Major histocompatibility complex class I diversity limits the repertoire of T cell receptors

Magdalena Migalska^a, Alvaro Sebastian^a, and Jacek Radwan^{a,1}

^aEvolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, 60-614 Poznań, Poland

Edited by Robert L. Coffman, Dynavax Technologies, Berkeley, CA, and approved January 24, 2019 (received for review May 8, 2018)

Major histocompatibility complex (MHC) genes encode proteins that initiate adaptive immune responses through the presentation of foreign antigens to T cells. The high polymorphism found at these genes, thought to be promoted and maintained by pathogen-mediated selection, contrasts with the limited number of MHC loci found in most vertebrates. Although expressing many diverse MHC genes should broaden the range of detectable pathogens, it has been hypothesized to also cause deletion of larger fractions of self-reactive T cells, leading to a detrimental reduction of the T cell receptor (TCR) repertoire. However, a key prediction of this TCR depletion hypothesis, that the TCR repertoire should be inversely related to the individual MHC diversity, has never been tested. Here, using high-throughput sequencing and advanced sequencing error correction, we provide evidence of such an association in a rodent species with high interindividual variation in the number of expressed MHC molecules, the bank vole (*Myodes glareolus*). Higher individual diversity of MHC class I, but not class II, was associated with smaller TCR repertoires. Our results thus provide partial support for the TCR depletion model, while also highlighting the complex, potentially MHC class-specific mechanisms by which autoreactivity may trade off against evolutionary expansion of the MHC gene family.

major histocompatibility complex | T cell receptor | TCR depletion | immunogenetic optimality | *Myodes glareolus*

The considerable empirical effort that has been devoted to understanding the evolution of the major histocompatibility complex (MHC) has shown that selective pressure from pathogens is a major driver of the MHC's extreme polymorphism (1). Parasites have been shown to evade recognition by MHC alleles prevailing in a population, giving an advantage to rare or novel MHC variants in a dynamic, frequency-dependent process (2, 3). Furthermore, heterozygosity at the MHC has been shown to give an advantage to hosts infected by multiple pathogens (4, 5), presumably because any given allele allows recognition of only a subset of antigens matching its binding groove (6). These mechanisms should favor MHC gene duplication and divergence, because expressing a larger number of MHC variants should increase the probability of responding to any encountered pathogen. However, the number of functional MHC loci per genome is typically orders of magnitude lower than the number of putatively adaptive variants segregating in a population. For example, in humans, there are three functional classical MHC class I loci (generally responsible for presenting peptides from pathogens that replicate in the cytoplasm to cytotoxic CD8⁺ T cells) and three to four classical MHC class II loci (generally presenting antigens originating from the extracellular space to helper CD4⁺ T cells) (7). This allows only 12–14 different alleles in a fully heterozygous individual, while the number of currently identified MHC alleles in the human population exceeds 17,000 [IPD (Immuno Polymorphism Database)–IMGT (International Immunogenetics Project)/HLA database] (ref. 8, release 3.30.0).

The prevailing explanation for this apparent paradox is the optimality (9), or TCR depletion, hypothesis (10), which proposes a trade-off between within-individual MHC diversity and T cell receptor (TCR) repertoire size (9–11). TCRs are responsible for specific recognition of peptide–MHC complexes

and initiation of the adaptive immune response. Gaps in the TCR repertoire have detrimental effects on the efficiency of the immune response (12). Primary TCR diversity, which is, to a large extent, random due to somatic V(D)J recombination and non-template addition/deletion of nucleotides, is censored through positive and negative selection in the thymus. Positive selection results in the retention of T cells that are able to interact with any MHC–self-peptide complex, while failure to engage in such interaction leads to death of neglect. In negative selection, T cells bearing TCRs with too strong an avidity to MHC–self-peptide complexes are deleted to prevent autoimmunity and assure self-tolerance (13). Theoretical models have predicted that expressing too many different MHC molecules will render more TCRs self-reactive, leading negative selection to drastically reduce the TCR repertoire (9, 10). However, Borghans et al. (14) pointed out that the effect of high MHC diversity on the TCR repertoire does not necessarily have to be negative: Additional MHC variants could enhance positive selection, leading to retention of a larger fraction of T cells. The net outcome of thymic selection in individuals with high MHC diversity may thus be depletion (9, 10) or relative enrichment (14) of the TCR repertoire.

Efforts to test the TCR depletion hypothesis have focused on species characterized by intraspecific variation in the number of MHC loci, where individuals with an average, presumably optimal, rather than maximal number of MHC loci are predicted to have the highest immune competence. However, indirect tests using either proxies of immunocompetence (e.g., parasite load,

Significance

The major histocompatibility complex (MHC) is central for self-/non-self-recognition and acquired immunity. The extreme polymorphism of MHC genes, promoted by parasite-mediated selection, contrasts with limited within-individual diversity. The prevailing explanation is a trade-off between increased pathogen recognition and the anti-autoimmune T cell receptor (TCR) depletion mechanism. However, the predicted inverse relationship between individual MHC diversity and TCR repertoire size has not yet been shown. Using a rodent species with a variable number of MHC genes, we detected such an effect for MHC class I, but not class II. Our results thus partially support the TCR depletion hypothesis, but also suggest additional, unexplored mechanisms that might be constraining expansion of the MHC gene family.

Author contributions: M.M. and J.R. designed research; M.M. performed research; M.M., A.S., and J.R. analyzed data; M.M. and J.R. wrote the paper; A.S. wrote the AmpliSAT suite and contributed to TCR repertoire analysis; and J.R. conceived the study.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: New nucleic acid sequences of MHC genes were deposited in the GenBank database (accession nos. [MH018253–MH018459](#)). MHC genotypes (with intermediate genotyping files) and MHC supertype assignments are deposited in the Open Science Framework repository (accession code [osf.io/r72x8](#)). TCR sequencing reads are deposited in the European Nucleotide Archive repository (accession no. [PRJEB25041](#)).

¹To whom correspondence should be addressed. Email: jradwan@amu.edu.pl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1807864116/-DCSupplemental.

Published online February 22, 2019.

diversity) or fitness and its correlates (e.g., reproductive success, body size) have yielded mixed results (15–26). The nature of the relationship between the number of MHC genes and TCR repertoire size thus remains a crucial, but open, empirical question, resolution of which has been hampered by the technical difficulties associated with assessing the immense sizes of TCR repertoires (27).

Here, we used high-throughput sequencing (HTS) and recent advances in sequencing error correction to investigate how the number of expressed MHC variants is associated with the TCR repertoire in the bank vole (*Myodes glareolus*). Individuals of this species vary widely in their number of expressed MHC variants (i.e., distinct amino acid sequences in functionally important exons at a number of duplicated and diversified loci) (28, 29), and previous studies on MHC class II DRB genes in this species have partly supported the immunogenetic optimality model: An intermediate number of MHC variants were associated with the lowest number of parasite species, but intensity of infection decreased with MHC variant number (19). Now, using a direct test of the optimality hypothesis, we show that high numbers of both MHC class I amino acid variants and supertypes are correlated with a decreased TCR repertoire, supporting the TCR depletion model. However, we did not detect such a relationship between the number of MHC class II supertypes (or amino acid variants) and the TCR repertoire, which suggests that MHC classes differ in the way they select T cells and/or deal with autoreactivity.

Results

We used animals from a large, genetically variable population founded as a laboratory model of adaptive radiation, which was kept under controlled, common garden conditions. We extracted total RNA from spleen samples from 156 voles of known age and sex, and estimated the number of functional MHC variants per individual by sequencing gene parts encoding the antigen-binding domains of class I (third exon) and class II (second exon of DQA, DQB, and DRB genes; *Materials and Methods* and *SI Appendix, Supplementary Text S1*). Multilocus typing of coamplifying genetic variants (referred to as “genotyping” throughout the text) was performed in two steps (details are provided in *Materials and Methods* and *SI Appendix, Supplementary Text S2* and Fig. S1). We first used two independent PCR assays to “MHC-genotype” all individuals (genotyping success: 155 of 156 individuals). We then selected 77 individuals from the tails of the expressed MHC copy number distribution for additional PCR replicates to improve genotyping accuracy (*SI Appendix, Fig. S2*). We used the distribution’s tails to increase the power of detecting a significant relationship between MHC diversity and TCR repertoire by reducing overrepresentation of individuals with an intermediate number of variants. Final, complete genotypes, based on the presence of an MHC variant in at least two of four independent PCR replicates, were obtained for 72 individuals. Summary statistics for MHC amino acid variants are provided in *SI Appendix, Table S1*. Our results corroborated the extent of the MHC copy number variation reported previously in the bank vole (28–31), and showed that it holds at the mRNA level.

We further chose 30 individuals, representing the whole range of expressed MHC variants (12–38 variants in MHC class I and II combined; *Materials and Methods* and *SI Appendix, Supplementary Text S2*) for in-depth TCR- β chain sequencing, and successfully sequenced the TCR repertoire for 28 voles (*SI Appendix, Supplementary Text S3*). The TCR repertoire size estimates were based on the number of unique amino acid variants of the hypervariable complementarity-determining region 3 (CDR3), the part of a TCR that determines the specificity of interactions between lymphocytes and MHC–peptide complexes. Individual repertoire size was calculated based on quadruplicate sequencing of CDR3 from cDNA tagged with unique molecular identifiers (UMIs) (32, 33), which allowed efficient error correction (details are provided in *Materials and Methods*). To

control for uneven sequencing depth, all amplicons were subsampled to 1 million reads. The numbers of variants we detected among the four replicates were highly repeatable (34) within individuals ($r = 0.99$). At the nucleotide level, the observed repertoire range was: 0.5×10^5 to 3.1×10^5 unique CDR3 sequences (mean: $1.6 \times 10^5 \pm \text{SD } 0.6 \times 10^5$). The observed repertoire at the amino acid level ranged from 0.4×10^5 to 2.3×10^5 unique CDR3 sequences (mean: $1.4 \times 10^5 \pm 0.4 \times 10^5$; *Dataset S1*). To correct for incomplete saturation of the TCR repertoire in a sample, we used an incidence-based richness estimator, Chao2 (35), which is commonly used to estimate total sizes of immune repertoires (36, 37). The estimated lower bound of the TCR- β amino acid repertoire size ranged from 0.5×10^5 to 2.9×10^5 CDR3 sequences (mean: $1.6 \times 10^5 \pm 0.6 \times 10^5$; *Dataset S1*). These values are of the same order of magnitude as reported for mice (38).

To estimate the relative proportions of CD4⁺ and CD8⁺ T cells, which may indicate the relative effects of MHC class II and class I, respectively, on the TCR repertoire, we performed a qPCR assay for markers specific to these cell types (*Materials and Methods* and *SI Appendix, Supplementary Text S4*). The mean CD4/CD8 ratio was 1.5 (median = 1.1, SD = 1.1; *SI Appendix, Table S3*), suggesting approximately similar effects of each MHC class on total TCR repertoire.

Not all amino acid substitutions among MHC variants translate to functionally significant differences that could affect antigen binding. Therefore, before analysis, we used clustering analyses to group MHC amino acid variants into supertypes (39): groups of alleles assumed to have similar peptide-binding capacities (*Materials and Methods* and *SI Appendix, Supplementary Text S5* and S6). For MHC class I, variants were classified into 22 supertypes; for MHC class II, variants were classified into 12 supertypes for DQA and 11 for DQB. Cluster analyses were problematic for DRB (details are provided in *SI Appendix, Supplementary Text S6*); thus, we grouped alleles into 17 “types” that consisted of alleles with identical amino acids at the residues likely to engage in peptide binding. Summary statistics on numbers of grouped variants per individual and allelic composition of each group are provided in *SI Appendix, Table S2*.

We used linear mixed models (*Materials and Methods*), controlling for sex, age, and line of origin, to test for a correlation between an individual’s number of MHC supertypes and its estimated TCR repertoire size. A model with total number of MHC supertypes resulted in a poorer model fit than when each MHC class was fitted as a separate predictor [ΔAIC (Akaike information criterion) = 2.01]. Analysis of the best-fit model (Table 1) showed that the TCR repertoire decreased with the number of MHC class I supertypes ($P = 0.01$; Fig. 1 and Table 1), and was lower in males compared with females ($P = 0.02$; Table 1). The number of MHC class II supertypes was not significantly correlated with the TCR repertoire (Table 1). Analogous analysis, performed with the sum of MHC amino acid variants (rather than supertypes) as predictors, concurred with the above findings ($P = 0.04$ and $P = 0.05$ for sum MHC class I variants and sex, respectively; *SI Appendix, Fig. S3* and Table S4). These results were robust to removal of potentially influential data points, MHC genotyping and supertyping protocols, and direct use of observed CDR3 diversity instead of Chao2 estimates (*SI Appendix, Supplementary Text S7*, Figs. S4 and S5, and Tables S5–S10).

Discussion

Our results are consistent with the key prediction of the TCR depletion hypothesis for MHC class I, but not for MHC class II. The two classes differ markedly in, among other traits, tissue distribution, antigen-processing pathway, and T cell types they interact with. MHC class I is expressed in every nucleated cell and presents antigens derived from intracellular pathogens to CD8⁺ cytotoxic T cells. MHC class II is mainly expressed by special antigen-presenting cells and presents peptides of extracellular

both MHC classes, which allowed us to uncover an unanticipated difference between MHC class I and MHC class II genes. This discrepancy does not rule out the TCR depletion hypothesis applying across MHC classes, but understanding the hypothesis' generalizability will require that the mechanism for the discrepancy be unpicked. We thus hope that the contrasting patterns revealed for the two MHC classes will encourage further research on the nature of evolutionary trade-offs shaping diversity of MHC and TCR repertoires, and will inspire experiments accounting for differences in which CD4⁺ and CD8⁺ T cells are recruited into the mature T cell pool.

Materials and Methods

Genetic Markers and Primer Design. To characterize individual diversity in expressed classical MHC genes (i.e., class I, class II) in bank voles, we sequenced the hypervariable exons (exon 3 for MHC class I and exon 2 for MHC class II) coding for antigen-binding domains of MHC molecules. MHC binding grooves are formed by a single α chain for MHC class I and by a dimer of α and β chains for MHC class II. For MHC class I, we amplified exon 3 using primers MyglMHC13_F and MyglMHC13_R. These primers were designed from de novo spleen transcriptome assembly (28), and should amplify all presumably classical variants of MHC I genes identified in the transcriptome. Unlike MHC class I genes that are characterized by fast allelic turnover and lack of orthology between species (51–53), MHC class II evolves more slowly, and orthologous relationships at divergent groups of loci (e.g., DR, DQ) can be inferred across mammalian orders (54). Because divergence between class II genes prevented design of universal primers, we designed separate primers for bank vole orthologs of DQ α and DQ β (hereafter, DQA and DQB), and DR β (hereafter, DRB) genes using the same transcriptome-based approach as for MHC class I (28). In brief, spleen transcriptomes from seven individuals were assembled de novo with Trinity (55), and contigs matching a desired target gene were recovered with blastn (56). A separate local blastn search was performed for each gene of interest, using a reference database with sequences from multiple rodent species (GenBank; the full list is provided in [Dataset S2](#)). Recovered transcripts were aligned with rodent MHC reference RNA sequences ([SI Appendix, Table S11](#)), and primers were designed in regions highly conserved among species, particularly so among the bank vole-transcribed sequences. At least one primer in each pair was placed in an adjacent exon (first and/or third exon) to avoid amplification of genomic DNA. We also identified DR α genes, but as they were nearly monomorphic (similar to murine ortholog H-2 IE α and human HLA-DR α) (57), we omitted this gene in further procedures. We used the following primer pairs: MyglDQA_F: RTCCTCGCCCTGACCACC + MyglDQA_R: GGGTGTGGGCTGACCCA; MyglDQB_F: AGCTGTGGTCTGATGGT + MyglDQB_R: ARTTGTGTCTGCACACCT; and MyglDRB_F: TGGCAGCTGTGATCCTGA + MyglDRB_R: AGCAGACCAGGAGTTGT. The amplified gene fragments had length of 328 bp (DQA), 253 bp (DQB), and 369 bp (DRB), excluding primers. The schematic location of primers is shown in [SI Appendix, Fig. S6](#). Details on MHC class I primers are provided elsewhere (28).

For TCR repertoire sequencing, we used a 5' rapid amplification of cDNA ends (RACE)-based library preparation protocol (58). A set of nested primers embedded in the 3' end of a constant region of the TCR- β chain was designed by Migalska et al. (33), using the same transcriptome-based approach as above.

We developed an additional qPCR assay to allow rough estimation of CD4/CD8 T cell ratio in the bank vole. As a reference gene, we used lymphocyte-specific protein tyrosine kinase (LCK), a kinase highly specific to T cells, involved in the intracellular signaling pathways from TCRs (59, 60). Gene-specific primers for the qPCR assay were designed based on de novo assembled bank vole transcriptomes, as explained previously. Primer sequences are listed in [SI Appendix, Table S12](#); details on primer design, assay optimization and PCR conditions, tests of specificity, calculations, and comparisons of amplification efficiencies are provided in [SI Appendix, Supplementary Text S4, Fig. S7, and Tables S13 and S14](#).

Samples. The study was performed using samples from a laboratory colony of the bank vole (*M. glareolus*) maintained at the Institute for Environmental Sciences (Jagiellonian University, Kraków, Poland). The colony was set up from over 320 wild-caught, genetically variable animals for a large-scale experiment to model adaptive radiation. The experiment consists of three artificial selection directions: predatory behavior (P), ability to maintain weight on herbivorous diet (H), aerobic performance (A), and an unselected control (C). Within each selection direction there are four replicated lines. The voles from all selection directions are kept under standardized conditions except for different types of short-term measurements they undergo.

Details about the animal initial source, maintenance, and selection protocols are available elsewhere (61).

Accurate TCR diversity estimation requires ultra-deep sequencing (32), which limited our analysis to a maximum of 30 individuals. However, to optimize the range of individual MHC diversity in the final sample, as well as to balance representation of sexes and selection directions, we screened MHC diversity in a larger set of voles from which we then selected a final subset of individuals for TCR sequencing. Spleens from bank voles from all selection directions ($n = 156$) were collected during necropsy in accordance with internationally recognized guidelines for research on animals, approved by the Kraków Ethical Committee for Experiments on Animals. Immediately after collection, spleens were cut into four to eight pieces (depending on spleen size), preserved in RNAlater (Sigma-Aldrich), and then stored at -20°C . The pieces of spleen were homogenized using FastPrep (MP Biomedicals); total RNA was extracted with RNeasyol RT (Sigma-Aldrich) according to the manufacturer's instructions and eluted in 50 μL of RNase-free water. For final MHC genotyping ([SI Appendix, Supplementary Text S2](#)), the aliquots of each isolate (20 μL) were pooled for reverse transcription. Samples selected for 5' RACE (and later used in qPCR assays) were further purified on NEXTflex Poly(A) Beads (Bioo Scientific) to remove abundant rRNAs and reduce sample volume (elution in 15 μL).

MHC Genotyping. Details on library preparation (reverse transcription, PCR conditions, purification of PCR products) and sequencing are provided in [SI Appendix, Supplementary Text S1 and Table S15](#). Genotyping was performed with AmpliSAS (Amplicon Sequencing Assignment tool) software (62). AmpliSAS de-multiplexes pooled samples, clusters putative sequencing artifacts with true variants, and filters variants according to user-specified parameters. In the analysis, we also used a number of accompanying tools from the AmpliSAT suite (Amplicon Sequencing Analysis Tools; [evobiolab.biol.amu.edu.pl/amplisat/index.php](#)). Detailed descriptions of the genotyping protocol and AmpliSAS parameters are provided in [SI Appendix, Supplementary Text S2, Fig. S8, and Tables S16 and S17](#).

For initial screening of the individual variation in the number of expressed MHC variants in the studied population, we used samples from 156 individuals (two independent PCR assays for each pair of primers, HTS by Ion Torrent PGM; Thermo Fisher Scientific). We successfully recovered complete MHC class I and class II genotypes from 155 bank voles, from which we chose 77 individuals for more accurate genotyping. Random sampling would mostly include individuals with an average number of expressed MHC genes; therefore, to increase the power of detecting a significant relationship between MHC diversity and TCR repertoire, we chose individuals from the first and fourth quantiles of the expressed MHC copy number distribution ([SI Appendix, Fig. S2](#)). For these individuals, we carried out two additional rounds of PCR and Illumina HTS (we changed the sequencing platform to avoid time-consuming manual curation, which turned out to be necessary for PGM data; [SI Appendix, Supplementary Text S2](#)) to improve precision of MHC typing. We succeeded in obtaining four replicates for 72 individuals.

We incorporated MHC variants into an individual's genotype only if present in at least two of four replicates. We opted for four replicates, because the use of RNA as the starting genetic material implied more expression level-induced variation in the sequencing reads coverage between MHC variants, compared with amplification from genomic DNA. This is a modification of standard approaches used in MHC typing from genomic DNA, for which genotyping strategies based on read abundance are well established (reviewed in ref. 63). Using four replicates allowed us to call variants with low per-amplicon frequency, on the additional condition that they occurred above minimal threshold frequency more than once in four replicates. For downstream analyses, we translated nucleotide sequences of focal MHC exons into amino acid variants.

TCR Repertoire Size Estimation. From 72 individuals with complete MHC genotypes, we chose 30 voles for TCR repertoire sequencing. The criteria for inclusion in the final set were good RNA quality from RNA isolates (assessed through agarose gel electrophoresis), balanced proportion of sexes (female = 18, male = 12), and balanced representation of selection regimes the animals originated from ($H = 14$, $P = 16$). Details on animals for which TCR repertoire sizes have been successfully sequenced ($n = 28$) and the repertoire size estimations (sufficient to recreate statistical models) are provided in [Dataset S1](#).

The protocol for TCR repertoire sequencing and bioinformatics analysis is described in detail in the quantitative analysis section of a study by Migalska et al. (33). The data from the three animals analyzed in that study were reanalyzed herein and included in the present study. Briefly, the library was prepared with 5' RACE (58) adapted to incorporate UMIs (32). UMIs uniquely

tag individual template DNA molecules during cDNA synthesis, and allow for accurate discrimination of sequencing errors from biological variants (32). For each bank vole, four libraries were prepared, originating from four PCR replicates. Samples were sequenced with dual indices allowing for de-multiplexing of the amplicons on Illumina HiSeq 2500 paired-end, 150-bp sequencing (Macrogen). A total of 120 amplicons were separated into three sequencing runs to assure sufficient per-amplicon sequencing depth. Bioinformatics analysis was performed with tools belonging to the AmpliSAT suite: AmpliMERGE, Ampli-CLEAN, and AmpliCDR3. To control for uneven sequencing depths between amplicons, we subsampled all amplicons to 1 million reads. For each individual with sufficient sequencing depth ($n = 28$; inclusion criteria are available in *SI Appendix, Supplementary Text S3*), we calculated the number of unique CDR3 sequences from four PCR replicates, which were subject to a UMI-based error correction protocol with the AmpliCDR3 tool (33). Because a directly observed number of CDR3s is likely a fraction of the total repertoire (64, 65), we estimated total TCR- β richness using estimators adapted from ecological census techniques. These approaches estimate missing observations ("unseen species"), and are often used to estimate the total size of immune repertoires (36, 37, 66, 67). Following methods used by other authors (36, 37), we used a nonparametric, incidence-based estimator, Chao2 (35). Chao2 uses the presence/absence data across replicates to estimate a lower boundary of TCR- β repertoire size. We focused on incidence data rather than abundance, because cDNA levels cannot be directly translated into absolute cell counts (identical CDR3s might derive from many mRNA transcripts of one T cell, or from different T cells with an identical CDR3). We did, however, check for disproportional expansion of particular CDR3s among sequenced variants, and none were observed. More than 97% of CDR3s were represented by rare variants (each represented <0.001% of the sequenced pool), and the largest expansions (1–2%) were represented by maximum of one to two variants per sample.

qPCR. A qPCR assay was designed based on minimum information for publication of quantitative real-time PCR experiments guidelines (68). The same type and amount of genetic material were used as for 5' RACE (~30 ng of purified mRNA as described in *Samples*). Details on PCR conditions are provided in *SI Appendix, Supplementary Text S4*. Comparable amplification efficiencies (*SI Appendix, Table S13*) allowed the use of a comparative method of relative quantification ($\Delta\Delta C_q$). Normalized (by LCK) target gene expression levels were calculated as $2^{-\Delta\Delta C_q}$. Results are shown in *SI Appendix, Table S3*. A proxy of the CD4/CD8 T cell ratio was calculated based on these expression levels.

MHC Supertyping. We classified MHC class I and class II variants into "supertypes" (39, 69), based on similarity of amino acid physicochemical properties (70) at putative peptide-binding residues (PBRs). Supertyping summarizes the functional diversity of the MHC by grouping alleles with similar predicted peptide binding potential. Therefore, supertypes may be better suited to investigate the effect of MHC diversity on the TCR repertoire than simple amino acid sequence diversity. Supertyping was performed for each gene separately. We first identified positively selected sites (PSSs). Such sites are known to play a role in the specificity of antigen binding (71, 72) and are commonly used as an approximation of polymorphic PBRs in species where MHC crystal structures are not resolved (73). PSSs for MHC class I were previously identified (28), whereas MHC class II sequences were longer than those analyzed by previous studies; thus, we inferred PSSs with CODEML from the PAML package (74) (details on the procedure, comparison of codon evolution models, and identification of codons for supertyping are provided in *SI Appendix, Supplementary Text S5, Fig. S9, and Tables S18–S21*). In supertyping, we used all alleles found in the present study together with MHC alleles deposited in the GenBank, provided they included all PSSs covered by the sequences we obtained (*SI Appendix,*

Supplementary Text S5 and S6). After assigning numerical values describing five physicochemical properties to amino acids at PSSs, we clustered alleles using discriminant analysis of principal components (75), as implemented in the adegenet (76, 77) package in R (78). A full description of the process, including parameter settings and cluster number selection, is provided in *SI Appendix, Supplementary Text S6 and Fig. S10*. We successfully clustered MHC variants into supertypes for MHC class I and MHC class II DQA and DQB genes. In DRB, only six sites exhibited signs of positive selection, which did not provide enough resolution for reliable supertyping. To deal with this issue in downstream analyses based on supertypes, we either used a simplified approach in which DRB variants were grouped into types consisting of variants with identical amino acids at PSSs (this analysis is reported in Table 1) or excluded DRB altogether (*SI Appendix, Table S9*).

Statistical Analysis. We used mixed-effect linear models implemented in the lmer4 (79) package in R (78) to assess the relationship between the number of expressed MHC alleles and the estimated size of individual TCR repertoires. To select the optimal model, we ranked models according to the corrected Akaike information criterion (AIC_c) with the MuMIn (80) package in R (78). Initially, fixed terms included the number of MHC class I supertypes, the number of MHC class II supertypes, sex, selection direction, age, and body mass at sampling, while random effects included selection line (within selection direction), date of sampling, and sequencing batch (however, sequencing batch yielded zero or minimal variance, and it was removed). To avoid overfitting, we constrained the number of fixed terms in the final model to a maximum of four. We also checked whether a model with MHC class I and class II supertype counts fitted simultaneously as two separate predictors showed a better fit than a summed count of supertypes across both MHC classes. P values for the optimal model were calculated using Kenward–Roger approximation of the degrees of freedom (package pbkrtest) (81) for t statistics, a method that is considered accurate, conservative, and robust to small sample sizes (82). Marginal and conditional R^2 values were calculated with MuMIn (80, 83). We further performed a series of additional statistical analyses (*SI Appendix, Supplementary Text S7 and Tables S5, S7, S9, and S10*) to check for influential points and robustness to genotyping and supertyping protocols, for example.

Analyses analogous to those carried out on supertypes were also performed with unique MHC amino acid variants (*SI Appendix, Fig. S3 and Tables S4, S6, and S8*). These corroborated the results obtained with supertypes, strengthening the main findings and supporting the robustness of the statistical analysis. In the main text, however, we present results only for supertypes, for two reasons. First, TCR selection is likely based on physicochemical properties of MHC residues involved in binding of oligopeptides more than on the amino acid sequence of the whole molecule (70). Second, the sampling strategy (individuals with either low or high combined MHC diversity) led to a higher correlation ($R^2 = 0.71$) between the number of MHC class I and II amino acid variants compared with that for supertypes ($R^2 = 0.45$). Thus, analyses for supertypes allowed a better assessment of independent effects of the two MHC classes.

ACKNOWLEDGMENTS. We thank P. Koteja for the donation of bank vole spleens; K. Dudek for running Illumina MiSeq sequencing; R. Ploski for advice on Illumina sequencing; W. Nowak for help with qPCR assays; J. Kaufman, W. Babik, D. S. Richardson, R. Martin, and K. P. Phillips for their comments on an earlier version of the manuscript; and C. Eizaguirre for an insightful question on data analysis during the XVI European Society for Evolutionary Biology Congress. This work was funded by Maestro Grant UMO-2013/08/ANZ8/00153 from the National Science Centre (NCN) (to J.R.) and supported by NCN Etiuda PhD Scholarship UMO-2017/24/TN28/00088 (to M.M.).

- Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc Biol Sci* 277:979–988.
- Kubinak JL, Ruff JS, Hyzer CW, Slev PR, Potts WK (2012) Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. *Proc Natl Acad Sci USA* 109:3422–3427.
- Phillips KP, et al. (2018) Immunogenetic novelty confers a selective advantage in host-pathogen coevolution. *Proc Natl Acad Sci USA* 115:1552–1557.
- Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci USA* 99:11260–11264.
- Oliver MK, Telfer S, Piernney SB (2009) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proc Biol Sci* 276:1119–1128.
- Falk K, Rötzschke O, Stevanović S, Jung G, Rammensee H-G (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351:290–296.
- Klein J (1986) *Natural History of the Major Histocompatibility Complex* (Wiley, New York).
- Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE (2013) IPD–The immuno polymorphism database. *Nucleic Acids Res* 41:D1234–D1240.
- Nowak MA, Tarczy-Hornoch K, Austyn JM (1992) The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci USA* 89:10896–10899.
- Woelfling B, Traulsen A, Milinski M, Boehm T (2009) Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philos Trans R Soc Lond B Biol Sci* 364:117–128.
- Vidović D, Matzinger P (1988) Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* 336:222–225.
- Nikolich-Zugich J, Slifka MK, Messaoudi I (2004) The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* 4:123–132.
- Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat Rev Immunol* 14:377–391.
- Borghans JAM, Noest AJ, De Boer RJ (2003) Thymic selection does not limit the individual MHC diversity. *Eur J Immunol* 33:3353–3358.

15. Madsen T, Ujvari B (2006) MHC class I variation associates with parasite resistance and longevity in tropical pythons. *J Evol Biol* 19:1973–1978.
16. Wegner KM, Kalbe M, Kurtz J, Reusch TBH, Milinski M (2003) Parasite selection for immunogenetic optimality. *Science* 301:1343.
17. Westerdahl H, Asghar M, Hasselquist D, Bensch S (2012) Quantitative disease resistance: To better understand parasite-mediated selection on major histocompatibility complex. *Proc Biol Sci* 279:577–584.
18. Kalbe M, et al. (2009) Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proc Biol Sci* 276:925–934.
19. Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Mol Ecol* 19:255–265.
20. Radwan J, et al. (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Mol Ecol* 21:2469–2479.
21. Sepil I, Lachish S, Hinks AE, Sheldon BC (2013) Mhc supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. *Proc Biol Sci* 280:20130134.
22. Harf R, Sommer S (2005) Association between major histocompatibility complex class II DRB alleles and parasite load in the hairy-footed gerbil, *Gerbillurus paeba*, in the southern Kalahari. *Mol Ecol* 14:85–91.
23. Biedrzycka A, et al. (2018) Blood parasites shape extreme MHC diversity in a migratory passerine. *Mol Ecol* 27:2594–2603.
24. Stiebens VA, Merino SE, Chain FJ, Eizaguirre C (2013) Evolution of MHC class I genes in the endangered loggerhead sea turtle (*Caretta caretta*) revealed by 454 amplicon sequencing. *BMC Evol Biol* 13:95.
25. Hablützel PI, et al. (2014) Intermediate number of major histocompatibility complex class IIB length variants relates to enlarged perivisceral fat deposits in the blunt-head cichlid *Tropheus moorii*. *J Evol Biol* 27:2177–2190.
26. Bonneaud C, Mazuc J, Chastel O, Westerdahl H, Sorci G (2004) Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution* 58:2823–2830.
27. Six A, et al. (2013) The past, present, and future of immune repertoire biology—The rise of next-generation repertoire analysis. *Front Immunol* 4:413.
28. Migalska M, Sebastian A, Konczal M, Kotlik P, Radwan J (2017) De novo transcriptome assembly facilitates characterisation of fast-evolving gene families, MHC class I in the bank vole (*Myodes glareolus*). *Heredity (Edinb)* 118:348–357.
29. Bryja J, Galan M, Charbonnel N, Cosson JF (2006) Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). *Immunogenetics* 58:191–202.
30. Axtner J, Sommer S (2007) Gene duplication, allelic diversity, selection processes and adaptive value of MHC class II DRB genes of the bank vole, *Clethrionomys glareolus*. *Immunogenetics* 59:417–426.
31. Scherman K, Råberg L, Westerdahl H (2014) Positive selection on MHC class II DRB and DQB genes in the bank vole (*Myodes glareolus*). *J Mol Evol* 78:293–305.
32. Shugay M, et al. (2014) Towards error-free profiling of immune repertoires. *Nat Methods* 11:653–655.
33. Migalska M, Sebastian A, Radwan J (2018) Profiling of the TCR β repertoire in non-model species using high-throughput sequencing. *Sci Rep* 8:11613.
34. Lessells CM, Boag PT (1987) Unrepeatable repeatabilities: A common mistake. *Auk* 104:116–121.
35. Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43:783–791.
36. Qi Q, et al. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci USA* 111:13139–13144.
37. Vanhanen R, et al. (2016) T cell receptor diversity in the human thymus. *Mol Immunol* 76:116–122.
38. Casrouge A, et al. (2000) Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. *J Immunol* 164:5782–5787.
39. Sidney J, Grey HM, Kubo RT, Sette A (1996) Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunol Today* 17:261–266.
40. Kindt TJ, Goldsby RA, Osborne BA, Kuby J (2007) *Kuby Immunology* (Freeman). Available at <https://books.google.com/books?id=oOsf2WfE5wC&pgis=1>. Accessed June 4, 2016.
41. Hsieh C-S, Lee H-M, Lio C-WJ (2012) Selection of regulatory T cells in the thymus. *Nat Rev Immunol* 12:157–167.
42. Sinclair C, Bains I, Yates AJ, Seddon B (2013) Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proc Natl Acad Sci USA* 110:E2905–E2914.
43. Yap JY (2017) Quantitative dissection of T cell negative selection mechanisms in the thymus. PhD dissertation (The Australian National University, Canberra, Australia).
44. Sakaguchi S (2004) Naturally arising CD4⁺ regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu Rev Immunol* 22:531–562.
45. Garg SK, et al. (2014) Aging is associated with increased regulatory T-cell function. *Aging Cell* 13:441–448.
46. Nishikawa H, Jäger E, Ritter G, Old LJ, Gnjatovic S (2005) CD4⁺ CD25⁺ regulatory T cells control the induction of antigen-specific CD4⁺ helper T cell responses in cancer patients. *Blood* 106:1008–1011.
47. Triguñaité A, Dima J, Jørgensen TN (2015) Suppressive effects of androgens on the immune system. *Cell Immunol* 294:87–94.
48. Foo YZ, Nakagawa S, Rhodes G, Simmons LW (2017) The effects of sex hormones on immune function: A meta-analysis. *Biol Rev Camb Philos Soc* 92:551–571.
49. Roden AC, et al. (2004) Augmentation of T cell levels and responses induced by androgen deprivation. *J Immunol* 173:6098–6108.
50. O'Connor EA, Cornwallis CK, Hasselquist D, Nilsson J-Å, Westerdahl H (2018) The evolution of immunity in relation to colonization and migration. *Nat Ecol Evol* 2:841–849.
51. Hughes AL, Nei M (1989) Evolution of the major histocompatibility complex: Independent origin of nonclassical class I genes in different groups of mammals. *Mol Biol Evol* 6:559–579.
52. Crew MD, Bates LM, Douglass CA, York JL (1996) Expressed *Peromyscus maniculatus* (Pema) MHC class I genes: Evolutionary implications and the identification of a gene encoding a Qa1-like antigen. *Immunogenetics* 44:177–185.
53. Hurt P, et al. (2004) The genomic sequence and comparative analysis of the rat major histocompatibility complex. *Genome Res* 14:631–639.
54. Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics* 56:683–695.
55. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652.
56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
57. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* 33:D256–D261.
58. Mamedov IZ, et al. (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol* 4:456.
59. Barber EK, Dasgupta JD, Schlossman SF, Trevillyan JM, Rudd CE (1989) The CD4 and CD8 antigens are coupled to a protein-tyrosine kinase (p56lck) that phosphorylates the CD3 complex. *Proc Natl Acad Sci USA* 86:3277–3281.
60. Smith-Garvin JE, Koretzky GA, Jordan MS (2009) T cell activation. *Annu Rev Immunol* 27:591–619.
61. Sadowska ET, Baliga-Klimczyk K, Chrzęściak KM, Koteja P (2008) Laboratory model of adaptive radiation: A selection experiment in the bank vole. *Physiol Biochem Zool* 81:627–640.
62. Sebastian A, Herdegen M, Migalska M, Radwan J (2016) AMPLISAS: A web server for multilocus genotyping using next-generation amplicon sequencing data. *Mol Ecol Resour* 16:498–510.
63. Biedrzycka A, Sebastian A, Migalska M, Westerdahl H, Radwan J (2017) Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Mol Ecol Resour* 17:642–655.
64. Laydon DJ, Bangham CRM, Asquith B, Crm B (2015) Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos Trans R Soc Lond B Biol Sci* 370:20140291.
65. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135:183–191.
66. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor β -chain diversity in alphabeta T cells. *Blood* 114:4099–4107.
67. Laydon DJ, et al. (2014) Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput Biol* 10:e1003646.
68. Bustin SA, et al. (2009) The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55:611–622.
69. Doytchinova IA, Flower DR (2005) In silico identification of supertypes for class II MHCs. *J Immunol* 174:7085–7095.
70. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* 41:2481–2491.
71. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170.
72. Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proc Natl Acad Sci USA* 86:958–962.
73. Schwensow N, Fietz J, Dausmann KH, Sommer S (2007) Neutral versus adaptive genetic variation in parasite resistance: Importance of major histocompatibility complex supertypes in a free-ranging primate. *Heredity (Edinb)* 99:265–277.
74. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
75. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet* 11:94.
76. Jombart T (2008) ADEGENET: A R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405.
77. Jombart T, Ahmed I (2011) ADEGENET 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071.
78. R Core Team (2017) R: A Language and Environment for Statistical Computing. Version 3.4.3. Available at <https://www.r-project.org/>. Accessed January 7, 2018.
79. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48.
80. Barton K (2016) MuMIn: Multi-Model Inference. Version 1.19. Available at <https://cran.r-project.org/web/packages/MuMIn/index.html>=MuMIn. Accessed February 8, 2018.
81. Halekoh U, Hojsgaard S (2014) A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package pbrtest. *J Stat Softw* 59:1–30.
82. Luke SG (2017) Evaluating significance in linear mixed-effects models in R. *Behav Res Methods* 49:1494–1502.
83. Nakagawa S, Schielzeth H (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol Evol* 4:133–142.