IDENTIFICATION, CHARACTERIZATION, AND MAPPING OF NOVEL MEMBERS OF THE SIGLEC FAMILY

by

George Foussias

A thesis submitted in conformity with the requirements for the Degree of Master of Science
Graduate Department of Laboratory Medicine and Pathobiology
Faculty of Medicine
University of Toronto

©Copyright by George Foussias 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-58828-9
IDENTIFICATION, CHARACTERIZATION, AND MAPPING OF NOVEL MEMBERS OF THE SIGLEC FAMILY

George Foussias
Master of Science, 2001
Department of Laboratory Medicine and Pathobiology
Faculty of Medicine
University of Toronto

ABSTRACT

The sialic acid binding immunoglobulin-like lectin (Siglec) family is a recently described member of the immunoglobulin superfamily. Within this Siglec family there exists a subgroup of molecules which bear a very high degree of homology with the molecule Siglec-3 (CD33), thus designated the Siglec-3-like subgroup of Siglec. Members of this subgroup have been localized to chromosome 19q13.4, primarily through in situ hybridization. Through our investigation of chromosome 19q13.3-13.4, we identified, and precisely characterized three genes belonging to this Siglec-3-like subgroup. We identified the novel Siglec-9, located adjacent to the human Kallikrein gene family on chromosome 19q13.4, and the previously reported Siglec-8 gene downstream of Siglec-9, for which we characterized a novel transcript. Finally, we identified and mapped another novel Siglec-like gene, and characterized two splice variants. From preliminary studies, all Siglec-3-like subgroup members may play an inhibitory role of haemopoietic cell proliferation, and have some utility in antineoplastic therapy.
ACKNOWLEDGEMENTS

This work could not have been accomplished without the guidance, enthusiasm, optimism, and continual support of my supervisor, Dr. Eleftherios P. Diamandis. Invaluable assistance was also provided by Dr. George M. Yousef and Steve M. Taylor. Discussions with Linda Grass, Angeliki Magklara, Albert Chang, Christina Obiezu, Dr. Bhupinder Bharaj, and Dr. Andreas Scorilas provided valuable suggestions and insights. I would also like to extend my gratitude to my thesis advisory committee, Dr. Dwayne Barber and Dr. Jeremy Squire, for their guidance and support. I am also indebted to my family and friends for their tremendous support and optimism.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1. The Immunoglobulin Superfamily</td>
<td>1</td>
</tr>
<tr>
<td>1.2. The Siglec Family of Genes</td>
<td>3</td>
</tr>
<tr>
<td>1.3. Evolution of the Siglec Family</td>
<td>5</td>
</tr>
<tr>
<td>1.4. The Physiological Roles of SiglecS</td>
<td></td>
</tr>
<tr>
<td>1.4.1. Siglec-1 (Sialoadhesin)</td>
<td>8</td>
</tr>
<tr>
<td>1.4.2. Siglec-2 (CD22)</td>
<td>9</td>
</tr>
<tr>
<td>1.4.3. Siglec-4a (Myelin-Associated Glycoprotein; MAG)</td>
<td>11</td>
</tr>
<tr>
<td>1.4.4. The Siglec-3-like Subgroup of SiglecS</td>
<td>12</td>
</tr>
<tr>
<td>1.5. Alternative Splicing of Siglec mRNA Transcripts</td>
<td>15</td>
</tr>
<tr>
<td>1.6. Siglec Specificities for Sialic Acids</td>
<td>17</td>
</tr>
<tr>
<td>2. Rationale</td>
<td>19</td>
</tr>
<tr>
<td>3. Hypothesis</td>
<td>20</td>
</tr>
<tr>
<td>4. Materials and Methods</td>
<td>21</td>
</tr>
<tr>
<td>4.1. Siglec-9</td>
<td></td>
</tr>
<tr>
<td>4.1.1. New Gene Identification</td>
<td>21</td>
</tr>
</tbody>
</table>
4.1.2. Expressed Sequence Tag (EST) Identification

4.1.3. Molecular Characterization of Siglec-9

4.1.4. Mapping and Chromosomal Localization of Siglec-9

4.1.5. Siglec-9 Tissue Expression

4.2. Siglec-8-Long (Siglec-8-L)

4.2.1. Identification of the Genomic Area containing Siglec-8

4.2.2. Molecular Characterization of Siglec-8-L

4.3. Siglec-Like Gene (SLG)

4.3.1. Identification of SLG

4.3.2. Cloning and Molecular Characterization of Two SLG Splice Forms

4.3.3. Tissue Expression of Two SLG Splice Variants

4.3.4. Mapping and Chromosomal Localization of SLG

4.3.5. Phylogenetic Analysis of the Expanded Siglec Family

5. RESULTS

5.1. Siglec-9

5.1.1. Identification of Siglec-9 on 19q13.4

5.1.2. Characterization of the Genomic Structure of the Siglec-9 Gene and its Protein Product

5.1.3. Mapping and Chromosomal Localization of Siglec-9

5.1.4. Homology with other Siglec Family Members

5.1.5. Tissue Expression Profile of Siglec-9

5.2. Siglec-8-Long (Siglec-8-L)

5.2.1. Identification of the Genomic Area Containing Siglec-8
5.2.2. Molecular Characterization of Siglec-8-L 48

5.3. Siglec-Like Gene (SLG)

5.3.1. Identification of SLG 53

5.3.2. Cloning and Molecular Characterization of Two SLG Splice Forms 53

5.3.3. Tissue Expression Profile of Two SLG Splice Variants 59

5.3.4. Mapping and Chromosomal Localization of SLG 65

5.3.5. Phylogenetic Analysis of the Expanded Siglec Family 68

6. DISCUSSION 70

7. REFERENCES 83
LIST OF FIGURES

1. Genomic Structure of Siglec-9 36
2. Hydrophobicity Plot of Siglec-9 38
3. Localization of the Siglec-9 Gene 40
4. Siglec-9 Multiple Alignment with other Siglec-3-like Subgroup Members 43
5. Tissue Expression Profile of Siglec-9 46
6. Genomic Organization of the Siglec-8 Gene 50
7. Protein Sequence Alignment for the Siglec-3-like Subgroup of Siglecs 51
8. Nucleotide and Protein Sequences of Both SLG Isoforms 55
9. Genomic Organization of the SLG Gene 58
10. Protein Sequence Alignment of Both SLG Splice Variants and the Siglec-3-like Subgroup of Siglecs 61
11. Hydrophobicity Plots of Both SLG Isoforms 64
12. Tissue Expression Profile of Both SLG Splice Forms 66
13. Physical Map of Chromosome 19q13.4 and the Siglec-3-like Subgroup Locus 67
14. Phylogenetic Analysis of the Siglec Family 69
LIST OF TABLES

1. Genomic Organization of Siglec-9 34
2. Overall Homology of Siglec-9 with the Siglec-3-like Subgroup of Siglecs 42
3. SLG Isoform Sequence Homology with the Siglec-3-like Subgroup of Siglecs 60
1. INTRODUCTION

1.1. The Immunoglobulin Superfamily

The immunoglobulin superfamily (IgSF) encompasses a large heterogeneous group of proteins with a diversity of functions. Initially, these molecules were thought to be involved only in immune responses, but with the expansion of this superfamily over the years it has become clear that members of the IgSF also play vital roles in the control of cellular behaviour in a wide variety of cells and biological systems. Despite the tremendous diversity in this superfamily, all of these molecules possess a characteristic structural feature, the immunoglobulin (Ig) fold (Williams and Barclay 1988; Halaby and Mornon 1998). Proteins possessing this characteristic Ig-fold are widely distributed in nature, having been identified in vertebrates, invertebrates, bacteria, viruses, fungi and plants (Halaby and Mornon 1998). As alluded to earlier, IgSF molecules have a diversity of functions including immune responses, as growth factor receptors, tumoral markers, as well as having enzymatic activity, being involved in neural cell development, the maintenance of tissue integrity, antitumoral activity, viral and bacterial pathogenesis, and cell adhesion (reviewed in (Halaby and Mornon 1998)). Despite this broad range of function, the common denominator appears to be involvement in homophilic or heterophilic recognition and adhesion processes (Williams and Barclay 1988; Halaby and Mornon 1998).

Given the wide species distribution of the Ig-fold, the evolutionary origin of this structural feature remains unclear. Initially it was believed that this characteristic fold arose from gene duplication of a primordial domain encoding about 100 amino acids (Williams and Barclay 1988; Williams, Davis et al. 1989). The functional diversity
observed within this superfamily was believed to have arisen through the process of divergent evolution. In keeping with this belief, the vast species distribution of the Ig-fold may have been the result of vertical or horizontal gene transfer, possibly through a viral vector (Halaby and Momon 1998). However, another possibility is that the Ig-fold present in so many species has arisen through convergent evolution. It may simply represent an energetically favourable structural framework, utilized by different species for a variety of biological functions (reviewed in (Halaby and Momon 1998)).

The diverse molecules belonging to the IgSF, as mentioned above, are characterized by the presence of an Ig-fold, a sandwich of two $\beta$-sheets stabilized by a conserved disulfide bond. The core of this domain is composed of $\beta$-strands A,B,E in one sheet and G,F,C in the other, and arise from the ends of the domain sequence (reviewed in (Williams and Barclay 1988)). In between, however, there is a great deal of sequence length variation, often including additional $\beta$-strands. These Ig domains occur in two types, the variable (V)-set and the constant (C)-set, and can be distinguished based on patterns of conserved amino acid residues responsible for forming the characteristic $\beta$-sheet sandwich. V-set domains consist of about 65-75 amino acid residues between conserved cysteines with four $\beta$-strands in each $\beta$-sheet, whereas C-set domains have about 55-60 residues giving rise to sheets of four and three $\beta$-strands (reviewed in (Williams and Barclay 1988)). The C-set domains can be further divided into C1- and C2-sets, and are distinguished by the fact that, although showing signs of a C-set domain, the latter half of C2-set domains exhibit sequence patterns more homologous to V-set rather than C1-set domains (Williams, Davis et al. 1989).
1.2. The Siglec Family of Genes

Recently, a novel family of structurally related IgSF molecules has been identified, which mediates protein-carbohydrate interactions through specific interactions with sialic acid-containing glycoproteins and glycolipids (Crocker, Kelm et al. 1996). This family was originally referred to as the sialoadhesins, but has recently been designated the sialic acid-binding Ig-like lectin (Siglec) family (Crocker, Clark et al. 1998). These molecules are characterized by the presence of one N-terminal V-set domain, and a variable number of downstream C2-set domains, ranging from 16 in sialoadhesin to 1 in Siglec-3 (Crocker, Kelm et al. 1996). Furthermore, these Ig-like domains possess some unique features. In the V-set domain, the conserved cysteine in β-strand F of classic V-set domains is absent, while a highly conserved cysteine is present in β-strand E in all Siglecs identified so far. This results in the cysteines in β-strands B and E being next to each other in one β-sheet, which likely results in an intrasheet disulfide bond (Williams, Davis et al. 1989; Crocker, Kelm et al. 1996). There is also an additional highly conserved cysteine residue in both the V-set and first C2-set domains of all Siglecs. In the V-set domain it is located at the beginning of β-strand B, while in the C2-set domain it is found between β-strands B and C. These two additional cysteines have been found to form an interdomain disulfide bond, a feature unique to Siglecs (Pedraza, Owens et al. 1990; Crocker, Kelm et al. 1996).

Currently, there exist eight members of the Siglec family in humans, each exhibiting a unique expression pattern. Siglec-1 (sialoadhesin), the prototypic member of this family is expressed exclusively on macrophages (Crocker, Mucklow et al. 1994). Meanwhile, Siglec-2 (CD22) is found on B lymphocytes (Stamenkovic and Seed 1990),
Siglec-3 (CD33) on myeloid progenitor cells and monocytes (Simmons and Seed 1988; Ulyanova, Blasioli et al. 1999) and Siglec-4a (myelin-associated glycoprotein (MAG)) on oligodendrocytes and Schwann cells (Kelm, Schauer et al. 1994). Further, Siglec-5, concurrently identified as a leptin binding protein (OB-binding protein 2) is found on neutrophils (Cornish, Freeman et al. 1998; Patel, Brinkman-Van der Linden et al. 1999), while Siglec-6, also identified as OB-binding protein 1, is found on B lymphocytes (Patel, Brinkman-Van der Linden et al. 1999). Siglec-7, originally identified as the inhibitory receptor p75/AIRM1 is expressed primarily on natural killer cells (Falco, Biassoni et al. 1999; Nicoll, Ni et al. 1999), and Siglec-8 is expressed on the cell surface of eosinophils (Floyd, Ni et al. 2000).

The genes encoding these members of the human Siglec gene family have been localized primarily through in situ hybridization techniques. Siglec-1 has been localized to chromosome 20p13 (Mucklow, Hartnell et al. 1995), while Siglec-2 and Siglec-4a have been mapped to Chromosome 19q13.1 (Barton, Arquint et al. 1987; Wilson, Najfeld et al. 1993). Meanwhile, Siglec-3 and Siglec-5, 6, 7, and 8 have all been localized to chromosome 19q13.4 (Angata and Varki 2000; Zhang, Nicoll et al. 2000). These Siglecs, which are localized to chromosome 19q13.4 exhibit extensive sequence homology between them, and have consequently been designated the Siglec-3-like subgroup of Siglecs (Angata and Varki 2000; Zhang, Nicoll et al. 2000).
1.3. Evolution of the Siglec Family

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-1 has been identified in *Homo sapiens* and *Mus musculus* (house mouse), in addition to its original identification in sheep (Crocker and Gordon 1986; Crocker, Mucklow et al. 1994; Mucklow, Hartnell et al. 1995). Siglec-2 has been found in humans, mice, as well as in several primates, including the orangutan (*Pongo pygmaeus*), the gorilla (*Gorilla gorilla*), the pygmy chimpanzee (*Pan paniscus*) and the chimpanzee (*Pan troglodytes*) (Stamenkovic and Seed 1990; Wilson, Fox et al. 1991; Brinkman-Van der Linden, Sjoberg et al. 2000). Siglec-4a has been identified in humans, mice, and rat (*Rattus norvegicus*) (Lai, Watson et al. 1987; Fujita, Sato et al. 1989). Further, Siglec-4b (Schwann cell myelin protein; SMP), a protein closely related to MAG, with identical binding specificities, has been identified in the quail and chicken (Dulac, Tropak et al. 1992; Kelm, Schauer et al. 1996).

With regards to the Siglec-3-related subgroup of Siglecs, Siglec-3 has been identified in humans and mice (Simmons and Seed 1988; Tchilian, Beverley et al. 1994). Recently, an additional Siglec-3-related Siglec has been reported in the mouse, mSiglec-E, which appears to recruit the tyrosine phosphatases SHP-1 and SHP-2 in a manner similar to that observed in human Siglec-3-related Siglecs (discussed below) (publication in press; GenBank Accession No. AF317298). Further, during their identification and characterization of Siglec-7, Falco *et al.* (1999) 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,
Biassoni et al. 1999). In light of the finding of Siglec-2 in various primates, it is quite likely that they may also possess additional Siglecs. To date, however, there are no other reports of Siglec-3-related Siglecs in any other organisms.

With regards to the human Siglec-3-related subgroup of Siglecs, as mentioned above, they are all localized to chromosome 19q13.4 (Angata and Varki 2000; Zhang, Nicoll et al. 2000). To date, this tight clustering of members of the Siglec family has only been observed in humans. This has raised this possibility that this subgroup has arisen through gene duplication and exon shuffling relatively recently in vertebrate evolution (Angata and Varki 2000). The mechanism proposed involves unequal crossing-over of sister chromatids during meiotic recombination. In fact, chromosome 19-specific minisatellites have been identified in the 19q13.3-qter region, which may have facilitated such gene duplication (Das, Jackson et al. 1987; Angata and Varki 2000). 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 of 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, Bailey et al. 1991)).

Through the determination of the chromosomal localization of members of the Siglec family, as mentioned above, it was noticed that the gene for Siglec-1 is located on mouse chromosome 2 and human chromosome 20 (Mucklow, Hartnell 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, Hartnell et al. 1995).

Further investigations into the evolution of the Siglec family have led to studies involving two extensively studied 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 2000). Interestingly, these species also lack the enzymes necessary for the synthesis of sialic acids. Based on these findings, the authors suggest 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 2000).
1.4. The Physiological Roles of Siglecs

1.4.1. Siglec-1 (Sialoadhesin)

Siglec-1 was initially identified as a sheep erythrocyte receptor and later found to be an Ig-like membrane protein restricted to macrophage cells, especially in bone marrow (Crocker and Gordon 1986). Further studies revealed that Siglec-1 is highly enriched at contact sites between macrophages and developing myeloid cells (Crocker, Werb et al. 1990). Cell binding experiments revealed that Siglec-1 preferentially bound myeloid cells from bone marrow (Crocker, Freeman et al. 1995). However, binding was also detected between Siglec-1 and lymphocytes derived from spleen and lymph nodes (van den Berg, Breve et al. 1992). Given the observed distribution within bone marrow macrophages and the cell binding results, it has been hypothesized that Siglec-1 is involved in the development of myeloid cells in bone marrow and in trafficking of leukocytes in lymphatic organs (Kelm, Schauer et al. 1996). Through studies with breast cancer tissues and cell lines, Siglec-1-expressing macrophages have been found to be closely associated with these cancer cells (Nath, Hartnell et al. 1999). Examination of this association revealed that Siglec-1 bound, in a sialic-acid dependent manner, to the mucin MUC-1, which is abundantly expressed on breast cancer cells. The consequences of such an interaction remain unclear, but may be involved in mediating, either positively or negatively, various macrophage effector functions (Nath, Hartnell et al. 1999).
1.4.2. Siglec-2 (CD22)

Siglec-2 has been one of the most extensively studied members of the Siglec family. It is expressed on B lymphocytes, and was initially thought to function as an adhesion molecule mediating both homotypic, B cell-B cell, and heterotypic, B cell-erythrocyte and B cell-monocyte, interactions (Stamenkovic and Seed 1990; Wilson, Fox et al. 1991). Based on these observations, it was hypothesized that Siglec-2 may serve to facilitate antigen recognition by promoting accessory cell adhesion. Further investigations into its physiological function revealed that it is also involved in B cell activation. Engagement of the B cell receptor (BCR) was shown to induce rapid tyrosine phosphorylation in the cytoplasmic domain of Siglec-2 (Schulte, Campbell et al. 1992; Peaker and Neuberger 1993). This raises the possibility that Siglec-2 may be involved in regulating both antigen-mediated B cell activation as well as the cell’s adhesiveness following activation.

Examination of the downstream events following tyrosine phosphorylation of the cytoplasmic tail of Siglec-2 revealed that it becomes associated with the SH2 domain-containing protein tyrosine phosphatase SHP1, the tyrosine kinase Syk, and phospholipase C-γ1 (PLC-γ1) (Law, Sidorenko et al. 1996). These observations suggest that Siglec-2 acts as a scaffolding protein to ensure efficient interaction and activation of PLC-γ1 by Syk. The recruitment of SHP1, however, may serve either of two functions: i) to downregulate signaling by the BCR; or ii) when Siglec-2 binds its extracellular ligand and consequently recruits SHP1 it may sequester the phosphatase away from the BCR, thus allowing enhanced BCR signaling (Law, Sidorenko et al. 1996; Tedder, Tuscano et al. 1997). Furthermore, tyrosine phosphorylated Siglec-2 has been found associated with
the tyrosine kinase Lyn and phosphatidylinositol-3 (PI-3) kinase (Tuscano, Engel et al. 1996). Overall, engagement of Siglec-2 and the consequent activation of these downstream signaling pathways has been shown to enhance B cell proliferation (Tuscano, Engel et al. 1996; Tuscano, Engel et al. 1996). The observed recruitment of SHP1 may serve to terminate the signaling pathways initiated by the associated kinases (Tedder, Tuscano et al. 1997). However, the presence of both immunoreceptor tyrosine kinase activation and inhibition motifs (ITAM and ITIM, respectively) in the cytoplasmic tail of Siglec-2, and the recruitment of both kinases and phosphatases, still leaves doubt as to whether this receptor functions as a positive or negative regulator of B cell activity (Tedder, Tuscano et al. 1997).
1.4.3. Siglec-4a (Myelin-Associated Glycoprotein; MAG)

Siglec-4a has been largely studied in mice where there exists two isoforms, a large and small isoform, L-Siglec-4a and S-Siglec-4a, respectively. Functional studies have revealed that, upon cross-linking of the large isoform (L-Siglec-4a) in mice, the tyrosine kinase Fyn is specifically activated (Umemori, Sato et al. 1994). This L-Siglec-4a associated kinase activity was highest during the initial stages of myelination, with a subsequent decrease in later stages. Fyn and L-Siglec-4a were also co-localized during this early myelination stage, and Fyn-deficient mice exhibited between 50-60% of the myelin per brain compared to wild-type mice (Umemori, Sato et al. 1994). However, no association has been detected for the short isoform, S-Siglec-4a. Further studies in Siglec-4a-deficient mice seem to indicate that, although not essential for myelin formation, it may facilitate glia-axon recognition and contact, enabling the proper formation and maintenance of periaxonal structures and optimal myelin formation (Li, Trapp et al. 1998). Such contact and engagement of Siglec-4a, particularly L-Siglec-4a, may cause the necessary cross-linking required to recruit the kinase Fyn and initiate appropriate intracellular signaling events for formation and maintenance of myelin.
1.4.4. The Siglec-3-like Subgroup of Siglecs

The Siglec-3-like subgroup of Siglecs, which currently includes Siglec-3 and Siglec-5 through -8, has expanded tremendously during the last few years. The receptors comprising this subgroup have also been the focus of several functional studies, albeit to a lesser degree than the Siglec family members discussed above. All the members of the Siglec-3-like subgroup possess two characteristic tyrosine-based motifs, with the possible exception of Siglec-8 (Floyd, Ni et al. 2000). The first of these contains the consensus sequence for the immunoreceptor tyrosine kinase inhibitory motif (ITIM), (ILV)xYxx(LV) (x being any amino acid) (Burshtyn, Yang et al. 1997; Vivier and Daeron 1997). This motif has been shown to be the binding site for the SH2 (src homology 2) domains of the SH2 domain-containing protein tyrosine phosphatases SHP-1 and SHP-2 (Borges, Hsu et al. 1997; Le Drean, Vely et al. 1998), as well as the SH2 domain-containing inositol phosphatases SHIP1 and SHIP2 (Muraille, Bruhns et al. 2000). The second motif displays homology to a tyrosine-based motif, TxYxx(IV), identified in the signaling lymphocyte activation molecule (SLAM) and is responsible for its association with the SLAM-associated protein (SAP), which in turn blocks the binding of SHP-2 to phosphorylated SLAM (Coffey, Brooksbank et al. 1998; Sayos, Wu et al. 1998).

Siglec-3, the prototypic member of this subgroup of Siglecs, was initially identified as a differentiation antigen of the monocytic/myeloid lineage (Griffin, Linch et al. 1984; Dinndorf, Andrews et al. 1986). It has been used for many years to distinguish more primitive myeloblastic forms of acute myelogenous leukemia (AML) from acute lymphocytic leukemia (Griffin, Linch et al. 1984; Dinndorf, Andrews et al. 1986;
Bernstein, Singer et al. 1992). However, its physiological function on the surface of these monocytes and myeloid cells remained unclear. With the identification of additional Siglecs bearing extensive homology to Siglec-3, as well as the progress in understanding the intracellular role of ITIM motifs, investigative efforts have focused again on the possible role of Siglec-3. Taylor et. al. (1999) have recently found that Siglec-3 in fact recruits the protein tyrosine phosphatases SHP1 and SHP2, both in vitro and in vivo. Further, this recruitment is the direct consequence of tyrosine phosphorylation in the ITIM motif. Moreover, the authors showed that mutation of the tyrosine in this motif resulted in increased red blood cell binding by Siglec-3-expressing COS cells (Taylor, Buckley et al. 1999). Pretreatment of wild type Siglec-3-expressing COS cells with a Src family kinase specific inhibitor, which inhibited Siglec-3 phosphorylation, resulted in a similar increase in red blood cell binding (Taylor, Buckley et al. 1999). Taken together, these results suggest that the ITIM motif has a negative regulatory role in Siglec-3-mediated sialic acid-dependent cell interactions.

More recently, Siglec-7, initially designated p75/AIRM1, was identified as a natural killer cell inhibitory receptor (Falco, Biassoni et al. 1999). It was observed that the cytoplasmic domain of Siglec-7, and in particular the ITIM motif, was tyrosine-phosphorylated, resulting in recruitment of SHP1 with a consequent inhibition of natural killer cell cytotoxicity. Although Siglec-7 was initially thought to be restricted to natural killer cells, it was also found to also be expressed in myeloid cells, at a later stage of differentiation than Siglec-3 (Vitale, Romagnani et al. 1999). In addition, Vitale et. al. (1999) found that engagement of Siglec-7 and Siglec-3, through the use of monoclonal antibodies, inhibits the proliferation of both normal and leukemic myeloid cells in vitro.
(Vitale, Romagnani et al. 1999). The observed inhibitory effects are believed to be the result of phosphorylation of the ITIM motif present in the cytoplasmic domain of both Siglec-3 and Siglec-7. These findings suggest that recruitment of SHP1 and SHP2 by these members of the Siglec-3-like subgroup may serve two roles: i) to inhibit the activating signaling pathways that lead to cell proliferation and survival; and ii) to modulate the receptor's ligand-binding activity (Taylor, Buckley et al. 1999).

The remainder of the members of the Siglec-3-like subgroup of Siglecs have not been the subject of any functional investigations to date. This is due primarily to their recent identification and characterization. However, as mentioned previously, with the exception of Siglec-8, they possess the ITIM and SLAM-like motifs. Based on the findings for Siglec-3 and Siglec-7, it is quite likely that these molecules also play an inhibitory role in the particular cell types in which they are expressed. However, detailed functional investigations are required to elucidate their precise roles.
1.5. Alternative Splicing of Siglec mRNA Transcripts

Alternative splice forms have been identified for several members of the Siglec family, in humans and other animals. In the mouse, Siglec-1 is usually found as a single pass transmembrane protein with 17 extracellular Ig-like domains (Crocker, Mucklow et al. 1994). However, two alternative splice forms have been identified, both lacking the transmembrane and cytoplasmic domains (Crocker, Mucklow et al. 1994). The first form possesses only the first 3 Ig-like domains, while the second contains the first 16 Ig-like domains, both of which are found as secreted proteins not associated with the cell surface. The alternative splicing in both cases results in the use of an unspliced intron as a variant exon, introducing a new stop codon with a small amount of novel sequence at the C-termini of these variants (Crocker, Mucklow et al. 1994).

For human Siglec-2, a small form (Siglec-2α) and a large form (Siglec-2β) have been identified (Wilson, Najfeld et al. 1993). Siglec-2α lacks exons 5 and 6, of a total of 15 exons, resulting in the absence of 2 C2-set Ig-like domains. The larger Siglec-2β, commonly referred to simply as Siglec-2, is the form whose physiological role has been extensively studied. In addition, alternative forms have been identified for both murine and rat Siglec-4a, both of which are encoded by a gene comprised of 13 exons. In the mouse, three forms have been identified: L-Siglec-4a, which lacks exon 12; S-Siglec-4a which lacks the non-coding exon 2; and a third form which lacks both exons 2 and 12 (Fujita, Sato et al. 1989). In the rat, only the two forms, L-Siglec-4a and S-Siglec-4a, with splicing pattern identical to those found in the mouse, have been identified (Lai, Watson et al. 1987).
Alternative forms have also been identified for both human and murine Siglec-3. In human Siglec-3, two transcripts of 1.5 kb and 1.8 kb have been found (Simmons and Seed 1988). Although there are no differences in the coding sequence, there is an extended 3' untranslated region, thought to result from the use of an alternative polyadenylation signal. This phenomenon has also been found in murine Siglec-3 transcripts (Tchilian, Beverley et al. 1994). However, in the mouse, there have also been two alternative splice variants identified. These are the result of inclusion of an 83 bp insert in the cytoplasmic domain of some Siglec-3 mRNA transcripts, resulting in a frame-shift and producing unique cytoplasmic domains for the two forms (Tchilian, Beverley et al. 1994). For another member of the Siglec-3-related subgroup of Siglecs, Siglec-7, alternative splicing has also been found. In this case, there exist two transcripts, of 1.4 kb and 1.1 kb. The 1.4 kb transcript, originally identified as AIRM-1, produces the Siglec-7 protein with one V-set Ig-like domain followed by two C2-set Ig-like domains (Falco, Biassoni et al. 1999; Nicoll, Ni et al. 1999). The smaller 1.1 kb transcript, designated AIRM-2, is identical to Siglec-7 over its entire sequence with the exception of the missing first C2-set Ig-like domain (Falco, Biassoni et al. 1999).
1.6. Siglec Specificities for Sialic Acids

One of the defining characteristics for members of the Siglec family of proteins is their sialic acid-dependent binding to glycoproteins and glycolipids (Crocker, Kelm et al. 1996). Sialic acids are themselves a family of nine-carbon sugars, all of which are derivatives of neuraminic acid (Schauer 1982). They occur primarily as terminal sugars in different linkages bound to various oligosaccharide glycans, with the most abundant sialic acid being N-acetylneuraminic acid (Neu5Ac). However, different Siglec family members exhibit diverse specificity for differently linked sialic acids. Siglec-1 was found to preferentially bind α2-3-linked sialic acids, with mild binding to α2-6-linked sialic acids (Brinkman-Van der Linden, Sjoberg et al. 2000). Siglec-2, both human and murine, was found to bind exclusively to α2-6-linked sialic acids (Kelm, Schauer et al. 1994; Brinkman-Van der Linden, Sjoberg et al. 2000). Siglec-3 was found to bind to both α2-3 and α2-6-linked sialic acids, with a preference for the latter (Brinkman-Van der Linden, Sjoberg et al. 2000). Meanwhile, Siglec-4a binds exclusively to α2-3-linked sialic acids, in both humans and rats (Kelm, Schauer et al. 1994; Brinkman-Van der Linden, Sjoberg et al. 2000). Siglec-4b, the avian protein highly similar to Siglec-4a, also exhibits identical sialic acid specificity to Siglec-4a (Kelm, Schauer et al. 1996). Siglec-5 and Siglec-7 have both been found to bind equally to α2-3 and α2-6-linked sialic acids (Cornish, Freeman et al. 1998; Nicoll, Ni et al. 1999; Brinkman-Van der Linden, Sjoberg et al. 2000). On the other hand, Siglec-6 is the only family member so far that has been found to bind to the sialyl-Tn epitope (Neu5Aca2-6GalNAc) (Patel, Brinkman-Van der Linden et al. 1999). Finally, Siglec-8 was shown to bind exclusively to α2-3 linked sialic acids (Floyd, Ni et al. 2000). Of interest, in addition to the diverse
specificity of these Siglecs for different sialic acid linkages, is the conservation of this specificity between different species, as seen for Siglec-2, -4a, and -4b. As further studies into the various Siglecs in other species continue, it will be interesting to see if this conservation holds true for the remainder of this family.
2. RATIONALE

At the onset of these investigations, the chromosome 19q13.3-13.4 region was not characterized in detail. Previous efforts in our laboratory focused on the human Kallikrein gene family and its characterization at both the nucleotide and protein level. Through this endeavour, a putative novel gene was identified which exhibited extensive homology to a different family of proteins, the sialic acid-binding immunoglobulin-like lectins (Siglecs). This homology was highest with Siglec-3 and a few other proteins highly homologous to Siglec-3. Further, the few known Siglec-3-like members known at the time were localized through in situ hybridization techniques to this region of chromosome 19. Given that this whole family of proteins was only recently identified and that little was known regarding the Siglec-3-like subgroup in particular, we decided to proceed with our investigation of this putative novel gene, and the surrounding genomic region, in an effort to expand the understanding of this Siglec family.
3. HYPOTHESIS

Through efforts to investigate the region of chromosome 19q13.3-13.4 and the Kallikrein gene family, our preliminary results indicated the existence of a novel gene on chromosome 19q13.4. This novel gene exhibited no homology to any known Kallikrein, but did bear striking resemblance to Siglec-3 and several other Siglec family members, particularly the Siglec-3-like subgroup. Further examination of the known members of this Siglec-3-like subgroup showed that they had been localized to chromosome 19q13.3-13.4. Based on these observations, we hypothesized that:

1. The putative novel gene we identified in fact represents a novel addition to the Siglec-3-like subgroup of Siglecs.
2. In this same region of chromosome 19q13.4 are located additional novel and known members of the Siglec-3-like subgroup of Siglecs.
4. MATERIALS AND METHODS

4.1. Siglec-9

4.1.1. New Gene Identification

Nucleotide sequencing data of approximately 130 Kb on chromosome 19q13.4 was obtained from the Lawrence Livermore National Laboratory (LLNL) web site (http://www-bio.llnl.gov/genome/genome.html), in the form of one contig (clone BC349142 contig 56). This genomic sequence was subjected to a number of computer algorithms (gene prediction programs) designed to predict the presence of putative new genes. All programs used were previously thoroughly evaluated in our laboratory using a large number of known genes (Yousef, Luo et al. 1999). Based on these results, we selected the most reliable algorithms - GeneBuilder (gene prediction) (http://125.itba.mi.cnr.it/~webgene/genebuilder.html) and GeneBuilder (exon prediction) (http://125.itba.mi.cnr.it/~webgene/genebuilder.html); Grail 2 (http://compbio.orl.gov); and GENEID-3 (http://apoio.imim.es/geneid.html) – for further use.

4.1.2. Expressed sequence tag (EST) identification

The genomic sequence of the putative new gene was subjected to a homology search against the human EST database using the BLASTN algorithm (http://www.ncbi.nlm.nih.gov/BLAST) (Altschul, Madden et al. 1997). Clones showing >95% homology were obtained from the I.M.A.G.E. consortium through Research Genetics Inc. (Huntsville, AL). The clone obtained was then propagated according to the suppliers instructions, purified, and sequenced from both directions with an automated sequencer, using the insert-flanking vector primers T3 and T7.
4.1.3. Molecular Characterization of Siglec-9

The sequence derived from the computer predicted exons of our putative new gene was also used to search the non-redundant protein sequence database, using the BLASTP algorithm (Altschul, Madden et al. 1997). Several proteins showing a high degree of homology were selected, and their nucleotide coding sequences were aligned with our predicted coding sequence using the ClustalX multiple alignment program (Jeanmougin, Thompson et al. 1998). From this, we selected regions on our putative gene which showed the least amount of homology to the others and designed PCR primers: S9-F1 (5'-TCACCGCTCTCTGTGAATG-3') and S9-R1 (5'-GTCTTCTGCCCCAGGTTCAG-3'). Using these primers, we performed PCR on bone marrow cDNA, prepared as discussed below, and chosen based on the tissue expression results. The PCR conditions were as follows: 2.5 units HotStarTaq polymerase (Qiagen, Valencia, CA), 1X PCR buffer with 1.5 mM MgCl₂ (Qiagen), 1 ul cDNA, 200 uM dNTPs (deoxynucleoside triphosphates), and 250 ng of primers, using the Mastercycler® gradient thermocycler (Eppendorf Scientific, Inc., Westbury, NY). The temperature profile was: denaturation at 95°C for 15 min. followed by 94°C for 30 s., annealing at 58°C for 30 s., and extension at 72°C for 1 min. for a total of 35 cycles. followed by a final extension at 72°C for 10 min. The PCR product was subjected to electrophoresis on a 2% agarose gel and stained with ethidium bromide. Aliquots of the PCR products were subsequently extracted from the gel and the purified DNA was directly sequenced using an automated sequencer.
In order to verify the sequence surrounding the proposed start codon, another set of primers were designed, again derived from regions showing low homology with other known genes: S9-F3 (5'-TCCTCTAAGTCTTGAGCCCG-3') and S9-R3 (5'CAGACGTGTGAGATGGACGGT-3'). PCR was performed using bone marrow cDNA, prepared as described below. The conditions used for the PCR reaction were identical to those discussed previously, with electrophoresis of the PCR product on a 2% agarose gel, gel extraction, and automated sequencing as before.

Following final characterization of the genomic structure of Siglec-9, the putative protein product was aligned with the protein sequences of the other Siglec family members using the ClustalX multiple sequence alignment tool.

Our putative protein was assessed for the presence of a possible signal peptide, using SignalP v1.1 (http://www.cbs.dtu.dk/) (Nielsen, Engelbrecht et al. 1997). Further, for the prediction of transmembrane domains, two independent algorithms were used, TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) and DAS (http://www.biokemi.su.se/~server/). In addition, the hydropathic profile of this novel Siglec was determined, using the Kyte-Doolittle method (http://bioinformatics.weizmann.ac.il/hydroph/plot_hydroph.html).

4.1.4. Mapping and Chromosomal Localization of Siglec-9

As mentioned previously, the contig on which the Siglec-9 gene was identified was obtained from the LLNL. EcoRI restriction maps were obtained from LLNL, and also generated using the Webcutter restriction analysis tool (http://www.first-market.com/cutter/cutter2.html), for both this contig, as well as the adjacent more
centromeric contigs, containing the recently identified kallikrein gene family (Yousef, Luo et al. 1999; Diamandis, Yousef et al. 2000). Overlapping restriction fragments were identified and used to order the contigs and determine the distance between KLK14, the most telomeric member of the kallikrein gene family, and this novel Siglec.

4.1.5. Siglec-9 Tissue Expression

Total RNA from 28 normal human tissues was obtained (Clontech, Palo Alto, CA, USA), and reverse transcription was performed using SuperScript II™, according to the manufacturer’s instructions (Gibco BRL, Gaithersburg, MD, USA). PCR was then performed using primers S9-F2 (5'-CGTGGAGATACGGGCATAG-3') and S9-R2 (5'-AAAAGGGAGGGCACAGTGTG-3'), using the same PCR conditions described previously. PCR for actin was also performed as describe elsewhere (Yousef, Obiezu et al. 1999), as a control for cDNA quality.
4.2. Siglec-8-Long (Siglec-8-L)

4.2.1. Identification of the Genomic Area containing Siglec-8

Genomic DNA sequences derived from BAC clones covering chromosome 19q13.4 were identified and obtained from the Lawrence Livermore National Laboratory (LLNL) Human Genome Center. These sequences were compared to the mRNA sequence for Siglec-8 (GenBank Accession No. AF195092), which has been reported to be linked to this area (Floyd, Ni et al. 2000), using the BLASTN nucleotide alignment tool (Altschul, Madden et al. 1997). In addition, genomic regions found to match Siglec-8 were also analyzed by the Grail exon prediction program (Murakami and Takagi 1998), in order to determine the existence of any new Siglecs, as well as possible additional exons for Siglec-8. Prediction results were compared to the human EST database by the BLAST alignment tool (Altschul, Madden et al. 1997). Further, the genomic region containing Siglec-8 was localized to a specific region of chromosome 19q13.4 through the aid of the WebCutter restriction analysis program and comparison of the fragments to the previously published EcoR1 map for chromosome 19q13.4 (Ashworth, Batzer et al. 1995).

4.2.2. Molecular Characterization of Siglec-8-L

Based on the results of exon prediction and the known sequence of the Siglec-8 mRNA, we designed PCR primers to determine the sequence of the Siglec-8-L mRNA, as well as to confirm the published Siglec-8 mRNA sequence, both through RT-PCR. The primers used were: S8-Forward (common), 5'-ACAAGTGACACTGGCAGCAG-3'; S8-L-Reverse, 5'-AGCTGAGGTTGCATAATGG-3'; S8-L-Reverse2, 5'-TACTGC-
ATAGCATGGGGGCTC-3'; S8-Reverse, 5'-AGAAGAGCAGGGAAACCAC-3'. The fetal liver cDNA used was prepared as described elsewhere (Foussias, Yousef et al. 2000). The PCR conditions were as follows: 2.5 units of HotStarTaq polymerase (Qiagen, Valencia, CA), 1X PCR buffer with 1.5 mM MgCl₂ (Qiagen), 1 µL cDNA, 200 µM dNTPs (deoxynucleoside triphosphates), and 250 ng of each primer, using the Mastercycler® gradient thermocycler (Eppendorf Scientific Inc., Westbury, NY). The temperature profile was: denaturation at 95°C for 15 min. followed by 94°C for 30 s., annealing at 60°C for 30 s., and extension at 72°C for 1 min. for a total of 35 cycles, followed by a final extension at 72°C for 10 min. The PCR product was subjected to electrophoresis on a 2% agarose gel containing ethidium bromide. The product bands were then extracted from the gel, and the purified DNA was directly sequenced using an automated sequencer.

We also utilized Marathon-Ready fetal liver cDNA (Clontech, Palo Alto, CA, USA) to perform nested 3'-RACE in order to verify the 3' end of the Siglec-8 mRNA. The procedure was carried out according to the manufacturer's instructions, with some minor modifications. Briefly, the first round of the 3' RACE reaction utilized the forward gene-specific primer (GSP1) which is identical to the above mentioned S8-Forward (common) primer and the provided adapter primer AP1. The nested 3' RACE reaction was carried out using GSP2 (CCTTCCTGCTCTTCATG), and the provided AP2. The touchdown PCR method was utilized as recommended by the manufacturer, with a slight temperature profile modification. The annealing temperatures used were: 70°C for 4 min. for 5 cycles, then 68°C for 4 min. for 5 cycles, followed by 66°C for 4 min. for 25 cycles. The denaturation temperature was set to 94°C for 5 s. for every cycle, with an
extension temperature of 72°C for 1 min. After all cycles were completed a final extension at 72°C for 10 min. was performed. The reaction was carried out using 2.5 units HotStarTaq (Qiagen, Valencia, CA), 1X PCR buffer with 1.5 mM MgCl₂ (Qiagen), 200 μM dNTPs, primer concentrations of 200ng (GSP1 and 2) and 1 μM (API and 2), 1 μL Marathon-Ready cDNA, or 5 μL of the first round 3’ RACE product (50 μL total volume). All 3’ RACE reactions were performed using the Perkin Elmer GeneAmp PCR System 9600 thermocycler (Perkin Elmer, Norwalk, CT, USA). The products of both rounds of the RACE reaction were subjected to electrophoresis on a 2% agarose gel containing ethidium bromide, with any bands evident being extracted and the DNA directly sequenced with an automated sequencer.

Based on RT-PCR and 3’ RACE results, as well as the initial alignment of the Siglec-8 mRNA species to the genomic sequence covering chromosome 19q13.4, we proceeded to map the exons of the Siglec-8 gene. This was achieved through the use of the BLASTN nucleotide alignment tool, which enabled us to localize the mRNA sequences to specific regions of genomic DNA.

Following final characterization of Siglec-8-L, the primary structure of the protein encoded by the Siglec-8-L mRNA was determined. This was compared to the published protein sequence for Siglec-8, as well as to all the members of the Siglec-3-like subgroup of Siglecs using the CLUSTALX multiple alignment tool (Jeanmougin, Thompson et al. 1998).
4.3. Siglec-Like Gene (SLG)

4.3.1. Identification of SLG

Based on the high degree of homology among the Siglec-3-like subgroup of Siglecs, in both the extracellular Ig-like domains as well as the cytoplasmic tyrosine-based motifs, we screened the human expressed sequence tag (EST) database with these sequences, using the BLAST alignment tool (Altschul, Madden et al. 1997). Any matching ESTs we identified were obtained from the I.M.A.G.E. consortium through Research Genetics Inc. (Huntsville, AL). These clones were then propagated according to the suppliers instructions, purified, and sequenced from both directions with an automated sequencer, using the flanking T3 and T7 vector primers. Further, we obtained genomic sequence derived from BAC clones which cover the area of chromosome 19q13.4, believed to contain the Siglec-3-like subgroup locus, from the Lawrence Livermore National Laboratory (LLNL) Human Genome Center. We utilized the BLAST alignment tool to determine the exact location of any EST identified above in this genomic sequence. We then proceeded to investigate the genomic sequence surrounding the EST sequence with an exon prediction program, Grail2Exons (Murakami and Takagi 1998).

4.3.2. Cloning and Molecular Characterization of Two SLG Splice Forms

Based on the alignment of the EST and the exon prediction results, we proceeded to design sets of primers to be used with reverse transcription-coupled polymerase chain reaction (RT-PCR) in order to determine the exact sequence of the SLG mRNA species. This design allowed for the production of overlapping RT-PCR fragments, thus enabling
us to determine the entire mRNA sequence. Based on results from RT-PCR with a panel of human tissues (see below), we used bone marrow as the tissue with which to work with. Bone marrow cDNA was prepared as described below. The primer combinations which we used were: i) SLG-F3: 5'-AGGAAGCCTCTGCCTCAGAG-3', and SLG-R3: 5'-CCTTCATCCTCCTGACATGCAC-3'; and ii) SLG-F2: 5'-ATCACTCGCTCCTCGATGCT-3', and SLG-R2: 5'-TCTCTCCTTCTCTCTGAGGAG-3'. Due to the high degree of homology, even at the nucleotide level, among the Siglec-3-like subgroup of Siglecs, we utilized semi-nested PCR, to ensure we were amplifying the correct mRNA species, using the above forward primers, SLG-F3 and SLG-F2, and the nested reverse primers SLG-R3N (5'-GAGGACTGTGAGGGGCTCAG-3') and SLG-R2N (5'-GATTCAATCAGGGGTCC-3'), respectively. The PCR conditions were as follows: 2.5 units HotStarTaq polymerase (Qiagen, Valencia, CA), 1X PCR buffer with 1.5 mM MgCl₂ (Qiagen), 1 μl cDNA, 200 μM dNTPs (deoxynucleoside triphosphates), and 200 ng of primers, using the Mastercycler® gradient thermocycler (Eppendorf Scientific, Inc., Westbury, NY). The temperature profile was: denaturation at 95°C for 15 min. followed by 94°C for 30 s., annealing at 62°C for 30 s., and extension at 72°C for 1 min., for a total of 35 cycles, followed by a final extension at 72°C for 10 min. The PCR product was subjected to electrophoresis on a 2% agarose gel containing ethidium bromide. The PCR products were extracted from the gel and the purified DNA was directly sequenced using an automated sequencer.

During our examination of this genomic area through exon prediction, an additional putative upstream exon was also identified, which also possessed a signal peptide and Ig-like sequence highly homologous to existing members of the Siglec-3-like
subgroup of Siglecs. In order to examine the possibility of an additional upstream exon with a slightly different signal peptide and Ig-like sequence, we designed a forward PCR primer specific for this exon, lying within its 5’ untranslated region. This upstream primer, SLG-FU (5’-TGGCACCTCCAACCCGACAC-3’), was used in combination with the reverse primer SLG-R3 for RT-PCR, under identical conditions to those described above. The PCR product was subjected to electrophoresis on a 2% agarose gel containing ethidium bromide. The resultant DNA bands were extracted from the gel, purified, and directly sequenced using an automated sequencer.

Following final characterization of both the SLG mRNA sequences as well as the genomic organization for SLG, the putative protein products were determined. These protein sequences were then aligned with those of the other known members of the Siglec-3-like subgroup of Siglecs using the ClustalX multiple alignment tool (Jeanmougin, Thompson et al. 1998). Shading of similar and identical residues was accomplished using the BOXSHADE alignment shading program (http://www.ch.embnet.org/software/BOX_form.html). Further, the SLG protein sequences were examined for the presence of transmembrane domains, with both TMpred (Hofmann and Stoffel 1993) and through their Kyte-Doolittle hydrophobicity profile (Kyte and Doolittle 1982). The presence of putative signal peptides was examined through the use of the SignalP v1.1 signal peptide prediction program (Nielsen et al 97).

4.3.3. Tissue Expression of Two SLG Splice Variants

The tissue expression profile for both SLG alternative splice forms was elucidated by performing RT-PCR using total RNA from 25 normal human tissues (Clontech, Palo
Alto, CA, USA). The PCR primers used were SLG-F3 and SLG-R3, and SLG-FU and SLG-R3, for the short and longs forms of SLG, respectively. The PCR conditions were the same as those described above, and reverse transcription was performed using SuperScript II™, according to the manufacturer’s instructions (Gibco BRL, Gaithersburg, MD, USA). The temperature profile for the PCR was identical to that described previously. From the cDNA that was produced we also performed a PCR for actin, as described elsewhere (Yousef, Obiezu et al. 1999), as a control for cDNA quality.

4.3.4. Mapping and Chromosomal Localization of SLG

As indicated above, the SLG gene was identified in the genomic sequence from a BAC clone covering chromosome 19q13.4. We subjected the sequence encompassing the SLG gene to the Webcutter restriction analysis tool to determine the size of the resultant EcoRI fragments. We then compared these results to the published EcoR1 map for chromosome 19 (Ashworth, Batzer et al. 1995), which is also available through the LLNL Human Genome Center. In addition, the SLG mRNA sequences were aligned with the final genomic sequence from this BAC clone (GenBank Accession No. AC020914) using the BLASTN nucleotide alignment program, allowing for precise localization of all exons. Further, by knowing the precise location of both the Siglec-9 gene (Foussias, Yousef et al. 2000) and the Siglec-8 gene (Foussias, Yousef et al. 2000), we determined the distance between these and the SLG gene.
4.3.5. Phylogenetic Analysis of the Expanded Siglec Family

In order to examine the evolutionary relationship among the members of the Siglec family, including the novel members identified in this work, we proceeded to perform phylogenetic analysis. This was achieved, initially, through use of the ClustalX multiple alignment tool, utilizing the protein sequence for all the human Siglec family members and those of Siglec-9, Siglec-8-L, and SLG-S and -L. Phylogenetic analysis was then performed with the Phylip software package, through the use of ClustalX and a distance matrix. Parsimony trees were constructed using neighbourhood-joining, UPGMA, and protpars parsimony methods. Selection of the appropriate tree was based on which of these was consistent with the findings by other groups (Angata and Varki 2000).
5. RESULTS

5.1. Siglec-9

5.1.1. Identification of Siglec-9 on 19q13.4

Computer analysis of the approximately 130 Kb contig predicted a putative new gene consisting of six exons. Five of these were predicted by at least three programs, with only one exon being predicted by two of the four programs (Table 1). Homology search for the putative new gene against the human EST database revealed the presence of one unique EST (GenBank accession # AA936059) which showed 98% identity to the sixth predicted exon.

We proceeded to sequence the entire insert of this EST, followed by alignment of this nucleotide sequence with the genomic sequence of our putative gene, using the “BLAST 2 sequences” program. This revealed the presence of an additional area, between predicted exons 5 and 6, with 98% identity to the EST. This suggested that there was an additional exon in this area which was not detected by the prediction algorithms used.

5.1.2. Characterization of the Genomic Structure of the Siglec-9 Gene and its Protein Product

Based on the existence of an EST almost identical to part of our putative gene, we postulated that this, in fact, is a novel gene. With the aid of unique primers, designed as discussed in the experimental section, we were able to perform RT-PCR on bone marrow cDNA and isolate two additional products, both encompassing multiple predicted exons. Upon sequencing of these PCR products, we confirmed the presence of all six predicted
Table 1: Genomic organization of Siglec-9.

<table>
<thead>
<tr>
<th>Exon No.</th>
<th>Coding Region$^1$</th>
<th>No. of base pairs</th>
<th>EST Match$^2$</th>
<th>Intron Phase</th>
<th>Exon Predicted$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1083</td>
<td>1591</td>
<td>509</td>
<td>-</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>1793</td>
<td>2071</td>
<td>279</td>
<td>-</td>
<td>I</td>
</tr>
<tr>
<td>3</td>
<td>2277</td>
<td>2324</td>
<td>48</td>
<td>-</td>
<td>I</td>
</tr>
<tr>
<td>4</td>
<td>3226</td>
<td>3492</td>
<td>267</td>
<td>-</td>
<td>I</td>
</tr>
<tr>
<td>5</td>
<td>4145</td>
<td>4235</td>
<td>91</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4610</td>
<td>4706</td>
<td>97</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6087</td>
<td>6503</td>
<td>417</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

1. The coding region shown includes the 5' untranslated region in exon 1, and the 3' untranslated region in exon 7. Numbers refer to GenBank accession no. AF135027.

2. EST; GenBank accession no. AA936059

3. The exon prediction programs are as follows: A) GeneBuilder (gene prediction); B) GeneBuilder (exon prediction); C) Grail 2; D) GENEID-3.
exons, as well as the newly identified exon, found from the EST sequence. With both cDNA and genomic sequence at hand, we determined the genomic organization of this new gene (Fig. 1). We found that the gene encoding this novel Siglec encompasses a genomic area of 5,421 bp. It is composed of seven exons, with six intervening introns. The lengths of the exons are 509, 279, 48, 267, 91, 97, and 417 bp, respectively. All the intron/exon splice sites and their flanking sequences are closely related to the consensus splice sites (-mGTAAGT...CAg-, where m is any base) (Iida 1990).

The proposed protein coding region of the Siglec-9 gene consists of 1,392 nucleotides, producing a 463 amino acid protein, with a predicted molecular mass of 50.1 kDa, excluding any post-translational modifications. The translation initiation codon (ATG) at position 1171 of the first exon (according to the numbering of our GenBank accession #AF135027), was chosen because: 1) the flanking region surrounding that codon closely matches the Kozak consensus sequence for translational initiation, particularly at position -3 (a purine), which appears to be the most highly conserved (Kozak 1991); 2) using this initiation codon, the proposed protein contains an N-terminal signal sequence which shows a high degree of homology to other similar proteins (see below). The 3' terminus of the Siglec-9 gene was verified by the presence of a poly dA tail present in the EST sequence. Further, the coding sequence of this gene possesses a 5' untranslated region of at least 88 nucleotides, as well as a 3' untranslated region of 228 nucleotides.

Examination of the hydrophobicity profile of the putative Siglec-9 protein revealed two regions with long stretches of hydrophobic residues (Fig. 2). The first of these occurs at the N-terminus, suggesting the presence of a signal peptide, and is
Figure 1: Genomic Structure of Siglec-9. Shown are the exon/intron boundaries, as well as the predicted protein sequence. The single underlined region is the 5' untranslated region, and the double underlined region is the 3' untranslated region. In the shaded box is the putative polyadenylation signal.
Figure 2: Hydrophobicity Plot of Siglec-9. This shows the regions of the Siglec-9 protein which contain stretches of hydrophobic amino acid residues. As is evident, there are two such regions, the first corresponding to the signal peptide, and the second, at around residues 350-370, the putative transmembrane region.
consistent with findings from a signal sequence prediction program (Nielsen, Engelbrecht et al. 1997), which predicts a 17 amino acid residue signal sequence. The second region occurs between residues 349 and 370, suggestive of a transmembrane domain, and is consistent with results from transmembrane region prediction programs. Based on this information, we postulate that the protein product of this novel gene is a type I transmembrane protein, after cleavage of the 17 residue signal sequence.

5.1.3. Mapping and Chromosomal Localization of Siglec-9

The contig in which we identified the gene encoding Siglec-9 is located at 19q13.4. telomeric to the kallikrein gene KLK3 (PSA). Previous studies in our laboratory have identified and mapped the kallikrein gene family locus on this region of chromosome 19 (Yousef, Luo et al. 1999; Diamandis, Yousef et al. 2000). The contig containing the novel Siglec gene was found, through EcoRI restriction mapping, to be located adjacent to this kallikrein gene family. The Siglec-9 gene is located 43.19 Kb more telomeric than KLK14, at 19q13.4. A detailed physical map of the area which contains some known genes and the newly identified Siglec gene is shown in Fig. 3. By computer analysis, we did not predict any other genes between KLK14 and this novel Siglec.

5.1.4. Homology with other Siglec Family Members

Using the predicted protein sequence, we performed a homology search against the GenBank database using the BLASTP program. Siglec-9 showed a high degree of homology to other known members of the Siglec family (Table 2). We further
Figure 3: Localization of the Siglec-9 Gene. The physical map of the 314kb genomic area around chromosome 19q13.3-q13.4 where the kallikrein gene family resides. Gene lengths are presented above each arrow, and distances between genes are shown below. Arrows denote the direction of transcription. The Siglec-9 gene resides 43.2 Kb telomeric to the KLK-L6 gene. For more details on this genomic region, see (Diamandis et al., 1999). KLK, kallikrein.
performed a multiple alignment of Siglec-9 with the other family members, using the ClustalX alignment program. As is evident in Fig. 4, the N-terminal signal sequence is highly conserved within this family of proteins. Furthermore, our protein contains Ig domains typically found in Siglec family members: an N-terminal V-set domain, followed by multiple C2-set domains (Crocker, Kelm et al. 1996). This novel Siglec contains a total of 3 Ig domains, a V-set and two C2-set domains, based on homology with known Ig domains. As shown in Figure 4, these V-set and C2-set domains are highly similar to Siglec-7 and -8. Siglec-9 exhibits conservation of the cysteine residues in the V-set and first C2-set domains, which form the two characteristic disulfide bridges in other Siglec family members. The V-set domain also possesses a conserved arginine which has been found to be essential for sialic-acid binding (van der Merwe, Crocker et al. 1996), as well as two conserved aromatic residues in β-strands A and G which have been found to make hydrophobic contacts with the N-acetyl and glycerol side groups of N-acetyl neuraminic acid (May, Robinson et al. 1998). As is evident from Figure 4, Siglec-9 also possesses this critical arginine, as well as one of the aromatic residues. The domain boundaries were determined based on the one domain: one exon rule (Williams and Barclay 1988), while taking into consideration the domain assignments of others (Cornish, Freeman et al. 1998; Crocker, Clark et al. 1998; Falco, Biassoni et al. 1999; Nicoll, Ni et al. 1999; Patel, Brinkman-Van der Linden et al. 1999).

Examination of the transmembrane and intracellular domains of Siglec family members reveals that it is more variable than the extracellular domain. However, there are regions that show a high level of conservation. As shown in Figure 4, all the Siglecs possess a single transmembrane domain, consisting of approximately 25 residues. In
Table 2: Overall homology of Siglec-9 with the Siglec-3-like subgroup of Siglecs.

<table>
<thead>
<tr>
<th>Siglec Family Member</th>
<th>Homology to Siglec-9</th>
<th>% identity</th>
<th>% similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siglec-7 (NM_014385)</td>
<td>75</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Siglec-8 (NM_014443)</td>
<td>65</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Siglec-5 (NM_003830)</td>
<td>52</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Siglec-3 (M23197)</td>
<td>52</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Siglec-6 (NM_001245)</td>
<td>49</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Siglec-1 (Z36293)</td>
<td>27</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>CD22 (X52785)</td>
<td>26</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Myelin associated glycoprotein (MAG) (M29273)</td>
<td>25</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>

1. GenBank accession numbers for each of the Siglec family members is also shown, in brackets.
2. Homology was determined using the BLASTP algorithm.
Figure 4: Siglec-9 Multiple Alignment with other Siglec-3-like Subgroup Members.

Siglec-9 was aligned with Siglec-5 to -8 and Siglec-3, using ClustalX (Jeanmougin, Thompson et al. 1998). The signal peptide was determined through computer prediction, and the Ig domain boundaries were assigned based on exon boundaries (shown with bent arrows). The transmembrane domain was also predicted, while taking into consideration exon boundaries as well. The ITIM and SLAM-like motifs are indicated, as are the conserved cysteines (*) which form the disulfide bonds of the Ig-like domains in Siglecs, and the conserved arginine and aromatic residues (□) which are responsible for sialic acid binding and specificity.
addition, within the cytoplasmic domain, we notice the presence of two highly conserved motifs. The first of these, L(HQ)YA(SV)L, exhibits similarity to an immunoreceptor tyrosine kinase inhibitory motif (ITIM), which has a 6 amino acid consensus sequence (ILV)xYxx(LV) (Burshtyn, Yang et al. 1997; Vivier and Daeron 1997). The second motif, TEYSE(IV), is homologous to a sequence (TxExx(IV)) recently found in the signaling lymphocyte activation molecule (SLAM) which is responsible for the binding of the SLAM-associated protein (SAP) (Coffey, Brooksbank et al. 1998; Sayos, Wu et al. 1998).

5.1.5. Tissue Expression Profile of Siglec-9

RT-PCR was performed on a panel of tissue-specific total RNA preparations (Fig. 5). We have found high levels of expression of Siglec-9 in bone marrow, placenta, spleen, and fetal liver. Lower levels of expression were also evident in fetal brain, stomach, lung, thymus, prostate, brain, mammary, adrenal gland, colon, trachea, cerebellum, testis, small intestine, and spinal cord. Expression of Siglec-9 was absent in heart, skeletal muscle, pancreas, and ovary. All PCR products obtained were of equal length, and corresponded to the length of the product obtained from overlapping EST (accession # AA936059). Sequencing of the PCR products ensured specificity.
Figure 5: Tissue Expression Profile of Siglec-9. RT-PCR was performed on 28 tissue total RNAs for Siglec-9 and actin (control gene). Siglec-9 is highly expressed in bone marrow, placenta, spleen, and fetal liver. There is also a lower degree of expression in many of the other tissues, while it is absent in ovary, pancreas, skeletal muscle, and heart.
5.2. Siglec-8-Long (Siglec-8-L)

5.2.1. Identification of the Genomic Area containing Siglec-8:

The Siglec-3-like subgroup of Siglecs has been mapped, through various means, to chromosome 19q13.4. The Siglec-9 gene has recently been characterized and localized to this area (Foussias, Yousef et al. 2000), immediately following the end of the kallikrein gene family. Further, the mRNA species for the other known members of this subgroup have also been characterized, and mapped to this region primarily through fluorescence in situ hybridization (FISH), as well as somatic cell hybridization (Cornish, Freeman et al. 1998; Falco, Biassoni et al. 1999; Nicoll, Ni et al. 1999; Patel, Brinkman-Van der Linden et al. 1999; Floyd, Ni et al. 2000). During our examination of this area, we identified a clone, CTD-3073N11, from the CalTech Human BAC library D, that contained the Siglec-8 gene. Upon exon prediction analysis of this genomic region, we identified two putative exons at the 3' terminus of the Siglec-8 gene which differed from the previously published mRNA sequence. One of these exons was much shorter in length and the other was not present at all. We proceeded to search the human EST database, both with the published mRNA sequence, as well as with the two putative exons, and were unable to find any significant matches.

Based on the sequence information from the clone on which Siglec-8 is localized, we used the WebCutter restriction analysis tool to determine the size of the EcoR1 fragments produced. By comparing these results with the available EcoR1 restriction map for chromosome 19q13.4, we determined that the Siglec-8 gene is located in the more centromeric region of 19q13.4, and is approximately 330 kb downstream from the Siglec-9 gene.
5.2.2. Molecular Characterization of Siglec-8-L

Using RT-PCR and primers derived from sequence shared by both Siglec-8 and Siglec-8-L, as well as from the 3' termini of Siglec-8 and the putative Siglec-8-L, we were able to confirm the existence of both putative exons mentioned above. Through automated sequencing and subsequent alignment of the mRNA sequence with the genomic sequence for our BAC clone, we were able to determine the exact genomic organization for these last two exons. However, using the PCR primers specific for the published sequence of Siglec-8 we obtained a very faint band, after agarose gel electrophoresis, which we were unable to sequence. In conjunction with 3' RACE and the Marathon-Ready fetal liver cDNA we identified additional sequence at the 3' terminus of Siglec-8-L, including both the termination codon and a portion of the 3' untranslated region. The nested 3' RACE reaction produced a major product with an approximate size of 300 bp, as well as a much fainter high molecular weight band. Attempts to sequence the latter band were unsuccessful.

Based on the results obtained through the initial alignment, as well as from the characterization of the last two exons of Siglec-8-L, we were able to characterize the genomic organization of the Siglec-8 gene. As is shown in Figure 6, both Siglec-8 and Siglec-8-L are comprised of seven exons. The first five exons are identical in both Siglec-8 and Siglec-8-L, with lengths of 502 bp, 279 bp, 48 bp, 270 bp, and 97 bp. The first exon of 502 bp contains a 5' untranslated region of 48 nucleotides, with the possibility that there is even more upstream sequence. For Siglec-8-L, exon six is 97 nucleotides long, while exon seven contains at least 299 bp, of which 252 bp code for
amino acid residues, and 44 bp being part of the 3' untranslated region. Exons six and seven of Siglec-8, on the other hand, are 895bp and 779bp long, respectively. Examining the splice donor and acceptor sites for each of these exons, we noticed that the first five exons, as well as exons six and seven of Siglec-8-L, all possessed sequences closely related to the consensus splice sites (-mGTAAGT...CAGm-, where m is any base) (Iida 1990). However, in the case of Siglec-8, these splice sites were not present. Examination of the open reading frame of Siglec-8-L revealed a 499 amino acid residue protein with a molecular weight of 54.04 kDa, excluding any post-translational modifications which may be present. The sequence of Siglec-8-L is identical to that of Siglec-8 until residue 415, after which Siglec-8-L contains a sequence homologous to the C-terminal sequences of the Siglec-3-like subgroup of Siglecs, including the two tyrosine-based motifs (Fig. 7).
**Figure 6: Genomic Organization of the Siglec-8 Gene.** Based on our own experimental results, as well as the previously published sequence for the Siglec-8 mRNA (GenBank Accession No. AF195092), we have established that both Siglec-8 and our Siglec-8-L are composed of seven exons (Arabic numerals). The two mRNA species are identical until exon six, where the Siglec-8 mRNA contains the whole of exon six (6a & 6b) and continues to exon 7a, indicated by the broken line, while the Siglec-8-L mRNA contains only exons 6b and 7b, shown by the solid line. The location of the stop codon (TGA) is shown for both mRNA species, and differs due to a change in the open reading frame.
Figure 7: Protein Sequence Alignment for the Siglec-3-like Subgroup of Siglecs. The sequence of Siglec-8-L was aligned to those of Siglec-8, as well as the remaining members of the Siglec-3-like subgroup of Siglecs, using the ClustalX multiple alignment tool (Jeanmougin et al. 1998). The solid vertical lines indicate the positions of the exon boundaries. In all but one case, shown by the broken vertical line, the exon boundaries match those found for Siglec-9 (Foussias et al. 2000). The conserved cysteine residues responsible for the intra- and interdomain disulfide bonds (Crocker et al. 1996; Pedraza et al. 1990; Williams et al. 1989) are indicated by the star, while the triangles denote the aromatic residues believed to be important for sialic acid binding, based on findings for Siglec-1 (May et al. 1998). The signal peptide cleavage site for Siglec-8-L, indicated by the solid circle, was predicted using the SignalP program (Nielsen et al. 1997). The Ig-like domain assignments, as well as those for the transmembrane and cytoplasmic domains, are based on previous reports (Floyd et al. 2000) and the one domain-one exon rule (Williams and Barclay 1988). The positions of the two tyrosine-based motifs, ITIM and SLAM-like, are indicated. The GenBank accession numbers are as follows: Siglec-8-L: AF287892; Siglec-8: AF195092; Siglec-9: AF135027; Siglec-7: AF170485; Siglec-6: NM_001245; Siglec-5: NM_003830; Siglec-3: M23197.
5.3. Siglec-Like Gene (SLG)

5.3.1. Identification of SLG

Our screening for ESTs homologous to previously published members of the Siglec-3-like subgroup revealed the presence of an EST (GenBank accession no. AI132995) which showed extensive homology to the tyrosine-based motifs found in the cytoplasmic tail of other members of this subgroup. This EST was then compared to genomic sequence derived from BAC clones covering chromosome 19q13.4. We identified one clone, CTD-3073N11, which contained this EST in the form of three exons. Subsequent exon prediction using the genomic sequence from this clone indicated the presence of an additional four putative exons.

5.3.2. Cloning and Molecular Characterization of Two SLG Splice Forms

Based on the results of exon prediction, followed by verification through RT-PCR and sequencing, we cloned and fully characterized the entire mRNA structure of SLG-Short (SLG-S). Through alignment with the genomic sequence we determined the precise genomic organization of the SLG gene. We found that this mRNA species, similar to our findings for Siglec-8 and -9 (Foussias, Yousef et al. 2000; Foussias, Yousef et al. 2000), is encoded by seven exons, with six intervening introns. The first two predicted exons mentioned above were found to be a single exon, based on our experimentally defined mRNA sequence. Further, the exon prediction did not detect the third exon, which was identified through RT-PCR. The lengths of the exons which encode the SLG-S mRNA are 474, 279, 48, 270, 97, 97, and 471bp (Fig. 8). All
intron/exon splice sites are closely related to the consensus splice sites (mGTAAGT...CAGm-, where m is any base) (Iida 1990).

The proposed open reading frame for the SLG-S mRNA transcript consists of 1736bp, which results in a 477 amino acid protein, with a molecular weight of 51.7 kDa, excluding any post-translational modifications. The translation initiation codon (ATG), located at position 21 (based on the numbering of our GenBank submission) was chosen for two reasons: i) the sequence surrounding this initiation codon conforms to the Kozak consensus sequence for translation initiation, especially with the most highly conserved purine at position -3 (Kozak 1991); ii) with this translation initiation codon the resultant protein product shows extensive homology with other members of the Siglec-3-like subgroup of Siglecs (see below), as well as the fact that no other initiation codon was found that produced a long, continuous open reading frame. The first exon contains a 5' untranslated region of at least 20bp, while in the seventh (last) exon there is a 3' untranslated region of 282bp. Through the presence of a poly dA tail at the 3' end of the EST, we were able to identify the end of the SLG-S mRNA transcript.

Through our examination of the additional putative upstream exon encoding an alternate signal peptide, we discovered that this produces an alternative mRNA species which differs from the SLG-S mRNA at its 5' end (Fig. 8). This novel SLG transcript, designated SLG-Long (SLG-L) is comprised of 8 exons, with 7 intervening introns. Through nucleotide alignment of this mRNA species with the genomic sequence from the BAC clone, we were able to determine the exon boundaries for this alternative splice form as well. The first exon present in the SLG-L mRNA is at least 447 bp in length, including at least 20 bp of 5' untranslated sequence, with a second exon of 381 bp. These
GTG AGG TCC TGC AGG AAG AAA TCG GCA AGG CCA GCA GTG GCC GTG GGC GAT ACA GCC ATG GAG GAC
VR SCR KKS ARPA VVG GD TGMED
GCA AAC GCT GTC AGG GCC TCA GCC TCT CAG gtga....INTRON 7....cag GGA CCC CTG AAT GAA TCC
ANAVRGSASQ
G PLIES
CCG GCA CAG AGC CCC CCA CAC CAT GCT CCG CCA GCC CTG GCC ACC CCC TCC CCA GAG GAA GGA
PAD D SPP HAPPA L A T PSPEEG
GAG ATC CAG TAT GCA TCC TTC AGC TTC CAC AAA GCC AGG CCT CAG TAC CCA CAG GAA CAG GAG GCC
EI QY ASL SFHK ARPQ YPQE QEA
ATC GCC TAT GAG TAG GCC GAG ATC AAC ATC CCC AAG TGA GAAAACGAGAGACCTCAGGCCCTGATGAGCTA
IGY EY SE I NIPK*
CGACCCCTCGAGCAAGGAGGGGAGAGTCATTTATGAATGGCCTGCCCATGCTGCGAGAGACTGAGACTATTGGCCG
TACCCACATATGAGAGAGTACTTCTTATGGTCTATTATGATATCTAGAGCT

Figure 8: Nucleotide and Protein Sequences of Both SLG Isoforms. The nucleotide sequences of both splice variants, SLG-L and SLG-S, which we verified experimentally, along with their amino acid translation are shown. The 5' untranslated region of the SLG-L mRNA transcript is indicated by the solid underline, while that of the SLG-S transcript is indicated by the dotted underline. Further, the predicted SLG-S protein is indicated by the broken underline, until the codon where both splice variants assume the same reading frame. The 3' untranslated region is indicated by the double underline. The intron/exon splice sites are also shown, in lower case letters. A putative polyadenylation signal is also indicated by the shaded sequence.
381 bp of the second SLG-L exon are identical to those of the SLG-S mRNA transcript. However, in the case of SLG-S, these are part of its first exon, which extends at least 93 bp, including at least 20 bp of 5'-untranslated sequence, upstream of the splice site for SLG-L (Fig. 9). The remainder of the SLG-L mRNA sequence is identical to that of SLG-S. As in the case of SLG-S, the SLG-L mRNA splice sites are all closely related to the consensus splice sites, as described above.

The putative protein product derived from the SLG-L mRNA transcript is composed of 2070 bp, producing a protein product of 595 amino acid residues, with a predicted molecular weight of 65.0 kDa, excluding post-translational modifications. The translation initiation codon (position 21 in the SLG-L GenBank submission) was chosen because it is the only one which produces a continuous open reading frame, maintaining this frame after splicing and assuming identity with the SLG-S mRNA sequence. Further, this open reading frame exhibits extensive homology to known members of the Siglec-3-related subgroup of Siglecs.

Examination of both SLG-S and SLG-L protein sequences revealed that they are highly homologous to other members of the Siglec-3-like subgroup of Siglecs (Fig. 10). Like other members of this subgroup, SLG-S and SLG-L also appear to possess similar N-terminal signal sequences. These hydrophobic N-terminal sequences are evident in the hydrophobicity plots for SLG-S and SLG-L, with the hydrophobic N-terminal sequence of SLG-L being longer than that of SLG-S (Fig. 11). This is also reflected in the signal peptide prediction results, which detects a strong signal putative signal peptide sequence only in the SLG-L protein sequence. These are followed by an N-terminal V-set Ig-like domain, two in the case of SLG-L, and two C2-set Ig-like domains, similar to other
Figure 9: Genomic Organization of the SLG Gene. Shown schematically above is the organization and splicing pattern for the two SLG mRNA transcripts we identified. The SLG-L transcript begins at exon 1, joining exon 2 as indicated by the solid line. The SLG-S mRNA transcript begins at the beginning of exon 2, with its corresponding ATG and continues through the remaining 6 exons, as does the SLG-L transcript. The shaded region denotes the nucleotide sequence that is contained in the SLG-S mRNA transcript, but absent in the SLG-L transcript. The sizes of the exons are indicated, with introns being denoted by roman numerals. The location of the stop codon (TGA), which is the same for both mRNA species, is also shown.
Siglec-3-like Siglecs. The single transmembrane domain, predicted by TMpred and evident in the Kyte-Doolittle hydrophobicity plots (Fig. 11), is in keeping with observations for other members of this subgroup. Furthermore, we noticed that both SLG splice forms also contain the two characteristic tyrosine-based motifs, ITIM and SLAM-like, noted in members of the Siglec-3-like subgroup of Siglecs. Further, as is evident in Figure 10, there is also conservation of all the key cysteine residues that are responsible for the characteristic folding of the extracellular Ig-like domains in all Siglecs (Pedraza, Owens et al. 1990; Crocker, Kelm et al. 1996). With regards to the residues believed to be responsible for sialic acid binding there is conservation of both aromatic residues, however in SLG-S there is a glutamine, and in SLG-L there is a cysteine, in place of the otherwise conserved arginine (van der Merwe, Crocker et al. 1996).

We also proceeded to perform a more detailed examination of the homology between both SLG splice forms and the other members of both the Siglec-3-like subgroup and the other Siglec family members. This was achieved through the use of the BLASTP protein alignment tool (Altschul, Madden et al. 1997). As shown in Table 3, SLG-S has over 70% similarity with Siglecs 7-9, in addition to slightly lower homology with other members of this subgroup. SLG-L also exhibits in excess of 70% homology with the Siglecs 7-9, with increased homology to Siglec-3, -5, and -6, compared to SLG-S.

5.3.3. Tissue Expression Profile of Two SLG Splice Variants

Through RT-PCR with a total RNA panel of 25 different normal human tissues we examined the tissue expression profile of both SLG splice forms (Fig. 12). SLG-S was
Table 3: SLG Isoform sequence homology with the Siglec-3-like subgroup of Siglecs.

<table>
<thead>
<tr>
<th>CD33-like Subgroup Member</th>
<th>GenBank Accession #</th>
<th>Homology to SLG-S (^1)</th>
<th>Homology to SLG-L (^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% Identity</td>
<td>% Similarity</td>
</tr>
<tr>
<td>Siglec-7</td>
<td>NM_014385</td>
<td>68</td>
<td>76</td>
</tr>
<tr>
<td>Siglec-8-L</td>
<td>AF287892</td>
<td>67</td>
<td>75</td>
</tr>
<tr>
<td>Siglec-8</td>
<td>NM_014442</td>
<td>66</td>
<td>74</td>
</tr>
<tr>
<td>Siglec-9</td>
<td>AF135027</td>
<td>64</td>
<td>73</td>
</tr>
<tr>
<td>Siglec-6</td>
<td>NM_001245</td>
<td>48</td>
<td>60</td>
</tr>
<tr>
<td>Siglec-3</td>
<td>M23197</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td>Siglec-5</td>
<td>NM_003830</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

3. Homology was determined using the BLASTP algorithm.
**Figure 10: Protein Sequence Alignment of both SLG Splice Variants and the Siglec-3-like Subgroup of Siglecs.** The sequences of both SLG-S and SLG-L were aligned to those of the entire Siglec-3-like subgroup of Siglecs, using the ClustalX multiple alignment tool (Jeanmougin, Thompson et al. 1998). The solid vertical lines indicate the positions of the exon boundaries, with the arrows indicating the different domains. The conserved cysteine residues responsible for the intra- and interdomain disulfide bonds are indicated by the star, while the triangles denote the residues believed to be important for sialic acid binding, based on findings for Siglec-1. The signal peptide is indicated and is based on homology with other members of the Siglec-3-like subgroup of Siglecs, as well as signal peptide prediction in the case of SLG-L. The Ig-like domain assignments, as well as those for the transmembrane and cytoplasmic domains, are based on previous reports (Foussias, Yousef et al. 2000) and the one domain-one exon rule (Williams and Barclay 1988). For the predicted transmembrane domain, the TMpred results indicate that it exists between residues 366 and 384 (based on the numbering of the SLG-S transcript). The positions of the two tyrosine-based motifs, ITIM and SLAM-like, are indicated.
Figure 11: Hydrophobicity Plots of Both SLG Isoforms. This shows the regions of the putative SLG-S and SLG-L proteins which contain stretches of hydrophobic amino acid residues. As is evident, there is a strong hydrophobic region around residues 360-390 of SLG-S, as well as around residues 475-505 for SLG-L, which contain the predicted transmembrane region. Further, the C-termini of both proteins, especially SLG-L, exhibit a smaller hydrophobic region, which likely correspond to the signal peptide as seen in other members of the Siglec-3-like subgroup of Siglecs.
found to be relatively highly expressed in spleen, small intestine and adrenal gland. Moderate to mild expression was apparent in thyroid, placenta, brain, stomach, bone marrow, spinal cord, and breast. SLG-L, on the other hand, exhibited relatively high levels of expression in spleen, small intestine, and bone marrow. Further, intermediate levels of expression were observed in thyroid, placenta, thymus, trachea, stomach, lung, adrenal gland, fetal brain, and testis. The PCR products obtained were all of equal size and corresponded to the length of the product obtained during the molecular characterization of the respective SLG mRNA transcript. Further, specificity was ensured through sequencing of RT-PCR products. In the case of SLG-L, where in some tissues two bands are observed, sequencing of both DNA bands revealed that they in fact are identical, and may possibly be due to different secondary structures for the same DNA sequence, resulting in different migration speeds.

5.3.4. Mapping and Chromosomal Localization of SLG

The contig on which we identified the SLG gene was located telomeric to the Siglec-9 gene, which was originally characterized and precisely mapped in our laboratory (Foussias, Yousef et al. 2000). Furthermore, this same contig contained the Siglec-8 gene, also described by others (Floyd, Ni et al. 2000) and by our group (Foussias, Yousef et al. 2000). Therefore, through EcoRI mapping, as described above, as well as the known locations of both the Siglec-8 and -9 genes, we were able to determine the location of the SLG gene. We found that the SLG gene is located 32.9 kb more telomeric than the Siglec-8 gene, and approximately 363 kb downstream of the Siglec-9 gene, on chromosome 19q13.4 (Fig. 13).
Figure 12: Tissue Expression Profile of Both SLG Splice Forms. RT-PCR was performed on 25 tissue total RNAs, for both SLG-S and SLG-L, as well as actin (control gene). SLG-S was found to be highly expressed in spleen, small intestine and adrenal gland. SLG-L, on the other hand, was found to be highly expressed in spleen, small intestine, and bone marrow. Relatively lower levels of expression were found for both these splice forms in various tissues. The molecular weights of the expected cDNA bands were 991 bp and 1345 bp, for SLG-S and SLG-L, respectively. The observed bands closely matched these molecular weights. The multiple bands observed for SLG-S in spleen, small intestine and adrenal gland were found to be of identical sequence, following purification and automated DNA sequencing. Some very faint bands may have been lost during reproduction of this figure.
Figure 13: Physical Map of Chromosome 19q13.4 and the Siglec-3-like Subgroup Locus. Shown schematically above is the developing physical map for the Siglec-3-like subgroup of Siglecs. Gene lengths are presented below each arrow, and distances between genes are shown above. Arrows denote the direction of transcription.
5.3.5. Phylogenetic Analysis of the Expanded Siglec Family

Phylogenetic analysis of the previously existing members of the Siglec family, as well as the novel Siglecs identified throughout this work (Siglec-9, Siglec-8-L, and SLG-S and -L) was accomplished through the Phylip software package, using the UPGMA method. As is evident in Figure 14, the novel Siglecs we have identified are all tightly clustered among the other previously known members of the Siglec-3-like subgroup. In particular, the novel Siglecs we have identified are closely related to Siglec-7, as well as the previously reported Siglec-8, while showing a somewhat more distant relationship with Siglec-3, -5, and -6.
Figure 14: Phylogenetic Analysis of the Siglec Family. Phylogenetic analysis was performed for all the known members of the Siglec family along with those we identified during these investigations. As seen by others, the Siglec-3-like subgroup exhibits a much closer relationship amongst themselves, than they do to either of Siglec-4a, Siglec-1, or Siglec-2. Siglec-9, Siglec-8-L, SLG-S, and SLG-L all appear tightly clustered, along with Siglec-7 and Siglec-8, suggesting a much greater evolutionary relationship amongst these members.
6. DISCUSSION

Using the positional cloning approach we have identified Siglec-9, a novel gene belonging to the Siglec family, in particular, the Siglec-3-like subgroup of Siglecs. This gene is comprised of 7 exons, with 6 intervening introns. The coding region of this gene is composed of 1,392 nucleotides, producing a 463 amino acid protein, with a predicted molecular mass of 50.1 kDa. This gene is located at 19q13.4, 43.19 Kb telomeric to the newly identified kallikrein KLK14. The high degree of homology between this novel Siglec and other members of the Siglec-3-related subgroup provides strong evidence that this protein also plays a role in sialic acid-dependent protein-glycoprotein or -glycolipid interactions. It possesses the unique pattern of conserved cysteine residues in its Ig-like domains, which are found only in members of the Siglec family. Further, Siglec-9 possesses the conserved arginine residue, which has been found to be essential for sialic acid binding (van der Merwe, Crocker et al. 1996). Of note, however, is that it only possesses one of the two conserved aromatic residues in the V-set domain, critical for sialic acid binding in Siglec-1 (May, Robinson et al. 1998), thus differing from other members of the Siglec-3-like subgroup. Following our identification of this novel Siglecs, investigations by other groups revealed that Siglec-9 does, in fact, mediate binding in a sialic acid-dependent manner, to both α2-3 and α2-6-linked sialic acids (Angata and Varki 2000; Zhang, Nicoll et al. 2000). These findings suggest that this aromatic residue, conserved in other members of the Siglec-3-like subgroup, may not be critical in mediating sialic acid-dependent binding by Siglecs.

We examined the tissue expression profile of Siglec-9 and found that it is highly expressed in bone marrow, placenta, spleen, and fetal liver. The high level of expression
of this novel Siglec in bone marrow, and tissues involved in stem cell differentiation, is consistent with findings from groups investigating the other Siglec family members. All currently known Siglecs have been found to be expressed in some type of bone marrow stem cell-derived cell, ranging from myeloid progenitor cells for Siglec-3 and eosinophils for Siglec-8, to natural killer cells for Siglec-7 and B lymphocytes for Siglec-2. We speculated that Siglec-9 is also predominantly expressed on a distinct subset of immune cells, where it plays an intercellular signaling role. Detailed expression studies by other groups, using flow cytometry techniques in conjunction with monoclonal antibodies directed against its extracellular domain, revealed that Siglec-9 is expressed on neutrophils, monocytes, and small population of poorly characterized natural killer cells (CD16+,CD56-) (Angata and Varki 2000; Zhang, Nicoll et al. 2000). Thus, Siglec-9 represents the first member of the Siglec family to be expressed on all three cell populations. It has been suggested that, given this wide distribution of Siglec-9 on cells involved in the innate immune response, along with its potential as a negative regulator of signal transduction due to the presence of the cytoplasmic tyrosine-based inhibitory motifs, this novel Siglec may function as a general negative regulator of “rapidly responding” cells (Angata and Varki 2000).

Through our continued investigation of chromosome 19q13.4 and the potential Siglec-3-like subgroup locus, we also identified the genomic area containing the Siglec-8 gene, located approximately 330 kb downstream of the Siglec-9 gene. Examination of this area revealed the existence of an alternative form of Siglec-8, designated Siglec-8-L. The protein product of this mRNA species differs markedly from the previously published Siglec-8 (GenBank Accession No. AF195092) at its C-terminus (Floyd, Ni et
al. 2000). Siglec-8-L is a 499 residue protein with a molecular weight of 54 kDa. It is encoded by seven exons, and unlike Siglec-8, it shows a high degree of homology at its C-terminus to the other members of the Siglec-3-like subgroup of Siglecs. Consistent with its inclusion in this subfamily of Siglecs, Siglec-8-L also possesses the two characteristic tyrosine-based motifs.

The Siglec-8 mRNA published by Floyd et. al. (2000) contains an abbreviated C-terminus, lacking the characteristic tyrosine-based motifs reported in other members of the Siglec-3-like subgroup of Siglecs (Floyd, Ni et al. 2000). Further, based on our characterization of its genomic structure, the Siglec-8 mRNA species reported by Floyd et. al. contains approximately 1.5 kb of untranslated sequence at its 3' end. By comparison, and in keeping with the hypothesis that this subgroup of Siglecs arose through gene duplication relatively recently in vertebrate evolution (Angata and Varki 2000; Zhang, Nicoll et al. 2000), none of the other Siglecs that belong to this subgroup have such an extensive 3’ untranslated region, or any untranslated exons at the 3’ end. Furthermore, the intron/exon splice sites for the last two exons of the Siglec-8 mRNA transcript (based on our genomic sequence and the mRNA sequence of Floyd et. al.) are not consistent with the characteristic splice donor and acceptor sites (Iida 1990), unlike the first five exons and the two exons identified by us for the Siglec-8-L transcript (Fig. 6).

The identification of Siglec-8 by Floyd et. al. (2000) was accomplished through the use of an EST that showed homology to Siglec-3 (Floyd, Ni et al. 2000). We have not as yet been able to identify this clone, and believe that the discrepancy between our sequence and theirs may be due to the inclusion of genomic sequence in the EST.
Wolfsberg and Landsman have reported that they have indeed found several instances in which ESTs contained intron sequences or unusual splice sites (Wolfsberg and Landsman 1997). If this is the case with Siglec-8, the EST used by Floyd et. al. (Floyd, Ni et al. 2000) may represent a partially spliced or incorrectly spliced sequence, which results in the inclusion of two pieces of genomic DNA (not present in any mRNA species we were able to identify) as well as the absence of an entire exon. During their identification of Siglec-8, the authors report an unsuccessful attempt of Northern blot analysis for Siglec-8 using a probe derived from the coding and 3' untranslated sequences of their mRNA species. Based on the above noted problems with ESTs, the lack of specific bands in the Northern blot may be due to such intronic sequence contamination. This contamination likely resulted in a change in the open reading frame of the Siglec-8 mRNA which caused premature termination and loss of the two tyrosine-based motifs. Despite efforts to identify the mRNA species reported by Floyd et. al. (Floyd, Ni et al. 2000), we have been unsuccessful in obtaining any PCR product that could be sequenced. However, we can still not rule out the possibility that this mRNA species does in fact represent an alternatively spliced form that lacks the two tyrosine motifs and is present in tissues not tested by us.

During their characterization of Siglec-8, the authors demonstrated eosinophil-specific expression for this member of the Siglec-3-like subgroup of Siglecs (Floyd, Ni et al. 2000). This was accomplished by generating monoclonal antibodies which recognize the extracellular domains of Siglec-8. Given that Siglec-8 and Siglec-8-L are identical in their extracellular domains, the expression of Siglec-8-L also appears to be restricted to eosinophils.
For several of the other members of the Siglec family, alternative splice forms have been described. Siglec-1 in mice has been found as either the 17 Ig-like domain form with a single transmembrane domain, or as one of two secreted forms with 3 or 16 Ig-like domains, both lacking the transmembrane domain (Crocker, Mucklow et al. 1994). For Siglec-2, two isoforms have been identified in humans, with either four or six C2-set Ig-like domains (Stamenkovic and Seed 1990; Wilson, Fox et al. 1991; Wilson, Najfeld et al. 1993). In the mouse and rat there have been three and two isoforms, respectively, of Siglec-4a identified (Lai, Watson et al. 1987; Fujita, Sato et al. 1988; Fujita, Sato et al. 1989). One of the murine isoforms lacks an untranslated exon in the 5' end of the mRNA, which has no effect on the size of the resultant polypeptide. The two other forms differ in the presence of exon 12, which is 45 nucleotides in length and introduces a termination codon when included in the mRNA, as is the case for rat Siglec-4a. In humans, however, there has been no report of any Siglec-4a isoforms. Further, two isoforms for Siglec-3 have been reported in the mouse, which differ by an 83 nucleotide in-frame insertion in the cytoplasmic domain (Tchilian, Beverley et al. 1994). Human Siglec-3 has also been reported to exist as two different size transcripts, which is believed to be through the use of alternative polyadenylation signals, with no change in the size of the polypeptide (Simmons and Seed 1988). In addition, two isoforms of human Siglec-7 have been reported, with the smaller isoform lacking the first C2-set Ig-like domain (Falco, Biassoni et al. 1999). By comparison, the drastic differences in splicing patterns of Siglec-8 and Siglec-8-L are not seen in other Siglecs. The only case in which there may be some similarity is murine Siglec-1. However, even there, the
unspliced intron does not appear to disrupt the preceding exon, nor is there an additional downstream untranslated exon.

Further investigations of chromosome 19q13.4 and attempts to identify additional novel members of the Siglec-3-like subgroup of Siglecs, led to the identification and characterization of a novel gene encoding a two putative Siglec splice variants, designated Siglec-like gene (SLG) –Short (SLG-S) and –Long (SLG-L). This novel Siglec gene was localized to chromosome 19q13.4, 32.9 kb downstream of the Siglec-8 gene, and approximately 370 kb telomeric to the Siglec-9 gene. The SLG gene we have characterized is comprised of 8 exons, with 7 intervening introns. All intron/exon splice sites are consistent with the consensus sequence for splice donor/acceptor sites (mGTAAGT...CAGm, where m is any base) (Iida 1990). The SLG-S mRNA transcript consists of exons 2 through 8, while the SLG-L transcript contains exons 1 through 8, with part of exon 2 being removed through splicing. The first exon of each of the two isoforms contains a 5' untranslated region of at least 20bp, while the last exon possesses a 282bp 3' untranslated region. The putative coding sequence of the SLG-S transcript consists of 1736 nucleotides, which encode for a 477 amino acid protein with a predicted molecular mass of 51.7 kDa. The putative translation initiation codon is consistent with the Kozak consensus sequence (Kozak 1991). The SLG-L transcript, on the other hand, has a 2070 bp coding sequence, producing a putative protein product of 595 amino acids with a predicted molecular weight of 65 kDa.

In the case of the SLG-L transcript, unlike that of SLG-S, there is no identifiable Kozak consensus sequence. However, it is the only potential initiation codon which not only produces an extensive open reading frame, but also produces a signal peptide highly
homologous to that seen in other Siglec-3-related Siglecs and at the same time maintains
the reading frame seen in SLG-S through exon 2 and onwards. Interestingly, although
the highly conserved purine at position -3 in the Kozak consensus sequence is absent,
this has been noted in a few other vertebrate mRNA species as well (reviewed in (Kozak
1991)). These mRNAs that lack the preferred nucleotide in both key positions, -4 and +3,
encode potent regulatory proteins such as growth factors and cytokines. This suggests
that the absence of strict adherence to the Kozak consensus sequence may provide a
mechanism for modulating the yield of proteins which may be harmful if overproduced
(Kozak 1991).

Examination of the tissue expression profile of both SLG transcripts revealed that
SLG-S was relatively highly expressed in spleen, small intestine, and adrenal gland. On
the other hand, the SLG-L transcript exhibited relatively high levels of expression in
spleen, small intestine, and bone marrow. From this it appears quite likely that both
transcripts are expressed in tissues, particularly spleen and bone marrow, in which
haemopoietic cell lineages are found. This is consistent with findings for other members
of the Siglec-3-related subgroup of Siglecs, as well as that of other members of the Siglec
family.

Based on examination of the homology between the putative protein products of
both the SLG-S and -L transcripts with other known members of the Siglec-3-like
subgroup, it is evident that this gene, and its two isoforms, likely represents the newest
addition to the expanding Siglec-3-like subgroup of Siglecs. As is evident in Figure 10,
both SLG-S and SLG-L protein products contain many of the structural characteristics
possessed by other Siglec-3-like Siglecs discovered thus far. They both contain the
distinctive distribution of cysteine residues found in all Siglecs, necessary for the unique folding pattern of their Ig-like domains (Pedraza, Owens et al. 1990; Crocker, Kelm et al. 1996). They both also contains the two highly conserved aromatic residues believed to be involved in sialic acid binding (van der Merwe, Crocker et al. 1996). The conserved arginine residue, which is present in all other Siglecs and believed to be essential for sialic acid binding, is replaced by glutamine in the case of the SLG-S protein, and cysteine in the SLG-L protein. Given that both conserved aromatic residues, found to be critical for sialic acid binding by Siglec-1 (May, Robinson et al. 1998), were not so critical in the case of Siglec-9, it is unclear what effects the absence of this conserved arginine residue will have on sialic acid-dependent binding by these SLG isoforms. Studies are currently in progress to assess the effect these changes have on sialic acid binding.

A unique feature evident in the putative SLG-L protein is the presence of an additional N-terminal V-set Ig-like domain. This is the first report of a putative Siglec in which more than one N-terminal V-set domain has been observed. Interestingly, the current hypothesis regarding the origin of this family in humans is that it developed through the process of gene duplication and exon shuffling through the aid of chromosome 19q13.1-qter - specific minisatellites (Das, Jackson et al. 1987; Angata and Varki 2000). If this is indeed how this subgroup of Siglecs evolved, then it is possible that during a recombination event in which there was unequal exchange of genetic information in this region of 19q13.4, a Siglec-3-like gene with one N-terminal V-set domain was inserted downstream of the first exon of another Siglec-3-like gene. This would result in generation of a Siglec-3-related gene capable of producing two mRNA
transcripts, the first of which would maintain the typical arrangement of Ig-like domains, as in the SLG-S transcript. However, an additional transcript would also be possible, in which transcription was initiated from the pre-existing first exon, yielding a mRNA encoding two N-terminal V-set Ig-like domains followed by two C2-set Ig-like domains, as seen in the SLG-L transcript.

The novel members of the Siglec-3-related subgroup of Siglecs we have identified, Siglec-9, Siglec-8-L, and the two SLG isoforms, all exhibit tissue specific expression, as discussed earlier. These findings are consistent with those observed for other members of this subgroup, which were found to be restricted to one or a few cell types (Simmons and Seed 1988; Stamenkovic and Seed 1990; Crocker, Mucklow et al. 1994; Kelm, Schauer et al. 1994; Cornish, Freeman et al. 1998; Falco, Biassoni et al. 1999; Nicoll, Ni et al. 1999; Patel, Brinkman-Van der Linden et al. 1999; Ulyanova, Blasioli et al. 1999; Angata and Varki 2000; Floyd, Ni et al. 2000; Foussias, Yousef et al. 2000; Zhang, Nicoll et al. 2000). Based on the gene duplication hypothesis, one would expect to see similar patterns of expression for all the members of the Siglec-3-like subgroup. However, through studies of the γ-globin gene in several species, it was found that such unequal exchanges may also introduce changes at the nucleotide level, in both coding and untranslated regions (reviewed in (Fitch, Bailey et al. 1991)). Such changes are believed to be involved in the regulatory changes that resulted in delayed expression of some globin genes from embryonic to fetal life. The model proposed for the evolution of the γ-globin gene is one of gene duplication with unequal cross-over events, similar to that suggested for the Siglec-3-like subgroup, mediated by interspersed repetitive elements (Fitch, Bailey et al. 1991). These findings for the γ-globin gene, and their
consequent regulatory effects, may explain how the highly homologous Siglec-3-related subgroup acquired their observed haemopoietic cell specificity.

The previously known members of the Siglec-3-related subgroup of Siglecs were all localized to chromosome 19q13.3-13.4 through in situ hybridization techniques (Angata and Varki 2000; Zhang, Nicoll et al. 2000). During our characterization of novel members of this subgroup, we were able to precisely localize them to chromosome 19q13.4, immediately following the human Kallikrein gene family (Yousef, Luo et al. 1999; Diamandis, Yousef et al. 2000). Based on this tight linkage among the novel Siglecs we have identified, as well as preliminary observations for the remainder of this region, it appears extremely probable that the entire Siglec-3-like subgroup locus is clustered in this region of chromosome 19q13.4 (our unpublished observations). This is consistent with both the chromosomal localization by other groups, and the gene duplication hypothesis for this subgroup of Siglecs.

The cytoplasmic tyrosine-based motifs, ITIM and SLAM-like, are present in all members of the Siglec-3-like subgroup of Siglecs, with the exception of Siglec-8 (Floyd, Ni et al. 2000). In addition, these characteristic motifs have been identified in all four of the novel Siglec mRNA transcripts we have identified, Siglec-9, Siglec-8-L, and both SLG isoforms. The functional significance of the more C-terminal SLAM-like motif remains unclear. What is known is that, upon phosphorylation, it recruits the SLAM-associated protein (SAP), blocking recruitment of the phosphatase SHP-2 (Coffey, Brooksbank et al. 1998; Sayos, Wu et al. 1998). The other cytoplasmic tyrosine-based motif, located upstream of the SLAM-like motif, is the ITIM motif, which as been more extensively studied. Investigations have revealed that it is the binding site for the SH2
domains of the tyrosine phosphatases SHP-1 and SHP-2, as well as the inositol phosphatases SHIP-1 and SHIP-2 (Borges, Hsu et al. 1997; Le Drean, Vely et al. 1998; Muraille, Bruhns et al. 2000).

Examination of the functional significance of these motifs, primarily the ITIM motif, have only been conducted for two Siglec-3-like subgroup members, Siglec-3 itself, and Siglec-7. In both cases, phosphorylation of this motif has been shown to be responsible for the recruitment of SHP-1, as well as SHP-2 in the case of Siglec-3 (Falco, Biassoni et al. 1999; Taylor, Buckley et al. 1999). Further, engagement of both these Siglecs has been found to result in tyrosine phosphorylation of this ITIM motif, with consequent inhibition of proliferation, in both normal and leukemic myeloid cells in vitro (Vitale, Romagnani et al. 1999). These findings strongly suggest that these Siglecs may normally play an inhibitory role in myeloid cells, and also in natural killer cells by Siglec-7. The mechanism by which this inhibition occurs is currently unclear, but preliminary investigations suggest that engagement may induce apoptosis (Vitale, Romagnani et al. 1999). By extensions, given that most of the members of the Siglec-3-related subgroup, including those identified by us, all possess these same tyrosine-based cytoplasmic motifs, the remainder of this subgroup may also be involved in regulating the proliferation of the particular haemopoietic cells on whose cell surface they are expressed.

The involvement of Siglec-3-like subgroup members in the development of haematological malignancies remains to be demonstrated. However, chromosomal aberrations involving the chromosome 19q13 band have been observed in a variety of both solid and haematological malignancies (Mitelman 1994). Several reports implicate
the band 19q13.3, including the adjacent Kallikrein gene family, in several solid tumours including ovarian cancer (Yousef and Diamandis; Mitelman 1994; Diamandis, Yousef et al. 2000). Further, several haematological malignancies exhibit chromosomal aberrations in the 19q13 band as well, including acute and chronic myeloid leukemias (AML and CML, respectively), acute lymphocytic leukemia (ALL), non-Hodgkin’s lymphoma, and chronic lymphoproliferative disorder (Cancer Chromosome Abnormalities Project (CCAP), and (Mitelman 1994)). Given the recent characterization of this family of cell surface receptors, continued investigation will be required in order to elucidate the precise involvement of this proteins in the development of malignancy.

In addition to its potential involvement in haematological malignancy development, Siglec-3 has been utilized for several years in differentiating primitive myelolastic forms of AML from ALL (Griffin, Linch et al. 1984; Dinndorf, Andrews et al. 1986; Bernstein, Singer et al. 1992). The expression of Siglec-7 on myeloid cells raises the possibility that it too may represent a useful marker for accurate leukemic cell typing as well (Vitale, Romagnani et al. 1999). Studies into the therapeutic utility of anti-Siglec-3 antibodies have shown some success in the treatment of myelogenous leukemias (Caron, Schwartz et al. 1994; Caron, Lai et al. 1995; Caron, Dumont et al. 1998). Further, phase I clinical studies for the treatment of AML have revealed that the use of monoclonal anti-Siglec-3 antibodies results in selective ablation of malignant haemopoiesis (Kossman, Scheinberg et al. 1999; Sievers, Appelbaum et al. 1999). In light of the inhibitory effect of Siglec-3 engagement on leukemic myeloid cell proliferation, it is possible that the observed selective ablation may result not only from antibody-dependent, cell-mediated cytotoxicity and phagocytosis, but also by through an
inhibitory effect on leukemic cell proliferation (Vitale, Romagnani et al. 1999). Based on the extensive homology between the novel members of the Siglec-3-like subgroup of Siglecs we have identified and the previously known members of this subgroup, it is possible that these novel Siglecs may also possess some utility as targets for immunological antineoplastic therapy of other haematological malignancies.
7. REFERENCES


Mucklow, S., A. Hartnell, et al. (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(2): 344-6.


