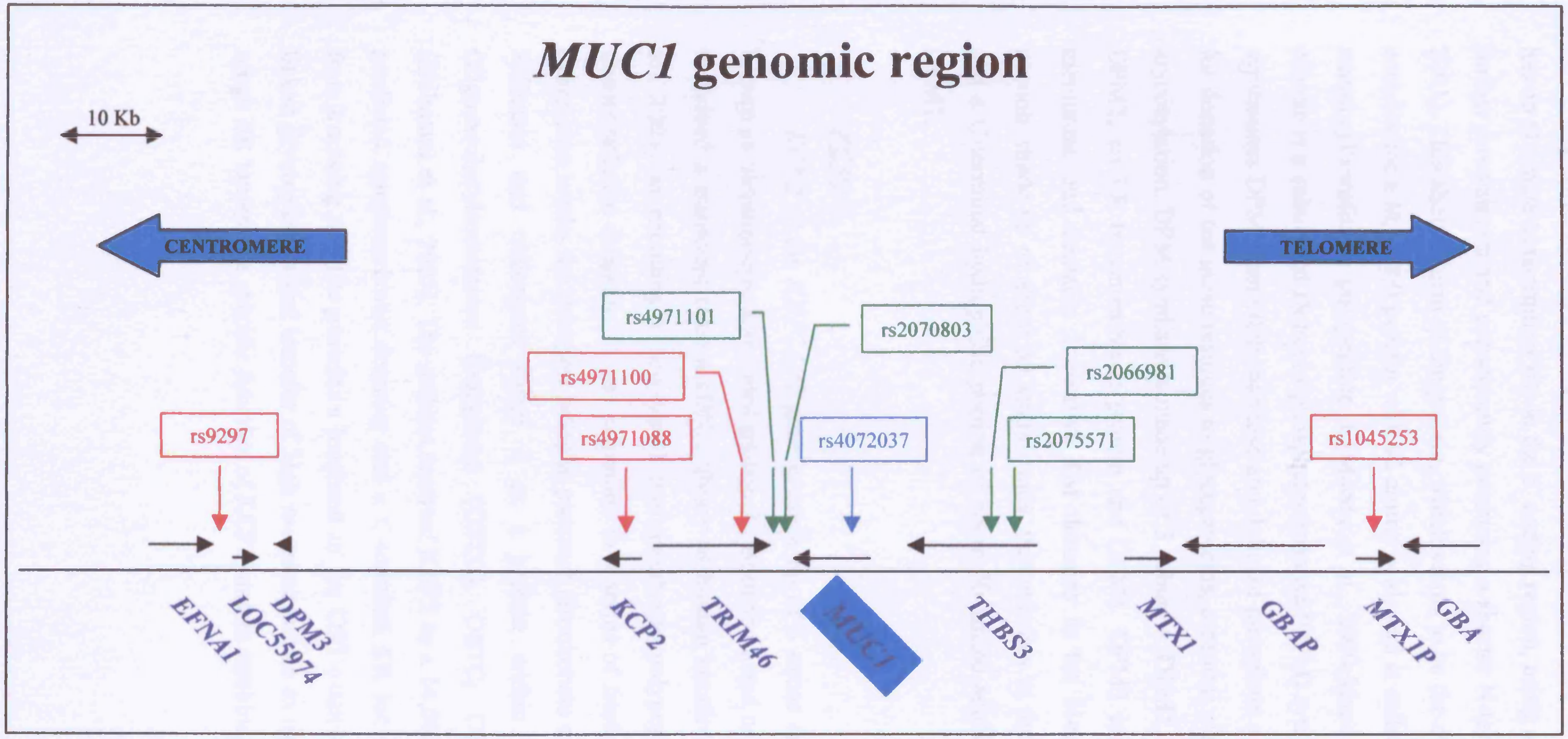


Figure 5.5: 100Kb genomic region surrounding *MUC1* displaying the 8 genes and 2 pseudogenes and their approximate size and orientation of transcription. *MUC1* is coloured in the middle of the figure in the blue box. The SNPs described in this chapter are represented as arrows facing downwards and coloured according to the way they were found: the blue arrow represents the only SNP known at the beginning of the project, and is located in exon 2 of *MUC1*; the green arrows indicate SNPs found using the NCBI dbSNP database, described later in this chapter; the red arrows indicate the 4 SNPs chosen from the HapMap database, whose analysis and characterisation will also be presented later in this chapter.

DPM3:

DPM3 is made up of one exon and can produce two isoforms, in which one has an alternate transcription site in the 5` coding region, using an initiating ATG site further downstream and consequently producing a shorter N-terminus (Manos et al., 2001). This shorter form of the protein, which seems to be the most common variant, encodes for a M_r 8,000 protein with 92 amino acids and is called dolichol-phosphate mannosyltransferase polypeptide 3 (Maeda et al., 2000;Manos et al., 2001). This protein is a subunit of Dolichol-phosphate-mannose (DPM) synthase, an enzyme that synthesises DPM from GDP-mannose and dolichol-phosphate a mechanism essential for donation of mannose residues to glycoproteins, a central mechanism in N-linked glycosylation. DPM synthase is made up of 3 subunits: DPM1, the catalytic subunit, DPM2, an ER transmembrane protein and DPM3. DPM3 is localised in the ER membrane and contains 2 putative TM domains in the N-terminal hydrophobic portion, made up of about 60 amino acids, that attaches to the TM subunit DPM2, and a C-terminal hydrophilic portion of about 30 amino acids that associates with DPM1.

KCP2:

KCP2 or the *KRTCAP2* gene is made up of 5 exons and encodes a protein known as 'keratinocyte associated protein 2'. Bonkobara and collaborators originally described a transcript from a cDNA library of human keratinocytes (Bonkobara et al., 2003) as encoding a non-type I transmembrane polypeptide with 2 putative transmembrane domains. It was expressed in a series of tissues, showing different expression levels, the strongest being in pancreas (Bonkobara et al., 2003). Later on, Shibatani and colleagues found it as a protein within the three types of Oligosaccharyltransferase Complexes (OSTC), $OSTC_I$ $OSTC_{II}$ and $OSTC_{III}$ (Shibatani et al., 2005). The authors reported KCP2 as a 14,000 MW protein with 4 predicted transmembrane domains and a C-terminal ER localisation motif. Apart from knowing that the protein is localised in the OST complexes, involved in N-linked glycosylation and transfer of high mannose sugars to nascent peptides in the rough ER lumen, the specific function of KCP2 protein remains unknown.

TRIM46:

TRIM46 gene is still not well described, though its cDNA sequence is available. The gene consists of 10 exons, and so far 4 isoforms have been reported (predicted from the cDNA), though only one is confirmed. The protein is called 'tripartite motif-containing 46' and belongs to the TRIM/RBCC family containing several of the conserved elements characteristic of this family: B-Box C-terminal domain, RING-finger containing E3 ubiquitin ligase, B-Box-type zinc finger, fibronectin type 3 domain and a SPRY domain (NCBI database, refseq NM_025058, 14th May, 2005). It is believed to be involved in protein ubiquitination with ligase activity within the ubiquitin ligase complex.

THBS3:

THBS3 is the gene located immediately upstream *MUC1*. It has 23 exons that span a 12kb region, most of it intronic sequence (Adolph et al., 1995). The exons in the gene are coded as letters (from A to F) and numbers (from 11 to 22), the numbered exons being homologous to exons in *THBS1* and *THBS2*. The 5` UTR is relatively small, made up of only 21 nucleotides. However, no TATA box was present in the region, suggesting that other transcription start sites (TSS) may occur. This hypothesis was confirmed later on by the same group in 1999 (Adolph and Bornstein, 1999), when they found mRNAs with transcription initiation sites in the THBS3/MTX1 intergenic region further 5` from that previously identified. A new exon was found 2498bp upstream exon A and was named exon A`. This new exon is 209bp long and 137bp away from the MTX1 TSS.

The protein Thrombospondin 3, encoded by *THBS3*, has 956 amino acids and is made up of a short hydrophobic region in the N-terminal signal peptide sequence and an extensive hydrophilic region (Adolph et al., 1995). It has 4 Type II (EGF-like) repeat domains and 7 Type III (Ca^{2+}-binding) repeat domains. Interchain disulphide bonds are likely to form through the two cysteines, and glycosylation of the protein may also occur via linkage to asparagine residues. Though the exact function of THBS3 in not known, it was hypothesised that it shares some function roles with

THBS1 and THBS2 in aggregation of activated platelets, cell proliferation, embryogenesis, morphogenesis and cell shape and movement, but possibly plays a more specialised role in cell-extracellular matrix interactions (Adolph et al., 1995).

MTX1:

MTX1 is localised immediately upstream the *THBS3* gene, but in the opposite orientation, therefore sharing a common upstream region. *MTX1* is made up of 8 exons, and is approximately 6kb long, with the translation initiation codon just 1374 nucleotides away from the ATG of *THBS3*. Adolph and colleagues, failed to find alternative transcripts at the 5`end of *MTX1*, suggesting that the initiation site for this gene is not variable (Adolph et al., 1995). Alignment of the longer transcripts for the 2 genes, *THBS3* and *MTX1*, showed overlap of about 90bp between the 5` ends, suggesting that if in that area some important transcriptional regulatory elements were present, that could have implications for tissue specific levels of expression, where the 2 genes would co-regulate each other.

The encoded protein, Metaxin 1 is 317 amino acids long and a mitochondrial outer membrane protein that functions as an import receptor for mitochondrial preproteins (Armstrong et al., 1997). Studies with *Mxt1* in mouse show that this seems to be an essential protein, since a knockout mouse exhibit embryo lethality phenotype (Bornstein et al., 1995). Also, mutations in *MTX1* have been reported to be associated with Gaucher disease, although the enzyme deficient in these patients being glucocerebrosidase, encoded by the *GBA* gene, immediately adjacent to *MTX1* (LaMarca et al., 2004).

GBA:

The β-glucocerebrosidase protein is encoded by *GBA*, a gene that contains 11 exons and is 7kb long (Horowitz et al., 1989;Reiner et al., 1988). The gene encodes an enzyme protein found in the lysosomal membrane responsible for cleavage of the β-glucosidic linkage of glycosylceramides (Grabowski et al., 1990). *GBA* has been widely studied due to its relationship with a lysosomal storage disorder, Gaucher disease, particularly frequent in the Ashkenazi Jewish population. Numerous mutations including point mutations, insertions, deletions and splicing mutations are described for the several types of the disease (Beutler, 1993;Stone et al., 2000).

Pseudogenes:

In the genomic area between *GBA* and *MTX1* there are two pseudogenes that arose as duplications of those two genes and were named *GBAP* and *MTX1P*, respectively (Long et al., 1996;Winfield et al., 1997). This duplication event is estimated to have occurred 40 million years ago (Winfield et al., 1997). The genes and respective pseudogenes are displayed as shown in figure 5.5. The 3` end of the duplication is within the *MTX1P* region and corresponds to exon 2 of the *MTX1* gene and is followed by a 5.5kb region homologous to the region immediately upstream *GBA*. However, the *GBAP* gene is separated from *MTXP1P* by a fragment of 11.8kb, corresponding to the 5.5kb region upstream *GBA* with a 6kb insert included containing 8 complete *Alu* sequences and several partial *Alu* sequences (Winfield et al., 1997).

Despite the high degree of homology between the genes and respective pseudogenes, unlike *GBAP* pseudogene, *MTX1P* does not seem to be transcribed, or if so, at very low levels (Long et al., 1996).

5.4 Search of databases for other SNPs in the 100Kb region

5.4.1 NCBI database- flanking markers

5.4.1.1 *THBS3*

Two polymorphisms were reported as characterised and validated for the *THBS3* gene sequence. The first SNP, entered with accession number rs2075571, is an A to G transition in intron 4 of the gene. It had, according to the database, an average estimated heterozygosity of 0.306 and a minor allele frequency (allele G) of 0.188. The second SNP, with accession number rs2066981, is a T to C transition in intron 8. The estimated average heterozygosity was 0.431 and the minor allele frequency (allele C) was 0.315.

Assays were developed for these two polymorphisms based on PCR and restriction enzyme digest, described in detail in Chapter 2. In figure 5.6 the localisation of each SNP is shown.

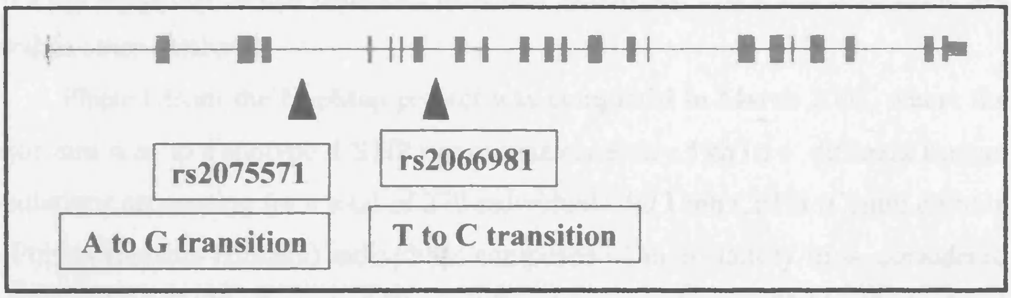


Figure 5.6: Schematic representation of *THBS3* gene and localisation of the 2 polymorphisms described, rs2075571 and rs2066981, in introns 4 and 8, respectively.

5.4.1.2 SNPs downstream of *MUC1*

Two SNPs located in the intergenic region between *MUC1* and *TRIM46* genes were reported as characterised and validated in the database (see figure 5.5). The SNP with accession number rs4971101 is a C to T transition and the two alleles have frequencies of about 0.5 and average heterozygosity of 0.499. The other SNP, rs2070803, is also a C to T transition. The estimated heterozygosity is reported as 0.500, and again the two alleles frequencies are close to 0.5.

These two polymorphisms, that are 81 nucleotides apart, can be typed using the same PCR product and two different restriction enzymes. The protocol developed to genotype the SNPs is described in detail in Chapter 2.

5.4.2 HapMap database- Selection of more distant markers using patterns of LD

The international HapMap Project (www.hapmap.org) is a database which main goal is to develop a haplotype map of the human genome. In this project, several countries around the world are involved, namely Japan, UK, Canada, China, Nigeria and USA, collaborating not only in the sequencing process but also in sample collection. The task of assembly and analysis of all the data is responsibility of several labs in UK and USA, sponsored by the multinational SNP consortium, "National Institutes of Health" (NIH) and "The Wellcome Trust". The aim of this project is to compile SNP information for the entire human genome and make it freely available to the research community. For this purpose, several thousands of SNPs are described in this database, including genotyped SNPs and others that are found in other databases.

Phase I from the HapMap project was completed in March 2005, where the major aim was to genotype 1 SNP approximately every 5kb in 4 different human populations accounting for a total of 270 individuals: 90 Utah CEPH (Centre d'Etude du Polymorphisme Humain) individuals composed of in 30 family trios, considered representative of a Northern and Western European population (CEU), 45 unrelated Han Chinese individuals collected in Beijing (HCB), 45 unrelated Japanese individuals collected in Tokyo (JPT), and 90 Yoruba individuals organised in family trios collected in Ibadan, Nigeria (YRI).

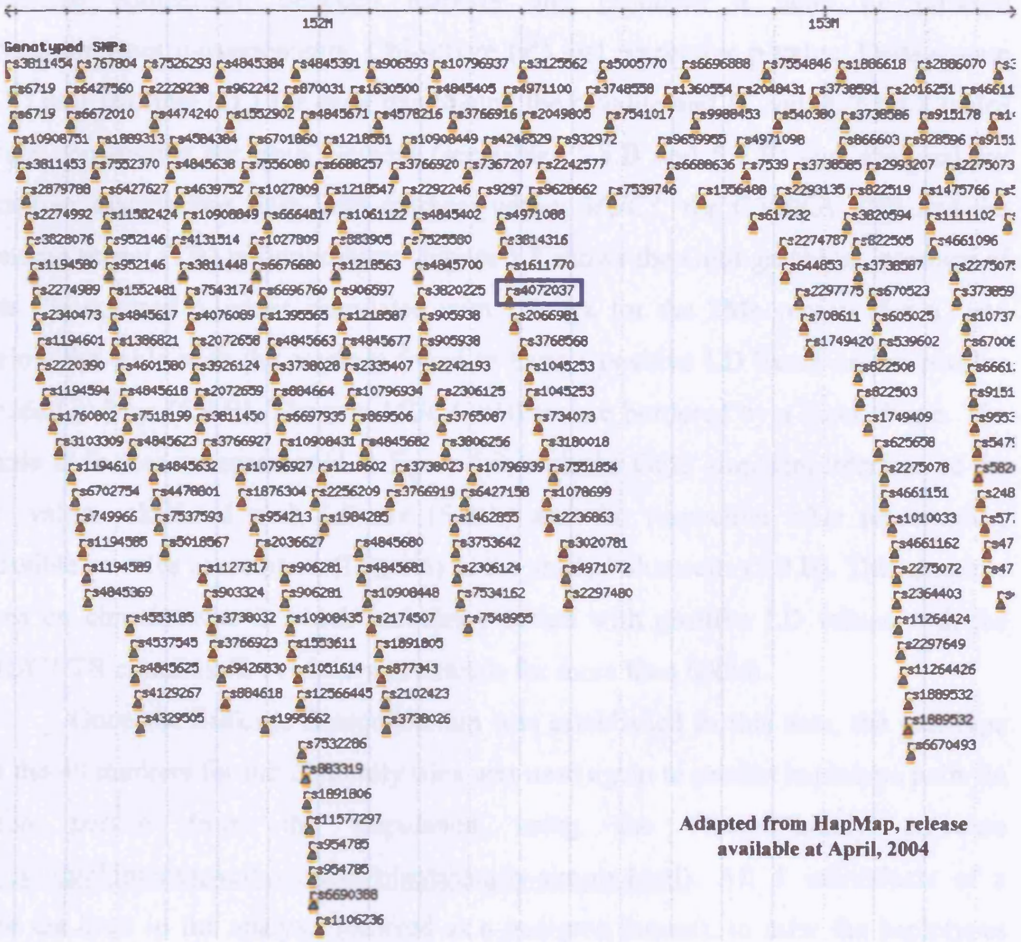
For each genotyped SNP frequency data as well as individual data for each individual are freely available.

5.4.2.1 HapMap in chromosome 1

For this project, HapMap was used to characterise the patterns of LD in the genomic area around *MUC1* and search for suitable polymorphisms to test. In the present study, only individuals from the CEPH cohort were used, because this work was done before the other 3 populations were included in the HapMap database. Also, the population for which this study was conducted, and that will be presented in Chapter 6, is of Northern European extraction, and therefore comparable to the Utah-CEPHs used in HapMap, thought to be originally from Northern European origin.

SNP data was retrieved from HapMap from an area of 2Mb, approximately 1Mb each side of *MUC1*, comprising a total of 175 SNPs, shown in figure 5.7. From these, 38 SNPs were excluded from the analysis: 20 SNPs were not polymorphic, at least for this population, 10 SNPs had a minor allele frequency of 0.03% or lower and 8 SNPs were repeated in the database and therefore one of the copies was excluded. The remaining 137 SNPs were used for the analysis.

Tandem Repeat size data produced previously in our laboratory by Ms Wendy Pratt was available for some of the CEPH families and integrated with the SNP data to improve the analysis. Because only 21 families in this study had this TR information available, the other 9 families were not used. Also, there was an apparent typing error in one of the remaining families for one of the markers, and consequently was excluded.



Adapted from HapMap, release available at April, 2004

Figure 5.7: HapMap database adaptation from the 2Mb region flanking the *MUC1* gene. The figure represents all the genotyped SNPs for the area and respective rs number. The G3506A SNP, present in *MUC1* is marked in a box for better recognition.

The final database analysed was comprised of 138 markers (137 SNPs plus the TR array) typed for 60 individuals, 20 Utah-CEPH family trios.

Data was arranged in a pedigree format and run in the LDmax program, part of GOLD statistical software package (http://www.sph.umich.edu/csg/abecasis/GOLD/index.html) to determine the Linkage Disequilibrium between the markers in the region. LDmax infers the haplotype frequencies in the population based on the genotypes from the founders using the EM algorithm (Excoffier and Slatkin, 1995) and uses these to run a pairwise comparison between markers and produces a table of Linkage Disequilibrium measurements: Chi-square (χ^2) and respective p-value, Delta-square (D^2) and D-prime (D`) for each pair. Using the p value and D` value, 2 by 2 tables were constructed for each element (see tables 5.8.B and 5.9.B) and analysed for positive associations with both markers within *MUC1*, the G3506A SNP and the tandem repeat (TR) polymorphism. Figure 5.8 shows the Gold graphical interface of the Chi-squared p-values calculated with LDmax for the 2Mb region (5.8.A) and below the table with the markers found to have a positive LD based on the p-value shaded in blue (5.8.B). The two *MUC1* markers are bordered by a black frame. The same 2Mb area is represented in figure 5.9, with the Gold graphical interface of the D` value calculated with LDmax (5.9.A) and the respective table representing possible positive association (D`$\geq$ 0.4) in the shaded blue cells (5.9.B). This genomic area on chromosome 1 which includes markers with positive LD values with the *MUC1* TR contains 40 markers and extends for more than 600kb.

Once the Linkage Disequilibrium was established in this area, the genotype of the 40 markers for the 20 family trios was used again to predict haplotype pairs in each person from the population using the Phase-Phamily software (http://archimedes.well.ox.ac.uk/pise/phamily-simple.html). All 3 individuals of a trio are used in the analysis (entered as a pedigree format), to infer the haplotypes from the parents, who can be considered as a non-related group of individuals. The children are used in this program as a way of improving the haplotype prediction of the parents. The program predicted 52 different haplotypes, some differing only in one or very few markers. To improve the analysis, the haplotypes were sorted for the *MUC1* markers, first by TR array and then by the G3506A SNP, as shown in figure 5.10, in an attempt to find markers that would correlate with the TR array or the G3506A SNP.

A

B

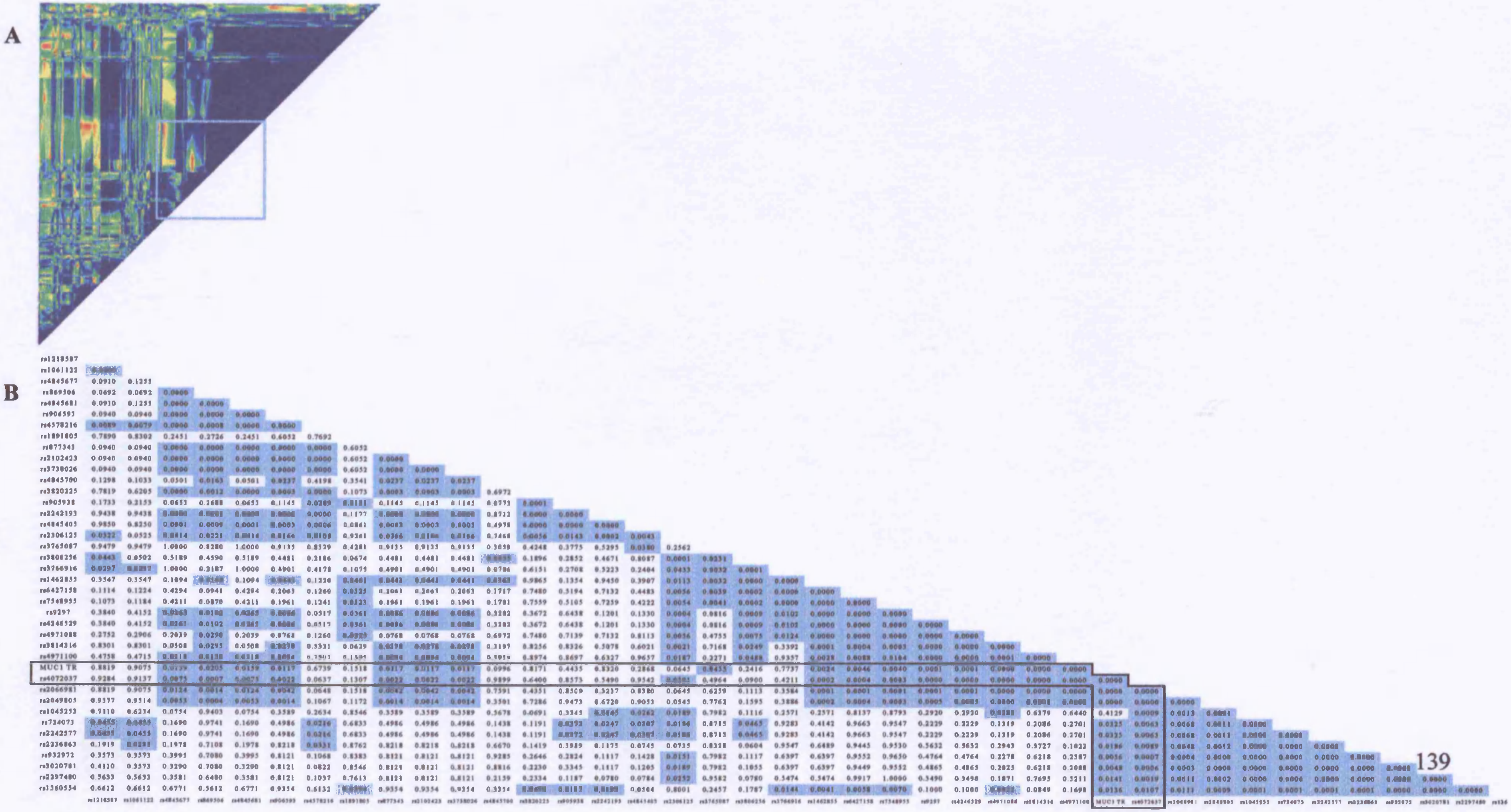
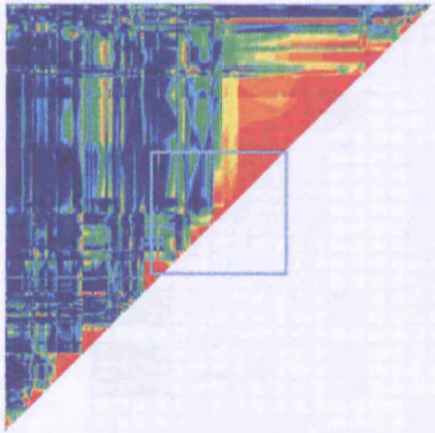


Figure 5.8: A. Chi-square p value diagram for significant LD between markers using GOLD graphical interface. The region within the lilac box is represented in detail in figure B, with significant associations between markers highlighted also in lilac. The 2 *MUC1* markers are surrounded by a black frame.

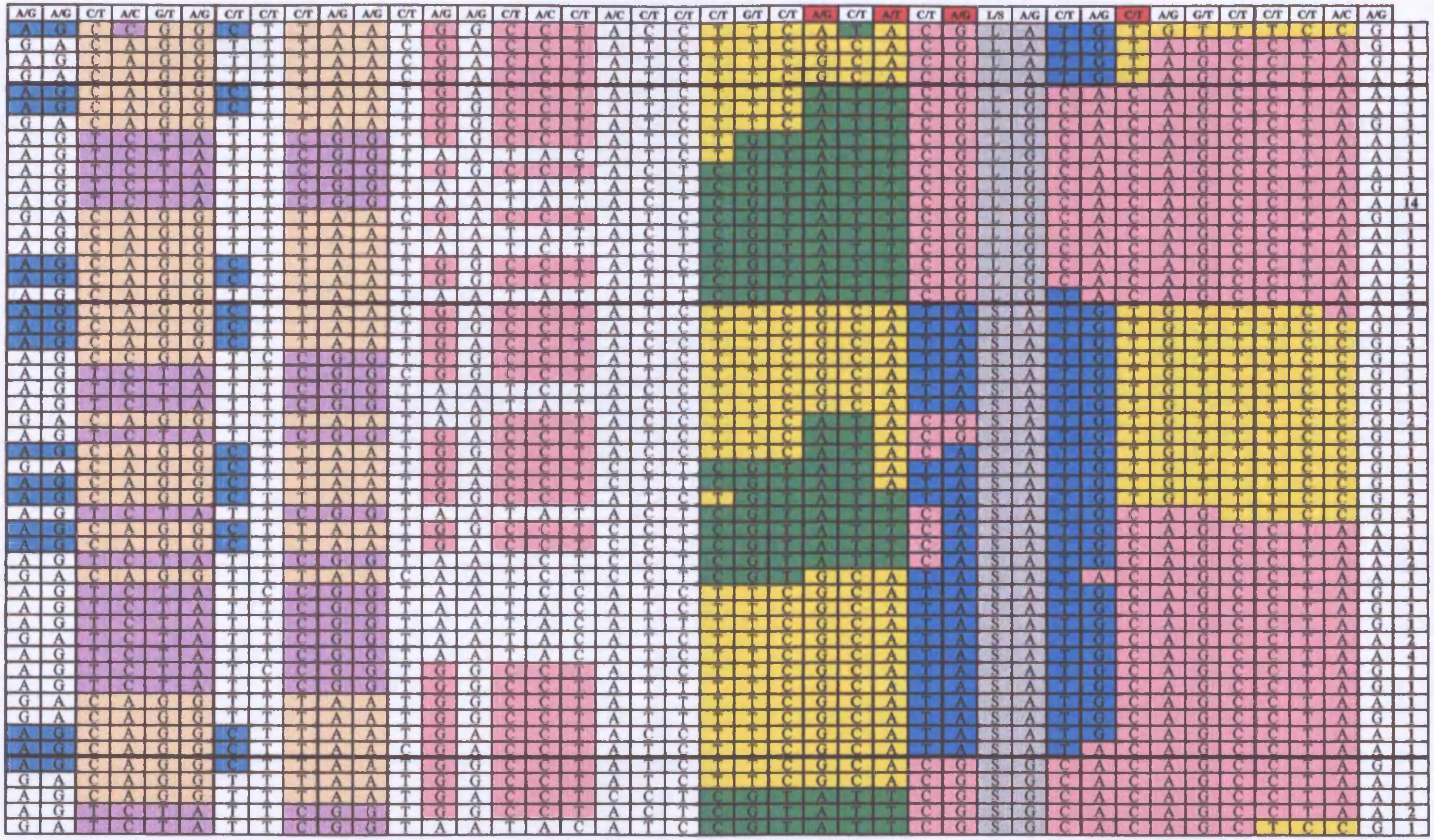
A



B

	rs1218587	rs1061122	rs4845677	rs869506	rs4845681	rs906593	rs4578216	rs1891805	rs877343	rs2102423	rs3738026	rs4845700	rs3820225	rs905938	rs2242193	rs4845405	rs2306125	rs3765087	rs3806256	rs3766916	rs1462855	rs6427158	rs7548955	rs9297	rs4246529	rs4971088	rs3814316	rs4971100	MUC1 TR	rs4072037	rs2066981	rs2049805	rs1045253	rs734073	rs2242577	rs2236863	rs932972	rs3020781	rs2297480	rs1360
rs1218587																																								
rs1061122	1.0000																																							
rs4845677	0.4790	0.4490																																						
rs869506	0.4990	0.4990	1.0000																																					
rs4845681	0.4790	0.4490	1.0000	1.0000																																				
rs906593	0.4750	0.4750	1.0000	1.0000	1.0000																																			
rs4578216	1.0000	1.0000	1.0000	0.8870	1.0000	1.0000																																		
rs1891805	0.4880	0.3980	1.0000	1.0000	1.0000	0.9980	0.2000																																	
rs877343	0.4750	0.4750	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980																																
rs2102423	0.4750	0.4750	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980	1.0000																															
rs3738026	0.4750	0.4750	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980	1.0000	1.0000																														
rs4845700	0.3240	0.3390	1.0000	1.0000	1.0000	1.0000	0.9700	1.0000	1.0000	1.0000	1.0000																													
rs3820225	0.0710	0.1370	0.6110	0.5580	0.6110	0.5780	1.0000	1.0000	0.5780	0.5780	0.5780	0.2710																												
rs905938	0.3960	0.3650	0.4260	0.2450	0.4260	0.3510	0.3380	1.0000	0.3510	0.3510	0.3510	0.6930	0.7920																											
rs2242193	0.0200	0.0200	0.6990	0.6880	0.6990	0.6840	1.0000	1.0000	0.6840	0.6840	0.6840	0.1130	1.0000	1.0000																										
rs4845405	0.0030	0.0580	0.7710	0.7330	0.7710	0.7490	1.0000	1.0000	0.7490	0.7490	0.7490	0.3720	1.0000	1.0000	1.0000																									
rs2306125	0.3610	0.3250	0.6810	0.7520	0.6810	0.7640	1.0000	0.0350	0.7640	0.7640	0.7640	0.3680	0.8000	1.0000	1.0000	0.6940																								
rs3765087	0.0580	0.0500	0.0000	0.1010	0.0000	0.0510	0.0620	1.0000	0.0510	0.0510	0.0510	1.0000	0.3190	0.4710	0.2650	1.0000	1.0000																							
rs3806256	[illegible]	[illegible]	0.1110	0.1400	0.1110	0.1480	0.3710	1.0000	0.1480	0.1480	0.1480	1.0000	0.3260	0.1960	0.1350	0.0530	1.0000	1.0000																						
rs3766916	[illegible]	0.1620	0.0000	0.3060	0.0000	0.2120	0.3250	1.0000	0.2120	0.2120	0.2120	1.0000	0.1270	0.3030	0.2020	0.4280	0.7220	1.0000	0.6680																					
rs1462855	0.3140	0.3140	0.3600	[illegible]	0.3600	0.4090	0.4470	1.0000	0.4090	0.4090	0.4090	1.0000	0.0020	0.3290	0.0150	0.2860	0.7760	1.0000	0.6770	0.5250																				
rs6427158	0.4720	0.4560	0.1560	0.3370	0.1560	0.2600	0.4210	1.0000	0.2600	0.2600	0.2600	1.0000	0.0560	0.1260	0.0780	0.2070	0.8000	1.0000	0.5320	0.5450	1.0000																			
rs7548955	0.4760	0.4400	0.1620	0.3440	0.1620	0.2670	0.4230	1.0000	0.2670	0.2670	0.2670	1.0000	0.0520	0.1280	0.0710	0.2190	0.8010	1.0000	0.5420	0.7410	1.0000	1.0000																		
rs9297	0.1810	0.1680	0.4430	0.5500	0.4430	0.5640	0.6130	1.0000	0.5640	0.5640	0.5640	0.4320	0.1780	0.1100	0.2970	0.2660	0.8470	0.6800	0.5320	0.1500	0.9220	0.9360	0.9360																	
rs4246529	0.1810	0.1680	0.4430	0.5500	0.4430	0.5640	0.6130	1.0000	0.5640	0.5640	0.5640	0.4320	0.1780	0.1100	0.2970	0.2660	0.8470	0.6800	0.5320	0.5500	0.9220	0.9360	0.9360	1.0000																
rs4971088	0.3120	0.3010	0.2550	0.4300	0.2550	0.3590	0.4210	1.0000	0.3590	0.3590	0.3190	0.2710	0.0560	0.0750	0.0780	0.0600	0.8000	0.2930	0.4180	0.5030	0.8190	0.8500	0.8470	1.0000	1.0000															
rs3814316	0.0660	0.0460	0.4340	0.4980	0.4340	0.5090	0.4800	1.0000	0.5090	0.5090	0.5030	0.4920	0.0410	0.0320	0.1160	0.0840	0.7030	0.1720	0.3900	0.2310	0.7360	0.6170	0.6210	0.7630	0.7630	0.8530														
rs4971100	0.1910	0.1930	0.4850	0.4760	0.4050	0.4950	0.1520	1.0000	0.4950	0.4950	0.4950	0.3390	0.0230	0.0290	0.0770	0.0090	0.6360	0.5730	0.3270	0.0160	0.3420	0.4170	0.4050	0.6350	0.6560	0.6290	1.0000													
MUC1 TR	0.0350	0.0260	0.4720	0.4370	0.4720	0.4920	0.1120	1.0000	0.4920	0.4920	0.4920	0.3360	0.0400	0.1910	0.0410	0.2210	0.6960	1.0000	0.2360	0.0610	0.5310	0.5450	0.5500	0.8160	0.8160	0.7910	1.0000	1.0000												
rs4072037	0.0270	0.0310	0.5850	0.6410	0.5850	0.6170	0.5170	1.0000	0.6170	0.6170	0.6170	0.0040	0.0930	0.0340	0.1440	0.0150	0.7250	0.3780	0.3050	0.1790	0.6090	0.6350	0.6390	0.8140	0.8140	0.8630	1.0000	1.0000	0.7690											
rs2066981	0.0350	0.0260	0.5370	0.5940	0.5370	0.5700	0.5290	1.0000	0.5700	0.5700	0.5700	0.1200	0.1560	0.0360	0.2330	0.0650	0.6960	0.2860	0.3270	0.2390	0.6760	0.7110	0.7150	0.8160	0.8160	0.8620	1.0000	1.0000	0.7370	1.0000										
rs2049805	0.0200	0.0150	0.5210	0.5870	0.5310	0.5640	0.4680	1.0000	0.5640	0.5640	0.5640	0.3770	0.0580	0.0140	0.0830	0.0260	0.7120	0.1820	0.2580	0.1910	0.5920	0.6090	0.6130	0.6670	0.6670	0.7500	0.8150	0.8630	0.7690	1.0000	1.0000									
rs1045253	0.2130	0.2760	0.5540	0.0260	0.5540	0.3350	0.2370	0.0670	0.3350	0.3350	0.3350	0.2050	0.4500	0.1460	0.6970	0.3800	1.0000	0.1680	0.5100	0.3800	0.5830	0.1770	0.0970	0.4630	0.4630	0.7310	0.2340	0.2140	0.3250	1.0000	1.0000	1.0000								
rs734073	1.0000	1.0000	0.4450	0.0090	0.4450	0.2270	[illegible]	0.1320	0.2270	0.2270	0.2270	1.0000	0.4040	0.3710	0.6500	0.7520	1.0000	0.0370	0.6200	0.0630	0.4230	0.0140	0.0210	0.4590	0.4590	0.5040	0.3630	0.4110	0.7810	1.0000	1.0000	1.0000	1.0000							
rs2242577	1.0000	1.0000	0.4450	0.0090	[illegible]	0.2270	[illegible]	0.1320	0.2270	0.2270	0.2270	1.0000	[illegible]	0.3710	0.6500	0.7520	1.0000	0.0370	0.6200	0.0630	0.4230	0.0140	0.0210	0.4590	0.4590	[illegible]	0.3630	0.4110	0.7810	1.0000	1.0000	1.0000	1.0000	1.0000						
rs2236863	0.5820	1.0000	0.3350	0.1130	0.3350	0.0630	0.3660	0.0570	0.0630	0.0630	0.0630	0.2580	0.3220	0.1300	0.4150	0.5020	0.6030	0.1400	0.4730	0.0180	0.1680	0.0170	0.0140	0.1630	0.1630	0.3140	0.1970	0.4640	[illegible]	0.8180	0.8340	0.8570	0.8150	1.0000	1.0000					
rs932972	0.4630	0.4430	0.2450	0.1170	0.2450	0.0680	0.3050	0.0730	0.0680	0.0680	0.0680	0.0360	0.2610	0.1710	0.4240	0.4450	1.0000	0.1680	0.4240	0.1790	0.1790	0.0150	0.0110	0.2220	0.2220	0.6020	0.1070	0.3820	0.8270	1.0000	1.0000	1.0000	0.8750	1.0000	1.0000	1.0000				
rs3020781	0.4240	0.4430	0.2750	0.1170	0.2750	0.0680	0.3300	0.0670	0.0680	0.0680	0.0680	0.0920	0.2810	0.1460	0.4240	0.4630	1.0000	0.1680	0.4350	0.1790	0.1790	0.0190	0.0150	0.2200	0.2200	0.4270	0.1070	0.4090	0.8540	1.0000	1.0000	1.0000	0.8860	1.0000	1.0000	0.9490	1.0000			
rs2297480	0.3190	0.3190	0.2980	0.1610	0.2980	0.0770	0.3420	0.1040	0.0770	0.0770	0.0770	0.9990	0.3020	0.2630	0.5080	0.5750	1.0000	0.0360	0.5080	0.2690	0.2680	0.0020	0.0000	0.3190	0.3190	0.4850	0.0720	0.2510	0.7980	1.0000	1.0000	1.0000	0.9310	0.9260	0.9260	0.9330	1.0000	1.0000		
rs1360554	0.1100	0.1100	0.1000	0.1470	0.1000	0.0180	0.1360	1.0000	0.0180	0.0180	0.0180	0.4820	0.3620	0.6260	0.4750	0.4480	0.0980	0.5450	0.2220	0.3530	0.5800	0.4940	0.4820	0.2460	0.2460	0.5650	0.2960	0.2820	0.5210	0.5960	0.5920	0.5610	0.6170	0.8820	0.8820	0.8040	0.8230	0.8230	1.0000	

Figure 5.9: A. LDmax D` value diagram for significant LD between markers using GOLD graphical interface. The region within the lilac box is represented in detail in figure B, with significant associations between markers highlighted also in lilac. The 2 *MUC1* markers are surrounded by a black frame.



A/G	A/G	C/T	A/C	G/T	A/G	C/T	C/T	C/T	A/G	A/G	C/T	A/G	A/G	C/T	A/C	C/T	A/C	C/T	C/T	C/T	G/T	C/T	A/G	C/T	A/T	C/T	A/G	L/S	A/G	C/T	A/G	C/T	A/G	G/T	C/T	C/T	C/T	A/C	A/G	
A	G	C	C	G	G	C	T	T	A	A	T	G	G	C	C	T	A	C	C	T	T	C	A	T	A	C	G	L	A	T	G	T	G	T	T	T	C	C	G	1
G	A	C	A	G	G	T	T	T	A	A	C	G	A	C	C	T	A	T	C	T	T	C	G	C	A	C	G	L	A	T	G	T	A	G	C	C	T	A	G	1
A	G	C	A	G	G	T	T	T	A	A	C	G	A	C	C	T	A	T	C	T	T	C	G	C	A	C	G	L	A	T	G	T	A	G	C	C	T	A	G	1
G	A	C	A	G	G	T	T	T	A	A	C	G	A	C	C	T	A	T	C	T	T	C	G	C	A	C	G	L	A	T	G	T	A	G	C	C	T	A	A	2
A	G	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	A	T	C	T	T	C	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	T	C	T	T	C	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
G	A	C	A	G	G	T	T	T	A	A	T	G	G	C	C	T	A	T	C	T	T	C	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	G	1
A	G	T	C	T	A	T	T	C	G	G	T	G	G	C	C	T	A	T	C	T	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	C	T	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	T	C	G	G	T	G	G	C	C	T	A	C	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	T	A	C	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	G	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	T	A	C	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	14
G	A	C	A	G	G	T	T	T	A	A	C	G	A	C	C	T	A	T	C	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	G	1
A	G	C	A	G	G	T	T	T	A	A	T	A	A	T	A	T	A	C	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	T	T	T	A	A	T	A	A	T	C	T	A	C	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	C	C	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	T	T	C	G	T	A	T	T	C	G	L	G	C	A	C	A	G	C	C	T	A	A	2
A	G	C	A	G	G	T	T	T	A	A	T	A	A	T	A	T	A	C	T	C	G	T	A	T	T	C	G	L	G	T	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	C	G	A	C	C	T	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	A	A	2
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	C	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	A	C	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	3
A	G	C	C	G	A	T	C	C	G	G	T	G	G	C	C	T	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	T	C	T	A	T	T	C	G	G	C	G	G	C	C	T	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	C	T	A	C	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	T	A	C	C	T	T	C	G	C	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
G	A	C	A	G	G	T	T	T	A	A	T	A	G	C	C	T	A	C	C	T	T	C	A	T	A	C	G	S	A	T	G	T	G	T	T	T	C	C	G	2
A	G	T	C	T	A	T	T	C	G	G	T	G	A	C	C	T	A	T	C	T	T	C	A	T	A	C	G	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	C	C	T	T	C	A	T	A	C	A	S	A	T	G	T	G	T	T	T	C	C	G	1
G	A	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	C	C	T	C	G	T	A	T	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	C	T	T	C	G	T	A	T	A	T	A	S	A	T	G	T	G	T	T	T	C	C	G	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	T	C	T	G	T	A	T	T	T	A	S	A	T	G	T	G	T	T	T	C	C	G	2
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	T	A	C	T	C	G	T	A	T	T	C	A	S	A	T	G	C	A	G	T	T	C	C	G	3
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	C	C	T	C	G	T	A	T	T	C	A	S	A	T	G	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	C	C	T	C	G	T	A	T	T	C	A	S	A	T	G	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	C	T	C	C	T	C	G	T	A	T	T	C	A	S	A	T	G	C	A	G	C	C	T	A	A	2
G	A	C	A	G	G	T	T	T	A	A	C	A	A	T	C	T	C	C	T	C	G	T	G	C	A	T	A	S	A	T	A	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	C	C	G	G	T	A	A	T	C	C	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	T	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
G	A	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	A	2
A	G	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	A	4
A	G	T	C	T	A	T	C	C	G	G	T	G	G	C	C	T	A	T	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
A	G	T	C	T	A	T	T	C	G	G	T	G	G	C	C	T	A	T	T	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
G	A	C	A	G	G	T	T	T	A	A	T	G	G	C	C	T	A	T	T	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	A	1
G	A	C	A	G	G	T	T	T	A	A	T	G	G	C	C	T	A	T	T	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	G	1
A	G	C	A	G	G	C	T	T	A	A	T	G	A	C	C	T	A	C	C	T	T	C	G	C	A	T	A	S	A	T	G	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	C	G	A	C	C	T	A	T	C	T	T	C	G	C	A	T	A	S	A	T	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	C	T	T	A	A	T	G	G	C	C	T	A	T	C	T	T	C	G	C	A	C	G	S	G	C	A	C	A	G	C	C	T	A	G	1
G	A	C	A	G	G	T	T	T	A	A	T	G	G	C	C	T	A	T	C	T	T	C	G	C	A	C	G	S	G	C	A	C	A	G	C	C	T	A	A	1
A	G	C	A	G	G	T	T	T	A	A	T	G	A	C	C	T	C	C	T	C	G	T	A	T	T	C	G	S	G	C	A	C	A	G	C	C	T	A	A	1
A	G	T	C	T	A	T	T	C	G	G	T	G	G	C	C	T	A	C	T	C	G	T	A	T	T	C	G	S	G	C	A	C	A	G	C	C	T	A	A	2
G	A	T	C	T	A	T	T	C	G	G	T	A	A	T	A	C	A	T	C	C	G	T	A	T	T	C	G	S	G	C	A	C	A	G	C	T	C	C	G	1

Figure 5.10: List of haplotypes predicted using Phase-Phamily software. The first 40 columns represent the 40 markers showing association with *MUC1* TR array and G3506A SNP. The right-hand column shows the number of chromosomes predicted for each haplotype. The cells are coloured showing blocks of LD present in this region. *MUC1* markers are shaded in grey and sorted by TR size and G3506A SNP.

Markers were chosen from the haplotype table (figure 5.10) to subdivide the simple 2 locus haplotypes (L A, L G, S A and S G) and to include one which appeared to associate better with LS than does G3506A SNP.

The 4 markers chosen from both sides of the *MUC1* gene are the ones marked in the first row in red in figure 5.10, covering an area of nearly 100kb and are described in more detail in table 5.11.

SNP reference number	Gene	Genomic region	Distance from MUC1 TR array	Alleles
rs9297	*EFNA1* - ephrin-A1	3`untranslated	-55kb	A/G
rs4971088	*KCP2* - keratinocyte associated protein 2	intronic	-19kb	A/T
rs4971100	*TRIM46* - tripartite motif-containing 46	intronic	-6kb	A/G
rs1045253	*MTX1P* - metaxin 1 pseudogene	not determined	+40kb	C/T

Table 5.11: Chromosome 1 markers chosen from the HapMap study to be genotyped. The table also displays the gene in the SNP area, genomic region in relation to the gene, distance from *MUC1* TR array and the two possible variants (alleles).

All new SNPs were typed using PCR followed by restriction enzyme digestion. The protocols used are described in detail in the methods chapter. Figure 5.12 illustrates as examples some of the PCR digests for the first three markers (rs9297, rs4971088 and rs4971100).

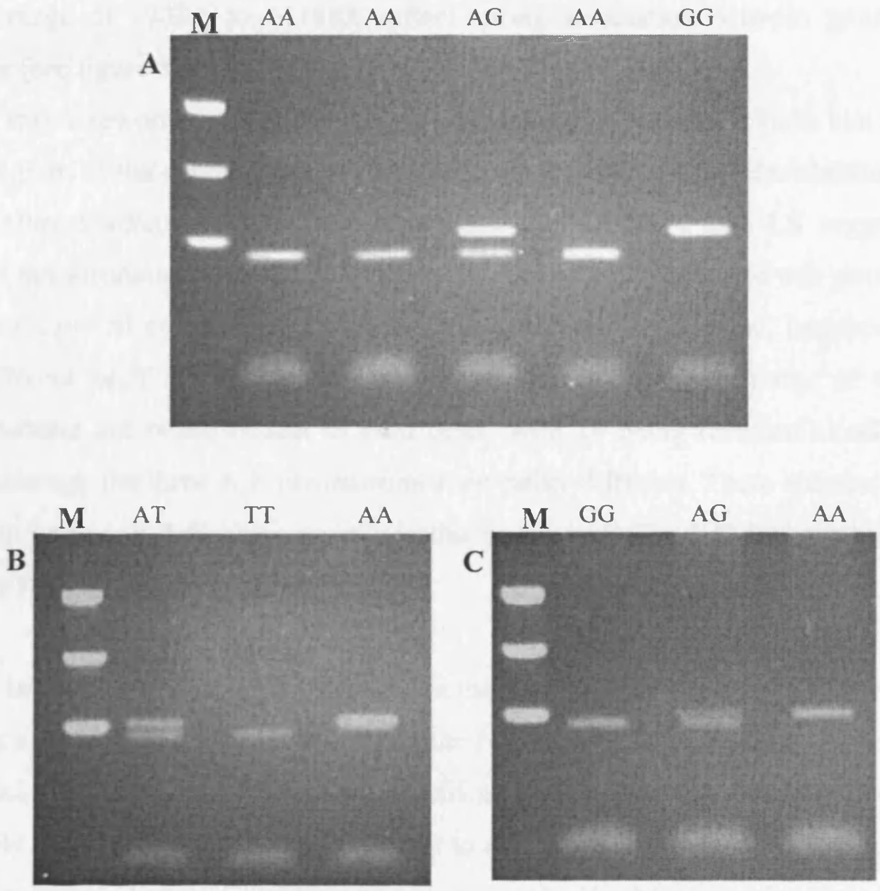


Figure 5.12: Photographs of agarose gels under UV light. The images illustrate the restriction enzyme digestion of the PCR products from the polymorphisms rs9297 (A), rs4971088 (B) and rs47971100 (C). All gels have a molecular weight marker on the left hand column, marked with a M. Figure A shows the genotypes of 5 different samples, containing the 3 possible variants, AA, AG or GG; figure B displays the 3 possible genotypes for the SNP rs4971088, AA, AT and TT; figure C also shows the 3 samples with the 3 possible variants for the polymorphism, AA, AG and GG.

5.5 Discussion:

It is clear from observation of the Linkage Disequilibrium maps (figures 5.8 and 5.9) that there is statistically significant positive association between the *MUC1* markers and other markers at a distance of about 300Kb each side.

Taking a closer look, and specifically analysing the possible haplotypes inferred with Phase-Phamily, blocks of haplotypes from between the markers that are in the range of –90Kb to +118Kb reflect strong association between groups of markers (see figure 5.10).

It is noteworthy that the *MUC1* gene region shows long LD blocks like many other regions of the genome, despite the existence of the potentially destabilising TR array. More markers are positively associated with G3506A than LS suggesting perhaps not surprisingly that much of the variability of the TR domain was generated by non-reciprocal events after the main haplotypes were established. Inspection of the different *MUC1* haplotype chromosomes show that the majority of the L chromosomes are rather similar to each other, with 14 being identical at all SNP sites, although the three A L chromosomes are rather different. There are also three common groups of A S chromosomes in this population. The 6 G S chromosomes, fall into two haplotype groups.

In Chapter 4 it has been shown that the two markers flanking the TR repeat domain are associated with each other in the Nigerians that were tested, suggesting LD blocks across this region also in an African group. Now that HapMap data are available for Nigerians it will be of interest to determine the pattern of LD blocks in this population. It should however be noted that the HapMap samples were taken from Yoruba, a different tribal group from the one available in our lab, that are mainly Ibibio.

In examining these extended regions of LD in the context of disease association studies it is important to consider other genes in the vicinity particularly in such a dense region. Functionally significant variations in these genes are likely to be or have been under selection and this will affect gene frequencies of whole haplotype blocks. It is noteworthy in this context that *GBA* allele frequencies are likely to be different in Ashkenazi Jewish populations where frequency for Gaucher disease is high and where these alleles may have been under selection. Such selection may have influenced *MUC1* allele frequencies.

One of the objectives of this analysis was to find markers that would highly correlate with the L/S variation of the *MUC1* TR array or with the recombinant G S haplotype. The other aim was to find markers to expand the haplotype analysis in the

Northern European gastritis cohort. For that, the new SNPs to be tested were chosen from within these LD block areas.

The first aim, choosing markers preferentially correlated with the TR was not successful, since all the SNPs in the area associated better with the G3506A SNP, even choosing a set of markers in both sides of *MUC1*.

The second aim, to select a suitable set of markers for extended haplotype analysis was successful, apart for one of the markers. The SNP with accession number rs1045253, located in the *MTX1P* pseudogene, was not used because the area where the variation occurred was 100% identical to the equivalent in the *MTX1* gene. In addition, the *MTX1* gene was not polymorphic in this site, making impossible to distinguish homozygous CC, the variant for which the restriction enzyme would cut, from the heterozygous CT.

CHAPTER 6

HELICOBACTER PYLORI GASTRITIS STUDY
A NEW COHORT FIVE YEARS AFTER THE ORIGINAL STUDY

Introduction

Results of previous studies showed that short alleles for the *MUC1* TR were more frequent in disease populations. Although not statistically significant in all groups there was in each case an increase in a particular *MUC1* haplotype: G S 11. The aim of this work was to see if these observations could be replicated in an independent study.

6.1 Characterisation of the samples

Volunteers were recruited over a period of 3 years, from 2002 to 2004, at the gastrointestinal clinic of Middlesex Hospital, part of University College London Hospitals. Volunteers were patients who had been referred for upper GI endoscopy for a variety of reasons, which included suspected gastritis/ulcer, oesophageal problems and less frequently for small intestinal complaints. Patients awaiting the procedure were approached and the project was explained to them, and they then gave written consent. A simple questionnaire was completed which covered age, sex, ancestry of parents/grandparents, smoking habit. In each case a blood sample was taken for the preparation of DNA and a series of biopsies was taken. According to the clinical needs of the patient, and the current circumstances in the clinic a variable number of biopsies were taken. These were used for conducting a CLO test (Marshall et al., 1987) which measures the amount of urease enzyme produced on a mounted gel pellet containing urea for routine histology, or frozen for future use for RNA.

6.1.1 Categorisation of the samples:

A total of 237 patients agreed to participate in the study, but of these 54 were not included in this part of the work because they were of non-Northern European ancestry. A further 29 were excluded because there was either no DNA (n=3) or no histology results (n=20) or for inconsistency in the collection of patients data.

Routine histology was used to determine gastritis status. A modified Sydney scoring method was used (Dixon et al., 1996), which assesses evidence of inflammation, atrophy and the occurrence of intestinal metaplasia. An illustration of the histological classification for the various disease groups is show in figure 6.1.

The presence of Helicobacter pylori was determined by routine histology, and where needed for confirmation, or in many cases where apparently negative, by immunohistology. Figure 6.2 shows gastric mucosa sections stained by immunohistochemistry for *Helicobacter pylori*. Figure 6.2 A shows a antral type mucosa with no *Helicobacter pylori* infection and figures 6.2 (B) show an example of a gastric mucosa positive for *H. pylori* infection, and a magnification of a gastric gland showing the bacilli very distinctly (C).

Patients were categorised into 4 groups, according to the histology report and *Helicobacter pylori* infection status. Individuals were assessed initially for *H. pylori* infection by CLO test and/or immunohistochemistry and considered as positive if positive for one or other test. In 4 cases there was a positive CLO test and *H. pylori* were not detected histologically. These slides were reviewed again and in one case a few organisms were detected.

The patients with *Helicobacter pylori* infection (n=34) all showed signs of gastric disease; 27 individuals with at least mild chronic inflammation also showed presence of atrophy and/or acute inflammation; presence of intestinal metaplasia was also recorded and scored, but all cases with IM (n=8) also had chronic inflammation and atrophy and/or acute inflammation of the stomach. Individuals with chronic inflammation and one of the other changes were considered in the same group, named '*H. pylori* gastritis'.

Patients showing the presence of no more than a few mononuclear cells (scored as mild, following the updated Sydney system) and negative for neutrophils, atrophy, intestinal metaplasia and *H. pylori* were considered normal. However, to eliminate the possibility of previous eradication therapy, most volunteers were also asked by telephone or in the clinic whether they had had previous antibiotic treatment. Those without current *H. pylori* who did not recall eradication therapy were considered as normal.

Seven people who had had previous eradication therapy were still positive for *H. pylori* and included in the *H. pylori* gastritis group. Those who had previously received eradication therapy and were free of *H. pylori* were considered as a separate group. It is noteworthy that 16 of this group of 44 individuals showed some evidence of gastritis, atrophy or intestinal metaplasia.

One patient had gastric cancer and *H. pylori* infection and was therefore included with the *H. pylori* gastritis disease group.

In summary, 59 were classified as normal, 33 as *H. pylori* gastritis, 44 as *H. pylori* eradicated and 18 as the fourth group, Gastritis with no *H. pylori* infection.

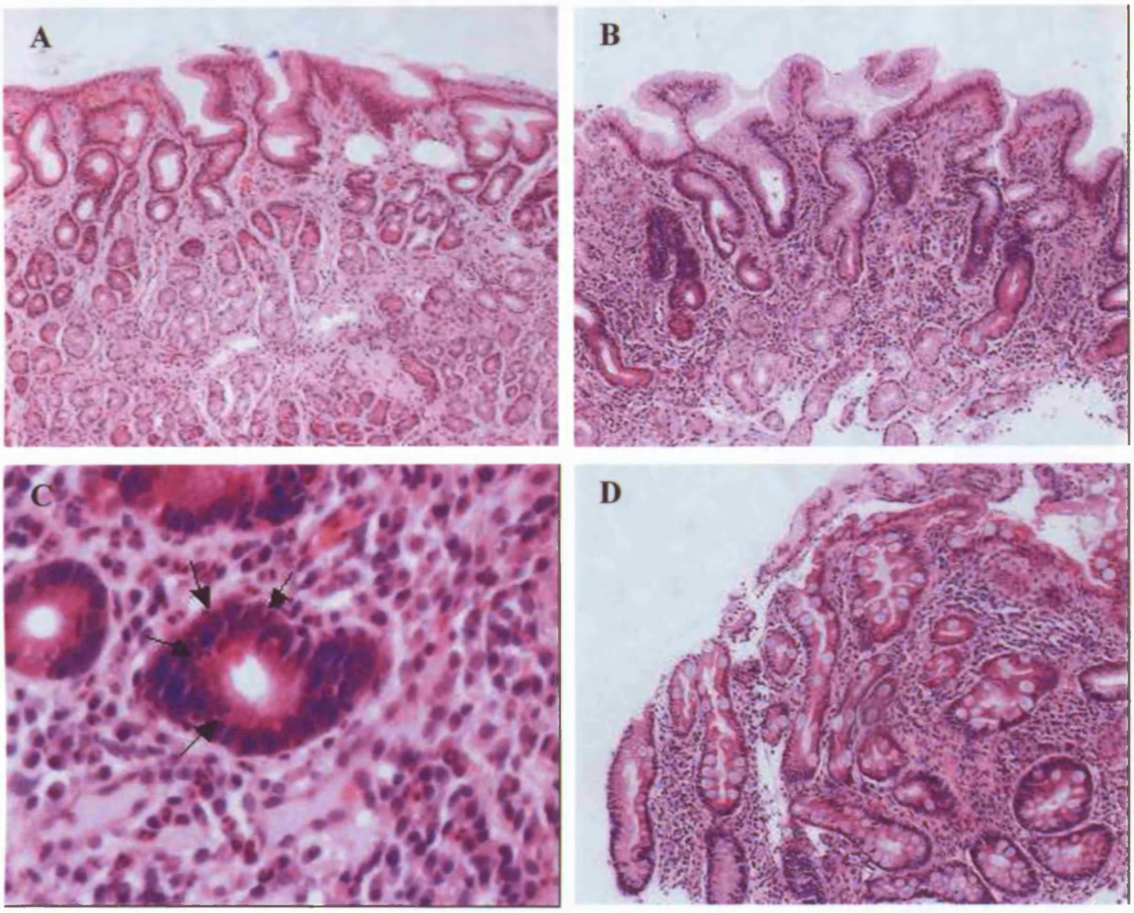


Figure 6.1: Haematoxylin-Eosin (HE) staining of human gastric mucosa showing progressive stages of gastric disease. (A) Antral type mucosa with mild reactive features but no inflammation, which was classified as normal. (B) Antral type mucosa exhibiting moderate chronic inflammation and moderate atrophy, classified as non-active chronic gastritis. (C) Antral type mucosa with acute chronic gastritis, with a large number of neutrophils marked with arrows. (D) Gastric mucosa with Intestinal Metaplasia (IM), exhibiting a colonic phenotype with a large number of goblet cells. Original magnification: x100 (A), x100 (B), x400 (C), x200 (D).

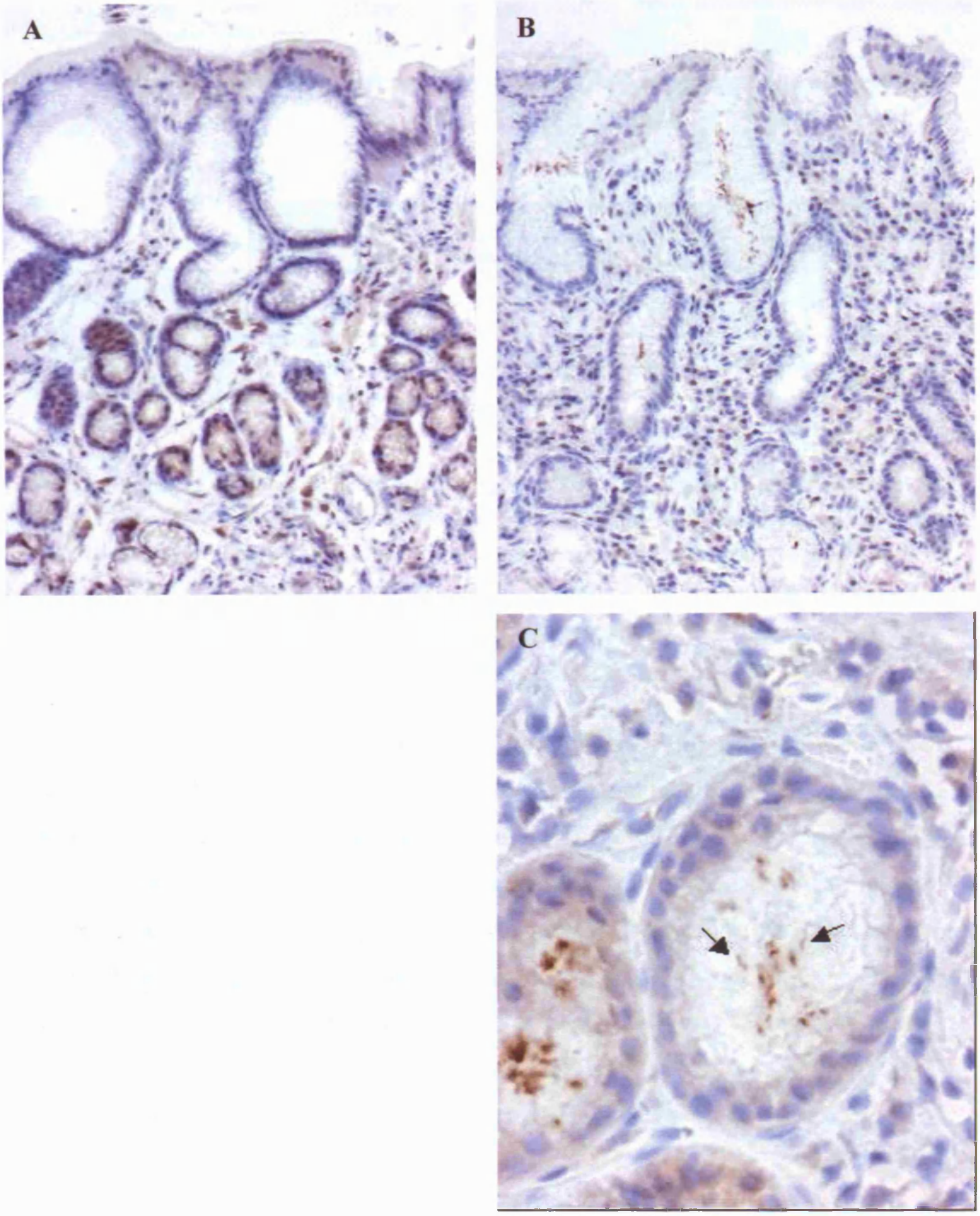


Figure 6.2: Human gastric mucosa stained by immunohistochemistry with a specific antibody against *Helicobacter pylori*. Figure A shows antral type mucosa negative for presence of *H. pylori*. Figures B and C show antral type mucosa exhibiting *Helicobacter pylori* organisms in gastric pits. Two *Helicobacter pylori* S-shaped bacilli are clearly shown by the arrows in the magnified gland in figure C. Original magnification: x200 (A) x200 (B) and x400 (C).

6.1.2 Characterisation of the patients within the groups

Table 6.3 summarises some of the patients' data on gender, age and smoking status.

	Gender		Age			Smoking status		
	Male	Female	Minimum	Maximum	Mean	Current	Ever	Never
H.pylori gastritis	0.55	0.45	26	84	59.72	0.52	0.09	0.39
Former *H. pylori* gastritis	0.52	0.48	18	88	52.14	0.37	0.12	0.51
Gastritis no *H. pylori*	0.56	0.44	25	80	54.71	0.33	0.00	0.67
Normal	0.49	0.51	18	79	52.22	0.41	0.05	0.53

Table 6.3: Summary of gender, age range and smoking status for the four groups: *H. pylori* gastritis (n=33), Former *H. pylori* gastritis (n=44), Gastritis no *H. pylori* (n=18) and Normal (n=59). For the smoking status, data was not available for 1 individual in the Former *H. pylori* gastritis and 1 individual in the Normal; Age was not available for 1 individual in the *H. pylori* gastritis group and 1 individual in the No *H. pylori* gastritis group.

For this study, the 3 disease groups show a slightly higher proportion of men, whereas in the Normal group, the proportion of men and women is very similar. The smoking status however, seems more variable, with the *H. pylori* group containing more smokers and almost 10% of ex-smokers. However, a Chi-square test revealed no significant differences between groups for the 2 variables. The *H. pylori* gastritis group showed a significant difference in age distribution from the Normal controls and Former *H. pylori* gastritis groups (T-test 2-tailed p= 0.0212 and p= 0.04463, respectively). The raw genetic data for each individual included in the 4 groups is shown in Appendix 2.

6.2 Tandem repeat array allele size distribution

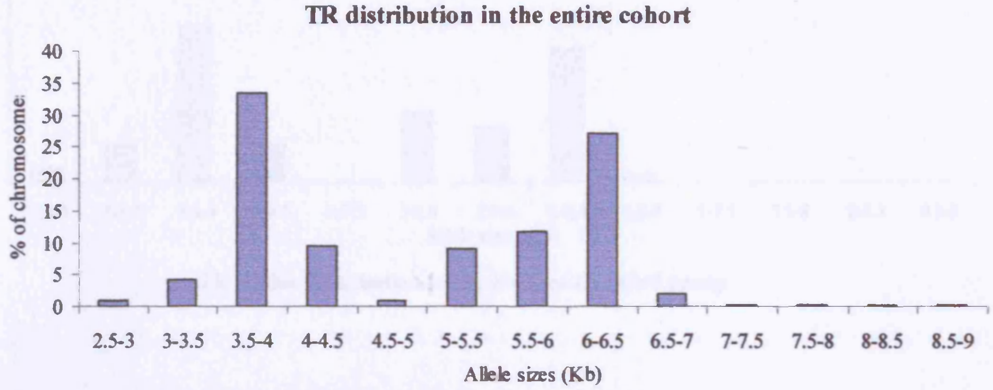


Figure 6.4: Tandem repeat array allele size distribution for the entire cohort, exhibiting a bimodal distribution of the alleles.

Figure 2.3 (Chapter 2) shows a representative Southern blot and Figure 6.4 shows the tandem repeat array allele size distribution for the whole cohort (n=154), which exhibits the typical bimodal distribution, observed in the previous gastritis cohort, and other European populations. There is a pronounced peak on the 3.5-4Kb range for the small alleles, and also a second peak with a mode of the 6-6.5Kb allele length and with quite a large number of alleles in the 5-6Kb range. Despite not being very obvious by observation of this chart, there are slightly more large alleles than small alleles in the cohort, partly a reflection of the high number of alleles binned in the 5 to 6Kb range.

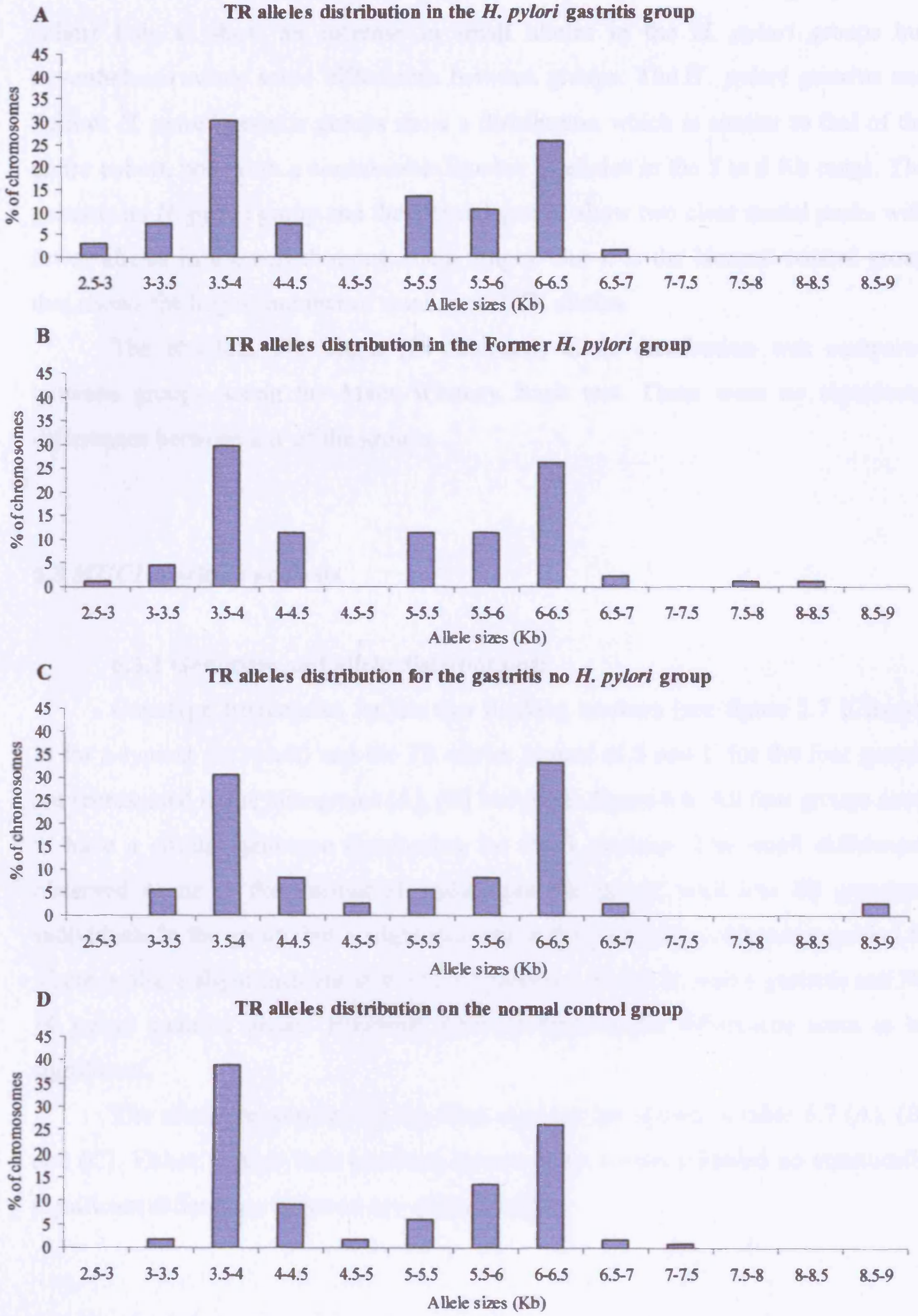


Figure 6.5: Tandem repeat array allele size distribution for each the four groups: (A) *H. pylori* gastritis, (B) Former *H. pylori* gastritis, (C) Gastritis no *H. pylori* and (D) Normal.

A closer examination of the TR distribution in the different subgroups of this cohort fails to show an increase in small alleles in the *H. pylori* groups but nevertheless r eveals some differences between groups. The *H. pylori* gastritis and Former *H. pylori* gastritis groups show a distribution which is similar to that of the entire cohort, both with a considerable number of alleles in the 5 to 6 Kb range. The gastritis no *H. pylori* group and the Normal groups show two clear modal peaks with fewer alleles in the neighbouring allele ranges. But it is the Normal control group that shows the largest number of small modal TR alleles.

The absolute TR length (ie unbinned) allele distribution was compared between groups, using the Mann-Whitney Rank test. There were no significant differences between any of the groups.

6.3 *MUC1* markers analysis

6.3.1 Genotype and allele distributions:

Genotype frequencies for the two flanking markers (see figure 2.7 (Chapter 2) for a typical gel result) and the TR alleles binned as S and L for the four groups are represented in the histograms (A), (B) and (C) in figure 6.6. All four groups seem to have a similar genotype distribution for the 3 markers. The small differences observed occur in the Former *H. pylori* gastritis group, with less SS genotype individuals in the group, but a slight increase in the percentage of heterozygotes LS. There is also a slight increase in the 12,12 genotype for the *H. pylori* gastritis and No *H. pylori* gastritis group. However, none of these small differences seem to be significant.

The allele frequencies for the three markers are shown in table 6.7 (A), (B) and (C). Fisher's exact tests based on chromosome counts revealed no statistically significant differences between any of the groups.

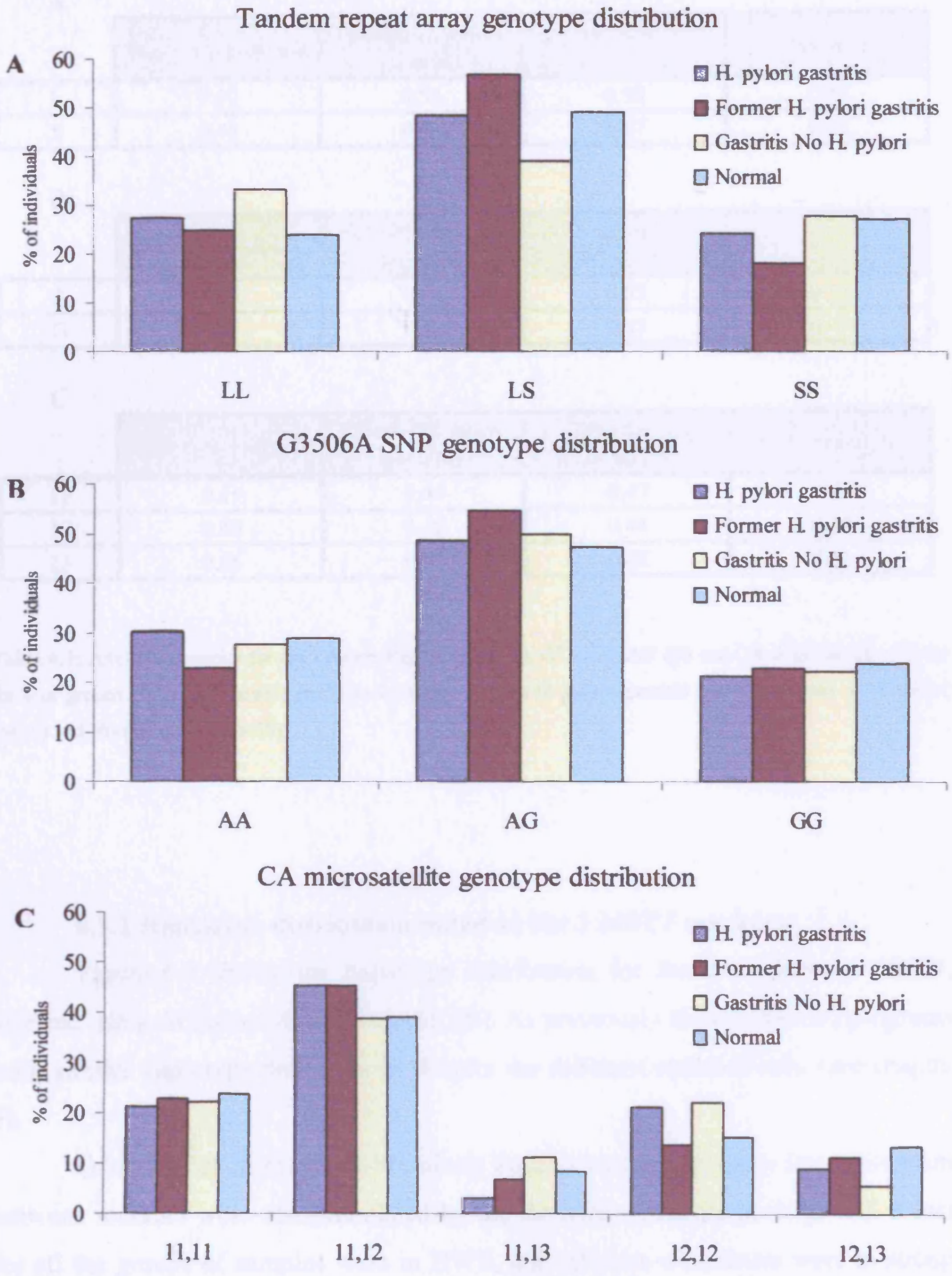


Figure 6.6: Genotype frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the four groups: *H. pylori* gastritis (n=33), Former *H. pylori* gastritis (n=44), Gastritis no *H. pylori* (n=18) and Normal control (n=59).

A

	H. pylori gastritis	Former *H. pylori* gastritis	Gastritis No *H. pylori*	Normal
L	0.52	0.53	0.53	0.48
S	0.48	0.47	0.47	0.52

B

	H. pylori gastritis	Former *H. pylori* gastritis	Gastritis No *H. pylori*	Normal
A	0.55	0.50	0.53	0.53
G	0.45	0.50	0.47	0.47

C

	H. pylori gastritis	Former *H. pylori* gastritis	Gastritis No *H. pylori*	Normal
11	0.45	0.49	0.47	0.47
12	0.48	0.42	0.44	0.42
13	0.06	0.09	0.08	0.11

Table 6.7: Allele frequencies for the Tandem Repeat array (A), G3506A SNP (B) and CA microsatellite (C) for the four groups: *H. pylori* gastritis (n=33 individuals), Former *H. pylori* gastritis (n=44), Gastritis no *H. pylori* (n=18) and Normal control (n=59).

6.3.2 Haplotype distribution based on the 3 *MUC1* markers:

Figure 6.8 shows the haplotype distribution for the 3 markers in *MUC1*, inferred using Arlequin (A) and PHASE (B). As previously shown, the two programs infer similar haplotype distributions, despite the different methods used (see chapter 3).

Goodness of fit to Hardy-Weinberg Equilibrium and Linkage Disequilibrium between markers were also calculated by the Arlequin software package. All 3 loci for all the groups of samples were in HWE, and all pairs of markers were in strong LD.

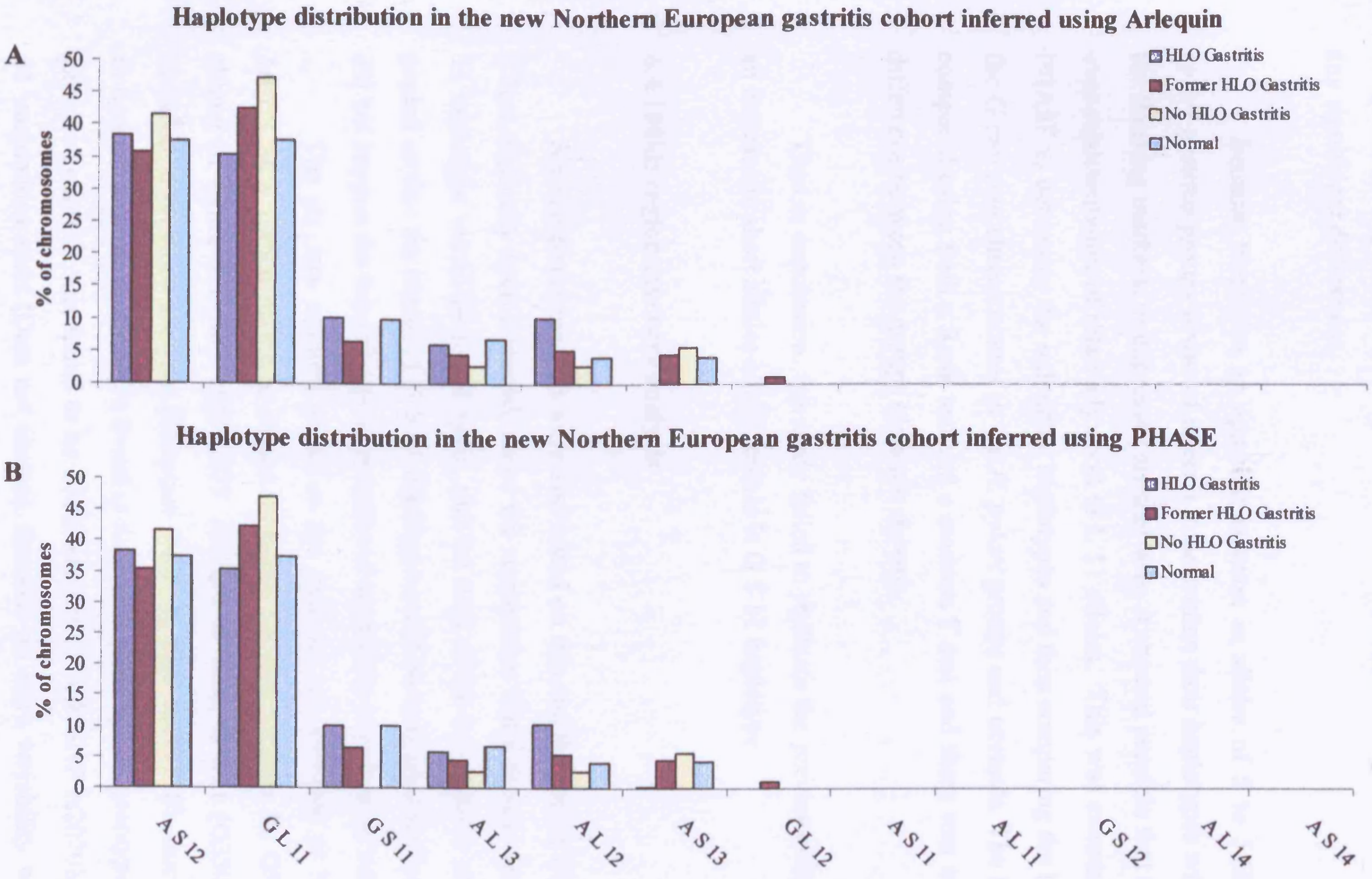


Figure 6.8: *MUC1* haplotype distribution for the four populations studied: *H. pylori* gastritis (n=33), Former *H. pylori* gastritis (n=44), Gastritis no *H. pylori* (n=18) and Normal control (n=59). The haplotypes were inferred using 2 different statistical methods: (A) a maximum-likelihood method, the EM algorithm, and (B) a Bayesian method.

Examining figure 6.8, it is notable that all four populations have very similar haplotype distributions, with no particular haplotypes over-represented in any group. Comparisons between pairs of populations using the "Exact test of population differentiation" (part of Arlequin package), revealed no significant differences between groups. Likewise, the differentiation test performed by PHASE, that computes the significant differences between haplotypes distribution did not reveal any significant differences.

Because there was an apparent increase in alleles of 5 to 5.5kb in the *H. pylori* gastritis groups it was of interest to determine their haplotypes with respect to the flanking markers, or expressed another way, it seemed possible that there was an over-representation of relatively short G L 11 alleles. This was examined by using PHASE to determine the individual haplotypes and then comparing the lengths of all the G carrying chromosomes in the *H. pylori* groups and normals. The lengths were compared using both a Rank test and a students T test and there was no significant difference between the groups (data not shown).

Thus in conclusion, this study failed to replicate the previous observations of an increase in short alleles or an increase in G S 11 haplotype.

6.4 100kb region markers analysis

Six additional markers were also tested on this and the earlier UK cohort and it had originally been intended, under the supposition that a disease associated G S 11 haplotype would be found again, that the data of the two cohorts would then be pooled so that the extended G S 11 haplotype could be examined further. Since this did not happen the two cohorts were analysed separately as presented below.

The six new markers added to the analyses are localised at both sides of *MUC1*, in a 100Kb region as shown in figure 5.5 (see Chapter 5). Of the 9 SNPs shown in figure 5.5, SNP rs4072037, marked in blue, is that (G3506A) already included in previous analyses (Chapters 3 and 4). SNP rs4971101, marked in green in figure 5.5 is one of the SNPs found in the NCBI dbSNP, was genotyped for the old gastritis cohort. It was found to be 100% associated with SNP rs2070803, localised 81 nucleotides apart (Data not shown). Because no extra variability was added by including the rs4971101 SNP, it was not genotyped for the second cohort, and is therefore not included in the analyses. Finally, the SNP with accession number rs1045253, was not possible to genotype because it is included in a pseudogene, *MTX1P* that highly resembles *MTX1*. The PCR amplified products of both genes, and the homologous region in *MTX1* does not contain the polymorphism.

Both gastritis cohorts were genotyped for the remaining 6 SNP by PCR and restriction enzyme digestion. Figure 5.12 (Chapter 5) shows some examples of restriction enzyme digestions run in agarose gels. The markers shown in figure 5.12 are the ones selected from HapMap: rs9297 (A), rs4971088 (B) and rs4971100 (C).

6.4.1 Extended haplotype distribution in the old and new gastritis cohorts

Figure 6.9 shows the 9 markers used for the extended haplotypes in a scheme of the genomic region.

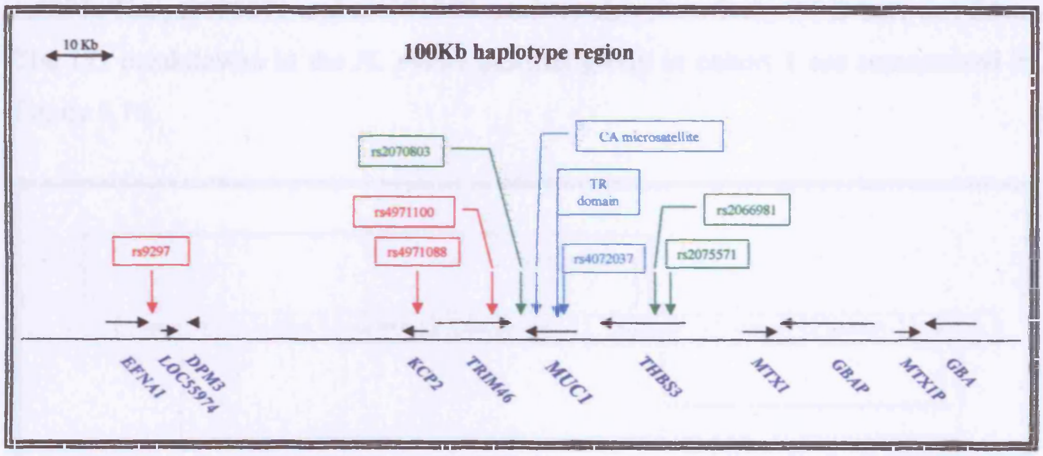


Figure 6.9: Schematic representation of the 100Kb genomic region and the markers genotyped for the construction of the extended haplotypes. The *MUC1* markers previously studied are shown in blue boxes, while the new markers are shown in green boxes (NCBI dbSNP) and red boxes (HapMap SNPs). *MUC1* size is only a rough approximation because of the polymorphic TR domain that can alter dramatically the size of *MUC1*.

The inferred extended 9 marker haplotypes for the two gastritis cohorts are shown in figure 6.11. In this figure, the new gastritis cohort only shows data for 2 groups: the *H. pylori* gastritis group and the normal controls. These two groups are the ones more comparable to the groups in the old gastritis cohort. Figure 6.12 shows the full data set for the new cohort.

The haplotypes for the old gastritis cohort (Figure 6.11 A) and new gastritis cohort (Figure 6.11 B) were both generated by Arlequin software package. Arlequin also calculates the goodness of fit to Hardy Weinberg equilibrium for the genotype frequencies for each locus. All loci were in Hardy-Weinberg Equilibrium, both in the old and in the new cohort.

Arlequin also runs a pairwise Linkage Disequilibrium test (LD) between pairs of markers. All pairs of markers were in LD in the new gastritis cohort, though the association between TR array and SNP rs4971088 (in *KCP2*, see figure 6.9) was of borderline significance (p=0.0457). However, in the old gastritis cohort, the LD breaks between several markers in the *H. pylori* gastritis group. In was previously shown in chapter 3 that the CA microsatellite in intron 6 is not in LD with the TR domain. In the 9 markers analyses the breakage extends further; the TR domain is not in LD with the CA microsatellite (as observed previously), the rs2075571 SNP and rs9297. The rs2075571 and rs4971088 markers also show lack of LD between them. The LD breakdowns in the *H. pylori* gastritis group in cohort 1 are summarised in Figure 6.10.

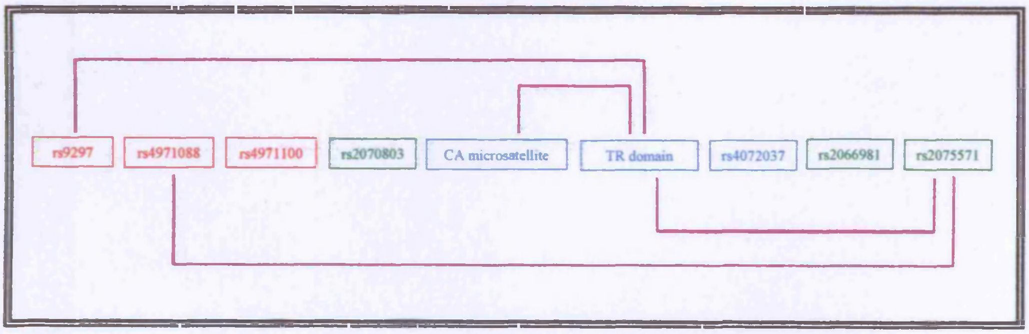


Figure 6.10: Schematic representation of the genomic region and markers genotyped for the extended haplotype, for *H. pylori* gastritis group from cohort 1. The lines in lilac connect pairs of markers for which LD is broken.

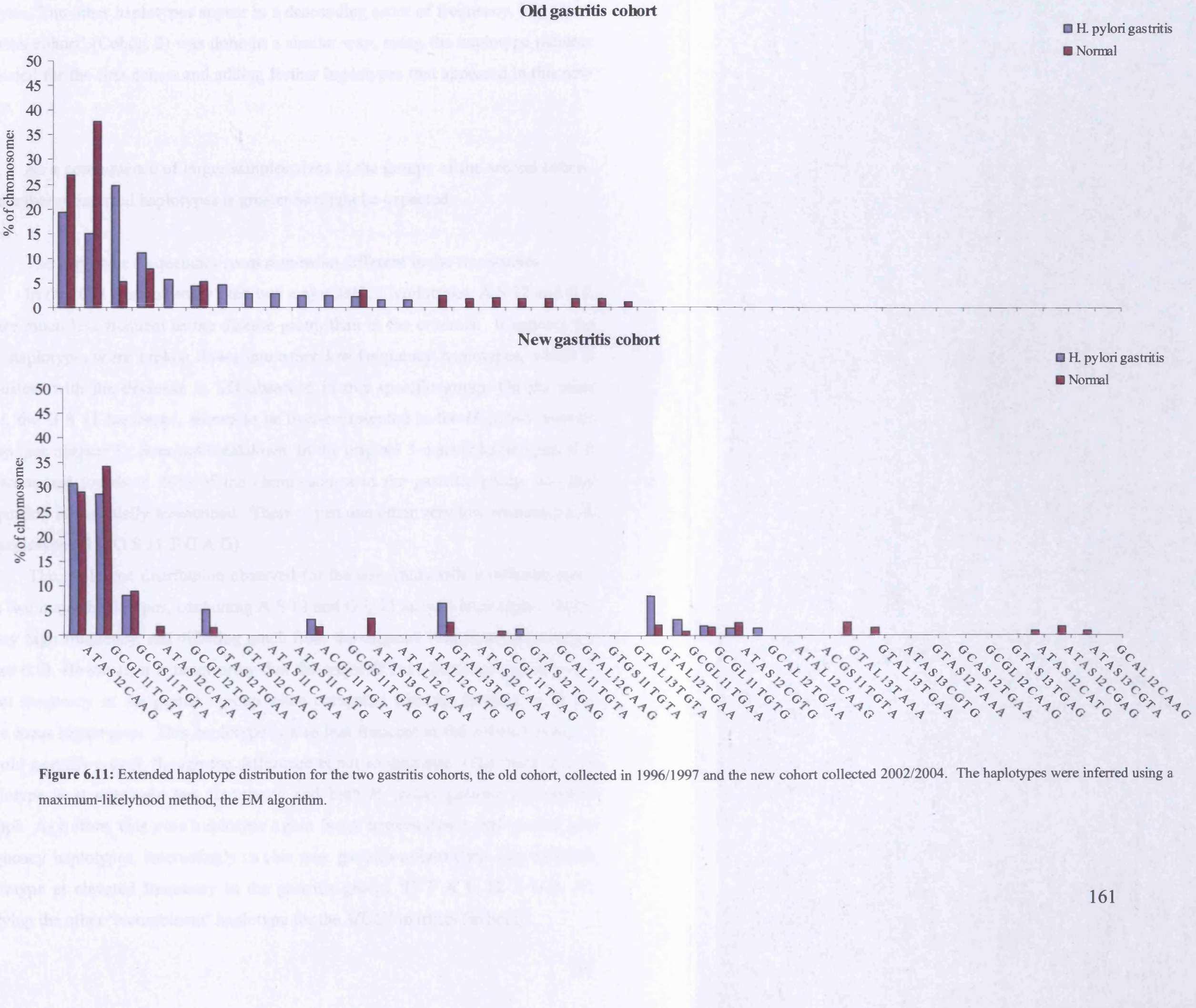


Figure 6.11: Extended haplotype distribution for the two gastritis cohorts, the old cohort, collected in 1996/1997 and the new cohort collected 2002/2004. The haplotypes were inferred using a maximum-likelyhood method, the EM algorithm.

The haplotype distribution for each cohort was generated in independent runs of Arlequin. Firstly, the 'Old gastritis cohort' (Cohort 1), was plotted into a histogram sorting out the haplotypes according to their frequencies. The first two haplotypes correspond to the two most frequent haplotypes observed in the 3-marker analysis. The other haplotypes appear in a descending order of frequency. The 'New gastritis cohort' (Cohort 2) was done in a similar way, using the haplotype patterns predicted for the first cohort and adding further haplotypes that appeared in this new study.

As a consequence of larger samples sizes in the groups of the second cohort, the number of inferred haplotypes is greater as might be expected.

The haplotype frequencies seem somewhat different in the two studies.

In the 'Old gastritis study', the two major *MUC1* haplotypes, A S 12 and G L 11 are much less frequent in the disease group than in the controls. It appears the core haplotypes were broken down into other low frequency haplotypes, which is consistent with the decrease in LD observed in this specific group. On the other hand, the G S 11 haplotype, shown to be over-represented in the *H. pylori* gastritis group (see chapter 3), does not breakdown. In the original 3-marker haplotypes, G S 11 accounted for about 24% of the chromosomes in the gastritis group, and that proportion is essentially maintained. There is just one other very low frequency G S 11 haplotype (G C G S 11 T G A G).

The haplotype distribution observed for the new study tells a different story. The two major haplotypes, containing A S 12 and G L 11 as core haplotypes, remain a very high frequency, not differing much from the original 3-markers analysis (see figure 6.8). However, it is noteworthy that the major G L 11 haplotype remains at a lower frequency in the gastritis group when compared with the controls, as in the three locus haplotypes. This haplotype is also less frequent in the gastritis group of the old gastritis cohort, though the difference is not so dramatic. The main G S 11 haplotype is at relatively low frequency, and both *H. pylori* gastritis and control groups. As before, this core haplotype again is not broken down into several low frequency haplotypes. Interestingly in this new gastritis cohort there is a different haplotype at elevated frequency in the gastritis group: 'G T **A L 12** T G A A', carrying the other 'recombinant' haplotype for the *MUC1* markers (in bold).

Despite these observations, the 'Exact test of population differentiation', performed by Arlequin, did not find any significant differences between the groups in either study. The haplotype differentiation test run by PHASE software did however find the difference between the two groups in the 'Old gastritis cohort' to be almost significant (p=0.052).

Inspection of the bar charts of the total population of patients and controls shows several noteworthy features. A large number of the G L 11 chromosomes are nearly identical for haplotype across the whole region. This concurs with what can be seen for this region of the CEPH chromosomes in Figure 5.10 (Chapter 5) that can be seen in a simplified table containing only the genotyped markers (Figure 6.13). The S chromosomes in contrast are more diverse, suggesting more recombination (Figure 6.13). Interestingly most of the G S 11 chromosomes are the same as G L 11 with respect to the flanking markers in both the London and CEPH cohorts.

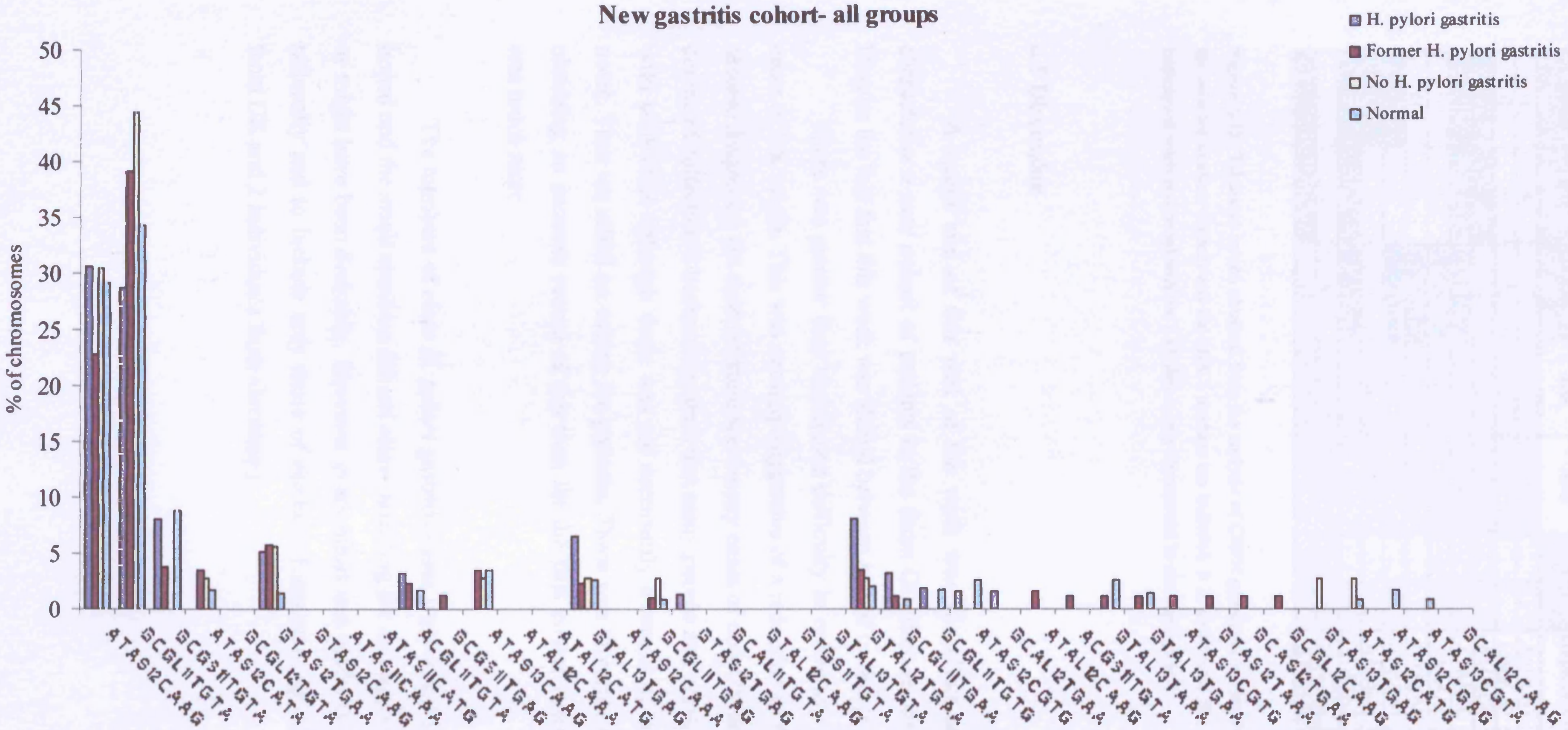


Figure 6.12: Extended haplotype distribution for the four populations studied: *H. pylori* gastritis (n=33), Former *H. pylori* gastritis (n=44), Gastritis no *H. pylori* (n=18) and Normal control (n=59). The haplotypes were inferred using) a maximum-likelihood method, the EM algorithm.

A/G	A/T	A/G	L/S	A/G	C/T	Haplotypes	Number of CEPH chromosomes
A	A	G	L	A	T	AAGLAT	1
G	A	G	L	A	T	GAGLAT	4
A	T	G	L	G	C	ATGLGC	28
A	T	A	S	A	C	ATASAC	7
G	A	A	S	A	C	GAASAC	17
G	A	A	S	A	T	GAASAT	10
A	A	A	S	A	T	AAASAT	3
A	A	G	S	A	T	AAGSAT	3
A	T	A	S	A	T	ATASAT	2
G	A	G	S	G	C	GACSGC	2
A	T	G	S	G	C	ATGSGC	4

Figure 6.13: Tabulated results obtained from the analysis of CEPH genotypes in HapMap. In this figure, only the selected markers (in red) and the *MUC1* markers are included. It is noteworthy the larger diversity of S A haplotypes when compared with the L G, that is only represented in one haplogroup.

6.5 Discussion

A major aim of this part of the work was to collect samples from and characterise a new cohort of patients in the same GI clinic as the previous cohort. Despite the fact that this work was shared between two of us this proved slow.

There was greater than anticipated difficulty in collecting *H. pylori* gastritis cases of UK origin. This was strongly suggestive of a reduction in the UK population disease frequency. (in contrast there were many cases of non-UK origin). During the course of collection it became apparent that many people had previously been treated with antibiotics although there was not necessarily a record of this in the hospital notes. Thus we relied on asking the patients. There was therefore more difficulty in obtaining an accurate record of this than for the first cohort where prior treatment was much rarer.

The numbers of clear *H. pylori* gastritis cases were many fewer than we had hoped and the small numbers did not allow matching for sex, age and smoking status as might have been desirable. However every effort was made to group the patients rationally and to include only those of northern European origin, (152 individuals from UK and 2 individuals from Germany).

Unfortunately it was not possible to replicate the original finding of an increase in S alleles and of a specific haplotype in *H. pylori* gastritis. This will be discussed in more detail in the next Chapter, Chapter 7.

Determination of extended haplotypes did however provide some interesting insights. In particular the very high frequency 100kb G L haplotype was noteworthy. It implies that the majority of long alleles originate from a single progenitor and that more recent slippages, gene conversions or other mutations have led to subsequent TR length diversity, but the existence of this 'haplogroup' at such high frequency suggests it might have been favoured by selection. Interestingly the most common G S 11 haplotype is identical with respect to the flanking markers, but the TR array is much shorter. Inspection of the precise allele lengths of the G S 11 alleles shows that they are quite variable in length (data not shown). This suggests perhaps that these alleles arose from several large deletions of common L alleles. As mentioned in Chapter 3 there was indication that the TR array of these alleles more closely resembles L alleles, which is consistent with this suggested origin. The long A L 12 alleles in contrast have a different haplotypic background. It will be of interest in the future to examine the more extended haplotypes in the Nigerians and other Africans as well as the Portuguese and other individuals of non-UK origin. From such studies it should be possible to chart the genealogy of the stretch of DNA more precisely. Examination of Ashkenazi Jewish populations will be of particular interest because of the location of *GBA*, the Gaucher gene, in this stretch of DNA (see Chapter 5 for discussion of this).

CHAPTER 7

DISCUSSION, CONCLUSIONS AND FUTURE PERSPECTIVES

At the outset of this project there was evidence for the involvement of *MUC1* in gastric disease. There was suggestive evidence, now published, that *H. pylori* binds MUC1 (Linden et al., 2004) and there was clear evidence for a change in the pattern of MUC1 glycoprotein expression in *H. pylori* gastritis (Vinall et al., 2002). This was complemented by early studies that showed evidence of association of *MUC1* allele length in gastric cancer (Carvalho et al., 1997) and the finding of the same pattern of association in a small UK *H. pylori* gastritis cohort collected in this laboratory (Vinall et al., 2002).

Since preliminary studies by Jo Fowler (PhD thesis, University of London, 2002) suggested that the markers flanking the TR domain were not disease associated, despite the fact that these markers had been shown to be associated with each other and TR array length in other cohorts, one of the first aims of my work was to test these markers in all the patient material available and examine the pattern of haplotype distribution in control and disease groups. This led to the observation that one particular haplotype (GS11) was over-represented in gastric disease patients or increased in frequency in parallel with disease severity, even though this was not statistically significant in the Portuguese patients.

The original aim of my work was to examine TR sequence variability in the context of gastric disease. The technique involved proved too cumbersome and required very high quality DNA, but preliminary examination suggested that the GS11 haplotype resembled long TR alleles in terms of repeat sequence array rather than short alleles (see Chapter 3).

In order to pursue these observations further, two approaches were taken. One was to determine the extent of linkage disequilibrium across the *MUC1* gene region and identify SNPs to examine the patterns of disease association over a larger genomic region. The other was to collect and characterise (together with Dr Adil Elamin) a new cohort of patients as a replication study. Despite the existence of the potentially destabilising TR array in the middle of *MUC1*, a region of LD of 600Kb was found. SNPs were selected to maximise haplotype diversity and with the hopes of providing particular association with the putative disease haplotype. Analysis of these markers in the first UK cohort still showed a clear disease haplotype but no single SNP associated better with disease than the original L/S variation.

DNA samples from a new cohort of 237 people were successfully collected and characterised. Of these 154 could be included in the subsequent analyses. The collection of samples from *H. pylori* positive individuals of UK origin proved to be a difficult task. In the early stages of the collection, and with a view to not showing discrimination, samples were collected from individuals regardless of their ancestry. Many individuals collected at that period were of non-Northern European ancestry, and many of these had *H. pylori* infection and gastritis. To maximise our chances of getting as many UK patients as possible, in the subsequent collection, UK or Northern European origin individuals and certain classes of patients were preferentially targeted. However in the 154 UK/Northern European samples, only 34 individuals had *H. pylori* infection. This proportion seems low when one considers that this was a selected population and, according to Vyse *et al* estimates, in 1996 14% of the England and Wales population was infected, though most of those individuals were born in the 1940s or pre-1940s (Vyse et al., 2002). However their data also suggests that *H. pylori* infection is becoming less frequent each decade, a possible explanation for the increased difficulty in collecting *H. pylori* positive samples from UK patients since the last gastritis cohort study, collected in 1996/1997. Other possible explanations for the decrease in the number of individuals infected is the rise in number of individuals treated with antibiotics to eradicate *H. pylori* infection

Using this new cohort the original observation of over-representation of the G S 11 and associated haplotypes found in the first even smaller cohort was not replicated. It is possible that the original associations were spurious resulting from small data sets in the first place (Cardon and Palmer, 2003;Tabor et al., 2002). It could also have resulted from population stratification which is very difficult indeed to avoid completely (Cardon and Palmer, 2003). We selected volunteers of northern European/UK ancestry but it was very clear that *H. pylori* is much more prevalent in people of non UK ancestry, possibly attributable to both cultural and genetic reasons (Vyse et al., 2002). The population of London is very mixed and there could be geographic differences in *Helicobacter pylori* frequency as well as different bacterial strains present at the population and individual level, that affect the disease outcome (Figueiredo et al., 2001;Nogueira et al., 2001). Particular strains, such as the ones carrying Cag PAI and VacA+ phenotype are particularly aggressive (Figueiredo et al., 2001). Kauser *et al* reported presence of less virulent strains as predominant, but also some strains that were typical of Asian populations (Kauser et al., 2005).

The presence of a geographically mixed population in London could also lead to independent differences in *MUC1* allele frequency. This however was considered to be an unlikely explanation because no significant difference in *MUC1* haplotype frequency was found across the UK (Vinall, Fowler unpublished). It should however be noted that one potential confounder could be difference in representation of patients of Ashkenazi Jewish extraction, who might possibly have different *MUC1* haplotype distribution due to selection (see Chapter 5 and 6 for discussion about *GBA*). The Portuguese gastric cancer study was likely on the other hand to have been better because the blood donors came from S. Joao Hospital, Porto, the same hospital as the cancer patients. The gastritis patients came from Viana do Castelo, a city in the far North of Portugal, where gastric cancer incidence (and consequently gastric disease in general) is known to be particularly high (Lunet and Barros, 2003) but it was not possible to make direct comparisons between these two groups.

Despite the negative findings, there were several curious features of the results which should be noted. There was an apparent increase in the currently and formerly H. pylori infected groups of alleles just above the 5kb cut point, a category that is rare in the population as a whole. Also, and perhaps more significant there was, in both gastritis cohorts, an increase in low frequency haplotypes in the *H. pylori* gastritis group (albeit non-significant in the second cohort). In each case there was a corresponding decrease in one extended haplotype (G C G L 11 T G T A) the one which is so frequent in the northern European population as a whole. The high frequency of this haplotype had suggested that this haplotype has, or has had in the past, a selective advantage (see Chapter 6). Perhaps this group of long *MUC1* alleles is protective against *H. pylori* colonisation?

Recent studies from the Hanisch group show that there is indeed likely to be a functional difference in different tandem repeat variants (Mensdorff-Pouilly et al., 2005). PESR repeats, which seem to be relatively more frequent in long TR alleles, show increased flexibility and possible reduced glycosylation. This comes from indirect evidence, showing that the PESR repeats were usually found in association with the A variant of the PAQT change site (Figure 1.3, Chapter 1), which was shown to have a reduced glycosylation density and even a negative effect on GalNAc-Transferase activity, affecting also some distant glycosylation sites (Mensdorff-Pouilly et al., 2005).

It should also be noted that the exon 2 GA SNP, like the TR array variation causes a functionally relevant change in *MUC1*. Presence of G at this position leads to the insertion of 27 nucleotides in the *MUC1* transcript. This leads to alteration of the signal sequence and abolition of one of the known proteolytic cleavage sites at the N-terminal end as depicted in Figure 7.1 and probable introduction of a new one (green arrow) (Parry et al., 2001), which leads to an extended N-terminal end as described in Hilkens and Buijs for the A splice variant (Hilkens and Buijs, 1988).

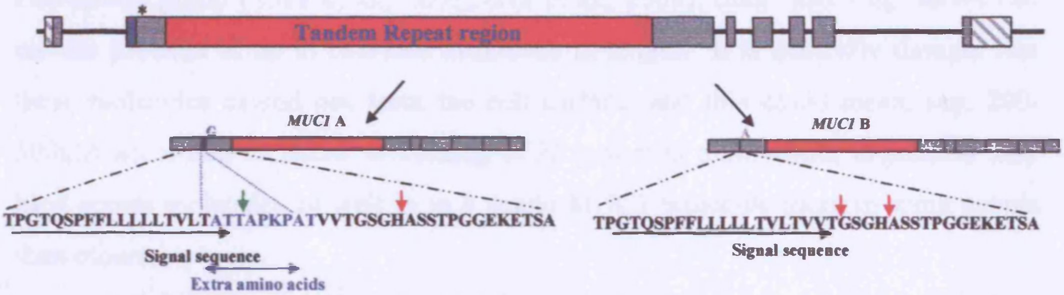

*
Tandem Repeat region
MUC1 A
MUC1 B
G
A
TPGTQSPFFLLLLLTVLTATTAPKPATVVTGSGHASSTPGGEKETSA
TPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSA
Signal sequence
Extra amino acids
Signal sequence

Figure 7.1: Schematic representation of *MUC1* alternative splicing event occurring in the 5` end of exon 2. *MUC1* A and B spliceoforms are associated with a SNP, G3506A, with spliceoform A associated with the presence of G nucleotide, and spliceoform B associated with the A nucleotide. This SNP causes a change in the proteolytic cleavage site as show by the red and green arrows. Adapted from diagram kindly supplied by Wendy Ng.

Although it is hard to predict what effect this extension will have on the molecule it is not unreasonable to suppose that it may affect intra-cellular transport. It is also worth noting that the B variant contains a peptide in its signal sequence which is absent in variant A, which has been shown in a cell culture system to provoke MUC1 specific cytotoxic T cells (Bohnenkamp, oral communication at 8th International Workshop on Carcinoma-associated Mucins, Mucins in Health and

Disease, Cambridge, 2005). Furthermore, preliminary studies in the lab indicate that the A transcripts (ones with the G allele) are expressed at lower levels than the B transcripts (Ng et al, unpublished data). This may be due to the associated variation in TR length (perhaps fewer long transcripts are transcribed) and this is currently under investigation, making use, in part, of data collected in this study. It is however possible to imagine that these various genetically determined variations as well as sequence variations in the TR array itself, act in different ways to alter susceptibility to *H. pylori* infection and colonisation. In other words there may be differences in transport, variability in the binding of *H. pylori* to MUC1, differences in consequent signalling, differences in the associated inflammatory and immune responses. At present not all of these variations can easily be measured. A simple PCR based method for examining TR sequence variation would certainly be helpful, so that a more in depth description of the TR variability could be obtained.

There is also the issue of heterozygosity that has been considered by the Portuguese group (Silva et al., 2001;Silva et al., 2003). Short and long alleles can encode proteins of up to two-fold difference in length. It is generally thought that these molecules extend out from the cell surface, and this could mean, say, 200-500nM according to allele. If binding of *H. pylori* is multivalent, organisms may bind across molecules as well as to a single MUC1 molecule more in some people than others.

All this potential diversity, together with the fact that in small data sets there are likely to be a number of confounders, could account for our not replicating the previously observed association.

The small number of individuals in each group in both cohorts may have lead to a bias in the results either by generating a false positive or false negative association. The best way of decreasing these types of errors is to increase the number of individuals, to at least a few hundred in each group. The best way to achieve this would be to collect more samples by taking advantage of the wide range of hospitals affiliated to UCL and do a UCLH (University College London Hospitals) multi- centre study, obtaining samples more quickly and still all in the same geographical area- London.

With a larger study, multiple regression analysis could be performed to take into account several other factors that may be involved in disease susceptibility. On the one hand, these include external factors, such as the particular *H. pylori* strain(s) present in the patients, in particular the virulence factors such as the Cag pathogenicity island, the VacA locus and the BabA2 adhesin, since the disease outcome and progression depends greatly on the strains that are present. Another external factor that may be involved is the smoking habits of the individuals in the study. On the other hand, we should also take into account other genetic loci (than the ones presented in this thesis) that are thought to alter *H. pylori* infectivity and disease consequences. These include the blood groups (e.g. Lewis b), that have been reported to be ligands to *H. pylori* or affect gastric disease susceptibility (Heneghan et al., 1998), and the inflammatory cytokines, such as IL-1β, IL-1RN and TNF-α, that were shown to have genetic variants that are overrepresented in gastric disease (El Omar et al., 2000;El Omar et al., 2001;El Omar, 2001;Machado et al., 2003).

An additional problem is that the Southern Blot technique not only requires large amounts of good quality DNA (preferably blood extracted DNA) but also it is very laborious and time consuming. To get around this, it was thought to be important to identify single nucleotide polymorphisms (SNPs) that would act as linked markers of the TR domain. However, none of the SNPs identified appeared to fulfil this role. In order to understand better how to proceed, I have analysed the haplotypes again using the 7 SNP markers together with the CA microsatellite and TR in *MUC1* using PHASE, but taking into account precise allele length. For this analysis, both cohorts (old and new cohort) were merged to increase sample size. The tandem repeat allele lengths were binned into 0.5kb classes, generating 12 alleles for the TR. With these 9 markers, 80 different haplotypes were predicted by PHASE. However, there are only 28 haplotypes based only on the SNPs, 5 of which (A, B, C, D and E) account for almost 90% of the chromosomes (figure 7.2).

To determine the relationship of the SNP haplotypes found, a phylogenetic tree was generated using Treeview software (version 1.6.6 uploaded in September 2001) (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html), which builds trees based on their similarity using a parsimonious principle (figure 7.3). Interestingly, the two most common haplotypes, A and B, are the exact opposite of each other. This phenomenon, "yin yang haplotypes", was previously observed for other loci (Zhang et al., 2003).

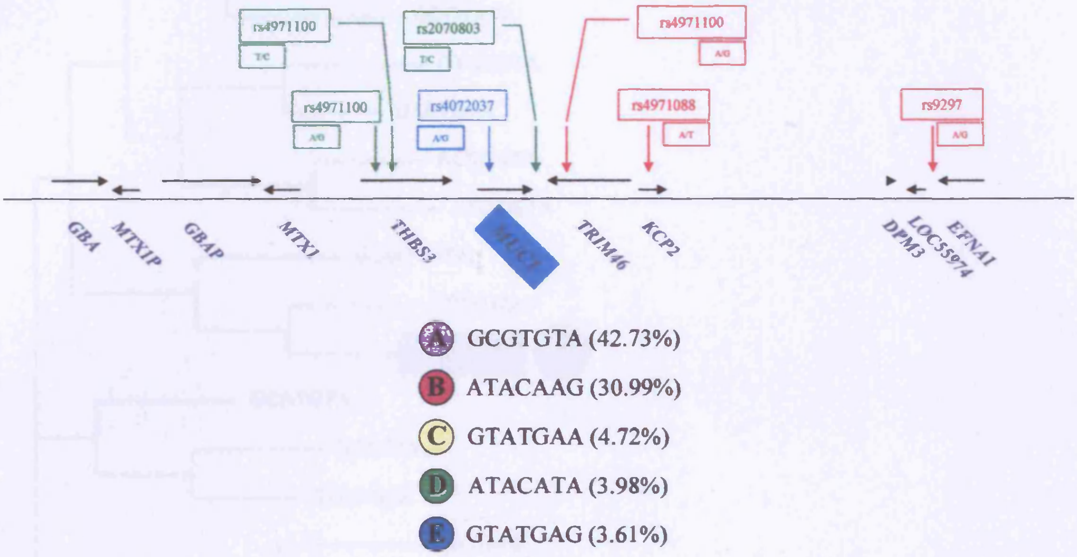


Figure 7.2: Schematic representation of *MUC1* genomic region and the 7 SNPs selected for the haplotype analysis. Below the diagram, the 5 major SNP haplotypes are shown (A, B, C D E) in order of frequency (shown in brackets) with the letter representing each haplotype surrounded by a circle coloured according to the colours used for the haplotypes in figures 7.3 and 7.4.

The TR allele distribution in relation to the common SNP haplotypes can be seen in figure 7.4, both for the *H. pylori* gastritis and normal control groups. As can readily be seen, each of the common haplotypes, in particular haplotype A (the only one carrying the G allele from the G3506A *MUC1* SNP), shows a considerable diversity in TR allele length, suggesting that the SNP haplotypes are much older than the derived TR alleles on each background. In view of this analysis, it seems that further SNP testing is unlikely to be useful for the disease association studies, but that testing microsatellites with their higher mutation rate might give insight into the evolutionary origins of the haplotypes and might even prove useful for the disease association studies.

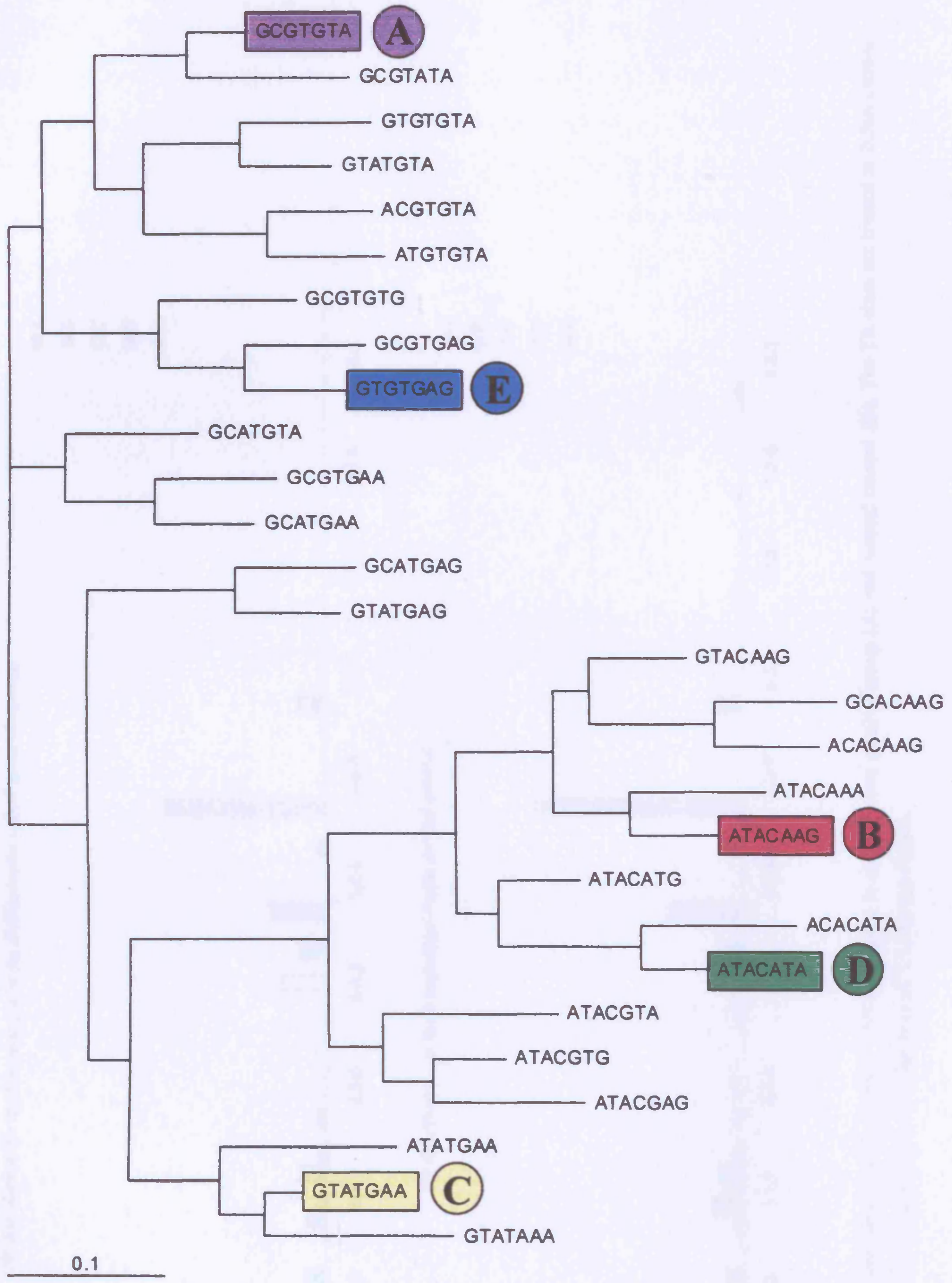


Figure 7.3: Phylogenetic tree for the SNPs in *MUC1* genomic region, generated using Treeview software. The 5 major haplotypes are displayed in coloured boxes next to the haplotype name (letters in circles). The order of the SNPs is in agreement with their distribution (left to right) shown in figure 7.2.

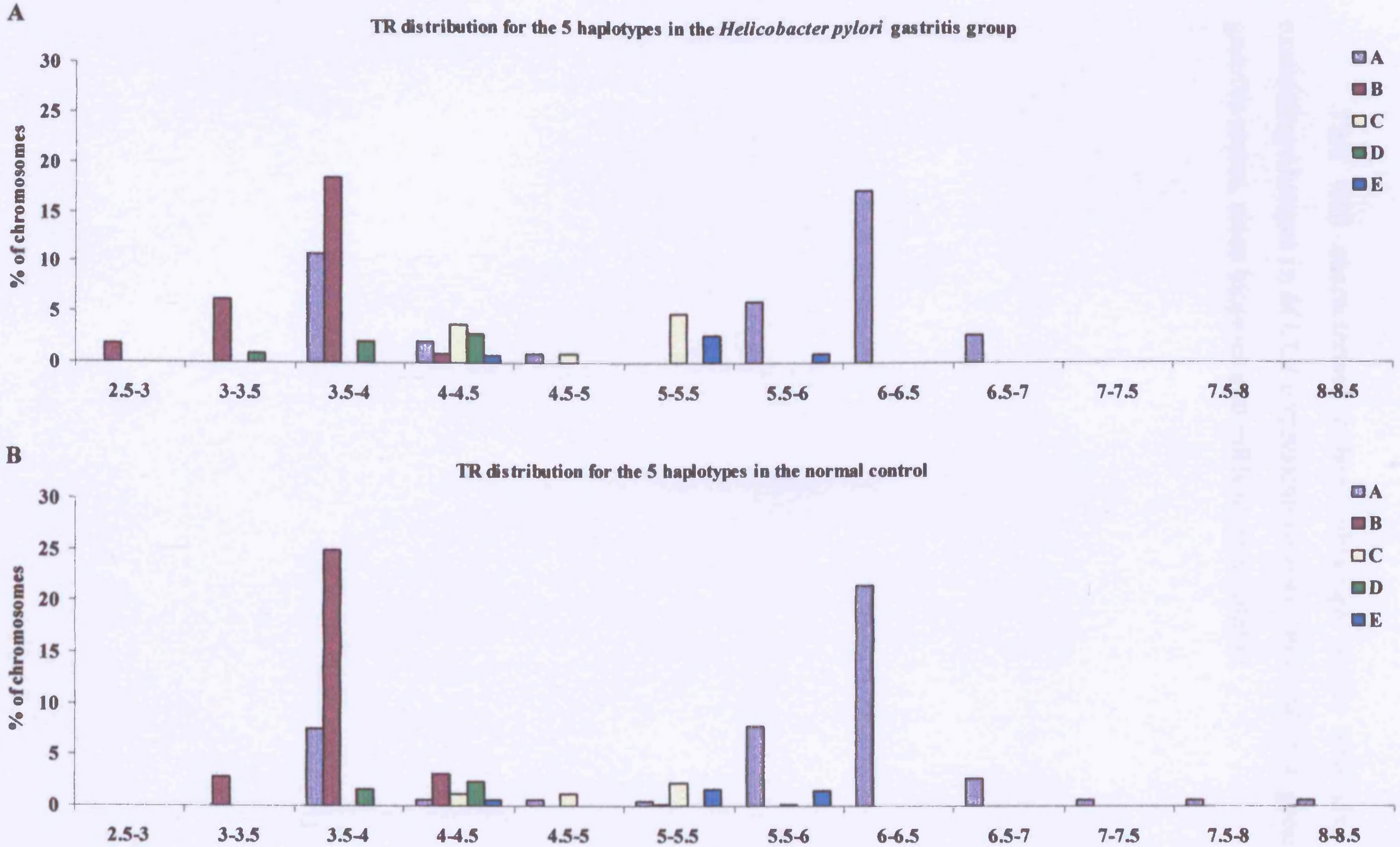


Figure 7.4: *MUC1* tandem repeat allele distribution for the 5 major SNP haplotypes in the *H. pylori* gastritis group (A) and normal control (B). The TR sizes are binned in 0.5kb classes though the L and S allele separation can be made between the 4.5-5kb (S) and 5-5.5kb (L) classes.

After completion of this study it still seems possible that the sequence variation of the tandem repeat array of MUC1 is the important factor in influencing gastric disease. One way forward might be to develop a protocol to look at the variant repeats using sequence specific hybridisation or minisequencing, though there would still be the need to separate the two alleles. It is possible that better DNA polymerases and protocols are now available for conducting these difficult long PCR reactions.

This well characterised cohort, although small, will also be useful for examining changes in *MUC1* expression in relation to *MUC1* gene haplotype and gastritis status, since biopsies and mRNA are available.

APPENDIX 1

Distribution of TR alleles within the 5Kb to 5.5Kb range in the Nigerian population

Summary of the results:

Haplotypes	Within 5-5.5kb	Possibly within 5-5.5Kb	Outside 5-5.5Kb
A L 12	18	9	4
non A L 12	3		

Raw data:

Individuals code	Haplotypes predicted		Haplotype probability	TR size (Kb)
IND: 1	AS 12	GS 11	1	4.683
				3.718
IND: 2	AS 12	AS 12	1	3.854
				3.744
IND: 3	AS 12	AL 12	1	5.179
				3.745
IND: 4	AL 12	AL 13	1	5.717
				5.128
IND: 5	AS 12	AL 12	1	5.256
				3.562
IND: 6	AS 12	AS 12	1	4.83
				3.411
IND: 7	AL 12	AL 12	1	5.138
				5.029
IND: 8	AS 12	GL 11	0.943	6.075
				4.241
IND: 9	AS 13	GL 11	0.816	5.807
				3.753
IND: 10	AS 12	GL 11	0.942	6.406
				3.715
IND: 11	AL 12	GL 11	0.998	6.021
				5.065
IND: 12	AS 12	AL 12	1	5.043
				3.669
IND: 13	AL 12	GL 11	0.998	5.04
				5.04
IND: 14	GL 11	GL 11	1	5.973
				5.973
IND: 15	AS 12	AL 12	1	5.608
				3.675
IND: 16	GL 11	GL 12	1	6.293
				6.07
IND: 17	AS 12	AL 12	1	5.149
				3.766
IND: 18	AS 12	GL 11	0.943	5.528
				3.974

Individuals code	Haplotypes predicted		Haplotype probability	TR size (Kb)
IND: 19	AS 12	AS 12	1	3.699
				3.699
IND: 20	GL 11	GL 11	1	5.978
				5.494
IND: 21	AS 12	GL 11	0.943	7.018
				3.645
IND: 22	AS 12	AS 12	1	4.8
				3.616
IND: 23	AS 12	AS 12	1	3.709
				3.596
IND: 24	AS 12	AS 12	1	3.671
				3.429
IND: 25	AS 12	AL 12	1	5.083
				3.674
IND: 26	AL 12	GL 11	0.998	6.078
				5.101
IND: 27	AL 12	AL 13	1	5.119
				5.119
IND: 28	AS 12	GL 11	0.942	6.056
				3.671
IND: 29	AS 12	AL 12	1	5.09
				3.667
IND: 30	AS 12	GL 11	0.942	6.151
				3.722
IND: 31	AS 12	AL 12	1	5.178
				3.597
IND: 32	AL 12	GL 11	0.998	6.054
				5.205
IND: 33	AS 12	AS 12	1	4.284
				3.889
IND: 34	AS 12	AL 12	1	5.151
				3.706
IND: 35	AS 12	AS 12	1	3.674
				2.913
IND: 36	GL 11	GL 11	1	6.294
				6.155
IND: 37	AL 12	GL 11	0.998	6.12
				5.429
IND: 38	AL 12	GL 11	0.998	6.056
				5.066
IND: 39	GL 11	GL 11	1	5.994
				5.994
IND: 40	AS 12	AL 12	1	5.151
				3.614
IND: 41	AL 12	AL 12	1	5.61
				5.091

Individuals code	Haplotypes predicted		Haplotype probability	TR size (Kb)
IND: 42	GL 11	GL 11	1	6.14
				6.14
IND: 43	AL 12	GL 11	0.998	6.055
				6.055
IND: 44	GL 11	GL 11	1	6.14
				6.14
IND: 45	AL 12	GL 11	0.998	6.313
				5.282
IND: 46	AS 12	GL 11	0.943	6.073
				3.725
IND: 47	AS 12	AS 12	1	3.759
				3.602
IND: 48	GL 11	GL 11	1	6.347
				6.072
IND: 49	AL 12	GL 11	0.998	5.988
				5.025
IND: 50	GL 11	GL 11	1	6.105
				6.105
IND: 51	AS 13	GL 11	0.832	5.538
				3.775
IND: 52	AS 12	AS 12	1	3.617
				3.617
IND: 53	AL 12	AL 12	1	5.309
				5.172
IND: 54	AS 12	GS 11	1	4.867
				4.172
IND: 55	AL 12	GL 11	0.998	6.311
				5.173
IND: 56	AL 12	GL 11	0.998	5.958
				5.638
IND: 57	AL 11	GL 11	1	6.143
				6.143
IND: 58	GL 11	GL 11	1	5.987
				5.987
IND: 59	AL 12	AL 12	1	5.151
				5.151
IND: 60	AS 12	GL 11	0.943	6.142
				4.513

Legend: TR size of the Nigerian individuals: The green boxes represent the A L 12 alleles that are without doubt in the 5 to 5.5Kb range; the yellow boxes represent the A L 12 alleles that have 50% chance of being in the 5 to 5.5Kb range; the purple boxes represent the A L 12 alleles that are not binned in the 5-5.5Kb range. In the TR size column, the number in bold represent non-A L 12 alleles in the 5 to 5.5Kb range.

APPENDIX 2

Raw data on the new gastritis cohort for the Northern European individuals

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M186	FORMER HLO GASTRITIS	5.109	3.722	LS	12,12	AA	AG	TT	CT	AG	AA	AG
M187	NORMAL	6.415	4.931	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M188	NORMAL	5.586	3.737	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M189	NORMAL	3.723	3.723	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M190	FORMER HLO GASTRITIS	5.007	3.729	LS	11,12	AG	AG	CT	CT	AG	AA	AG
M191	FORMER HLO GASTRITIS	6.278	6.076	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M194	HLO GASTRITIS	6.24	3.665	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M195	NORMAL	5.551	3.695	LS	12,13	AA	AG	TT	CT	AG	AA	AG
M196	HLO GASTRITIS	6.293	3.057	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M199	NORMAL	6.065	4.181	LS	11,12	AG	AG	CT	CT	AG	TT	AA
M204	NORMAL	3.712	3.54	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M205	FORMER HLO GASTRITIS	5.168	5.168	LL	11,12	AG	GG	CT	TT	GG	AT	AA
M206	HLO GASTRITIS	3.775	3.626	SS	11,12	AG	AG	CT	n.d.	AG	AT	AG
M207	NORMAL	5.146	3.682	LS	12,12	AA	AG	TT	n.d.	AG	AA	AG
M208	HLO GASTRITIS	5.144	3.693	LS	12,12	AA	AG	TT	CT	AG	AA	AG
M210	HLO GASTRITIS	6.01	4.192	LS	11,12	AG	AA	CT	n.d.	AG	TT	AA
M211	HLO GASTRITIS	3.705	3.524	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M212	FORMER HLO GASTRITIS	3.695	3.695	SS	12,12	AA	AA	TT	n.d.	AA	AA	GG
M213	FORMER HLO GASTRITIS	5.086	3.604	LS	11,12	AG	GG	CC	TT	GG	AT	AA
M215	GASTRITIS NO HLO	5.846	3.675	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M217	FORMER HLO GASTRITIS	5.878	3.192	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M218	HLO GASTRITIS	5.308	3.564	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M219	FORMER HLO GASTRITIS	6.169	6.169	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M220	FORMER HLO GASTRITIS	5.106	3.613	LS	12,12	AA	AG	TT	CT	AG	AA	AG
M221	HLO GASTRITIS	6.281	3.658	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M226	FORMER HLO GASTRITIS	6.299	6.064	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M227	FORMER HLO GASTRITIS	6.121	3.591	LS	11,12	AG	AG	CT	CT	AG	TT	AA
M229	FORMER HLO GASTRITIS	6.981	6.12	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M232	NORMAL	5.981	3.635	LS	11,12	AG	AG	CT	CT	AG	AT	AG

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M233	NORMAL	6.322	3.685	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M234	NORMAL	6.357	5.952	LL	11,11	GG	GG	CC	TT	GG	TT	AG
M235	FORMER HLO GASTRITIS	5.823	3.649	LS	11,12	AG	AG	CT	CT	AG	AA	GG
M237	FORMER HLO GASTRITIS	8.447	5.716	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M238	FORMER HLO GASTRITIS	3.726	3.606	SS	11,12	AG	AG	CT	CT	AG	AA	GG
M239	FORMER HLO GASTRITIS	4.124	3.454	SS	12,12	AA	AA	TT	CC	AA	AT	AG
M240	FORMER HLO GASTRITIS	6.323	3.636	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M241	FORMER HLO GASTRITIS	5.456	3.698	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M242	HLO GASTRITIS	5.1	3.672	LS	12,12	AA	AG	CT	CT	AG	AA	AG
M246	NORMAL	5.707	3.679	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M248	GASTRITIS NO HLO	3.726	3.682	SS	12,12	AA	AA	TT	CC	AA	AT	AG
M249	HLO GASTRITIS	4.087	3.68	SS	12,12	AA	AG	TT	CT	AG	AA	AG
M250	NORMAL	5.993	3.697	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M253	FORMER HLO GASTRITIS	5.737	4.088	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M254	FORMER HLO GASTRITIS	6.094	3.576	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M255	FORMER HLO GASTRITIS	6.231	3.219	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M264	FORMER HLO GASTRITIS	3.782	2.943	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M265	FORMER HLO GASTRITIS	5.734	4.149	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M267	FORMER HLO GASTRITIS	5.755	3.722	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M269	GASTRITIS NO HLO	6.399	5.603	LL	11,11	GG	GG	CC	TT	GG	AT	AG
M273	NORMAL	6.307	3.684	LS	11,12	AG	AG	CT	CT	AG	AT	GG
M274	HLO GASTRITIS	5.197	3.432	LS	12,12	AA	AG	TT	CT	AG	AA	AG
M277	FORMER HLO GASTRITIS	5.385	3.674	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M278	FORMER HLO GASTRITIS	6.076	3.667	LS	11,12	AG	AG	CT	CT	AG	TT	AA
M279	GASTRITIS NO HLO	6.112	5.369	LL	11,13	AG	GG	CT	TT	GG	AT	AG
M281	NORMAL	5.906	4.111	LS	11,12	AG	AG	CT	CT	AG	TT	AA
M282	HLO GASTRITIS	5.928	3.743	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M284	NORMAL	6.004	4.069	LS	11,12	AG	GG	CT	TT	GG	AT	AA

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M285	GASTRITIS NO HLO	5.945	3.626	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M289	HLO GASTRITIS	6.279	6.088	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M291	HLO GASTRITIS	5.212	5.212	LL	12,13	AA	GG	TT	TT	GG	AA	AG
M293	NORMAL	4.419	3.747	SS	11,11	GG	GG	CC	TT	GG	TT	AA
M294	FORMER HLO GASTRITIS	6.131	4.066	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M295	FORMER HLO GASTRITIS	6.004	4.138	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M296	GASTRITIS NO HLO	4.118	3.672	SS	12,13	AA	AA	TT	CC	AA	AA	GG
M297	GASTRITIS NO HLO	6.113	3.627	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M298	NORMAL	6.306	5.398	LL	11,13	AG	GG	CT	TT	GG	AA	GG
M298A	GASTRITIS NO HLO	6.222	4.781	LS	11,13	AG	GG	CT	TT	GG	AT	AG
M301	NORMAL	6.394	6.084	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M303	GASTRITIS NO HLO	6.542	4.111	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M312	NORMAL	6.196	5.975	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M313	GASTRITIS NO HLO	6.196	3.005	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M317	NORMAL	6.064	6.064	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M318	GASTRIC CANCER	5.647	3.737	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M320	NORMAL	6.257	5.704	LL	11,12	AG	GG	CC	CT	AG	AT	AG
M322	NORMAL	5.591	3.77	LS	12,13	AA	AG	TT	CT	AG	AA	GG
M323	NORMAL	6.491	3.689	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M324	HLO GASTRITIS	5.647	5.317	LL	11,12	AG	GG	CT	TT	GG	AT	AA
M327	HLO GASTRITIS	6.162	5.913	LL	11,11	GG	GG	CC	TT	GG	TT	AG
M330	NORMAL	6.03	6.03	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M334	HLO GASTRITIS	6.048	5.498	LL	11,13	AG	GG	CT	TT	GG	AT	AG
M335	NORMAL	6.252	3.648	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M336	NORMAL	6.399	3.712	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M338	HLO GASTRITIS	6.204	4.133	LS	11,12	AG	GG	CT	TT	GG	AT	AG
M339	NORMAL	5.432	3.801	LS	11,13	AG	AG	CT	CT	AG	AT	AG
M340	NORMAL	6.325	5.989	LL	11,11	GG	GG	CC	TT	GG	TT	AA

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M341	NORMAL	6.292	3.733	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M342	FORMER HLO GASTRITIS	6.408	6.408	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M344	NORMAL	6.135	3.726	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M345	FORMER HLO GASTRITIS	3.752	3.752	SS	12,13	AA	AA	TT	CC	AA	AA	GG
M346	HLO GASTRITIS	3.641	3.046	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M347	NORMAL	3.742	3.742	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M348	HLO GASTRITIS	5.843	2.902	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M349	NORMAL	3.668	3.668	SS	12,12	AA	AA	TT	CC	GG	AT	GG
M350	NORMAL	7.365	4.14	LS	11,12	AG	AG	CT	CT	GG	TT	AG
M351	FORMER HLO GASTRITIS	5.741	4.055	LS	11,12	AG	GG	CT	TT	GG	AT	AA
M352	GASTRITIS NO HLO	4.071	3.661	SS	12,12	AA	AG	TT	CT	AG	AA	AG
M353	NORMAL	4.972	3.594	SS	12,12	AA	AG	TT	CT	AG	AA	AG
M354	FORMER HLO GASTRITIS	6.039	4.09	LS	11,13	AG	AG	CT	CT	AG	AT	AG
M355	NORMAL	4.157	3.675	SS	12,12	AA	AA	TT	CC	AA	AT	GG
M356	NORMAL	5.422	3.688	LS	12,13	AA	AG	TT	CT	AA	AA	AG
M357	HLO GASTRITIS	6.549	6.208	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M358	NORMAL	6.526	5.502	LL	11,13	AG	GG	CT	TT	AG	AT	AA
M359	HLO GASTRITIS	6.142	3.824	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M360	FORMER HLO GASTRITIS	7.983	3.421	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M361	HLO GASTRITIS	3.75	3.75	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M362	HLO GASTRITIS	5.908	3.165	LS	11,12	AG	AA	CT	CT	GG	TT	AG
M363	HLO GASTRITIS	6.306	3.788	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M365	NORMAL	3.664	3.122	SS	11,12	AG	AG	CT	CT	AG	AT	AG
M367	FORMER HLO GASTRITIS	4.02	3.732	SS	12,12	AA	AG	TT	CT	AA	AA	AG
M368	NORMAL	5.951	5.103	LL	11,12	AG	GG	CT	TT	GG	AT	AA
M369	NORMAL	6.734	4.229	LS	11,12	AG	AG	CT	CT	GG	TT	AG
M371	FORMER HLO GASTRITIS	6.799	5.361	LL	12,13	AA	AG	TT	CT	AA	AA	AG
M373	GASTRITIS NO HLO	3.705	3.705	SS	12,12	AA	AA	TT	CC	AA	AA	GG

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M374	NORMAL	3.689	3.689	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M375	HLO GASTRITIS	6.122	5.058	LL	11,12	AG	GG	CT	TT	GG	AT	AA
M376	NORMAL	3.652	3.069	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M377	NORMAL	6.144	3.717	LS	11,11	GG	GG	CC	TT	GG	AT	AA
M378	HLO GASTRITIS	6.332	6.332	LL	11,11	GG	GG	CC	TT	GG	AT	AA
M379	FORMER HLO GASTRITIS	3.763	3.763	SS	11,12	AG	AA	CT	CT	AG	AT	AG
M382	NORMAL	4.217	3.722	SS	12,13	AA	AA	TT	CC	AA	AA	GG
M383	FORMER HLO GASTRITIS	6.212	5.235	LL	11,13	AG	GG	CT	TT	GG	AT	AA
M384	HLO GASTRITIS	4.143	3.866	SS	12,12	AA	AG	TT	CT	AG	AA	AG
M385	FORMER HLO GASTRITIS	6.212	3.645	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M386	GASTRITIS NO HLO	6.067	5.12	LL	11,12	AG	GG	CT	TT	GG	AT	AA
M387	NORMAL	6.251	3.79	LS	11,11	GG	GG	CC	TT	GG	TT	AA
M389	NORMAL	5.121	3.726	LS	12,12	AA	AG	TT	CT	GG	AA	AG
M390	GASTRITIS NO HLO	6.347	6.19	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M391	FORMER HLO GASTRITIS	6.219	4.187	LS	11,13	AG	AA	CT	CT	AG	AT	AG
M392	NORMAL	5.872	3.708	LS	11,12	AG	AA	CT	CT	AG	AT	AG
M393	FORMER HLO GASTRITIS	6.363	3.57	LS	11,12	AG	AA	CT	CT	AG	AT	AG
M394	NORMAL	5.767	5.522	LL	11,13	AG	GG	CT	TT	AG	AT	AA
M395	GASTRITIS NO HLO	3.596	3.067	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M396	NORMAL	6.25	6.099	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M397	NORMAL	6.125	3.612	LS	11,12	AG	AA	CT	CT	AG	AT	AG
M398	HLO GASTRITIS	4.058	3.675	SS	11,12	AG	GG	CT	TT	GG	AT	AA
M399	HLO GASTRITIS	6.026	3.678	LS	11,12	AG	AA	CT	CT	AG	AT	AG
M400	NORMAL	3.894	3.789	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M401	FORMER HLO GASTRITIS	4.264	3.76	SS	12,13	AA	AA	TT	CC	AG	AT	GG
M402	NORMAL	6.094	6.094	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M403	FORMER HLO GASTRITIS	6.37	5.966	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M404	FORMER HLO GASTRITIS	6.326	5.925	LL	11,11	GG	GG	CC	TT	GG	TT	AA

Survey number	Categorisation	MUC1 TR size (Kb)		TR size L/S	CA repeat	g/a SNP	THBS3 PCR1	THBS3 PCR2	rs2070803	497 SNP	KCP2	EFNA1
M406	NORMAL	4.279	3.804	SS	12,13	AA	AA	TT	CC	AA	AA	GG
M407	NORMAL	4.312	3.757	SS	12,13	AA	AA	TT	CC	AA	AA	GG
M408	NORMAL	6.023	5.267	LL	11,13	AG	GG	CT	TT	GG	AT	AA
M409	HLO GASTRITIS	6.122	5.974	LL	11,11	GG	GG	CC	TT	GG	AT	AA
M410	NORMAL	6.097	3.802	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M411	NORMAL	6.167	3.815	LS	11,12	AG	AG	CT	CT	AG	AT	AG
M412	GASTRITIS NO HLO	8.822	6.371	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M415	FORMER HLO GASTRITIS	5.685	3.8	LS	12,12	AG	AG	CT	CC	AA	AA	GG
M416	GASTRITIS NO HLO	6.294	6.294	LL	11,11	GG	GG	CC	TT	GG	TT	AA
M417	HLO GASTRITIS	3.055	2.931	SS	12,12	AA	AA	TT	CC	AA	AA	GG
M418	NORMAL	4.277	3.788	SS	12,13	AA	AA	TT	CC	AG	AT	AG
M419	NORMAL	3.848	3.73	SS	11,11	GG	GG	CC	TT	GG	TT	AA
M420	GASTRITIS NO HLO	6.358	3.863	LS	11,12	AG	AG	CT	CT	AG	TT	AG

References

1. 1994a. Infection with Helicobacter pylori. *IARC Monogr Eval. Carcinog. Risks Hum.* 61:177-240.

2. 1994b. Schistosomes, liver flukes and Helicobacter pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7-14 June 1994. *IARC Monogr Eval. Carcinog. Risks Hum.* 61:1-241.

3. Adolph, K.W. and P.Bornstein. 1999. The human thrombospondin 3 gene: analysis of transcription initiation and an alternatively spliced transcript. *Mol. Cell Biol. Res. Commun.* 2:47-52.

4. Adolph, K.W., G.L.Long, S.Winfield, E.I.Ginns, and P.Bornstein. 1995. Structure and organization of the human thrombospondin 3 gene (THBS3). *Genomics* 27:329-336.

5. Akada, J.K., M.Shirai, H.Takeuchi, M.Tsuda, and T.Nakazawa. 2000. Identification of the urease operon in Helicobacter pylori and its control by mRNA decay in response to pH. *Mol. Microbiol.* 36:1071-1084.

6. Ando, I., A.Kukita, G.Soma, and H.Hino. 1998. A large number of tandem repeats in the polymorphic epithelial mucin gene is associated with severe acne. *J. Dermatol.* 25:150-152.

7. Armstrong, L.C., T.Komiya, B.E.Bergman, K.Mihara, and P.Bornstein. 1997. Metaxin is a component of a preprotein import complex in the outer membrane of the mammalian mitochondrion. *J. Biol. Chem.* 272:6510-6518.

8. Ashall, F., M.E.Bramwell, and H.Harris. 1982. A new marker for human cancer cells. 1 The Ca antigen and the Ca1 antibody. *Lancet* 2:1-6.

9. Baruch, A., M.Hartmann, M.Yoeli, Y.Adereth, S.Greenstein, Y.Stadler, Y.Skornik, J.Zaretsky, N.I.Smorodinsky, I.Keydar, and D.H.Wreschner. 1999. The breast cancer-associated MUC1 gene generates both a receptor and its cognate binding protein. *Cancer Res.* 59:1552-1561.

10. Baruch, A., M.Hartmann, S.Zrihan-Licht, S.Greenstein, M.Burstein, I.Keydar, M.Weiss, N.Smorodinsky, and D.H.Wreschner. 1997. Preferential expression of novel MUC1 tumor antigen isoforms in human epithelial tumors and their tumor-potentiating function. *Int. J. Cancer* 71:741-749.

11. Beutler, E. 1993. Gaucher disease as a paradigm of current issues regarding single gene mutations of humans. *Proc. Natl. Acad. Sci. U. S. A* 90:5384-5390.

12. Bonkobara, M., A.Das, J.Takao, P.D.Cruz, and K.Ariizumi. 2003. Identification of novel genes for secreted and membrane-anchored proteins in human keratinocytes. *Br. J. Dermatol.* 148:654-664.

13. Boren, T., P.Falk, K.A.Roth, G.Larson, and S.Normark. 1993. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* 262:1892-1895.

14. Bork, P. and L.Patthy. 1995. The SEA module: a new extracellular domain associated with O-glycosylation. *Protein Sci.* 4:1421-1425.

15. Bornstein, P., C.E.McKinney, M.E.LaMarca, S.Winfield, T.Shingu, S.Devarayalu, H.L.Vos, and E.I.Ginns. 1995. Metaxin, a gene contiguous to both thrombospondin 3 and glucocerebrosidase, is required for embryonic development in the mouse: implications for Gaucher disease. *Proc. Natl. Acad. Sci. U. S. A* 92:4547-4551.

16. Boyle, P. and J.Ferlay. 2005. Cancer incidence and mortality in Europe, 2004. *Ann. Oncol.* 16:481-488.

17. Bramwell, M.E., V.P.Bhavanandan, G.Wiseman, and H.Harris. 1983. Structure and function of the Ca antigen. *Br. J. Cancer* 48:177-183.

18. Bramwell, M.E., A.K.Ghosh, W.D.Smith, G.Wiseman, A.I.Spriggs, and H.Harris. 1985. Ca2 and Ca3. New monoclonal antibodies evaluated as tumor markers in serous effusions. *Cancer* 56:105-110.

19. Bry, L., P.G.Falk, T.Midtvedt, and J.I.Gordon. 1996. A model of host-microbial interactions in an open mammalian ecosystem. *Science* 273:1380-1383.

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

21. Burchell, J., H.Durbin, and J.Taylor-Papadimitriou. 1983. Complexity of expression of antigenic determinants, recognized by monoclonal antibodies HMFG-1 and HMFG-2, in normal and malignant human mammary epithelial cells. *J. Immunol.* 131:508-513.

22. Burchell, J.M., A.Mungul, and J.Taylor-Papadimitriou. 2001. O-linked glycosylation in the mammary gland: changes that occur during malignancy. *J. Mammary. Gland. Biol. Neoplasia.* 6:355-364.

23. Cardon, L.R. and L.J.Palmer. 2003. Population stratification and spurious allelic association. *Lancet* 361:598-604.

24. Carraway, K.L., V.P.Ramsauer, B.Haq, and C.A.Carothers Carraway. 2003. Cell signaling through membrane mucins. *Bioessays* 25:66-71.

25. Cartron, J.P., M.Kornprobst, M.Lemonnier, P.Lambin, F.Piller, and C.Salmon. 1982. Isolation from human urines of a mucin with blood group SDa activity. *Biochem. Biophys. Res. Commun.* 106:331-337.

26. Carvalho, F., A.Peixoto, R.Steffensen, A.Amorim, L.David, and M.Sobrinho-Simoes. 1999. MUC1 gene polymorphism does not explain the different incidence of gastric cancer in Portugal and Denmark. *Ann. Hum. Genet.* 63 (Pt 3):187-191.

27. Carvalho, F., R.Seruca, L.David, A.Amorim, M.Seixas, E.Bennett, H.Clausen, and M.Sobrinho-Simoes. 1997. MUC1 gene polymorphism and gastric cancer--an epidemiological study. *Glycoconj. J.* 14:107-111.

28. Censini, S., C.Lange, Z.Xiang, J.E.Crabtree, P.Ghiara, M.Borodovsky, R.Rappuoli, and A.Covacci. 1996. cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. U. S. A* 93:14648-14653.

29. Clayton, C.L., M.J.Pallen, H.Kleanthous, B.W.Wren, and S.Tabaqchali. 1990. Nucleotide sequence of two genes from Helicobacter pylori encoding for urease subunits. *Nucleic Acids Res.* 18:362.

30. Clayton, C.L., B.W.Wren, P.Mullany, A.Topping, and S.Tabaqchali. 1989. Molecular cloning and expression of Campylobacter pylori species-specific antigens in Escherichia coli K-12. *Infect. Immun.* 57:623-629.

31. Correa, P. 2005. New strategies for the prevention of gastric cancer: Helicobacter pylori and genetic susceptibility. *J. Surg. Oncol.* 90:134-138.

32. Correa, P., W.Haenszel, C.Cuello, S.Tannenbaum, and M.Archer. 1975. A model for gastric cancer epidemiology. *Lancet* 2:58-60.

33. Covacci, A., S.Censini, M.Bugnoli, R.Petracca, D.Burroni, G.Macchia, A.Massone, E.Papini, Z.Xiang, N.Figura, and . 1993. Molecular characterization of the 128-kDa immunodominant antigen of Helicobacter pylori associated with cytotoxicity and duodenal ulcer. *Proc. Natl. Acad. Sci. U. S. A* 90:5791-5795.

34. Cover, T.L. and M.J.Blaser. 1992. Purification and characterization of the vacuolating toxin from Helicobacter pylori. *J. Biol. Chem.* 267:10570-10575.

35. Dixit, V.M., S.Green, V.Sarma, L.B.Holzman, F.W.Wolf, K.O'Rourke, P.A.Ward, E.V.Prochownik, and R.M.Marks. 1990. Tumor necrosis factor-alpha induction of novel gene products in human endothelial cells including a macrophage-specific chemotaxin. *J. Biol. Chem.* 265:2973-2978.

36. Dixon, M.F., R.M.Genta, J.H.Yardley, and P.Correa. 1996. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *Am. J. Surg. Pathol.* 20:1161-1181.

37. El Omar, E.M. 2001. The importance of interleukin 1beta in Helicobacter pylori associated disease. *Gut* 48:743-747.

38. El Omar, E.M., M.Carrington, W.H.Chow, K.E.McColl, J.H.Bream, H.A.Young, J.Herrera, J.Lissowska, C.C.Yuan, N.Rothman, G.Lanyon, M.Martin, J.F.Fraumeni, Jr., and C.S.Rabkin. 2000. Interleukin-1 polymorphisms associated with increased risk of gastric cancer. *Nature* 404:398-402.

39. El Omar, E.M., M.Carrington, W.H.Chow, K.E.McColl, J.H.Bream, H.A.Young, J.Herrera, J.Lissowska, C.C.Yuan, N.Rothman, G.Lanyon,

M.Martin, J.F.Fraumeni, Jr., and C.S.Rabkin. 2001. The role of interleukin-1 polymorphisms in the pathogenesis of gastric cancer. *Nature* 412:99.

40. Engelmann, K., S.E.Baldus, and F.G.Hanisch. 2001. Identification and topology of variant sequences within individual repeat domains of the human epithelial tumor mucin MUC1. *J. Biol. Chem.* 276:27764-27769.

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

42. Falush, D., T.Wirth, B.Linz, J.K.Pritchard, M.Stephens, M.Kidd, M.J.Blaser, D.Y.Graham, S.Vacher, G.I.Perez-Perez, Y.Yamaoka, F.Megraud, K.Otto, U.Reichard, E.Katzowitsch, X.Wang, M.Achtman, and S.Suerbaum. 2003. Traces of human migrations in Helicobacter pylori populations. *Science* 299:1582-1585.

43. Figueiredo, C., L.J.van Doorn, C.Nogueira, J.M.Soares, C.Pinho, P.Figueira, W.G.Quint, and F.Carneiro. 2001. Helicobacter pylori genotypes are associated with clinical outcome in Portuguese patients and show a high prevalence of infections with multiple strains. *Scand. J. Gastroenterol.* 36:128-135.

44. Filipe, M.I., N.Munoz, I.Matko, I.Kato, V.Pompe-Kirn, A.Jutersek, S.Teuchmann, M.Benz, and T.Prijon. 1994. Intestinal metaplasia types and the risk of gastric cancer: a cohort study in Slovenia. *Int. J. Cancer* 57:324-329.

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

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

47. Freeman, B., J.Powell, D.Ball, L.Hill, I.Craig, and R.Plomin. 1997. DNA by mail: an inexpensive and noninvasive method for collecting DNA samples from widely dispersed populations. *Behav. Genet.* 27:251-257.

48. Gabriel, S.B., S.F.Schaffner, H.Nguyen, J.M.Moore, J.Roy, B.Blumenstiel, J.Higgins, M.DeFelice, A.Lochner, M.Faggart, S.N.Liu-Cordero, C.Rotimi, A.Adeyemo, R.Cooper, R.Ward, E.S.Lander, M.J.Daly, and D.Altshuler. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225-2229.

49. Geis, G., S.Suerbaum, B.Forsthoff, H.Leying, and W.Opferkuch. 1993. Ultrastructure and biochemical studies of the flagellar sheath of Helicobacter pylori. *J. Med. Microbiol.* 38:371-377.

50. Gendler, S.J. 2001. MUC1, the renaissance molecule. *J. Mammary. Gland. Biol. Neoplasia.* 6:339-353.

51. Gendler, S.J., J.M.Burchell, T.Duhig, D.Lamport, R.White, M.Parker, and J.Taylor-Papadimitriou. 1987. Cloning of partial cDNA encoding differentiation and tumor-associated mucin glycoproteins expressed by human mammary epithelium. *Proc. Natl. Acad. Sci. U. S. A* 84:6060-6064.

52. Gendler, S.J., C.A.Lancaster, J.Taylor-Papadimitriou, T.Duhig, N.Peat, J.Burchell, L.Pemberton, E.N.Lalani, and D.Wilson. 1990. Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* 265:15286-15293.

53. Girling, A., J.Bartkova, J.Burchell, S.Gendler, C.Gillett, and J.Taylor-Papadimitriou. 1989. A core protein epitope of the polymorphic epithelial

mucin detected by the monoclonal antibody SM-3 is selectively exposed in a range of primary carcinomas. *Int. J. Cancer* 43:1072-1076.

54. Goldstein, D.B., G.L.Cavalleri, and K.R.Ahmadi. 2003. The genetics of common diseases: 10 million times as hard. *Cold Spring Harb. Symp. Quant. Biol.* 68:395-401.

55. Goldstein, D.B. and M.E.Weale. 2001. Population genomics: linkage disequilibrium holds the key. *Curr. Biol.* 11:R576-R579.

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

57. Grabowski, G.A., S.Gatt, and M.Horowitz. 1990. Acid beta-glucosidase: enzymology and molecular biology of Gaucher disease. *Crit Rev. Biochem. Mol. Biol.* 25:385-414.

58. Griffiths, B., L.G.Bobrow, L.Happerfield, and D.M.Swallow. 1988a. Expression of the hypervariable PUM locus in normal and malignant lung: the tumor-associated epitopes are present but masked in normal tissue. *Dis. Markers* 6:195-202.

59. Griffiths, B., A.Gordon, J.Burchell, M.E.Bramwell, A.Griffiths, M.Price, J.Taylor-Papadimitriou, D.Zanin, and D.M.Swallow. 1988b. The breast tumour-associated epithelial mucins and the peanut lectin binding urinary mucins are coded by a single highly polymorphic gene locus 'PUM'. *Dis. Markers* 6:185-194.

60. Gross, M.S., V.Guyonnet-Duperat, N.Porchet, A.Bernheim, J.P.Aubert, and V.C.Nguyen. 1992. Mucin 4 (MUC4) gene: regional assignment (3q29) and RFLP analysis. *Ann. Genet.* 35:21-26.

61. Hanisch, F.G. and S.Muller. 2000. MUC1: the polymorphic appearance of a human mucin. *Glycobiology* 10:439-449.

62. Hanisch, F.G., T.R.Stadie, F.Deutzmann, and J.Peter-Katalinic. 1996. MUC1 glycoforms in breast cancer--cell line T47D as a model for carcinoma-associated alterations of 0-glycosylation. *Eur. J. Biochem.* 236:318-327.

63. Hareuveni, M., I.Tsarfaty, J.Zaretsky, P.Kotkes, J.Horev, S.Zrihan, M.Weiss, S.Green, R.Lathe, I.Keydar, and . 1990. A transcribed gene, containing a variable number of tandem repeats, codes for a human epithelial tumor antigen. cDNA cloning, expression of the transfected gene and over-expression in breast cancer tissue. *Eur. J. Biochem.* 189:475-486.

64. Hartman, M., A.Baruch, I.Ron, Y.Aderet, M.Yoeli, O.Sagi-Assif, S.Greenstein, Y.Stadler, M.Weiss, E.Harness, M.Yaakubovits, I.Keydar, N.I.Smorodinsky, and D.H.Wreschner. 1999. MUC1 isoform specific monoclonal antibody 6E6/2 detects preferential expression of the novel MUC1/Y protein in breast and ovarian cancer. *Int. J. Cancer* 82:256-267.

65. Heneghan, M.A., A.P.Moran, K.M.Feeley, E.L.Egan, J.Goulding, C.E.Connolly, and C.F.McCarthy. 1998. Effect of host Lewis and ABO blood group antigen expression on Helicobacter pylori colonisation density and the consequent inflammatory response. *FEMS Immunol. Med. Microbiol.* 20:257-266.

66. Hilkens, J. and F.Buijs. 1988. Biosynthesis of MAM-6, an epithelial sialomucin. Evidence for involvement of a rare proteolytic cleavage step in the endoplasmic reticulum. *J. Biol. Chem.* 263:4215-4222.

67. Hilkens, J., F.Buijs, J.Hilgers, P.Hageman, J.Calafat, A.Sonnenberg, and d.van, V. 1984. Monoclonal antibodies against human milk-fat globule membranes detecting differentiation antigens of the mammary gland and its tumors. *Int. J. Cancer* 34:197-206.

68. Ho, S.B., L.L.Shekels, N.W.Toribara, Y.S.Kim, C.Lyftogt, D.L.Cherwitz, and G.A.Niehans. 1995. Mucin gene expression in normal, preneoplastic, and neoplastic human gastric epithelium. *Cancer Res.* 55:2681-2690.

69. Hollox, E.J., M.Poulter, M.Zvarik, V.Ferak, A.Krause, T.Jenkins, N.Saha, A.I.Kozlov, and D.M.Swallow. 2001. Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* 68:160-172.

70. Holzman, L.B., R.M.Marks, and V.M.Dixit. 1990. A novel immediate-early response gene of endothelium is induced by cytokines and encodes a secreted protein. *Mol. Cell Biol.* 10:5830-5838.

71. Horowitz, M., S.Wilder, Z.Horowitz, O.Reiner, T.Gelbart, and E.Beutler. 1989. The human glucocerebrosidase gene and pseudogene: structure and evolution. *Genomics* 4:87-96.

72. Jeffreys, A.J., J.K.Holloway, L.Kauppi, C.A.May, R.Neumann, M.T.Slingsby, and A.J.Webb. 2004. Meiotic recombination hot spots and human DNA diversity. *Philos. Trans. R. Soc. Lond B Biol. Sci.* 359:141-152.

73. Jeffreys, A.J., A.MacLeod, K.Tamaki, D.L.Neil, and D.G.Monckton. 1991. Minisatellite repeat coding as a digital approach to DNA typing. *Nature* 354:204-209.

74. Jeffreys, A.J., R.Neumann, M.Panayi, S.Myers, and P.Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.* 37:601-606.

75. Jeffreys, A.J., R.Neumann, and V.Wilson. 1990. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60:473-485.

76. Josenhans, C., L.Vossebein, S.Friedrich, and S.Suerbaum. 2002. The neuA/flmD gene cluster of Helicobacter pylori is involved in flagellar biosynthesis and flagellin glycosylation. *FEMS Microbiol. Lett.* 210:165-172.

77. Karlsson, S., D.M.Swallow, B.Griffiths, G.Corney, D.A.Hopkinson, A.Dawnay, and J.P.Cartron. 1983. A genetic polymorphism of a human urinary mucin. *Ann. Hum. Genet.* 47 (Pt 4):263-269.

78. Kauser, F., M.A.Hussain, I.Ahmed, S.Srinivas, S.M.Devi, A.A.Majeed, K.R.Rao, A.A.Khan, L.A.Sechi, and N.Ahmed. 2005. Comparative genomics of Helicobacter pylori isolates recovered from ulcer disease patients in England. *BMC. Microbiol.* 5:32.

79. Kostrzynska, M., J.D.Betts, J.W.Austin, and T.J.Trust. 1991. Identification, characterization, and spatial localization of two flagellin species in Helicobacter pylori flagella. *J. Bacteriol.* 173:937-946.

80. Kovarik, A., P.J.Lu, N.Peat, J.Morris, and J.Taylor-Papadimitriou. 1996. Two GC boxes (Sp1 sites) are involved in regulation of the activity of the epithelium-specific MUC1 promoter. *J. Biol. Chem.* 271:18140-18147.

81. Kovarik, A., N.Peat, D.Wilson, S.J.Gendler, and J.Taylor-Papadimitriou. 1993. Analysis of the tissue-specific promoter of the MUC1 gene. *J. Biol. Chem.* 268:9917-9926.

82. Kozak, M. 1987a. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15:8125-8148.

83. Kozak, M. 1987b. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* 196:947-950.

84. Kufe, D., G.Inghirami, M.Abe, D.Hayes, H.Justi-Wheeler, and J.Schlom. 1984. Differential reactivity of a novel monoclonal antibody (DF3) with human malignant versus benign breast tumors. *Hybridoma* 3:223-232.

85. Labigne, A., V.Cussac, and P.Courcoux. 1991. Shuttle cloning and nucleotide sequences of Helicobacter pylori genes responsible for urease activity. *J. Bacteriol.* 173:1920-1931.

86. LaMarca, M.E., M.Goldstein, N.Tayebi, M.Arcos-Burgos, B.M.Martin, and E.Sidransky. 2004. A novel alteration in metaxin 1, F202L, is associated with N370S in Gaucher disease. *J. Hum. Genet.* 49:220-222.

87. Lamarque, D. and M.Peek. 2003. Pathogenesis of Helicobacter pylori infection. *Helicobacter.* 8 Suppl 1:21-30.

88. Lan, M.S., S.K.Batra, W.N.Qi, R.S.Metzgar, and M.A.Hollingsworth. 1990. Cloning and sequencing of a human pancreatic tumor mucin cDNA. *J. Biol. Chem.* 265:15294-15299.

89. Lancaster, C.A., N.Peat, T.Duhig, D.Wilson, J.Taylor-Papadimitriou, and S.J.Gendler. 1990. Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochem. Biophys. Res. Commun.* 173:1019-1029.

90. LAUREN, P. 1965. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. *Acta Pathol. Microbiol. Scand.* 64:31-49.

91. Leunk, R.D., P.T.Johnson, B.C.David, W.G.Kraft, and D.R.Morgan. 1988. Cytotoxic activity in broth-culture filtrates of Campylobacter pylori. *J. Med. Microbiol.* 26:93-99.

92. Leying, H., S.Suerbaum, G.Geis, and R.Haas. 1992. Cloning and genetic characterization of a Helicobacter pylori flagellin gene. *Mol. Microbiol.* 6:2863-2874.

93. Li, Y., A.Bharti, D.Chen, J.Gong, and D.Kufe. 1998. Interaction of glycogen synthase kinase 3beta with the DF3/MUC1 carcinoma-associated antigen and beta-catenin. *Mol. Cell Biol.* 18:7216-7224.

94. Ligtenberg, M.J., A.M.Gennissen, H.L.Vos, and J.Hilkens. 1991. A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA. *Nucleic Acids Res.* 19:297-301.

95. Ligtenberg, M.J., L.Kruijshaar, F.Buijs, M.van Meijer, S.V.Litvinov, and J.Hilkens. 1992. Cell-associated episialin is a complex containing two proteins derived from a common precursor. *J. Biol. Chem.* 267:6171-6177.

96. Ligtenberg, M.J., H.L.Vos, A.M.Gennissen, and J.Hilkens. 1990. Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini. *J. Biol. Chem.* 265:5573-5578.

97. Lillehoj, E.P., S.W.Hyun, B.T.Kim, X.G.Zhang, D.I.Lee, S.Rowland, and K.C.Kim. 2001. Muc1 mucins on the cell surface are adhesion sites for Pseudomonas aeruginosa. *Am. J. Physiol Lung Cell Mol. Physiol* 280:L181-L187.

98. Lillehoj, E.P., B.T.Kim, and K.C.Kim. 2002. Identification of Pseudomonas aeruginosa flagellin as an adhesin for Muc1 mucin. *Am. J. Physiol Lung Cell Mol. Physiol* 282:L751-L756.

99. Lillehoj, E.P., H.Kim, E.Y.Chun, and K.C.Kim. 2004. Pseudomonas aeruginosa Stimulates Phosphorylation of the Epithelial Membrane Glycoprotein Muc1 and Activates MAP Kinase. *Am. J. Physiol Lung Cell Mol. Physiol.*

100. Linden, S., J.Mahdavi, J.Hedenbro, T.Boren, and I.Carlstedt. 2004. Effects of pH on Helicobacter pylori binding to human gastric mucins: identification of binding to non-MUC5AC mucins. *Biochem. J.* Pt.

101. Litvinov, S.V. and J.Hilkens. 1993. The epithelial sialomucin, episialin, is sialylated during recycling. *J. Biol. Chem.* 268:21364-21371.

102. Lloyd, K.O., J.Burchell, V.Kudryashov, B.W.Yin, and J.Taylor-Papadimitriou. 1996. Comparison of O-linked carbohydrate chains in MUC-1 mucin from normal breast epithelial cell lines and breast carcinoma cell lines. Demonstration of simpler and fewer glycan chains in tumor cells. *J. Biol. Chem.* 271:33325-33334.

103. Long, G.L., S.Winfield, K.W.Adolph, E.I.Ginns, and P.Bornstein. 1996. Structure and organization of the human metaxin gene (MTX) and pseudogene. *Genomics* 33:177-184.

104. Lunet, N. and H.Barros. 2003. Helicobacter pylori infection and gastric cancer: facing the enigmas. *Int. J. Cancer* 106:953-960.

105. Machado, J.C., C.Figueiredo, P.Canedo, P.Pharoah, R.Carvalho, S.Nabais, A.C.Castro, M.L.Campos, L.J.van Doorn, C.Caldas, R.Seruca, F.Carneiro, and M.Sobrinho-Simoes. 2003. A proinflammatory genetic profile increases the risk for chronic atrophic gastritis and gastric carcinoma. *Gastroenterology* 125:364-371.

106. Machado, J.C., P.Pharoah, S.Sousa, R.Carvalho, C.Oliveira, C.Figueiredo, A.Amorim, R.Seruca, C.Caldas, F.Carneiro, and M.Sobrinho-Simoes. 2001. Interleukin 1B and interleukin 1RN polymorphisms are associated with increased risk of gastric carcinoma. *Gastroenterology* 121:823-829.

107. Maeda, Y., S.Tanaka, J.Hino, K.Kangawa, and T.Kinoshita. 2000. Human dolichol-phosphate-mannose synthase consists of three subunits, DPM1, DPM2 and DPM3. *EMBO J.* 19:2475-2482.

108. Manos, E.J., M.L.Kim, J.Kassis, P.Y.Chang, A.Wells, and D.A.Jones. 2001. Dolichol-phosphate-mannose-3 (DPM3)/prostin-1 is a novel phospholipase C-gamma regulated gene negatively associated with prostate tumor invasion. *Oncogene* 20:2781-2790.

109. Marshall, B.J. and J.R.Warren. 1984. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* 1:1311-1315.

110. Marshall, B.J., J.R.Warren, G.J.Francis, S.R.Langton, C.S.Goodwin, and E.D.Blincow. 1987. Rapid urease test in the management of Campylobacter pyloridis-associated gastritis. *Am. J. Gastroenterol.* 82:200-210.

111. Mensdorff-Pouilly, S., L.Kinarsky, K.Engelmann, S.E.Baldus, R.H.Verheijen, M.A.Hollingsworth, V.Pisarev, S.Sherman, and F.G.Hanisch. 2005. Sequence-variant repeats of MUC1 show higher conformational flexibility, are less densely O-glycosylated and induce differential B lymphocyte responses. *Glycobiology* 15:735-746.

112. Meulenbelt, I., S.Droog, G.J.Trommelen, D.I.Boomsma, and P.E.Slagboom. 1995. High-yield noninvasive human genomic DNA isolation method for genetic studies in geographically dispersed families and populations. *Am. J. Hum. Genet.* 57:1252-1254.

113. Middleton-Price, H., S.Gendler, and S.Malcolm. 1988. Close linkage of PUM and SPTA within chromosome band 1q21. *Ann. Hum. Genet.* 52 (Pt 4):273-278.

114. Mobley, H.L., M.J.Cortesia, L.E.Rosenthal, and B.D.Jones. 1988. Characterization of urease from Campylobacter pylori. *J. Clin. Microbiol.* 26:831-836.

115. Morris, J.R. and J.Taylor-Papadimitriou. 2001. The Spl transcription factor regulates cell type-specific transcription of MUC1. *DNA Cell Biol.* 20:133-139.

116. Mullenbach, R., P.J.Lagoda, and C.Welter. 1989. An efficient salt-chloroform extraction of DNA from blood and tissues. *Trends Genet.* 5:391.

117. Muller, S., K.Alving, J.Peter-Katalinic, N.Zachara, A.A.Gooley, and F.G.Hanisch. 1999. High density O-glycosylation on tandem repeat peptide from secretory MUC1 of T47D breast cancer cells. *J. Biol. Chem.* 274:18165-18172.

118. Nguyen, V.C., J.P.Aubert, M.S.Gross, N.Porchet, P.Degand, and J.Frezal. 1990. Assignment of human tracheobronchial mucin gene(s) to 11p15 and a tracheobronchial mucin-related sequence to chromosome 13. *Hum. Genet.* 86:167-172.

119. Nogueira, C., C.Figueiredo, F.Carneiro, A.T.Gomes, R.Barreira, P.Figueira, C.Salgado, L.Belo, A.Peixoto, J.C.Bravo, L.E.Bravo, J.L.Realpe, A.P.Plaisier, W.G.Quint, B.Ruiz, P.Correa, and L.J.van Doorn. 2001. Helicobacter pylori genotypes may determine gastric histopathology. *Am. J. Pathol.* 158:647-654.

120. Obermair, A., B.C.Schmid, M.Stimpfl, B.Fasching, O.Preyer, S.Leodolter, A.J.Crandon, and R.Zeillinger. 2001. Novel MUC1 splice variants are expressed in cervical carcinoma. *Gynecol. Oncol.* 83:343-347.

121. Oluwasola, A.O. and J.O.Ogunbiyi. 2003. Gastric cancer: aetiological, clinicopathological and management patterns in Nigeria. *Niger. J. Med.* 12:177-186.

122. Oluwasola, A.O. and J.O.Ogunbiyi. 2004. Chronic gastritis and Helicobacter pylori infection in University College Hospital Ibadan, Nigeria--a study of 85 fibre optic gastric biopsies. *Niger. J. Med.* 13:372-378.

123. Oosterkamp, H.M., L.Scheiner, M.C.Stefanova, K.O.Lloyd, and C.L.Finstad. 1997. Comparison of MUC-1 mucin expression in epithelial and non-epithelial cancer cell lines and demonstration of a new short variant form (MUC-1/Z). *Int. J. Cancer* 72:87-94.

124. Parkin, D.M., F.Bray, J.Ferlay, and P.Pisani. 2005. Global cancer statistics, 2002. *CA Cancer J. Clin.* 55:74-108.

125. Parry, S., H.S.Silverman, K.McDermott, A.Willis, M.A.Hollingsworth, and A.Harris. 2001. Identification of MUC1 proteolytic cleavage sites in vivo. *Biochem. Biophys. Res. Commun.* 283:715-720.

126. Parsonnet, J., D.Vandersteen, J.Goates, R.K.Sibley, J.Pritikin, and Y.Chang. 1991. Helicobacter pylori infection in intestinal- and diffuse-type gastric adenocarcinomas. *J. Natl. Cancer Inst.* 83:640-643.

127. Pelham, H.R. 1996. The dynamic organisation of the secretory pathway. *Cell Struct. Funct.* 21:413-419.

128. Pemberton, L.F., A.Rughetti, J.Taylor-Papadimitriou, and S.J.Gendler. 1996. The epithelial mucin MUC1 contains at least two discrete signals specifying membrane localization in cells. *J. Biol. Chem.* 271:2332-2340.

129. Pigny, P., W.S.Pratt, A.Laine, A.Leclercq, D.M.Swallow, V.C.Nguyen, J.P.Aubert, and N.Porchet. 1995. The MUC5AC gene: RFLP analysis with the Jer58 probe. *Hum. Genet.* 96:367-368.

130. Pratt, W.S., I.Islam, and D.M.Swallow. 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.

131. Price, M.R., S.Edwards, M.Powell, and R.W.Baldwin. 1986. Epitope analysis of monoclonal antibody NCRC-11 defined antigen isolated from human ovarian and breast carcinomas. *Br. J. Cancer* 54:393-400.

132. Raymond, M. and F.Rousset. 1995. An Exact Test for Population Differentiation. *Evolution* 49:1280-1283.

133. Reiner, O., M.Wigderson, and M.Horowitz. 1988. Structural analysis of the human glucocerebrosidase genes. *DNA* 7:107-116.

134. Reis, C.A., L.David, P.Correa, F.Carneiro, C.De Bolos, E.Garcia, U.Mandel, H.Clausen, and M.Sobrinho-Simoes. 1999. Intestinal metaplasia of human stomach displays distinct patterns of mucin (MUC1, MUC2, MUC5AC, and MUC6) expression. *Cancer Res.* 59:1003-1007.

135. Reis, C.A., L.David, M.Seixas, J.Burchell, and M.Sobrinho-Simoes. 1998. Expression of fully and under-glycosylated forms of MUC1 mucin in gastric carcinoma. *Int. J. Cancer* 79:402-410.

136. Ren, J., Y.Li, and D.Kufe. 2002. Protein kinase C delta regulates function of the DF3/MUC1 carcinoma antigen in beta-catenin signaling. *J. Biol. Chem.* 277:17616-17622.

137. Rokkas, T., M.I.Filipe, and G.E.Sladen. 1991. Detection of an increased incidence of early gastric cancer in patients with intestinal metaplasia type III who are closely followed up. *Gut* 32:1110-1113.

138. Roussel, P., G.Lamblin, M.Lhermitte, N.Houdret, J.J.Lafitte, J.M.Perini, A.Klein, and A.Scharfman. 1988. The complexity of mucins. *Biochimie* 70:1471-1482.

139. Schut, I.C., P.M.Waterfall, M.Ross, C.O'Sullivan, W.R.Miller, F.K.Habib, and C.W.Bayne. 2003. MUC1 expression, splice variant and short form transcription (MUC1/Z, MUC1/Y) in prostate cell lines and tissue. *BJU. Int.* 91:278-283.

140. Shibatani, T., L.L.David, A.L.McCormack, K.Frueh, and W.R.Skach. 2005. Proteomic analysis of mammalian oligosaccharyltransferase reveals multiple subcomplexes that contain Sec61, TRAP, and two potential new subunits. *Biochemistry* 44:5982-5992.

141. Shimizu, M. and K.Yamauchi. 1982. Isolation and characterization of mucin-like glycoprotein in human milk fat globule membrane. *J. Biochem. (Tokyo)* 91:515-524.

142. Shiraga, T., D.Smith, H.N.Nuthall, M.A.Hollingsworth, and A.Harris. 2002. Identification of two novel elements involved in human MUCI gene expression in vivo. *Mol. Med.* 8:33-41.

143. Siddiqui, J., M.Abe, D.Hayes, E.Shani, E.Yunis, and D.Kufe. 1988. Isolation and sequencing of a cDNA coding for the human DF3 breast carcinoma-associated antigen. *Proc. Natl. Acad. Sci. U. S. A* 85:2320-2323.

144. Silva, E., A.Teixeira, L.David, F.Carneiro, C.A.Reis, J.Sobrinho-Simoes, J.Serpa, E.Veerman, J.Bolscher, and M.Sobrinho-Simoes. 2002. Mucins as

key molecules for the classification of intestinal metaplasia of the stomach. *Virchows Arch.* 440:311-317.

145. Silva, F., F.Carvalho, A.Peixoto, M.Seixas, R.Almeida, F.Carneiro, P.Mesquita, C.Figueiredo, C.Nogueira, D.M.Swallow, A.Amorim, and L.David. 2001. MUC1 gene polymorphism in the gastric carcinogenesis pathway. *Eur. J. Hum. Genet.* 9:548-552.

146. Silva, F., F.Carvalho, A.Peixoto, A.Teixeira, R.Almeida, C.Reis, L.E.Bravo, L.Realpe, P.Correa, and L.David. 2003. MUC1 polymorphism confers increased risk for intestinal metaplasia in a Colombian population with chronic gastritis. *Eur. J. Hum. Genet.* 11:380-384.

147. Silva, S., M.I.Filipe, and A.Pinho. 1990. Variants of intestinal metaplasia in the evolution of chronic atrophic gastritis and gastric ulcer. A follow up study. *Gut* 31:1097-1104.

148. Spicer, A.P., T.Duhig, B.S.Chilton, and S.J.Gendler. 1995. Analysis of mammalian MUC1 genes reveals potential functionally important domains. *Mamm. Genome* 6:885-888.

149. Spicer, A.P., G.Parry, S.Patton, and S.J.Gendler. 1991. Molecular cloning and analysis of the mouse homologue of the tumor-associated mucin, MUC1, reveals conservation of potential O-glycosylation sites, transmembrane, and cytoplasmic domains and a loss of minisatellite-like polymorphism. *J. Biol. Chem.* 266:15099-15109.

150. Spratt, B.G. 2003. Microbiology. Stomachs out of Africa. *Science* 299:1528-1529.

151. Stephens, M., N.J.Smith, and P.Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978-989.

152. Stone, D.L., N.Tayebi, E.Orvisky, B.Stubblefield, V.Madike, and E.Sidransky. 2000. Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. *Hum. Mutat.* 15:181-188.

153. Stumpf, M.P. and D.B.Goldstein. 2003. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.* 13:1-8.

154. Suerbaum, S., C.Josenhans, and A.Labigne. 1993. Cloning and genetic characterization of the Helicobacter pylori and Helicobacter mustelae flaB flagellin genes and construction of H. pylori flaA- and flaB-negative mutants by electroporation-mediated allelic exchange. *J. Bacteriol.* 175:3278-3288.

155. Swallow, D.M., S.Gendler, B.Griffiths, G.Corney, J.Taylor-Papadimitriou, and M.E.Bramwell. 1987a. The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature* 328:82-84.

156. Swallow, D.M., S.Gendler, B.Griffiths, A.Kearney, S.Povey, D.Sheer, R.W.Palmer, and J.Taylor-Papadimitriou. 1987b. The hypervariable gene locus PUM, which codes for the tumour associated epithelial mucins, is located on chromosome 1, within the region 1q21-24. *Ann. Hum. Genet.* 51 (Pt 4):289-294.

157. Swallow, D.M., B.Griffiths, M.Bramwell, G.Wiseman, and J.Burchell. 1986. Detection of the urinary 'PUM' polymorphism by the tumour-binding monoclonal antibodies Ca1, Ca2, Ca3, HMFG1, and HMFG2. *Dis. Markers* 4:247-254.

158. Tabor, H.K., N.J.Risch, and R.M.Myers. 2002. Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* 3:391-397.

159. Taylor-Papadimitriou, J., J.A.Peterson, J.Arklie, J.Burchell, R.L.Ceriani, and W.F.Bodmer. 1981. Monoclonal antibodies to epithelium-specific components of the human milk fat globule membrane: production and reaction with cells in culture. *Int. J. Cancer* 28:17-21.

160. Teixeira, A., L.David, C.A.Reis, J.Costa, and M.Sobrinho-Simoes. 2002. Expression of mucins (MUC1, MUC2, MUC5AC, and MUC6) and type 1 Lewis antigens in cases with and without Helicobacter pylori colonization in metaplastic glands of the human stomach. *J. Pathol.* 197:37-43.

161. Tishkoff, S.A., A.J.Pakstis, M.Stoneking, J.R.Kidd, G.Destro-Bisol, A.Sanjantila, R.B.Lu, A.S.Deinard, G.Sirugo, T.Jenkins, K.K.Kidd, and A.G.Clark. 2000. Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. *Am. J. Hum. Genet.* 67:901-925.

162. Tredaniel, J., P.Boffetta, E.Buiatti, R.Saracci, and A.Hirsch. 1997. Tobacco smoking and gastric cancer: review and meta-analysis. *Int. J. Cancer* 72:565-573.

163. Vinall, L.E., A.S.Hill, P.Pigny, W.S.Pratt, N.Toribara, J.R.Gum, Y.S.Kim, N.Porchet, J.P.Aubert, and D.M.Swallow. 1998. Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Hum. Genet.* 102:357-366.

164. Vinall, L.E., M.King, M.Novelli, C.A.Green, G.Daniels, J.Hilkens, M.Sarner, and D.M.Swallow. 2002. Altered expression and allelic association of the hypervariable membrane mucin MUC1 in Helicobacter pylori gastritis. *Gastroenterology* 123:41-49.

165. Vyse, A.J., N.J.Gay, L.M.Hesketh, N.J.Andrews, B.Marshall, H.I.Thomas, P.Morgan-Capner, and E.Miller. 2002. The burden of Helicobacter pylori infection in England and Wales. *Epidemiol. Infect.* 128:411-417.

166. Wang, H., E.P.Lillehoj, and K.C.Kim. 2003. Identification of four sites of stimulated tyrosine phosphorylation in the MUC1 cytoplasmic tail. *Biochem. Biophys. Res. Commun.* 310:341-346.

167. Weale, M.E., C.Depondt, S.J.Macdonald, A.Smith, P.S.Lai, S.D.Shorvon, N.W.Wood, and D.B.Goldstein. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* 73:551-565.

168. Weeks, D.L., S.Eskandari, D.R.Scott, and G.Sachs. 2000. A H+-gated urea channel: the link between Helicobacter pylori urease and gastric colonization. *Science* 287:482-485.

169. Wen, Y., T.C.Caffrey, M.J.Wheelock, K.R.Johnson, and M.A.Hollingsworth. 2003. Nuclear association of the cytoplasmic tail of MUC1 and beta-catenin. *J. Biol. Chem.* 278:38029-38039.

170. Wesseling, J., S.W.van der Valk, and J.Hilkens. 1996. A mechanism for inhibition of E-cadherin-mediated cell-cell adhesion by the membrane-associated mucin episialin/MUC1. *Mol. Biol. Cell* 7:565-577.

171. Winfield, S.L., N.Tayebi, B.M.Martin, E.I.Ginns, and E.Sidransky. 1997. Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. *Genome Res.* 7:1020-1026.

172. Wreschner, D.H., M.Hareuveni, I.Tsarfaty, N.Smorodinsky, J.Horev, J.Zaretsky, P.Kotkes, M.Weiss, R.Lathe, A.Dion, and . 1990. Human epithelial tumor antigen cDNA sequences. Differential splicing may generate multiple protein forms. *Eur. J. Biochem.* 189:463-473.

173. Wreschner, D.H., M.A.McGuckin, S.J.Williams, A.Baruch, M.Yoeli, R.Ziv, L.Okun, J.Zaretsky, N.Smorodinsky, I.Keydar, P.Neophytou, M.Stacey,

H.H.Lin, and S.Gordon. 2002. Generation of ligand-receptor alliances by "SEA" module-mediated cleavage of membrane-associated mucin proteins. *Protein Sci.* 11:698-706.

174. Xiang, Z., S.Censini, P.F.Bayeli, J.L.Telford, N.Figura, R.Rappuoli, and A.Covacci. 1995. Analysis of expression of CagA and VacA virulence factors in 43 strains of Helicobacter pylori reveals that clinical isolates can be divided into two major types and that CagA is not necessary for expression of the vacuolating cytotoxin. *Infect. Immun.* 63:94-98.

175. Yamamoto, M., A.Bharti, Y.Li, and D.Kufe. 1997. Interaction of the DF3/MUC1 breast carcinoma-associated antigen and beta-catenin in cell adhesion. *J. Biol. Chem.* 272:12492-12494.

176. Yonezawa, S., J.C.Byrd, R.Dahiya, J.J.Ho, J.R.Gum, B.Griffiths, D.M.Swallow, and Y.S.Kim. 1991. Differential mucin gene expression in human pancreatic and colon cancer cells. *Biochem. J.* 276 (Pt 3):599-605.

177. Zaretsky, J.Z., R.Sarid, Y.Aylon, L.A.Mittelman, D.H.Wreschner, and I.Keydar. 1999. Analysis of the promoter of the MUC1 gene overexpressed in breast cancer. *FEBS Lett.* 461:189-195.

178. Zaridze, D., E.Borisova, D.Maximovitch, and V.Chkhikvadze. 2000. Alcohol consumption, smoking and risk of gastric cancer: case-control study from Moscow, Russia. *Cancer Causes Control* 11:363-371.

179. Zhang, J., W.L.Rowe, A.G.Clark, and K.H.Buetow. 2003. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am. J. Hum. Genet.* 73:1073-1081.

180. Zrihan-Licht, S., A.Baruch, O.Elroy-Stein, I.Keydar, and D.H.Wreschner. 1994a. Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins. Cytokine receptor-like molecules. *FEBS Lett.* 356:130-136.

181. Zrihan-Licht, S., H.L.Vos, A.Baruch, O.Elroy-Stein, D.Sagiv, I.Keydar, J.Hilkens, and D.H.Wreschner. 1994b. Characterization and molecular cloning of a novel MUC1 protein, devoid of tandem repeats, expressed in human breast cancer tissue. *Eur. J. Biochem.* 224:787-795.

Hum Genet (2003) 113 : 473–479
DOI 10.1007/s00439-003-1011-8

ORIGINAL INVESTIGATION

Joanna C. Fowler · Ana S. Teixeira · Lynne E. Vinall
Dallas M. Swallow

Hypervariability of the membrane-associated mucin and cancer marker MUC1

Received: 6 June 2003 / Accepted: 25 July 2003 / Published online: 27 August 2003