THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 276, No. 43, Issue of October 26, pp. 40282–40287, 2001

Printed in U.S.A.

A Second Uniquely Human Mutation Affecting Sialic Acid Biology*[S]

Received for publication, June 26, 2001, and in revised form, August 1, 2001
Published, JBC Papers in Press, August 23, 2001, DOI 10.1074/jbc.M105926200

Takashi Angata, Nissi M. Varki, and Ajit Varki‡

From the Glycobiology Research and Training Center, Departments of Medicine and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, California 92093-0687

Siglecs are immunoglobulin superfamily member lectins that selectively recognize different types and linkages of sialic acids, which are major components of cell surface and secreted glycoconjugates. We report here a human Siglec-like molecule (Siglec-L1) that lacks a conserved arginine residue known to be essential for optimal sialic acid recognition by previously known Siglecs. Loss of the arginine from an ancestral molecule was caused by a single nucleotide substitution that occurred after the common ancestor of humans with the great apes but before the origin of modern humans. The chimpanzee Siglec-L1 ortholog remains fully functional and preferentially recognizes *N*-glycolylneuraminic acid, which is a common sialic acid in great apes and other mammals. Reintroducing the ancestral arginine into the human molecule regenerates the same properties. Thus, the single base pair mutation that replaced the arginine on human Siglec-L1 is likely to be evolutionarily related to the previously reported loss of *N*-glycolylneuraminic acid expression in the human lineage. Siglec-L1 and its chimpanzee Siglec ortholog also have a different expression pattern from previously reported Siglecs because they are found on the lumenal edge of epithelial cell surfaces. Notably, the human genome contains several Siglec-like pseudogenes that have independent mutations that would have replaced the arginine residue required for optimal sialic acid recognition. Thus, additional changes in the biology of sialic acids may have taken place during human evolution.

The great apes are the closest evolutionary relatives of humans, with the chimpanzee/bonobo clade likely having shared a last common ancestor with humans about 6–7 million years ago (1–5). Human genomic DNA sequences differ on average by only 1–2% from those of these great apes (3–5). Thus, the altered expression of relatively few gene products may underlie some of the obvious morphological and functional differences between the species. We and others recently discovered an inactivating mutation in the human gene encoding CMP-*N*-acetylneuraminic acid hydroxylase (6, 7), an enzyme that is functional in the great apes (6). This explains the human-specific loss of a major sialic acid, *N*-glycolylneuraminic acid (Neu5Gc)[1] (8).

The sialic acids are a family of 9-carbon sugars that are abundantly expressed on cell surfaces and secreted glycoconjugates of animals of the deuterostome lineage (9–11). Located mostly at the outer end of glycan chains on glycoproteins and glycolipids, sialic acids mediate a variety of recognition events involving pathogenic microbes and toxins, as well as endogenous animal lectins (12, 13). Siglecs are the largest family of such endogenous sialic acid-recognizing lectins defined to date (14, 15). All 10 reported human Siglecs are type I membrane proteins, consisting of an amino-terminal Ig V-set domain, variable numbers of Ig C2-set domains, a single-pass transmembrane domain, and a cytoplasmic tail typically containing tyrosine-based signaling motifs. Sialic acid recognition is mediated by the first Ig V-set domain (15–19), and certain amino acid residues invariant to this domain are known to be involved in interactions with the sialic acid ligand (20). In particular, all Siglec V-set domains have a conserved arginine residue that forms a salt bridge with the carboxylate group of sialic acids. Experimental mutation of this residue markedly diminishes binding in all Siglecs studied to date (18, 19, 21–23).

Here we report a human molecule that has many features of Siglecs but lacks robust sialic acid recognition because of a specific mutation of the "essential" arginine residue, which remains conserved in its great ape orthologs. We also consider potential connections to the previously described human-specific mutation involving sialic acid biology and search the human genome for Siglec-like pseudogenes, several of which turn out to have independent mutations replacing the "essential" arginine residue. After this work was completed, another group independently reported the cloning of the same molecule, which they called S2V, a "putative" Siglec (24). Our study shows that the "essential" arginine residue, which was mutated to cysteine specifically in the human lineage, is required for easily detectable sialic acid-dependent recognition typical of previously reported Siglecs. By comparison, we demonstrate robust sialic acid-dependent recognition by the fully functional chimpanzee ortholog and by the human molecule when the arginine is restored. Thus, we do not consider the native human molecule a *bona fide* Siglec; instead, we call it Siglec-like molecule 1 (Siglec-L1).

EXPERIMENTAL PROCEDURES

cDNA Cloning and Preparation of Expression Constructs—A human expressed sequence tag clone (GenBank™ accession number AI132995) encoding transmembrane and cytosolic domains of Siglec-L1 was iden-

* This work was supported by United States Public Health Service Grants R01-GM323373 and P01-HL57345 and by the G. Harold and Leila Y. Mathers Charitable Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplementary information.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF282256 (human), AF293372 (chimpanzee), AY029752 (bonobo), AY029754 (gorilla), and AY029756 (orangutan).

‡ To whom correspondence should be addressed: University of California at San Diego School of Medicine, 9500 Gilman Dr., MC 0687, La Jolla, CA 92093-0687. Tel.: 858-534-3296; Fax: 858-534-5611; E-mail: avarki@ucsd.edu.

[1] The abbreviations used are: Neu5Gc, *N*-glycolylneuraminic acid; Neu5Ac, *N*-acetylneuraminic acid; PCR, polymerase chain reaction; Siglec-L1, Siglec-like molecule 1; UTR, untranslated region; LacNAc, *N*-acetyllactosamine.

tified by BLAST search (25) of the GenBank™ data base. A cDNA for the Siglec-L1 full-length coding sequence was obtained by 5′-rapid amplification of cDNA ends from human fetal liver Marathon-Ready cDNA library (CLONTECH) using primers 5′R-SP1 (5′-GGGAGAGTTTGGTCATCAGGCACG-3′) and AP1 (included in the kit), followed by a nested PCR using primers 5′R-SP2 (5′-CTGATGTCCTGTTGCACAGCATGG-3′) and AP2 (included in the kit). The 3′-untranslated region (UTR) was obtained likewise, using 3′R-SP1 (5′-GTGGGCGTGGGGGATACAGGCATGG-3′) and AP1 for the first PCR, followed by a nested PCR using 3′R-SP2 (5′-CCAGTATGCATCCCTCAGCTTCCAC-3′) and AP2. The DNA fragments were cloned into pCRII-TOPO (Invitrogen), and the sequences were analyzed. Based on these sequences, new primers were designed: 5′Expr (5′-CCTTCTAGAGCCACCATGCTACTGCTGCTGCTACTGCTGCC-3′; the *Xba*I site is *underlined*), 3′-UTR (5′-CTGAAGCTTCTAAAGGAATCAGTCTCGGG-3′; the *Hin*dIII site is *underlined*), and 3′Chi 3D (5′-ATCTCCTTGGAAGACAGTCATGGTC-3′). The PCR product with 5′Expr and 3′-UTR (containing full coding sequence) was digested with *Xba*I and *Hin*dIII and ligated into appropriate sites of pcDNA3.1 (Invitrogen) for functional analysis. The PCR product with 5′Expr and 3′Chi 3D was digested with *Xba*I and ligated into the *Xba*I-*Eco*RV sites of EK-Fc-pEDdC vector for recombinant fusion protein production. Mammalian cell transfection with this construct gives a fusion protein of Siglec-L1 (first three Ig-like domains) and human IgG Fc tail (22, 23, 26).

The chimpanzee Siglec-L1 ortholog cDNA was obtained by reverse transcription-PCR from spleen total RNA (kindly provided by Dr. Elaine Muchmore, San Diego VA Hospital, San Diego, CA) using primers 5′-UTR (5′-TCAGCAACCCTGGCACCTCCAACCC-3′) and 5′R-SP2, followed by nested PCR using primers c5′Expr (5′-CCCTCTAGAGCCACCATGCTACTGCTGCTGCTACTGCTACTGCTGCC-3′; the *Xba*I site is *underlined*) and c3′-UTR (5′-CTGAAGCTTCTAAAGGAATCAGCCTCGGG-3′; the *Hin*dIII site is *underlined*). The DNA fragments were digested with *Xba*I and *Hin*dIII and ligated into appropriate sites of pBluescript II (Stratagene), and the sequences were analyzed. A sequence-verified full-length cDNA was subcloned to pcDNA3.1 for functional analysis.

The C122R "reverse" mutation was introduced into human Siglec-L1/pcDNA3.1 using the QuickChange Site-Directed Mutagenesis Kit (Invitrogen) with primers C122R S (5′-GGGACATACGTCTTTCGTGTAGAGAGAGG-3′; Cyt^{364} is *underlined*) and C122R AS (complementary to C122R S).

Sequence Analysis of Genomic DNA Fragments Encoding Human Siglec-L1 and Its Great Ape Orthologs—Genomic DNA from chimpanzee, bonobo, gorilla, and orangutan was prepared from Epstein-Barr virus-transformed lymphoblastoid cell lines of respective species (kindly provided by Dr. Peter Parham, Stanford University). Genomic DNA from human subjects was either obtained from the peripheral blood of healthy donors or kindly provided by Dr. Stephen Warren (Emory University). DNA fragments containing the first four exons of Siglec-L1 orthologs (~2.5 kb) were amplified using primers 5′-UTR and 3′Chi 3D, and the first two exons were sequenced using the following primers: 5′-UTR, Intr1 AS (5′-TTTGGCCTCTCTTGGAGCCC-3′), and Intr1 AS2 (5′-CTTTGGCCTCTCTTGGAGCC-3′; for some human samples) for exon 1, and Ig2F (5′-GGGCTGACCCTCGTTTCTCC-3′) and Ig2R (5′-GGCTGGTCCTAGAGCCAGAG-3′) for exon 2.

Synthesis of Probes Containing Neu5Gc or Neu5Ac—A biotinylated linear polyacrylamide (PAA-Bio) polyvalently conjugated with *N*-acetyllactosamine (LacNAc; 20% mol/mol) was purchased from Glycotech (Rockville, MD). Sialyltransferase ST6Gal-I was a kind gift from Amgen Inc. (Thousand Oaks, CA). CMP-Neu5Gc was synthesized as described previously (27) from 5 μmol of CTP and 2.5 μmol of Neu5Gc (Calbiochem; >99% pure), using 2 units of *Haemophilus ducreyi* CMP-sialic acid synthase (kindly provided by Dr. Bradford Gibson, University of California, San Francisco, CA). The reaction product was treated with alkaline phosphatase and purified by ultrafiltration (Microcon-10; Millipore), anion exchange (AG-1x8; Bio-Rad) and charcoal cartridge (Glyko) chromatography. LacNAc-PAA-Bio (50 μg; containing ~50 nmol of LacNAc) was incubated with 250 nmol of CMP-Neu5Gc or CMP-Neu5Ac (Sigma) and 80 milliunits of ST6Gal-I (28). The reaction products were ultrafiltered to remove substrates, recovered in water, and stored frozen until use. Equivalent substitution with both sialic acids was obtained, as determined by fluorohetric HPLC analysis (29).

Analysis of Sialic Acid Preference of Siglec-L1 by Flow Cytometry—Bicistronic expression constructs with green fluorescent protein in the second open reading frame were prepared by subcloning full-length cDNA into pIRES2-EGFP (CLONTECH). The resulting constructs were transfected to 293H cells using LipofectAMINE 2000 (Life Technologies, Inc.). The cells were dispersed in 5 mM EDTA in phosphate-buffered saline 48 h after transfection, treated with *Arthrobacter ureafaciens* sialidase (10 milliunits/10^6 cells) at 37 °C for 1 h, and washed extensively with 1% bovine serum albumin in phosphate-buffered saline. The cells (10^6) were incubated with 1 μg of PAA-Bio probes carrying Neu5Ac-LacNAc, Neu5Gc-LacNAc, or nonsialylated LacNAc at 4 °C for 1 h, washed, and then incubated with 1 μg of streptavidin-phycoerythrin conjugate (Jackson ImmunoResearch) at 4 °C for 30 min. The cells were washed again, suspended in 1% bovine serum albumin-phosphate-buffered saline, and subjected to flow cytometry using FACScan (Becton Dickinson). Green fluorescence-positive cells ($FL1^{high}$, the cells successfully transfected) were gated and analyzed for probe binding (phycoerythrin fluorescence, FL2).

Preparation of a Monospecific Chicken Polyclonal Antibody against the Extracellular Domain of Siglec-L1—Recombinant human Siglec-L1-Fc was prepared and used for immunization of Rhode Island Red hens (23). To subtract immunoglobulin subfractions that cross-react with human IgG Fc or Siglec-7 (the closest paralog of Siglec-L1), the preimmune and postimmune sera (500 μl) were incubated with 200 μl of IgG-Sepharose (Amersham Pharmacia Biotech), followed by incubation with Siglec-7-Fc (100 μg) immobilized on protein A-Sepharose (Amersham Pharmacia Biotech). The post-immune serum processed as described above was specific for Siglec-L1, as judged by an enzyme-linked immunosorbent assay using Siglec-L1-Fc, Siglec-7-Fc, and L-selectin-Fc as test samples (data not shown).

Immunohistochemical Analysis of Human and Chimpanzee Tissues—Chimpanzee tissues were kindly provided by Yerkes Primate Center. Paraffin-embedded formalin-fixed tissue sections mounted on glass slides were deparaffinized and incubated overnight at 4 °C with 1:200 diluted chicken preimmune or postimmune sera prepared as described above, followed by washing and incubation with 1:200 diluted peroxidase-conjugated $F(ab')_2$ fragment of donkey anti-chicken IgY antibody (Jackson ImmunoResearch). The slides were then washed and developed using TSA-Direct Cyanine 3 kit (PerkinElmer Life Sciences). Sections were viewed using epifluorescence under a Zeiss microscope fitted with a red barrier filter for Cyanine 3 emission. Specificity of the staining was confirmed by subjecting aliquots of postimmune serum (processed as described above) to preadsorption with Chinese hamster ovary cells expressing human Siglec-L1-C122R or parental Chinese hamster ovary K1 cells (10^7 cells/100 μl serum) and using the supernatant for tissue staining.

Phylogenetic Analysis of V-set Domains of Siglec Genes and Pseudogenes—Amino acid sequences of known Siglec V-set domains were used to find similar sequences (E values < 1) in the human genome data base in GenBank™ using the TBLASTN program. Relevant DNA sequences were analyzed for putative splice donor/acceptor sites using a web server (www.fruit fly.org/seq_tools/splice.html). This information and DNA sequence alignments with known Siglec genes were used to predict the borders of exon fossils in the putative pseudogenes. Most of the putative pseudogenes were confirmed by PCR amplification and sequencing of corresponding genomic DNA fragments. Thirteen pseudogenes on the chromosome 19 q arm were numbered in order of their location: the one closest to the centromere was numbered P1. The only pseudogene outside the cluster was numbered P14. This numbering system is as recommended by the Human Gene Nomenclature Committee. One DNA segment that we could not unambiguously assign as a gene or pseudogene was named X. The precise order of some pseudogenes was determined with the help of the publicly accessible portion of the Celera human genome data base (30).

DNA sequences of the exons and exon fossils of V-set domains were aligned by Clustal W (www2.ebi.ac.uk/clustalw/) and minimally adjusted using six anchor points (*i.e.* codons for two aromatic amino acid residues and the arginine residue involved in sialic acid recognition, and three cysteines conserved in known Siglecs (20)). Coding sequences for the signal peptide were then removed because some pseudogenes lack a well-defined signal sequence, and the remaining sequences were analyzed for phylogenetic relationship using PAUP 4.0 (Sinauer Associates). Phylogenetic trees were constructed using the neighbor-joining method (31) with a distance matrix based on absolute distance. Pseudogene P12 was excluded from the analysis because its coding sequence was disrupted by an insertion of Alu element.

Genomic Structures of Siglec-L1 and Its Close Paralogs—Genomic structures of Siglec-L1, Siglec-7, Siglec-8, Siglec-9, and mouse Siglec-E were deduced by comparing the cDNA (Siglec-7, NM_014385; Siglec-8, NM_014442 and AF287892; Siglec-9, AF227924; mouse Siglec-E, AF317298) and genomic DNA sequences of these genes in the GenBank™ data base. The "coding" sequence of pseudogene Siglec-P4 was reconstructed by comparing Siglec-8 cDNA (the closest paralog) with the genomic DNA region containing Siglec-P4. The Siglec-7 genomic

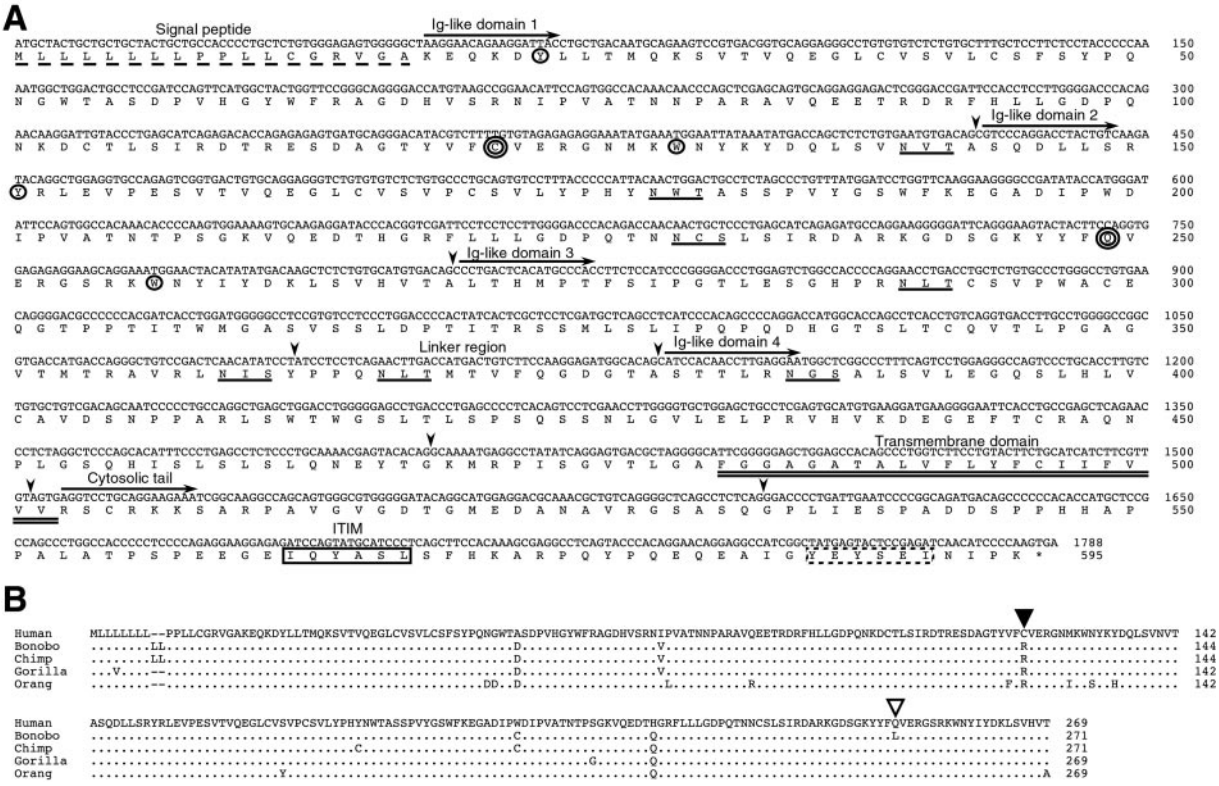

FIG. 1. **Sequences of human Siglec-L1 and its great ape orthologs.** *A,* cDNA and deduced amino acid sequences of human Siglec-L1 (GenBank™ AF282256). *Double circles,* amino acids occupying the expected positions of the "essential" arginine in typical Siglec V-set domains; *circles,* aromatic amino acid residues in the two V-set domains typical of Siglec V-set domains; *underline* with *hatched* and *double lines*, signal peptide and transmembrane domain, respectively; *arrowheads,* exon junctions; *underline,* potential *N*-glycosylation sites; *boxes* with *solid* and *hatched lines,* putative ITIM motif and another tyrosine-based motif conserved among Siglecs, respectively. *B,* alignment of deduced amino acid sequences of the two amino-terminal V-set domains of human Siglec-L1 and the great ape orthologs. Amino acid residues that differ from human Siglec-L1 are indicated. The expected positions of the "essential" arginine in the first and second V-set domains are indicated with *filled* and *open triangles*, respectively.

DNA region in the data base appeared to be misassembled; it was therefore reassembled from some overlapping fragments in draft-stage sequences of BAC clones RP11–423F16 and CTD-3187F8. Repetitive elements were identified by using RepeatMasker (repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker).

RESULTS AND DISCUSSION

A Human Siglec-like Molecule Lacks Arginine Residue(s) Essential for Robust Sialic Acid Recognition—A sequence in the GenBank™ expressed sequence tag data base (GenBank™ accession number AI132995) showed close similarity to the coding regions for carboxyl-terminals of known Siglecs. A complete coding region was obtained via 5′-rapid amplification of cDNA ends. The cDNA encodes a new Siglec-like molecule with 595 amino acids (Fig. 1*A*), predicting a type I membrane protein with four Ig-like domains, a single pass transmembrane domain, and a cytosolic tail with two tyrosine-based putative signaling motifs. All these features are typical of the Siglec-3-related subgroup of the Siglecs (15, 23). Unlike previously known Siglecs that have only one amino-terminal V-set domain, the new molecule has two tandem V-set domains that are similar to each other and to those of other Siglecs. However, both of these V-set domains lack the "essential" arginine residue known to be indispensable for robust sialic acid recognition by previously reported Siglecs.

A classic assay for sialic acid recognition by Siglecs is the sialidase-sensitive rosetting of human erythrocytes onto transiently transfected COS-7 cells expressing the molecule (32). As expected from the lack of the "essential" arginine residues, the newly discovered Siglec-like molecule failed to produce distinct rosetting, as seen in Siglecs we have reported previously (22, 23). A recombinant soluble Fc-fusion protein incorporating the three amino-terminal Ig-like domains of the molecule also failed to show sialic acid binding (data not shown) in a standard enzyme-linked immunosorbent assay (23, 26). Because designation as a Siglec requires sialic acid recognition (14), we have named this protein Siglec-like molecule-1 (Siglec-L1). Given the close sequence similarity of the first V-set domain of Siglec-L1 to the corresponding domains of Siglec-7 and Siglec-9, we hypothesized that the lack of the "essential" arginine residue (universal in humans, see below) represents a derived change from the ancestral state. To explore this possibility, we studied Siglec-L1 orthologs in the great apes.

Great Ape Orthologs Have the Arginine Residue in the First Domain and Can Recognize Sialic Acids—DNA fragments including the first four exons of the Siglec-L1 orthologs of great apes were amplified from genomic DNA of the chimpanzee, bonobo, gorilla, and orangutan, and the exon sequences were analyzed. All great ape orthologs have the "essential" arginine residue in the first Ig-like domain (Fig. 1*B*), encoded by a CG(T/C) codon. Thus, a single nucleotide substitution (C to T) at the first residue of the Arg codon replaced the ancestral Arg with a Cys residue in humans. On the other hand, all great apes as well as humans lack the "essential" arginine residue in the second Ig-like domain (Fig. 1*B*). To confirm that the great apes orthologs are indeed fully functional Siglecs, we cloned a full-length cDNA encoding the chimpanzee ortholog (GenBank™ accession number AF293372). When expressed in COS-7 cells by transient transfection, this molecule gave robust sialic acid-dependent erythrocyte rosetting (Fig. 2*A*).

Restoration of the Arginine Residue in the First V-set Domain of the Human Siglec-L1 Regenerates Robust Sialic Acid Recognition—A reverse mutation (C122R) restoring the ancestral

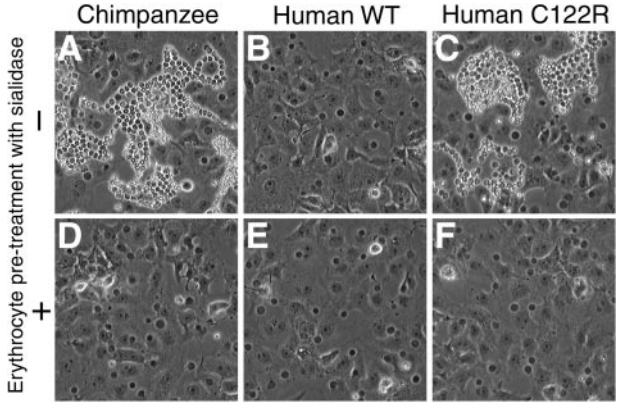

FIG. 2. **Erythrocyte rosetting by transfected COS-7 cells.** COS-7 cells were transiently transfected with expression constructs for the chimpanzee Siglec-L1 ortholog (*A* and *D*), human wild-type Siglec-L1 (*B* and *E*), and human Siglec-L1 C122R reverse mutant (*C* and *F*) and studied for erythrocyte binding (rosetting) as described previously (22, 23). Human erythrocytes were treated (*D*–*F*) or mock-treated (*A*–*C*) with sialidase before the assay.

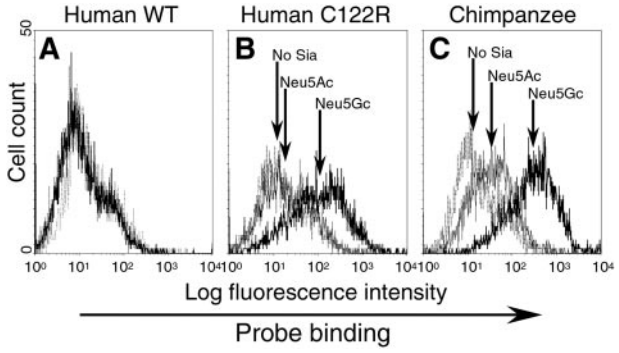

FIG. 3. **Sialic acid binding preference.** Polymeric probes containing equal amounts of terminal Neu5Ac or Neu5Gc were incubated with 293H cells expressing human wild-type Siglec-L1 (*A*), the C122R reverse mutant (*B*), or the chimpanzee Siglec-L1 ortholog (*C*) and analyzed for binding by flow cytometry as described under "Experimental Procedures." Signals are shown for the nonsialylated control probes (*dotted line*), Neu5Ac-containing probes (*thin line*), and Neu5Gc-containing probes (*thick line*).

arginine residue was introduced in the human Siglec-L1 cDNA and studied in the same rosetting assay as described above. As shown in Fig. 2*C,* the "restored" human molecule also showed robust sialic acid-dependent erythrocyte binding. Thus, the single nucleotide substitution that changed the Arg codon to a Cys codon was primarily responsible for diminishing the sialic acid binding property of the human molecule.

There is an apparent discrepancy with another recent study (24) that claimed sialic acid recognition by the native human molecule. This might be explained by the fact that those authors counted transfected COS cells associated with only a few red cells in an assay that seemed to have a signal:noise ratio of about 4:1. In contrast, we saw robust sialic acid-dependent recognition by the chimpanzee ortholog as well as the arginine-restored human molecule, with large clusters of red cells attached to the transfected COS cells (see Fig. 2). Thus, we are confident that the native human molecule does not have robust sialic acid binding properties typical of other Siglecs. It is possible that the binding reported in the other study represents some residual recognition of other aspects of the sialic acid molecule that can occur even in the absence of the arginine residue.

The Arginine Mutation in the First Domain Occurred before the Common Origin of Modern Humans—There is now almost universal agreement that all living humans are very closely related genetically (33, 34). To determine when the arginine-replacing nucleotide change in the first Ig-like domain occurred relative to the common origin of modern humans, we studied human genomic DNA samples from six Asians, eight Europeans/European-Americans, five African-Americans, and five Africans (two Biaka and three Mbuti pygmies). All individuals studied had the same point mutation changing the Arg codon to a Cys codon, indicating that this substitution occurred before the common origin of modern humans. We also noted that some human alleles have a frameshift mutation that must have occurred after the common origin of modern humans. This is currently being investigated further.

Both Chimpanzee and Arginine-restored Human Siglecs Preferentially Recognize Neu5Gc, the Sialic Acid That Is Missing in Humans—This finding represents the second human-specific genetic mutation that involves the biology of sialic acids; the first results in the loss of Neu5Gc expression (6). To explore whether the two events are functionally related, we studied the relative preference of the arginine-restored human Siglec-L1 and the chimpanzee ortholog for their ability to recognize Neu5Ac and Neu5Gc. Probes carrying equal amounts of sialic acids in the Neu5Ac or Neu5Gc form were used to study cells transfected with human wild-type Siglec-L1, the human C122R reverse mutant, or the chimpanzee ortholog. As shown in Fig. 3, the human wild-type Siglec-L1-expressing cells showed no binding of either probe in this assay. In contrast, both the human C122R molecule and the chimpanzee ortholog bound both sialic acids. However, both clearly preferred Neu5Gc over Neu5Ac. Because fluorescence intensity is recorded on a log scale, we can also say that any binding of the wild-type molecule to either sialic acid (if present) is markedly diminished.

Tissue Expression Profiles of Human Siglec-L1 and the Chimpanzee Ortholog—A specific polyclonal chicken antibody against Siglec-L1 stained the lumenal edge of epithelial cells in organs such as the prostate, ileum, stomach and tonsil, and collecting duct cells in the kidney medulla (see Fig. 4 for examples). A similar expression pattern was noted in chimpanzee tissues. To date, there are only two examples of Siglec expression outside the hematopoietic system: Siglec-4 in the nervous system (35), and Siglec-6 in the placenta (26). This is the first example of a Siglec or Siglec-like molecule that is predominantly expressed in epithelial cells. Not all human samples showed positive staining, which may be explained by the presence of frameshift mutations in some human alleles.

The Human Genome Contains Several Siglec Pseudogenes with Independent Mutations Replacing the Essential Arginine Residue—The characteristic amino acid motifs of the V-set domain of known Siglecs allowed us to search for other Siglecs and Siglec-like pseudogenes in the human genome sequence (36). Most of the Siglec-3-related Siglec genes are contained within a ~0.5-Mb region of cytological band 19q13.3–19q13.4, suggesting that they arose by duplications and exon reassortments from an ancestral gene. We found that this region also contains many Siglec-like pseudogenes. The phylogenetic relationship of the V-set domains of these genes and pseudogenes is shown in Fig. 5*A*. In this phylogram, many pseudogenes are found clustering with different functional Siglec genes, suggesting that they were independently derived from functional Siglec genes and eventually inactivated over the course of evolution. Interestingly, 7 of these 14 pseudogenes have mutations at the codon that would have replaced the "essential" arginine residue. Several of these "arginine-replacing" mutations seemed to have been discrete events (Fig. 5*A*). The currently available (incomplete) mouse genome sequence shows evidence for four Siglec-3-like Siglec genes and two Siglec-like

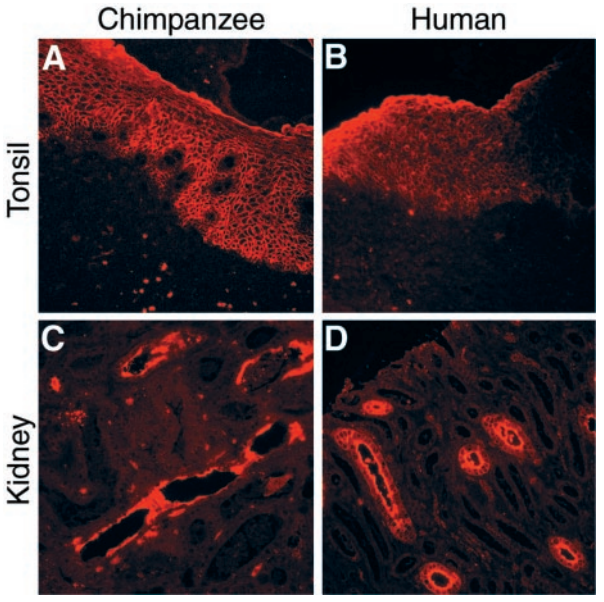

FIG. 4. **Immunohistochemical analysis of Siglec-L1 expression.** *Top panels* show sections of chimpanzee (*A*) and human (*B*) tonsils. Squamous epithelial cells overlying the follicles of the tonsils are stained. *Bottom panels* show sections of chimpanzee (*C*) and human (*D*) renal medulla, with staining of the lumenal edges of collecting ducts. Adjacent sections were analyzed with preimmune serum at the same dilution and showed no staining (data not shown). Magnification, ×200.

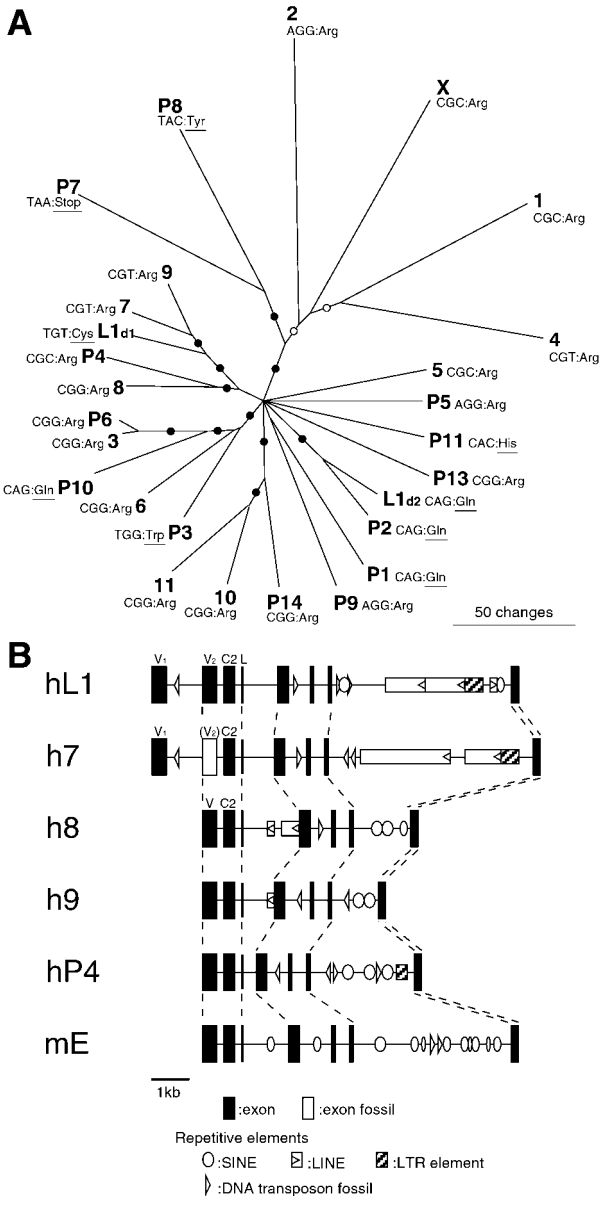

FIG. 5. **Phylogenetic relationships and genomic structures of human Siglecs.** *A,* phylogenetic relationships of V-set domains of human Siglec genes and pseudogenes. Siglecs 1–10 are in the literature (15), Siglec-11 represents our unpublished data, and pseudogenes were identified and numbered as described under "Experimental Procedures." The tree was constructed using the neighbor-joining method (31) with a distance matrix based on absolute distance. The internodes supported by bootstrap values >90 and 75–90 (for 1000 replications) are indicated with *filled* and *open circles,* respectively. Both maximum parsimony and maximum likelihood methods gave trees with very similar topology (data not shown). Note that Siglec-L1 is represented twice in this tree by virtue of having two different V-set domains (d1 and d2). P12 was omitted from this analysis because of an Alu element disrupting the sequence. The sequence alignments used are provided as supplementary material. *B,* genomic structures of Siglec-L1, Siglec-7, Siglec-8, Siglec-9, and mouse Siglec-E. *V,* Ig V-set domain; *C2,* Ig C2-set domain; *L,* linker region.

pseudogenes in the syntenic regions of mouse chromosome, and one of the latter has an "arginine-replacing" mutation. Taken together, these data suggest that the Siglec-3-related genes underwent extensive duplications during mammalian evolution, and several of the resulting pseudogenes may have been initially inactivated for sialic acid binding by mutations changing the "essential" arginine residue. A less likely explanation for the high frequency of "arginine-replacing" mutations is that inactivation of functional Siglec genes by some other mutations was followed by CpG dinucleotide mutations to TpG/CpA (Ref. 36 and the references therein) to which the arginine codons (CGN) are susceptible. However, the frequency of these "arginine-replacing" mutations (in ~50% of the pseudogenes) seems to be too high to be explained by the latter mechanism. Additional studies of the corresponding regions of various primate genomes (as well as complete sequence data from the syntenic regions of other genomes such as those of the mouse and rat) are needed to define the series of genetic events that lead to the current human condition.

Genomic Structural Analysis Suggests a Complex Ancestry of Siglec-L1—The genomic structure of human Siglec-L1 was deduced from the human genome sequence and compared with those of its close paralogs. As shown in Fig. 5*B,* there is a high degree of similarity with the genomic structure of the Siglec-7 gene, suggesting that Siglec-7 and Siglec-L1 are sibling molecules generated by a relatively recent gene duplication. Of particular interest, the second Ig-like domain of Siglec-L1 is very similar to an exon fossil in the Siglec-7 gene (P2 in Fig. 5*A*; the *open box* in front of the Siglec-7 C2-set domain in Fig. 5*B*). The other close human paralogs (Siglec-8, Siglec-9, and pseudogene P4) and the mouse gene orthologous to this group, called "Siglec-E" (37), do not share this exon configuration (Fig. 5*B*). The tight association of three adjacent exons (V or V_2-C2-linker) in all of these genes and the presence of a DNA transposon fossil (type MER20) between exons V_1 and V_2 of Siglec-L1 (and the corresponding regions of the Siglec-7 gene) suggest that the ancestral V_1 exon of Siglec-L1 and Siglec-7 may have been initially inserted in 5′ region of the ancestral gene, possibly in association with the DNA transposon insertion (38). This V-set exon could have then functionally replaced the original one (now V_2), allowing the arginine codon in V_2 to mutate to glutamine in Siglec-L1 (and allowing this exon to become inactive in Siglec-7).

Conclusions and Future Prospects—We have described the second human-specific and human-universal genetic mutation that affects the biology of sialic acids, the first of which leads to the loss of expression of the sialic acid Neu5Gc. Given the fact

that the chimpanzee Siglec-L1 and the "arginine restored" human Siglec-L1 show a preference for binding Neu5Gc over Neu5Ac, it seems particularly likely that the two events are evolutionarily linked. If so, which came first? One possibility is that the "arginine-replacing" mutation in human Siglec-L1 occurred first, thereby making conditions more permissive for the loss of Neu5Gc synthesis. The other possibility is that the loss of Neu5Gc resulted in the effective loss of function of this Siglec, thereby making conditions permissible for it to accumulate mutations. However, in the latter scenario, it is much more likely that a random gene inactivation event would have occurred, rather than this highly specific functional inactivation, mutating the arginine residue required for optimal sialic acid binding.

It is of course possible that human Siglec-L1 has other functions that are unrelated to the ability to recognize sialic acids and that such function(s) resulted in the continued maintenance of the open reading frame. The selective expression of these molecules in epithelial cells also raises some interesting issues. The presence of putative tyrosine-based signaling motifs in the cytosolic tail suggests that the great ape orthologs are (and that the human ancestral Siglec-L1 could have been) involved in the regulation of homeostasis of sialylated glycoconjugates within the lumen of some organs. Indeed, others have shown that the membrane-proximal tyrosine-based motif of this molecule can interact with the tyrosine phosphatases SHP-1 and SHP-2 (24). The expression of this molecule on surfaces that come into contact with the environment also suggests its possible involvement in microbial infection (*e.g.* viral particles carrying sialic acids might be accidentally captured). Whereas these possibilities are obviously speculative, they must be taken into consideration when trying to explain the original selection pressure that resulted in the arginine-eliminating mutation in humans.

Many other questions arise from this work. When exactly did this mutation occur during the last ~6 million years since our common ancestor split from the chimpanzee/bonobo clade? Can the unusual expression profile of this Siglec-like molecule on epithelial surfaces explain apparent differences in the incidence of diseases like epithelial cancers between humans and chimpanzees (39)? Why do some human alleles have an additional frameshift mutation? Why are there so many Siglec-like pseudogenes with mutated "essential" arginines? Could these be the result of repeated fine-tuning of endogenous sialic acid recognition? Do these also represent human-specific mutations related to the loss of Neu5Gc in humans, or did some occur earlier during mammalian evolution? We are currently pursuing some of these questions, particularly by comparison of the corresponding regions of the great ape and other primate genomes. Such studies may also determine whether this mutation can explain some of the obvious morphological and functional differences between humans and our closest evolutionary cousins.

Acknowledgments—We thank Pascal Gagneux and Els Brinkman-van der Linden for helpful discussions.

Note Added in Proof—Another group has also reported the cloning of two splice variants of the same human gene (Foussias, G., Taylor, S. M., Yousef, G. M., Tropak, M. B., Ordon, M. H., and Diamandis, E. P. (2001) *Biochem. Biophys. Res. Commun.* **284,** 887–899).

REFERENCES

1. Darwin, C. (1871) *The Descent of Man, and Selection in Relation to Sex*, D. Appleton and Co., New York
2. King, M. C., and Wilson, A. C. (1975) *Science* **188,** 107–116
3. Sibley, C. G., Comstock, J. A., and Ahlquist, J. E. (1990) *J. Mol. Evol.* **30,** 202–236
4. Ruvolo, M. (1997) *Mol. Biol. Evol.* **14,** 248–265
5. Goodman, M. (1999) *Am. J. Hum. Genet.* **64,** 31–39
6. Chou, H. H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., Wright, K. L., Muchmore, E. A., Nelson, D. L., Warren, S. T., and Varki, A. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 11751–11756
7. Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T., and Suzuki, A. (1998) *J. Biol. Chem.* **273,** 15866–15871
8. Muchmore, E. A., Diaz, S., and Varki, A. (1998) *Am. J. Phys. Anthropol.* **107,** 187–198
9. Varki, A. (1992) *Glycobiology* **2,** 25–40
10. Schauer, R. (1982) *Sialic Acids: Chemistry, Metabolism and Function, Cell Biology Monographs,* Vol. 10, Springer-Verlag, New York
11. Kelm, S., and Schauer, R. (1997) *Int. Rev. Cytol.* **175,** 137–240
12. Varki, A. (1997) *FASEB J.* **11,** 248–255
13. Karlsson, K. A. (1998) *Mol. Microbiol.* **29,** 1–11
14. Crocker, P. R., Clark, E. A., Filbin, M., Gordon, S., Jones, Y., Kehrl, J. H., Kelm, S., Le Douarin, N., Powell, L., Roder, J., Schnaar, R. L., Sgroi, D. C., Stamenkovic, K., Schauer, R., Schachner, M., Van den Berg, T. K., Van der Merwe, P. A., Watt, S. M., and Varki, A. (1998) *Glycobiology* **8,** V
15. Crocker, P. R., and Varki, A. (2001) *Trends Immunol.* **22,** 337–342
16. Law, C. L., Aruffo, A., Chandran, K. A., Doty, R. T., and Clark, E. A. (1995) *J. Immunol.* **155,** 3368–3376
17. Nath, D., Van der Merwe, P. A., Kelm, S., Bradfield, P., and Crocker, P. R. (1995) *J. Biol. Chem.* **270,** 26184–26191
18. Van der Merwe, P. A., Crocker, P. R., Vinson, M., Barclay, A. N., Schauer, R., and Kelm, S. (1996) *J. Biol. Chem.* **271,** 9273–9280
19. Vinson, M., Van der Merwe, P. A., Kelm, S., May, A., Jones, E. Y., and Crocker, P. R. (1996) *J. Biol. Chem.* **271,** 9267–9272
20. May, A. P., Robinson, R. C., Vinson, M., Crocker, P. R., and Jones, E. Y. (1998) *Mol. Cell* **1,** 719–728
21. Tang, S., Shen, Y. J., DeBellard, M. E., Mukhopadhyay, G., Salzer, J. L., Crocker, P. R., and Filbin, M. T. (1997) *J. Cell Biol.* **138,** 1355–1366
22. Angata, T., and Varki, A. (2000) *Glycobiology* **10,** 431–438
23. Angata, T., and Varki, A. (2000) *J. Biol. Chem.* **275,** 22127–22135
24. Yu, Z., Lai, C.-M., Maoui, M., Banville, D., and Shen, S.-H. (2001) *J. Biol. Chem.* **276,** 23816–23842
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
26. Patel, N., Brinkman-Van der Linden, E. C. M., Altmann, S. W., Gish, K., Balasubramanian, S., Timans, J. C., Peterson, D., Bell, M. P., Bazan, J. F., Varki, A., and Kastelein, R. A. (1999) *J. Biol. Chem.* **274,** 22729–22738
27. Tullius, M. V., Munson, R. S., Wang, J., and Gibson, B. W. (1996) *J. Biol. Chem.* **271,** 15373–15380
28. Weinstein, J., de Souza, E., Silva, U., and Paulson, J. C. (1982) *J. Biol. Chem.* **257,** 13835–13844
29. Hara, S., Yamaguchi, M., Takemori, Y., Furuhata, K., Ogura, H., and Nakamura, M. (1989) *Anal. Biochem.* **179,** 162–166
30. Venter, J. C., *et al.* (2001) *Science* **291,** 1304–1351
31. Saitou, N., and Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425
32. Crocker, P. R., Mucklow, S., Bouckson, V., McWilliam, A., Willis, A. C., Gordon, S., Milon, G., Kelm, S., and Bradfield, P. (1994) *EMBO J.* **13,** 4490–4503
33. Cann, R. L. (2001) *Science* **291,** 1742–1748
34. Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P. A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O. A., and Woodruff, D. S. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 5077–5082
35. Schachner, M., and Bartsch, U. (2000) *Glia* **29,** 154–165
36. Lander, E. S., *et al.* (2001) *Nature* **409,** 860–921
37. Yu, Z. B., Maoui, M., Wu, L. T., Banville, D., and Shen, S. H. (2001) *Biochem. J.* **353,** 483–492
38. Smit, A. F. (1996) *Curr. Opin. Genet. Dev.* **6,** 743–748
39. Varki, A. (2000) *Genome Res.* **10,** 1065–1070