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

1Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Kiel, Germany
2Department of Biology, University of Munich, Planegg-Martinsried, Germany
3Research Institute, Puget Sound Blood Center, Seattle, WA
4Division of Hematology, Department of Medicine, University of Washington, Seattle, WA
5Max 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 (Andrè 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 RIII/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...
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 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

![Image of Fig. 1](image-url)

**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.

### Table 1. Linkage Disequilibrium.

<table>
<thead>
<tr>
<th>Population</th>
<th>Informative Sites</th>
<th>Average $r^2$</th>
<th>Pairwise Comparisons</th>
<th>Significant Comparisons (%)</th>
<th>Range of Significant SNPs (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>51</td>
<td>0.9</td>
<td>1128</td>
<td>1035 (91.76)</td>
<td>18878</td>
</tr>
<tr>
<td>IR</td>
<td>35</td>
<td>0.78</td>
<td>561</td>
<td>440 (87.43)</td>
<td>12141</td>
</tr>
<tr>
<td>KH</td>
<td>68</td>
<td>0.55</td>
<td>2211</td>
<td>1322 (59.8)</td>
<td>18933</td>
</tr>
<tr>
<td>IN</td>
<td>63</td>
<td>0.3</td>
<td>1953</td>
<td>472 (24.17)</td>
<td>17239</td>
</tr>
<tr>
<td>SP</td>
<td>75</td>
<td>0.43</td>
<td>2278</td>
<td>1288 (56.54)</td>
<td>13655</td>
</tr>
</tbody>
</table>

**Note.**—aSites 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 $\chi^2$ test.
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. 1a). In three of the five populations, the fragments with elevated polymorphism also display significantly positive Tajima’s D (Tajima 1989) values (fig. 1b). 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 CS7BL6/J inbred mouse strains are ubiquitously present (supplementary fig. S2, Supplementary Material online). However, the extent of linkage disequilibrium (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. 2a and c). However, the pattern observed ~10 kb upstream of the start codon is particularly striking (fig. 2b). 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 CS7 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 CS7 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 CS7 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,
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 β-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


