Gene conversions in the Siglec and CEA immunoglobulin gene families of primates

Mouldi Zid

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of the University of Ottawa in partial fulfillment of the requirements for the M.Sc. degree in the Ottawa Carlton Institute of Biology

Thèse soumise à la Faculté des études supérieurs et postdoctorales de l'Université d'Ottawa en vue de l'obtention de la maîtrise en biologie de l'Institut pour la biologie Ottawa Carlton

Examiners

The following members of the Ottawa Carlton Institute of Biology

Dr. Guy Drouin

Supervisor

Dr. Michel Dumontier

Examining Member (Carleton University)

Dr. Stéphane Aris-Brosou

Examining Member (University of Ottawa)

Dr. Marcel Turcotte

Examining Member (University of Ottawa)

Table of contents

Acknowledgements

I would like to thank my supervisor, Dr. Guy Drouin, for the opportunity that he gave me to do this MSc. degree at the University of Ottawa. His cheerful guiding throughout my program was an extraordinary experience and gift for me. As well as, I would sincerely like to thank him for reviewing my thesis.

I would also like to thank Dr. Stéphane Aris-Brosou who provided me with valuable advice and guidelines.

In addition, I would like to thank you Dr. Marcel Turcotte and Dr. Michel Dumontier for agreeing to be a members of my graduate committee.

Finally, I would like to thank all of my fellow graduate students that helped me in various ways throughout my MSc. degree.

List of tables

List of figures

List of abbreviations

aa	amino acid
bp	base pair
CEA	Carcinoembryonic antigen
CEACAM	CEA-related cell adhesion molecule
CD	Cluster of differentiation
dHJs	double Holliday junctions
DNA	Deoxyribonucleic acid
DSB	Double strand DNA breaks
DSBR	Double strand break repair
gBGC	GC-biased gene conversion
HBG	γ-globin gene
ITIM	Immunoreceptor tyrosine-based inhibitory motif
ITAM	Immunoreceptor tyrosine-based activation motif
I-CAM	Intercellular cellular adhesion molecule
IR	Ionizing radiation
JAC	Junctional adhesion molecule
LD	Linkage disequilibrium
MAG	Myelin-associated glycoprotein
Mb	Megabase
MAdCAM	Mucosal addressin cellular adhesion molecule
NK	Natural killer
N-CAM	Neural cellular adhesion molecule
Vascular-CAM	Vascular cellular adhesion molecule
SLAM	Signaling lymphocytic activation molecule
SHP	Protein tyrosine phosphatase
SLGS	Siglec-like gene short
SLGL	Siglec-like gene long
SHM	Somatic hypermutation
SIGLECS	Sialic acid-binding immunoglobulin-like lectins

Abstract

Siglecs and CEA are two families of cell surface proteins belonging to the immunoglobulin superfamily. They are thought to be involved in cell-cell interactions and have various other biological functions. We used the GENECONV program that applies statistical tests to detect gene conversion events in each family of five primate species. For the Siglec family, we found that gene conversions are frequent between CD33rSiglec genes, but are absent between their conserved Siglec genes. For the CEA family, half of gene conversion events detected are located in coding regions. A significant positive correlation was found between the length of the conversions and the similarity of the converted regions only in the Siglec gene family. Moreover, we found an increase in GC-content and similarity in converted regions compared to non-converted regions of the two families. Furthermore, in the two families, gene conversions occur more frequently in the extracellular domains of proteins, and rarely in their transmembrane and cytoplasmic regions. Finally, these two families appear to be evolving neutrally or under negative selection.

Résumé

Siglecs et CEA sont deux familles de protéines de surface cellulaire appartenant à la superfamille des immunoglobulines. Elles sont impliquées dans les interactions cellule-cellule et dans diverses autres fonctions biologiques. Nous avons utilisé le programme GENECONV qui applique des tests statistiques pour détecter les événements de conversion génique dans chaque famille chez cinq espèces de primates. Pour la famille Siglec, nous avons trouvé que les conversions géniques sont fréquentes entre les gènes CD33rSiglec, mais sont absent entre les gènes Siglec conservés. Pour la famille CEA, la moitié des conversions géniques détectées sont situées dans les régions codantes. Une corrélation positive significative a été trouvée entre la longueur des conversions et la similarité des régions converties seulement dans la famille Siglec. Par contre, nous avons trouvé une augmentation du contenu en GC et de la similarité dans les régions converties par rapport aux régions non-converties dans les deux familles. De plus, dans les deux familles, les conversions géniques se produisent plus fréquemment dans les domaines extracellulaires des proteines, et rarement dans leurs regions transmembranaires et cytoplasmiques. Enfin, ces deux familles semblent évoluer de façon neutre ou sous l'effet de la sélection négative.

Chapter 1. Introduction

1. Gene conversions

1.1. Mechanism of gene conversion

Gene conversion is considered one of the major homologous recombination mechanisms responsible for generating genetic variation in various species ranging from bacteria to eukaryotes (Chen et al. 2007). Unlike crossover, gene conversion is the nonreciprocal exchange of genetic information between two genes (acceptor and donor) that share high degree of homology (Strachan and Read 1999). There are two known types of gene conversions that have been observed based on the conversion targets. A conversion between alleles on sister chromatids or homologous chromosomes referred to as allelic gene conversion whereas conversion between paralogous sequences either on the same chromosome or on different chromosomes are called ectopic gene conversion (Petes and Hill 1988). Gene conversion is well known as an important mechanism in the evolution of multigene families, mediating, for example, the concerted evolution of ribosomal RNA (rRNA) gene family in Eubacteria, Archaea, and Eukaryotes (Liao 2000; Rooney 2004; Eickbush and Eickbush 2007).

Gene conversion was originally observed in meiotic products of fungi where it refers to a non-Mendelian segregation of genes in tetrads of a heterozygous diploid strain (Radding 1978). In 1980, Slightom et al. reported a gene conversion event between the highly similar human HBG1 and HBG2 fetal globin genes (see also Scott et al. 1984). They showed that the first two thirds of adjacent non-allelic human $^{G}\gamma$- and $^{A}\gamma$-globin genes had almost identical sequences and that the high similarity found in this region was greater than that observed between two $^{A}\gamma$-globin alleles.

There are two major non-exclusive models that can account for gene conversion events: the double strand break repair (DSBR) model of Szostak and colleagues (1983), and the synthesis-dependent strand annealing (SDSA) model (Figure 1.1). In both models, gene conversion occurs at high frequencies in germ cells during meiosis division and in somatic cells during mitosis division (Szostak et al. 1983; Iwase et al. 2010). During meiosis, double strand DNA breaks (DSBs) is catalyzed by a type II topoisomerase-like enzyme (SPO11). While, during mitosis, DSBs can be created by various genotoxic agents including ionizing radiation (IR), stalled replication forks or by the induction of specialized endonucleases such as I-*Sce*I and HO mega-endonucleases, which have been extensively used to study the repair of DSBs in the yeast *Saccharomyces cerevisiae* (Chen et al. 2007; Kee and D'Andrea 2010). In the DSBR model, the 5' ends of DSBs are degraded by endonucleases releasing 3' single DNA strands of several hundred base pairs (Figure 1.1). After finding the homologous sequences, one single-stranded DNA invades the homologous DNA duplex creating a structure D-loop, which is extended by repair synthesis (Figure 1.1). Once repair synthesis is complete, the D-loop migrates and allows the other 3' single stranded DNA tail to pair (Figure 1.1). After DNA synthesis, double Holliday junctions (dHJs) are created by repair synthesis and then cleaved by an HJ. If the dHJs are symmetrically resolved, this results in a gene conversion with non-crossover. In contrast, if dHJs are asymmetrically resolved, this results in a gene conversion with crossover (Chen et al. 2009; Figure 1.1). The SDSA pathway allows the repair of the DSB by moving the invading DNA strand outside the duplex and annealing to the other 3' end of the DSB. After the strands anneal, steps of DNA synthesis and cleavage of mismatched sequences may be necessary (Figure 1.1).

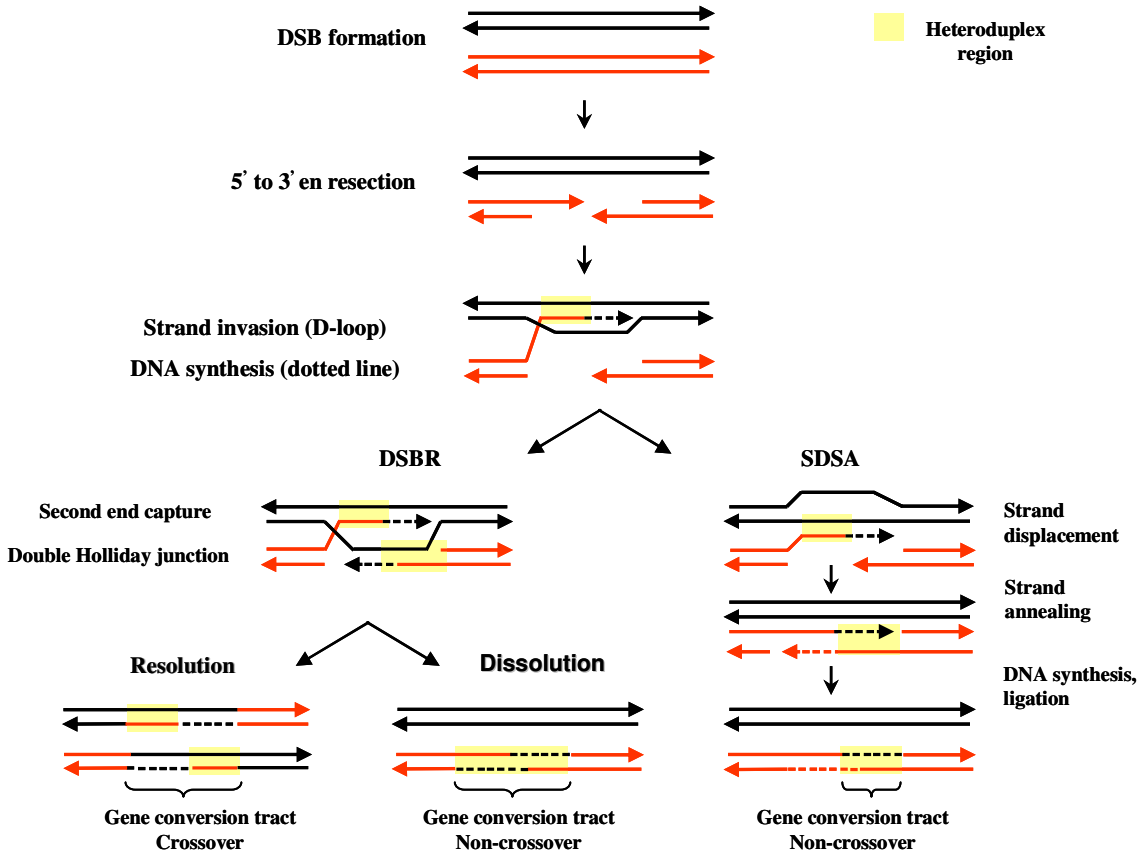


Figure 1.1. Mechanisms of gene conversion. Three models have been proposed to explain these different pathways: Double-Strand-Break Repair (DSBR), dissolution of the double Holliday junctions and Synthesis-Dependent Strand-Annealing (SDSA). Adapted from Duret and Galtier (2009).

1.2. Roles of gene conversion

The process of gene conversion becomes especially critical for a number of reasons. Gene conversion has a different effect on linkage disequilibrium (LD) in several genomes, such as humans (Ardlie et al. 2001) and *Drosophila melanogaster* (Langley et al. 2000; Ardlie et al. 2002). On the other hand, it seems that gene conversion associated with recombination can be biased toward GC nucleotides, as has been shown in several organisms (Pessia et al. 2012). GC-biased gene conversion (gBGC) is a process that can increase the GC-content of converted DNA sequences of a wide variety of eukaryotes and prokaryotes species. This process may explain the positive correlation between the recombination rate and GC-content observed in many organisms such as human (Meunier and Duret 2004; Chen et al. 2006), *Saccharomyces cerevisiae* (Gerton et al. 2000), *Mus musculus* (Perry and Ashworth 1999), *Drosophila melanogaster* (Singh et al. 2005), and birds (Hurst et al. 1999) genomes. Other studies have shown that gene conversion increases the GC-content of tandemly repeated sequences during evolution (Noonan et al. 2004; Backström et al. 2005). Furthermore, the study of Benovoy et al. (2005) showed that gene conversion affects the GC-content between dispersed duplicated genes in the yeast and *Arabidopsis* genomes.

Moreover, gene conversion also plays a major role for the molecular evolution of several organisms by contributing to both maintenance of genetic information and generation of genetic variability, by introducing multiple substitutions into a gene family member (Hwang and Green 2004). The result of such mutagenic gene conversion events is the generation of different polymorphic alleles within populations or mutations that lead to various inherited diseases. Consequently, studying gene conversion is important to identify

the genetic diseases and disorders such as congenital adrenal hyperplasia and spinal muscular atrophy (Lawson et al. 2009).

As mentioned above, gene conversions have been shown to be implicated in the concerted evolution of many gene families of various organisms from bacteria to humans (Santoyo and Remero 2005; Benovoy and Drouin 2009). A total of 526 gene conversion events were detected involving 143 (2%) of the 7829 pairs of duplicated genes in *Caenorhabditis elegans* (Semple and Wolfe 1999). Drouin (2002) characterized 151 gene conversions within 202 yeast gene families. Ezawa et al. (2006) studied 2,641 gene quartets, each consisting of two pairs of orthologous genes in mouse and rat, and found that 488 (18%) appear to have undergone gene conversion. In the rice genome, 377 gene conversion events were detected within 626 multigene families (Xu et al. 2008). However, these studies investigated gene conversion events only between pairs of protein-coding genes, although conversion can occur between any pair of highly similar regions (Chen et al. 2007). In humans, gene conversions between multigene family members have been described in a wide variety of protein coding genes. For example, gene conversions have been shown to occur between human genes coding for the β subunit of gonadotropin hormones (Hallast et al., 2005) and growth hormones (Petronella and Drouin 2011). The large palindromic sequences found in the human Y chromosome have also been shown to be subject to frequent gene conversion events (Rozen et al. 2003). Gene conversions also occur between Alu and LINE-1 elements (Chen et al. 2007).

1.3. Methods for detecting gene conversion

Several methods for detecting gene conversion and homologous recombination have been developed over the last 15 years and are diverse in terms of their underlying approaches. The method of Stephens (1985) attempts to examine sites that are associated with particular phylogenetic partitions of the set of sequences into two groups: thus a site is associated with a partition if it has one allele in one subset and a different allele in the other. This method is a statistical test to determine whether a set of sites associated with a particular partition are clustered. Sawyer's (1989) developed a statistical method that overcomes some shortcomings of Stephens' method by analyzing the distribution of maximal-length segments common to all pairs of sequences and control for variable mutation rates along the genome. In this method, a Monte Carlo test involving various permutations of silent sites is used to estimate the significance of the distribution of polymorphic sites. McGuire et al. (1997) have developed an alternative approach that employs a least-square and maximum likelihood method for determining the possibility of recombination within a given window. Other phylogenetic methods include the phylogenetic scanning method of Fitch and Goodman (1991), the parsimony-based approach of Hein (1993), and a Bayesian method devised by McGuire et al. (2000). Many of these and other methods have been reviewed extensively elsewhere (Weiller 1998; Drouin et al. 1999; Wiuf et al. 2001).

2. Immunoglobulin superfamily

The immunoglobulin superfamily (IgSF) is one of the largest groups of glycoproteins, including thousands of different proteins expressed by diverse immune cells of eukaryote and prokaryote species (Barrow and Trowsdale 2008; Wai et al. 2012). This family contains several members, such as the major histocompatibility complex type I and

type II molecules, virus receptors, T cell receptors complex, cell surface antigen receptors, co-stimulatory molecules, co-receptors, binding molecules, and some cytokine receptors (Bork et al. 1994; Dermody et al. 2009). The majority of these proteins are linked to cell surface and they are also secreted molecules. Members of the IgSF are characterized by the presence of one or more extracellular Ig-like domains, a single transmembrane region, and a cytoplasmic tail (Juliano 2002; Cheng et al. 2012). The Ig domains possess a characteristic sandwich structure, which is formed by two opposing antiparallel β-strands stabilized by a conserved disulfide bonds between cysteine residues (Dermody et al. 2009).

IgSF proteins play a crucial role in diverse cellular phenomena, including recognition process, cytoskeleton organisation, cell motility, binding and/or adhesion processes of cells (Holness and Simmons 1994; Walsh and Doherty 1997). Moreover, several IgSF proteins are implicated in diverse aspects in the central and peripheral nervous system, including brain development, axon guidance, neuronal migration, as well as in the formation and function of synapse in adult (Yamagata et al. 2002; Rougon and Hobert 2003). Other IgSF proteins such as basigin, nectin-2, and nectin-3 are expressed on the testicular cells and have a crucial function in spermatogenesis (Toshimori et al. 2006). Another IgSF member, Izumo, has also been identified as sperm membrane protein required for sperm to fuse with eggs in mouse and human (Inoue et al. 2005). Furthermore, IgSF proteins are involved in multiple interactions between immune cells and their cellular partners. Interestingly, the domains of these proteins exhibit a homophilic (e.g., N-CAM, JAC, and the myelin protein P0) or heterophilic interactions (e.g., I-CAM, MAdCAM, and Vascular-CAM; Steffe et al. 1996; Aricescu and Jones 2007).

The diversity of immunoglobulin genes is generated by three different processes, which are switch recombination, somatic hypermutation (SHM) and gene conversion (Sun

et al. 2012). In such species as chickens and rabbits, the diversification of the primary antibody repertoire is achieved in specialized tissues such as the bursa of Fabricius by a gene conversion mechanism that uses a set of immunoglobulin variable pseudogenes as templates (Weinstein et al. 1994; Arakawa and Buerstedde 2004). In sheep and cows, the diversity in the genes involved in immune system results in somatic hypermutation that occurs in the ileal Peyer's patches (Reynaud et al. 1991; Lucier et al. 1998).

2.1. Siglec family

2.1.1. Definition and structure of Siglecs

The sialic acid-binding immunoglobulin-like lectins (Siglecs) are a family of type I transmembrane proteins that recognise different fragments of sialic acid present on glycolipids and glycoproteins (Varki and Angata 2006). Their unique structural properties include a single N-terminal V-set domain that binds sialic acid, followed by a variable number of C2-set domains that can vary in number from one (*SIGLEC3* and *SIGLEC15*) to sixteen (*SIGLEC1*) (Crocker et al. 2007; Varki 2009; Figure 1.2). One peculiar member designated *SIGLEC12* contains two N-terminal IgV domains (Figure 1.2). Most Siglecs proteins contain one or more conserved immunoreceptor tyrosine-based inhibitory motifs (ITIMs) and/or ITIM-like motifs in their cytosolic tails (see Figure 1.2). These motifs are associated with protein tyrosine phosphatases containing SHP-1 and SHP-2 domains when they are phosphorylated by Src family tyrosine kinases, as well as with the SH2-domain-containing inositol polyphosphate 5-phosphatase (SHIP; Crocker and Redelinghuys 2008).

In humans, the siglecs family can be subdivided in two subgroups on the basis of similarities in sequence and evolutionary conservation (Crocker et al. 2007). An evolutionary conserved subgroup comprises *SIGLEC1* (Sialoadhesin, located on human

chromosome 20p13), *SIGLEC2* (CD22, located on human chromosome 19q13.1), *SIGLEC4* (Myelin-associated glycoprotein (MAG), located on human chromosome 19q13.1), and *SIGLEC15* (located on human chromosome 18q12.3). The members of this subgroup do not present a specific name and are quite distantly related (~25-30% sequence similarity) (Crocker et al. 2007; Pillai et al. 2012). The second subgroup of Siglecs is known as CD33-related Siglecs (CD33rSiglecs), because the members of this group possess a very high degree of homology with the molecule *SIGLEC3* (CD33), and comprise *SIGLEC5* through *SIGLEC12*, *SIGLEC14* and *SIGLEC16*. In humans, CD33rSiglec genes have been mapped to chromosome 19q13.3-13.4 and are located near the KLK gene cluster (Angata et al. 2004).

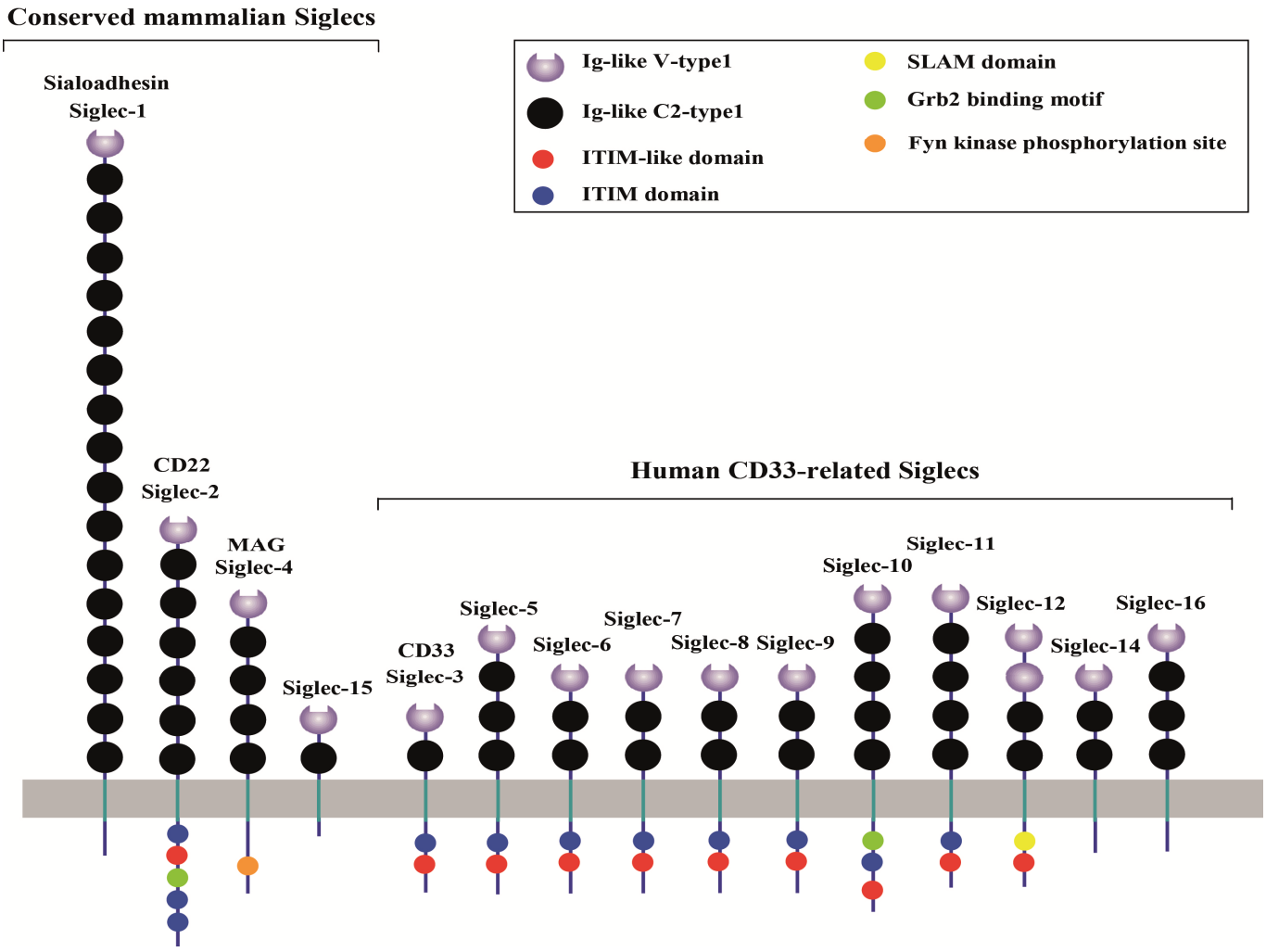


Figure 1.2. Structure of human Siglecs genes. Adapted from Cao and Crocker (2011) and Jandus et al. (2011).

In mammals, members of CD33rSiglecs share about 50-85% of sequence similarity and seem to evolve rapidly by multiple processes, including essentially inverse duplication (e.g. *SIGLEC11/SIGLEC16)*, gene conversion *(*e.g. *SIGLEC11/SIGLEC16* and *SIGLEC5/SIGLEC14)*, exon shuffling and exon loss (Angata et al. 2004; Crocker et al. 2007; Pillai et al. 2012). In rodents, the CD33rSiglecs subgroup has undergone a dramatic loss of genes (Cao et al. 2009).

2.1.2. Expression pattern and functions of Siglecs

Siglecs members are expressed on diverse cells of the immune system and involved in distinct functions. The *SIGLEC1* gene was characterized as the first family member of Siglecs that have the ability to bind glycoconjugate ligands in a sialic acid-dependent manner. It is highly expressed on tissue macrophages in the spleen, bone marrow, lungs, lymph node, colon and liver (Hartnell et al. 2001; Figure 1.3). Furthermore, macrophages also express five members of the Siglec family, *SIGLEC3*, *SIGLEC5*, *SIGLEC11*, *SIGLEC15* and *SIGLEC16* (Figure 1.3). *SIGLEC2* and *SIGLEC4* are restricted to mature B cells, and glial cells, respectively (Li et al. 1998; Nitschke and Tsubata 2004; Figure 1.3). *SIGLEC2* molecule inhibits B-cell signaling (Nitschke and Tsubata 2004). Quarles (2007) has reported that *SIGLEC4* is able to be implicated in the process of myelination during nerve regeneration. *SIGLEC3* is also highly expressed in human mature monocytes and myeloid precursor cells (Freeman et al. 1995; Figure 1.3). *SIGLEC5* is the first Siglec detected by bioinformatics analyses based on sequence similarity database searches, and it is expressed in several immune cell types including B cells, monocytes, neutrophils, basophils and mast cells. *SIGLEC5* has been shown to function as an inhibitory receptor in the absence of tyrosine phosphorylation, and it has been demonstrated that *SIGLEC5* plays a role in the diagnosis and monitoring of acute myelogenous leukemia

(Connolly et al. 2002; Avril et al. 2005; Figure 1.3). *SIGLEC6* was found to be expressed highly in placental trophoblast cells and lowly in B cells. Also, it was recognized as a leptin binding protein and regulates body weight (Patel et al. 1999; Figure 1.3). Additionally, *SIGLEC7* is an inhibitory receptor expressed at high levels in NK cells, CD8+ T lymphocytes and dendritic cells, but also it is found at low levels in monocytes and granulocytes (Nicoll et al. 1999). Moreover, *SIGLEC7* can to be phosphorylated in tyrosine and recruit SHP-1 (Angata and Varki 2000). *SIGLEC8* has a distinct expression profile on eosinophils, mast cells, as well as on basophiles at low levels (Kikly et al. 2000). Nutku et al. (2005) have reported that *SIGLEC8* induces eosinophil apoptosis. *SIGLEC9* is a close homolog of *SIGLEC7* (80% identity), and it is expressed by T cells, B cells, neutrophils, monocytes, subsets of dendritic cells and NK cells (Foussias et al. 2000; Zhang et al. 2004; Figure 1.3). *SIGLEC10* has low levels expression profile in B cells, dendritic cells, eosinophils, monocytes and subpopulation of NK cells (Li et al. 2001; Munday et al. 2001; Figure 1.3). *SIGLEC10* associates with CD24 in humans and the CD24-*SIGLEC10* association protect the host against lethal response to pathogenic cell death (Chen et al., 2009). *SIGLEC11* is expressed on tissue macrophage, including brain microglia, as well as on monocytes (Figure 1.3). During immune response, *SIGLEC11* interacts with SHP-1 and/or SHP-2, protein tyrosine phosphatases (also known as Src homology region 2 domain containing phosphatase 1 and/or 2), which are known to modulate microglia biology (Angata et al. 2002). *SIGLEC12* has two isoforms. The short isoform, SLGS, is highly expressed in spleen, small intestine and adrenal gland. In contrast, the long isoform, SLGL, shows high levels of expression in spleen, small intestine and bone marrow (Foussias et al. 2001). In humans, *SIGLEC12* has an Arg→Cys mutation in their V-type1 domain, which results in an inactive protein unable to recognize sialic acid (Angata et al. 2001). In

humans, the *SIGLEC13* gene is absent, but it is present in the chimpanzee genome (Angata et al. 2004). *SIGLEC14* is expressed on monocytes and increases respiratory burst in neutrophils. *SIGLEC14* is highly homologous to *SIGLEC5,* with 83% identity in their amino acid sequences, and it interacts with the activating adapter protein DAP12 (Angata et al. 2006). *SIGLEC15* is expressed on macrophages, monocytes, and/or dendritic cells of human spleen and lymph nodes (Angata et al. 2007). *SIGLEC15* interacts with the activating adapter proteins DAP10 and DAP12 via its lysine residue in the transmembrane domain (Angata et al. 2007). Lastly, *SIGLEC16* appears to be expressed on macrophages and normal human brain. It is newly activating human receptor, which is reported to contain functional and non-functional alleles in the UK population (Cao et al. 2008). *SIGLEC16* is reported to have arisen from a *SIGLEC11* by inverse duplication and conversion. The extracellular domain of human *SIGLEC16* share a high percentage (99%) of amino acid identity with extracellular domain of human *SIGLEC11* (Cao et al. 2008).

The family members *SIGLEC5*, *SIGLEC7*, *SIGLEC8*, *SIGLEC9*, *SIGLEC10* and *SIGLEC14* bind glycoconjugates carrying sialic acid in either α-2,3- and α-2,6 linkages. *SIGLEC1* and *SIGLEC4* bind preferentially to α-2,3- linked sialic acid. *SIGLEC11* and *SIGLEC16* have the ability to bind to α-2,8-linked sialic acids. *SIGLEC2*, *SIGLEC3* and *SIGLEC6* bind only to α-2,6-linked sialic acids.

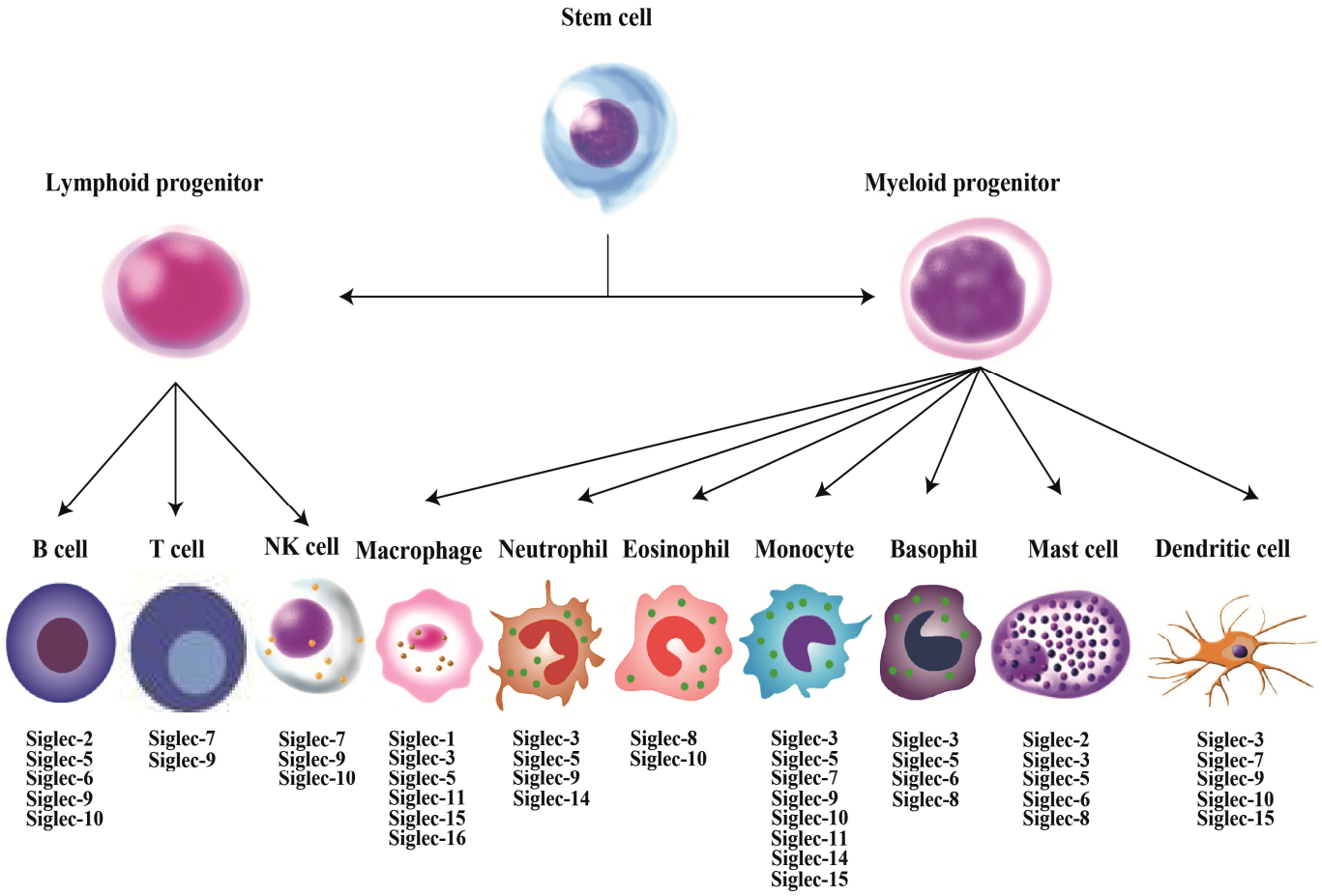


Figure 1.3. Expression pattern of human Siglecs in immune cells. Figure is modified from Cao and Crocker (2011) and Jandus et al. (2011). Abbreviation: NK, natural killer.

2.2. CEA family

2.2.1. Definition and structure of CEA

The human carcinoembryonic antigen (CEA) multigene family is a subgroup of the immunoglobulin superfamily, which consists of 22 related genes and 12 pseudogenes located within a 1.5 Mb cluster on the long arm of chromosome 19, 19q13.1 thought 19q13.4 (Olsen et al. 1994; Naghibalhossaini et al. 2007). The members of this family are highly glycosylated including some proteins linked to the cytoplasmic membrane and others are secretory proteins. Several studies have demonstrated that members of the CEA gene family are subdivided into three subgroups based on sequence similarity, developmental expression patterns and their biological functions: the CEA-related Cell Adhesion Molecule (CEACAM) subgroup containing twelve genes (*CEACAM1*, *CEACAM3-CEACAM8*, *CEACAM16* and *CEACAM18-CEACAM21*), the Pregnancy Specific Glycoprotein (PSG) subgroup containing eleven closely related genes (*PSG1-PSG11*) and a subgroup of eleven pseudogenes (*CEACAMP1-CEACAMP11*) (Zhou et al. 2000; Gray-Owen and Blumberg 2006; Figure 1.4-1.5). Most members of the CEACAM subgroup have similar structures consist of an extracellular Ig-like domains composed of a single N-terminal V-set domain, with structural homology to the immunoglobulin variable domains, followed by varying numbers of C2-set domains of A or B subtypes, a transmembrane domain and a cytoplasmic domain (Beauchemin et al. 1999; Kammerer et al., 2007). There are two members of CEACAM subgroup (*CEACAM16* and *CEACAM20*) that show a few exceptions in the organization of their structures. *CEACAM16* contains two Ig-like V-type domains at its N and C termini and *CEACAM20* contains a truncated Ig-like V-type 1 domain (Figure 1.4). The CEACAM molecules can be anchored to the cell surface via their transmembrane

domains (*CEACAM5* thought *CEACAM8*) or directly linked to glycophosphatidylinositol (GPI) lipid moiety (*CEACAM5*, *CEACAM18* thought *CEACAM21*).

Unlike the CEACAMs, all PSG members are secreted proteins (Beauchemin et al., 1999, Zheng et al., 2011, Figure 1.4-1.5). As shown in Figure 1.4 and Figure 1.5, all known members of the PSG subgroup have a similar domains organization. Each PSG protein contains a single N-terminal V-set domain (N1) domain linked to two or three C2-set domains of A, B and/or C subtypes (Zhou and Hammarström 2001; Figure 1.5). Most PSGs (*PSG2*, *PSG3*, *PSG5*, *PSG6*, *PSG7*, *PSG9*, and *PSG11*) have a conserved motif of Arginine-Glycine-Aspartic acid (RGD) residues, which is present in the N-terminal V-set domains, and they function as an adhesion recognition signal for several integrins, such as fibrinogen and vitronectin (Ruoslahti and Pierschbacher 1987; Teglund et al. 1994).

2.2.2. Expression and biological functions of CEA

Members of CEA family are known to be expressed in diffrent cell types and have a wide rang of biological functions. CEACAMs are found prominently on most epithelial cells and are present on different leucocytes. In humans, *CEACAM1*, the ancestor member of CEA family, is expressed on the apical side of epithelial and endothelial cells as well as on lymphoid and myeloid cells. *CEACAM1* mediates cell-cell adhesion through homophilic (*CEACAM1* to *CEACAM1*) as well as heterophilic (e.g. *CEACAM1* to *CEACAM5*) interactions (Kuespert et al. 2006; Gray-Owen and Blumberg 2006). In addition, *CEACAM1* is involved in many other biological processes, such as angiogenesis, cell migration, and immune functions (Gray-Owen and Blumberg 2006). *CEACAM3* and *CEACAM4* expression is restricted to granulocytes, and they are able to convey uptake and destruction of several bacterial pathogens including *Neisseria*, *Moraxella*, and *Haemophilus* species (Schmitter et al. 2004).

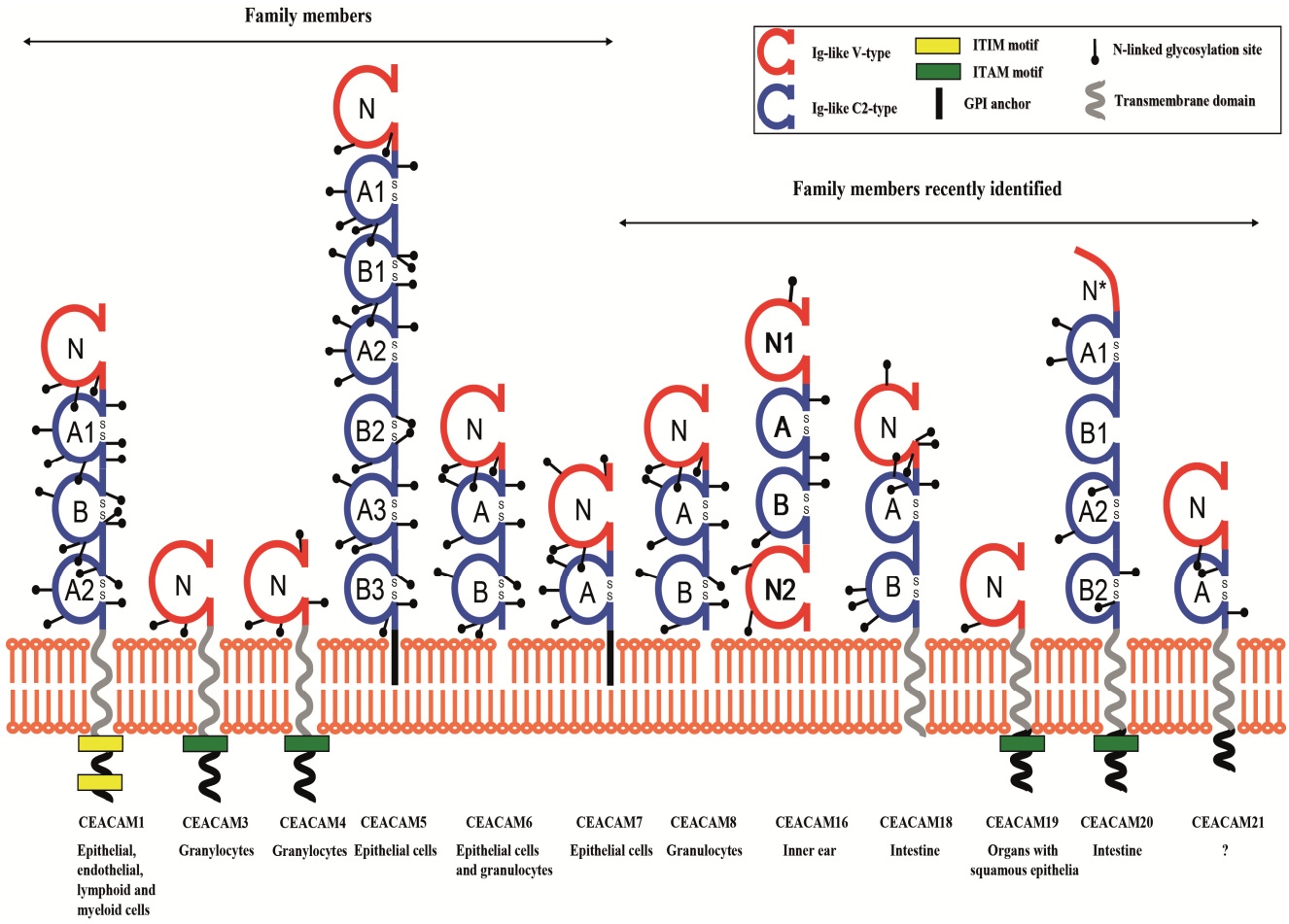


Figure 1.4. Organisation and tissue expression of the human CEACAM genes. Adapted from http://www.carcinoembryonic-antigen.de/.

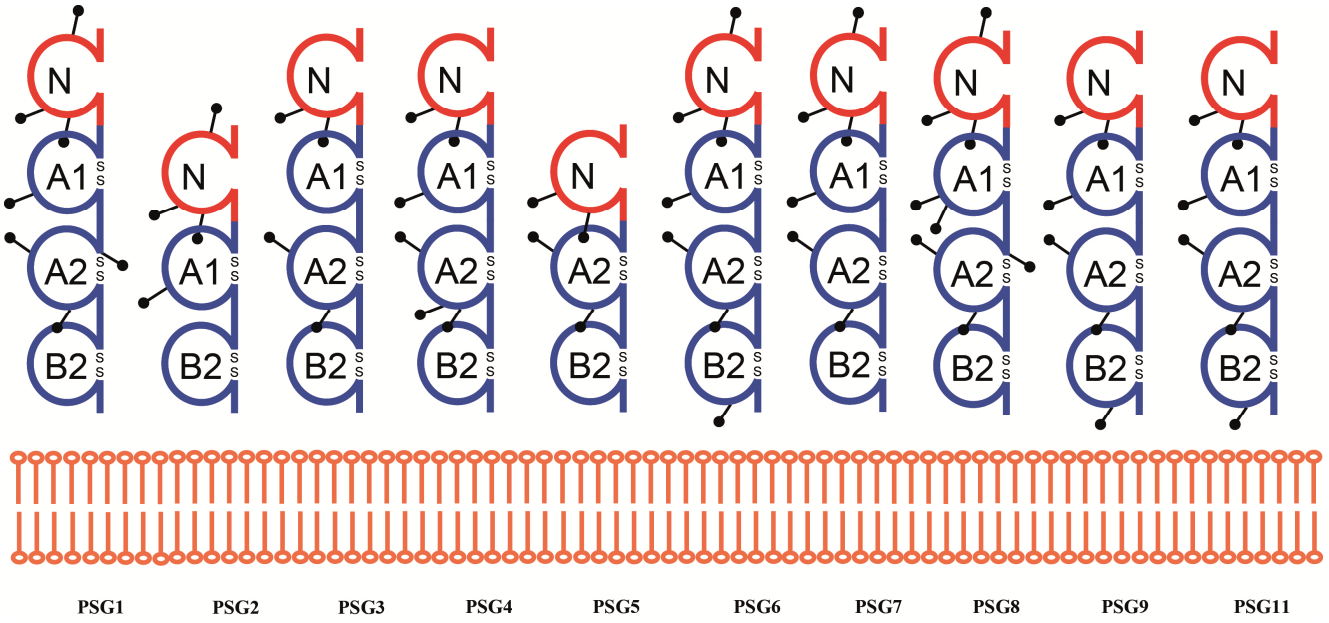


Figure 1.5. Organisation of the human PSG genes. Adapted from http://www.carcinoembryonic-antigen.de/.

CEACAM5 is expressed primarily in epithelial cells, playing a role as a calcium-independent adhesion molecule through homophilic (*CEACAM5* to *CEACAM5*) and heterophilic associations (*CEACAM5* to *CEACAM1* or *CEACAM6*; Zheng et al. 2011). Schölzel et al. (2000) have reported that *CEACAM6* is expressed by granulocytes and epithelia of various organs and play a role in colorectal oncogenesis. Expression of *CEACAM7* has been observed on the epithelium of the colon and on pancreatic ducts (Schölzel et al. 2000). *CEACAM8* expression has been shown to occur in granulocytes (neutrophils and eosinophils) and in acute leukemia (Lasa et al. 2008). Functionally, *CEACAM8* serves as a binding partner for *CEACAM6* and a natural ligand Galectin-3 (Yoon et al. 2007). The recently discovered genes (*CEACAM16*, *CEACAM18* thought *CEACAM21*) have a distinct expression pattern compared to all other CEA family members. *CEACAM16* is expressed only in the inner ear (Kammerer et al. 2012). Zebhauser et al. (2005) have reported that *CEACAM18* and *CEACAM20* are expressed in the small and large intestine and at lower levels in thymus. *CEACAM19* expression is restricted to organs with squamous epithelia including the skin, esophagus, tongue, eye, uterus, and stomach (Zebhauser et al. 2005). The function of these recently discovered genes has not been described.

PSG molecules are mainly expressed in syncytiotrophoblasts (STBs) of primate and rodent placentas and play a critical role in the maintenance of pregnancy (Zhou et al. 1997). Previous studies have shown that PSGs are immuno-modulatory proteins which are able to modulate activity of T-lymphocytes and regulate anti-inflammatory cytokines secretion in human monocytes and macrophages. For example, Snyder et al. (2001) reported the ability of human *PSG1*, *PSG6* and *PSG11* to stimulate the secretion of transforming growth factor beta 1 (TGFβ1), interleukin 6 (IL-6), and interleukin 10 (IL-10) in

macrophages and monocytes. Also, some PGSs such as *PSG11* can bind to the surface of promonocyte cells via its RGD motif (Rutherfurd et al. 1995).

3. Objectives

The first objective of this work was to characterize gene conversion events that occurred in two members of the immunoglobulin superfamily (Siglecs and CEA families) across five primate species (*Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Pongo abelii* and *Nomascus leucogenys*). We first use the GENECONV software to detect gene conversions events. We then examine the length, polarity and frequency of conversions, GC-content and sequence similarity in converted and non converted regions.

The second objective was to assess the impact of gene conversions events on the evolution of these two families.

Literature cited

Angata T, Varki NM, Varki A. 2001. A second uniquely human mutation affecting sialic acid biology. *J Biol Chem.* 27: 40282–40287.

Angata T, Varki A. 2000. Siglec-7: a sialic acid-binding lectin of the immunoglobulin superfamily. *Glycobiology.* 10: 431–138.

Angata T, Kerr SC, Greaves DR, Varki NM, Crocker PR, Varki A. 2002. Cloning and characterization of human Siglec-11. A recently evolved signaling molecule that can interact with SHP-1 and SHP-2 and is expressed by tissue macrophages, including brain microglia. *J Biol Chem.* 277: 24466–22474.

Angata T, Margulies EH, Green ED, Varki A. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci U S A.* 101: 13251–13256.

Angata T, Hayakawa T, Yamanaka M, Varki A, Nakamura M. 2006. Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. *FASEB J.* 20: 1964–1973.

Angata T, Tabuchi Y, Nakamura K, Nakamura M. 2007. Siglec-15: an immune system Siglec conserved throughout vertebrate evolution. *Glycobiology*. 17: 838–846.

Arakawa H, Buerstedde JM. 2004. Immunoglobulin gene conversion: insights from bursal B cells and the DT40 cell line. *Dev Dyn*. 229: 458–464.

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L. 2001. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am J Hum Genet.* 69: 582–589.

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet.* 3: 299–309.

Aricescu AR, Jones EY. 2007. Immunoglobulin superfamily cell adhesion molecules: zippers and signals. *Curr Opin Cell Biol*. 19: 543–550.

Avril T, Freeman SD, Attrill H, Clarke RG, Crocker PR. 2005. Siglec-5 (CD170) can mediate inhibitory signaling in the absence of immunoreceptor tyrosine-based inhibitory motif phosphorylation. *J Biol Chem*. 280: 19843–19851.

Backström N, Ceplitis H, Berlin S, Ellegren H. 2005. Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. *Mol Biol Evol*. 22: 1992–1999.

Barrow AD, Trowsdale J. 2008. The extended human leukocyte receptor complex: diverse ways of modulating immune responses*. Immunol Rev*. 224:98–123.

Beauchemin N, Draber P, Dveksler G, Gold P, Gray-Owen S, Grunert F, Hammarström S, Holmes KV, Karlsson A, Kuroki M, Lin SH, Lucka L, Najjar SM, Neumaier M, Obrink B, Shively JE, Skubitz KM, Stanners CP, Thomas P, Thompson JA, Virji M, von Kleist S, Wagener C, Watt S, Zimmermann W. 1999. Redefined nomenclature for members of the carcinoembryonic antigen family. *Exp Cell Res*. 252: 243–249.

Benovoy D, Morris RT, Morin A, Drouin G. 2005. Ectopic gene conversions increase the G + C content of duplicated yeast and Arabidopsis genes. *Mol Biol Evol*. 22: 1865–1868.

Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics*. 93: 27–32.

Bork P, Holm L, Sander C. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol*. 242: 309–320.

Cao H, Lakner U, de Bono B, Traherne JA, Trowsdale J, Barrow AD. 2008. SIGLEC16 encodes a DAP12-associated receptor expressed in macrophages that evolved from its inhibitory counterpart SIGLEC11 and has functional and non-functional alleles in humans. *Eur J Immunol*. 38: 2303–2315.

Cao H, de Bono B, Belov K, Wong ES, Trowsdale J, Barrow AD. 2009. Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. *Immunogenetics*. 61: 401–417.

Cao H, Crocker PR. 2011. Evolution of CD33-related siglecs: regulating host immune functions and escaping pathogen exploitation? *Immunology*. 132: 18–26.

Chen JF, Lu F, Chen SS, Tao SH. 2006. Significant positive correlation between the recombination rate and GC content in the human pseudoautosomal region. *Genome*. 49: 413–419.

Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8: 762–775.

Chen GY, Tang J, Zheng P, Liu Y. 2009. CD24 and Siglec-10 selectively repress tissue damage-induced immune responses. *Science*. 323:1722–1725.

Cheng SF, Hu YH, Sun BG, Zhang M, Chi H, Sun L. 2012. A single immunoglobulin-domain IgSF protein from Sciaenops ocellatus regulates pathogen-induced immune response in a negative manner. *Dev Comp Immunol*. 38: 117–127.

Connolly NP, Jones M, Watt SM. 2002. Human Siglec-5: tissue distribution, novel isoforms and domain specificities for sialic acid-dependent ligand interactions. Br. J. Haematol. 119: 221–238.

Crocker PR. 2005. Siglecs in innate immunity. *Curr Opin Pharmacol*. 5:431-437.

Crocker PR, Paulson JC, Varki A. 2007. Siglecs and their roles in the immune system. *Nat Rev Immunol*. 7: 255–266.

Crocker PR, Redelinghuys P. 2008. Siglecs as positive and negative regulators of the immune system. *Biochem Soc Trans*. 36: 1467–1471.

Dermody TS, Kirchner E, Guglielmi KM, Stehle T. 2009. Immunoglobulin superfamily virus receptors and the evolution of adaptive immunity. *PLoS Pathog*. 5: e1000481.

Drouin G, Prat F, Ell M, Clarke GD. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol*. 16: 1369–1390.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. *J Mol Evol*. 55:14–23.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 10: 285–311.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics*.175: 477–485.

Ezawa K, OOta S, Saitou N. 2006. Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol Biol Evol*. 23: 927–940.

Fitch DH, Goodman M. 1991. Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversions and other recombinational events. *Comput Appl Biosci*. 7: 207–215.

Foussias G, Yousef GM, Diamandis EP. 2000. Identification and molecular characterization of a novel member of the siglec family (SIGLEC9). *Genomics*. 67: 171–178.

Foussias G, Taylor SM, Yousef GM, Tropak MB, Ordon MH, Diamandis EP. 2001. Cloning and molecular characterization of two splice variants of a new putative member of the Siglec-3-like subgroup of Siglecs. *Biochem Biophys Res Commun*. 284: 887–899.

Freeman SD, Kelm S, Barber EK, Crocker PR. 1995. Characterization of CD33 as a new member of the sialoadhesin family of cellular interaction molecules. *Blood*. 85: 2005–2012.

Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U.S.A*. 97: 11383–11390.

Gray-Owen SD, Blumberg RS. 2006. CEACAM1: contact-dependent control of immunity. *Nat Rev Immunol*. 6: 433–446.

Hallast P, Nagirnaja L, Margus T, Laan M. 2005. Segmental duplications and gene conversion: human luteinizing hormone/chorionic gonadotropin beta-gene cluster. *Genome Research*.15: 1535–1546.

Hartnell A, Steel J, Turley H, Jones M, Jackson DG, Crocker PR. 2001. Characterization of human sialoadhesin, a sialic acid binding receptor expressed by resident and inflammatory macrophage populations. *Blood*. 97: 288–296.

Hein J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol*. 36: 396–405.

Holness CL, Simmons DL. 1994. Structural motifs for recognition and adhesion in members of the immunoglobulin superfamily. *J Cell Sci*. 107: 2065–2070.

Hurst LD, Brunton CF, Smith NG. 1999. Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends Genet*. 15: 437–439.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*. 101: 13994–14001.

Iwase M, Satta Y, Hirai H, Hirai Y, Takahata N. 2010. Frequent gene conversion events between the X and Y homologous chromosomal regions in primates. *BMC Evolutionary Biology*.10: 225.

Inoue N, Ikawa M, Isotani A, Okabe M. 2005. The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs. *Nature*. 434: 234-238.

Jandus C, Simon HU, von Gunten S. 2011. Targeting siglecs-a novel pharmacological strategy for immuno- and glycotherapy. *Biochem Pharmacol*. 82: 323–332.

Juliano RL. 2002. Signal transduction by cell adhesion receptors and the cytoskeleton: functions of integrins, cadherins, selectins, and immunoglobulin-superfamily members. *Annu Rev Pharmacol Toxicol*. 42: 283–323.

Kammerer R, Popp T, Härtle S, Singer BB, Zimmermann W. 2007. Species-specific evolution of immune receptor tyrosine based activation motif-containing CEACAM1-related immune receptors in the dog. *BMC Evol Biol.* 7: 196.

Kammerer R, Rüttiger L, Riesenberg R, Schäuble C, Krupar R, Kamp A, Sunami K, Eisenried A, Hennenberg M, Grunert F, Bress A, Battaglia S, Schrewe H, Knipper M, Schneider MR, Zimmermann W. 2012. Loss of mammal-specific tectorial membrane component carcinoembryonic antigen cell adhesion molecule 16 (CEACAM16) leads to hearing impairment at low and high frequencies. *J Biol Chem*. 287: 21584–21598.

Kee Y, D'Andrea AD. 2010. Expanded roles of the Fanconi anemia pathway in preserving genomic stability. *Genes Dev*.24: 1680–1694.

Kikly KK, Bochner BS, Freeman SD, Tan KB, Gallagher KT, D'alessio KJ, Holmes SD, Abrahamson JA, Erickson-Miller CL, Murdock PR, Tachimoto H, Schleimer RP, White JR. 2000. Identification of SAF-2, a novel siglec expressed on eosinophils, mast cells and basophils. *J Allergy Clin Immunol*. 105: 1093–1100.

Kuespert K, Pils S, Hauck CR. 2006. CEACAMs: their role in physiology and pathophysiology. *Curr Opin Cell Biol*. 18: 565–571.

Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM. 2000. Linkage disequilibria and the site frequency spectra in the su (s) and su (wa) regions of the Drosophila melanogaster X chromosome. *Genetics*. 156: 1837–1852.

Lasa A, Serrano E, Carricondo M, Carnicer MJ, Brunet S, Badell I, Sierra J, Aventín A, Nomdedéu JF. 2008. High expression of CEACAM6 and CEACAM8 mRNA in acute lymphoblastic leukemias. *Ann Hematol*. 87: 205–211.

Lawson MJ, Jiao J, Fan W, Zhang L. 2009. A Pattern Analysis of Gene Conversion Literature. Comparative and Functional Genomics 2009. 11 pages.

Li C, Trapp B, Ludwin S, Peterson A, Roder J. 1998. Myelin associated glycoprotein modulates glia-axon contact in vivo. *J. Neurosci. Res.* 51: 210–217.

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

Liao D. 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol*. 51: 305–317.

Lucier MR, Thompson RE, Waire J, Lin AW, Osborne BA, Goldsby RA. 1998. Multiple sites of VL diversification in cattle. *J Immunol*. 161: 5438–5444.

McGuire G, Wright F, Prentice MJ. 1997. A graphical method for detecting recombination in phylogenetic data sets. *Mol Biol Evol*. 14: 1125–1131.

McGuire G, Wright F, Prentice MJ. 2000. A Bayesian model for detecting past recombination events in DNA multiple alignments. *J Comput Biol*. 7: 159-170.

Meunier J, Duret L. 2004. Recombination drives the Evolution of GC-content in the human genome. *Mol Biol Evol*. 21: 984–990.

Munday J, Kerr S, Ni J, Cornish AL, Zhang JQ, Nicoll G, Floyd H, Mattei MG, Moore P, Liu D, Crocker PR. 2001. Identification, characterization and leucocyte expression of Siglec-10, a novel human sialic acid-binding receptor. *Biochem J*. 355: 489–497.

Naghibalhossaini F, Yoder AD, Tobi M, Stanners CP. 2007. Evolution of a tumorigenic property conferred by glycophosphatidyl-inositol membrane anchors of carcinoembryonic antigen gene family members during the primate radiation. *Mol Biol Cell.* 18: 1366–1374.

Nicoll G, Ni J, Liu D, Klenerman P, Munday J, Dubock S, Mattei MG, Crocker PR. 1999. Identification and characterization of a novel siglec, siglec-7, expressed by human natural killer cells and monocytes. *J Biol Chem*. 274: 34089–34095.

Nitschke L, Tsubata T. 2004. Molecular interactions regulate BCR signal inhibition by CD22 and CD72. *Trends Immunol*. 25: 543–550.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*. 14: 354–366.

Nutku E, Hudson SA, Bochner BS. 2005. Mechanism of Siglec-8-induced human eosinophil apoptosis: role of caspases and mitochondrial injury. *Biochem Biophys Res Commun*. 336: 918–924.

Olsen A, Teglund S, Nelson D, Gordon L, Copeland A, Georgescu A, Carrano A, Hammarström S. 1994. Gene organization of the pregnancy-specific glycoprotein region on human chromosome 19: assembly and analysis of a 700-kb cosmid contig spanning the region. *Genomics*. 23: 659–668.

Patel N, Brinkman-Van der Linden EC, Altmann SW, Gish K, Balasubramanian S, Timans JC, Peterson D, Bell MP, Bazan JF, Varki A, Kastelein RA. 1999. OBBP1/ Siglec-6. A leptin- and sialic acid-binding protein of the immunoglobulin superfamily. *J Biol Chem*. 274: 22729–22738.

Perry J, Ashworth A. 1999. Evolutionary rate of a gene affected by chromosomal position. *Curr Biol*. 9: 987–989.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol*. 4: 675–682.

Petes TD, Hill CW. 1988. Recombination between repeated genes in microorganisms. *Annu Rev Genet*. 22:147–168.

Pillai S, Netravali IA, Cariappa A, Mattoo H. 2012. Siglecs and immune regulation. *Annu Rev Immunol*. 30: 357–392.

Quarles RH. 2007. Myelin-associated glycoprotein (MAG): past, present and beyond. *J Neurochem*. 100: 1431–1448.

Radding C. 1978. Genetic recombination: strand transfer and mismatch repair. *Annu Rev Biochem*. 47: 847–880.

Reynaud CA, Mackay CR, Muller RG, Weill JC. 1991. Somatic generation of diversity in a mammalian primary lymphoid organ: the sheep ileal Peyer's patches. *Cell*. 64: 995.

Rooney AP. 2004. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans. *Mol Biol Evol*. 21: 1704–1711.

Rutherfurd KJ, Chou JY, Mansfield BC. 1995. A motif in PSG11s mediates binding to a receptor on the surface of the promonocyte cell line THP-1. *Mol Endocrinol*. 9:1297–12305.

Rougon G, Hobert O. 2003. New insights into the diversity and function of neuronal immunoglobulin superfamily molecules. *Annu Rev Neurosci*. 26: 207–238.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*. 423: 873–876.

Santoyo G, Romero D. 2005. Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol Rev*. 29: 169–183.

Sawatzki H, Krawczak M, Cooper DN. 2009. A gene conversion hotspot in the human growth hormone (GH1) gene promoter. *Hum Mutat*. 30: 239–247.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6: 526–538.

Schmitter T, Agerer F, Peterson L, Munzner P, Hauck CR. 2004. Granulocyte CEACAM3 is a phagocytic receptor of the innate immune system that mediates recognition and elimination of human-specific pathogens. *J Exp Med*. 199: 35–46.

Schölzel S, Zimmermann W, Schwarzkopf G, Grunert F, Rogaczewski B, Thompson J. 2000. Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are

differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. *Am J Pathol*. 156: 595–605.

Scott AF, Heath P, Trusko S, Boyer SH, Prass W, Goodman M, Czelusniak J, Chang LY, Slightom JL, Blechl AE, Smithies O. 1980. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell*. 2: 627–638.

Semple C, Wolfe KH. 1999. Gene duplication and gene conversion in the Caenorhabditis elegans genome. *J Mol Evol*. 48: 555–564.

Singh ND, Davis JC, Petrov DA. 2005. Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. *J Mol Evol*. 61: 315–324.

Slightom JL. 1984. The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. *Mol Biol Evol*. 1: 371–389.

Snyder SK, Wessner DH, Wessells JL, Waterhouse RM, Wahl LM, Zimmermann W, Dveksler GS. 2001. Pregnancy-specific glycoproteins function as immunomodulators by inducing secretion of IL-10, IL-6 and TGF-beta1 by human monocytes. *Am J Reprod Immunol*. 45: 205–216.

Stephens J. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol Biol Evol*. 2: 539–556.

Steffen BJ, Breier G, Butcher EC, Schulz M, Engelhardt B. 1996. ICAM-1, VCAM-1, and MAdCAM-1 are expressed on choroid plexus epithelium but not endothelium and mediate binding of lymphocytes in vitro. *Am J Pathol*. 148: 1819–1838.

Strachan T, Read AP. 1999. Human Molecular Genetics. 2nd edition. New York: Wiley-Lis. 576 pages.

Sun Y, Liu Z, Ren L, Wei Z, Wang P, Li N, Zhao Y. 2012. Immunoglobulin genes and diversity: what we have learned from domestic animals. *J Anim Sci Biotechnol*. 3: 18.

Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. 1983. The double-strand-break repair model for recombination. *Cell*. 33: 25–35.

Teglund S, Olsen A, Khan WN, Frängsmyr L, Hammarström S. 1994. The pregnancy-specific glycoprotein (PSG) gene cluster on human chromosome 19: fine structure of the 11 PSG genes and identification of 6 new genes forming a third subgroup within the carcinoembryonic antigen (CEA) family. *Genomics*. 23: 669–684.

Toshimori K, Maekawa M, Ito C, Toyama Y, Suzuki-Toyota F, Saxena D. 2006. The involvement of immunoglobulin superfamily proteins in spermatogenesis and sperm-egg interaction. *Reprod Med Biol*. 5: 87–93.

Varki A, Angata T. 2006. Siglecs–the major sub-family of I-type lectins. *Glycobiology*. 16: 1R–27R.

Varki A. 2009. Natural ligands for CD33-related Siglecs? *Glycobiology* 19:810-812.

Wai Wong C, Dye DE, Coombe DR. 2012. The role of immunoglobulin superfamily cell adhesion molecules in cancer metastasis. *Int J Cell Biol*. 2012: 340296.

Weiller G. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol Biol Evol*. 15: 326–335.

Weinstein PD, Mage RG, Anderson AO. 1994. The appendix functions as a mammalian bursal equivalent in the developing rabbit. *Adv Exp Med Biol*. 355: 249–253.

Wiuf C, Christensen T, Hein J. 2001. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*. 18: 1929–1939.

Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK, Wang J, Zheng X. 2008. Gene conversion in the rice genome. *BMC genomics*. 9: 93–100.

Yamagata M, Weiner J, Sanes J. 2002. Sidekicks. Synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell*. 110: 649–660.

Yoon J, Terada A, Kita H. 2007. CD66b regulates adhesion and activation of human eosinophils. *J Immunol*. 179: 8454–8462.

Yousef GM, Ordon MH, Foussias G, Diamandis EP. 2002. Genomic organization of the siglec gene locus on chromosome 19q13.4 and cloning of two new siglec pseudogenes. *Gene*. 286: 259–270.

Zebhauser R, Kammerer R, Eisenried A, McLellan A, Moore T, Zimmermann W. 2005. Identification of a novel group of evolutionarily conserved members within the rapidly diverging murine Cea family. *Genomics*. 86: 566–580.

Zhang JQ, Biedermann B, Nitschke L, Crocker PR. 2004. The murine inhibitory receptor mSiglec-E is expressed broadly on cells of the innate immune system whereas mSiglec-F is restricted to eosinophils. *Eur J Immunol*. 34: 1175–1184.

Zheng C, Feng J, Lu D, Wang P, Xing S, Coll JL, Yang D, Yan X. 2011. A novel anti-CEACAM5 monoclonal antibody, CC4, suppresses colorectal tumor growth and enhances NK cells-mediated tumor immunity. *PLoS One*. 6: e21146.

Zheng J, Miller KK, Yang T, Hildebrand MS, Shearer AE, DeLuca AP, Scheetz TE, Drummond J, Scherer SE, Legan PK, Goodyear RJ, Richardson GP, Cheatham MA, Smith RJ, Dallos P. 2011. Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with α-tectorin and is mutated in autosomal dominant hearing loss (DFNA4). *Proc Natl Acad Sci*.108: 4218–4223.

Zhou GQ, Baranov V, Zimmermann W, Grunert F, Erhard B, Mincheva-Nilsson L, Hammarström S, Thompson J. 1997. Highly specific monoclonal antibody demonstrates that pregnancy-specific glycoprotein (PSG) is limited to syncytiotrophoblast in human early and term placenta. *Placenta*. 18: 491–501.

Zhou GQ, Zhang Y, Hammarström S. 2000. The carcinoembryonic antigen (CEA) gene family in non-human primates. *Gene*. 264: 105–112.

Zhou GQ, Hammarström S. 2001. Pregnancy-specific glycoprotein (PSG) in baboon (Papio hamadryas): family size, domain structure, and prediction of a functional region in primate PSGs. *Biol Reprod*. 64: 990–999.

Chapter 2. Gene conversions are frequent but not under positive selection in the Siglec gene families of primates

Abstract

Siglecs are a family of cell surface proteins which bind sialic acids and belong to the immunoglobulin superfamily. They are thought to be involved in cell-cell interactions and signalling functions such as self-recognition. They are composed of two groups, the conserved Siglecs and the CD33-related Siglecs. The four genes of the first group are present in all mammalian species, but the number of members of the second group varies between mammalian species. Previous studies have reported the occurrence of gene conversions between human CD33-related Siglecs and suggested that these conversions are adaptive because they increase the diversity of these immunoglobulin-related genes. Here, we analyze the Siglec genes of five primate species (*Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Pongo abelii* and *Nomascus leucogenys*) and show that gene conversions are frequent between primate CD33-related Siglecs. However, although gene conversions only occur in the extracellular region of primate CD33-related Siglecs, they also only occur between closely related genes and they occur equally frequently in sialic acid binding and non-binding domains. Our results therefore suggest that the high frequency of gene conversions between the extracellular regions of CD33-related Siglec genes is simply a consequence of their high degree of sequence similarity and is not the result of positive or diversifying selection.

2.1. Introduction

Sialic acids binding immunoglobulin-like lectins (Siglecs) are a family of cell surface proteins belonging to the immunoglobulin superfamily. They are thought to be involved in cell-cell interactions and signalling functions such as self-recognition (Crocker 2002; Varki 2010). They are characterized by the presence of one or two N-terminal V-like immunoglobulin domain followed by varying numbers of C2-set domains, a transmembrane domain and a cytoplasmic domain (Crocker et al. 2007; Varki 2009; fig. 2.1). They are subdivided into two groups based upon sequence similarity and evolutionary relatedness. The first group is called the conserved Siglecs because they are found in rodents and primates. It is composed of four genes: *SIGLEC1* (sialoadhesin, located on human chromosome 20p13), *SIGLEC2* (CD22, located on human chromosome 19q13.1), *SIGLEC4* (myelinassociated glycoprotein (MAG), located on human chromosome 19q13.1) and *SIGLEC15* (located on human chromosome 18q12.3, Figure 2.1, Angata et al. 2004). Sialoadhesin is restricted to macrophages, CD22 to B cells and Mast cells, MAG to glial cells, and *SIGLEC15* to macrophages, monocytes and dendritic cells (Figure 2.1). Members of this group are quite distantly related, sharing only about 25-30% sequence similarity in their protein sequences (Crocker 2002; Supplemental Table 2.1). The second group is called the CD33-related Siglecs (CD33rSiglecs) because they are similar to *SIGLEC3* (CD33). Members of this group are not conserved in all species (Angata et al. 2004). In humans, eleven hCD33rSiglec genes have been characterized. They are clustered on chromosome 19q13.3-13.4 and include *SIGLEC3*, *SIGLEC5* to *SIGLEC12*, *SIGLEC14* and *SIGLEC16* (Figure 2.2). In comparison, there are 12 CD33rSiglec genes in the chimpanzee genome and only 8 in the rhesus monkey genome (Figure 2.2). Furthermore, *SIGLEC13* is present in the chimpanzee genome but not in the human genome (Angata et al. 2004).

SIGLEC12, *SIGLEC14* and *SIGLEC16* are partially pseudogenized in the human population (Angata et al. 2004; Cao et al. 2008). The protein sequences of human CD33rSiglecs share about 41-88% identity, with the greatest similarity being in their extracellular region (Figure 2.3 and Supplemental Table 2.1).

The CD33rSiglecs are widely distributed throughout in various cell types in the immune system of primates (Angata et al. 2004; Figure 2.1). *SIGLEC3* and *SIGLEC5* through *SIGLEC11* are expressed on monocytes, NK cells, granulocytes cells, Mast cells, dendritic cells, T cells, B cells, and macrophages (Crocker et al. 2007). *SIGLEC14*, *SIGLEC15* and *SIGLEC16* have been discovered only recently, and their functions are poorly understood. In humans, *SIGLEC12* has lost its lectin activity (Jandus et al. 2007).

Gene conversions are the non-reciprocal exchange of genetic information between two genes. They have been observed is a wide variety of eukaryotes, from yeasts to mammals, and usually occur between sequences sharing a high degree of sequence similarity (Benovoy and Drouin 2009; McGrath et al. 2009; Chen et al. 2010). Although some gene conversions are known to be responsible for some human genetic diseases, most deleterious gene conversions are eliminated by purifying selection (Chen et al. 2010; Petronella and Drouin 2011; and references therein). Therefore, most gene conversions are selectively neutral, i.e., they are not eliminated by purifying selection because they represent selectively neutral events.

Previous studies have shown that gene conversions occur between some human CD33rSiglec genes. For example, Hawakawa et al. (2005) showed that a 2 kb-long region, including exons 1 to 5, of the human *SIGLEC11* gene was converted by the human *SIGLEC16* pseudogene and these authors argued that this gene conversion was evolutionarily significant. Here, in order to address the evolutionary significance of gene

conversions among Siglec genes, we studied the presence, location and characteristics of the gene conversions that occurred between the Siglec genes of five primate species (human, chimpanzee, Sumatran orang-utan, Northern white-cheeked gibbon and rhesus monkey). We found 33 gene conversion events between CD33rSiglec genes but none between conserved Siglec genes. Furthermore, we found that all the conversions occurring in the coding regions of the CD33rSiglec genes occur only in the extracellular domain of these proteins. However, they occur equally frequently in the sialic acid binding and non-binding domains of the proteins encoded by these genes and the length of the conversions is strongly correlated with sequence similarity. Our results therefore suggest that the high frequency of gene conversions between CD33-related Siglec genes is simply a consequence of their high degree of sequence similarity and is not the result of positive or diversifying selection.

2.2. Materials and methods

Gene sequences and their analyses

Sequences of five primate species were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) and Ensembl (http://www.ensembl.org/). The list of sequences used, and their accession numbers, is shown in Supplemental Table 2.2. Sequence alignments were performed using MUSCLE v3.8.31 (Edgar 2004) and then verified and refined using BioEdit v7.0.5.3 (Hall 1999). Sequence similarities were calculated using MEGA5 (Tamura et al. 2011). GC-contents were calculated using Seqool (http://www.biossc.de). Standard deviations and *t*-test were calculated using Excel 2010. Spearman rank correlations tests were calculated using R version 2.15 (R Development Core Team 2006).

Detection of gene-conversion events

The GENECONV v.1.81 (Sawyer 1989) computer program was used to identify the gene conversions. This method was chosen because it has one of the highest probabilities of correctly inferring gene conversions when they are present (Posada and Crandall 2001). The GENECONV program computes global and pairwise *p*-values and allows mismatches within converted regions. Global and pairwise *p*-values are calculated using two methods. The first method is based on 10,000 permutations of the original data, and the second is based on a method similar to that used by the BLAST database searching algorithm. Here, we only used *p*-values from permutations because they are more conservative and accurate. We also only considered *p*-values ($p < 0.05$) from global inner fragments because their *p*-values are corrected for multiple comparisons whereas the *p*-values of pairwise fragments are not corrected for multiple comparisons. Analyses were

performed using the “g2” parameter, to allow mismatches within converted fragments and to take into account that substitutions do occur after conversions have occurred (Drouin 2002).

The directions of the conversions, i.e., from the donor gene to the acceptor gene, were determined based on the patterns of nucleotide variation inside and outside of the converted regions when compared to non-converted sequences from closely related species. For example, if there was a gene conversion from nucleotides 4 to 7 between genes 1 and 2 of species A, with sequences CATCGGCTGT and AGCCGGCTAG, respectively, and that a related gene from a closely related species has the sequence CATCGGCTGT, or AGCTATATAG, then it is gene 1 of species A that converted gene 2 of species A and not the reverse.

Phylogenetic analyses

Phylogenetic trees were built using the maximum likelihood method implemented in the PhyML v3.0 program, using the Blosum62 model of amino acid substitutions (Guindon and Gascuel 2003). Trees were visualized using the TreeView program (Page 1996).

Tests of selection

In order to measure the selective pressures acting on different gene regions, the number of nonsynonymous substitutions per non-synonymous site (dN) and the number of synonymous substitutions per synonymous site (dS) of Ig-like V-type 1 and Ig-like C2-type1 domains were calculated using the maximum likelihood method implemented in the codeml program of the PAML software package version 4.5 (Yang 2007). These values were calculated using the options seqtype = 1, runmode = -2, CodonFreq = 2 and fix_omega = 0

in the codeml.ctl files of this package. We calculated dN and dS values between human, chimpanzee and Sumatran orang-utan sequences because these three genomes share the largest number of orthologous genes and that such orthologous genes are necessary to obtain meaningful dN and dS values. We tested whether the dN/dS ratios we obtained were significantly different from 1 by calculating the likelihood of these ratios being equal to 1 and performing a likelihood ratio test against a χ^2 distribution with one degree of freedom. The values used for these likelihood ratio tests were calculated as twice the difference between the likelihood of the calculated dN/dS ratios and the likelihood of these ratios being equal to 1.

2.3. Results

2.3.1. Number, lengths and polarity of gene conversions

We detected 33 significant gene conversion events in the five species we analyzed: 11 in human, 9 in chimpanzee, 4 in Sumatran orang-utan, 5 in Northern white-cheeked gibbon, and 4 in rhesus monkey (Table 2.1). All of these conversions are between CD33rSiglec genes and none are between conserved Siglec genes. A few of these conversions are found in more than one species. For example, a conversion of about 200 bp is found in the region coding for the C2-type 1 domain of the human, chimpanzee and rhesus monkey *SIGLEC3* and *SIGLEC6* genes. In all three cases, *SIGLEC3* converted *SIGLEC6* (Table 2.1 and Figure 2.2). Furthermore, a conversion ranging from 503 bp to 1060 bp, and starting in exon 1, is found between the *SIGLEC5* and *SIGLEC14* genes of all five species (Table 2.1). In all cases, *SIGLEC14* converted *SIGLEC5* (Table 2.1 and Figure 2.2).

The average length (±standard deviation) of all conversions is 590.21 ± 429.34 nucleotides and they range from 100 to 2286 nucleotides long. Of these 33 gene conversions, 12 started at an exon and ended at another exon (average size = 822.17 ± 671.55 nucleotides), 7 started at an intron and ended at another intron (average size = 185.43 ± 53.79 nucleotides) and 14 started at an intron and ended at exon, or vice versa (average size = 593.78 ± 279.07 nucleotides).

As shown in Figure 2.1, of the eleven CD33rSiglec genes involved in gene conversion events, 5 are donor only (*SIGLEC3*, *SIGLEC8*, *SIGLEC10*, *SIGLEC12* and *SIGLEC14*), 2 are acceptor only (*SIGLEC6* and *SIGLEC7*) and 5 are both acceptor and donor (*SIGLEC5*, *SIGLEC9*, *SIGLEC11*, *SIGLEC16*, and *SIGLEC16P*). The *SIGLEC13*

gene is only found in the chimpanzee genome and it was never involved in gene conversions. Gene conversions occur between both adjacent genes and distant genes (Figure 2.2). In fact, in humans, there is no correlation between the distance between genes and the frequency of conversions (Spearman's rank correlation test, ρ = -0.16, *p*-value = 0.23). The same is true for the other four species (results not shown). However, there is a weak, but significant, correlation between the distance between genes and the frequency of conversions when considering the data from all species (Spearman's rank correlation test, ρ = -0.15, *p*-value = 0.03). The proximity of genes therefore only explains some 2% of the variation in gene conversion frequency.

2.3.2. Sequence similarities and GC-content of converted and non-converted regions

For the primate CD33rSiglec genes, there is a strong and significant correlation between the length of the converted regions and the sequence similarity of the converted regions (r = 0.52, *p* = 0.002). Furthermore, the average similarity (± standard deviation) between the converted regions (94.85% ± 3.27) is also significantly higher than the average similarity between the non-converted regions (61.09% ± 13.74; *t*-test, $p = 8.68 \times 10^{-16}$). Moreover, there is also a significant difference in GC-content between converted and non-converted regions. The average GC-content (± standard deviation) of the converted regions is 61.4% ± 3.18 whereas that of the non-converted regions is 50% ± 4.17 (*t*-test, $p = 2.08 \times 10^{-18}$).

2.3.3. Biased distribution of converted regions

Except for *SIGLEC12*, all CD33rSiglec genes contain a single Ig-like V-type 1 domain followed by 1 to 16 Ig-like C2-type domains (Figure 2.1). These two types of

domains constitute the extracellular portion of these proteins, the rest being made up of a transmembrane and a cytoplasmic domain. All of the conversions affecting the protein coding regions are only found in the gene regions coding for the extracellular part of these proteins (Table 2.1 and Figure 2.4). This suggests that the conversions which are present in the extracellular part of these proteins are not deleterious. The absence of conversions in the regions coding for the transmembrane and cytoplasmic domains of these proteins is likely due to the low degree of sequence similarity in these regions (see below).

2.3.4. Tests of selection

Table 2.2 shows the dN and dS values in the Ig-like V-type 1 domain and Ig-like C2-type 1 domain in the conserved Siglec genes and in the CD33rSiglecs genes. The unusually large (more than 15 substitutions per sites) dS values observed for the comparisons involving *SIGLEC4* and *SIGLEC7* chimpanzee sequences are due to the fact that the chimpanzee sequences are very different from the human and orang-utan sequences. Similarly, the unusually large (more than 10 substitutions per site) dN and dS values observed for the comparisons involving the orang-utan *SIGLEC10* sequence results from the fact that this sequence is very different from the human and chimpanzee sequences.

All of the significant dN/dS ratios of the Ig-like V-type 1 domain of conserved Siglecs are smaller than 1. Although two of these ratios are very large (138 for the *SIGLEC2* human-chimpanzee comparison and 68 for the *SIGLEC15* human-chimpanzee comparison) these are simply the result of the very low numbers of synonymous substitutions between these two pairs of sequences and these ratios are not significantly greater than 1 (Table 2.2). Similarly, all of the significant dN/dS ratios of the Ig-like C2-type

1 domains are smaller than 1. This suggests that, in conserved Siglecs, both of these domains evolve either neutrally (when dN/dS ratios are not significantly different from 1) or by purifying (negative) selection (when dN/dS ratios are significantly smaller than 1). Similarly, in CD33rSiglecs, except for the Ig-like V-type 1 domain of the *SIGLEC9* human-orang-outan comparison, all of the significant dN/dS ratios of the Ig-like V-type 1 and the Ig-like C2-type 1 domains are smaller than 1 (Table 2.2). This suggests that, in this sub-family, only the Ig-like V-type 1 domain of *SIGLEC9* is evolving under positive selection. All other domains evolve either neutrally or under purifying selection.

2.4. Discussion

Our results show that gene conversions are frequent between CD33rSiglecs genes but they are not observed between conserved Siglec genes (Table 2.1). This suggests that gene conversions do not occur between the conserved genes or that, if they do occur, they are removed by purifying selection. Of these two possibilities, the former is likely responsible for the absence of gene conversions between conserved Siglec genes. As can be seen in Figure 2.3, the sequences of the conserved Siglecs are much more divergent from one another than those of CD33rSiglecs. In fact, at the protein level, the conserved Siglecs share only an average of 21% similarity between one another (Supplemental Table 2.1). At the DNA level, the conserved Siglecs share only an average of 53% similarity between one another (Figure 2.3). Since gene conversions are more frequent between more similar genes, and that gene conversions are very infrequent between genes sharing less than 80% nucleotide sequence similarity, the low similarity between conserved Siglec genes, and between conserved Siglec and CD33rSiglec genes, is likely responsible for the absence of gene conversions in conserved Siglec genes (Benovoy et al. 2005, Chen et al. 2007, Benovoy and Drouin 2009). This suggestion is also consistent with the pattern of gene conversions observed between CD33rSiglec genes where gene conversions occur almost exclusively between related genes. For example, in humans, the eight conversions between CD33rSiglec genes occur exclusively between closely related genes, i.e., the *SIGLEC7* and *SIGLEC9* genes, the *SIGLEC3* and *SIGLEC6* genes, the *SIGLEC5* and *SIGLEC14* genes and the *SIGLEC10*, *SIGLEC11* and *SIGLEC16* genes (Table 2.1 and Figure 2.3). Apart from the 199 bp conversion between the *SIGLEC5* and *SIGLEC6* genes of chimpanzee, the same pattern is observed for the gene conversions between chimpanzee genes; gene conversions occur exclusively between closely related

genes. Therefore, gene conversion are limited to very similar genes which a part of the same CD33rSiglec subfamilies. This is supported by the fact that there is a strong positive correlation between the length of the conversions and the similarity of the converted regions ($r = 0.52$, $p = 0.002$).

Another result is that, in CD33rSiglec genes, all gene conversions that span exon sequences are found exclusively in the extracellular region of the protein coded by these genes, and never in their transmembrane or cytoplasmic regions (Table 2.1 and Figure 2.4). Again, this bias is most readily explained by the fact that gene conversions only occur between sequences sharing at least 80% similarity (see above). Since the similarity (and standard deviation) of the transmembrane and cytoplasmic regions (35.7% ± 19.6%) is significantly smaller than that of extracellular regions (56.4% ± 13.4%, *t*-test, $p = 4.2 \times 10^{-9}$), gene conversions are unlikely to occur in transmembrane and cytoplasmic regions (Supplemental Table 2.1). Furthermore, the frequent conversions between the extracellular region of the *SIGLEC5* and *SIGLEC14* genes, and the long conversions between the extracellular regions of the *SIGLEC11* and *SIGLEC16* genes, are likely simply the result of the fact that the extracellular regions of these two pairs of genes are 96% identical at the protein level, and 95% similar at the DNA level (Table 2.1, Figure 2.3, Supplemental Table 2.1). The effect of sequence similarity on the occurrence of gene conversions is also evident when one compares the human CD33rSiglec genes which converted one another with those that did not. At the protein level, the average similarity (± standard deviation) between the extracellular regions of the genes which did convert one another (i.e., the *SIGLEC3-SIGLEC6*, *SIGLEC5-SIGLEC14*, *SIGLEC7-SIGLEC9*, *SIGLEC10-SIGLEC11*, *SIGLEC10-SIGLEC16* and *SIGLEC11-SIGLEC16* gene pairs, 83.7% ± 11.2%) is significantly higher than the average similarity of the extracellular regions between the

genes which did not convert one another (i.e., all other gene pairs; 54.4% ± 11.3%; *t*-test, $p = 0.0007$; Supplemental Table 2.1).

Our data also suggest that the higher similarity caused by conversions favours the occurrence of more conversions. This suggestion, that gene conversions have occurred repeatedly during the evolution of the extracellular domain of CD33rSiglecs, is supported by the fact the GC-content of the converted regions is significantly higher than that of non-converted regions. Several studies have shown that increased recombination, such as gene conversions, leads to increases in GC-content (Galtier et al. 2001; Birdsell 2002; Meunier and Duret 2004; Benovoy et al. 2005). Therefore, the fact that converted regions have significantly higher GC-contents than non-converted regions, suggests that repeated recombination events occurred in the converted regions.

Our results show that gene conversions are indeed very frequent between CD33rSiglecs. Given that we observed 11 conversions in humans, 9 in chimpanzees, 4 in Sumatran orang-utan, 4 in rhesus monkey and 5 in Northern white-cheeked gibbon (Table 2.1) and that these species have, respectively, 11, 12, 10, 7 and 8 CD33rSiglec genes (Figure 2.2), the frequency of gene conversions between the CD33rSiglec genes of these respective species, calculated as number of conversions per number of gene pairs compared, is therefore 20%, 14%, 9%, 19% and 18%. This is much higher than the average frequency of gene conversion events between human genes which we calculated to be 0.88% using the same methodology as we used in this study (Benovoy and Drouin 2009).

As mentioned above, Hawakawa et al. (2005) previously observed the ~ 2 kb-long conversion present between the human *SIGLEC11* gene and the *SIGLEC16P* pseudogene and they suggested that this conversion was potentially adaptive (Hayakawa et al. 2005; Table 2.1). Note that this conversion is not present in other primate species

because the *SIGLEC16P* pseudogene (allele) in only found in human genomes (Figure 2.2). Wang et al. (2012) also recently suggested that this conversion is evolutionary significant, but they suggested that this conversion was in fact made up of two tandem gene conversions. Our results support both the presence of a conversion between these two sequences and the fact that it might be composed of two tandem gene conversions (Table 2.1, results not shown). In fact, although the genomic DNA (i.e., including exons and introns) of the coding regions of these two sequences is only 70% similar, the 2286 bp contained in the converted region are 99% similar and this region is followed by a region of 490 bp with 88% similarity between these two sequences. This second region of high similarity might therefore represent the remnants of an older conversion. Our results also complement those of previous studies by showing that gene conversions are not limited to those between the human *SIGLEC11* and *SIGLEC16P* sequences. In fact, in both humans and chimpanzees, gene conversions also occurred between the *SIGLEC10* and *SIGLEC11*, the *SIGLEC10* and *SIGLEC16* and the *SIGLEC11* and *SIGLEC16* gene pairs (Table 2.1).

The conversion between the 5'-end of the *SIGLEC5* and *SIGLEC14* genes has previously been detected in 5 primate species (Angata et al. 2006). We extend this finding by showing that it is also present in the Northern white-cheeked gibbon (Table 2.1 and Figure 2.2). However, contrary to this previous study, our results show that this conversion is always from the *SIGLEC14* gene to *SIGLEC5* gene (Table 2.1 and Figure 2.2). Again, given that this conversion is found in the same region in all species, it likely occurred in their common ancestor. However, given that the increased similarity brought about by this initial gene conversion event facilitates further conversion events, we cannot rule out that subsequent conversion events occurred independently in the same region in some or all these species. The conversion between the *SIGLEC10* and *SIGLEC11* genes of human

and chimpanzee was also previously reported in humans (Angata et al. 2002; Cao and Crocker 2010; Table 2.1).

Are the frequent gene conversions present in the extracellular region of CD33rSiglec genes adaptive or not? Since the extracellular region of CD33rSiglecs is exposed outside the cells, previous studies have suggested that the gene conversions occurring in this region were subject to positive selection (Angata et al. 2004, Crocker et al. 2007, Varki 2010). Furthermore, since the V-type 1 Ig domain binds sialic acid, it has been suggested that positive selection should act on V-type 1 Ig domains and not the adjacent C2-type 1 Ig domains (Altheide et al. 2006; Figure 2.1). This prediction has been convincingly confirmed for the *SIGLEC9* gene (Sonnenburg et al. 2004). Therefore, if the gene conversions observed between CD33rSiglec genes were selected to increase sequence diversity, we would expect them to occur between the most divergent gene sequences. We would also expect them to be more frequent within V-type 1 domains than within the adjacent C2-type 1 domains. As discussed above, our results show that there is a strong correlation between sequence similarity and the length of the converted regions. In other words, gene conversions only occur between very similar sequences. Furthermore, gene conversions are not more frequent within V-type 1 Ig domains than within the adjacent C2-type 1 Ig domains. In fact, if we consider the data for all primate species, there are 10 gene conversions within V-type 1 domains and 9 within the adjacent C2-type 1 domains (Table 2.1, Figure 2.4). These observations therefore do not support the hypothesis that diversifying selection is responsible for gene conversions observed between CD33rSiglec genes. Since we cannot reject the neutral hypothesis that gene conversions have no selective impact on the evolution of CD33rSiglec genes, our results suggest that the gene conversions observed between them are selectively neutral.

The fact that gene conversions occur equally frequently in Ig-like V-type 1 and the Ig-like C2-type 1 domains of CD33rSiglec genes suggests that both these regions evolve under similar selective constraints. We therefore tested the claim that the Ig-like V-type 1 domains of Siglec genes evolve under positive selection and that the Ig-like C2-type 1 domains do not. Our results clearly show that, except for the Ig-like V-type 1 domain of the *SIGLEC9* gene, both the Ig-like V-type 1 and the Ig-like C2-type 1 domains of Siglec genes do not evolve under positive selection (Table 2.2). In these cases the dN/dS ratios of these domains are either not significantly different than 1 or are significantly smaller than one. This suggests that both domains evolve neutrally or under purifying selection. These results are contrary to those of previous studies which suggested that the high degree of variation observed in Siglec genes is adaptive and that the Ig-like V-type 1 domains of all Siglec genes evolve under positive selection (Angata et al. 2004; Altheide et al. 2006; Varki 2010; Jandus et al. 2011). This discrepancy can be due to the fact that previous studies did not assess the statistical significance of the dN/dS ratios they calculated, that they used inappropriate methodologies to calculate these ratios, that they concatenated numerous sequences or that they mistakenly interpreted higher dN/dS ratios smaller than 1 as representing positive selection rather than relaxed purifying selection.

In conclusion, our results suggest that the evolution of Siglec genes is different from the currently accepted view which posits that the high degree of variation observed in Siglec genes is adaptive (Angata et al. 2004; Altheide et al. 2006; Varki 2010; Jandus et al. 2011). If the observed variation was adaptive, one would expect that there would be conversions between the conserved Siglec genes. The fact that there are none suggests that it is not adaptive. We suggest that the absence of gene conversions is simply the result that these genes are too different from one another, and from the CD33rSiglec genes, to

convert them or be converted by them. Conversely, the fact that conversions are frequent between closely related CD33rSilec genes is likely the result of the fact that they very similar to one another. This suggestion is supported by the fact that conversions occur almost exclusively between closely related CD33rSiglec genes (Table 2.1, Figure 2.3, Supplemental Table 2.1). It is also supported by the fact that gene conversions only occur in highly similar regions of CD33rSiglecs (Table 2.1; Figure 2.3, Supplemental Table 2.1). Furthermore, the fact that gene conversions are not more frequent within V-type 1 Ig domains than within the adjacent C2-type 1 Ig domains suggest that these conversions are not subject to positive selection. Therefore, the frequent conversions observed between primate CD33rSiglec genes likely represent neutral events that are not selected against.

Acknowledgements

This work was supported by a discovery grant from the Natural Science and Engineering Research Council of Canada to G. D. M. Z. also received a tuition fee exemption scholarship from the Ministry of Higher Education, Scientific Research and Technology of the Republic of Tunisia.

Literature cited

Angata T, Kerr SC, Greaves DR, Varki NM, Crocker PR, Varki A. 2002. Cloning and characterization of human Siglec-11. A recently evolved signaling molecule that can interact with SHP-1 and SHP-2 and is expressed by tissue macrophages, including brain microglia. *J Biol Chem*. 277: 24466–24474.

Angata T, Margulies EH, Green ED, Varki A. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci*. 101: 13251–13256.

Angata T, Hayakawa T, Yamanaka M, Varki A, Nakamura M. 2006. Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. *FASEB J*. 20: 1964–1973.

Altheide TK, Hayakawa T, Mikkelsen TS, Diaz S, Varki N, Varki A. 2006. System-wide genomic and biochemical comparisons of sialic acid biology among primates and rodents: Evidence for two modes of rapid evolution. *J Biol Chem.* 281: 25689–25702.

Benovoy D, Morris RT, Morin A, Drouin G. 2005. Ectopic gene conversions increase the GC-content of duplicated yeast and Arabidopsis genes. *Mol Biol Evol*. 22: 1865–1868.

Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics* 93: 27–32.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*. 19: 1181–1197.

Cao H, Lakner U, de Bono B, Traherne JA, Trowsdale J, Barrow AD. 2008. SIGLEC16 encodes a DAP12 associated receptor expressed in macrophages that evolved from its inhibitory counterpart SIGLEC11 and has functional and non-functional alleles in humans. *Eur J Immunol.* 38: 2303–2315.

Cao H, Crocker PR. 2010. Evolution of CD33-related siglecs: regulating host immune functions and escaping pathogen exploitation? *Immunology* 132: 18–26.

Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8: 762–775.

Chen JM, Férec C, Cooper DN. 2010. Gene conversion in human genetic disease. *Genes* 1: 550–563.

Crocker PR. 2002. Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling. *Curr Opin Struct Biol.* 12: 609–615.

Crocker PR, Paulson JC, Varki A. 2007. Siglecs and their roles in the immune system. *Nat Rev Immunol.* 7: 255–266.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. *J Mol Evol.* 55: 14–23.

Edgar R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159: 907–911.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acid Symp Ser.* 41: 95–98.

Hayakawa T, Angata T, Lewis AL, Mikkelsen TS, Varki NM, Varki A. 2005. A human-specific gene in microglia. *Science* 309: 1163.

Jandus C, Simon HU, von Gunten S. 2011. Targeting siglecs a novel pharmacological strategy for immuno- and glycotherapy. *Biochem Pharmacol*. 82: 323–332.

McGrath CL, Casola C, Hahn MW. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics* 182: 615–622.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*. 21: 984–990.

Page RDM. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci*. 12: 357–358.

Petronella N, Drouin G. 2011. Gene conversions in the growth hormone gene family of primates: Stronger homogenizing effects in the Hominidae lineage. *Genomics* 98: 173–181.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci.* 98: 13757–13762.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6:526–538.

Sonnenburg JL, Altheide TK, Varki A. 2004. A uniquely human consequence of domain-specific functional adaptation in a sialic acid-binding receptor. *Glycobiology* 14: 339–346.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 28: 2731–2739.

Varki A. 2009. Natural ligands for CD33-related Siglecs? *Glycobiology* 19:810–812.

Varki A. 2010. Uniquely Human Evolution of Sialic Acid Genetics and Biology. *Proc Natl Acad Sci*. 107: 8939–8946.

Wang X, Mitra N, Cruz P, Deng L; NISC Comparative Sequencing Program, Varki N, Angata T, Green ED, Mullikin J, Hayakawa T, Varki A. 2012. Evolution of Siglec-11 and Siglec-16 Genes in Hominins. *Mol Biol Evol*. 29: 2073–2086.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24: 1586–1591.

Table 2.1. Gene conversions, and their location, in the CD33rSiglec genes of five primate species

Species	Donor (D) and acceptor (A) genes	Length (bp)	Location			
			Genomic DNA (bp)		Protein (domains)	
			From	To	From	To
Homo sapiens	Siglec-11 (A) /Siglec-16P* (D)	2286	Exon1 (134)	Exon7 (2420)	Signal peptide	C2-type 3
(Human)	Siglec-11 (D) /Siglec-16 (A)	1628	Exon1 (92)	Exon5 (1720)	Signal peptide	C2-type 2
	Siglec-10 (D) /Siglec-16P* (A)	847	Exon2 (708)	Intron4 (1555)	V-type1	C2-type1
		694	Intron4 (1684)	Exon6 (2378)	C2-type 3	C2-type 3
	Siglec-10 (D) /Siglec-16 (A)	235	Exon2 (703)	Intron2 (938)	V-type1	V-type1
		427	Exon5 (1819)	Intron6 (2246)	C2-type 2	C2-type 3
	Siglec-11 (A) /Siglec-10 (D)	852	Exon2 (611)	Intron4 (1463)	C2-type 1	C2-type 1
		1236	Exon5 (1599)	Exon8 (2835)	C2-type 3	C2-type 3
	Siglec-14 (D) /Siglec-5 (A)	1060	Exon1 (121)	Intron3 (1181)	Signal peptide	C2-type 1
	Siglec-6 (A) /Siglec-3 (D)	208	Exon3 (960)	Exon3 (1168)	C2-type 1	C2-type 1
	Siglec-7 (A) /Siglec-9 (D)	145	Intron1 (1985)	Exon2 (2130)	C2-type 1	C2-type 1
Pan troglodytes	Siglec-11 (A) /Siglec-16 (D)	2168	Exon1 (281)	Exon7 (2449)	Signal peptide	C2-type 3
(Chimpanzee)	Siglec-10 (D) /Siglec-16 (A)	644	Exon6 (2085)	Exon8 (2729)	C2-type 2	C2-type 3
	Siglec-11 (A) /Siglec-10 (D)	685	Exon5 (1611)	Exon7 (2296)	C2-type 2	C2-type 3
		417	Exon7 (2461)	Intron8 (2878)	C2-type 3	C2-type 3
	Siglec-6 (A) /Siglec-3 (D)	192	Exon3 (1017)	Intron3 (1209)	C2-type 1	C2-type 1
	Siglec-9 (A) /Siglec-12 (D)	200	Intron5 (4691)	Intron5 (4891)	NA	NA
	Siglec-7(A) /Siglec-12 (D)	131	Intron3 (4274)	Intron3 (4405)	NA	NA
	Siglec-5 (A) /Siglec-14 (D)	503	Exon1 (126)	Intron2 (629)	Signal peptide	V-type /C2-type 1
	Siglec-6 (A) /Siglec-5 (D)	199	Exon5 (1905)	Exon5 (2104)	C2-type 2	C2-type 2
Macaca mulatta	Siglec-9 (D) /Siglec-11 (A)	137	Intron6 (6623)	Intron6 (6760)	NA	NA
(Rhesus	Siglec-6 (A) /Siglec-3 (D)	235	Exon4 (1034)	Exon4 (1269)	C2-type 1	C2-type 1
monkey)	Siglec-14 (D) /Siglec-5 (A)	1022	Exon1 (21)	Intron5 (1043)	Signal peptide	V-type/C2-type 1
	Siglec-9 (A) /Siglec-3 (D)	100	Intron6 (6447)	Intron6 (6547)	NA	NA
Pongo abelii	Siglec-14 (D) /Siglec-5 (A)	1029	Exon1 (115)	Intron3 (1144)	Signal peptide	C2-type 1
(Sumatran	Siglec-6 (A) /Siglec-5 (D)	116	Exon5 (1864)	Exon5 (1980)	C2-type 2	C2-type 2
orang-utan)	Siglec-7 (A) /Siglec-9 (D)	209	Intron1 (2451)	Intron1 (2660)	NA	NA
	Siglec-7 (A) /Siglec-12 (D)	268	Intron4 10243)	Intron4(10511)	NA	NA
Nomascus leucogenys	Siglec-10 (D) /Siglec11 (A)	296	Exon2 (544)	Exon3 (840)	C2-type 1	C2-type 1
		284	Exon5 (1543)	Intron5 (1827)	C2-type 2	C2-type 3
*(*Northern white-	Siglec-5 (A) /Siglec-14 (D)	606	Exon1 (245)	Intron2 (851)	Signal peptide	C2-type 1
cheeked gibbon)	Siglec-5 (D) /Siglec-6 (A)	165	Exon4 (2348)	Exon4 (2513)	C2-type 2	C2-type 2
	Siglec-7 (A) /Siglec-8 (D)	253	Intron5 (4808)	Intron5 (5061)	NA	NA

NA: not applicable. Asterisks (*) indicate that this sequence is a pseudogene.

Table 2.2. dN and dS values for Ig-like V-type1 and Ig-like C2-type 1 domains

Genes	Pairwise species	Ig-like V-type1 domain			Ig-like C2-type 1 domain		
		dN	dS	dN/dS	dN	dS	dN/dS
Conserved Siglecs							
Siglec1	Human/Chimpanzee	0.0068	0.0217	0.31	0.0062	0.0236	0.26***
	Human/Orang-utan	0.0216	0.1332	0.16**	0.0276	0.0765	0.36***
	Chimpanzee/Orang-utan	0.0212	0.1116	0.19*	0.0269	0.0615	0.44***
Siglec2	Human/Chimpanzee	0.0138	0.0001	138	0	0	0
	Human/Orang-utan	0.0188	0.0371	0.51	0.0371	0.0132	2.81
	Chimpanzee/Orang-utan	0.0181	0.0423	0.43	0.0371	0.0132	2.81
Siglec4	Human/Chimpanzee	0.1565	0.4614	0.34	0.99	93.4478	0.01***
	Human/Orang-utan	0.0045	0.1238	0.03***	0.0041	0.2132	0.02***
	Chimpanzee/Orang-utan	0.1618	0.6644	0.24***	1.0124	87.4077	0.01***
Siglec15	Human/Chimpanzee	0.0068	0.0001	68	0.0101	0.0409	0.25
	Human/Orang-utan	0.0067	0.04	0.17	0.0203	0.1167	0.17*
	Chimpanzee/Orang-utan	0	0.0479	0**	0.0204	0.1189	0.17*
CD33rSiglecs							
Siglec3	Human/Chimpanzee	0.0324	0.0109	2.97	0.0106	0.0167	0.63
	Human/Orang-utan	0.0414	0.0432	0.96	0.0326	0.0166	1.96
	Chimpanzee/Orang-utan	0.0544	0.0319	1.70	0.0213	0.0002	106.5
Siglec5	Human/Chimpanzee	0.0394	0.0116	3.40	0.0158	0.0171	0.92
	Human/Orang-utan	0.1003	0.0686	1.46	0.0209	0.0562	0.37
	Chimpanzee/Orang-utan	0.0913	0.0828	1.10	0.0156	0.0373	0.42
Siglec6	Human/Chimpanzee	0.0053	0.0001	53	0.005	0.0674	0.07*
	Human/Orang-utan	0.0649	0.0752	0.86	0.0422	0.1076	0.39
	Chimpanzee/Orang-utan	0.0707	0.0743	0.95	0.0367	0.0825	0.44
Siglec7	Human/Chimpanzee	0.0492	17.5484	0.003***	0.0841	43.1122	0.002***
	Human/Orang-utan	0.0279	0.0295	0.95	1.0024	2.4523	0.41
	Chimpanzee/Orang-utan	0.0501	15.1777	0.003***	1.0158	5.8636	0.17***
Siglec8	Human/Chimpanzee	0.0056	0.0001	56	0.021	0.0353	0.60
	Human/Orang-utan	0.0155	0.0193	0.80	1.4138	3.8808	0.36***
	Chimpanzee/Orang-utan	0.0212	0.0181	1.17	1.5404	2.7574	0.56***
Siglec9	Human/ Chimpanzee	0.0285	0.0003	95	0.0048	0.076	0.06**
	Human/Orang-utan	0.1146	0.0162	7.07*	0.0216	0.1276	0.17*
	Chimpanzee/ Orang-utan	0.0867	0.0169	5.13	0.0268	0.1541	0.17**
Siglec10	Human/Chimpanzee	0.0047	0.0247	0.19	0.0108	0.0212	0.51
	Human/Orang-utan	18.9852	10.7467	1.77	0.6126	1.7758	0.34**
	Chimpanzee/Orang-utan	18.9598	10.8299	1.75	0.6065	1.7911	0.34**
Siglec11	Human/Chimpanzee	0.0119	0.037	0.32	0.005	0.0566	0.09*
	Human/Orang-utan	0.0357	0.0561	0.64	0.0155	0.0559	0.28
	Chimpanzee/Orang-utan	0.043	0.0477	0.90	0.0204	0.0382	0.53
Siglec12	Human/Chimpanzee	0.0116	0.0001	116	0.0199	0.0665	0.30
	Human/Orang-utan	0.0418	0.0551	0.76	0.0253	0.1569	0.16**
	Chimpanzee/Orang-utan	0.0382	0.0533	0.72	0.0257	0.1264	0.20*
Siglec14	Human/Chimpanzee	0.0394	0.0116	3.40	0.0104	0.0175	0.60
	Human/Orang-utan	0.0924	0.081	1.14	0.0103	0.077	0.13*
	Chimpanzee/Orang-utan	0.0834	0.0949	0.88	0.0104	0.0934	0.11**
Siglec16	Human/Chimpanzee	0.0073	0.0587	0.12**	0.0101	0.0374	0.27

Notes. Siglec13 is not present in the human genome. Siglec16 is only present in the human and chimpanzee genomes. * $P < 5\%$; ** $P < 1\%$; *** $P < 0.1\%$.

Figure legends

Figure 2.1. Structure and tissue expression of Hominid Siglecs. This figure is modified from Crocker and Varki (2001) and Jandus et al. (2011).

Figure 2.2. Organisation of CD33rSiglec genes and gene conversions in five primate species. Arrows indicate gene conversions and their direction. Note that, in humans, the Siglec16 locus contains either a functional gene (*Siglec16*) or a pseudogene (*Siglec16P*).

Figure 2.3. Phylogenetic tree of human Siglecs proteins. The scale bar represents 10% difference and the numbers next to nodes are bootstrap values. Percent (%) values indicate the percentage sequence identity of the DNA sequences between the genes indicated by arrows.

Figure 2.4. Schematic representation of the structure of CD33rSiglec proteins and location of the gene conversions detected between the CD33rSiglec genes of five primate species. The different domains of the proteins are indicated by boxes. Boxes with a full line represent domains found in all Siglec proteins whereas those with a dashed line represent domains found only in some Siglec proteins (see Figure 2.1). The range of lengths of these different domains, in amino acids, is indicated below each box. Only conversions affecting protein coding regions are indicated.

FIGURE 2.1

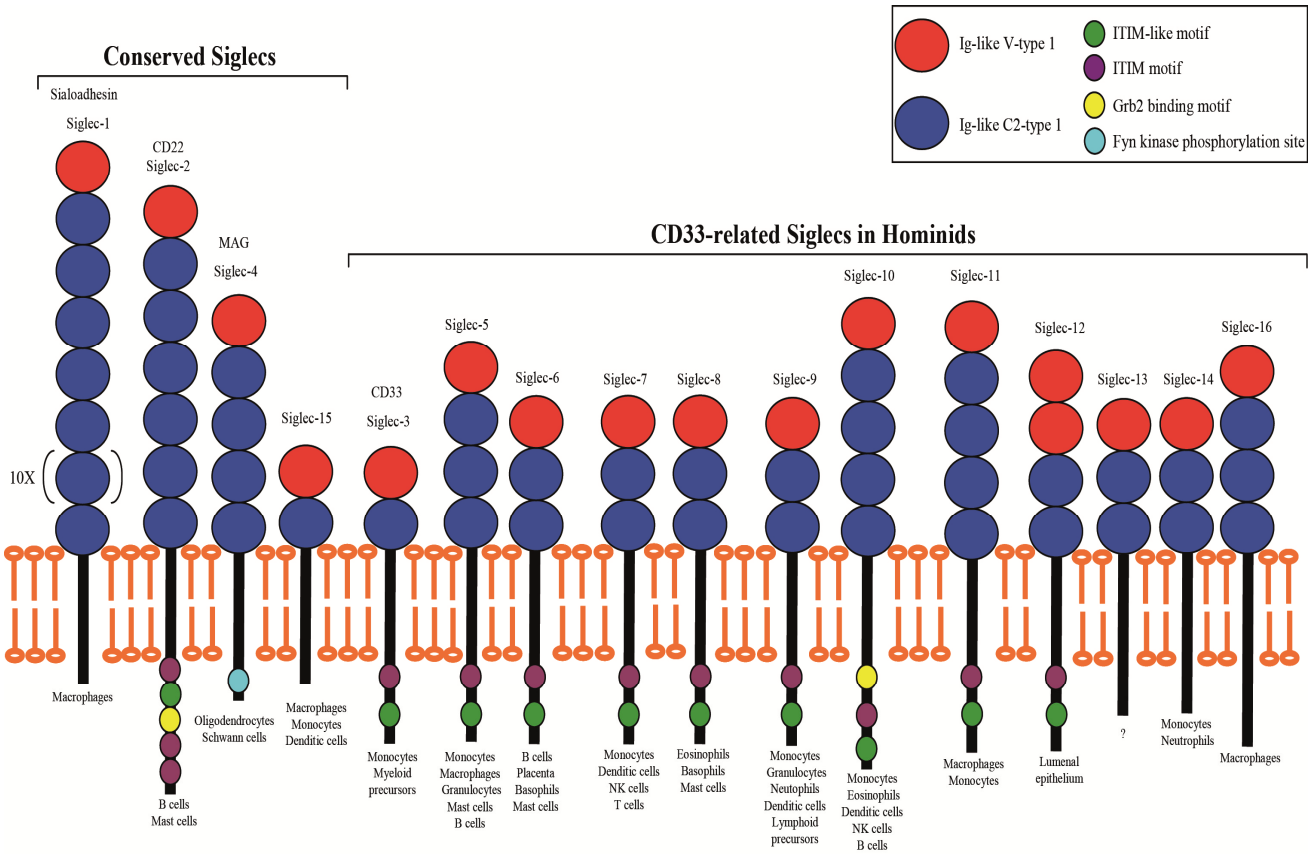

Ig-like V-type 1
Ig-like C2-type 1
ITIM-like motif
ITIM motif
Grb2 binding motif
Fyn kinase phosphorylation site
Conserved Siglecs
Sialoadhesin
Siglec-1
CD22
Siglec-2
MAG
Siglec-4
Siglec-15
CD33
Siglec-3
10X
Macrophages
B cells
Mast cells
Oligodendrocytes
Schwann cells
Macrophages
Monocytes
Dendritic cells
Monocytes
Myeloid
precursors
CD33-related Siglecs in Hominids
Siglec-5
Siglec-6
Siglec-7
Siglec-8
Siglec-9
Siglec-10
Siglec-11
Siglec-12
Siglec-13
Siglec-14
Siglec-16
Monocytes
Macrophages
Granulocytes
Mast cells
B cells
B cells
Placenta
Basophils
Mast cells
Monocytes
Dendritic cells
NK cells
T cells
Eosinophils
Basophils
Mast cells
Monocytes
Granulocytes
Neutophils
Dendritic cells
Lymphoid
precursors
Monocytes
Eosinophils
Dendritic cells
NK cells
B cells
Macrophages
Monocytes
Lumenal
epithelium
?
Monocytes
Neutrophils
Macrophages

FIGURE 2.2

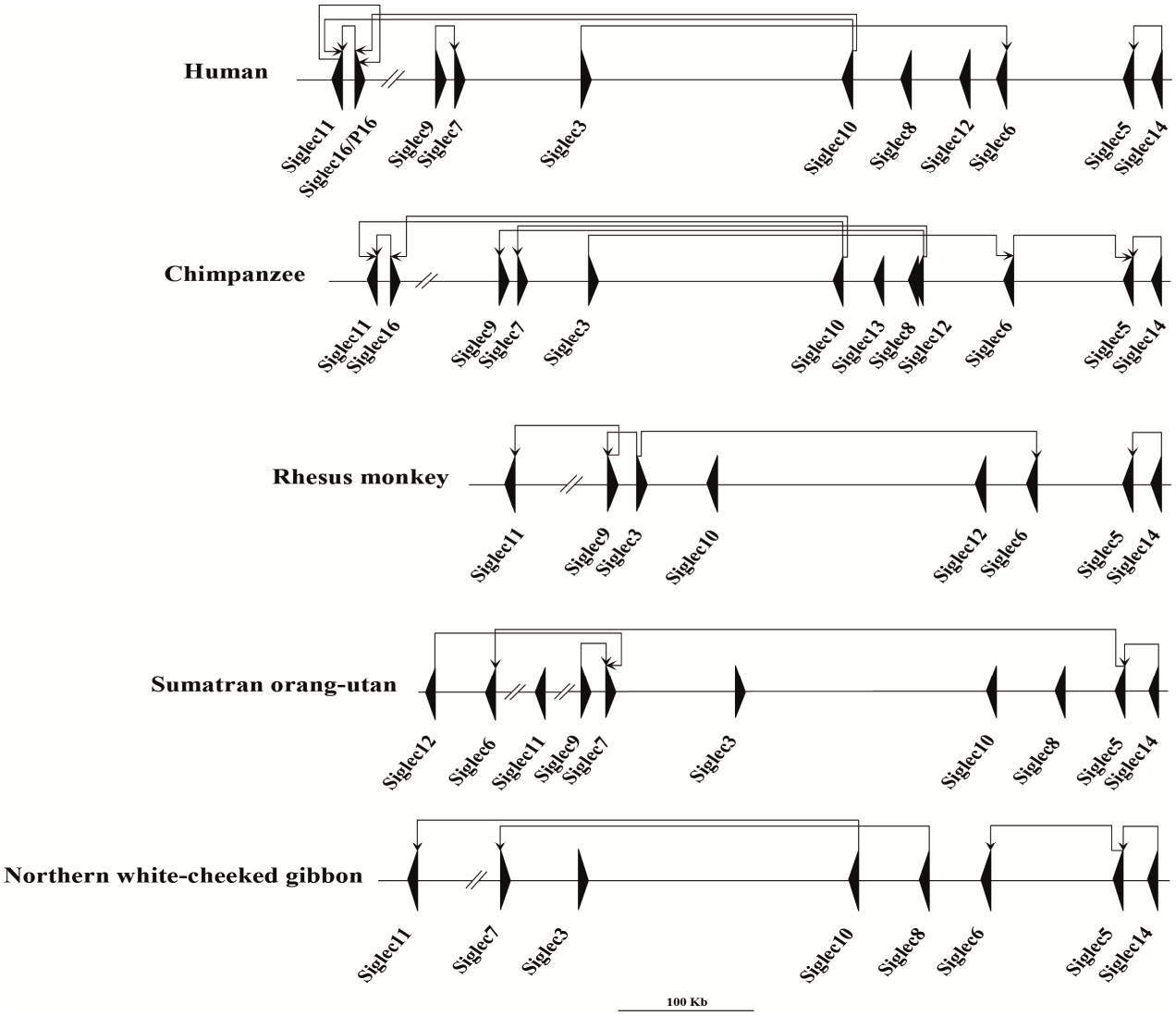

Human
Siglec11
Siglec16/P16
Siglec9
Siglec7
Siglec3
Siglec10
Siglec8
Siglec12
Siglec6
Siglec5
Siglec14
Chimpanzee
Siglec11
Siglec16
Siglec9
Siglec7
Siglec3
Siglec10
Siglec13
Siglec8
Siglec12
Siglec6
Siglec5
Siglec14
Rhesus monkey
Siglec11
Siglec9
Siglec3
Siglec10
Siglec12
Siglec6
Siglec5
Siglec14
Sumatran orang-utan
Siglec12
Siglec6
Siglec11
Siglec9
Siglec7
Siglec3
Siglec10
Siglec8
Siglec5
Siglec14
Northern white-cheeked gibbon
Siglec11
Siglec7
Siglec3
Siglec10
Siglec8
Siglec6
Siglec5
Siglec14
100 Kb

FIGURE 2.3

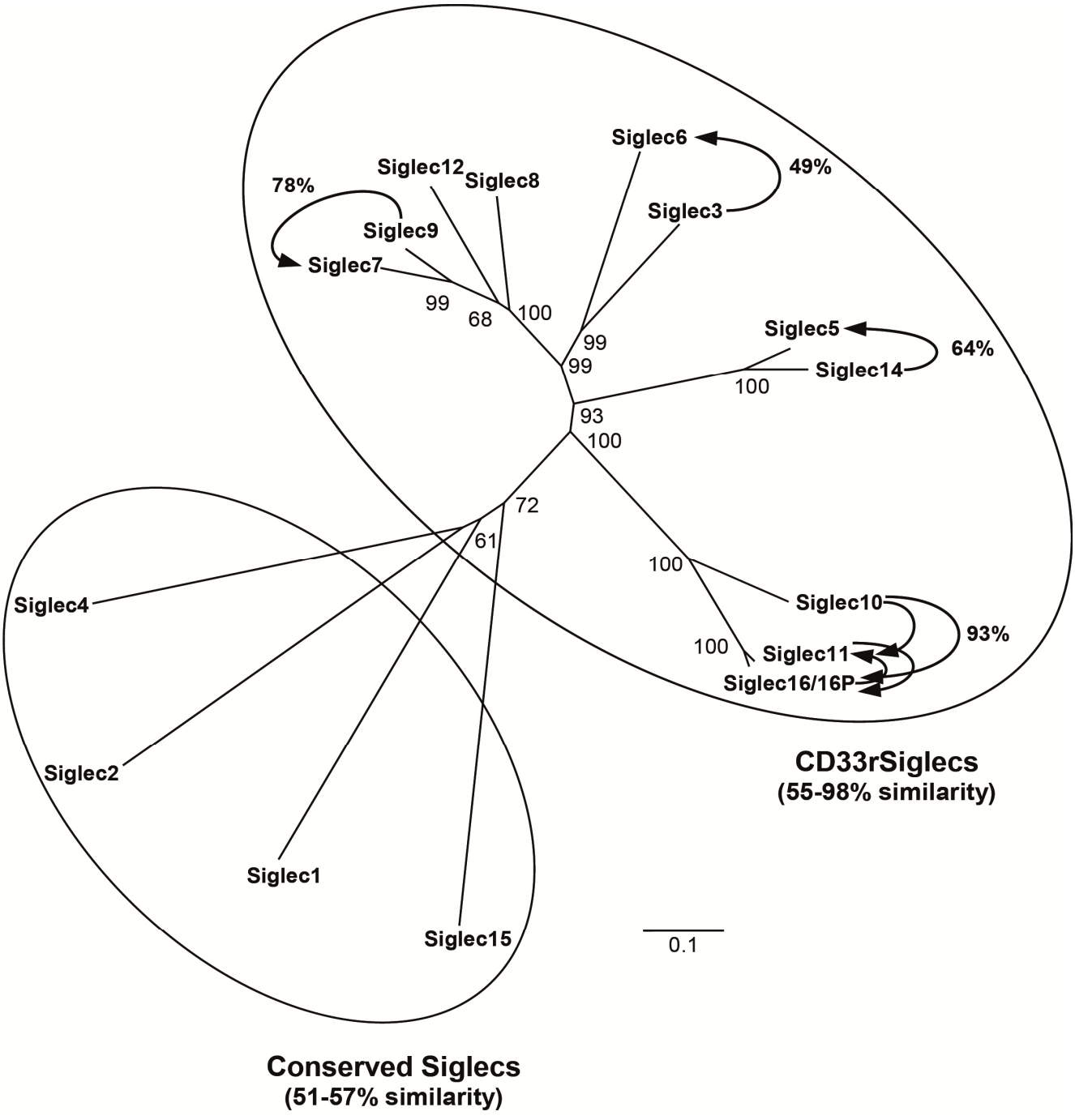
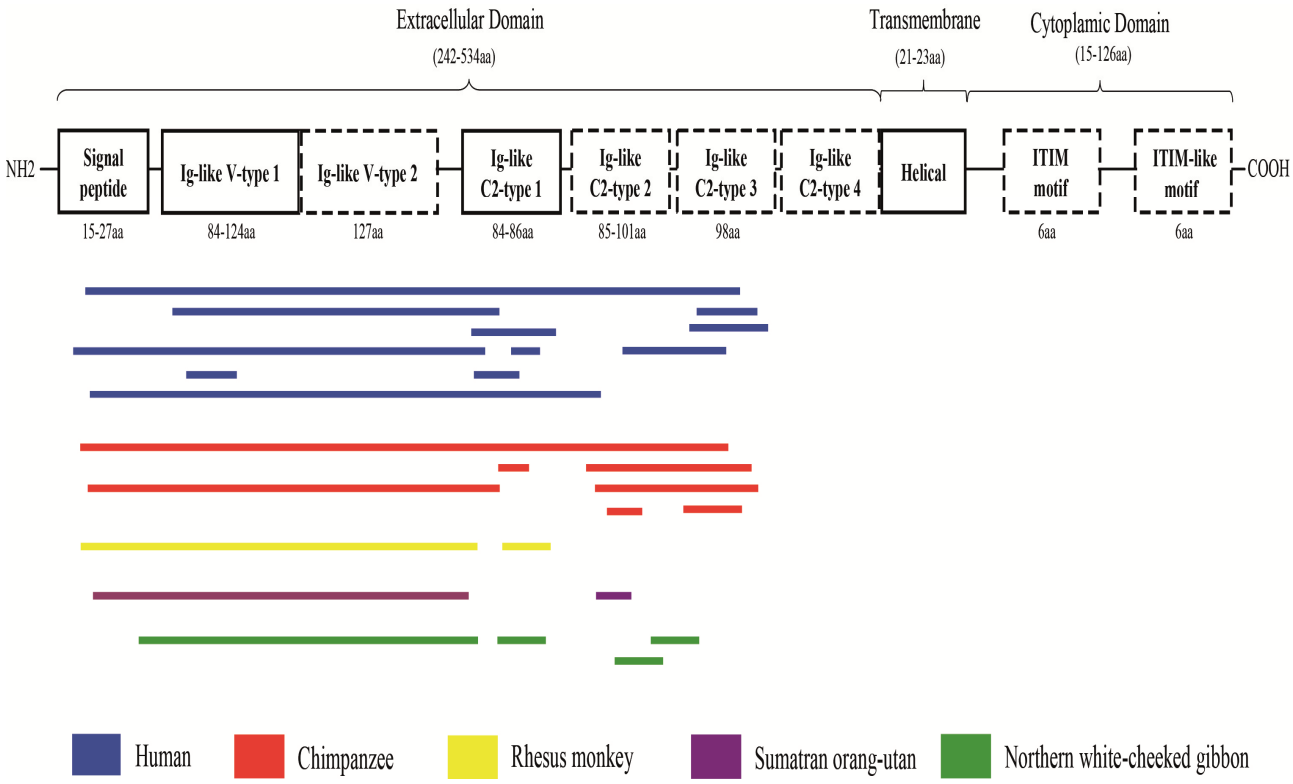

Siglec6
49%
Siglec12
Siglec8
78%
Siglec3
Siglec9
Siglec7
99
68
100
99
99
Siglec5
64%
Siglec14
100
93
100
72
61
100
Siglec10
93%
100
Siglec11
Siglec16/16P
Siglec4
Siglec2
Siglec1
Siglec15
0.1
CD33rSiglecs
(55-98% similarity)
Conserved Siglecs
(51-57% similarity)

FIGURE 2.4


Extracellular Domain
(242-534aa)
Transmembrane
(21-23aa)
Cytoplamic Domain
(15-126aa)
NH2
Signal peptide
Ig-like V-type 1
Ig-like V-type 2
Ig-like C2-type 1
Ig-like C2-type 2
Ig-like C2-type 3
Ig-like C2-type 4
Helical
ITIM motif
ITIM-like motif
COOH
15-27aa
84-124aa
127aa
84-86aa
85-101aa
98aa
6aa
6aa
Human
Chimpanzee
Rhesus monkey
Sumatran orang-utan
Northern white-cheeked gibbon

Supplemental Table 2.1. Similarity between human Siglec protein sequences

Pairwise comparisons		Similarity (%)		
Gene1	Gene2	Complete proteins	Extracellular regions	Transmembrane and cytoplasmic regions
Conseved Siglecs				
Siglec1	Siglec2	31	28	25
Siglec1	Siglec4	22	30	10
Siglec1	Siglec15	18	29	25
Siglec2	Siglec4	24	26	24
Siglec2	Siglec15	15	22	31
Siglec4	Siglec15	18	23	19
Average		**21**	**27**	**23**
CD33 related Siglecs				
Siglec3	Siglec5	57	55	46
Siglec3	Siglec6	69	67	46
Siglec3	Siglec7	60	59	34
Siglec3	Siglec8	62	62	40
Siglec3	Siglec9	60	61	37
Siglec3	Siglec10	47	48	54
Siglec3	Siglec11	46	45	40
Siglec3	Siglec12	56	55	31
Siglec3	Siglec14	53	56	11
Siglec3	Siglec16	46	45	17
Siglec5	Siglec6	51	54	43
Siglec5	Siglec7	54	55	37
Siglec5	Siglec8	54	56	46
Siglec5	Siglec9	54	55	43
Siglec5	Siglec10	48	46	37
Siglec5	Siglec11	47	46	54
Siglec5	Siglec12	51	53	37
Siglec5	Siglec14	78	96	11
Siglec5	Siglec16	45	45	14
Siglec6	Siglec7	56	57	51
Siglec6	Siglec8	57	57	66
Siglec6	Siglec9	59	60	60
Siglec6	Siglec10	43	43	40
Siglec6	Siglec11	43	41	54
Siglec6	Siglec12	55	55	46
Siglec6	Siglec14	49	54	14
Siglec6	Siglec16	42	47	11
Siglec7	Siglec8	72	73	74
Siglec7	Siglec9	78	85	63
Siglec7	Siglec10	43	45	31
Siglec7	Siglec11	43	44	46
Siglec7	Siglec12	76	76	69
Siglec7	Siglec14	52	56	14
Siglec7	Siglec16	43	46	9
Siglec8	Siglec9	72	73	80
Siglec8	Siglec10	47	50	40
Siglec8	Siglec11	46	48	54
Siglec8	Siglec12	72	76	66
Siglec8	Siglec14	53	56	14
Siglec8	Siglec16	47	48	14
Siglec9	Siglec10	43	47	37
Siglec9	Siglec11	44	46	51
Siglec9	Siglec12	76	75	57
Siglec9	Siglec14	53	56	11

Siglec9	Siglec16	45	46	14
Siglec10	Siglec11	70	80	34
Siglec10	Siglec12	45	46	26
Siglec10	Siglec14	45	47	9
Siglec10	Siglec16	76	78	14
Siglec11	Siglec12	43	43	43
Siglec11	Siglec14	45	47	17
Siglec11	Siglec16	88	96	11
Siglec12	Siglec14	50	54	17
Siglec12	Siglec16	41	44	14
Siglec14	Siglec16	43	46	14
Average		**55**	**57**	**36**

Note. The fact that the similarity of complete sequences is not simply an average of the similarities of extracellular and transmembrane and cytoplasmic regions is a consequence of the fact that these values were calculated using different protein alignments.

Supplemental Table 2.2. List of the Siglecs genes that were used in this study and their corresponding GenBank/Ensembl accession numbers

Species	Accession number	Species	Accession number
Homo sapiens Siglec1	AF230073	*Macaca mulatta* Siglec4	XM_001111459
Homo sapiens Siglec2	X52785	*Macaca mulatta* Siglec5	NM_001109886
Homo sapiens Siglec3	M23197	*Macaca mulatta* Siglec6	XM_001116391
Homo sapiens Siglec4	M29273	*Macaca mulatta* Siglec9	XM_001114560
Homo sapiens Siglec5	U71383	*Macaca mulatta* Siglec10	XM_001116298
Homo sapiens Siglec6	D86358	*Macaca mulatta* Siglec11	XM_001115795
Homo sapiens Siglec7	AF170485	*Macaca mulatta* Siglec12	XM_001116366
Homo sapiens Siglec8	AF195092	*Macaca mulatta* Siglec14	XM_001114886
Homo sapiens Siglec9	AF135027	*Macaca mulatta* Siglec15	XM_001089000
Homo sapiens Siglec10	AF310233	*Pongo abelii* Siglec1	XM_002830081
Homo sapiens Siglec11	AF337818	*Pongo abelii* Siglec2	ENSPPYG00000009858
Homo sapiens Siglec12	AF282256	*Pongo abelii* Siglec3	XM_002829641
Homo sapiens Siglec14	AY854038	*Pongo abelii* Siglec4	NM_001134188
Homo sapiens Siglec15	AK095432	*Pongo abelii* Siglec5	ENSPPYG00000010324
Homo sapiens Siglec16	JQ045129	*Pongo abelii* Siglec6	XM_002834514
Homo sapiens Siglec16P	BC030222	*Pongo abelii* Siglec7	ENSPPYG00000010312
Pan troglodytes Siglec1	XM_525251	*Pongo abelii* Siglec8	XM_002829656
Pan troglodytes Siglec2	AF199415	*Pongo abelii* Siglec9	ENSPPYG00000010311
Pan troglodytes Siglec3	XM_512850	*Pongo abelii* Siglec10	XM_002829668
Pan troglodytes Siglec4	ENSPTRG00000024187	*Pongo abelii* Siglec11	XM_002829588
Pan troglodytes Siglec5	XM_512860	*Pongo abelii* Siglec12	XM_002834513
Pan troglodytes Siglec6	XM_003316575	*Pongo abelii* Siglec14	XM_002829657
Pan troglodytes Siglec7	XM_524360	*Pongo abelii* Siglec15	ENSPPYG00000009127
Pan troglodytes Siglec8	ENSPTRG00000011384	*Nomascus leucogenys* Siglec1	XM_003278003
Pan troglodytes Siglec9	XM_003316566	*Nomascus leucogenys* Siglec2	XM_003280021
Pan troglodytes Siglec10	XM_524361	*Nomascus leucogenys* Siglec3	XM_003269838
Pan troglodytes Siglec11	XM_001158083	*Nomascus leucogenys* Siglec4	ENSNLEG00000011129
Pan troglodytes Siglec12	AF293372	*Nomascus leucogenys* Siglec5	XM_003269706
Pan troglodytes Siglec13	AY485345	*Nomascus leucogenys* Siglec6	XM_003269848
Pan troglodytes Siglec14	ENSPTRG00000024249	*Nomascus leucogenys* Siglec7	XM_003269835
Pan troglodytes Siglec15	XM_512109	*Nomascus leucogenys* Siglec8	XM_003269847
Pan troglodytes Siglec16	XM_001173557	*Nomascus leucogenys* Siglec10	XM_003269704
Macaca mulatta Siglec1	XM_001115256	*Nomascus leucogenys* Siglec11	XM_003269785
Macaca mulatta Siglec2	ENSMMUG00000017621	*Nomascus leucogenys* Siglec14	XM_003269707
Macaca mulatta Siglec3	XM_001114616	*Nomascus leucogenys* Siglec15	XM_003267532

Chapter 3. Gene conversions are under purifying selection in the carcinoembryonic antigen immunoglobulin gene families of primates

Abstract

The carcinoembryonic antigen (CEA) family contains a large number of glycoproteins belonging to the immunoglobulin superfamily. The majority of these proteins are involved in cell-cell interactions and antigen recognition. Previous studies reported the presence of gene conversions between the CEA genes of some mammalian species. Here, we investigate whether gene conversions also occur between primate CEA genes and whether these conversions could be adaptive. Our results show primate CEA genes are subject to frequent gene conversion events that roughly half of them affect coding regions and that most of them are in regions coding for the extracellular portion of CEA proteins. The fact that converted regions are significantly more GC-rich than non-converted regions suggests that repeated gene conversion events occurred in the regions where they were detected. Furthermore, gene conversions occur most frequently between closely related genes, are not more frequent in Ig-like V-type 1 domains than in the Ig-like C2-type 1 domains and dN/dS ratio tests shown that both these domains evolve either neutrally or under purifying selection. Our results therefore suggest that CEA genes evolve under purifying selection and the gene conversions events we observed likely represent selectively neutral events between genes having similar sequences and functions.

3.1. Introduction

The human carcinoembryonic antigen (CEA) multigene family is a subgroup of the immunoglobulin superfamily which includes 34 related genes and pseudogenes located within a 1.5 Mb region of chromosome 19 (Naghibalhossaini et al. 2007). This family is involved in cell–cell adhesion and affects the growth and differentiation of various normal and pathogenic human tissues (Gray-Owen and Blumberg 2006). At present, the CEA family can be divided into three subgroups based on sequence similarity, developmental expression patterns, and their functions characteristics. The CEACAM (CEA-related cell adhesion molecule) subgroup contains twelve genes (*CEACAM1*, *CEACAM3* through *CEACAM8*, *CEACAM16*, and *CEACAM18* through *CEACAM21*), the PSG (pregnancy specific glycoprotein) subgroup consists of eleven closely related genes (*PSG1* through *PSG11*), and the third subgroup encloses eleven pseudogenes (*CEACAMP1* through *CEACAMP11*; Gray-Owen and Blumberg 2006; Figures 3.1-3.3). The extracellular regions of the CEACAM subgroup are heavily glycosylated and share 17-88% similarity in their protein sequences (Supplemental Table 3.1). The members of the PSG subgroup are more conserved; they share 81-94% similarity in their extracellular protein sequences (Supplemental Table 3.1). The common characteristic among CEACAM proteins is that they are made up of single 26-160 amino acids long Ig-like V-type 1 domains, followed by zero to six Ig-like C2-type domains of A and B subtypes in their extracellular region (Kammerer et al. 2007). However, *CEACAM16* contains two Ig-like V-type domains at its NH_2-terminal (N1) and COOH-terminal (N2; Zheng et al. 2011), whereas *CEACAM20* contains a truncated Ig-like V-type 1 domain (Gires and Seliger 2009; Figure 3.1). Proteins encoded by members of the PSG subgroup have a structure similar to that encoded by the CEACAM members (Figure 3.2). The extracellular domain of PSG proteins contains

one Ig-like V-type domain (N1), followed by three Ig-like C2-type domains of A, B and C subtypes (Olsen et al. 1994; Figure 3.2). The majority of CEACAM members are anchored in the cell membrane, whereas *CEACAM16* and all of the PSGs appear to be secreted glycoproteins with no glycosyl phosphatidyl inositol (GPI) membrane anchor (Beauchemin et al. 1999; Zheng et al. 2011; Figures 3.1 and 3.2).

Various CEA-family members exhibit diverse functions, including controlling tissue homeostasis, insulin metabolism, modulation of angiogenesis, tumour development (Ergun et al. 2000; Gray-Owen et Blumberg 2006, Kuespert et al., 2006), modulation of apoptosis (Nittka et al. 2004), and regulation of immune response during the period of pregnancy (Wessells et al. 2000). Interestingly, two members of this family, *CEACAM1* and *CEACAM3,* act as receptors for pathogenic bacteria and viruses. Both genes are expressed on the surface of immune cells such as granulocytes. Their role is to destroy the specific pathogens via their signaling motifs, yet diverse pathogenic bacteria, including *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Moraxella catarrhalis* and *Haemophilus influenzae*, as well as some strains of *Escherichia coli* utilize CEACAMs during the course of infection of humans (Pils et al. 2008).

The number of CEACAM and PSG genes varies between mammalian species. To date, humans and mice CEA gene family includes 34 and 32 members, which are localized on chromosome 19 and chromosome 7, respectively (Zebhauser et al. 2005; McLellan et al. 2005; Naghibalhossaini et al. 2007). Thirteen genes were identified in the rat genome. There are 12 functional CEACAM genes and 7 PSG genes in chimpanzee genome, 8 of which are CEACAM-like and PSG-like genes.

Gene conversions are non-reciprocal homologous recombination events where a DNA subsequence of one gene is replaced by the DNA subsequence from

another gene. Gene conversions can either generate diversity between genes or maintain sequence similarity between related genes (Arakawa and Buerstedde 2004, Petronella and Drouin 2011). They can also lead to genetic diseases (Chen et al. 2010).

A recent study on the evolutionary forces acting on the CEA family (Kammerer and Zimmermann 2010) demonstrated that the ancestral CEA gene family in mammals consisted of five closely genes, including the *CEACAM1* ancestor. The CEA family is also hypothesized to have evolved by multiple phases of *CEACAM1* duplication and gene conversion events. Furthermore, Kammerer et al. (2007) suggested that the 99% DNA sequence identity observed between the first 2,332 bp of the *CEACAM1* and *CEACAM28* genes of dogs was the result of a gene conversion between these genes.

Here, we report a complete analysis of gene conversions that occurred between the CEA genes of five primate species (*Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Pongo abelii*, and *Nomascus leucogenys*). Since CEA genes are part of the immunoglobulin superfamily, and that adaptive gene conversions are involved in increasing the diversity of many members of this family, our goal was to determine whether adaptive gene conversions also occur between primate CEA genes (Maizel 2005, Das et al. 2011). Our results demonstrate that 55 gene conversion events occurred between primate CEA genes and that the roughly half these conversions are located in exons. Interestingly, most of the 24 conversions affecting the coding regions are localized in the extracellular, transmembrane and cytoplasmic domains but do not show any evidence of having been positively selected. This suggests that CEA genes are evolving under strong purifying selection and gene conversion events are not involved in increasing the sequence diversity of these genes.

3.2. Materials and methods

Database searches and sequence analyses

Protein and genomic sequence of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), Rhesus monkey (*Macaca mulatta*), Sumatran orang-utan (*Pongo abelii*), and Northern white-cheeked gibbon (*Nomascus leucogenys*) were obtained from NCBI (http://www.ncbi.nlm.nih.gov/genbank/) and Ensembl (Hubbard et al. 2005) or by conducting BLAST searches (Altschul 1990). The list of the GenBank and Ensembl accession numbers of the 104 CEA genes we analyzed is shown in Supplemental Table 3.2. Sequences were aligned with the MUSCLE program (Edgar 2004) and then checked and refined by eye using BioEdit (Hall 1999). The sequences similarities between genes, i.e., the uncorrected proportion of identical nucleotides, were calculated using the pairwise-distance computing function of MEGA 5 (Tamura et al. 2011). The GC-contents of the sequences were measured using Seqool (http://www.biossc.de). Average, standard deviations and *t*-test were calculated using Excel 2010. Spearman rank correlations tests were calculated using R version 2.15 (R Development Core Team 2006).

Detecting Gene Conversion

GENECONV version 1.81 (Sawyer 1989) was used to identify gene conversion events using the options g-scale= 2, maxSimPairPval= 0.05, maxSimGlobPval= 0.05 and simPvals based on 10,000 permutations. We used GENECONV because it is a reliable tool for detecting gene conversion events (Drouin et al. 1999; Posada and Crandall 2001; Drouin 2002; Posada 2002). Based on an alignment of DNA sequences, GENECONV detects gene conversion events by

seeking for sequence segments which have unusually high similarity between each pair of sequences.

The directionality of the gene conversion events between gene pairs (donor and acceptor) for each species were determined based on the patterns of nucleotide variation inside and outside of the converted regions when compared to non-converted sequences from closely related species. For example, if there was a gene conversion from nucleotides 4 to 9 between genes 1 and 2 of species A, with sequences CATCGGCGCTGT and AGCCGGCGCTAG, respectively, and that a related gene from a closely related species has the sequence CATCGGCGCTGT, or AGCTATATATAG, then it is gene 1 of species A that converted gene 2 of species A and not the reverse.

Phylogenetic analyses and Test of selection

Phylogenetic trees were built using the neighbor-joining method implemented in MEGA 5 with p-distance on amino acids and 500 bootstrap trees (Tamura et al. 2011).

The ratio of nonsynonymous/synonymous substitution rates (dN/dS= ω) of Ig-like V-type 1 and Ig-like C2-type 1 domains were analyzed using the maximum likelihood method implemented in the codeml program of the PAML software package version 4.5 (Yang 2007). The codon substitution model M0 was used to evaluate the general sequence substitution pattern of coding sequences, specifying seqtype = 1, runmode = -2, CodonFreq= 2 and fix_omega= 0. We calculated and compared the dN/dS ratios between human, chimpanzee and Sumatran orang-utan coding sequences because these three genomes share the largest number of orthologous genes. If dN/dS = 1, the sequences are evolving neutrally, i.e., without selection. If

dN/dS < 1, the coding region is evolving under purifying selection. If dN/dS > 1, the coding region is evolving under positive selection. We tested whether these dN/dS ratios were significantly different than 1 by performing likelihood ratio tests based on the Chi-square distribution with one degree of freedom. These tests compared the likelihood of the dN/dS ratios we obtained when ω was set as a free parameter with those obtained by setting ω equal to 1 (using fix_omega= 1 and omega = 1).

3.3. Results

3.3.1. Number, lengths and direction of gene conversion

Using the GENCONV program, 55 gene conversions events were detected between genome sequences of the CEA multigene family of five primate species. These involved 21 genes, 11 predicted genes and 1 pseudogene (*PSG10P*). There are 28 conversions in human, 12 in chimpanzee, 4 in rhesus monkey, 5 in Sumatran orang-utan and 6 in Northern white-cheeked gibbon (Table 3.1). The average length (± standard deviation) of these gene conversions is 522.76 ± 296.32 nucleotides and they range from 121 to 1130 nucleotides long. It is interesting to note that no conversion was detected between CEACAM genes and the eleven pseudogenes found in the human genome.

Table 3.1 shows the lengths of 55 gene conversions events and their respective locations. There is a 276 nucleotides long conversion between the second exon of the human *CEACAM1* and *CEACAM3* genes. There are 31 conversions located in one intron (introns 1, 2, 3, 4, 5 or 8) with the smallest being 160 nucleotides long and the largest being 1130 nucleotides long. There are 7 conversions starting in an intron and ending in a second intron; they range from 419 to 621 nucleotides long. The average length (± standard deviation) of conversions limited to one intron (564.4 ± 353.38 nucleotides) and those starting in an intron and ending in a second intron (503± 81.14 nucleotides) are not statistically different (*t*-test, *p*-value= 0.39). There are 16 conversions starting in an intron and ending in an exon, or vice versa. The smallest of these conversions is 121 nucleotides long and the largest is 853 nucleotides long. The average length (± standard deviation) of these conversions (462.56 ± 244.55 nucleotides) is not statistically different from the average length of the conversions

which occurred in one intron or those starting in an intron and ending in a second intron (*t*-tests, $p = 0.26$ and $p = 0.56$, respectively). Finally, there is a single (276 nucleotide long) conversion limited to an exon. Of the 55 conversions we detected, roughly half (24) affect coding regions. Moreover, all PSG genes and the seven members of the CEACAM subgroup (*CEACAM1*, *CEACAM3* thought *CEACAM8*) are involved in gene conversions (Figure 3.3, Table 3.1). In contrast, the five members recently identified (*CEACAM16*, *CEACAM18*, *CEACAM19*, *CEACAM20* and *CEACAM21*) did not undergo any gene conversions. This is likely due to their high degree of sequence divergence (Figure 3.4).

The directions of gene conversion between converted genes are shown in Figures 3.3 and 3.4. Eight CEA genes are donor only (*CEACAM5A*, *CEACAM5B*, *PSG2A*, *PSG2C*, *PSG6-likeA*, *PSG8*, *PSG9* and *PSG11-like*), two predicted CEACAM genes and five PSG genes are acceptor only (*CEACAM5-likeA*, *CEACAM5-likeB*, *PSG2*, *PSG3-like*, *PSG4*, *PSG7* and *PSG7-like*) and fifteen genes are both acceptor and donor (*CEACAM1*, *CEACAM1-like*, *CEACAM3* through *CEACAM8*, *PSG1*, *PSG2-like*, *PSG3*, *PSG6*, *PSG6-like*, *PSG6-likeB*, and *PSG8-like*). The polarity between the other genes members of this family could not be determined (*PSG5*, *PSG10P* and *PSG11*). For example, for the human lineage, the direction between *PSG11* and *PSG10P* was not established due to the absence of the *PSG10P* sequence in other primate species.

As shown in Figure 3.3, gene conversions occur between both adjacent and distant genes for all primate species. However, there is a significant negative correlation between the distance between CEA genes and the frequency of conversions in the genome of human, chimpanzee, rhesus monkey and Sumatran orang-utan (Spearman's rank correlation test, $\rho = -0.16$, *p*-value $= 2.27 \times 10^{-04}$;

ρ = -0.27, *p*-value = 6.19 x 10^{-05}; ρ = -0.26, *p*-value = 0.01 and ρ = -0.25, *p*-value = 3.43 x 10^{-03}, respectively). In the Northern white-cheeked gibbon genome, the negative correlation between the frequency of conversions and distance is not significant (Spearman's rank correlation test, ρ = -0.13, *p*-value = 0.29). However, a significant negative correlation is observed between the distance between genes and the frequency of conversions when analyzing the data from all species (Spearman's rank correlation test, ρ = -0.21, *p*-value = 4.9 x 10^{-12}).

3.3.2. Sequence similarities and GC-content of converted and non-converted regions

We calculated the average GC-content (± standard deviation) and the average similarity (± standard deviation) of all converted regions and non-converted regions. The average GC-content (± standard deviation) of converted regions (48.70 % ± 6.08) is significantly higher than that of non-converted regions (45.22 % ± 4.14; *t-test*, *p* = 0.0007). Moreover, there is a consistent pattern between the average similarity of the converted regions and the similarity of the non-converted regions. The average similarity (± standard deviation) between the converted regions is 95.22% ± 2.99, which is significantly higher than the average similarity between the non-converted regions (85.21% ± 12.11; *t*-test, *p* = 1.49 x 10^{-07}).

3.3.3. Biased distribution of converted regions

The extracellular regions of CEACAM proteins extend from a signal peptide to the last Ig-like C2-type domain, the rest being made up of a transmembrane and cytoplasmic domain. Except *CEACAM16*, this region contains a single Ig-like V-type 1 domain followed by 0 to six Ig-like C2-type domains. For this group, the distribution of

17 gene conversions events affecting protein coding regions in the 5 primate species ranges from a signal peptide to the cytoplasmic domain (Figure 3.5A). There are seven gene conversions located in the Ig-like V-type 1 domain, six in the Ig-like C2-type 1, two in the Ig-like C2-type 2, one in the transmembrane domain and one in the cytoplasmic domain. For the PSG subgroup, there are seven gene conversions affecting protein coding regions in human and northern white-cheeked gibbon species. There is one conversion in the peptide signal, one in the Ig-like V-type 1 domain, one in the Ig-like C2-type 1, one in the transmembrane domain and three in the cytoplasmic domain (Figure 3.5B). Overall, of the 24 gene conversions affecting protein coding regions, 18 occurred in the extracellular regions, 2 occurred in the transmembrane domain and 4 occurred in the cytoplasmic domain.

3.3.4. Tests of selection

Table 3.2 shows the ratios of nonsynonymous substitution rates and synonymous substitution rates (dN/dS) in the Ig-like V-type 1 and Ig-like C2-type 1 domains in various CEACAM and PSG genes. For human-chimpanzee comparison, the high dN/dS ratios observed for the *PSG3* (80.5), *CEACAM5* (81.5), *CEACAM7* (56) and *CEACAM18* (77) are due to the very low number of synonymous substitutions between each pairs of the sequences and these values are not significantly greater than 1 (Table 3.2).

In the Ig-like V-type 1 domains, among the seven converted genes (*CEACAM1*, *CEACAM3* though *CEACAM8*), only the *CEACAM6* appear to be under purifying selection for the chimpanzee-orang-utan comparison (dN/dS ratio <1). For the non-converted genes (*CEACAM16*, *CEACAM18*, and *CEACAM19*), purifying selection was found only in the *CEACAM16* gene for the three comparisons and in the

CEACAM18 gene for the human-orang-utan and chimpanzee-orang-utan comparisons (Table 3.2). The values of dN and dS are not shown for *CEACAM20* because it has a truncated Ig-like V-type 1 domain.

In the Ig-like C2-type 1 domains, only the non-converted gene (*CEACAM16*) evolve mainly by purifying selection for the human-orang-utan and chimpanzee-orang-utan comparisons (dN/dS ratios significantly smaller than 1). These results suggest that most CEA family members evolve neutraly, that a few may evolve under negative selection but that none evolve under positive selection.

Discussion

During mammalian evolution, the CEA gene family has been reported to be subject to multiple phases of duplication followed by possible gene conversion events (Kammerer et al. 2007, Kammerer and Zimmermann 2010). Based on the analysis of similarity of N domains sequences between five mammal species, Kammerer et al. (2007) reported that *CEACAM1* and *CEACAM28* genes in dogs have undergone concerted evolution by gene conversions. In the present study, for the first time, we applied the GENECONV statistical method to detect gene conversion events in the CEA gene family of five primate species: *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Pongo abelii*, and *Nomascus leucogenys*. Our results show that gene conversions occur between similar genes, but not occur between more dissimilar genes (Table 3.1 and Figure 3.4). The absence of gene conversion between the group made up of the *CEACAM16*, *CEACAM18*, *CEACAM19*, *CEACAM20,* and *CEACAM21* genes is likely due to their low degree of similarity in their proteins (25% average similarity) and genomic (40-75% similarity) sequences in human genome (Supplemental table 3.1 and Figure 3.4). In contrast, nine gene conversions occurred between the seven human CEACAM genes (*CEACAM1*, *CEACAM3-CEACAM8*) and these genes have DNA sequence similarities in the range of 60-86% (Figure 3.4). Finally, in humans, most gene conversions (19) occurred between the genes coding for PSG proteins, and these genes share a high degree of DNA sequence similarity (92-95% similarity). Our findings are therefore consistent with previous studies which showed that gene conversions occur most often between genes having high nucleotide sequence similarity (Chen et al. 2007, Xu et al. 2008, Benovoy and Drouin 2009, Duvvuri and Wu 2012).

Most of the conversion we detected occurred in, or affected, intron sequences (Table 3.1). High frequencies of gene conversions between the introns of the genes coding for the L and M opsin genes of humans and gibbons have also been observed (Shyue et al. 1994, Zhao et al. 1998; Hiwatashi et al. 2011). This relatively high frequency of gene conversions in introns is likely the result of the fact that they are not eliminated by purifying natural selection because they do not affect protein functions.

Eighteen of the 24 gene conversion events which affected protein coding regions are located in the extracellular region of the CEA proteins. Two conversions occurred in transmembrane regions, and four conversions occurred in cytoplasmic regions (Table 3.1 and Figure 3.5). This high number of conversions in the extracellular regions is likely due to a greater sequence similarity observed between these regions. For example, in the converted CEACAM genes, the average protein similarity (± standard deviation) of the extracellular regions (69.23% ± 13.65%) is significantly higher than that of the transmembrane and cytoplasmic regions (31.57% ± 23.68%, *t*-test, $p = 4.4 \times 10^{-7}$; Supplemental table 3.1).

In CEA genes, the average GC-content of the converted regions is significantly higher than that of non-converted regions ($p = 0.0007$). Since gene conversions are known to increase the GC-content of recombining DNA sequences, this suggests that these regions have been subject to repeated gene conversion events (Galtier et al. 2001; Birdsell 2002; Meunier and Duret 2004; Benovoy et al. 2005). Therefore, the converted regions we detected likely only represent the latest gene conversion events between these sequences.

The frequency of gene conversions (number of conversion events/number of gene pairs compared) in humans (4.99%), chimpanzees (5.71%), rhesus monkey

(4.39%), Sumatran orang-utan (3.67%), and Northern white-cheeked gibbon (9.23%) CEA genes, is higher than that of the average gene families found in the human genome (0.88%; Benovoy and Drouin 2009). This is likely the result of the fact that CEA genes are more similar to one another that the sequences of the genes found in other human gene families.

Our analyses of dN/dS ratios in the CEA family suggest that both Ig-like V-type 1 and the Ig-like C2-type 1 domains evolve neutrally or under purifying selection (Table 3.2). These results are consistent with those of Kammerer and Zimmermann (2010) which showed the extracellular domains of various *CEACAM* members from different mammalian species are under neutrally or purifying selection. Furthermore, within each gene, gene conversions are not more frequent within Ig-like V-type1 domains (8 conversions) than in the adjacent Ig-like C2-type1 domains (7 conversions) (Table 3.1 and Figure 3.5).

Previous studies analyzed the evolution of paired immune receptors within the CEA gene family in mammals and have shown the presence of gene conversion tracts between some CEA paired receptors based on the analysis of phylogeny of N domains. Kammerer and Zimmermann (2010) showed the presence of gene conversion between the human *CEACAM1* and *CEACAM3* genes. Our result supports this conversion, which has been detected in the Ig-like V-type1 domain of these two genes (Table 3.1 and Figure 3.5A). Moreover, Kammerer et al. (2007) suggested that the high sequence similarity of the N-domains of the *CEACAM1* and the *CEACAM28* genes (99%) is due to a partial gene conversion in which the inhibitory receptor *CEACAM1* gene converted the activating receptor *CEACAM28* gene. Note that we did not detect this conversion in primates because the *CEACAM28* gene is not present in the 5 primate species we anaylzed.

Given that CEA genes are part of the immunoglobulin superfamily, and that gene conversions are one of the mechanism known to be involved in the generation of diversity in immunoglobulin genes, one could expect that gene conversions would also be involved in generating sequence diversity in CEA genes (Maizel 2005, Das et al. 2011). However, our results suggest that this is not the case. Although gene conversions are relatively frequent between CEA genes, and lead to significantly higher GC-content in converted regions, most of roughly half of them occur in introns and the gene conversions observed between coding regions are all between similar genes and gene regions. This is the opposite of what would be expected if these gene conversions were positively selected to increase sequence diversity. Furthermore, within each gene, gene conversions are not more frequent within Ig-like V-type1 domains than in the adjacent Ig-like C2-type1 domains. Again, since the Ig-like V-type1 domains are thought to be more functionally important than the adjacent Ig-like C2-type1 domains, this is the opposite of what would be expected if these gene conversions were positively selected to increase sequence diversity. On the other hand, these results are consistent with our analyses of dN/dS ratios which suggest that both Ig-like V-type 1 and the Ig-like C2-type 1 domains evolve either neutrally or under purifying selection. This suggests that CEA genes are evolving under purifying selection and gene conversion events observed between these genes are not involved in increasing their sequence diversity.

Acknowledgements

We thank Stéphane Aris-Brosou (Biology Department, University of Ottawa) for his help with the PAML software. This work was supported by a discovery grant from the Natural Science and Engineering Research Council of Canada to G. D. M. Z.

also received a tuition fee exemption scholarship from the Ministry of Higher Education, Scientific Research and Technology of the Republic of Tunisia.

Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol*. 215: 403–410.

Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics.* 93: 27–32.

Benovoy D, Morris RT, Morin A, Drouin G. 2005. Ectopic gene conversions increase the G+C content of duplicated yeast and Arabidopsis genes. *Mol. Biol. Evol*. 22: 1865–1868.

Beauchemin N, Draber P, Dveksler G, Gold P, Gray-Owen S, Grunert F, Hammarström S, Holmes KV, Karlsson A, Kuroki M, Lin SH, Lucka L, Najjar SM, Neumaier M, Obrink B, Shively JE, Skubitz KM, Stanners CP, Thomas P, Thompson JA, Virji M, von Kleist S, Wagener C, Watt S, Zimmermann W. 1999. Redefined nomenclature for members of the carcinoembryonic antigen family. *Exp. Cell. Res.* 252: 243–249.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* 19: 1181–1197.

Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet*. 8: 762–775.

Chen JM, Férec C, Cooper DN. 2010. Gene conversion in human genetic disease. *Genes.* 1: 550–563.

Das S, Hirano M, McCallister C, Tako R, Nikolaidis N. 2011. Comparative genomics and evolution of immunoglobulin-encoding loci in tetrapods. *Adv Immunol*. 111: 143–178.

Drouin G, Prat F, Ell M, Clarke GD. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol*. 16: 1369–1390.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. *J. Mol. Evol*. 55: 14–23.

Duvvuri B, Wu GE. 2012. Gene conversion-like events in the diversification of human rearranged IGHV3-23*01 gene sequences. *Front Immunol*. 3: 158.

Ergün S, Kilik N, Ziegeler G, Hansen A, Nollau P, Götze J, Wurmbach JH, Horst A, Weil J, Fernando M, Wagener C. 2000. CEA-related cell adhesion molecule 1: a potent angiogenic factor and a major effector of vascular endothelial growth factor. *Mol. Cell.* 5: 311–320.

Edgar R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5: 113.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*. 159:907–911.

Gray-Owen SD, Blumberg RS. 2006. CEACAM1: contact-dependent control of immunity. *Nat. Rev. Immunol.* 6: 433–446.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41: 95–98.

Hiwatashi T, Mikami A, Katsumura T, Suryobroto B, Perwitasari-Farajallah D, Malaivijitnond S, Siriaroonrat B, Oota H, Goto S, Kawamura S. 2011. Gene conversion and purifying selection shape nucleotide variation in gibbon L/M opsin genes. *BMC Evol. Biol*.11: 312.

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E. 2005. Ensembl. *Nucleic Acids Res.* 33: 447–453.

Kammerer R, Popp T, Härtle S, Singer BB, Zimmermann W. 2007. Species-specific evolution of immune receptor tyrosine based activation motif-containing CEACAM1-related immune receptors in the dog. *BMC Evol. Biol.* 7: 196.

Kammerer R, Zimmermann W. 2010. Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. *BMC Biol*. 8: 12.

Kuespert, K, Pils S, Hauck CR. 2006. CEACAMs: Their role in physiology and pathophysiology. *Curr. Opin. Cell Biol.* 18: 565–571.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol*. 21:984–990.

Maizels N. 2005. Immunoglobulin gene diversification. *Annu Rev Genet*. 39: 23-46.

McLellan AS, Fischer B, Dveksler G, Hori T, Wynne F, Ball M, Okumura K, Moore T, Zimmermann W. 2005. Structure and evolution of the mouse pregnancy-specific glycoprotein (Psg) gene locus. *BMC Genomics.* 6: 4.

Naghibalhossaini F, Yoder AD, Tobi M, Stanners CP. 2007. Evolution of a tumorigenic property conferred by glycophosphatidyl-inositol membrane anchors of

carcinoembryonic antigen gene family members during the primate radiation. *Mol. Biol. Cell.* 18: 1366–1374.

Nittka, S, Günther J, Ebisch C, Erbersdobler A, Neumaier M. 2004. The human tumor suppressor CEACAM1 modulates apoptosis and is implicated in early colorectal tumorigenesis. *Oncogene.* 23: 9306–9313.

Olsen A, Teglund S, Nelson D, Gordon L, Copeland A, Georgescu A, Carrano A, Hammarström S. 1994. Gene organization of the pregnancy-specific glycoprotein region on human chromosome 19: assembly and analysis of a 700-kb cosmid contig spanning the region. *Genomics*. 23: 659–668.

Gires O, Seliger B. 2009. Tumor-Associated Antigens: Identification, characterization, and clinical applications. 1 edition, John Wiley & Sons, 383 pages.

Pils S, Gerrard DT, Meyer A, Hauck CR. 2008. CEACAM3: an innate immune receptor directed against human-restricted bacterial pathogens. *Int. J. Med. Microbiol.* 298: 553–560.

Petronella N, Drouin G. 2011. Gene conversions in the growth hormone gene family of primates: Stronger homogenizing effects in the Hominidae lineage. *Genomics.* 98: 173–181.

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol*. 19: 708–717.

Posada D., Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA.* 98:13757–13762.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6: 526–538.

Shyue SK, Li L, Chang BH, Li WH. 1994. Intronic gene conversion in the evolution of human X-linked color vision genes. *Mol. Biol. Evol*. 11: 548–-551.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731–2739.

Wessells J, Wessner D, Parsells R, White K, Finkenzeller D, Zimmermann W, Dveksler G. 2000. Pregnancy specific glycoprotein 18 induces IL-10 expression in murine macrophages. *Eur. J. Immunol.* 30: 1830–1840.

Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK, Wang J, Zheng X. 2008. Gene conversion in the rice genome. *BMC Genomics*. 9: 93.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586–1591.

Zebhauser R, Kammerer R, Eisenried A, McLellan A, Moore T, Zimmermann W. 2005. Identification of a novel group of evolutionarily conserved members within the rapidly diverging murine Cea family. *Genomics.* 86: 566–580.

Zheng J, Miller KK, Yang T, Hildebrand MS, Shearer AE, DeLuca AP, Scheetz TE, Drummond J, Scherer SE, Legan PK, Goodyear RJ, Richardson GP, Cheatham MA, Smith RJ, Dallos P. 2011. Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with α-tectorin and is mutated in autosomal dominant hearing loss (DFNA4). *Proc. Natl. Acad. Sci.*108: 4218–4223.

Zhao Z, Hewett-Emmett D, Li WH. 1998. Frequent gene conversion between human red and green opsin genes. *J. Mol. Evol*. 46: 494–496.

Table 3.1. Gene conversions, and their location, in the CEA genes of five primate species.

Species	Donor and acceptor genes	Length (bp)	Location		
			Genomic DNA (bp)		Protein
			From	To	
Homo sapiens	Ceacam3(A)/Ceacam1(D)	276	Exon2 (1083)	Exon2 (1359)	Ig-like V-type1
	Ceacam3(A)/Ceacam4(D)	191	Intron1 (603)	Intron1 (794)	NA
	Ceacam3(D)/Ceacam5(A)	419	Intron1 (981)	Intron2 (1400)	Ig-like V-type1
	Ceacam3(?)/Ceacam6(?)	574	Exon2 (1119)	Intron2 (1693)	Ig-like V-type1
	Ceacam1(D)/Ceacam4(A)	211	Intron1 (598)	Intron1 (809)	NA
	Ceacam1(A)/Ceacam5(D)	424	Intron1 (1028)	Exon2 (1452)	Ig-like V-type1
	Ceacam1(D)/Ceacam7(A)	466	Exon3 (6376)	Intron3 (6842)	Ig-like C2-type1
	Ceacam6(D)/Ceacam1(A)	433	Intron2 (5707)	Intron3 (6140)	Ig-like C2-type1
	Ceacam6(D)/Ceacam8(A)	121	Intron2 (5643)	Exon3 (5764)	Ig-like C2-type1
	PSG1(D)/PSG2(A)	329	Intron1 (1426)	Exon2 (1755)	Ig-like V-type1
	PSG1(?)/PSG4(?)	812	Intron5 (11763)	Exon6 (12575)	Cytoplasmic
	PSG1(A)/PSG8(D)	822	Intron5 (11755)	Exon6 (12577)	Cytoplasmic
	PSG1(?)/PSG10P(?)	570	Intron1 (832)	Intron1 (1402)	NA
	PSG2(A)/PSG3(D)	1040	Intron4 (11152)	Intron4 (12192)	NA
	PSG2(?)/PSG5(?)	864	Intron4 (11185)	Intron4 (12049)	NA
	PSG2(?)/PSG9(?)	894	Intron4 (11152)	Intron4 (12046)	NA
	PSG2(?)/PSG11(?)	239	Intron2 (3710)	Intron2 (3949)	NA
	PSG3(?)/PSG5(?)	929	Intron5 (11553)	Intron5 (12482)	NA
	PSG3(?)/PSG8(?)	886	Intron3 (8111)	Intron3 (8997)	NA
	PSG3(?)/PSG9(?)	1036	Intron5 (11443)	Intron5 (12479)	NA
	PSG3(?)/PSG11(?)	945	Intron5 (11506)	Intron5 (12451)	NA
	PSG4(?)/PSG5(?)	726	Exon3 (7470)	Intron3 (8196)	Ig-like C2-type1
	PSG5(?)/PSG9(?)	162	Intron4 (12986)	Intron4 (13148)	NA
	PSG5(?)/PSG10P(?)	856	Intron4(11717)	Intron4 (12573)	NA
	PSG6(D)/PSG7(A)	219	Intron2 (2261)	Intron2 (2480)	NA
	PSG6(?)/PSG9(?)	161	Intron5 (13003)	Intron5 (13164)	NA
	PSG7(?)/PSG8(?)	549	Intron5 (11822)	Exon6 (12371)	Cytoplasmic
	PSG11(?)/PSG10P(?)	481	Exon4 (13879)	Intron4 (14360)	Transmembrane
Pan troglodytes	Ceacam6(D)/Ceacam1-like(A)	533	Intron1 (964)	Intron2 (1497)	Ig-like V-type1
	Ceacam5(D)/Ceacam1-like(A)	853	Intron1 (591)	Exon2 (1444)	Ig-like V-type1
	Ceacam6(D)/Ceacam5(A)	216	Intron1 (656)	Intron1 (872)	NA
	Ceacam6(?)/Ceacam7(?)	498	Intron2 (5587)	Intron3 (6085)	Ig-like C2-type1
	Ceacam6(D)/Ceacam8(A)	429	Intron2 (5636)	Intron3 (6065)	Ig-like C2-type1
	Ceacam7(A)/Ceacam1-like(D)	207	Exon3 (4420)	Intron3 (4627)	Ig-like C2-type1
	PSG3(A)/PSG8(D)	274	Intron3 (9766)	Intron3(10040)	NA
	PSG4(A)/PSG6-like(D)	1130	Intron2 (4900)	Intron2 (6030)	NA
	PSG6(A)/PSG8(D)	972	Intron2 (4569)	Intron2 (5541)	NA
	PSG1(A)/PSG11-like(D)	899	Intron4 (13979)	Intron4 (14878)	NA
	PSG9(D)/PSG4(A)	919	Intron2 (4796)	Intron2 (5715)	NA
	PSG3(D)/PSG1(A)	897	Intron5 (11496)	Intron5 (12393)	NA
Macaca mulatta	Ceacam4(D)/Ceacam5-likeA(A)	588	Intron5 (9386)	Intron6 (9974)	Cytoplasmic
	Ceacam5-likeB(A)/Ceacam8(D)	162	Intron3 (4249)	Exon4 (4411)	Ig-like C2-type2
	PSG6-like(A)/PSG2C(D)	338	Intron2 (4447)	Intron2 (4785)	NA
	PSG7-like(A)/PSG2A(D)	380	Intron2 (4873)	Intron2 (5253)	NA
Pongo abelii	Ceacam5A(D)/Ceacam6(A)	220	Exon4 (7904)	Intron4 (8124)	Ig-like C2-type2
	Ceacam5B(D)/Ceacam3(A)	622	Intron8 (97361)	Intron8 (97983)	NA
	Ceacam6(A)/Ceacam3(D)	621	Intron1 (1546)	Intron2 (2167)	Ig-like V-type1
	Ceacam5A(D)/Ceacam3(A)	181	Intron1 (661)	Intron1 (842)	NA
	PSG6-like(D)/PSG3-like(A)	541	Intron3 (8802)	Intron3 (9343)	NA
Nomascus leucogenys	Ceacam7(D)/Ceacam8(A)	236	Intron2 (4713)	Exon3 (4949)	Transmembrane
	Ceacam4(A)/Ceacam8(D)	276	Intron2 (63787)	Intron2 (64063)	NA
	PSG6-likeA(D)/PSG8-like(A)	419	Exon1 (1)	Intron1 (420)	Signal peptide
	PSG2-like(D)/PSG6-likeB(A)	225	Intron2 (2235)	Intron2 (2460)	NA
	PSG6-likeB(D)/PSG8-like(A)	321	Intron2 (2187)	Intron2 (2506)	NA
	PSG2-like(A)/PSG8-like(D)	160	Intron2 (2251)	Intron2 (2411)	NA

Notes. (A) and (D) indicate acceptor and donor sequences respectively. Interrogation points (?) indicate conversions for which the polarity could not be determined. NA: not applicable.

Table 3.2. dN/dS ratios (ω) of Ig-like V-type 1 and Ig-like C2-type 1 domains.

Genes	Pairwise species	Ig-like V-type1 domain			Ig-like C2-type 1 domain		
		dN	dS	ω	dN	dS	ω
CEACAM subgroup							
Ceacam1	Human/Chimpanzee	0.0178	0.0111	1.60	0.0105	0.0302	0.35
	Human/Orang-utan	0.1044	0.0587	1.78	0.0321	0.0149	2.15
	Chimpanzee/Orang-utan	0.1072	0.0442	2.42	0.0426	0.0154	2.77
Ceacam3	Human/Chimpanzee	0.0459	0.022	2.09	NA	NA	NA
	Human/Orang-utan	0.1201	0.0573	2.10	NA	NA	NA
	Chimpanzee/Orang-utan	0.0949	0.0565	1.68	NA	NA	NA
Ceacam4	Human/Chimpanzee	0	0	0	NA	NA	NA
	Human/Orang-utan	0.005	0.0194	0.26	NA	NA	NA
	Chimpanzee/Orang-utan	0.005	0.0194	0.26	NA	NA	NA
Ceacam5	Human/Chimpanzee	0.0318	0.0204	1.56	0.0163	0.0002	81.5
	Human/Orang-utan	0.1202	0.0849	1.42	0.0604	0.0159	3.80
	Chimpanzee/Orang-utan	0.1222	0.0816	1.50	0.0665	0.0155	4.29
Ceacam6	Human/Chimpanzee	0.0044	0.0225	0.20	0	0.0131	0
	Human/Orang-utan	0.0445	0.0985	0.45	0.0261	0.0663	0.39
	Chimpanzee/Orang-utan	0.0409	0.1194	0.34*	0.026	0.0494	0.53
Ceacam7	Human/Chimpanzee	0.0119	0.0161	0.74	0.0056	0.0001	56
	Human/Orang-utan	0.0492	0.0318	1.55	0.0293	0.0121	2.42
	Chimpanzee/Orang-utan	0.0449	0.0483	1	0.0355	0.0121	2.93
Ceacam8	Human/Chimpanzee	0	0.0092	0	0	0.0123	0
	Human/Orang-utan	0.0404	0.0237	1.70	0.0166	0.0561	0.30
	Chimpanzee/Orang-utan	0.0408	0.0121	3.37	0.0164	0.0424	0.39
Ceacam16	Human/Chimpanzee	0	0.0241	0*	0.0049	0.0244	0.2
	Human/Orang-utan	0.0106	0.1071	0.10**	0.0001	0.1038	0**
	Chimpanzee/Orang-utan	0.0105	0.1389	0.08**	0.0049	0.126	0.04**
Ceacam18	Human/Chimpanzee	0.0188	0.0349	0.54	0.0077	0.0001	77
	Human/Orang-utan	0.1019	0.1983	0.51*	0.0156	0.0542	0.29
	Chimpanzee/Orang-utan	0.0952	0.1813	0.52*	0.0236	0.0546	0.43
Ceacam19	Human/Chimpanzee	0	0.0108	0	NA	NA	NA
	Human/Orang-utan	0.0045	0.0216	0.21	NA	NA	NA
	Chimpanzee/Orang-utan	0.0045	0.0106	0.42	NA	NA	NA
PSG subgroup							
PSG3	Human/Chimpanzee	0.0161	0.0002	80.5	0.0056	0.0123	0.45
	Human/Orang-utan	0.0537	0.0451	1.19	0.0469	0.1131	0.41
	Chimpanzee/Orang-utan	0.0397	0.0456	0.87	0.0543	0.0826	0.66
PSG6	Human/Chimpanzee	0.0596	0.1004	0.59	0.0898	0.0329	2.73
	Human/Orang-utan	0.0518	0.1068	0.48	0.0731	0.049	1.50
	Chimpanzee/Orang-utan	0.0406	0.0354	1.15	0.1328	0.047	2.82

Notes. NA: not applicable.*, Significant at the 5% level; **, significant at the 0.1% level.

Figure legends

Note that Figures 3.1, 3.2, 3.3 and 3.5 are intended to be reproduced in color on the Web (free of charge) and in black-and-white in print.

Figure 3.1. Structure and tissue expression of human CEACAM proteins. Adapted from http://www.carcinoembryonic-antigen.de/.

Figure 3.2. Structure of human PSG proteins. Adapted from http://www.carcinoembryonic-antigen.de/.

Figure 3.3. Gene organisation of the CEA genes on human, chimpanzee, rhesus monkey, sumatran orang-utan and northern white-cheeked gibbon genomes. The organisation of the Ceacam and Ceacam-like genes (red), Ceacam pseudogenes (green) and PSG and PSG-like genes (blue) are shown for the five genomes. The location and orientation of all CEA genes are indicated below the map. Arrows indicate the polarity of gene conversions. Lines without arrows represent gene conversions for which the polarity could not be determined.

Figure 3.4. Phylogenetic tree of human CEA proteins. The scale bar represents 5% difference and the numbers on the nodes represent the bootstrap values. Similarity values refer to the pairwise similarity of the DNA sequences which are part of each of the groups indicated. Arrows indicate the polarity of gene conversions. Lines without arrows represent gene conversions for which the polarity could not be determined.

Figure 3.5. Graphical representation of the structure of CEACAM (A) and PSG (B) proteins and the distribution of gene conversions affecting protein coding regions between the five primate species. The domain organisations of proteins are built based on protein information coming from protein annotations in Uniprot database. Boxes with a full line represent domains found in all CEACAM (A) and PSG (B) proteins whereas those with a dashed line represent domains found only in some CEA proteins. TM: Transmembrane region. ITIM: Immunoreceptor tyrosine-based inhibition motif. ITAM: Immunoreceptor tyrosine-based activation motif.

Figure 3.1

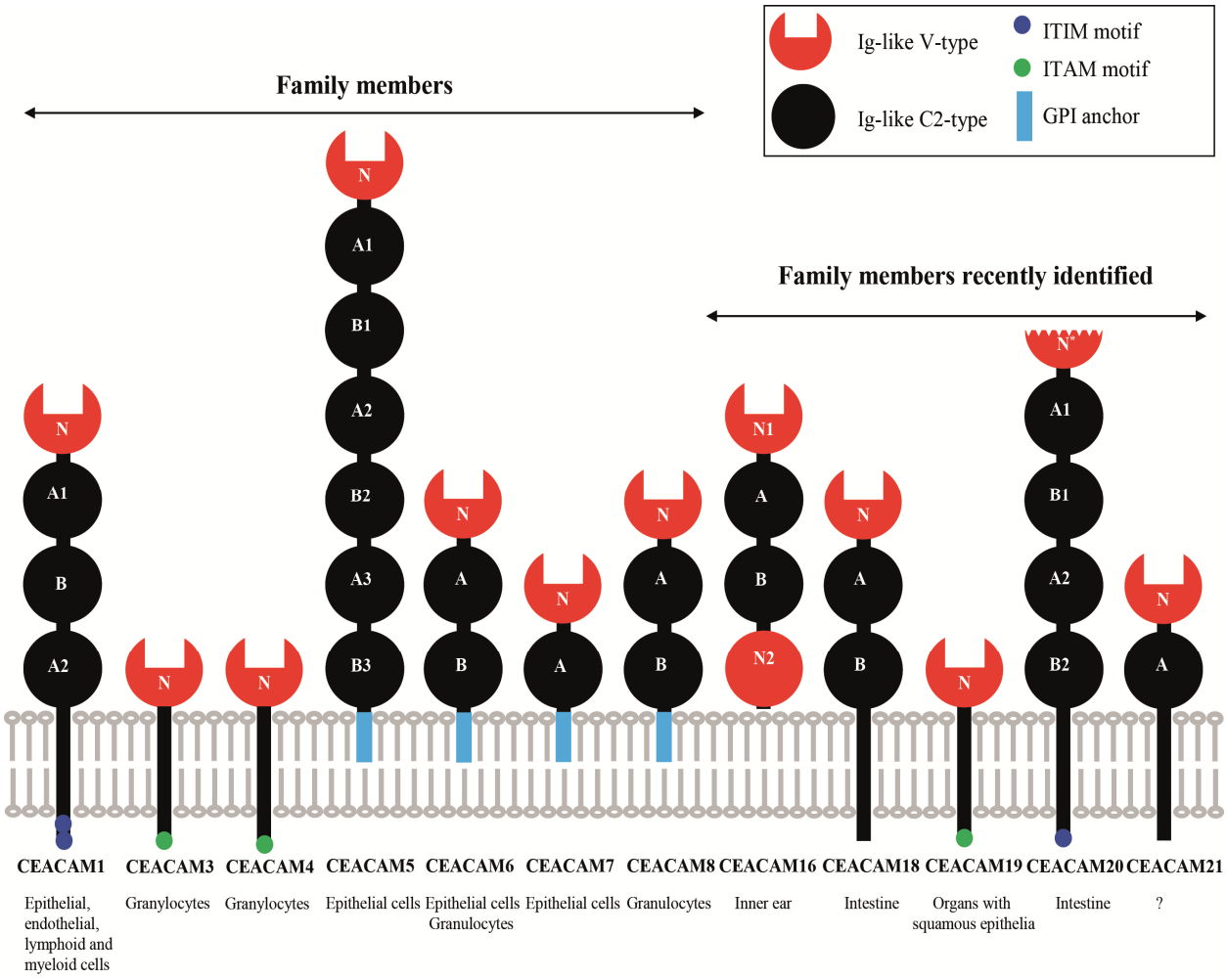

Ig-like V-type
Ig-like C2-type
ITIM motif
ITAM motif
GPI anchor
Family members
Family members recently identified
CEACAM1
CEACAM3
CEACAM4
CEACAM5
CEACAM6
CEACAM7
CEACAM8
CEACAM16
CEACAM18
CEACAM19
CEACAM20
CEACAM21
Epithelial, endothelial, lymphoid and myeloid cells
Granylocytes
Granylocytes
Epithelial cells
Epithelial cells Granulocytes
Epithelial cells
Granulocytes
Inner ear
Intestine
Organs with squamous epithelia
Intestine
?

Figure 3.2

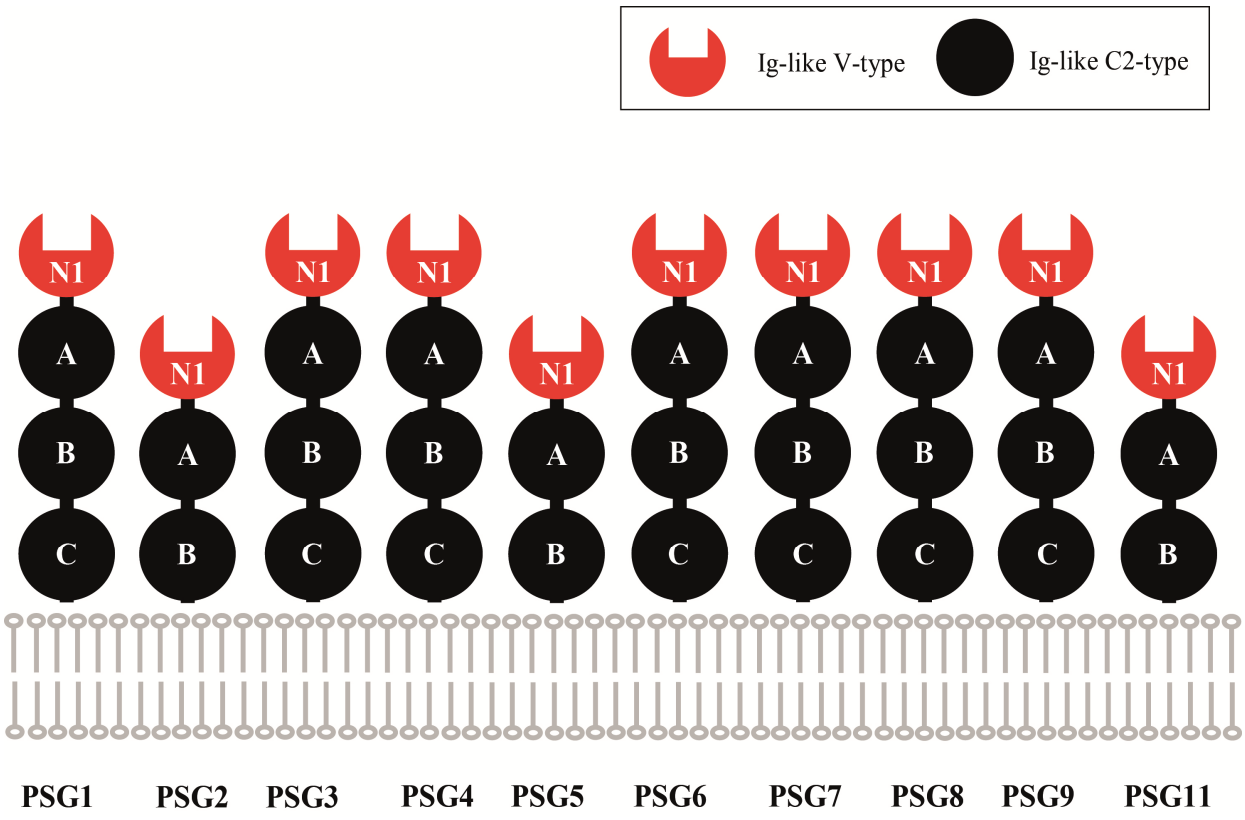

Ig-like V-type
Ig-like C2-type
N1
A
B
C
PSG1
PSG2
PSG3
PSG4
PSG5
PSG6
PSG7
PSG8
PSG9
PSG11

Figure 3.3

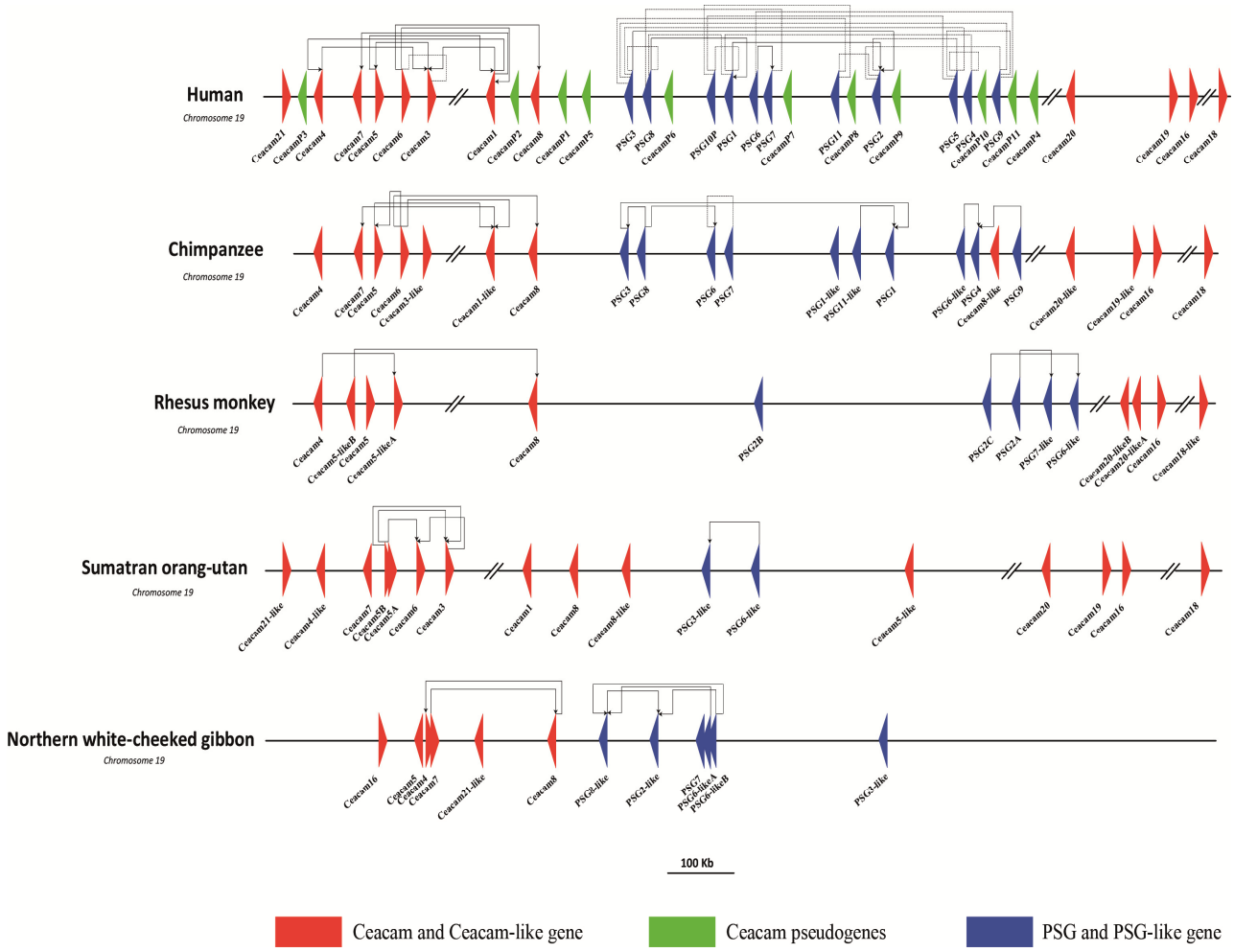

Human
Chromosome 19
Chimpanzee
Chromosome 19
Rhesus monkey
Chromosome 19
Sumatran orang-utan
Chromosome 19
Northern white-cheeked gibbon
Chromosome 19
100 Kb
Ceacam and Ceacam-like gene
Ceacam pseudogenes
PSG and PSG-like gene

Figure 3.4

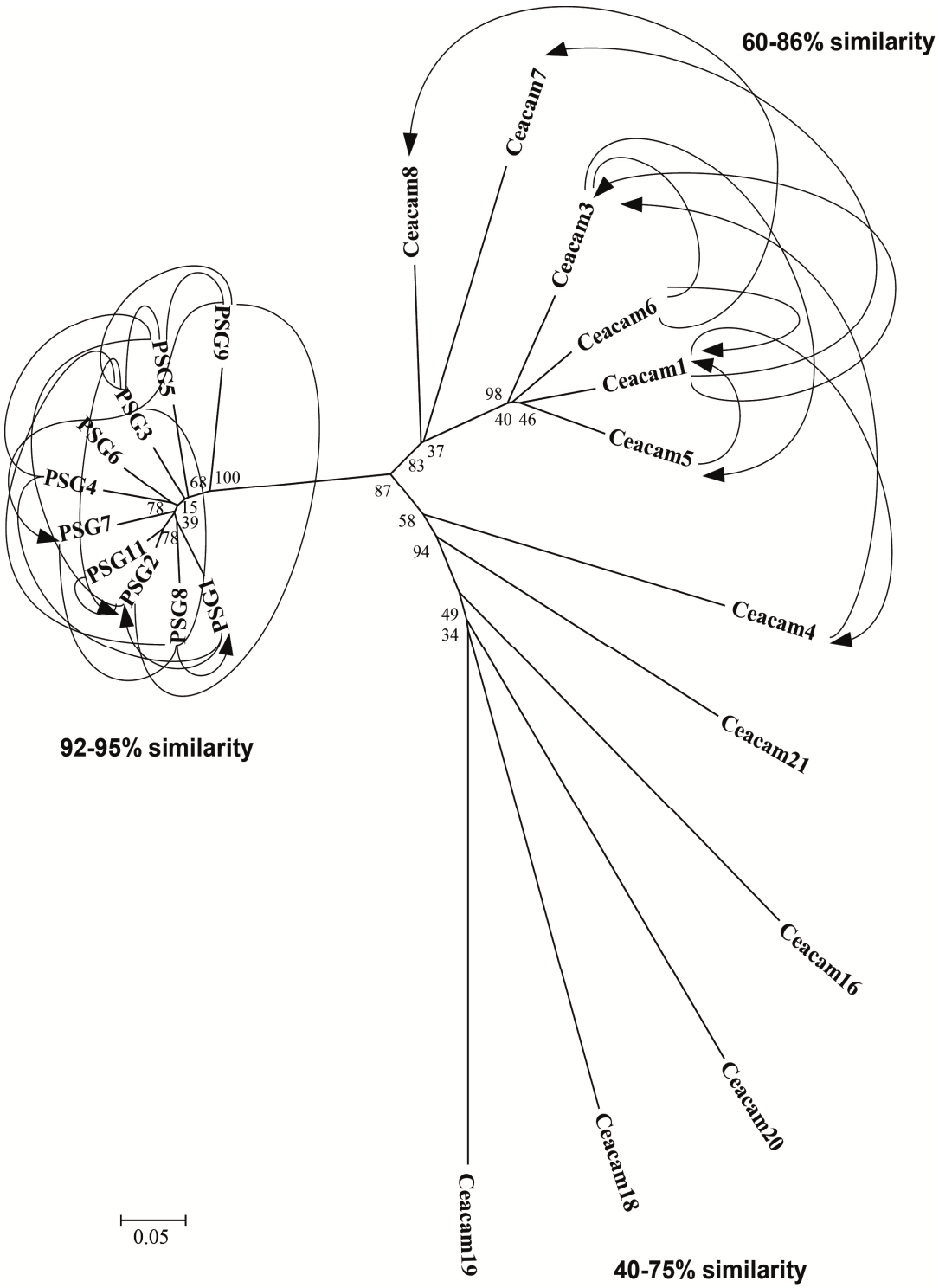
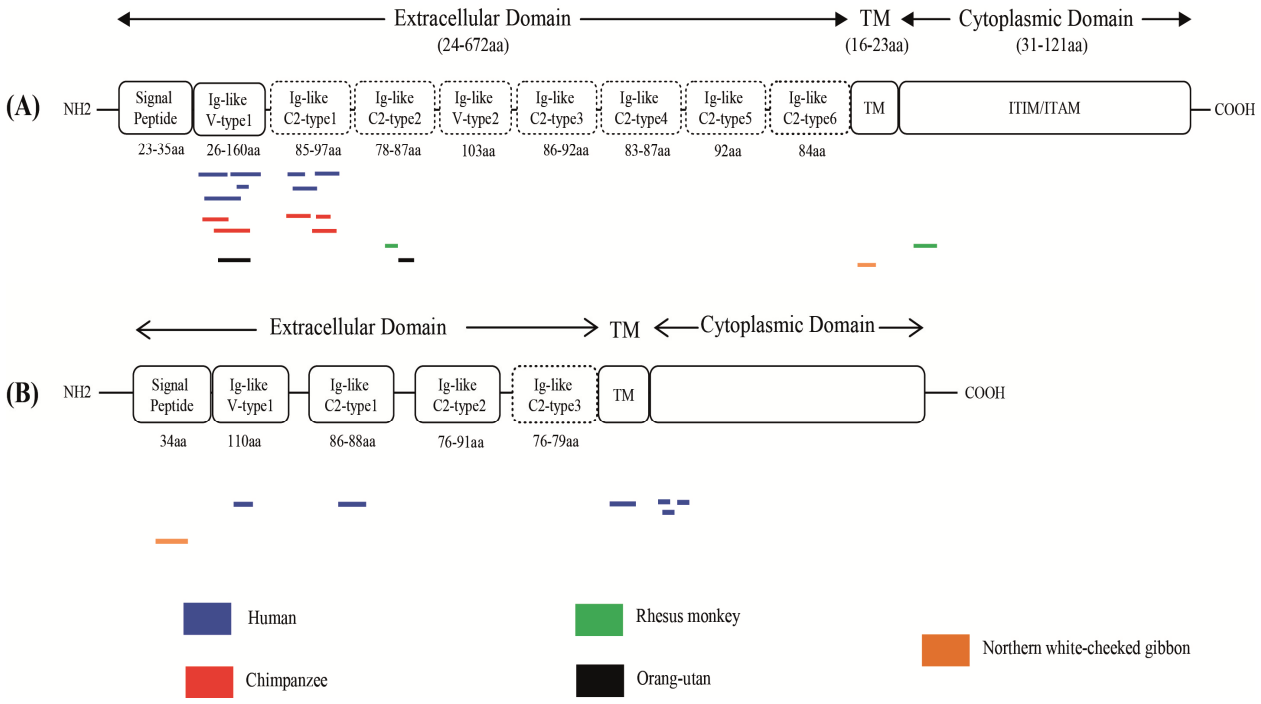

60-86% similarity
Ceacam7
Ceacam8
Ceacam3
Ceacam6
Ceacam1
Ceacam5
Ceacam4
Ceacam21
Ceacam16
Ceacam20
Ceacam18
Ceacam19
40-75% similarity
PSG9
PSG5
PSG3
PSG6
PSG4
PSG7
PSG11
PSG2
PSG8
PSG1
92-95% similarity
98
40
46
37
83
87
58
94
49
34
68
100
78
15
39
78
0.05

Figure 3.5


Extracellular Domain
(24-672aa)
TM
(16-23aa)
Cytoplasmic Domain
(31-121aa)
(A)
NH2
Signal Peptide
Ig-like V-type1
Ig-like C2-type1
Ig-like C2-type2
Ig-like V-type2
Ig-like C2-type3
Ig-like C2-type4
Ig-like C2-type5
Ig-like C2-type6
TM
ITIM/ITAM
COOH
23-35aa
26-160aa
85-97aa
78-87aa
103aa
86-92aa
83-87aa
92aa
84aa
Extracellular Domain
TM
Cytoplasmic Domain
(B)
NH2
Signal Peptide
Ig-like V-type1
Ig-like C2-type1
Ig-like C2-type2
Ig-like C2-type3
TM
COOH
34aa
110aa
86-88aa
76-91aa
76-79aa
Human
Chimpanzee
Rhesus monkey
Orang-utan
Northern white-cheeked gibbon

Supplemental Table 3.1. Similarity between human CEA protein sequences.

Pairwise comparisons		Similarity (%)							
Gene1	**Gene2**	**CP**	**ECD**	**TCD**	**Gene1**	**Gene2**	**CP**	**ECD**	**TCD**
CEACAM subgroup									
Ceacam1	Ceacam3	61	85	25	Ceacam4	Ceacam19	20	27	12
Ceacam1	Ceacam4	40	51	25	Ceacam4	Ceacam20	24	22	12
Ceacam1	Ceacam5	88	88	25	Ceacam4	Ceacam21	37	48	62
Ceacam1	Ceacam6	89	87	25	Ceacam5	Ceacam6	87	86	62
Ceacam1	Ceacam7	70	69	12	Ceacam5	Ceacam7	68	69	62
Ceacam1	Ceacam8	81	76	25	Ceacam5	Ceacam8	78	75	80
Ceacam1	Ceacam16	35	35	31	Ceacam5	Ceacam16	35	36	28
Ceacam1	Ceacam18	32	29	25	Ceacam5	Ceacam18	35	32	12
Ceacam1	Ceacam19	25	26	25	Ceacam5	Ceacam19	27	28	37
Ceacam1	Ceacam20	32	23	12	Ceacam5	Ceacam20	30	24	25
Ceacam1	Ceacam21	51	56	37	Ceacam5	Ceacam21	50	57	12
Ceacam3	Ceacam4	65	55	77	Ceacam6	Ceacam7	67	65	37
Ceacam3	Ceacam5	60	84	12	Ceacam6	Ceacam8	81	74	62
Ceacam3	Ceacam6	61	84	12	Ceacam6	Ceacam16	34	35	24
Ceacam3	Ceacam7	45	65	12	Ceacam6	Ceacam18	34	31	25
Ceacam3	Ceacam8	50	71	12	Ceacam6	Ceacam19	26	26	37
Ceacam3	Ceacam16	24	36	13	Ceacam6	Ceacam20	31	21	25
Ceacam3	Ceacam18	21	27	25	Ceacam6	Ceacam21	52	56	12
Ceacam3	Ceacam19	21	26	12	Ceacam7	Ceacam8	68	68	50
Ceacam3	Ceacam20	19	19	12	Ceacam7	Ceacam16	34	34	19
Ceacam3	Ceacam21	41	56	62	Ceacam7	Ceacam18	31	28	12
Ceacam4	Ceacam5	41	50	12	Ceacam7	Ceacam19	22	25	37
Ceacam4	Ceacam6	41	52	12	Ceacam7	Ceacam20	31	53	50
Ceacam4	Ceacam7	36	48	12	Ceacam7	Ceacam21	48	21	12
Ceacam4	Ceacam8	41	52	12	Ceacam8	Ceacam16	33	34	27
Ceacam4	Ceacam16	25	35	12	Ceacam8	Ceacam18	33	30	12
Ceacam4	Ceacam18	23	28	25	Ceacam8	Ceacam19	25	26	37

Ceacam8	Ceacam20	33	26	25	Ceacam18	Ceacam19	21	23	25
Ceacam8	Ceacam21	49	53	12	Ceacam18	Ceacam20	25	17	12
Ceacam16	Ceacam18	29	25	11	Ceacam18	Ceacam21	30	29	37
Ceacam16	Ceacam19	24	26	12	Ceacam19	Ceacam20	20	22	25
Ceacam16	Ceacam20	27	20	20	Ceacam19	Ceacam21	20	24	25
Ceacam16	Ceacam21	38	38	25	Ceacam20	Ceacam21	32	23	12
PSG subgroup									
PSG1	PSG2	86	88	44	PSG3	PSG11	86	88	33
PSG1	PSG3	88	90	44	PSG4	PSG5	87	88	44
PSG1	PSG4	86	94	67	PSG4	PSG6	87	88	44
PSG1	PSG5	84	85	55	PSG4	PSG7	90	90	78
PSG1	PSG6	85	88	44	PSG4	PSG8	90	91	89
PSG1	PSG7	87	88	78	PSG4	PSG9	82	82	56
PSG1	PSG8	91	92	78	PSG4	PSG11	83	87	22
PSG1	PSG9	80	81	56	PSG5	PSG6	86	88	33
PSG1	PSG11	82	86	22	PSG5	PSG7	86	88	44
PSG2	PSG3	87	87	89	PSG5	PSG8	85	87	44
PSG2	PSG4	86	88	55	PSG5	PSG9	82	84	44
PSG2	PSG5	86	86	88	PSG5	PSG11	83	86	33
PSG2	PSG6	85	87	44	PSG6	PSG7	88	91	44
PSG2	PSG7	88	88	44	PSG6	PSG8	87	89	56
PSG2	PSG8	87	88	56	PSG6	PSG9	86	86	78
PSG2	PSG9	79	81	56	PSG6	PSG11	85	88	11
PSG2	PSG11	90	93	33	PSG7	PSG8	91	91	59
PSG3	PSG4	89	91	55	PSG7	PSG9	82	84	56
PSG3	PSG5	89	89	78	PSG7	PSG11	86	88	33
PSG3	PSG6	90	91	44	PSG8	PSG9	83	85	67
PSG3	PSG7	89	90	44	PSG8	PSG11	84	87	22
PSG3	PSG8	90	91	56	PSG9	PSG11	79	81	11
PSG3	PSG9	84	86	44					

Notes. **CP**: Complete proteins; **ECD**: Extracellular domain; **TCD**: Transmembrane and cytoplasmic domains.

Supplemental Table 3.2. List of the CEA genes that were used in this study and their corresponding GenBank/Ensembl accession numbers.

Species	Accession number	Species	Accession number
Homo sapiens Ceacam1	M72238	*Pan troglodytes* PSG6	XM_512707
Homo sapiens Ceacam3	E03349	*Pan troglodytes* PSG6-like	XM_003316408
Homo sapiens Ceacam4	D90276	*Pan troglodytes* PSG7	XM_003316388
Homo sapiens Ceacam5	M17303	*Pan troglodytes* PSG8	XM_524284
Homo sapiens Ceacam6	M29541	*Pan troglodytes* PSG9	XM_001144286
Homo sapiens Ceacam7	X98311	*Pan troglodytes* PSG11-like	XM_003316397
Homo sapiens Ceacam8	D90064	*Macaca mulatta* Ceacam4	XM_001102867
Homo sapiens Ceacam16	BC144608	*Macaca mulatta* Ceacam5	NM_001047125
Homo sapiens Ceacam18	NM_001080405	*Macaca mulatta* Ceacam5-likeA	XM_001096258
Homo sapiens Ceacam19	AF406955	*Macaca mulatta* Ceacam5-likeB	XM_002801262
Homo sapiens Ceacam20	AY358129	*Macaca mulatta* Ceacam8	ENSMMUG00000013218
Homo sapiens Ceacam21	AK023602	*Macaca mulatta* Ceacam16	ENSMMUG00000010037
Homo sapiens CeacamP1	M96921	*Macaca mulatta* Ceacam18-like	XM_001114788
Homo sapiens CeacamP2	M96922	*Macaca mulatta* Ceacam20-likeA	XM_001103639
Homo sapiens CeacamP3	M96923	*Macaca mulatta* Ceacam20-likeB	XM_001103562
Homo sapiens CeacamP4	M96923	*Macaca mulatta* PSG2A	XM_001099377
Homo sapiens CeacamP5	U06672	*Macaca mulatta* PSG2B	XM_001098661
Homo sapiens CeacamP6	U06674	*Macaca mulatta* PSG2C	XM_001106054
Homo sapiens CeacamP7	U06675	*Macaca mulatta* PSG6-like	XM_001099683
Homo sapiens CeacamP8	U06676	*Macaca mulatta* PSG7-like	XM_001099482
Homo sapiens CeacamP9	U06677	*Pongo abelii* Ceacam1	NM_001132423
Homo sapiens CeacamP10	U06678	*Pongo abelii* Ceacam3	ENSPPYG00000010026
Homo sapiens CeacamP11	U06679	*Pongo abelii* Ceacam4-like	ENSPPYG00000010021
Homo sapiens PSG1	NM_006905	*Pongo abelii* Ceacam5A	XM_002829277
Homo sapiens PSG2	NM_031246	*Pongo abelii* Ceacam5B	XM_002829277
Homo sapiens PSG3	NM_021016	*Pongo abelii* Ceacam5-like	ENSPPYG00000010053
Homo sapiens PSG4	NM_213633	*Pongo abelii* Ceacam6	ENSPPYG00000010024
Homo sapiens PSG5	NM_002781	*Pongo abelii* Ceacam7	XM_002829276
Homo sapiens PSG6	NM_002782	*Pongo abelii* Ceacam8	ENSPPYG00000010051
Homo sapiens PSG7	NM_001206650	*Pongo abelii* Ceacam8-like	XM_003779192
Homo sapiens PSG8	M74106	*Pongo abelii* Ceacam16	XM_002829364
Homo sapiens PSG9	M34481	*Pongo abelii* Ceacam18	XM_003780746
Homo sapiens PSG10P	L14724	*Pongo abelii* Ceacam18-like	ENSPPYG00000010323
Homo sapiens PSG11	U25988	*Pongo abelii* Ceacam19	XM_003779199
Pan troglodytes Ceacam1-like	ENSPTRG00000011061	*Pongo abelii* Ceacam20	XM_002829372
Pan troglodytes Ceacam3-like	XM_003316369	*Pongo abelii* Ceacam21	XM_002829273
Pan troglodytes Ceacam4	XM_512688	*Pongo abelii* PSG3-like	ENSPPYG00000010052
Pan troglodytes Ceacam5	XM_003316366	*Pongo abelii* PSG6-like	XM_002829303
Pan troglodytes Ceacam6	XM_003316368	*Nomascus leucogenys* Ceacam5	XM_003270326
Pan troglodytes Ceacam7	XM_003316365	*Nomascus leucogenys* Ceacam4	ENSNLEG00000013590
Pan troglodytes Ceacam8	XM_512705	*Nomascus leucogenys* Ceacam7	ENSNLEG00000013606
Pan troglodytes Ceacam8-like	XM_003316414	*Nomascus leucogenys* Ceacam8	ENSNLEG00000010809
Pan troglodytes Ceacam16	XM_003316443	*Nomascus leucogenys* Ceacam16	ENSNLEG00000001672
Pan troglodytes Ceacam18	XM_00331665	*Nomascus leucogenys* Ceacam21-like	XM_003270513
Pan troglodytes Ceacam19-like	XM_003316471	*Nomascus leucogenys* PSG2-like	XM_003280969
Pan troglodytes Ceacam20-like	XM_003316470	*Nomascus leucogenys* PSG3-like	XM_003280963
Pan troglodytes PSG1	XM_001155353	*Nomascus leucogenys* PSG6-likeA	XM_003280972
Pan troglodytes PSG1-like	XM_003316393	*Nomascus leucogenys* PSG6-likeB	XM_003280971
Pan troglodytes PSG3	XM_00331638	*Nomascus leucogenys* PSG7	XM_003280970
Pan troglodytes PSG4	XM_003339321	*Nomascus leucogenys* PSG8-like	XM_003280968

Chapter 4. General conclusions

The present study examines the effect of gene conversion events in two members of the immunoglobulin superfamily (Siglec and CEA) of five primate species. Analysis of Siglec family shows that gene conversions are frequent events in five primate species within the CD33-related Siglec genes, but not in other conserved Siglec genes (Table 2.1 Chapter 2). This finding is due to the fact that the CD33-related Siglec genes share a high degree of sequence similarity (about 55–98%). This is supported by the fact that there is a strong positive correlation between the length of the conversions and the similarity of the converted regions ($r = 0.52$, $p = 0.002$). In a total of 77 Siglec genes, we detected 33 gene conversion events in the five studied primate species and their average size (± standard deviation) was 590.21 ± 429.34 nucleotides (Table 2.1 Chapter 2). While, in the CEA gene family, a total of 55 gene conversion events were detected and their average length (± standard deviation) was 523 ± 296 nucleotides (Table 3.1 Chapter 3). In this family, half of gene conversion events are located in the coding regions and these conversions occur between PSG genes, seven CEACAM genes (*CEACAM1*, *CEACAM3* thought *CEACAM8*) and never between the five recently identified CEACAM genes (*CEACAM16*, *CEACAM18* thought *CEACAM21*). The absence of conversions between the five CEACAM genes and the conserved Siglec genes is likely due to their high degree of sequence divergence. Additionally, the frequency of gene conversion is higher for CD33-related Siglec genes (16%) than for carcinoembryonic antigen genes in five primate species (5%). These two frequencies are much higher than the average frequency of gene conversion events that occurred between all human genes (0.88%, Benovoy and Drouin 2009).

In both families, gene conversions are more frequent between closely linked genes (Chapter 2 and Chapter 3). For the Siglec family, there is a weak, but significant, correlation between the distance between genes and the frequency of conversions when considering the data from all species (Chapter 2). The same result was found in the CEA genes where significant correlations were found between the distance between CEA genes and the frequency of conversions in the human, chimpanzee, rhesus monkey and Sumatran orang-utan genomes (Chapter 3).

In each gene family studied here, there is an increase in GC-content in converted regions compared to non-converted regions (Chapter 2 and Chapter 3). An explanation for this finding is that the DNA repair machinery is baised towards the incorporation of guanine and cytosine over adenine and thymine (Eyre-Walker 1993; Galtier et al. 2001). Our findings are consistent with previous studies that have shown biased gene conversion (BGC) leads to increases in the GC content (Benovoy et al. 2005; Pessia et al. 2012).

In the two families, most of gene conversions occur more frequently in the extracellular Ig-like domains, and rarely in the transmembrane regions and in the cytoplasmic tail (Figure 2.4, Chapter 2 and Figure 3.5, Chapter 3). These finding suggest that the high frequency of gene conversions in the extracellular Ig-like domains is only a consequence of their high degree of sequence similarity, and is not the result of positive or diversifying selection, as suggested by previously studies (Flajnik 2002; Varki 2010; Wang et al. 2012).

Finally, our results suggest that gene conversions that occur between CD33-related Siglec genes and CEA genes are not adaptive, because gene conversion events only occur between highly similar gene sequences and are not

more frequent within Ig-like V-type 1 domains than within the adjacent Ig-like C2-type 1 domains. Furthermore, except for one Siglec gene, our results show that selective pressure acting on Ig-like V-type 1 and Ig-like C2-type 1 domains are either negative or neutral. Our results therefore do not support previous suggestions that gene conversions between Siglec and CEA genes are adaptive.

Literature cited

Benovoy D, Morris RT, Morin A, Drouin G. 2005. Ectopic gene conversions increase the G + C content of duplicated yeast and Arabidopsis genes. *Mol Biol Evol*. 22: 1865–1868.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc Biol Sci.* 252: 237–243.

Flajnik MF. 2002. Comparative analyses of immunoglobulin genes: surprises and portents. *Nat Rev Immunol*. 9: 688–698.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159: 907–911.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol*. 4: 675–682.

Varki, A. 2010. Uniquely Human Evolution of Sialic Acid Genetics and Biology. *Proc Natl Acad Sci*. 107: 8939–8946.

Wang X, Mitra N, Cruz P, Deng L; NISC Comparative Sequencing Program, Varki N, Angata T, Green ED, Mullikin J, Hayakawa T, Varki A. 2012. Evolution of Siglec-11 and Siglec-16 Genes in Hominins. *Mol Biol Evol*. 29: 2073–2086.