Genomic Organization of the 3′ Region of the Human Mucin Gene MUC5B*

(Received for publication, January 28, 1997, and in revised form, April 28, 1997)

Jean-Luc Desseyn‡, Jean-Pierre Aubert‡§, Isabelle Van Seuningen‡, Nicole Porchet‡§, and Anne Laine¶¶

From the ‡Unité 377 INSERM, Place de Verdun, 59045 Lille Cedex, and §Laboratoire de Biochimie et de Biologie Moléculaire de l’Hôpital C. Huriez, Centre Hospitalier Régional et Universitaire, 59037 Lille Cedex, France

MUC5B, mapped clustered with MUC6, MUC2, and MUC5AC to chromosome 11pter15.5, is a human mucin gene of which the genomic organization is being elucidated. We have recently published the sequence and the peptide organization of its huge central exon, 10,713 base pairs (bp) in length. We present here the genomic organization of its 3′ region, which encompasses 10,690 bp. The genomic sequence has been completely determined. The 3′ region of MUC5B is composed of 18 exons ranging in size from 32 to 781 bp, contrasting thus with the very large central exon. The sizes of the 18 introns range from 114 to 1118 bp. Some repetitive sequences were identified in four introns. The peptide deduced from the sequence of the 18 exons consists of an 808-amino acid peptide. This carboxyl-terminal region exhibits extensive sequence similarity to MUC2, MUC5AC, and von Willebrand factor, particularly the number and the positions of the cysteine residues, suggesting that this domain may be derived from a common ancestral gene. The presence in these components of a cystine knot also found in growth factors such as transforming growth factor-β is of particular interest. Moreover, one part of this peptide is identical to the 196-amino acid sequence deduced from the cDNA clone pSM2-1, which codes for a part of the high molecular weight mucin MG1 isolated from human sublingual gland. Considering the expression pattern of MUC5B and the origin of MG1, we can thus conclude that MUC5B encodes MG1.

Mucus is the layer that covers, protects, and lubricates the luminal surfaces of epithelial respiratory, gastrointestinal, and reproductive tracts. These basic properties are due to the viscous and viscoelastic properties of mucins, the major glycoprotein components of mucus. Mucins constitute a family of high molecular mass glycoproteins synthesized by the goblet cells of the epithelia and in some cases by submucosal glands (for more complete reviews, see Refs. 1–3).

Alterations of the biosynthesis of mucins affecting the protein core and/or the carbohydrate content linked to the peptide have been observed in numerous pathological situations such as various adenomas and carcinomas, inflammatory diseases such as cystic fibrosis, asthma, chronic bronchitis, or inflammatory bowel diseases (4–7). Moreover, the hypersecretion of mucins and the presence of alternating hydrophobic and hydrophilic domains in mucins have been shown to play a central role in the pathogenesis of cholesterol gallstones (8, 9).

All apomucins contain tandemly repeated sequences rich in threonine and/or serine. Due to the high carbohydrate content, the peptide moiety of mucins has been difficult to characterize. cDNA cloning has enabled researchers to approach the study of the mucins over the past decade. Today, the membrane-associated mucin MUC1 and the secreted MUC7 are the only mucins for which the full-length cDNA and the genomic organization have been reported (10–13). Both were revealed to be, in fact, small mucins. A complete cDNA of the large secreted mucin MUC2 (14–17) has been described. Partial cDNAs have been identified for the other human mucin genes that code for secreted mucins: MUC3 (18), MUC4 (19), MUC5AC (20–24), MUC5B (25), and MUC6 (26).

Four mucin genes are mapped to 11p15.5: MUC5AC, MUC5B, MUC2, and MUC6 (26–28). Recently, we have determined that the order of the four clustered 11p15.5 human mucin genes is tel-MUC6/MUC2/MUC5AC/MUC5Bcen (29). We have also established that MUC2, MUC5AC, and MUC5B have a consensus cysteine-rich domain found twice in MUC2 (16), at least four times in MUC5AC (21, 22, 24), and seven times in MUC5B (30).

MUC5B is expressed mainly in bronchus glands and also in submaxillary glands, endocervix, gall bladder, and pancreas (31–35). The structural organization of the peptide deduced from the nucleotide sequence of the central region of MUC5B has been published recently (30). The single large exon of 10,713 bp, containing all the tandem repeat domain, is, to our knowledge, the biggest described for a vertebrate gene. It codes for a 3570-amino acid peptide. Nineteen subdomains have been individualized. Most of the MUC5B subdomains show similarity to each other, creating four larger composite super-repeat units of 528 amino acids. Each super-repeat is made up of repeats consisting of an irregular repeat of 29 amino acids, one cysteine-rich subdomain (10 cysteine residues, 108 aa), and one unique sequence of 111 amino acid residues also rich in serine and threonine. The complete organization of the region downstream of the central region of the human MUC5B gene, i.e. its complete 3′ region, is reported in this paper; we present here the complete genomic nucleotide sequence, the exon-intron organization, and the full cDNA sequence coding for the carboxyl-

* This work was supported by La Ligue contre le Cancer, L’Association de Recherche contre le Cancer, and L’Association François Aupetit. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) Y10080 and Y09788.

† To whom correspondence should be addressed: INSERM U-377, 59045 Lille Cedex, France. Tel.: 33-3-20-29-88-59; Fax: 33-3-20-53-85-62; E-mail: laine@lille.inserm.fr.

1 The abbreviations used are: bp, base pair(s); aa, amino acid(s); BSM, bovine submaxillary gland mucin-like; CK, cystine knot; FIM, frog integumentary mucin; PCR, polymerase chain reaction; PSM, porcine submaxillary mucin; TGF, transforming growth factor; RACE, rapid amplification of cDNA ends; RT, reverse transcription; vWF, von Willebrand factor; ORF, open reading frame.
terminal domain of the human MUC5B apomucin. This domain stretches 808 amino acid residues and can be divided into six subdomains. The last five cysteine-rich subdomains exhibit extensive sequence similarity to MUC2, MUC5AC, and vWF (17, 22, 23, 36), particularly the number and the positions of the cysteine residues, suggesting that this domain may be derived from a common ancestral gene. Moreover, with the exception of one substitution, which does not change the coded amino acid, one part of the cDNA sequence we determined is identical to the nucleotide sequence of pSM2-1. This cDNA codes for 196 amino acids in the carboxy-terminal region of the high molecular weight mucin MG1 isolated from human sublingual gland (37). Considering the expression pattern of MUC5B (31) and the origin of MG1, we can thus conclude that MUC5B encodes MG1.

**EXPERIMENTAL PROCEDURES**

**Screening of cDNA and Genomic Libraries**—A agt11 cDNA library constructed from human tracheal mucosa was screened with rabbit antibodies raised to deglycosylated Pronase glycopeptides from bronchial mucins (38). Among the various positive clones obtained, the one designated TH71 and containing a poly(A) tail was of particular interest in the present study.

A human genomic αEMBL4 phage library was screened using hybridization with the JER57 probe (25). One positive clone, CEL5, was isolated and studied.

A human placental genomic DNA library in pWE15 cosmid provided by Stratagene was screened using the JER57 probe. Two positive clones, BEN1 and BEN2, were obtained (30). BEN2 was the useful clone in the present study.

**Oligonucleotide Primers**—Oligonucleotide primers used in PCR, RACE-PCR, RT-PCR, and sequencing experiments were synthesized by Eurogentec (Liège, Belgium). Their sequences and locations are indicated in Table I.

**3'-End Amplification of cDNAs**—The 5'-AmpliFINDER RACE kit (CLONTECH) was used to synthesize first-strand cDNA from human trachea poly(A)⁺ RNA (1 μg) obtained from CLONTECH using NAU61 as a primer (Table I), followed by ligation of the 5'-ANCHOR adapter. The PCR was then performed using the nested primer NAU67 (Table I and Fig. 1) and the 5'-ANCHOR primer. Nested PCRs involving a second or third round amplification were carried out with 1 μl of the reaction mixture obtained from each previous round of PCR as template.

**RT-PCR Amplification**—Total RNA of human gall bladder was extracted as described previously (39). Single-stranded cDNA was performed using the 1st STRAND Synthesis kit (CLONTECH), random hexamers and human trachea poly(A)⁺ mRNA (0.5 μg) (CLONTECH) or total gall bladder RNA (1 μg). PCR amplification reaction mixtures (50 μl) contain 0.3 mM dNTPs, 2.5 units of Taq DNA polymerase (Boehringer Mannheim), 15 pmol of the appropriate primers, the buffer system purchased with Taq DNA polymerase, and an aliquot of CDNA. The PCR was performed using a Perkin-Elmer Thermal Cycler 480. PCR parameters were 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, 60 °C for 1 min, and 72 °C for 2 min, followed by a final extension at 72 °C for 15 min. The amplified products were electrophoresed on a 1% SeaPlaque gel (FMC, Rockland, ME) and stained with ethidium bromide. The band was cut out, purified using Preps DNA purification resin (Promega), and subcloned into the T/A cloning vector, pMOSBlue T-vector (Amersham). Thereafter, cDNA clones were subcloned into pBluescript KS⁺ vector (Stratagene) using the restriction enzymes (Boehringer Mannheim) PstI, SacI, and/or SmaI. Subclones were sequenced as described below using either universal primers or a series of oligonucleotides specific for both strands of the inserts (Table I).

**Isolation and Sequencing of MUC5B Genomic Clones and Sequence Analyses**—Fragments of the genomic clones CEL5 and BEN2 corresponding to the region downstream of the central exon were subcloned into pBluescript KS⁺ vector as described previously (30). The double-stranded plasmid inserts were sequenced manually using the dideoxynucleotide chain termination method (40) using [α-³²P]dATP from Amersham. The clones were sequenced on both strands several times. Direct DNA sequencing on cosmids was performed as described previously (30). Computer analyses were performed using PC/GENE Software. The whole genomic sequence reported in this paper has been submitted to the EMBL Data Bank with accession number Y09788. The sequence of TH71 has been submitted to the EMBL Data Bank with accession number Y10080.

**Study of the Intron G**—To determine the exact number of repeats in the intron G, first we cut the genomic subcloned BglII-BglII fragment using ScaI and RsaI that flank the region containing these direct 59-bp repeats. The complete digestion with SmaI was obtained using 10 units/μg DNA for 3 h. The partial digestions were performed using 1 unit/μg DNA and 0.25 unit/μg DNA for 1 h. After electrophoretic separation on 1.5% agarose gel, the blot analysis was conducted using the antisense oligonucleotide NAU199 (5'- AAGAGCCAGGGGTCTGGG-G'), which had been previously radiolabeled using T4 polynucleotide kinase (Boehringer Mannheim) and [γ-³²P]ATP from Amersham.

**RESULTS AND DISCUSSION**

**Characterization of the Genomic Fragments of the 3′ Region of MUC5B**—The partial restriction maps of the genomic clones CEL5 and BEN2 were determined. Their overlapping parts present the same restriction map. The partial restriction map of the 3′ region of MUC5B is shown in Fig. 1 together with the

**FIG. 1. Restriction map of the 3′ region of MUC5B and sequencing strategy.** The fragments indicated were subcloned into pBluescript KS(+) vector. Some primers and their directions are indicated (not to scale) by horizontal arrows and their NAU numbers (their sequences are given in Table I).
overlapping fragments, which were separated and subcloned into pBluescript KS(+) vector. The fragments BamHI-SacII and PstI-SacII (in the left part of Fig. 1) contain the 3' end of the central exon. All these clones were entirely sequenced after restriction digestion and subcloning. Primer walking using specific oligonucleotides (Table I) was also performed.

3' Region of MUC5B — Several cDNA-positive clones were obtained by screening the λgt11 cDNA library using antibodies as described previously (38). The clone designated TH71 is 380 bp in length. Its sequence (Fig. 2), submitted to the EMBL data bank with accession number Y10080, revealed a poly(A) tail 16 bp downstream from a polyadenylation signal. (Table I) shows that the clone BEN2 contains the 3' end sequence of the MUC5B gene. Using the two synthesized oligonucleotides NAU61 and NAU67 chosen in this sequence, a 5'-RACE-PCR experiment was performed. After cloning of the fragment obtained, the insert of 88 bp designated RACE67 was sequenced. This sequence is identical to the 88-bp sequence determined in the TH71 clone (Fig. 2). In contrast, the first 34 nucleotides differ from the sequence of TH71. The TH71 clone, which has been found using the antibodies directed against the repeat part of the MUC5B apomucin (38), begins with a 132-bp sequence we found in the central exon. Between this sequence and the 3' end identical to the RACE67, TH71 seems to have been rearranged; moreover, the following results show that an important part of the cDNA has been lost. We will discuss these data below.

The NotI-BglII fragment from BEN2 contains two other clustered canonical polyadenylation signals, AATAAA. The first was located about 2 kilobase pairs downstream from the first polyadenylation signal and the second 298 base pairs downstream from this latter AATAAA. The significance of these two additional polyadenylation signals is not known. It will be interesting to determine if several forms of MUC5B mRNA can be transcribed by selection of alternative polyadenylation signals.

The dinucleotides TG and GT were found with oligo(T) stretches in the region downstream from the first AATAAA motif within the PstI-BglII subclone. This region, referred to as "GT cluster," is important for 3' processing of polyadenylated mRNAs (41). Moreover, the pentanucleotide CATTG was found between the AATAAA sequence and the poly(A) site addition (Fig. 3). This CATTG recognition element has been described to be related to cleavage site selection by Berget (42). The author suggested that pre-polyadenylated RNA hybridized with the AAUA structure, and with CAYUG recognition element within the U4 small nuclear ribonucleoproteins as related to cleavage site selection. Hence, MUC5B combines some common features of the 3' mRNA processing. From this nucleotide sequence, the new oligonucleotide NAU102 was synthesized to perform RT-PCR.

**RT-PCR** — Two specific overlapping cDNAs were synthesized by RT-PCR experiments. The locations of the oligonucleotides used in these experiments are indicated in Fig. 1. The oligonucleotide primer NAU151 was designed on the basis of the sequence determined for the BglII-BglII cosmid fragment (Fig. 1). This fragment hybridized with human tracheal RNA on Northern blot and probably contains coding sequences. An amplification product was obtained when the RT-PCR was performed with the two primers NAU151 and NAU102 using human tracheal first-strand cDNA as template. It was designated RT151-102 and is 2209 nucleotides in length. An other RT-PCR was then performed with the following oligonucleotide primers: NAU152, designed with the sequence of RT151-102, and NAU128, chosen in the 3' end sequence of the MUC5B central exon (30). The resultant 1166-bp amplification product, called RT128-152, and the RT151-102 were cloned into pMOS-
The 2423 nucleotides open reading frame (Fig. 3) encodes a 808-amino acid peptide rich in cysteine (10.1%) and proline (9.5%). This region was predicted to contain 62% coils structures. The rigid conformation could be essential for the oligomerization process. In fact 88% of the cysteine residues are located in or near a β turn as well as 11 out of 15 potential N-glycosylation sites. Moreover, a serumalbumin family signature with the consensus sequence YXGXXC-X has also been found between residues 658 and 682. As mucins are well known to bind various hydrophobic substances such as cholesterol, fatty acids, or bilirubin, this small region could be important in the formation of gallstones for example in which mucins have been described to be involved (8, 9, 48, 49).

As far as TH71 is concerned, we can now evaluate that more than 2000 bp have been lost in this last cDNA. We were unable to reproduce this cDNA using RT-PCR. It may be concluded that there has been a problem when producing this clone, which has been otherwise of great interest in determining the location of the polyadenylation signal in the genomic DNA.

**Deduced Amino Acid Sequence of MUC5B Carboxyl-terminal Region: Comparison with Other Proteins**—Some partial alignments with the sequence of vWF have been made by other authors, for example for MUC2 (16) and for MUC5AC-related cDNA clones, like NP3a (22) and L31 (23). Fig. 4 shows that an alignment on longer sequences can be accomplished. The conservation of nearly all of the cysteine residues and of several other amino acids of vWF and of the three 11p15.5 human mucins MUC2, MUC5AC, and MUC5B is readily apparent, suggesting a very similar tertiary structure. The comparison of the deduced carboxyl-terminal MUC5B peptide with vWF especially allows us to dissect this region into six domains: one domain called MUC11p15-type, which follows the central exon, one 56-amino acid domain with similarities to what we called MUC11p15-type, which follows the central exon, one 56-amino acid domain with similarities to what we called.
Fig. 3. Sequence of the 3' region of the MUC5B gene. The sequence shows the entire region beginning at the sequence of the oligonucleotide NAU128 used for the RT-PCR. The sequences of exons are indicated in uppercase letters, and the sequences of introns are indicated in lowercase letters. Encoded amino acids are shown in single-letter code and numbered 1–845 in the right margin. Bold letters on the right indicate the names of the introns. GC boxes are shaded. The polyadenylation signal is bold boxed. GT stretches and the pentanucleotide CATTG that may be involved in processing or polyadenylation are underlined. The position of the signal of poly(A) addition is marked by an arrow. The splice donor and acceptor sites are in bold.
Carboxyl-terminal Region of MUC5B

Fig. 3—continued
the A3uD4 domain (located between the A3 and the D4 domains in vWF), one D4-like domain, one B-like domain, one C-like domain, and one CK domain of 86 amino acid residues (see Fig. 5A).

The domain called MUC11p15-type (Fig. 5A) from aa 38 to 84 in MUC5B follows the central domain described previously (30). It shows similarities to MUC2 and MUC5AC, particularly with regard to the cysteine residues. This domain is somewhat different from the A3 domain of vWF.

The second domain, called A3uD4, also present in vWF, is 69 aa in length in MUC5B and spans aa 85–154. This domain is also found in MUC2 and L31, i.e. MUC5AC.

The vWF-D4 domain was found in MUC5AC, MUC2, and MUC5B (aa 155–533). D4 was found in zonadhesin, a sperm protein residue motif CGXG, which also exists in vitellogenin and zonadhesin. Mayadas et al. (52) showed that these motifs are similar to the amino acid sequences at the active site of disulfide isomerases that catalyze thiol protein disulfide interchanges. These vicinal cysteines may have the capacity to catalyze disulfide interchain formation, but Voorberg et al. (53) indicated that the dimerization resides in the last 151 residues for the vWF. Recently Perez-Vilar et al. (54) validated this hypothesis for PSM, showing that this apomucin can very likely form dimers between its carboxyl-terminal domains. Hence, our sequencing data suggest that the MUC5B apomucin is also able to form dimers between its carboxyl-terminal domains. The presence of disulfide-linked subunits was already predicted by Loomis et al. (46) in MG1. Moreover, Kawagishi et al. reported that MG1 contains at least two subunits (55), one of which is the salivary link component that weakly cross-reacts with antiserum to the human small intestinal link component, which contains N-linked carbohydrate (56).

Following the D4-like domain, one B-like domain of 40 aa residues was found in MUC5B (aa 534–574), MUC2, and MUC5AC instead of the three B domains defined for vWF (36). This B-like domain has not yet been found in other protein sequences.

Instead of the two C domains (C1 and C2) present in vWF, only one C-like domain was found in the three 11p15.5 mucins MUC5B, MUC5AC, and MUC2 (aa 575–740 in MUC5B). This C-like domain has been described to be more related to the C1 than to the C2 domain. In contrast to vWF, in which the homologous C1 and C2 domains arose by duplication, this duplication did not occur in FIM-B.1.
Fig. 4. Comparison of MUC5B carboxyl-terminal protein with other proteins: vWF (36), MUC2 (16), and L31 (23). Dashes indicate gaps introduced in the sequence for alignment purposes. Cysteine residues are shaded. Potential N-glycosylation sites are indicated by bold italic letters. Black down-pointing arrowheads indicate the positions of the introns of human vWF and are named according to Mancuso et al. (36). Black up-pointing arrowheads indicate the positions of MUC5B introns and are marked with the intron letter according to the Table II. Identical amino acids in vWF, MUC2, L31 (MUC5AC-39 end), and MUC5B are bold boxed. Thin boxes indicate identical amino acids in at least three proteins.
Carboxyl-terminal Region of MUC5B

A

Carboxyl-terminal Region of MUC5B

**Fig. 5.** Organization of the 3' region of MUC5B. A, schematic diagram of the carboxyl-terminal MUC5B core protein and similar proteins. B, organization of exons and introns in the 3' region of MUC5B. Exons are indicated by open boxes and numbered consecutively with 0 for the central exon (light gray box). Black box indicates the 3'-untranslated region. Introns as well as 3'-flanking sequence are indicated by lines. Each intron is named with a letter according to Table II.

(57). In fact, our genomic study shows that the C-like domain of MUC5B is more related to the C2 domain. Extremely conserved intron positions can be shown for introns 45 and L, and for introns 46 and M. Thus C1 and C2 in vWF, and C-like domains in 11p15.5 mucin genes, probably have a common ancestor domain, which has duplicated into C1 and C2 in vWF.

The last domain found in MUC5B is the CK domain (for cystine knot) from aa 741 to 826. The CK domain was also found in the 3' end of the secreted proteins MUC2, MUC5AC, FIM-B.1, and rat-Muc2 (58, 59). The CK domain exists in other secreted proteins (60–62). Eleven cysteine residues and some other amino acid residues within this CK domain are nearly invariant. Molecular modeling of the Norrie disease protein (61) predicts that this domain has a tertiary structure similar to that of transforming growth factor β (TGF-β). In TGF-β, seven cysteine residues, corresponding to cysteines 741, 764, 768, 787, 788, 818, and 820 in MUC5B, are nearly invariant (61). Crystallography studies of TGF-β2 have shown that six of these cysteine residues are closely grouped to make a rigid structure called the cystine knot (for reviews see 60, 63). Moreover, the determination of the crystal structure of dimeric nerve growth factor and platelet-derived growth factor revealed a structure similar to the one of TGF-β2. Tertiary structure similarities probably account for a strong resistance, as suggested by these authors, to heat, denaturants, and extremes of pH. The remaining cysteine residue in each monomer, corresponding in MUC5B to Cys-787, forms an additional disulfide bond that was found to link two TGF-β2 monomers into a dimer. Hence, the human mucins MUC5B, MUC5AC, and MUC2, the animal mucins PSM, BSM, rat-Muc2, and FIM-B.1 and Norrie disease protein may be members, with their 11 cysteine residues, of a new CK subfamily.

Out of the 15 consensus sequences for attachment of N-linked oligosaccharides (italic in Fig. 4), 10 sites are close to those observed in the 3' ends of MUC2, and/or MUC5AC (L31) and/or vWF. Four of them (positions aa 179, 298, 299, and 569 in MUC5B) have the same positions in the deduced peptides of the three human mucin genes mapped on chromosome 11p15.5. One site has the same position in the three mucins and in vWF (aa 2223 in vWF and aa 463 in MUC5B). In addition to the typical and expected O-glycosylation that occurs in MUC5B, it is very tempting to speculate that this apomucin, synthesized in the endoplasmic reticulum, is rapidly N-glycosylated. The polypeptide might fold to form intramolecular interactions and then dimers through intermolecular disulfide bridges within the carboxyl-terminal region. Although previous studies on bovine vWF suggested that N-glycosylation is not necessary for dimerization (64), Wagner et al. (65) more recently reported that N-linked carbohydrate addition onto human vWF is important for dimerization. In contrast, Perez-Vilar et al. (54) demonstrated that PSM dimerization is not dependent on the N-linked oligosaccharides within its carboxyl-terminal domain. Further studies on MUC5B apomucin using recombinant proteins synthesis to obtain antibodies and culture of mucous-secreting cells such as HT29-MTX in presence of tunicamycin will be required to clarify the role of N-glycosylation.

**Sequence Analyses of the Introns**—The schematic organization of the carboxyl-terminal MUC5B gene is given in Fig. 5B. No alternative splicing was found using total RNA from gall bladder or poly(A) mRNA from trachea and the following pairs of primers: NAU128/NAU152, NAU151/NAU203, NAU140/NAU219, NAU227/NAU208, and NAU233/NAU67 (Table I).

Introns A, C, E, G, I, K, L, and Q are class 1, where each intron interrupts the coding sequence between the first and second bases of the codon (66). Introns F, H, and R are class 2 (the intron interrupts the second and the third bases of the codon), and introns B, D, J, M, N, O, and P are class 0 (the intron occurs between codons). Introns B, C, E, G, J, K, L, M, and N have the same positions as the introns 33, 34, 36, 37, 38, 40/41, 45, 46, or 43/47, respectively, in vWF (Fig. 4). It must be emphasized that introns C and 34, E and 36, G and 37, K and
40 or 41, L and 45 are class 1, while introns J and 38, M and 46, and N and 43 or 47 are class 0. We can then observe that the ORFs between the symmetrical introns C and E, between E and G, between G and K, between K and L, between J and M and between M and N have flanking introns of the same phase class at both their ends. Consequently these ORFs are good candidates for exon shuffling especially when both ORF flanking introns are class 1 (67). It would be interesting to determine if MUC2, MUC5AC, and MUC5B have a common 3' end gene organization. Then it would be proposed that exon shuffling mechanisms may have played an important role in the formation of genes coding for proteins with D/B/C/CK domains, while a single ancestral gene may have evolved by successive duplications to give rise to the 11p15.5 human mucin gene family. Clearly, much work remains to be done and new data have to be collected concerning the three other 11p15.5 mucin genes MUC2, MUC5AC, and MUC6 to confirm our hypothesis; in particular, the exon-intron repartitions have to be elucidated.

In some introns, unique tandemly repeated sequences that are more or less perfect are found: 23 copies of an imperfect 20-bp repeat in intron A, 11 copies of an imperfect 10-bp repeat in intron C, 9 copies of a perfect 59-bp repeat in intron G, and 12 copies of an imperfect 20-bp repeat in intron P. Searching of the GenBank database indicated that the consensus sequences of these four distinct repeats were not identical with any registered sequence. It is striking that intron G is 75% (G+C)-rich and it is almost entirely built up of copies of a perfect 59-bp repeat, CCTGGCGGTTAGTGGGGCGGCCGGCCCGCCAGACCCCCCCCTCGTCTGAGTG. Each repeat contains one GC box binding site and one SmaI enzyme recognition site. The first copy of this repeat begins in the 3' end of the exon 6. To determine the exact number of repeats in this intron, first we cut the genomic subcloned RgII-BglII fragment using ScaI and RsaI enzymes that flank the region containing these perfect 59-bp repeats. Then we performed a complete and partial restriction digestions with SmaI (for details see “Experimental Procedures”). NAU199 is an oligonucleotide that recognizes a part of the 59-bp repeat. The results shown in Fig. 6 led us to conclude that there are nine 59-bp repeats. In a previous study where single-stranded oligonucleotides were used, we found that a nuclear factor called NF1-MUC5B (68), extracted from the colonic mucus-secreting subclone HT29-MTX, binds this GC site. This factor, with a molecular weight of 42,000, has been demonstrated not to be Sp1. Biochemical studies are currently in progress in our laboratory to characterize this nuclear factor.

In summary, we have cloned and sequenced the whole genomic 3' region of MUC5B and defined the exon-intron repartition. We have proved that this gene codes for the high molecular weight salivary mucin MG1. MUC5B is expressed essentially at high levels in acini mucous cells of salivary and respiratory submucosal glands and in epithelial cells of gall bladder, endocervix, and pancreas. This study provides the first genomic organization of the 3' region for a large size secreted gel-forming mucin gene. Our recent work showed that the central domain of MUC5B is encoded by a single large exon (10,713 bp), the largest one described to date in vertebrates. The deduced protein contains 19 subdomains. Some of them show similarity to each other, creating repeat units called super-repeats of 528 amino acid residues, which are the biggest ever determined in mucin genes. Each super-repeat comprises a 108-amino acid cysteine-rich subdomain. This last subdomain, found seven times in MUC5B, has thus been found several times in at least three of the four human mucin genes mapped to 11p15.5, (30). It seems that 11p15.5 human mucin genes are characterized by (i) a large exon encoding the repetitive domain as demonstrated for MUC5B and as suggested by Torihara et al. for MUC2 (15), (ii) the presence of Cys subdomains with 10% Cys residues (30), and (iii) a unique sequence just downstream from the repetitive domain typical of the 11p15.5 mucin genes and a cysteine-rich region, which is divided in several subdomains similar to vWF-D4, vWF-B, vWF-C, and CK domains (Fig. 4). It will be interesting to determine if the three other mucin genes MUC2, MUC5AC, and MUC6 have the same 3' end genomic organization as MUC5B. Moreover, it is clear with our previously published data (21, 29, 30) and with our present results, that MUC5AC and MUC5B are two distinct genes; therefore, it would be preferable for all authors to be precise in specifying which gene is concerned when they write MUC5.

Acknowledgments—We are grateful for the technical assistance of Michel Crépin, Evelyne Destailleur, Viviane Mortelec, and Danièle Petitprez.

Note Added in Proof—While this manuscript was in review, Nielsen, P. A., Bennett, E. P., Wandall, H. H., Therkildsen, M. H., Hannibal, J., and Clausen, H. (1997) Glycobiology 7, 413–419 identified MG1 as tracheobronchial mucin MUC5B. On the other hand, Keates, A. C., Nunes, D. P., Afshah, N. H., Troxler, R. F., and Offner, G. D. (1997) Biochem. J. 324, 295–303 published a partial genomic organization of the 3' end of MUC5B with some differences from our data.

REFERENCES
1. Gendler, S. J., and Spicer, A. P. (1995) Annu. Rev. Physiol. 57, 607–634
2. Bansil, R., Stanley, E., and LaMont, J. T. (1995) Annu. Rev. Physiol. 57, 635–657
3. Forstner, G. (1995) Annu. Rev. Physiol. 57, 585–605
4. Verma, M. (1994) Cancer Biochem. Biophys. 14, 151–162
