[Frontiers in Bioscience 6, d1207-1215, October 1, 2001]

POLYMORPHISM OF THE HUMAN *MUC* GENES

Joanna Fowler, Lynne Vinall, Dallas Swallow

Galton Laboratory, Department of Biology, Wolfson House, 4 Stephenson Way, London NW1 2HE

TABLE OF CONTENTS

1. ABSTRACT

Mucins encoded by the *MUC* genes share the common feature of having an extensive tandem repeat region that encompasses a large proportion of the coding sequence. In many of the genes this tandem repeat region shows a great deal of allelic length variation and recently studies have demonstrated person to person variation in pattern of nucleotide or amino-acid changes in the repeat units. The length and sequence variability will be discussed in this review, as will its role in disease susceptibility.

2. INTRODUCTION

Fourteen human mucin genes have so far been reported that encode mucin-like glycoproteins expressed in epithelia and have been given the symbol *MUC* (www.hugo-international.org/hugo/). These genes do not fall into one simple family. Four located in a cluster on chromosome 11p15.5 encode classical secreted gel forming mucins and are clearly related (1). A further small secreted mucin, is encoded by a single gene, *MUC7*, on chromosome 4 (2). Several others are membrane bound, some falling in a cluster on chromosome 7 (3). The *MUC* genes do however share an important common feature. They all contain at least one extended exonic region of repetitive sequence, which in most cases comprises 50% or more of the polypeptide. These domains generally contain tandemly repeated coding sequence that show, in most mucins, length polymorphism due to Variation in the Number of Tandem Repeats (VNTR). Tandem repeats also occur in the introns of the mucin genes but probably no more frequently than elsewhere in the genome.

Repetitive sequences are widely dispersed in the genome but are usually non-coding. These range from microsatellites with simple 2-4 base pair repeats to alpha satellites that are made up of monomers of 171 base pairs, the latter forming blocks of repetitive DNA often ranging from 100kb to several megabases in length. The best-characterised repetitive sequences in the genome are that of the so-called minisatellites. These sequences consist of long expanses of simple repeats usually 30 base pairs in length that show high levels of heterozygosity due to VNTR polymorphism (4). This is thought to have arisen due to high levels of replication slippage, recombination and gene conversion. These minisatellite regions tend to be clustered towards the telomeric end of the

Table 1. Chromosomal localisation and details of the different repeat arrays of the MUC genes

	Chromosome location	Length of tandem repeat unit (amino acids)	Tandem repeat sequence	Tandem repeat region length variation
MUC1	1q21	20	PDTRPAPGSTAPPAHGVSTA	2.8-8kb
MUC2	11p15.5	23	PTTTPITTTTTVTPTPTPTGTQT	3.3-11.4kb
MUC3A and B	7q22	17 (in both genes)	HSTPSFTSSITTTTETTS	7-15kb,20-50kb
MUC4	3q29	16	TSSASTGHATPLPVTD	6.5-27kb
MUC5AC	11p15.5	8 (interrupted)	TTSTTSAP	6.6-7.4kb
MUC5B	11p15.5	29 (interrupted)	ATGSTATPSSTPGTTHTPPVLTTTATTPT	16kb
MUC6	11p15.5	169	SPFSSTGPMTATSFQTTTTYPTPSHPQTTLPTH VPPFSTSLVTPSTGTVITPTHAQMATSASIHST PTGTIPPPTTLKATGSTHTAPPMTPTTSGYSQA HSSTSTAAKTSTSLHSHTSSTHHPEVTPTSTTT ITPNPTSTGTSTPVAHTTSATSSRLPTPFTTHSP PTGS	8-13.5kb
MUC7	4	22	TTAAPPTPSATTPAPPSSSAPG	5/6/8 repeats
MUC8	12q24.3	41 base pairs	TSCPRPLQEGTRV or TSCPRPLQEGTPGSRAAHALSRRGHRVHELPT SSPGGDTGF	?
MUC9	1p13	15	VGHQSVTPGEKTLTS	4 alleles
MUC11	7q22	28	SGLSEESTTSHSSPGSTHTTLSPASTTT	?
MUC12	7q22	28	SGLSQESTTFHSSPGSTHTTLSPASTTT	?

chromosomes (5). Most of these sequences are non-coding and the repeat units are not multiples of three.

Coding sequence tandem repeat variation was first shown for *MUC1*(6-8) and was demonstrated by the observation of the same length variation using different restriction enzymes. The *MUC1* length polymorphism could be seen at the DNA, mRNA and protein levels. It then became clear that this kind of length variation was a general feature of the mucin genes. *MUC1*, *MUC2*, *MUC3A*, *MUC3B*, *MUC4*, *MUC5AC* and *MUC6* are all highly polymorphic. In contrast *MUC5B* shows very little length variation, *MUC7* only shows two common alleles and *MUC9* (recently renamed OVGP1) shows 4 different length alleles (9, 10). The repeat units vary in size from 8 amino acid residues in *MUC5AC* to 169 amino acid residues in *MUC6* (Table 1).

Several other genes, as well as the mucins, also contain coding tandem repeat domains. For example the dopamine D4 receptor gene (DRD4) located on chromosome 11p15.5. This receptor shows length variation in the third cytoplasmic loop of the protein, relating to an imperfect 48 base pair repeat in exon 3 (11). The repeats can vary in number between 2 to 10. Different alleles show different drug binding affinities, which may imply potential differences in the efficacy of drug treatment for example clozapine for schizophrenia (12). There is evidence to suggest that in some people the variation in this gene may affect the traits of novelty seeking and alcoholism (13,14). In addition to the number of repeats, the variation in the sequence of the repeat region of the DRD4 receptor has also been studied in detail (11).

The involucrin gene and the salivary proline rich genes are further examples of a coding sequence containing repetitive DNA (15,16).

Also in this category is the apo (a) gene which produces a protein that is major component of lipoprotein a (Lp(a)). This gene contains 10 different plasminogen-like Kringle IV units. One of these units (Kringle IV-2) is polymorphic and can range in number from 2 to 41. The tandem repeat array is unusual, since intronic sequence as well as exonic sequence forms a major part of each tandem repeat unit. The number of Kringle IV-2 repeats is inversely related to the Lp(a) plasma concentration (17). Other examples are the androgen receptor and the HD gene, which can cause spinal and bulbar atrophy and Huntingtons disease respectively (18,19). These genes contain repetitive sequences but unlike the mucins these are simple tri-nucleotide repeats.

3. MUCIN POLYMORPHISM

3.1. Membrane bound mucins (*MUC1, MUC3A, MUC3B, MUC4, MUC11, MUC12*)

The mucin glycoproteins encoded by these genes share the property of having a C-terminal membrane-spanning region and are expressed on the surface of epithelial cells.

3.1.1. *MUC1*

MUC1 protein is widely distributed in normal tissues and is present in many bodily fluids (e.g. urine) (7). It is expressed by most epithelial cells and also in some other cells e.g. T cells and fibroblasts (20,21; Swallow DM unpublished). In many early studies MUC1 was identified as a cancer antigen to which many monoclonal antibodies had been raised (22).

Figure 1 shows a diagram of the *MUC1* gene. The gene structure is typical of all of the mucin genes in that there is a large central exon containing a tandem repeat region. The tandem repeats are 60 base pairs in length and have a high guanine/cytosine

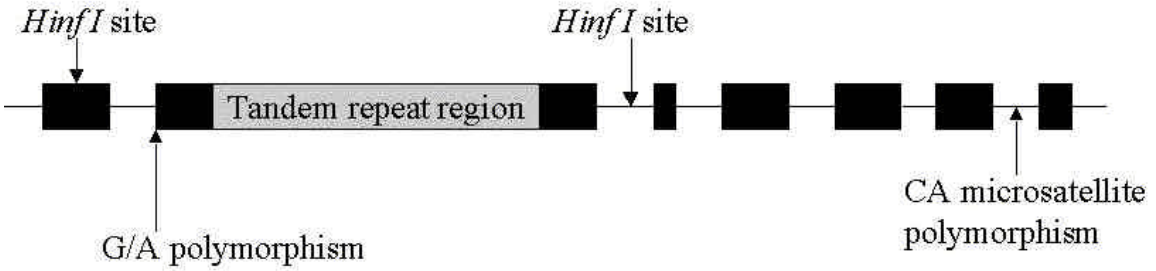

Figure 1. Structure of the *MUC1* gene showing the position of the tandem repeat array and the two flanking polymorphisms. The positions of the *HinfI* sites used for the VNTR analysis are also shown.

content (approximately 80 %) and encode a 20 amino acid repeat rich in serine, threonine and proline. Within each repeat unit there is a highly immunogenic region to which most of the MUC1 antibodies are directed. The antibody binding site has in many cases been finely mapped and almost always overlaps the amino acid sequence PDTR (proline, aspartic acid, threonine, arginine).

The VNTR polymorphism of *MUC1* can be detected in genomic DNA using a wide range of restriction enzymes. Arrows indicate the position of the *HinfI* restriction sites used in our laboratory to study the length polymorphism. The allelic band sizes with *HinfI* range in size from 2.8-8kb and show a bimodal distribution with modal sizes of 3.75kb and 6.75kb (23,24) (Figure 2). This same polymorphism is detectable at the RNA level on Northern blots (6) and on SDS gels using the monoclonal antibodies directed against the TR region, or lectins (25). Peanut lectin25 and the monoclonal antibody Ca1 are the best tools for showing the polymorphism (8).

A number of other polymorphisms have been identified. There is a guanine to adenine transition, 5' to the tandem repeat region within the second exon at nucleotide position 3506 (Gen bank ref. M6110) (Figure 1). This transition is associated with alternative splicing of the mRNA (26). Transcripts with an adenine at position 3506 are spliced to give a product that is 27 nucleotides longer than the one produced if there is a guanine at position 3506. This leads to variation in the proteolytic cleavage of the signal peptide giving either a N-terminal serine for the A allele and a 10 amino acid longer sequence with an alanine for the G allele (27). A CA microsatellite has been found in the sixth intron (28) and alleles with 11, 12 and 13 CA repeats have so far been identified. Analysis of the CEPH family pedigrees has given haplotype information and shown that there is a high degree of linkage disequilibrium across this part of the *MUC1* gene (28). This means that individuals with an adenine at nucleotide position 3506 tend to have smaller VNTR alleles and larger CA repeat alleles but individuals with a guanine at position 3506 have longer tandem repeat alleles and smaller CA repeat alleles. The degree of association along the *MUC1* gene indicates that unequal reciprocal recombination was not the major method by which the length polymorphism of the VNTR region evolved (28).

3.1.2. *MUC3A* and *MUC3B*

Much of the early work on *MUC*3 considered it as a single gene though it is now very clear that mRNA transcripts can be detected that are encoded by two separate highly homologous loci in close proximity, *MUC3A* and *MUC3B* (3,29). Each gene contains a major VNTR domain containing variable numbers of a 17 amino acid motif. The two VNTR domains share the same consensus sequence although in *MUC3B* there is apparently a more frequent proline at amino acid position 11 in the repeat sequence. In the initial reports in which a tandem repeat probe was used to detect *MUC*3 it was thought to occur mainly in the intestine (30). However it is also expressed in the gall bladder (31,32) and in hepatocytes (33,34). A recent report suggests that *MUC3A* is expressed in several tissues while *MUC3B* in contrast was shown to be restricted to intestine (29).

Southern blot analysis using the restriction enzyme PvuII and the 51bp tandem repeat probe shows two polymorphic VNTR domains. One of the regions varies from 20-48.5kb with the most frequent allele being 24kb (0.67 heterozygosity in the U.K. population). The distribution appears to be multimodal. The other region varies in size from 7-15kb with a unimodal distribution and a peak at approximately 12kb (0.51 heterozygosity in the U.K. population) (23). It is not currently known whether the larger or smaller set of bands relates to *MUC3A* or *MUC3B*.

MUC3A also contains a domain of 1kb tandem repeats (35) though it is not known if this domain is polymorphic. A large number of SNPs (Single Nucleotide Polymorphisms) have also been reported several of which cause amino acid substitutions (3,29).

3.1.3. *MUC4*

MUC4 is expressed in the trachea, bronchus, colon and a few other tissues (20). *MUC4* shows more allelic diversity than the other mucin genes with a very large uninterrupted tandem repeat region with repeat units of 48bp which can be detected using a double digest with *PstI* and *EcoRI* (36) or a single

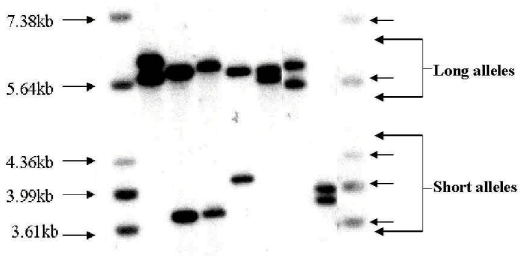

Figure 2 Southern blot of *Hinf1* digested genomic DNA from seven different individuals probed with a *MUC1* tandem repeat probe (Pum24P). End lanes contain the Raoul marker (Oncor Appligene) run as a size standard, sizes as indicated.

digest with *PvuII* (23). The distribution is bimodal with length variation of between 6.5kb and 27kb when digested with *PvuII* (23). The tandem repeat region shows a heterozygosity of 0.78 in the U.K. population. There are also three minisatellite polymorphisms in introns. The first is a 15 base pair tandem repeat in intron 3 varying in length from 1.5-6.5kb (36), the second a 26-32bp imperfect TR in intron 4 and the third a 32bp repeat in intron 5 (37).

3.1.4. *MUC11* and *MUC12*

These two genes have only recently been reported and are found on chromosome 7 in close proximity to *MUC3A* and *MUC3B* (38) suggesting that there may also be a cluster of genes on 7q22. The *MUC11* and *MUC12* clones both contain a TR which shows a high level of homology, and one has associated C-terminal sequences which show homology to the *MUC*3 membrane spanning domain. In fact there is little formal evidence that these clones come from different genes the only indication so far being that they appear to be differently expressed in different tissues. So far no variation in the tandem repeat array has been reported.

3.2. Gel forming mucins (*MUC2, MUC5AC, MUC5B, MUC6*)

The gel forming mucins are located in a cluster on chromosome 11p15.51. These proteins have cysteine rich regions that appear to be involved in disulphide crosslinking. It is the polymerisation of these proteins that forms the mucus network.

3.2.1. *MUC2*

The *MUC2* gene contains two tandem repeat domains, which show no homology to each other. The main repeat array contains 69 base pair tandem repeats. This region shows length variation with a bimodal distribution in some populations. In the U.K. population the majority of individuals have alleles between 6.5-8kb (when restriction digested with *HinfI*) (23). The mean length of the group of shorter allele is around 3.5-4kb. The heterozygosity of this tandem repeat region in the U.K. population is 0.59.

5' to the main tandem repeat region there is a smaller repetitive region with poorly conserved 48 base pair repeats which is 385 amino acids long and does not show common length variation (23). The amino acid repeats vary in size from 7-40 amino acids but the modal size is 16 amino acids (48 base pairs). The first repeats (1-9) at the 5' end of the tandem repeat show high homology with the homology decreasing as one moves 5' to 3'. It seems plausible that this tandem repeat region arose by a series of duplications 5' to 3' accounting for the high homology at the 5' end (39).

Within intron 6 of *MUC2* there is a polymorphic minisatellite (D11S150) (Swallow DM, Pratt WS, Aubert JP and Gum JR unpublished). The repeat array varies in length between 0.9 and 8kb (40). The individual repeats are either 33 or 34bp long.

3.2.2. *MUC5B*

The *MUC5B* gene has a particularly large central exon (41) which contains the tandem repeats but is complex in structure. It contains a mixture of domains, which include 7 copies of a cysteine rich sequence, interspersed with five domains of 87bp tandem repeats as well as other serine and threonine rich sequences. Much of the exon is made up of 4 super-repeats. No evidence has been found for person to person length variation of this region (23).

MUC5B does however show a VNTR polymorphism in intron G where there is a region containing a 59bp repeating unit (42). Each repeat contains an Sp1 binding site. So far 4 alleles have been identified containing either 3,5,7 or 8 repeats.

3.2.3. *MUC5AC*

MUC5AC has a similar tandem repeat structure to *MUC5B* a 24bp tandemly repeated sequence and interspersed cysteine rich domains (43). Restriction digestion of the *MUC5AC* gene with several different restriction enzymes shows multiple alleles. *HinfI* digestion yields two classes of alleles: a and b. However digestion with *PstI* gives 4 distinct alleles (44). The nature of the polymorphism is not yet known though correspondence of the patterns with different restriction enzymes suggests at least some length variation. It could however represent the presence of absence of a whole 'super-repeat unit' and perhaps also polymorphism of restriction sites in the cysteine rich domains. Using *HinfI* digests the frequency of a and b alleles in the U.K. population is 0.79 and 0.21 respectively (23).

3.2.4. *MUC6*

This gene has a tandem repeat domain made up of very large tandem repeat units (169 amino acids, 507bp). As with all mucin proteins it is rich in serine, threonine and proline (45). Restriction digestion with *HindIII* and *EcoRI* yields a large band (>>30kb) too big to be able to detect different sized alleles accurately. *PvuII* detects the length polymorphism clearly and shows the person to person variation. Eleven or more

distinct alleles have been identified giving a unimodal distribution with a peak at 10kb. Alleles range in size from 8-13.5kb. 70% heterozygosity was observed in the unrelated (European) chromosomes from the CEPH families (23).

MUC6, *MUC2*, *MUC5AC* and *MUC5B* are in close proximity on chromosome 11 in band p15.5 within 400kb (1). However this is a very recombination rich region1and there is little evidence of any association between any of the polymorphisms in these genes (1,44).

3.3. Other mucins

3.3.1. *MUC7*

The *MUC7* gene has a relatively small coding sequence of 377 amino acids (2). It does not apparently undergo disulphide crosslinking. As with the other mucin tandem repeat regions the 23 amino acid *MUC7* central tandem repeat array is rich in serine, threonine and proline. Two common alleles have been identified with 5 or 6 repeats. The frequency of the 5 repeat and 6 repeat alleles is in the UK population is approximately 0.1 and 0.9 respectively (46). A rare 8 repeat allele has also been identified (1 heterozygote from 202 individuals) (46).

3.3.2. *MUC8*

Little is known about *MUC8*, a gene which has been reported to map to chromosome 12 (47). It is said to contain an imperfect 41 base pair tandem repeat region, however unlike the other mucin genes the repeat region comprises of two repeat units of different lengths (48) which causes a shift in reading frame.

3.3.3. *MUC9* (*OVGP1*)

MUC9 encodes a protein that is more commonly referred to as oviductin and has been mapped to chromosome 1 (10). *MUC9* is a secreted glycoprotein that is expressed solely by the secretory epithelial cells of the oviduct. Four length variants have been identified.

3.3.4. *MUC13*

MUC13 is the most recent human *MUC* gene to be characterised and shows the same domain organisation as *MUC*3 and *MUC4*. It is a small mucin with 10 degenerate tandem repeats rich in the amino acids serine and threonine (49) and may well not show VNTR variation (see section 5 below).

3.3.5. Other *MUC* gene symbols

The gene symbol *Muc10* has been used in the mouse Mus musculus (Gen bank reference: 20005630 and 6678961) and thus has been reserved in humans but a human homologue has not been found. Several other *MUC* gene symbols have been used but have now been renamed (e.g *MUC18*, *MUC24*) and yet others have been reserved (*MUC14*-*MUC16*), but no information is at present available for these.

4. POLYMORPHISM AND DISEASE ASSOCIATION

Mucins are expressed at the surface of epithelia and are also secreted into mucus. They play a role in protecting and lubricating epithelial surfaces and are thought to be involved in cell-cell interactions and signalling. Variation in the length of the coding region of most of the *MUC* genes predicts substantial interallelic differences in the length of the mucin glycoprotein. Indeed this can be seen directly from the proteins MUC1, MUC2 and MUC7 (8,46,50). The variation will lead to quantitative differences in the number of carbohydrate side chains and in the case of the gel forming mucins the distance between the cross links, which will effect the biophysical properties of the gel. In the membrane bound mucins this variation will alter the distance that the mucin protein protrudes into the lumen. Thus it seems likely that these differences are not functionally silent and these polymorphisms may play a role in disease susceptibility.

Changes in mucus are frequent in inflammatory diseases of the epithelia. For example high levels of secretion of the mucin proteins is a common factor in asthma and chronic bronchitis, Crohns disease and cystic fibrosis. In contrast, a thin mucosal layer is detected in ulcerative colitis (51).

Several *MUC* gene specific associations have so far been reported. For example individuals with gastric cancer have been shown to have a higher proportion of short *MUC1* alleles compared to a control population (24). Work from our laboratory has recently shown that individuals with atopy but not asthma have a significantly different distribution such that the group with atopy and no asthma have a shorter median *MUC2* allele length (52). It has also been shown that the 5 repeat allele in the *MUC7* gene is significantly rarer in atopic asthmatic individuals than individuals with atopy but no asthma (46). Rare alleles of the *MUC*3 with the shorter tandem repeat bands have been shown to be associated with ulcerative colitis (53), though the possible mechanism for this was not obvious. Recently an association has also been shown, of a non-synonymous polymorphism at nucleotide position 2557 in *MUC3A*, encoding a tyrosine in the cytoplasmic domain with a predisposition to familial Crohn's disease (29).

5. VARIATION IN THE SEQUENCE OF THE TANDEM REPEATS

The tandem repeat region of the mucin proteins is very heavily glycosylated and alterations in the pattern of glycosylation have frequently been reported in inflammatory disease and cancer (54).

In most of the mucin genes the individual repeats in the repeat array differ in sequence. These differences can potentially alter the extended structure

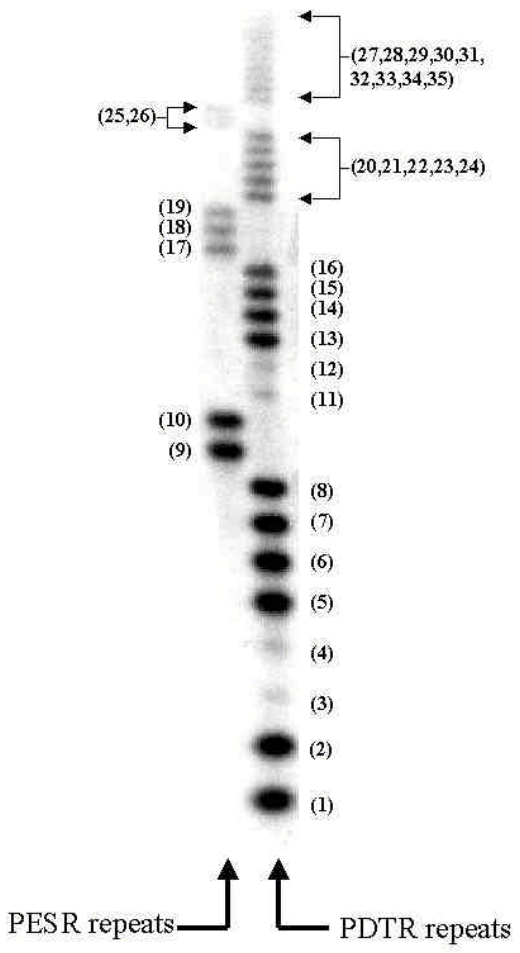

Figure 3. Southern blot, probed with Pum24P, showing a typical forward MVR map. The two lanes represent two PCR reactions using primers specific either for the sequence PDTR or PESR. Numbers in parenthesis indicate repeat number from the 5' end of the array. Repeat numbers 3, 4, 11 and 12 are designated null repeats since they amplify with neither of the primers due to the presence of a further nucleotide change 'under' the primer.

of the apoprotein and the pattern of glycosylation. Until recently, the differences in sequence of the tandem repeats had not been fully investigated, either across the tandem repeat array or in different people. Work in our laboratory sequencing the tandem repeat region of *MUC7* confirmed repeat specific differences but little person to person variation was found. Indeed only one variant was found; an individual with a 5 repeat allele was shown to have an entire duplication of the first two repeats and a novel 5th repeat (TR12127) (46). Evidence from clones and restriction digest of genomic DNA using *TaqI* have shown that the repeats of *MUC2* vary in sequence across the array and in different individuals (39, 44). Likewise clone sequences show repeat sequence differences across the array for *MUC*3, *MUC4*, *MUC5AC*, *MUC5B*, and *MUC6* (30, 41, 45, 55, 56).

Using the technique of MVR (ministellite variant repeat) mapping (57) we have obtained clear evidence of person to person variation tandem repeat sequence differences in the *MUC1* gene. A typical map can be seen in Figure 3 using primers designed around the nucleotide changes described by Siddiqui and colleagues (58) that agreed with changes found in our own cDNA clone Pum24P. These nucleotide changes alter the amino acids PDTR (which comprises the epitope of many of the monoclonal antibodies) to PESR. Müller and colleagues have confirmed the occurrence of PESR in some tandem repeats at the amino acid level (59). The importance of these amino acid changes is not certain but it is possible that the immunogenicity and/or glycosylation of alleles with many copies of this motif (PESR) is significantly altered. Furthermore there is clear evidence of polymorphism of other amino acid substitutions which may be functionally relevant.

This technique can potentially be applied to the tandem repeat regions of the other *MUC* genes though some of the longer alleles may have to be done in stages. This should help to elucidate person to person functional and /or glycosylation differences in the mucins and to understand the evolutionary origin of the enormous diversity.

6. PERSPECTIVE

Mucins show an enormous diversity of structures in a single individual as well as substantial differences between individuals. Genetically determined polymorphism of the multiple *MUC* genes contributes to this variability, adding to that generated by variation in expression and polymorphism of glycosyl transferases. This high level of diversity probably plays an important role in the defence of epithelial surfaces from microorganisms. It is not understood why some genes e.g. *MUC1*, *MUC2*, *MUC4* and *MUC6* show such high levels of variation while others such as *MUC5B* and *MUC7* show such little length polymorphism. The mutational events presumably result from some kind of non-reciprocal (28) recombination/gene conversion events brought about by misalignment of repeat units. The complicated structure of *MUC5B* with each super-repeat having so many parts may assist in correct alignment and discourage length variation mutations. Poorly conserved repeats as exist in the short repeat region of *MUC2*, seem to have the same effect (39).

It appears that the *MUC7* gene is located in an intron of another gene, and this may place constraints on the length variation of the tandem repeat region. This situation is found for the minisatellite MS205 (D16S309) which is located in the second intron of a calcium channel (60) and shows

rather limited length polymorphism (61).

It will be interesting to determine the extent of person to person tandem repeat sequence variation in the different mucin genes and how this relates to the length variation.

7. REFERENCES

1 Pigny P, Guyonnet-Duperat V, Hill AS, Pratt WS, Galiegue-Zouitina S, Collyn D'Hooge M, Laine A, Van Sueningen I, Degand P, Gum JR, Kim YS, Swallow DM, Aubert JP, Porchet N Human mucin genes assigned to 11p15.5: Identification and organisation of a cluster of genes. *Genomics* 38: 340-352 (1996)

2 Bobek L, Liu J, Sait SN, Shows TB, Bobek Y, Levine MJ Structure and chromsomal localization of the human salivary mucin gene, *MUC7*. *Genomics* 31: 277-282 (1996)

3 Pratt W, Crawley S, Hicks J, Ho J, Nash M, Kim YS, Gum JR, Swallow DM Multiple transcripts of *MUC*3: Evidence for two genes *MUC3A* and *MUC3B*. *Biochem. Biophys. Res. Comm.* 275: 916-923 (2000)

4 Jeffreys A, Wilson V, Thein SL Hypervariable 'minisatellite' regions in human DNA. *Nature* 314: 67-73 (1985)

5 Royle N, Clarkson RE, Wong Z, Jeffreys AJ Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* 3: 352-60 (1988)

6 Hareuveni M, Tsarfarty I, Zaretsky J, Horev J, Zrihan S, Weiss M, Green S, Lathe R, Keydar I, Wreschner DH A transcribed gene, containing a variable number of tandem repeats, codes for a human epithelial tumour antigen. *Eur. J. Biochem.* 189: 475-486 (1990)

7 Karlsson S, Swallow DM, Griffiths B, Corney G, Hopkinson DA, Dawnay A, Cartron JP A genetic polymorphism of a human urinary mucin. *Ann. Hum. Genet* 47: 263-269 (1983)

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

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

10 Lapensée L, Paquette Y, Bleau G Allelic polymorphism and chromosomal localization of the human oviductin gene (*MUC9*). *Fert. & Ster.* 65: 702-708 (1997)

11 Lichter J, Barr C, Kennedy J, Van Tol H, Kidd K, Livak K A hypervariabale segment in the human dopamine receptor D4 (DRD4) gene. *Hum. Mol. Genet* 2: 767-773 (1993)

12 Van Tol H, Wu CM, Guan HC, Ohara K, Bunzow J, Civelli O, Kennedy J, Seeman P, Niznik H, Jovanovic V Multiple dopamine D4 receptor variants in the human population. *Nature* 358: 149-152 (1992)

13 Hill S, Zerra N, Wipprecht G, Locke J, Neiswanger K Personality traits and dopamine receptors (D2 and D4) linkage studies in families of alcoholics. *Am. J. Med. Genet.* 15: 634-641 (1999)

14 Noble E, Ozkaragoz TZ, Rithie TL, Zang X, Belin TR, Sparkes RS D2 and D4 dopamine receptor polymorphisms and personality. *Am. J. Med. Genet.* 8: 257-67 (1998)

15 Simon M, Phillips M, Green H Polymorphism due to variable number of repeats in the human involucrin gene. *Genomics* 9: 576-80 (1991)

16 Azen EA Genetics of salivary protein polymorphisms. *Crit. Rev. Oral. Biol. Med.* 4: 479-85 (1993)

17 Brunner C, Lobentaz E-M, Petho-Schramm A, Ernts A, Kang C, Dieplinger H, Muller H-J, Utermann G The number of identical Kringle repeats in Apolipoprotein (a) affects its proccessing and secretion by HepG2 cells. *J. Biol. Chem.* 271: 32403-32410 (1996)

18 Merry D, Kobayashi Y, Bailey CK, Taye AA, Fischbeck KH Cleavage, aggregation and toxicity of the expanded androgen receptor in spinal and bulbar muscular atrophy. *Hum. Mol. Genet.* 7: 693-701 (1998)

19 Ho L, Carmichael J, Swartz J, Wyttenbach A, Rankin J, Rubinsztein DC The molecular biology of Huntington's disease. *Psychol. Med.* 31: 3-14 (2001)

20 Van Klinken B, Dekker J, Buller HA, Einerhand AW Mucin gene structure and expression: protection vs. adhesion. *Am. J. Physiol.* 269: G613-27 (1995)

21 Agrawal B, Krantz MJ, Parker J, Longenecker BM Expression of *MUC1* mucin on activated human T cells: implications for a role of MUC1 in normal immune regulation. *Cancer Res.* 15: 4079-81 (1998)

22 Taylor-Papadimitriou J, Burchell J, Miles DW, Dalziel M *MUC1* and cancer. *Biochim. Biophys. Acta.* 8: 301-13 (1999)

23 Vinall L, Pratt WS, Swallow DM Detection of mucin polymorphism. *Methods Mol. Biol.* 125: 337-50 (2000)

24 Carvalho F, Seruca R, David L, Amorim A, Seixas M, Bennet E, Clausen H, Sobrinho-Simoes M *MUC1* gene polymorphism and gastric cancer - an epidemiological study. *Glycon. J.* 13: 1-6 (1996)

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

26 Lightenberg M, Gennissen AMC, Vos HL, Hilkens J A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA. *Nuc. Acids Res.* 19: 297 -301 (1990)

27 Ligtenberg M, Vos HL, Gennissen AMC, Hilkens J Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variant with alternative amino termini. *J. Biol. Chem.* 265: 5573-5578 (1990)

28 Pratt W, Islam I, Swallow DM Two additional polymorphisms within the hypervariable *MUC1* gene: association of alleles either side of the VNTR region. *Ann. Hum. Genet.* 60: 21-28 (1996)

29 Kyo K, Muto T, Nagawa H, Lathrop GM, Nakamura Y Associations of distinct variants of the intestinal mucin gene *MUC3A* with ulcerative colitis and Crohn's disease. *J. Hum. Genet.* 46: 5-20 (2001)

30 Gum J, Ho JL, Pratt WS, Hicks JW, Hill AS, Vinall LE, Roberton, Swallow DM, Kim YS *MUC*3 Human intestinal mucin. *J. Biol. Chem.* 272: 26678-26686 (1997)

31 Baeckstrom D, Karlsson N, Hansson GC Purification and characterization of sialyl-Le(a) carrying mucins of human bile; evidence for the presence of *MUC1* and *MUC*3 apoproteins. *J. Biol. Chem.* 269: 14430-14437 (1994)

32 Van Klinken J, Van Dijken TC, Oussen E, Buller HA, Dekker J, Einerhand HWC Molecular cloning of human *MUC*3 cDNA reveals a novel 59 amino acid tandem repeat region. *Biochem. Biophys. Res. Comm.* 238: 143-148 (1997)

33 Vandenhaute B, Buisine MP, Debailleul V, Clement B, Moniaux N, Dieu MC, Degand P, Porchet N, Aubert JP Mucin gene expression in biliary epithelial cells. *J. Hepatol* 27: 1057-66 (1997)

34 Buisine M, Devisme L, Degand P, Dieu MC, Gosselin B, Copin MC, Aubert JP, Porchet N Developmental mucin gene expression in the gastroduodenal tract and accessory digestive glands. II. Duodenum and liver, gallbladder, and pancreas. *J. Histochem. Cytochem.* 48: 1667-1676 (2000)

35 Gum J, Ho JJ, Pratt WS, Hicks JW, Hill AS, Vinall LE, Roberton AM, Swallow DM, Kim YS *MUC*3 human intestinal mucin. Analysis of gene structure, the carboxy terminus and a novel upstream region. *J. Biol. Chem.* 272: 26678-86 (1997)

36 Nollet S, Moniaux N, Maury J, Petitprez D, Degand P, Laine A, Porchet N, Aubert JP Human mucin gene *MUC4*: organization of its 5'-region and polymorphism of its central tandem repeat array. *Biochem. J.* 332: 739-748 (1998)

37 Moniaux N, Nollet S, Porchet N, Degand P, Laine A, Aubert JP Complete sequence of the human mucin *MUC4*: a putative cell membrane-associated mucin. *Biochem J.* 338: 325-333 (1999)

38 Williams S, McGuckin MA, Gotley DC, Eyre HJ, Sutherland GR, Antalis TM Two novel mucin genes down-regulated in colorectal cancer identified by differential display. *Cancer Res.* 59: 4083-4089 (1999)

39 Toribara N, Gum JR, Culhane PJ, Lagace RE, Hicks JW, Peterson GM, Kim YS *MUC*-2 human small intestinal mucin gene structure. *J. Clin. Invest.* 88: 1005-1013 (1991)

40 Brookes A, Hedge PH, Solomon E A highly polymorphic locus on chromosome 11 which has homology to a collagen triple-helix coding sequence. *Nucleic Acids Res.* 17: 1792 (1989)

41 Desseyn J-L, Guyonnet-Dupérat V, Porchet N, Aubert JP, Laine A Human mucin gene *MUC5B*, the 10.7kb large central exon encodes various alternate subdomains resulting in a super-repeat. *J. Biol. Chem.* 272: 3168-3178 (1997)

42 Desseyn J, Rousseau K, Laine A Fifty-nine bp polymorphism in the uncommon intron 36 of the human mucin gene *MUC5B*. *Electrophoresis* 20: 493-6 (1999)

43 Escande F, Buisine MP, Porchet N, Aubert JP Human mucin gene *MUC5AC*: organisation of the amino-terminal and central repetitive region Mucins in health and disease. 6th internation workshop on carcinoma-associated mucins, Cambridge 1999

44 Vinall L, Hill AS, Pigny P, Pratt WS, Toribara N, Gum JR, Kim Y, Porchet N, Aubert JP, Swallow DM Variable number tandem repeat polymorphism of the mucin genes located in the complex 11p15.5. *Hum. Genet* 102: 357-366 (1998)

45 Toribara N, Roberton AM, Ho SB, Kuo WL, Gum E, Hicks JW, Gum JR, Byrd JC, Siddiki B, Kim YS Human gastric mucin. *J. Biol. Chem.* 268: 5879-5885 (1993)

46 Kirkbride H, Bolscher JG, Nazmi K, Vinall L, Nash MW, Moss FM, Mitchell DM, Swallow DM

Genetic polymorpism of *MUC7*: Allele frequencies and association with asthma. *Euro. J. Hum. Genet* 9: 347-354 (2001)

47 Shankar V, Pichan P, Eddt RL, Tonk V, Nowak N, Sait SNJ, Shows TB, Schultz RE, Gotway G, Elkins RC, Gilmore MS, Sachdev GP Chromsomal location of a human mucin gene (*MUC8*) and cloning of the cDNA corresponding to the carboxy terminus. *Am. J. Resp. Cell. Mol. Biol.* 16: 232-241 (1997)

48 Shankar V, Gilmore MS, Elkins RC, Sachdev GP A novel human airway mucin cDNA encodes a protein with unique tandem-repeat organisation. *Biochem. J.* 300: 295-8 (1994)

49 Williams S, Wreschner DH, Tran M, Eyre HJ, Sutherland GR, McGuckin MA *MUC13*- A novel cell surface mucin expressed by epithelial and hemopoietic cells. *J. Biol. Chem.* epub (2001)

50 Herrmann A, Davies JR, Lindell G, Martensson S, Packer NH Swallow DM, Carlstedt I Studies on the "insoluble" glycoprotein complex from human colon. Identification of reduction-insensitive *MUC2* oligomers and C-terminal cleavage. *J. Biol. Chem.* 274: 15828-36 (1999)

51 Pullan R, Thomas GA, Rhodes M, Newcombe RG, Williams GT, Allen A, Rhodes J Thickness of adherent mucus gel on colonic mucosa in humans and its relevance to colitis. *Gut* 35: 353-9 (1994)

52 Vinall L, Fowler J, Jones A, Kirkbride H, de Bolos C, Laine A, Porchet N, Gum J, Kim Y, Moss F, Mitchell D, Swallow DM Polymorphism of human mucin genes in chest disease. Possible significance of *MUC2*. *Am. J. Respir. Cell Mol. Biol.* 23: 678-686 (2000)

53 Kyo K, Parkes M, Takei Y, Nishimori H, Vyas P, Satsangi J, Simmons J, Nagawa H, Baba S, Jewell D, Muto T, Lathrop M, Nakamura Y Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin gene, *MUC*3. *Hum. Mol. Genet.* 8: 307-311 (1999)

54 Brockhausen I Pathways of O-glycan synthesis in cancer cells. *Biochim. Biophys. Acta* 1473: 67-95 (1999)

55 Porchet N, Cong NV, Dufosse J, Audie JP, Guyonnet-Dupérat V, Gross MS, Denis C, Degand P, Bernheim A, Aubert JP Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem. Biophys. Res.Comm.* 175: 414-422 (1991)

56 Guyonnet Duperat, Audie JP, Debailleu V, Laine A, Buisine MP, Zouitina-Galiegue S, Pigny P, Degand P, Aubert JP, Porchet N Characterization of the human mucin gene *MUC5AC*: a consensus cysteine-rich domain for 11p15 mucin genes? *Biochem. J.* 304: 1-9 (1994)

57 Jeffreys A, MacLeod A, Tamaki K, Neil DL, Monckton DG Minisatellite repeat coding as a digital approach to DNA typing. *Nature* 354: 204-209 (1991)

58 Siddiqui J, Abe M, Hayes D, Shani E, Yunis E, Kufe D Isolation and sequencing of a cDNA coding for the human DF3 breast carcinoma antigen. *Proc. Natl. Acad. Sci. USA* 85: 2320-3 (1988)

59 Müller S, Alving K, Peter-Katalinic J, Zachara N, Gooley AA, Hanisch FG High density O-glycosylation on tandem repeat petides from secretory *MUC1* of T47D breast cancer cells. *J. Biol. Chem.* 274: 18165-18172 (1999)

60 Badge R, Yardley J, Jeffreys AJ, Armour AJ Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.* 1: 1239-44 (2000)

61 May C, Jeffreys AJ, Armour JA Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205. *Hum. Mol. Genet.* 5: 1832-33 (1996)

Key Words: Mucin, Polymorphism, Glycoprotein, Review

Send correspondence to: Dallas Swallow, Galton Laboratory, Biology Department, Wolfson, House, 4 Stephenson Way, London, NW1 2HE, Tel: 020 76795040, Fax: 020 73873496, E-mail: dswallow@hgmp.mrc.ac.uk