1 **Signatures of diversifying selection and convergence acting on passerine Toll-like**
2 **receptor 4 in an evolutionary context**

3
4 Tereza Králová^{1,2}, Tomáš Albrecht^{1,3}, Josef Bryja^{1,2}, David Hořák⁴, Arild Johnsen⁵, Jan T.
5 Lifjeld⁵, Marian Novotný⁶, Ondřej Sedláček⁴, Hana Velová³ and Michal Vinkler³

6
7 ¹The Czech Academy of Sciences, Institute of Vertebrate Biology, Květná 8, 603 65 Brno,
8 Czech Republic

9 ²Masaryk University, Department of Botany and Zoology, Faculty of Science, Kotlářská 2,
10 611 37 Brno, Czech Republic

11 ³Charles University, Faculty of Science, Department of Zoology, Viničná 7, 128 44 Prague,
12 Czech Republic

13 ⁴Charles University, Faculty of Science, Department of Ecology, Viničná 7, 128 44 Prague,
14 Czech Republic

15 ⁵Natural History Museum, University of Oslo, 0318 Oslo, Norway

16 ⁶Charles University, Faculty of Science, Department of Cell Biology, Viničná 7, 128 44
17 Prague, Czech Republic

18
19 **Correspondence**

20 Tereza Králová, The Czech Academy of Sciences, Institute of Vertebrate Biology,
21 Studenec 122, 675 02 Koněšín, Czech Republic, e-mail: kralova.tereza@mail.muni.cz

22
23 **Running title:** TLR4 evolution in passerines

24 **Keywords:** TLR, surface charge distribution, functional evolution, host-pathogen interaction,
25 TLR4 molecular phenotype

26

27 **Abstract**

28 Positive selection acting on Toll-like receptors (TLRs) has been recently investigated to
29 reveal evolutionary mechanisms of host-pathogen molecular co-adaptation. Much of this
30 research, however, has focused mainly on the identification of sites predicted to be under
31 positive selection, bringing little insight into the functional differences and similarities among
32 species and a limited understanding of convergent evolution in the innate immune molecules.
33 In this study we provide evidence of phenotypic variability in the avian TLR4 ligand-binding
34 region (LBR), the direct interface between host and pathogen molecular structures. We show
35 that 55 passerine species vary substantially in the distribution of electrostatic potential on the
36 surface of the receptor, and based on these distinct patterns we identified four species clusters.
37 Seven out of the 34 evolutionarily non-conservative and positively selected residues
38 correspond topologically to sites previously identified as being important for
39 lipopolysaccharide, lipid IVa or MD-2 binding. Five of these positions co-determine the
40 identity of the charge clusters. Groups of species that host related communities of pathogens
41 were predicted to cluster based on their TLR4 LBR charge. Despite some evidence for
42 convergence among taxa there were no clear associations between the TLR4 LBR charge
43 distribution and any of the general ecological characteristics compared (migration, latitudinal
44 distribution, and diet). Closely related species, however, mostly belonged to the same surface
45 charge cluster indicating that phylogenetic constraints are key determinants shaping TLR4

46 adaptive evolution. Our results suggest that host innate immune evolution is consistent with
47 Fahrenholz's rule on the co-speciation of hosts and their parasites.

48

49 **Introduction**

50 Immune genes frequently exhibit remarkable levels of adaptive variation as a result of
51 pathogen-mediated diversifying selection (Schmid-Hempel, 2011). In animal genomes, the
52 most conspicuous and well-studied adaptive immune genetic variation is undoubtedly found
53 in the Major Histocompatibility Complex (*MHC*; Hughes & Yeager, 1998). Research on the
54 *MHC*, however, has not yet provided a full understanding of the basic principles of host-
55 pathogen co-adaptation, for several possible reasons. Firstly, *MHC* variation is typically
56 highly complex, comprising in most species tens to hundreds of alleles encoded at a variable
57 number of loci (Pirotney & Oliver, 2006), making locus-specific research difficult in most
58 taxa. Secondly, conformational plasticity responsible for structurally flexible ligand binding
59 has been documented for some (though not all) *MHC* variants, further increasing the
60 complexity of the pathogen detection system (Koch et al., 2007). Finally, *MHC* is not a
61 pathogen-recognising receptor itself, but rather the detection system is dependent on T-cell
62 receptor clonal somatic variability, which is virtually unlimited and uncharacterizable based
63 on the sequencing methods currently available. Taken together, this model system may be
64 simply too complex to allow the resolving of some particular basic questions of molecular co-
65 adaptations between hosts and their pathogens. It has also been suggested that the *MHC* is not
66 sufficient for an understanding wildlife immunogenetics since approximately half of the
67 genetic variability for resistance to infection is attributable to non-*MHC* genes (Acevedo-
68 Whitehouse & Cunningham, 2006).

69 Considering their level of interspecific and intraspecific variability with potential functional
70 significance, innate immune receptors, such as Toll-like receptors (TLRs), may be convenient
71 models for describing the basic features of host-pathogen molecular co-evolution (Vinkler &
72 Albrecht, 2009). TLRs are transmembrane proteins that directly bind various host-derived
73 damage-associated molecular patterns and pathogen-derived microbe-associated molecular
74 patterns (MAMPs), both representing important danger signals to immunity (Akira, Uematsu,
75 & Takeuchi, 2006). From an evolutionary perspective, MAMPs are particularly interesting.
76 Typically, MAMPs are structures essential for the survival of broad ranges of potential
77 pathogens (bacteria, viruses, etc.). Although generally conservative, important variability has
78 been described in these structures across pathogen taxa (Maeshima & Fernandez, 2013;
79 Needham & Trent, 2013; Nijland, Hofland, & van Strijp, 2014; Wang, Quinn & Yan, 2015),
80 including adaptations allowing the avoidance of immune detection. In response to changeable
81 pathogen selective pressure, hosts may be predicted to adaptively modify their TLR
82 structures. Several independent lines of evidence using several tests of positive selection
83 support this assumption in phylogenetically diverse animals (Alcaide & Edwards, 2011;
84 Areal, Abrantes, & Esteves, 2011; Grueber, Wallis, & Jamieson, 2014; Wlasiuk & Nachman,
85 2010; Yilmaz, Shen, Adelson, Xavier, & Zhu, 2004).

86 The TLR protein family (typically comprising about a dozen members per vertebrate species)
87 is strictly conservative in its general structure. The TLR ligand-binding N-terminal
88 ectodomain (ECD) consists of several leucine-rich repeat motifs, resulting in its typical
89 horseshoe shape. The ECD is followed by a short transmembrane domain and a C-terminal
90 Toll/interleukin-1 receptor (TIR) cytosolic domain responsible for signal transduction
91 (Takeda, Kaisho, & Akira, 2003). Despite negative selection acting on most parts of the
92 molecule, evolutionarily variable regions interspace the conservative leucine-rich repeat

93 motifs in the ECD to provide variability to the ligand-binding surfaces. It has been shown that
94 the predominantly bacterial-sensing TLRs are more genetically polymorphic than the viral-
95 sensing TLRs, due to their ability to recognise several structurally different ligands and their
96 greater redundancy in bacterial recognition (Vinkler, Bainová, & Bryja, 2014; Wlasiuk &
97 Nachman, 2010; Yilmaz et al., 2004). TLR4 in particular binds a broad spectrum of
98 structurally diverse ligands including lipopolysaccharides (LPS). It exhibits high levels of
99 putatively adaptive variation, mainly in the ligand-binding region (LBR) that is located in the
100 central part of the ECD (Alcaide & Edwards, 2011; Areal et al., 2011; Fornůsková, Bryja,
101 Vinkler, Macholán, & Piálek, 2014; Grueber et al., 2014; Tong et al., 2015; Vinkler, Bryjová,
102 Albrecht, & Bryja, 2009; Vinkler et al., 2014; Wlasiuk & Nachman, 2010). TLR4 ligand
103 variation is increased by strain- or species-specific modifications such as acylation/de-
104 acylation, phosphorylation or hydroxylation of the LPS in Gram-negative bacteria (Needham
105 & Trent, 2013). Consequently, various forms of LPS can bind with different affinity to
106 different TLR4 variants, and it has been shown that even slight differences in TLR4 structure
107 can have dramatic effects on its signalling (Walsh et al., 2008). For example, it has been
108 suggested that at least two single point substitutions in human *TLR4* evolved in response to
109 distinct pathogen pressures in Africa and Eurasia (Ferwerda et al., 2007). Substitutions in
110 *TLR4* can have either structural effects, changing the shape of the LBR, or functional effects,
111 e.g. altering surface molecular features such as its charge (Walsh et al., 2008). For example, in
112 Galloanserae birds the electrostatic potential distribution on the surface of the TLR4 molecule
113 was found to be more significant than changes in tertiary structure (Vinkler et al., 2014).

114 Perching birds (order Passeriformes), with their high species richness and large ecological and
115 phylogenetic diversification (Jetz, Thomas, Joy, Hartmann, & Mooers, 2012), are an ideal
116 model taxon for research on host-pathogen molecular co-evolutionary dynamics. Some of the

most important ecological characteristics associated with parasite diversity in birds are latitudinal distribution, seasonal movements and diet (e.g. Hannon, Kinsella, Calhoun, Joseph, & Johnson, 2015; Schemske, Mittelbach, Cornell, Sobel, & Roy, 2009; Waldenström, Bensch, Kiboi, Hasselquist, & Ottosson, 2002). It has been shown that in several vertebrate species, the parasite load and disease incidence increase towards the equator, following differences in the extent of host-pathogen interactions with changing latitude (reviewed in Schemske et al., 2009). With respect to this fact, the migratory species inhabiting diverse habitats during their annual cycle expose themselves to increased parasite diversity compared to both temperate and tropical residents (Gutiérrez, Rakhimberdiev, Piersma, & Thieltges, 2017; Hannon et al., 2015; Leung & Koprivnikar, 2016; Waldenström et al., 2002). Dietary preferences for more diverse food may also increase the parasite richness with trophically transmitted parasite species, i.e. carnivorous and omnivorous birds are expected to host a wider range of parasites than herbivores. However, evidence supporting this hypothesis has only been obtained through studies focusing on macroparasites like helminths (Hannon et al., 2015; Santoro, Kinsella, Galiero, Uberti, & Aznar, 2012).

The theory of adaptive evolution driven by pathogen selective pressure predicts that species exposed to similar pathogens should show similarities in the binding properties of their immune receptors. Since microbial-richness species-specific differences are scarcely known in birds (but see Kropáčková et al., 2017), analyses of host associated ecological variables together with their neutral phylogeny may reveal factors driving molecular convergence in immune receptors. Furthermore, the consistency of the immune receptor phenotype clustering with the neutral phylogeny of the host species may provide support for the co-diversification of parasites with their hosts, as hypothesised by Fahrenholz's rule (Eichler, 1948). Closely related species of hosts are frequently colonized by closely related communities of microbiota

(Kropáčková et al., 2017). As a result of their intimate associations, many pathogens and their hosts continually co-adapt in phylogenetically corresponding patterns.

Combining sequence analyses with protein structural modelling, we applied a novel methodological approach to explore the molecular evolution of the most functionally relevant part of the TLR4 LBR. Patterns of TLR4 LBR adaptive evolution were investigated in a carefully selected dataset of 55 species representing both tropical and temperate passerine birds with diverse ecological characteristics. The main aim of this study was to test the adaptive value of sequence variation encoding molecular phenotypic traits attributable to microbiota-determining traits. More specifically, we compared the effects of ecological traits, such as migration, latitudinal distribution and diet, on TLR4 LBR variation, while controlling for phylogenetic relatedness. Our approach should be generally applicable to comparative analyses of functional consequences at the protein level arising from interspecific nucleotide sequence variation shaped by natural selection.

Material and methods

Target region of the TLR4 and studied taxa

In this study we performed targeted amplification of the whole LBR-coding region (681 bp) in the *TLR4* gene, which encodes for all putative LPS-binding sites. For polymerase chain reaction (PCR), a new pair of *TLR4* LBR-specific primers was designed based on an alignment of previously known sequences of passerine *TLR4*s. These primers, passerTLR4fw-764 (5'-GTTTGCAGGTCAGCAGACTC-3') and passerTLR4rev-1444 (5'-GCAAAGAGTTCAAGCCACAAAA-3'), successfully amplified the required gene region in all 46 *de novo* sequenced passerine species from the Czech Republic, Norway, Cameroon and

Nigeria selected for this study (samples obtained from European biobanks as scientific loans, see Table S1 in the Supplement; for PCR conditions see below). Several of the analysed species form phylogenetically related pairs inhabiting distinct geographical zones. The newly produced sequence data were added to nine currently available passerine *TLR4* sequences freely available in genome data deposited at NCBI GenBank. The complete list of species including the origin of the representative samples and sequences is provided in Table S1. Furthermore, the sequence of the chicken *TLR4* (*Gallus gallus*, *Galgal*; GenBank acc. no. AY064697.1) was used as an outgroup in both phylogenetic and structural analyses, and the crystal structure of the mouse *TLR4* (*Mus musculus*; GenBank acc. no. NM_021297.2; PDB ID 2Z64) was used in the structural analysis.

TLR4 sequencing

For the sequencing step, special fusion primers were synthesized that consisted of the *TLR4* LBR-specific primer (see above), a unique 12bp long barcode (Caporaso et al., 2012), a key sequence (TCAG) and a specific 21bp-long next-generation sequencing adaptor (Roche, Basel, Switzerland). Blood samples were stored frozen (-20°C) in 96% ethanol or refrigerated (4°C) in Queens Lysis Buffer (QLB) until the DNA extraction using DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. PCR was first performed with only the *TLR4* LBR-specific primers to confirm successful amplification, and then with fusion primers required for the sequencing step on the 454-pyrosequencing platform, with a final PCR product length of 755 bp. The PCR was mixed using a Multiplex PCR Kit (QIAGEN) under the following conditions: 0.3 µM of each primer, 1× Multiplex PCR Master Mix (including HotStarTaq® Polymerase and 3mM Mg²⁺), 1 µl of purified DNA

with the addition of deionised water up to 10 µl. The reaction was performed on a Mastercycler ep (Eppendorf, Hamburg, Germany) with initial heat activation at 94°C (15 min) followed by 30 cycles of denaturation at 94°C (30 s), annealing at 56°C (40 s) and extension at 72°C (1 min), ending with a final extension step at 72°C (10 min). All amplified samples were then purified using magnetic beads (Agencourt AMPure XP, Beckman Coulter Life Sciences, Brea, CA, USA), pooled in equimolar amounts and sequenced with an FLX+ instrument (Roche). Application of the next-generation sequencing (NGS) approach allowed us to avoid the costly cloning step in our effort to obtain single-haplotype sequences. Each individual was identified according to its unique combination of the forward and reverse barcode and the data were de-multiplexed using the R packages seqinr, Biostrings and ShortRead in the open statistical software and programming language R v. 3.1.2 (R Core Team, 2015). All nucleic acid sequences were aligned, edited and translated into protein sequences using Geneious v. 9.1.5. (<http://www.geneious.com>; Kearse et al., 2012). No interspecifically shared variation was observed among the species in our dataset. For methodological reasons, the final dataset consisted of one randomly-selected haplotype sequence per species (GenBank acc. nos. MG209180-MG209225). Sequences of the Hooded crow (*Corvus cornix cornix*) and Atlantic canary (*Serinus canaria*) had to be excluded from further analyses because of their identity with sequences of the American crow (*Corvus brachyrhynchos*) and European serin (*Serinus serinus*), respectively. Thus, the final dataset consisted of haplotype sequences from 53 passerine species covering 28 families across the passerine phylogeny.

Analysis of functional variation

A basic description of the structural and physical features of the sequences was performed in Geneious. The molecular weight and charge of the linear sequences at pH=7 were computed using the web-based tool LRRfinder (www.lrrfinder.com), which compares input protein sequence to those already present in a Protein Data Bank (PDB) database. The protein structure of the whole extracellular domain of TLR4 in the Zebra finch (*Taeniopygia guttata*) was modelled using the I-TASSER server (Yang et al., 2015). The LBR of this protein structure served as a PDB model to identify regions on the surface of the molecule that may be potentially involved in host-pathogen co-evolution based on their level of evolutionary non-conservatism using the online tool ConSurf (<http://consurf.tau.ac.il>; Glaser et al., 2003). The advantage of ConSurf compared to other similar tools is that it takes into account the phylogenetic relationships among the sequences, which is especially important for the interspecific comparison of species differing in their levels of relatedness. For this analysis, the alignment of the 53 unique amino acid sequences (Fig. S1) was uploaded to ConSurf along with the phylogenetic species tree extracted from <http://birdtree.org/> (Jetz et al., 2012; see below for more details). We applied a Bayesian calculation method with an LG evolutionary substitution model for the analysis (Le & Gascuel, 2008). Since in immune receptors any host adaptations to diversified pathogen structures can occur only at the phylogenetically non-conservative sites, we consider evolutionary non-conservatism as evidence strengthening indication for positive selection.

According to recent research in Galloanserae birds (Vinkler et al., 2014), one of the most important variable features of TLRs is the electrostatic potential distribution on the molecular surface, so we performed a closer examination of this property. To evaluate the anticipated functional differences in TLR4 LBRs between selected passerine species, we predicted the tertiary structure of each species-specific protein using homology modelling tool Modeller

v. 9.14 (Eswar et al., 2006; pairwise sequence alignment performed in Geneious), with mouse TLR4 as a template with known crystal structure (Kim et al., 2007). We generated two models for each sequence, and the one with the lower DOPEScore was selected for further analysis. The quality of the selected model for each species was evaluated using the ModFOLD Model Quality Assessment Server v. 4.0 and 6.0 (McGuffin, Buenavista, & Roche, 2013) and DALI pairwise comparison tool computing similarity of two proteins based on their structural alignment (DaliLite v. 3, http://ekhidna.biocenter.helsinki.fi/dali_lite/start; Hasegawa & Holm, 2009).

All structures (including the chicken as an outgroup) were superimposed in UCSF Chimera v. 1.10.1 using the “Match->Align” function (<http://www.cgl.ucsf.edu/chimera/>; Pettersen et al., 2004) and the whole set was uploaded to PIPSA (Protein Interaction Property Similarity Analysis; Richter, Wenzel, Stein, Gabdoulline, & Wade, 2008) to acquire a matrix of species with pairwise comparisons of their surface electrostatic potential distances. The visualization of the protein surface charge variation in 3D space was displayed in UCSF Chimera using input files generated in APBS v. 1.4.1. (Adaptive Poisson-Boltzmann Solver; Baker, Sept, Joseph, Holst, & McCammon, 2001), which solves the equations of continuum electrostatics using structures adapted for it in PDB2PQR v. 2.0.0. (Dolinsky et al., 2007; Dolinsky, Nielsen, McCammon, & Baker, 2004) based on the PARSE force-field (PDB2PQR Server, http://nbc-222.ucsd.edu/pdb2pqr_2.0.0/). To estimate the surface receptiveness of individual residues in the TLR4 LBR, we calculated the relative solvent accessibility (RSA) from the residue’s solvent accessibility surface area (ASA) computed from the PDB models of four selected species using the web server xssp (<http://www.cmbi.ru.nl/xssp/>; based on Kabsch & Sander, 1983) divided by a corresponding maximum possible ASA for the given amino acid (Tien, Meyer, Sydykova, Spielman, & Wilke, 2013).

Finally, a cluster analysis was performed in R v. 3.4.1. (R Core Team, 2017) using RStudio v. 1.0.153 (RStudio Team, 2015) based on the matrix of pairwise electrostatic distances from PIPSA and using the `upgma` function (package `phangorn` v. 2.2.0, `method` = `ward`; Schliep, 2011). This phenetic tree (dendrogram) of surface charge distribution was then visually compared with a neutral phylogenetic species tree using R packages `ape` v. 4.1 (Paradis, Claude, & Strimmer, 2004), `phangorn` v. 2.2.0 (Schliep, 2011) and `dendextend` v. 1.5.2 (Galili, 2015). Both dendrograms were also compared statistically by the “entanglement” function (`E`) in the package `dendextend`, resulting in a value between 1 (different dendrograms) and 0 (identical dendrograms). The phylogenetic tree was extracted from the global phylogeny of birds using a web-based tool at <http://birdtree.org/> (Jetz et al., 2012; Fig. S2). We generated 1000 random trees (Stage2 Mayr All Hackett backbone), which were then summarized into a single consensus tree with `TreeAnnotator` v. 1.8.2 implemented in BEAST software package v. 1.8.2 (Drummond, Suchard, Xie, & Rambaut, 2012). The *TLR4* gene tree was constructed using the maximum likelihood principle in a web-based phylogeny software `PhyML` 3.0 (Guindon et al., 2010) and it mostly resembled the phylogenetic species tree (data not shown). All trees were visualized using `FigTree` v. 1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Selection analysis

First, the possibility of recombination influencing evolution of the *TLR4* LBR in the studied species was excluded based on the results of GARD analysis (Genetic Algorithms for Recombination Detection, <http://www.datamonkey.org>; Kosakovsky Pond, Posada, Gravenor, Woelk, & Frost, 2006). Then, we applied four standard codon-based methods to find

signatures of diversifying selection acting on the *TLR4* LBR in passerines: (1) FUBAR - A Fast, Unconstrained Bayesian AppRoximation for inferring selection (Murrell et al., 2013), (2) MEME - Mixed Effects Model of Episodic Diversifying Selection (Murrell et al., 2012), (3) SLAC - Single Likelihood Ancestor Counting (Pond & Frost, 2005b) – all three available on the Datamonkey webserver (<http://www.datamonkey.org>; Delpont, Poon, Frost, & Kosakovsky Pond, 2010; Pond & Frost, 2005a), and (4) the codon-based substitution model (codeml) M8a (null model, $\omega=1$) in comparison with M8 (an alternative model, $\beta\omega$) implemented in PAML v. 4.8 (Phylogenetic Analysis by Maximum Likelihood; Yang, 2007). Using several alternative selection tests allowed us to evaluate the selective pressure occurring in our diverse dataset with slightly different assumptions and constraints. The SLAC model combines maximum-likelihood (ML) and counting approaches to find nonsynonymous (dN) and synonymous (dS) substitution rates based on the reconstructed ancestral sequences. This is different from the FUBAR method that infers dN and dS substitution rates using Bayesian approach with the assumption of constant selective pressure along the entire phylogeny, while MEME allows for selection under the proportion of branches to detect sites evolving under episodic selection. Evolutionary signatures in the input sequences were also searched using the identification of conserved/variable physicochemical properties at individual amino acid sites in PRIME (PRoperty Informed Models of Evolution) running on the Datamonkey server (Delpont et al., 2010; Pond & Frost, 2005a). We applied a set of five physicochemical properties reflecting the polarity, secondary structure, molecular volume, refractivity/heat capacity and electrostatic charge of the amino acids (Atchley, Zhao, Fernandes, & Drücke, 2005). The location and variability of positively selected sites were then compared with the functional sites of the *TLR4* LBR found in previous vertebrate studies using crystallographic methods (Table S2; Kim et al., 2007; Ohto, Fukase, Miyake, &

Shimizu, 2012; Park et al., 2009). The variability of the positively selected amino acids with indications of diversity in their physicochemical properties was visualized using WebLogo v. 3 (Crooks, Hon, Chandonia, & Brenner, 2004; <http://weblogo.threeplusone.com/>). Throughout the study, sites are numbered according to the Zebra finch *TLR4* coding sequence (GenBank acc. no. FJ695612).

Identifying convergence in structural adaptive evolution

The bioinformatic approach described above allows searching for evidence on evolutionary convergence even in phylogenetically distantly related taxa. In an attempt to interpret the convergence in structural adaptive evolution of the TLR4 LBR, we compared our surface-charge clustering with several broad ecological characteristics of the species. We focused on basic ecological factors and characterised the studied species according to: (1) migration and latitudinal distribution (hereafter called latitudinal dispersion: temperate resident, temperate partial or short-distance migrant, temperate long-distance migrant and tropical resident species) and (2) diet (predominantly plant, animal, or mixed diet). These traits represent intersection between those characteristics to which data are presently available for all the analysed species (including the tropical species) and the traits that were shown in earlier studies to exhibit associations with parasite abundance or prevalence (Gutiérrez et al., 2017; Hannon et al., 2015; Leung & Koprivnikar, 2016; Santoro et al., 2012; Waldenström et al., 2002). The categories of each characteristic were mapped onto the dendrogram of electrostatic potentials using a colour code to compare TLR4 phenotypic clustering with the ecological variable distribution. To examine the significance of associations between surface-charge clusters and selected ecological categories, we performed the Fisher's exact test for count data

in R v. 3.4.1. (R Core Team, 2017). In addition, we applied phylogenetic generalized least squares (PGLS) analysis which corrects for the relatedness of taxa using the electrostatic distances of individual species from the chicken outgroup (caper package v. 0.5.2 in R; Orme et al., 2013). Using the maximum likelihood method, we also estimated the phylogenetic scaling parameter lambda (λ) and tested its value against $\lambda = 0$ (indicating phylogenetic independence) and $\lambda = 1$ (phylogenetic dependence). The complex PGLS model (with both predictors included), models containing only one predictor and the null model were then compared using the Second-order Akaike information criterion (AICc) to identify the best-fit model (i.e. the model with the lowest AICc value; MuMIn package v. 1.15.6; Barton, 2016).

Results

All 53 unique nucleotide sequences of the *TLR4* LBR encode polypeptides without any stop codons, and with only one single codon deletion in the Medium ground finch (*Geospiza fortis*) (Table S3, Fig. S1). The mean interspecific identity to the Zebra finch (*Taeniopygia guttata*) amino acid sequence was 84.3% (range 73.0-94.2%). Interestingly, this sequence variation affected the charge variation of the linearized polypeptide molecules (values from -9.1 to 8.0; Table S3). We also found non-synonymous substitutions at several LPS-binding positions (267, 344, 368, 392, 396; Table S2) identified in previous studies (Kim et al., 2007; Ohto et al., 2012; Park et al., 2009).

No interspecific variation in TLR4 LBR tertiary structures

To assess the effect of sequence variation on the protein structural properties, we modelled the tertiary structures of the TLR4 LBRs. The accuracy of our models was supported with ModFOLD (quality scores of all models with $p \ll 0.001$). The interspecific pairwise comparison performed in DALI showed only low variability in TLR4 LBR structures, with RMSD values below 1.0 Å.

Distinct patterns of TLR4 LBR surface charge distribution in passerines

Contrary to the conservative tertiary protein structure, the distribution of electrostatic potential on the polypeptide surfaces varied substantially among the species (Fig. S3). The electrostatic distance (calculated as the absolute difference between the surface electrostatic potential values of two different molecules) ranged from 0.307 (*Phoenicurus phoenicurus* vs. *Ficedula albicollis*) to 1.284 (*Erithacus rubecula* vs. *Certhia familiaris*) along the whole TLR4 LBR (Fig. 1). Based on PIPSA clustering (Fig. S4), we divided the whole dataset into four distinct groups differing in their patterns of surface charge distribution (Fig. 1, Table S3). The PGLS analysis with implemented phylogeny correction showed that PIPSA clustering corresponded to the charges of the linearized polypeptides calculated in LRRfinder (pgls, $F = 9.26$, $p \ll 0.001$; this pattern was less obvious without phylogeny correction, see Table S3, Fig. S5). We found no clear distinctions between surface charge clusters with respect to the physicochemical features of the predicted LPS-binding sites.

Positive selection shaped TLR4 LBR surface charge clustering

370 In order to show the evolutionary significance of the variation in TLR4 LBR physicochemical
371 properties, we predicted the key amino acid positions under positive selection in our dataset.
372 We assumed that sites under positive (diversifying) selection should be also the least
373 evolutionarily conservative sites. Using ConSurf, 18 sites were identified as evolutionarily
374 non-conservative in our dataset, four of which lie in close topological proximity to the
375 potentially important LPS-binding sites (Fig. S6, Table S4). Furthermore, based on five
376 independent approaches (FUBAR, MEME, SLAC, PAML, PRIME) we identified a total of
377 32 amino acid sites predicted to be under positive/diversifying selection (Fig. 2, Fig. S7, for
378 details see Table S5 and Table S6). Sixteen sites out of these 32 were also detected in
379 ConSurf as evolutionarily non-conservative and six sites corresponded to the amino acid
380 positions defined as potentially important for LPS or lipid IVa binding. Altogether we found
381 34 evolutionarily non-conservative and positively selected residues in the studied region
382 (Table S7). To avoid false positive results, before proceeding with further analysis we applied
383 a criterion that only those positions identified by at least three independent approaches
384 (including the ConSurf analysis) would be considered as reliable positively selected sites.
385 Only 11 sites fulfilled this criterion: positions 267, 269, 270, 272, 298, 308, 331, 344, 351,
386 374, and 402 (for details and consistency with other vertebrate studies see Table S7). On these
387 11 positively selected sites, we found distinct patterns in amino acid characteristics
388 differentiating surface charge cluster A from clusters B, C and D (Table 1). Most of the
389 evaluated positions, however, were variable within individual clusters, suggesting functional
390 diversification in various taxa (Table S8, Fig. 3). Out of 20 positively selected sites identified
391 by MEME, which tests for the clade-specific selection, three sites clearly differed between
392 charge clusters (351, 366, and 374). While positions 351 and 374 were found by multiple
393 selection tests, position 366 was recognized only by MEME suggesting its association with

branch-specific positive selection acting on charge cluster A. PRIME results showed that the physicochemical properties that are variable among passerine birds were connected with the secondary structure factor (265, 298), refractivity/heat capacity (270, 371) and charge/isoelectric point (267, 374) (Table S6). All these sites except two were also identified by at least two other methods that either identify putatively positively selected sites based on Ka/Ks approaches or estimate evolutionary non-conservatism (for the sites 265 and 371 that do not match this criterion further evidence supporting their functional relevance is shown in Table S7). Interestingly, positions 267 and 374 seem to divide the clusters into two distinct groups: A+D and B+C. While clusters B+C are polar uncharged at position 267 and positive at position 374, clusters A+D are variable at 267 and possess mainly negative residues at 374, (Fig. 3, Table S8). Hence, despite the indicated relatedness of the deep branches of the surface charge dendrogram (Fig. S5), judging by the physicochemical features at the key sites, cluster A is closer to cluster D, and cluster B to C. The importance of these positions is also supported by the fact that site 267 is directly involved in LPS and MD-2 binding and site 374 lies in close proximity to the lipid IVa recognition site (Table S2). Cluster A is then differentiated from cluster D by the properties of position 351 (polar vs. hydrophobic) and cluster B from cluster C mainly by the properties of position 298 (where B is negatively charged, while C is mostly not; Fig. 3). Moreover, according to the relative solvent accessibility analysis, the amino acids on most of the identified positively selected sites tend to be exposed on the surface of the molecule, allowing for direct interaction with the ligand (Table S9).

Surface physicochemical features reflect phylogeny more than ecology of the species

417 Based on the theory of adaptive evolution driven by the pathogen-mediated natural selection,
 418 we predicted that the TLR4 LBR clusters A-D would group passerine species having distinct
 419 pathogen spectra. Therefore, we hypothesized that the clusters would be related to traits like
 420 latitudinal dispersion or diet. In our dataset there was no association of surface-charge clusters
 421 with latitudinal dispersion (Fisher's Exact Test, $p = 0.527$), but diet might have had an effect
 422 on the physicochemical features of TLR4 LBR molecules (Fisher's Exact Test, $p = 0.022$),
 423 which can also be seen on the projection of these characteristics onto a dendrogram of the
 424 TLR4 LBR surface charge distribution (Fig. S8). A visual comparison of neutral phylogeny
 425 with the surface charge clustering indicates, however, that closely related species mostly
 426 belong to the same surface charge group (Fig. 4). In accordance with this, the PGLS analysis
 427 showed no influence of either latitudinal dispersion or diet on the TLR4 LBR electrostatic
 428 distance of individual species from the chicken outgroup (the complex model with both
 429 predictors involved; $F_{5,46} = 0.833$, $p = 0.533$, $\lambda = 0.41^{0.61; 0.01}$, $AICc = -86.85$). In concordance
 430 with this finding, the null model had a lower $AICc$ than the full model ($AICc = -94.93$).
 431 Models containing only diet or only latitudinal dispersion were also non-significantly
 432 different from the null model ($F_{2,49} = 1.895$, $p = 0.161$, $\lambda = 0.42^{0.52; 0.01}$, $AICc = -93.64$ and
 433 $F_{3,48} = 0.164$, $p = 0.92$, $\lambda = 0.66^{0.04; 0.03}$, $AICc = -88.69$, respectively; also see Table S10).
 434 These results imply that phylogeny is most likely the key factor influencing the charge of the
 435 TLR4 LBR. On the one hand, the well-distinguished functional cluster A is also the only
 436 cluster that is composed of species belonging to the Muscicapoidea monophyletic clade
 437 (Fig. S9). On the other hand, the least well-defined and relatively heterogeneous cluster D
 438 consists of phylogenetically unrelated species, including domestic chicken that was used as an
 439 outgroup. In addition, the statistical comparison of both dendrograms showed an incomplete
 440 effect of phylogeny on the surface charge clustering ($E = 0.578$).

441

442 **Discussion**

443 Despite the earlier view of TLRs as highly conservative proteins (Medzhitov & Janeway,
444 1998), an increasing body of evidence has supported the idea that certain parts of these innate
445 immune receptors are evolutionarily plastic and presumably shaped by pathogen-mediated
446 adaptive evolution (*e.g.* Smirnova, Poltorak, Chan, McBride, & Beutler, 2000; Wang, Zhang,
447 Liu, Zhao, & Yin, 2016). Our analysis of 53 unique passerine TLR4 LBRs, covering species
448 inhabiting both the temperate and tropical climatic zones, revealed relatively high levels of
449 sequence variability in *TLR4*, including substitutions at the eleven positions previously shown
450 to bind LPS. Given that six out of these sites were at least in close topological proximity to
451 sites also identified as positively selected, our data support the notion of this genetic variation
452 having functional significance. Among the 34 sites detected in the passerine TLR4 LBR as
453 positively selected based on the multiple criteria adopted in this study, ten have also been
454 identified as positively selected in other avian studies (Alcaide & Edwards, 2011; Grueber et
455 al., 2014; Vinkler et al., 2009; Vinkler et al., 2014), and another eleven were consistent with
456 sites previously detected as being under positive selection in both birds and mammals (Areal
457 et al., 2011; Fornůšková et al., 2013; Grueber et al., 2014; Vinkler et al., 2009; Vinkler et al.,
458 2014; Wlasiuk & Nachman, 2010; for the consensual sites see Table S7). This may suggest
459 convergence in TLR4 molecular evolution among various vertebrate taxa.

460 Interestingly, the presumably functional sequence variability in *TLR4* is not reflected in the
461 shapes of the predicted tertiary structures of the TLR4 LBR polypeptides, which were found
462 to be invariant. Although we are aware that our models are only predictions, given the
463 structural conservatism in vertebrate TLRs and the high accuracy of our models, we consider

464 them to be trustworthy. Together with similar results previously reported for non-passerine
465 birds and mammals (Fornůsková et al., 2013; Vinkler et al., 2014), our data thus indicate a
466 high structural uniformity in TLRs, which appears to be an evolutionarily stable feature. In
467 contrast (but again similar to non-passerine birds and mammals; Fornůsková et al., 2013;
468 Vinkler et al., 2014), TLR4 LBR surface charge distributions showed very distinct patterns
469 across passerines, suggesting the importance of this physicochemical feature for ligand
470 binding. We were able to identify four basic clusters of surface charge distribution patterns in
471 passerines. These surface charge clusters were not identified by amino acid substitutions at
472 previously defined mammalian LPS-binding sites (Kim et al., 2007; Ohto et al., 2012; Park et
473 al., 2009), but were clearly recognisable by variation at 5 of the 11 well-supported positively
474 selected positions (267, 298, 308, 351, and 374) that are candidate for functional effects.

475 Among these candidate sites, position 374 in particular distinguishes the surface charge
476 clusters A and D from the clusters B and C. It is positioned in close topological proximity to
477 residue 375, which has already been shown to be crucial for lipid IVa recognition in mammals
478 (Ohto et al., 2012). Therefore, substitutions at this position altering the molecular charge from
479 positive to negative may profoundly affect the recognition of diverse pathogens. Given that
480 analogous positive selection has been shown in Galloanserae birds (Vinkler et al., 2014) and
481 cyprinid fish (Tong et al., 2015), it appears likely that charge-changing substitutions at
482 position 267, which is directly involved in LPS and MD-2 binding, may also be of particular
483 evolutionary significance. Furthermore, the B-cluster-determining position 298, which was
484 predicted to affect the receptor's secondary structure, has been identified as a positively
485 selected site in mammals and birds (Areal et al., 2011; Vinkler et al., 2009; Wlasiuk &
486 Nachman, 2010). Similarly, cluster A is differentiated from other charge clusters based on the
487 properties of position 351, a hydrophobic-altering site that has previously been identified as

positively selected in other avian studies (Alcaide & Edwards, 2011; Grueber et al., 2014), along with an adjacent site in various mammals (Areal et al., 2011; Fornůsková et al., 2013; Wlasiuk & Nachman, 2010). The interspecific differences in charge on the receptor surface may reflect the result of long lasting co-evolutionary processes between hosts and their pathogens (Woolhouse, Webster, Domingo, Charlesworth, & Levin, 2002). It appears that also residues other than those previously identified in mammals may play an important role in shaping TLR4 LBR binding specificity in birds, which is consistent with empirical results published earlier (Leveque et al., 2003). We recommend that the sites 267, 298, 308, 351, 374 be subject to closer investigation in further and especially functional studies.

Given the lack of published data on avian pathogens present in the environments inhabited by the wild passerine host species investigated here (Poulin & Morand, 2000), our effort to explain the observed patterns of variation in *TLR4* focused on possible connections between the TLR4 LBR surface charge clusters and basic ecological characteristics that should be related to the pathogen pressure. However, we were unable to find any associations between TLR4 surface charge clustering and the ecology of these species. Although both studied parameters have been reported to be associated with parasite richness in previous studies (e.g. Hannon et al., 2015; Schemske et al., 2009; Waldenström et al., 2002) and TLR4 is also capable of detecting parasite ligands from tropical malaria (Eriksson, Sampaio, & Schofield, 2014), the possible effects of latitude and migration are not mirrored in the interspecific TLR4 surface charge distribution. The apparent link of surface charge with the type of diet reflects only the misbalance in numbers of species in each group and ecological similarities between phylogenetically closely-related species. Therefore, our analysis suggests that TLR4 surface physicochemical features reflect the phylogeny of species more than their ecology. This is mostly consistent with the results of a recent study in passerines reporting a high association

of species phylogeny with their gut microbiota (Kropáčková et al., 2017). According to Fahrenholz's rule (Eichler, 1948), phylogenetically related host species may share related species of pathogens. Thus, a similar selection mode can be predicted in individual clades of related species, allowing for clade-specific molecular evolutionary interactions between the hosts and their parasites. As a result, host compatibility rather than ecological factors may determine the avian host susceptibility to pathogens (Medeiros, Hamer, & Ricklefs, 2013). Nevertheless, Kropáčková et al. (2017) also found significant (though weaker) effects of host ecology and geography on gut microbiota content. In addition, in TLR4 the consistency between the surface charge dendrogram and passerine neutral phylogeny is incomplete, indicating a contribution of some non-phylogenetic factors to the formation of TLR4 molecular properties. The TLR4 LBR charge cluster A was the only well-distinguished cluster formed by species belonging to the monophyletic clade of Muscipoidea. In contrast, cluster D consists of a heterogeneous spectrum of unrelated species including chicken, which suggests convergence resulting from sharing pathogens with common immunogenic profiles. A very similar result was obtained in MHC class I investigated across distantly related passerine species (Follin et al., 2013). At a larger evolutionary scale, a similar pattern was documented by our previous findings in TLR4 showing a surface charge similarity of chicken LBR to mouse LBR that is distinct from human LBR charge (Vinkler et al., 2014). The adaptive significance of this finding is highlighted by functional test results documenting similar ligand-specificity of the chicken and mouse recombinant proteins when experimentally expressed *in-vitro* in model cells (Keestra & Putten, 2008). Therefore, in our opinion, it may be difficult to distinguish ecological signals from phylogenetic signals in traits that may be involved in host pathogen co-speciation according to the Fahrenholz's rule, since

these traits adaptively evolve in response to parasite ecology, yet in the same time they generally follow the host phylogeny.

To conclude, our study provides an insight into the potential functionality of evolutionary adaptations in TLR4 that may contribute to understanding of the interaction patterns in the presumed host-parasite molecular co-evolution. We found no associations between TLR4 functional variation and ecological characteristics. Further research on selective pressures acting on the innate immune receptors would benefit from linking immunogenetic population data with a wider set of ecological parameters including data on microbial communities. Since the interspecific differences in charge on the receptor surface were found mainly around positively selected sites, indicating functional significance, our results point to the value of *in silico* structural analyses. We suggest that the methodological approach adopted in this study can have general applicability for studies of the forces driving host-pathogen co-evolution based on sequence data.

Acknowledgements

We are grateful to David Hardekopf and anonymous reviewers for helpful comments on an earlier version of the manuscript. We thank Jiří Reif for providing the samples for analysis, as well as Anna Bryjová, Dagmar Čížková and Jana Nekvindová for their appreciated help in the lab. We also thank Jakub Kreisinger for his help with bioinformatics. This study was supported through Czech Science Foundation Projects P506/15-11782S and GA17-24782S, the Grant Agency of Masaryk University (Grant no. MUNI/A/0888/2013), the Grant Agency of Charles University (Grant no. 540214), through Institutional Research Support (Grant SVV 260 434 / 2018); M.V. was further supported by Charles University Research Centre program

No. 204069 and project PRIMUS/17/SCI/12, and M.N. by ELIXIR CZ research infrastructure project (MEYS Grant No: LM2015047).

References

- Acevedo-Whitehouse, K., & Cunningham, A. (2006). Is MHC enough for understanding wildlife immunogenetics? *Trends in Ecology & Evolution*, 21(8), 433–438. doi:10.1016/j.tree.2006.05.010
- Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen Recognition and Innate Immunity. *Cell*, 124(4), 783–801. doi:10.1016/j.cell.2006.02.015
- Alcaide, M., & Edwards, S. V. (2011). Molecular Evolution of the Toll-Like Receptor Multigene Family in Birds. *Molecular Biology and Evolution*, 28, 1703–1715. doi:10.1093/molbev/msq351
- Areal, H., Abrantes, J., & Esteves, P. J. (2011). Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. *BMC Evolutionary Biology*, 11(1), 368. doi:10.1186/1471-2148-11-368
- Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18), 6395–6400. doi:10.1073/pnas.0408677102
- Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18), 10037–10041. doi:10.1073/pnas.181342398
- Barton, K. (2016). *MuMIn: Multi-Model Inference*. Retrieved from <https://CRAN.R-project.org/package=MuMIn>
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., ... Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, 6(8), 1621–1624. doi:10.1038/ismej.2012.8
- Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6), 1188–1190. doi:10.1101/gr.849004
- Delpont, W., Poon, A. F. Y., Frost, S. D. W., & Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26(19), 2455–2457. doi:10.1093/bioinformatics/btq429
- Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., & Baker, N. A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(suppl 2), W522–W525. doi:10.1093/nar/gkm276
- Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(suppl 2), W665–W667. doi:10.1093/nar/gkh381
- Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8), 1969–1973. doi:10.1093/molbev/mss075

- Eichler, W. (1948). XLI.—Some rules in Ectoparasitism. *Annals and Magazine of Natural History*, 1(8), 588–598. doi:10.1080/00222934808653932
- Eriksson, E. M., Sampaio, N. G., & Schofield, L. (2014). Toll-Like Receptors and Malaria – Sensing and Susceptibility. *Journal of Tropical Diseases*, 02(01). doi:10.4172/2329-891X.1000126
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., ... Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]*, 0 5, Unit-5.6. doi:10.1002/0471250953.bi0506s15
- Ferwerda, B., McCall, M. B. B., Alonso, S., Giamarellos-Bourboulis, E. J., Mouktaroudi, M., Izagirre, N., ... Netea, M. G. (2007). TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans. *Proceedings of the National Academy of Sciences*, 104(42), 16645–16650. doi:10.1073/pnas.0704828104
- Follin, E., Karlsson, M., Lundegaard, C., Nielsen, M., Wallin, S., Paulsson, K., & Westerdahl, H. (2013). In silico peptide-binding predictions of passerine MHC class I reveal similarities across distantly related species, suggesting convergence on the level of protein function. *Immunogenetics*, 65(4), 299–311. doi:10.1007/s00251-012-0676-3
- Fornůsková, A., Bryja, J., Vinkler, M., Macholán, M., & Piálek, J. (2014). Contrasting patterns of polymorphism and selection in bacterial-sensing toll-like receptor 4 in two house mouse subspecies. *Ecology and Evolution*, 4(14), 2931–2944. doi:10.1002/ece3.1137
- Fornůsková, A., Vinkler, M., Pagès, M., Galan, M., Jouselin, E., Cerqueira, F., ... Cosson, J.-F. (2013). Contrasted evolutionary histories of two Toll-like receptors (Tlr4 and Tlr7) in wild rodents (MURINAE). *BMC Evolutionary Biology*, 13(1), 194.
- Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22), 3718–3720. doi:10.1093/bioinformatics/btv428
- Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., & Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics*, 19(1), 163–164. doi:10.1093/bioinformatics/19.1.163
- Grueber, C. E., Wallis, G. P., & Jamieson, I. G. (2014). Episodic Positive Selection in the Evolution of Avian Toll-Like Receptor Innate Immunity Genes. *PLoS ONE*, 9(3), e89632. doi:10.1371/journal.pone.0089632
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321. doi:10.1093/sysbio/syq010
- Gutiérrez, J. S., Rakhimberdiev, E., Piersma, T., & Thieltges, D. W. (2017). Migration and parasitism: habitat use, not migration distance, influences helminth species richness in Charadriiform birds. *Journal of Biogeography*, 44(5), 1137–1147. doi:10.1111/jbi.12956
- Hannon, E. R., Kinsella, J. M., Calhoun, D. M., Joseph, M. B., & Johnson, P. T. J. (2015). Endohelminths in Bird Hosts from Northern California and an Analysis of the Role of Life History Traits on Parasite Richness. *Journal of Parasitology*, 102(2), 199–207. doi:10.1645/15-867
- Hasegawa, H., & Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology*, 19(3), 341–348. doi:10.1016/j.sbi.2009.04.003

- Hughes, A. L., & Yeager, M. (1998). Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics*, 32, 415–435. doi:10.1146/annurev.genet.32.1.415
- Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, 491(7424), 444–448. doi:10.1038/nature11631
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637. doi:10.1002/bip.360221211
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)*, 28(12), 1647–1649. doi:10.1093/bioinformatics/bts199
- Keestra, A. M., & Putten, J. P. M. van. (2008). Unique Properties of the Chicken TLR4/MD-2 Complex: Selective Lipopolysaccharide Activation of the MyD88-Dependent Pathway. *The Journal of Immunology*, 181(6), 4354–4362. doi:10.4049/jimmunol.181.6.4354
- Kim, H. M., Park, B. S., Kim, J.-I., Kim, S. E., Lee, J., Oh, S. C., ... Lee, J.-O. (2007). Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell*, 130(5), 906–917. doi:10.1016/j.cell.2007.08.002
- Koch, M., Camp, S., Collen, T., Avila, D., Salomonsen, J., Wallny, H.-J., ... Kaufman, J. (2007). Structures of an MHC Class I Molecule from B21 Chickens Illustrate Promiscuous Peptide Binding. *Immunity*, 27(6), 885–899. doi:10.1016/j.immuni.2007.11.007
- Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*, 23(10), 1891–1901. doi:10.1093/molbev/msl051
- Kropáčková, L., Těšický, M., Albrecht, T., Kubovčíak, J., Čížková, D., Tomášek, O., ... Kreisinger, J. (2017). Codiversification of gastrointestinal microbiota and phylogeny in passerines is not explained by ecological divergence. *Molecular Ecology*, 26(19):5292–5304. doi:10.1111/mec.14144
- Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7), 1307–1320. doi:10.1093/molbev/msn067
- Leung, T. L. F., & Koprivnikar, J. (2016). Nematode parasite diversity in birds: the role of host ecology, life history and migration. *Journal of Animal Ecology*, 85(6), 1471–1480. doi:10.1111/1365-2656.12581
- Leveque, G., Forgetta, V., Morroll, S., Smith, A. L., Bumstead, N., Barrow, P., ... Malo, D. (2003). Allelic variation in TLR4 is linked to susceptibility to *Salmonella enterica* serovar Typhimurium infection in chickens. *Infection and Immunity*, 71(3), 1116–1124.
- Maeshima, N., & Fernandez, R. (2013). Recognition of lipid A variants by the TLR4-MD-2 receptor complex. *Frontiers in Cellular and Infection Microbiology*, 3. doi:10.3389/fcimb.2013.00003
- McGuffin, L. J., Buenavista, M. T., & Roche, D. B. (2013). The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Research*, 41(Web Server issue), W368–W372. doi:10.1093/nar/gkt294
- Medeiros, M. C. I., Hamer, G. L., & Ricklefs, R. E. (2013). Host compatibility rather than vector-host-encounter rate determines the host range of avian *Plasmodium* parasites.

- 696 *Proceedings of the Royal Society B: Biological Sciences*, 280(1760), 20122947–
 697 20122947. doi:10.1098/rspb.2012.2947
- 698 Medzhitov, R., & Janeway Jr, C. A. (1998). An ancient system of host defense. *Current*
 699 *Opinion in Immunology*, 10(1), 12–15. doi:10.1016/S0952-7915(98)80024-1
- 700 Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., &
 701 Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for
 702 inferring selection. *Molecular Biology and Evolution*, 30(5), 1196–1205.
 703 doi:10.1093/molbev/mst030
- 704 Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S.
 705 L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS*
 706 *Genetics*, 8(7), e1002764. doi:10.1371/journal.pgen.1002764
- 707 Needham, B. D., & Trent, M. S. (2013). Fortifying the barrier: the impact of lipid A
 708 remodelling on bacterial pathogenesis. *Nature Reviews Microbiology*, 11(7), 467–481.
 709 doi:10.1038/nrmicro3047
- 710 Nijland, R., Hofland, T., & van Strijp, J. A. G. (2014). Recognition of LPS by TLR4:
 711 Potential for Anti-Inflammatory Therapies. *Marine Drugs*, 12(7), 4260–4273.
 712 doi:10.3390/md12074260
- 713 Ohto, U., Fukase, K., Miyake, K., & Shimizu, T. (2012). Structural basis of species-specific
 714 endotoxin sensing by innate immune receptor TLR4/MD-2. *Proceedings of the*
 715 *National Academy of Sciences of the United States of America*, 109(19), 7421–7426.
 716 doi:10.1073/pnas.1201193109
- 717 Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2013).
 718 *caper: Comparative Analyses of Phylogenetics and Evolution in R*. Retrieved from
 719 <http://CRAN.R-project.org/package=caper>
- 720 Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and
 721 Evolution in R language. *Bioinformatics*, 20(2), 289–290.
 722 doi:10.1093/bioinformatics/btg412
- 723 Park, B. S., Song, D. H., Kim, H. M., Choi, B.-S., Lee, H., & Lee, J.-O. (2009). The structural
 724 basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature*,
 725 458(7242), 1191–1195. doi:10.1038/nature07830
- 726 Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C.,
 727 & Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory
 728 research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612.
 729 doi:10.1002/jcc.20084
- 730 Piertney, S. B., & Oliver, M. K. (2006). The evolutionary ecology of the major
 731 histocompatibility complex. *Heredity*, 96(1), 7–21. doi:10.1038/sj.hdy.6800724
- 732 Pond, S. L. K., & Frost, S. D. W. (2005a). Datamonkey: rapid detection of selective pressure
 733 on individual sites of codon alignments. *Bioinformatics (Oxford, England)*, 21(10),
 734 2531–2533. doi:10.1093/bioinformatics/bti320
- 735 Pond, S. L. K., & Frost, S. D. W. (2005b). Not So Different After All: A Comparison of
 736 Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and*
 737 *Evolution*, 22(5), 1208–1222. doi:10.1093/molbev/msi105
- 738 Poulin, R., & Morand, S. (2000). The diversity of parasites. *The Quarterly Review of Biology*,
 739 75(3), 277–293.
- 740 R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna,
 741 Austria. Retrieved from <https://www.R-project.org/>

- R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Richter, S., Wenzel, A., Stein, M., Gabdoulline, R. R., & Wade, R. C. (2008). webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Research*, 36(suppl 2), W276–W280. doi:10.1093/nar/gkn181
- RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. Retrieved from <http://www.rstudio.com/>
- Santoro, M., Kinsella, J. M., Galiero, G., Uberti, B. degli, & Aznar, F. J. (2012). Helminth Community Structure in Birds of Prey (Accipitriformes and Falconiformes) in Southern Italy. *Journal of Parasitology*, 98(1), 22–29. doi:10.1645/GE-2924.1
- Schemske, D. W., Mittelbach, G. G., Cornell, H. V., Sobel, J. M., & Roy, K. (2009). Is There a Latitudinal Gradient in the Importance of Biotic Interactions? *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 245–269. doi:10.1146/annurev.ecolsys.39.110707.173430
- Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593. doi:10.1093/bioinformatics/btq706
- Schmid-Hempel, P. (2011). *Evolutionary Parasitology: The Integrated Study of Infections, Immunology, Ecology, and Genetics*. OUP Oxford.
- Smirnova, I., Poltorak, A., Chan, E. K., McBride, C., & Beutler, B. (2000). Phylogenetic variation and polymorphism at the Toll-like receptor 4 locus (TLR4). *Genome Biology*, 1(1), research002.1-research002.10.
- Takeda, K., Kaisho, T., & Akira, S. (2003). Toll-like receptors. *Annual Review of Immunology*, 21, 335–376. doi:10.1146/annurev.immunol.21.120601.141126
- Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilities of Residues in Proteins. *PLOS ONE*, 8(11), e80635. doi:10.1371/journal.pone.0080635
- Tong, C., Lin, Y., Zhang, C., Shi, J., Qi, H., & Zhao, K. (2015). Transcriptome-wide identification, molecular evolution and expression analysis of Toll-like receptor family in a Tibet fish, *Gymnocypris przewalskii*. *Fish & Shellfish Immunology*, 46(2), 334–345. doi:10.1016/j.fsi.2015.06.023
- Vinkler, M., Bryjová, A., Albrecht, T., & Bryja, J. (2009). Identification of the first Toll-like receptor gene in passerine birds: TLR4 orthologue in zebra finch (*Taeniopygia guttata*). *Tissue Antigens*, 74(1), 32–41. doi:10.1111/j.1399-0039.2009.01273.x
- Vinkler, M., & Albrecht, T. (2009). The question waiting to be asked: Innate immunity receptors in the perspective of zoological research. *Folia Zoologica*, 58, 15–28.
- Vinkler, M., Bainová, H., & Bryja, J. (2014). Protein evolution of Toll-like receptors 4, 5 and 7 within Galloanserae birds. *Genetics Selection Evolution*, 46(1), 72.
- Waldenström, J., Bensch, S., Kiboi, S., Hasselquist, D., & Ottosson, U. (2002). Cross-species infection of blood parasites between resident and migratory songbirds in Africa. *Molecular Ecology*, 11(8), 1545–1554. doi:10.1046/j.1365-294X.2002.01523.x
- Walsh, C., Gangloff, M., Monie, T., Smyth, T., Wei, B., McKinley, T. J., ... Bryant, C. (2008). Elucidation of the MD-2/TLR4 Interface Required for Signaling by Lipid IVa. *The Journal of Immunology*, 181(2), 1245–1254. doi:10.4049/jimmunol.181.2.1245
- Wang, J., Zhang, Z., Liu, J., Zhao, J., & Yin, D. (2016). Ectodomain Architecture Affects Sequence and Functional Evolution of Vertebrate Toll-like Receptors. *Scientific Reports*, 6. doi:10.1038/srep26705

- Wang, X., Quinn, P. J., & Yan, A. (2015). Kdo₂-lipid A: structural diversity and impact on immunopharmacology: Kdo₂-lipid A. *Biological Reviews*, 90(2), 408–427. doi:10.1111/brv.12114
- Wlasiuk, G., & Nachman, M. W. (2010). Adaptation and Constraint at Toll-Like Receptors in Primates. *Molecular Biology and Evolution*, 27(9), 2172–2186. doi:10.1093/molbev/msq104
- Woolhouse, M. E., 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(4), 569–577.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12(1), 7–8.
- Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591. doi:10.1093/molbev/msm088
- Yilmaz, A., Shen, S., Adelson, D. L., Xavier, S., & Zhu, J. J. (2004). Identification and sequence analysis of chicken Toll-like receptors. *Immunogenetics*, 56(10), 743–753. doi:10.1007/s00251-004-0740-8

Data Accessibility

DNA sequences: Genbank accessions MG209180-MG209225
All data are presented within this manuscript or in online Supplemental Information.

Author Contributions

T.A., J.B. and M.V. performed the study design and provided funding. T.A., D.H., A.J., J.T.L., O.S., M.V. provided the samples for analysis. T.K. and H.V. performed laboratory analysis. T.K., H.V., M.N. and M.V. analysed the data. T.K. and M.V. drafted the manuscript. All authors provided helpful comments and recommendations and approved the final version of the manuscript.

Tables and Figures

Table 1 Predominant characteristics of amino acid variants represented at the well-supported positively selected sites in the four distinct TLR4 LBR surface charge clusters. Only those sites that were identified as positively selected (or non-conservative) by at least three independent tests (PAML, FUBAR, MEME, SLAC, PRIME, ConSurf) are displayed (n=11; for details see Table S7). A characteristic was considered predominant for a specific cluster only when present in more than 75% of the species belonging to the particular cluster (Table S8). POS = positive charge, NEG = negative

charge, PLR = polar uncharged, SPE = special cases, HYD = hydrophobic; insufficient data are indicated by a dash (-).

Figure 1 Variability in the surface charge of the TLR4 LBR in four different passerine species. One species from each charge cluster (Figure S3) is displayed – *Erithacus rubecula* (cluster A), *Hirundo rustica* (cl. B), *Certhia familiaris* (cl. C), and *Carpodacus erythrinus* (cl. D). *Certhia familiaris* and *Erithacus rubecula* are the two species with the greatest distance based on the surface electrostatic potential of the whole TLR4 LBR ($d = 1.284$, Figure S3). The inner (left-hand side) and outer (right-hand side) surface of each TLR4 LBR is shown, with the ligand-binding surface shown in the left-hand side view. Positive charge (10) is highlighted in red and negative (-10) in blue (colours in electronic version only). Positively selected amino acids confirmed by three independent tests (Table 1, Table S7) are labelled (the site numbering corresponds to the Zebra finch protein sequence; positions 272 and 351 are hidden). The position of the LBR within the ECD is indicated by the blue colour in the schematic representation of Zebra finch TLR4. For the variability in the surface electrostatic potential of all studied species see Figure S2.

Figure 2 Positions of positively selected sites visualized on the 3D protein model of the Zebra finch (*Taeniopygia guttata*; *Taegut*) TLR4 LBR. Positively selected sites (PAML/FUBAR/MEME/SLAC/PRIME) are highlighted in red (colours in electronic version only); those sites that are also non-conservative (ConSurf) are labelled with an orange background; the sites that are in close proximity of potentially important LPS binding sites defined in previous studies (Kim et al. 2007, Park et al. 2009, Ohto et al. 2012) are indicated with black arrows; the blue arrow shows the position of the positively selected site potentially important for lipid IVa recognition (Table S7). Site numbering corresponds to the *Taegut* protein sequence. The position of the ligand-binding region (in blue) on the receptor is shown in the small model of the *Taegut* TLR4 extracellular domain.

Figure 3 Projection of ecological characteristics to TLR4 LBR surface charge clustering. The dendrogram showing the surface charge clustering of the TLR4 LBR (Fig. S3, Fig. S4) was collapsed into four main clusters (A-D), and for each group we determined the proportion of two basic ecological characteristics: (1) migration and latitude, (2) diet (Table S1). The letter diagrams show the proportion of specific amino acids on the 11 positively selected sites with their physicochemical properties indicated by colour (in electronic version only; red – positive charge, blue – negative charge, orange – polar uncharged, grey – special cases, black – hydrophobic; Table S8). The arrows specify those amino acid sites that appear to be important for the differentiation of individual clusters.

Figure 4 Consistency of neutral phylogeny with surface charge clustering. A comparison of the neutral phylogeny (left-hand side) and the surface charge dendrogram (right-hand side) shows that closely related species mostly cluster together in both trees, but the deeper branches do not match. The end branches are labelled with the species name abbreviations (Table S1). The chicken sequence (*Galgal*, GenBank Acc.No. AY064697.1) was used as an outgroup. The end branches of the same species are connected with lines. Coloured lines (electronic version only) indicate common sub-trees

863 of both dendrograms. Branches leading to distinct sub-trees are highlighted with a dashed line. The
864 surface charge clusters are indicated in the dendrogram by capital letters and miniature of the
865 representative structural models.

Cluster	267	269	270	272	298	308	331	344	351	374	402
A	-	PLR	NEG	HYD	SPE	-	HYD	PLR	PLR	NEG	-
B	PLR	-	-	-	NEG	-	HYD	-	HYD	POS	-
C	PLR	PLR	NEG	HYD	-	PLR	HYD	-	HYD	POS	PLR
D	-	-	-	-	-	PLR	HYD	-	HYD	-	PLR

Figure 1 Variability in the surface charge of the TLR4 LBR in four different passerine species. One species from each charge cluster (Figure S3) is displayed – *Erithacus rubecula* (cluster A), *Hirundo rustica* (cl. B), *Certhia familiaris* (cl. C), and *Carpodacus erythrinus* (cl. D). *Certhia familiaris* and *Erithacus rubecula* are the two species with the greatest distance based on the surface electrostatic potential of the whole TLR4 LBR ($d = 1.284$, Figure S3). The inner (left-hand side) and outer (right-hand side) surface of each TLR4 LBR is shown, with the ligand-binding surface shown in the left-hand side view. Positive charge (10) is highlighted in red and negative (-10) in blue (colours in electronic version only). Positively selected amino acids confirmed by three independent tests (Table 1, Table S7) are labelled (the site numbering corresponds to the Zebra finch protein sequence; positions 272 and 351 are hidden). The position of the LBR within the ECD is indicated by the blue colour in the schematic representation of Zebra finch TLR4. For the variability in the surface electrostatic potential of all studied species see Figure S2.

179x80mm (300 x 300 DPI)

Figure 2 Positions of positively selected sites visualized on the 3D protein model of the Zebra finch (*Taeniopygia guttata*; Taegut) TLR4 LBR. Positively selected sites (PAML/FUBAR/MEME/SLAC/PRIME) are highlighted in red (colours in electronic version only); those sites that are also non-conservative (ConSurf) are labelled with an orange background; the sites that are in close proximity of potentially important LPS binding sites defined in previous studies (Kim et al. 2007, Park et al. 2009, Ohto et al. 2012) are indicated with black arrows; the blue arrow shows the position of the positively selected site potentially important for lipid IVa recognition (Table S7). Site numbering corresponds to the Taegut protein sequence. The position of the ligand-binding region (in blue) on the receptor is shown in the small model of the Taegut TLR4 extracellular domain.

152x129mm (300 x 300 DPI)

Figure 3 Projection of ecological characteristics to TLR4 LBR surface charge clustering. The dendrogram showing the surface charge clustering of the TLR4 LBR (Fig. S3, Fig. S4) was collapsed into four main clusters (A-D), and for each group we determined the proportion of two basic ecological characteristics: (1) migration and latitude, (2) diet (Table S1). The letter diagrams show the proportion of specific amino acids on the 11 positively selected sites with their physicochemical properties indicated by colour (in electronic version only; red – positive charge, blue – negative charge, orange – polar uncharged, grey – special cases, black – hydrophobic; Table S8). The arrows specify those amino acid sites that appear to be important for the differentiation of individual clusters.

152x85mm (300 x 300 DPI)

Figure 4 Consistency of neutral phylogeny with surface charge clustering. A comparison of the neutral phylogeny (left-hand side) and the surface charge dendrogram (right-hand side) shows that closely related species mostly cluster together in both trees, but the deeper branches do not match. The end branches are labelled with the species name abbreviations (Table S1). The chicken sequence (Galgal, GenBank Acc.No. AY064697.1) was used as an outgroup. The end branches of the same species are connected with lines. Coloured lines (electronic version only) indicate common sub-trees of both dendrograms. Branches leading to distinct sub-trees are highlighted with a dashed line. The surface charge clusters are indicated in the dendrogram by capital letters and miniature of the representative structural models.

152x82mm (300 x 300 DPI)