Molecular Cloning and Analysis of the Mouse Homologue of the Tumor-associated Mucin, MUC1, Reveals Conservation of Potential O-Glycosylation Sites, Transmembrane, and Cytoplasmic Domains and a Loss of Minisatellite-like Polymorphism*

(Received for publication, March 7, 1991)

Andrew P. Spicer‡, Gordon Parry§, Stuart Patton¶, and Sandra J. Gendler‡∥

From the ‡Molecular Epithelial Cell Biology Laboratory, Imperial Cancer Research Fund, P. O. Box 123, Lincoln’s Inn Fields, London WC2A 3PX, United Kingdom; the §Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720, and the ¶Department of Neurosciences, School of Medicine, Center for Molecular Genetics, La Jolla, California 92037.

We present here the full-length cDNA sequence and genomic structure of the mouse homologue of the tumor-associated mucin, MUC1. This mucin (previously called polymorphic epithelial mucin) is present at the apical surface of most glandular epithelial cells. The mouse gene, Muc-1, encodes an integral membrane protein with 40% of its coding capacity made up of serine, threonine, and proline, a composition typical of a highly O-glycosylated protein. The mucin core protein consists of an amino-terminal signal sequence, a tandem repeat domain encoding 16 repeats of 20–21 amino acids, and a unique sequence containing transmembrane and cytoplasmic domains. Homology with the human protein is only 34% in the tandem repeat domain, mainly showing conservation of serines and threonines, presumed sites of O-linked carbohydrate attachment. Homology rises to 87% in the transmembrane and cytoplasmic domains, suggesting that these regions may be functionally important. The pattern of expression of the mouse mucin is very similar to that of its human counterpart and accordingly the two promoter regions share homology. 74%, although previously identified potential hormone-responsive elements are not conserved. Interestingly, the mouse homologue, unlike its human counterpart does not exhibit a variable number tandem repeat polymorphism. We present evidence that suggests that the mouse gene was at one time polymorphic but has mutated away from this state.

High molecular weight mucin glycoproteins are expressed by a wide variety of epithelial tissues and are often important differentiation markers in the development of these tissues. MUC1, previously referred to as the polymorphic epithelial mucin, is a highly glycosylated membrane glycoprotein expressed by a large number of simple secretory epithelial tissues, e.g. mammary gland, pancreas, lung, fallopian tube, salivary gland, and chief cells of stomach as well as by certain carcinomas (Zotter et al., 1988) where it shows aberrant expression (Burchell et al., 1987; Girling et al., 1989; Byrd et al., 1989).

Monoclonal antibodies (mAbs)† HMFG-1, HMFG-2, and SM-3 directed to normal and/or malignant human mammary epithelial cells have been found to recognize epitopes present within this mucin (Taylor-Papadimitriou et al., 1981; Burchell et al., 1983, 1989; Gendler et al., 1988; Girling et al., 1989). Subsequently, partial cDNA clones revealed the presence of a 60-bp GC-rich tandem repeat coding for a potentially highly glycosylated 20-amino acid repeat (Gendler et al., 1987, 1988; Siddiqui et al., 1988). The high degree of polymorphism observed at the DNA, RNA, and protein level, which led to the protein initially being called the polymorphic epithelial mucin, was found to be due to differing numbers of tandem repeats, the numbers/allele ranging from 20 to 125 (Swallow et al., 1987; Gendler et al., 1988, 1990); the gene can thus be described as being an expressed variable number tandem repeat gene.

Recent cloning of the porcine submaxillary (Timp et al., 1988), human intestinal MUC2 and MUC3 (Gum et al., 1989, 1990), human tracheobronchial MUC4 (Porchet et al., 1991), and Xenopus integumentary mucin cDNAs (Hoffmann, 1988; Probst et al., 1990) has revealed the presence of tandemly repeated domains made up of 81, 23, 17, 16, 9, and 11 amino acid repeats, respectively. The presence of a tandemly repeated domain thus appears to be a characteristic feature of mucins, and this feature has been shown to give several of the mucin genes a variable number tandem repeat polymorphism (Swallow et al., 1987; Gendler et al., 1988, 1990; Griffiths et al., 1990; Hauser et al., 1990; Porchet et al., 1991).

Periodic acid-Schiff silver-stained gels identify an equivalent high molecular weight glycosylated protein, present in the milk fat globule, in a wide spectrum of mammals ranging from human to mouse (Patton et al., 1989; Patton and Patton, 1990). However, mAbs directed to the human MUC1 tandem repeat core protein have shown cross-reactivity only with the mucin of higher primates, suggesting that the sequence of the repeat unit has altered during evolution. As much as 50% of the molecular weight of the human MUC1 protein is made up of carbohydrate attached mainly to sites present within the tandem repeats by O-linked glycosylation. If the primary function of mucin tandemly repeated domains is to provide a

† 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™/EMBL Data Bank with accession number(s) M64928 and MM65132.

∥ To whom reprint requests should be addressed.

* The abbreviations used are: mAbs: monoclonal antibodies; PCR: polymerase chain reaction; GTC: guanidinium isothiocyanate; MFG: milk-fat-globule; bp, base pair(s); kb, kilobase(s); kbp, kilobase pair(s); SDS, sodium dodecyl sulfate.

§. S. Patton, personal communication.
carbohydrate scaffold, then divergence of the repeat sequence through evolution should be "allowed" within certain limits, those limits being the maintenance of potential O-glycosyla-
tion.

The MUC1 protein is located exclusively in the apical domain of the plasma membrane of highly polarized epithelial cells. Full-length human cDNA clones revealed regions encoding a 31-amino acid transmembrane and 69-amino acid cytoplasmic tail domain (Gendler et al., 1990; Lan et al., 1990; Litgenberg et al., 1990; Wreschner et al., 1990). Recent studies on apical membrane polarity, examining in particular the distribution of MUC1 in mammary epithelial cell cultures, have suggested that the protein interacts either directly or indirectly with the actin cytoskeleton, presumably by means of its cytoplasmic tail (Parry et al., 1990).

In order to investigate the function of this protein and to look at evolutionary changes in the tandem repeat, the mouse homologue has been cloned. Full-length cDNA and genomic sequencing has revealed high overall homology between the two species. In particular, the mouse sequence contains a repeat domain encoding a variable 20–21-amino acid tandem repeat with a very high potential for O-glycosylation, although homology between the human and mouse repeats is less than 40% at the protein level. Regions of highest homology between the human and mouse sequences include the 31-amino acid transmembrane and 69-amino acid cytoplasmic tail domain. This result suggests that this region may be very important functionally in its interaction with the actin cytoskeleton. The promoter region immediately upstream of the TATAA box also exhibits high homology, 74%, which may indicate that this region plays a role in the epithelial specific expression of this gene.

Southern blots indicate that the rodent MUC1 homologue is not polymorphic. However, sequence analysis within the repeat domain of the mouse MUC-1 gene suggests that at one time the rodent gene may have exhibited a variable number tandem repeat polymorphism and that this polymorphism has subsequently been lost.

MATERIALS AND METHODS

Genomic DNA Isolation and Southern Blot Analysis—Genomic DNA was prepared from T47D (human mammary carcinoma (Keydar et al., 1979), C57 MG (mouse mammary carcinoma (Vaidya et al., 1978)), and HC1 (COMMA-1D-mouse mammary epithelial (Bali et al., 1988)) cell lines with an Applied Biosystems (Foster City, CA) 340A DNA Extractor. Samples (10–15 μg) were digested with the appropriate restriction endonucleases (New England Biolabs and Northumbria Biologicals Ltd.) under conditions recommended by the manufacturer, prior to electrophoresis through a 0.7% agarose gel and subsequent hybridization/washing procedures. The human probes, pMUC7, corresponding to 500 bp of the repeat domain encoding a variable 20–21-amino acid tandem repeat with a very high potential for O-glycosylation, although homology between the human and mouse repeats is less than 40% at the protein level. Regions of highest homology between the human and mouse sequences include the 31-amino acid transmembrane and 69-amino acid cytoplasmic tail domain. This result suggests that this region may be very important functionally in its interaction with the actin cytoskeleton. The promoter region immediately upstream of the TATAA box also exhibits high homology, 74%, which may indicate that this region plays a role in the epithelial specific expression of this gene.

Southern blots indicate that the rodent MUC1 homologue is not polymorphic. However, sequence analysis within the repeat domain of the mouse MUC-1 gene suggests that at one time the rodent gene may have exhibited a variable number tandem repeat polymorphism and that this polymorphism has subsequently been lost.
Sandy, Faray, and Eday (in the Orkney Isles), Skokholm Island (off the Welsh coast), Burton-On-Trent, Taunton, John O'Groats, Belfast, and Denmark and inbred laboratory strains, C57BL X CBA, the rat, Rattus norvegicus, Mus spretus from southern Spain, the bank vole, Clethrionomys glareolus, and the short-tailed vole, Microtus agrestis. Genomic DNA was also prepared from C57 MG, HC11, and C6 (rat glial fibroblast) cell lines as previously described. Samples (10-15 µg) were digested with Tsp45I restriction endonuclease under conditions recommended by the manufacturer, prior to electrophoresis through a 1.2% agarose gel. Southern blotting and washing conditions were carried out as previously described. The mouse Muc-1 probe utilized, pMuc2TR, consisted of 550 bp of repeat and was labeled by random priming as described previously.

Milk-fat-globule (MFG) proteins were isolated as described (Patton and Huston, 1986) from human, rhesus monkey, bovine, mouse, guinea pig, cat, dog, horse, and goat milk. Approximately 5-µg samples of total MFG protein from the various species were size separated as described (Patton and Patton, 1990). Proteins were visualized according to Morrissey (1981).

DNA Sequencing—Denatured double-stranded DNA was sequenced fully in both directions by the di-deoxy chain termination method using primers directed to vector sequence or synthetic oligonucleotides. Computer analyses of DNA and amino acid sequences were performed on a VAX computer using Intelligenetics and Genetics Computer Group Software.

RESULTS

Cloning Strategy—Southern blots of mouse genomic DNA cut with various restriction endonucleases were screened with cDNA probes corresponding to the human mucin repeat domain, pMUC7, and the human mucin 3'-coding region. No cross-hybridization was observed with the repeat probe at low stringency, whereas the 3'-probe cross-hybridized at high stringency revealing a single EcoRI fragment at 10 kb and a single BamHI fragment at 15 kb (data not shown). The human 3'-probe, pGEM-16.2, was used to screen an amplified Xgt10 cDNA library constructed from poly(A)+ mouse lactating mammary gland RNA, where the mucin is known to be expressed at high levels in other animals. All the initial clones obtained represented 3'-clones stretching from the poly(A) tail to a position corresponding 3' of the repeat domain of the human cDNA sequence. The primary cDNA library was re-screened with this initial clone, and resulting positives were analyzed using a PCR approach to determine if they contained sequence further 5' of the original clone. Clones containing approximately 550 bp of further 5'-sequence were identified.

To obtain the remaining 5'-cDNA sequence, a mouse cosmid library was screened with the mouse cDNA clone, pMuc2TR, in order to obtain genomic sequence in the 5'-region from which to construct specific oligonucleotide primers to carry out PCR amplification. First-strand cDNA synthesis was carried out using C57BL mouse lactating mammary gland RNA and a synthetic antisense oligonucleotide primer directed to the furthest 5'-sequence of the cDNA clone, pMuc2. Sequence obtained from the genomic clones, corresponding to the potential translation start site, was used to synthesize a specific primer in order to amplify the product of the first-strand reaction by PCR, using this primer and the antisense primer mentioned.

3'- and Repeat cDNA Clones—Twenty-two positive clones were isolated from the amplified library, all of which were observed to contain an insert of 1.2 kb (Fig. 1B). Two of these clones were selected for sequencing, the sequences from which were found to be identical. The initial clone, when used to screen the primary library, identified five positive clones. These were analyzed by a PCR approach using synthetic oligonucleotide primers directed to both vector sequence on either side of the cloning site, and to the mouse mucin sequence previously determined. In this way it was deduced that two of the five clones contained further 5'-sequence amounting to about 550 bp (Fig. 1B). Sequencing of the 550-bp fragment revealed that it was entirely made up of a 60-63-bp degenerate repeat, the degeneracy of which increased from 5' to 3'. The repeat encoded a 20-21-amino acid repeat extremely rich in serine and threonine, >50% in some repeats. Sequence obtained from the PCR-generated clones and the clone obtained by conventional methods matched exactly.

5'-cDNA and Genomic Clones—Genomic clones were obtained from a mouse cosmId library constructed from a partial Sau3A digest of mouse genomic DNA. Six positive clones were selected and characterized, two of which appeared to be rearranged, this perhaps being due to the repetitive region. 10-kb EcoRI and 15-kb BamHI fragments were isolated and subcloned into pBS-KSI+ for sequencing. Sequence obtained from the genomic clones indicated that these fragments contain the entire coding region (Fig. 1C). Sequencing of the genomic clones enabled the synthesis of a potential translation start site oligo. First-strand cDNA synthesis using C57BL-lactating mammary gland RNA followed by PCR amplification, as described, resulted in the amplification of a diffuse band at about 150-300 bp. Restriction enzyme sites located at the ends of the primers enabled the product to be digested and cloned into pBS-KSI+ for sequencing. Six colonies were picked, three of which were identical in sequence, except for a PCR introduced point mutation in one (C to A at nucleotide 78) and three of which differed in the number of repeats (Fig. 1B). Apart from the single point mutation mentioned, all sequences from PCR-derived clones matched with both the genomic sequence and the cDNA sequence previously described.

To confirm that the three clones, 5', 3', and repeat, were part of the same transcript, Northern blots of mouse-lactating mammary gland RNA, T47D human mammary carcinoma cell line RNA, and ICRF23 human lung fibroblast cell line RNA were probed with regions of each of the clones (Fig. 1). An identical pattern of hybridization was observed in the mouse mammary RNA with a single transcript at about 2.3 kb in each case. The 3'-probe also showed weak cross-reactivity with T47D at high stringency where transcripts from the two different sized alleles were observed; this correlated well as the 3'-clones were selected by their strong cross-hybridization with the human probe pGEM-16.2.

Nucleotide Sequence and Genomic Structure—The composite DNA sequence from the 5'-PCR-derived clone, the tandem repeat clone, the genomic clones and the 3' -Agt10 clone is shown in Fig. 2. Sequence shown was determined in both directions. The predicted leader sequence was derived from sequence of the two genomic clones. Genomic sequence suggests that the transcription start site of the mouse gene occurs some 72 bp away from the translation start site, as is the case in the human mucin (Gendler et al., 1990). This is suggested by the fact that the alignment between the two sequences is good in this region and that a TATAA box is located at the same position 23 bp away (Figs. 2 and 6). Repeated attempts to precisely mapping the transcription start site using a variety of oligonucleotide primers and conditions in primer extension and nuclease protection experiments met with no success. All cDNA sequence corresponded exactly with sequence determined from genomic clones apart from the point mutation previously mentioned.

The genomic structure of the mouse homologue (Fig. 3) is very similar to that of its human counterpart, there being seven exons and five introns of similar sizes in both genes; mouse intron one, however, is significantly longer than its human equivalent. All exon/intron boundaries are conserved. A dot-matrix plot comparing the human and mouse genomic
Mouse Homologue of a Tumor-associated Mucin

sequences, performed using the COMPARE and DOTPLOT Genetics Computer Group software with a window of 21 and stringency of 14, illustrated the high overall similarity and demonstrated very well the difference between introns and exon and promoter; introns were observed to correspond to gaps in the main diagonal (Fig. 4).

Predicted Amino Acid Sequence and Composition of the Mouse MUC1 Core Protein—The deduced sequence of the mouse mucin core protein (Fig. 2) encodes an integral membrane protein with 40% of its coding capacity made up of serine, threonine, and proline, a composition typical of a protein that is highly O-glycosylated. The protein appears to consist of three distinct regions: (a) an amino-terminal region containing a hydrophobic signal sequence preceding a short stretch of unique sequence; (b) a tandem repeat region encoding 16 degenerate repeats (underlined), five of which are 21 amino acids in length, the remaining 11 repeats being 20 amino acids; and (c) a carboxy-terminal region containing unique sequence followed by a hydrophobic membrane spanning domain of 31 amino acids and a 69-amino acid cytoplasmic tail. According to the predicative method of von Heijne (1986), the signal peptide is 11 amino acids long and is cleaved between the glycine and phenylalanine residues (Fig. 5A).

The sequence of the 20–21-amino acid tandem repeat unit corresponds to what might be expected for a protein that is extensively glycosylated. On average there are 9 serine/threonine residues/repeat with eight of these being found as doublets. The predicted molecular mass for this core protein is 65 kDa, yet mouse milk fat globule proteins when run on protein gels indicate that the fully glycosylated protein is greater than 200 kDa in size (see Fig. 8). This would imply that up to 75% of the molecular weight of the fully glycosylated protein is made up of carbohydrate. As well as there being multiple potential O-linked glycosylation sites, the sequence also contains 10 possible N-linked glycosylation sites (Asn-X-Ser/Thr) in the extracellular domain, five of which are found within the last six repeats.

Homology with Human MUC1 Protein—Alignment of the amino acid sequence of the human and mouse genes revealed the most significant homology to be centered around the membrane-spanning and cytoplasmic tail domains, 87%.
Other significant areas of homology noted were the NH₂-terminal signal sequence, the serines and threonines (potential attachment sites for O-linked carbohydrate) found within the repeated region, and the potential N-linked glycosylation sites. The most significant difference between the two sequences occurred at the 5'-end where a large region of the human sequence is not represented in the mouse sequence. Fig. 5A depicts the alignment between the two amino acid sequences, and Fig. 5B summarizes homology levels over the various domains. Alignment of the mouse amino acid sequence to a human sequence containing 12 consensus repeats was carried out by a VAX computer using the Genetics Computer Group GAP program.

Promoter Analysis—Several potential hormone response elements have been identified within 500 bp of the human mucin transcription start site including potential progesterone, glucocorticoid (Lancaster et al., 1990), and estrogen (Tsarfaty et al., 1996) response elements. An alignment of the corresponding sequence of the mouse mucin promoter suggests that these potential elements are not functional. In addition, a potential enhancer sequence, described by Tsarfaty et al. (1990) within the first intron of the human MUC1 and which showed significant homology to a murine cellular enhancer was also not conserved. The two promoter regions within 500 bp of the transcription start site are well conserved, however, displaying 74% identity (Fig. 6). In particular a region immediately upstream of the TATAA box is very highly conserved, displaying close to 100% identity. This may imply that this region may be functioning in the epithelial-specific expression of this gene. No significant homology was found with any other promoter sequences in EMBL or GenBank databases.

Analysis of Polymorphism—Southern blots containing both wild and inbred rodent DNA samples revealed that all individuals possessed Muc-1 alleles of the same length (16 repeats) (Fig. 7). The restriction endonuclease TaqI, from a knowledge of the sequence data, should yield a fragment of approximately 1.6 kb, and indeed this fragment was observed in all mouse, M. musculus, samples as well as Mus spretus, the rat and two other rodent species, the bank vole and short-tailed vole. This fragment has been observed in a sample of greater than 40 wild rodents (data not shown), suggesting that the Muc-1 gene may exhibit a fixed length in the rodents in general. This appears to be in contrast to the situation observed for this locus in a large number of other mammalian species. As the polymorphic region of this gene occurs in a region coding for protein, it is possible to visualize the length polymorphism at the protein level as well as at the nucleic acid level. Using milk-fat-globule preparations and silver-stained SDS-protein gels, it was possible to visualize the equivalent mammary protein from a variety of mammals, including human, rhesus monkey, bovine, horse, goat, dog, cat, guinea pig, and mouse (Fig. 8). These studies indicated that all species considered, apart from mouse and guinea pig (members of two rodent subgroups), appear to be polymorphic at this locus, i.e. exhibited two protein bands as opposed to a single band.

Detailed sequence analysis of the mouse Muc-1 repeat region indicated that on average repeats shared 75% homology, as opposed to 97–100% in the human MUC1 gene (GenBank).
function of the protein. It is feasible that this region of the protein is binding to a link protein which in turn binds to the actin cytoskeleton. Other short stretches of high homology occur between the repeat and transmembrane/cyttoplasmic tail region (see Fig. 5). Analysis of the full-length cDNA revealed the presence of what can be described as a degenerate tandem repeat domain, repeats being 60 or 63 bp, which would allow for up to half the amino acids to be O-glycosylated (the serines and threonines). The predicted size of the mouse mucin core protein is 65 kDa yet SDS-protein gels indicate that the mature glycosylated protein present in mouse milk is greater than 200 kDa, suggesting that up to 75% of the molecular mass of the mature protein is made up of carbohydrate. Alignment between the human and mouse repeat domains revealed significant homologies to occur at the serine/threonine doublets, the prolines, the single histidine, and the central single threonine found in the human PDTRP motif (Fig. 5, A and B). From the deduced mouse mucin amino acid sequence and the alignment with the human sequence we can see why mAbs HMFG-2 and SM-3 directed to the repeat core protein showed no cross-reactivity with the mouse mucin; the epitopes for these two mAbs being DTR and PDTRP, respectively (Burchell et al., 1989) (see Fig. 5, A and B); the 2 charged residues contained within these motifs, glutamate (D) and arginine (R), are not conserved, these residues becoming uncharged alanine (A) and serine (S) or threonine (T) in the mouse.

The 5'-portion of the mouse cDNA shows significant differences from the human 5'-cDNA (see Fig. 5, A and B). Where the human 5'-region encodes a hydrophobic signal peptide, a short stretch of unique sequence followed by three degenerate repeats, the mouse sequence codes for a signal peptide, which shows high homology with that of the human, and a much shorter stretch of unique sequence which has no significant homology with that region of the human cDNA. The splice site at the end of the mouse mucin first intron appears to correspond to the alternate splice site reported for the human mucin by several groups (Ligtenberg et al., 1990; Wreschner et al., 1990). However, our sequence suggests that an alternative splicing mechanism does not operate in the mouse homologue as there does not appear to be any potential alternative splice acceptor within the next 20–40 bp which would result in the reading frame being maintained.

In addition to the many potential O-glycosylation sites, the deduced amino acid sequence we describe contains 10 potential N-glycosylation sites, five of which are found within the repeat domain. This number is compared with the five found in the human mucin. Four of these five sites are precisely conserved between the two sequences, the other one having a potential site in the mouse sequence located within 1 amino acid residue of its position in the human sequence.

The tandem repeat domain in the human MUC1 gene shows allelic variations in length which result in such a high degree of polymorphism that the sequence can be considered to be a variable number tandem repeat locus (Swallow et al., 1987). However, when the mouse repeat probe was used to investi-

Fig. 3. Genomic structure of the mouse Muc-1 gene. The seven exons and six introns span approximately 4.4 kb. I, 33-bp signal sequence; II, 93-bp transmembrane signal; III, 16 repeats, 60–63 bp.

Fig. 4. Dot-matrix plot showing the homology between the human and mouse mucin genomic sequences. A straight line passing through the diagonal would be expected if there was a perfect match between the two sequences. Lines parallel to the main diagonal are indicative of repetitive domains. From this plot the various levels of homology can be clearly seen. In particular, regions corresponding to intron sequences can be seen as gaps in the diagonal (Intron VI has been deleted for the purposes of this plot). Analysis was performed using Genetics Computer Group Software and the Compare and Dotplot programs, with a Window:21 and Stringency:14.0. Axes are in base pairs.
## Mouse Homologue of a Tumor-associated Mucin

### A. MOUSE 1 MTPGIRAPFLLLLALSKGF

| 21 |
|---|

HUMAN 1 MTPGQSEFILLLTVTVTGVCGNASSTPGGEKETSATGKSSTVPSSTTE

| 50 |

### Panel A

Comparison of human and mouse MUC1 amino acid sequence. Conserved amino acids are represented by vertical bars and gaps by dots. The signal peptides, potential N-glycosylation sites and the membrane-spanning domain are underlined. Arrows indicate the predicted signal peptide cleavage sites obtained by the predective method of von Heyne (1986).

### Panel B

Bar diagram to summarize the various levels of homology for the respective regions.

---

**Fig. 5.** Panel A, comparison of human and mouse MUC1 amino acid sequence. Conserved amino acids are represented by vertical bars and gaps by dots. The signal peptides, potential N-glycosylation sites and the membrane-spanning domain are underlined. Arrows indicate the predicted signal peptide cleavage sites obtained by the predictive method of von Heyne (1986). Panel B, bar diagram to summarize the various levels of homology for both the DNA and protein level for the respective regions.
**Mouse Homologue of a Tumor-associated Mucin**

**FIG. 6. Sequence of the promoter region within 500 bp of the mouse mucin transcription start site.** Asterisks (*) indicate nucleotides that are conserved in the human promoter region; dashes indicate missing bases in the homologue. It is of interest to note that two sequences present within the human mucin promoter previously described as being potential response elements for progesterone and estrogen are not conserved. Twelve extra bases (~414 to ~403) are present in the mouse promoter corresponding to the position of the potential progesterone response element whereas the potential estrogen response element has been deleted (~311). The highly conserved sequence located between two potential Sp1-binding sites has been underlined.

**FIG. 7. Variation at the mouse Muc-1 locus.** 10–15 µg cell line or mouse tail genomic DNA was digested with TaqI restriction endonuclease, treated as described under “Materials and Methods” and hybridized to mouse Muc-1 probe pMuc2TR. In all cases a single hybridizing fragment can be seen migrating at 1.6 kbp. This fragment was also observed in greater than 40 wild mouse samples (data not shown), indicating that the mouse Muc-1 gene is not polymorphic. **Panel** is a composite of two autoradiographs. Alignment of all bands was justified as both gels contained the samples C57 MG, HC11, and C6 at each side.

Gate polymorphisms occurring in the mouse, none were found in either laboratory or wild mice (Fig. 7). These results indicate that the polymorphic nature of the gene has been lost in the mouse and other rodents. Milk-fat-globule proteins from a variety of other mammalian species, when run on SDS-protein gels and silver stained, identified an equivalent highly glycosylated protein migrating in the size range >180 kDa. Such gels indicated that species such as chimpanzee, rhesus monkey, horse, cow, goat, dog, and cat are polymorphic with respect to this protein (Patton et al., 1985). However, in good agreement with our southern blot data, the rodents appeared to be non-polymorphic (Fig. 8). In order to confirm that the proteins we were observing were indeed homologous, the polyclonal antibody, CT-1, directed to the last 17 amino acids of the human MUC1 cytoplasmic tail was utilized. However, although this antibody reacted very strongly against the equivalent mucin from a variety of other species on sections, it does not blot well (even with the human mucin).

Sequence comparison within the mouse mucin repeat domain revealed the presence of a degenerate repeat with on average 15/60 mismatches, or approximately 75% homology between repeats (Fig. 9). This is in contrast to the human MUC1 gene in which repeat homology levels range from 97 to 100% (Gendler et al., 1987, 1988, 1990; Siddiqui et al., 1988; Abe et al., 1989; Ligtenberg et al., 1990; Wreschner et al., 1990; Lan et al., 1990). The generation of new length minisatellite alleles is based upon misalignment of repeat arrays followed by an unequal exchange event (Jarman and Wells, 1989). This is now thought most likely to take place during sister-chromatid exchange. In order for repeat arrays to recognize each other, a minimum level of repeat homology is required, and in general the most polymorphic minisatellite loci (of which the human MUC1 gene is one) are those with the most precise repeats (Jeffreys et al., 1985).

Although the mouse Muc-1 gene is not polymorphic, sequence comparison of repeats provides very strong evidence for the gene having at one time undergone some kind of unequal exchange event/s. For instance, five of the 16 repeats possess an extra codon, in each case at the same position within the repeat (Fig. 9), and several copies of this length variant could only have been generated by a process of duplication. Indeed, one such duplication event can be quite finely localized to have occurred between the ends of ancestral repeats numbers one and six resulting in a duplication of five repeats (Fig. 10). From the number of mismatches over this region, eight out of 204 (or extrapolating, approximately 40/kilobase pair), we estimate this duplication occurred approximately 8–10 million years ago (assuming an average mutation rate of approximately 1-amino acid substitution/kbp every 200,000 years (Alberts et al., 1989)). It may be that the accumulation of five repeats containing an extra codon exerted pairing constraints on any subsequent misaligned duplexes and thus reduced the rate of unequal exchange to such an extent that base substitutions accumulated and the level of repeat homology dropped to a point where polymorphism was lost. The rate of nucleotide substitution in the rodents has been found to be significantly higher than that

---

3 L. Pemberton, manuscript in preparation.
Mouse Homologue of a Tumor-associated Mucin

| Codon Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|--------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| DNA Consensus | GCC TCC ACC AGC A/TCT GTC | | | | | | | | | | | | | | | | | | | |
| Repeat No. 1 | -CA -G -T -C | | | | | | | | | | | | | | | | | | | |
| Repeat No. 2 | --- --- --- --- --- | | | | | | | | | | | | | | | | | | | |
| Repeat No. 3 | --- --- --- --- --- | | | | | | | | | | | | | | | | | | | |
| Repeat No. 4 | A-T G-A | | | | | | | | | | | | | | | | | | | |
| Repeat No. 5 | --- --- --- --- --- A -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 6 | --- --- --- --- --- A -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 7 | --- --- --- --- --- A -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 8 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 9 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 10 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 11 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 12 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 13 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 14 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 15 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |
| Repeat No. 16 | --- --- --- --- --- G -T | | | | | | | | | | | | | | | | | | |

Mouse RepeatAA Consensus

| Residue Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| AA Consensus   | DST | | | | | | | | | | | | | | | | | | | |

---

FIG. 9. Alignment of mouse Muc-1 repeats with the derived consensus repeats at both the DNA and protein level. Dashed lines indicate a conserved base or amino acid whereas a differing base or amino acid is shown. It can be seen that repeats 4-6 and 9-10 possess an extra codon at the same position and that homology decreases markedly in the last six of the 16 repeats. Overall, homology to the derived consensus rarely exceeds 90% at the DNA level. Comparison of any one of the 16 repeats with all others gives an average 15 mismatches over 60 bp.

---

FIG. 10. A diagrammatic representation of the misalignment/unequal exchange event that is proposed to have taken place to generate the modern mouse Muc-1 gene. The misalignment is proposed to have been followed by a recombination event between ancestral repeats number 1 and 6 such that repeats 2-5 were tandemly duplicated. All sites for the restriction endonucleases, M (MspI), T (TagI), H (HinfI), S (SacI), and P (PstI) are shown. Solid boxes indicate identical repeats, striped boxes represent very nearly identical repeats, 21s indicate those repeats with 21 amino acid residues as opposed to 20. Alignment between the modern sequences stretching from the end of repeat number 1 to the end of repeat number 4 and from the end of repeat number 6 to the end of repeat number 9 indicates very high homology with only eight mismatches over 204 bps whereas alignment of any one repeat with another yields an average 15 mismatches over 60 bps.
observed in humans or bovine (Wu and Li, 1985), but whether or not this is a factor in the rodent gene losing polymorphism it is as yet unknown.

In other large structures with a high content of O-linked carbohydrate, exact repeats of short stretches of amino acids have also been found to occur. This is found to be the case for the human intestinal mucin, MUC2, (Gum et al., 1989) the porcine submaxillary gland mucin (Timpoe et al., 1988) and the two Xenopus integumentary mucins (Hoffmann, 1988; Probst et al., 1990). These proteins show a large variation in size, and in the case of the human intestinal and Xenopus integumentary mucins this has been found to be due to a repeat polymorphism similar to that observed for the human MUC1 gene (Griffiths et al., 1990; Hauser et al., 1990). Gendler et al. (1990) suggest that variations in size mean that length is not crucial to the function, but rather that the core exists in an extended form as a scaffold for the attachment of O-linked carbohydrate. We hypothesized that one of the factors that might be acting to regulate the evolutionary divergence of mucin repeat sequences is the maintenance of these potential O-glycosylation sites. The amino acid alignment of the human and mouse mucin repeat domains suggests that this may indeed be a factor with greater than 80% of the conserved amino acids in the repeat domain being serine/threonine or proline. Maintenance of the potential O-glycosylation sites within the repeats suggests that the attached carbohydrate side chains are an important functional part of the external domain of this protein. This idea is reinforced by the observation that where antibodies directed to the human mucin core protein show no cross-reactivity with the mouse mammary mucin, antibodies directed to the carbohydrate moieties of the human mammary mucin (Moss et al., 1988) do cross-react. The recently cloned human pancreatic mucin (Lan et al., 1990) which is identical in sequence to MUC1 has been shown to be differently glycosylated as compared with the mammary mucin (Khorrarni et al., 1989). From these different lines of evidence we can envision a system in which the external repeat domain is glycosylated in a manner specific to its function or the glycosylation machinery in that particular tissue and the transmembrane and cytoplasmic tail domains act to localize and maintain the specific membrane distribution of the mucin in the epithelial cells by way of an interaction with the actin cytoskeleton.

Northern blots and antibody staining with antibodies directed to the human MUC1 cytoplasmic tail indicate that the pattern of expression of the mouse mucin is very similar to that observed in the human. An analysis of the two promoter regions within 500 bp of the TATAA box indicated a very high level of similarity, greater than 70% (Fig. 6). However, potential responsive elements described for this region of the human promoter do not appear to be conserved, except for potential binding sites for Sp1. In particular, the potential estrogen response element identified by Tsarfaty et al., (1990) is absent as is the potential sequence in the first intron which shows homology to a murine cellular enhancer. Potential progesterone-responsive elements and glucocorticoid-responsive elements described by Lancaster et al., (1990) are also not conserved in the mouse homologue. However, a region immediately upstream of the TATAA box located between two potential Sp1-binding sites (−92 to −34) was found to be almost 100% conserved. A search of the KeyBank promoter and enhancer element sequences found no significant matches within this region. The possibility therefore exists that this region may be playing a role in the tissue-specific expression of this gene.

Clones of the mouse homologue of the tumor-associated mucin, MUC1, will allow us to investigate the expression of this gene during embryonic and mammary gland development using in situ hybridization and by immunohistochemistry and it is hoped will also allow us to elucidate the function of this protein through gene-targeting experiments in mouse embryonic stem cells.

Acknowledgments—We thank Trevor Duhig, Nigel Peat, Rosalind Smith, and Vern Pantl for helpful discussion and technical advice, Dr. Malcolm Parker and Dr. Peter Goodfellow for critical reading of the manuscript, Dr. Dimitris Kioussis for the mouse genomic library, Dr. Alistair Lammie for the kind supply of the mouse genomic library, and in particular Peter King for the wild rodent samples and Dr. Ian Goldsmith for the synthesis of primers used in both sequencing and PCR.

REFERENCES

Abe, M., Siddiqui, J., and Kufe, D. (1989) Biochem. Biophys. Res. Commun. 165, 644–646
Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1989) Molecular Biology of the Cell, 2nd Ed., pp. 220–221, 599, Garland Publishing Inc., New York
Bali, R. K., Friis, R. R., Schönenberger, C. A., Doppler, W., and Cerny, V. (1987) EMBO J. 6, 2093–2096
Burchell, J., Durbin, H., and Taylor-Papadimitriou, J. (1983) J. Immunol. 131, 508–513
Burchell, J., Taylor-Papadimitriou, J., Boshell, M., Gendler, S., and Duhig, T. (1989) Int. J. Cancer 44, 691–696
Byrd, J. C., Lampert, D. T. A., Siddiqui, B., Kuan, S-F, Erikson, R., Twitchit, S. H., and Kim, Y. S. (1989) Biochem. J. 261, 617–625
Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) Biochemistry 18, 5294–5299
Chirgwin, G. M., and Gilbert, W. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 1991–1995
Feinstein, A. P., and Vogelstein, B. (1984) Anal. Biochem. 137, 266–267
Gendler, S. J., Burchell, J. M., Duhig, T., Lampert, D. R., White, R., Parker, M., and Taylor-Papadimitriou, J. (1987) Proc. Natl. Acad. Sci. U. S. A. 84, 6060–6064
Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J., and Burchell, J. (1988) J. Biol. Chem. 263, 12820–12823
Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Laiini, E-N., and Wilson, D. (1990) J. Biol. Chem. 265, 15286–15293
Girling, A., Bartkova, J., Burchell, J., Gendler, S., Gillett, C., and Taylor-Papadimitriou, J. (1990) Int. J. Cancer 43, 1072–1076
Griffiths, B., Matthews, D. J., West, L., Attwood, J., Povey, S., Swallow, D. M., Gum, J. R., and Kim, Y. S. (1990) Ann. Hum. Genet. 54, 277–285
Gum, J. R., Byrd, J. C., Hicks, J. W., Turbinah, N. W., Lampert, D. T. A., and Kim, Y. S. (1989) J. Biol. Chem. 264, 6480–6487
Gum, J. R., Hicks, J. W., Swallow, D. M., Lagueca, R. L., Byrd, J. C., Lampert, D. T. A., Siddiki, B., and Kim, Y. S. (1990) Biochem. Biophys. Res. Commun. 171, 407–415
Hauser, F., Gertzen, E. M., and Hoffmann, W. (1990) Exp. Cell Res. 189, 157–162
Hoffman, W. (1988) J. Biol. Chem. 263, 7686–7690
Hogan, B., Constantini, F., and Lacy, E. (1986) Manipulating the Mouse Embryo: A Laboratory Manual, pp. 174–176, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
Jarman, A. P., and Wells, R. A. (1989) Trends Genet. 5, 367–371
Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985) Nature 314, 67–73
Keydar, I., Chen, L., Karby, S., Weiss, F. R., Delarea, J., Radu, M., Chaitcik, S., and Brenner, H. J. (1979) Eur. J. Cancer 15, 659–670
Khorrarni, A., Lan, M. S., Metzgar, R. S., and Kaufman, B. (1989) Biochemistry J. 6, 428
Kioussis, D., Wilson, F., Daniels, C., Levoton, C., Taverne, J., and Playfair, J. H. L. (1987) EMBO J. 6, 355–361
Lan. M. S., Batra, S. K., Qi, W-N., Metzgar, R. S., Hollingsworth, M. A. (1990) J. Biol. Chem. 265, 15294–15299
Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990) Biochem. Biophys. Res. Commun. 173, 1019–1029
Ligtenberg, M. J. L., Vos, H. L., Gennisen, A. M. C., and Hilkeins,
Mouse Homologue of a Tumor-associated Mucin

J. (1990) J. Biol. Chem. 265, 5573–5578
Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
Morrissette, J. (1981) Anal. Biochem. 117, 307–310
Moss, L., Greenwalt, D., Cullen, B., Dinh, N., Ranish, R., and Parry, G. (1988) J. Cell. Physiol. 137, 310–320
Parry, G., Beck, J. C., Moss, L., Bartley, J., and Ojakian, G. K. (1990) Exp. Cell Res. 188, 302–311
Patton, S., and Parry, G. E. (1986) Lipids 21, 170–174
Patton, S., Huston, G. E., Jennes, R., and Vaucher, Y. (1989) Biochim. Biophys. Acta 980, 333–338
Patton, S., and Patton, R. S. (1990) J. Dairy Sci. 73, 3567–3574
Porchet, N., Cong, N-V., Dufosse, J., Audie, J. P., Guyonnet-Duperat, V., Gross, M. S., Denis, C., Degand, P., Bernheim, A., and Aubert, J. P. (1991) Biochem. Biophys. Res. Commun. 175, 414–422
Probst, J. C., Gertzen, E. M., and Hoffmann, W. (1990) Biochemistry 29, 6240–6244
Schimizu, M., and Yamauchi, K. (1982) J. Biochem. (Tokyo) 91, 515–519
Siddiqui, J., Abe, M., Hayas, D., Shani, E., Yunis, E., and Kufe, D. (1988) Proc. Natl. Acad. Sci. U. S. A. 85, 2329–2333
Stubbs, J. D., Lekutis, C., Singer, K. L., Bui, A., Yuzuki, D., Srini-vasan, U., and Parry, G. (1990) Proc. Natl. Acad. Sci. U. S. A. 87, 8417–8421
Swallow, D. M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J., and Bramwell, M. (1987) Nature 328, 82–84
Taylor-Papadimitriou, J., Peterson, J. A., Arkle, J., Burchell, J., Ceriani, R. L., and Bodmer, W. F. (1981) Int. J. Cancer 28, 17–21
Timpte, C. S., Eckhardt, A. E., Abernethy, J. L., and Hill, R. L. (1988) J. Biol. Chem. 263, 1081–1088
Tsarfaty, I., Hareuveni, M., Horev, J., Zaretsky, I., Weiss, M., Jeltsch, J. M., Garnier, J. M., Lathe, R., Keydar, I., and Wreschner, D. H. (1990) Gene (Amst.) 93, 313–318
Vaidya, A. B., Lasfargues, R. Y., Sheffield, J. B., and Coutinho, W. G. (1978) Virology 90, 12–22
von Heijne, G. (1986) Nucleic Acids Res. 14, 4683–4691
Wreschner, D. H., Hareuveni, M., Tsarfaty, I., Smorodinsky, N., Horev, J., Zaretsky, J., Kotkes, P., Weiss, M., Lathe, R., Dion, A., and Keydar, I. (1990) Eur. J. Biochem. 189, 463–473
Wu, C-L., and Li, W-H. (1985) Proc. Natl. Acad. Sci. U. S. A. 82, 1741–1745
Zotter, S., Hageman, P. C., Lossnitzer, A., Mooi, W. J., and Hilgers, J. (1988) Cancer Res. 11–12, 55–100