Combining GWAS and F_{ST}-based approaches to identify targets of *Borrelia*-mediated selection in natural rodent hosts

Luca Cornetti[1,2] & Barbara Tschirren[3*]

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[2]Zoological Institute, University of Basel, Basel, Switzerland

[3]Centre for Ecology and Conservation, University of Exeter, Treliever Road, Penryn, TR10 9FE, United Kingdom

*Correspondence:
Barbara Tschirren, Email: b.tschirren@exeter.ac.uk

Running title: Targets of *Borrelia*-mediated selection

Abstract

Recent advances in high-throughput sequencing technologies provide opportunities to gain novel insights into the genetic basis of phenotypic trait variation. Yet to date, progress in our understanding of genotype-phenotype associations in non-model organisms in general and natural vertebrate populations in particular has been hampered by small sample sizes typically available for wildlife populations and a resulting lack of statistical power, as well as a limited ability to control for false positive signals. Here we propose to combine a genome-wide association (GWAS) and F_{ST}-based approach with population-level replication to partly overcome these limitations. We present a case study in which we used this approach in combination with Genotyping-by-Sequencing (GBS) SNP data to identify genomic regions associated with *Borrelia afzelii* resistance or susceptibility in the natural rodent host of this Lyme disease-causing spirochete, the bank vole (*Myodes glareolus*). Using this combined approach we identified four consensus SNPs located in exonic regions of the genes *Slc26a4, Tns3*, *Wscd1* and *Espnl*, which were significantly associated with the voles' *Borrelia* infectious status within and across populations. Functional links between host responses to bacterial infections and most of these genes have previously been demonstrated in other rodent systems, making them promising new candidates for the study of evolutionary host responses to *Borrelia* emergence. Our approach is applicable to other systems and may facilitate the identification of genetic variants underlying disease resistance or susceptibility, as well as other ecologically relevant traits, in wildlife populations.

Keywords: host-parasite interactions, wild immunogenetics, pathogen-mediated selection, evolutionary change, RAD sequencing (RAD-seq), conservation genetics

Introduction

Testing evolutionary theories of host-parasite interactions and resistance evolution is hampered by a lack of understanding of the genetic architecture of host defence and susceptibility (Lazzaro & Little, 2009). This is particularly the case for natural populations of non-model organisms, for which very little functional genetic information is currently available (Spurgin & Richardson, 2010). Infectious diseases are a major cause of wildlife population declines (Smith, Sax, & Lafferty, 2006) and pose a substantial threat to global biodiversity (MacPhee & Greenwood, 2013). At the same time, wildlife diseases can spillover into human populations and are thus of public health concern (Daszak, Cunningham, & Hyatt, 2000; Jones et al., 2008; Wiethoelter, Beltrán-Alcrudo, Kock, & Mor, 2015). A better understanding of the genetic basis of wildlife disease resistance or susceptibility is thus crucial for conservation efforts (Jones et al., 2007; Margres et al., 2018) but also to understand and predict the dynamics of zoonotic diseases (Beldomenico & Begon, 2010; Price, Spencer, & Donnelly, 2015).

Recent advances in high-throughput sequencing technologies have made it possible to obtain extensive genomic information for non-model organisms (Tagu, Colbourne, & Nègre, 2014). Yet challenges to associate genetic variants with phenotypic traits of interest remain formidable (e.g. Hong & Park, 2012). Two main approaches are typically used to identify genomic regions of interest. The first approach, genome-wide association (GWAS)(Amos, Driscoll, & Hoffman, 2011), tests for associations between phenotypic traits of interest and genomic variants across the whole genome (Petersen, Fredrich, Hoeppner, Ellinghaus, & Franke, 2017). A second approach (F_{ST} outlier approach; Vitti, Grossman, & Sabeti, 2013) used to identify genomic regions associated with phenotypic trait variation is based on the premise that natural selection acting on a locus of interest will result in differences in allele frequencies among populations subject to different environmental conditions or showing different phenotypes. The F_{ST} outlier approach identifies regions with unusually large (when compared to the genome-wide F_{ST} distribution) genetic differentiation between populations suggesting that they are under selection.

Although these two approaches have been successfully used to identifying genomic regions of interest in humans (e.g. Visscher et al., 2017) and model organisms (e.g. Flint & Eskin, 2012; Togninalli et al., 2018; Wangler, Hu, & Shulman, 2017), so far outcomes were mixed for non-model organisms in general, and natural vertebrate populations in particular (Santure & Garant, 2018). A key limitation is that sample sizes available for GWAS in wildlife populations are typically several magnitudes smaller than for humans or model organisms (Amos et al., 2011; Hong & Park, 2012). Furthermore, it remains notoriously difficult to control for false positive signals when testing for associations between genetic variants and phenotypes (Amos et al., 2011; McCarthy et al., 2008), and distinguishing between molecular signals of natural selection and genetic drift with F_{ST} outlier approaches is non-trivial (Hoban et al., 2016; Vitti et al., 2013).

In order to increase statistical power to identify genomic regions associated with phenotypic traits of interest, and at the same time control for false positive signals, it has been suggested to combine approaches and to apply population level replication (Chanock et al., 2007; Santure & Garant, 2018; Schielzeth, Rios, & Burri, 2018). Yet, to our knowledge this approach (i.e. the combination of GWAS and F_{ST} -based tests and application of population-level replication) has not been used to identify targets of pathogen-mediated selection in wildlife disease systems to date.

Borrelia afzelii is the most common *Borrelia* genospecies in Europe and one of the causative agents of human Lyme disease (Steere, Coburn, & Glickstein, 2004). It is transmitted by ticks (*Ixodes* sp.) and rodents, such as the bank vole (*Myodes glareouls*), are its main natural hosts (Kurtenbach et al., 2006; Mannelli, Bertolotti, Gern, & Gray, 2012). Recently, it has been experimentally shown that *B. afzelii* has negative fitness consequences for bank voles (Cayol et al., 2018). Defense mechanisms that prevent or control *Borrelia* infections in natural hosts will thus be favored by natural selection.

Using a candidate gene approach, we have previously demonstrated that naturally occurring genetic variants of Toll-like receptor 2 (*TLR2*) are associated with the *Borrelia* infection status of bank voles (Cornetti et al., 2018; Tschirren et al., 2013). Yet, *Borrelia* susceptibility is most likely influenced by many genes and the variation explained by a single candidate gene remains limited (Wilfert & Schmid-Hempel, 2008). In this study, we performed genome-wide scans to identify potential targets of *Borrelia*-mediated selection using a combination of complementary approaches and by applying a population replication criterion. Specifically, we used (1) a GWAS approach to identify genetic variants associated with *Borrelia* infection status, and (2) a F_{ST}-based analysis between *Borrelia*-infected and -uninfected bank voles within seven independent populations to identify outlier SNPs. The latter was combined with (3) a population level replication criterion in which we considered only SNPs that were identified as outliers in multiple populations. Finally, we (4) overlaid the results of these complementary approaches to identify consensus candidate SNPs, which were found to explain significant amount of variation in bank vole *Borrelia* infection status.

Materials and Methods

Field sampling

Bank voles (*Myodes glareolus*) were captured during summer 2014 at seven locations in the Kanton Graubünden, Switzerland (Table 1; Supporting information Figure S1) using live-traps (Longworth Mammal Traps, Anglian Lepidopterist Supplies). A high *Borrelia* infection prevalence has previously been documented in bank voles at these sampling sites (Cornetti et al., 2018). Caught bank voles (N = 177; Table 1) were weighed (to the nearest g), aged following Gliwicz (1988)(adults (>20 g), subadults (15-20 g), and juveniles (< 15 g)), and a small ear biopsy was collected and stored in 95% ethanol. The animals were then released at their capture site. Vole capture, handling and tissue sampling complied with the current laws of Switzerland and were performed under a license issued by the Department of Food Safety and Animal Health of the Kanton Graubünden, Chur, Switzerland (permit number 2012_17).

Borrelia infection in bank voles

Genomic DNA was extracted from the ear biopsy using the QIAGEN DNeasy Blood & Tissue Kit (Qiagen, Venlo, the Netherlands). To determine the *B. afzelii* infection status of bank voles, we used a highly sensitive quantitative real-time PCR (qPCR) assay using the *flaB B. afzelii*-specific primers Fla5F: 5‘-CACCAGCATCACTTTCAGGA-3‘ and Fla6R: 5‘-CTCCCTCACCAGCAAAAAGA-3‘ (Råberg, 2012) on a StepOnePlus real-time qPCR machine (Applied Biosystems, Foster City, CA, USA). We focused on *B. afzelii* because a pilot study using reverse line blot (Herrmann et al., 2013) had revealed that *B. afzelii* is the only *Borrelia* genospecies present in bank voles at our study sites (unpublished data).

The amplification was carried out in a total volume of 20 µl, including 10 µl SYBR® Select Master Mix (2x, Applied Biosystems), 0.8 µl of each primer (10 µM) and 4 µl extracted genomic DNA. The qPCR protocol consisted of two initial holding steps first at 50 °C and then at 95 °C for 2 min each, followed by 42 cycles of 95 °C for 15 sec, 59 °C for 30 sec, and 72 °C for 30 sec (Råberg, 2012). Eight negative controls and eight serially diluted positive controls were included on each plate. Samples with a cycle threshold (Ct) value > 0 and a melting temperature between 76.4 °C and 77.8 °C were considered to be *B. afzelii*-positive (Råberg, 2012). All samples were analyzed in duplicate on two different plates (see Cornetti et al., 2018 for details).

Genotyping-by-Sequencing and SNPs calling

Samples of 118 adult bank voles were used for Genotyping-by-Sequencing (GBS). Only adult bank voles were included because variation in *Borrelia* infection status among juveniles and subadults is likely due to variation in exposure rather than resistance (Tschirren et al., 2013). An equal number of *Borrelia*-infected and uninfected individuals were randomly selected for each site, whenever possible. Overall, 45% of the sequenced individuals were *B. afzelii* infected (Table 1).

Extracted genomic DNA was sent to the GBS platform (http://www.biotech.cornell.edu) of Cornell University, USA in July 2015. GBS libraries were prepared using a double digest protocol with *SbfI* and *HpaII* as restriction enzymes (Poland, Brown, Sorrells, & Jannink, 2012). Sequencing (100-bp single-end reads) was performed on IlluminaHiSeq 2500.

In total 264,483,546 reads were obtained. Illumina adapters were removed from raw sequences using Trimmomatic 0.33 (Bolger, Lohse, & Usadel, 2014) . Sequences were aligned to the prairie vole (*Microtus ochrogaster*) reference genome (MicOch1.0, (McGraw, Davis, Young, & Thomas, 2011)) using Bowtie2 (Langmead & Salzberg, 2012). The prairie vole and the bank vole are members of the same subfamily (Arvicolinae) and their divergence time has been estimated to be 5.9±0.8 Mya (95% confidence interval: 4.6-7.6 Mya) based on nuclear genes (Abramson, Lebedev, Tesakov, & Bannikova, 2009). To date, the prairie vole is the closest relative of the bank vole with a high quality genome assembly (McGraw et al., 2011). The current version of the prairie vole reference genome consists of 28 main scaffolds, corresponding to 17 autosomes, the X chromosome and ten linkage groups (Zerbino et al., 2018). The average mapping rate to the prairie vole genome was about 80%. Samtools 1.3 was used to filter the BAM alignments for quality (-q 20) before SNP calling was performed with GATK 3.7 (Van der Auwera et al., 2013). The final set of SNPs was obtained with VCFtools 0.1.15 (Danecek et al., 2011) by filtering for quality (minimum genotype quality score of 20), coverage (minimum genotype depth of 6 per individual) and rare variants (minor allele frequency of 0.01), requiring that at least 70% of all individuals passed the filters.

Bank vole population structure

Population structuring was assessed using a Multi-Dimensional Scaling (MDS) approach implemented in Plink 1.90 (Chang et al., 2014), as well as using the software Structure 2.3.4 (Pritchard, Stephens, & Donnelly, 2000). As input for Structure we used a reduced dataset of 1555 SNPs, which included variants with no missing data, and, to fulfill Structure model assumptions of independence of loci (i.e. no linkage disequilibrium within populations), only one SNP per read. Analyses were performed using an admixture model with correlated allele frequencies for ten independent runs. We determined the most likely number of genetic clusters (*K*) exploring *K* values between one and seven (i.e. the number of sampling sites). Burn-in periods of 100,000 were used, followed by 500,000 iterations. The most plausible number of genetically well-defined groups was determined by comparing the likelihood at different *K* values (Pritchard et al., 2000) using Structure Harvester (Earl & VonHoldt, 2012).

Furthermore, we calculated the fixation index F_{ST} among populations as a measure of population differentiation using the software Arlequin version 3.5 (Excoffier & Lischer, 2010) and tested for isolation-by-distance (IBD), by correlating the pairwise genetic differentiation and geographic distance among populations using Mantel test (Mantel, 1967). Linear geographic distances between locations were calculated with the Geographic Distance Matrix Generator version 1.2.3 (Ersts, 2017). The relationship between the linearized

F_{ST} (F_{ST} /(1- F_{ST})) and the log-transformed linear geographic distance (log(km)) was estimated using the Isolation-by-Distance Web Server (Jensen, Bohonak, & Kelley, 2005).

Identifying targets of Borrelia-mediated selection

To identify putative targets of *Borrelia*-mediated selection we used a combination of two approaches. First, we tested for an association between bank vole SNPs and *Borrelia* infection status using the *R* package GenAbel (Aulchenko, Ripke, Isaacs, & van Duijn, 2007). GenAbel allows performing genome-wide association (GWAS) between SNPs and a phenotype while correcting for population structure. We first computed a kinship matrix for our samples using the function *ibs*. Then, using an Eigenstrat method, we calculated the probability of each SNP to be associated with the phenotype (i.e. *Borrelia* infection status), after correcting for kinship within the whole dataset (Aulchenko et al., 2007) using the function *egscore*. This method uses the genomic kinship matrix to derive axes of genetic variation (principal components) and adjusts both the trait (i.e. *Borrelia* infection status) and the genotypes onto these axes (Price et al., 2006). Corrected genotypes are defined as residuals from regression of genotypes onto axes. Correlation between corrected genotypes and the phenotype is computed, and test statistics is defined as the square of this correlation times (N - K - 1), where N is the number of genotyped subjects and K is the number of axes (Aulchenko et al., 2007). The association analysis was performed using the combined dataset (i.e. all individuals, N = 118, across all populations and all 21,811 SNPs).

Second, we complemented the GWAS approach with an independent analysis based on F_{ST} to identify outlier SNPs between *Borrelia*-infected and uninfected bank voles within populations. In each of the seven populations, we calculated for each SNP the F_{ST} value between *Borrelia*-infected and uninfected individuals using VCFtools 0.1.15 (Danecek et al., 2011). The rationale of this approach is that no neutral population differentiation is expected between infected and uninfected animals within a population, and significantly differentiated SNPs suggest an association with *Borrelia* resistance or susceptibility. Within each population, we selected the outlier SNPs that lay within the top 10% of the population-specific F_{ST} distribution (Bankers et al., 2017; Myles, Davison, Barrett, Stoneking, & Timpson, 2008; Zueva et al., 2014). Given the significant population differentiation and distinct genetic clustering of the sampled bank vole populations (see Results), these populations represent independent replicates.

To control for false positive F_{ST} outliers within populations, we applied a population replication criterion and only considered F_{ST} outliers that were among the top 10% of the population-specific F_{ST} distribution in at least three of the seven populations (40%) (see Simulations below). This approach is conservative because it assumes that the same genetic mechanisms underlie variation in *Borrelia* resistance or susceptibility in different populations. Because genetic variants underlying variation in resistance or susceptibility to infectious diseases are typically found to be exonic (Hill, 2012), we specifically focused on SNPs located in exons in both approaches.

In a final step, we overlapped the results of the GWAS approach and of the population-level replicated F_{ST}-based approach to identify consensus candidate SNPs that were associated with *Borrelia* infection status in both analyses. Candidate SNPs were annotated with SNPdat (Doran & Creevey, 2013), a tool specifically designed for non-model organisms. We then used a generalized linear mixed model with a binomial error structure and site included as a random effect to test for associations between these consensus candidate SNPs and *Borrelia* infection status. Significance of fixed effects was determined by comparing nested models (with and without the variable of interest) using likelihood ratio tests. We considered both additive and dominant modes of gene action of candidate SNPs. Analyses were performed using the package lme4 (Bates, Maechler, & Bolker, 2011) in *R* version 3.6.1. (R Core Team, 2014).

Neutral simulations

We performed neutral simulations to quantify the false-positive rates when using a GWAS approach alone, a F_{ST} approach without population replication, a F_{ST} approach with a two population replication criterion and a F_{ST} approach with a three population replication criterion, as well as a combined approach that focuses on consensus SNPs identified with both approaches to identify putative targets of *Borrelia*-mediated selection. We used the forward-time genetic simulator SLiM version 3.3 (Haller & Messer, 2019) to generate neutral polymorphisms based on the number of SNPs (N = 21’811) and individuals (N = 118) included in the empirical dataset, repeated 100 times. The commented SLIM script describing the simulation process is presented in the Supporting information. In short, we generated sequences of 27’999 bp in length, with a mutation rate of 2×10^{-4} per base per generation and a recombination rate of r = 0.05. At generation 1, seven subpopulations appear and evolve independently for 4999 generations with some gene flow among them. At the end of the simulations (i.e. after 4999 generations) a subsample of each population is randomly selected according to the actual sample size of the empirical data. The data is written in seven VCF files that are used for further analyses.

The parameters used in the simulations were selected, after many pilot runs, by taking into account computational and temporal constraints and in order to obtain a total number of SNPs and population differentiation (in particular in term of F_{ST} values) similar to the ones observed in the real dataset. For each of the 100 simulated datasets, we randomly assigned infected and uninfected individuals within each population, according to the real data (for example, in Sagong, 10 infected and 9 uninfected individuals were defined). Then, the seven VCF files were merged using GATK version 4 (McKenna et al., 2010); during this step, we also constrained the total number of SNPs and the amount of missing data to be comparable to the observed data. The resulting 100 VCF files, including 118 samples from seven populations, were used to perform GWAS with Genabel (Aulchenko, Ripke, Isaacs, & van Duijn, 2007) based on the full dataset (118 samples), and F_{ST}-based calculation between infected and uninfected individuals within each of the populations.

Results

Bank vole population structure

A total of 21,811 SNPs were retained after quality filtering. Population differentiation (F_{ST}) was comparably high considering the relatively small size of the study area (maximum linear distance between sampling sites: 34 km), and varied from 0.052 (Sagogn-Flims, the second closest locations) to 0.111 (Malans-Sagogn, the most distant locations, Supporting information Table S1). All F_{ST} values were statistically significant (Supporting information Table S1).

Within the study area, population structure was well defined. The MDS analysis highlighted seven distinct groups corresponding to the seven sampling locations (Figure 1). Similarly, using Structure we found that $K = 7$ was the most supported partition (Supporting information Figure S2), with the seven well defined genetic clusters corresponding to the seven sampling locations (Figure 1, Supporting information Figure S3). The relationship between genetic and geographic distance was positive (r = 0.87, Supporting information Figure S4) and statistically significant (Mantel test, $p < 0.001$), suggesting pronounced isolation-by-distance across populations.

Neutral simulations

The neutral simulations demonstrated the weaknesses of using GWAS and F_{ST} -based tests in isolation (Supporting information Figures S5 and S6) and the significant reduction in false-positive rates when using a three-population replication criterion for the F_{ST}-based test (Supporting information Figure S5). The simulations furthermore showed that combining GWAS and F_{ST} -based tests with population replication substantially reduces false-positive rates, and thus increases the power to identify core candidate SNPs (Figure 2).

Identifying targets of Borrelia-mediated selection

a) GWAS approach

Using a GWAS approach that corrects for population structure, we identified 1065 SNPs that were associated ($p < 0.05$) with the *Borrelia* infection status of bank voles (Supporting information Figure S6). As expected given the comparably small number of individuals included in this GWAS, none of these associations reached statistical significance when accounting for multiple testing using the Benjamini-Hochberg procedure (Supporting information Figure S6). The identified SNPs were distributed across the whole genome (Supporting information Figure S7). After correcting for chromosome size, chromosome 15 and chromosome 19 showed the highest and lowest number of putatively *Borrelia*-associated SNPs, respectively (chr 15: 1.02 SNPs/MB; chr 19: 0.17 SNPs/MB; Supporting information Figure S7). Among the 1065 SNPs, 53 were located in exonic regions of 48 unique genes (Supporting information Table S2).

b) F_{ST}-based approach with population-level replication

In a second step, we performed a F_{ST}-based analysis between *Borrelia* infected and non-infected individuals within each of the seven populations and applied a population-level replication criterion, by considering only F_{ST} outliers that were among the top 10% of the population-specific F_{ST} distribution in at least three of the seven population. The top 10% F_{ST} threshold ranged from

0.079 (Flims) to 0.145 (Rodels, Supporting information Figure S8). We obtained 305 SNPs that showed consistent differentiation between *Borrelia*-infected and uninfected bank voles in at least three independent population replicates, 70 of which were also identified using the GWAS approach (Figure 2). Among the 305 SNPs, nine were located in exonic regions of nine unique genes (Table 2).

c) Consensus candidate loci

When overlapping the results of the GWAS and the F_{ST}-based test, there were four SNPs in coding regions that were associated with *Borrelia* infection status in both analyses (Table 2). These four SNPs were located in the exonic regions of the genes *Slc26a4* (Solute carrier family 26, member 4), *Tns3* (Tensin 3), *Wscd1* (WSC domain containing 1) and *Espnl* (Espin-like). The less frequent allele of *Slc26a4* (allele A; χ^2 = 4.414, *DF* = 1, *P* = 0.036) and *Tns3* (allele A; GLMM: χ^2 = 6.859, *DF* = 1, *P* = 0.009) were associated with a lower probability of *Borrelia* infection (Figure 3, Supporting information Table S3), whereas the less frequent allele of *Wscd1* (allele G; χ^2 = 5.790, *DF* = 1, *P* = 0.016) and *Espnl* (allele C; χ^2 = 6.488, *DF* = 1, *P* = 0.011) were associated with a higher probability of *Borrelia* infection (Figure 3, Supporting information Table S3). Models that treated the heterozygous and homozygous state of the less frequent allele as separate genotypes are presented here. Models that combined the heterozygous and homozygous state of the less frequent allele are presented in Supporting information Table S4.

Discussion

For most non-model organisms, and wild-living vertebrates in particular, the genetic basis underlying infectious disease resistance or susceptibility is poorly understood (Spurgin & Richardson, 2010), which hampers progress in our understanding of the eco-evolutionary dynamics of host-parasite interactions and wildlife disease. Here we present a case study in a natural rodent-*Borrelia* system in which we combined a GWAS and a F_{ST}-based approach with population-level replication to identify genomic regions associated with variation in *Borrelia* infection status.

Borrelia prevalence in bank voles was high at all seven study sites, with 31-61% of adult bank voles being *Borrelia* infected across sites. Given the negative fitness consequences a *Borrelia* infection has for the rodent host (Cayol et al., 2018), natural selection is expected to favour mechanism that control or prevent infection. Using a GWAS approach we identified a large number (1065) of SNPs that were associated ($p < 0.05$) with the voles' *Borrelia* infection status while controlling for population structure. Given the relatively relaxed criteria that were applied to identify genotype-phenotype associations in this GWAS approach, many of these SNPs were likely false positives. Indeed, the risk of false positive signals in GWAS increases when a large number of SNPs but a small number of individuals are included (Hong & Park, 2012), which was the case in our study, and is typical for most studies on wildlife populations.

In order to increase statistical power to detect true signals while controlling for false positives, we complemented the GWAS approach with an F_{ST}-based analysis, replicated across populations. We observed pronounced (given the relatively small geographical distances among study sites) genetic differentiation of bank voles across study sites and strong isolation-by-distance. Furthermore, Bayesian assignment analysis identified seven well defined genetic clusters, corresponding to the seven sampling sites. These results indicate that the seven bank vole populations are sufficiently isolated to represent independent replicates for the identification of targets of *Borrelia*-mediated selection. Within each population we identified SNPs lying within the top 10% of the population-specific F_{ST} distribution and considered SNPs that were among these extreme 10% in multiple populations to be putatively associated with *Borrelia* infection status. This F_{ST} -based approach assumes that the same mechanisms underlie *Borrelia* resistance or susceptibility in multiple populations, which does not necessarily need to be the case. Indeed, quantitative trait loci (QTL) are often found to be population-specific (Schielzeth et al., 2018; Tschirren & Bensch, 2010), and previous work in other systems has demonstrated variation in host responses to pathogen infection across populations (Bankers et al., 2017; Kurtz et al., 2014). Applying a population-level replication criterion is conservative and will prevent identifying such population-specific SNPs associated with *Borrelia* infection status, yet it also allows us to control for false positive signals. Indeed, the neutral simulations demonstrated the power of using a population replication criterion to reduce false positive signals.

The simulations furthermore showed that false-positive rates are further reduced when GWAS and F_{ST}-based approaches are combined. Using such a combined apporach, we identified four consensus polymorphisms located in exonic regions representing promising targets of *Borrelia*-mediated selection in bank voles. Interestingly, for most of the genes in which the consensus SNPs were located, a functional association with response to bacterial infection has previously been demonstrated in rodents: *Slc26a4* encodes the anion exchanger Pendrin and is expressed in membranes of ion-transporting epithelia where it regulates luminal pH and fluid transport (Royaux et al., 2000). *Slc26a4* has been found to be significantly upregulated in mouse macrophages experimentally stimulated with live *B. burgdorferi* (Gautam et al., 2011). Similarly, mice infected with the bacterial pathogen *Bordetella pertussis*, the etiological agent of whooping cough, exhibited significant *Slc26a4* upregulation (Scanlon et al., 2014). *Tns3* encodes a phosphoprotein thought to act as a link between the extracellular matrix and the cytoskeleton (Lo, 2004). This gene was found to be significantly upregulated in mice experimentally infected with the bacterium *Mycobacterium bovis*, the main etiological agent of bovine tuberculosis (Aranday-Cortes et al., 2012). Another member of the tensin gene family, *Tensin 1*, was downregulated and upregulated 4 h and 24 h respectively, after stimulation of mouse macrophages with live *B. burgdorferi* (Gautam et al., 2011), suggesting a general role of the tensin gene family in the response to bacterial infections. *Wscd1* encodes a protein with sulfotransferase activity (Smith, Blake, Kadin, Richardson, & Bult, 2018). *Wscd1* has been found to be significantly downregulated in mice five days after infection with the bacterium *Yersinia pseudotuberculosis* (Heine et al., 2018). Taken together, these previous findings suggest that several genes identified as candidates in our study play a role in the response to bacterial infections in rodents. However, the exact mechanisms by which these genes may confer resistance or susceptibility to *Borrelia* in bank voles are currently unknown and will be the focus of future work.

Previously, we have demonstrated that naturally occurring polymorphisms at the innate immune receptor Toll-like receptor 2 (*TLR2)* are significantly associated with the *Borrelia* infection status of bank voles in multiple populations, including the current study population (Cornetti et al., 2018; Tschirren et al., 2013; Tschirren, 2015). These findings were in line with biomedical research that has identified *TLR2* as a candidate gene for *Borrelia* resistance in laboratory mice (Alexopoulou et al., 2002; Dennis et al., 2009; Singh & Girschick, 2006; Wooten et al., 2002). Interestingly, however, *TLR2* was not identified as a potential candidate gene in this study, also not when considering the results of GWAS and F_{ST}-based approaches separately. It demonstrates the limitations of reduced-representation sequencing approaches, which do not cover the whole genome (Davey & Blaxter, 2010). Depending on the density and distribution of SNPs, as well as recombination rates in regions of interest, the power to detect signals might be low. In our study, no GBS SNP was close enough to possibly pick up signals of selection acting on *TLR2*. In fact the closest SNP was about 46kb away from *TLR2*, a physical distance larger than the linkage disequilibrium decay estimated for natural rodent populations (r^2<0.2 ~20 kb; Staubach et al, 2012). This suggests that some other putative candidate regions were likely missed with our approach because of insufficient SNP coverage. At the same time, we can exclude the possibility that the SNPs identified in our study were physically linked to *TLR2*. *TLR2* is located on chromosome 1 in the prairie vole, whereas *Tns3, Wscd1* and *Espnl* are located on chromosome 7, linkage group 1 and linkage group 4, respectively. *Slc26a4* is located on chromosome 1 as well, but separated by more than 40 Mb from *TLR2*. One possible way to obtain a more conclusive list of candidate genes associated with *Borrelia* infectious status would be to increase SNP density or sequence the whole genome. However, costs associated with especially the latter approach are still excruciatingly high, in particular when large numbers of individuals are included and the study species has a large genome size, as is the case for mammals (Catchen et al., 2017).

In conclusion, by combining GWAS and F_{ST}-based approach with population-level replication we identified consensus SNPs in exonic regions of genes for which a functional association with host responses to bacterial infections has previously been demonstrated. These loci thus represent promising new candidate genes that may allow tracking evolutionary changes in host populations in response to *Borrelia* emergence. More generally, the combined approach used in this case study can be applied to other systems and may contribute to a better understanding of genotype-phenotype associations in wildlife populations.

Acknowledgements

This study was supported by the University of Zurich Research Priority Program “Evolution in Action: from Genomes to Ecosystems”, the Faculty of Science of the University of Zurich, the Baugarten Stiftung and the Stiftung für wissenschaftliche Forschung an der Universität Zürich. We thank the numerous people who contributed to sample collection in the field, Peter

Fields for suggestions on genetic simulations, and Jacek Radwan and two anonymous reviewers for their constructive comments on the manuscript.

Author Contributions

LC and BT designed the research, LC performed laboratory work and analysed the data, LC and BT wrote the paper.

Data accessibility

The Genotyping-by-Sequencing data are deposited in NCBI BioProject PRJNA306409, SRA experiment SRR3031372. The quality filtered SNP file used for population genomic analyses and information on population and infection status of individual bank voles are deposited in the Dryad repository doi:10.5061/dryad.c866t1g3t.

References

Abramson, N. I., Lebedev, V. S., Tesakov, A. S., & Bannikova, A. A. (2009). Supraspecies relationships in the subfamily Arvicolinae (Rodentia, Cricetidae): An unexpected result of nuclear gene analysis. *Molecular Biology*, *43*(5), 834–846. doi:10.1134/S0026893309050148

Alexopoulou, L., Thomas, V., Schnare, M., Lobet, Y., Anguita, J., Schoen, R. T., … Flavell, R. A. (2002). Hyporesponsiveness to vaccination with *Borrelia burgdorferi* OspA in humans and in *TLR1*- and *TLR2*-deficient mice. *Nature Medicine*, *8*(8), 878–84. doi:10.1038/nm732

Amos, W., Driscoll, E., & Hoffman, J. I. (2011). Candidate genes versus genome-wide associations: Which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society B: Biological Sciences*, *278*(1709), 1183–1188. doi:10.1098/rspb.2010.1920

Aranday-Cortes, E., Hogarth, P. J., Kaveh, D. A., Whelan, A. O., Villarreal-Ramos, B., Lalvani, A., & Vordermeier, H. M. (2012). Transcriptional profiling of disease-induced host responses in bovine tuberculosis and the identification of potential diagnostic biomarkers. *PLoS ONE*, *7*(2). doi:10.1371/journal.pone.0030626

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, *23*(10), 1294–1296. doi:10.1093/bioinformatics/btm108

Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., & Neiman, M. (2017). Genomic evidence for population-specific responses to co-evolving parasites in a New Zealand freshwater snail. *Molecular Ecology*, *26*(14), 3663–3675. doi:10.1111/mec.14146

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2011). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Beldomenico, P. M., & Begon, M. (2010). Disease spread, susceptibility and infection intensity: vicious circles? *Trends in Ecology and Evolution*, *25*(1), 21–27. doi:10.1016/j.tree.2009.06.015

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken : RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*, 362–365. doi:10.1111/1755-0998.12669

Cayol, C., Giermek, A., Gomez-Chamorro, A., Hytönen, J., Kallio, E. R., Mappes, T., … Koskela, E. (2018). *Borrelia afzelii* alters reproductive success in a rodent host. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1884), 20181056. doi:10.1098/rspb.2018.1056

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2014). Second-generation PLINK: rising to the challenge of larger and richer datasets, 1–16. doi:10.1186/s13742-015-0047-8

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., … Collins, F. S. (2007). Replicating genotype–phenotype associations. *Nature*, *447*, 655-660. doi: 10.1038/447655a

Cornetti, L., Hilfiker, D., Lemoine, M., & Tschirren, B. (2018). Small-scale spatial variation in infection risk shapes the evolution of a *Borrelia* resistance gene in wild rodents. *Molecular Ecology*, *27*(17), 3515–3524. doi:10.1111/mec.14812

Cornetti, L. & Tschirren, B. (2020) Data from: Combining GWAS and F_{ST}-based approaches to identify targets of *Borrelia*-mediated selection in natural rodent hosts; Dryad; doi: 10.5061/dryad.c866t1g3t.

Cornetti, L. & Tschirren, B. (2020). NCBI Sequence Read Archive (BioProject ID: PRJNA306409, SRA experiment SRR3031372). https://www.ncbi.nlm.nih.gov/bioproject/PRJNA306409.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2000). Emerging infectious diseases of wildlife - threats to biodiversity and human health. *Science*, *287*, 443–449. doi:10.1126/science.287.5452.443

Davey, J. L., & Blaxter, M. W. (2010). RADseq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. doi:10.1093/bfgp/elq031

Dennis, V. A., Dixit, S., O'Brien, S. M., Alvarez, X., Pahar, B., & Philipp, M. T. (2009). Live *Borrelia burgdorferi* spirochetes elicit inflammatory mediators from human monocytes via the toll-like receptor signaling pathway. *Infection and Immunity*, *77*(3), 1238–1245. doi:10.1128/IAI.01078-08

Doran, A. G., & Creevey, C. J. (2013). Snpdat: easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC Bioinformatics*, *14*(1), 45. doi:10.1186/1471-2105-14-45

Earl, D. A., & VonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. doi:10.1007/s12686-011-9548-7

Ersts, P. J. (2017). Geographic Distance Matrix Generator (version 1.2.3). Retrieved July 18, 2017, from http://biodiversityinformatics.amnh.org/open_source/gdmg

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Flint, J., & Eskin, E. (2012). Genome-wide association studies in mice. *Nature Reviews. Genetics*, *13*(11), 807–817. doi:10.1038/nrg3335

Gautam, A., Dixit, S., Philipp, M. T., Singh, S. R., Morici, L. A., Kaushal, D., & Dennis, V. A. (2011). Interleukin-10 alters effector functions of multiple genes induced by *Borrelia burgdorferi* in macrophages to regulate lyme disease inflammation. *Infection and Immunity*, *79*(12), 4876–4892. doi:10.1128/IAI.05451-11

Gliwicz, J. (1988). Seasonal dispersal in non-cyclic populations of *Clethrionomys glareolus* and *Apodemus flavicollis*. *Acta Theriologica*, *33*, 263–272. doi:10.4098/AT.arch.88-20

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. doi:10.1093/molbev/msy228

Heine, W., Beckstette, M., Heroven, A. K., Thiemann, S., Heise, U., Nuss, A.

M., … Dersch, P. (2018). Loss of CNFYtoxin-induced inflammation drives *Yersinia pseudotuberculosis* into persistency. *PLoS Pathogens 14*(2), e1006858. doi:10.1371/journal.ppat.1006858

Hill, A. V. S. (2012). Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1590), 840–849. doi:10.1098/rstb.2011.0275

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *American Naturalist*, *188*(4), 379–397. doi:10.1086/688018

Hong, E. P., & Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, *10*(2), 117–122. doi:10.5808/GI.2012.10.2.117

Jensen, J. L., Bohonak, A. J., & Kelley, S. T. (2005). Isolation by distance, web service. *BMC Genetics*, *6*, 13. doi:10.1186/1471-2156-6-13

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993. doi:10.1038/nature06536

Jones, M. E., Jarman, P. J., Lees, C. M., Hesterman, H., Hamede, R. K., Mooney, N. J., … McCallum, H. (2007). Conservation management of Tasmanian devils in the context of an emerging, extinction-threatening disease: Devil facial tumor disease. *EcoHealth*, *4*(3), 326–337. doi:10.1007/s10393-007-0120-6

Kurtenbach, K., Hanincová, K., Tsao, J. I., Margos, G., Fish, D., & Ogden, N.

H. (2006). Fundamental processes in the evolutionary ecology of Lyme borreliosis. *Nature Reviews Microbiology*, *4*(9), 660–669. doi:10.1038/nrmicro1475

Kurtz, J., Behrens, S., Schulenburg, H., Bornberg-Bauer, E., Peuß, R., Milutinović, B., … Esser, D. (2014). Infection routes matter in population-specific responses of the red flour beetle to the entomopathogen *Bacillus thuringiensis*. *BMC Genomics*, *15*(1), 445. doi:10.1186/1471-2164-15-445

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9. doi:10.1038/nmeth.1923

Lazzaro, B. P., & Little, T. J. (2009). Immunity in a variable world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1513), 15–26. doi:10.1098/rstb.2008.0141

Lo, S. H. (2004). Tensin. *International Journal of Biochemistry and Cell Biology*, *36*(1), 31–34. doi:10.1016/S1357-2725(03)00171-7

MacPhee, R. D. E., & Greenwood, A. D. (2013). Infectious disease, endangerment, and extinction. *International Journal of Evolutionary Biology*, *2013*, 1–9. doi:10.1155/2013/571939

Mannelli, A., Bertolotti, L., Gern, L., & Gray, J. (2012). Ecology of *Borrelia burgdorferi* sensu lato in Europe: transmission dynamics in multi-host systems, influence of molecular processes and effects of climate change. *FEMS Microbiology Reviews*, *36*(4), 837–61. doi:10.1111/j.1574-6976.2011.00312.x

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209-220.

Margres, M. J., Jones, M., Epstein, B., Kerlin, D. H., Comte, S., Fox, S., …

Storfer, A. (2018). Large-effect loci affect survival in Tasmanian devils (*Sarcophilus harrisii*) infected with a transmissible cancer. *Molecular Ecology*, *27*(21), 4189–4199. doi:10.1111/mec.14853

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. a, & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics*, *9*(5), 356–369. doi:10.1038/nrg2344

McGraw, L. A., Davis, J. K., Young, L. J., & Thomas, J. W. (2011). A genetic linkage map and comparative mapping of the prairie vole (*Microtus ochrogaster*) genome. *BMC Genetics*, *12*(1), 60. doi:10.1186/1471-2156-12-60

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi:10.1101/gr.107524.110

Myles, S., Davison, D., Barrett, J., Stoneking, M., & Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics*, *1*(1), 22. doi:10.1186/1755-8794-1-22

Petersen, B. S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, *18*(1), 1–13. doi:10.1186/s12863-017-0479-5

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, *7*(2). doi:10.1371/journal.pone.0032253

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. doi:10.1038/ng1847

Price, A. L., Spencer, C. C. A., & Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1821), 20151684. doi:10.1098/rspb.2015.1684

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. doi:10.1111/j.1471-8286.2007.01758.x

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Råberg, L. (2012). Infection intensity and infectivity of the tick-borne pathogen *Borrelia afzelii*. *Journal of Evolutionary Biology*, *25*(7), 1448–53. doi:10.1111/j.1420-9101.2012.02515.x

Royaux, I. E., Suzuki, K., Mori, A., Katoh, R., Everett, L. A., Kohn, L. D., & Green, E. D. (2000). Pendrin, the protein encoded by the pendred syndrome gene (PDS), is an apical porter of iodide in the thyroid and is regulated by thyroglobulin in FRTL-5 cells. *Endocrinology*, *141*(2), 839–845. doi:10.1210/endo.141.2.7303

Santure, A. W., & Garant, D. (2018). Wild GWAS - association mapping in natural populations. *Molecular Ecology Resources*, *18*(4), 729–738. doi:10.1111/1755-0998.12901

Scanlon, K. M., Gau, Y., Zhu, J., Skerry, C., Wall, S. M., Soleimani, M., &

Carbonetti, N. H. (2014). Epithelial anion transporter Pendrin contributes to inflammatory lung pathology in mouse models of *Bordetella pertussis* infection. *Infection and Immunity*, *82*(10), 4212–4221. doi:10.1128/iai.02222-14

Schielzeth, H., Rios, A., & Burri, R. (2018). Success and failure in replication of genotype-phenotype associations: How does replication help in understanding the genetic basis of phenotypic variation in outbred populations? *Molecular Ecology Resources*, *18*(4), 739–754. doi:10.1111/1755-0998.12780

Singh, S. K., & Girschick, H. J. (2006). Toll-like receptors in *Borrelia burgdorferi*-induced inflammation. *Clinical Microbiology and Infection*, *12*(8), 705–717. doi:10.1111/j.1469-0691.2006.01440.x

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., & Bult, C. J. (2018). Mouse Genome Database (MGD)-2018: Knowledgebase for the laboratory mouse. *Nucleic Acids Research*, *46*(D1), D836–D842. doi:10.1093/nar/gkx1006

Smith, K. F., Sax, D. F., & Lafferty, K. D. (2006). Evidence for the role of infectious disease in species extinction and endangerment. *Conservation Biology*, *20*(5), 1349–1357. doi:10.1111/j.1523-1739.2006.00524.x

Spurgin, L. G., & Richardson, D. S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1684), 979–88. doi:10.1098/rspb.2009.2084

Staubach, F., Lorenc, A., Messer, P. W., Tang, K., Petrov, D. A., & Tautz, D. (2012). Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics*, *8*(8), e1002891. doi:10.1371/journal.pgen.1002891

Steere, A. C., Coburn, J., & Glickstein, L. (2004). The emergence of Lyme disease. *Journal of Clinical Investigation*, *113*(8), 1093–1101. doi:10.1172/JCI200421681

Tagu, D., Colbourne, J. K., & Nègre, N. (2014). Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC Genomics*, *15*(1), 490. doi:10.1186/1471-2164-15-490

Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., … Grimm, D. G. (2018). The AraGWAS Catalog: A curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Research*, *46*(D1), D1150–D1156. doi:10.1093/nar/gkx954

Tschirren, B., Andersson, M., Scherman, K., Westerdahl, H., Mittl, P. R., & Råberg, L. (2013). Polymorphisms at the innate immune receptor *TLR2* are associated with *Borrelia* infection in a wild rodent population. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20130364. doi:10.1098/rspb.2013.0364

Tschirren, B. (2015). *Borrelia burgdorferi* sensu lato infection pressure shapes innate immune gene evolution in natural rodent populations across Europe. *Biology Letters*, *11*, 20150263. doi:10.1098/rsbl.2015.0263

Tschirren, B. & Bensch, S. (2010). Genetics of personalities: no simple answers for complex traits. *Molecular Ecology*, *19*(4), 624–626. doi: 10.1111/j.1365-294X.2009.04519.x

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. doi:10.1002/0471250953.bi1110s43

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97–120. doi:10.1146/annurev-genet-111212-133526

Wangler, M. F., Hu, Y., & Shulman, J. M. (2017). *Drosophila* and genome-wide association studies: a review and resource for the functional dissection of human complex traits. *Disease Models & Mechanisms*, *10*(2), 77–88. doi:10.1242/dmm.027680

Wiethoelter, A. K., Beltrán-Alcrudo, D., Kock, R., & Mor, S. M. (2015). Global trends in infectious diseases at the wildlife–livestock interface. *Proceedings of the National Academy of Sciences USA*, *112*(31), 9662–9667. doi:10.1073/pnas.1422741112

Wilfert, L., & Schmid-Hempel, P. (2008). The genetic architecture of susceptibility to parasites. *BMC Evolutionary Biology*, *8*(1), 1–8. doi:10.1186/1471-2148-8-187

Wooten, R. M., Ma, Y., Yoder, R. A., Brown, J. P., Weis, J. H., Zachary, J. F., … Weis, J. J. (2002). Toll-like receptor 2 is required for innate, but not acquired, host defense to *Borrelia burgdorferi*. *Journal of Immunology,* *168*(1), 348–355.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. doi:10.1093/nar/gkx1098

Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., & Primmer, C. R. (2014). Footprints of directional selection in wild atlantic salmon populations: Evidence for parasite-driven evolution? *PLoS ONE*, *9*(3). doi:10.1371/journal.pone.0091672

Tables

Table 1. Sampling locations and number of analysed bank voles

Elevation and study site coordinates, the number of genotyped adult bank voles (N), the number of genotyped *Borrelia*-free bank voles (N uninf) and the number of genotyped *Borrelia*-infected bank voles (N inf) and *Borrelia* prevalence in adult bank voles at the study sites are reported.

Location	Label	Elevation (masl)	North	East	N	N uninf	N inf	*Borrelia* prevalence (%)
Bonaduz	BON	944	46.799	9.352	16	8	8	50.0
Rodels	ROD	630	46.760	9.425	17	13	4	31.2
Sagogn	SAG	693	46.783	9.233	19	10	9	48.4
Flims	FLI	1138	46.827	9.280	15	9	6	54.5
Malans	MAL	560	46.992	9.557	19	10	9	44.8
Passugg	PAS	732	46.840	9.538	13	6	7	61.5

Trimmis	TRI	762	46.882	9.559	19	9	10	44.8

Table 2. F_{ST} outlier SNPs

Exonic SNPs that were identified as outliers in multiple populations when comparing *Borrelia*-infected and *Borrelia*-free bank voles using a F_{ST}-based approach. SNPs in bold were also identified to be associated with *Borrelia*-infection status using a GWAS approach. The SNP position refers to the prairie vole genome version MicOch1.0.

Chromosome Number	SNP Position	Start of exon	End of exon	Protein ID	Gene description	Number of populations in which the SNP was an outlier
1	**82812170**	**82812122**	**82812285**	**ENSMUSP00000001253**	**Solute carrier family 26, member 4**	3
5	90628057	90626456	90630654	ENSMUSP00000081880	Golgin subfamily A member 4	3
7	81147668	81147563	81147724	ENSMUSP00000081864	Dynein, axonemal, heavy chain 17	3

7	**27136917**	**27136736**	**27136920**	**ENSMUSP00000104150**	**WSC domain containing 1**	4
7	25811226	25810812	25812290	ENSMUSP00000055806	Haspin	3
8	72655777	72655627	72655797	ENSMUSP00000059419	Sideroflexin 3	3
LG1	**8299897**	**8299839**	**8299951**	**ENSMUSP00000020695**	**Tensin 3**	4
LG4	**60445822**	**60445116**	**60446729**	**ENSMUSP00000086294**	**Espin-like**	3
LG5	38783881	38783857	38783985	ENSMUSP00000030187	Talin-1	3

Figures

Figure 1. Multi-dimensional scaling of bank vole genetic diversity.

Different colours represent different sampling sites. The inset shows the proportion of ancestry for each sampled bank vole (N = 118) for seven genetic clusters inferred with STRUCTURE (see Supplementary Figures S2 and S3 for additional information).

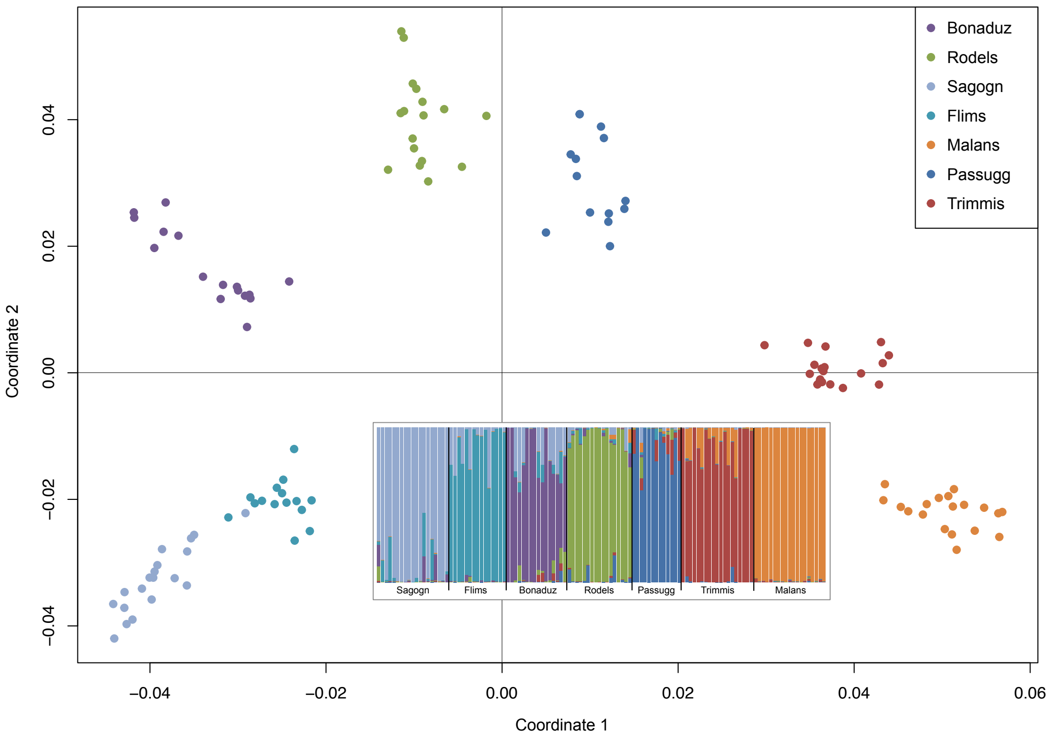


Figure 2. Simulations of false-positive rates.

We used a simulation approach to quantify the false positive rate of the GWAS approach (a), the F_{ST}-based approach with a three population replication criterion (b), and (c) the combined approach (i.e. consensus SNPs identified in (a) and (b)). The red asterisk indicates the number of identified SNPs observed in the real data using the respective approach.

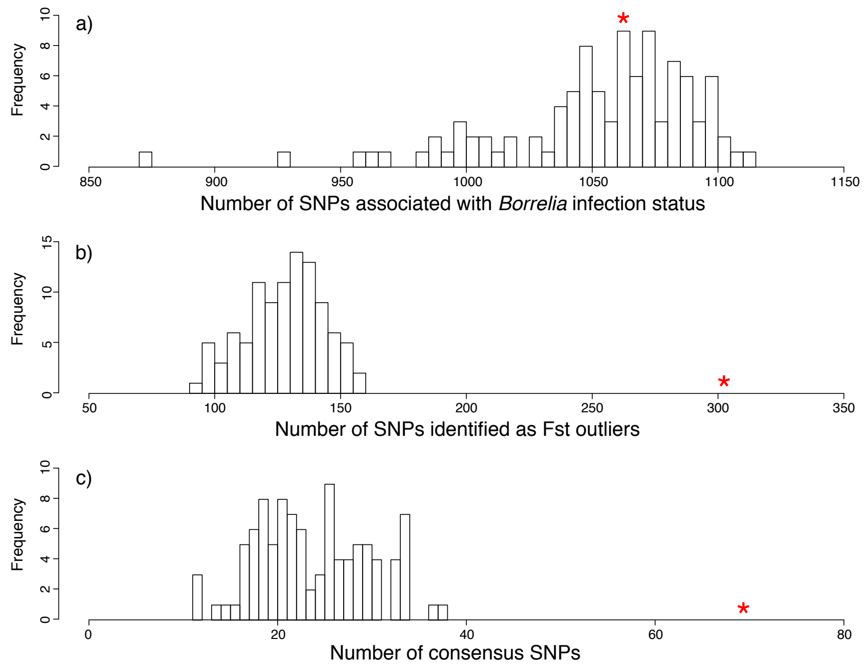
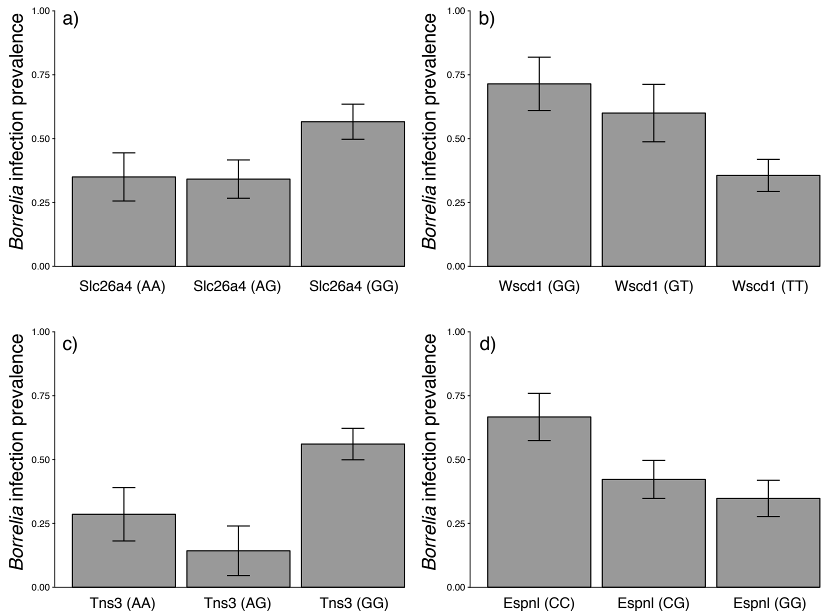


Figure 3. Genetic polymorphisms at the four consensus candidate loci are associated with *Borrelia* infection status in bank voles.

Animals that carried the rarer allele of *Slc26a4* (a, allele A) and *Tns3* (c, allele A) were less likely to be *Borrelia*-infected, whereas animals that carried the rarer allele of *Wscd1* (b, allele G) and *Espnl* (d, allele C) were more likely to be *Borrelia*-infected. Error bars represent standard errors.



Supporting information

Combining GWAS and F_{ST}-based approaches to identify targets of *Borrelia*-mediated selection in natural rodent hosts

L. Cornetti & B. Tschirren

1. Supporting Methods

Neutral simulations

Script used for the simulations

```
initialize() {
  initializeMutationRate(2e-4);  ## mutation rate
  initializeMutationType("m1", 0.5, "f", 0.0);  ## mutation type description: non-coding or synonymous
  initializeGenomicElementType("g1", c(m1), c(100));  ## mutation occurrence
  initializeGenomicElement(g1, 0, 27999);  ## size of the simulated chromosome
  initializeRecombinationRate(0.05); ## recombination rate
}
1 { ## at generation 1 seven subpopulations appear
  sim.addSubpop("p1", 100);  ## population size of p1
  sim.addSubpop("p2", 100);  ## population size of p2
  sim.addSubpop("p3", 100);  ## population size of p3
  sim.addSubpop("p4", 100);  ## population size of p4
  sim.addSubpop("p5", 100);  ## population size of p5
  sim.addSubpop("p6", 100);  ## population size of p6
  sim.addSubpop("p7", 100);  ## population size of p7
  p1.setMigrationRates(c(p2,p3,p4,p5,p6,p7), c(0.05,0.05,0.05,0.01,0.03,0.01));  ## migration rates into population p1 from the others
  p2.setMigrationRates(c(p1,p3,p4,p5,p6,p7), c(0.05,0.03,0.03,0.01,0.05,0.03));  ## migration rates into population p2 from the others
  p3.setMigrationRates(c(p1,p2,p4,p5,p6,p7), c(0.05,0.03,0.05,0.01,0.01,0.01));  ## migration rates into population p3 from the others
  p4.setMigrationRates(c(p1,p2,p3,p5,p6,p7), c(0.05,0.03,0.05,0.01,0.03,0.03));  ## migration rates into population p4 from the others
  p5.setMigrationRates(c(p1,p2,p3,p4,p6,p7), c(0.01,0.01,0.01,0.01,0.03,0.03));  ## migration rates into population p5 from the others
```

```
  p6.setMigrationRates(c(p1,p2,p3,p4,p5,p7),
c(0.03,0.05,0.01,0.03,0.01,0.05));  ## migration rates into population p6 from
the others
  p7.setMigrationRates(c(p1,p2,p3,p4,p5,p6),
c(0.01,0.03,0.01,0.03,0.05,0.03));  ## migration rates into population p7 from
the others
}
4999 late() {  ## number of generation simulated
  bonaduz = p1.sampleIndividuals(16).genomes;  ## number of samples
selected from p1 according to the sample size of Bonaduz
  bonaduz.outputVCF(filePath="/home/p1.vcf");  ## the SNPs are written in a
VCF file
  rodels = p2.sampleIndividuals(17).genomes;  ## number of samples
selected from p2 according to the sample size of Rodels
  rodels.outputVCF(filePath="/home/p2.vcf");  ## the SNPs are written in a
VCF file
  sagogn = p3.sampleIndividuals(19).genomes;  ## number of samples
selected from p3 according to the sample size of Sagogn
  sagogn.outputVCF(filePath="/home/p3.vcf");  ## the SNPs are written in a
VCF file
  flims = p4.sampleIndividuals(15).genomes;  ## number of samples selected
from p4 according to the sample size of Flims
  flims.outputVCF(filePath="/home/p4.vcf");  ## the SNPs are written in a VCF
file
  malans = p5.sampleIndividuals(19).genomes;  ## number of samples
selected from p5 according to the sample size of Malans
  malans.outputVCF(filePath="/home/p5.vcf");  ## the SNPs are written in a
VCF file
  passugg = p6.sampleIndividuals(13).genomes;  ## number of samples
selected from p6 according to the sample size of Passugg
  passugg.outputVCF(filePath="/home/p6.vcf");  ## the SNPs are written in a
VCF file
  trimmis = p7.sampleIndividuals(19).genomes;  ## number of samples
selected from p7 according to the sample size of Trimmis
  trimmis.outputVCF(filePath="/home/p7.vcf");  ## the SNPs are written in a
VCF file
}
```

Supporting Figures

Figure S1. Map of the sampling sites in the Kanton Graubünden, Switzerland.

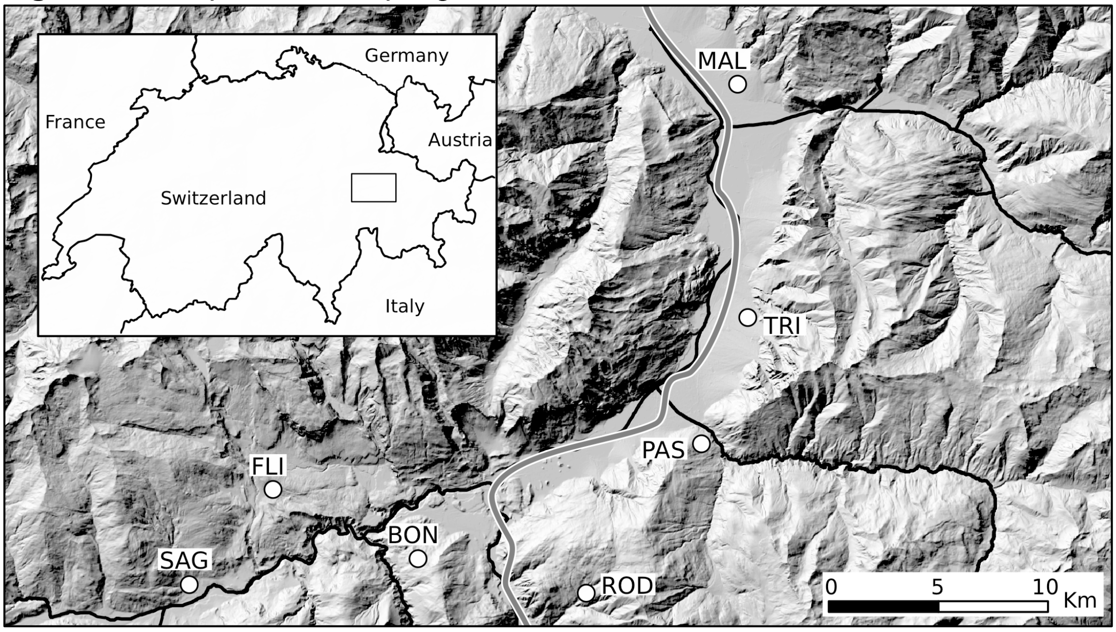


Figure S2. Estimate of the number of genetically well defined groups (*K*) based on mean likelihood (Pritchard, Stephens, & Donnelly, 2000).

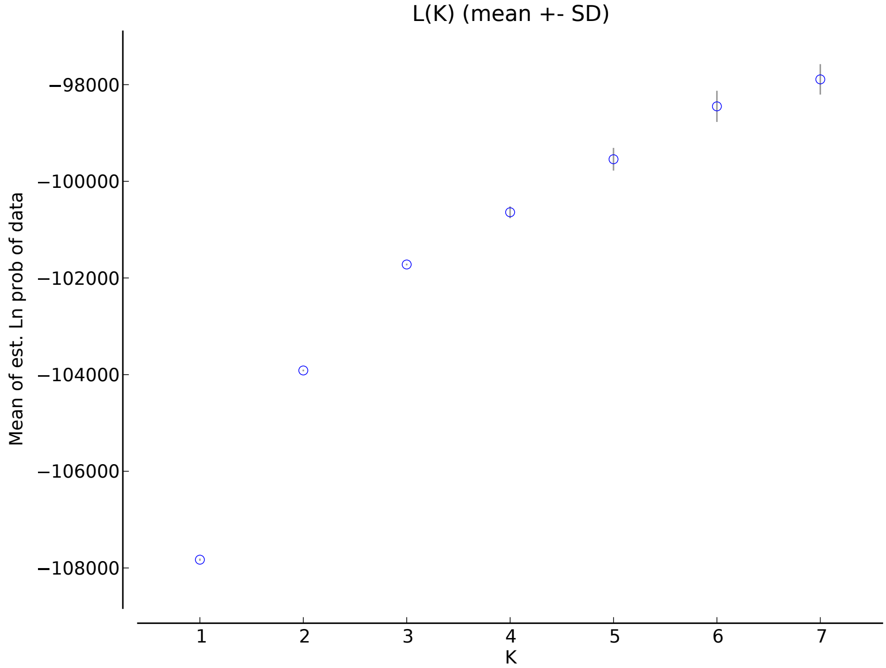


Figure S3. STRUCTURE plot describing the bank vole population structure in the study area using K=2 to K=7 as most probable number of genetic groups.

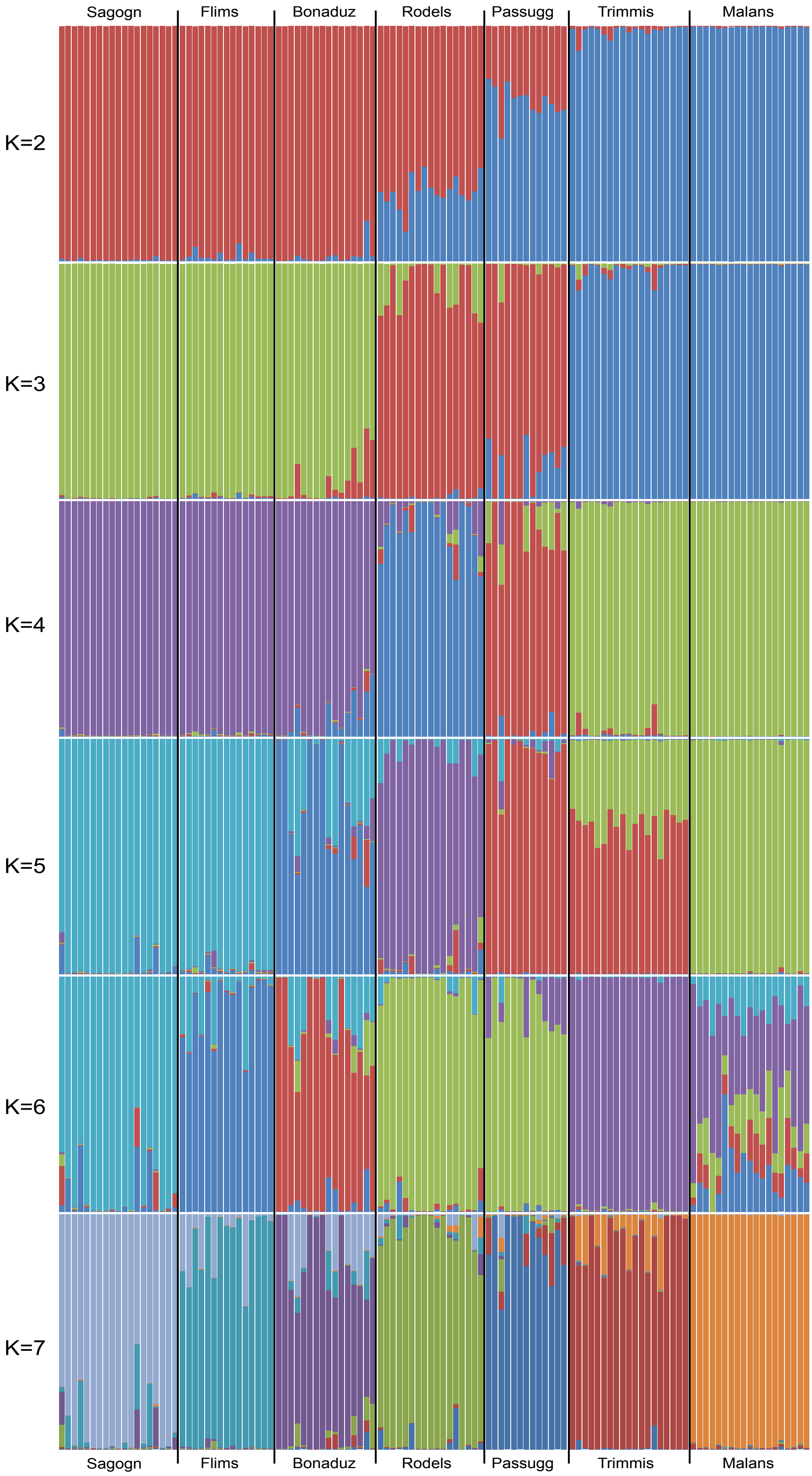


Figure S4. Isolation-by-distance of bank vole populations across our study sites.

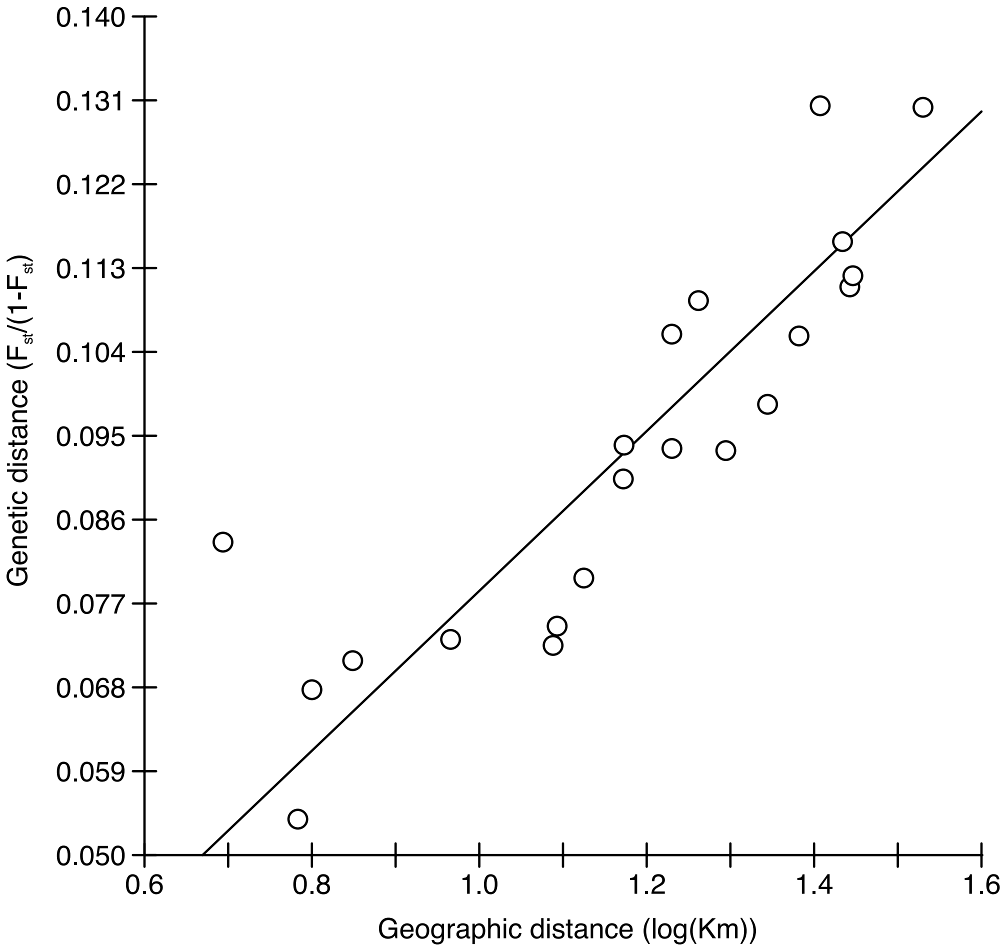


Figure S5 False positive rates of F_{ST} -based approach.
We used a simulation approach to quantify the false positive rate of the F_{ST}-based approach when using no population replication (a), a two population replication criterion (b), and a three population replication criterion (as used in the main study) (c). The red asterisk indicates the number of outliers observed in the real data using the respective approach.

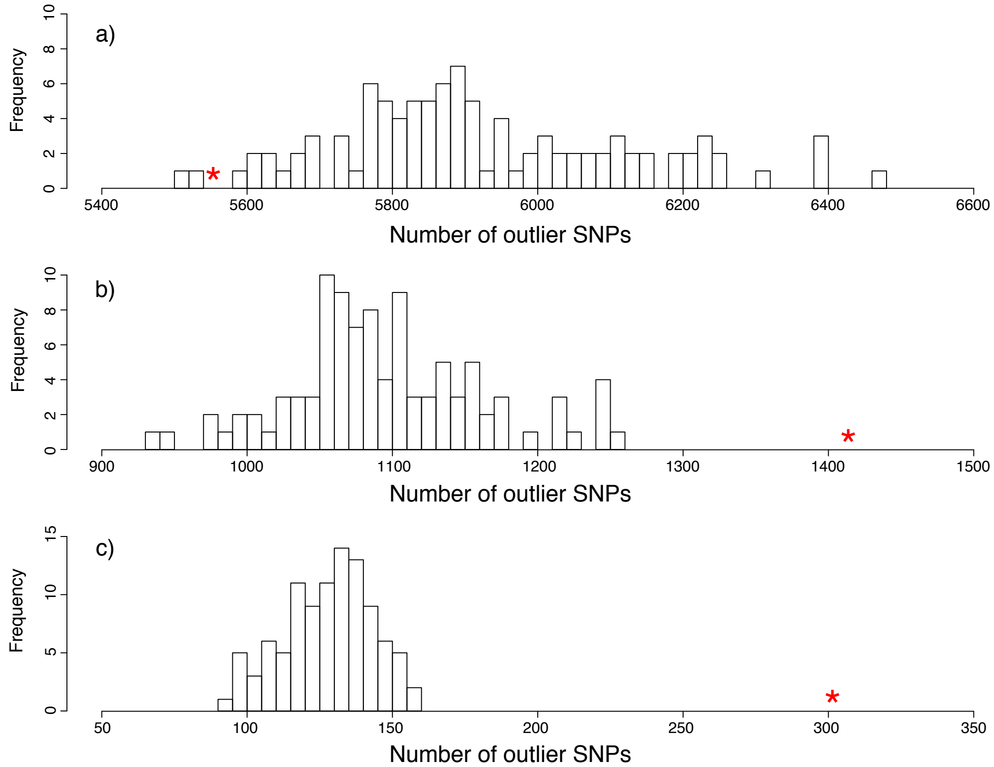
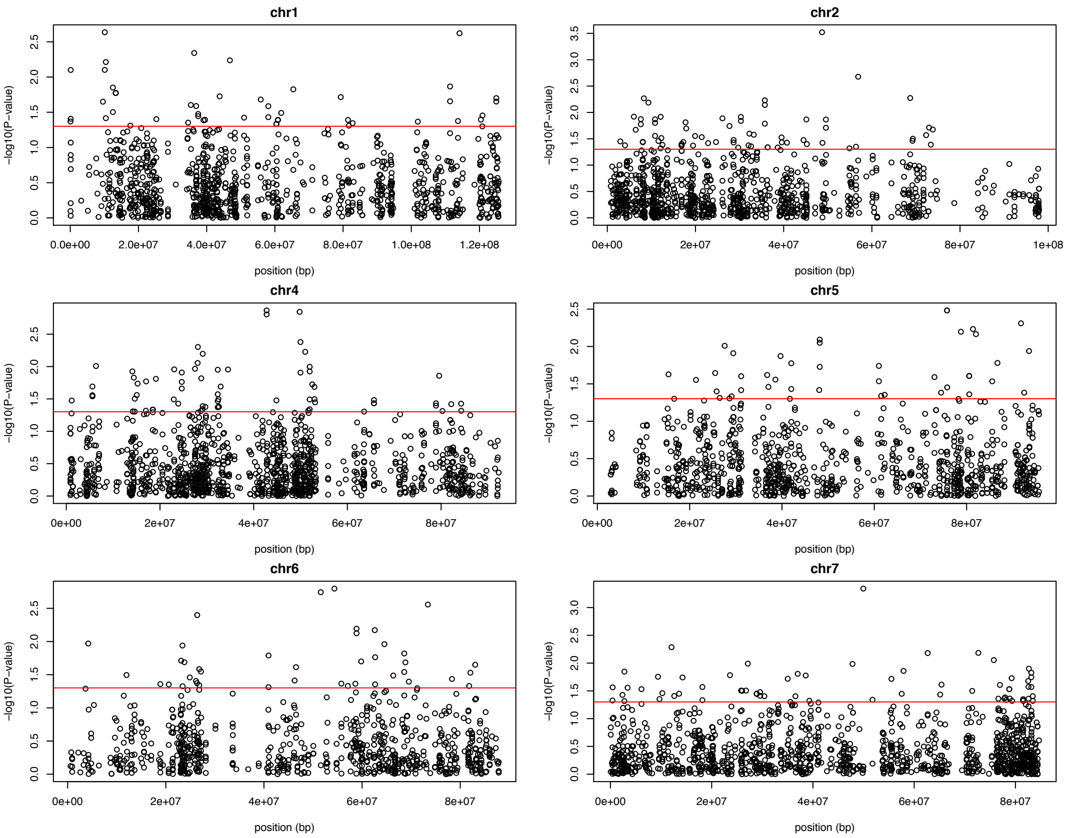
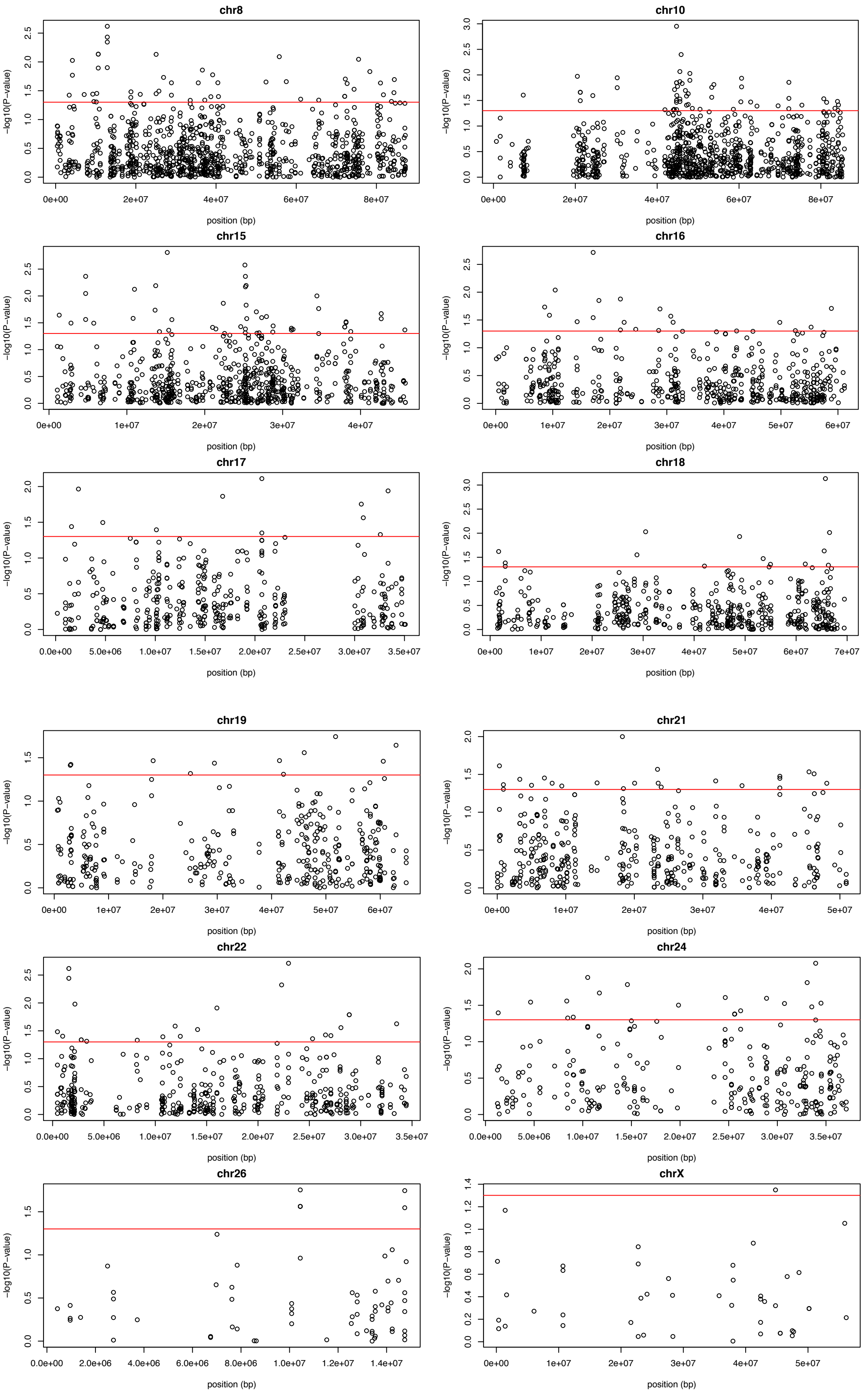


Figure S6. Association scan for *Borrelia* infection status in bank voles using a GWAS approach. Associations between SNPs and bank vole *Borrelia* infection status after correcting for population structure (–log10(P-value)) are plotted for each chromosome and linkage group. The red line represents the *P* = 0.05 threshold. Dots above this line represent SNPs possibly associated with *Borrelia* infection status.




chr8
chr10
chr15
chr16
chr17
chr18
chr19
chr21
chr22
chr24
chr26
chrX
−log10(P−value)
position (bp)

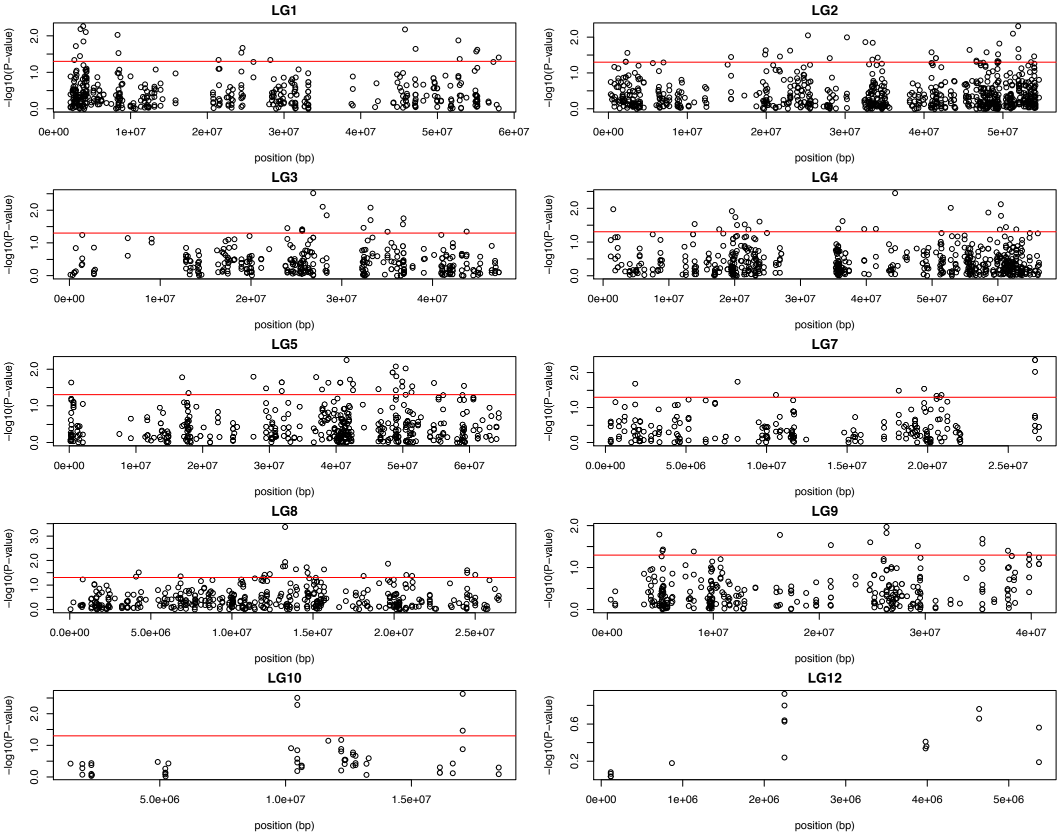

LG1
LG2
LG3
LG4
LG5
LG7
LG8
LG9
LG10
LG12
−log10(P−value)
position (bp)

Figure S7. Number of putatively *Borrelia* infection status associated SNPs per MB across chromosomes and linkage groups. Candidate SNPs were identified using a GWAS approach (see Figure S6). The main 28 scaffolds of the prairie vole (*Microtus ochrogaster*) reference genome, version MicOch1.0, are shown.

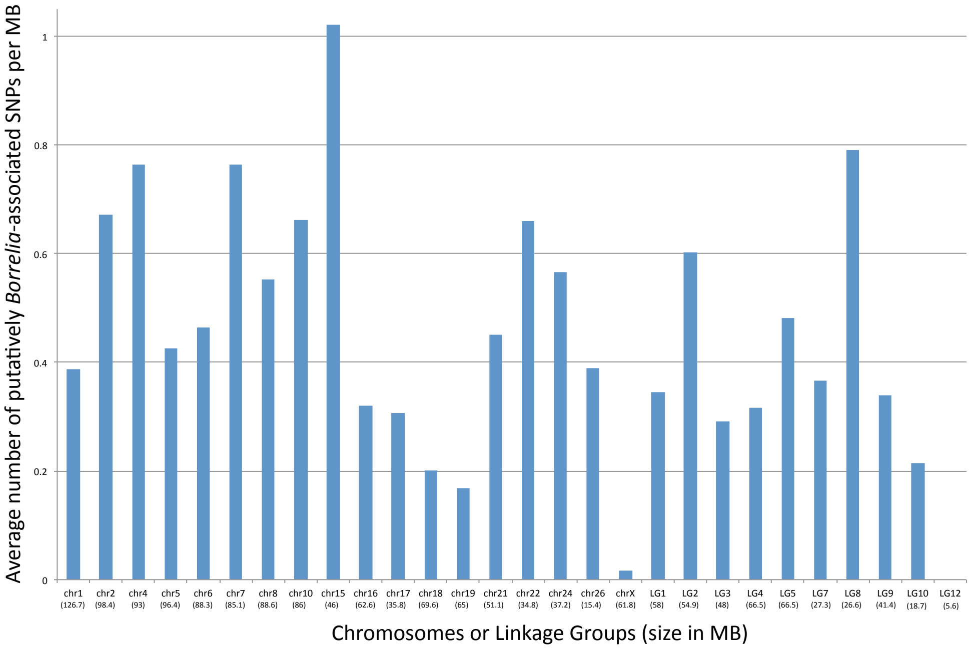
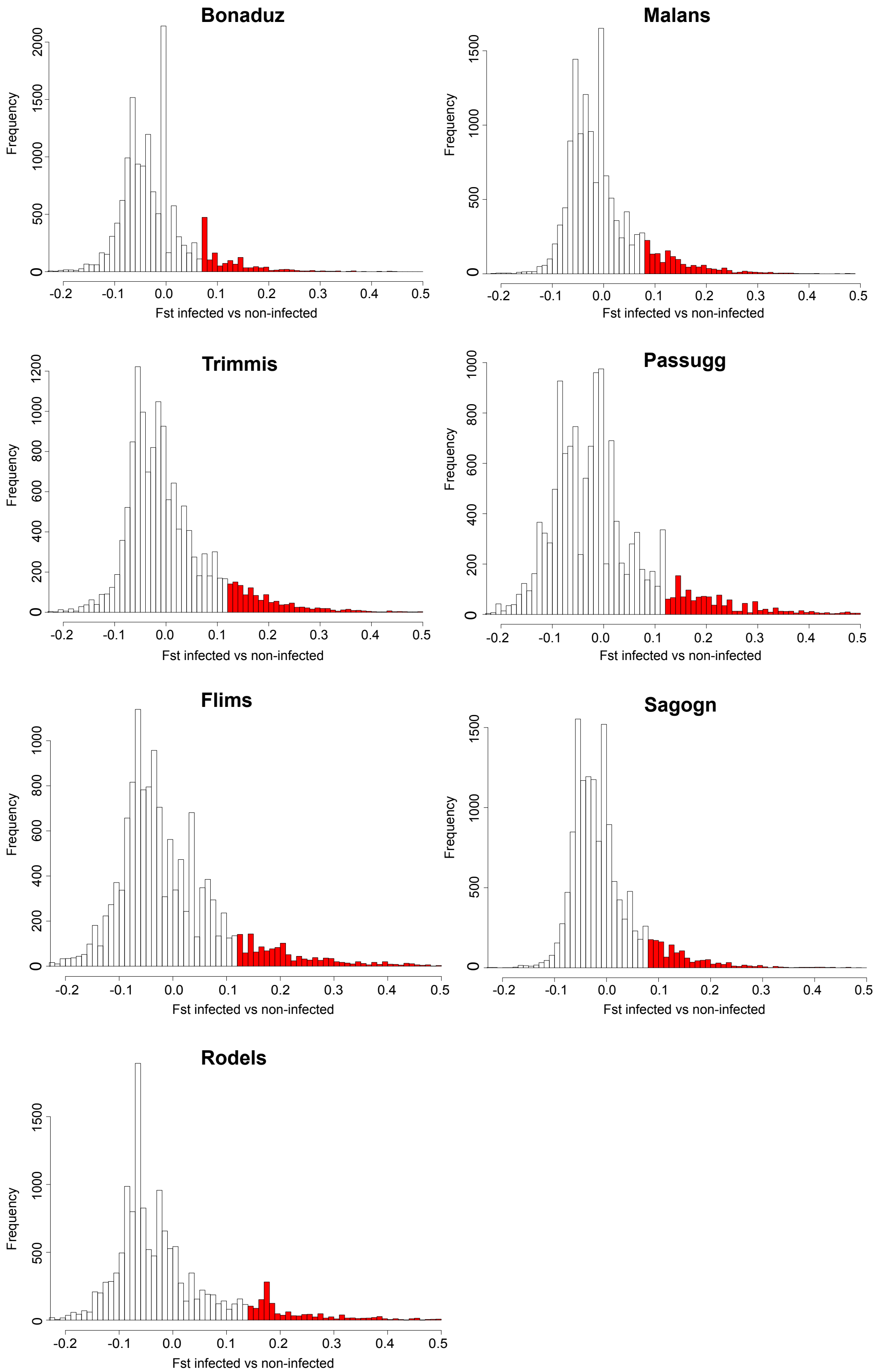


Figure S8. F_{ST} distribution of SNPs across the seven study populations. The red colour represents the SNPs falling into the 10% highest F_{ST} values when comparing *Borrelia*-infected vs uninfected bank voles within each population.



Supporting Tables

Table S1. Pairwise F_{ST} based on 21,811 genome-wide SNPs (above the diagonal) and linear distance (km) between locations (below the diagonal). All F_{ST} values are statistically significant ($p < 0.05$).

	Bonaduz	Rodels	Sagogn	Flims	Malans	Passugg	Trimmis
Bonaduz	-	0.063	0.068	0.062	0.110	0.083	0.092
Rodels	7.06	-	0.084	0.073	0.101	0.069	0.083
Sagogn	9.24	14.86	-	0.052	0.111	0.095	0.099
Flims	6.31	13.33	6.07	-	0.094	0.086	0.087
Malans	25.55	27.71	33.90	27.95	-	0.096	0.066
Passugg	14.89	12.39	24.09	19.70	16.98	-	0.076
Trimmis	18.27	16.99	27.16	22.10	12.25	4.94	-

Table S2. List of 53 exonic SNPs putatively associated with *Borrelia* infection status in bank voles, identified using a GWAS approach. Genes with more than one SNP per GBS read are underlined, and genes for which SNPs were observed in two different exons and two different GBS reads are highlighted with an asterisk (*). Candidate SNPs identified with both the GWAS and the F_{ST} -based approach with population replication are highlighted in bold. The SNP position refers to the prairie vole genome version MicOch1.0.

Chromosome Number	SNP Position	Start of exon	End of exon	Protein ID	Gene description
LG2	46643179	46643140	46643278	ENSMUSP00000001183	formiminotransferase cyclodeaminase
1	**82812170**	**82812122**	**82812285**	**ENSMUSP00000001253**	**solute carrier family 26, member 4**
21	900079	899816	900277	ENSMUSP00000005015	papillary renal cell carcinoma (translocation-associated)
21	900160	899816	900277	ENSMUSP00000005015	papillary renal cell carcinoma (translocation-associated)
10	43932666	43932619	43932690	ENSMUSP00000010007	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
15	25218452	25218178	25218816	ENSMUSP00000016172	cadherin, EGF LAG seven-pass G-type receptor 1
15	25218467	25218178	25218816	ENSMUSP00000016172	cadherin, EGF LAG seven-pass G-type receptor 1
15	25218528	25218178	25218816	ENSMUSP00000016172	cadherin, EGF LAG seven-pass G-type receptor 1
LG8	4254355	4253368	4254449	ENSMUSP00000016399	tubulin, beta 1 class VI
LG1	**8299897**	**8299839**	**8299951**	**ENSMUSP00000020695**	**tensin 3**
7	29565141	29565139	29565229	ENSMUSP00000021259	guanylate cyclase 2e
15	2823221	2823152	2823685	ENSMUSP00000023720	keratin 84
24	8453004	8453003	8453147	ENSMUSP00000026500	advillin
22	11931974	11931956	11932112	ENSMUSP00000026922	homer scaffolding protein 2
LG5	54679380	54679319	54680259	ENSMUSP00000027035	SRY (sex determining region Y)-box 17
LG2	23030453	23030389	23030539	ENSMUSP00000027257	MIT, microtubule interacting and transport, domain containing 1
LG4	60476015	60476012	60476048	ENSMUSP00000027532	selenocysteine lyase
6	23429687	23428593	23430689	ENSMUSP00000027706	leucine rich repeat protein 2, neuronal
5	74480869	74480842	74481033	ENSMUSP00000034961	immunoglobulin superfamily, DCC subclass, member 3

5	81380322	81380142	81380335	ENSMUSP00000035232	kelch domain containing 8B
LG4	19571145	19570722	19571251	ENSMUSP00000036699	immunity-related GTPase family, Q
4	17036091	17035966	17036136	ENSMUSP00000039271	fucokinase
7	83424112*	83423459	83425210	ENSMUSP00000043643	BAH domain and coiled-coil containing 1
7	83439688*	83439647	83439722	ENSMUSP00000043643	BAH domain and coiled-coil containing 1
2	2946490	2946456	2946531	ENSMUSP00000046544	small G protein signaling modulator 1
24	33941679	33941663	33941821	ENSMUSP00000048309	stabilin 2
21	14567230	14566282	14567610	ENSMUSP00000052306	phospholipid phosphatase related 4
7	25811175	25810812	25812290	ENSMUSP00000055806	germ cell associated 2, haspin
22	25313647	25313056	25313865	ENSMUSP00000073855	synemin, intermediate filament protein
2	6446782	6446705	6447009	ENSMUSP00000075488	serine/arginine repetitive matrix 4
LG4	13915780	13915721	13915902	ENSMUSP00000076093	utrophin
<u>2</u>	<u>16862787</u>	<u>16862779</u>	<u>16862865</u>	<u>ENSMUSP00000078198</u>	<u>rabphilin 3A</u>
<u>2</u>	<u>16862810</u>	<u>16862779</u>	<u>16862865</u>	<u>ENSMUSP00000078198</u>	<u>rabphilin 3A</u>
15	26699638	26699630	26699696	ENSMUSP00000079575	RIKEN cDNA 1810041L15 gene
4	79583497	79583457	79583535	ENSMUSP00000079752	low density lipoprotein receptor-related protein 2
7	77798465	77798372	77798499	ENSMUSP00000081398	kinesin family member 19A
LG4	**60445822**	**60445116**	**60446729**	**ENSMUSP00000086294**	**espin-like**
15	28737242	28737225	28737356	ENSMUSP00000086582	meiotic double-stranded break formation protein 1
7	78870838	78870772	78870933	ENSMUSP00000091439	myosin XVB
6	23134593	23134517	23134713	ENSMUSP00000092148	neurofascin
LG9	16293329	16293222	16293422	ENSMUSP00000093587	spectrin repeat containing, nuclear envelope 1
LG8	25004183	25003387	25004577	ENSMUSP00000096813	zinc finger, CCHC domain containing 3
8	74431200	74431102	74432722	ENSMUSP00000096972	cyclin M2
2	4016724	4016690	4016844	ENSMUSP00000099642	acetyl-Coenzyme A carboxylase beta
10	75362360	75362298	75362460	ENSMUSP00000102320	podocan
22	2174486	2174390	2174566	ENSMUSP00000102745	myosin VIIA

7	**27136917**	**27136736**	**27136920**	**ENSMUSP00000104150**	**WSC domain containing 1**
7	30401083	30400232	30401249	ENSMUSP00000104314	netrin 1
5	15411871	15411823	15411894	ENSMUSP00000111093	adaptor protein complex AP-1, mu 2 subunit
6	55777259	55777095	55777561	ENSMUSP00000113964	leucine-rich repeat, immunoglobulin-like and transmembrane domains 1
8	38709839	38709811	38709966	ENSMUSP00000115356	aldehyde dehydrogenase 3 family, member B2
6	40936018	40935948	40936067	ENSMUSP00000126464	acyl-Coenzyme A oxidase 2, branched chain
15	10977462	10977321	10977467	ENSMUSP00000129100	5-oxoprolinase (ATP-hydrolysing)

Table S3. Allele frequencies at the four consensus candidate genes identified with both the GWAS and the F_{ST}-based approach. Frequencies of the reference allele are shown.

			Gene			
	N	*Borrelia*-infection status	Slc26a4	Tns3	WSC1	Espin-like
Chromosome			chr1	chr7	LG1	LG4
Position			82812170	27136917	8299897	60445822
Alleles			A / G	G / A	G / T	G / C
Reference allele			A	G	G	G
	118	all	35.5	83.9	19.8	58.0
	53	infected	27.5	92.7	28.9	48.1
	65	uninfected	42	76.1	12.5	66.1

Table S4. Analysis of consensus candidate loci

We used generalized linear mixed models with a binomial error structure and location included as a random effect to test for associations between consensus SNPs and *Borrelia* infection status of bank voles. Models that combined the heterozygous and homozygous state of the less frequent allele are presented here. Models that treated the heterozygous and homozygous state of the less frequent allele as separate genotypes are presented in the main text (see Figure 3 in main text for a visualization of the effects).

SNP	*Test statistics*
Slc26a4	$\chi^2 = 5.679$, $DF = 1$, $P = 0.017$
Tns3	$\chi^2 = 9.346$, $DF = 1$, $P = 0.002$
Wscd1	$\chi^2 = 5.639$, $DF = 1$, $P = 0.018$
Espin-like	$\chi^2 = 3.162$, $DF = 1$, $P = 0.075$

References

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. doi:10.1111/j.1471-8286.2007.01758.x