Long-Term Balancing Selection at the Blood Group-Related Gene *B4galnt2* in the Genus *Mus* (Rodentia; Muridae)

Miriam Linnenbrink,[1,2] Jill M. Johnsen,[3,4] Inka Montero,[5] Christine R. Brzezinski,[3,4] Bettina Harr,[5] and John F. Baines*[,1,5]

[1]Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Kiel, Germany

[2]Department of Biology, University of Munich, Planegg-Martinsried, Germany

[3]Research Institute, Puget Sound Blood Center, Seattle, WA

[4]Division of Hematology, Department of Medicine, University of Washington, Seattle, WA

[5]Max Planck Institute for Evolutionary Biology, Plön, Germany

***Corresponding author:** E-mail: baines@evolbio.mpg.de.

Associate editor: Matthew Hahn

Abstract

Recent surveys of the human genome have highlighted the significance of balancing selection in relation to understanding the evolutionary origins of disease-associated variation. Cis-regulatory variation at the blood group–related glycosyltransferase *B4galnt2* is associated with a phenotype in mice that closely resembles a common human bleeding disorder, von Willebrand disease. In this study, we have performed a survey of the 5′ flanking region of the *B4galnt2* gene in several *Mus musculus* subspecies and *Mus spretus*. Our results reveal a clear pattern of trans-species polymorphism and indicate that allele classes conferring alternative tissue-specific expression patterns have been maintained for >2.8 My in the genus *Mus*. Furthermore, analysis of *B4galnt2* expression patterns revealed the presence of an additional functional class of alleles, supporting a role for gastrointestinal phenotypes in the long-term maintenance of expression variation at this gene.

Key words: *B4galnt2*, balancing selection, *Mus spretus*, *Mus musculus*, blood group, von Willebrand disease.

The phenomenon of balancing selection, whereby natural selection acts to maintain multiple alleles in a population, may arise by a number of diverse processes including a heterozygote advantage, frequency-dependent selection (Kojima 1971), and temporal or spatial variation in selective pressures (Hedrick et al. 1976; Gillespie 1978). Although the frequency and overall impact of balancing selection on the levels of diversity in natural populations have been a subject of debate since the collection of the first polymorphism data (Lewontin and Hubby 1966), a diverse set of individual examples exists (see Charlesworth 2006 for a review). Some of the first genome-level studies cast doubt on the existence of appreciable balancing selection in the human genome (Asthana et al. 2005; Bubb et al. 2006). However, a recent landmark study of polymorphism data in the human genome demonstrates the clear significance of this mode of selection (Andrés et al. 2009). Among their conservative list of 60 targets of balancing selection, roughly a third are known to be associated with human disease.

Studies of DNA sequence variation surrounding the blood group–related glycosyltransferase *β-1,4-N-acetylgalactosaminyl transferase-2* (*B4galnt2*) gene in house mice have uncovered the presence of two divergent haplotypes, one corresponding closely to the sequence of the RIIIS/J (RIII) inbred mouse strain and the other to the C57BL6/J (C57) strain (Johnsen et al. 2008, 2009). The RIII allele carries a cis-regulatory mutation that directs a remarkable tissue-specific switch in *B4galnt2* gene expression from its more common site in intestinal epithelium (observed in C57) to vascular endothelium. Vascular expression of *B4galnt2* results in the aberrant glycosylation of the clotting protein von Willebrand Factor (VWF), leading to accelerated VWF clearance from circulation and low VWF levels (Mohlke et al. 1999) similar to the common human bleeding disorder, Type 1 von Willebrand disease (Sweeney et al. 1990).

Both the RIII allele and the C57 allele are found in inbred mouse strains and natural *Mus musculus domesticus* populations, where the relationship between *B4galnt2* genotype and tissue-specific expression was confirmed (Johnsen et al. 2008, 2009). Striking signatures of natural selection were present in the populations studied by Johnsen et al. (2009), and simulation analyses revealed introgression alone as an unlikely explanation for these patterns. We proposed long-term balancing selection as the most likely explanation, but direct support for this hypothesis was lacking as the study surveyed only a single subspecies.

To determine whether long-term balancing selection played a role in the generation of extreme sequence divergence (up to 8%) between *B4galnt2* haplotypes, we extended our previous survey to ancestral populations of three house mice subspecies (*M. m. musculus*, *M. m. domesticus*, and *M. m. castaneus*) and the more distantly related *Mus spretus* (see animal material, Supplementary Material online). The haplotypes previously described in *M. m. domesticus* from France extended over ~60 kb. However, due to the ample opportunity for recombination, the signatures of long-term balancing selection are predicted to localize to

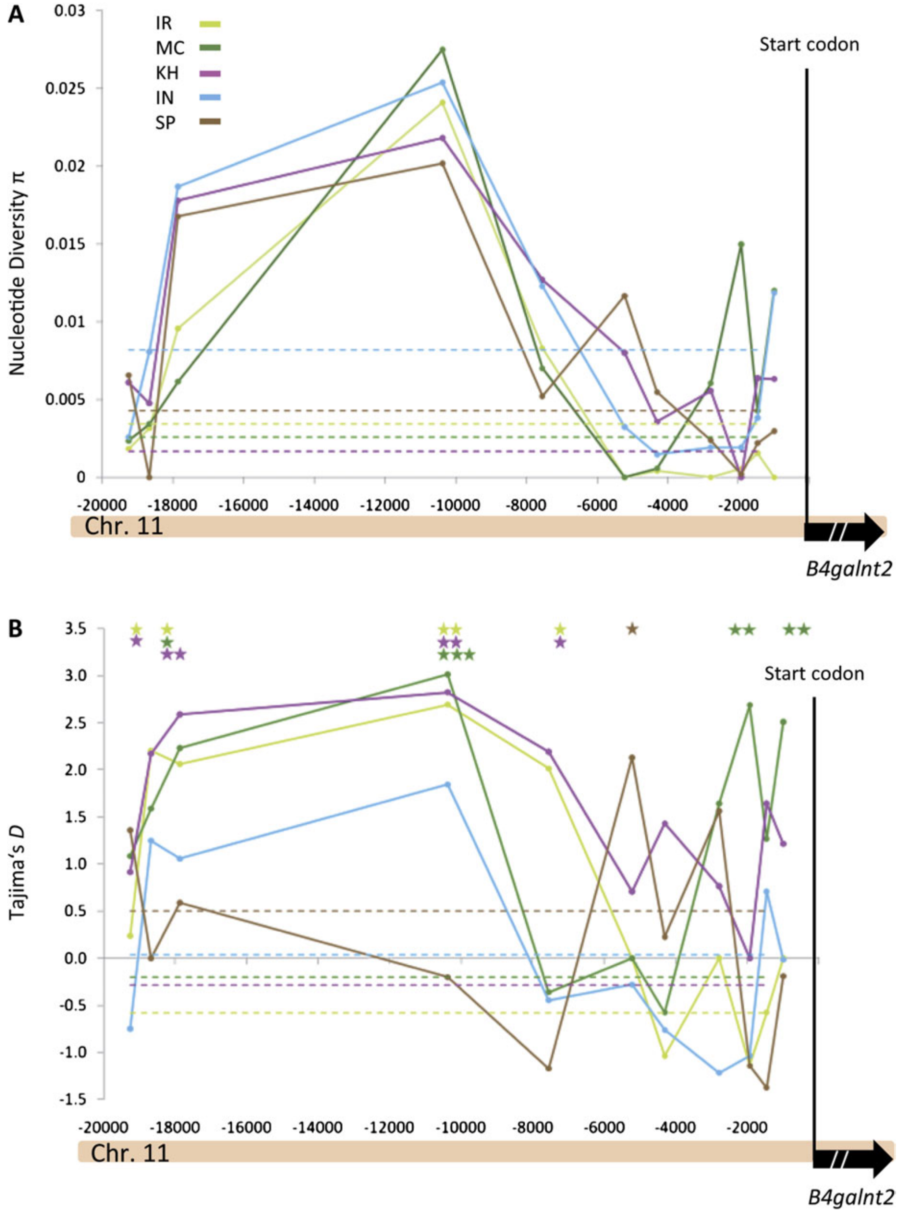

Fig. 1. (*a*) Nucleotide diversity and (*b*) Tajima's *D* across the *B4galnt2* upstream gene region. Populations analyzed were *M. m. domesticus* from Iran (IR, light green) and France (MC, dark green), *M. m. musculus* (KH, purple), *M. m. castaneus* (IN, blue), and *M. spretus* (SP, brown). Dashed lines represent average values at seven autosomal reference loci (Baines and Harr 2007) and this study (*M. spretus*). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

narrow regions (Charlesworth 2006). Thus, we increased the density of sequence fragments spanning the peak of polymorphism observed by Johnsen et al. (2009) and added the additional fragments to our previous data (supplementary fig. S1 and supplementary table S1, Supplementary Material online). In addition, we sequenced seven reference loci in the *M. spretus* sample (supplementary table S2, Supplementary Material online) for which

Table 1. Linkage Disequilibrium.

Population	Informative Sites[a]	Average r^{2b}	Pairwise Comparisons	Significant Comparisons[c] (%)	Range of Significant SNPs (bp)
MC	51	0.9	1128	1035 (91.76)	18878
IR	35	0.78	561	440 (87.43)	12141
KH	68	0.55	2211	1322 (59.8)	18933
IN	63	0.3	1953	472 (24.17)	17239
SP	75	0.43	2278	1288 (56.54)	13655

Note.—[a]Sites with at least two copies of the rarer variant present in the sample.

[b] Composite genotypic r^2, the squared correlation of genotypic indicators at two loci in diploid individuals, was calculated using the "composite_LD" function submitted to the Bioperl project (Stajich et al. 2002) as described in Johnsen et al. (2009).

[c] Based on the χ^2 test.

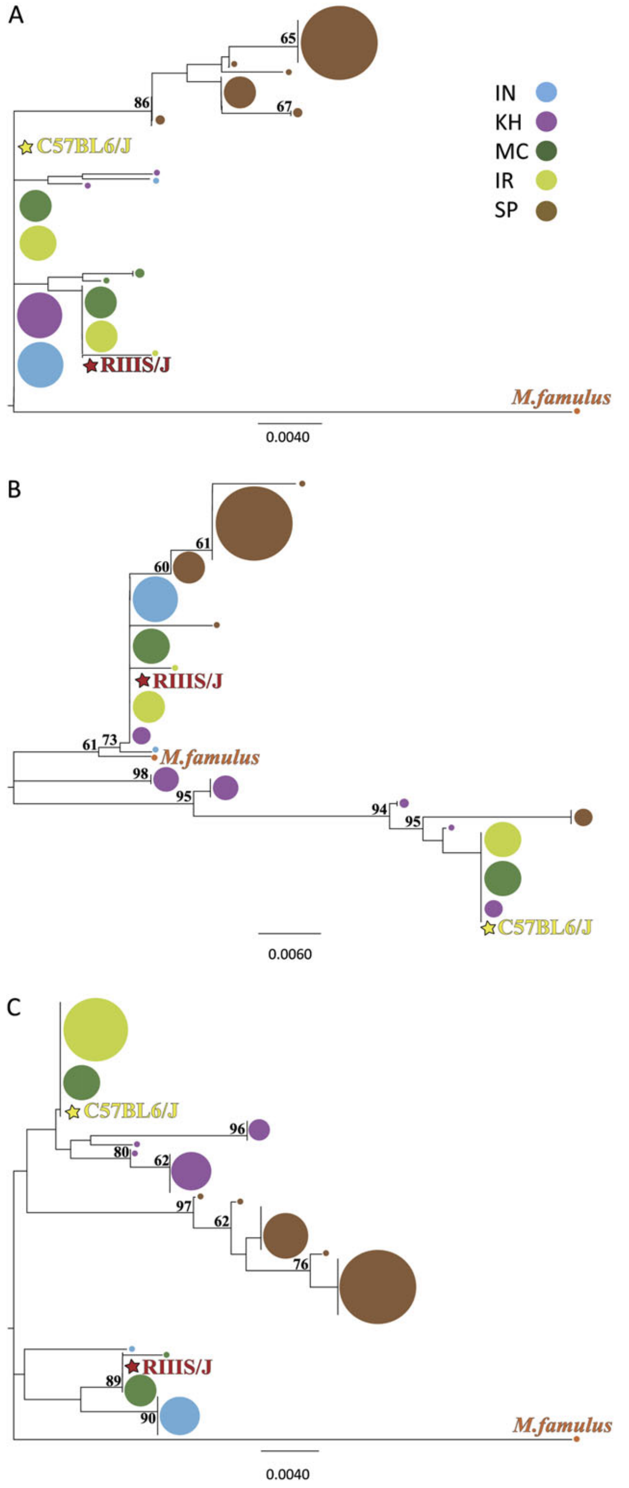

Fig. 2. Neighbor-Joining trees of *B4galnt2* upstream regions. Sizes of the circles are proportional to the number of occurrences. Sequences of RIIIS/J and C57BL6/J were included for reference and *M. famulus* as an outgroup. Trees are from sequences located (*a*) ~2, (*b*) ~10, and (*c*) ~20 kb upstream of *B4galnt2* (fragments 6.2, 5, and 3.5 in supplementary table S1, Supplementary Material online).

data from all other populations were available (Baines and Harr 2007) and thus provide the first information on DNA sequence polymorphism in this species.

As previously observed in *M. m. domesticus* from France, a peak of polymorphism approximately 10 kb upstream of the *B4galnt2* start codon is present in all species and populations, which displays a minimum of 3-fold up to ~13-fold higher levels compared with the panel of reference loci (fig. 1*a*). In three of the five populations, the fragments with elevated polymorphism also display significantly positive Tajima's *D* (Tajima 1989) values (fig. 1*b*). After closer inspection, the difference in Tajima's *D* between these and the remaining two populations is clearly due to differences in the frequency of divergent haplotypes (see below).

To analyze the pattern of haplotype variation, we estimated the phase of diploid sequences (Stephens et al. 2001; Stephens and Donnelly 2003). Two divergent haplotype classes similar in sequence to the RIIIS/J and C57BL6/J inbred mouse strains are ubiquitously present (supplementary fig. S2, Supplementary Material online). However, the extent of linkage disequlibrium (LD) differed by population (table 1). LD was highest in the *M. m. domesticus* population from France (average $r^2 = 0.9$), followed by Iran (average $r^2 = 0.78$). Values from the other subspecies/species were comparatively lower (range: 0.3–0.55). Due to the high number of pairwise comparisons, no association is significant after Bonferroni correction.

To further investigate the relationship between haplotypes, we examined representative sequence fragments using phylogenetic analysis. For this, we included a single individual of *M. famulus*, which shares a common ancestor with the studied taxa approximately 2.8 Mya (Ferguson et al. 2008). The sequences at approximately −2 and approximately−20 kb cluster largely according to species (fig. 2*a* and *c*). However, the pattern observed ~10 kb upstream of the start codon is particularly striking (fig. 2*b*). Although the three house mouse subspecies and *M. spretus* share a common ancestor 1.4 Mya (Ferguson et al. 2008), the sequences clearly cluster according to allele class rather than by species, demonstrating a clear pattern of trans-species polymorphism. Furthermore, the single *M. famulus* is most closely related to the RIII allele class.

To test whether the *B4galnt2* expression patterns conferred by the RIII and C57 haplotype classes are conserved in other populations and species, we performed *Dolichos biflorus* (DBA) lectin staining, which is specific for B4galnt2-carbohydrate residues (Johnsen et al. 2008, 2009). In wild-derived mice from *M. m. domesticus* (Iran) and *M. spretus* (Spain), we compared staining patterns with respect to *B4galnt2* genotype using amplicon #5 (~10 kb upstream) as a diagnostic marker (table 2). Wild-derived *M. m. musculus* (Kazakhstan) individuals harbored a recombinant C57 and RIII haplotype, which we termed "CRK" class. Thus, additional amplicons (as in the population data; supplementary table S1, Supplementary Material online) were sequenced in those individuals in order to correlate haplotype classes with their expression phenotype (table 2).

As expected, all individuals homo- or heterozygous for the C57 allele class exhibited lectin staining in the gastrointestinal tract (GI) tract, and all individuals homozygous for the RIII allele class exhibited loss of GI staining. However,

Table 2. Relationship Between Genotype (allele class) and DBA Lectin Staining Pattern.

Population	Allele Class	Gut+/Vessel−	Gut−/Vessel+	Gut+/Vessel+	Gut−/Vessel−
MC (Johnsen et al. 2009)	C57/C57	8[a]			
	C57/RIII			10	
	RIII/RIII		5		
	CRK/CRK				
IR	C57/C57	2			
	C57/RIII	5			
	RIII/RIII				10
	CRK/CRK				
SP	C57/C57	1			
	C57/RIII	7			
	RIII/RIII				19
	CRK/CRK				
KH	C57/C57	1			
	C57/RIII			11	
	RIII/RIII		6		
	CRK/CRK			3	

NOTE.—[a]Numbers indicate the sample size of each genotype tested per population-/species-of-origin.

in contrast to the blood vessel positive DBA lectin pattern observed in RIII-homozygous individuals from France (Johnsen et al. 2009) and Kazakhstan, all individuals with the RIII allele class from Iran and *M. spretus* failed to display a blood vessel staining pattern. Thus, a third functional class of alleles is present, which confers neither GI expression nor blood vessel *B4galnt2* expression but is related to the RIII-class alleles at the sequence level. To investigate whether this loss of expression might be due to a loss-of-function mutation, we sequenced all *B4galnt2* exons, intronic flanking sequences, and UTRs of four *M. spretus* (one C57 and three RIII homozygotes) and five *M. m. domesticus* (Iran) individuals (two C57 and three RIII homozygotes) but detected no variants predicted to disrupt the transcript or protein (supplementary fig. S3, Supplementary Material online). Surprisingly, CRK homozygotes displayed both bowel and blood vessel lectin staining indistinguishable from RIII-C57 heterozygotes, supporting a modular model of tissue-specific *B4galnt2* gene regulation in which two or more GI or vascular-specific regulatory elements lie within distinct genomic regions.

We here report a pattern of both great haplotypic and functional diversity at *B4galnt2* across the genus *Mus*. These results have important implications regarding the nature of the selective forces maintaining variation at *B4galnt2*. Together with extreme divergence between haplotypes, elevated polymorphism, and significantly positive Tajima's *D* values, the presence of these two distinct haplotype classes in all subspecies of *M. musculus* and the closely related species *M. spretus* provides strong evidence of long-term balancing selection. This adds to a growing list of examples of this mode of selection in mice such as ß-globin (Storz et al. 2007) and the *Oas1b* locus associated with West Nile virus infection (Ferguson et al. 2008).

Interestingly, the common pattern across all species, subspecies, and populations included in this study is the presence of B4galnt2-GalNAc residues on the GI tracts of any individual with the C57 allele class and the loss of these residues in all individuals homozygous for the RIII allele class. Although the function of *B4galnt2* is unknown, gene expression is conserved in the GI tract in vertebrates from fish (Stuckenholz et al. 2009) to humans (Montiel et al. 2003). We speculate that selection on glycosylation in the gut is a contributing factor to the long-term maintenance of this variation, likely by altering glycan-specific host–pathogen interactions.

Supplementary Material

Supplementary tables S1–S2 and figs. S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Anja Hörger for helpful discussion and Katja Cloppenborg-Schmidt for technical assistance and an anonymous reviewer for helpful comments. This work was supported by American Heart Association Award 0575033N (JMJ), Puget Sound Blood Center (JMJ and CRB), and Deutsche Forschungsgemeinschaft (DFG) Grant BA2863/2-1 and Excellence Cluster "Inflammation at Interfaces" (JFB). The GenBank accession numbers for the sequences reported in this paper are JN128296–JN128464.

References

Andrés AM, Hubisz MJ, Indap A, et al. (12 co-authors). 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26:2755–2764.

Asthana S, Schmidt S, Sunyaev S. 2005. A limited role for balancing selection. *Trends Genet.* 21:30–32.

Baines JF, Harr B. 2007. Reduced X-linked diversity in derived populations of house mice. *Genetics* 175:1911–1921.

Bubb KL, Bovee D, Buckley D, et al. (12 co-authors). 2006. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics.* 173:2165–2177.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2:e64.

Ferguson W, Dvora S, Gallo J, Orth A, Boissinot S. 2008. Long-term balancing selection at the west nile virus resistance gene, *Oas1b*, maintains transspecific polymorphism in the house mouse. *Mol Biol Evol.* 25:1609–1618.

Gillespie JH. 1978. A general model to account for enzyme variation in natural populations. V. The SAS-CFF model. *Theor Popul Biol*. 14:1–45.

Hedrick PW, Ginevan M, Ewing E. 1976. Genetic polymorphism in heterogeneous environments. *Annu Rev Ecol Syst*. 7:1–32.

Johnsen JM, Levy GG, Westrick RJ, Tucker PK, Ginsburg D. 2008. The endothelial-specific regulatory mutation, Mvwf1, is a common mouse founder allele. *Mamm Genome*. 19:32–40.

Johnsen JM, Teschke M, Pavlidis P, McGee BM, Tautz D, Ginsburg D, Baines JF. 2009. Selection on cis-regulatory variation at *B4galnt2* and its influence on von Willebrand factor in house mice. *Mol Biol Evol*. 26:567–578.

Kojima K. 1971. Is there a constant fitness for a given genotype? No!. *Evolution* 25:281–285.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and the degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54:595–609.

Mohlke KL, Purkayastha AA, Westrick RJ, Smith PL, Petryniak B, Lowe JB, Ginsburg D. 1999. Mvwf, a dominant modifier of murine von Willebrand factor, results from altered lineage-specific expression of a glycosyltransferase. *Cell* 96:111–120.

Montiel MD, Delannoy P, Harduin-Lepers A. 2003. Molecular cloning, gene organization and expression of the human UDP-GalNAc: neu5Acα2-3Galβ-R β1, 4-N-acetyl-galactosaminyltransferase responsible for the biosynthesis of the blood group Sd a/Cad antigen: evidence for an unusual extended cytoplasmic domain. *Biochem J*. 373:369–379.

Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: perl modules for the life sciences. *Genome Res*. 12:1611–1618.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 68:978–989.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. 73:1162–1169.

Storz JF, Baze M, Waite JL, Hoffmann FG, Opazo JC, Hayes JP. 2007. Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. *Genetics* 177:481–500.

Stuckenholz C, Lu L, Thakur P, Kaminski N, Bahary N. 2009. FACS-assisted microarray profiling implicates novel genes and pathways in zebrafish gastrointestinal tract development. *Gastroenterology* 137:1321–1332.

Sweeney JD, Novak EK, Reddington M, Takeuchi KH, Swank RT. 1990. The RIIIS/J inbred mouse strain as a model for von Willebrand disease. *Blood* 76:2258–2265.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.