The Impact of Natural Selection on the Evolution and Function of Placentally Expressed Galectins

Zackery A. Ely[1], Jiyun M. Moon[1], Gregory R. Sliwoski[1], Amandeep K. Sangha[2,3], Xing-Xing Shen[1], Abigail L. Labella[1], Jens Meiler[2,3], John A. Capra[1,4], and Antonis Rokas [1,4,*]

[1]Department of Biological Sciences, Vanderbilt University

[2]Department of Chemistry, Vanderbilt University

[3]Center for Structural Biology, Vanderbilt University

[4]Department of Biomedical Informatics, Vanderbilt University School of Medicine

*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu.

Abstract

Immunity genes have repeatedly experienced natural selection during mammalian evolution. Galectins are carbohydrate-binding proteins that regulate diverse immune responses, including maternal–fetal immune tolerance in placental pregnancy. Seven human galectins, four conserved across vertebrates and three specific to primates, are involved in placental development. To comprehensively study the molecular evolution of these galectins, both across mammals and within humans, we conducted a series of between- and within-species evolutionary analyses. By examining patterns of sequence evolution between species, we found that primate-specific galectins showed uniformly high substitution rates, whereas two of the four other galectins experienced accelerated evolution in primates. By examining human population genomic variation, we found that galectin genes and variants, including variants previously linked to immune diseases, showed signatures of recent positive selection in specific human populations. By examining one nonsynonymous variant in Galectin-8 previously associated with autoimmune diseases, we further discovered that it is tightly linked to three other nonsynonymous variants; surprisingly, the global frequency of this four-variant haplotype is ~50%. To begin understanding the impact of this major haplotype on Galectin-8 protein structure, we modeled its 3D protein structure and found that it differed substantially from the reference protein structure. These results suggest that placentally expressed galectins experienced both ancient and more recent selection in a lineage- and population-specific manner. Furthermore, our discovery that the major Galectin-8 haplotype is structurally distinct from and more commonly found than the reference haplotype illustrates the significance of understanding the evolutionary processes that sculpted variants associated with human genetic disease.

Key words: galectins, population genetics, positive selection, comparative modeling, human evolution, mammalian evolution.

Introduction

Galectins are carbohydrate-binding proteins that play diverse and crucial roles in immunity and development (Houzelstein et al. 2004; Rabinovich and Toscano 2009; Than et al. 2015). The immunoregulatory functions of galectins span both innate and adaptive immunity and include—but are not limited to—pattern recognition, T cell-mediated immunosuppression, regulation of acute and chronic inflammation, and selective autophagy in response to pathogen invasion (Rabinovich and Toscano 2009; Vasta 2012). In mammalian development, galectins facilitate implantation and angiogenesis, and they regulate immune tolerance at the maternal–fetal interface (Than et al. 2015).

Of the 13 galectins found in humans, seven are expressed in the placenta and regulate both immunity and development. All seven placentally expressed galectins fulfill critical roles in pregnancy, from facilitating implantation events (such as blastocyst attachment to the uterine epithelium) to regulating maternal–fetal immune tolerance through apoptosis of activated T cells (Than et al. 2015). Through this regulation of activated T cells, galectins help protect fetal tissues from attack by

maternal immune cells (Than et al. 2012), thereby establishing a healthy environment for fetal development.

The seven placentally expressed galectins can be divided into two groups on the basis of their phylogenetic conservation and patterns of tissue expression. The first group, which we term the "ancient" group, is comprised of *LGALS1* (which makes the Galectin-1 protein), *LGALS3* (Galectin-3), *LGALS8* (Galectin-8), and *LGALS9* (Galectin-9); all these genes are conserved across vertebrates and expressed in most human tissues, including the placenta (Houzelstein et al. 2004). In contrast, the second group, which we term the "placental cluster" group, is comprised of *LGALS13* (Galectin-13), *LGALS14* (Galectin-14), and *LGALS16* (Galectin-16); these three genes are only present in primates, formed via gene duplication, are found next to each other in a region of human chromosome 19 (hence the "cluster" term), and appear to be specifically expressed in the placenta (Than et al. 2009).

All proteins in the galectin family mediate their functions through a conserved carbohydrate recognition domain (CRD) comprised of ~130 amino acids (Rabinovich and Toscano 2009). CRD structure and organization divide the galectin family into three major types: the prototype, the tandem-repeat type, and the chimera type. The seven placentally expressed galectins span all three subtypes (fig. 1A). Prototype galectins (all three in the placental cluster group and Galectin-1) contain a single CRD, whereas tandem-repeat type galectins (Galectin-8 and Galectin-9) contain two CRDs with distinct binding profiles connected by a peptide linker. The chimera type Galectin-3 contains one CRD joined to a collagen-like N-terminal domain. Although CRD structure and organization differs among galectin types, all galectin CRDs share eight, mostly invariant residues involved in carbohydrate binding (fig. 1B) (Cummings et al. 2017). CRD topology is well-conserved among galectins, with most family members possessing a CRD comprised of a β-sandwich of five-stranded and six-stranded antiparallel β-sheets (Than et al. 2015). Despite high structural conservation, the CRD amino acid sequence differs enough among family members to confer distinct carbohydrate-binding preferences (Rabinovich and Toscano 2009).

Based on the structural conservation of the galectin CRD and the crucial roles galectins play in fundamental cellular processes, previous studies have hypothesized that mammalian galectins evolved under strong functional constraints that prevented dramatic changes in the CRD (Vasta 2012). In contrast, based on the crucial roles galectins play in placental development, other studies have hypothesized that galectins experienced positive selection in early placental mammalian evolution, followed by strong evolutionary conservation thereafter (Than et al. 2008). At present, these hypotheses have been tested via examination of evolutionary rates in only Galectin-1 and the placental cluster galectins (Than et al. 2008, 2009). For example, evidence for positive selection at the origin of placental mammals followed by purifying

selection thereafter has been reported for Galectin-1 (Than et al. 2008), but this analysis did not include the other ancient galectins: Galectin-3, Galectin-8, and Galectin-9. Thus, a significant part of placental galectin evolutionary history remains unexamined, and a comprehensive analysis of the evolutionary rates of all placentally expressed galectins could inform us of the tempo and mode of placental galectin evolution during placental mammalian evolution.

Little is known about how galectins evolved *within* the human species. This is surprising given that DNA variants of human immune genes often bear strong signatures of positive selection due to the selective pressures applied by pathogens on the human genome (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2011; Vitti et al. 2013). For example, DARC encodes a membrane protein used by the malaria parasite, *Plasmodium vivax*, for entry into red blood cells (Barreiro and Quintana-Murci 2010). A null mutation in this gene eliminates protein expression and thereby confers full resistance to *P. vivax* in homozygotes (Hamblin et al. 2002). Positive selection has targeted the null allele in a population-specific manner, bringing it to fixation in sub-Saharan Africa while it remains virtually absent elsewhere (Hamblin et al. 2002). Human population genetic analyses, such as detection of gene variants highly differentiated between populations, can link disease-associated variants to signatures of natural selection as well as identify tradeoffs (Raj et al. 2013; Ramos et al. 2014, 2015), but such approaches have not been used to study placental galectin evolution. This is surprising given that several galectin gene variants have known links to many diseases such as breast cancer and rheumatoid arthritis (fig. 1C) (Balan et al. 2008; Hu et al. 2011). Examination of the population genetics of galectins in humans could connect these known disease associations to signatures of natural selection and pinpoint new directions for functional studies of galectin gene variants and their role in disease.

To gain insight into the evolution of placental galectins both across mammals as well as in the human population, we performed a comprehensive evolutionary analysis on the seven human galectin genes that regulate both immunity and placental development. We found that members of the ancient group had distinct patterns of nonsynonymous to synonymous substitution rates, suggesting that the functional constraints on galectin evolution were neither uniform across galectin family members nor constant throughout mammalian evolution. By examining patterns of genetic variation in human galectin genes, we identified gene-wide signatures of positive selection on two human galectins. We further found variant-specific signatures of population differentiation, suggestive of geographically restricted positive selection, for several single nucleotide polymorphisms (SNPs) found in galectin genes.

Although studying galectin gene variants, we also observed that the reference sequence commonly used when studying Galectin-8 is not the most common haplotype in
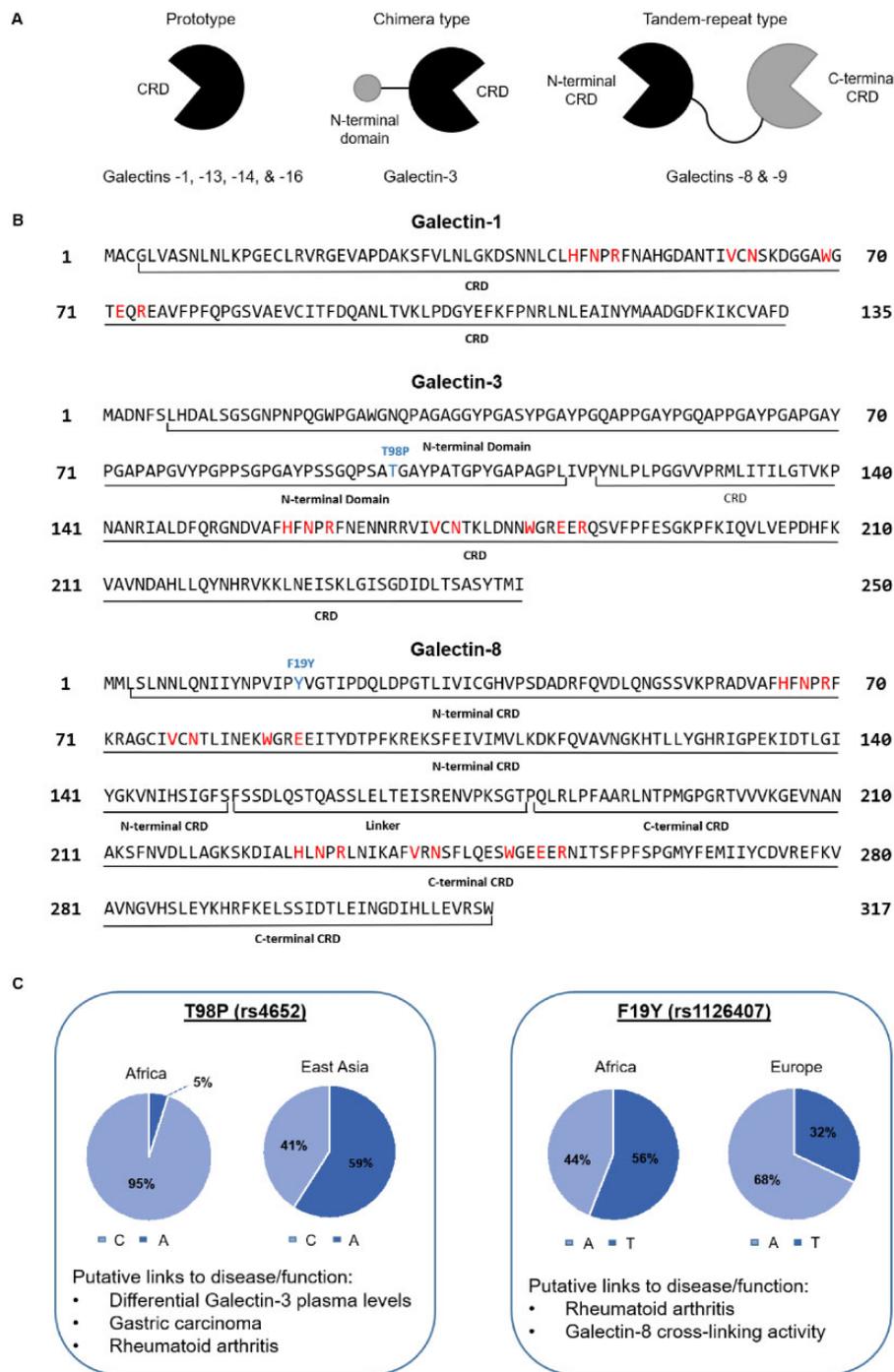
FIG. 1.—Galectin protein carbohydrate recognition domain (CRD) organization, classification, and sequence. (A) Classification of galectin subtypes based on CRD organization. Prototype galectins are defined by a single CRD, chimera type galectins are defined by a single CRD fused to an additional collagen-like domain, and tandem-repeat type galectins consist of two CRDs joined by a peptide linker of up to 70 amino acids. (B) Amino acid sequences of representatives of each of the three galectin classes. Residues directly involved in carbohydrate binding are highlighted in red and are mostly invariant across all galectin family members. The two residues highlighted in blue represent missense variants described in panel (C). (C) Population disparities and putative disease links of the common Galectin-3 single nucleotide variant, rs4652; the C allele, which is at 95% frequency in African genomes, is part of a codon for proline (Pro), whereas the A allele is part of a codon for threonine (Thr). Similarly, the A allele of the common Galectin-8 single nucleotide variant, rs1126407, is part of a codon for tyrosine (Tyr), whereas the T allele is part of a codon for phenylalanine (Phe).

the human population. The most common haplotype (hereafter the major haplotype) differs from the current reference sequence by four missense variants. These variants frequently co-occur, and one variant, rs1126407 (hereafter F19Y), is associated with rheumatoid arthritis (Pal et al. 2012). Interestingly, even though F19Y is most often found alongside the three other variants on the major haplotype, it is typically studied on the background of the reference haplotype. To examine differences between proteins encoded by the reference haplotype, major haplotype, and F19Y on the background of reference haplotype, we predicted and compared the protein structures of all three sequences using comparative modeling. The predicted structure of the major haplotype is structurally distinct from that of the reference haplotype as well as from that of F19Y on the background of the reference haplotype. This suggests that the co-occurrence of four variants on the major haplotype leads to structural differences not captured by studies examining F19Y in isolation, highlighting the importance of understanding the genomic background and evolutionary history of loci associated with human complex genetic disease.

Materials and Methods

Ortholog Retrieval and Sequence Alignment

We examined seven human galectins that regulate both immunity and placental development (supplementary table 5, Supplementary Material online). To construct their gene phylogenies, we retrieved orthologous protein sequences of all seven human genes across 21 placental mammals and five outgroup species from Ensembl (human genome assembly GRCh38.p12) (supplementary table 6, Supplementary Material online). By default, Ensembl provided the longest available isoform for each sequence. This can be problematic when evaluating signatures of positive selection, because sequence alignments based on the longest available isoforms tend to have increased amounts of misaligned positions that can lead to false detection of positive selection (Villanueva-Canas et al. 2013). Isoforms of similar length generate alignments of better quality and increase the accuracy of downstream molecular evolutionary analyses. Thus, we examined sequences that differed in length from the longest human isoform reported by Ensembl. These sequences were replaced if 1) Ensembl listed another isoform with a length more similar to the human isoform, or if 2) the NCBI Gene database listed an isoform with a more similar length. The accession numbers of all final isoforms included in the analysis are listed in supplementary table 7, Supplementary Material online.

Some species lacked annotated orthologs in the Ensembl database, presumably due to gene loss, so the final sequence files for each galectin varied in the total number of sequences (supplementary table 8, Supplementary Material online). This

was most common for placental cluster galectins, which have experienced dramatic episodes of gene loss and gain during the evolution of primates (Than et al. 2009). Some species had multiple co-orthologs reported for the same human gene; for example, Ensembl reported three horse genes as co-orthologs for human *LGALS9*. All co-orthologs were included in the final sequence files (supplementary table 6, Supplementary Material online).

Several nonprimate galectins, mostly in artiodactyl species, were reported as orthologs for all three placental cluster galectins. These genes were included in the phylogenetic analysis to evaluate the evolutionary relationship between primate placental galectins and their nonprimate counterparts. *LGALS13*, *LGALS14*, and *LGALS16* sequence files were combined into one sequence file with the nonprimate sequences. In rare cases, placental cluster orthologs reported by Ensembl included genes previously identified as galectin pseudogenes by Than et al. (2009). These genes were not included in the final sequence files (see supplementary table 9, Supplementary Material online, for a detailed description of all these exceptions and the corresponding reasons for their exclusion).

Protein sequences were aligned using the multiple sequence alignment software, MAFFT, version 7 (Katoh and Standley 2013). We produced five protein alignments: four for each galectin of the ancient group separately and one for the placental galectins and their nonprimate orthologs together. Corresponding codon-based alignments were obtained using the sequence alignment webserver, PAL2NAL, version 14 (Suyama et al. 2006). Gene trees were reconstructed for each codon alignment using RAxML, version 8.2.4 (Stamatakis 2014). All trees for the ancient galectins were rooted using the sequence from the frog, *Xenopus tropicalis*, as the outgroup. The placental cluster gene tree was rooted with the opossum, *Monodelphis domestica*, as the outgroup.

We examined whether species trees were congruent with the gene sequence trees. Species trees were obtained using the Common Tree tool from the NCBI Taxonomy Database. The species trees contained two polytomies resulting from ambiguities in the placement of horse and green monkey on the mammalian phylogeny. We resolved these polytomies into bipartitions by enforcing horse to be a sister group to a clade containing pig, cow, sheep, and goat and by enforcing green monkey to be a sister group to a clade containing olive baboon and Macaques, in accordance with Perelman et al. (2011) and Shen et al. (2016). We applied the approximately unbiased test to evaluate whether the gene trees differed significantly from the corresponding species trees (Shimodaira 2002). Only the *LGALS8* gene tree agreed with its species tree (supplementary table 10, Supplementary Material online). Thus, the *LGALS8* species tree was tested during the evolutionary rate analysis, while the gene tree

Table 1

Two Sets of Evolutionary Hypotheses Tested for the Rate of Evolution of Placental Galectin Genes

Hypotheses for ancient galectins	Parameters
Null: Uniform rate for all branches	$\omega_{\text{outgroups}}=\omega_{\text{placental stem}}=\omega_{\text{placental clade}}=\omega_{\text{primate clade}}$
2: Unique rates for placental mammals and placental stem branch	$\omega_{\text{placental stem}}\neq\omega_{\text{outgroups}}\neq\omega_{\text{placental clade}}=\omega_{\text{primate clade}}$
3: Unique rate for placental mammals	$\omega_{\text{placental stem}}=\omega_{\text{outgroups}}\neq\omega_{\text{placental clade}}=\omega_{\text{primate clade}}$
4: Unique rate for placental stem branch	$\omega_{\text{placental stem}}\neq\omega_{\text{outgroups}}=\omega_{\text{placental clade}}=\omega_{\text{primate clade}}$
5: Unique rate for primates	$\omega_{\text{placental stem}}=\omega_{\text{outgroups}}=\omega_{\text{placental clade}}\neq\omega_{\text{primate clade}}$
6: Unique rates for placental mammals and primates	$\omega_{\text{placental stem}}=\omega_{\text{outgroups}}\neq\omega_{\text{placental clade}}\neq\omega_{\text{primate clade}}$

Hypotheses for placental cluster galectins	Parameters
Null: Uniform rate for all branches	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}=\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-14}}=\omega_{\text{Galectin-16}}$
2: Unique rate for *LGALS13*	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}=\omega_{\text{Galectin-14}}=\omega_{\text{Galectin-16}}\neq\omega_{\text{Galectin-13}}$
3: Unique rate for *LGALS14*	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}=\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-16}}\neq\omega_{\text{Galectin-14}}$
4: Unique rate for *LGALS16*	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}=\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-14}}\neq\omega_{\text{Galectin-16}}$
5: Unique rates for *LGALS13*, *LGALS14*, and *LGALS16*	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}\neq\omega_{\text{Galectin-13}}\neq\omega_{\text{Galectin-14}}\neq\omega_{\text{Galectin-16}}$
6: Unique rate for primate stem branch	$\omega_{\text{primate stem}}\neq\omega_{\text{outgroups}}=\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-14}}=\omega_{\text{Galectin-16}}$
7: Unique rates for primate stem branch and primate clade	$\omega_{\text{primate stem}}\neq\omega_{\text{outgroups}}\neq\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-14}}=\omega_{\text{Galectin-16}}$
8: Unique rate for primate clade	$\omega_{\text{primate stem}}=\omega_{\text{outgroups}}\neq\omega_{\text{Galectin-13}}=\omega_{\text{Galectin-14}}=\omega_{\text{Galectin-16}}$

was tested for all other galectins (supplementary fig. 8, Supplementary Material online).

Evolutionary Rate Analysis

To estimate galectin nucleotide substitution rates across different timeframes in the evolution of placental mammals, we employed the codeML program from the phylogenetic analysis package, PAML, version 4.9c (Yang 2007). Codon frequency, transition/transversion rate ratio, and the nonsynonymous/synonymous rate ratio ω were estimated by codeML in all tests. ω is also annotated as d_N/d_S and was used to estimate the strength of selection. Classic branch tests were used to evaluate different evolutionary scenarios in which ω varied between the placental stem branch, the clade of placental mammals, the clade of primates, and nonplacental outgroup species (supplementary fig. 9A, Supplementary Material online). For tests involving the placental cluster galectins, branch tests were also used to evaluate scenarios in which ω varied between the primate stem branch, the nonprimate outgroup clade, and the clades containing *LGALS13*, *LGALS14*, and *LGALS16* (supplementary fig. 9B, Supplementary Material online). Following Than et al. (2009), we assigned the placental stem branch the same ω as the outgroup species when not assigning it a unique ω. Similarly, we assigned the primate stem branch the same ω as the outgroup genes when not assigning it a unique ω. Six hypotheses were tested for each of the ancient galectin genes, and eight hypotheses were tested for the placental cluster galectin genes (table 1). We used likelihood ratio tests to evaluate significance of each of the different evolutionary scenarios.

Genotype Data for Population Genetics Analyses

To examine human galectin genes for signatures of recent positive selection and population differentiation, we retrieved human gene variant data from the 1000 Genomes Project Phase 3 release (1000 Genomes Project Consortium 2015). Identifiers and genomic coordinates were downloaded for all seven human galectin genes from Ensembl (human genome assembly GRCh37.p13) via the python module PyEnsembl (https://www.hammerlab.org/2015/02/04/exploring-the-genome-with-ensembl-and-python/; last accessed September 4, 2019). These coordinates were then used to extract corresponding genotype data for 1,668 individuals from the 1000 Genomes Project. These individuals comprised three different continental populations: African (661 individuals, including admixed ASW and ACB populations), European (503 individuals), and East Asian (504 individuals). We analyzed these three populations specifically because demographic parameters have been previously estimated for each of them (Gravel et al. 2011), thereby enabling the simulations of neutral evolution discussed in subsequent sections. We retrieved these gene variant data using a combination of the data retrieval tool, Tabix, version 0.2.6 (Li 2011), and a custom python script. We removed indels from these genotype data and excluded them from all subsequent analyses.

Detection of Recent Positive Selection

We calculated three different metrics (Tajima's D, F_{ST}, and *H12*) to evaluate different signatures of positive selection during recent human evolution. All three metrics were calculated for the genic region—defined by the genomic coordinates retrieved by PyEnsembl—of all seven galectin genes. Using the R package *PopGenome*, version 2.1.6 (Pfeifer et al. 2014), we calculated Tajima's D to evaluate signatures left

from selective events such as hard selective sweeps (Tajima 1989; Sabeti et al. 2006). We then calculated weighted Weir and Cockerham's fixation index (F_{ST}) to detect signatures of population differentiation (Weir and Cockerham 1984). Using the program *VCFtools* (Danecek et al. 2011), we calculated F_{ST} both gene-wise (i.e., for entire genes) and SNP-wise (i.e., for individual SNPs). One global comparison (i.e., considering all three populations at once) and three pairwise population comparisons (i.e., Africa vs. East Asia, Africa vs. Europe, and East Asia vs. Europe) were used to calculate four different F_{ST} values for every galectin gene and SNP.

Metrics like Tajima's D are based on allele frequency spectra and detect signals left by a specific mode of positive selection known as a hard selective sweep, in which a single beneficial haplotype rises to near or full fixation in a population (Barreiro and Quintana-Murci 2010). Hard selective sweeps reduce nucleotide diversity in the genomic region neighboring the selected variant, as linked variants sweep through a population in a phenomenon called genetic hitch-hiking (Barreiro and Quintana-Murci 2010). Alternatively, positive selection may simultaneously act on *multiple* beneficial haplotypes through a soft selective sweep, which can bring several haplotypes to high frequency in a population (Pennings and Hermisson 2006a,b; Messer and Petrov 2013; Garud et al. 2015). Soft selective sweeps do not reduce nucleotide diversity and are better detected by metrics like the *H12* statistic, an index that considers the frequency of the two most common haplotypes in a population (Garud et al. 2015). Following Moon et al. (2018), we used a modified version of the original python script written by Garud et al. (2015) to calculate a gene-wise *H12* for all seven analyzed galectin genes.

Construction of Null Distributions

Demographic events and stochastic processes, such as genetic drift, can shape patterns of genetic variation and generate signatures similar to those of selection (Vitti et al. 2013), influencing the performance of each of the three metrics. To determine whether galectin gene variation can be accounted for by positive selection—as opposed to neutral evolution—we simulated null distributions for each metric based on a neutral model of human evolution. We compared observed galectin values against these null distributions to compute *P* values. Galectins with a *P* value <0.05 (i.e., greater than the 5th percentile of simulated values) for a given statistic were considered significant and likely to have experienced positive selection.

To construct null distributions, we first simulated neutrally evolved gene sequences with *SLiM*, version 2.4.1 (Haller and Messer 2017). Following Moon et al. (2018), we accounted for past demographic events by employing *SLiM*'s implementation of human demographic parameters previously calculated by Gravel et al. (2011) for Africans, East Asians, and

Europeans. This simulation model first designates that the ancestral African population expanded from a size of 7,310–14,474 approximately 148,000 years ago. It further designates that the out-of-Africa migration occurred ~51,000 years ago, which was followed by the Eurasian split ~23,000 years ago. A population bottleneck followed the Eurasian split and reduced the European population to 1,032 individuals, and European and East Asian populations experienced exponential growth in population size over the last 23,000 years. The exponential coefficient was set at 0.0038 for Europeans and at 0.0048 for East Asians. Finally, the model assumes fixed rates for both recombination (1.0×10^{-8} recombination events per base pair) and mutation (2.4×10^{-8} mutations per base pair). More detailed demographic parameters regarding migration rates were also set in accordance with Moon et al. (2018).

We simulated the genotypes of 661, 503, and 504 individuals to reflect the number of individuals analyzed from the African, European, and East Asian populations, respectively. For each of the seven galectin genes, we simulated 1,000 sequences of the same length and then used these length-matched simulated sequences to calculate Tajima's D, Weir and Cockerham's F_{ST}, and *H12*. These metrics were calculated for simulated sequences exactly as described for galectin genes.

To compare population differentiation of galectin genes and SNPs to genome-wide signatures of population differentiation, we constructed empirical distributions of F_{ST} for all protein-coding genes in the genome. The variant data for these regions were obtained exactly as described for galectin genes, and genic F_{ST} for all four population comparisons was also calculated as described for galectin genes. Galectin genic F_{ST} was evaluated against these distributions to obtain empirical *P* values (supplementary table 2, Supplementary Material online).

Analysis of Galectin SNPs

Examination of gene-wise patterns of selection may miss positively selected SNPs if they reside in otherwise neutrally evolving regions; however, site frequency spectrum-based metrics like Tajima's D and haplotype-based metrics like *H12* were not designed for analysis of individual SNPs. By contrast, F_{ST} compares variation in SNP allele frequencies between populations (Barreiro and Quintana-Murci 2010) and can thus be applied to individual SNPs. To identify individual SNPs that are highly differentiated between populations, we compared the F_{ST} values of all galectin SNPs to the distributions of F_{ST} values of simulated SNPs using a *P* value of 0.05 as the significance threshold (e.g., a galectin SNP F_{ST} value calculated by comparing East Asians and Africans would be evaluated against an F_{ST} distribution for SNPs from neutral sequences simulated for East Asia and Africa). To focus on SNPs for which evidence of important functional effects is more likely, we filtered the

SNPs with significant F_{ST} values by retaining only those previously cited in the literature. We created this list by searching the Ensembl BioMart database (assembly GRCh38.p12) to identify which SNPs were previously cited. We further filtered this list by only retaining SNPs with a *putative* association to disease (i.e., prior studies explicitly claimed an association or functional link to a disease or phenotype). To evaluate how the F_{ST} values of these final SNPs compare with genome-wide population differentiation, we assessed empirical *P* values of the SNPs by generating F_{ST} values for all SNPs located in protein-coding regions genome-wide (table 3). For preliminary effect prediction, we obtained gene consequence information from Ensembl, and we submitted missense variants from the final list to the PolyPhen-2 webserver (Adzhubei et al. 2010).

Estimation of Allele Ages and Evolutionary History

To estimate the age of each allele for SNPs with significant or near significant F_{ST} values that have previously been associated with human complex genetic disease, we generated minimum allele ages using parsimony ancestral reconstruction across the mammalian phylogenetic tree. We obtained sequence data for 30 mammals for each SNP from the Multiz Alignment of 30 mammals track in the UCSC genome browser using the UCSC Table Browser (Kent et al. 2002; Karolchik et al. 2004). We included additional allelic information from six great ape species using data from the Great Ape Project (Prado-Martinez et al. 2013). We further obtained ancient human sequencing information for six additional genomes from the Max Planck Institute for Evolutionary Anthropology (Meyer et al. 2012). A phylogenetic tree for all species was obtained from the TimeTree database (Kumar et al. 2017). We conducted parsimony ancestral state reconstruction for each allele using the phangorn package in R (Schliep 2011). The minimum age of each allele was assigned as the age of the first node ancestral to humans inferred not to contain the specified allele.

It is possible that recurrent mutation could be the source of apparently old polymorphisms. We assessed an additional set of metrics to evaluate the likelihood of this scenario and to further explore the evolutionary history of these final SNPs (supplementary table 4, Supplementary Material online). To assess conservation of the investigated SNPs, we obtained PhyloP (Pollard et al. 2010) and phastCons (Siepel and Haussler 2005) values from the UCSC Genome Browser release GRCh37/hg19 (Kent et al. 2002; Hubisz et al. 2011). To assess linkage disequilibrium surrounding the ancestral and derived alleles for the variants of interest, we calculated the Integrated Haplotype Score (*iHS*) using data from the 1000 Genomes Project Phase 3 release, version 5a. The *iHS* values were computed from VCF files using the R package rehh, version 2.0.2 (Gautier et al. 2017). Finally, to evaluate the

possibility of balancing selection occurring on these variants, we obtained the precomputed β statistic for each site (Siewert and Voight 2017).

Structural Analysis with RosettaCM

To predict the structural effects of missense variants, we used a comparative modeling approach with the RosettaCM application, which requires experimentally determined structures that can serve as templates. Comparative modeling uses the templates, which have a high degree of sequence similarity to the sequence of interest, to approximate the overall fold of the protein. Rounds of energetic minimization allow for minor conformational changes to accommodate the sequence differences and produce more accurate models (Bender et al. 2016). Currently available structures of most galectin proteins are incomplete, thereby limiting which variants we were able to successfully model. For example, Galectin-3's N-terminal domain—containing both rs4644 and rs4652—lacks a published structure because it contains Pro-Gly repeats that make it highly flexible and difficult to crystallize (Ahmad et al. 2004). Of the four galectin candidates with variants passing previously discussed filters, Galectin-8 was found to be the most suitable protein for predicting variant effects. We thus employed RosettaCM from Rosetta version 3.8 to model the structural effects of a Galectin-8 variant associated with rheumatoid arthritis, along with three other tightly linked variants (Song et al. 2013).

We retrieved the PDB file for the template structure 4FQZ from the Protein Data Bank. 4FQZ was published in a study that structurally characterized Galectin-8 in complex with different common oligosaccharide ligands (Yoshida et al. 2012). 4FQZ represents a protease-resistant mutant form of Galectin-8 containing both the C-terminal and N-terminal CRDs. In this mutant form, the Galectin-8's linker peptide was replaced by two amino acids (His-Met) to enable in vivo and in vitro experiments that are otherwise infeasible due to the linker's susceptibility to proteolysis (Kim et al. 2013). Despite these mutations, this linker-shortened form serves as an excellent template because it represents a structure of Galectin-8 containing both CRDs. In vivo studies show that this mutated protein retains full activity and binds several of Galectin-8's common binding partners (Yoshida et al. 2012; Kim et al. 2013), further supporting the use of 4FQZ as an appropriate template for comparative modeling.

We isolated Chain A from 4FQZ and retrieved its FASTA sequence from the PDB. The amino acid sequence of 4FQZ exactly matches the Galectin-8 reference sequence at all four variant positions, and thereby serves as the protein's reference sequence. Two target sequences were aligned to this reference sequence: a sequence excluding the linker amino acids with four residue changes (F19>Y, R36>C, M56>V, and R158>S) to reflect the major Galectin-8 haplotype, and a sequence changing only F19>Y to reflect the hybrid

haplotype. The latter change reflects the substitution caused by rs1126407, the variant usually studied in isolation. This experimental design allowed us to directly compare changes in the protein's structure between the reference, hybrid, and major haplotypes.

We used Rosetta's partial thread application to thread all target sequences over 4FQZ, providing an initial approximation of the protein structures for each target-template alignment by assigning coordinates from the template PDBs to the aligned residues in the target sequences. The threaded PDBs were then used as single inputs in separate iterations of the rest of the comparative modeling procedure. We used the RosettaCM application to model any missing residue densities, generating 10,000 all-atom models for each threaded PDB. All generated models were further minimized and refined with Rosetta's dualspace relax protocol, bringing them closer to energetic minima. Parameters for both the initial comparative modeling and relaxation procedures are described in supplementary file 1, Supplementary Material online. We initially repeated the relaxation protocol on chain A from 4FQZ directly in order to compare it with the new models, but the resulting structure diverged dramatically from the best predicted model (root-mean-square deviation ~1.8). To resolve this issue, we subjected the reference sequence to the full procedure described earlier, generating and relaxing 10,000 hybridized structures based on the reference sequence.

After sorting models by overall energy (in Rosetta Energy Units), we selected the top 5,000 from each condition and clustered them with the clustering algorithm, Calibur (Li and Ng 2010). The cutoff was adjusted until the largest cluster contained 20–25% of the input models. We then selected the central model listed for the largest cluster as the best comparative model. These final models were compared with each other to evaluate structural differences between the reference haplotype, the major haplotype, and the hybrid haplotype. We used the PyMOL Molecular Graphics System, Version 2.1 Schrödinger, LLC., for all visual analyses, alignments, and root-mean-square deviation (RMSD) calculations.

Ensemble Analysis of Comparative Models

To account for structural variation among the top-scoring models of each condition and reflect the conformational flexibility of Galectin-8, we performed a clustering analysis of the top 100 scoring models from the reference, major, and hybrid haplotypes, respectively. Using Calibur, we examined whether the top-scoring models of each haplotype grouped into different structure-based clusters. Model distances were represented by the RMSD in Å of Galectin-8's N-terminal CRD spanning residues R45 to E89, the region containing eight residues known to directly interact with carbohydrate-bearing ligands (Yoshida et al. 2012). The cutoff was adjusted

until ~50% of all tested models were assigned to a single cluster. We then aligned the central models of the two largest clusters from residues R45 to E89.

Results

Ancient Galectin Genes Varied in Their Evolutionary Rates during Mammalian Evolution

To examine the evolutionary rates and strength of selection on ancient group galectins, we estimated the ratio of nonsynonymous to synonymous nucleotide substitutions ($\omega=d_N/d_S$) under several different evolutionary hypotheses (table 1). These hypotheses were designed to test whether galectins experienced accelerated evolutionary rates during the origin of placental mammals (e.g., Hypothesis 4), during the evolution of placental mammals generally (e.g., Hypothesis 3), and/or during the evolution of primates specifically (e.g., Hypotheses 5 and 6). Using the Likelihood Ratio Test, we evaluated the statistical significance of these hypotheses against each other and against a null hypothesis that assumed a uniform ω. Results for hypotheses that differed significantly from the null are summarized in table 2; all results are detailed in supplementary table 1, Supplementary Material online. Across all of our analyses, we found that galectins always exhibited ω values <1, which suggests that galectins have experienced (different degrees of) purifying selection throughout the evolution of placental mammals.

All alternative hypotheses differed significantly from the null hypothesis for *LGALS3* (table 2 and supplementary table 1, Supplementary Material online). However, the model assigning a unique ω to the placental stem branch, placental mammals (including primates) and outgroups did not differ significantly from the model in which the placental stem had the same ω as the outgroups (Likelihood Ratio Test or LRT; *P* value = 0.23), so we conclude that Galectin-3 did not experience a different evolutionary rate at the origin of placental mammals. The model assigning a different ω to outgroups, placental mammals (excluding primates), and primates was significantly better than the one assigning different ω to outgroups and placental mammals (including primates) (LRT; *P* value = 0.01), so we conclude that Galectin-3 experienced different evolutionary rates in placental mammals generally and primates specifically. Based on these tests, we conclude that Galectin-3 experienced a 5-fold increase in ω in placental mammals (other than primates) relative to outgroups ($\omega_{placental\ mammals}=0.35$; $\omega_{outgroups}=0.07$), and experienced a further increase in ω in primates ($\omega_{primates}=0.57$) relative to other placental mammals (fig. 2B).

Similar to *LGALS3*, all alternative models for *LGALS9* differed significantly from the null hypothesis. The model assigning a unique rate to the placental stem branch, placental mammals (including primates), and outgroups was significantly better than other models assigning any of these

Table 2

Statistically Significant Differences in the Rate of Evolution of Placental Galectin Genes

Gene	Hypothesis[a]	Likelihood (−lnL)	P value	ω Values			
				Outgroups	Placental Clade	Placental Stem	Primates
LGALS1	Null	5,353.76	N/A	0.15	0.15	0.15	0.15
	2	5,344.95	1.5E-04	0.20	0.16	0.01	0.16
	4	5,345.53	5.0E-05	0.17	0.17	0.01	0.17
LGALS3	Null	7,886.70	N/A	0.25	0.25	0.25	0.25
	2	7,834.11	1.44E-23	0.07	0.39	0.21	0.39
	3	7,834.84	2.33E-24	0.08	0.39	0.08	0.39
	4	7,884.15	2.39E-02	0.26	0.26	0.08	0.26
	5	7,873.09	1.82E-07	0.22	0.22	0.22	0.58
	6	7,831.69	1.29E-24	0.07	0.35	0.07	0.57
LGALS8	Null	8,161.84	N/A	0.19	0.19	0.19	0.19
	5	8,153.47	4.29E-05	0.17	0.17	0.17	0.36
	6	8,153.36	2.1E-04	0.18	0.17	0.18	0.36
LGALS9	Null	14,016.54	N/A	0.34	0.34	0.34	0.34
	2	14,007.70	1.4E-04	0.28	0.36	0.11	0.36
	3	14,010.44	4.74E-04	0.24	0.36	0.24	0.36
	4	14,009.79	2.36E-04	0.35	0.35	0.11	0.35
	5	14,014.56	0.0461	0.32	0.32	0.32	0.40
	6	14,009.85	0.0012	0.24	0.35	0.24	0.40

[a]Numbers correspond to the hypotheses described in table 1.
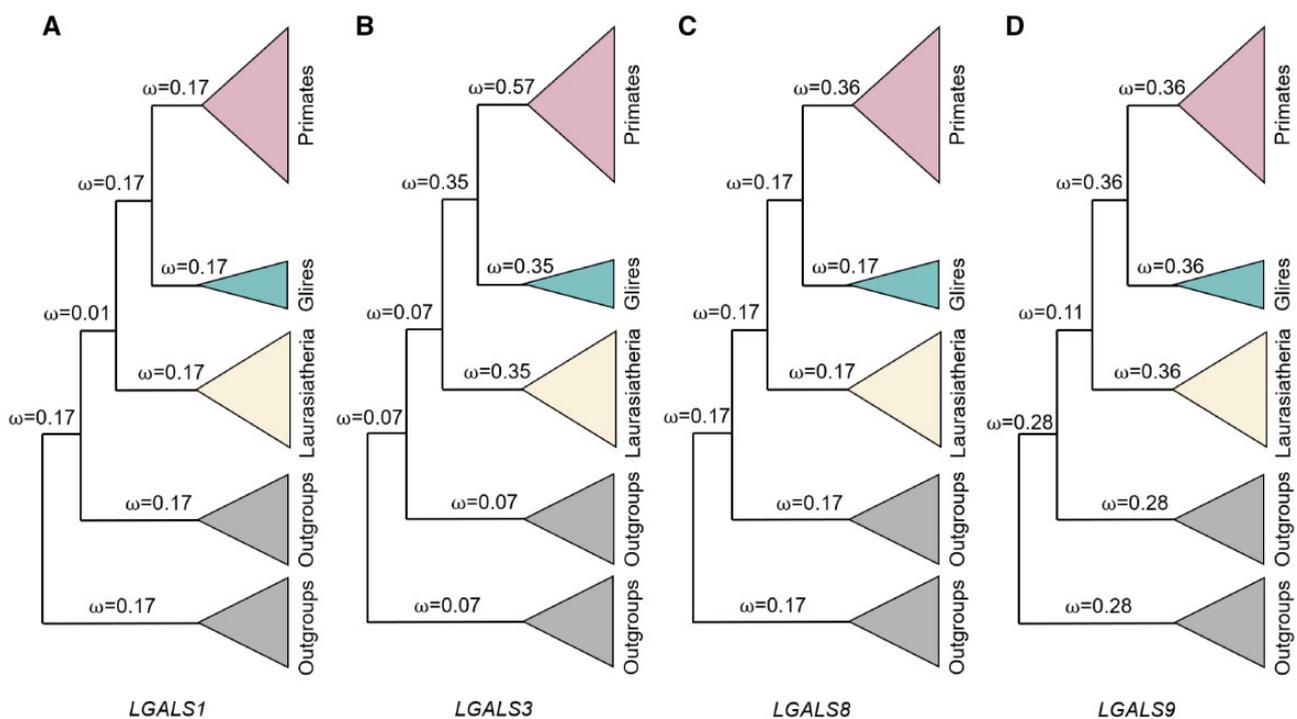
Fig. 2.—Changes in the evolutionary rate (measured by the ω ratio) in ancient placentally expressed galectin genes. The name of each ancient galectin gene is provided below each topology. The placental stem branch is located at the node representing the split between Laurasiatheria and other placental mammals (Glires and primates). All analyses were performed using codeML (Yang 2007). The complete set of hypotheses tested for each of the ancient galectin genes are provided in table 1 and their results in table 2. The full gene trees are provided in supplementary figure 8, Supplementary Material online. (A) *LGALS1* corresponds to hypothesis 4. (B) *LGALS3* corresponds to hypothesis 6. (C) *LGALS8* corresponds to hypothesis 5. (D) *LGALS9* corresponds to hypothesis 2.

variables the same rate, so we conclude that *LGALS9* experienced changes in evolutionary rate at the origin of placental mammals and during the evolution of placental mammals thereafter. Specifically, *LGALS9* experienced an episodic reduction from $\omega=0.28$ to $\omega=0.11$ on the placental stem branch, and then a long-term increase to $\omega=0.36$ during the evolution of placental mammals (fig. 2D).

In contrast to *LGALS3* and *LGALS9*, *LGALS8* and *LGALS1* only had two alternative hypotheses that were significantly better than the null hypothesis (table 2 and supplementary table 1, Supplementary Material online). For *LGALS8*, we found that the model assigning different ω ratios to both placental mammals (excluding primates) and primates was not significantly better than the model assigning a single ω ratio across primates (LRT; *P* value = 0.64). Thus, we conclude that *LGALS8* experienced accelerated evolution in primates (from $\omega=0.17$ to $\omega=0.36$; fig. 2C).

For *LGALS1*, we found that the hypothesis assigning different ω ratios to the placental stem, outgroups, and placental mammals (including primates) did not differ significantly from the hypothesis assigning a different ω ratio to only the placental stem (LRT; *P* value = 0.28). Therefore, we conclude that the ω ratio of *LGALS1* was episodically reduced at the origin of placental mammals (from $\omega=0.17$ to $\omega=0.01$; fig. 2A).

Placental Cluster Galectins Experienced Similar Evolutionary Rates

To examine the evolutionary rates and strength of selection in placental cluster galectins, we compared the null hypothesis of a uniform ω ratio to alternative hypotheses that posited 1) a different ω for Galectin-13 alone, 2) a different ω for Galectin-14 alone, 3) a different ω for Galectin-16 alone, and 4) a different ω ratio for each of the three genes. None of the alternative hypotheses differed significantly from the null (supplementary table 1, Supplementary Material online), suggesting that all three placental cluster galectins evolved at a similar rate of $\omega=0.60$.

Retrieval of placental cluster galectin genes from Ensembl also identified closely related genes named *LGALS13* in pig, *LGALS16* in goat, and *LGALS15* in goat and sheep, as well as additional genes lacking formal names in cow, goat, and opossum (fig. 3). Our phylogenetic analysis suggests that all these nonprimate genes are placed outside the clade formed by placental cluster galectins present in primates (fig. 3A and supplementary fig. 1, Supplementary Material online). Therefore, none of the nonprimate genes named *LGALS13* and *LGAS16* are orthologs of the primate *LGALS13* and *LGAS16* genes, respectively. Interestingly, the three goat galectin genes also form a cluster in ~200 kb region on chromosome 18 and are flanked by genes *EID2* and *DYRK1B*, the same genes used to demarcate and define the cluster of placental galectins found in primates (Than et al. 2009) (fig. 3B).

The same is true for the cow galectins, which also form a cluster and are flanked by the same two genes. Therefore, it appears that the placental cluster of galectin genes in primates originated independently of the cluster of galectin genes observed in cow and goat, as has been observed for other clusters of tandemly duplicated genes (Ortiz and Rokas 2017). We hypothesize that an ancestral galectin gene was likely present at this locus in both lineages and underwent independent duplications to form nonorthologous gene clusters. We tested whether placental cluster galectins specific to primates experienced a different evolutionary rate than these closely related nonprimate sequences. We further tested whether the ancestral gene of the placental cluster galectins experienced a short-term change in ω on the primate stem branch. These tests also failed to reject the null hypothesis (supplementary table 1, Supplementary Material online), suggesting that the primate-specific galectins evolved at a rate similar to their nonprimate relatives.

LGALS3 and *LGALS13* Experienced Recent Positive Selection in Humans

To examine more recent signatures of positive selection that occurred specifically within the human lineage, we calculated three different metrics (Tajima's *D*, F_{ST}, and *H12*) for galectin genes. Tajima's *D* is optimal for detecting hard selective sweeps. Hard sweeps take place when a beneficial variant spreads to very high frequency or fixation alongside the neutral variants linked to it, resulting in the presence of one dominant extended haplotype in the population (Tajima 1989; Sabeti et al. 2006). The fixation index F_{ST} is optimal for detecting population differentiation, which can result from local adaptation (Weir and Cockerham 1984). Finally, the *H12* index is optimal for detecting soft selective sweeps, which take place when multiple beneficial variants of the same locus spread into the population, resulting in two or more dominant extended haplotypes increasing in frequency in the population (Pennings and Hermisson 2006a,b; Messer and Petrov 2013; Garud et al. 2015).

To assess statistical significance, we simulated sequences of similar length to galectins based on a neutral evolution model that considered human population demography (see Materials and Methods). We specifically simulated the genotypes of 661, 503, and 504 individuals to reflect the number of individuals analyzed from the African, European, and East Asian populations, respectively, in the 1000 Genomes Project. Analysis of the simulated sequences enabled us to generate null distributions of values for each metric across each simulated galectin gene (supplementary fig. 2, Supplementary Material online). For F_{ST}, we also generated null empirical distributions of population differentiation for all protein-coding genes across all individuals in the 1000 Genomes Project (supplementary fig. 3, Supplementary Material online). We then compared the galectin genes' actual values against
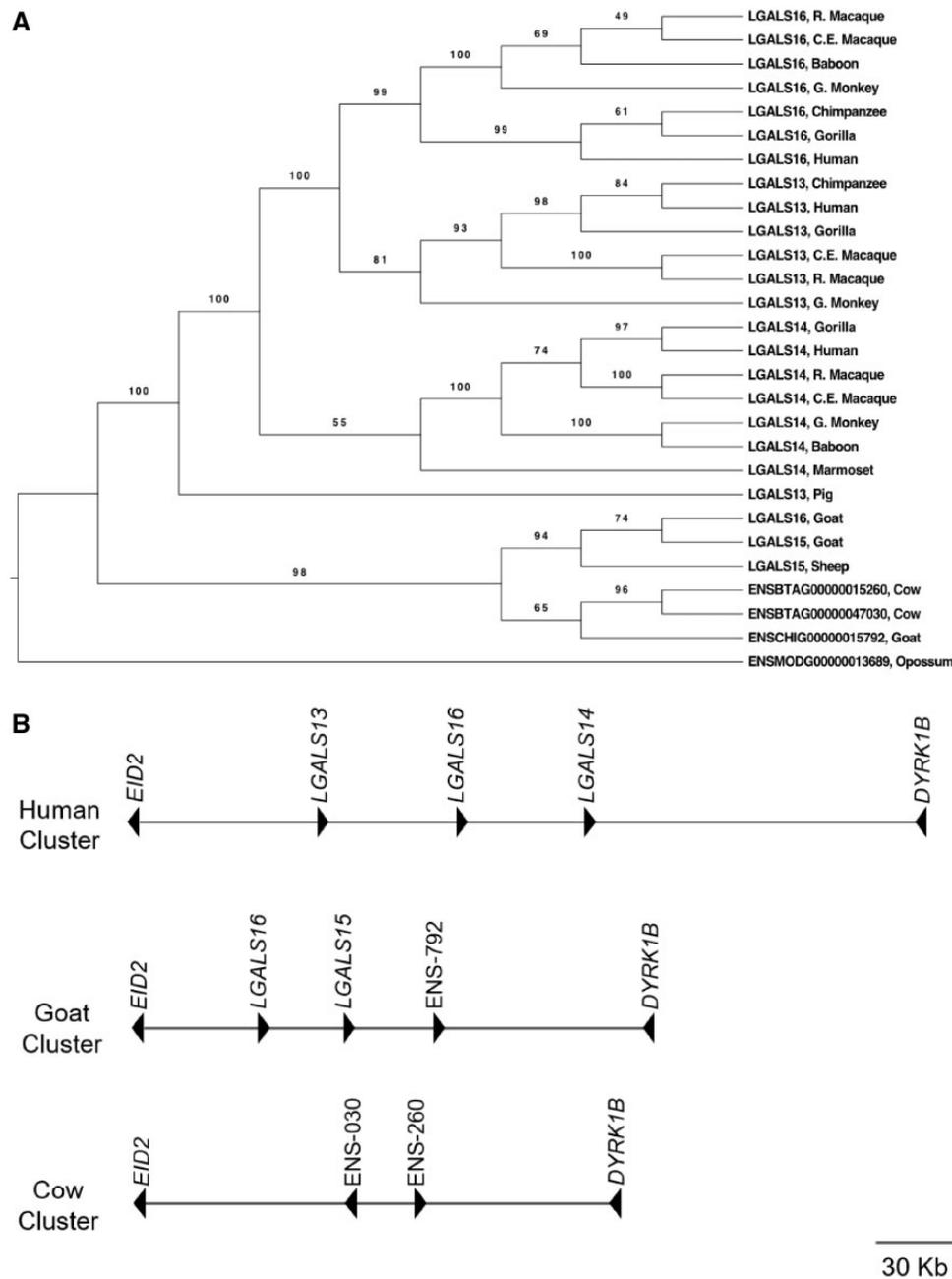
FIG. 3.—Evolution and genomic organization of placental cluster galectins and related genes. (A) Cladogram of the placental cluster galectins. Bootstrap support values are included as branch labels. Taxa are labeled as the gene names reported by Ensembl. All alternative models failed to reject the null hypothesis of an equal evolutionary rate ($\omega = 0.60$) for the placental cluster galectins. (B) Illustration of genomic organization of galectin gene clusters in human, cow, and goat. Remarkably, even though the cow and goat galectin gene clusters originated independently from the human galectin cluster, all three clusters are flanked by the same two genes: EID2 on the 5′ end and DYRK1B on the 3′ end. Cow and goat genes without official gene names are abbreviated with the last three numbers of their Ensembl accession numbers: ENSBTAG00000015260, ENSBTAG00000047030, and ENSCHIG00000015792. A phylogram depicting relative branch lengths is provided in supplementary figure 1, Supplementary Material online.

the null distributions to evaluate significance. We found that LGALS3 experienced positive selection in the form of a soft selective sweep ($H12 = 0.12$; P value < 0.01) (fig. 4A). We calculated a significantly negative Tajima's D value for LGALS13 (Tajima's $D = -2.32$; P value $= 0.01$) (fig. 4B). This is traditionally interpreted as a signature of a hard selective sweep, though we note that negative selection can also generate reduced nucleotide diversity consistent with this result. The other galectin genes, such as LGALS1 (fig. 4C and D), did not show signatures of positive selection for any of the three
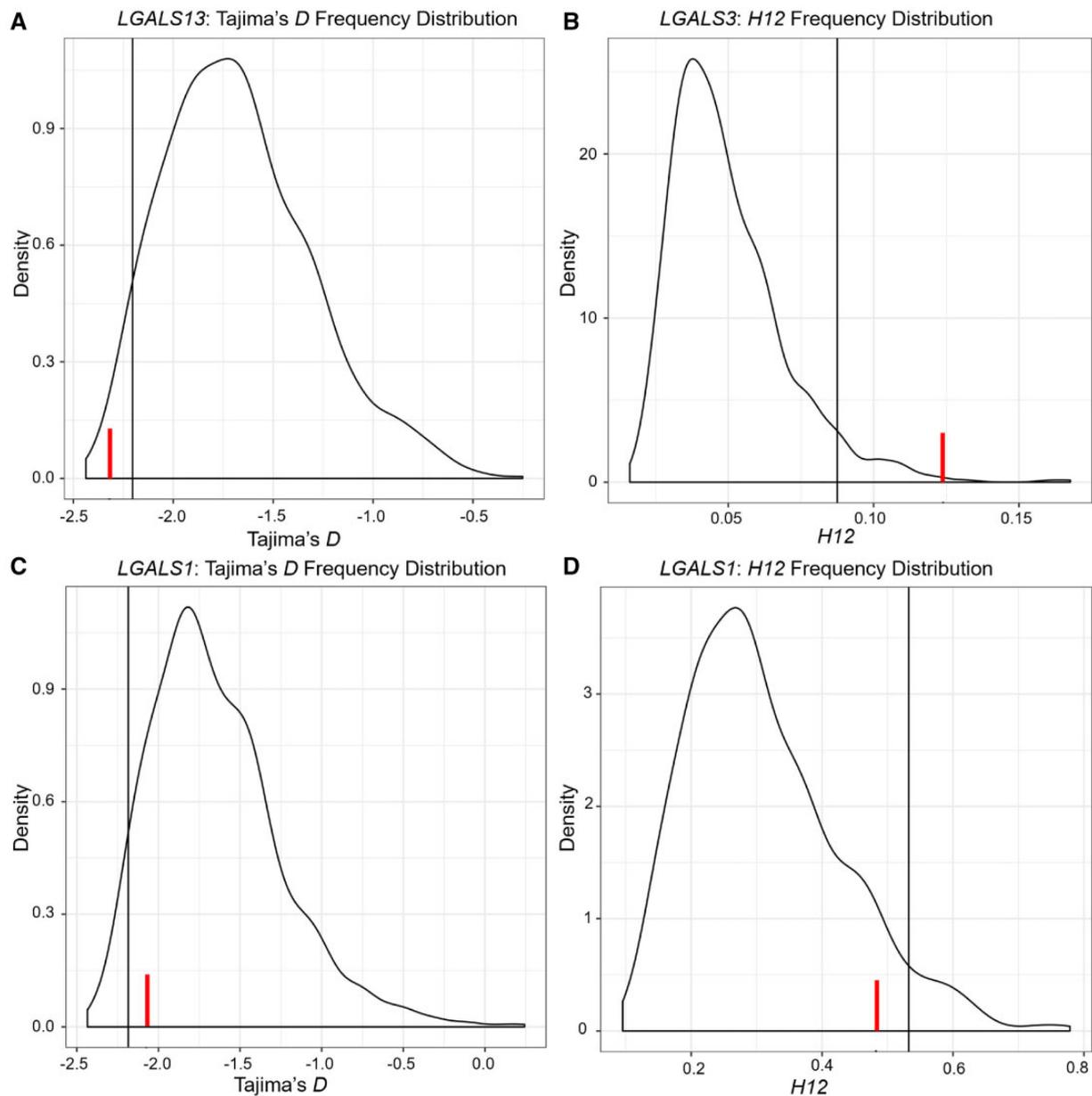
Fig. 4.—Representative distributions of Tajima's *D* and *H12*. Vertical lines indicate the significance cutoff corresponding to the 95th percentile of the null distribution. Red bars indicate where the actual values calculated for galectin genes fall along the distribution. Significantly negative values of Tajima's *D* are interpreted as evidence of positive selection via hard selective sweeps, whereas significantly positive values of *H12* are interpreted as evidence of positive selection via soft selective sweeps. (*A*) A significantly negative value of Tajima's *D* is found for *LGALS13*. (*B*) A significantly positive value of H12 is found for *LGALS3*. (*C*) and (*D*) *LGALS1* did not show significant results for these statistics.

metrics, suggesting that they evolved under purifying selection in the human lineage (supplementary table 2, Supplementary Material online).

Several Galectin SNPs Previously Linked to Disease Show Extreme Population Differentiation

Calculating F_{ST} for specific SNPs can identify signals of positive selection (Barreiro and Quintana-Murci 2010) that may be masked in gene-wide analyses due to the presence of other, neutrally evolving SNPs in a gene. To detect SNP-specific extremes in F_{ST}, we compared galectin SNP F_{ST} values to neutral SNP F_{ST} distributions based on the same population comparison. We found that 138 galectin SNPs (4.5% of those 3,084 tested after filtering), exceeded the 95th percentile of simulated F_{ST} for at least one population comparison (supplementary table 3, Supplementary Material online). Since the functional consequences of the majority of these SNPs are

unknown, we further filtered the SNPs to include only those seven that had a putative link to a disease or phenotype, and we estimated their minimum allele ages using parsimony ancestral reconstruction across a phylogenetic tree of 30 mammals (see Materials and Methods for detail).

For these final SNPs, we evaluated their F_{ST} values against empirical distributions of SNP-wise F_{ST} genome-wide, and we calculated an additional sequence of evolutionary statistics to assess their linkage disequilibrium (captured by the Integrated Haplotype Score, iHS), their degree of evolutionary conservation (captured by PhyloP scores), and the possibility of balancing selection (captured by β statistics) (see Materials and Methods for detail). The F_{ST} values, PolyPhen values, and allele ages of these SNPs are described in table 3; all other SNP-related data are presented in supplementary table 4, Supplementary Material online. We found seven galectin SNPs that exhibit significant population differentiation and that are linked to disease or another phenotype.

Of these, rs4652, a LGALS3 variant causing a proline to threonine substitution, showed the highest F_{ST} value: 0.52 between African and East Asian populations. It is associated with rheumatoid arthritis (Hu et al. 2011), risk for gastric carcinoma (Shi et al. 2011), differential LGALS3 gene expression, and differential Galectin-3 circulation levels in the blood (de Boer et al. 2012). We found that both rs4652 alleles are ancient: The C allele is found across eutherian mammals and is estimated to be ~105 Myr old, whereas the A allele is found across the great apes and is estimated to be ~16 Myr old (table 3 and supplementary fig. 4, Supplementary Material online). We calculated low β statistics for this SNP across all subcontinental populations (supplementary table 4, Supplementary Material online), suggesting that balancing selection has not been acting on this SNP. We also calculated a positive PhyloP score (1.10) for this variant, further suggesting that it occurs in a conserved element that is evolving slowly. Given that the great apes share this polymorphism and that it is embedded within a slowly evolving element, we conclude that the rs4652 locus has been polymorphic for millions of years. This provides interesting evolutionary context for the SNP's population disparity, in which the derived A allele exists at very low frequency (~5%) in African populations but has risen to >50% frequency in non-African populations. We calculated negative iHS scores for this variant in East Asian (−1.56) and European (−0.78) populations, which suggest that the derived allele is under positive selection.

LGALS8 Has Two Major Haplotypes Defined by Four Common Variants

Another interesting variant, rs1126407 (F19Y), causes a phenylalanine to tyrosine substitution in Galectin-8. This variant narrowly missed the threshold for significant population differentiation according to simulated null distributions (F_{ST}

Table 3

F_{ST} Values, Phenotype Links, and Allele Ages of Placental Galectin Variants with Significant Population Differentiation and Putative Links to Disease or Other Phenotypes

SNP	Gene	P Value	Empirical P Value	Functional Consequence	Highest Calculated F_{ST}	PolyPhen Score	Putative Disease/Phenotype Link	ID and Age of Oldest Allele[a]	ID and Age of Youngest Allele
rs4652	LGALS3	0.004	0.002	Missense (T > P)	0.52	0 (benign)	Several	C; 105	A; 15.8
rs10403702	LGALS16	0.03	0.017	Missense (L > P)	0.20	0	Acute insulin response to glucose	C; 43	T; human-specific allele
rs4830	LGALS14	0.02	0.009	Missense (C > S/R)	0.32	0.02	Epistatic effect in breast cancer	A; human-specific	T; human-specific
rs1009977	LGALS3	0.01	0.008	Intron variant	0.35	N/A	Cognitive function in old age	G; 29.4	T; 15.8
rs10498474	LGALS3	0.02	0.012	Intron variant	0.24	N/A	Body weight	A; 43	G; human-specific
rs2273865	LGALS8	0.048	0.029	Nonsense variant	0.14	N/A	Differential expression of Galectin-8 isoforms	T; 105	A; human-specific
rs4659682	LGALS8	0.01	0.008	Intron variant	0.34	N/A	ADHD-IV impulsivity	A; 105	G; human-specific
rs4644[b]	LGALS3	0.056	0.032	Missense (P > H)	0.11	1 (deleterious)	Several	C; 105	A; human-specific
rs1126407[b]	LGALS8	0.066	0.041	Missense (F > Y)	0.10	1	Rheumatoid arthritis	A; 105	T; 6.23

[a] All ages are given in Myr; human-specific alleles are identified as such.
[b] These SNPs are marginally nonsignificant according to simulated distributions but are linked to disease, so we opted to include them.
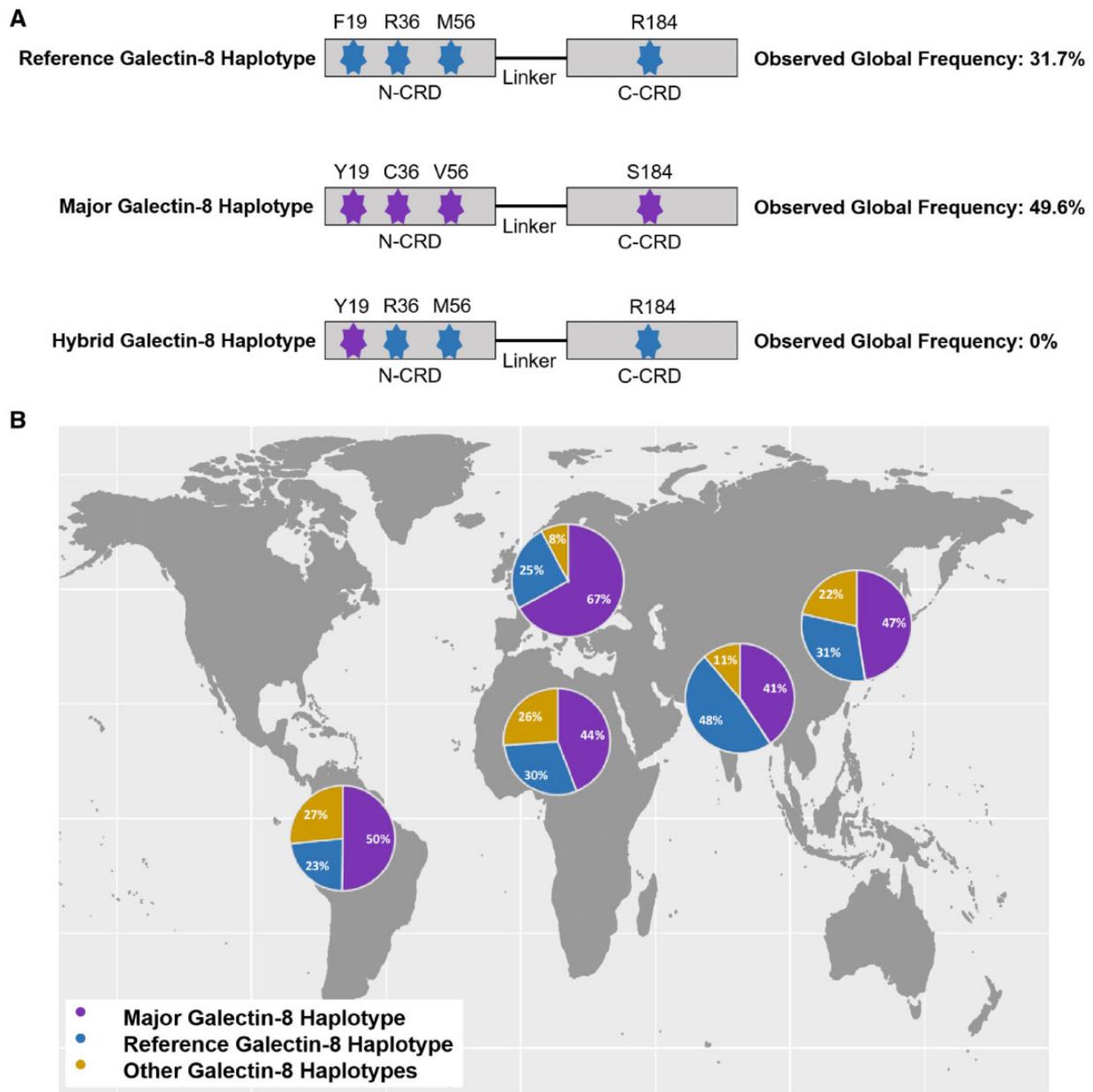
FIG. 5.—Protein haplotype frequencies of the canonical Galectin-8 isoform across and within five major human populations. (*A*) Global frequency of two common Galectin-8 protein haplotypes defined by four missense variants. The reference haplotype commonly used in Galectin-8 functional studies is not the most common protein haplotype. None of the 2,504 human genomes in the 1000 Genomes Project contains the hybrid haplotype. (*B*) Population frequencies of different Galectin-8 protein haplotypes mapped to five major continental populations defined in the 1000 Genomes Project: Africa, America (admixed populations), East Asia, Europe, and South Asia. Frequencies are based on data from the 1000 Genomes Project compiled in Ensembl for the canonical isoform. The Ensembl accession number of the corresponding *LGALS8* transcript is ENST00000341872.10.

value = 0.1; *P* value = 0.07), but passed the threshold according to empirical null distributions (empirical *P* value = 0.04) (table 3). The A allele, which encodes tyrosine, has existed for ~105 Myr, while the younger T allele, which encodes phenylalanine, has existed for ~6 Myr (table 3 and supplementary fig. 5, Supplementary Material online). Similar to rs4652, we found a positive PhyloP score (2.10) that suggests

this SNP occurs within a slowly evolving element. We also calculated remarkably high β statistics (all >54) for this SNP (supplementary table 4, Supplementary Material online), indicating that this polymorphism has been maintained over time due to balancing selection. The A allele is present at 68% frequency in European populations, while it remains <50% frequency in Africa (44%) and East Asia (48%). F19Y was
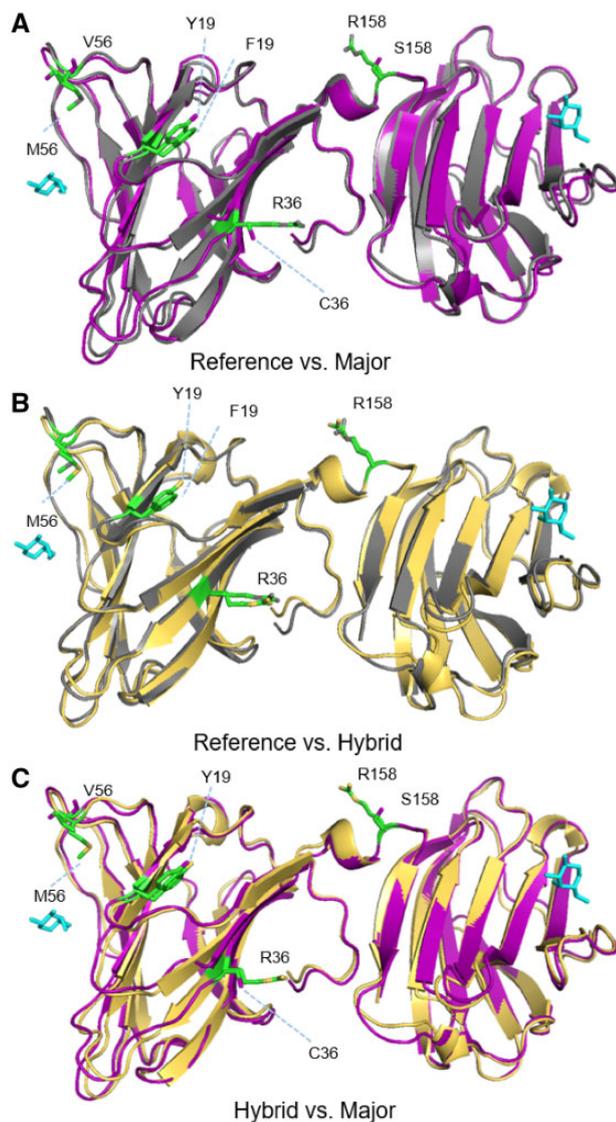
FIG. 6.—Protein structural differences between the central models of the major haplotype, the reference haplotype, and the hybrid haplotype. The protein structure for the reference haplotype is colored gray, the major haplotype purple, and the hybrid haplotype yellow, respectively. Variant residues are colored green. Cyan-colored structures illustrate the approximate position typically occupied by carbohydrate ligands that bind the carbohydrate recognition domain. RMSD values are given in units of angstroms. (A) Model based on the major haplotype superimposed with the model based on the reference haplotype (RMSD = 0.79). (B) Model based on the hybrid haplotype superimposed with the model based on the reference haplotype (RMSD = 0.70). (C) Model based on the hybrid haplotype superimposed with model based on the major haplotype (RMSD = 0.74).

associated with rheumatoid arthritis in a previous study (Pal et al. 2012). Two functional studies that followed the association study elucidated both structural and functional impacts of this variant on Galectin-8 carbohydrate-binding activity (Ruiz et al. 2014; Zhang et al. 2015). Importantly, both studies

focused on the F19Y variant alone by mutating only F19 to Y19 on a background of the human reference sequence; however, our analysis revealed another variant, R36C (rs1041935), that is both highly differentiated and in complete linkage disequilibrium ($D' = 1$; $r^2 = 1$) with F19Y. Further investigation of 1000 Genomes Project data revealed that Y19 frequently co-occurs with three other missense variants (C36, V56, and S184) in over 80% of Galectin-8 protein haplotypes. We also found a global frequency of 31.7% for the reference haplotype (F19, R36, M56, and R184) of the canonical isoform typically used in Galectin-8 functional studies; in contrast, the global frequency of the major haplotype (Y19, C36, V56, S184) is 49.6%. Notably, Y19 has never been observed in the reference background (i.e., with R36, M56, and R184) (fig. 5A), suggesting that the "hybrid" haplotype (Y19, R36, M56, and R184) examined by the studies of Ruiz et al. (2014) and Zhang et al. (2015) either does not exist in the human population or is present at very low frequency.

Galectin-8 Sequence Differences Influence Structure

To evaluate potential structural differences between the major, reference, and hybrid haplotypes, we used comparative modeling with RosettaCM to predict their structures (see Materials and Methods). We compared the models of the three haplotypes using global RMSD, a metric that quantifies the mean distance between atoms of superimposed protein structures. Our comparisons revealed structural differences between the reference and major haplotypes (RMSD = 0.79), between the reference and hybrid haplotypes (RMSD = 0.70), and between the hybrid and major haplotypes (RMSD = 0.74) (fig. 6A–C, respectively).

These results suggest that the major haplotype (Y19, C36, V56, S184) induces different structural changes than the hybrid haplotype (Y19, R36, M56, and R184). However, it is important to note that considerable structural variation exists in the top ten scoring models of all three sequences (supplementary fig. 6, Supplementary Material online). RosettaCM is designed to predict the most likely conformations for a protein of interest. However, it is possible that a single model cannot accurately capture the conformational flexibility of Galectin-8. To address this issue, we employed the Calibur clustering algorithm to examine whether the carbohydrate binding region of the top 100 scoring models of each haplotype (i.e., the reference haplotype, the major haplotype, and the hybrid haplotype) group into different structure-based clusters. This would indicate that, despite the variation within each set, the different haplotypes are still likely to adopt different structural conformations in the natural protein at the major carbohydrate binding site.

Focusing on residues spanning the binding region of the N-terminal CRD, models based on the hybrid haplotype tend to cluster with models based on the reference haplotype, which together comprise roughly equal amounts of 86% of the

models present in the largest cluster (supplementary fig. 7A, Supplementary Material online). In contrast, models based on the major haplotype tend to cluster on their own, comprising 70% of the models in the second largest cluster. This reinforces the hypothesis that the major haplotype has different structural impacts than the hybrid haplotype, and it also suggests that these differences are pronounced at the N-terminal CRD (supplementary fig. 7B, Supplementary Material online), one of two major functional sites for Galectin-8's carbohydrate-binding activities.

Discussion

We found that seven placentally expressed galectin genes involved in immunity evolved at different rates during the evolution of placental mammals. *LGALS3*, *LGALS8*, and *LGALS9* experienced periods of accelerated evolution that reflect changes in functional constraints possibly driven by positive selection. We found further evidence for selection in the human lineage, including gene-wide signatures of positive selection acting on *LGALS3* and *LGALS13*, as well as SNP-wise signatures of population differentiation. By combining these signatures with findings from previous studies, we identified seven galectin gene variants that experienced positive selection and that are linked to disease. Taken together, these results suggest that the evolutionary history of placentally expressed galectin genes has been dynamic, both across mammals as well as within humans.

Prior studies have hypothesized strong evolutionary constraints on the basis of galectins' fundamental and evolutionary ancient immune functions (Vasta 2012). This hypothesis is reinforced by our finding that galectins show ω values <1 across all branches and clades analyzed by our study. This suggests the presence of purifying selection acting to constrain galectins' evolutionary rates throughout placental mammalian evolution; however, our results also show that galectins evolved at different evolutionary rates and experienced both episodic and long-term rate changes, suggesting that functional constraint varied across genes, lineages, and time windows. In the case of *LGALS3*, *LGALS8*, and *LGALS9*, periods of accelerated evolution are suggestive of relaxed purifying selection. This apparent relaxation in selective pressure may have been driven by an increase in positive selection. This is especially interesting considering the multifunctional nature of mammalian galectins, which have fundamental immunoregulatory functions that likely predate the origin of placental mammals, but which also play important roles in placental development (Vasta 2012; Than et al. 2015). Periods of increased ω ratios may reflect evolutionary modifications that occurred as these ancient immunity genes adapted their immunity functions for new roles in placental development.

Phylogenetic evidence also suggests that *LGALS1* and *LGALS9* experienced increased purifying selection (decreased ω) at the origin of placental mammals. This result is in contrast to previous findings by Than et al. (2008), who found evidence for positive selection acting on *LGALS1* at the origin of placental mammals. This contradiction is explained by differences in methodology: Than et al. inferred site-specific positive selection through the use of branch-site models, whereas we inferred negative selection through the use of branch models (Yang 2007). Branch models can miss positive selection that acted on only a subset of codons; branch-site models would not, but have been shown to suffer from high false-positive rates (Pond et al. 2011). By utilizing only branch models, we avoided the risk of false-positives and detected negative selection acting on *LGALS1* and *LGALS9* in the genes at-large.

It is interesting to note the different patterns of selection acting on these ancient galectins in the context of their shared structural homology, overlapping roles in innate and adaptive immunity, and signature binding preference for β-galactoside-bearing carbohydrates. These shared attributes raise the hypothesis that these family members likely experienced similar patterns of selection, yet our results suggest otherwise. With this in mind, it is interesting to consider the differences that define unique family members within a shared structural, biochemical, and phenotypic framework. For example, Rabinovich and Toscano (2009) argue that while all galectins specifically bind N-acetyllactosamine (LacNAc) as a "minimal structure," there is variation in their binding preferences for complex carbohydrate structures that is influenced by numerous factors, such as the degree of carbohydrate branching and modifications. These binding preferences are likely responsible for differences in functional activity. Along these lines, it has also been noted that while galectins sometimes overlap in their highly diverse immunity functions, several family members also carry out unique functions (Rabinovich and Toscano 2009). For example, Galectin-3 is uniquely involved in alternative macrophage activation, while Galectin-8 is uniquely involved in an autophagic response to salmonella infection. Taking these points together, we hypothesize that the differences in selection patterns likely reflect the different evolutionary histories galectin family members underwent in the course of their functional divergence.

In contrast to the variable evolutionary rates found in ancient galectin genes, we found evidence for a constant evolutionary rate in placental cluster galectin genes (figs. 2 and 3). Previously established evolutionary and functional evidence may help explain this difference between the ancient and placental cluster galectins. For example, all three of the placental cluster galectins are solely expressed in the placenta (Than et al. 2009), and they evolved relatively recently by tandem duplication in the primate lineage and share specific carbohydrate binding preferences and apoptotic functions (Than et al. 2009). Thus, shared tissue environment and biological function could have produced a similar selection

pressure for all three placental galectins, leading to their similar rates of evolution observed in our analysis.

The finding that placental cluster galectins exhibit highly similar evolutionary rates contradicts previous findings from Than et al. (2009), who proposed highly variable evolutionary rates for these placental galectins. This inference was based on the free-ratio model that assigns a unique ω to every branch in the phylogeny. This test has received criticism because it is parameter-rich and prone to overfitting; its use is also discouraged in the PAML manual (Yang 2007). Than et al. also performed their analysis on a phylogeny containing several paralogs of *LGALS13*, *LGALS14*, and *LGALS16* that exist in expanded gene clusters in other primate species. By contrast, our analysis only focused on orthologs of the three genes of this cluster that are present in humans and solely expressed in the placenta. Thus, Than et al.'s inference of variable evolutionary rates applies to a larger set of genes than the one included in our analysis; however, inferences based on a free-ratio model are still parameter-rich and unlikely to produce accurate estimates of all ω values (Yang 2007), and they are partly contradicted by the failure of branch model tests to reject the null hypothesis in our analysis.

Our analysis widens the lens on the evolution of the placental cluster galectins. By including closely related nonprimate genes, we discovered that galectin gene clustering is not unique to primates: both goat and cow contain small clusters of galectin genes in a chromosomal region homologous to the region containing primate placental galectins. Interestingly, *LGALS*15, a galectin gene present in both sheep and goat, is solely expressed in the uterus and supports trophoblast attachment in these species (Lewis et al. 2007). Our results also show high similarity between duplicates in both primate and artiodactyl lineages, which could be explained by either gene conversion or independent duplication. It is challenging to determine which process is responsible for the similarity of the duplicates. We believe this represents a striking case of convergent evolution of a cluster of tandem duplicated genes (Ortiz and Rokas 2017) between artiodactyls and primates: both lineages evolved gene clusters of variable size comprised of galectins with unique expression in reproductive tissues and putative functions in pregnancy.

Human immune genes often show strong signatures of positive selection as our genomes evolved in response to selective pressures applied by pathogens (Barreiro and Quintana-Murci 2010). We found that *LGALS3* and *LGALS13* exhibit signatures of recent positive selection in the human lineage. *LGALS3* specifically experienced a soft selective sweep, in which multiple beneficial alleles rise to relatively high frequencies in a population (Pennings and Hermisson 2006a,b; Messer and Petrov 2013; Garud et al. 2015). *LGALS3* encodes a highly multifunctional protein with nearly ubiquitous tissue expression (Dumic et al. 2006), and it is highly likely that the selected variants may have impacted several Galectin-3-related processes at once;

consequently, identifying which function might have been the driving force behind selection is challenging. It is notable, though, that *LGALS3* also showed accelerated evolution in placental mammals and primates. Taken together, these results suggest that Galectin-3 has long been a substrate for evolutionary modifications, undergoing accelerated substitution rates in ancient mammals and significant changes in haplotype frequencies in modern humans.

We also discovered SNP-specific signatures of positive selection through analysis of the F_{ST} statistic, which can detect geographically restricted positive selection (Barreiro and Quintana-Murci 2010). Seven of these SNPs are putatively linked to diseases, such as breast cancer and rheumatoid arthritis, as well as other phenotypes, such as ADHD impulsivity (Kim et al. 2012). Of these, rs4652, a *LGALS3* variant, showed the strongest signature of population differentiation ($F_{ST}=0.52$ between African and East Asian populations). rs4652 has been implicated in diseases such as rheumatoid arthritis and gastric carcinoma (Hu et al. 2011; Shi et al. 2011). After finding that Galectin-3 deletion in mice conferred partial protection to malaria, Oakley et al. (2009) hypothesized that this SNP could influence malaria resistance by disrupting Galectin-3 multimerization, an idea strengthened by the fact that rs4652 disrupts a key repeat in the domain required for Galectin-3 multimerization. In addition to influencing Galectin-3's multimerization, rs4652 may also affect its gene expression and circulation levels. de Boer et al. (2012) reported that rs4652 is strongly associated with Galectin-3 circulating levels in the blood. Galectin-3 circulating levels are implicated in several diseases that span immunological diseases and cardiovascular disorders such as arteriosclerosis (de Boer et al. 2012; Madrigal-Matute et al. 2014). Interestingly, Banfer et al. (2018) recently elucidated a mechanism for Galectin-3's exosomal secretion that involves direct binding to Tsg101, a component of the endosomal sorting complex required for transport. Galectin-3 interacts with Tsg101 via a P(S/T)AP peptide motif that is disrupted by rs4652, and a coimmunoprecipitation assay confirmed that the variant altered Galectin-3's capacity to precipitate Tsg101 (Banfer et al. 2018). We propose that this impact may represent a functional link to the variant's association with Galectin-3 circulation levels in the blood.

Another variant, rs1126407 (F19Y), lies within *LGALS8* and passed the significance threshold for population differentiation based on empirical distributions (empirical P value $= 0.04$). This variant is associated with rheumatoid arthritis and has been the subject of several functional analyses (Ruiz et al. 2014; Zhang et al. 2015). These studies analyzed F19Y in the background of the reference human haplotype, which resulted in the exclusion of three other nonsynonymous variants that are linked with it in over 80% of Galectin-8 haplotypes in the human population. This exclusion means that the functional effects of the F19Y variant were examined in the context of a "hybrid" haplotype that

is not found in the 5,008 human haplotypes present in the 1000 Genomes Project. Our results predict that the major haplotype, containing all four variants, impacts the protein structure differently than the hybrid haplotype, and they predict that the major haplotype induces a greater impact on the key binding region of Galectin-8's N-terminal CRD. This could thus impact Galectin-8's binding affinity with a variety of ligands that bind its N-terminal CRD. These results illustrate the importance of considering human haplotype diversity when investigating natural protein function and disease-associated genetic variants. By referring only to the reference sequence, investigators may not capture the properties of far more common, nonreference sequences that may harbor several amino acid differences in the protein of interest.

Future studies of Galectin-8 sequence and structure must consider its haplotype diversity, in particular the major haplotype, which is more common globally than the reference sequence. Future studies of the structure of Galectin-8 in complex with the glycoprotein CD44 would also be valuable. Galectin-8 normally executes an anti-inflammatory function by inducing apoptosis of synovial fluid (SF) cells through binding with CD44 at the cell surface (Sebban et al. 2007). In rheumatoid arthritis, SF cells express a unique CD44 variant (CD44vRA) that has been shown to ligate and sequester Galectin-8 in patients with rheumatoid arthritis (Sebban et al. 2007), thereby reducing its anti-inflammatory effects. We hypothesize that major Galectin-8 haplotypes have differential affinity for binding CD44vRA, and that this explains the association of F19Y with rheumatoid arthritis.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Literature Cited

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248–249.

Ahmad N, et al. 2004. Galectin-3 precipitates as a pentamer with synthetic multivalent carbohydrates and forms heterogeneous cross-linked complexes. J Biol Chem. 279(12):10841–10847.

Balan V, et al. 2008. Racial disparity in breast cancer and functional germ line mutation in galectin-3 (rs4644): a pilot study. Cancer Res. 68(24):10045–10050.

Banfer S, et al. 2018. Molecular mechanism to recruit galectin-3 into multivesicular bodies for polarized exosomal secretion. Proc Natl Acad Sci U S A. 115(19):E4396–E4405.

Barreiro L, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet. 11(1):17–30.

Bender BJ, et al. 2016. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. Biochemistry. 55(34):4748–4763.

Cummings RD et al. 2017. Essentials of glycobiology. In: Varki A, Cummings RD, Esko JD, editors. Galectins. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

de Boer R, et al. 2012. A genome-wide association study of circulating galectin-3. PLoS One 7(10):e47385.

Dumic J, Dabelic S, Flögel M. 2006. Galectin-3: an open-ended story. Biochim Biophys Acta. 4:616–635.

Fumagalli M, et al. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 7(11):e1002355.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 11(2):e1005004.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 17(1):78–90.

Gravel S, et al. 2011. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108(29): 11983–11988.

Haller BC, and Messer PW. 2017. SLiM 2: Flexible, interactive forward genetic simulations. Mol. Biol. Evol. 34(1):230–240.

Hamblin M, Thompson E, Di Rienzo A. 2002. Complex signatures of natural selection at the duffy blood group locus. Am J Hum Genet. 2:369–383.

Houzelstein D, et al. 2004. Phylogenetic analysis of the vertebrate galectin family. Mol Biol Evol. 21:117–1187.

Hu C, Chang SK, Wu CS, Tsai WI, Hsu PN. 2011. Galectin-3 gene (*LGALS3*) +292C allele is a genetic predisposition factor for rheumatoid arthritis in Taiwan. Clin Rheumatol. 9:1227–1233.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinformatics 12:41–51.

Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32(Database issue):D493–D496.

Katoh S, Standley D. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12(6):996–1006.

Kim BW, Hong SB, Kim JH, Kwon DH, Song HK. 2013. Structural basis for recognition of autophagic receptor NDP52 by the sugar receptor Galectin-8. Nature Commun. 4:1613–1613.

Kim DS, et al. 2012. Results of genome-wide analyses on neurodevelopmental phenotypes at four-year follow-up following cardiac surgery in infancy. PLoS One 7(9):e45936.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812–1819.

Lewis SK, et al. 2007. Galectin 15 (*LGALS15*): a gene uniquely expressed in the uteri of sheep and goats that functions in trophoblast attachment. Biol Reprod. 77(6):1027–1036.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27(5):718–719.

Li S, Ng Y. 2010. Calibur: a tool for clustering large numbers of protein decoys. BMC Bioinformatics. 11:25.

Madrigal-Matute J, et al. 2014. Galectin-3, a biomarker linking oxidative stress and inflammation with the clinical outcomes of patients with atherothrombosis. J Am Heart Assoc. 3:e000785.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 28:10.

Meyer M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338(6104):222–226.

Moon JM, Aronoff DM, Capra JA, Abbot P, Rokas A. 2018. Examination of Signatures of Recent Positive Selection on Genes Involved in Human Sialic Acid Biology. G3 (Bethesda). 8(4):1315–1325.

Oakley M, et al. 2009. Pathogenic roles of CD14, galectin-3, and OX40 during experimental cerebral malaria in mice. PLoS One 4(8):e6793.

Ortiz JF, Rokas A. 2017. CTDGFinder: a novel homology-based algorithm for identifying closely spaced clusters of tandemly duplicated genes. Mol Biol Evol. 34(1):215–229.

Pal Z, et al. 2012. Non-synonymous single nucleotide polymorphisms in genes for immunoregulatory galectins: association of Galectin-8 (F19Y) occurrence with autoimmune diseases in a Caucasian population. Biochim Biophys Acta. 10:1512–1518.

Pennings PS, Hermisson J. 2006a. Soft sweeps II – molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 23(5):1076–1084.

Pennings PS, Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2(12):e186.

Perelman P, et al. 2011. A molecular phylogeny of living primates. PLoS Genet. 7(3):e1001342.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, and Lercher MJ. 2014 PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Mol. Biol. Evol. 31(7):1929–1936.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of non-neutral substitution rates on mammalian phylogenies. Genome Res. 20(1):110–121.

Pond S, et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 11:3033–3043.

Prado-Martinez J, et al. 2013. Great ape genome diversity. Nature 499(7459):471–475.

Rabinovich G, Toscano M. 2009. Turning 'sweet' on immunity: galectin–glycan interactions in immune tolerance and inflammation. Nat Rev Immunol. 9(5):338–352.

Raj T, et al. 2013. Common risk alleles for inflammatory diseases are targets of recent positive selection. Am J Hum Genet. 92(4):517–529.

Ramos PS, Shaftman SR, Ward RC, Langefeld CD. 2014. Genes associated with SLE are targets of recent positive selection. Autoimmune Dis. 2014:1.

Ramos PS, Shedlock AM, Langefeld CD. 2015. Genetics of autoimmune diseases: insights from population genetics. J Hum Genet. 60(11):657–664.

Ruiz F, et al. 2014. Natural single amino acid polymorphism (F19Y) in human galectin-8: detection of structural alterations and increased growth-regulatory activity on tumor cells. FEBS J. 5:1446–1464.

Sabeti PC, et al. 2006. Positive natural selection in the human lineage. Science 312(5780):1614–1620.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592.

Sebban L, et al. 2007. The involvement of CD44 and its novel ligand Galectin-8 in apoptotic regulation of autoimmune inflammation. J Immunol. 179:1225–1235.

Shen X, Salichos L, Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol Evol. 8(8):2565–2280.

Shi Y, et al. 2011. Galectin-3 rs4652 A>C polymorphism is associated with the risk of gastric carcinoma and P-glycoprotein expression level. Oncol Lett. 14:8144–8149.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51(3):492–508.

Siepel A, Haussler D. 2005. Phylogenetic hidden Markov models. In: Statistical Methods in Molecular Evolution. Statistics for Biology and Health. New York, NY: Springer.

Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency correlation. Mol Biol Evol. 34(11):2996–3005.

Song Y, et al. 2013. High-Resolution Comparative Modeling with RosettaCM. Structure. 21(10):1735-42

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server issue):W609–W612.

Than NG, et al. 2008. Emergence of hormonal and redox regulation of galectin-1 in placental mammals: implication in maternal–fetal immune tolerance. Proc Natl Acad Sci U S A. 105(41):15819–15824.

Than NG, et al. 2009. A primate subfamily of galectins expressed at the maternal–fetal interface that promote immune cell death. Proc Natl Acad Sci U S A. 106(24):9731–9736.

Than NG, et al. 2012. Galectins: guardians of eutherian pregnancy at the maternal–fetal interface. Trends Endocrinol Metab. 23(1):23–32.

Than NG, et al. 2015. Galectins: double-edged swords in the cross-roads of pregnancy complications and female reproductive tract inflammation and neoplasia. J Pathol Transl Med. 49(3):181–208.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74.

Vasta GR. 2012. Galectins as pattern recognition receptors: structure, function, and evolution. Adv Exp Med Biol. 946:21–36.

Villanueva-Canas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol. 5(2):457–467.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet. 47:97–120.

Weir B, Cockerham C. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Yoshida H, et al. 2012. X-ray structure of a protease-resistant mutant form of human Galectin-8 with two carbohydrate recognition domains. FEBS J. 279(20):3937–3951.

Zhang S, et al. 2015. Unraveling functional significance of natural variations of a human galectin by glycodendrimersomes with programmable glycan surface. Proc Natl Acad Sci U S A. 18:5585–5590.

Associate editor: Soojin Yi