THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 281, NO. 35, pp. 25689–25702, September 1, 2006

System-wide Genomic and Biochemical Comparisons of Sialic Acid Biology Among Primates and Rodents

EVIDENCE FOR TWO MODES OF RAPID EVOLUTION[*S]

Received for publication, May 3, 2006, and in revised form, June 12, 2006 Published, JBC Papers in Press, June 12, 2006, DOI 10.1074/jbc.M604221200

Tasha K. Altheide[‡§1], Toshiyuki Hayakawa[‡§1,2], Tarjei S. Mikkelsen[¶], Sandra Diaz[‡§], Nissi Varki[‖], and Ajit Varki[‡§3]**

*From the [‡]Glycobiology Research and Training Center and the Departments of [§]Cellular and Molecular Medicine, **Medicine, and [‖]Pathology, University of California San Diego, La Jolla, California 92093-0687 and the [¶]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02141*

Numerous vertebrate genes are involved in the biology of the oligosaccharide chains attached to glycoconjugates. These genes fall into diverse groups within the conventional Gene Ontology classification. However, they should be evaluated together from functional and evolutionary perspectives in a "biochemical systems" approach, considering each monosaccharide unit's biosynthesis, activation, transport, modification, transfer, recycling, degradation, and recognition. Sialic acid (Sia) residues are monosaccharides at the outer end of glycans on the cell-surface and secreted molecules of vertebrates, mediating recognition by intrinsic or extrinsic (pathogen) receptors. The availability of multiple genome sequences allows a system-wide comparison among primates and rodents of all genes directly involved in Sia biology. Taking this approach, we present further evidence for accelerated evolution in Sia-binding domains of CD33-related Sia-recognizing Ig-like lectins. Other gene classes are more conserved, including those encoding the sialyltransferases that attach Sia residues to glycans. Despite this conservation, tissue sialylation patterns are shown to differ widely among these species, presumably because of rapid evolution of sialyltransferase expression patterns. Analyses of *N*- and *O*-glycans of erythrocyte and plasma glycopeptides from these and other mammalian taxa confirmed this phenomenon. Sia modifications on these glycopeptides also appear to be undergoing rapid evolution. This rapid evolution of the sialome presumably results from the ongoing need of organisms to evade microbial pathogens that use Sia residues as receptors. The rapid evolution of Sia-binding domains of the inhibitory CD33-related Sia-recognizing Ig-like lectins is likely to be a secondary consequence, as these inhibitory receptors presumably need to keep up with recognition of the rapidly evolving "self"-sialome.

Sialic acid (Sia)[4] residues are negatively charged nine-carbon sugars typically occupying the distal ends of glycan chains on the cell-surface and secreted molecules in the deuterostome lineage of animals (1, 2). The two most common forms of Sia in vertebrates are *N*-acetylneuraminic acid (Neu5Ac) and *N*-glycolylneuraminic acid (Neu5Gc), which differ by a single oxygen atom that is added by the CMP-Neu5Ac hydroxylase enzyme CMAH. A third basic type of Sia is 2-keto-3-deoxynonulosonic acid, which is less common in mammals. Rarely, the amino group of neuraminic acid can remain unmodified. These four forms of Sia are subject to a variety of modifications (most prominently *O*-acetylation at the 4-, 7-, 8-, or 9-position) and can be presented in many different linkages to the underlying sugar chain (1–3). The sum total of this diversity has been termed the "sialome" (4).[5]

Sia residues are involved in many biological processes, often involving binding by intrinsic and extrinsic Sia-recognizing proteins. As an example of intrinsic Sia recognition, complement factor H uses Sia as a means to identify "self" and to prevent autoimmune attack by the alternate complement pathway; in contrast, foreign cells lacking Sia are not protected (5). Current evidence suggests that the CD33-related subset of Sia-recognizing Ig-like lectins (Siglecs) may also serve to recognize host Sia residues as self (4), thus dampening autoreactivity of cells of the innate immune response. Meanwhile, numerous vertebrate pathogens recognize and bind to glycan structures containing Sia residues (2), using them as portals to gain entry. Elimination of vertebrate host Sia production to avoid such pathogens is not an option, as this results in embryonic lethality (6). Complicating matters, many successful microbes express surface Sia residues, mimicking the host and avoiding recogni-

* This work was supported by National Institutes of Health Grants R01GM32373 and P01HL57345 (to A. V.), a Japan Society for the Promotion of Science postdoctoral fellowship for research abroad (to T. H.), and an American Cancer Society postdoctoral fellowship (to T. K. A.). 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 supplemental text, references, and supplemental Tables 1–3.

[1] Both authors contributed equally to this work.

[2] Present address: Research Inst. for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan.

[3] To whom correspondence should be addressed: Dept. of Cellular and Molecular Medicine, Mail Code 0687, 9500 Gilman Dr., University of California, San Diego, La Jolla, CA 92093-0687. Tel.: 858-534-2214; Fax: 858-534-5611; E-mail: a1varki@ucsd.edu.

[4] The abbreviations used are: Sia, sialic acid; Neu5Ac, *N*-acetylneuraminic acid; Neu5Gc, *N*-glycolylneuraminic acid; Siglecs, Sia-recognizing Ig-like lectins; CD33rSiglecs, CD33-related Sia-recognizing Ig-like lectins; SNA, *S. nigra* agglutinin; MAH, *M. amurensis* hemagglutinin; HPAEC-PAD, high performance anion exchange chromatography with pulsed amperometric detection.

[5] The term sialome denotes the total complement of Sia types and linkages and their modes of presentation on a particular organelle, cell, tissue, organ, or organism as found at a particular time and under specific conditions (4).

tion by many arms of the immune system (7). Taken together, these data suggest that Sia residues are involved in an ongoing biochemical "arms race" between hosts and pathogens, driven to diversify by "Red Queen" effects,[6] even while conserving critical endogenous functions (4, 8).

Taking a "biochemical systems" approach to analyzing all extant data (see Fig. 1), we found that there are <60 known genetic loci directly involved in the biosynthesis, activation, transport, modification, transfer, recycling, degradation, and recognition of Sia residues within humans and other vertebrates (see Fig. 1 and Table 1). With the exception of the CD33-related Siglecs (CD33rSiglecs), all these loci are conserved between the human and mouse genomes, indicating their functional importance. Precursor molecules are first converted into Sia residues, which are then activated to CMP-Sia and transported to the Golgi apparatus, where members of a family of sialyltransferases transfer them onto the terminal ends of glycan chains in various types of structurally distinct linkages (see Fig. 1). Sia residues may also be modified from one form to another such as from Neu5Ac to Neu5Gc or by the addition of *O*-acetyl groups, and these alterations are differentially recognized by receptors on the same cell surface or on other cells. Sia residues attached to macromolecules are eventually cleaved from glycan chains in the lysosome, actively returned to the cytosol, and then recycled or degraded (see Fig. 1).

Humans and chimpanzees share >99% identity in typical protein sequences (9–12). Of the few published genetic differences between humans and chimpanzees and other "great apes"[7] with known/potential functional consequences, several involve genes related to Sia biology: a human-specific exon deletion in *CMAH* resulting in the inability to convert Neu5Ac to Neu5Gc (13–16); a human-specific point mutation in *SIGLEC12* (previously called *Siglec-L1*) eliminating its Sia recognition property (17); a human-specific up-regulation of α2–6-linked Sia expression on selected cell types, presumably because of changed expression of the sialyltransferase *ST6GAL1* (18); human-specific changes in one *SIGLEC9* exon associated with the accommodation of Neu5Ac recognition by SIGLEC9 (19); human-specific loss of an entire primate-specific Siglec gene (*SIGLEC13*) (20); a human-specific gene conversion of *SIGLEC11* causing changes in binding properties and newly induced expression in the brain (21); and selective down-regulation of CD33rSiglecs in human T cells (22). Additional studies suggest other species-specific gene conversion events among some hominid Siglecs (23) and other examples of human-specific changes in Siglec gene expression (22). The finding of so many human-specific functional differences from chimpanzees and other great apes within one biochemical/biological system suggests that it was subjected to major selective pressure(s) at some point(s) in human evolution.

Although all these genes are part of a well defined system (Sia metabolism and function), they are not represented as a single biological process in widely used genomic classification systems such as the Gene Ontology system (24) or PANTHER (25), which is also true of most other genes involved in glycan biology. These functionally related genes actually fall into diverse groups within the conventional Gene Ontology classification. They are also (with the exception of the CD33rSiglecs) randomly distributed throughout the genome. We suggest that all these genes should be evaluated together in a biochemical systems approach, considering the biosynthesis, activation, transport, modification, transfer, recycling, degradation, and recognition of Sia residues. Here, we undertake such an approach toward understanding the evolution of Sia biology in primates, rodents, and other mammals in combination with selected biochemical studies. We first investigate whether specific loci or functional classes of loci in this system have been subjected to adaptive selective pressures, whether any common principles emerge, and whether differences between chimpanzees and humans are more significant despite a shorter divergence time. We then take a biochemical approach to put the genomic data into context.

EXPERIMENTAL PROCEDURES

Human Loci—We identified 55 loci in humans that are known to be (or to potentially be) involved in Sia biology (see Table 1). Human RefSeqs and the genome sequences of target loci were obtained from the NCBI LocusLink web site (available at www.ncbi.nlm.nih.gov/LocusLink/). Some loci have several RefSeqs representing splice variants and/or show some sequence differences between the RefSeqs and the genome sequences. In the former case, the sequence with the most inclusive number of exons was used as a representative of the locus. In the latter case, the actual genome sequences were used for the analyses.

Identification of Chimpanzee Orthologs—Human RefSeqs or human genome sequences from 43 human loci (excluding the CD33rSiglecs) were used to extract orthologous coding sequences from the chimpanzee genome assembly (NCBI Build 1 Version 1) as identified by reciprocal best BLASTZ alignments (12). Phred quality scores for each site in chimpanzee sequences were also provided by the Chimpanzee Sequencing and Analysis Consortium (12). For eight of the human CD33rSiglecs (*SIGLEC3*, *SIGLEC5–10*, and *SIGLEC12*), high quality sequences were also obtained from our independent high resolution comparative analyses of human, chimpanzee, baboon, mouse, and rat (20). One more Siglec locus (*SIGLEC13*) is found in the chimpanzee and baboon genomes, but its complete deletion in humans was reported previously (20). *SIGLEC13* was therefore used only for the domain-specific comparative analyses.

Mouse and Rat Orthologs—RefSeqs of mouse and rat orthologs were obtained from the NCBI LocusLink web site. Reliable mouse sequences were obtained for all but four loci; reliable rat sequences were obtained for all but nine sequences. All rodent loci obtained have one sequence as the RefSeq, with

[6] The Red Queen effect in evolution refers to the observation to Alice by the Red Queen that "it takes all the running you can do, to keep in the same place." Complex multicellular animals with long life cycles must evolve rapidly to survive the attacks of microbial pathogens that can replicate much faster (57, 58).

[7] The term great apes (including chimpanzees, bonobos, gorillas, and orangutans) is used here in the colloquial sense, as phylogenetic analysis of genomic information no longer supports this species grouping (59). Under the currently common classification, these species are now grouped together with humans in the family Hominidae.

the exception of the mouse *St3gal2* locus. The representative sequence of the mouse *St3gal2* locus was selected by following the procedure described for human loci. The high quality sequences of mouse and rat CD33rSiglec orthologs (*Siglec3* and *SiglecE–G*) were obtained as described (20).

Evolutionary Analysis—Sequence alignments of coding regions were performed in ClustalW (26) and manually checked to see whether chimpanzee sequences had insertions or deletions causing frameshifts in the aligned open reading frame. These were handled with reference to human and mouse sequences, which showed identical open reading frames for all loci studied, except for one locus (*NAGK*; see supplemental text). Chimpanzee-specific insertions were assumed to be errors and were deleted to maintain an open reading frame even if they had high quality scores. Frameshifts caused by deletions were left in the alignments as gaps, but the codons they were located in were removed from the analyses (see supplemental text). The sequences modified by these processes are referred to as "modified" sequences in supplemental Table 1. In the alignments, some sites that are substitutions or indels between the human and chimpanzee sequences show low quality Phred scores in chimpanzee. Because these low quality sites could be artifacts from the sequencing and base-calling process, a second round of analyses were done in which such low quality chimpanzee sites were changed to match the human sequences at the sites in question. The chimpanzee sequences in which substitutions were modified to match the human sequences are referred to as "humanized" sequences in supplemental Table 1. Several chimpanzee sequences also show regions of non-called bases (represented by "N" in supplemental Table 1). Gene sequence regions that had non-called bases in the chimpanzee sequence were excluded from analyses.

The evolutionary parameters shown in Table 1 were calculated in multiple species comparisons using human, chimpanzee, mouse, and rat. The numbers of synonymous (*Ks*) and nonsynonymous (*Ka*) substitutions per site were estimated by the method of Nei and Gojobori (27) with the Jukes-Cantor correction (28). Values for *Ka* and *Ks* were calculated using DnaSP Version 3.51 (29) or MEGA2 (30). Statistical tests were performed to assess the significance of evolutionary differences obtained in the analyses by using InStat Version 0.6 (GraphPad Software) or MEGA2.

Protein Secondary Structure Prediction—For secondary structure prediction of the sialyltransferase loci, new joint method analysis was performed using web-based software at the Parallel Protein Information Analysis System (PAPIA) web site (available at www.cbrc.jp/papia-cgi/ssp_menu.pl) (61).

Lectin Staining of Sialic Acids on Tissue Sections—Paraffin sections of lung, kidney, and spleen samples from seven humans, eight chimpanzees, six rats, and six mice were deparaffinized, blocked, and overlaid with predetermined concentrations of biotinylated *Sambucus nigra* agglutinin (SNA) lectin or biotinylated *Maackia amurensis* hemagglutinin (MAH) lectin or with control reagent. Binding was detected by alkaline phosphatase-labeled streptavidin using Vector blue substrate, nuclear fast red counterstaining, and aqueous mounting. Samples were washed with Tris-buffered saline containing 0.2% Tween and 1% bovine serum albumin to block nonspecific binding. Digital photomicrographs were taken while viewing with an Olympus BH2 microscope with a MacroFire camera and Adobe Photoshop.

Preparation of Erythrocyte Ghosts and Plasma—Blood from multiple taxa was collected directly into BD Vacutainer tubes containing EDTA, stored overnight at 4 °C, and then spun at 2000 × *g* for 10 min at 4 °C. The plasma was removed and stored frozen until further work-up. The buffy coat was removed, and the erythrocyte pellet was washed twice with 10 volumes of ice-cold phosphate-buffered saline (pH 7.4). Lysis of erythrocytes was accomplished by adding 15 volumes of ice-cold 10 mM Tris-HCl (pH 7.5) and 1 mM EDTA. The sample was transferred into glass Sorvall tubes and centrifuged at 10,000 × *g* for 20 min. The supernatant was carefully aspirated so as not to disturb the remaining "soft" pellet. The creamy particulate material that did not diffuse easily (representing contaminating white cells) was also removed. The tubes were filled with lysis buffer and centrifuged again. The process was repeated until ghosts were white. The last wash was made with ice-cold water containing 0.01% butylated hydroxytoluene as a preservative.

Glycopeptide Preparation from Erythrocyte Ghosts and Plasma—Plasma (0.5 ml) was lyophilized in a glass conical tube, and 250 μl of water was added. 0.5 ml of the ghosts was transferred into a glass conical tube, assuming ~50% water. The lipids were extracted from each of the above samples with 20 volumes of 2:1 (v/v) chloroform/methanol using a Brinkmann Instruments Polytron at a high setting for 30–60 s. The samples were centrifuged at 800 × *g* for 5 min after each extraction. All supernatants containing the lipids were pooled into a single glass vessel. Each sample was extracted again with 2:1 (v/v) chloroform/methanol. The pellets were extracted twice with 1:1 (v/v) chloroform/methanol and twice with 1:2 (v/v) chloroform/methanol. The remaining glycoprotein pellet was extracted with 95% ethanol, and the supernatant was also added to the pool. The glycoprotein pellet was immediately dissolved in 100 mM Tris-HCl (pH 6.5). Low molecular weight molecules were removed from the samples by performing dialysis using M_r 3500 cutoff tubing against a 500-fold volume of 100 mM Tris-HCl (pH 6.5) and 2 mM EDTA overnight at 4 °C. The retentate was recovered and digested with 0.1 volume of 20 mg/ml proteinase K made in 50 mM Tris-HCl (pH 8.0) and 2 mM calcium acetate, followed by incubation at 50 °C for 8 h. At the end of the day, another aliquot of the 10× proteinase K solution was added to the sample, and the digestion mixture was allowed to incubate overnight. The enzyme was inactivated by boiling for 10 min; the sample was centrifuged to remove particulates; and the resulting supernatant was loaded onto a 1-ml column of DEAE-Sephacel (GE Healthcare) equilibrated in 20 mM Tris-HCl (pH 6.5) and 0.1 M NaCl (62). The column run-through fraction was collected and reloaded onto the column. The column was washed with 30 ml of 20 mM Tris-HCl (pH 6.5) and 0.1 M NaCl. The column run-through fractions containing glycopeptides were pooled with the wash, and dialysis was performed against a 100-fold volume of water at 4 °C using M_r 1000 cutoff tubing for 12–16 h. The dialysis solution was changed to 2 mM EDTA for 8–12 h and changed back to water overnight. The sample from the dialysis tubing was recovered, frozen, and

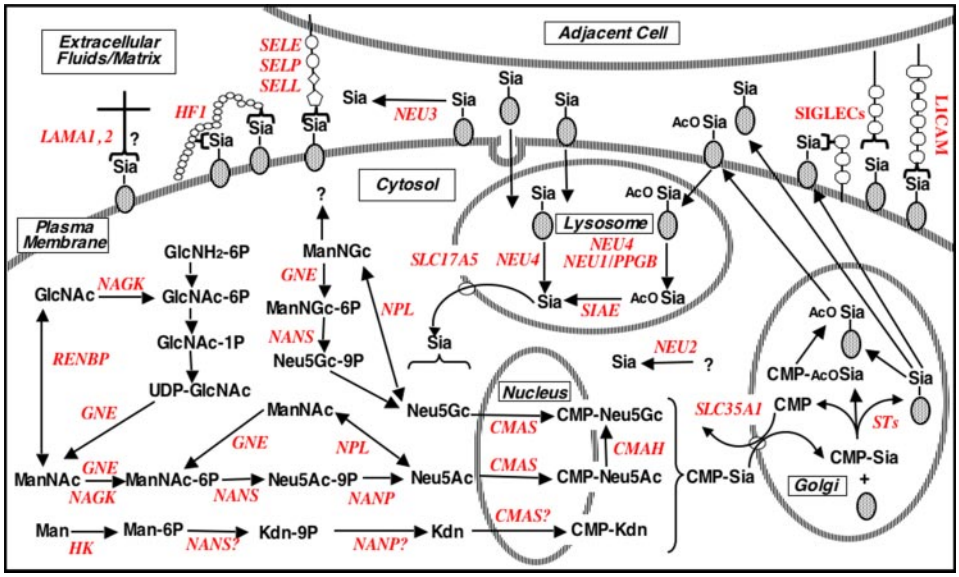


FIGURE 1. **Genes involved in sialic acid biochemistry and biology.** Shown are all genes or groups of genes (in *red italics*) thought to be directly involved in Sia biochemistry and biology, *i.e.* biosynthesis, activation, transport, modification, transfer, recycling, degradation, and recognition (see Table 1 for a full listing). The *question marks* indicate unknown or hypothetical pathways. *Kdn*, 2-keto-3-deoxynonulosonic acid; *STs*, sialyltransferases.

lyophilized. The resulting powder was dissolved in 1 ml of water, transferred to a smaller container, and frozen and lyophilized again. The resulting glycopeptides were recovered and weighed.

Release of N- and O-Glycans from Glycopeptides by Automated Hydrazinolysis—The glycopeptides were dissolved in 500 μl of water, and a 2-mg equivalent was transferred to a GlycoPrep reactor vial, frozen, and lyophilized. Hydrazinolysis was performed in the N + O mode using an automated hydrazinolysis instrument (GlycoPrep 1000, Oxford GlycoSciences, Abingdon, UK), which was set to heat at 95 °C for 4 h, followed by automated purification (16–24 h). The released glycans were filtered through 0.5-μm polytetrafluoroethylene filters to remove silica gel particles and lyophilized.

Analysis of N- and O-Glycans by High Performance Anion Exchange Chromatography with Pulsed Amperometric Detection (HPAEC-PAD)—Free oligosaccharides were analyzed by HPAEC-PAD (31) on a CarboPac PA1 column (4 × 250 mm) in-line on a DX500 HPLC system equipped with a pulsed amperometric detector and a Thermo Separations AS3500 autosampler. The various oligosaccharides were eluted with a linear gradient of sodium acetate from 20 to 250 mM over 60 min in 100 mM sodium hydroxide. Data acquisition and processing were performed with Dionex PeakNet software. Elution profiles of the glycans were compared with those of standard *N*- and *O*-glycans of known elution behavior.

Determination of the Sialic Acid Types in Erythrocyte Ghosts and Plasma Samples—Sialic acids were released from the erythrocyte ghost or plasma glycopeptides by hydrolysis in 2 M acetic acid at 80 °C for 3 h. The released Sia residues were separated from high molecular weight proteins by passage through an Amicon Microcon-10 filter. The flow-through fraction was derivatized with an equal volume of 2× 1,2-diamino-4,5-methylenedioxybenzene reagent (32) and heated at 50 °C for 2.5 h. The fluorescently tagged sialic acids were separated on a Varian Microsorb-MV 100-5 C18 column (4 × 250 mm) in the isocratic mode using 85% water, 8% acetonitrile, and 7% methanol and detected with a SpectroVision FD-300 fluorescence detector with emission set at 373 nm and excitation at 448 nm. Elution profiles were compared with those of standard sialic acids of known elution behavior.

RESULTS

Identification of Loci and Analysis of Functional Categories—Genes involved in Sia biology encode proteins with widely differing functions, ranging from cell-surface receptors that recognize Sia residues, to enzymes cleaving Sia residues in the lysosome, to transporters making them available for reuse by the cell (Fig. 1 and Table 1). Chimpanzee orthologs were identified for all 55 human loci known to be involved in Sia biology (see supplemental "Experimental Procedures"), indicating that there have been no major chimpanzee-specific deletions in this system. Although identifiable, not all loci were analyzable (see Table 1 and "Experimental Procedures"). Additionally, the presence of all corresponding loci in the mouse genome (with the exception of some primate-specific CD33rSiglecs; see supplemental "Experimental Procedures") suggests that these loci are generally conserved in mammals. The loci fall into different functional biochemical categories, which we have termed as "biosynthesis"; "activation, transport, and transfer"; "modification"; "recognition"; and "recycling and degradation." Biosynthesis refers to loci involved in the production of Sia residues from precursor molecules such as UDP-GlcNAc and ManNAc and includes epimerases, kinases, and phosphatases. Activation, transport, and transfer refer to loci that activate free Sia into the nucleotide donor CMP-Sia and transport it into the lumen of the Golgi, where multiple sialyltransferases then transfer the Sia residues from the CMP donors to newly synthesized glycoconjugates. Modification of CMP-Neu5Ac to CMP-Neu5Gc occurs by the action of the *CMAH* locus, which is a pseudogene in humans but is functional in chimpanzees (13). Additional modification genes presumed to be involved in other modifications of Sia residues such as *O*-acetylation, *O*-methylation, *O*-sulfation, etc., have yet to be identified. Recycling and degradation genes encode sialidases, which release Sia residues from glycan chains; a stabilizer protein; a lysosomal sialic acid *O*-acetylesterase; a lysosomal Sia exporter; and sialate pyruvate-lyase, which cleaves free Sia residues in the cytosol into pyruvate and acylmannosamines. Recognition molecules do not directly participate in the Sia biochemical life cycle, but act as receptors for Sia residues. The major category of these molecules is the Siglecs, a family of cell-surface receptors that recognize and bind to different linkages and structural variants of Neu5Ac and

TABLE 1

Genes involved in sialic acid biology

All human loci known or thought to be involved in aspects of Sia biology are listed. The human sequences used to identify homologous sequences in the chimpanzee genome are listed by accession number. The primary function of each gene as pertaining to Sia biology is indicated.

Human gene	NCBI accession no.	Molecular function(s)
Biosynthesis		
GNE	NM_005476	Regulates and initiates biosynthesis of Neu5Ac
NANP	NM_152667	Dephosphorylates Neu5Ac-9-P to Neu5Ac
NAGK	NM_017567	Converts GlcNAc into GlcNAc-6-P
NANS	NM_018946	Converts ManNAc to Neu5Ac-9-P
RENBP	NM_002910	Catalyzes the interconversion of GlcNAc to ManNAc
Activation, transport, and transfer		
CMAS	NM_018686	Catalyzes the activation of Sia to CMP-Sia
SLC35A1	NM_006416	Transports CMP-Sia from cytosol to Golgi
ST3GAL1	NM_003033	Attaches α2–3-linked Sia residues to glycolipids and *O*-glycans
ST3GAL2	NM_006927	Attaches α2–3-linked Sia residues to glycolipids and glycoproteins
ST3GAL3	NM_174963	Attaches α2–3-linked Sia residues to *N*-glycans
ST3GAL4	NM_006278	Attaches α2–3-linked Sia residues to *N*- and *O*-glycans
ST3GAL5	NM_003896	Attaches α2–3-linked Sia to lactosylceramide
ST3GAL6	NM_006100	Attaches α2–3-linked Sia residues to glycolipids and glycoproteins
ST6GAL1	NM_173216	Attaches α2–6-linked Sia residues to *N*-glycans
ST6GAL2	NM_032528	Attaches α2–6-linked Sia residues to free glycans and glycoproteins
ST6GALNAC1	NM_018414	Attaches α2–6-linked Sia residues to *O*-glycans
ST6GALNAC2[a]	NM_006456	Attaches α2–6-linked Sia residues to *O*-glycans
ST6GALNAC3	NM_152996	Attaches α2–6-linked Sia residues to glycolipids and *O*-glycans
ST6GALNAC4	NM_014403	Attaches α2–6-linked Sia residues to *O*-glycans
ST6GALNAC5	NM_030965	Attaches α2–6-linked Sia residues to glycolipids
ST6GALNAC6	NM_013443	Attaches α2–6-linked Sia residues to glycolipids
ST8SIA1	NM_003034	Attaches α2–8-linked Sia residues to glycolipids
ST8SIA2	NM_006011	Attaches α2–8-linked Sia residues to *N*-glycans
ST8SIA3	NM_015879	Attaches α2–8-linked Sia residues to glycoproteins
ST8SIA4	NM_005668	Attaches α2–8-linked Sia residues to *N*-glycans
ST8SIA5	NM_013305	Attaches α2–8-linked Sia residues to glycolipids
ST8SIA6[b]	XM_291725	Attaches α2–8-linked Sia residues to *O*-glycans
Modification		
CMAH	NM_003570	Converts CMP-Neu5Ac to CMP-Neu5Gc
Recognition		
SIGLEC1	NM_023068	Sia recognition and potential phagocytosis
CD22	NM_001771	Sia recognition and signaling
CD33	NM_001772	Sia recognition and potential signaling
MAG[a]	NM_080600	Sia recognition and signaling
SIGLEC5[c]	NM_003830	Sia recognition and potential signaling
SIGLEC6[d]	NM_198845	Sia recognition and potential signaling
SIGLEC7[d]	NM_016543	Sia recognition and potential signaling
SIGLEC8[d]	NM_014442	Sia recognition and potential signaling
SIGLEC9[d]	NM_014441	Sia recognition and potential signaling
SIGLEC10[d]	NM_033130	Sia recognition and potential signaling
SIGLEC11[c]	NM_052884	Sia recognition and potential signaling
SIGLEC12[d]	NM_053003	Sia recognition and potential signaling
HF1	NM_000186	Sia recognition and regulation of alternate complement pathway
SELE	NM_000450	Sia recognition mediates endothelial binding to leukocytes
SELL	NM_000655	Sia recognition mediates leukocyte binding to endothelium
SELP	NM_003005	Sia recognition mediates binding of platelets and endothelium to leukocytes
L1CAM	NM_000425	Cell-cell interactions in brain
LAMA1-G	NM_005559	Sia recognition necessary for α-dystroglycan binding?
LAMA2-G	NM_000426	Sia recognition necessary for α-dystroglycan binding?
Recycling and degradation		
NEU1	NM_000434	Intralysosomal catabolism of sialylated glycoconjugates
NEU2	NM_005383	Unknown function in cytosol
NEU3	NM_006656	Releases Sia residues from gangliosides in lipid bilayer
NEU4[a]	NM_080741	Membrane-associated sialidase activity
NPL	NM_030769	Breakdown of Sia into acylmannosamines and pyruvate
PPGB	NM_000308	Essential for stability/activity of lysosomal sialidase
SIAE	NM_170601	Removes 9-*O*-acetyl esters from Sia residues
SLC17A5	NM_012434	Lysosomal transporter delivering free Sia residues to cytosol

[a] Locus not analyzed because of poor chimpanzee sequence.
[b] Locus not analyzed because of multiple human sequences in data bases.
[c] Locus not analyzed because of gene conversion events.
[d] Actual sequence from Angata *et al.* (Ref. 20 and chr19:55978691–56924690; International Human Genome Sequencing Consortium 2001, Ref. 60).

Neu5Gc both in *cis* and in *trans* on cell surfaces (4, 33). Other known Sia-recognizing intrinsic receptors include E-, P-, and L-selectins (34–36); factor H (5); and *L1CAM* (37). We also included the G domains of two laminin loci (*LAMA1* and *LAMA2*) in this classification because these domains are thought to recognize Sia residues (38), although conclusive proof is lacking.

Some of these functional labels correspond generally to those listed for loci in the Gene Ontology (24) or PANTHER (25) data bases, but most are more specific in the context of Sia biology. A few loci appear to have additional functions or capabilities external to the Sia biology pathway, *e.g.* the *RENBP* gene product is also a renin-binding protein, and the *PPGB* gene product

TABLE 2

Summary statistics comparing human and chimpanzee gene categories involved in sialic acid biology

The average nucleotide divergence, average amino acid divergence, average *Ka* values, average *Ks* values, and average *Ka*/*Ks* values are shown for human-chimpanzee comparisons. The single locus in the modification category (*CMAH*) is included in the overall estimation.

Biological function	No. loci analyzed	Average nucleotide divergence	Average amino acid divergence	Average *Ka*	Average *Ks*	Average *Ka*/average *Ks*	Average *Ka*/*Ks*
		%	%				
Overall	49	0.84	1.18	0.006	0.017	0.319	0.326
Activation, transport, and transfer	20	0.72	0.82	0.004	0.018	0.213	0.205
Biosynthesis	5	0.58	0.78	0.004	0.012	0.293	0.344
Recognition	16	1.09	1.85	0.009	0.019	0.464	0.484
Recycling and degradation	7	0.81	1.02	0.005	0.017	0.292	0.301

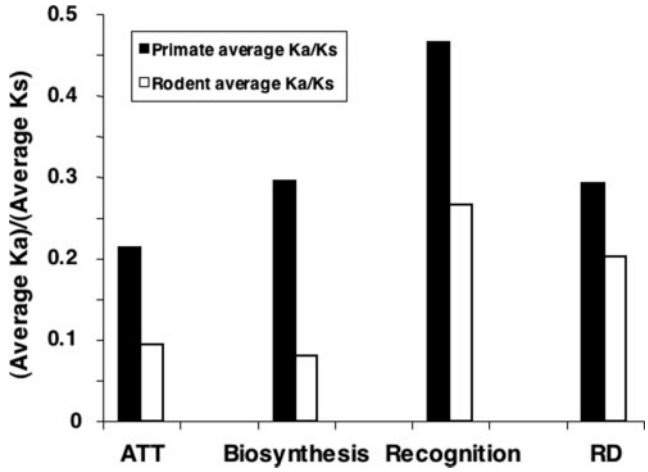


FIGURE 2. **Comparison of average *Ka*/*Ks* ratios between Sia biology categories in primates and rodents.** Average *Ka*/average *Ks* values for the activation, transport, and transfer (*ATT*); biosynthesis; recognition; and recycling and degradation (*RD*) categories are shown. Different numbers of loci were used between primates and rodents for each category. For details, see supplemental Tables 1 and 2.

TABLE 3

Comparison of loci within the recognition category shows rapid evolution in the Siglecs

The average *Ka*/average *Ks* ratios are shown for the Siglec and non-Siglec loci in the recognition category in human-chimpanzee and mouse-rat comparisons. The number of Siglec loci differs between primates and rodents because of lineage-specific duplication and loss events.

	No. loci used	Average *Ka*	Average *Ks*	Average *Ka*/average *Ks*
Siglec loci				
Primates	9	0.012	0.019	0.614
Rodents	5	0.119	0.194	0.612
Non-Siglec loci				
Primates	7	0.005	0.018	0.263
Rodents	7	0.067	0.252	0.265

is a cathepsin protease that also serves to stabilize lysosomal β-galactosidase. By taking a systematic "sialic acid biochemistry-based" approach to grouping these loci rather than a strict categorical label via the current Gene Ontology scheme,[8] we hoped to uncover information that is specifically relevant to the evolution and diversity of Sia biology in humans and other mammals.

Differences in Evolution Rates among Functional Gene Categories Indicate Rapid Evolution in Genes That Recognize Sialic Acids—We found overall differences in the evolutionary rates between functional categories of human and chimpanzee loci involved in Sia biology (Table 2). The amino acid divergence between human and chimpanzee ranged from 0% at several loci to 4.40% (*CD33*) (supplemental Table 1). The recognition category ($n = 16$) had the highest average amino acid divergence across categories (1.84%), followed by the recycling and degradation category ($n = 7$) with 1.02% divergence. The activation/transport/transfer ($n = 20$) and biosynthesis ($n = 5$) categories had lower levels of amino acid divergence (0.82 and 0.78%, respectively).

As a measure of evolutionary rates, we used the *Ka*/*Ks* ratio, a commonly used statistic that provides an indication of selection for amino acid changes during evolution This ratio is based on the rate of the nonsynonymous (amino acid changing) nucleotide substitutions compared with the rate of synonymous (non-amino acid changing) nucleotide substitutions between two taxa. Both numbers need to be normalized to the number of possible events that could have occurred. Thus, the ratio is calculated as the number of nonsynonymous substitutions/total number of nonsynonymous sites in the sequence of interest divided by the number of synonymous substitutions/total number of synonymous sites in the same sequence. The underlying assumption is that synonymous changes (*Ks*) are neutral with regard to evolutionary selection and should occur at a fixed rate in a given region of the genome. In contrast, the nonsynonymous changes (*Ka*) could represent selection, if they occur at a higher rate than expected from the background *Ks* rate. A ratio >1 is thus commonly used as an indicator of strong positive selection and accelerated evolution, and the higher the ratio is, the greater the relative number of nonsynonymous substitutions (and hence potential adaptive evolution). However, many protein sequences simultaneously experience strong purifying selection (disallowing deleterious amino acid changes) over most of their length and can thus be targeted by adaptive evolution (positive selection) at only a few sites. Thus, a *Ka*/*Ks* ratio taken over all sites in a given protein can result in a value much less than 1 even if strong positive selection has occurred at one or a few sites. Taking this approach, the average ratio across all 49 loci is 0.322, slightly greater than the human-chimpanzee genome-wide average of 0.23 (12). Within humans and chimpanzees, the recognition category had the greatest

[8] Sometimes the molecular function or biological process listed in Gene Ontology is not wholly descriptive of a gene product's function. For example, searching *CMAS* returns biological processes of "CMP-*N*-acetylneuraminate biosynthesis," "lipopolysaccharide biosynthesis," molecular function of "*N*-acylneuraminate cytidylyltransferase activity," and cellular component of "nucleus." But no statement indicates that this gene is involved in the biosynthesis of sialylated glycans.

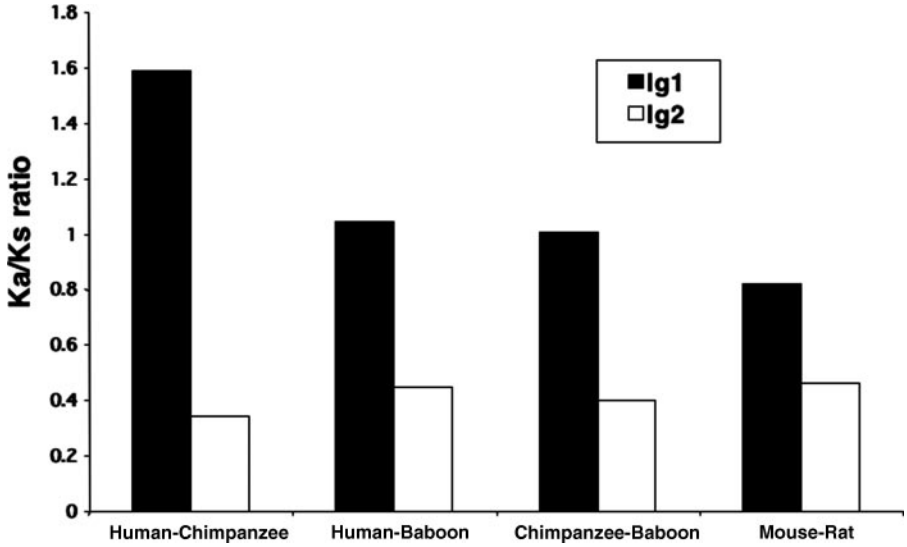


FIGURE 3. **Comparison of *Ka*/*Ks* ratios between concatenated Siglec Ig domains among species pairs.** Siglec Ig1 and Ig2 domains were concatenated, and *Ka*/*Ks* ratios were calculated over the total concatenated length of each domain. See "Results" for the Siglec loci used in each comparison.

average *Ka*/*Ks* ratio (0.465), followed by the recycling and degradation category (0.292); biosynthesis (0.293); and activation, transport, and transfer (0.213). The one currently known modification gene (*CMAH*) has been pseudogenized in humans by a 92-bp exon deletion (13, 14), so comparisons between human and chimpanzee are not appropriate. The average *Ka*/*Ks* values between the recognition category and the activation, transport, and transfer category, the two extremes, are significantly different from each other ($p = 0.003$, t test), and a comparison between the recognition category and the recycling and degradation category approaches significance ($p = 0.08$).

Average *Ka*/*Ks* ratios for the Sia functional categories are greater for primates compared with rodents across all categories (Fig. 2), a general finding consistent with other studies. For 38 loci for which reliable rat orthologs are available (see supplemental Table 2), rodent *Ka*/*Ks* ratios range from 0.005 to 0.757, and the average *Ka*/*Ks* ratio from mouse-rat comparisons is 0.156, a value smaller than that from human-chimpanzee comparisons. Primates have significantly greater average *Ka*/*Ks* values than rodents for the activation, transport, and transfer ($p = 0.005$, t test) and recycling and degradation ($p = 0.009$, t test) categories. As with the human-chimpanzee pair, the mouse-rat comparisons showed the highest average *Ka*/*Ks* ratio for the recognition group, although there is no statistically significant difference between taxa for this category. Of 33 orthologous loci examined between primates and rodents (excluding the CD33rSiglecs, which are not strictly orthologous), 22 (67%) showed greater *Ka*/*Ks* ratios in human-chimpanzee comparisons than in mouse-rat comparisons ($p = 0.05$ by a binomial test) (data not shown), suggesting an overall acceleration in primates compared with rodents.

Overall, the high rate of substitution and the relatively high *Ka*/*Ks* values suggest that the recognition category is evolving more rapidly than the others. This difference between gene categories may reflect a difference in evolutionary environment. Previous work in the anthocyanin pathway (39) suggested that genes upstream in a biosynthetic pathway tend to evolve more slowly than downstream genes. Although the Sia biology pathway as we have defined it here is not strictly linear, there is a general trend toward early acting loci such as those involved in biosynthesis and activation/transport/transfer evolving under more constraint than downstream loci such as those in the recognition category (Table 2).

Within the recognition group, Siglecs account for 56% (9 of 16) of human genes and 42% (5 of 12) of mouse and rat genes. *Ka*/*Ks* ratios for Siglec loci are significantly greater than those for non-Siglec members of the recognition group in both primates ($p = 0.006$) and rodents ($p = 0.007$) (Table 3), indicating that Siglecs are driving the higher values for this category. This difference appears to come mainly from an increase in *Ka* values rather than *Ks* values (Table 3), consistent with the notion that Siglecs may be undergoing adaptive evolution in humans and primates (19, 20). Indeed, comparisons of the chimpanzee and human genomes indicated that CD33rSiglecs are among the fastest evolving groups of genes in the entire genome (12).

The Sia-binding V-set Ig-like Domains of Siglecs Are Evolving Most Rapidly—Ka/*Ks* ratios calculated across the entire coding regions of genes can miss important changes because of substitutions at a relatively small number of sites. We therefore looked for domain-specific evolutionary changes between humans and chimpanzees. Such potentially important changes were found in Siglecs, sialyltransferases, and *HF1*. Details regarding the first two are presented here, and evaluation of *HF1* will be reported elsewhere.[9]

Siglecs have multiple extracellular Ig-like domains, followed by a single transmembrane domain and a short cytoplasmic tail (4, 33). The first Ig-like domain (Ig1, V-set Ig-like domain) is known to be responsible for Sia recognition. Prior analyses have suggested domain-specific accelerated evolution associated with a functional change in the Ig1 domain of human *SIGLEC9* (19), as well as a more rapid accumulation of nonsynonymous substitutions compared with an adjacent domain (Ig2, C2-set Ig-like domain) (20). These data indicate that Ig1 might be the target for evolutionary change in the primate lineage. We therefore re-examined our prior analyses (20) by examining non-CD33rSiglec loci (*SIGLEC1* and *CD22*) and excluding *SIGLEC11* and *SIGLEC5* (because of evidence of gene conversion) (21, 23). The Siglec loci thus used were as follows: human *SIGLEC1*, *CD22*, *CD33*, *SIGLEC6–10*, and *SIGLEC12* (*SIGLEC13* is deleted in the human genome (20)); chimpanzee *SIGLEC1*, *CD22*, *CD33*, *SIGLEC6–10*, *SIGLEC12*, and *SIGLEC13*; baboon *CD33*, *SIGLEC6*, *SIGLEC8–10*, and *SIGLEC13* (*SIGLEC7* and *SIGLEC12*

[9] R. E. Taylor, T. K. Altheide, A. Varki, unpublished data.

A

		SNA: Siaα2-6Gal(NAc)				MAH: Siaα2-3Gal			
		Hsa	Ptr	Rno	Mmu	Hsa	Ptr	Rno	Mmu
	Number of individuals	7	8	6	6	7	8	6	6
Lung	Bronchial Epithelium, Apical	Strong							
	Mucin in Lumen	Strong	Strong				Weak		
	Bronchial Epithelium, Goblet cells	Weak	Strong				Strong		
	Stroma	Weak	Weak	Weak	Weak	Weak	Weak	Weak	Weak
	Vascular endothelium	Strong	Strong	Strong	Strong	Weak	Weak	Weak	Weak
	Alveolar Epithelium		Weak	Strong	Strong	Weak	Weak	Weak	Weak
Kidney	Glomeruli	Strong	Strong	Strong	Strong	Strong	Strong	Strong	Strong
	Bowman's capsule	Weak	Strong	Strong	Strong	Strong	Weak		
	Proximal tubules								
	Distal tubules	Weak	Weak	Strong			Weak		
	Collecting ducts	Weak	Weak	Strong	Strong		Strong	Strong	
	Vascular endothelium	Strong	Strong	Strong	Strong	Strong	Strong	Strong	Strong
Spleen	Periarteriolar lymphoid tissue	Strong	Strong	Strong		Strong	Strong	Strong	Strong
	T cell zone	Weak	Strong	Strong		Strong	Strong	Strong	Strong
	B cell zone	Strong	Strong		Strong		Weak		
	Red pulp cords		Strong				Strong		
	Sinusoidal endothelium	Strong	Strong	Strong	Strong	Weak	Strong	Strong	Strong

Staining

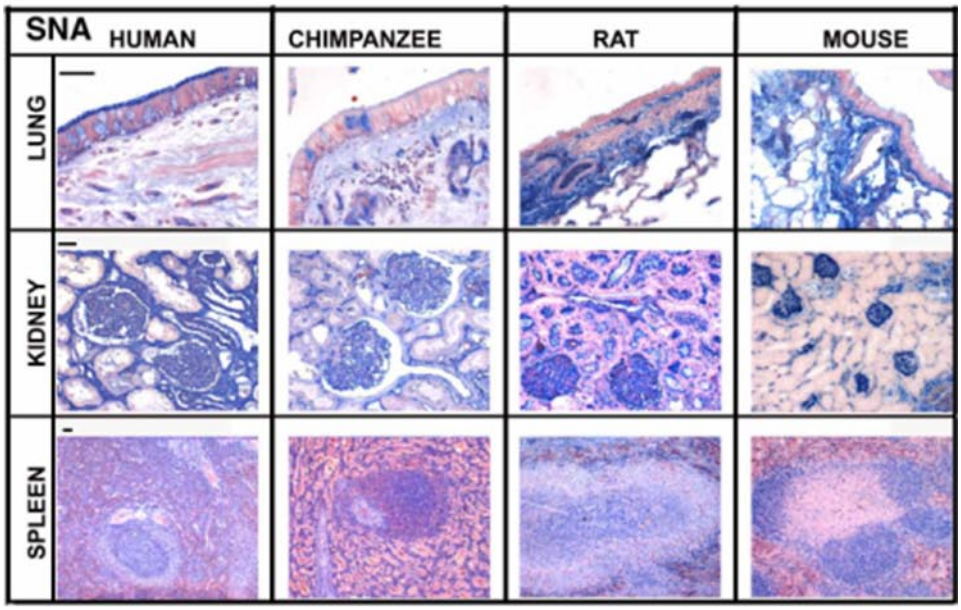
Strong	
Weak	
Absent	

Species Code	
Human	Hsa
Chimpanzee	Ptr
Rat	Rno
Mouse	Mmu

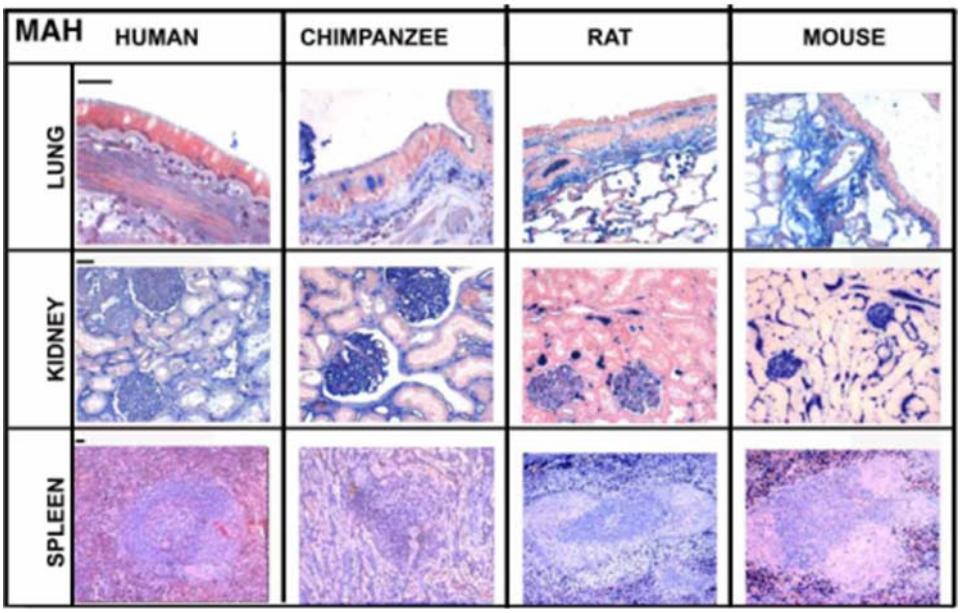
MAH = *Maackia amurensis* hemagglutinin

SNA = *Sambucus nigra* agglutinin

B



C



are deleted in the baboon genome; baboon *SIGLEC1* and *CD22* are not available (20)). For mouse and rat, we used the available reliable sequences, which were *Cd33* and *SiglecE–G*. Orthology between primate and rodent CD33rSiglecs is unclear because several exon/domain-shuffling events appear to have occurred in the primate lineage (20). Thus, we could not reliably compare individual primate and rodent CD33rSiglecs genes.

In comparisons between closely related species such as human and chimpanzee, a lack of nucleotide substitution can result in $Ks = 0$, which renders the Ka/Ks ratio statistically meaningless. Thus, we used the statistic $Ka - Ks$ to detect the signature of natural selection ($Ka - Ks > 0$, $= 0$, and < 0 are consistent with positive selection, neutral selection, and purifying selection, respectively). In human-chimpanzee comparisons, six of nine V-set Ig1-coding sequences showed $Ka - Ks > 0$, and only one C2-set Ig2-coding sequence showed $Ka - Ks > 0$ (see supplemental Table 3). Fisher's exact test supported the significance of this difference ($p = 0.0498$), indicating that $Ka > Ks$ is found more frequently in Ig1 domains than in Ig2 domains. In human-baboon and chimpanzee-baboon comparisons, the proportion of genes showing $Ka > Ks$ was nearly equal between Ig1 and Ig2 domains (supplemental Table 3). The mean $Ka - Ks$ value of each Ig1- and Ig2-coding sequence was calculated in every comparison. Mann-Whitney tests (paired) indicated a significant difference of mean $Ka - Ks$ values between Ig1- and Ig2-coding sequences in human-chimpanzee comparisons ($p = 0.0195$), supporting the hypothesis that Sia-binding Ig1 is the target of accelerated evolution in human and chimpanzee lineages. Similar tests in human-baboon, chimpanzee-baboon, and mouse-rat comparisons gave mean $Ka - Ks$ values between Ig1- and Ig2-coding regions that showed the same trend, but were not statistically significant ($p > 0.05$).

The above approach compares Ig1- and Ig2-coding sequences that are only ~400 and ~300 bp, respectively. To obtain more robust statistical power, we concatenated all available Siglec Ig1- or all Ig2-coding sequences for each species. For concatenated Ig1, all primate comparisons showed $Ka/Ks > 1$, indicating rapid evolution (Fig. 3). The mouse-rat comparisons did not show $Ka/Ks > 1$ (0.821), but the value is still rather high. In contrast, all comparisons of concatenated Ig2 sequences gave relatively low Ka/Ks ratios. We performed Fisher's exact tests to compare rates of synonymous and nonsynonymous evolution between concatenated Ig1 and concatenated Ig2. Concatenated Ig1 domains had a greater number of total substitutions than concatenated Ig2 domains in all species pairs. All species pairs also had significant differences in the proportions of nonsynonymous and synonymous substitutions between concatenated Ig1 and concatenated Ig2 ($p < 0.010$ for all four comparisons), with more nonsynonymous changes in concatenated Ig1 (data not shown). Taken together, the above findings indicate that an accelerated accumulation of nonsynonymous substitutions has occurred in Ig1 compared with Ig2 and that the Sia recognition function of the Siglec Ig1 domains is more rapidly evolving in at least two different mammalian clades, primates and rodents, with the highest rate in humans.

Sialyltransferase Sequences Are Highly Conserved, but Their Tissue Expression Patterns Are Not—Sialyltransferases are responsible for the formation of sialylglycoconjugates by transferring the Sia group from CMP-Sia to one of many possible glycoconjugate acceptors. In striking contrast to the Sia-recognizing proteins, sialyltransferase sequences were found to be highly conserved among primates and rodents (Table 2 and supplemental Tables 1 and 2). Despite this, we found that the actual tissue pattern of Sia linkages generated by these enzymes varies widely across different tissue types among humans, chimpanzees, mice, and rats (Fig. 4). Using the lectins SNA and MAH to detect α2–6- and α2–3-Sia linkages, respectively, we found many interspecies differences and only a few consistent similarities (Fig. 4). For example, expression of SNA-positive α2–6-linked Sia in lung bronchioles was human-specific. α2–6-Linked Sia is also expressed in B cell areas of the spleen in human, chimpanzee, and mouse, but not in rat. SNA reactivity in the red pulp area of the spleen was seen only in chimpanzee. In kidney distal tubules, expression of α2–6-linked Sia was found discordantly in human and rat. However, expression of this linkage is preserved across all four species in endothelial cells and kidney glomeruli. Expression of MAH-positive α2–3-linked Sia was also found in T cell areas of the spleen and in kidney glomeruli of all four species examined. In contrast, it was seen only in chimpanzee lung bronchial epithelium goblet cells and chimpanzee spleen red pulp. Thus, each species appears to have experienced specific gains and losses of Sia expression, despite general conservation of sialyltransferase sequences.

Species-specific Changes in Sialylmotifs—Although the causes of species-specific differences in sialylation are mostly unclear, a few focused sequence changes in sialyltransferase catalytic domains could have effects on sialyltransferase action. All eukaryotic sialyltransferases have four conserved peptide regions in their catalytic domains, referred to as sialylmotifs L (long) and S (short) (40), 3 (41), and VS (very short) (42). Sialylmotif L is involved mainly in donor substrate binding (43), and sialylmotif S is important for binding to both donor and acceptor substrates (44). We identified a number of species-specific amino acid changes in the sialylmotif regions of several sialyltransferases. Because crystal structures of sialyltransferases are not currently available, protein secondary structure prediction was performed to obtain information about consequences of these species-specific amino acid changes. Comparison of predicted locations of helix, coil, and sheet structures among primates and rodents suggests that one locus (*ST8SIA3*) has potentially important structural changes between rodents because of both mouse- and rat-specific amino acid changes and that two additional loci (*ST6GALNAC3* and *ST8SIA2*) show potentially major structural changes in primates resulting from human-specific amino acid changes (Fig. 5, *A* and *B*). Of these, the human-specific change in *ST8SIA2* is of particular

FIGURE 4. **Multispecies comparisons of Sia linkage patterns across different tissue types.** *A*, summary of comparative expression patterns in various tissues among humans (*Homo sapiens* (*Hsa*)), chimpanzees (*Pan troglodytes* (*Ptr*)), rats (*Rattus norvegicus* (*Rno*)), and mice (*Mus musculus* (*Mmu*)). Reactivities (*blue*) of the various cell types in tissue were recorded using a relative color scale as indicated. *B*, examples of staining with the SNA lectin (specific for α2–6-linked Sia residues). *C*, examples of staining with the MAH lectin (specific for α2–3-linked Sia residues).

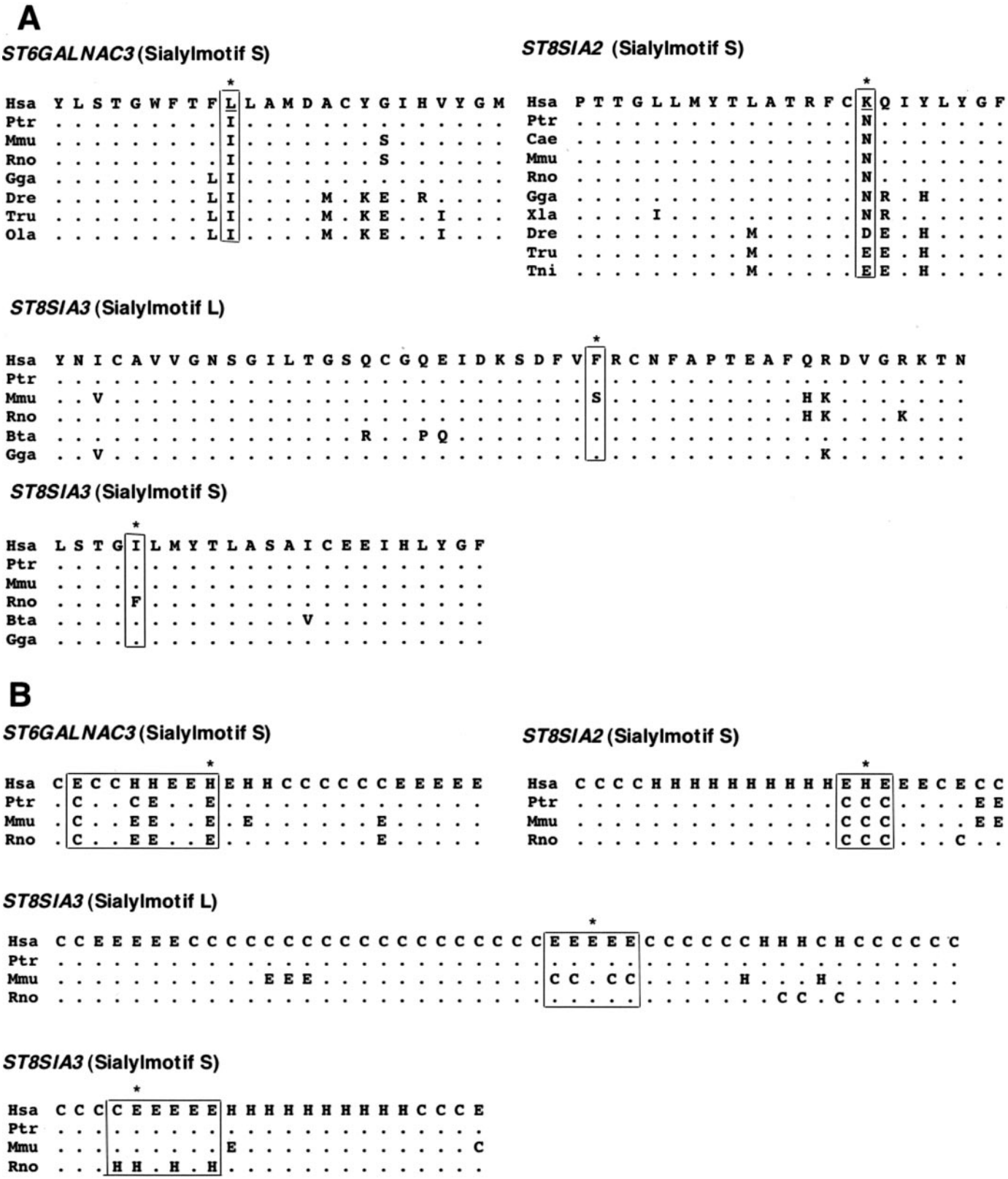


FIGURE 5. **Comparative analysis of sialylmotifs.** *A*, species-specific amino acid changes in the sialylmotifs of sialyltransferase loci. *Asterisks* represent amino acid differences that predict possible structural changes. *Hsa*, *H. sapiens*; *Ptr*, *P. troglodytes*; *Cae*, *Cercopithecus aethiops*; *Mmu*, *M. musculus*; *Rno*, *R. norvegicus*; *Bta*, *Bos taurus*; *Gga*, *Gallus gallus*; *Xla*, *Xenopus laevis*; *Dre*, *Danio rerio*; *Tru*, *Takifugu rubripes*; *Tni*, *Tetraodon nigroviridis*; *Ola*, *Oryzias latipes*. *B*, comparison of predicted structural changes in sialylmotifs from *A*. The prediction of secondary structure was performed using the entire amino acid sequence as described under "Experimental Procedures." Only the sialylmotif region is represented here. *Asterisks* denote sites of structurally important amino acid changes. *Boxes* represent the predicted structural differences caused by species-specific amino acid changes. *H*, *E*, and *C* represent predicted α-helix, β-sheet, and coil, respectively.

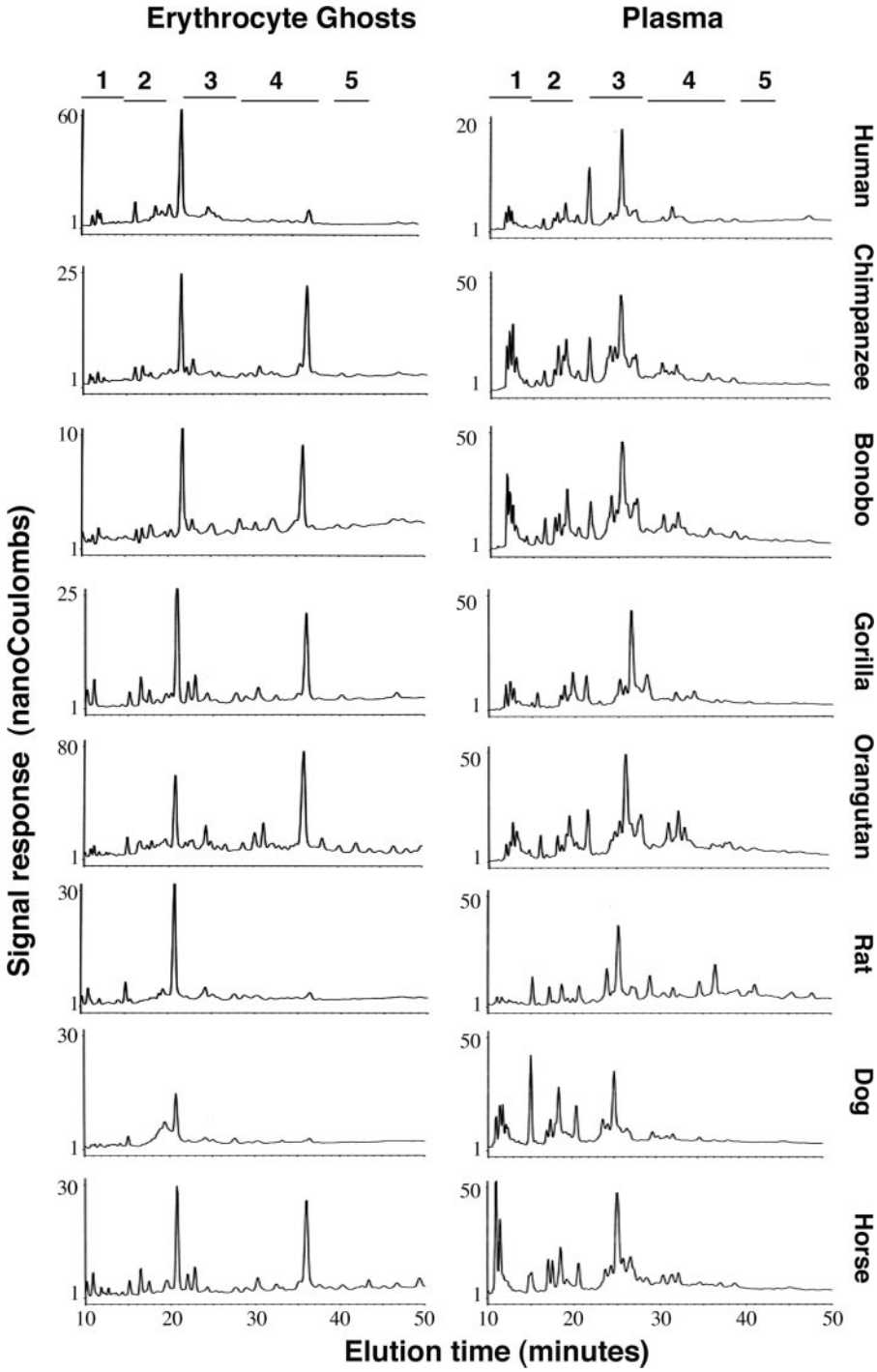


FIGURE 6. **Multispecies comparison of *N*- and *O*-glycan elution profiles by HPAEC-PAD.** The erythrocyte ghost and blood plasma oligosaccharide profiles of eight mammalian species are shown. The general elution positions for standard sialylated *N*- and *O*-glycans are shown as monosialyl-*O*-glycans (*1*), disialyl-*O*-glycans (*2*), disialyl-*N*-glycans (*3*), trisialyl-*N*-glycans (*4*), and tetrasialyl-*N*-glycans (*5*).

interest because it appears to be expressed mainly in fetal brain (45) and generates polysialic acid chains, which are known to be involved in regulating neural plasticity and neurite outgrowth (46–48).

Multispecies Comparisons of Erythrocyte and Plasma Protein N- and O-Glycans Confirm Rapid Evolution of Sialylation Patterns—The above tissue sialylation patterns were determined using linkage-specific lectins, which do not differentiate among different classes of glycans and do not provide information about underlying glycan structure. To obtain further biochemical evidence for the diversity of tissue sialylation, we studied glycoproteins from erythrocyte ghosts and plasma proteins in several mammalian taxa. (Mice were not studied because of the small quantities of material obtainable.) Total *N*- and *O*-glycans were released by hydrazinolysis and profiled by Dionex HPAEC-PAD. As shown in Fig. 6, the elution profiles of negatively charged (sialylated) glycans from each species were unique, indicating that sialylation patterns are also unique. (We did not further study the other potential cause of diversity, varying *N*-glycan branching.) *N*- and *O*-glycan diversity is pronounced between taxa, as evidenced by gains, losses, and shifts of various peaks. For example, both ghost and plasma proteins showed differences, with peak shifts in gorilla and orangutan compared with the other primates, a relative lack of tri- and tetrasialyl-*N*-glycans in rat and dog ghosts compared with other taxa, and differing amounts of mono- and disialyl-*O*-glycans in plasma between taxa.

Species-specific Diversity of Sialic Acid Types—Although the above profiling method has many advantages, one limiting factor is the fact that two common types of Sia residues (Neu5Ac and Neu5Gc) can cause significantly different elution properties for glycans to which they are attached. Indeed, some of the most striking differences between human and great ape samples could be partly due to the human lack of Neu5Gc. Another problem is that the hydrazinolysis procedure can result in some loss of *N*-glycolyl groups (converted into *N*-acetyl groups upon re-acetylation) and complete loss of *O*-acetyl esters on sialic acids. Thus, we also quantified the relative amount of different kinds of sialic acids in the erythrocyte ghost and plasma glycopeptides (Table 4). Although human ghost and plasma glycans contain primarily Neu5Ac, great apes contain predominantly Neu5Ac in plasma but mostly Neu5Gc in ghosts. Only small amounts of 9-*O*-acetylated Neu5Ac were seen in these hominids. Rat and horse exhibited high levels of *O*-acetylated Neu5Ac, whereas the other taxa showed lit-

TABLE 4

Differences in sialic acid types found on plasma and erythrocyte ghosts of various mammalian species

Plasma and erythrocyte ghosts were subjected to mild acid hydrolysis, and the released sialic acids were studied by 1,2-diamino-4,5-methylenedioxybenzene derivatization and high pressure liquid chromatography as described under "Experimental Procedures." Most *O*-acetylated sialic acids had a single *O*-acetyl ester at the 7-, 8-, or 9-position, and these are lumped together, as the esters can easily migrate from the 7- or 8-position to the 9-position.

Sample	*n*	% of total sialic acids			
		Neu5Ac	*O*-Acetylated Neu5Ac	Neu5Gc	*O*-Acetylated Neu5Gc
Human					
Ghosts	8	94–97	3–6	<0.5	<0.5
Plasma	8	94–97	3–6	<0.5	<0.5
Chimpanzee					
Ghosts	5	9–31	0	69–92	<0.5
Plasma	6	64–85	2–6	15–32	<0.5
Bonobo					
Ghosts	5	9–19	<0.5	81–91	<0.5
Plasma	5	69–80	3–5	16–27	<0.5
Gorilla					
Ghosts	5	8–12	<0.5	88–92	<0.5
Plasma	5	65–82	2–6	16–29	<0.5
Orangutan					
Ghosts	3	1–5	<0.5	90–98	4–5
Plasma	3	64–72	2–5	25–31	<0.5
Rat					
Ghosts	1	46	52.2[a]	1.7	<0.5
Plasma	1	31.2	65.9[a]	2.9	<0.5
Dog					
Ghosts	1	86.1	11.3	2.6	<0.5
Plasma	1	81	0.2	18.7	<0.5
Horse					
Ghosts	1	16	<0.5	74.8	9.2[b]
Plasma	1	55.8	30.5[b]	13.7	<0.5

[a] Samples included a significant amount of 7(8),9-di-*O*-acetylated species.
[b] Samples had the *O*-acetyl groups at the 4-position.

tle to none. Only orangutan appeared to have *O*-acetylated Neu5Gc, in contrast to the other primates as well as other mammals. Overall, we can conclude that both sialic acid diversity and expression are rapidly evolving among different taxa.

DISCUSSION

Combined genomic and biochemical data from multiple closely related species can provide new insights into evolution and functionality of particular biological systems and of broader classes of taxa. Large-scale genome-wide comparisons between humans and chimpanzees have suggested that positive selection may have influenced the evolution of human loci involved in processes such as sensory perception, olfaction, hearing, and brain growth (49–52). Previous work utilizing candidate gene approaches in comparative Sia biology has also suggested that real biological differences exist between humans and great apes for some loci and phenotypes (13, 15, 18, 19). Here, we have taken the reverse approach, namely a comparative genomic analysis of the previously candidate-targeted Sia biology system, particularly using the *Ka*/*Ks* ratio, a commonly used measure of protein evolutionary rate (see "Results" for a detailed description of this ratio). This systematic comparative analysis has expanded prior findings of multiple potentially functionally significant genetic and biochemical differences between humans and chimpanzees affecting Sia biology.

There appear to be different selective pressures between gene categories involved in Sia biology, as evidenced by differences in divergence rates and rates of evolution as measured by *Ka*/*Ks* ratios across categories. Sia-recognizing molecules in particular appear to be very rapidly evolving, and the acceleration in Siglec molecules that recognize Sia residues is consistent with the hypothesis that these loci play important roles in host immune modulation. Loci involved in Sia biosynthesis appear to be under stronger functional constraint in both primates and rodents. However, there is a striking disparity between the level of coding sequence conservation and species-specific expression of sialyltransferase products. The precise mechanisms and consequences of these unique species-specific expression patterns are currently unknown. Previous work has suggested that the expression of one sialyltransferase (*ST6GAL1*) may be regulated either by differential promoter usage or by changes in the expression of transcription factors (53–55). This may be the case for all sialyltransferase loci, as their generally high level of coding sequence conservation suggests that factors other than simple amino acid changes may be responsible for the patterns of interspecific expression variation. As for the additional rapid evolution of sialic acid types, most of the relevant genes have not yet been identified and cloned, so it cannot yet be determined whether coding sequence changes or regulatory changes are responsible for these patterns. Regardless of the underlying cause(s) of this rapid evolution, the fact that tissue sialylation patterns differ so widely among such closely related taxa raises caution about the use of animal model systems to understand human glycosylation-related disorders.

Overall, it appears that two distinct modes of rapid evolution are taking place in Sia biochemistry and biology. Within the CD33rSiglecs, there are ongoing changes in the actual amino acid sequences of the Sia-binding Ig-like domain associated with changes in binding activity. In contrast, the expression patterns of the sialyltransferases (and glycans in general) are rapidly diverging within mammals, even while their primary amino acid sequences remain conserved. Although these are different classes of loci that operate in different parts of the Sia life cycle, these two phenomena are related by the fact that the CD33rSiglecs recognize Sia residues originally placed onto glycan chains by the sialyltransferases. Overall, the current data are consistent with a recently proposed evolutionary scenario (4) predicting that terminal sialylation would have evolved more rapidly than other systems to evade pathogenic infections. Thus, whereas the sialyltransferase expression patterns defining the host sialome are rapidly evolving to evade pathogens that use Sia residues as targets for binding (a Red Queen effect) (8), the Sia-binding sites of CD33rSiglecs (which are thought to have the ability to recognize the self-sialome) are also rapidly evolving to keep up with the constantly changing sialome, resulting in a secondary Red Queen effect (4). It is also possible that CD33rSiglec Sia-binding sites need to simultaneously evolve rapidly to directly evade pathogens that express Sia residues, another primary Red Queen effect (4).

Interestingly, a second class of Sia-recognizing molecules, the selectins, did not show a similar rapid evolution of their Sia-binding C-type lectin domains. Although both the Siglecs and the selectins bind Sia residues, the selectins differ from the Siglecs in their recognition specificity and functions. Siglecs discriminate subtle differences in the specific Sia involved, such

as its underlying linkage, charge, and side chain type (4, 33). Unlike Siglecs, however, selectins do not require the entire sialic molecule for recognition, just the negative charge, which can even be provided by a sulfate ester at the same 3-position of galactose (35, 36). Thus, selectins should be under less pressure to evolve rapidly to match the host sialome. Also, whereas Siglecs appear to have both intrinsic and extrinsic recognition functions, selectins are thought to act primarily in intrinsic recognition processes in vascular biology. This suggests that intrinsic recognition is under stronger constraint than extrinsic recognition in Sia biology. Indeed, there are no amino acid substitutions between humans and chimpanzees in any of the selectin C-type lectin domains (data not shown), suggesting that these regions are under stronger functional constraint and less diversifying pressure.

Recent studies have suggested examples of domain-specific rapid evolution in settings in which *Ka*/*Ks* ratios for the entire genes showed no significant differences (19, 20). This hypothesis is supported by domain-specific analyses of the Siglec loci, which suggest more rapid evolution of the functional Sia-binding domain than adjacent domains. Also of note is the fact that we have so far not found as many major differences in Sia biology-related genes in rodents as in primates, despite the much greater time since their evolutionary divergence. Taken together, the data imply that the primate lineage, specifically the human lineage, has experienced differential selection pressures affecting Sia biology. More focused study of candidate loci and biochemical differences may help elucidate the causative mechanisms.

It has been suggested that the majority of gene expression differences between species are not necessarily functional adaptations, but rather the consequence of neutral or nearly neutral substitutions (56). If few gene expression changes are adaptive, then it may be even harder to see signatures of selection at the genomic DNA level. This underscores the important role that functional and biochemical studies must play in validating the existence and importance of biological changes between species. Our biochemical data underscore this fact, as we see marked species-specific differences in sialylation profiles between mammalian taxa that would not be predictable from sequence data alone. One additional way to help clarify genomic evidence for natural selection will be a population genetic approach, placing intraspecies polymorphism data in the context of divergence, to detect the footprint(s) of natural selection in these species. Functional and population genetic studies on several of these loci are underway to determine how these genetic and biochemical differences among primates and rodents may have contributed to functional phenotypic consequences relevant to the biological evolution of our species.

Acknowledgment—We thank the Glycotechnology Core Facility at the University of California, San Diego, for assistance with the HPAEC-PAD profiles.

REFERENCES

1. Traving, C., and Schauer, R. (1998) *CMLS Cell. Mol. Life Sci.* **54,** 1330–1349
2. Angata, T., and Varki, A. (2002) *Chem. Rev.* **102,** 439–470
3. Beyer, T. A., Rearick, J. I., Paulson, J. C., Prieels, J. P., Sadler, J. E., and Hill, R. L. (1979) *J. Biol. Chem.* **254,** 12531–12534
4. Varki, A., and Angata, T. (2006) *Glycobiology* **16,** 1R–27R
5. Pangburn, M. K. (2000) *Immunopharmacology* **49,** 149–157
6. Schwarzkopf, M., Knobeloch, K. P., Rohde, E., Hinderlich, S., Wiechens, N., Lucka, L., Horak, I., Reutter, W., and Horstkorte, R. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5267–5270
7. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., and Steenbergen, S. M. (2004) *Microbiol. Mol. Biol. Rev.* **68,** 132–153
8. Gagneux, P., and Varki, A. (1999) *Glycobiology* **9,** 747–755
9. Goodman, M., Bailey, W. J., Hayasaka, K., Stanhope, M. J., Slightom, J., and Czelusniak, J. (1994) *Am. J. Phys. Anthropol.* **94,** 3–24
10. Satta, Y., Klein, J., and Takahata, N. (2000) *Mol. Phylogenet. Evol.* **14,** 259–275
11. Chen, F. C., and Li, W. H. (2001) *Am. J. Hum. Genet.* **68,** 444–456
12. Chimpanzee Sequencing and Analysis Consortium (2005) *Nature* **437,** 69–87
13. 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
14. Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T., and Suzuki, A. (1998) *J. Biol. Chem.* **273,** 15866–15871
15. Chou, H. H., Hayakawa, T., Diaz, S., Krings, M., Indriati, E., Leakey, M., Paabo, S., Satta, Y., Takahata, N., and Varki, A. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11736–11741
16. Hayakawa, T., Satta, Y., Gagneux, P., Varki, A., and Takahata, N. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11399–11404
17. Angata, T., Varki, N. M., and Varki, A. (2001) *J. Biol. Chem.* **276,** 40282–40287
18. Gagneux, P., Cheriyan, M., Hurtado-Ziola, N., Brinkman Van Der Linden, E. C., Anderson, D., McClure, H., Varki, A., and Varki, N. M. (2003) *J. Biol. Chem.* **278,** 48245–48250
19. Sonnenburg, J. L., Altheide, T. K., and Varki, A. (2004) *Glycobiology* **14,** 339–346
20. Angata, T., Margulies, E. H., Green, E. D., and Varki, A. (2004) *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13251–13256
21. Hayakawa, T., Angata, T., Lewis, A. L., Mikkelsen, T. S., Varki, N. M., and Varki, A. (2005) *Science* **309,** 1693
22. Nguyen, D. H., Hurtado-Ziola, N., Gagneux, P., and Varki, A. (2006) *Proc. Natl. Acad. Sci. U. S. A.* **103,** 7765–7770
23. Angata, T., Hayakawa, T., Yamanaka, M., Varki, A., and Nakamura, M. (2006) *FASEB J.*, in press
24. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) *Nat. Genet.* **25,** 25–29
25. Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., and Thomas, P. D. (2005) *Nucleic Acids Res.* **33,** 284–288
26. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998) *Trends Biochem. Sci* **23,** 403–405
27. Nei, M., and Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426
28. Jukes, T. H., and Cantor, C. R. (1969) in *Mammalian Protein Metabolism* (Munro, H. N., ed.) pp. 21–132, Academic Press, Inc., New York
29. Rozas, J., and Rozas, R. (1999) *Bioinformatics* **15,** 174–175
30. Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001) *Bioinformatics* **17,** 1244–1245
31. Hardy, M. R., and Townsend, R. R. (1994) *Methods Enzymol.* **230,** 208–225
32. Hara, S., Yamaguchi, M., Takemori, Y., Nakamura, M., and Ohkura, Y. (1986) *J. Chromatogr.* **377,** 111–119
33. Crocker, P. R. (2005) *Curr. Opin. Pharmacol.* **5,** 431–437
34. Varki, A. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 7390–7397
35. McEver, R. P. (2002) *Curr. Opin. Cell Biol.* **14,** 581–586
36. Rosen, S. D. (2004) *Annu. Rev. Immunol.* **22,** 129–156
37. Kleene, R., Yang, H., Kutsche, M., and Schachner, M. (2001) *J. Biol. Chem.* **276,** 21656–21663

38. Chiba, A., Matsumura, K., Yamada, H., Inazu, T., Shimizu, T., Kusunoki, S., Kanazawa, I., Kobata, A., and Endo, T. (1997) *J. Biol. Chem.* **272,** 2156–2162
39. Rausher, M. D., Miller, R. E., and Tiffin, P. (1999) *Mol. Biol. Evol.* **16,** 266–274
40. Datta, A. K., and Paulson, J. C. (1997) *Indian J. Biochem. Biophys.* **34,** 157–165
41. Jeanneau, C., Chazalet, V., Auge, C., Soumpasis, D. M., Harduin-Lepers, A., Delannoy, P., Imberty, A., and Breton, C. (2004) *J. Biol. Chem.* **279,** 13461–13468
42. Geremia, R. A., Harduin-Lepers, A., and Delannoy, P. (1997) *Glycobiology* **7,** v–vii
43. Datta, A. K., and Paulson, J. C. (1995) *J. Biol. Chem.* **270,** 1497–1500
44. Datta, A. K., Sinha, A., and Paulson, J. C. (1998) *J. Biol. Chem.* **273,** 9608–9614
45. Giordanengo, V., Bannwarth, S., Laffont, C., Van Miegem, V., Harduin-Lepers, A., Delannoy, P., and Lefebvre, J. C. (1997) *Eur. J. Biochem.* **247,** 558–566
46. Fujimoto, I., Bruses, J. L., and Rutishauser, U. (2001) *J. Biol. Chem.* **276,** 31745–31751
47. Angata, K., Long, J. M., Bukalo, O., Lee, W., Dityatev, A., Wynshaw-Boris, A., Schachner, M., Fukuda, M., and Marth, J. D. (2004) *J. Biol. Chem.* **279,** 32603–32613
48. Seidenfaden, R., Krauter, A., Schertzinger, F., Gerardy-Schahn, R., and Hildebrandt, H. (2003) *Mol. Cell. Biol.* **23,** 5908–5918
49. Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2003) *Science* **302,** 1960–1963
50. Gilad, Y., Man, O., Paabo, S., and Lancet, D. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100,** 3324–3327
51. Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2005) *PLoS Biol.* **3,** e170
52. Dorus, S., Vallender, E. J., Evans, P. D., Anderson, J. R., Gilbert, S. L., Mahowald, M., Wyckoff, G. J., Malcom, C. M., and Lahn, B. T. (2004) *Cell* **119,** 1027–1040
53. Appenheimer, M. M., Huang, R. Y., Chandrasekaran, E. V., Dalziel, M., Hu, Y. P., Soloway, P. D., Wuensch, S. A., Matta, K. L., and Lau, J. T. (2003) *Glycobiology* **13,** 591–600
54. Takashima, S., Kurosawa, N., Tachida, Y., Inoue, M., and Tsuji, S. (2000) *J. Biochem.* (*Tokyo*) **127,** 399–409
55. Dalziel, M., Lemaire, S., Ewing, J., Kobayashi, L., and Lau, J. T. Y. (1999) *Glycobiology* **9,** 1003–1008
56. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Paabo, S. (2004) *PLoS Biol.* **2,** e132
57. Van Valen, L. (1974) *Nature* **252,** 298–300
58. Hamilton, W. D., Axelrod, R., and Tanese, R. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87,** 3566–3573
59. Goodman, M. (1999) *Am. J. Hum. Genet.* **64,** 31–39
60. International Human Genome Sequencing Consortium (2004) *Nature* **431,** 931–945
61. Akiyama, Y., Onizuka, K., Noguchi, T. and Ando, M. (1998) *Genome Inform. Ser. Workshop Genome Inform.* **9,** 131–140
62. Bame, K. J. and Esko, J. D. (1989) *J. Biol. Chem.* **264,** 8059–8065