ELSEVIER

Gene 286 (2002) 259–270

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Genomic organization of the siglec gene locus on chromosome 19q13.4 and cloning of two new siglec pseudogenes

George M. Yousef[a,b], Michael H. Ordon[a], George Foussias[a,b], Eleftherios P. Diamandis[a,b,*]

[a]*Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, Canada M5G 1X5*
[b]*Department of Laboratory Medicine and Pathobiology, University of Toronto, ON, Canada*

Received 2 November 2001; received in revised form 2 January 2002; accepted 24 January 2002
Received by R. Di Lauro

Abstract

The sialic acid binding immunoglobulin-like lectin (Siglec) family of genes is a recently described member of the immunoglobulin superfamily. Within this Siglec family there is a subgroup of genes which bear a high degree of homology with Siglec-3 (CD33), thus designated the Siglec-3-like subgroup of Siglecs. While their mRNA structure has been reported, the full genomic organization of these genes, is not known. Genes of this subgroup have been mapped to chromosome 19q13.4, primarily through in situ hybridization. Through analysis of several bacterial artificial chromosome (BAC) clones, we studied an approximate 700 kb region that encompasses the putative Siglec gene locus on chromosome 19q13.4. We established the first detailed map of the locus, which contains 8 Siglec and Siglec-like genes. Our map shows the relative position of all genes and the precise distances between them, along with the direction of transcription of each gene. To our knowledge, this is the first report that describes the full genomic organization of all members of the CD33-like subgroup of Siglecs, including the promoter sequences of all genes. Members of this subfamily exhibit two patterns of organization of the signal peptide, which is followed by one V-set domain (except for the long form of the *siglecL1* gene). Exons containing the C2-set domains are all comparable in size and are separated by linker exons. The transmembrane domain is encoded for by a separate exon of almost the same size in all genes. The total number of exons differs according to the number of C2-set Ig domains, but intron phases are identical. The cytoplasmic domain is always encoded by two exons. We further identified two new Siglec pseudogenes in this locus, and analyzed their tissue expression pattern and their structural features.

Keywords: Siglec; CD33-like-Siglec; Immunoglobulin superfamily; Sialoadhesin; Chromosome 19; Pseudogene

1. Introduction

The immunoglobulin superfamily is composed of a large number of cell surface proteins that play a vital role not only in immunity, but also in controlling the behavior of cells in different tissues, through their ability to mediate cell surface recognition events. Sialic acid binding immunoglobulin-like lectins (Siglecs) are a recently defined subset of this superfamily. The first Siglecs to be characterized were sialoadhesin (Siglec-1), CD22 (Siglec-2), CD33 (Siglec-3), and myelin-associated glycoprotein (Siglec-4A). We and others have uncovered the presence of a cluster of genes that encode novel Siglecs, which are highly related to CD33. These genes include Siglec-3, Siglecs 5–10 (Cornish et al., 1998; Patel et al., 1999; Angata and Varki, 2000a,b; Floyd et al., 2000; Foussias et al., 2000a,b; Zhang et al., 2000; Li et al., 2001; Munday et al., 2001; Whitney et al., 2001; Yousef et al., 2001c) and a Siglec-like gene, SiglecL1 (also known as SLG) (Angata et al., 2001; Foussias et al., 2001; Yu et al., 2001) and fluorescence in-situ hybridization data from previous reports indicate that members of the CD33-like subgroup of Siglecs are all located on the long arm of chromosome 19. Determination of the size of this gene subfamily and cloning of all its members is important to our understanding of their contribution to human biology. Also, obtaining detailed information about the relative positions of all members and identifying the promoter sequences for each gene is the first step towards understanding the

Abbreviations: Siglec, sialic acid binding immunoglobulin-like lectin; RT–PCR, reverse transcription–polymerase chain reaction; Ig, immunoglobulin; EST, expressed sequence tag; ITIM, immunoreceptor tyrosine kinase inhibitory motif; *SiglecL1*, Siglec-like gene (also known as SLG) (GenBank accession no. AF277806); Siglec10, also known as Siglec-like gene-2 (GenBank accession no. AY029277); *ΨSP-1*, Siglec pseudogene-1; *ΨSP-2*, Siglec pseudogene-2; LMIP, Lens fiber membrane intrinsic protein (GenBank accession no. P55344); ZFP, zinc finger protein 175 (GenBank accession no. NM007147); BAC, bacterial artificial chromosome; TIGR, The Institute of Genomic Research; HGP, Human Genome Project; IgSF, immunoglobulin superfamily

* Corresponding author. Tel., +1-416-586-8443; fax: +1-416-586-8628.
E-mail address: ediamandis@mtsinai.on.ca (E.P. Diamandis).

0378-1119/02/$ - see front matter
PII: S0378-1119(02)00432-8

Table 1
Primers used for reverse transcription polymerase chain reaction (RT–PCR) analysis for the ΨSP-1 pseudogene

Gene	Primer name	Sequence[a]	Direction
ΨSP-1	SN4-F1	GGACAGGGACAGCAAGAGAA	Forward
	SN4-F3	CCTCACAGTCCTCGAACCAT	Forward
	SN4-GF	CCTCCCGATCCTGCACCCAA	Forward
	SN4-R1	CCTCACAGTCCTCGAACCAT	Reverse
	SN4-R3	GGAAAGGGGAGATGTTGGTC	Reverse
	SN4-GR	CTGTGGGTCAGGGCTGGTGA	Reverse
ACTIN	ACTINS	ACAATGAGCTGCGTGTGGCT	Forward
	ACTINAS	TCTCCTTAATGTCACGCACGA	Reverse

[a] All primer sequences are presented in the 5′-3′ direction.

regulatory mechanisms that control gene expression in various physiological and pathological conditions.

Recent estimates indicate that the size of chromosome 19 is 58–72 Mb. Unique features of chromosome 19 include the highest density of CpG islands (43 islands/megabase), the highest density of genes and the harboring of many clusters of gene families.

In our previous work, we characterized the human kallikrein gene locus on chromosome 19q13.3–q13.4 (Yousef et al., 2000; Yousef and Diamandis, 2001). In this paper, we describe the first detailed map of the cluster of the CD-33-like human Siglec genes which is located adjacent to the kallikrein locus. We define the order of genes along the chromosomal region, the direction of transcription, and estimated the distances between genes. We also analyze in detail their common structural features, at the genomic level and identify potential promoter sequences upstream of each gene. We further cloned two Siglec-like pseudogenes located in the same chromosomal region. These pseudogenes share the same structural features with other Siglec genes but they encode for a predicted truncated proteins since the sequence is interrupted by in-frame stop codons.

2. Materials and methods

2.1. DNA sequences on chromosome 19

Genomic sequences generated by the Human Genome Project (HGP) were obtained from the web site of the Lawrence Livermore National Laboratory (LLNL) (http://www-bio.llnl.gov/genome/genome.html). These sequences encompass a region of 700 kb and include the putative locus of the CD33-like subgroup of Siglecs on chromosome 19q13.4. The data were in the form non-directional finished and unfinished BAC clones.

2.2. End sequencing of BAC clones

Purification of BAC DNA was done by a rapid alkaline lysis miniprep method, which is a modification of the standard Qiagen-Tip method. The ends of these clones were sequenced using vector-specific primers, with an automated DNA sequencer. Additional information for end-sequencing for some clones were obtained from The Institute of Genomic Research (TIGR) data base (http://www.tigr.org).

2.3. Reverse transcriptase–polymerase chain reaction (RT–PCR)

Two micrograms of total RNA was reverse-transcribed into first strand cDNA using the Superscript preamplification system (Gibco BRL). The final volume was 20 μl.

Gene-specific primers for each of the Siglec pseudogenes were used for PCR-based amplification of a human tissue panel described below. PCR was carried out in a reaction mixture containing 1 μl of cDNA,10 mM Tris–HCl (pH 8.3), 50 mM KCl, 2 mM $MgCl_2$, 200 μM dNTPs (deoxynucleoside triphosphates), 100 ng of primers and 2.5 units of HotStar Taq polymerase (Qiagen, Valencia, CA) on an Eppendorf thermal cycler. The cycling conditions were 95 °C for 15 min to activate the HotStar Taq polymerase, followed by 40 cycles of 94 °C denaturation for 30 s, annealing (temperature varied according to primer combination used) for 30 s, and 72 °C extension for 1 min, and a final extension at 72 °C for 10 min. Equal amounts of PCR products were electrophoresed on 1.5% agarose gels and visualized by ethidium bromide staining. Primers used for the different reactions are summarized in Tables 1 and 2.

2.4. Tissue expression

Total RNA isolated from 24 different human tissues was purchased from Clontech, Palo Alto, CA. We prepared cDNA as described above and used it for PCR reactions. Tissue cDNAs were amplified at various dilutions using gene-specific pairs of primers for Siglec pseudogene-1 (SN4-F3 and SN4-R3, see Table 1).

Table 2
EST clones with > 98% homology to ΨSP-2 pseudogene

GenBank no.	Tissue of origin	I.M.A.GE. ID	Matching exons[a]
AF150431	Umbilical cord blood (cd34 + stem cell)	–	2–6
AI801574	Stomach	2185632	6
AI040646	Fetal liver and spleen	1657109	6
AV742547	Hematopoietic stem cell	CBCBQF09	6

[a] Exon numbers refer to our GenBank accession no. AY040545.

Table 3
Primers used for RT–PCR analysis for the *ΨSP-2*

Primer name	Sequence[a]	Direction	Template
SN-F1	CTGCTGCTGCCCCTGCTGTG	Forward	cDNA
SN-F2	ATGCAAGATTCCGGCTGGAG	Forward	cDNA
SN-F3	AGGACCCTCCACTCCTCAGA	Forward	cDNA
SN-G1	GGATGACCTCTGACCACGTG	Forward	Genomic
SN-R1	GGAACCAGTGGCCATAAGCA	Reverse	cDNA
SN-R2	ACGGTTCCTGATTTCCAGGA	Reverse	cDNA
SN-R3	ACGGAAATGCTGGCACATAG	Reverse	cDNA
SN-G2	GCTCAGGAGCAGTGTTCCTT	Reverse	Genomic

[a] All primer sequences are presented in the 5′–3′ direction.

2.5. Expressed sequence tag (EST) searching

The predicted exons of the putative new genes (pseudogenes) were subjected to homology search using the BLASTN algorithm (Altschul et al., 1997)on the National Center for Biotechnology Information web server (http://www ncbi.nlm.nih.gov/BLAST/) against the human EST database (dbEST). Clones with 98% identity were obtained from the I.M.A.G.E. consortium through Research Genetics, Huntsville, AL (Table 3). The clones were propagated, purified and sequenced from both directions with an automated sequencer, using insert-flanking vector primers.

2.6. Cloning and sequencing of the PCR products

Due to the high degree of homology between the genes in this genomic region, all primers were designed to be specific for each gene, annealing away from conserved regions. To further verify the identity of the PCR products, they were cloned into the pCR 2.1-TOPO vector (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The inserts were sequenced from both directions using vector-specific primers, with an automated DNA sequencer.

2.7. Structure analysis

A number of computer programs were used to predict the presence of putative new genes in the genomic area of interest. We initially tested these programs using the genomic sequences of known Siglec genes. The most reliable computer programs, GeneBuilder (gene prediction/exon prediction) (http://l25.itba.mi.cnr.it/~webgene/genebuilder.html), Grail 2 (http://compbio.ornl.gov) and GENESCAN (http://genes.mit.edu/GENSCAN.html) were selected for further use. Multiple alignment was performed using the 'Clustal X' software package. Phylogenetic studies were performed using the 'Phylip' software package. Distance matrix analysis was performed using the 'Neighbor-Joining/UPGMA' program and parsimony analysis was done using the 'Protpars' program. Hydrophobicity study was performed using the Baylor College of Medicine search launcher. Signal peptide was predicted using the 'SignalP' server. Protein structure analysis was performed by 'SAPS' (structural analysis of protein sequence) program. Conserved domain search was performed using the 'Conserved Domain' (CD) and 'ProDom' programs.

3. Results

3.1. Assembly of a continuous genomic sequence encompassing the Siglec gene locus

Genomic sequences from the region encompassing the putative Siglec gene locus on chromosome 19q13.4 were obtained from the draft sequences of the Human Genome Project (Lander et al., 2001). About 700 kb were analyzed in this study. These sequences were in the form of six BAC clones with different lengths. Four of these clones were fully sequenced, and the remaining two were in the form of separate 'contigs'. The orientation of each clone (centromeric to telomeric) was not determined, and the sequenced strand (sense or anti-sense) was not defined. Four complementary approaches were followed to obtain a contiguous, directional region representing the locus:

1. We performed an *Eco*RI restriction analysis study of the BAC clones, using the *Eco*RI restriction enzyme, and the results were compared to that of the *Eco*RI restriction map of the region, available from the HGP. This enabled us to align the clones along the chromosomal region.
2. A pairwise 'BLAST' analysis was performed for the sequences of the six clones. The ends of some clones were overlapping, enabling us to construct a contiguous segment.
3. We screened the RPCI-11 human BAC library, using gene-specific probes obtained from the mRNA sequences of known Siglecs in this region, and four positive clones (10I11, 615L12, 261J23 and 99I13) were obtained and end-sequenced. Additional end-sequencing data from other libraries were obtained from the TIGR end-sequence database.
4. Long PCR reactions were performed to estimate the missing parts of the unfinished clones using primers located at the ends of the adjacent clones.

These strategies, in addition to our previous data which determined the orientation of the human BAC clone (BC349142) (Yousef et al., 2000; Yousef and Diamandis, 2000), and the direction of transcription of the most telomeric kallikrein gene (KLK14) (Yousef et al., 2001b), enabled us to construct a directional map of the Siglec locus. Fig. 1 shows the localization of the BAC clones in the region of interest on chromosome 19q13.4.

3.2. Organization of the Siglec gene locus on chromosome 19q13.4

As shown in Fig. 2, the Siglec gene locus on chromosome

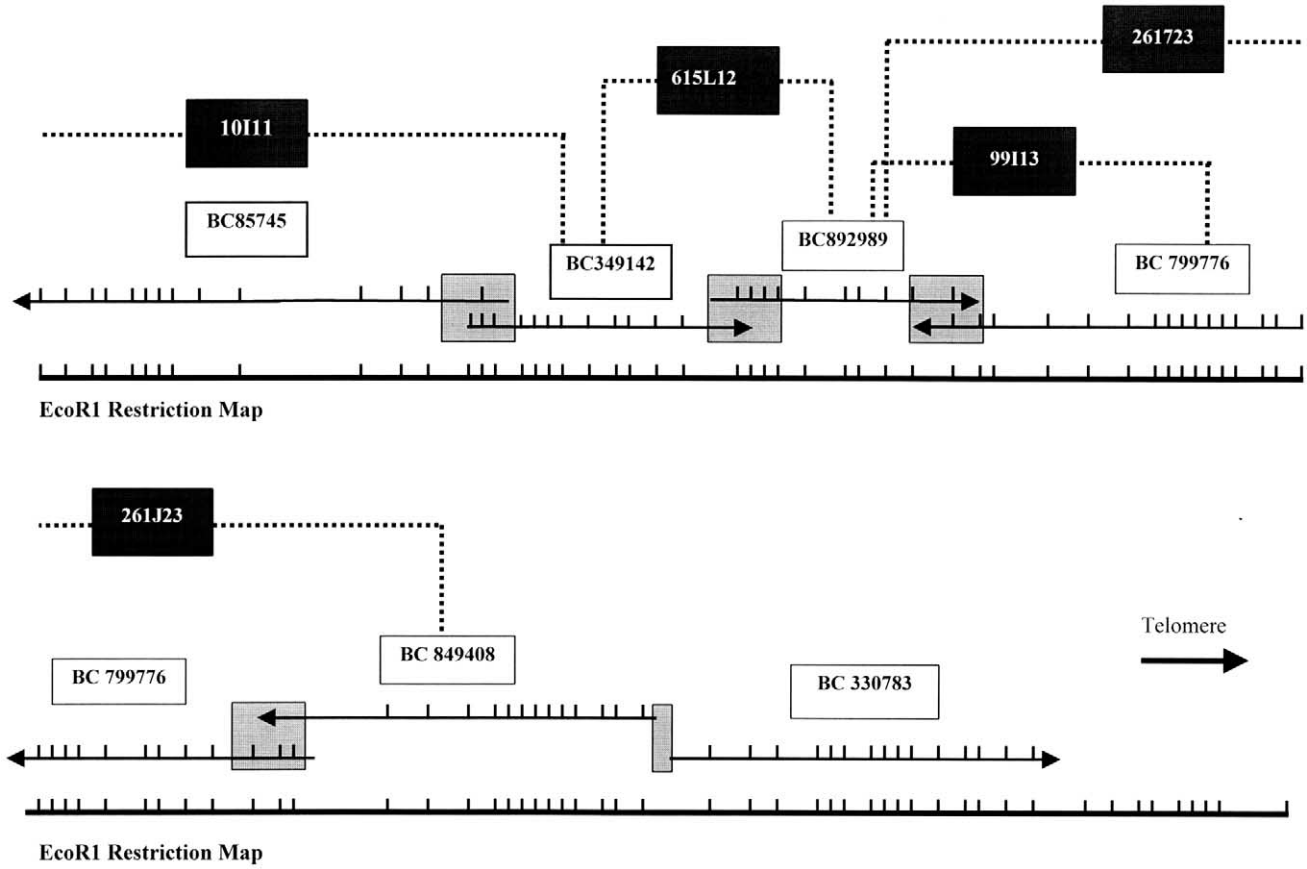

Fig. 1. Schematic representation of the alignment of BAC clones along the *Eco*RI restriction map of the Siglec gene locus on human chromosome 19 (presented as a horizontal line at the bottom of the figure). The *Eco*RI restriction sites are presented by small vertical lines. BAC clones are represented by horizontal arrows with the arrowhead representing clone direction. Clone ID is indicated inside the open boxes. End-sequenced clones are represented by dotted lines with the clone ID inside shaded boxes. Areas of overlap are shown in gray.

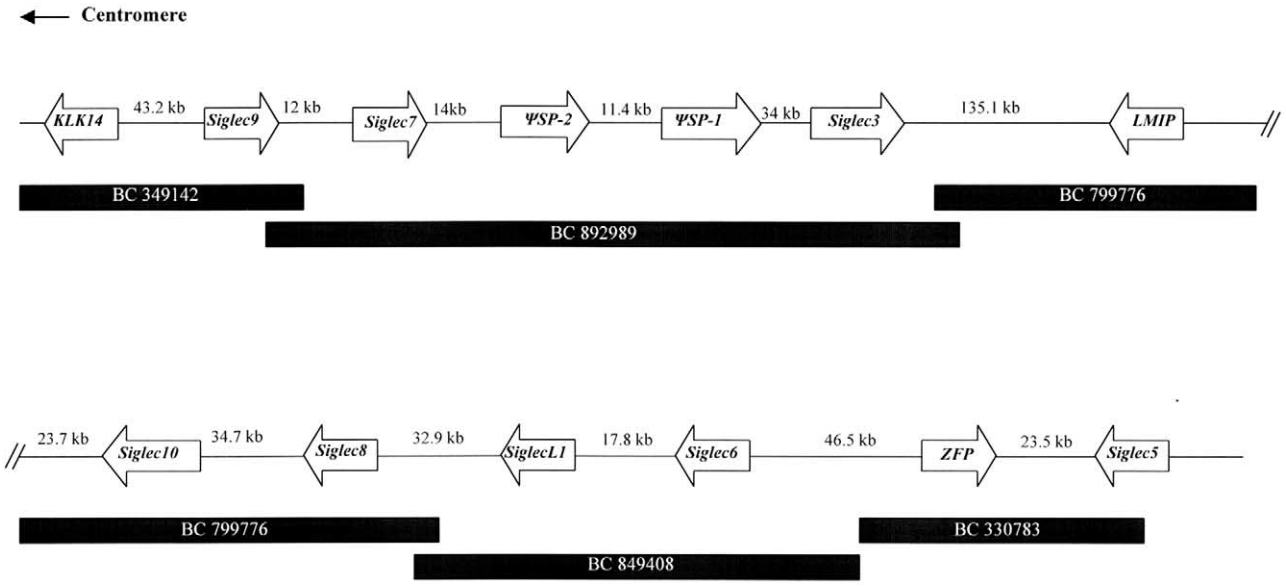

Fig. 2. An approximate 700 kb region of contiguous genomic sequence around chromosome 19q13.4. Overlapping BAC clones are presented by black boxes. Genes are represented by horizontal arrows denoting the direction of the coding sequence. Distances between genes are mentioned in kilobases. For complete gene names and GenBank accession numbers, see legend of Fig. 8. Figure is not drawn to scale.

19q13.4 spans a region of approximately 506.2 kb of genomic sequences and is located 43.2 kb more telomeric to the adjacent kallikrein gene locus. The locus contains the six known members of the CD33-like subgroup of siglecs (*Siglec-3*, *Siglecs 5–10*) (Cornish et al., 1998; Patel et al., 1999; Angata and Varki, 2000a,b; Floyd et al., 2000; Foussias et al., 2000a,b; Zhang et al., 2000; Li et al., 2001; Munday et al., 2001; Whitney et al., 2001; Yousef et al., 2001c) and a Siglec-like gene, *SiglecL1* (also known as *SLG*) (Angata et al., 2001; Foussias et al., 2001; Yu et al., 2001) in addition to two non-siglec genes, Lens Fiber Membrane Intrinsic Protein (LMIP, GenBank accession no. P55344) and Zinc Finger Protein 175 (ZFP, GenBank accession no. NM007147). *Siglec-9* was found to be the most centromeric gene, transcribed from centromere to telomere and separated by 12 kb from *Siglec7*, which is transcribed in the same direction. The *Siglec-3* gene is located 69.4 kb further telomeric and is separated by 166 kb from *Siglec10* (previously known as *SLG2*) which is transcribed in the opposite direction. *Siglec-8*, *SiglecL1*, *Siglec-6* and *Siglec-5* are all clustered in the telomeric region of the locus and separated from each other by 32.9, 17.8 and 85.1 kilobases, respectively. Two non-Siglec genes are present in the locus; LIMP telomeric to Siglec-3, and *ZFP* between *Siglec-6* and *Siglec-5*. In addition to the known Siglec genes, we cloned two novel Siglec pseudogenes, as described below.

3.3. Common structural features of the CD33-like subgroup of Siglecs

The genomic organization of all members of the CD33-like subgroup of Siglecs was obtained by comparing the published mRNA structures of these genes with the genomic sequence of the locus, and the exon/intron boundaries were precisely defined (for our GenBank accession numbers, see the legend of Fig. 8). In addition, we identified a region of about 1 kb upstream of each gene that contains the putative proximal promoter. All published mRNAs match exactly with the genomic sequences with the exception of *Siglec-5*. The 3′untranslated region of the *Siglec-5* mRNA submitted by Cornish et al. (Cornish et al., 1998) has 20 additional nucleotides that were not found in the genomic sequences (bases 19391–19410 according to our GenBank accession no. AY040820). Human EST database search indicated the presence of six clones of *Siglec-5* mRNA. Multiple alignment of sequences of all EST clones and the genomic sequence of *Siglec-5* indicated that these extra bases exist in all sequences except the genomic sequence (data not shown), raising the possibility of an error in the HGP sequence.

The CD33-like locus on chromosome 19q13.4 consists of eight Siglec genes and two Siglec pseudogenes, described in this paper. Exon/intron boundaries of all genes are in agreement with the consensus splice sites sequences, in particular the presence of the highly conserved GT/AG donor/acceptor site at the beginning and end of introns, respectively. All Siglec genes in this locus consist of a signal peptide, one V-set immunoglobulin domain (with the exception of the long form of the SiglecL1 gene (Foussias et al., 2001) which has an extra V-set domain), followed by a variable number of C2-set immunoglobulin domains. As shown in Fig. 3, the relation between the signal peptide and the V-set

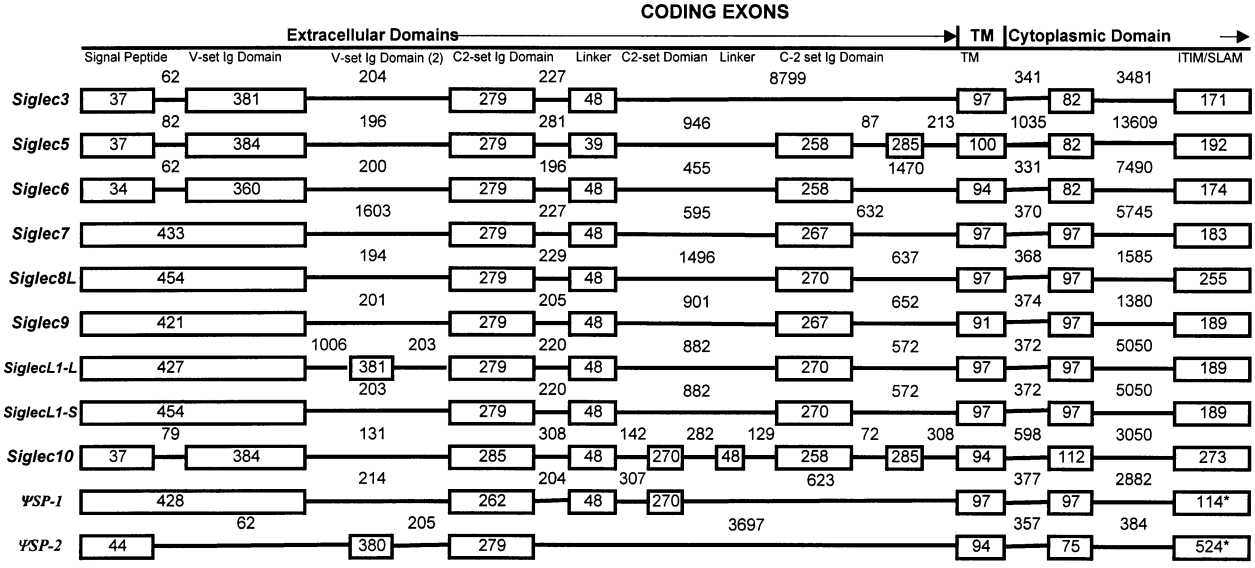

Fig. 3. Schematic diagram showing the comparison of the genomic structure of the CD33-like subgroup of Siglecs. Exons are shown by solid bars with exon length indicated inside in base pairs and introns by the connecting lines with length indicated above each line. First exon length was calculated by convention from the start codon and last exon length is shown up to the point of the stop codon. The upper panel indicates domain names. SiglecL1, siglec-like gene-1, SiglecL1-L, long isoform of SiglecL1. SiglecL1-S, short isoform of SiglecL1. Siglec10, siglec-like gene-2; ΨSP-1, Siglec pseudogene-1; ΨSP-2, Siglec pseudogene-2; TM, transmembrane region; (*) does not contain the conserved ITIM/ITIM-like (SAP-binding consensus) motifs. ITIM/ITIM-like, see under abbreviations and text. Figure is not drawn to scale.

Ig domain falls into one of two patterns; in the first, the signal peptide is encoded by a separate exon of about 34–37 nucleotides, followed by a short intron of 62–82 bp and the V-set domain is encoded by another exon of about 360–384 bp. Siglec-3, Siglecs-5–6 and Siglec10 fall into this category. In the remaining genes, the signal peptide and the V-set Ig domain are included in one large exon. The V-set domain is followed by a C2-set Ig domain of almost invariable length of 279 nucleotides (except for Siglec10 which is 285 bp in length). The following exon always represents a linker peptide encoded for by 48 nucleotides (except for Siglec-5, where it is 39 bp in length). Siglec10 has a unique extra C2-set Ig domain followed by another linker peptide. The next exon represents a C2-set Ig domain which exists in all genes except for Siglec-3. The length of this exon ranges from 258 to 270 bp. An additional unique exon of 285 bp exists before the transmembrane domain in Siglec-5 and the Siglec10 genes only. The transmembrane domain is encoded by a separate exon in all Siglecs, and is followed by a cytoplasmic domain which is encoded by two exons. The second of these exons contains the characteristic ITIM and ITIM-like (SAP-binding consensus) motifs which occur in a similar position in all genes (see also Fig. 4). It is interesting that not only the exons, but also many intron lengths are comparable in the Siglec gene, with the last intron being the largest, ranging from 3 kb in the Siglec10 gene to about 13 kb in Siglec-5. Fig. 4 shows that the intron phases are conserved in all members of this subgroup of Siglecs, except for the two pesudosiglec genes. All exons end with an intron phase of I except for the last two exons, where the exon containing the transmembrane domain ends with an intron phase of 'II' and the following exon always has an intron phase of '0'.

3.4. Phylogenetic analysis

Phylogenetic analysis of all protein members of the Siglec family, including the hypothetical proteins of the two Siglec pseudogenes, was accomplished through the Phylip software package. Different trees were constructed using the UPGMA, neighbor-joining and parsimony methods. All trees agreed, as is evident in Fig. 5, that the two newly cloned pseudogenes are tightly clustered among the other previously known members of the Siglec-3-like subgroup of Siglecs, where ΨSP-1 was close to Siglecs 7–9 and SiglecL1 and Siglec-3 was the closest gene to ΨSP-2. Other Siglecs (Siglec- 1–2 and Siglec-4) were phylogenitically distinct from members of the CD33-like subgroup.

3.5. Cloning of two novel pseudo-Siglec genes in the Siglec locus

A number of computer programs were used to predict the presence of putative new genes within the contiguous genomic region and two new putative genes were predicted. In order to verify the existence of these genes and to obtain their full genomic structure, different approaches were used, including EST sequencing and homology search and PCR

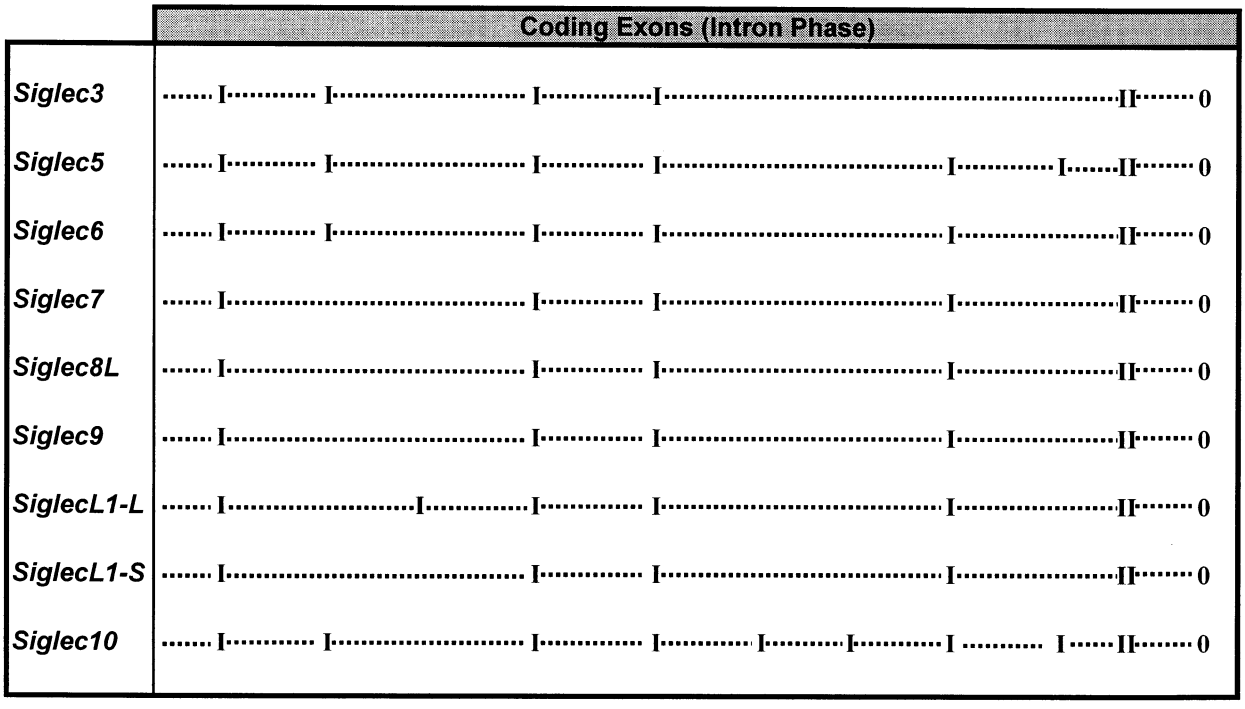

Fig. 4. Schematic diagram showing the intron phases of the CD33-like subgroup of Siglecs. Exons are represented by dotted lines and intron phases are indicated by Roman letters between exons. Exons were aligned as shown in Fig. 3. The intron phase refers to the location of the intron within the codon; I denotes that the intron occurs after the first nucleotide of the codon, II the intron occurs after the second nucleotide, 0 the intron occurs between codons. Intron phases were found to be identical in all genes.

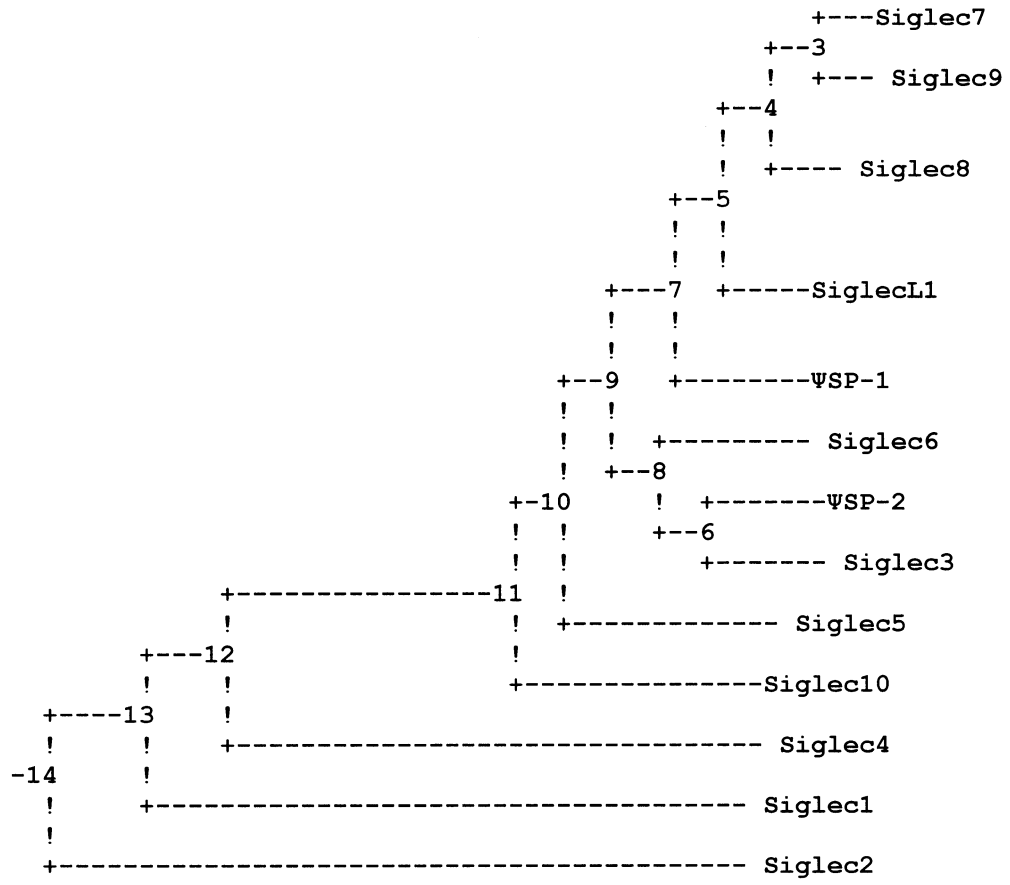

Fig. 5. Dendrogram of the predicted phylogenetic tree for human Siglec proteins using the UPGMA method. As expected, the tree grouped the CD33-like group of Siglecs together, away from Siglecs 1, 2 and 4.

screening of tissues, as described elsewhere (Yousef and Diamandis, 1999; Yousef et al., 2001d).

3.6. Genomic organization of the Siglec Pseudogene-1 (ΨSP-1)

The *ΨSP-1* gene spans an area of 6221 nucleotides of genomic sequence on chromosome 19q13.4, 34 kb centromeric to Siglec-3 (Fig. 2). The gene is formed of 7 exons and 6 intervening introns. Exon lengths are 422, 265, 48, 270, 97, 97 and 361 bp, respectively (Fig. 7). All intron/exon splice sites are closely related to the consensus splice sites (-m**GT**AAGT…C**AG**m-, where m is any base) (Iida, 1990). Verification of the structure of this pseudogene was obtained by screening a panel of 24 tissues by RT–PCR using different combinations of primers (Table 1). The mRNA of this gene was found to be highly expressed in bone marrow, spleen, colon and testis and to a lower extent in other tissues (Fig. 6). The predicted translation initiation codon (ATG) is located at position 2125 (numbers refer to

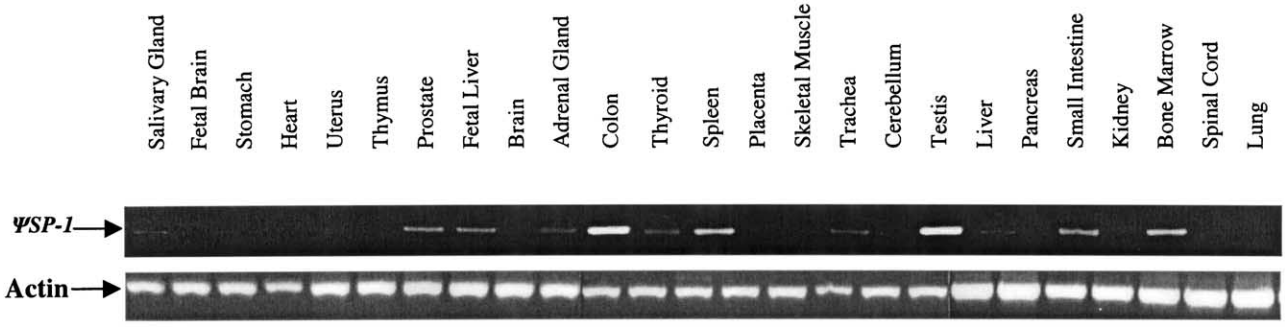

Fig. 6. Tissue expression of the ΨSP-1 pseudogene. Actin, a housekeeping gene, was used as a control. The pseudogene was highly expressed in the testis, colon, spleen and bone marrow, and to a lesser extent in other organs.

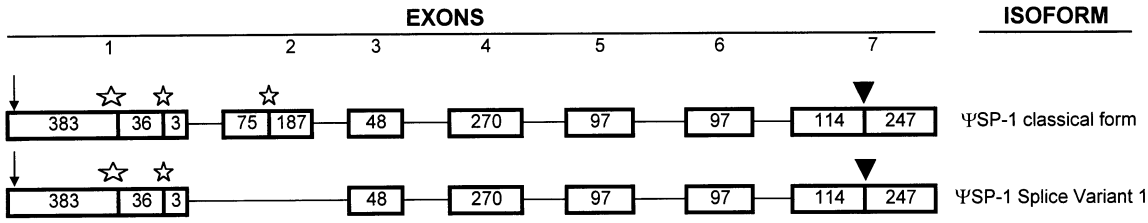

Fig. 7. A diagram representing the splice variants of the ΨSP-1 pseudogene. Exons are represented by solid boxes with lengths mentioned in base pairs and introns by the connecting lines. The position of the start codon is denoted by an arrow, and the hypothetical stop codon in exon 7 by an arrowhead. In-frame stop codons are marked by asterisks. Full genomic sequences of all splice variants are described in GenBank accession no. AY029277.

our GenBank submission no. AY040544). The sequences surrounding this start codon match with a Kozak consensus sequence for translation initiation, especially the most highly conserved purine at position −3 that occurs in 97% in eukaryotic mRNAs (Kozak, 1991). In addition, there is a conserved 'C' in position +4 that is found in other members of this family, including Siglecs-5–9. Furthermore, using this start codon, the resulting predicted protein product shows extensive homology with other members of the CD33-like subgroup of Siglecs.

As shown schematically in Fig. 7, attempts to translate the gene in all possible reading frames show that there is only one frame that preserves the genomic organization of a typical Siglec, as described earlier. However, this open reading frame was interrupted by three stop codons; two are in the first exon and the third is in the second exon (Fig. 7), giving rise to a truncated protein product of 127 amino acids, with a predicted molecular weight of 14.3 kDa. Screening a panel of tissues, did not allow us to identify any other alternatively-spliced form of this gene that might be encoding for a non-truncated protein, except for one form (splice variant-1) which is missing exon 2 but still has two of the three stop codons (Fig. 7).

In order to reveal the structural identity of this pseudogene, we performed an analysis of its hypothetical protein product after removal of the three interrupting stop codons. Like other members of the CD33-like subgroup of Siglecs, ΨSP-1 also possesses an N-terminal signal sequence. Using neural networks and hidden Markov models trained on eukaryotes, a cleavage site was predicted between amino acids 16 and 17 (VEG-QG), in a position that is comparable to the cleavage site for other Siglecs (Fig. 8). The presence of a signal sequence is also supported by the hydrophobicity pattern of the protein. Conserved Domain (CD) and ProDom searches together with homology alignment indicated the presence of 3 conserved immunoglobulin domains (residues: 24–121, 173–242 and 251–329), representing the V-set domain, followed by the 2 C-set domains present in other CD33-like Siglecs (Fig. 8). The transmembrane domain, predicted by TMpred and also evident in the Kyte–Doolittle hydrophobicity plot (data not shown) is in keeping with observations for other members of this subgroup. However, this pseudogene does not contain the two characteristic tyrosine-based motifs, ITIM and ITIM-like (SAP-binding consensus), noted in other members of the CD33-like subgroup of Siglecs (Fig. 8). The overall degree of similarity of this hypothetical protein with other members of the CD33-like subfamily of Siglecs is 50–75%. Taken together, these data suggest that ΨSP-1 should be considered a Siglec pseudogene.

3.7. Genomic organization of the Siglec Pseudogene-2 (ΨSP-2)

Using gene prediction programs, we also identified another potential Siglec gene in the locus, centromeric to *ΨSP-1* and 14 kb more telomeric to Siglec-7 (Fig. 2). In order to verify the structure of this potential gene, we screened the human EST database and we were able to identify four EST clones with > 98% homology with predicted sequence of the gene (Table 2). These clones were isolated from tissues known to express Siglecs like the hematopoietic stem cells, fetal liver, spleen and stomach. PCR amplification of different tissues using gene-specific primers (Table 3) in addition to sequencing of EST clones, allowed us to reveal the full genomic structure of the gene. The ΨSP-2 gene is 6101 bp in length, and is formed of six exons and five introns, with the exon/intron boundaries in agreement with the consensus sequence for splice junctions. Exon lengths are 44, 380, 279, 94, 75 and 524 bp. The full genomic structure is submitted to GenBank (GenBank accession no. AY040545). Attempts to translate the mRNA in all reading frames indicated the presence of a potential methionine start codon at position number 2815 of the genomic sequence, which fits well in the consensus sequence for translation initiation, especially the most highly conserved purine at position −3, and the conserved 'C' in position +4 that is found in other members of this family. This is the only start codon that is followed by the characteristic signal peptide of Siglecs. We classified this potential Siglec as a pseudogene based on the fact that all three reading frames encode for truncated protein products that are interrupted by stop codons. Structural analysis indicated that one of these frames (frame 2) is missing only one critical nucleotide before position 2964 of the genomic structure and this leads to frame shifting and predicted truncation of the protein product. Inserting a nucleotide would lead to a full length hypothetical Siglec protein. Re-sequen-

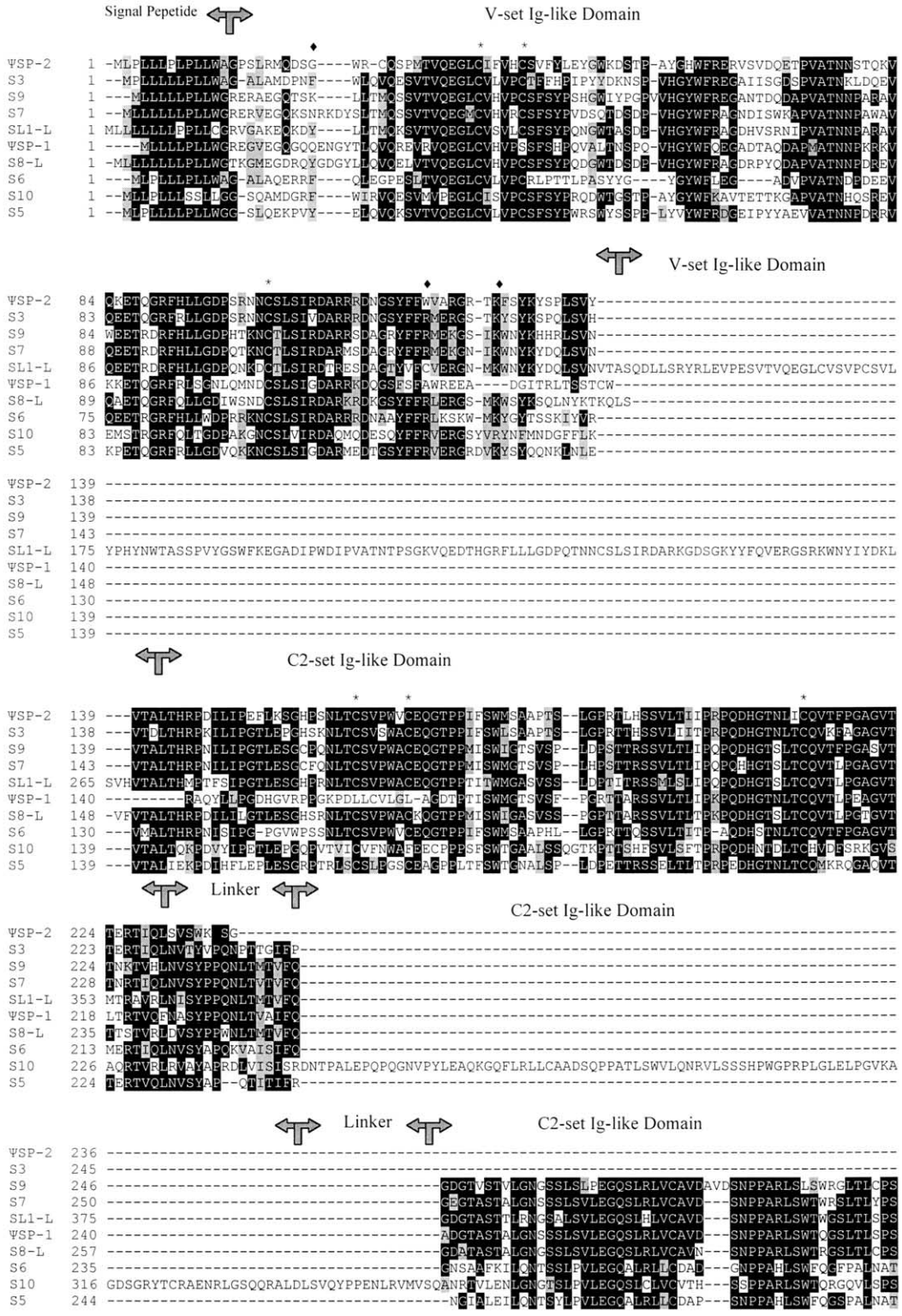

Fig. 8. Multiple alignment of the amino acid sequences of members of the CD33-like subgroup of Siglecs. Numbers of the amino acid residues of each protein are shown on the left of each row. Identical residues are highlighted in black, and similar residues in gray. Ig domains are indicated above the corresponding sequence. Exon boundaries were determined based on the genomic structure and are shown with bent arrows. The transmembrane domain, ITIM and ITIM-like motifs are indicated, as are the conserved cysteine residues (*) that form the disulfide bonds of the Ig-like domains in siglecs. The amino acid residues essential for Sialic acid binding are indicated by (◆). Gene names and GenBank accession numbers are as follows: ΨSP-1, Siglec pseudogene-1 (AY040544); ΨSP-2, Siglec pseudogene-2 (AY040545); S3, Siglec-3 (AY040541); S9, Siglec-9 (AF135027); S7, Siglec- 7 (AY040543); SLI-L, the long form of Siglec-like gene-1 (AF277806); S8, Siglec-8 (AF287892); S6, Siglec-6 (AY040542); Siglec10 (also known as Siglec-like gene-2) (AY029277); S5, Siglec-5 (AY040820). For more on boundaries details, see Fig. 3.

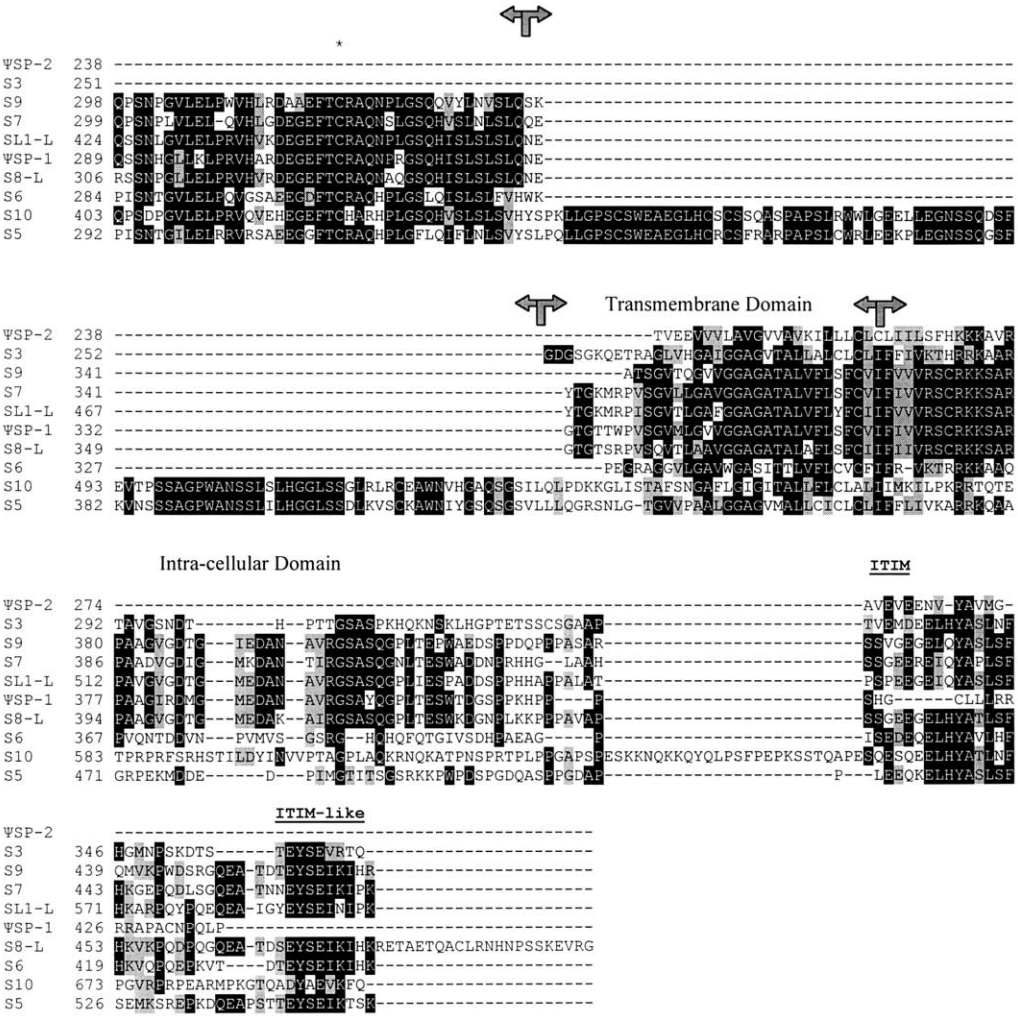

Fig. 8. (*continued*)

cing of the EST clones from both directions, mRNA from different human tissues and genomic DNA from 10 different individuals using the genomic gene-specific primers (SN-G1 and SN-G2) (Table 3) failed to reveal any polymorphic pattern that contains an extra nucleotide in this region.

Structural analysis of the hypothetical protein product after addition of one nucleotide was performed, to verify the identity of this pseudogene. Like other members of the CD33-like subgroup of Siglecs, ΨSP-2 has an N-terminal signal sequence with a predicted cleavage site between amino acids 13 and 14 (LWA-GP), in a position that is comparable to the cleavage site for other Siglecs (Fig. 8). The presence of a signal sequence is also supported by the hydrophobicity pattern of the protein (data not shown). The protein has also two conserved immunoglobulin domains (residues 27–139 and 162–234), representing the V-set domain, followed by a C-set domain (Fig. 8). The single transmembrane domain (amino acid residues 242–264) of the hypothetical protein, predicted by TMpred and also evident in the Kyte–Doolittle hydrophobicity plot, is in keeping with observations for other members of this subgroup (data not shown). Like ΨSP-1, this pseudo gene was also lacking the two characteristic tyrosine-based motifs, ITIM and ITIM-like (SAP-binding consensus). The overall degree of similarity of this hypothetical protein with other members of the CD33-like subfamily of Siglecs is ~65%.

We compared the 5′ and 3′ flanking ends of each of the pseudogenes against each other and against all direct repeat sequences reported by Vanin for all known processed pseudogenes (Vanin, 1985). No evidence for the presence of direct repeats was found in either of the two pseudogenes.

4. Discussion

The CD33-like subgroup of Siglecs is a recently identified group of the immunoglobulin superfamily (IgSF) of genes. In this report, we describe a characterization of the human Siglec gene locus on chromosome 19q13.4 and

present the first detailed map which shows the relative positions of all members of the CD33-like subgroup of Siglecs and their direction of transcription (Fig. 2). This map is consistent with previous reports on the localization of these genes using fluorescence in situ hybridization (FISH) and other techniques (Cornish et al., 1998). It should be noted, however, that the estimated intervals between genes depends on the published mRNA sequences, and might change slightly in the future, since the transcription initiation site of these genes is not well defined and some genes may have extra 5′ untranslated exon(s) that have not as yet been identified. In addition, some of these genes (e.g., SiglecL1 gene) have two or more splice variants with different lengths. The map is directional. Our results indicate that the centromeric group of Siglecs (including *Siglec-3*, *-9* and *-7* and the two Siglec pseudogenes are all transcribed from centromere to telomere, while the more telomeric group is transcribed in the opposite direction (Fig. 2). *Siglec-9* is likely the most centromeric member of this family, as it is in close proximity to *KLK14*, the most centromeric member of the kallikrein gene family (Yousef and Diamandis, 2001). However, the possibility still exists that the locus is extended further telomerically, and that other Siglec genes may be located downstream of *Siglec-5*. This possibility seems less likely since we could not detect any Siglec genes by gene prediction programs in an area of about 100 kb further telomeric to *Siglec-5*.

The term 'pseudogene' comprises a wide group of non-functional loci with a marked diversity of characteristics (Martinez-Arias et al., 2001). Vanin (1985) defined a pseudogene by two major features: being *related* to one or more paralogous genes, and being *defective* in function. The lack of function results from either failure of transcription, translation, or production of a protein that does not have the same functional repertoire as that encoded by the normal paralog gene (Mighell et al., 2000). Most pseudogenes are created by one of two mechanisms: tandem duplication (non-processed pseudogenes) or retrotransposition from a functional gene (processed pseudogenes) (Cooper, 1999). Our newly identified pseudogenes (ΨSP-1 and ΨSP-2) both lack the characteristic criteria for processed pseudogenes including the lack of intronic sequences in the genomic structure, cessation of homology with the functional gene at the start and end of transcription, extra poly(A) sequences and the presence of flanked direct repeats (Vanin, 1985). They, however, meet the criteria of non-processed pseudogenes, including retaining the exon-intron structure and the close proximity to the functional ortholog (Martinez-Arias et al., 2001). It should be noted, however, that the possibility still exists that there are other functional splice variants of these genes that exist only in certain tissues or certain developmental stages or pathological situations. Also, as is the case with other pseudogenes, the possibility still exists that they can encode for a truncated, yet functional protein (Karin and Richards, 1982; Scarpulla, 1984; Vanin, 1985).

In keeping with the wide species distribution of the IgSF, members of the Siglec family have been identified in a wide variety of organisms. Siglec-3 has also been identified in mice (Tchilian et al., 1994). Further, during their identification and characterization of Siglec-7, Falco et. al. reported that under low stringency conditions, a Siglec-7-specific probe hybridized with genomic DNA from Rhesus monkey, suggesting a cross-species conservation between humans and monkeys (Falco et al., 1999).

With regard to the human Siglec-3-related subgroup of Siglecs, we have shown here that they are all localized to chromosome 19q13.4. To date, this tight clustering of members of the Siglec family has only been observed in humans and mice. This has raised the possibility that this subgroup has arisen through gene duplication and exon shuffling relatively recently in vertebrate evolution (Angata and Varki, 2000a). The mechanism proposed involves unequal crossing-over of sister chromatids during meiotic recombination. In fact, chromosome 19-specific minisatellites have been identified in the long arm of chromosome 19, which may have facilitated such gene duplication (Angata and Varki, 2000a; Yousef et al., 2001a). Studies of the γ-globin gene in a wide range of species have led to the development of a gene duplication model similar to that proposed for the Siglec-3-related subgroup of Siglecs. Fitch et al. (1991) found that interspersed repetitive elements may act as nucleation sites for unequal cross-over events. Further, these exchanges can also introduce nucleotide changes in the coding and untranslated regions of these genes, which have been implicated in the regulatory changes that delayed expression of some globin genes from embryonic to fetal life (reviewed in Fitch et al., 1991).

The gene for *Siglec-1* is located on mouse chromosome 2 and human chromosome 20 (Mucklow et al., 1995). This is in striking contrast to the remainder of the Siglec family members, which are all localized to mouse chromosome 7 and human chromosome 19q. This absence of linkage between *Siglec-1* and the remainder of the members of the Siglec family, in both humans and mice, suggests that the *Siglec-1* locus was separated from other Siglecs prior to mammalian speciation (Mucklow et al., 1995).

The evolution of the Siglec family has also been studied in two model organisms of the proteostome lineage, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. In both of these organisms, no obvious Siglec family members were detected (Angata and Varki, 2000a). Interestingly, these species also lack the enzymes necessary for the synthesis of sialic acids. Based on these findings, it was suggested that the emergence of Siglecs during evolution appears to be dependent on the constitutive expression of sialic acids in animals of the deuterostome lineage (Angata and Varki, 2000a).

In conclusion, we characterized the Siglec gene locus on chromosome 19 and revealed the genomic organization of each of its members. We also cloned two new Siglec pseudogenes in this locus. This information will be useful in

further understanding the evolution and function of these genes in humans.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Angata, T., Varki, A., 2000a. Cloning, characterization and phylogenetic analysis of Siglec-9, a new member of the CD33-related group of Siglecs. Evidence for co-evolution with sialic acid synthesis pathways. J. Biol. Chem. 275, 22127–22135.

Angata, T., Varki, A., 2000b. Siglec-7: a sialic acid-binding lectin of the immunoglobulin superfamily. Glycobiology 10, 431–438.

Angata, T., Varki, N.M., Varki, A., 2001. A second uniquely human mutation affecting sialic acid biology. J. Biol. Chem. 276, 40282–40287.

Cooper, D.N., 1999. Pseudogenes and their formation. In: Cooper, D.N. (Ed.). Human Gene Evolution, BIOS Scientific Publishers, Oxford, pp. 265–293.

Cornish, A.L., Freeman, S., Forbes, G., Ni, J., Zhang, M., Cepeda, M., Gentz, R., Augustus, M., Carter, K.C., Crocker, P.R., 1998. Characterization of siglec-5, a novel glycoprotein expressed on myeloid cells related to CD33. Blood 92, 2123–2132.

Falco, M., Biassoni, R., Bottino, C., Vitale, M., Sivori, S., Augugliaro, R., Moretta, L., Moretta, A., 1999. Identification and molecular cloning of p75/AIRM1, a novel member of the sialoadhesin family that functions as an inhibitory receptor in human natural killer cells. J. Exp. Med. 190, 793–802.

Fitch, D.H., Bailey, W.J., Tagle, D.A., Goodman, M., Sieu, L., Slightom, J.L., 1991. Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc. Natl. Acad. Sci. USA 88, 7396–7400.

Floyd, H., Ni, J., Cornish, A.L., Zeng, Z., Liu, D., Carter, K.C., Steel, J., Crocker, P.R., 2000. Siglec-8. A novel eosinophil-specific member of the immunoglobulin superfamily. J. Biol. Chem. 275, 861–866.

Foussias, G., Yousef, G.M., Diamandis, E.P., 2000a. Identification and molecular characterization of a novel member of the siglec family (SIGLEC9). Genomics 67, 171–178.

Foussias, G., Yousef, G.M., Diamandis, E.P., 2000b. Molecular characterization of a Siglec8 variant containing cytoplasmic tyrosine-based motifs, and mapping of the Siglec8 gene. Biochem. Biophys. Res. Commun. 278, 775–781.

Foussias, G., Taylor, S.M., Yousef, G.M., Tropak, M.B., Ordon, M.H., Diamandis, E.P., 2001. Cloning and molecular characterization of two splice variants of a new putative member of the siglec-3-like subgroup of siglecs. Biochem. Biophys. Res. Commun. 284, 887–899.

Iida, Y., 1990. Quantification analysis of 5′-splice signal sequences in mRNA precursors. Mutations in 5′-splice signal sequence of human beta-globin gene and beta-thalassemia. J. Theor. Biol. 145, 523–533.

Karin, M., Richards, R.I., 1982. Human metallothionein genes–primary structure of the metallothionein- II gene and a related processed gene. Nature 299, 797–802.

Kozak, M., 1991. An analysis of vertebrate mRNA sequences: intimations of translational control. . J. Cell Biol. 115, 887–903.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, N., Zhang, W., Wan, T., Zhang, J., Chen, T., Yu, Y., Wang, J., Cao, X., 2001. Cloning and characterization of Siglec-10, a novel sialic acid binding member of the Ig superfamily, from human dendritic cells. J. Biol. Chem. 276, 28106–28112.

Martinez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andres, A., Bertranpetit, J., 2001. Sequence variability of a human pseudogene. Genome Res. 11, 1071–1085.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. Vertebrate pseudogenes. FEBS Lett. 468, 109–114.

Mucklow, S., Hartnell, A., Mattei, M.G., Gordon, S., Crocker, P.R., 1995. Sialoadhesin (Sn) maps to mouse chromosome 2 and human chromosome 20 and is not linked to the other members of the sialoadhesin family, CD22, MAG, and CD33. Genomics 28, 344–346.

Munday, J., Kerr, S., Ni, J., Cornish, A.L., Zhang, J.Q., Nicoll, G., Floyd, H., Mattei, M.G., Moore, P., Liu, D., Crocker, P.R., 2001. Identification, characterization and leucocyte expression of Siglec-10, a novel human sialic acid-binding receptor. Biochem. J. 355, 489–497.

Patel, N., Brinkman-Van der Linden, E.C., Altmann, S.W., Gish, K., Balasubramanian, S., Timans, J.C., Peterson, D., Bell, M.P., Bazan, J.F., Varki, A., Kastelein, R.A., 1999. OB-BP1/Siglec-6. a leptin- and sialic acid-binding protein of the immunoglobulin superfamily. J. Biol. Chem. 274, 22729–22738.

Scarpulla, R.C., 1984. Processed pseudogenes for rat cytochrome c are preferentially derived from one of three alternate mRNAs. Mol. Cell. Biol. 4, 2279–2288.

Tchilian, E.Z., Beverley, P.C., Young, B.D., Watt, S.M., 1994. Molecular cloning of two isoforms of the murine homolog of the myeloid CD33 antigen. Blood 83, 3188–3198.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 19, 253–272.

Whitney, G., Wang, S., Chang, H., Cheng, K.Y., Lu, P., Zhou, X.D., Yang, W.P., McKinnon, M., Longphre, M., 2001. A new siglec family member, siglec-10, is expressed in cells of the immune system and has signaling properties similar to CD33. Eur. J. Biochem. 268, 6083–6096.

Yousef, G.M., Diamandis, E.P., 1999. The new kallikrein-like gene, KLK-L2. Molecular characterization, mapping, tissue expression, and hormonal regulation. J. Biol. Chem. 274, 37511–37516.

Yousef, G.M., Diamandis, E.P., 2000. The expanded human kallikrein gene family: locus characterization and molecular cloning of a new member, KLK-L3 (KLK9). Genomics 65, 184–194.

Yousef, G.M., Diamandis, E.P., 2001. The new human tissue kallikrein gene family: structure, function, and association to disease. Endocr. Rev. 22, 184–204.

Yousef, G.M., Chang, A., Scorilas, A., Diamandis, E.P., 2000. Genomic organization of the human kallikrein gene family on chromosome 19q13.3–q13.4. Biochem. Biophys. Res. Commun. 276, 125–133.

Yousef, G.M., Bahraj, B.S., Yu, H., Poulopoulos, J., Diamandis, E.P., 2001a. Sequence analysis of the human kallikrein gene locus identifies a unique polymorphic minisatellite element. Biochem. Biophys. Res. Commun. In press.

Yousef, G.M., Magklara, A., Chang, A., Jung, K., Katsaros, D., Diamandis, E.P., 2001b. Cloning of a new member of the human kallikrein gene family, klk14, which is down-regulated in different malignancies. Cancer Res. 61, 3425–3431.

Yousef, G.M., Ordon, M.H., Diamandis, E.P., 2001c. Molecular characterization, tissue expression, and mapping of a novel siglec-like gene (slg2) with three splice variants. Biochem. Biophys. Res. Commun. 284, 900–910.

Yousef, G.M., Scorilas, A., Jung, K., Ashworth, L.K., Diamandis, E.P., 2001d. Molecular cloning of the human kallikrein 15 gene (KLK15). Up-regulation in prostate cancer. J. Biol. Chem. 276, 53–61.

Yu, Z., Maoui, M., Wu, L., Banville, D., Shen, S., 2001. mSiglec-E, a novel mouse CD33-related siglec (sialic acid-binding immunoglobulin-like lectin) that recruits Src homology 2 (SH2)-domain- containing protein tyrosine phosphatases SHP-1 and SHP-2. Biochem. J. 353, 483–492.

Zhang, J.Q., Nicoll, G., Jones, C., Crocker, P.R., 2000. Siglec-9. A novel sialic acid binding member of the immunoglobulin superfamily expressed broadly on human blood leukocytes. J. Biol. Chem. 275, 22121–22126.