Balancing selection in Pattern Recognition Receptor signalling pathways is associated with gene function and pleiotropy in a wild rodent

Max Lundberg | Xiuqin Zhong | Anna Konrad | Remi-André Olsen | Lars Råberg

Abstract
Pathogen-mediated balancing selection is commonly considered to play an important role in the maintenance of genetic diversity, in particular in immune genes. However, the factors that may influence which immune genes are the targets of such selection are largely unknown. To address this, here we focus on Pattern Recognition Receptor (PRR) signalling pathways, which play a key role in innate immunity. We used whole-genome resequencing data from a population of bank voles (Myodes glareolus) to test for associations between balancing selection, pleiotropy and gene function in a set of 123 PRR signalling pathway genes. To investigate the effect of gene function, we compared genes encoding (a) receptors for microbial ligands versus downstream signalling proteins, and (b) receptors recognizing components of microbial cell walls, flagella and capsids versus receptors recognizing features of microbial nucleic acids. Analyses based on the nucleotide diversity of full coding sequences showed that balancing selection primarily targeted receptor genes with a low degree of pleiotropy. Moreover, genes encoding receptors recognizing components of microbial cell walls etc. were more important targets of balancing selection than receptors recognizing nucleic acids. Tests for localized signatures of balancing selection in coding and noncoding sequences showed that such signatures were mostly located in introns, and more evenly distributed among different functional categories of PRR pathway genes. The finding that signatures of balancing selection in full coding sequences primarily occur in receptor genes, in particular those encoding receptors for components of microbial cell walls etc., is consistent with the idea that coevolution between hosts and pathogens is an important cause of balancing selection on immune genes.

Keywords
balancing selection, gene conversion, host-pathogen coevolution, innate immune genes, protein–protein interactions
INTRODUCTION

Several of the classical examples of balancing selection (i.e., selection that maintains genetic variation; Fijarczyk & Babik, 2015) concern genes directly involved in immune defence, for example MHC genes in vertebrates and R genes in plants (Bergelson, Kreitman, Stahl, & Tian, 2001; Hedrick, 2002). In addition, genome-wide scans and analyses of candidate genes—mainly in human populations but also in other organisms—have identified numerous other immune defence genes with signatures of balancing selection (Andrés et al., 2009; Bakker, Toomajian, Kreitman, & Bergelson, 2006; Bitarrello et al., 2018; DeGiorgio, Lohmueller, & Nielsen, 2014; Fijarczyk, Dudek, & Babik, 2016; Quintana-Murci & Clark, 2013; Tennessen et al., 2015; Unkless et al., 2016). In humans, for example, there is evidence of balancing selection on a number of immune genes other than MHC genes, including NOD2 which encodes an innate immune receptor (Nakagome et al., 2012), DEF81 which encodes an antimicrobial peptide (Cagliani et al., 2008) and TRIM5 which encodes a retroviral restriction factor (Cagliani et al., 2010). Taken together, this clearly supports the view, originally proposed by Haldane (1949), that host-pathogen interactions are important for the maintenance of genetic diversity (Lazzaro & Little, 2009; Woolhouse, Webster, Domingo, Charlesworth, & Levin, 2002). However, while there are numerous examples of immune genes with signatures of balancing selection, little is as yet known about the factors (if any) influencing which immune genes are the targets of such selection. In other words, do immune genes under balancing selection share certain properties?

One factor that could conceivably affect the likelihood of pathogen-mediated balancing selection on an immune gene is its function. Here we consider two broad functional categories of immune genes: those encoding receptors directly involved in the recognition of pathogens, and those encoding signal transduction proteins downstream of receptors. Host-pathogen interactions can result in balancing selection through several different processes (Hedrick, 2002; Karasov et al., 2014). One particularly influential idea is that balancing selection is a result of coevolution between a host–pathogen pair driven by negative frequency-dependent selection, often referred to as Red Queen dynamics (Woolhouse et al., 2002). Models of this process—“gene-for-gene” and “matching-allele” models, and variants of them—typically assume that the host genes involved encode proteins that interact directly with pathogens (Agrawal & Lively, 2003; Dybdahl, Jenkins, & Nußmer, 2014). Consequently, if host–pathogen coevolution is an important cause of balancing selection, genes encoding innate immune receptors involved in pathogen recognition could be expected to be more important targets than genes encoding signal transduction proteins. Moreover, receptors can be divided into two types: those recognizing components in bacterial or fungal cell walls, bacterial flagella, or viral capsids, and those recognizing features of microbial nucleic acids (e.g., dsRNA and unmethylated CpG sequences; Takeuchi & Akira, 2010). This classification is motivated because components of microbial cell walls, flagella and capsids are likely to be evolutionarily more dynamic than features of microbial nucleic acids. Thus, if balancing selection is driven by coevolution, we would expect it to primarily target receptors recognizing components of cell walls etc.

Another factor that could affect the likelihood of balancing selection on a gene is its degree of pleiotropy. In analyses across species, genes with a high degree of pleiotropy, measured for example as the number of protein–protein interactions, are more constrained (Chesmore, Bartlett, Cheng, & Williams, 2016; Fraser, Hirsh, Steimmetz, Scharfe, & Feldman, 2002). It seems likely that this effect can operate also at the intraspecific level, so that more pleiotropic genes are less likely to evolve under balancing selection. Gene function and pleiotropy are probably associated with each other (receptors having lower pleiotropy than signalling genes), and as far as we are aware no study has disentangled their relative importance for balancing selection.

To investigate if and how gene function and pleiotropy are associated with balancing selection, here we focus on a particular set of immune pathways—Pattern Recognition Receptor (PRR) signalling pathways. PRRs play a key role in the initiation of immune responses by recognizing molecular structures that are unique to microbes, so called Microbe-Associated Molecular Patterns (MAMPs). In vertebrates, there are four main families of PRRs: Toll-like receptors (TLRs), NOD-like receptors (NLRs), RIG-I-like receptors (RLRs) and C type lectin receptors (CLRs) (Takeuchi & Akira, 2010). All these PRRs have similar and partly overlapping signalling pathways. Briefly, ligand (MAMP) binding to a PRR leads to the recruitment of an adaptor protein. This initiates a signalling cascade involving phosphorylation and ubiquitination of signalling proteins, and ultimately activation and nuclear translocation of transcription factors. These transcription factors induce expression of inflammatory cytokines, chemokines, costimulatory molecules, etc. (Caruso, Warner, Inohara, & Nüñez, 2014; Gay, Symmons, Gangloff, & Bryant, 2014; Geijtenbeek & Gringhuis, 2009; Loo & Gale, 2011). Some NLRs also activate inflammatory cytokines such as interleukin-1β (IL-1β) and IL-18 (which are initially present as inactive precursors) through the formation of inflammasomes, instead of inducing transcription of cytokines through activation of transcription factors (Wen, Miao, & Ting, 2013).

We used whole-genome resequencing data from a population of bank voles (Myodes glareolus) to analyse patterns of intraspecific diversity in a set of 123 genes in PRR pathways, from receptors to transcription factors. First, we analysed which factors (gene function, pleiotropy) influence the amount of variation (nucleotide diversity) in coding sequences, and performed population genetic analyses to test if genes with exceptionally high diversity were under balancing selection. Second, we scanned for signatures of balancing selection in both protein coding and noncoding sequences of PRR pathway genes using betascan (Siewert & Voight, 2017), and then tested if these signatures were associated with gene function and pleiotropy. In both cases we employed tests designed to detect long-term (but not recent) balancing selection (Fijarczyk & Babik, 2015; Siewert & Voight, 2017).
Figure 1  Pattern Recognition Receptor (PRR) signalling pathway genes included in the analyses. PRR pathway genes were divided into two main functional categories: genes encoding receptors and genes encoding signal transduction proteins ("receptor genes" and "signalling genes," respectively). Receptors were divided into two subcategories depending on what types of ligands (Microbe-Associated Molecular Patterns; MAMPs) they recognize: components of bacterial and fungal cell walls, bacterial flagella or viral capsids versus features of microbial nucleic acids. Signal transduction genes include genes encoding adaptor proteins, protein kinases and other enzymes, transcription factors, etc. [Colour figure can be viewed at wileyonlinelibrary.com] [Colour figure can be viewed at wileyonlinelibrary.com]
2 | MATERIAL AND METHODS

2.1 | PRR signalling pathway genes

Based on KEGG pathways (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017) and a survey of recent review papers (Bowie & Unterholzner, 2008; Caruso et al., 2014; Dambuza & Brown, 2015; Franchi, Warner, Viani, & Nuñez, 2009; Gay et al., 2014; Geijtenbeek & Gringhuis, 2009; Hancock, Nijnik, & Philpott, 2012; Hoving, Wilson, & Brown, 2014; Lee, Avalos, & Ploegh, 2012; Loo & Gringhuis, 2009; Hancock, Nijnik, & Philpott, 2012; Hoving, Wilson, & Brown, 2014; Lee, Avalos, & Ploegh, 2012; Loo & Gale, 2011; Moynagh, 2014; Neill & Bowie, 2007; Neill, Golenbock, & Bowie, 2013; Takeuchi & Akira, 2010; Theofilopoulos et al., 2010; Wen et al., 2013) we compiled a list of 155 genes that have a well-described function in PRR pathways. Of these, 123 were annotated in our bank vole genome (see below) and had a coding sequence that covered at least 75% of the mouse orthologue (Figure 1; see Table S1 for a list of genes that were included, and Table S2 for a list of genes that could not be included).

PRR pathway genes were divided into two main categories—receptors and signalling genes—depending on the function of the encoded protein. Receptors include all genes encoding proteins directly involved in detection of MAMPs. Receptors were further divided into two groups depending on what type of MAMP they recognize: receptors that bind ligands from viral capsids, bacterial flagella, and bacterial or fungal cell walls (e.g., lipoproteins, lipopolysaccharides, zymosan, flagellin); and receptors that recognize features of microbial nucleic acids (e.g., dsRNA). Signalling genes encode proteins involved in the signalling cascade downstream of receptors. This category includes several different types of proteins; what they have in common is that they are not directly involved in recognition of MAMPs. The main types of signalling proteins are: (a) adaptor/scaffold proteins, which mediate protein-protein interactions and the formation of protein complexes, and typically lack enzymatic activity; (b) protein kinases, i.e., enzymes that phosphorylate other proteins; (c) ubiquitination enzymes, which are involved in the modification of proteins by the addition of ubiquitin to specific sites; (d) modulators, i.e., molecules regulating (usually inhibiting) the signalling cascade, without enzymatic activity; and (e) transcription factors. Not all signalling proteins could be assigned to the above categories. Note that NLRP10 is classified as a signalling gene—despite its name implying it is a receptor—because it lacks LRR motifs, which form the domain involved in MAMP recognition in other NLRs. PRR pathway genes were divided into functional categories based on information in the surveyed literature (see above). UniProt (The UniProt Consortium, 2017) and PhosphoSitePlus (Hornbeck et al., 2015) about the function of the encoded protein.

2.2 | Study species and field work

The bank vole (Myodes glareolus, Schreber 1780) is a small rodent in the subfamily Arvicolinae (lemmings, voles and muskrats). It occurs in forests and meadows with thick ground cover, from western Europe to central Siberia (Wilson, Lacher, & Mittermeier, 2017). Bank voles for this study were trapped at six different sites (N = 2–8 per site, depending on the area of each site) at Reningehed, an area comprising 43 km² of grazed grassland with small woods, marshes and meadows, situated 20 km east of Lund, southern Sweden. All sites are well connected, without any obvious dispersal barriers. Previous analyses based on microsatellite loci showed weak differentiation among the sites (FST 0.022–0.055; Tschirren, Andersson, Scherman, Westerdahl, & Råberg, 2012). To avoid including siblings we only sampled adults (i.e., after post-juvenile dispersal). Samples were collected with permission from the Malmö/Lund board for animal experiment ethics (permission M47-14).

2.3 | De novo genome sequencing, assembly and annotation

DNA was extracted from the spleen of a male bank vole. Two paired-end libraries with targeted insert sizes of 180 and 670 bp and two mate pair libraries with target insert sizes of 3 and 20 kb were sequenced (Table S3). All libraries were sequenced on a Hiseq X (Illumina). The reads were assembled de novo using a combination of ALLPATHS-LG (Gnerre et al., 2011), GMAP (Boetzer, Henkel, Jansen, Butler, & Pirovano, 2011) and BEAST (Sahlín, Chikhi, & Arvestad, 2016), with gap closing provided by GAPCLOSER in the SOAPDENOVovo package (Li et al., 2010) (Table S4). The initial assembly was filtered for contaminating sequences and short and redundant scaffolds. We further assembled scaffolds containing partial coding sequences of our focal immune genes by using synteny to the prairie vole or mouse genome. Gene annotation was performed on a repeatmasked version of the genome. Briefly, AUGUSTUS (Stanke, Diekhans, Baertsch, & Haussler, 2008) was used to generate gene models guided by RNAseq data from six tissues (Table S5) mapped to the genome with GSNAP (Wu, Reeder, Lawrence, Becker, & Brauer, 2016) and rodent proteins mapped to the reference genome with EXONERATE (Slater & Birney, 2005). We further performed manual curation of the focal genes in WEBAPOLLO (Lee et al., 2013). See the Supporting Information and Table S6 for further details on de novo genome assembly and annotation.

2.4 | Whole-genome resequencing and variant calling

DNA from spleens of 30 bank voles (Table S7) were extracted. Sequence libraries with a targeted insert size of 350 bp were prepared and sequenced on 30 Hiseq X (Illumina) lanes. For mapping we also included the 670-bp library used in the de novo genome assembly. The reads were mapped to the reference assembly using BWA (Li & Durbin, 2009) and read duplicates were removed with PICARDTOOLS (http://broadinstitute.github.io/picard/). Variants were called using FREEBAYES (Garrison & Marth, 2012). The raw variants were filtered
in several steps, including for low quality score, coverage and overlap with annotated repeats. See the Supporting Information and Table S8 for further details on whole-genome resequencing and variant calling.

We performed a phylogenetic clustering of mitochondrial haplotypes by downloading a set of complete bank vole mitochondrial genomes originating from several bank vole populations in Europe (Filipí, Marková, Searle, & Kotlík, 2015). We aligned the mitochondrial sequences of our samples with those of the European population set using ClustalW and created a neighbor-joining tree in the R package ape (Paradis, Claude, & Strimmer, 2004).

2.5 | Analyses of balancing selection in coding regions

We used PopGenome (Pfeifer, Wittelsbu, Ramos-onsins, & Lercher, 2014) to calculate nucleotide diversity (π) for the coding sequence of each of the 123 PRR pathway genes. To obtain an empirical background distribution of nucleotide diversity, we extracted coding sequences from a set of control genes in the bank vole genome. We initially considered all genes (N = 12,314) that could be uniquely assigned to a mouse protein through blast (excluding, for example, genes split over multiple scaffolds) and that were not included in our focal set of immune genes. We also required that the same bank vole gene was the reciprocally best blast hit of the mouse gene and that the highest scoring segment pair (HSP) of the coding sequence of the bank vole gene covered at least 90% of the mouse protein. This resulted in a total of 9,246 control genes.

For PRR pathway genes with exceptionally high nucleotide diversity in the coding region (“outlier genes,” with π above the 95th percentile of the empirical distribution based on 9,246 control genes), we used Hudson–Kreitman–Aguadé (HKA) tests (Hudson, Kreitman, & Aguadé, 1987) to investigate whether the high nucleotide diversity was a result of balancing selection (Fijarczyk & Babik, 2015). HKA tests were performed with the software MLHKA (Wright & Charlesworth, 2004), where polymorphism/divergence ratios (at synonymous sites) for a focal gene are compared to polymorphism/divergence ratios for a set of neutrally evolving genes. We first used either house mouse (Mus musculus) or prairie vole (Microtus ochrogaster) as the outgroup for a few of the highly polymorphic genes. In all cases the analyses yielded identical conclusions and we ultimately decided to perform all analyses with the house mouse as outgroup because its reference genome is better annotated. We used blast to match the coding sequence of the outlier bank vole genes to protein-coding transcripts of mouse and extracted the highest scoring transcript to be used for alignment with the bank vole sequences. The sequence alignment was performed with Geneious R11 (Biomatters) using “translation align.” The alignments were manually inspected and fixed insertions or deletions between the species were removed. We used DNAsp version 6 (Rozas et al., 2017) to obtain polymorphism and divergence estimates for the aligned data. Specifically, for each gene we calculated the number of synonymous segregating sites and the total number of synonymous sites in the bank vole samples. We also estimated the population-scaled mutation rate (θ) for each gene, to use as starting value for the MLHKA analysis. For divergence estimates, we calculated the number of synonymous differences between the mouse sequence and one of the haplotypes of the bank vole reference genome. To obtain a set of putatively neutrally evolving control genes for the MLHKA analyses, we first filtered the 9,246 control genes according to the following criteria: (a) ≥3 synonymous sites, (b) Tajima’s D of nonsynonymous and synonymous sites between −1 and 1, (c) size range within the distribution of the outlier genes (498–5,169 bp) and (d) the mouse orthologue is not located on a sex chromosome. From this set of suitable control genes (N = 3,340) we randomly selected 60 genes. We then performed McDonald–Kreitman tests on these genes (McDonald & Kreitman, 1994), using DNAsp (Rozas et al., 2017). We ranked the 60 genes according the Neutrality Index (NI). From this list we selected a set of 20 genes with a median NI as close as possible to 1 (NI median = 1.04, range: 0.7–2.0, all p ≥ .36), which were then used as control genes in the MLHKA tests. The alignment of these sequences to the mouse orthologue and the calculation of divergence and polymorphism followed the same approach as described above for the outlier genes (Table S9). For each outlier gene, we ran the MLHKA analysis at least four times with chain length = 100,000 and varying input parameters (divergence time [T] and mutation rate [θ]) to ensure convergence. We performed the HKA test on individual genes separately instead of groups of genes (e.g., comparing all receptor genes versus the 20 control genes) because we see little reason to expect that all (or even most) genes in a certain category should be under balancing selection.

Some of the focal genes are paralogues situated close to each other; they are therefore prone to paralogous gene conversion (PGC), which could confound signatures of balancing selection. To test for PGC, we used Geneious R11 to construct dot plots, and DNAsp version 6 (Rozas et al., 2017) to perform sliding window analyses of coding sequence nucleotide diversity (π) and nucleotide divergence between the paralogues (Dxy). We used Popart (Leigh & Bryant, 2015) to construct median-joining networks (Bandelt, Forster, & Röhl, 1999).

2.6 | Analyses of balancing selection in coding and noncoding regions

We used betascan (Siewert & Voight, 2017) to search for more localized signatures of long-term balancing selection in coding and noncoding sequences of the 123 PRR pathway genes. Betascan calculates a summary statistic (β) that quantifies allele frequency correlations between single nucleotide polymorphisms (SNPs). Under long-term balancing selection there is expected to be a cluster of SNPs with similar allele frequencies, which will cause β to be high (Siewert & Voight, 2017). In the software we used the folded allele frequency spectrum, and set a minimum minor allele frequency of 0.15 (to reduce the number of false positive signals) as recommended (Siewert & Voight, 2017). As the average recombination rate in rodents is
about half that in humans (Jensen-Seaman et al., 2004), we used a window size of 2 kb, twice of that used by Siewert and Voight (2017). We also evaluated window sizes of 1 and 5 kb and found the results qualitatively similar to those based on 2 kb. To reduce the risk of spurious signals caused by copy number variation and incorrectly mapped paralogues, we filtered beta values for SNPs located within 1 kb of a duplication or deletion called by delly (Rausch et al., 2012) (see Supporting Information). For each PRR pathway gene we selected the highest $\beta$ value within the annotated gene interval $\pm$ 2 kb ($\beta_{\text{max}}$). $\beta$ could only be calculated for 119 of 123 PRR pathway genes because some lacked SNPs with sufficiently high minor allele frequency. To obtain a null distribution for $\beta_{\text{max}}$ we calculated the empirical distribution of $\beta_{\text{max}}$ within the annotated gene interval $\pm$ 2 kb of the control genes ($N = 8,818$; fewer than in the set of control genes for $\pi$ because of the filtering steps above; Figure S1), and considered PRR pathway genes with a $\beta_{\text{max}}$ $\geq$ the 95th percentile ($\beta = 235.5$) as potential targets of long-term balancing selection.

In analyses of $\beta_{\text{max}}$ we included gene length as a covariate, as longer genes provide a larger target for balancing selection. Gene length was calculated as the length of the annotated gene interval plus 2,000 bp at either end. For genes with signatures of balancing selection according to betascan, we used popgenome (Pfeifer et al., 2014) to perform sliding window analyses with a window size of 2,000 bp and a step size of 500 bp for nucleotide diversity and Tajima’s $D$ (where a high value is indicative of balancing selection; Fijarczyk & Babik, 2015; Tajima, 1989) across the entire gene and in up- and downstream regions.

2.7 | Sanger sequencing

Copy-number variation and incorrect mapping of paralogues can lead to artefactually high nucleotide diversity and indicate that a locus is under balancing selection. To avoid this, the genome resequencing data were subjected to a number of filtering steps before the analyses above (Supporting Information). To ensure these filtering steps were sufficient, we performed Sanger sequencing of a selection of highly polymorphic genes (CARD6, CD36, CLEC5A, DDX58, IFI1H1, IRAK3, LY75, NLRP10, NOD1, TLR1, TLR2, TLR5, TLR6; see Table S10 for primers). Primers were designed to amplify a fragment covering the peak (or, in the case of NOD1, as close as possible to the peak) in nucleotide diversity and/or $\beta$. The PCR products were sequenced in both directions on a Genetic Analyser 3500 (Applied Biosystems) using Big Dye terminator chemistry (Applied Biosystems). Sequences were processed, assembled and aligned using geneious R11. For each locus, we Sanger sequenced 5–38 individuals from an independent set of bank voles (sampled from the same population as the genome resequencing data set; see Table S10 for sample sizes for each gene), so that we could confirm that SNPs with relatively even allele frequencies were present as both homozygotes and heterozygotes (which should rule out copy-number variation and incorrect mapping of paralogues). Sanger sequencing confirmed the SNPs observed in the genome resequencing data set for all the tested genes. Thus, we are confident that the filtering steps were sufficient to eliminate confounding effects of copy-number variation and incorrect mapping.

2.8 | Pleiotropy

We used the number of protein–protein interactions (PPI, i.e., the number of other host proteins a given protein in PRR signalling pathways interacts with) as a proxy for pleiotropy (Papakostas et al., 2014; Zhang & Yang, 2015). The numbers of PPIs for Mus musculus orthologues of the bank vole PRR pathway genes were retrieved from the STRING database (Szklarczyk et al., 2015), and were based on the sources “experiments” and “databases,” with high confidence (0.7).

2.9 | Statistical analyses

Analyses of factors influencing nucleotide diversity and $\beta_{\text{max}}$ were performed with proc glm in SAS 9.4 (SAS Institute). Nonsignificant terms ($p > .1$) were deleted in a step-wise manner. Nucleotide diversity, $\beta_{\text{max}}$, PPI and gene length were log$_{10}$ transformed to normalize the distribution of residuals and improve homoscedasticity.

3 | RESULTS

3.1 | De novo genome assembly

We constructed a de novo genome assembly from a male bank vole caught in southern Sweden using four paired-end short read libraries (Table S3). Following several assembly and filtering steps we obtained an assembly comprising 50,926 scaffolds with a scaffold N50 of 1.6 Mb and a total scaffold length of 2.4 Gb (Table S4). Using RNAseq data from six different tissues (Table S5) and protein homology data from other rodents we annotated 24,940 gene models that correspond to 16,983 unique mouse genes.

3.2 | Whole-genome resequencing

To assay variation in a population of bank voles from southern Sweden we performed whole-genome resequencing of an additional 30 individuals and mapped these data to the reference genome with an average coverage of mapped reads of 44× (Table S8). Using the whole-genome resequencing data, together with a paired-end library from the reference genome sample, we obtained a set of 21.6 million filtered SNPs of which 21.4 million were bi-allelic. To provide a basic characterization of the samples included in the resequencing data set from a phylogeographic perspective, we performed a phylogenetic analysis of full-length mitochondrial haplotypes. This showed that our samples formed a monophyletic group with shallow divergence belonging to the “Carpathian” mitochondrial DNA
(mtDNA) clade (Figure S2), confirming previous analyses of bank vole populations in southern Sweden based on smaller data sets (Filipi et al., 2015; Morger et al., 2015).

### 3.3 • PRR gene set

We obtained complete or nearly complete coding sequences (at least 75% of the mouse orthologues) for 123 PRR signalling pathway genes: 35 encoding receptors and 88 encoding signal transduction proteins (Figure 1; Table S1). Of the receptors, 22 recognize components of microbial cell walls etc., 10 recognize nucleic acids, and for three the ligands are unclear (Figure 1).

### 3.4 • Associations between gene function, pleiotropy and coding sequence nucleotide diversity of PRR pathway genes

To test for effects of gene function (receptor versus signalling) and number of protein–protein interactions (PPI; as a proxy for pleiotropy; see Material and Methods) on nucleotide diversity, we performed a general linear model with nucleotide diversity against gene function, PPI and their interaction. There were significant main effects of both gene function (\( F = 15.65, df = 1, 119, p < .0001 \)) and PPI (\( F = 22.14, df = 1, 119, p < .0001 \)), and also a significant interaction between gene function and PPI (\( F = 8.13, df = 1, 119, p = .0051 \)). Inspection of Figure 2a shows that receptor genes on average had higher diversity than signalling genes, but only at low PPI. This conclusion is robust to restricting the analysis to subsets of the data (e.g., PPI ≤ 10 or PPI ≤ 40) to eliminate problems with heteroscedasticity and different ranges of PPI between receptor and signalling genes (data not shown).

To test for a difference in diversity between receptors recognizing different types of ligands (proteins etc. versus nucleic acids), we performed a general linear model with nucleotide diversity against receptor type, PPI and their interaction (excluding the three receptors with unknown ligands; Figure 1). This analysis showed a main effect of receptor type (\( F = 6.61, df = 1, 27, p = .016 \); Figure 2b) and a marginally nonsignificant interaction (\( F = 3.87, df = 1, 27, p = .06 \)), but no overall effect of PPI (PPI: \( F = 1.48, df = 1, 27, p = .23 \)).

### 3.5 • Signatures of balancing selection in coding sequences of PRR pathway genes

High nucleotide diversity can be a result of balancing selection, but it could also simply be a result of a high mutation rate and/or relaxed
constraint. To investigate if the high nucleotide diversity observed in the coding sequence of some PRR genes (see Figure 2a) was a result of balancing selection, we first identified which of the 123 PRR genes had exceptionally high nucleotide diversity. For this, we obtained SNP variation of the coding sequence for a set of 9,246 high-quality control genes (see Materials and Methods for criteria) to use as an empirical background distribution. Both receptor and signalling genes had higher median nucleotide diversity than control genes (Mann–Whitney tests, receptor versus control: \( p < .001 \); signalling versus control: \( p < .006 \); Figure S3). We used the 95th percentile of nucleotide diversity (\( \pi = 0.0037 \)) in the control gene data set as a cut-off to define which of the focal immune genes had exceptionally high levels of variation. Nine receptors and three signalling genes had a coding sequence diversity above this threshold (Figure S3; Table 1). We then performed the HKA test (see Materials and Methods), a classical population genetic test for balancing selection in coding and noncoding sequences, on these 12 outlier genes.

Six receptors (CD207, CLEC4E, CLEC5A, TLR1, TLR2, TLR6) and one signalling gene (NLRP10) showed signatures of balancing selection based on the HKA test (Table 1). Thus, the patterns of polymorphism/divergence of seven out of the 12 outlier genes were consistent with their high nucleotide diversity being a result of balancing selection. Note, however, that in the case of TLR1 and TLR6 there is evidence for paralogous gene conversion, which complicates the interpretation of these tests, as discussed further below.

The proportion of receptor genes with signatures of balancing selection in the coding sequence was significantly higher than the proportion of signalling genes (not considering TLR1 and TLR6; 4/35 versus 1/88; Fishers’s exact test: \( p = .023 \)). The difference between receptors recognizing cell wall components etc. versus receptors recognizing nucleic acids was, however, not statistically significant (not considering TLR1 and TLR6; 4/22 versus 0/10; \( p = .28 \)). Signatures of balancing selection in four out of 22 genes encoding receptors for components of microbial cell walls etc. is significantly more than the 0.05 × 22 = 1.1 expected by chance (binomial test for \( \geq 4/22: p < .022 \)).

### 3.6 Associations between gene function, pleiotropy and signatures of balancing selection in coding and noncoding sequences

The HKA test of whole coding sequences described above pick up signals of balancing selection that span a substantial part of the coding sequence of a gene, but has little power to detect signals of more localized selection on the coding sequence (in particular in large multi-exon genes), and selection on noncoding regions of a gene. We therefore used BETASCAN to test for more localized signatures of selection in coding and noncoding sequences.

A general linear model with \( \beta_{\text{max}} \) against gene function, PPI, gene length and their two-way interactions showed that receptors had higher \( \beta_{\text{max}} \) than signalling genes (\( F = 5.07, df = 1, 115, p = .026 \); Figure 3a), while controlling for negative effects of PPI (\( F = 4.28, df = 1, 115, p = .041 \); Figure 3a) and gene length (\( F = 14.3, df = 1, 115, p = .001 \); Figure 3b; all interactions deleted at \( p > .55 \)). A separate analysis of receptor genes showed that receptors recognizing components of microbial cell walls etc. had higher \( \beta_{\text{max}} \) than receptors recognizing nucleic acids (\( F = 5.4, df = 1, 28, p = .027 \)), while controlling for gene length (\( F = 10.78, df = 1, 28, p = .003 \); PPI and all interactions deleted at \( p > .13 \); Figure S4).

Of the 119 PRR pathway genes for which we could calculate \( \beta \), seven (four receptors and three signalling genes; 4/34 versus 3/85, Fisher’s exact test: \( p = .1 \)) had a \( \beta_{\text{max}} \) greater than or equal to the 95th percentile of the empirical distribution (Table 2). In most cases, \( \beta_{\text{max}} \) was located in noncoding regions (mostly introns) and coincided with peaks in nucleotide diversity and Tajima’s D (Figure S5). There was a

| Gene  | \( \pi \) (CDS) | \( \pi \) percentile rank | Bank vole–Mus alignment length (bp) | HKA \( k \) | HKA 2\( \Delta \)LL |
|-------|----------------|---------------------------|-------------------------------------|----------|----------------|
| CARD6 | 0.00562        | 98.59                     | 3,123                               | 1.18     | -0.5          |
| CD207 | 0.00466        | 97.44                     | 966                                 | 2.85     | 5.24          |
| CLEC4D| 0.00817        | 99.56                     | 657                                 | 2.68     | 3.73          |
| CLEC4E| 0.00861        | 99.60                     | 642                                 | 3.31     | 6.07          |
| CLEC4N| 0.00499        | 97.93                     | 627                                 | 2.97     | 3.7           |
| CLEC5A| 0.00723        | 99.33                     | 495                                 | 4.28     | 6.76          |
| IRAK3 | 0.00384        | 95.37                     | 1,827                               | 1.65     | 1.41          |
| NLRP10| 0.00569        | 98.68                     | 2,019                               | 2.25     | 3.92          |
| TLR1  | 0.00521        | 98.15                     | 2,385                               | 2.12\( ^a \) | 4.3         |
| TLR2  | 0.00580        | 98.75                     | 2,349                               | 1.89     | 3.93          |
| TLR5  | 0.00433        | 96.87                     | 2,574                               | 1.38     | 0.64          |
| TLR6  | 0.00603        | 98.88                     | 2,418                               | 2.32\( ^a \) | 4.95        |

Note: Genes with a signature of balancing selection based on Hudson–Kreitman–Aguadé (HKA) tests (performed with MLHKA; Wright & Charlesworth, 2004) are indicated in bold (\( p < .05 \) based on likelihood ratio test with 1 df). CDS, coding sequence.

\( ^a \)Value questionable because of paralogous gene conversion.
positive correlation between $\beta_{\text{max}}$ and nucleotide diversity of the full coding sequence ($r_s = 0.648$, $N = 119$, $p < .01$).

3.7 | Confounding effects of paralogous gene conversion?

One potentially confounding factor in analyses of balancing selection is PGC. PGC can occur between paralogues with high sequence identity situated in tandem (Chen, Cooper, Chuzhanova, Férec, & Patrinos, 2007). TLR1 and TLR6 are estimated to be 26 kb apart in the bank vole genome and the coding sequences have 71.7% identity at the nucleotide level; these two genes are therefore prone to PGC. Indeed, previous analyses have shown that PGC constrains divergence of TLR1 and TLR6 between species (Kruithof, Satta, Liu, Dunoyer-geindre, & Fish, 2007). The rate of PGC is typically at least an order of magnitude higher than single nucleotide mutations; consequently, PGC can inflate the nucleotide diversity (Innan, 2003; Thornton, 2007). The high nucleotide diversity of TLR1 and TLR6 could therefore be a result of PGC instead of balancing selection.

Detailed analyses of diversity within and divergence between TLR1 and TLR6 provided strong evidence for PGC between these genes also in the bank vole, and nucleotide diversity peaked in the region affected by PGC (Figure 4a). PGC violates assumptions of neutrality tests traditionally used to detect balancing selection, such as the HKA test (Innan, 2003; Thornton, 2007). We therefore regard the results of the HKA tests for both TLR1 and TLR6 as dubious. Still, haplotype networks based on the regions subject to PGC showed that in both TLR1 and TLR6 there were divergent haplotype groups at relatively even frequencies (Figure 4b); this pattern suggests that the high diversity is not just an effect of PGC in combination with drift, but that balancing selection has also played a role (Storz et al., 2007).
the bank vole than TLR1 and TLR6. However, there are some potential candidates. First, some species (e.g., Homo sapiens, but not Mus musculus) have a TLR2 pseudogene (Roach et al., 2005). In humans, TLR2 and TLR2P are <3 kb apart and the TLR2 coding region has 67% sequence identity with the corresponding part of TLR2P, which makes PGC a possibility. Given the apparent effect of PGC on nucleotide diversity in bank vole TLR1 and TLR6, we considered the possibility that the reason TLR2 is so polymorphic in bank voles is that there is a pseudogene in this species (or perhaps voles in general), and PGC between TLR2 and TLR2P. However, based on BLASTN searches we could not find a TLR2 pseudogene in either our bank vole genome assembly or the prairie vole vole genome (MichOch1.0), so there is no indication that PGC plays a role in bank vole TLR2 diversity (analyses not shown). Second, several of the C type lectin receptors with signatures of balancing selection are situated closely to each other (CLEC4N, CLEC4D and CLEC4E within 76 kb) or to other paralogues (CD207 in tandem with CLEC4F; CD209B between CD209D and CD209C). However, in these cases, the sequence identities of coding regions are relatively low (≤49%) and dot plots of whole genes (coding and noncoding regions) revealed no indications of PGC (data not shown).

4 | DISCUSSION

Based on analyses of both coding and noncoding regions, we identified a set of PRR pathway genes with signatures of balancing selection. Human is the only other mammal where comprehensive analyses of intraspecific diversity of innate immune genes have been performed. Human orthologues to at least three of the 123 genes investigated here have signatures of balancing selection in at least one population, including IFIH1, CD207 and NOD2 (Forni et al., 2014;
Fumagalli et al., 2010; Nakagome et al., 2012). Of these, CD207 was under balancing selection also in bank voles.

Tests for signatures of balancing selection are sensitive to confounding effects of demographic processes (e.g., bottlenecks), gene conversion, admixture and introgression (Fijarczyk & Babik, 2015). The inclusion of a set of control genes in the HKA analyses makes these relatively robust to confounding effects of demography (Wright & Charlesworth, 2004). We addressed the potential effects of gene conversion and found evidence that it has contributed to the high diversity of TLR1 and TLR6, but not in other cases. Further analyses based on more than one bank vole population, and other closely related species, are required to investigate if admixture and introgression have contributed to the high levels of polymorphism in some PRR pathway genes in the bank vole, as appear to be the case for some immune genes in humans (Dannemann, Andrés, & Kelso, 2016; Fumagalli et al., 2010). With these caveats, we regard the genes identified here (Tables 1 and 2) as highly likely targets of balancing selection.

Analyses of coding sequences showed that genes encoding receptors had higher nucleotide diversity than genes encoding signaling proteins, although only at low levels of pleiotropy (measured as the number of protein–protein interactions). Moreover, genes encoding receptors recognizing components of microbial cell walls etc. had higher diversity than receptors recognizing nucleic acids. Four of the 22 genes encoding receptors for components of microbial cell walls etc. (or six if TLR1 and TLR6 are included) had signatures of balancing selection based on HKA tests, but only one of the other 101 genes (encoding signal transduction proteins and receptors for nucleic acids). Taken together, these results indicate that gene function is an important determinant of balancing selection, with genes encoding receptors recognizing components of microbial cell walls etc. being particularly probable targets. A similar pattern, with higher diversity of TLR genes encoding receptors for components of microbial cell walls etc. than for those recognizing nucleic acids, has previously been found in other rodents and humans (Fornusková et al., 2013; Wlasiuk & Nachman, 2010). The pattern we found also mirrors results from analyses of positive selection on coding sequences of immune genes in Drosophila, where immune receptor genes are more often under positive selection than other immune genes (Sackton et al., 2007).

Analyses of more localized signals of balancing selection (based on $β$) in coding and noncoding regions showed that receptor genes had higher values of $β_{max}$, while controlling for PPI and gene length. Moreover, receptor genes recognizing components of microbial cell walls etc. had higher values of $β_{max}$ than receptors recognizing nucleic acids. Seven genes had $β_{max}$ above the 95th percentile of the genome-wide distribution of $β_{max}$. Of these, four were receptors (CD36, CLEC5A, DDX58, LY75) and three were signalling genes (CARD6, NLRP10, PRKCE). Of the receptors, two (CD36, CLEC5A) recognize components of microbial cell walls etc. and one (DDX58) recognizes nucleic acids. The signalling genes belong to different categories: CARD6 is an adaptor, PRKCE is a protein kinase, while NLRP10 is considered a (negative) regulator of innate immune responses. Thus, analyses of $β_{max}$ yielded similar patterns as the analyses of coding sequence nucleotide diversity, in the sense that receptors on average had higher $β_{max}$ than signalling genes, and receptors recognizing components of cell walls etc. on average had higher $β_{max}$ than receptors recognizing nucleic acids. However, unlike in the analyses of overall coding sequence nucleotide diversity (where signatures of balancing selection where mainly found in genes encoding receptors for cell wall components etc.), genes belonging to all categories were fairly evenly represented among the genes considered to have signatures of balancing selection based on $β_{max}$ (i.e., with $β_{max}$ above the 95th percentile). This is possibly an effect of $β_{max}$ in these cases being mostly located in noncoding sequences (in particular in introns; Table 2), suggesting that $β_{max}$ primarily picked up signatures of selection acting on regulatory variation (although we note that it is difficult to determine the location of the target of selection based on these analyses alone). The prediction, based on models of coevolution (Agrawal & Lively, 2003; Dybdahl et al., 2014), that receptors recognizing cell wall components etc. should be key targets of balancing selection primarily concerns selection on coding sequences. In contrast, when it comes to balancing selection on regulatory variation, it is more difficult to see why genes in a certain functional group should be particularly important targets.

The analyses of whole coding sequences showed a negative association between pleiotropy and nucleotide diversity. Similarly, the analyses of localized signatures of balancing selection in coding and noncoding sequences showed negative associations between pleiotropy and $β_{max}$. Again, these patterns mirror results of analyses across species, where genes with a higher number of protein–protein interactions are more constrained (Chesmore et al., 2016; Fraser et al., 2002). However, in the case of divergence across species, the most important factor is the level of expression (highly expressed genes are more constrained), and it is debatable whether there is an independent effect of pleiotropy (Alvarez-Ponce, 2012; Fraser & Hirsh, 2004; Pål, Papp, & Lercher, 2006; Zhang & Yang, 2015). Further studies are required to determine if the association between nucleotide diversity and pleiotropy found in the present study is a direct effect, or rather a result of both pleiotropy and diversity being associated with the level of gene expression.

Our finding that genes encoding receptors for pathogen recognition, in particular receptors recognizing components of microbial cell walls etc., are primary targets of balancing selection is consistent with the idea that coevolution is an important driver of pathogen-mediated balancing selection (Haldane, 1949; Woolhouse et al., 2002). However, it is important to recognize that this pattern does not exclude other causes of balancing selection, such as heterozygote advantage (Doherty & Zinkernagel, 1975), environmental heterogeneity in pathogen abundance (Hedrick, 2002; Lazzaro & Little, 2009) and “diffuse interactions” (i.e., negative frequency-dependent selection involving a multitude of pathogens, but without coevolution; Karasov et al., 2014). Ultimately, the relative importance of different processes needs to be determined gene by gene through analyses of functional effects of polymorphisms. One key
ACKNOWLEDGEMENTS

This work was supported by grants from the Swedish Research Council (2015-05418 to L.R.), the Crafoord foundation (20150741 to L.R.), the Erik Philip-Sörensen foundation (to L.R.) and the Royal Physiographic Society (to X.Z.). We are grateful to Emily O’Connor for comments on the manuscript. Kristoffer Sahlin provided useful comments on the use of the betascan software and scaffolding of genomes in general, Mario Stanke provided suggestions on how to run the Augustus software, and Katherine Siewert provided useful comments about the usage of the Betascan software. Anna Bergqvist, Melissa Chan, Johanna Hedenström and Ilona Szücs helped with Sanger sequencing. The authors would like to acknowledge support from the Science for Life Laboratory, the National Genomics Infrastructure, NGI, Stockholm, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure.

AUTHOR CONTRIBUTIONS

L.R. and M.L. designed the study. L.R. performed fieldwork and DNA/RNA extraction. A.K. performed library preparations for the de novo genome sequencing. R.A.O. performed the initial de novo genome assembly and quality control. M.L. performed the main bioinformatics analyses. L.R. performed statistical and population genetic analyses. X.Z. performed Sanger-sequencing and associated analyses. M.L. and L.R. wrote the manuscript with input from X.Z. All authors approved the manuscript.

DATA AVAILABILITY STATEMENT

Sequence data have been deposited in SRA under bioproject PRJNA335935. The genome assembly has been deposited in GenBank with accession no. MULK00000000.1. The data set with population genetic parameters and PPI for each PRR pathway gene, as well as the genome annotation, has been provided at Dryad (https://doi.org/10.5061/dryad.2bqv83bms). Code for genome analyses is available upon request.

ORCID

Max Lundberg https://orcid.org/0000-0002-1895-3622
Xiurin Zhang https://orcid.org/0000-0002-4772-4255
Lars Råberg https://orcid.org/0000-0001-5219-7448

REFERENCES

Agrawal, A. F., & Lively, C. M. (2003). Modelling infection as a two-step process combining gene-for-gene and matching-allele genetics. Proceedings of the Royal Society of London. Series B, 270, 323–334. https://doi.org/10.1098/rspb.2002.2193

Alvarez-Ponce, D. (2012). The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. BMC Evolutionary Biology, 12, 192. https://doi.org/10.1186/1471-2148-12-192

Andres, A. M., Hubisz, M. J., Indap, A., Torger son, D. G., Degenhardt, J. D., Boyko, A. R., ... Nielsen, R. (2009). Targets of balancing selection in the human genome. Molecular Biology and Evolution, 26, 2755–2764. https://doi.org/10.1093/molbev/msp190

Bakker, E. G., Toomajian, C., Kreitman, M., & Bergelson, J. (2006). A genome-wide survey of R gene polymorphisms in Arabidopsis. The Plant Cell, 18, 1803–1818.

Bandelt, H.-J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Molecular Biology, 16, 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Bergelson, J., Kreitman, M., Stahl, E. A., & Tian, D. (2001). Evolutionary dynamics of plant R-genes. Science, 292, 2281–2285. https://doi.org/10.1126/science.1061337

Bitarello, B. D., de Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & André, A. M. (2018). Signatures of long-term balancing selection in human genomes. Genome Biology and Evolution, 10, 939–955. https://doi.org/10.1093/gbe/evy054

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27, 578–579. https://doi.org/10.1093/bioinformatics/btq683

Bowie, A. G., & Unterholzner, L. (2008). Viral evasion and subversion of pattern-recognition receptor signalling. Nature Reviews Immunology, 8, 911–922. https://doi.org/10.1038/nri2436

Caglìani, R., Fumagalli, M., Biasin, M., Piacentini, L., Riva, S., Pozzoli, U., ... Sironi, M. (2010). Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Human Genetics, 128, 577–588. https://doi.org/10.1007/s00439-010-0884-6

Caglìani, R., Fumagalli, M., Riva, S., Pozzoli, U., Comi, G. P., Menozzi, G. et al (2008). The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biology, 9, R143.

Caruso, R., Warner, M., Inohara, N., & Núñez, G. (2014). NOD1 and NOD2: Signaling, host defense, and inflammatory disease. Immunity, 41, 898–908. https://doi.org/10.1016/j.immuni.2014.12.010

Chen, J. M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: Mechanisms, evolution and human disease. Nature Reviews Genetics, 8, 762–775. https://doi.org/10.1038/nrg2193

Chesmore, K. N., Bartlett, J., Cheng, C., & Williams, S. M. (2016). Complex patterns of association between pleiotropy and transcription factor evolution. Genome Biology and Evolution, 8, 3159–3170. https://doi.org/10.1093/gbe/evw228

Dambuza, I. M., & Brown, G. D. (2015). C-type lectins in immunity: Recent developments. Current Opinion in Immunology, 32, 21–27. https://doi.org/10.1016/j.coi.2014.12.002

Dannemann, M., Andrés, A. M., & Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. American Journal of Human Genetics, 98, 22–33. https://doi.org/10.1016/j.ajhg.2015.11.015

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genetics, 10, e1004561. https://doi.org/10.1371/journal.pone.1004561

Doherty, P. C., & Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature, 256, 50–52. https://doi.org/10.1038/256050a0

Dybå, M. F., Jenkins, C. E., & Nielsen, S. L. (2014). Identifying the molecular basis of host-parasite coevolution: Merging models and mechanisms. American Naturalist, 184, 1–13. https://doi.org/10.1086/676591

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: Limits and prospects. Molecular Ecology, 24, 3529–3545. https://doi.org/10.1111/mec.13226

Fijarczyk, A., Dudek, K., & Babik, W. (2016). Selective landscapes in new immune genes inferred from patterns of nucleotide variation. Genome
Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Hoving, J. C., Wilson, G. J., & Brown, G. D. (2014). Signalling C-Type lectin receptors, microbial recognition and immunity. Molecular Biology and Evolution, 8, 946–479. https://doi.org/10.1093/molbev/msu309

Fornósková, A., Vinkler, M., Pagés, M., Galan, M., Jousselin, E., Cerqueira, F., ... Cosson, J.-F. (2013). Contrastressed evolutionary histories of two Toll-like receptors (Tlr4 and Tlr7) in wild rodents (Murinae). BMC Evolutionary Biology, 13, 194. https://doi.org/10.1186/1471-2148-13-194

Franchi, L., Warner, N., Viani, K., ... Nuñez, G. (2009). Function of Nod-like receptors in microbial recognition and host defense. Immunological Reviews, 227, 106–128. https://doi.org/10.1111/j.1600-065X.2008.00734.x

Fraser, H. B., & Hirsh, A. E. (2004). Evolutionary rate depends on number of protein–protein interactions independently of gene expression level. BMC Evolutionary Biology, 5, 13.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. Science, 296, 750–752. https://doi.org/10.1126/science.1068696

Fumagalli, M., Caglioni, R., Riva, S., Pozzoli, U., Blasin, M., Piacentini, L., ... Sironi, M. (2010). Population genetics of IFIH1: Ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. Molecular Biology and Evolution, 27, 2555–2566. https://doi.org/10.1093/molbev/msq141

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, 1207.3907v.

Gay, N. J., Symmons, M. F., Gangloff, M., & Bryant, C. E. (2014). Assembly and localization of Toll-like receptor signalling complexes. Nature Reviews Immunology, 14, 546–558. https://doi.org/10.1038/nri3713

Geijtenbeek, T. B. H., & Gringhuis, S. I. (2009). Signalling through C-type lectin receptors: Shaping immune responses. Nature Reviews Immunology, 9, 465–479. https://doi.org/10.1038/nri2569

Gnerre, S., MacCallum, I., Przybilsky, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., ... Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences, 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

Haldane, J. B. S. (1949). Disease and evolution. Suppl. to La Ric. Sci., 19, 68–76.

Hancock, R. E. W., Nijink, A., & Philpott, D. J. (2012). Modulating immunity as a therapy for bacterial infections. Nature Reviews Microbiology, 10, 243–254. https://doi.org/10.1038/nrmicro2745

Hedrick, P. W. (2002). Pathogen resistance and genetic variation at MHC loci. Evolution (N. Y.), 56, 1902–1908.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., ... Skrzyypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Research, 43, D512–D520. https://doi.org/10.1093/nar/gku1267

Hoving, J. C., Wilson, G. J., & Brown, G. D. (2014). Signalling C-Type lectin receptors, microbial recognition and immunity. Cellular Microbiology, 16, 185–194. https://doi.org/10.1111/cmi.12249

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics, 116, 153–159.

Innan, H. (2003). The coalescent and infinite-site model of a small multigene family. Genetics, 810, 803–810.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C., ... Jacob, H. J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. Genome Research, 14, 528–538. https://doi.org/10.1101/gr.1970304

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomics, pathways, diseases and drugs. Nucleic Acids Research, 45, 353–361. https://doi.org/10.1093/nar/gkw1092

Karasov, T. L., Kniskern, J. M., Gao, L., DeYoung, B. J., Ding, J., Dubiella, U., ... Bergelson, J. (2014). The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature, 512, 436–440. https://doi.org/10.1038/nature13439

Krühlhof, E. K. O., Satta, N., Liu, J. W., Dunoyer-geindre, S., & Fish, R. J. (2007). Gene conversion limits divergence of mammalian TLR1 and TLR6. BMC Evolutionary Biology, 10, 148. https://doi.org/10.1186/1471-2148-10-148

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

Lee, C. C., Avalos, A. M., & Ploegh, H. L. (2012). Accessory molecules for Toll-like receptors and their function. Nature Reviews Immunology, 12, 168–179. https://doi.org/10.1038/nri3599

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., ... Lewis, S. E. (2013). Web Apollo: A web-based genomic annotation editing platform. Genome Biology, 14, R93. https://doi.org/10.1186/gb-2013-14-8-r93

Leigh, J. W., & Bryant, D. (2015). POPART: Full-feature software for haplotype network construction. Methods in Ecology and Evolution, 6, 1110–1116.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows – Wheeler transform. Bioinformatics, 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., ... Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Research, 20, 265–272. https://doi.org/10.1101/gr.097261.109

Loo, Y. M., & Gale, M. (2011). Immune signaling by RIG-I-like receptors. Immunity, 34, 680–692. https://doi.org/10.1016/j.immuni.2011.05.003

McDonald, J. H., & Kreitman, M. (1994). Adaptive protein evolution at the Adh locus in Drosophila. Nature, 367, 532–538.

Morger, J., Råberg, L., Hille, S. M., Helsen, S., Štefka, J., Al-Sabi, M. M., ... Tschirren, B. (2015). Distinct haplotype structure at the innate immune receptor Toll-like receptor 2 across bank vole populations and lineages in Europe. Biological Journal of the Linnean Society, 82, 941–959. https://doi.org/10.1111/bij.12593

Moynagh, P. N. (2014). The roles of Pellino E3 ubiquitin ligases in immunity. Nature Reviews Immunology, 14, 122–131. https://doi.org/10.1038/nri3599

Nakagome, S., Mano, S., Kozlowski, L., Bujnicki, J. M., Shibata, H., Fukushima, Y., ... Oota, H. (2012). Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. Molecular Biology and Evolution, 29, 1569–1585. https://doi.org/10.1093/molbev/msq006

Neill, L. A. J. O., & Bowie, A. G. (2007). The family of five: TIR-domain-containing adapters in Toll-like receptor signalling. Nature Reviews Immunology, 7, 353–364. https://doi.org/10.1038/nri2079

Neill, L. A. J. O., Golenbock, D., & Bowie, A. G. (2013). The history of Toll-like receptors - redefining innate immunity. Nature Reviews Immunology, 13, 453–460. https://doi.org/10.1038/nri3446

Pál, C., Papp, B., & Lercher, M. J. (2006). An integrated view of protein evolution. Nature Reviews Genetics, 7, 337–348. https://doi.org/10.1038/nrg1838
Papakostas, S., Vallestad, L. A., Bruneaux, M., Aykanat, T., Vanoverbeke, J., Ning, M., ... Leder, E. H. (2014). Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. *Nature Communications*, 5, 4071. https://doi.org/10.1038/ncomms5071

Paradis, E., Claude, J., & Strimmer, K. (2004). *ape*: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289–290.

Pfeifer, B., Wittelsbue, U., Ramos-onsins, S. E., & Lercher, M. J. (2014). *POPGENOME*: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, 31, 1929–1936. https://doi.org/10.1093/molbev/msu136

Quintana-Murci, L., & Clark, A. G. (2013). Population genetic tools for dissecting innate immunity in humans. *Nature Reviews Immunology*, 13, 280–293. https://doi.org/10.1038/nri3421

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). *delly*: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28, 333–339.

Roach, J. C., Glusman, G., Rowen, L., Kaur, A., Purcell, M. K., Smith, K. D., ... Aderem, A. (2005). The evolution of vertebrate Toll-like receptors. *Proceedings of the National Academy of Sciences USA*, 102, 9577–9582. https://doi.org/10.1073/pnas.0502272102

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, 34, 3299–3302. https://doi.org/10.1093/molbev/msx248

Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, M. S., & Clark, A. G. (2015). Hyperdiverse gene cluster in snail host conveys resistance to human schistosome parasites. *Nature Communications*, 6, 11771. https://doi.org/10.1038/ncomms11771

Sahlén, K., Chikhi, R., & Arvestad, L. (2016). Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics*, 32, 1925–1932. https://doi.org/10.1093/bioinformatics/btw064

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, 34, 2996–3005. https://doi.org/10.1093/molbev/msx209

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 11, 31.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using naïve protein interaction networks, integrated over the tree of life. *Current Biology*, 18, 283–334. https://doi.org/10.1016/j.cub.2007.06.070

Sahlin, K., Chikhi, R., & Arvestad, L. (2016). Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics*, 32, 1925–1932. https://doi.org/10.1093/bioinformatics/btw064

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, 34, 2996–3005. https://doi.org/10.1093/molbev/msx209

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 11, 31.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using naïve and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24, 637–644. https://doi.org/10.1093/bioinformatics/btn13

Storz, J. F., Baze, M., Waite, J. L., Hoffmann, F. G., Opazo, J. C., & Hayes, J. P. (2007). Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. *Genetics*, 177, 481–500. https://doi.org/10.1534/genetics.107.078550

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... von Mering, C. (2015). STRING v10: Protein – protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43, 447–452. https://doi.org/10.1093/nar/gku1003

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 195, 585–595.

Takeuchi, O., & Akira, S. (2010). Pattern recognition receptors and inflammation. *Cell*, 140, 805–820. https://doi.org/10.1016/j.cell.2010.01.022

Tennessen, J. A., Thérèon, A., Marine, M., Yeh, J.-Y., Rognon, A., & Blouin, M. S. (2015). Hyperdiverse gene cluster in snail host conveys resistance to human schistosome parasites. *PLoS Genetics*, 11, e1005067. https://doi.org/10.1371/journal.pgen.1005067

The UniProt Consortium (2017). *uniprot*: The universal protein knowledgebase. *Nucleic Acids Research*, 45, D158–D169.

Thielenpoulos, A. N., Gonzalez-Quintal, R., Lawson, B. R., Koh, Y. T., Stern, M. E., Kono, D. H., ... Baccala, R. (2010). Sensors of the innate immune system: Their link to rheumatic diseases. *Nature Reviews Rheumatology*, 6, 146–156. https://doi.org/10.1038/nrrheum.2009.278

Thornton, K. R. (2007). The neutral coalescent process for recent gene duplications and copy-number variants. *Genetics*, 1000, 987–1000. https://doi.org/10.1534/genetics.107.074948

Tschirren, B., Andersson, M., Scherman, K., Westerdahl, H., & Råberg, L. (2012). Contrasting patterns of diversity and population differentiation at the innate immunity gene toll-like receptor 2 (tlr2) in two sympatric rodent species. *Evolution (N. Y)*, 66, 720–731. https://doi.org/10.1111/j.1558-5646.2011.01473.x

Unckless, R. L., Howick, V. M., Lazzaro, B. P., Unckless, R. L., Howick, V. M., & Lazzaro, B. P. (2016). Convergent balancing selection on an antimicrobial peptide in Drosophila. *Current Biology*, 26, 257–262. https://doi.org/10.1016/j.cub.2015.11.063

Wen, H., Miao, E. A., & Ting, J. P. Y. (2013). Mechanisms of NOD-like receptor-associated inflammasome activation. *Immunity*, 39, 432–441. https://doi.org/10.1016/j.immuni.2013.08.037

Wilson, D. E., Lacher, T. E., & Mittermeier, R. A. (2017). *Handbook of the mammals of the world*: Rodents II. Barcelona, Spain: Lynx Edicions.

Wlasiuk, G., & Nachman, M. W. (2010). Adaptation and constraint at Toll-like receptors in primates. *Molecular Biology and Evolution*, 27, 2172–2186. https://doi.org/10.1093/molbev/msq104

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., & Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32, 569–577. https://doi.org/10.1038/ng1202-569

Wright, S. I., & Charlesworth, B. (2004). The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics*, 1076, 1071–1076. https://doi.org/10.1534/genetics.104.026500

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., & Brauer, M. J. (2016). GMAP and GSnap for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. *Statistical Genomics: Methods and Protocols, Methods in Molecular Biology*, 1418, 283–334.

Zhang, J., & Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16, 409–420. https://doi.org/10.1038/nrg3950

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.