MUC-2 Human Small Intestinal Mucin Gene Structure

Repeated Arrays and Polymorphism

Neil W. Toribara,*‡ James R. Gum, Jr.,* Patrick J. Culhane,* Robert E. Lagace,* James W. Hicks,* Gloria M. Petersen,§ Young S. Kim*‡

*Gastrointestinal Research Laboratory, Veterans Administration Medical Center, San Francisco, California 94121;
‡Department of Medicine, Gastroenterology Division, University of California, San Francisco, California 94143;
§Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California 90048

Abstract

MUC-2, the first described intestinal mucin gene, has become important as a prototype for secreted mucins in several organ systems. However, little is known about its protein backbone structure and hence its role in diseases such as colon cancer, ulcerative colitis, and cystic fibrosis, which are known to have mucin abnormalities. Studies in this manuscript show that MUC-2 contains two distinct regions with a high degree of internal homology, but the two regions bear no significant homology to each other. Region 1 consists mostly of 48-bp repeats which are interrupted in places by 21–24-bp segments. Several of these interrupting sequences show similarity to each other, creating larger composite repeat units. Region 1 has no length polymorphisms. Region 2 is composed of 69-bp tandem repeats arranged in an uninterrupted array of up to 115 individual units. Southern analysis of genomic DNA samples using TaqI and HinfI reveals both length and sequence polymorphisms which occur within region 2. The sequence polymorphisms have different ethnic distributions, while the length polymorphisms are due to variable numbers of tandem repeats. (*J. Clin. Invest.* 1991. 88:1005–1013.) Key words: glycoprotein • tandem repeats • *O*-glycosylation • colonic mucin • mucus

Introduction

Mucin glycoproteins are extremely large and heavily glycosylated structures that consist primarily of a polypeptide backbone and *O*-linked oligosaccharide chains. Approximately 80% of the mass of mucins are carbohydrates, which gives them the high density, hydrodynamic volume, and viscosity necessary for the formation of mucus gels whose biological functions include maintenance of tissue hydration, lubrication, and cytoprotection against proteases, pH extremes, chemical irritants, and biological agents (1).

Four human mucins have been described and at least partially characterized (2–14). All of these genes have been found to contain tandemly repeated sequences that encode peptides rich in the hydroxy-amino acids threonine and/or serine. These repetitive arrays are thought to be heavily glycosylated and to make up a major portion of the overall structure of mucins. This is certainly the case with MUC-1 which is known to be highly polymorphic due to variable numbers of tandem repeats (VNTRs)[1] in different alleles.

Our laboratory has cloned partial cDNAs that are derived from two different human intestinal mucin genes, MUC-2 and MUC-3 (9, 13). These genes appear to encode very large polymorphic proteins, but precise structural information has been lacking. The MUC-2 cDNAs were the first isolated and are the better characterized of the two. This gene is expressed in colon, small intestine, colonic tumors, bronchus, cervix, gall bladder, and possibly other tissues, suggesting that the MUC-2 gene product is physiologically important in many organ systems (9–12). In addition, mucin abnormalities have been described in a number of gastrointestinal diseases including cystic fibrosis, ulcerative colitis, and colon cancer (15–18). Mucinous colon adenocarcinomas appear to be more biologically aggressive with resulting poorer prognosis than histologically nonmucinous tumors (19). These two histologically different groups of colon carcinomas are also reported to have different patterns of chromosomal abnormalities (20).

Further information pertaining to mucin structure and expression may lead to a better understanding of the pathology of several diseases. This paper gives a detailed analysis of the MUC-2 gene and the differences between its various alleles.

Methods

DNA isolation. Leukocytes were isolated from the peripheral blood of 156 healthy Caucasians and 15 healthy Asians by centrifugation and DNA was extracted by the method of Blin and Stafford (21).

DNA (Southern) blotting. DNA was digested with various restriction enzymes and the fragments separated by electrophoresis through 0.8% agarose (22). Blotting, hybridization, and washing were conducted as described (9) except that the final wash (in 0.1× standard saline citrate, 0.1% SDS) was conducted at 65°C rather than 55°C.

Four different hybridization probes were used. The cDNA clone SMUC 40 was used as a probe for the MUC-2 tandem repeats. For the 5′-region of the gene, a 152-bp BamHI/NcoI fragment of the GMUC clone (described later) was used. Two probes derived from the unique portion of the SMUC 41 cDNA clone were used to examine the 3′-region of the gene. These included a 262-bp ApaI/KpnI fragment and a 204-bp KpnI/EcoRI fragment (9). Probes were labeled with [^{32}P]dCTP by random hexamer priming.

Genomic DNA clone isolation. Human peripheral leukocyte DNA from a homozygous individual was digested to completion with

This work has appeared in abstract form (1990. *Gastroenterology.* 98:315*a*.).

Address correspondence and reprint requests to Dr. Neil W. Toribara, Gastrointestinal Research Laboratory (151M2), Veterans Administration Medical Center, 4150 Clement Street, San Francisco, CA 94121. Dr. Petersen's present address is Department of Epidemiology, School of Hygiene and Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205.

Received for publication 3 October 1990 and in revised form 1 May 1991.

1. *Abbreviations used in this paper:* VNTR, variable numbers of tandem repeats.

BamHI and size fractionated by sucrose density gradient centrifugation (23). The DNA was then precipitated from the fractions and examined by dot blot hybridization with the MUC-2 tandem repeat probe. Fractions giving the highest signal were pooled and used to construct a library in the Lambda-GEM-11 vector system (Promega Corp., Madison, WI). The library was then plated in soft agar at a density of ~ 25,000 plaques/150 mm plate using *Escherichia coli* strain KW251 as a host. Plates were then incubated at 37°C until plaques formed, then overlayed for 2 min with nylon membranes. The membranes were then treated with 0.5 M sodium hydroxide/1.5 M sodium chloride to denature the DNA, followed by neutralization with 1.0 M Tris, pH 7.6/1.5 M sodium chloride. The membranes were then baked in a vacuum oven for 1 h at 80°C and screened by hybridization with the MUC-2 tandem repeat probe under the conditions described above. Plaques giving a positive signal were isolated, plated, and rescreened, with this process being repeated until clonality was finally obtained.

DNA sequencing. Sequencing was done using M13mp18 and M13mp19 vectors as single-stranded templates. The dideoxynucleotide chain termination method was used with a modified T7 DNA polymerase (Sequenase, version 2.0; United States Biochemical Corp., Cleveland, OH) (24).

Thermal amplification. The tandem repeat regions of genomic DNA samples were amplified using low concentrations of DNA (50 ng/reaction) along with a high concentration of primers (25 pmol each/reaction). The oligonucleotide primer flanking the 5′ side of the tandem repeats was 5′-TGCCTCAACTACGAGATCAAC-3′ (+ strand). The primer on the 3′ side was 5′-ATTGGATGTGGTCAACTCAGC-3′ (− strand). Reaction conditions were 67 mM Tris, pH 7.6, 16.6 mM $(NH_4)_2SO_4$, 6.7 mM $MgCl_2$, 10% DMSO, 200 μM dNTPs, and 2.5 U of AmpliTaq (Perkin-Elmer Cetus Corp., Norwalk, CT) per reaction. DNA was denatured for 40 s at 94°C, followed by 1 min of annealing at 55°C and 15 min of extension at 72°C. 14 cycles of amplification were used, because higher cycle numbers increased the possibility of self priming of the amplified product by the 69-bp tandem repeat regions.

Results

MUC-2 polymorphisms. Restriction fragment length polymorphisms of the MUC-2 gene have been noted in this laboratory and others (8, 9). Further characterization of the structural basis for this polymorphism was felt to be necessary in order to better understand the individual differences that exist in this gene. TaqI polymorphisms were examined by blot analysis using genomic DNA isolated from the lymphocytes of 171 individuals. The vast majority of these samples contained either two or three bands in the size range between 1.7 and 2.3 kb (Fig. 1 *A*). All of the samples also contained numerous bands of 0.5 kb and below, as will be discussed later. When only the larger TaqI fragments were considered, the MUC-2 genotype in the population examined consisted primarily of just two alleles which were inherited in a Mendelian codominant fashion (8). These alleles were designated either "A" or "B". Individuals homozygous for A alleles exhibited two bands at 2.3 and 2.1 kb, while those homozygous for the B allele had bands at 2.3 and 1.7 kb (Fig. 1 *A*). Most of the DNA samples examined were from Caucasians living in the United States or Western Europe. 156 such individuals were examined, with 25% of the population homozygous for pattern A, 23% for pattern B, and 42% being combinations of patterns A and B. About 10% of individuals appear to be heterozygotes of A or B and "other" patterns with only one individual having neither pattern A nor B.

The distribution of alleles in Asians appears to be quite different than in Caucasians. 15 Asians were studied of whom 14 were homozygous for pattern A and one was heterozygous for pattern A/B. The difference in pattern distribution between

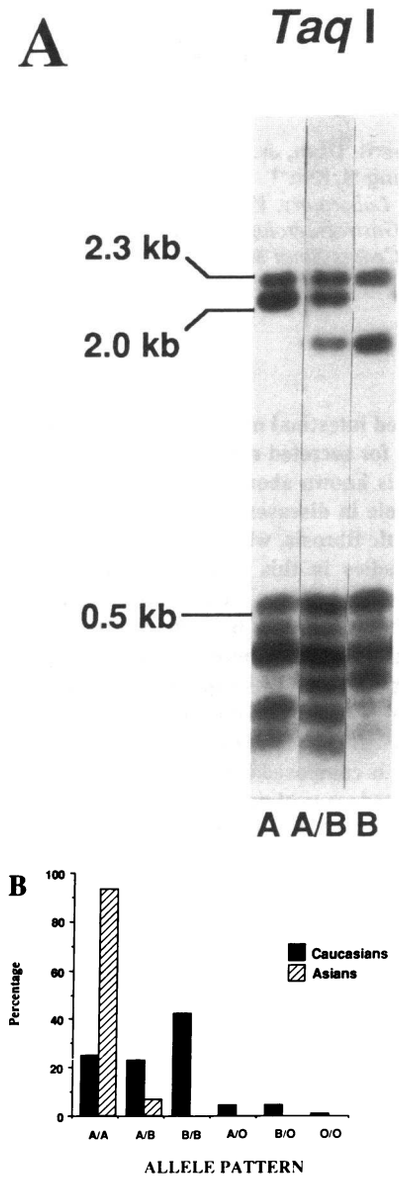

Figure 1. Major MUC-2 allele patterns. (*A*) Restriction fragment length polymorphisms with TaqI using the MUC-2 tandem repeat probe. The two major homozygous patterns are designated *A* and *B*. The middle lane represents the progeny of parents with patterns *A* and *B*. (*B*) Distribution of MUC-2 alleles in Caucasians and Asians. *A* and *B* refer to patterns "A" and "B" and *O* refers to "other" patterns.

Caucasians and Asians is significant at a $P < 0.005$ by Chi square analysis. The allele frequencies in both Caucasians and Asians are in Hardy-Weinberg equilibrium.

Southern analysis using the restriction enzyme HinfI has previously shown the MUC-2 gene to be polymorphic (8, 9). Extending these observations to a large number of genomic DNA samples, we found that the HinfI-digested samples contained either one or two major bands. These polymorphisms were independent of the homozygosity or heterozygosity of the same samples when digested with TaqI (Fig. 2), suggesting different reasons for the polymorphisms seen with these two restriction enzymes. In order to further characterize the MUC-2 gene and its polymorphisms, it was necessary to isolate and sequence a genomic DNA clone.

MUC-2 genomic DNA clone. Because BamHI-digested human genomic DNA contains a single fragment of ~ 15 kb that hybridizes to SMUC 40, this enzyme was used to generate a library from the DNA of an individual homozygous for the A allele (defined using TaqI). Approximately 3×10^6 plaques were screened to obtain a single stable clone, despite the use of

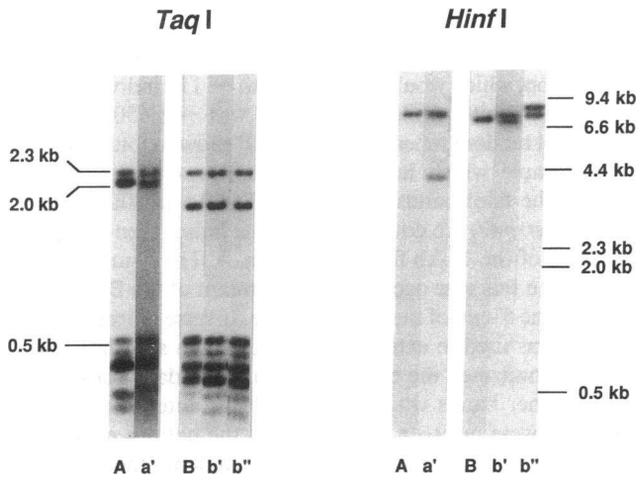

Figure 2. Comparison of TaqI alleles to HinfI alleles. Upper case letters refer to samples which are homozygous on both TaqI and HinfI digests. Lower case letters refer to samples which appeared to be homozygous on TaqI digest, but heterozygous on HinfI digest. Southern analyses performed with MUC-2 tandem repeat probe.

size selected DNA in the construction of the library. Furthermore, portions of this clone proved to be unstable upon attempts at subcloning into plasmid vectors. This appears to be due to the unstable nature of DNA segments containing short tandem repeats in bacterial host strains (9, 13). The size of the insert (designated GMUC) was determined to be 11 kb, i.e., 4 kb shorter than the BamHI fragment from which it was derived. Thus, a sizable portion of this clone was deleted through host strain-mediated recombination, with the deletion occurring within the tandem repeat array.

Structural basis of polymorphisms in MUC-2. The GMUC clone was restriction mapped as shown in Fig. 3. Parts of the clone were then sequenced. The regions immediately inside of both boundaries of the cross-hatched box in Fig. 3 were determined to consist of tandem repeats for at least several hundred nucleotides from either end with the direction of transcription determined to be from left to right. The bulk of the 3,000-bp DNA fragment within the cross-hatched box could not be successfully cloned into M13mp18, M13mp19, pSP6, or pBluescript, making direct sequence determination impossible. However, this region appears to consist entirely of tandem repeats as indicated by the experiment shown in Fig. 4. The restriction enzyme BstEII was used to produce either total or partial restriction digests of the GMUC clone. The resulting fragments were separated by agarose gel electrophoresis and hybridized with the tandem repeat probe following blotting.

Because over 85% of the individual MUC-2 tandem repeats have a BstEII recognition site (GGTNACC) within their most highly conserved region, complete digestion would be expected to break the tandem repeat region into small pieces consisting mainly of individual units of 69 bp, although a small number of fragments would have two or more units if no BstE II site was present. Fig. 4 *A* shows a complete BstEII digest of the 11-kb GMUC insert. The bands at ~ 1,850 and 800 bp represent the fragments produced at the 5′ and 3′ ends of the tandem repeat region, respectively. These fragments contain approximately one-half of a tandem repeat and would be expected to produce a faint signal on hybridization to the tandem repeat probe. The fact that these bands could be detected is significant because it indicates that any unique sequence interposed in the tandem repeat region would also produce a detectable fragment. The fact that no such fragment was observed suggests that the tandem repeat region of the GMUC clone is not interrupted by any unique segment of 500 bp or larger.

Partial digestion, where only a fraction of the total number of sites are actually cleaved, would be expected to cut the tandem repeat region into fragments consisting of multiples of individual tandem repeat units. Because the BstEII sites would be digested randomly, all multiples of tandem repeat units should be represented, creating a ladder effect where the individual rungs consist of 69-bp multiples. Total Lambda-GEM-11 DNA containing the GMUC insert was digested with BstEII (0.5 U/μg of DNA) for various lengths of time. Fig. 4 *B* shows the 2-, 5-, and 10-min time points. Careful examination of the Southern blot gives at least 36 rungs, placing the minimum number of uninterrupted tandemly repeated units at 36. Restriction mapping has shown that HinfI digestion sites flank the tandem repeat region by ~ 300 bp on the 5′ side and 65 bp on the 3′ side. HinfI digestion of the Lambda-GEM-11 DNA followed by Southern analysis with the MUC-2 tandem repeat probe showed a single band whose size was estimated to be

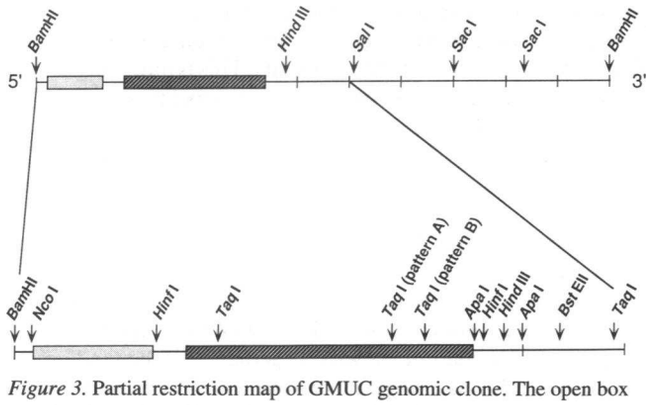

Figure 3. Partial restriction map of GMUC genomic clone. The open box encodes the region of imperfect repeats and the hatched box encodes the tandem repeat region. Both of these regions have a high concentration of proline and potential *O*-glycosylation sites. (▩) Tandem repeats. (▨) Region of imperfectly conserved repeats.

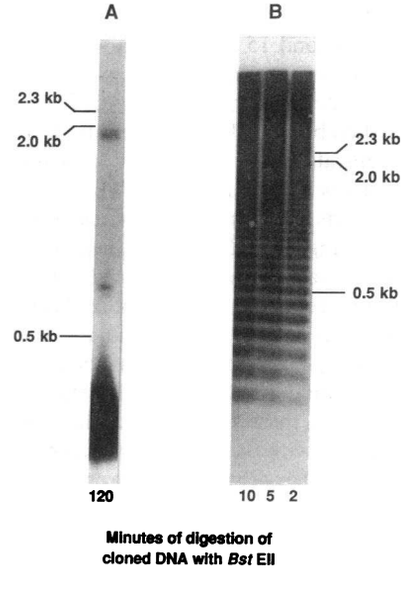

Figure 4. Characterization of the tandem repeat region in the GMUC clone. (*A*) GMUC 11-kb fragment DNA (1 μg) was digested with 20 U of the restriction enzyme BstEII for 2 h. (*B*) Partial digests of Lambda-GEM-11 DNA containing the GMUC insert. Each lane represents 2 μg of DNA digested for the specified time with BstEII at a concentration of 0.5 U/μg of DNA. Blot analysis was conducted using the MUC-2 tandem repeat probe.

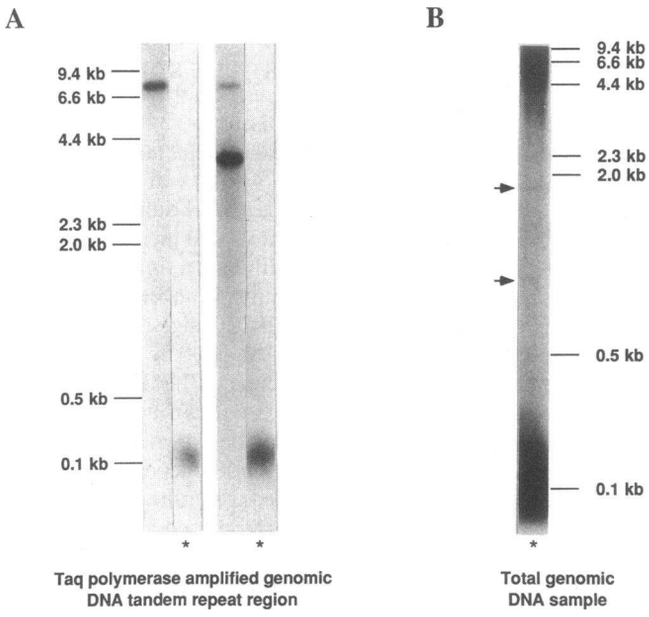

Figure 5. Characterization of the tandem repeat region in genomic DNA. (*A*) Taq polymerase amplification of genomic DNA from a homozygous individual (*left*) and heterozygous individual (*right*) using primers immediately flanking the tandem repeat region of the GMUC clone. Template DNA samples were from the same individuals seen in lanes *A* and *a'* in Fig. 2. The amplified products were divided in half, with half of the sample being digested with 20 U of BstEII for 2 h (*) and the other half being untreated. (*B*) Total genomic DNA (2 μg of sample *A*) digested with 20 U of BstEII for 2 h). Faint bands indicated by arrows are at ~ 1,900 and 800 bp.

3,300 bp (data not shown). Subtracting the flanking unique sequences, the tandem repeat region was estimated to be 2,900 bp in length, corresponding to ~ 42 to 43 tandem repeats. These results strongly suggest that the tandem repeat region of the cloned DNA consists of 42 to 43 uninterrupted individual tandem repeat units.

Attempts to apply the same type of analysis to obtain a direct estimate of the number of individual repeat units in genomic DNA samples were unsuccessful because the lower resolution of bands on the Southern blots made counting of "rungs" on partial digests impossible beyond 15 (data not shown), necessitating the use of more indirect methods.

Thermal amplification of the tandem repeat region of the MUC-2 gene was accomplished by using low concentrations of genomic DNA, high concentrations of primers flanking the tandem repeat region, and a low cycle number (see Methods). Southern analysis of the products was then performed using the MUC-2 tandem repeat probe (Fig. 5 *A*). These amplified DNA samples from homozygous (*left*) and heterozygous (*right*) individuals corresponded well in size to the bands seen on HinfI digests of these same individuals (Fig. 2, lanes *A* and *a'*). The position of the oligonucleotide primers predicts that the thermal amplification products would be about 300 bp shorter than the HinfI fragments. When these thermal amplification products are completely digested with BstEII, the result is small fragments of ~ 50–200 bp. No other discrete bands were noted on overexposure of the autoradiogram, supporting the view that the tandem repeat portion of the HinfI band or bands is uninterrupted by intervening unique sequences. The largest amplified fragment determined in this experiment was ~ 8,000 bp, which would correspond to ~ 115 individual tandem repeat units, while the smallest was ~ 3,650 bp or 51 individual tandem repeat units. Fig. 5 *B* shows a total genomic DNA sample which has been digested to completion with BstEII. The faint bands visible at ~ 1,900 and 800 bp (indicated by *arrows*) are consistent with the bands seen on BstEII digestion of the 11-kb fragment in Fig. 4. The 5'-most BstEII site used in this case occurs 65 bp upstream of the BamHI site defining the 5'-end of the GMUC clone. Inverse thermal amplification was used to extend the cloned region shown in Fig. 3 and has confirmed the existence of this site (data not shown). Again, other bands which might indicate interposed unique sequence were not seen. The smear at the bottom of the lane runs from ~ 50 to 250 bp. Thus, the tandem repeat array in genomic DNA appears to contain between ~ 50 and 115 uninterrupted tandem repeats, depending on the allele.

We next wanted to correlate the bands observed with TaqI digestion with their location within the gene. To accomplish this, probes were constructed which were located 5' (BamHI/NcoI fragment of GMUC) and 3' (ApaI/KpnI fragment of the cDNA clone SMUC 41) in relation to the tandem repeat region. Fig. 6 shows the results of hybridization of these probes to the same TaqI-digested genomic DNA blot. Hybridization to the 5' probe, seen in the left panel of Fig. 6, clearly shows that the 2.3-kb band seen on all of the TaqI digests corresponds to the region 5' to the first TaqI site in the tandem repeat region, and that this fragment does not show any length polymorphism. The middle panel shows the results of hybridization to the tandem repeat probe. Sequence analysis of the published tandem repeat units indicates that the ladder effect seen at the bottom (and in Figs. 1 *A* and 2) is due to the cleavage of fragments containing multiples of the tandem repeat unit. Hybridization to the 3' probe in the right panel shows that the fragments responsible for the observed TaqI polymorphisms lie downstream of the most 3' TaqI site in the tandem repeat region. Additional analysis showed no polymorphism in the ApaI/ApaI restriction fragment immediately 3' of the tandem repeat region (using the 3' probe) nor any in the immediately downstream ApaI/TaqI restriction fragment using the KpnI/EcoR1 fragment from SMUC-41 as a probe (data not shown). These results show that the sequences flanking the tandem repeat region in the 3' direction are not polymorphic.

Thus, the polymorphisms in the 3' TaqI restriction fragment are due to variations in the length of the tandem repeat containing TaqI/ApaI fragment rather than the ApaI/TaqI fragment lying 3' of the tandem repeats. This is due to sequence variations within the individual tandem repeats, which place the 3' terminal TaqI site in different positions within the tandem repeat region (Fig. 6). The 3' TaqI fragment from pattern A is ~ 350 bp longer than that from pattern B, which corresponds to about five extra tandem repeats.

Structure of the GMUC clone. The sequence of the 5' region of the GMUC clone is shown in Fig. 7. This region of the gene consists of a single long open reading frame as demonstrated by reverse transcription of colonic RNA with subsequent thermal amplification (data not shown). Together with the tandem repeat array, it forms a single large exon which in the most common alleles in genomic DNA extends over 8,700 bp. This 5' region was found to encode a segment of 385 amino acids whose composition was 47.8% threonine, 35.6% proline, and 10.6% serine. Closer examination reveals that this segment has

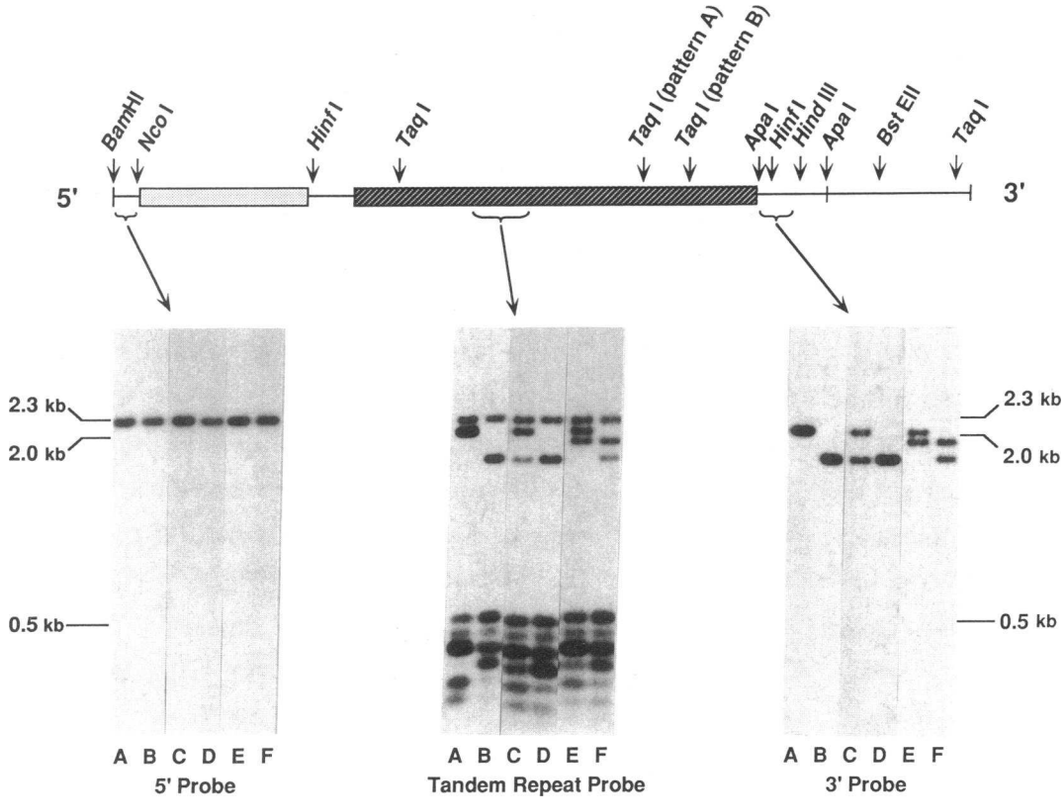

Figure 6. Localization of sequence polymorphisms. The same blot of TaqI digested DNA samples was hybridized sequentially with the 5′ probe (*left*), the MUC-2 tandem repeat probe (*center*), and the 3′ probe (*right*). The location of the probes on the genomic GMUC clone is indicated by the brackets. The arrows indicate the results when the blot was hybridized with the respective probe.

a high degree of internal homology with repeated units having lengths varying from 7 to 40 amino acids (Fig. 8). The most common length appears to be 16 amino acids with the first 9 showing a particularly high degree of conservation at both the amino acid (96.4%) and the nucleotide (89.6%) levels. This high degree of conservation suggests that the repeats arose by a series of duplication events which would have involved varying lengths of sequence at the 3′ end of the segment resulting in units with similar 5′ ends but different 3′ ends. According to this scheme, repeats 7–11 contain 39 or 40 amino acids because the conservation of nucleotide as well as amino acid sequence makes it probable that they underwent duplication in the past as 39 or 40 amino acid units. However, they can be further divided into two 16 amino acid repeats with a 7 or 8 amino acid "tail" at the 3′ end of the second repeat. The sequence for this region of imperfect repeats bears no significant homology to the 69-bp tandem repeats or to any other form of mucin. It is also noteworthy that the 5′ region of imperfect repeats appears to have no length polymorphisms. This may be due to the interspersed repeats of varying length, which would make unequal crossover of large homologous segments of DNA (thought to be the genesis of VNTR length polymorphisms) very difficult.

In addition to this large region of repetitive DNA and the tandem repeat array, there are two 63-bp regions which have a high degree of homology to each other at both the amino acid (87%) and nucleotide level (81%). These are located ~ 80 bp upstream of the threonine/serine and proline rich regions containing the two repetitive arrays (Fig. 7, *enclosed box*). These segments contain one threonine and one serine residue, and hence are not heavily glycosylated. Their role in the overall structure of mucin is not currently understood.

The sequence of the 3′ portion of the GMUC clone immediately downstream of the tandem repeats was the same as the unique sequence of the cDNA clone SMUC 41 for 253 bp followed by a 727-bp intron before the coding sequence was encountered again. The second exon contains only 199 bp before another intron interrupts the sequence. While analysis of the gene 3′ of the tandem repeats is not yet complete, comparison with cDNA sequences indicates that there are at least 7 introns in this region and that no length or sequence polymorphisms have yet been detected (data not shown). The coding regions of the GMUC clone both 5′ (Fig. 7, *underlined*) and 3′ (data not shown) of the tandem repeat array contain multiple cysteine residues.

Discussion

MUC-2 polymorphisms. All four of the currently known human mucin genes have been found to be polymorphic (2–14). The MUC-1 gene has been the most thoroughly examined in this respect. Alleles of this gene have been identified that appear to contain from as few as 20 to as many as 125 tandem repeats (6). mRNA and protein species whose lengths correspond to these alleles have also been identified. However, it has not been possible to demonstrate that large MUC-2 alleles give rise to comparably large gene products and vice versa, due mainly to two factors. First, the MUC-2 protein is far too large to be accurately sized by electrophoresis. Second, the MUC-2 message as determined by blot analysis, is quite polydisperse. Therefore, more indirect methodology was required to examine the allelic variations within the MUC-2 gene.

The MUC-2 gene exhibits a sequence polymorphism that is

```
1     G GAT CCC CAC CTC AGC TTG GAG CAG CTA GGC CAG AAG GTG CAG TGT GAT GTC TCT GTT GGG TTC ATT TGC AAG AAT GAA GAC CAG TTT GGA AAT GGA CCA
        Asp Pro His Leu Ser Leu Glu Gln Leu Gly Gln Lys Val Gln Cys Asp Val Ser Val Gly Phe Ile Cys Lys Asn Glu Asp Gln Phe Gly Asn Gly Pro

101     TTT GGA CTG TGT TAC GAC TAC AAG ATA CGT GTC AAT TGT TGC TGG CCC ATG GAT AAG TGT ATC ACC ACT CCC AGC CCT CCA ACT ACC ACT CCC AGC CCT
        Phe Gly Leu Cys Tyr Asp Tyr Lys Ile Arg Val Asn Cys Cys Trp Pro Met Asp Lys Cys Ile Thr Thr Pro Ser Pro Pro Thr Thr Thr Pro Ser Pro

200     CCA CCA ACC AGC ACG ACC ACC CTT CCA CCA ACC ACC ACC CCC AGC CCT CCA ACC ACC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACC CCC AGC CCT CCA
        Pro Pro Thr Ser Thr Thr Thr Leu Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro

299     ATA ACC ACC ACG ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ATA AGC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACC
        Ile Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Ile Ser Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr

398     ACC ACT CCC AGC CCT CCA ACC ACC ACT CCC AGC CCT CCA ACA ACC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACG ACT ACG CCC
        Thr Thr Pro Ser Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Pro

497     ATC ACT CCA CCA GCC AGC ACT ACC ACC CTT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACA ACC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGT
        Ile Thr Pro Pro Ala Ser Thr Thr Thr Leu Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser

596     CCT CCA ACG ACT ACG CCC ATC ACT CCA CCA ACC AGC ACT ACT ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA CCA ACC ACC ACA ACC ACC CCT CCA
        Pro Pro Thr Thr Thr Pro Ile Thr Pro Pro Thr Ser Thr Thr Thr Leu Pro Pro Thr Thr Thr Pro Ser Pro Pro Pro Thr Thr Thr Thr Thr Pro Pro

695     CCA ACC ACC ACT CCC AGC CCT CCA ACA ACC ACC ACT CCC AGT CCT CCA ACA ATC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACA
        Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Pro Ser Pro Pro Thr Ile Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr

794     ACG ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACG ACT ACA CCC ATC ACT CCA CCA ACC AGC ACT ACC ACC CTT CCA CCA ACC ACC
        Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Pro Ile Thr Pro Pro Thr Ser Thr Thr Thr Leu Pro Pro Thr Thr

893     ACT CCC AGC CCC CCA CCA ACC ACC ACA ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC CCT CCA ACA ACC ACC ACT CCC AGC CCT CCA ATA ACC ACC ACA
        Thr Pro Ser Pro Pro Pro Thr Thr Thr Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Thr Pro Ser Pro Pro Ile Thr Thr Thr

992     ACC ACC CCT CCA CCA ACC ACC ACT CCC AGC TCT CCA ATA ACC ACC ACT CCC AGC CCT CCA ACA ACC ACC ATG ACC ACC CCT TCA CCA ACC ACC ACC CCC
        Thr Thr Pro Pro Pro Thr Thr Thr Pro Ser Ser Pro Ile Thr Thr Thr Pro Ser Pro Pro Thr Thr Thr Met Thr Thr Pro Ser Pro Thr Thr Thr Pro

1091    AGC TCT CCA ATA ACC ACC ACA ACC ACC CCT TCC TCA ACT ACC ACT CCC AGC CCT CCA CCA ACC ACC ATG ACC ACC CCT TCA CCA ACC ACC ACT CCC AGC
        Ser Ser Pro Ile Thr Thr Thr Thr Thr Pro Ser Ser Thr Thr Thr Pro Ser Pro Pro Pro Thr Thr Met Thr Thr Pro Ser Pro Thr Thr Thr Pro Ser

1190    CCT CCA ACA ACC ACC ATG ACC ACC CTT CCA CCA ACC ACC ACT TCC AGC CCT CTA ACA ACT ACT CCT CTA CCT CCA TCA ATA ACT CCT CCT ACA TTT TCA
        Pro Pro Thr Thr Thr Met Thr Thr Leu Pro Pro Thr Thr Thr Ser Ser Pro Leu Thr Thr Thr Pro Leu Pro Pro Ser Ile Thr Pro Pro Thr Phe Ser

1289    CCA TTC TCA ACG ACA ACC CCT ACT ACC CCA TGC GTG CCT CTC TGC AAT TGG ACT GGC TGG CTG GAT TCT GGA AAA CCC AAC TTT CAC AAA CCA GGT GGA
        Pro Phe Ser Thr Thr Thr Pro Thr Thr Pro Cys Val Pro Leu Cys Asn Trp Thr Gly Trp Leu Asp Ser Gly Lys Pro Asn Phe His Lys Pro Gly Gly

1388    GAC ACA GAA TTG ATT GGA GAC GTC TGT GGA CCA GGC TGG GCA GCT AAC ATC TCT TGC AGA GCC ACC ATG TAT CCT GAT GTT CCC ATT GGA CAG CTT GGA
        Asp Thr Glu Leu Ile Gly Asp Val Cys Gly Pro Gly Trp Ala Ala Asn Ile Ser Cys Arg Ala Thr Met Tyr Pro Asp Val Pro Ile Gly Gln Leu Gly

1487    CAA ACA GTG GTG TGT GAT GTC TCT GTG GGG CTG ATA TGC AAA AAT GAA GAC CAA AAG CCA GGT GGG GTC ATC CCT ATG GCC TCC TGC CTC AAC TAC GAG
        Gln Thr Val Val Cys Asp Val Ser Val Gly Leu Ile Cys Lys Asn Glu Asp Gln Lys Pro Gly Gly Val Ile Pro Met Ala Phe Cys Leu Asn Tyr Glu

1586    ATC AAC GTT CAG TGC TGT GAG TGT GTC ACC CAA CCC ACC ACC ATG ACA ACC ACC ACC ACA GAG AAC CCA ACT CCG ACA CCA ATC ACC ACC ACC ACT ACG
        Ile Asn Val Gln Cys Cys Glu Cys Val Thr Iln Pro Thr Thr Met Thr Thr Thr Thr Thr Glu Asn Pro Thr Pro Thr Pro Ile Thr Thr Thr Thr Thr

1685    GTG ACC CCA ACC CCA ACA CCC ACC AGC ACA CAG AGT
        Val Thr Pro Thr Pro Thr Pro Thr Ser Thr Gln Ser 100
```

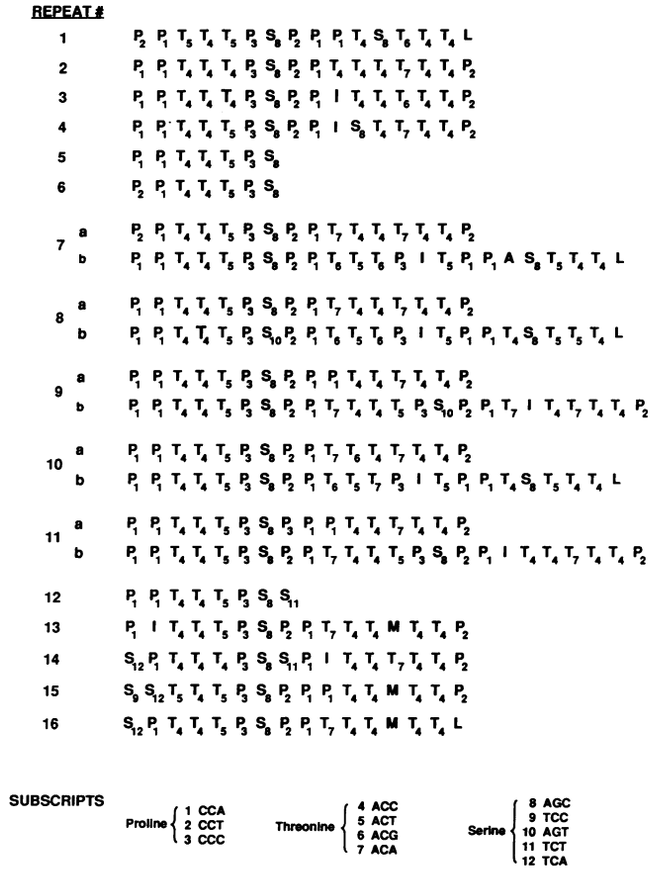

Figure 8. Protein sequence of the region of imperfectly conserved repeats. The protein sequence enclosed in the round brackets in Fig. 7 is presented in repeated units, which are numbered. The longer 39 and 40 amino acid repeats (Nos. *7–11*) are divided into "a" and "b" segments in order to show the internal 16 amino acid repeats. The subscripts indicate the actual codon used for the particular amino acid residue and are listed at the bottom of the figure.

detectable using the enzyme TaqI. Most of the 171 individual human DNA samples analyzed with this enzyme gave two or three bands between 1.7 and 2.3 kb which hybridized to the MUC-2 tandem repeat probe. The 2.3-kb band was constant in more than 90% of all alleles examined and was found to be derived from the 5′ portion of the tandem repeats and upstream unique sequences. The common polymorphism arose from the 2.1- and 1.7-kb bands. These bands were excised from the 3′ portion of the tandem repeat array and accompanying unique sequences, with the TaqI site giving rise to the polymorphism being located within the tandem repeat region and not in the downstream sequence. Thus, the common TaqI polymorphism is due to a sequence variation in the 3′ region of the tandem repeat array. An ethnic difference in this polymorphism was also detected. Almost all Asians are homozygous for the pattern A TaqI allele, while Caucasians have approximately equal numbers of A and B alleles.

More significantly in terms of protein structure, the MUC-2 gene was also found to exhibit length polymorphism in its tandem repeat array. This was suggested by blot analysis using HinfI (Fig. 2) and confirmed by thermal amplification using primers flanking the tandem repeats (Fig. 5). The MUC-2 gene therefore appears to have alleles with VNTRs as has been reported for MUC-1 (2). In this study, alleles were detected which varied between 51 and 115 tandem repeat units (calculated from the data in Fig. 2). The size distribution of MUC-2 alleles does not appear to vary as dramatically as is the case with MUC-1 alleles, with the majority of the alleles examined containing ~ 100–115 tandem repeat units. Thus, the most common size for the tandem repeat domain in the MUC-2-type intestinal mucin is slightly more than 2,300 amino acids. Even without taking into account the sizes of the carboxy- and amino-termini, the MUC-2 protein is very large, indeed.

MUC-2 gene and GMUC clone. The GMUC genomic DNA clone is 11 kb in length, or 4 kb shorter than the BamHI fragment from which it was derived, due to a deletion occurring within the tandem repeat array. Several lines of evidence support this conclusion. First, the sequence of the GMUC clone immediately downstream from the tandem repeat array is the same as the corresponding sequence in the SMUC 41 cDNA (9), implying that the 3′ portion of the sequence has not been artifactually recombined. Second, thermal amplification using primers based upon sequences flanking the tandem repeats in GMUC gives a 7-kb band with genomic DNA, whereas this distance is ~ 3 kb in GMUC. This strongly suggests that 4 kb has been deleted from the tandem repeats in GMUC. Finally, blot analysis using HinfI indicates that the tandem repeat array in both alleles of the genomic DNA sample used for library preparation is 7 kb, rather than the 3 kb which would be expected from analysis of the GMUC clone. Given the high degree of sequence similarity between repeat units, this deletion is likely to have occurred via an unequal crossover during phage propagation. The high number of plaques that had to be screened to obtain this single stable clone reflects its rare and serendipitous nature.

The experiments shown in Figs. 4 and 5 suggest that the tandem array of both the GMUC clone and the MUC-2 gene is uninterrupted. Thus, the tandem repeat region appears to be contained within a single exon. Clearly, it would be more desirable to demonstrate this by cloning and sequencing the entire tandem repeat array; however, given the problems encountered in cloning this region, this does not appear to be practical.

The presence of the 1,077-bp repetitive segment in the 5′ portion of the GMUC clone is quite intriguing. This DNA segment encodes a peptide that is very rich in threonine and proline and in this respect is similar to the 69-bp tandem repeats. This is perhaps further evidence for the role of proline as a recognition factor for the glycosyltransferases responsible for *O*-linked oligosaccharide synthesis (25, 26). This region is dissimilar to the 69-bp tandem repeats however, in that its 16

Figure 7. Nucleotide sequence for the 5′ portion of the GMUC clone. The cysteine residues are underlined. The repetitive portion of the 5′ region containing a high threonine/serine and proline concentration (referred to in the text as the region of imperfectly conserved repeats) is enclosed in round brackets. The tandem repeat region is enclosed in square brackets. The boxed-in regions contain the two 63-bp repeats. These sequence data are available from EMBL/GenBank/DDBJ under accession number M74027.

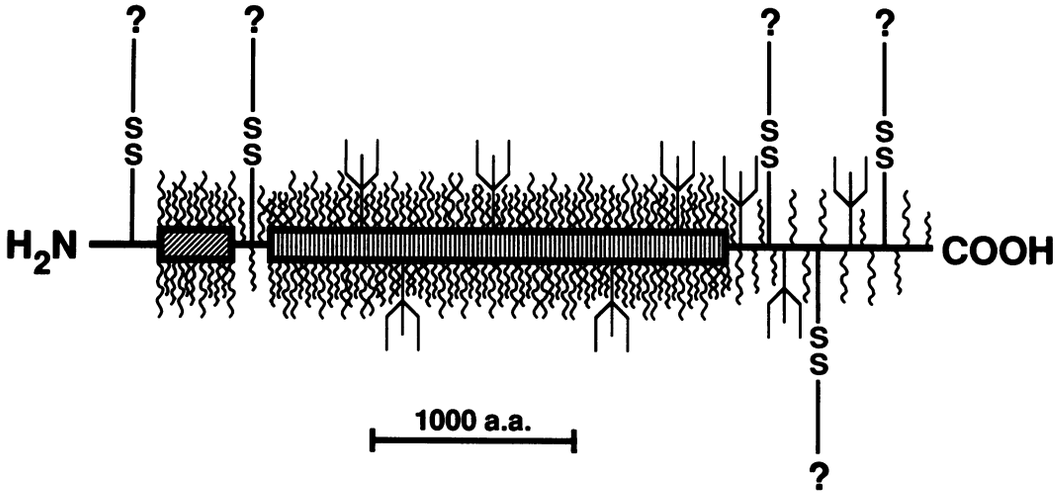

Figure 9. Current model of MUC-2 glycoprotein structure. (ψ) Potential *N*-glycosylation site. (⦚) Potential *O*-glycosylation site. (▨) Region of imperfectly conserved repeats. (▩) Tandem repeat region.

amino acid repeat units are often noncontinuous. Furthermore, it should be noted that several of the interrupting sequences found in this region of the GMUC clone share considerable sequence similarity (Fig. 8, repeats *5, 6,* and *11;* the terminal 7 amino acids of 7b, 8b, and 10b; and the terminal 8 amino acids of 9b and 11b).

Thus, the MUC-2 gene has at least two different sizable repetitive domains. Both of these domains are found on the same large exon, separated by ~ 600 bp. This is the first mucin found to have two repetitive domains rather than just one. It will be interesting to determine if these domains developed separately or whether they evolved as a continuous unit over time. The current working model of the MUC-2 glycoprotein present in the GMUC clone is shown in Fig. 9.

The presence of multiple cysteine residues both 5′ (Fig. 7, *underlined residues*) and 3′ (data not shown) strongly supports the idea that the MUC-2 codes for a secreted mucin, because disulfide bonding is necessary for the formation of a mucus gel. The potential of forming multiple disulfide bonds in the mucin molecule makes possible head-to-tail, head-to-head, and tail-to-tail arrangements of mucin molecules. This could allow both linear arrays and the "windmill" type conformation without the need for a "link" peptide (1, 27).

This study provides the first characterization of the major structural features of the MUC-2 gene including the tandem repeat region, which is the source for the polymorphisms seen in the MUC-2 gene of normal subjects. A separate highly repetitive region encoding a high concentration of potential *O*-glycosylation sites and having a highly unusual structure is also described. The structure of this widely distributed mucin may provide important insights into the mechanisms by which secreted mucins perform their functions.

Acknowledgments

The authors would like to thank Jane Brown for her invaluable assistance in obtaining subjects and Devina Magalong for her help with the DNA isolations.

This work was supported by the Veterans Administration Medical Research Service.

References

1. Neutra, M. R., and J. F. Forstner. 1987. Gastrointestinal mucus: synthesis, secretion, and function. *In* Physiology of the Gastrointestinal Tract. L. R. Johnson, editor. Raven Press, New York. 975–1009.

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

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

4. 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. USA.* 85:2320–2323.

5. Wreschner, D. H., M. Hareuveni, H. Tsarfaty, N. Smorodinsky, J. Horev, J. Zaretsky, P. Kotkes, M. Weiss, R. Lathe, A. Dion, and I. Keydar. 1990. Human epithelial tumor antigen cDNA sequences. *Eur. J. Biochem.* 189:463–473.

6. 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.

7. Ligtenberg, M. J. L., H. L. Bos, A. M. C. Gennisser, 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.

8. Griffiths, B., D. J. Matthews, L. West, J. Attwood, S. Povey, D. M. Swallow, J. R. Gum, and Y. S. Kim. 1991. Assignment of the polymorphic intestinal mucin gene (MUC-2) gene to chromosome 11p. *Ann. Hum. Genet.* 54:277–286.

9. Gum, J. R., J. C. Byrd, J. W. Hicks, N. W. Toribara, D. T. A. Lamport, and Y. S. Kim. 1989. Molecular cloning of human intestinal mucin cDNAs. *J. Biol. Chem.* 264:6480–6487.

10. Gerard, C., R. L. Eddy, and T. B. Shows. 1990. The core polypeptide of cystic fibrosis tracheal mucin contains a tandem repeat structure. *J. Clin. Invest.* 86:1921–1927.

11. Jany, B. H., M. W. Gallup, P.-S. Yan, J. R. Gum, Y. S. Kim, and C. B. Basbaum. 1991. Human bronchus and intestine express the same mucin gene. *J. Clin. Invest.* 87:77–82.

12. Van Cong, N., J. P. Aubert, M. S. Gross, N. Porchet, P. Degand, and J. Frezal. 1991. Assignment of human tracheobronchial mucin gene(s) to 11p15 and a tracheobronchial mucin-related sequence to chromosome 13. *Hum. Genet.* 86:167–172.

13. Gum, J. R., J. W. Hicks, D. M. Swallow, R. L. Lagace, J. C. Byrd, D. T. A. Lamport, B. Siddiki, and Y. S. Kim. 1990. Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem. Biophys. Res. Commun.* 171:407–415.

14. Porchet, N., N. Van Cong, J. Dufosse, J. P. Audie, V. Guyonnet-Duperat, M. S. Gross, C. Denis, P. Degand, A. Bernheim, and J. P. Aubert. 1991. Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem. Biophys. Res. Commun.* 175:414–422.

15. Mantle, M., and G. Stewart. 1989. Intestinal mucins from normal subjects and patients with cystic fibrosis. *Biochem. J.* 259:243–253.

16. Podolsky, D. K., and K. J. Isselbacher. 1984. Glycoprotein composition of colonic mucosa: Specific alterations in ulcerative colitis. *Gastroenterology.* 87:991–998.

17. Itzkowitz, S. H., M. Yuan, C. K. Montgomery, T. Kjeldsen, H. K. Takahashi, W. L. Bigbee, and Y. S. Kim. 1989. Expression of Tn, Sialosyl-Tn, and T antigens in human colon cancer. *Cancer Res.* 49:197–204.

18. Boland, C. R., C. K. Montgomery, and Y. S. Kim. 1982. Alterations in human colonic mucin occurring with cellular differentiation and malignant transformation. *Proc. Natl. Acad. Sci. USA.* 79:2051–2055.

19. Bresalier, R. S., and Y. S. Kim. 1989. Malignant neoplasms of the large and small intestine. *In* Gastrointestinal Disease. M. H. Sleisenger and J. S. Fordtran, editors. W. B. Saunders, Philadelphia. 1519–1560.

20. Kern, S. E., E. R. Fearon, K. W. F. Tersmette, J. P. Enterline, M. Leppert, Y. Nakamura, R. White, B. Vogelstein, and S. R. Hamilton. 1989. *J. Am. Med. Assoc.* 261:3099–3103.

21. Blin, N., and D. W. Stafford. 1976. Isolation of high molecular weight DNA. *Nucleic Acids Res.* 3:2303–2308.

22. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98:503–517.

23. Sambrook, J., E. F. Fritsch, and T. Maniatas. 1989. Molecular Cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

24. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74:5463–5467.

25. Hanover, J. A., W. J. Lennarz, and J. D. Young. 1980. Synthesis of *N*- and *O*-linked glycopeptides in oviduct membrane preparations. *J. Biol. Chem.* 255:6713–6716.

26. Briand, J. P., S. P. Andrews, Jr., E. Cahill, N. A. Conway, and J. D. Young. 1981. Investigation of the requirements for *O*-glycosylation by bovine submaxillary gland UDP-*N*-acetylgalactosamine:polypeptide *N*-acetylgalactosamine transferase using synthetic peptide substrates. *J. Biol. Chem.* 256:12205–12207.

27. Carlstedt, I., and J. K. Sheehan. 1984. Macromolecular properties and polymeric structure of mucous glycoproteins. *Ciba Found. Symp.* 109:157–172.