Cloning and Characterization of a Novel Mouse Siglec, mSiglec-F

DIFFERENTIAL EVOLUTION OF THE MOUSE AND HUMAN (CD33) Siglec-3-RELATED GENE CLUSTERS*

Received for publication, September 6, 2001, and in revised form, September 27, 2001 Published, JBC Papers in Press, September 28, 2001, DOI 10.1074/jbc.M108573200

Takashi Angata‡, Ravi Hingorani§, Nissi M. Varki‡, and Ajit Varki¶

From the ‡Departments of Medicine and Cellular and Molecular Medicine, the Glycobiology Research and Training Center, University of California, San Diego, La Jolla, California 92093-0687 and ¶BD Biosciences Pharmingen, Research Immunocytometry, San Diego, California 92121

A novel mouse Siglec (mSiglec-F) belonging to the subfamily of Siglec-3-related Siglecs has been cloned and characterized. Unlike most human Siglec-3 (hSiglec-3)-related Siglecs with promiscuous linkage specificity, mSiglec-F shows a strong preference for α2-3-linked sialic acids. It is predominantly expressed in immature cells of the myelomonocytic lineage and in a subset of CD11b (Mac-1)-positive cells in some tissues. As with previously cloned Siglec-3-related mSiglecs, the lack of strong sequence similarity to a singular hSiglec made identification of the human ortholog difficult. We therefore conducted a comprehensive comparison of Siglecs between the human and mouse genomes. The mouse genome contains eight Siglec genes, whereas the human genome contains 11 Siglec genes and a Siglec-like gene. Although a one-to-one orthologous correspondence between human and mouse Siglecs 1, 2, and 4 is confirmed, the Siglec-3-related Siglecs showed marked differences between human and mouse. We found only four Siglec genes and two pseudogenes in the mouse chromosome 7 region syntenic to the Siglec-3-related gene cluster on human chromosome 19, which, in contrast, contains seven Siglec genes, a Siglec-like gene, and thirteen pseudogenes. Although analysis of gene maps and exon structures allows tentative assignments of mouse-human Siglec ortholog pairs, the possibility of unequal genetic recombination makes the assignments inconclusive. We therefore support a temporary lettered nomenclature for additional mouse Siglecs. Current information suggests that mSiglec-F is likely a hSiglec-5 ortholog. The previously reported mSiglec-3/CD33 and mSiglec-E/MIS are likely orthologs of hSiglec-3 and hSiglec-9, respectively. The other Siglec-3-like gene in the cluster (mSiglec-G) is probably a hSiglec-10 ortholog. Another mouse gene (mSiglec-H), without an apparent human ortholog, lies outside of the cluster. Thus, although some duplications of Siglec-3-related genes predated separation of the primate and rodent lineages (about 80–100 million years ago), this gene cluster underwent extensive duplications in the primate lineage thereafter.

The abbreviations used are: Siglec, sialic acid-binding lectins of immunoglobulin superfamily; hSiglec and mSiglec, human and mouse Siglec, respectively; Ig, immunoglobulin; ITIM, immunoreceptor tyrosine-based inhibitory motif; PAA-Bio, biotinylated polyacrylamide; ETFB, electron transfer flavoprotein; RACE, rapid amplification of cDNA ends; sialyl Lex, Neu5Ac2–3Galβ1–4(Fucp1–3)GlcNAc; EST, expressed sequence tag; BAC, bacterial artificial chromosome; kb, kilobase(s); UTR, untranslated region; GlcNAc, N-acetylgalactosamine; RACE, rapid amplification of cDNA ends; PCR, polymerase chain reaction; DHFR, dihydrofolate reductase; ELISA, enzyme-linked immunosorbent assay; TBS, Tris-buffered saline; NK cells, natural killer cells; HAS1, hyaluronan synthase 1; ETFB, electron transfer flavoprotein β subunit; MAG, myelin-associated glycoprotein.

1 The abbreviations used are: Siglec, sialic acid-binding lectins of immunoglobulin superfamily; hSiglec and mSiglec, human and mouse Siglec, respectively; Ig, immunoglobulin; ITIM, immunoreceptor tyrosine-based inhibitory motif; PAA-Bio, biotinylated polyacrylamide; RACE, rapid amplification of cDNA ends; sialyl Lex, Neu5Ac2–3Galβ1–4(Fucp1–3)GlcNAc; EST, expressed sequence tag; BAC, bacterial artificial chromosome; kb, kilobase(s); UTR, untranslated region; GlcNAc, N-acetylgalactosamine; RACE, rapid amplification of cDNA ends; PCR, polymerase chain reaction; DHFR, dihydrofolate reductase; ELISA, enzyme-linked immunosorbent assay; TBS, Tris-buffered saline; NK cells, natural killer cells; HAS1, hyaluronan synthase 1; ETFB, electron transfer flavoprotein β subunit; MAG, myelin-associated glycoprotein.

2 T. Angata, S. Kerr, N. M. Varki, P. R. Crocker, and A. Varki, manuscript in preparation.
analysis of Siglec-3-related mouse Siglecs.  

Here we report the molecular cloning and characterization of a novel mSiglec, as well as a genomic overview of Siglec-3-related gene clusters on mouse chromosome 7 and human chromosome 19. Through multiple approaches, the human and mouse Siglec orthology issue is addressed. Our analyses provide evidence of complex and differential evolutionary paths that CD33-related Siglecs have taken in the rodent and primate lineages.

**EXPERIMENTAL PROCEDURES**

**Materials**—Reagents were purchased from Fisher or Sigma, unless otherwise stated. Access rights to the assembled mouse genomic DNA sequence was purchased from Celera Genomics (Rockville, MD), under an agreement with the University of California System. Corresponding genomic DNA sequences in GenBank™ deposited by publicly funded sequencing projects were also utilized.  

**Cloning of a Novel mSiglec (mSiglec-F) cDNA—**An EST clone (GenBank™ accession number AA672534), which shows sequence similarity to hSiglec-5, was identified and kindly provided by Dr. Paul Crocker (Duke University). The clone contained the 3′-end of the cDNA (~1.6 kb), including the coding sequence for the third (partial) and fourth Ig-like domains, a transmembrane domain and cytosolic tail, and a 3′-untranslated region (UTR). This putative mSiglec was tentatively given the letter designation mSiglec-F,

because it is the sixth mSiglec reported so far and because of the difficulty in unequivocally identifying the human ortholog of this molecule. In an independent attempt to clone an ortholog in our laboratory, the cDNA fragments corresponding to the first Ig-like domain of mSiglec-F were cloned, one of which was later revealed to be identical to mSiglec-F.  

Mouse bone marrow total RNA (C57Bl/6 strain) was prepared with an RNasey Midi kit (Qiagen) and used for the cDNA cloning. The 5′-end cDNA sequence of mSiglec-F was elucidated by 5′-rapid amplification of cDNA ends (RACE), as follows: the first-strand cDNA was synthesized with Superscript II (Invitrogen) using mSL5-SP1 (5′-CTGCTAATCGTTGCTATGCTACC-3′) as primer, and oligo(dA) was added to the 3′-end of the first-strand cDNA using terminal deoxynucleotidyl transferase (Invitrogen) and dATP. Gene-specific cDNA fragments were amplified using the Expand Long Template PCR system (Roche Molecular Biochemicals), with first-strand PCR using oligo(dT)-BAMHI (5′-CTGCTAGCTGGATCCTCAGCTCCT-3′) and mSL5-SP2 (5′-CTGCTATCGAGGATCCAGGCAGGACACTCAG-3′) as primers. DNA fragments obtained were cloned into pCRII-TOPO TA cloning vector (Invitrogen), and sequences were analyzed. Primers were designed based on the revealed sequence, and a full-length cDNA was cloned by reverse transcription-PCR, as follows: random-primer first-strand cDNA was synthesized using Superscript II with random DNA primers, and sequences were analyzed. Primers were designed based on the revealed sequence, and a full-length cDNA was cloned by reverse transcription-PCR, as follows: random-primer first-strand cDNA was synthesized using Superscript II with random DNA primers, and sequences were analyzed. 

**Preparation of Rat Monoclonal Antibodies against mSiglec-F—**Rat bone marrow, spleen, blood, and lymph node cells obtained from Balb/c, C57Bl/6, and C3H mice (Jackson Laboratories, Bar Harbor, ME) were labeled with either E50 (BD Biosciences Pharmingen, San Diego, CA). All clones gave a similar staining pattern on primary cells. The selection of clone E50 (BD Biosciences Pharmingen, San Diego, CA). All clones gave a similar staining pattern on primary cells. The selection of clone E50 (BD Biosciences Pharmingen, San Diego, CA). All clones gave a similar staining pattern on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells. The selection of clone E50 was based on its staining intensity on primary cells.
Differential Evolution of Human and Mouse Siglec Gene Clusters

RESULTS

Cloning of a Novel mSiglec cDNA—A partial mouse EST clone showing high sequence similarity to hSiglec-5 was identified, and full-length cDNA of this molecule was cloned through 5’-RACE and reverse transcription-PCR from mouse bone marrow. The cDNA encodes a Siglec-like molecule (569 amino acids) with four Ig-like domains, a single-pass transmembrane domain and cytosolic tail (Fig. 1A). The amino acid sequence of the translated product contains the defining structural features of Siglecs, i.e. three amino acids whose side chains are known to contact sialic acid in Siglec-1/Sialoadhesin (a conserved arginine residue and two aromatic amino acids, one near the N-terminal and the other near the conserved arginine residue) (24), as well as three conserved cysteine residues in each of the first and second Ig-like domains (Fig. 1A). Judging from the sequence similarity of the first two Ig-like domains and cytosolic tail, this new molecule belongs to a group of Siglecs related to Siglec-3, which typically contains two putative tyrosine-based signaling motifs in both human and mouse (3, 4). However, despite high sequence similarity to hSiglec-5 in the fourth Ig-like domain and shared Ig-like domain numbers, the similarity is not exclusive to hSiglec-5 in the other regions of the molecule (Fig. 1B). Because this new mouse molecule is the sixth mSiglec to be reported and its human ortholog could not be unequivocally established at this point, it was tentatively given the letter designation mSiglec-F. The previously reported mSiglec-3/CD33 (13) and mSiglec-E/MIS (14, 15) also lack clear orthologous correspondence to hSiglecs. As discussed below, this issue is addressed here by multiple approaches, including phylogenetic analysis and whole-genome survey of human and mouse Siglecs, as well as analyses of gene structure and chromosomal localization.

Stable Binding Specificity of the Novel mSiglec—COS-7 cells transiently transfected with mSiglec-F full-length expression construct showed clear sialic acid-dependent binding of human erythrocytes (Fig. 2), justifying the denomination of this molecule as a Siglec (1). A point mutation at the conserved arginine residue (R114A) abolished this binding, proving that this arginine residue is essential in sialic acid recognition by this molecule (Fig. 2), as has been shown in all other Siglecs studied to date (6, 16, 18, 25–28). Unlike some previously reported Siglecs, pretreatment of the COS cells was not essential to liberate the cell surface Siglec from masking by cis-ligands (Fig. 2). To determine a more detailed ligand binding specificity, a recombinant fusion protein mSiglec-F:Fc was prepared and used in an ELISA assay, with biotinylated polycrylamide arrays carrying multiple copies of sialylated oligosaccharides as ligands (16, 18, 19). As shown in Fig. 3, mSiglec-F showed a strong preference toward α2–3–linked sialic acids. This is in contrast to many Siglec-3-related hSiglecs, which show rather promiscuous binding to both α2–3- and α2–6–linked sialic acids (3, 29). Substitution at the proximal GlcNAc with fucose residue (i.e. sialyl Leα strongly interfered with the interaction. The latter finding is similar to that for hSiglecs recognizing α2–3–linked sialic acids (29), with the known exception of hSiglec-9 (16).

Expression Pattern of the Novel mSiglec—Rat monoclonal antibodies against mSiglec-F were prepared as described under “Experimental Procedures” and used for flow-cytometry and immunohistochemical analyses. Four clones with identical staining pattern were obtained, and the clone (E50–2440) with the highest staining intensity on primary cells was selected for the phenotyping studies. The staining profile in bone marrow shows that mSiglec-F is present on CD11blo as well as on very short cytosolic domain, which greatly restricts the amount of information usable for phylogenetic analysis if included.

5 T. Angata and A. Varki, unpublished.
CD11b\textsuperscript{hi} cell populations (Fig. 4A). However, the expression of mSiglec-F is considerably higher on the CD11b\textsuperscript{lo} population. A similar staining profile for mSiglec-F was seen in cells isolated from mouse blood and spleen, although the actual percentage of CD11b-positive cells was lower in these compartments (data not shown). Detailed analysis with appropriate markers (see “Experimental Procedures”) indicated that mSiglec-F is not expressed on T cells, B cells, or NK cells, thus restricting the expression of these cells to the myeloid and monocytic lineages (CD11b\textsuperscript{+/H11001} or Ly6G\textsuperscript{+/H11001}, Fig. 4A and data not shown). Cells expressing mSiglec-F are also positive for CD54 and CD95 but negative for CD14, CD11c, CD80, CD117, Ly6A/E, Ly6C, and Mac-3. Identical staining profiles for all these tissues was seen in C3H, Balb/c, and C57BL/6 strains of mice, indicating that expression of mSiglec-F is not restricted by strain differences (data not shown).

To further characterize the cell populations expressing mSiglec-F, bone marrow cells that had been sorted by fluorescence-assisted cell sorting were cytospun onto slides, air-dried, then stained with the Wright-Giemsa stain, and then studied under oil-immersion microscopy. Cells that were CD11b\textsuperscript{lo}/mSiglec-F\textsuperscript{hi} were identified as being predominantly immature cells of myelomonocytic lineage (Fig. 4B). Cells that were CD11b\textsuperscript{lo}/mSiglec-F\textsuperscript{lo} were mostly mature or immature neutrophils and monocytes (data not shown), including leukocytes with ring-shaped nuclei that are now thought to include murine granulocytes, monocytes, and their immediate precursors (30). In contrast, the CD11b\textsuperscript{lo}/mSiglec-F\textsuperscript{neg} population consisted mainly of lymphocytes and erythrocyte precursors (data not shown). The cells showing high Ly6G (Gr-1) expression (presumably the most mature neutrophils) were not strongly mSiglec-F-positive (data not shown), suggesting that the level of this Siglec falls as the cells of the myelomonocytic lineage reach full maturity. However, this conclusion remains uncertain, because of the current controversy over the precise morphological basis for differentiating between mouse neutrophils and monocytes and their immediate precursors (30).

Immunohistochemical analysis of tissue sections also supports the conclusion that mSiglec-F expression is restricted to the cells of myelomonocytic lineage (Fig. 5). In the thymic
medulla (Fig. 5, A–C), most of the cells that expressed mSiglec-F also expressed CD11b (Mac-1); however, there were also rare single positive mSiglec-F cells seen. In the spleen, the expression of mSiglec-F did not completely overlap with that of CD11b (Fig. 5, D–F), whereas it showed better correlation with that of Ly6G (Gr-1) (data not shown). Scattered mSiglec-F cells were also noted in the lung parenchyma. No staining was seen in the heart, liver, kidney, pancreas, lymph nodes, brain, salivary gland, or gastrointestinal tract. Taken together, the data indicate that there is a prominent expression of mSiglec-F on immature bone marrow cells of the myelomonocytic lineage, and a decreased expression in mature neutrophils and monocytes. The nature of the mSiglec-F-positive cells found in thymus, spleen, and lung requires further investigation. Of particular note, the high level of mSiglec-F expression on immature bone marrow cells of myelomonocytic lineage (see Fig. 4) provides the first known marker that is selective for these cell populations in the mouse.

Comparison of Mouse and Human Siglec-3-related Gene Clusters—To gain an overview of all mSiglecs and to compare the structure of mouse and human Siglec-3-related gene clusters, we analyzed mouse genomic DNA sequences recently assembled to near completion by Celera Genomics. The previously known mSiglecs with clear human orthologs (Siglec-1/Sialoadhesin, Siglec-2/CD22, and Siglec-4/myelin-associated glycoprotein) were confirmed to be single-copy genes. We also found a mouse genomic region (chromosome 7, cytological band...
7B2) syntenic to the hSiglec-3-like gene cluster (chromosome 19, cytological band 19q13.3–13.4) in which a number of mSiglec genes and pseudogenes are clustered (Fig. 6). Although the same genomic region is also being sequenced by the publicly funded mouse genome sequencing project, the relevant scaffolds are not assembled into a contiguous entity as of the time of this writing. In the human genome, the Siglec-3-related gene cluster is flanked by the kallikrein gene cluster on the centromeric side by ∼40 kb (the closest one is KLK-L6/KLK14), and by a hyaluronic acid synthetase gene (hyaluronan synthase 1 (HAS1)) at the telomeric side by ∼80 kb. The mSiglec-3-related gene cluster is also flanked by the kallikrein gene cluster on one side. On the other hand, the HAS1 ortholog was not found near the cluster. According to a recent study (31), there is a discontinuity of mouse-human synteny between ETFB (electron transfer flavoprotein β subunit) and HAS1. In fact, a region of ∼3 Mb containing the mouse Siglec-3-related gene cluster appears to have undergone intrachromosomal translocation, whereas the genomic segment containing mouse HAS1 is found on mouse chromosome 17 (31). Nevertheless, all but one canonical Siglec-3-like sequence (mSiglec-H) are found in the above cluster in the mouse genome. Thus, we consider the mouse Siglec-3-related gene cluster on chromosome 7 to be unique and the only cluster equivalent to that in the human genome.

Besides mSiglec-F, there is only one other Siglec-like putative gene (mSiglec-G) in the cluster yet to be formally reported. It appears to be an ortholog of hSiglec-10, judging from its high sequence identity to hSiglec-10 and similar chromosomal localization in relation to the proximal genes (ETFB, NKG7, and LIM2). Including this putative gene (which is a functional gene according to Dr. P. Crocker⁶), there are four Siglec genes and two pseudogenes in this region. This is in striking contrast to the syntenic region of the human genome, which contains seven Siglec genes, one Siglec-like gene (Siglec-L1), and thirteen pseudogenes (exon fossils of Siglec-like Ig V-set domains). As mentioned earlier, the Siglec-3-related mouse genes do not show strong one-to-one correspondence to hSiglec, with the only exception of mSiglec-G, the putative hSiglec-10 ortholog. However, comparison of the gene maps of Siglec-3-related human and mouse gene clusters, as well as the gene structures of individual Siglec (Fig. 7), provide clues to address the orthology between mouse and human Siglec. First, the previously reported mSiglec-3/CDD3 (13) has a shared domain structure with hSiglec-3 (Fig. 7) and a similar chromosomal localization in relation to other Siglec genes (Fig. 6). Thus, it is the best candidate for an ortholog of hSiglec-3/CDD3. By the same criteria, mSiglec-EMIS (14, 15) is a probable ortholog of hSiglec-9. Although hSiglecs 7 and 9 are localized close together in the human gene cluster and show a high degree of similarity at the mRNA and polypeptide levels, their genomic structures are different: Although hSiglec-7 gene contains a remnant of V-set-domain exon (Siglec-P2) between the exons encoding the first and second Ig-like domains of the protein product, hSiglec-9 and mSiglec-E lack this signature motif (Fig. 7). Identification of the mSiglec-F ortholog in humans is not as straightforward, due to the lack of useful genomic landmarks and the presence of two candidate genes, namely hSiglecs 5 and 6; the former has exons for four Ig-like domains, whereas the latter has exons for three Ig-like domains and one exon fossil similar to the fourth Ig-like domain exon in mSiglec-F and hSiglec-5 (Fig. 7). However, in view of the fact that, unlike hSiglec-6, mSiglec-F is not expressed on B lymphoid cells but restricted to cells of myeloid lineage, we consider that hSiglec-5 is more likely to be the ortholog of mSiglec-F. Further support comes from the finding that mSiglec-F is not expressed in the mouse placenta (data not shown), whereas hSiglec-6 is expressed in this fetal organ (19).

**Phylogenetic Analysis of Human and Mouse Siglecs**—To further understand the phylogenetic relationships of the mSiglecs and hSiglecs, we separately compared the N-terminal (first two Ig-like domains) and C-terminal (the transmembrane domain plus the cytosolic tail) amino acid sequences of the Siglec. As shown in Fig. 8A, Siglecs 1, 2, and 4 of mouse and human show a clear one-to-one correspondence in the phylogenetic tree based on the amino acid sequences of the first two Ig-like domains. On the other hand, the Siglec-3-related Siglecs do not show such a clear relationship. However, mSiglec-G appears to form a monophyletic clade with hSiglecs 10 and 11, and mSiglec-E with hSiglecs 7, 8, and 9 (Fig. 8A). In these cases, the lack of one-to-one orthologous correspondence can be attributed to specific gene duplication(s) that took place in the lineage leading to humans after the split of the primate and rodent lineages (estimated to be 80–100 million years ago) (32). This interpretation is also supported by the phylogenetic tree reconstructed from the amino acid sequences of transmembrane plus cytosolic domains of Siglecs (Fig. 8B). Here again mSiglec-E

---

⁶ P. Crocker, personal communication.
forms a monophyletic clade with hSiglec-7, 8, and 9 (and L1), and mSiglec-G groups with hSiglec-10. The fact that hSiglec-11 does not belong to "Siglec-10 clade" in this tree suggests that this gene is a product of a gene duplication and recombination. On the other hand, the evolutionary relationships of other Siglec-3-related genes in this analysis is not quite consistent with our interpretation of orthology: Thus, in the tree based on N-terminal amino acid sequences (Fig. 8A), mSiglec-3 and hSiglec-3 are paraphyletic rather than monophyletic, and the same is true of mSiglec-F and hSiglec-5. Also, this clade contains Siglecs with 2 and 4 Ig-like domains (hSiglec-6 has an exon fossil corresponding to the fourth Ig-like domain, as discussed above). These anomalies may be due to unequal genetic recombination(s), including gene conversion of some Siglec genes, which consist mostly of modular exons encoding one Ig-like domain each, with the potential for "exon shuffling." The phylogenetic tree of C-terminal amino acids (Fig. 8B) does not provide significant clues to resolve this problem, either. A more precise understanding of the evolutionary paths taken by this gene clade will require more data on related genomic sequences from other mammalian species.

Another clear implication of this phylogenetic analysis is that the ancestral Siglec-3-related genes (proto-Siglecs 3, 9, 10, and possibly 5) were already present in the common ancestor of
Differential Evolution of Human and Mouse Siglec Gene Clusters

proposed previously, based on their distinct expression patterning suggests that, despite common basic properties (sialic acid residues).7 This strongly suggests that the Siglec family of molecules is present in all vertebrates. On the other hand, we failed to find Siglec orthologs in the Drosophila and Caenorhabditis elegans genomes (16). Thus, true Siglecs may have emerged only in the deuterostome lineage of animals, where sialic acids are prominently expressed.

The presence of putative intracellular signaling motifs in the cytosolic tail of mSiglec-F suggests that it is involved in intracellular signaling, as has been shown in some other Siglec-3-related Siglecs (7, 14, 15, 28, 35, 36). Although it is beyond the scope of this report, the identification of proteins interacting with the cytosolic tail of mSiglec-F is of interest, especially those that interact with the distal motif, which does not conform to canonical ITIM motif but is highly conserved among all Siglec-3-related Siglecs. We are also pursuing natural ligand(s) for the extracellular domain of mSiglec-F, which may regulate intracellular signaling. Although α2–3-linked sialic acids are abundantly expressed in mouse tissues, mSiglec-F may prefer a limited number of specific ligands expressing this structure in a preferred manner, or in a particular polypeptide structural context, as in the case of the specific interaction of P-selectin with P-selectin glycoprotein ligand-1 (37). Alternatively, α2–3-linked sialic acid recognition by mSiglec-F, regardless of other structural contexts, may constitutively send suppressive signals via its cytosolic tail. As in the interaction of killer cell Ig-like receptors and major histocompatibility complex molecules (38), such a suppressive signal may cease in the absence of ligand, e.g. upon contact with abnormally under-sialylated cells (e.g. virus-infected cells), which may in turn lead to elevation of the cellular activity (16). The apparent absence of mSiglec-F masking by cis-ligands conforms to this hypothesis. We are currently pursuing the functional aspects of mSiglec-F, especially the importance of sialic acid recognition by this molecule, via a targeted exon-replacement approach in mice.

Although we consider our genomic sequence-based analysis of human and mouse Siglecs “comprehensive,” there are several caveats. First, we are assuming the human and mouse genomic data bases we have probed are complete. Second, we have noted a genomic DNA segment (~1.2 kb) that shows >99% sequence identity to a 5′-region of the hSiglec-5 gene, containing coding sequences for the signal peptide and Ig-like domains 1 and 2 (indicated as a triangle labeled “5′” in Fig. 6). This segment is upstream of the true hSiglec-5 gene, is found in both the public and Celera human genome data bases, and could be confirmed by PCR analysis.8 Given the extremely high sequence identity to the Siglec-5 genomic region, this DNA segment is likely to be a relic of a very recent partial gene duplication. However, due to the lack of coding sequences for the downstream components (Ig-like domains, transmembrane domain, or cytosolic tail), we cannot be certain if this genomic DNA segment yields a functional mRNA or translated polypeptide. On the other hand, we cannot confidently assign it as a pseudogene. Third, we have noted a GenBankTM entry by the Human Genome Project derived from an unfinished human

7 T. Angata and A. Varki, unpublished.
8 T. Angata and A. Varki, unpublished.
BAC clone sequence (accession number AC026638, clone RP11-423F16) that contains several putative Siglec genes and pseudogenes. Although it is assigned to chromosome 8, we have found that many fragments (contigs) of this clone show extremely high sequence identity (>99.5%) to another unfinished BAC clone CTD-3187F8 (GenBank™ accession number AC063977), which encompasses a part of the Siglec-3-related gene cluster on chromosome 19. Given the existence of so-called “segmental duplication” (34), or “paralogous genome evolution” (39) in which long genomic DNA segments (up to 200 kb) duplicate in as yet unknown mechanism, it is possible that the chromosome 8 region covered by the clone RP11-423F16 actually represents another Siglec-3-related gene cluster, which was duplicated very recently. However, evidence against this possibility is provided by fluorescence in situ hybridization studies of Siglec-3/CD33 (40) and Siglec-7 (41). In both cases hybridization signals were localized exclusively on chromosome 19 (cytological band 19q13.3–13.4). If clone RP11-423F16 was indeed derived from chromosome 8, the DNA sequences nearly identical (>99%) to these genes in the chromosome 8 region represented by this clone should have shown an additional signal in the fluorescence in situ hybridization analysis. Also, there is only one unique Siglec-3-related gene cluster in the Celera Genomics human genome data base. Overall, it seems more likely that this is a technical artifact, such as misidentification of a clone or cross-contamination by an adjacent BAC clone, among other possibilities. Finally, there is one more unique Siglec-like putative gene in the human and mouse genomes that has not been studied. This hypothetical molecule was identified by a homology search of genomic and EST data bases, sequencing of a mouse EST clone, and by mutual subtractions of mouse and human genomic DNA sequences. However, the number and positions of cysteine residues in the first Ig-like domain of this molecule is different from those of typical Siglecs, and one of the cysteines is situated immediately adjacent to the “conserved arginine” residue involved in sialic acid recognition by known Siglecs. Thus, it is questionable if this hypothetical protein will recognize sialic acids. Nevertheless, this gene, which is highly conserved between human and mouse (~85% nucleotide identity), is likely to be of interest in deciphering the relationship of Siglec family to other Ig-like molecules and perhaps the origin of Siglec family gene.

Acknowledgments—We thank Els Brinkman-Van der Linden and Pascal Gagneux for general advice and helpful discussions, and Hui-Ling Han for technical assistance. Some of the mouse genomic data used for this report were generated through the use of the Celera Discovery System, under a Client agreement with the University of California System.

REFERENCES


* T. Angata and A. Varki, unpublished.