

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

Correspondence to: Dr Yoonsoo Hahn, Department of Life Science, Chung-Ang University, 47 Heukseok-ro, Dongjak-gu, Seoul 156-756, Republic of Korea
E-mail: hahny@cau.ac.kr

Key words: *MSLN*, mesothelin, megakaryocyte-potentiating factor, Laurasiatheria, exon deletion

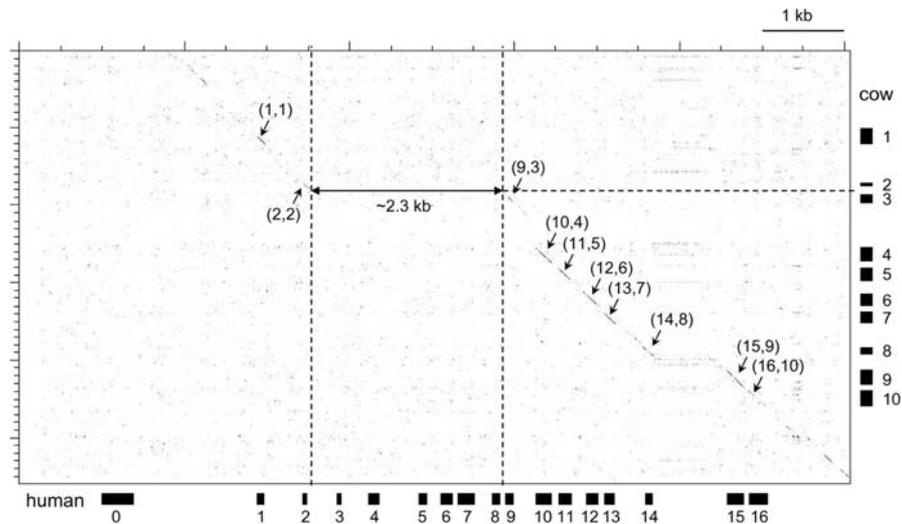


Figure 1. A dot-plot comparing human (horizontal) and cow (vertical) *MSLN* genomic sequences. Regions of sequence similarity between the two sequences appear as diagonal lines. The matching exon pairs are presented in parentheses. The ~2.3-kb, six exon deletion in the cow gene is indicated by a double-headed arrow. The increased length of human intron 14 compared to its orthologous cow intron 8 is due to an expansion of a tandem repeat in humans. The exon locations are marked by black boxes, with the corresponding exon numbers at the right (cow) and at the bottom (human). The human 5'-UTR exon is numbered 0, with coding exons labeled from 1-16.

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*, ailMe11); 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://sonnhammer.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

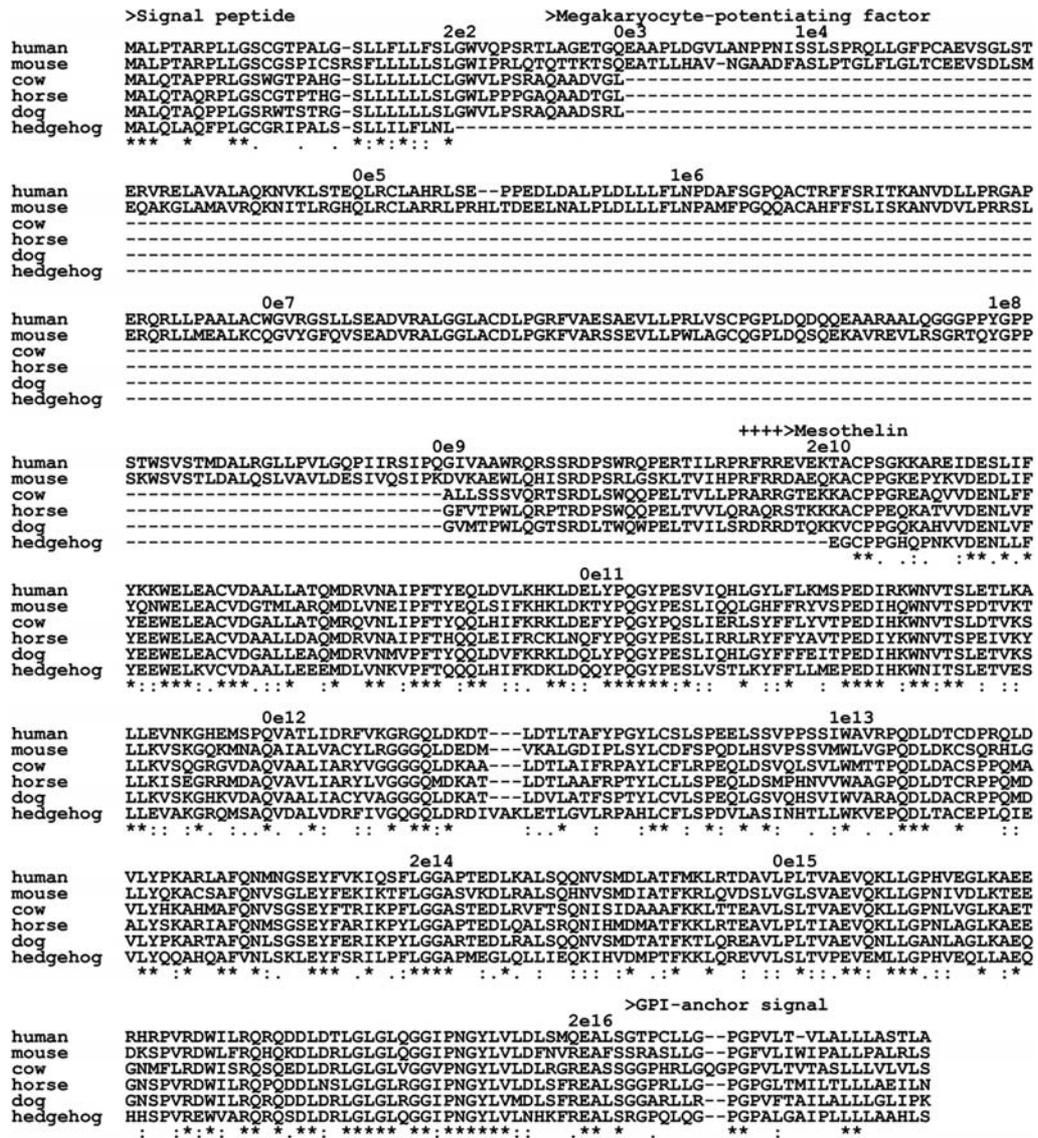


Figure 2. Multiple alignments of mesothelin precursor protein sequences. The RefSeq and predicted mesothelin precursor protein sequences from representative mammalian species were aligned. Just above the sequences are the start positions of each exon of the human *MSLN* gene, marked by the letter e followed 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.

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 orienta-

tions. 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.

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,

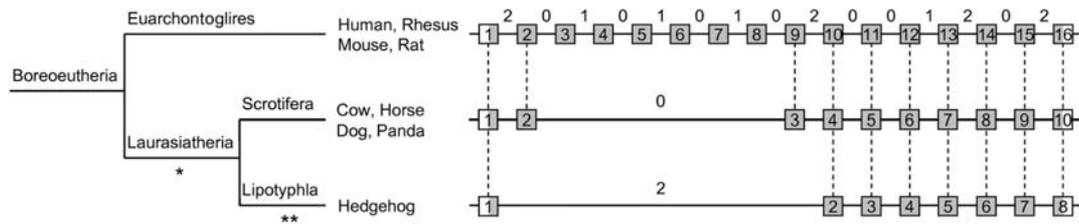


Figure 3. The genomic organizations of *MSLN* genes. Exon numbers are in the boxes: gray, coding region; white, UTR. The orthologous exons are connected by dotted vertical lines. The numbers above the lines are the intron phases. The major mammalian clades are indicated above the branch. The initial deletion event that removed the ancestral exons 3-8 and the subsequent deletion that removed two more exons, 2 and 9, are indicated by * and **, respectively. The figure is not drawn to scale.

```

cow      1  GTAAGGA-CCCTCTGGGGAGTCAGAGGAACGTTTCTGGCCCTGGCTGGGGGGGTTCCCTC 59
pig      1  GTAAGGC-CCCTCTGGGGAAATCAGGGGAGGGTGTCTGGCCCTGGGTG---GGGGTCCCTC 56
horse    1  GTAAGTG--CTTCTGGGGAAACAG--GAAGGAATCC---CCCTGGT---GAGGTTCCT 48
dog      1  GTAAGGT-CCCTCTGGGGAA--CAGAGGAAGGTGTCT-GACCCCTGGGTG---AGGGTCCAT 53
panda    1  GTAAGGT-CCCTCTGGGGAAACAGAAATACATGTCTGGCCCTGGGTG---GGGGTCCCTC 56
cat      1  GTAAGGT-CCCTCTGGGGAAACAGAGGAAGGTGTCT-GGCCCCAGGTA---GGGGTCCCTC 55
human.i2 1  GTAAGGTCCCTCTGGGGAAACAGGGGAGGGTCTTCAGGTCCAGGTG---GGGGTGT+++ 54
          ***** * ***** ** * * ** *

```

```

cow      60  GGACCCCAAGGA-----CACTCCCGCTCACTCCAGCCCTGCTCTGTCTCTCAG 107
pig      57  GAACCCCAAGGC-----CCCTTGGGTTTCATTC-CAGCCCAAGCTCTGTCTCTTAG 103
horse    49  GGATGCAGGGC-----CACACGCAAGCTCAGCCACAACCCCTGCTCTGTCTCTCAG 96
dog      54  GCATGCATGGCTCATGCATGGCTCAGCCACAGCCGCACTCTGCCCTCAG 106
panda    57  GGACAATGGGC-----CACCAAGGCTCAGCCACAGCCCTGCTCTGTCTCTCAG 104
cat      56  GGACCCCTGGC-----CATGCACAGCTCAGCTACAGCCCTGCTCTGTCTCTCAG 103
human.i8 27  GGATGCCCGGC-----CACCCGAGGCTCAGCCAGGTCTCTGCTC-GTCTCTCAG 73
          ** ** * * ** * ** **

```

Figure 4. Sequence comparison of introns involving the breakage and rejoining event. The intron 2 sequences of the cow, pig, horse, dog, panda, and cat *MSLN* genes were aligned with the 5' part of human intron 2 (human.i2, from 1-54) concatenated with the 3' part of intron 8 (human.i8, from 27-73). The two human intron parts are delimited by 3 plus signs (+). Note that the cow, pig, horse, dog, panda, and cat sequences are similar to the artificially fused human sequence. The identical residues in all sequences are marked by asterisks (*).

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).

In the Euarchontoglian *MSLN* genes, which retained the ancestral structure, the introns flanking the deleted exons from the Laurasiatherians had the same phases: phase 0 for the cow, pig, horse, dog, and panda gene, and phase 2 for the



SPANDIDOS PUBLICATIONS gene (Figs. 2 and 3). Since the introns that contained on 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 the 5' part of intron 2 and the 3' part of intron 8 of the human *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 responsible for the exon deletion and inactivation of the human *CMAH* gene (22). However, many other segmental deletions are not due to homologous recombination. For example, the human-specific exon deletions of *MOXD2* and *S100A15A* were not associated with any homologous sequences (16). This type of deletion was caused by the non-homologous end joining (NHEJ) pathway of DNA double-strand break repair (23). In the human *MSLN* gene locus, there was no detectable sequence similarity between introns 2 and 8, indicating that a homologous recombination was not involved in the deletion of the Laurasiatherian *MSLN* gene. Instead, a breakage and rejoining of the ancestral *MSLN* gene introns 2 and 8 via NHEJ seemed to have been responsible for the initial deletion of the Laurasiatherian *MSLN* gene.

In conclusion, we identified full-length *MSLN* genes encoding mesothelin precursor proteins from several Laurasiatherian mammals, including the pig, horse, dog, panda, and hedgehog, by analyzing their genome assemblies and WGS contigs. Comparisons of sequences and genomic organizations revealed that the Laurasiatherian *MSLN* genes experienced a segmental deletion which removed the exons coding for the MPF part of the protein. This observation demonstrated that an exon deletion 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

- Argani P, Iacobuzio-Donahue C, Ryu B, *et al*: Mesothelin is overexpressed in the vast majority of ductal adenocarcinomas of the pancreas: Identification of a new pancreatic cancer marker by serial analysis of gene expression (SAGE). *Clin Cancer Res* 7: 3862-3868, 2001.
- Chang K and Pastan I: Molecular cloning of mesothelin, a differentiation antigen present on mesothelium, mesotheliomas, and ovarian cancers. *Proc Natl Acad Sci USA* 93: 136-140, 1996.
- Ho M, Bera TK, Willingham MC, *et al*: Mesothelin expression in human lung cancer. *Clin Cancer Res* 13: 1571-1575, 2007.
- Hassan R and Ho M: Mesothelin targeted cancer immunotherapy. *Eur J Cancer* 44: 46-53, 2008.
- Hassan R, Bera T and Pastan I: Mesothelin: a new target for immunotherapy. *Clin Cancer Res* 10: 3937-3942, 2004.
- Yamaguchi N, Yamamura Y, Konishi E, *et al*: Characterization, molecular cloning and expression of megakaryocyte potentiating factor. *Stem Cells* 14 (Suppl 1): S62-S74, 1996.
- Yamaguchi N, Hattori K, Oh-eda M, Kojima T, Imai N and Ochi N: A novel cytokine exhibiting megakaryocyte potentiating activity from a human pancreatic tumor cell line HPC-Y5. *J Biol Chem* 269: 805-808, 1994.
- Wang T, Kajino K, Abe M, *et al*: Suppression of cell death by the secretory form of N-terminal ERC/mesothelin. *Int J Mol Med* 26: 185-191, 2010.
- Rump A, Morikawa Y, Tanaka M, *et al*: Binding of ovarian cancer antigen CA125/MUC16 to mesothelin mediates cell adhesion. *J Biol Chem* 279: 9190-9198, 2004.
- Kaneko O, Gong L, Zhang J, *et al*: A binding domain on mesothelin for CA125/MUC16. *J Biol Chem* 284: 3739-3749, 2009.
- Bera TK and Pastan I: Mesothelin is not required for normal mouse development or reproduction. *Mol Cell Biol* 20: 2902-2906, 2000.
- Daniels RJ, Peden JF, Lloyd C, *et al*: Sequence, structure and pathology of the fully annotated terminal 2 MB of the short arm of human chromosome 16. *Hum Mol Genet* 10: 339-352, 2001.
- Jovine L, Park J and Wassarman PM: Sequence similarity between stereocilin and otoancorin points to a unified mechanism for mechanotransduction in the mammalian inner ear. *BMC Cell Biol* 3: 28, 2002.
- Sathyanarayana BK, Hahn Y, Patankar MS, Pastan I and Lee B: Mesothelin, stereocilin, and otoancorin are predicted to have superhelical structures with ARM-type repeats. *BMC Struct Biol* 9: 1, 2009.
- Forrest SM, Cross GS, Speer A, Gardner-Medwin D, Burn J and Davies KE: Preferential deletion of exons in Duchenne and Becker muscular dystrophies. *Nature* 329: 638-640, 1987.
- Hahn Y, Jeong S and Lee B: Inactivation of *MOXD2* and *S100A15A* by exon deletion during human evolution. *Mol Biol Evol* 24: 2203-2212, 2007.
- Stedman HH, Kozyak BW, Nelson A, *et al*: Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428: 415-418, 2004.
- Asher RJ, Bennett N and Lehmann T: The new framework for understanding placental mammal evolution. *Bioessays* 31: 853-864, 2009.
- Asher RJ and Helgen KM: Nomenclature and placental mammal phylogeny. *BMC Evol Biol* 10: 102, 2010.
- Chen KS, Manian P, Koeuth T, *et al*: Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat Genet* 17: 154-163, 1997.
- Sen SK, Han K, Wang J, *et al*: Human genomic deletions mediated by recombination between alu elements. *Am J Hum Genet* 79: 41-53, 2006.
- Hayakawa T, Satta Y, Gagneux P, Varki A and Takahata N: Alu-mediated inactivation of the human CMP-N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci USA* 98: 11399-11404, 2001.
- Burma S, Chen BP and Chen DJ: Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair* 5: 1042-1048, 2006.