MUC3 Human Intestinal Mucin

ANALYSIS OF GENE STRUCTURE, THE CARBOXYL TERMINUS, AND A NOVEL UPSTREAM REPETITIVE REGION

James R. Gum, Jr.‡§‡, Jenny L. Hoš, Wendy S. Pratt, James W. Hicks§, Alex S. Hill, Lynne E. Vinall, Anthony M. Roberton§§, Dallas M. Swallow, and Young S. Kim‡‡

From the Gastrointestinal Research Laboratory (151M2), Department of Veterans Affairs Medical Center and Departments of Anatomy, Medicine, and Pathology, University of California, San Francisco, California 94121 and the MRC Human Biochemical Genetics Unit, The Galton Laboratory, University College London, Warden House, 4 Stephenson Way, London, NW1 2HE, United Kingdom

MUC3 is a large mucin glycoprotein expressed by the human intestine and gall bladder. In this manuscript, we present details of the deduced protein structure of MUC3. The MUC3 carboxyl-terminal domain is 617 residues in length, including 511 residues of a non-repetitive mucin-like domain (27% Thr, 22% Ser, and 11% Pro) and a 106-residue Cys-rich domain with homology to the epidermal growth factor (EGF) -like structural motifs found in many proteins. The region of MUC3 located upstream of the previously described 51-base pair (bp) tandem repeats, which encode a major Ser and Thr-rich domain, consists of a second type of repetitive structure with an imperfect periodicity of approximately 1125 bp. This domain is also mucin-like and appears to be considerably larger than 2000 residues (6000 bp). The MUC3 gene itself is large and complex. Using pulse field gel electrophoresis and blot analysis, the smallest fragment found that contained all human genomic DNA hybridizing to the 51-bp tandem repeat probe was 200 kilobases with restriction enzyme ScaI. Both PvuII and PstI produced two sets of hybridizing fragments that were hypervariable within the human population with a pattern suggestive of both a variation in the number of tandem repeats (VNTR) and sequence polymorphism. These fragments varied independently of each other, but no genetic recombination was detected in a study of 40 human families. Thus, the MUC3 gene encodes a very large glycoprotein with a structure very different from that of any mucin currently described.

The mucosal surfaces of the gastrointestinal, respiratory, and reproductive tracts secrete mucus, a hydrophilic gel that serves to coat and protect delicate epithelial linings. Mucus owes its hydrophilic and viscoelastic properties to its content of the glycoproteins collectively known as mucins. Mucins are large, heterogeneous and highly glycosylated proteins (1–4) in the glycoproteins collectively known as mucins. Mucins are owes its hydrophilic and viscoelastic properties to its content of

which the vast majority of the carbohydrate side chains are O-linked to Thr and Ser residues of the polypeptide backbone. Mucin heterogeneity is due partly to the glycosylation, partly to variability of polymerization (5, 6), and partly due to the fact that mucin polypeptides are encoded by a number of distinct and highly polymorphic genes that are variably expressed (1, 2). The identification of individual mucins has been greatly assisted by the isolation of partial cDNA clones that have usually contained tandem repeat domains, the sequence of which is unique to a particular mucin gene.

In humans, at least nine different mucin genes encode epithelial mucins (1, 2, 7–9). Of these, the complete cDNA sequence is available for only three, the membrane mucin MUC1 and two that encode secretory mucins, MUC2 which is expressed in the intestine and MUC7 which is expressed in the salivary glands (9–11). The MUC1 cDNA ranges in size from 800 to 1700 amino acid residues depending upon its content of a 20 residue tandemly repeated peptide unit that comprises much of the coding region and exhibits a variable number of tandem repeats (VNTR) polymorphism within the population (12–14). This integral membrane protein is associated with many different types of epithelium. The deduced MUC2 polypeptide, which also shows VNTR polymorphism, of a 23 amino acid unit has a monomeric size of more than 5000 amino acids in most alleles (10), has been shown to undergo dimerization (16–18), and is capable of higher order polymerization (6). It has a central repetitive segment consisting of two separate regions of tandem repeats that both contain more than 50% content of Ser and Thr residues, i.e. potential O-glycosylation sites (19). In addition, MUC2 contains four conserved cysteine-rich domains that have sequence similarity to the D-domains of pro-von Willebrand clotting factor (10, 20). MUC7 is a much smaller molecule than MUC2 with a coding sequence of only 377 residues (9). This mucin contains a central segment consisting of six tandemly repeated peptides of 23 residues, each rich in Thr, Ser, and Pro. The central repetitive region is flanked by short, hydrophilic domains (9). The mature MUC7 peptide contains only 27 cysteine residues. While MUC2 and MUC7 are the only secretory mucins for which the entire sequence has been determined, portions of the unique sequences of several other mammalian mucins are known (21–25).

1 The abbreviations used are: VNTR, variable number of tandem repeats; RACE, rapid amplification of cDNA ends; PCR, polymerase chain reaction; EGF, epidermal growth factor; bp, base pair(s); kb, kilobase(s); mb, megabase(s); VECs, vectorette sequence; contig, group of overlapping clones; PFGE, pulsed field gel electrophoresis.

2 L. E. Vinall, A. S. Hill, P. Figny, W. S. Pratt, N. Toribara, J. R. Gum, Y. S. Kim, N. Forchett, J. P. Aubert, and D. M. Swallow (1997) Hum. Genet., submitted for publication.
**MUC3 Human Intestinal Mucin**

**MUC3** was identified through the cloning and characterization of partial intestinal cDNAs. These cDNAs encode a peptide domain containing 17 residue Thr- and Ser-rich tandem repeats (26). The corresponding gene is located on chromosome 7q22 and is thus clearly a separate entity from **MUC2**, which is located on chromosome 11p15.5 (27). Also, **MUC3** exhibits a different pattern of expression from **MUC2**. In the intestines, **MUC2** is expressed solely in goblet cells while **MUC3** is expressed in both goblet cells and enterocytes (28, 29). **MUC3** is also expressed at high levels in the gall bladder epithelium (30).

Other than the initial description of the partial cDNAs containing 51 residue tandem repeats (26), details of the structure of **MUC3** have not yet been published. In this manuscript, we describe the complete structure of the carboxyl-terminal domain of **MUC3**. We also describe additional cDNA clones we have isolated and genomic DNA sequence immediately upstream from **MUC3** tandem repeats, which provide additional information pertaining to the complexity and polymorphism of this gene.

**EXPERIMENTAL PROCEDURES**

**cDNA and Genomic DNA Cloning and RACE PCR**—The human intestine Agt11 cDNA library used previously for the isolation of **MUC2** and **MUC3** cDNA clones was also used for this study (31). Immunoblotting was performed as described (11) using polyclonal antiserum to deglycosylated pancreatic cancer cell line SW1990 xenograft mucin (32). A human placental genomic DNA library in the **FIXII** vector was obtained from Stratagene. Hybridization probe screening was done as described (20). RACE PCR of the 3' terminus of the **MUC3** message was performed as described previously using total RNA prepared from human small intestine (20). The reverse transcription reaction was conducted using the (dT)17 adapter 5' primer (Table I). The primary PCR amplification utilized the R1 primer and 3' RACE primer, as a gene-specific primer. A secondary amplification was conducted using the R1 primer and a second gene-specific primer, **MUC3R3**. A fragment of approximately 1000 bp was obtained and cloned into the pCRII vector (Invitrogen) for sequencing.

**Vectorette-PCR**—Specific fragments of DNA upstream from known sequence were obtained by Vectorette-PCR. A HindIII Vectorette library was constructed using genomic DNA from a single individual and the pre-digested vector provided in the Vectorette I starter pack S (Genosys Biotechnologies, Inc.). One μg of restricted DNA (50 μl) was added to 5 μl of vectorette (3 pmol) together with 1 μl of T4 DNA ligase (1 unit/μl), 1 μl of ATP (100 mM), and 1 μl of dithiothreitol (100 mM). The reaction was subsequently cycled 3 times between 20 °C for 60 min and 37 °C for 60 min and 1 h.

The PCR was carried out using a hot start method, with 1 μl of library/100 μl of reaction using the universal vectorette primer (Genosys) together with a gene-specific primer. The reaction was prepared as two layers with a combined volume of 100 μl separated by a layer of wax. The PCR conditions were as recommended by the suppliers of the AmpliWax PCR Gem (Perkin-Elmer) and Taq polymerase (Promega). The lower layer contained 500 μM nucleotides, 50 pmol of each primer, 1 × Taq polymerase buffer and water to a volume of 40 μl. A single AmpliWax ball was then added to each tube, and the contents were heated to 78 °C for 5 min and then cooled to form an impermeable barrier. The upper reaction mix contained 1 to 2 units of Taq polymerase, 1 × buffer, target DNA, and water to a volume of 60 μl. Subsequently, 10 cycles of amplification were conducted that consisted of denaturation for 1 min at 94 °C, annealing for 30 s, and elongation at 72 °C for 3 min, reducing the annealing temperature by 1 °C each cycle from 70 to 61 °C. The remaining 20 cycles consisted of denaturation at 94 °C for 1 min, annealing at 60 °C for 30 s, and elongation at 72 °C for 3 min. If necessary to obtain a specific product, the reaction was repeated using a nested primer and 1 μl of a 1:1000 dilution of the first PCR product. The reactions were performed using a Perkin-Elmer DNA thermal cycler.

**DNA Sequencing and Analysis**—Most of the sequencing was performed using Sequenase (U. S. Biochemical Corp.) with M13mp18 or M13mp19 single strand templates except as noted. The exonuclease III sequential deletion method was used to generate most sequencing templates (33). cDNA clones SIB227, SIB238, SIB276, SIB297, and the 3'-RACE PCR fragment were sequenced by the UCSF Biomolecular Center using an Applied Biosystems model 373A automated sequencer with standard protocols. Custom primers were utilized with the automated sequencer to obtain some internal sequences of the cloned fragments, and in some cases individual subclones were prepared. Vector-X products were sequenced on both strands by cycle sequencing using the Thermo Sequenase cycle sequencing kit (Amersham Corp.). Enzyme sequencing was performed by the manufacturer and using specific primers and the universal vectorette sequencing primer (UVseqP, Genosys). Comparison of sequences to data bases was performed using FASTSD software (Intelligenetics, Inc.), the BlastP and BlastN Service provided by the National Center for Biotechnology Information and the GCG Wisconsin package using the UK MRC HGMP Resource Center computing facility. Multiple sequence alignment was performed using the ClustalW routine of Macvector 6.0.

**DNA Blot Analysis**—Agarose electrophoresis of restriction enzyme-digested DNA was carried out in 1 × Tris-borate-EDTA buffer and capillary transfer to Hybond N+ was performed as recommended by the manufacturer (Amersham Corp.). For sizing the large bands, 24 cm, 0.5% agarose gels were run at 50 volts for 24 h and then for a further 19 h at 30 volts (long electrophoresis system). Blots containing restriction enzyme-digested genomic DNA obtained from members of the CEPH (Center d'Etude du Polymorphisme Humain) families and prepared by the EUROGEM (European genome mapping initiative) consortium were also used (34). The blots were washed to a final stringency of 0.1 × SSC at 65 °C for 10 min. Sizes were determined by comparison with Raoul molecular weight markers (Appligene).

Pulsed field gel electrophoresis (PFGE) on the CHEF-DR II pulsed field electrophoresis system (Bio-Rad) was carried out using 1% agarose gels on which the DNA was applied as agarose blocks. For separations of fragments from 50 kb to 2 mb, the pulse time was increased linearly from 10 to 250 s over the course of the run and at a potential difference of 150 volts for 40 h. For separations of fragments from 5 to 200 kb, the pulse time was ramped from 1 to 20 s for a run time of 20 h. Sizing was done using Bio-Rad Saccharomyces cerevisiae and λ ladder size markers. DNA from the erythroid cell line K562 was used for most of the PFGE analysis because it showed a relatively lower level of DNA methylation than other cell lines (7).

**RESULTS**

**Isolation of 3'-Terminus Clones**—Screening of the human small intestine cDNA library with antiserum directed against deglycosylated pancreatic cancer xenograft mucin gave the 460 bp cDNA clone designated Puck in Fig. 1. The protein sequence encoded by Puck was mucin-like but contained no tandem repeats and was initially assumed to encode a portion of a previously unidentified mucin. Rescreening of the library using Puck as a hybridization probe resulted in the isolation of clones 20 and 23. Together the three cDNAs encoded a 625-residue mucin-like peptide with a substantial 3'-untranslated region. These cDNAs possessed no discernible repetitive sequence nor did they have any unambiguous similarity to any known sequence.

In an unrelated experiment, a human genomic DNA library was screened using the tandem repeat clone SIB124 (26) as a hybridization probe. A single intensely positive plaque, designated clone GM3, containing a 20-kb insert was isolated from a genomic library of Macvector 6.0. The sequence were obtained by Vectorette-PCR. A single intensely positive plaque, designated clone GM3, containing a 20-kb insert was isolated from a genomic library of Macvector 6.0. The sequence was identical to that defined by clones 23, 20, and PUCK, and they did not have any unambiguous similarity to any known sequence.

In situ hybridization using both the cDNA clone 23 and the genomic clone GM3 revealed a signal at a single locus on chromosome 7q22 (data not shown). The genomic sequences were fully sequenced (Figs. 1 and 2). The position and size of the introns were confirmed by PCR of genomic DNA and of cDNA prepared from small intestine, colon, and HT29 cells (data not shown).
**Structure of the Tandem Repeats and 3' Terminus**—Fig. 2 shows the combined coding sequence of the clones described in Fig. 1. This segment of MUC3 contains 20 tandem repeats with 88% similarity to the consensus sequence, as shown in Fig. 3. While the individual consensus sequence of the tandem repeats found in clone GM3 matched the overall consensus sequence, the individual consensus sequences of both SIB139 and SIB124 were different (Ref. 26, Fig. 3), suggesting regional variation in consensus sequence. Downstream of the tandem repeats 1851 nucleotides of open reading frame encode a stretch of 511 residues of mucin-like non-repetitive sequence, containing 27% Thr, 22% Ser, and 11% Pro, and only three cysteines followed by 106 residues, containing 9% Cys, 3% Thr, 4% Ser, and 8% Pro. A small segment of this cysteine-rich region exhibited sequence similarity to the unique sequence of rat intestinal mucin cDNA clone RMUC176 that we isolated previously (35).

Moreover, the spacing of the cysteines in this terminal region, and to a lesser extent the overall sequence, were similar to the EGF-like structural motifs of several proteins (Fig. 4). There was no hydrophobic domain in this carboxyl-terminal membrane region.

The stop codon at nucleotide 2873 is followed by more than 1800 bases of putative 3'-untranslated sequence (Fig. 2). However, nucleotides 2736 to 3488 (Fig. 2) constitute an alternative, partially overlapping open reading frame, and thus extreme care was taken in checking the sequence between base 2736 and base 2873. Neither substitution of dGTP with dITP, nor use of the Taqtrack system (Promega), revealed any errors in this sequence. A portion of this sequence initiating at base 3583 was nearly identical to that of a 320-bp expressed sequence tag obtained by Greer and co-workers (40) (Genbank accession number Z36791), which was interestingly obtained in a study aimed at defining transcripts preferentially expressed in pancreatic cancer.

**Sequences 5' to the Tandem Repeats**—More than 30 cDNA clones were isolated by screening the human intestinal library with SIB124. Many of these were rearranged or were chimeric, but one, SIB172 (598 bp), had a 380 bp unique segment that encoded a Thr-, Ser-, and Pro-rich unique segment upstream and in frame with its tandem repeats. A probe (SIB172U) made from the first 357 bases of this sequence was used to again screen the library. Importantly, three independent clones were obtained that bridged the interface between the unique sequence and the tandem repeats, confirming the authenticity of clone SIB172. Other clones were found that were similar but not identical to SIB172U, suggesting that the “unique” sequence is itself repeated. The longest of these, SIB227 (1079 bp), was used to rescreen the cDNA library. Several clones were obtained including SIB238 (818 bp) that contained a region of identical overlap with SIB227 allowing the two sequences to be combined to form a contig (SIB227C, 1555 bp). Two other clones, SIB297 (1031 bp) and SIB276 (1779 bp) had sequences that were similar to SIB227C but were not identical (Fig. 5A). Furthermore, SIB227C and SIB276 both have an internal duplication of a long stretch of sequence indicative of a tandem repeat unit of approximately 1125 nucleotides in which individual repeat units show some amino acid substitutions as well as deletions and insertions. The repetitive nature of these sequences is readily apparent in the similarity matrices shown in Fig. 5B.

In parallel, upstream genomic sequence was obtained using vectorette PCR with total genomic DNA as template using primers designed from the sequence of SIB172 (MUC3FP3A and MUC3FP2A, Table I). The first product, VEC1 (600 bp), overlapped SIB172 by 265 nucleotides. A further 350-bp product was made using two more primers designed from the 5' end of the VEC1 sequence, MUC3FP4A and the nested primer MUC3FP5A (Table I). To confirm the continuity of these segments, we sequenced PCR products made from genomic DNA to span the junction between these new sequences and SIB172. The combined vectorette sequence (VECS) shows a high level of identity with the large tandem repeat containing cDNAs (Fig. 5A). VECS is identical to SIB172 in the region of overlap, is clearly located immediately adjacent to a region of 51-bp tandem repeats, and does not contain introns. The combined coding sequences obtained from VECS and SIB172 are shown in Fig. 6, designated V172.

The domain of 1125-bp repeat encodes a peptide sequence that is mucin-like, comprising 31% Thr, 21% Ser, 8% Pro, 6% Leu, and less than 5% of any other amino acid residue. In total, approximately 6000 bp of the 5'-repetitive sequence has been isolated without extensive overlap being evident, suggesting a very large size for this domain. Interestingly, all of the clones and contigs exhibit at least one region of similarity to the 51 bp
MUC3 tandem repeats. In all clones identified so far, however, (with the exception of SIB172), these regions of similarity are one repeat unit or less in length. Other than this similarity to the 51 bp MUC3 tandem repeats, no similarity to other sequences in the GenBankTM/EBI Data Bank was detected. Interestingly however, we did note that the initial 101 residues of the carboxyl terminus of MUC3 shown in Fig. 2 had 59 identical residues as a region of SIB227C (residues 158–258, alignment not shown).

DNA blot analysis of MUC3 gene structure—We have previously reported that the tandem repeat probe SIB124 recognizes the 51 bp tandem repeats, no similarity to other sequences in the GenBankTM/EBI Data Bank with accession numbers AF007195 and AF007196. Tandemly repeated peptides are indicated by underlined text. Clone GM3–1 contains a 106-base intron located at base 2735. Intronic sequences are deposited in the GenBankTM/EBI Data Bank with accession numbers AF007195 and AF007196.

The sizes of the PstI bands were estimated in comparison with the Raoul size markers, which were applied twice to each gel. More than 80 unrelated individuals were tested. The apparent size range of the large set varies from 20 kb to greater than 48.5 kb. The smaller set of polymorphic fragments detected vary in size from 9.4 to 13.5 kb (Fig. 7). The PstI bands described previously were not sized accurately using the long electrophoresis system, but they are very similar in size to the PstI bands. The broad similarity of the patterns observed in the upper set of fragments is not independent on that seen in the lower set. This suggests that the polymorphic regions of DNA are physically separated in the genome and do not arise from common restriction sites. A similar situation was observed with PstI. The sequences are, however, close in the genome since no genetic recombination was detected in these families as previously reported in abstract form (41).

The consensus sequences defined from the overall consensus sequence. Clone GM3 was isolated in this study, and clones SIB139 and SIB124 were described previously (26). The consensus sequences defined from the overall consensus sequence. Clone GM3 was isolated in this study, and clones SIB139 and SIB124 were described previously (26).
In addition, a 1.8-kb \textit{Pvu}II fragment is detected with both clone 20 and SIB172U. Weaker additional bands were also detected with \textit{Pvu}II in some individuals. It should also be noted that the 2.8-, 2.6-, and 1.5-kb fragments detected with SIB172U in \textit{PstI}-digested DNA show some person to person variation that may be related to the proposed polymorphic \textit{PstI} site mentioned previously.

**Pulsed Field Gel Electrophoresis**—Blots of K562 DNA digested with a variety of enzymes and separated by PFGE were

**FIG. 5**—continued

The other sequences are cDNAs. Bold letters, background, and numbers are as in Fig. 4. Arrows demarcate the 375 amino acid repeats. Sequence with significant homology to the previously identified 17 amino acid tandem repeats (26) is underlined. The nucleic acid sequences for SIB227C, SIB276, SIB297, and VECS are deposited in the GenBank\textsuperscript{130} EBI Data Bank with accession numbers AF007190, AF007191, AF007192, and AF007197. Panel B, similarity matrices illustrating the amino acid similarity (identical and chemically similar residues) between SIB227C and SIB276 (top) and the internal similarity present in clone SIB276 (bottom). Dots, which combine to form lines, represent regions of similarity. The \textit{parallel lines} are indicative of repeats with a periodicity of 375 amino acids.

**FIG. 5**—

**Panel A**, alignment of sequences containing the 1125-bp (375 amino acids) repeats. VECS represents the contig formed from the genomic vectorette clones, and
analyzed to gain insight into the overall size and structure of the MUC3 gene. Blots probed with SIB124 exhibited multiple bands in some cases, but a single large fragment was seen with NotI and BssHII (370 and 355 kb, respectively, Fig. 10A). In separate experiments we noted that SwaI produced a single 200-kb fragment. The size of this band did not vary significantly when DNA from five donors was examined, indicating a lack of large scale size differences in the region around MUC3 between individuals (Fig. 10B). Similar band patterns were observed with NotI, BssHII, and SwaI when the blots were probed with clone 23 and SIB172U, indicating that these single fragments contain the flanking regions of the MUC3 gene (data not shown). With at least two enzymes (Smal and NcoI), there was evidence of two bands of approximately equal intensity (Fig. 10A), and it seems possible that these correspond to the two “zones” separated by PouII and PstI (Fig. 7).

**DISCUSSION**

Southern blot analysis of MUC3 reveals that it covers a large genetic region. Two sets of very large polymorphic restriction fragments are detected with the small tandem repeat CDNA probe, SIB124, and the smallest single fragment we could detect with this probe covered approximately 200 kb (Fig. 10). Previously only two short cDNAs consisting of 51-bp tandem repeats have been described for MUC3 (26). Considerable efforts have been made in two laboratories to isolate further cDNAs and genomic clones. Extensive screening of cosmids and YAC libraries yielded only a few rearranged clones, but a λ phage clone corresponding to the 3’ end of the gene and many partial cDNAs were isolated along with vectorette PCR products. Taken together, these clones and PCR products offer considerable insight into the structure of the MUC3 gene and its message.

Here, we report the complete sequence of the MUC3 carboxyl terminus and a partial structure for a large domain of 375-amino acid residue tandem repeats located immediately upstream of the 51-bp (17-amino acid residue) repeats. The carboxyl terminus contains a mucin-like domain and a cysteine-rich motif. The cysteine-rich sequence lacks detectable similarity to other mucins (20–24), and neither does it have detectable similarity to the various trefoil factors abundantly present in many proteins, especially in terms of the spacing of the cysteines. EGF-like motifs may serve to expose ligand binding sites on the exterior regions of a tri-lobe structure, but MUC3 ligands, if present, are currently unidentified. In the case of porcine submaxillary mucin, a cysteine-rich region in the extreme carboxyl terminus has been shown to be important for dimerization (24, 45). Similar regions are present in MUC2 and other mucins (17, 20, 22, 23, 46) and are probably important for processing and assembly in these mucins as well. It will be important to determine the function of the cysteine-rich motif of MUC3. The 375-residue repeats located upstream of the tandem repeats are mucin-like with their high content of Thr, Ser, and Pro. The sequence of this region of the macromolecule has not been completely determined although it appears to be very large. Fig. 11 represents a diagram of MUC3 containing all currently elucidated domains. Other possibilities exist that may complicate this model.

The precise role of the mucin encoded by the MUC3 gene is unknown. MUC3 is expressed in the small intestine, in both goblet cells and villus columnar cells (28, 29). It does not appear to be highly expressed in the colon and may not form a significant component of colonic mucus, which may be mainly MUC2 (5, 47). There is very little indication of its size. RNA blot analysis suggests a transcript of 13 kb, but this value is at or above the limit of resolution of the agarose gels used for analysis (48). As yet there is no published information as to the size of the polypeptide. There is no suggestion of a hydrophobic membrane anchor at the C-terminal end, but the pattern of cellular expression does suggest the possibility that it may not be secreted.

The physical mapping of the rest of the MUC3 genetic region has been complicated by the lack of large genomic clones located upstream of the 51-bp tandem repeats. However, analysis of cDNAs and PCR-amplified fragments of genomic DNA indicates that this region is quite complex and contains at least one tandem array of large repeats, of which at least five have been completely or partially sequenced. Given the relative paucity of overlapping clones, it is not unreasonable to speculate that this region could contain substantially more repeats. The sequences obtained by vectorette and standard PCR from genomic DNA failed to reveal the presence of introns in the part of this region that is immediately upstream of the 51-bp tandem repeats.

The presence of two major zones of 51 bp tandem repeats as indicated by the restriction enzyme digestion patterns obtained with a variety of enzymes suggests that there may be a major genomic duplication. The facts that these zones were both detected with 3' “unique” and the large tandem repeat-containing cDNAs and that the same 1.8-kb band is detected by clone 20 and SIB172U suggest that variants of these sequences may occur in close proximity. However, these interpretations must remain speculative because of small patches of homology to the 51-bp tandem repeats. For example, each of the large tandem repeats appears to contain at least a partial 51-bp repeat unit within it, which may allow some cross hybridization. It is formally possible that there are two tandemly arrayed MUC3 genes, or a gene and a pseudogene, but because of the close proximity it seems more likely that there is a major internal duplication. There is precedent for this in the case of MUC5B, which shows four super repeats within its central exon (21), and in MUC5AC, which has at least two regions of tandem repeat and multiple copies of certain cysteine-rich regions (49).

The explanation for the very large size of the zones that contain 51-bp tandem repeats, detected with PouII and PstI, is unclear. In the case of MUC1 and MUC2, which serve as models for VNTR-containing genes, the tandem repeats are
contained within a single large exon and are not interrupted by intronic sequences. Because of their repetitive nature, they lack restriction sites for many enzymes that cut with moderate frequency in the genome, such as EcoRI, PstI, BamHI, HindIII, and HinfI. Thus, VNTR-restriction fragments obtained using enzymes that cut with moderate frequency provide a reasonable estimate of transcript size. In the case of MUC3, the total size of two haploid sets of polymorphic PvuII fragments can exceed 50 kb, and this does not even cover the whole transcript since the large tandem repeat zone contains regular PvuII sites. If in fact most of this sequence is expressed, a very large message would be required. Indeed, the largest and smallest alleles of the upper set of fragments detected with PvuII differ in size by approximately 30,000 bp. These observations possibly indicate that the VNTR zones contain intronic sequences. It is also possible that the total length of intronic sequences differ in different alleles, perhaps due to VNTR polymorphism. A precedent for this general idea is to be found in the gene encoding apolipoprotein(a) (50), though in this case introns constitute the major component of each tandem repeat unit. It is tempting to speculate that the unusual gene structure of

\[ \text{FIG. 7. Long electrophoresis on 0.5\% agarose gels of PvuII digests of DNA from six unrelated individuals. Fragment sizes were determined using Raoul molecular size standards (mobilities shown), and the blot was probed with SIB124.} \]

\[ \text{FIG. 8. Standard agarose electrophoresis of PstI and PvuII digests of two families showing hybridization of the 5'-, tandem repeat, and 3' -probes. Blots were digested and probed as indicated. Mobilities of the Raoul molecular weight markers are shown.} \]

\[ \text{FIG. 9. Short electrophoresis on 1\% agarose gels showing hybridization of the 5'-, tandem repeat, and 3' -probes. Note the hybridization of the Clone 20 and SIB172U probes to small fragments not detected by SIB124, the tandem repeat probe. Mobilities of the Raoul molecular weight markers are shown.} \]
FIG. 10. PFGE of the MUC3 gene. Panel A, K562 DNA digested with five different restriction enzymes. Single bands are observed with BsaHII and NofI, while the other enzymes exhibit two bands that are not well resolved in the lanes digested with SfiI and NcoI. Mobilities of the S. cerevisiae and a ladder size standards are shown. Panel B, DNA from five unrelated human donors was digested with StuI and subjected to PFGE and blot analysis. A single band of 200 kb (indicated by line) was observed with each sample. Both bands were probed with SIB12.

FIG. 11. Molecular model of MUC3 transcript structure. The region of 17-amino acid tandem repeats exhibits VNTR polymorphism. Sequences located upstream of the 375 amino acid tandem repeats are undetermined at present.

MUC3 adds to the difficulties encountered in obtaining large genomic clones for this region.

Acknowledgments—We thank the Center d’Etude du Polymorphisme Humain and the EUROGEM consortium for family DNA samples and genomic clones for this region. We thank the Center d’Etude du Polymorphisme Humain and the EUROGEM consortium for family DNA samples and genomic clones for this region.

REFERENCES

1. Gum, J. R. (1994) Biochem. Soc. Trans. 23, 795–799
2. Gessler, S., and Spier, A. P. (1995) Annu. Rev. Physiol. 57, 607–634
3. Neutra, M. R., and Forstner, J. F. (1987) in Physiology of the Gastrointestinal Tract (Johnson, L. R., ed) 2nd Ed., pp. 627–639, Raven Press, New York
4. Allen, A. (1981) in Physiology of the Gastrointestinal Tract (Johnson, L. R., ed) 1st Ed., pp. 627–639, Raven Press, New York
5. Carlstedt, I., Herrmann, A., Hoovenberg, H., Lindell, G., Nordman, H., Wickstrom, C., and Davies, J. R. (1995) Biochem. Soc. Trans. 23, 845–851
6. Sheehan, J. K., Thornton, D. J., Howard, M., Carlstedt, I., Corfield, A. P., and Paraskeva, C. (1996) Biochim. Biophys. Acta 1298, 1055–1069
7. Pigny, P., Guyonnet-Duperat, V., Hill, A. S., Laine, A., Van-Seuningen, I., Degand, P., Gum, J. R., Kim, Y. S., Swallow, D. M., Lagace, R. E., Byrd, J. C., Lampert, D. T., Siddiki, B., and Kim, Y. S. (1996) Biochim. Biophys. Res. Commun. 231, 407–415
8. Fox, M. F., Laihii, F., Pratt, W., Attwood, J., Gum, J. R., Kim, Y. S., and Swallow, D. M. (1992) Annu. Hum. Genet. 56, 281–287
9. Chang, S. K., Dohrmann, A. B., Bassham, C. B., Ho, S. B., Tuesta, T., Toribara, N. W., Gum, J. R., and Kim, Y. S. (1994) Gastroenterology 107, 28–36
10. Audie, J. P., Janin, A. P., Corfield, A. P., and Laine, A. (1996) Eur. J. Hum. Genet. 2, 193–252
11. Gum, J. R., Jr., Hicks, J. W., Lagace, R. E., Byrd, J. C., Toribara, N. W., Siddiki, B., Fearney, P. J., Lampert, D. T. A., and Kim, Y. S. (1991) J. Biol. Chem. 266, 9678–9686
12. Klop, L. W., Van Bens, L., and Strous, G. J. (1995) Biochem. J. 308, 831–838
13. Gum, J. R., Jr., Hicks, J. W., Lagace, R. E., Byrd, J. C., Lampert, D. T. A., Siddiki, B., and Kim, Y. S. (1996) Biochem. Biophys. Res. Commun. 221, 407–415
14. MacConaill, A., Herrmann, A., Chamberlain, H., Lindell, G., Nordman, H., Wickett, C., and Davies, J. R. (1995) Biochem. Soc. Trans. 23, 845–851
15. Sheehan, J. K., Thornton, D. J., Howard, M., Carlstedt, I., Corfield, A. P., and Paraskeva, C. (1996) Biochim. Biophys. Acta 1298, 1055–1069
16. Pigny, P., Guyonnet-Duperat, V., Hill, A. S., Laine, A., Van-Seuningen, I., Degand, P., Gum, J. R., Kim, Y. S., Swallow, D. M., Lagace, R. E., Aubert, J. P., and Laine, A. (1996) Eur. J. Hum. Genet. 23, 1698–1707
17. Sato, T. N., Qin, Y., Kozak, C. A., and Audus, K. L. (1993) Proc. Natl. Acad. Sci. U.S.A. 90, 9355–9358
18. Bell, G. I., Fong, M. N., Stempien, M. M., Wormstad, M. A., Caput, D., Ku, L. L., Urdea, M. S., Rall, L. B., and Sanchez-Pescador, R. (1986) Nucleic Acids Res. 14, 8427–8446
19. Derynck, R., Roberts, A. B., Winkler, M. E., Chen, Y. E., and Goeddel, D. V. (1984) Cell 38, 287–297
20. Grese, F. M., Muller-Pillusch, F., Geng, M., Zimmerhakl, F., Zehetner, G., Fries, H., Buchler, M., Adler, G., and Lehrach, H. (1996) Oncogene 13, 1819–1830
21. Hill, A. S., Gum, J. R., Ho, J., Pratt, W. S., Beck, S. K., Kim, Y. S., and Swallow, D. M. (1995) Cytoengen. Genet. Cell. Biol. 71, 25–26
22. Sand, B. M., Ogata, H., Lynch-Danevsky, K., deBeaumont, M., Erez, R., and Podolsky, D. K. (1995) J. Biol. Chem. 270, 9353–9361
23. Jeffrey, G. P., Oates, P. S., Wang, T. C., Babitsky, M. W., and Brand, S. J. (1994) Gastroenterology 106, 336–345
24. Hoffmann, W., and Hauser, F. (1993) Trends Biochem. Sci. 18, 239–243
25. Perez-Vilar, J., Eckhardt, A. E., and Hill, R. L. (1996) J. Biol. Chem. 271, 9845–9850
26. Toribara, N. W., Ho, S. B., Gum, J. R., Swallow, D. M., Lagace, R. E., Byrd, J. C., Fearney, P. J., Lampert, D. T. A., and Kim, Y. S. (1997) J. Biol. Chem. 272, 16398–16403
27. Ho, S. B., Niehans, G. A., Limnoff, C., Van, P. S., Chernitz, D. L., Gum, E. T., Dahiya, R., and Kim, Y. S. (1995) Cancer Res. 55, 641–651
28. Lesaffre, T., Porchet, N., Aubert, J. P., Swallow, D., Gum, J. R., Kim, Y. S., Real, F. X., and Zwiebaum, A. (1993) J. Cell Sci. 106, 771–783
29. Guyonnet-Duperat, V., Audie, J. P., Debaillieu, V., Laine, A., Bozon, M., Galiegou-Zouitina, S., Pigny, P., Degand, P., Aubert, J. P., and Porchet, N. (1995) Biochem. J. 305, 211–219
30. Laenker, C., Boerwinkle, E., Leffert, C., Rahming, T., and Hobs, H. (1991) J. Clin. Invest. 87, 2153–2161

MUC3 Human Intestinal Mucin