Exon deletion in the MSLN gene encoding MPF/mesothelin precursor protein during Laurasiatherian mammal evolution

DONG SEON KIM and YOONSOO HAHN
Department of Life Science (BK21 Program) and Research Center for Biomolecules and Biosystems, Chung-Ang University, Seoul 156-756, Republic of Korea

Received September 17, 2010; Accepted November 19, 2010

DOI: 10.3892/ijmm.2010.573

Abstract. Mesothelin is a cell surface glycoprotein that is present in normal mesothelial cells and is highly expressed in several human cancers, including mesotheliomas and ovarian, pancreatic, and lung cancers. The human mesothelin gene (MSLN) encodes a precursor protein, which is processed into 2 mature polypeptides: the N-terminal soluble megakaryocyte-potentiating factor (MPF) and the C-terminal membrane-bound mesothelin which functions as a cell adhesion molecule. In this study, we report the identification and sequence comparison of the MSLN genes in various mammalian species. We found that multiple exon deletion occurred in the Laurasiatherian MSLN genes including 6 exons in the cow, pig, horse, cat, dog, and panda genes and 8 exons in the hedgehog gene. The genomic deletion did not change the open reading frame of the resulting Laurasiatherian MSLN genes, producing internally deleted precursor proteins. The modified precursor was still able to produce the intact cell surface mesothelin protein but would not confer the MPF activity. Genomic sequence comparison showed that a breakage and rejoining event of ancestral introns 2 and 8 was responsible for the deletion. The present findings support that exon deletion is one of the molecular mechanisms underlying gene evolution in mammalian genomes.

Introduction

Mesothelin is a glycosylated cell surface protein that is present in normal mesothelial cells and that is abundantly overexpressed in several human cancers, including malignant mesotheliomas and ovarian, pancreatic, and lung cancers (1-3). Mesothelin has been considered the most promising target for immunotherapeutic treatment of mesotheliomas and other mesothelin-presenting cancers (4). The precursor protein for mesothelin, which is encoded by the mesothelin gene (MSLN), is synthesized as a 622-aa polypeptide with a calculated molecular mass of about 68 kDa. The precursor protein undergoes several post-translational modifications, including N-glycosylations, proteolytic cleavages, and a glypiation [attachment of a glycosylphosphatidylinositol (GPI) moiety] (5). A site-specific proteolytic cleavage of the precursor protein by a furin-like convertase produces 2 mature forms, the N-terminal 31-kDa soluble fragment known as the megakaryocyte-potentiating factor (MPF), which is released from the cell, and the C-terminal 40-kDa mesothelin polypeptide, which is bound to the cytoplasmic membrane via the GPI anchor (2,6). The secreted N-terminal MPF protein functions as a cytokine that can stimulate megakaryocyte colony formation in mouse bone marrow culture (7). Recently, the MPF protein has been reported to suppress cell death (8). The cell surface C-terminal mesothelin can mediate heterotypic cell adhesion by binding to the membrane-associated mucin CA125/MUC16 and may play a role in tumor metastasis (9,10). The MSLN knockout mice do not show any detectable phenotypic abnormalities, and both male and female mutant mice reproduce normally, suggesting that MPF and mesothelin proteins are not essential for growth or reproduction in mice (11).

The mesothelin precursor protein is closely homologous to the hypothetical protein MSLNL (MSLN-like, also known as MPFL, for megakaryocyte-potentiating factor-like), the function of which is unknown. The human MSLN and MSLNL genes are located on chromosome 16p13.3 in opposing orientations and share a similar genomic organization, suggesting that these 2 genes were generated from a common ancestral gene via duplication (12). These 2 proteins are remotely related to stereocilin and otoancorin, which are involved in mechanotransduction in the mammalian inner ear (13,14). Structural analyses have shown that the mesothelin precursor, MSLNL, stereocilin, and otoancorin have superhelical structures with ARM-type repeats that function to bind the carbohydrate moieties of extracellular glycoproteins (10,14). Each of the human, mouse, and rat MSLN genes is composed of 16 coding exons, and they all share a virtually identical genomic structure, except that the human MSLN has an additional 5’ untranslated (UTR) exon. Interestingly, the cow MSLN gene has only 10 coding exons, missing the 6 exons corresponding to the human coding exons 3-8, suggesting a genomic deletion in the cow genome. A genomic deletion that removes 1 or more exons generally causes...
inactivation or malfunction of the affected genes, often resulting in genetic diseases in humans (15). In some cases during human evolution, gene inactivation due to exon deletions occurred without any obvious harmful effects and actually might have given rise to advantageous phenotypic traits (16,17). In the case of the cow MSLN gene, the putative genomic deletion does not disrupt the open reading frame (ORF). The cow MSLN mRNA produces biologically functional proteins with possibly altered activity compared to that of the human MPF/mesothelin precursor. In this study, we identify and compare MSLN genes from various mammalian species by analyzing the genome assemblies and whole genome shotgun (WGS) contigs. The structural evolution of MSLN genes during mammalian radiation and its possible functional consequence are discussed.

**Materials and methods**

**Identification of MSLN genes from various mammalian species.** The National Center for Biotechnology Information (NCBI) protein database (http://www.ncbi.nlm.nih.gov/) was searched using the BLASTP program with the human MPF/mesothelin precursor protein sequence in order to collect mammalian orthologs. The genome assemblies of mammalian species available at the University of California Santa Cruz (UCSC) Genome Browser Database (http://genome.ucsc.edu/) were searched using the BLAT program in order to identify additional mammalian MSLN genes and to elucidate their genomic organizations. The genome assemblies searched include those of humans (Homo sapiens, hg18), chimpanzees (Pan troglodytes, panTro2); rhesus macaques (Macaca mulatta, rheMac2); mice (Mus musculus, mm9); rats (Rattus norvegicus, rn4); cats (Felis catus, felCat4); pandas (Ailuropoda melanoleuca, ailMel1); dogs (Canis lupus familiaris, canFam2); horses (Equus caballus, equCab2); pigs (Sus scrofa, susScr2); cows (Bos taurus, bosTau4) and chickens (Gallus gallus, galGal3). Additionally, the whole genome shotgun (WGS) contigs for the European hedgehog (Erinaceus europaeus) were searched using BLASTN at the NCBI. Exons were predicted through the sequence comparison between each of the human exons and the target genome assembly or WGS contig. Exons were concatenated into a virtual cDNA contig to deduce amino acid sequences.

**Bioinformatics analyses.** Multiple sequence alignments were generated using MUSCLE (http://www.drive5.com/muscle/) and were decorated using the BOXSHADE web server (http://www.ch.embnet.org/software/BOX_form.html). The SignalP server (http://www.cbs.dtu.dk/services/SignalP/) was used to predict signal peptides, and the GPI-anchor signal was identified using the GPI-SOM server (http://gpi.unibe.ch/). Annotated features of the human mesothelin precursor were obtained from the UniProt database (http://www.uniprot.org/uniprot/Q13421), and genomic structures of MSLN genes were examined by aligning the cDNAs and the corresponding genomic sequences using SIM4 (http://globin.cse.psu.edu/html/docs/sim4.html). Dot-plot analyses were conducted using DOTTER (http://somhammer.sbc.su.se/Dotter.html).

**Results and Discussion**

**Loss of 6 exons in the cow MSLN gene.** Currently, the MSLN gene RefSeq transcript and genomic structure have been reported in humans (RefSeq mRNA accession numbers NM_005823, NM_013404, and NM_001177355), mice (NM_018857), rats (NM_031658), and cows (NM_001100374). The MPF/mesothelin precursor proteins encoded by the human, mouse, and rat MSLN genes have amino acid lengths of 622, 625, and 625, respectively. However, the cow MSLN gene encodes a 403-aa protein. Each of the human, mouse, and rat MSLN genes has 16 coding exons, and they share a nearly identical exon/intron organization, except that the human MSLN has 1
more exon in its 5' UTR. In contrast, the cow MSLN gene is composed of only 10 coding exons; the 6 exons corresponding to the human coding exons 3-8 are not present.

A dot-plot analysis of the human and cow MSLN genomic sequences confirmed that the cow gene has fewer exons, suggesting that a segmental deletion occurred in the cow genome (Fig. 1). Alternatively, it is possible that an insertion occurred in the genome of an ancestor of humans, mice, and rats, introducing the additional 6 exons into the sequences of these species. To determine which scenario was more probable, we analyzed the MSLNL gene, a paralog of the MSLN gene, which was generated via a duplication of an ancestral gene within the mammalian genome. The MSLN/MSLNL gene duplication event seems to have occurred early in the vertebrate evolution, since the chicken genome also contains the 2 genes, MSLN (accession no. XM_414835) and MSLNL (XM_001234086) on chromosome 14 in opposing orientations. A comparison of the genomic organizations of the human and cow MSLN and MSLNL genes demonstrated that the 6 exons were part of the ancestral MSLN gene, confirming that the cow MSLN gene lost these 6 coding exons during evolution.

A comparison of the genomic organizations of the human and cow MSLN and MSLNL genes demonstrated that the 6 exons were part of the ancestral MSLN gene, confirming that the cow MSLN gene lost these 6 coding exons during evolution.

Figure 2. Multiple alignments of mesothelin precursor protein sequences. The RefSeq and predicted mesothelin precursor protein sequences from representative mammalian species were aligned. Just above the sequences are the start positions of each exon of the human MSLN gene, marked by the letter e followed by the exon number and preceded by the phase of the preceding intron. At the top are the domains (signal peptide, MPF, mesothelin, and GPI anchor signal) of the human MPF/mesothelin precursor protein, indicated by greater-than (>) signs. The furin-like convertase target site is marked by 4 plus signs (+). Note that the cow, horse, dog, and hedgehog sequences lack the region corresponding to the human MPF due to the exon deletion.

Comparison of MPF/mesothelin precursor proteins in mammals. Sequence comparisons of the human, mouse, rat, and cow MSLN genes and proteins revealed a segmental deletion in the cow gene. To determine whether this deletion was unique to the cow gene or was shared by other mammals, we collected mesothelin precursor protein sequences from various mammalian species. We analyzed mammalian genome assemblies from the UCSC Genome Database and the WGS contigs at the NCBI.

We successfully identified full-length MPF/mesothelin precursor protein sequences from various mammalian species,
including the rhesus macaque, pig, horse, dog, panda, and hedgehog. A multiple sequence alignment of the collected MPF/mesothelin precursor proteins revealed that not only the cow protein but also the pig, horse, dog, panda, and hedgehog proteins lacked a substantial portion of the sequence compared to those of the human, rhesus macaque, mouse, and rat orthologs (Fig. 2).

The human MPF/mesothelin precursor has been reported to undergo several steps of post-translational processing. First, the signal peptide at the N-terminus is cleaved after being transported into the endoplasmic reticulum (ER). Then, the C-terminal GPI-anchor signal is cleaved and replaced by the GPI anchor that binds the MPF/mesothelin protein to the cell surface, and 4 asparagine residues become glycosylated. Finally, furin-like convertase cleaves the recognition sequence (canonically Arg-X-(Arg/Lys)-Arg) in the middle of the precursor protein to release the MPF part from the mesothelin part.

A multiple sequence alignment revealed that the cow, pig, horse, dog, and panda proteins lacked a large portion of the putative MPF part. However, the cleavage site of the signal peptide and the furin-like convertase target sequence were retained in these proteins, indicating that the precursor proteins could still be processed to produce 2 mature proteins. In the case of the hedgehog protein, the entire MPF region is missing. The conserved signal peptide cleavage site and the furin-like convertase target sequences were also removed. However, the hedgehog protein was predicted using the SignalP program to have a highly potential alternative cleavage site, suggesting that the precursor protein could still be processed to expose the mesothelin part on the hedgehog cell surface as in other mammals. All proteins analyzed here had a GPI-anchor signal at the C-terminus that would bind them to the plasma membrane via a GPI anchor.

Deletion in the MSLN gene occurred in a common ancestor of Laurasiatheria. Each of the human, rhesus macaque, mouse, and rat MSLN genes has 16 coding exons. In contrast, the cow, pig, horse, dog, and panda MSLN genes are composed of only 10 coding exons. Also, the hedgehog MSLN gene has 2 additional missing exons resulting in 8 exons, which correspond to the human coding exons 2-9 that are not present, as confirmed through a dot-plot analysis (data not shown). Humans, rhesus macaques, mice, and rats are members of the Euarchontoglires clade, while cows, pigs, horses, dogs, pandas, and hedgehogs belong to the Laurasiatheria clade (18,19). Therefore, we concluded that the initial deletion of exons 3-8 occurred in a common ancestor of Laurasiatheria, and a subsequent deletion of 2 additional exons occurred in the hedgehog lineage (Fig. 3).
hedgehog gene (Figs. 2 and 3). Since the introns that contained the deletion boundaries had the same phase, the ORF of the modified gene was not changed. Although the remodeled MSLN genes encoded uninterrupted precursor proteins, they should have no megakaryocyte-potentiating activities because they lack a substantial or entire part of the MPF region compared to the Euarchontoglires orthologs. The cell adhesion function of the cell surface mesothelin protein should still be retained in these organisms.

Exon deletion of the MSLN gene was caused by an intron breakage and rejoining event. Genome sequence comparison showed that the Laurasiatherian MSLN gene lost a genomic segment that was flanked by introns 2 and 8 of the ancestral MSLN gene. When we compared the intron 2 sequences of the cow, pig, horse, dog, panda, and cat with those of the human introns 2 and 8, we found that the 5' and 3' halves of the former species intron 2 sequences showed similarity to human introns 2 and 8, we found that the 5' and 3' halves of the ancestral MSLN gene (Fig. 4). This observation suggested that a recombination event between introns 2 and 8 in the ancestral MSLN gene caused the deletion in the Laurasiatherians.

Genomic deletions are usually involved with homologous recombination between genomic segments with highly similar sequences (20). In humans, the most common homologous recombination-driven deletion is the Alu recombination-mediated deletion (ARMD) (21). For example, ARMD was recombination-driven deletion is the Alu recombination-sequences (20). In humans, the most common homologous recombination was one of the mechanisms driving gene evolution in the mammalian genome.

Acknowledgments

This research was supported by a Chung-Ang University Research Grant in 2010.

References