Biochemical and Biophysical Research Communications **284,** 900–910 (2001)
doi:10.1006/bbrc.2001.5053, available online at http://www.idealibrary.com on IDEAL®

Molecular Characterization, Tissue Expression, and Mapping of a Novel Siglec-like Gene (SLG2) with Three Splice Variants

George M. Yousef,*,† Michael H. Ordon,* George Foussias, and Eleftherios P. Diamandis*,†,[1]
**Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada; and †Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada*

Received May 7, 2001

The sialic acid binding immunglobulin-like lectin (Siglec) family is a recently described member of the immunoglobulin superfamily. Within the Siglec family, there exists a subgroup, which bears a high degree of homology with the molecule CD33 (Siglec-3), and has thus been designated the CD33-like subgroup of Siglecs. Members of this subgroup have been localized to chromosome 19q13.4. Through the positional candidate approach, we identified a novel potential member of this subgroup of Siglecs. We have characterized the complete genomic structure of this gene, determined its chromosomal localization, its homology to other members of the Siglec family, and its tissue expression profile. This new Siglec-like gene is comprised of 11 exons, with 10 intervening introns, and is localized 278 kb telomeric to Siglec-9 and 35 kb centromeric to Siglec-8 and on chromosome 19q13.4. The coding region consists of 2094 base pairs, and encodes for a putative 76.6 kDa protein. All Siglec-conserved structural features, including V-set domain, three C-set domains, transmembrane domain, ITIM and SLAM motifs, were found in this Siglec-like gene. Also, it has the conserved amino acids essential for sialic acid binding. The Siglec-like gene has 40–66% homology with members of the CD33-like subgroup, including Siglecs 5–9. Through RT-PCR we have examined the expression profile of this new gene in a panel of human tissues and found it to be primarily expressed in the bone marrow, spleen, brain, small intestine, colon, and spinal cord. We were also able to identify three different splice variants of the new gene. This gene may represent the latest novel member of the CD33-like subgroup of Siglecs, and, given its high degree of homology, it may also serve a regulatory role in the proliferation and survival of a particular hematopoietic stem cell lineage, as has been found for CD33 and Siglec-7. © 2001 Academic Press

Key Words: **Siglec; Siglec gene family; CD33-like subgroup; immunoglobulin superfamily; sialoadhesins; immunoreceptor tyrosine kinase inhibition motif (ITIM); alternative splicing; gene mapping.**

Sialic acids are a family of α-keto acids with 9-carbon backbones that are expressed abundantly in animals of the deuterosome lineage (1). Siglecs (sialic acid-binding immunoglobulin-like lectins) are a family of the sialic acid recognizing lectins that have been recently defined (2). Members of this family are transmembrane polypeptides with an N-terminal Ig V-set domain, a variable number of C2 set domains, a transmembrane domain and a cytoplasmic tail. The V-set Ig-like domain is the most important in carbohydrate recognition, and the second Ig-like domain may also contribute to the binding (3).

Among the Siglecs, members of the CD33 subgroup have expanded tremendously during the past few years. Proteins of this subgroup, including Siglec-3, and 5 (4), 6 (5), 7 (6, 7), 8 (8), 8-L (9), 9 (10, 11) and 10 (12), share a high degree of sequence homology to Siglec-3, and are clustered on chromosome 19q13.4 just telomeric to the kallikrein family of genes (9, 11, 13). Also, all members of the CD33-like subgroup possess two characteristic tyrosine-based motifs, with the exception of Siglec-8 (8). The first of these contains the consensus immunoreceptor tyrosine kinase inhibitory motif (ITIM), and the second motif displays homology

GenBank Submission No. AY029277.

Abbreviations used: 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; SLAM, signaling lymphocyte activation molecule; SAP, SLAM-associated protein; SH2, src homology 2; SHP, SH2 domain-containing protein tyrosine phosphatase; SHIP, SH2 domain-containing inositol phosphatase; SLG, Siglec-like gene (AF277806); SLG2, Siglec-like gene described in this study; EST, expressed sequence tag.

[1] To whom correspondence and reprint requests should be addressed at Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. Fax: 416-586-8628. E-mail: ediamandis@mtsinai.on.ca.

0006-291X/01 $35.00

TABLE 1

Primers Used for Reverse Transcription Polymerase Chain Reaction (RT-PCR) Analysis

Gene	Primer name	Sequence[1]	Primer direction
Putative new Siglec-like gene (SLG2)	PS-F1	AACAGGCCTGTCTCAGGCAG	Forward
	PS-FM	CTGGAAAACCTTGGGAACGG	Forward
	PS-F2	CATTCTCCAACGGAGCGTTT	Forward
	PS-RU	CTCCTCTGTGCTGCTGACAG	Reverse
	PS-RM	CTCTGCCTGGCCCTGATCAT	Reverse
	PS-CR	CGGAATCAGAAAGCCACACC	Reverse
	PS-R1	CCTAGGATGATGCTGGGTGT	Reverse
ACTIN	ACTINS	ACAATGAGCTGCGTGTGGCT	Forward
	ACTINAS	TCTCCTTAATGTCACGCACGA	Reverse

[1] All primer sequences are written in the 5′–3′ direction.

to the SLAM (signaling lymphocyte activation molecule) tyrosine-based motif. Using the positional candidate gene approach, we previously identified 3 members of the Siglec family of genes (9, 11 and GenBank Accession No. AF277806). In this paper we describe the cloning of a putative new member of the Siglec family of genes, and describe its genomic structure, chromosomal localization and tissue expression pattern. We also describe three splice variants of this gene, of which two of them are predicted to encode for an active protein.

MATERIALS AND METHODS

Gene prediction programs. We used a number of computer programs 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.

Identification of the new gene. Through analysis of the chromosomal region 19q13.4, we have previously cloned 3 new genes that belong to the CD33 subgroup of Siglecs (9, 11 and GenBank submission No. AF277806). Overlapping bacterial artificial chromosome (BAC) clones spanning this area were identified by screening of a human BAC library using different radiolabeled gene-specific probes. Restriction digestion analysis and end sequencing were used to align these clones in the proper orientation, by comparing the results with the *Eco*RI restriction map and the raw sequences of this chromosomal region (available from the Lawrence Livermore National Laboratory, LLNL). A BAC clone that extends more telomerically (BC 799776) from Siglecs 8 and 9 was chosen for further analysis. Initially, we analyzed contigs of linear genomic sequences from this clone, which are available from the LLNL web site, to predict the presence of novel genes. Next, bioinformatic approaches, as previously described (14, 15), were used to identify a putative new Siglec. The sequence of the putative gene was then verified by different experimental approaches, including sequencing of fragments of this BAC clone, EST database search and PCR screening of tissues, as described below.

Expressed sequence tag (EST) searching. The predicted exons of the putative new gene were subjected to homology search using the BLASTN algorithm (16) 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 IMAGE consortium (17) through Research Genetics Inc. (Huntsville, AL). The clones were propagated, purified as described elsewhere (18) and sequenced from both directions with an automated sequencer, using insert-flanking vector primers.

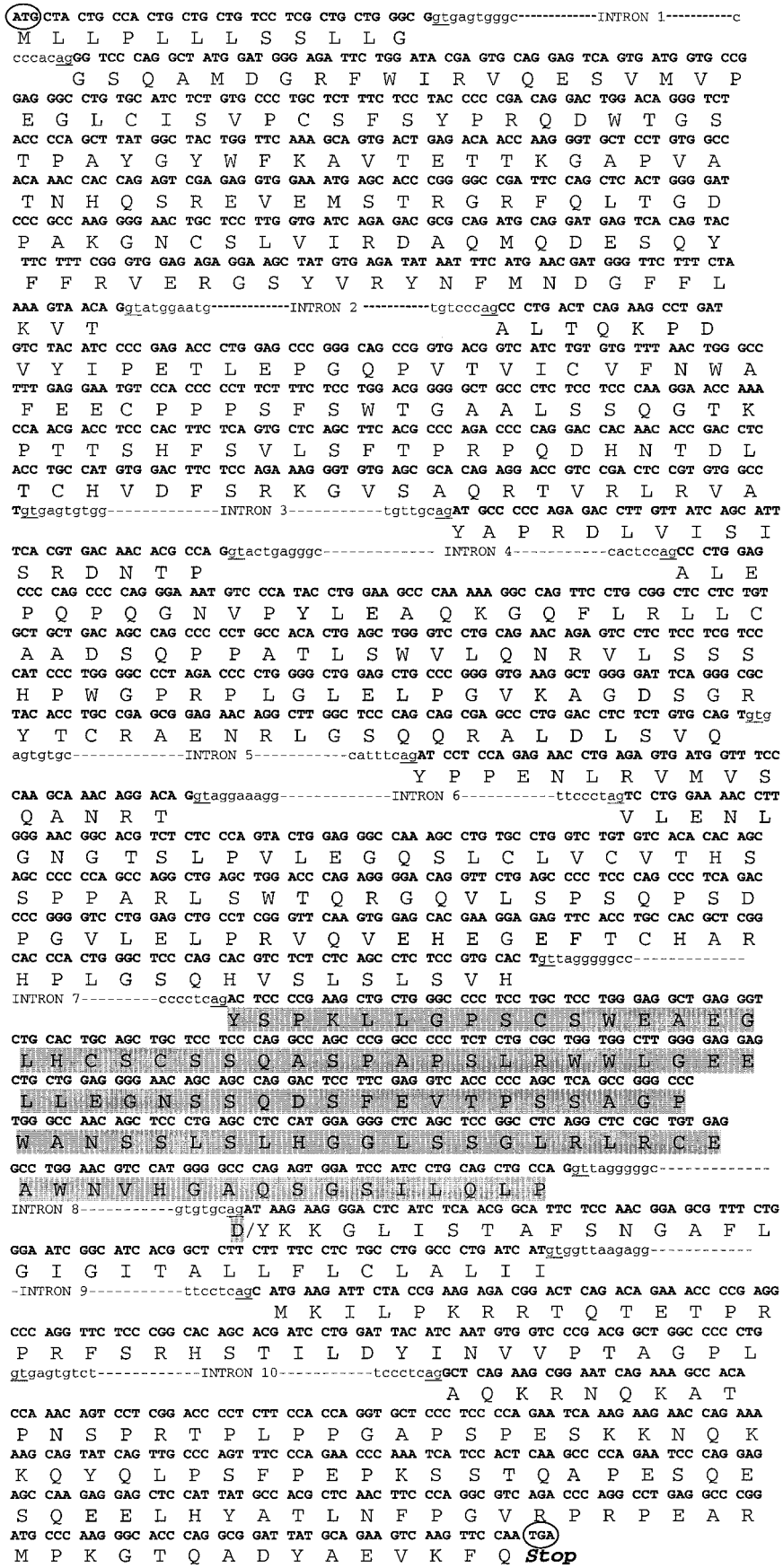
Reverse transcriptase–polymerase chain reaction (RT-PCR). 2 μg of total RNA was reverse-transcribed into first strand cDNA using the Superscript preamplification system (Gibco BRL). The final volume was 20 μl. Based on the combined information obtained from the predicted genomic structure of the new gene and the EST sequences, different combinations of forward primers (PS-F1, PS-FM, PS-F2) and reverse primers (PS-RU, PS-RM, PS-R1, PS-CR) were used to obtain the full mRNA sequence of the gene (Table 1). PCR was carried out in a reaction mixture containing 1 μl of cDNA, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 200 μM dNTPs (deoxynucleoside triphosphates), 150 ng of primers and 2.5 units of HotStarTaq DNA polymerase (Qiagen Inc., Valencia, CA) on a Perkin–Elmer 9600 thermal cycler. The cycling conditions were 95°C for 15 min to activate the *Taq* DNA polymerase, followed by 40–45 cycles of 94°C for 30 s, 60–65°C (depending on primer combinations) for 30 s, 72°C for 1 min and a final extension step at 72°C for 10 min. Equal amounts of PCR products were electrophoresed on 1.5% agarose gels and visualized by ethidium bromide staining. To verify the identity of the PCR products, they were cloned into the pCR 2.1-TOPO vector (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The inserts were sequenced from both directions using vector-specific primers, with an automated DNA sequencer.

Tissue expression. Total RNA isolated from 25 different human tissues was purchased from Clontech (Palo Alto, CA). We prepared cDNA as described above for the tissue culture experiments and used it for PCR reactions. Tissue cDNAs were amplified at various dilutions using two gene-specific primers (PS-FM and PS-RM) that recognize both the long and short forms of the gene (Table 1). Due to the high degree of homology between Siglecs, and to exclude nonspecific amplification, PCR products were verified by sequencing.

TABLE 2

EST Clones with >98% Homology to Our Siglec-like Gene, SLG2

GenBank No.	Tissue of origin	IMAGE ID	Gene form
AI880327	Brain—astrocytoma	1957140	Splice V-2
BE858523	Brain—glioblastoma	3308709	Splice V-2
BF663289	Primary B-cells from tonsils	4297851	Long form


ATG CTA CTG CCA CTG CTG CTG TCC TCG CTG CTG GGC Ggtgagtgggc----------INTRON 1----------c
M L L P L L L S S L L G
cccacagGG TCC CAG GCT ATG GAT GGG AGA TTC TGG ATA CGA GTG CAG GAG TCA GTG ATG GTG CCG
G S Q A M D G R F W I R V Q E S V M V P
GAG GGC CTG TGC ATC TCT GTG CCC TGC TCT TTC TCC TAC CCC CGA CAG GAC TGG ACA GGG TCT
E G L C I S V P C S F S Y P R Q D W T G S
ACC CCA GCT TAT GGC TAC TGG TTC AAA GCA GTG ACT GAG ACA ACC AAG GGT GCT CCT GTG GCC
T P A Y G Y W F K A V T E T T K G A P V A
ACA AAC CAC CAG AGT CGA GAG GTG GAA ATG AGC ACC CGG GGC CGA TTC CAG CTC ACT GGG GAT
T N H Q S R E V E M S T R G R F Q L T G D
CCC GCC AAG GGG AAC TGC TCC TTG GTG ATC AGA GAC GCG CAG ATG CAG GAT GAG TCA CAG TAC
P A K G N C S L V I R D A Q M Q D E S Q Y
TTC TTT CGG GTG GAG AGA GGA AGC TAT GTG AGA TAT AAT TTC ATG AAC GAT GGG TTC TTT CTA
F F R V E R G S Y V R Y N F M N D G F F L
AAA GTA ACA Ggtatggaatg----------INTRON 2 ----------tgtcccagCC CTG ACT CAG AAG CCT GAT
K V T A L T Q K P D
GTC TAC ATC CCC GAG ACC CTG GAG CCC GGG CAG CCG GTG ACG GTC ATC TGT GTG TTT AAC TGG GCC
V Y I P E T L E P G Q P V T V I C V F N W A
TTT GAG GAA TGT CCA CCC CCT TCT TTC TCC TGG ACG GGG GCT GCC CTC TCC TCC CAA GGA ACC AAA
F E E C P P P S F S W T G A A L S S Q G T K
CCA ACG ACC TCC CAC TTC TCA GTG CTC AGC TTC ACG CCC AGA CCC CAG GAC CAC AAC ACC GAC CTC
P T T S H F S V L S F T P R P Q D H N T D L
ACC TGC CAT GTG GAC TTC TCC AGA AAG GGT GTG AGC GCA CAG AGG ACC GTC CGA CTC CGT GTG GCC
T C H V D F S R K G V S A Q R T V R L R V A
Tgtgagtgtgg----------INTRON 3----------tgttgcagAT GCC CCC AGA GAC CTT GTT ATC AGC ATT
Y A P R D L V I S I
TCA CGT GAC AAC ACG CCA Ggtactgagggc---------- INTRON 4----------cactccagCC CTG GAG
S R D N T P A L E
CCC CAG CCC CAG GGA AAT GTC CCA TAC CTG GAA GCC CAA AAA GGC CAG TTC CTG CGG CTC CTC TGT
P Q P Q G N V P Y L E A Q K G Q F L R L L C
GCT GCT GAC AGC CAG CCC CCT GCC ACA CTG AGC TGG GTC CTG CAG AAC AGA GTC CTC TCC TCG TCC
A A D S Q P P A T L S W V L Q N R V L S S S
CAT CCC TGG GGC CCT AGA CCC CTG GGG CTG GAG CTG CCC GGG GTG AAG GCT GGG GAT TCA GGG CGC
H P W G P R P L G L E L P G V K A G D S G R
TAC ACC TGC CGA GCG GAG AAC AGG CTT GGC TCC CAG CAG CGA GCC CTG GAC CTC TCT GTG CAG Tgtg
Y T C R A E N R L G S Q Q R A L D L S V Q
agtgtgc----------INTRON 5----------catttcagAT CCT CCA GAG AAC CTG AGA GTG ATG GTT TCC
Y P P E N L R V M V S
CAA GCA AAC AGG ACA Ggtaggaaagg----------INTRON 6----------ttccctagTC CTG GAA AAC CTT
Q A N R T V L E N L
GGG AAC GGC ACG TCT CTC CCA GTA CTG GAG GGC CAA AGC CTG TGC CTG GTC TGT GTC ACA CAC AGC
G N G T S L P V L E G Q S L C L V C V T H S
AGC CCC CCA GCC AGG CTG AGC TGG ACC CAG AGG GGA CAG GTT CTG AGC CCC TCC CAG CCC TCA GAC
S P P A R L S W T Q R G Q V L S P S Q P S D
CCC GGG GTC CTG GAG CTG CCT CGG GTT CAA GTG GAG CAC GAA GGA GAG TTC ACC TGC CAC GCT CGG
P G V L E L P R V Q V E H E G E F T C H A R
CAC CCA CTG GGC TCC CAG CAC GTC TCT CTC AGC CTC TCC GTG CAC Tgttagggggcc----------
H P L G S Q H V S L S L S V H
INTRON 7---------ccctcagAC TCC CCG AAG CTG CTG GGC CCC TCC TGC TCC TGG GAG GCT GAG GGT
Y S P K L L G P S C S W E A E G
CTG CAC TGC AGC TGC TCC TCC CAG GCC AGC CCG GCC CCC TCT CTG CGC TGG TGG CTT GGG GAG GAG
L H C S C S S Q A S P A P S L R W W L G E E
CTG CTG GAG GGG AAC AGC AGC CAG GAC TCC TTC GAG GTC ACC CCC AGC TCA GCC GGG CCC
L L E G N S S Q D S F E V T P S S A G P
TGG GCC AAC AGC TCC CTG AGC CTC CAT GGA GGG CTC AGC TCC GGC CTC AGG CTC CGC TGT GAG
W A N S S L S L H G G L S S G L R L R C E
GCC TGG AAC GTC CAT GGG GCC CAG AGT GGA TCC ATC CTG CAG CTG CCA Ggttagggggc----------
A W N V H G A Q S G S I L Q L P
INTRON 8----------gtgtgcagAT AAG AAG GGA CTC ATC TCA ACG GCA TTC TCC AAC GGA GCG TTT CTG
D/Y K K G L I S T A F S N G A F L
GGA ATC GGC ATC ACG GCT CTT CTT TTC CTC TGC CTG GCC CTG ATC ATgtggttaagagg----------
G I G I T A L L F L C L A L I I
-INTRON 9----------ttcctcagC ATG AAG ATT CTA CCG AAG AGA CGG ACT CAG ACA GAA ACC CCG AGG
M K I L P K R R T Q T E T P R
CCC AGG TTC TCC CGG CAC AGC ACG ATC CTG GAT TAC ATC AAT GTG GTC CCG ACG GCT GGC CCC CTG
P R F S R H S T I L D Y I N V V P T A G P L
gtgagtgtct----------INTRON 10----------tccctcagGCT CAG AAG CGG AAT CAG AAA GCC ACA
A Q K R N Q K A T
CCA AAC AGT CCT CGG ACC CCT CTT CCA CCA GGT GCT CCC TCC CCA GAA TCA AAG AAG AAC CAG AAA
P N S P R T P L P P G A P S P E S K K N Q K
AAG CAG TAT CAG TTG CCC AGT TTC CCA GAA CCC AAA TCA TCC ACT CAA GCC CCA GAA TCC CAG GAG
K Q Y Q L P S F P E P K S S T Q A P E S Q E
AGC CAA GAG GAG CTC CAT TAT GCC ACG CTC AAC TTC CCA GGC GTC AGA CCC AGG CCT GAG GCC CGG
S Q E E L H Y A T L N F P G V R P R P E A R
ATG CCC AAG GGC ACC CAG GCG GAT TAT GCA GAA GTC AAG TTC CAA TGA
M P K G T Q A D Y A E V K F Q Stop

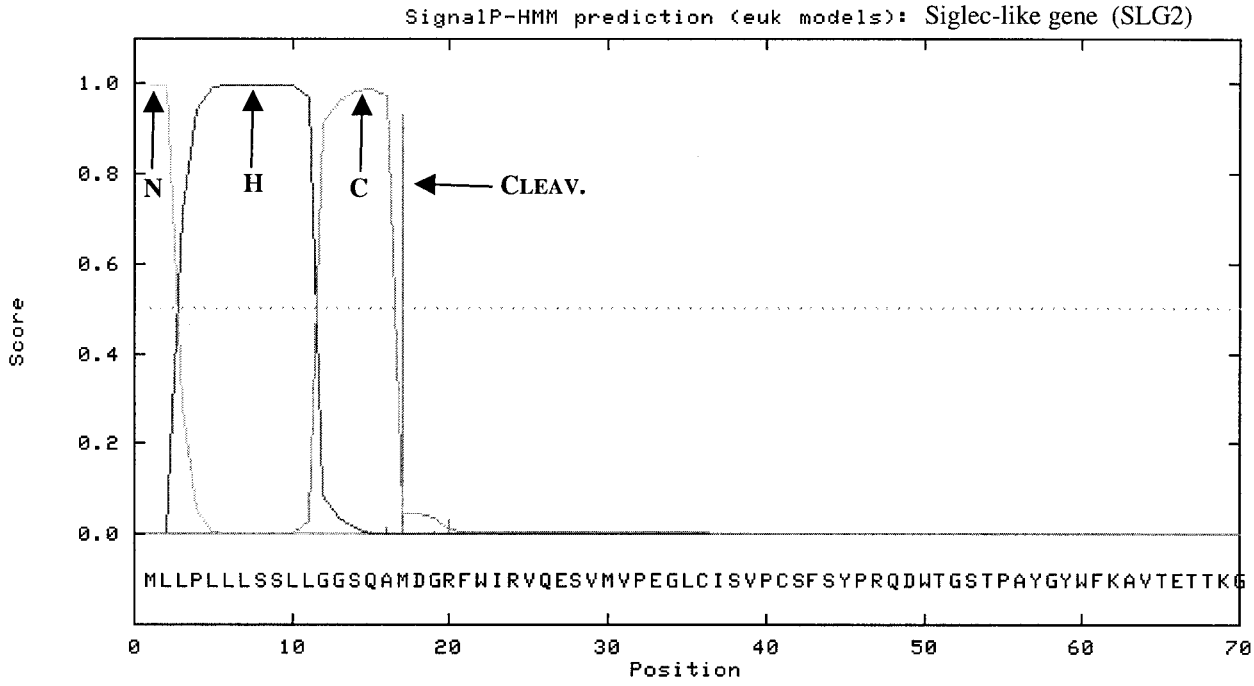


FIG. 2. Plot illustrating the different structural regions of the signal peptide of the new Siglec-like gene SLG2. N, basic N-terminal region; H, central hydrophobic region; C, polar C-terminal region; Cleav., cleavage site (SQA ↓ MD).

Structure analysis. Multiple alignment was performed using the "Clustal X" software package and the multiple alignment program available from the Baylor College of Medicine (Houston, TX). 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.

RESULTS

Cloning of the Novel Siglec-like Gene (SLG2)

Computer analysis of BAC clone (BC 799776) revealed the presence of a putative novel gene that shows a high degree of homology with other members of the Siglec family of genes. Homology searches, carried out using the BLAST algorithm against the human EST database, revealed the presence of 3 EST clones with >98% identity with the predicted exons (Table 2). These ESTs were obtained, purified and sequenced. The sequence of the new gene was further confirmed by RT-PCR, using a panel of different tissues as templates.

Genomic Organization of the Siglec-like Gene (SLG2)

The SLG2 gene spans an area of 6,554 bp of genomic sequence on chromosome 19q13.4. The gene is formed of 11 exons and 10 intervening introns. Exon lengths are 37, 384, 285, 48, 270, 48, 258, 285, 94, 112 and 273bp (Fig. 1). All intron/exon splice sites are closely related to the consensus splice sites (-m**GT**AAGT . . . C**AG**m-, where m is any base) (19).

The open reading frame of the long form of the gene consists of 2094bp, and encodes for a putative 697 amino acid polypeptide with a predicted molecular weight of 76.6 kDa, excluding any posttranslational modifications. The predicted translation initiation codon (ATG) is located at position 1664 (numbers refer to our GenBank submission No. AY029277). The sequences surrounding this start codon match with a Kozak consensus sequence for translation initiation, especially the most highly conserved purine at position

FIG. 1. Genomic organization and partial genomic sequence of the Siglec-like gene SLG2. Intronic sequences are not shown except for the splice junctions. Introns are shown with lowercase letters and exons with capital letters. For full sequence, see GenBank Accession No. AY029277. The start and stop codons are encircled and the exon–intron junctions are underlined. The translated amino acids of the coding region are shown underneath by a single letter abbreviation. Amino acids of exon 8, which are missing in the short form of the gene are highlighted in gray.

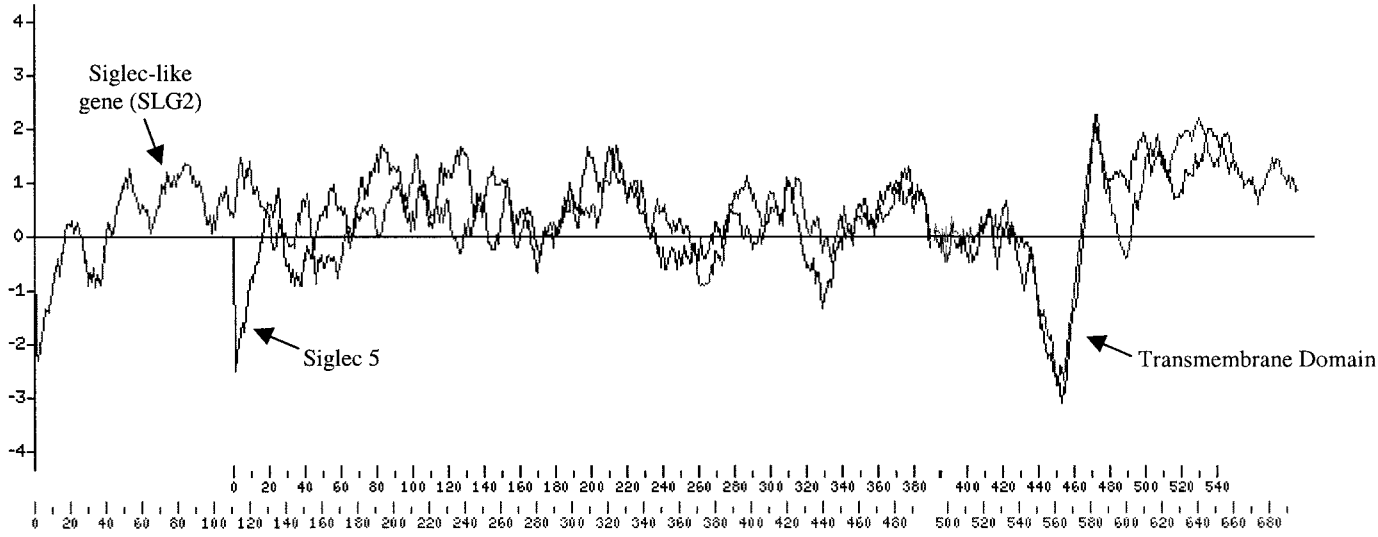


FIG. 3. Plot comparing the hydrophobicity and hydrophilicity patterns of the Siglec-like gene SLG2 and Siglec-5. Note that the N-terminal region is quite hydrophobic. The hydrophobic transmembrane domain is recognized close to the C-terminal region.

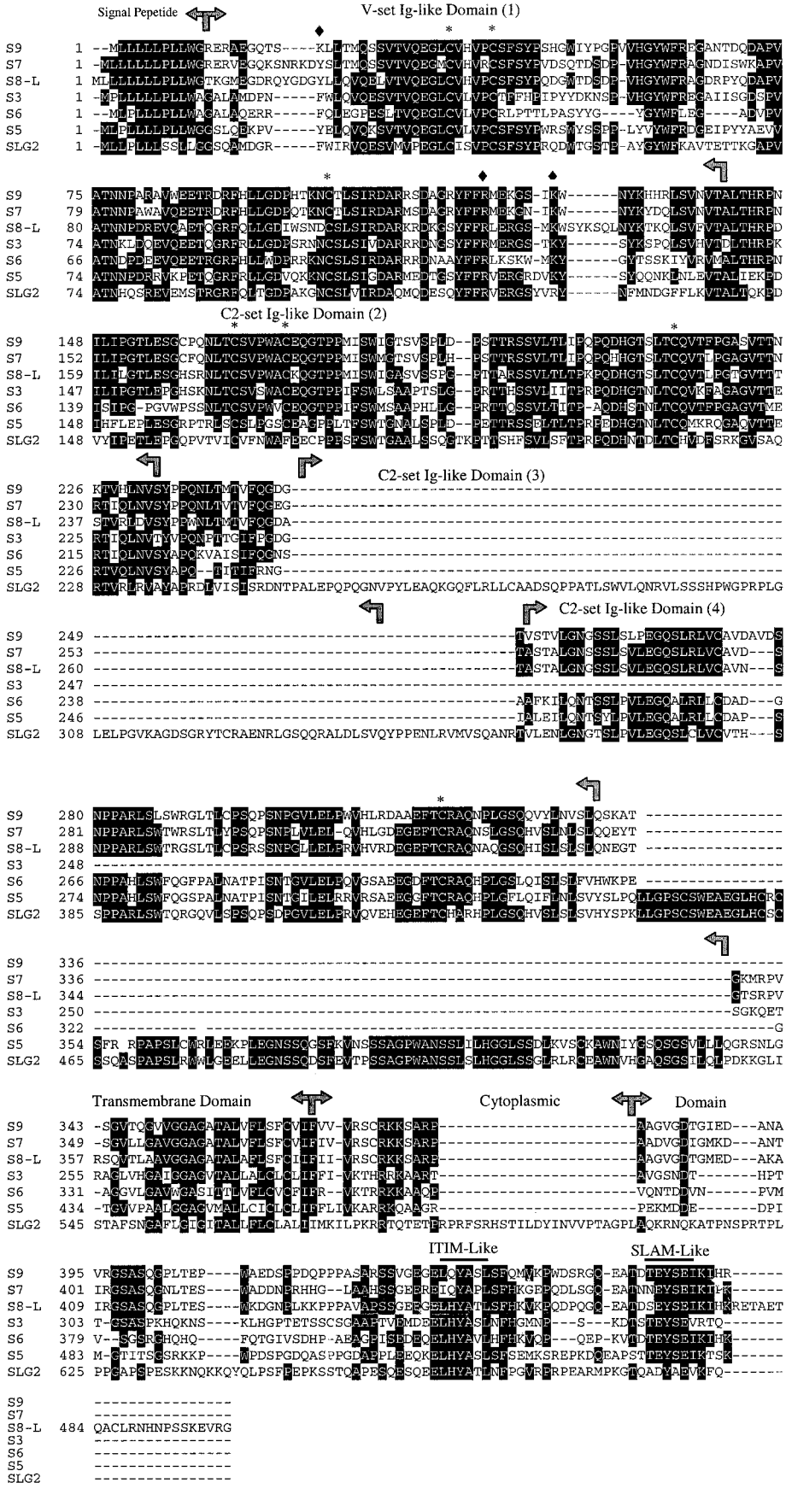
−3 that occurs in 97% in eukaryotic mRNAs (20). In addition, there is a conserved "C" in position +4, that is found in other members of this family, including Siglecs 5–9 (20). Furthermore, using this start codon, the resultant protein product shows extensive homology with other members of the CD33-like subgroup of Siglecs (see below). No other initiation codon was found that produced a long, continuous open reading frame.

Structural analysis indicated, like other members of this subgroup, that the SLG2 gene 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 (SQA-MD) (Fig. 2). The presence of a signal sequence is also supported by the hydrophobicity pattern of the protein (Fig. 3). Conserved Domain (CD) and ProDom searches together with homology alignment indicated the presence of 3 conserved immunoglobulin domains (residues: 27–132, 164–237 and 370–443), representing the V-set domain, followed by the 2 C-set domains present in other CD33-like Siglecs (Fig. 4). In addition, an extra Ig domain, which is formed mainly by exon 5 (amino acid residues 262–341) was identified, representing an extra C-set domain. The single transmembrane domain, predicted by TMpred and also evident in the Kyte–Doolittle hydrophobicity plot (Figs. 3 and 5), is in keeping with observations for other members of this subgroup. Furthermore, the new putative Siglec contains the two characteristic tyrosine-based motifs, ITIM and SLAM-like, also noted in other members of the CD33-like subgroup of Siglecs (Fig. 4).

Homology with Other Siglecs

Although the structure of SLG2 is unique, it has a high degree of homology with other members of CD33-like subgroup of Siglecs. At the protein level, it shows 48% identity and 66% similarity with Siglec-5, and 40–60% similarity with Siglecs 6–9. Multiple alignment was used to compare the protein sequence of the Siglec-like gene with other members of the CD33-like subgroup and the other Siglec family members. As is evident in Fig. 4, there is conservation of the key cysteine residues that are responsible for the characteristic folding of the extracellular Ig-like domains in all Siglecs (21, 22), except for one cysteine at position 170, which might be an alignment artifact because of the presence of another closely located cysteine residue (position 173). The residues believed to be responsible for sialic acid binding, in particular the critical arginine at position 119 and the two aromatic residues at positions 21 and 128 on the A and G strands of the V-set domains (23), are also conserved in the new

FIG. 4. Multiple alignment of the amino acid sequences of members of the CD33-like subgroup of siglecs. The newly characterized putative Siglec (SLG2) was aligned with Siglec-3 and Siglecs-5 to -9. Numbers of the amino acid residues of each protein are shown on the left of each row. Identical residues are highlighted in black. The signal peptide was determined through computer prediction (see also Figs. 2 and 3). 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-like and SLAM-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 (◆).


Signal Pepetide
V-set Ig-like Domain (1)
C2-set Ig-like Domain (2)
C2-set Ig-like Domain (3)
C2-set Ig-like Domain (4)
Transmembrane Domain
Cytoplasmic
Domain
ITIM-Like
SLAM-Like
S9
S7
S8-L
S3
S6
S5
SLG2

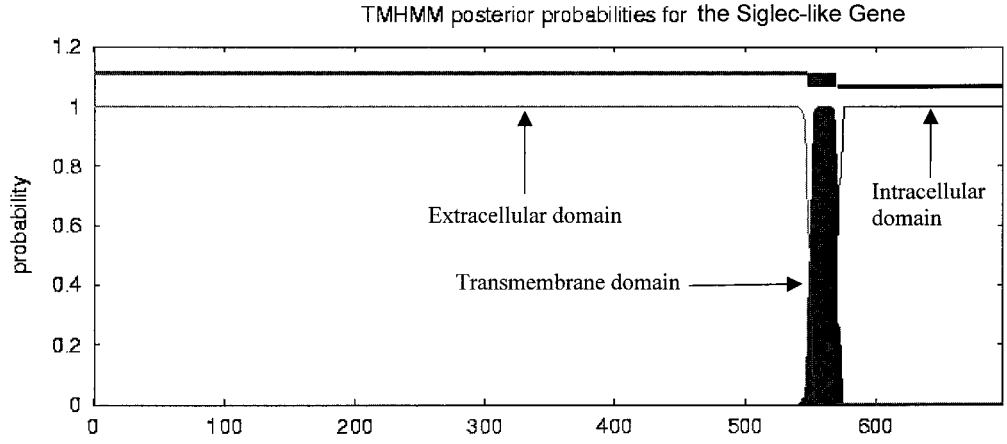


FIG. 5. A plot illustrating the predicted domains of Siglec-like gene SLG2 using the "TMHMM" server. The extracellular, transmembrane and intracellular domains are shown.

gene (Fig. 4). Multiple alignment indicates that exon 8 of the gene encodes for a unique domain that is highly homologous to a similar domain in the Siglec-5 gene, but was not found in other members of this subfamily of Siglecs.

Phylogenetic Analysis

Phylogenetic analysis was done for all reported human Siglecs using both parsimony and distance matrix methods. Results indicate a close association of the Siglec-like gene SLG2 with other members of the CD33 subgroup of Siglecs (Fig. 6). As expected from the high sequence identity mentioned above, the new gene is most closely related to Siglec-5, and shows a somewhat more distant relationship with other, non-CD33, Siglecs.

Splice Variants of the Siglec-like Gene (SLG2)

In our attempts to reveal the full genomic structure of the gene, multiple bands were obtained from some tissues during RT-PCR analysis. Sequencing of these PCR products revealed the presence of two additional splice variants of the gene in addition to the form described above (which is referred to as the "long form" in our GenBank submission). The different splice variants of the gene are represented schematically in Fig. 7. The short form of the gene is the same as the long form, except for a missing exon 8. Multiple alignment shows that exon 8 is a unique exon that occurs only in the long form of the new gene and the Siglec-5 gene, but not in any of the other members of this siglec subfamily. There is a high degree of sequence homology of this exon in these two genes. Removal of this exon in the short form of the gene results in a smaller predicted protein product that has all the conserved features of this subfamily. In the third splice variant of the gene (splice variant-2, according to our GenBank submission) exons 2 and 3 are combined together with the intronic area in-between, and the same for exons 4 and 5, 7–9, and exon 10 is extended. This form is predicted to encode for a truncated protein of 1086 amino acids.

Tissue Expression

The tissue expression profile for the three splice forms of the SLG2 gene was elucidated by performing RT-PCR using total RNA from 25 normal human tissues. The long form of the gene was found to be relatively highly expressed in bone marrow, spleen, spinal cord, brain, colon and small intestine. Lower expression was apparent in a few other tissues (Fig. 8). The short form was expressed mainly in the bone marrow, spinal cord, spleen and brain, but at lower levels compared with the long form. Splice variant-2 was expressed mainly in the small intestine and to a lower degree in a few other tissues (Fig. 8).

Mapping and Chromosomal Localization of SLG2

Restriction digestion analysis of three overlapping BAC clones spanning the chromosomal region 19q13.4 enabled us to precisely locate the relative position of the SLG2 gene in relation to other Siglecs in the region (Fig. 9). Also, guided by the EcoRI restriction map of the region available from the LLNL, the three BAC clones were aligned in the correct orientation, and thus we were able to establish the direction of transcription of all known Siglec genes in this area. The new gene is transcribed from the telomere towards the centromere, and is located 34,702 bp more centromeric to Siglec-8, which is transcribed in the same direction. Another newly identified Siglec gene (GenBank Accession No. AF277806) is located 32,925 bp further telomerically and transcribed in the same direction. On the centromeric side, Siglec-9 is the closest gene to the new Siglec, separated by 278 kb and running on the oppo-

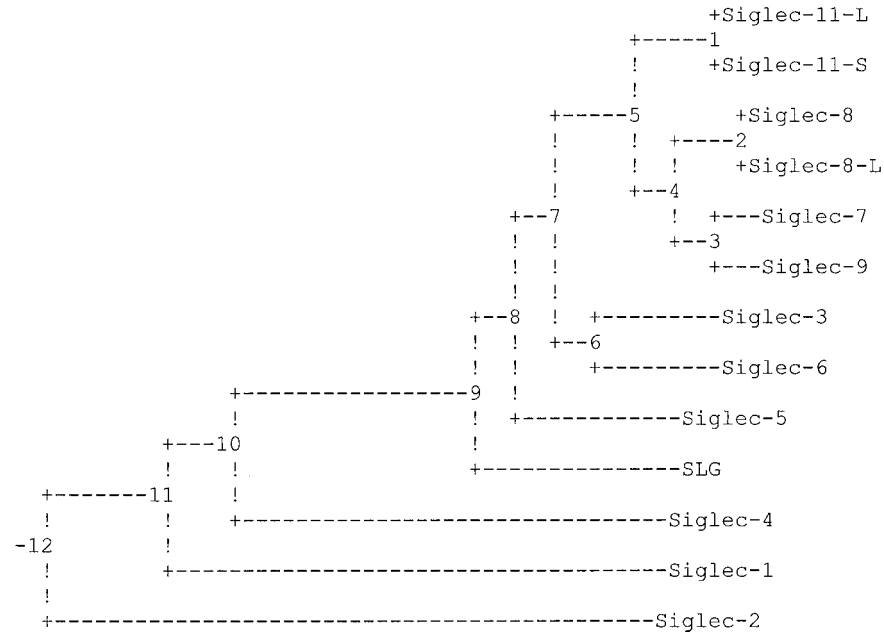


FIG. 6. Dendrogram of the predicted phylogenetic tree for human siglec proteins using the UPGMA method. As expected, the tree grouped the CD33-like group of siglecs together. Siglec-5 is the most related protein to the Siglec-like gene (SLG2).

site direction, followed by KLK14 (a member of the kallikrein family of genes) (24) which is located 43,187 bp more centromeric to Siglec-9.

DISCUSSION

The positional candidate cloning is a new approach for gene discovery that combines the knowledge of map position with the increasingly dense human transcript maps and the available expressed sequence tags (ESTs) (25). Using this approach, we were able to clone a new putative Siglec gene.

Although experimental evidences are lacking, this gene likely represents a new member of the expanding CD33-like subgroup of Siglecs. As is evident in Fig. 4, the predicted protein contains all the structural characteristics possessed by other CD33-like Siglecs discovered so far. It has the distinctive distribution of cysteine residues found in all Siglecs, necessary for the unique folding pattern of their Ig-like domains (21, 22). There is also conservation of key amino acids involved in sialic acid binding, specifically the conserved arginine and the two aromatic residues on the A and G strands of the V-set domains (23).

The cytoplasmic tyrosine-based motifs, ITIM and SLAM-like, found in all other members of the CD33-like subgroup of Siglecs, are also present in our putative member of this subgroup. These motifs have been the focus of investigations in order to elucidate the functional role of these Siglecs within the cell. The primary emphasis has been on the ITIM motif, which has been found to be involved in recruitment of the tyrosine phosphatases SHP1 and 2, and the inositol phosphatases SHIP1 and 2 (26–28). Siglec7, originally identified as a natural killer cell inhibitory receptor, was found to recruit the tyrosine phosphatase SHP1 following tyrosine phosphorylation of its ITIM motif,

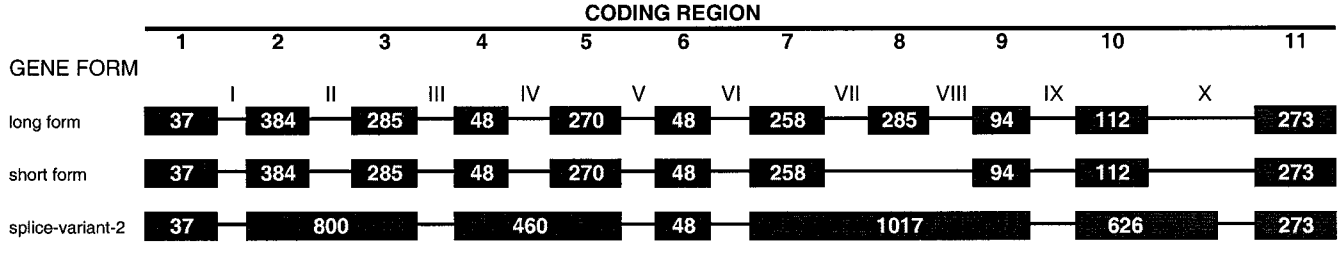


FIG. 7. A diagram representing the 3 splice variants of the Siglec-like gene SLG2. Exons are represented by solid boxes with lengths mentioned in base pairs and introns by the connecting lines. Full genomic sequences of all splice variants are available in GenBank Accession No. AY029277.

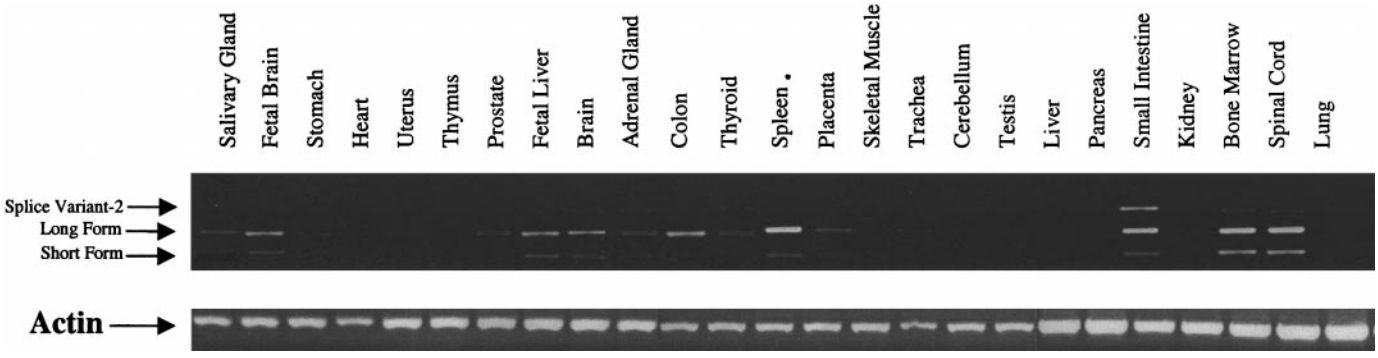


FIG. 8. Tissue expression of the Siglec-like gene SLG2 as determined by RT-PCR. Actin was used as a control gene.

leading to the inhibition of natural killer cell cytotoxicity (29). In addition, CD33 has also been found to recruit SHP1 and 2, both *in vitro* and *in vivo*, as a result of phosphorylation of the tyrosine in its ITIM motif (30). Further, mutation of this tyrosine results in increased red blood cell binding by CD33-expressing COS cells. More recently, it has been reported that engagement of Siglec-7 and CD33 with monoclonal antibodies results in the inhibition of proliferation of both normal and leukemic myeloid cells *in vitro* (31). Although Siglec-7 was initially thought to be expressed exclusively in natural killer cells, it has also been found in myeloid cells, at a later stage of differentiation than CD33. The observed inhibitory effects are believed to be the result of phosphorylation of the ITIM motif present in the cytoplasmic domains of both CD33 and Siglec-7. These findings suggest that recruitment of SHP1 and SHP2 by members of the CD33-like subgroup of Siglecs may serve to: (i) inhibit the activating signaling pathways that lead to cell proliferation and survival; and (ii) to modulate the receptor's ligand-binding activity (30).

More supportive evidence that this new gene is a member of the CD33-like subfamily of Siglecs comes from the chromosomal localization of the gene. As is the case with other members of this subfamily, the new gene is located in the chromosomal region 19q13.4 in close proximity to Siglecs-8, 9 and SLG (Fig. 9). Interestingly, this chromosomal region harbors a large number of other hematopoitically expressed Ig superfamily members. These include a family of genes encoding killer cell inhibitory receptors expressed on natural killer cells and subsets of T-lymphocytes and immunoglobulin-like transcripts (ILT-1, 2 and 3) expressed on myeloid cells (32). An emerging theme for these receptors is that they are involved in either stimulatory or inhibitory signalling functions in hematopoietic cells.

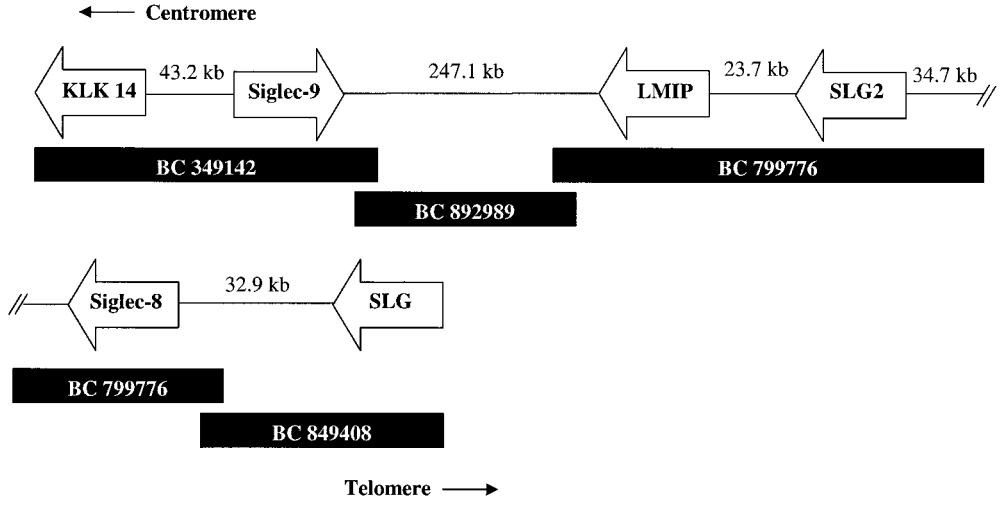


FIG. 9. The relative locations of some members of the CD33-like subgroup of Siglecs on chromosome 19q13.4. This partial map of the region was constructed by aligning 4 overlapping BAC clones (shown as solid black boxes). Genes are represented by horizontal arrows denoting the direction of transcription. Distances between genes are mentioned in kilobases. LMIP, lens fiber membrane intrinsic protein (GenBank Accession No. P55344). SLG2, Siglec like gene described in this paper. SLG (GenBank Accession No. AF277806). KLK14, kallikrein gene 14 (24). Figure is not drawn to scale.

Our tissue expression results indicate that the gene is expressed, at the mRNA level, in a variety of tissues; high levels of expression are found in tissues where the hemopoietic cell lineage are found, e.g., spleen and bone marrow. This is consistent with the expression patterns of other members of the Siglec-3-like subgroup of Siglecs. In addition, the EST clone that was found to encode for the long form was isolated from B-cells.

The removal of intervening RNA sequences (introns) from the pre-messenger RNA in eukaryotic nuclei is a major step in the regulation of gene expression (33). RNA splicing provides a mechanism whereby protein isoform diversity can be generated and the expression of particular proteins with specialized functions can be restricted to certain cell or tissue types during development (33). The sequence elements in the pre-mRNA at the 5′ and 3′ splice sites in metazoans have very loose consensus sequence; only the first and the last two bases (GT . . . AG) of the introns are highly conserved (18). These sequences cannot be the sole determinants of splice site selection, since identical, but not ordinarily active, consensus sequences can be found within both exons and introns of many eukaryotic genes. Other protein factors and sequences downstream of the splice sites are also involved.

Here, we describe the isolation of 2 additional splice variants of the Siglec-like gene SLG2 (in addition to the long form). The existence of multiple splice forms is not unpreceeded among Siglecs. We have recently isolated a long splice-variant of Siglec-8 (9), and also a Siglec-like gene, SLG, with two splice variants (GenBank Accession No. AF277806).

A major challenge for the future will be to elucidate the function of the CD33-like subgroup of sialic acid binding receptors and to determine the significance of sialic acid binding.

REFERENCES

1. Schauer, R. (1982) Chemistry, metabolism, and biological functions of sialic acids. *Adv. Carbohydr. Chem. Biochem.* **40,** 131–234.
2. Crocker, P. R., Clark, E. A., Filbin, M., Gordon, S., Jones, Y., Kehrl, J. H., Kelm, S., Le Douarin, N., Powell, L., Roder, J., Schnaar, R. L., Sgroi, D. C., Stamenkovic, K., Schauer, R., Schachner, M., van den Berg, T. K., van der Merwe, P. A., Watt, S. M., and Varki, A. (1998) Siglecs: A family of sialic-acid binding lectins [letter]. *Glycobiology* **8,** v.
3. Vinson, M., van der Merwe, P. A., Kelm, S., May, A., Jones, E. Y., and Crocker, P. R. (1996) Characterization of the sialic acid-binding site in sialoadhesin by site-directed mutagenesis. *J. Biol. Chem.* **271,** 9267–9272.
4. Cornish, A. L., Freeman, S., Forbes, G., Ni, J., Zhang, M., Cepeda, M., Gentz, R., Augustus, M., Carter, K. C., and Crocker, P. R. (1998) Characterization of siglec-5, a novel glycoprotein expressed on myeloid cells related to CD33. *Blood* **92,** 2123–2132.
5. 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., and 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.
6. Angata, T., and Varki, A. (2000) Siglec-7: A sialic acid-binding lectin of the immunoglobulin superfamily. *Glycobiology* **10,** 431–438.
7. Nicoll, G., Ni, J., Liu, D., Klenerman, P., Munday, J., Dubock, S., Mattei, M. G., and Crocker, P. R. (1999) Identification and characterization of a novel Siglec, Siglec-7, expressed by human natural killer cells and monocytes. *J. Biol. Chem.* **274,** 34089–34095.
8. Floyd, H., Ni, J., Cornish, A. L., Zeng, Z., Liu, D., Carter, K. C., Steel, J., and Crocker, P. R. (2000) Siglec-8. A novel eosinophil-specific member of the immunoglobulin superfamily. *J. Biol. Chem.* **275,** 861–866.
9. Foussias, G., Yousef, G. M., and Diamandis, E. P. (2000) Molecular characterization of a Siglec8 variant containing cytoplasmic tyrosine-based motifs, and mapping of the Siglec8 gene. *Biochem. Biophys. Res. Commun.* **278,** 775–781.
10. Zhang, J. Q., Nicoll, G., Jones, C., and 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.
11. Foussias, G., Yousef, G. M., and Diamandis, E. P. (2000) Identification and molecular characterization of a novel member of the siglec family (SIGLEC9). *Genomics* **67,** 171–178.
12. Munday, J., Kerr, S., Ni, J., Cornish, A. L., Zhang, J. Q., Nicoll, G., Floyd, H., Mattei, M. G., Moore, P., Liu, D., and Crocker, P. R. (2001) Identification, characterization and leucocyte expression of Siglec-10, a novel human sialic acid-binding receptor. *Biochem. J.* **355,** 489–497.
13. Yousef, G. M., Chang, A., Scorilas, A., and 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.
14. Yousef, G. M., Obiezu, C. V., Luo, L. Y., Black, M. H., and Diamandis, E. P. (1999) Prostase/KLK-L1 is a new member of the human kallikrein gene family, is expressed in prostate and breast tissues, and is hormonally regulated. *Cancer Res.* **59,** 4252–4256.
15. Yousef, G. M., and 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.
16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
17. Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M. B. (1996) The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33,** 151–152.
18. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory, NY.
19. 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.
20. Kozak, M. (1991) An analysis of vertebrate mRNA sequences: Intimations of translational control. *J. Cell Biol.* **115,** 887–903.
21. Crocker, P. R., Kelm, S., Hartnell, A., Freeman, S., Nath, D., Vinson, M., and Mucklow, S. (1996) Sialoadhesin and related

cellular recognition molecules of the immunoglobulin superfamily. *Biochem. Soc. Trans.* **24,** 150–156.

22. Pedraza, L., Owens, G. C., Green, L. A., and Salzer, J. L. (1990) The myelin-associated glycoproteins: Membrane disposition, evidence of a novel disulfide linkage between immunoglobulin-like domains, and posttranslational palmitylation. *J. Cell Biol.* **111,** 2651–2661.
23. van der Merwe, P. A., Crocker, P. R., Vinson, M., Barclay, A. N., Schauer, R., and Kelm, S. (1996) Localization of the putative sialic acid-binding site on the immunoglobulin superfamily cell-surface molecule CD22. *J. Biol. Chem.* **271,** 9273–9280.
24. Yousef, G. M., Magklara, A., Chang, A., Jung, K., Katsaros, D., and Diamandis, E. P. (2001) Cloning of a new member of the human kallikrein gene family, klk14, which is down-regulated in different malignancies. *Cancer Res.* **61,** 3425–3431.
25. Collins, F. S. (1995) Positional cloning moves from perditional to traditional [published erratum appears in *Nat. Genet.* **11**(1), 104, 1995]. *Nat. Genet.* **9,** 347–350.
26. Borges, L., Hsu, M. L., Fanger, N., Kubin, M., and Cosman, D. (1997) A family of human lymphoid and myeloid Ig-like receptors, some of which bind to MHC class I molecules. *J. Immunol.* **159,** 5192–5196.
27. Le Drean, E., Vely, F., Olcese, L., Cambiaggi, A., Guia, S., Krystal, G., Gervois, N., Moretta, A., Jotereau, F., and Vivier, E. (1998) Inhibition of antigen-induced T cell response and antibody-induced NK cell cytotoxicity by NKG2A: Association of NKG2A with SHP-1 and SHP-2 protein-tyrosine phosphatases [published erratum appears in *Eur. J. Immunol.* **28**(3), 1122, 1998]. *Eur. J. Immunol.* **28,** 264–276.
28. Muraille, E., Bruhns, P., Pesesse, X., Daeron, M., and Erneux, C. (2000) The SH2 domain containing inositol 5-phosphatase SHIP2 associates to the immunoreceptor tyrosine-based inhibition motif of Fc gammaRIIB in B cells under negative signaling. *Immunol. Lett.* **72,** 7–15.
29. Falco, M., Biassoni, R., Bottino, C., Vitale, M., Sivori, S., Augugliaro, R., Moretta, L., and 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.
30. Taylor, V. C., Buckley, C. D., Douglas, M., Cody, A. J., Simmons, D. L., and Freeman, S. D. (1999) The myeloid-specific sialic acid-binding receptor, CD33, associates with the protein-tyrosine phosphatases, SHP-1 and SHP-2. *J. Biol. Chem.* **274,** 11505–11512.
31. Vitale, C., Romagnani, C., Falco, M., Ponte, M., Vitale, M., Moretta, A., Bacigalupo, A., Moretta, L., and Mingari, M. C. (1999) Engagement of p75/AIRM1 or CD33 inhibits the proliferation of normal or leukemic myeloid cells. *Proc. Natl. Acad. Sci. USA* **96,** 15091–15096.
32. Dupont, B., Selvakumar, A., and Steffens, U. (1997) The killer cell inhibitory receptor genomic region on human chromosome 19q13.4. *Tissue Antigens* **49,** 557–563.
33. Adams, M. D., Rudner, D. Z., and Rio, D. C. (1996) Biochemistry and regulation of pre-mRNA splicing. *Curr. Opin. Cell Biol.* **8,** 331–339.