The T-cell receptor repertoire of wild mice

Jacob Cohen[1,3,4], Simon Hunter-Barnett[1,3], Gayathri Nageswaran[2], Suzanne Byrne[2], Gemma Freeman[2], Matthew Cowley[2], Benny Chain[2] & Mark Viney[1*]

[1]Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool, L69 7ZB, UK

[2]Division of Infection and Immunity, University College London, WC1E 6BT, UK

[3]These authors contributed equally to this work.

[4]Present address, Department of Biology, University of Oxford, Oxford, OX1 3RB.

*** Correspondence:**
Corresponding Author
Mark.Viney@liverpool.ac.uk

Keywords: mouse, wild, T-cell, receptor, diversity

ABSTRACT

Here we characterize the T-cell receptor (TCR) repertoire of wild mice to provide a resource for eco-immunology and to better understand wild animals' immune state. While laboratory mice are central to immunological research, their immune systems differ significantly from those of their wild counterparts who are exposed to more intense immunological stimulation from a broader range of infections. We performed high-throughput sequencing of the TCR alpha and beta chains of $CD4^+$ and $CD8^+$ T-cells isolated from 65 wild *Mus musculus domesticus* captured at two UK sites. We analysed repertoire richness and diversity in relation to mouse age, sex, and sample location. The results show that wild mice have large TCR repertoires. We found that repertoire richness, which measures the breadth of the repertoire, was not significantly affected by mouse age or sex, suggesting that wild mice maintain the capacity to respond to novel antigens throughout their lives. In contrast, repertoire diversity (measured by Shannon's index) was affected by mouse sex and age. This low diversity, coupled with constant richness, points to older mice having comparatively more highly abundant clones in their repertoires, likely because of chronic exposure to persistent pathogens in their environment. Individual mice shared a considerable number of TCR sequences, with greater sharing observed between mice from the same location, suggesting that local environmental pressures shape the TCR repertoire. These findings provide a novel and valuable description of the wild mouse TCR, revealing an immune system that balances maintaining a broad response capacity with developing strong, lasting responses to infections in the natural environment.

INTRODUCTION

The vertebrate immune system protects individuals from infection and disease. Most of our description and understanding of this system comes from study of the laboratory mouse, but laboratory mice aren't exposed to the infections and other threats of the natural environment, which is the context in which the immune system and its function have evolved. There has been recent interest in studying the immune system and immune state of wild mice and other rodents (Viney & Riley 2017), finding that in wild mice the immune system is in an altered, often more activated, state compared with that of laboratory animals (Beura *et al*. 2016; Abolins *et al*. 2017; Abolins *et al*. 2018). Other studies have used rewilded mice, that is laboratory mice released into outdoor enclosures, and have found that they have an altered immune response that, for example, makes them susceptible to a parasitic nematode infection, while laboratory-maintained mice are resistant to it (Leung *et al.* 2018). Transplanting laboratory mice with wild mouse microbiomes also brings about profound immunological changes, reducing inflammation and, for example, increasing susceptibility to infection with influenza virus (Rosshart *et al*. 2017). The unifying thread from these studies is that laboratory mice are underexposed to infections compared with wild / rewilded / wild microbiome-exposed mice, and that greater infection stimulates the immune system into a different functional state, compared with laboratory mice.

B and T-cells are the principal players of the adaptive immune response and the dominant cell types in the lymphatic system, together (in humans) accounting for 40 % of immune cells (*c.* 25 % T-cells, 15 % B-cells) (Sender *et al*. 2023). T-cells bind to antigenic peptide molecules that cause the T-cells to proliferate and then to act directly against the antigenic source. Antigens are perceived by heterodimeric (in mice and humans principally α and β chains) antigen-specific T-cell receptors (TCRs), coded for by somatically recombined TCR genes. Each pathogen antigen is recognised by a small number of specific clones of T-cells, so that an individual animal contains a large, diverse repertoire of T-cell clones. T-cell responses to individual pathogens are idiosyncratic (Weng 2023); for example, in people infected with *Mycobacterium tuberculosis* individuals' $CD4^+$ T-cells recognised an average of 24 different epitopes, but individuals varied extensively in the epitopes that they recognised (Lindestam Arlehamn *et al*. 2016).

Next generation parallel sequencing of TCR-coding genes and transcripts has been used to investigate the TCR-coding sequences present among circulating T-cells in people and in laboratory mice. Theoretically, the TCR repertoire is enormous: in people potentially up to 10^{15} alpha and beta TCRs, though reduced to some 10^{13} after thymic selection, and then absolutely limited by the number of T-cells present in an individual, some 10^{12} in a person and 10^{8} in a mouse (Qi *et al*. 2014; Nikolich-Žugich *et al.* 2004). Sequence-based analyses are therefore only sampling a small fraction of the actual T-cells present, and so only sampling a small sub-fraction of the TCR repertoire (Laydon *et al*. 2015). Many studies have confirmed that people and laboratory mice have large TCR repertoires. The clone size distribution of the repertoire is approximated well by a power law, with the vast majority of T-cell receptors found at very low frequency (typically observed once in a sample), but a proportion of large clones taking up a disproportionate part of the overall repertoire (Oakes *et al*. 2017; Sun *et al*. 2017). These distribution profiles can vary quantitatively between $CD4^+$ and $CD8^+$ cells, between naïve and memory cells, and as individuals age (Britanova *et al*. 2016; Oakes *et al*. 2017). Specifically, as people age there is a reduction in overall TCR diversity, driven in part by a decrease in the number of TCR-defined clones in naïve, but not in memory CD4+ cells, with similar changes in CD8+ cells, though they have a comparatively smaller TCR repertoire (Song *et al*. 2025; Weng 2023). The decrease in diversity of memory cells is driven by an increase in the relative abundance of some clones; indeed, in some individuals individual clones specific to a chronic virus

such as Cytomegalovirus can occupy a significant proportion of the total repertoire (Lindau *et al.* 2019; Khan *et al.* 2002).

Here we sought to investigate the TCR repertoire of wild mice to complement other studies that have investigated wild mouse immune state. We present these data as a biological resource that is particularly relevant to eco-immunology and to laboratory mouse immunology. We relate the wild TCR repertoires to mouse age, sex, and other metadata.

MATERIALS AND METHODS

Animals

We live trapped mice, *Mus musculus domesticus* between May and September 2023 at two sites: (i) a pig farm in Nottinghamshire and (ii) a dairy farm on the Wirral peninsula, both in the UK. Mice were killed by cervical dislocation, sexed, and weighed. We aged mice by determining the mass of their dried eye lenses after Rowe *et al*. 1985; this method has been validated against mice of known age. We also analysed the repertories of 6 female CD-1 laboratory mice.

Splenocytes, FACS, and isolation of RNA

From each mouse we aseptically isolated the spleen, determined its mass and then passed it through a 70 µm sieve to create a single-cell suspension. We resuspended the splenocytes in blood cell lysis buffer (Invitrogen) held at room temperature for 5 minutes, supplemented with an excess of phosphate buffered saline (PBS), sedimented the cells by centrifugation, washed them in PBS by resuspension and sedimentation, after which the cells were resuspend in 5 mL of staining buffer (Biolegend). We determined the number of splenocytes using a haemocytometer. We removed a portion of the cells as the FACS negative control. To the remainder we blocked Fc receptors using TruStain FcX™ PLUS on ice for 10 minutes and then stained the cells by addition of the following conjugated antibodies: anti-CD3-PE, anti-CD4-FITC, and CD8-APC for 20 minutes in the dark, after which the cells were washed twice in staining buffer. Immediately before FACS the cells were additionally stained with DAPI.

Samples were analysed and separated using a FACS Aria III Cell Sorter (BD Bioscience, Oxford, UK). The gating strategy was incorporated from dot plots with the following hierarchy: scatter gates (SSC/FCS) to identify mammalian cell clusters, distinct from debris with low particle size and granularity that were expected to be excluded from the gate. A singlet gate was applied (FSC-H / FSC-A) to remove doublets / multiplets that deviated from cluster linearity. From the singlet population, particles with negative signal toward the DAPI stain were considered membrane-intact and hence gated as live cells. The $CD3^+$ population was used to identify T-cells, and finally $CD4^+$ (negative to CD8) or $CD8^+$ (negative to CD4) to detect and sort corresponding T-cell subsets (**Supplementary Material 1**). On-site data visualization, gating, and sort command were performed using FACS Diva software, version 8.1.

The resulting $CD4^+$ and $CD8^+$ sorted cells were concentrated by centrifugation, and the supernatant removed and RLT buffer (Qiagen) with 1 % (v/v) 2 mercaptoethanol added to the pellet of cells, and mixed by pipetting to form a lysate, which was then processed through a QIAshredder homogeniser spin column following the manufacturer's methods, with the resultant RNA stored at -80°C.

TCR sequencing and analysis

From the isolated RNA, TCR alpha and beta genes were sequenced using an established quantitative sequencing pipeline, which uses unique molecular barcodes to correct for PCR bias, and sequencing errors (Uddin *et al.* 2019; Oakes *et al.* 2017). The fastq files were processed and annotated using an in-house open source computational pipeline Decombinator V4 (Peacock *et al.* 2021; Decombinator 2024). We excluded any TCR nucleotide sequences that did not contain a CDR3 sequence.

Data Analysis

Each TCR nucleotide sequence defines a different T-cell clone, where the number of occurrences of an identical nucleotide sequence is a measure of the abundance of that T-cell clone. Different TCR nucleotide sequences can code for the same TCR amino acid sequence. In the analyses that follow

we therefore refer to, and differentiate, nucleotide-defined TCRs and amino acid-defined TCRs. Amino acid-defined TCR sequences were the predicted translation of the concatenation of the nucleotide sequences of the variable gene segment (V), joining gene segment (J) and CDR3 region.

We quality controlled the amino acid-defined TCR sequence data by examining the number of sequences obtained for alpha and beta chains of $CD4^+$ and $CD8^+$ cells and then excluding data for a chain where there were fewer than 5,000 TCR molecules that passed our quality control for that chain. We also examined the frequency distribution of amino acid sequences and fitted a power law distribution to it to generate an alpha coefficient, and then examined the relationship of this with wild mouse age. To test whether the frequency distribution of amino acid sequences for individual mice matched the power law distribution we used the Kolmogorov-Smirnov test with 500 simulations. Mice were classified into three categories based on the resulting p values: Fits power law ($p > 0.1$), Borderline fit to power law ($0.1 > p > 0.05$), Does not fit power law ($p < 0.05$). Results were summarised for each receptor type. The slope of the distribution did not vary with age.

We define TCR repertoire size as the sum of the number of occurrences of all amino acid-defined TCR sequences, separately for alpha and beta genes from $CD4^+$ and $CD8^+$ cells. We also calculated the number of unique amino acid-defined TCR sequences for each of the alpha and beta chains of $CD4^+$ and $CD8^+$ cells.

We calculated the richness of the TCR repertoire as the number of unique amino acid-defined TCRs as a proportion of the repertoire size, which corrects for the depth of sequencing of the samples. For the wild mice we tested for differences in the richness of the TCR repertoire between the alpha and beta chains of $CD4^+$ and $CD8^+$ cells using a gamma-distributed generalised linear model (GLM) with TCR richness as the response variable and RECEPTOR TYPE (*i.e.* $CD4^+$ alpha, $CD4^+$ beta, $CD8^+$ alpha, $CD8^+$ beta), SAMPLE SITE (Nottingham *vs*. Wirral), AGE, SEX, and an AGE*SEX interaction as explanatory variables. We calculated estimated marginal means (EMMs), which corrects for multiple comparisons, from the GLMs to test *post hoc* for differences between receptor types.

We calculated two diversity indices of the wild mouse TCRs: Shannon's diversity index and Simpson's diversity index. These metrics have recently been identified as being appropriate to quantify richness and evenness in TCR repertories (Mika *et al.* 2025). For both we down-sampled each mouse's repertoire to that of the smallest, relevant repertoire to account for differences in repertoire size between mice. To do this, we first replicated all TCR amino acid-defined sequences by their number of occurrences, then randomly sampled this list without replacement to the size of the smallest, relevant repertoire, and then re-compiled the TCR amino acid-defined sequences with their total number of occurrences. This down-sampling process was done separately for the two diversity measures.

Shannon's diversity index was calculated as

$$Shannon's = -\sum_{i=1}^{R} p_i \ln(p_i)\,, \tag{1}$$

where R is the repertoire size, p_i the proportional abundance of the ith TCR amino acid-defined sequence.

Simpson's diversity index was calculated as

$$\lambda = \sum_{i=1}^{R} p_i^2 , \quad (2)$$

where R is the repertoire size and p_i is the proportional abundance of the ith TCR amino acid-defined sequence. Analyses were conducted as for Shannon's diversity index. We log transformed Simpson's diversity index as

$$Simpson's = \log\left(\frac{1}{\lambda}\right). \quad (3)$$

Thus, larger values indicate higher diversity.

We tested for differences in Shannon's and Simpson's diversity among the four receptor types (above) using a GLM (gaussian distribution) with Shannon or Simpson's index as the response variable and RECEPTOR TYPE, SAMPLE SITE, AGE, SEX and an AGE*SEX interaction as explanatory variables. We calculated EMMs from the GLMs to test *post hoc* for differences between: (i) receptor types and (ii) diversity indices in male and female mice.

To investigate the degree of sharing of amino acid-defined TCR sequences between all pairwise combinations of the wild and laboratory mice, we first down-sampled each mouse's TCR amino acid-defined repertoire to that of the smallest, relevant repertoire, as described above. We then compiled these repertoires into lists of unique amino acid-defined TCRs and performed pairwise comparisons between mice to determine the number of shared amino acid sequences between mouse pairs. From this we generated heatmaps of the degree of sharing, supplemented with dendrograms calculated from distance matrix data, which then arranged more similar individual mice together. We also examined these data using a principal coordinate analysis (PCoA) by calculating a re-scaled pairwise dissimilarity (with the most similar pair having a dissimilarity of 0), which was used in a weighted PCoA separately for alpha and beta chains of $CD4^+$ and $CD8^+$ cells.

To investigate whether mouse sex and age affected the extent to which mice shared amino acid-defined TCR sequences, we generated pairwise sharing data for all combinations of mice; we restricted this to only the Nottingham-derived mice to avoid any potentially effects of sample site. We then tested the effects of (i) pairwise mouse sex (same sex *vs.* different sex), (ii) pairwise difference in mouse age, and (iii) pairwise sum of mouse age on the number of pairwise shared amino acid-defined TCR sequences with a GLM (gaussian distribution). The response variable for the GLM was the number of pairwise shared amino acid-defined TCR sequences, while the explanatory variables were RECEPTOR TYPE, pairwise SEX DIFFERENCE, pairwise AGE DIFFERENCE, pairwise AGE SUM, a pairwise AGE DIFFERENCE*RECEPTOR TYPE interaction, and a pairwise AGE SUM*RECEPTOR TYPE interaction. We calculated slope estimates for the explanatory variables from the GLM using EMMs to test *post hoc* for differences between receptor types.

We conducted all data analysis in R (R Core Team 2025). Model diagnostics were carried out using the 'DHARMa' package (Hartig 2024). EMMs *post hoc* testing was carried out using the 'emmeans' package (Lenth 2025). PCoAs were conducted using the 'vegan' package (Oksanen *et al*. 2025). In addition, we used the 'dplyr' (Wickham *et al*. 2023), 'stringr' (Wickham 2023), 'purrr' (Wickham & Henry 2025), 'tidyr' (Wickham *et al*. 2024) poweRlaw (Gillespie 2015), here (Müller 2020), and 'modeest' (Poncet, 2019) packages. Plots were generated in R using packages ggplot2 (Wickham 2016), gplots (Warnes *et al*. 2024), showtext (Qiu 2024), patchwork (Pederson 2024), ggpubr (Kassambara 2023) and ggbeeswarm (Clarke *et al.* 2023).

RESULTS

Animals

We caught 69 mice (30 male, 39 female), 65 from the Nottingham site and four from the Wirral site. The average age of the wild mice was 13 weeks (SD 10, median 9, range 2 – 49); average body mass was 15.5 g (SD 4, median 11, range 7 – 27). Among the wild mice, the spleens had an average mass of 53 mg (SD 33, range 10 – 170), yielding an average of 16 million splenocytes (SD 13, range 8.7×10^4 – 5.1×10^7). There was a positive relationship between mouse age and spleen mass (R^2 = 0.3, $p < 0.0001$) and number of splenocytes (R^2 = 0.13, $p < 0.0025$), and between spleen mass and number of splenocytes (R^2 = 0.49; p , 0.0001) (**Supplementary Material 2**). Previous comparisons of wild and laboratory mice have found that wild mice are smaller than laboratory mice, and that their spleens are absolutely smaller (approximately one third of the mass) and with commensurate fewer cells, but also proportionally smaller (compared with body mass), compared with those of laboratory mice (Abolins *et al*. 2017).

The FACS of splenocytes resulted in an average of 8 million $CD3^+$ T-cells per wild mouse (SD 8 million, range 1.1×10^4 – 4.9×10^7). There were proportionally more $CD4^+$ cells than $CD8^+$ cells, with an average of 62 % (SD 12, range 34 – 82), giving an average $CD4^+$: $CD8^+$ ratio of 2.4 (SD 1.4, range 0.2 - 6.9). For four female mice from the Nottingham site insufficient cells were collected for TCR sequence analysis, refining our sample size to 65 (30 male, 35 female) wild mice and 6 lab mice.

TCR sequences

To quality control the data we examined the number of alpha and beta chain coding sequences for $CD4^+$ and $CD8^+$ cells separately, and excluded the data for one cell type if there were fewer than 5,000 TCR molecules of either the alpha or beta chain, which resulted in the exclusion of 9 and 13 mice, respectively. This resulted in final sample sizes of 62 (56 wild, 6 lab) for $CD4^+$ alpha and beta, and 58 (52 wild, 6 lab) $CD8^+$ alpha and beta chain TCR sequences that were analysed further.

After our data quality control the minimum TCR repertoire size among the wild mice was 54,908, 58,772, 66,881 and 64,224, and for the laboratory mice the minimum 68,637, 79,052, 49,379 and 75,640 for $CD4^+$ alpha and beta and $CD8^+$ alpha and beta chains, respectively. We examined the frequency distribution of the amino acid-defined TCR sequences, which showed that many of these followed a power law distribution, which has frequently been observed on other TCR datasets (**Supplementary Material 3**).

TCR repertoire richness and diversity

We investigated the richness of the TCR repertoire, finding that among the wild mice beta chain richness was greater than that of the alpha chain, for both within and between $CD4^+$ and $CD8^+$ cell comparisons (**Supplementary Material 4**). The comparatively larger repertoire of beta chains, observed previously in human T-cells (Oakes *et al*. 2017), occurs because they are encoded by recombination of V, D and J genes, unlike the V and J recombination of alpha chain coding genes.

There was no significant effect of wild mouse AGE, SAMPLE SITE, SEX, or AGE*SEX on the TCR repertoire richness. However, the TCR diversity measured by Shannon's index (after down-sampling each mouse's repertoire to that of the smallest, relevant repertoire) differed significantly between the $CD4^+$ beta and the $CD8^+$ alpha chains (**Figure 1; Supplementary Material 5**), but also that there were significant effects of SEX ($p = 0.0058$), SAMPLE SITE ($p = 0.0063$), SEX*AGE ($p = 0.0396$), and a marginally non-significant effect of AGE ($p = 0.0762$) with older mice tending to have a lower diversity. The difference between Shannon's index of male and female mice was marginally non-significant (p

= 0.0859). We obtained broadly similar results using Simpson's diversity index though with more significant differences in diversity between receptor types (**Supplementary Material 6**), significant effects of SEX (p = 0.0452), and AGE (p = 9.91 x 10-5), with older mice having lower diversity (**Supplementary Material 7**). The difference between Simpson's index of male and female mice was also marginally non-significant (p = 0.0884).

TCR networks

We investigated how much different wild and laboratory mice shared the same amino acid-defined TCR sequence, which we did by counting the number of common sequences between every pairwise combination of mice. This showed that there was considerable sharing of TCR sequences, with this greatest for alpha chains (5.9 and 5.2 % for $CD4^+$ and $CD8^+$, respectively) than for beta chains (2.9 and 2.2 % for $CD4^+$ and $CD8^+$, respectively) (**Figure 2**; **Supplementary Material 8**). We asked how many TCR sequences were public, which we defined as those shared among $\geq$75 % of the mice (both wild and laboratory). We found that this ranged from 157 (for $CD8^+$ beta) to 1,113 (for $CD4^+$ alpha), but that this was always less than 0.25 % of unique TCRs (**Supplementary Material 8**). Previous studies of public TCR sequences in 28 laboratory mouse found that ~0.5 % of sequences were shared among 75 % of them (Madi *et al*. 2014). A PCoA analysis showed that there are differences in the sharing of sequences among mice from the different wild sampling sites, and between wild mice and laboratory mice (**Supplementary Material 9**).

We investigated whether mouse sex and age affected the extent to which mice shared amino acid-defined TCR sequences. Mindful of the differences in sharing among mice from the different wild sites, we focussed this analysis on only the mice from the Nottingham site, and used a GLM that analysed the number of shared sequences by (i) same sex *vs.* different sex pairs, (ii) sum of age of mouse pairs, and (ii) difference in age of mouse pairs. This showed that overall there was no difference in the number of shared TCR sequences between same sex and different sex pairs (p = 0.69), and that the sum of mouse pair age was not significant (p = 0.47), nor for the difference in mouse pair age (p = 0.122), (**Figure 3; Supplementary Material 10**). However, *post* hoc analysis of the four different receptor types showed that for $CD8^+$ alpha and beta receptors there were consistent negative relationships of the number of shared sequences and the sum or difference of age (**Supplementary Material 11**). This was also observed for CD4+ beta chains, but not alpha.

DISCUSSION

The aim of this work was to describe the TCR repertoire of wild mice, to act as a biological resource in support of the fields of eco-immunology and mouse immunology more generally. We successfully sampled wild mice and obtained TCR sequences from $CD4^+$ and $CD8^+$ cells. Despite wild mice being smaller than laboratory mice and having absolutely and proportionally smaller spleens with fewer splenocytes (Abolins *et al*. 2017) we successfully obtained sufficient $CD4^+$ and $CD8^+$ cells from 65 wild mice for TCR sequence analysis.

Wild mice have large TCR repertoires and their richness was not affected by individuals' sex, age, or the site from which the mice were obtained. Richness (which is the normalised number of unique TCRs in a repertoire) measures the breadth of the TCR repertoire. We hypothesise, therefore, that both male and female mice maintain the capacity to respond to a wide range of antigens in both $CD4^+$ and $CD8^+$ T-cell compartments throughout their natural lifespan in the wild. This contrasts with some studies in humans that report declining richness with age (Weng, 2023). These differences between wild mice and people may reflect some immunological differences between wild mouse and human populations. But in addition to this (i) our sampling of wild mice will have under-sampled very young mice (which typically will not leave their nests), so that we may not have been able to account for TCR richness in very young individuals, and (ii) the maximum lifespan of wild mice is not equivalent to that of elderly people included in these studies and / or the relative rarity of old mice limits our statistical power.

In contrast to TCR repertoire richness, measures of diversity (such as Shannon and Simpson's indices) assess both the number of different TCRs and their relative abundance of those, such that when there are more abundant clones then the diversity is lower. We found that the TCR diversity was affected by sex (but that there was no significant difference between males and females), and declined with age. In the context of constant richness, decreased diversity must reflect an increase in the abundance of some clones, which we hypothesise may arise from chronic exposure to a limited set of infections prevalent in wild mouse populations. An influence of sex on diversity has been reported previously in humans, with greater diversity in females (Trofimov *et al.* 2022). The mechanism behind these effects remains unclear, but could reflect hormonal differences leading to differences in thymic selection (Trofimov *et al.* 2022). Decreased diversity with age has also been described repeatedly in humans (Britanova *et al.* 2016). In wild mice the effects of sex and age may reflect differences among mice in their exposure to infection, with higher diversity a consequence of less exposure to infection. Thus young mice will not yet have had time to experience many infections unlike older mice. Male house mice may have more limited social interactions because they typically disperse from the site of their birth (Pocock *et al.* 2005) leading to comparatively less intense or less chronic exposure to infection in their environment. In contrast, females breed communally (Manning *et al*. 1995), which may expose them to greater levels of infection and antigenic challenge. Taken together, our results suggest that wild mice maintain a breadth of TCR repertoire that allows them to react to a breadth of infections but, particularly in older, female mice, this is evidence of strong, life-long immune responses to a limited set of persistent infections in their environment.

We found that individual mice shared a considerable number of TCR sequences; specifically *c.*5 % of alpha chains and *c.*2 % of beta chains. The differences in the amount of TCR sharing between alpha and beta chains is, presumably, because beta chains are comparatively more diverse, making the chance of them being shared less likely compared with alpha chains. We found a pattern of greater sharing between mice from within the same group (from Nottingham, from the Wirral, or laboratory mice) than between groups, which is suggestive of diverse antigenic experience of mice – leading to different TCR sequences – in the different environments of the groups of mice. This is

consistent with previous work that found immunological differences between wild mice from different sample sites (Abolins *et al.* 2018).

We also found that wild mouse age affects the number $CD8^+$ TCR sequences that are shared among mice. We suggest that this occurs because as mice age they become more idiosyncratic in their amino acid-defined TCR sequences, so that older mice share fewer sequences with younger mice (the effect of the difference in mouse age) and that older mice share fewer sequences with other, older mice (the effect of the sum of mouse age). Idiosyncratic responses of laboratory mice to protein antigens and viral infection has been reported previously (Sun *et al.* 2017; Milighetti *et al.* 2023). Study of wild mouse (*Mus musculus* and *Apodemus sylvaticus*) gut microbiomes show that they differ significantly among individuals (Marsh *et al.* 2022; Hanski *et al*. 2025), and this may result in idiosyncratic antigenic experience in the gut, which may contribute to the individuality of wild mouse TCRs that we have observed.

Our study has a number of limitations. One, we obtained our samples from live-trapped mice, and this method does not effectively sample very young animals, so that these are under-represented in our study. Two, we caught wild mice at only two sample sites, and most of the caught mice came from one site. This did not give us much statistical power to investigate how TCR repertoires may vary among individuals from different sample sites. This could be addressed by further work in the future. Three, we analysed only a small number of control, laboratory mice. Surprisingly, given the central role of laboratory mouse models in the development of modern immunology, there appear to be a dearth of publicly available data on the TCR repertoires of the common strains of laboratory mice. We have therefore been unable to make a comprehensive comparison of TCR repertoires between wild and laboratory mice. Given the already established differences in immune state between wild and laboratory mice, a future, in-depth comparison between the TCR repertoires of wild and laboratory mice would be of interest.

In summary, here we describe the TCR repertoire of wild mice and present these data as a research resource to the wider community. Our analyses of the wild mouse TCR repertoire finds some patterns that are similar to human TCR repertories, though we find differences between wild male and female mice that may reflect their different lives and exposure to infection.

CONFLICT OF INTEREST

The authors have no conflicts of interest.

AUTHOR CONTRIBUTIONS

BC, MV: conceived the study and obtained the funding; SHB: undertook the field and laboratory work; SB, GN: undertook the sequencing; BC, JC, MC, GF, MV: contributed to the data analysis; BC, JC, MV: drafted the paper; JC, SHB, GN, SB, GF, MC, BC, MV edited the paper.

FUNDING

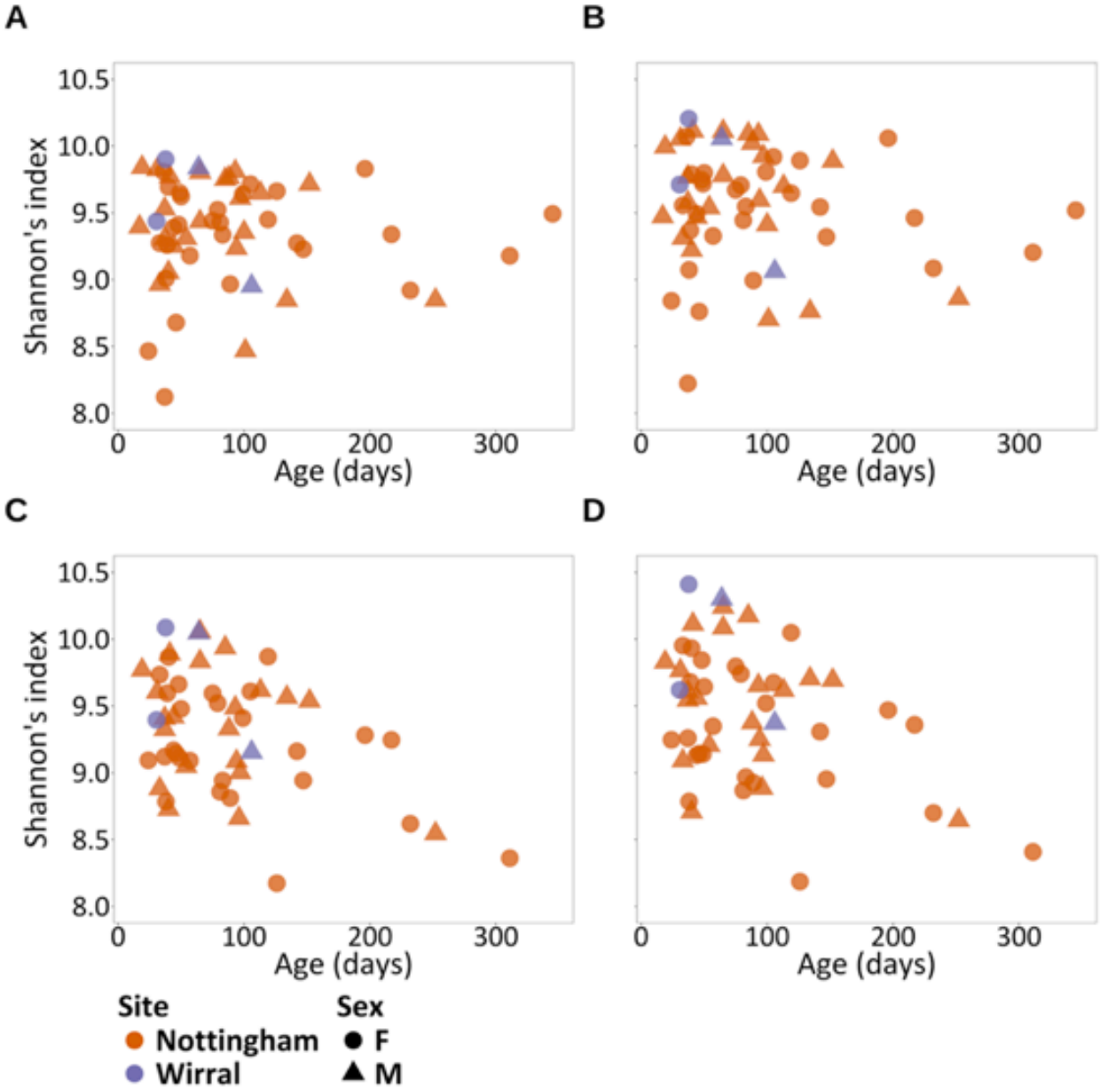
This work was funded by a NERC Exploring the Frontiers grant, NE/X010295/1.

ACKNOWLEDGMENTS

We would like to thank landowners for access; in Liverpool the staff of the Biomedical Services Unit, and Chris Law of the Cell Sorting Facility for their help.

416
417

Figure 1



418
419
420 **Figure 1**. Alpha diversity measured by Shannon's index of wild mouse, amino acid-defined TCRs
421 for (A) $CD4^+$ alpha, (b) $CD4^+$ beta, (C) $CD8^+$ alpha and $CD8^+$ beta cells.

Figure 2

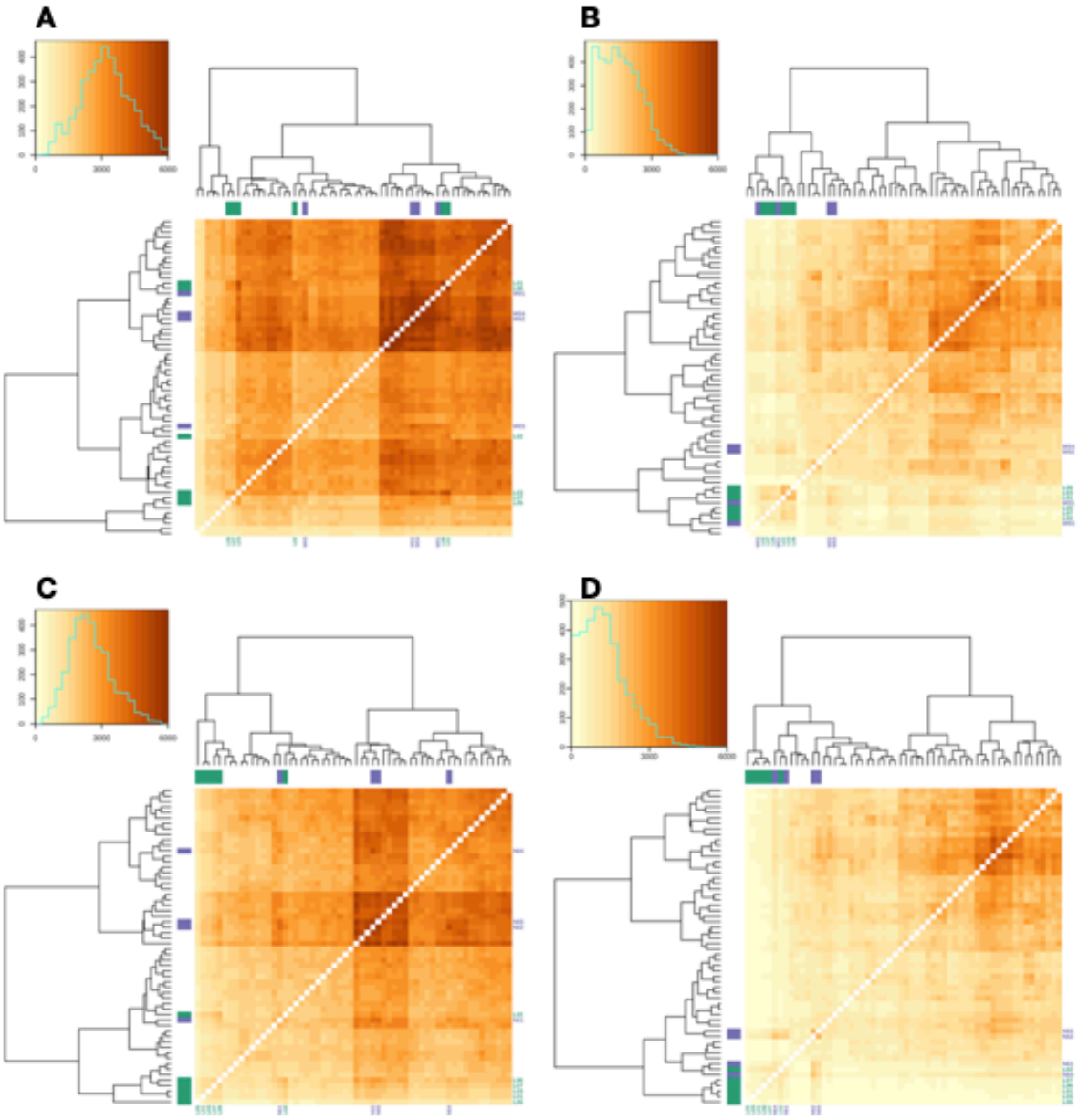


Figure 2. The sharing of amino acid-defined TCR sequences among wild and laboratory mice, with the colour scale showing the number of shared sequences for (A) $CD4^+$ alpha, (B) $CD4^+$ beta, (C) $CD8^+$ alpha, (D) $CD8^+$ beta. Against the plots, purple indicates wild mice from the Wirral site and green laboratory mice; mice where no colour is shown are wild mice from the Nottingham site. In all plots the brown scale is the same. In each, the insert plot shows a frequency distribution of the number of shared sequences; note that the y-axis is a different scale in these four plots, though the x-axis is consistent.

Figure 3

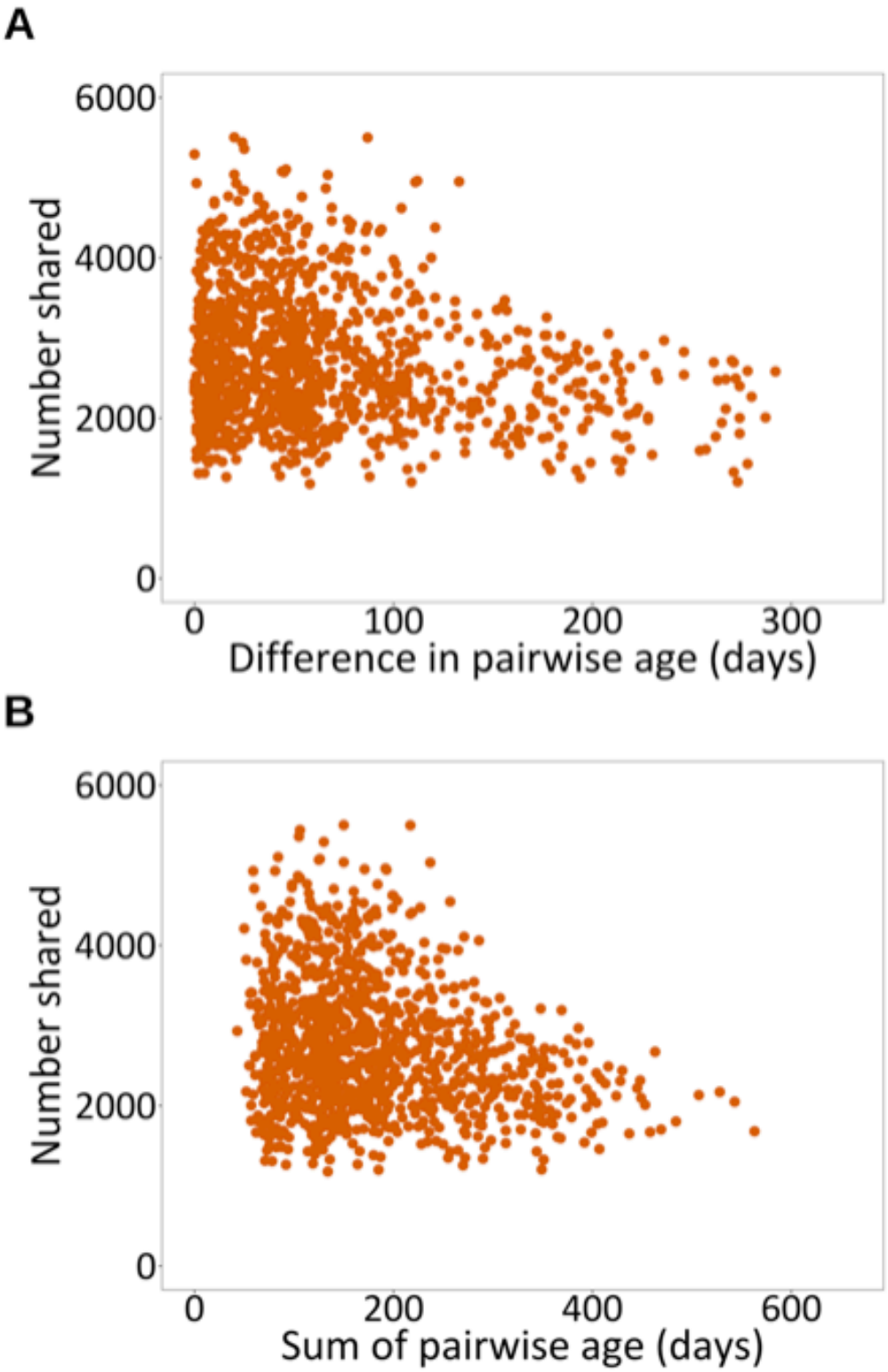


Figure 3. The number of shared amino acid-defined TCR sequences for $CD8^+$ alpha between all pairwise combinations of wild mice from the Nottingham site against (A) the difference in pairwise age and (B) the sum of the pairwise age, where age is in days. The results for $CD4^+$ alpha and beta, and $CD8^+$ beta are shown in **Supplementary Material 10**.

REFERENCES

Abolins S, King EC, Lazarou L, Weldon L, Hughes L, Drescher P, *et al.* (2017) The comparative immunology of wild and laboratory mice, *Mus musculus domesticus*. *Nature Communications* **8**:14811.

Abolins S, Lazarou L, Weldon L, Hughes L, King EC, Drescher P, *et al.* (2018) The ecology of immune state in a wild mammal, *Mus musculus domesticus. PLoS Biology* **16**:e2003538.

Beura LK, Hamilton SE, Bi K, Schenkel JM, Odumade OA, Casey KA, *et al.* (2016) Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**:512-516.

Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, *et al.* (2016) Dynamics of individual T cell repertoires: from cord blood to centenarians. *Journal of Immunology* **196**:5005-13.

Clarke E, Sherrill-Mix S, Dawson C. (2023) ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.7.2.

Decombinator (2024) https://github.com/innate2adaptive/Decombinator

Gillespie CS (2015) Fitting heavy tailed distributions: The poweRlaw Package. *Journal of Statistical Software* **64**:1-16.

Hanski E, Joseph S, Curtis MA, Swann JW, Vallier M, Linnerbrink M. *et al*. (2025) Wild house mice have a more dynamic and aerotolerant gut microbiota than laboratory mice. *BMC Microbiology* **25**:204.

Hartig F. (2024) DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.4.7.

Kassambara A (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0.

Khan N, Shariff N, Cobbold M, Bruton R, Ainsworth JA, Sinclair AJ, *et al.* (2002) Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality in healthy elderly individuals. *Journal of Immunology* **169**:1984-92.

Laydon DJ, Bangham CR, Asquith B. (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**:20140291.

Lenth R. (2025) emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.11.1.

Leung JM, Budischak SA, Chung The H, Hansen C, Bowcutt R, Neill R, *et al.* (2018) Rapid environmental effects on gut nematode susceptibility in rewilded mice. *PLoS Biology* **16**:e2004108.

Lindau P, Mukherjee R, Gutschow MV, Vignali M, Warren EH, Riddell SR, *et al.* (2019) Cytomegalovirus exposure in the elderly does not reduce CD8 T cell repertoire diversity. *Journal of Immunology* **202**:476-483.

Lindestam Arlehamn CS, McKinney DM, Carpenter C, Paul S, Rozot V, Makgotlho E, *et al.* (2016) A quantitative analysis of complexity of human pathogen-specific CD4 T cell responses in healthy *M. tuberculosis* infected South Africans. *PLoS Pathogens* **12**:e1005760.

Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, *et al.* (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Research* **24**:1603-12.

Manning JC, Dewsbury DA, Wakeland EK, Potts WK. (1995) Communal nesting and communal nursing in house mice, *Mus musculus domesticus*. *Animal Behaviour* **50**:741-751.

Marsh KJ, Raulo AM, Brouard M, Troitsky T, English HM, Allen B. *et al.* (2022) Synchronous seasonality in the gut microbiota of wild mouse populations. *Frontiers in Microbiology* **13**:809735.

Mika J, Polanska A, Blenman KR, Pusztai L, Polanska J, Candéias S, *et al.* (2025) A comprehensive evaluation of diversity measures for TCR repertoire profiling. *BMC Biology* 23:133.

Milighetti M, Peng Y, Tan C, Mark M, Nageswaran G, Byrne S, *et al.* (2023) Large clones of pre-existing T cells drive early immunity against SARS-COV-2 and LCMV infection. *iScience* **26**:106937.

Müller K (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1.

Nikolich-Žugich J, Slifka M, Messaoudi I. (2004) The many important facets of T-cell repertoire diversity. *Nature Reviews Immunol*ogy **4**:123-132.

Oakes T, Heather JM, Best K, Byng-Maddick R, Husovsky C, Ismail M, *et al.* (2017) Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Frontiers in Immunology* **8**:1267.

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, *et al.* (2025) vegan: Community Ecology Package. R package version 2.6-10.

Peacock T, Heather JM, Ronel T, Chain B. (2021) Decombinator V4: An improved AIRR compliant-software package for T-cell receptor sequence annotation? *Bioinformatics* **37**:876-878.

Pedersen T (2024). patchwork: The Composer of Plots. R package version 1.3.0.

Pocock MJO, Hauffe HC, Searle JB. (2005) Dispersal in house mice. *Biological Journal of the Linnean Society* **84**:565-583

Poncet P (2019). modeest: Mode Estimation. R package version 2.4.0.

Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, *et al.* (2014) Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences USA* **111**:13139-13144.

Qiu Y (2024) showtext: Using fonts more easily in R Graphs. R package version 0.9-7.

R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R version 4.5.0

Rosshart SP, Vassallo BG, Angeletti D, Hutchinson DS, Morgan AP, Takeda K, *et al.* (2017) Wild mouse gut microbiota promotes host fitness and improves disease resistance. *Cell* **171**:1015-1028.e13.

Rowe FP, Bradfield A, Quy RJ, Swinney T. (1985) Relationship between eye lens weight and age in the wild house mouse (*Mus musculus*). *Journal of Applied Ecology* **22**:55-61.

Sender R, Weiss Y, Navon Y, Milo I, Azulay N, Keren L, *et al.* (2023) The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences USA* **120**:e2308511120.

Song N, Elbahnasawy MA, Weng NP. (2025) General and individualized changes in T cell immunity during aging. *Journal of Immunology* **214**:872-879.

Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E *et al.* (2017) Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. *Frontiers in Immunology* **8**:430.

Trofimov A, Brouillard P, Larouche JD, Séguin J, Laverdure JP, Brasey A, *et al*. (2022) Two types of human TCR differentially regulate reactivity to self and non-self antigens. *iScience* **25**:104968.

Uddin I, Woolston A, Peacock T, Joshi K, Ismail M, Ronel T, *et al.* (2019) Quantitative analysis of the T cell receptor repertoire. In *Methods in Enzymology* vol. 629 465-492 (Academic Press Inc., 2019).

Viney ME, Riley EM. (2017) The immunology of wild rodents : current status and future prospects. *Frontiers in Immunology* **8**:1481.

Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A. *et al* (2024) gplots: Various R Programming Tools for Plotting Data. R package version 3.2.0.

Weng NP. (2023) Numbers and odds: TCR repertoire size and its age changes impacting on T cell functions. *Seminars in Immunology* **69**:101810.

Wickham H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham H. (2023) stringr: Simple, consistent wrappers for common string operations. R package version 1.5.1.

Wickham H, François R, Henry L, Müller K, Vaughan D. (2023) dplyr: A grammar of data manipulation. R package version 1.1.4.

Wickham H, Henry L. (2025) purrr: Functional programming tools. R package version 1.0.4.

Wickham H, Vaughan D, Girlich M. (2024). tidyr: Tidy messy data. R package version 1.3.1.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains 11 items in a single file.

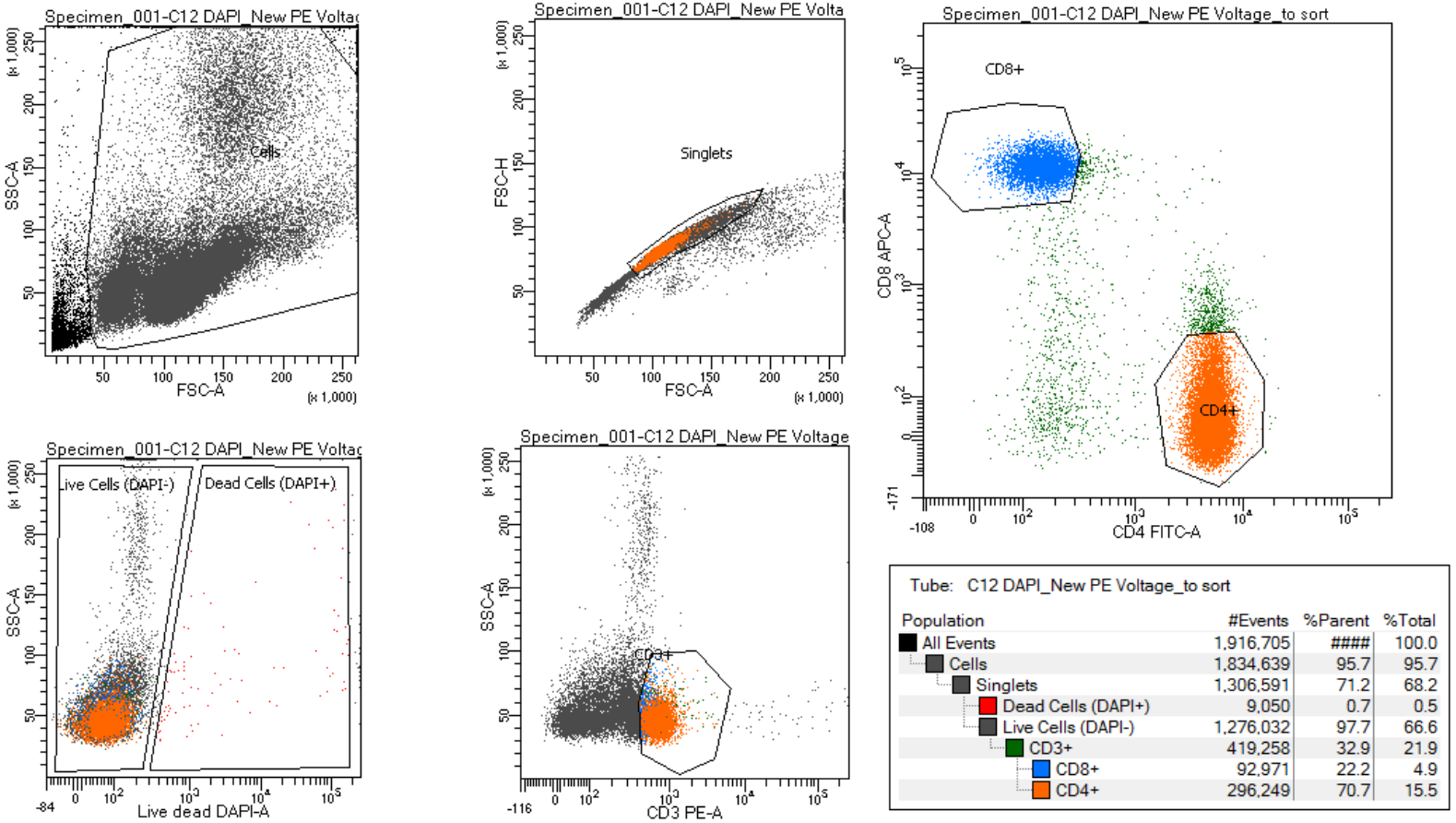
DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Sequence Read Archive [www.ncbi.nlm.nih.gov/sra] as SUB15234028 / PRJNA1249624 and our processed repertories and

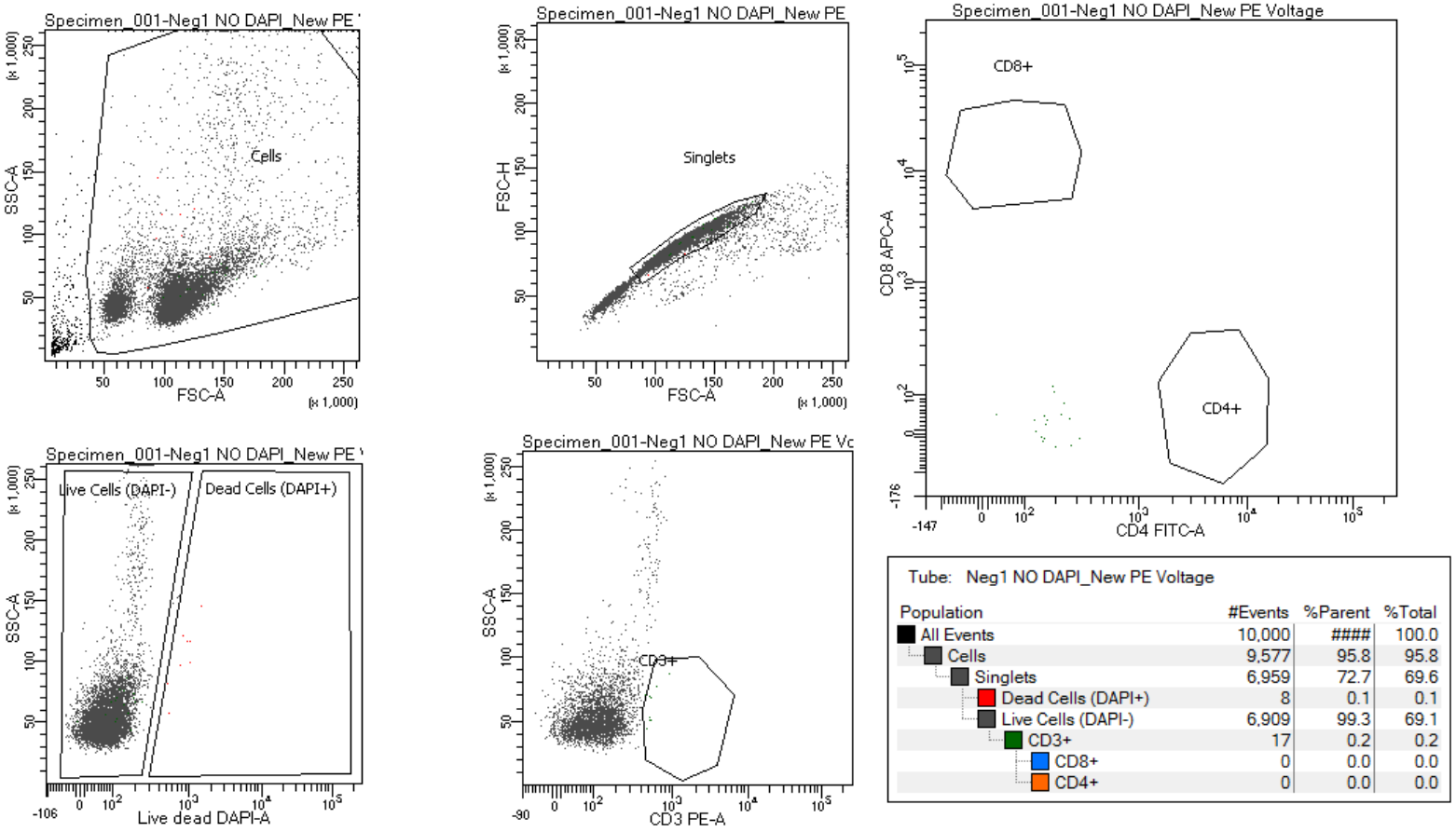
590 mouse metadata are in Zenodo at https://doi.org/10.5281/zenodo.17241409. Code used to analyse
591 data for this study can be found at https://github.com/jacohen1/Wild-mouse-TCR.

Supplementary Material 1. Example FACS plots of a mouse splenocyte sample where (A) cells were stained with DAPI, CD3-PE, CD4-FITC, and CD8-APC and (B) where for the same sample DAPI, CD3-PE, CD4-FITC, and CD8-APC were omitted from the staining protocol. For each, the gating hierarchy are plotted sequentially as (i) FSC-A *vs*. SSC-A, (ii) FSC-A *vs.* FSC-H, (iii) DAPI *vs.* SSC, (iv) CD3-PE *vs.* SSC, (v) CD4-FITC *vs.* CD8-APC, and the subsequent $CD3^+$ $CD4^+$ and $CD3^+$ $CD8^+$ were defined from dot plots of pair-matched negative controls (no antibody staining) from each individual (B), where maximum tolerance to false positive was <0.3 %.

A



B

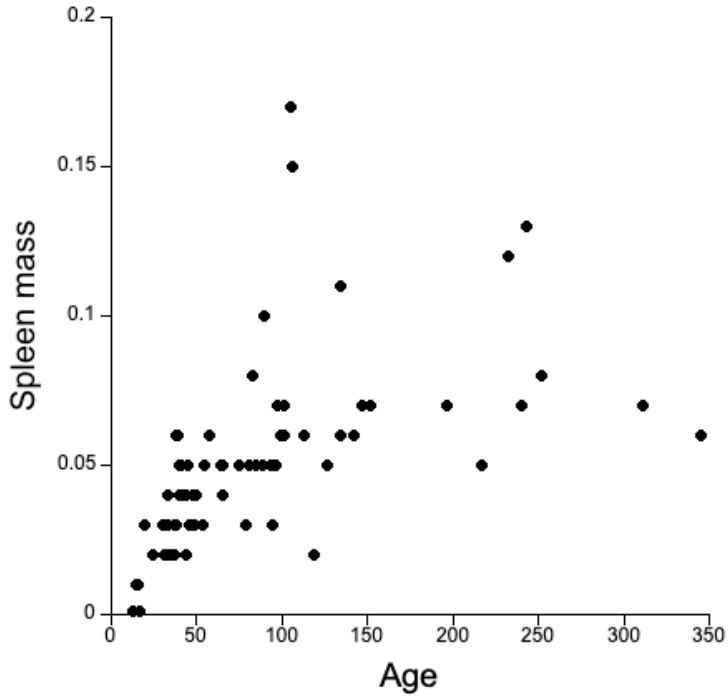


Tube: Neg1 NO DAPI_New PE Voltage

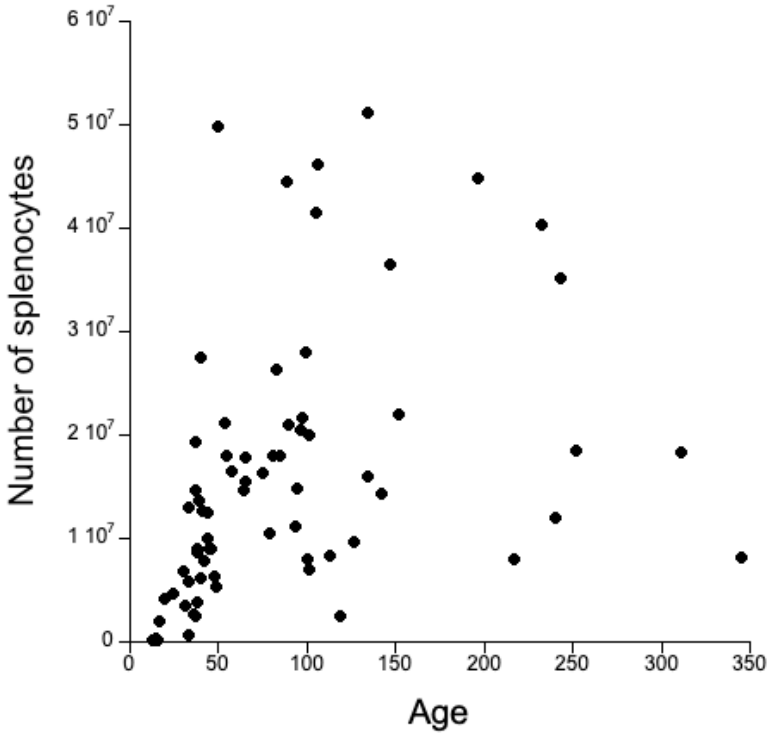
Population	#Events	%Parent	%Total
All Events	10,000	####	100.0
Cells	9,577	95.8	95.8
Singlets	6,959	72.7	69.6
Dead Cells (DAPI+)	8	0.1	0.1
Live Cells (DAPI-)	6,909	99.3	69.1
CD3+	17	0.2	0.2
CD8+	0	0.0	0.0
CD4+	0	0.0	0.0

Supplementary Material 2. Mouse age in days and (A) spleen mass in grammes (Spleen mass = 0.03 + 0.0002.Age, $R^2 = 0.3$, $p < 0.0001$) and (B) number of splenocytes (Number of splenocytes = $1.02 \times 10^6 + 6.37 \times 10^4$.Age, $R^2 = 0.13$, $p < 0.0025$).

A

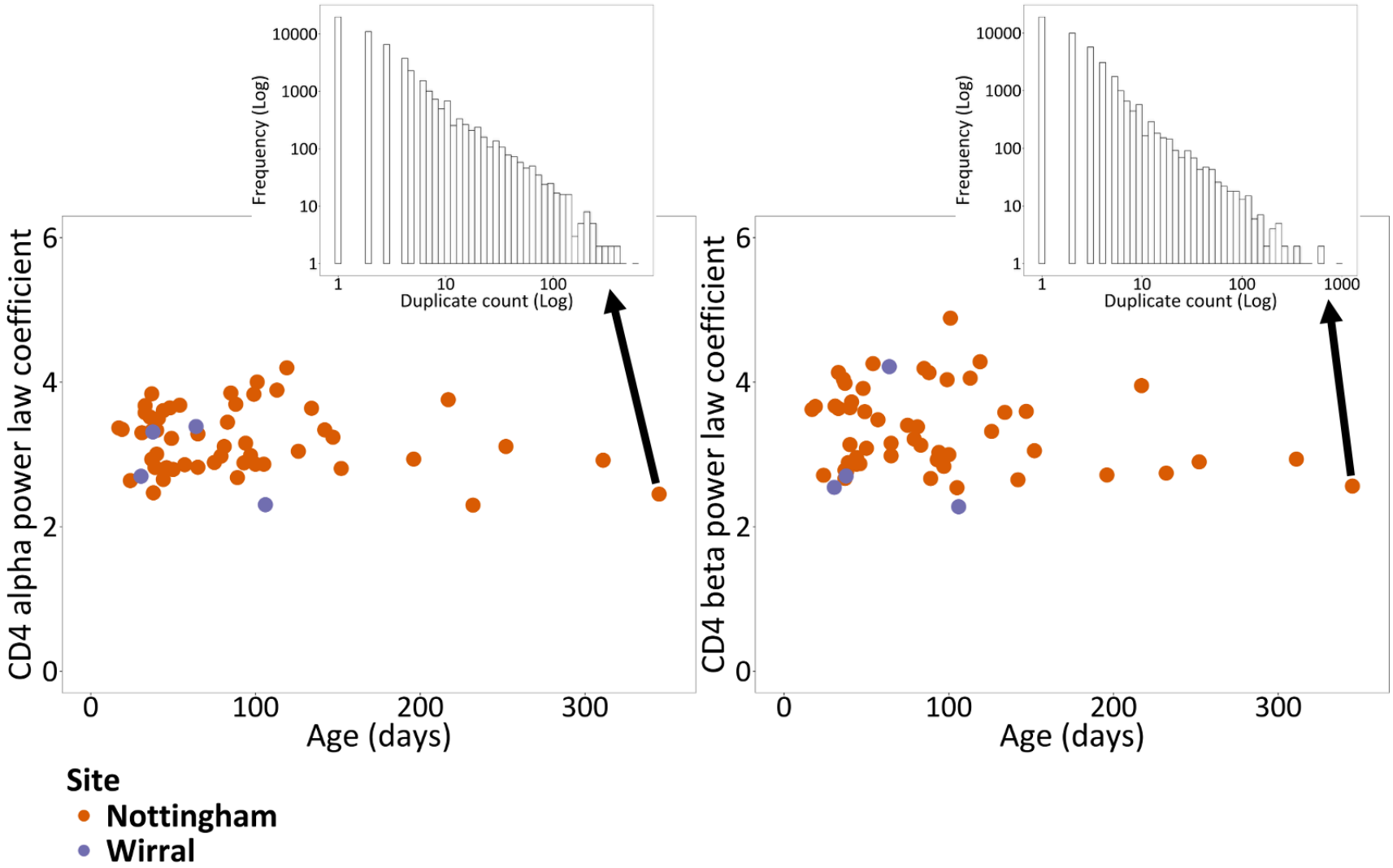


B

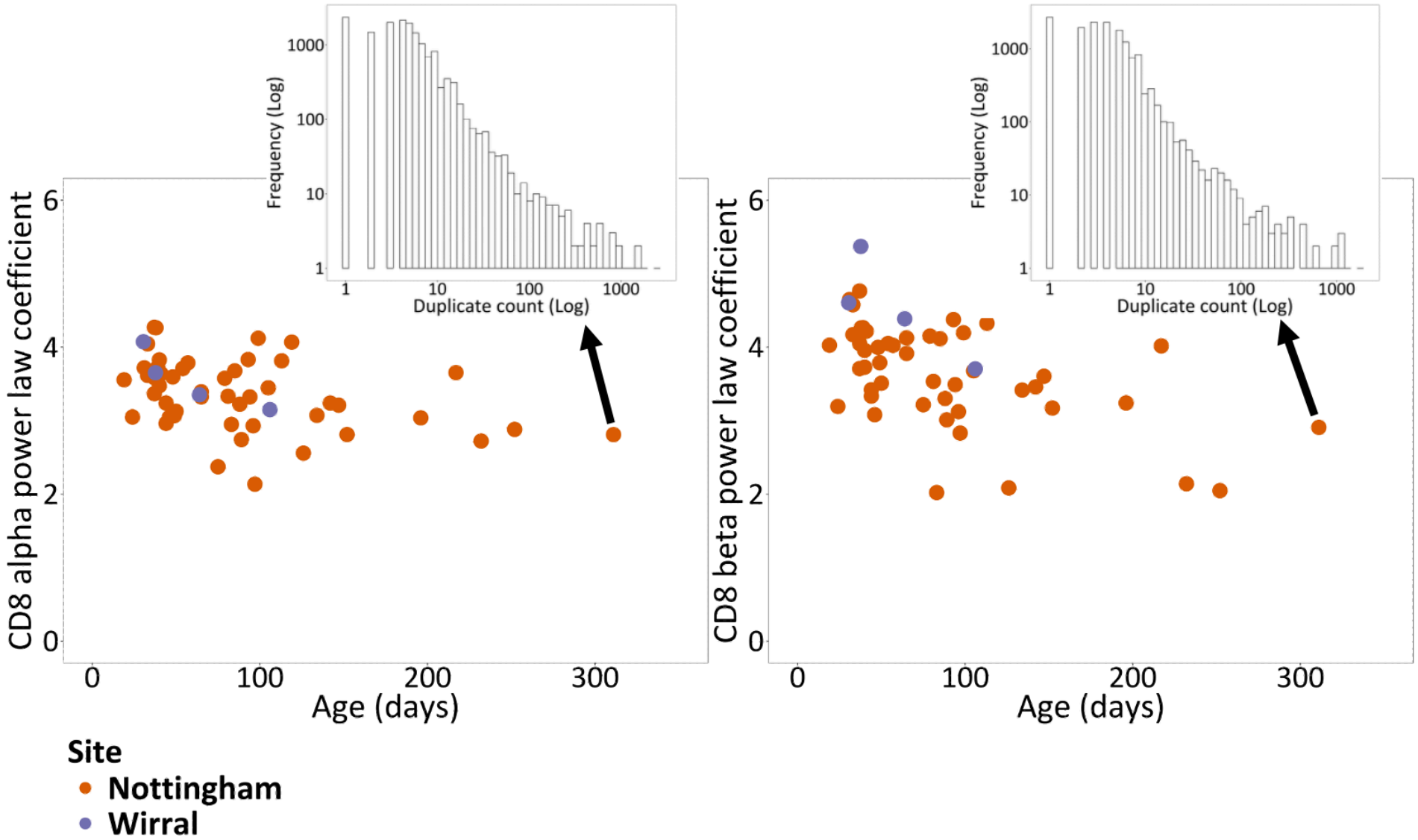


Supplementary Material 3. The power law coefficients of the frequency distribution of amino acid-defined TCR sequences for wild mice from the two sample sites for (A) $CD4^+$ and (B) $CD8^+$ cells; within each an example Log Frequency *vs.* Log Number of amino acid-defined sequences is shown. (C) Result of Kolmogorov-Smirnov test with 500 bootstraps to test that the data follows a power law distribution, showing the number of mice in each category.

A


Frequency (Log)
10000
1000
100
10
1
1
10
100
Duplicate count (Log)
CD4 alpha power law coefficient
6
4
2
0
0
100
200
300
Age (days)
Frequency (Log)
10000
1000
100
10
1
1
10
100
1000
Duplicate count (Log)
CD4 beta power law coefficient
6
4
2
0
0
100
200
300
Age (days)
Site
Nottingham
Wirral

B

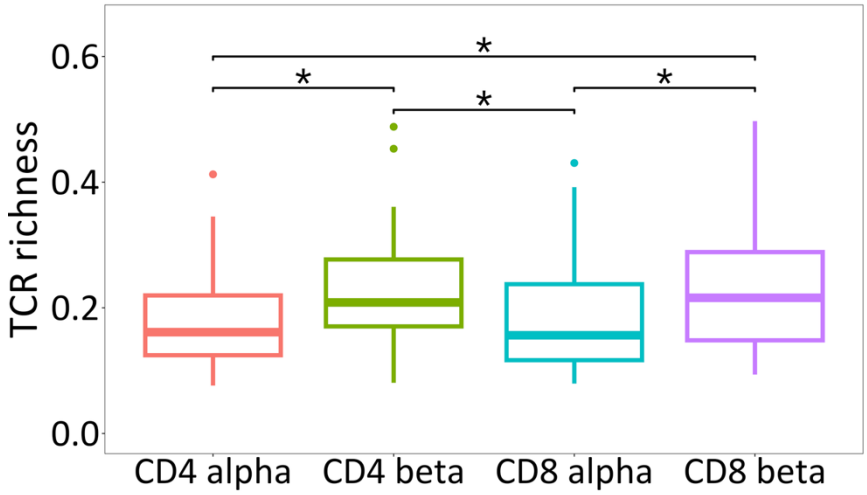

CD8 alpha power law coefficient
Age (days)
Frequency (Log)
Duplicate count (Log)
CD8 beta power law coefficient
Age (days)
Frequency (Log)
Duplicate count (Log)
Site
Nottingham
Wirral

C

Receptor type	Fit to power law distribution		
	Fits ($p > 0.1$)	**Borderline fit ($0.05 < p < 0.1$)**	**Does not fit ($p < 0.05$)**
$CD4^+$ alpha	n = 45 (72.6%)	n = 3 (4.8%)	n = 14 (22.6%)
$CD4^+$ beta	n = 24 (38.7%)	n = 3 (4.8%)	n = 35 (56.5%)
$CD8^+$ alpha	n = 24 (41.4%)	n = 4 (6.9%)	n = 30 (51.7%)
$CD8^+$ beta	n = 10 (17.2%)	n = 2 (3.4%)	n = 46 (79.3%)

Supplementary Material 4. (A) Wild mouse TCR richness (calculated as the number of unique amino acid-defined TCRs as a proportion of the total number of amino acid-defined TCR sequences); the whiskers are 1.5 x the inter-quartile range; * is $p < 0.05$. (B) The p values are for pairwise comparisons of the marginal means estimated from a GLM of the form VALUE ~ RECEPTOR TYPE + SAMPLE SITE + AGE + SEX; $p < 0.05$ are in bold.

A

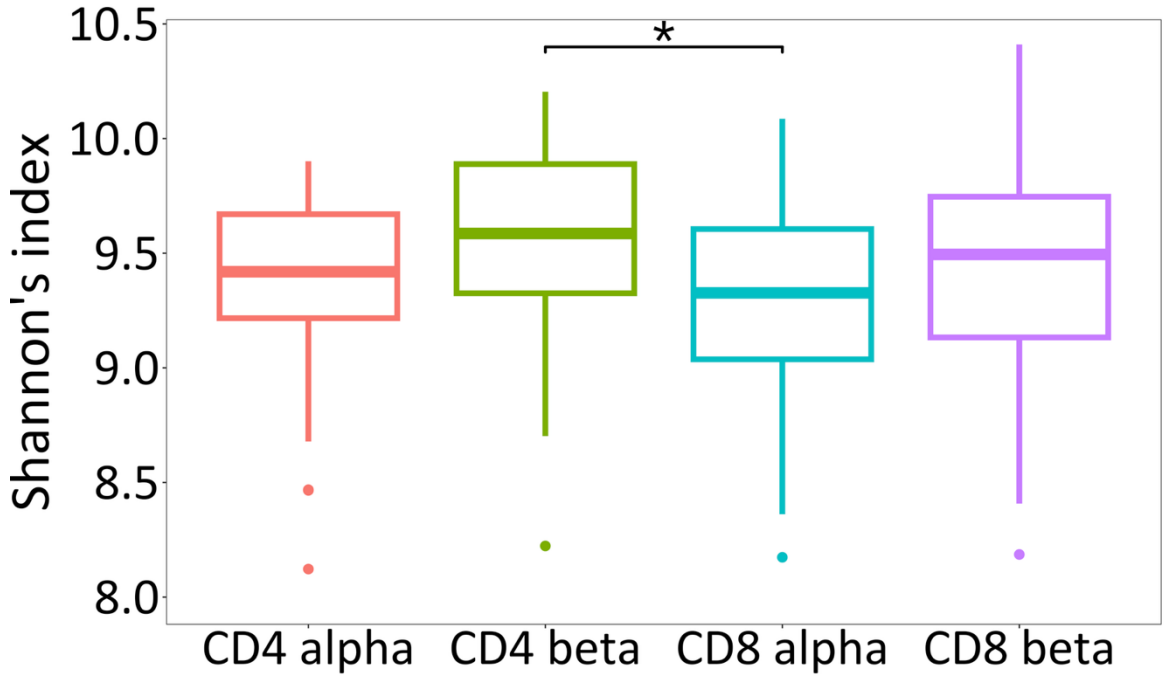


B

	CD4⁺ alpha	**CD4⁺ beta**	**CD8⁺ alpha**
CD4⁺ beta	**0.011**		
CD8⁺ alpha	0.9994	**0.0275**	
CD8⁺ beta	**0.0165**	0.9997	**0.0275**

Supplementary Material 5. (A) Wild mouse amino acid-defined TCR alpha diversity using Shannon's index; the whiskers are 1.5 x the inter-quartile range; * is $p < 0.05$. (B) The p values are for pairwise comparisons of the marginal means estimated from a GLM of the form VALUE ~ RECEPTOR TYPE + SAMPLE SITE + AGE + SEX + AGE*SEX ; $p < 0.05$ are in bold.

A

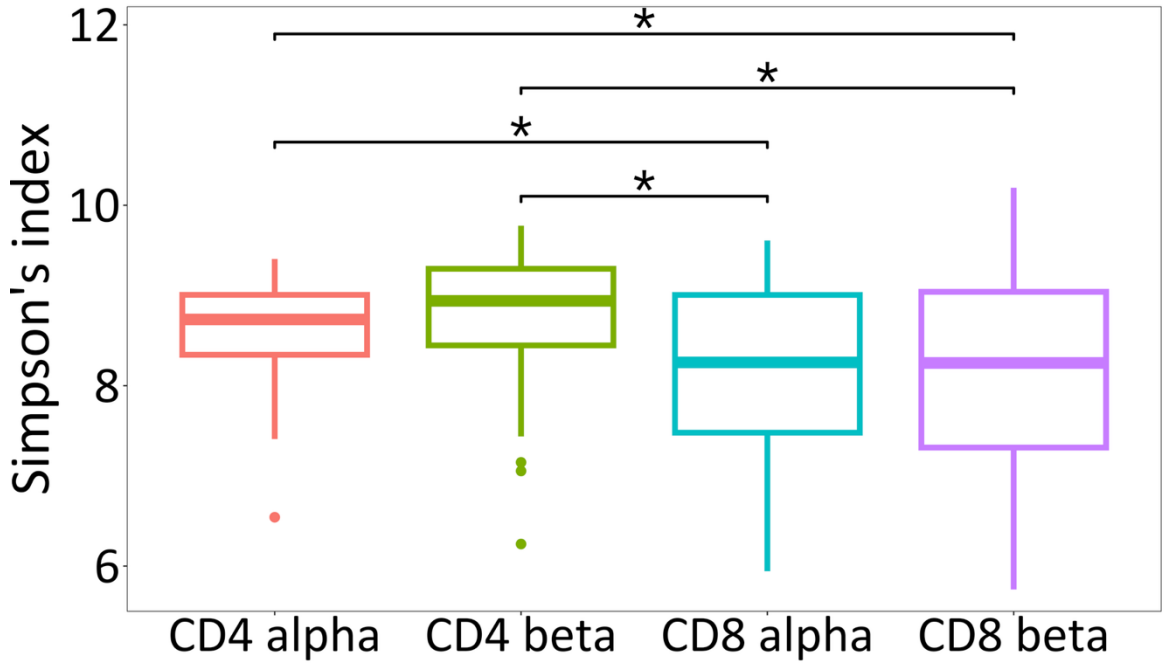


B

	CD4⁺ alpha	**CD4⁺ beta**	**CD8⁺ alpha**
CD4⁺ beta	0.0876		
CD8⁺ alpha	0.8197	**0.0089**	
CD8⁺ beta	0.8576	0.4244	0.3614

Supplementary Material 6. (A) Wild mouse amino acid-defined TCR alpha diversity using Simpson's index; the whiskers are 1.5 x the inter-quartile range; * is $p < 0.05$. (B) The p values are for pairwise comparisons of the marginal means estimated from a GLM of the form VALUE ~ RECEPTOR TYPE + SAMPLE SITE + AGE + SEX + AGE*SEX ; $p < 0.05$ are in bold.

A



B

	CD4⁺ alpha	**CD4⁺ beta**	**CD8⁺ alpha**
CD4⁺ beta	0.6853		
CD8⁺ alpha	**0.0240**	**0.0006**	
CD8⁺ beta	**0.0195**	**0.0005**	0.9999

Supplementary Material 7. The results of GLM analysis of effects on wild mouse TCR diversity measured using Shannon and Simpson's diversity indices, of the form VALUE ~ RECEPTOR TYPE + SITE + AGE + SEX + SEX*AGE, where for each term or the interaction the parameter estimate and the p values are shown. NS is not significant at p = 0.05.

	Shannon's	**Simpson's**
Age	-0.00090, p = 0.0762	-0.00380, p = 9.91e-05
Sex (Male)	0.2803, p = 0.0058	0.3827, p = 0.0452
Sample site (Wirral)	0.30279, p = 0.0063	NS
Age*Sex (Male)	-0.00205, p = 0.0396	NS

Supplementary Material 8. (A) The mean and median number of amino acid-defined TCR sequences shared among wild and laboratory mice and, in parentheses, these expressed as a percentage of the number of down-sampled sequences. (B) The number of amino-acid defined TCR sequences shared among ≥75 % of mice (both wild and laboratory) and this expressed as a percent of all the unique TCR sequences.

A

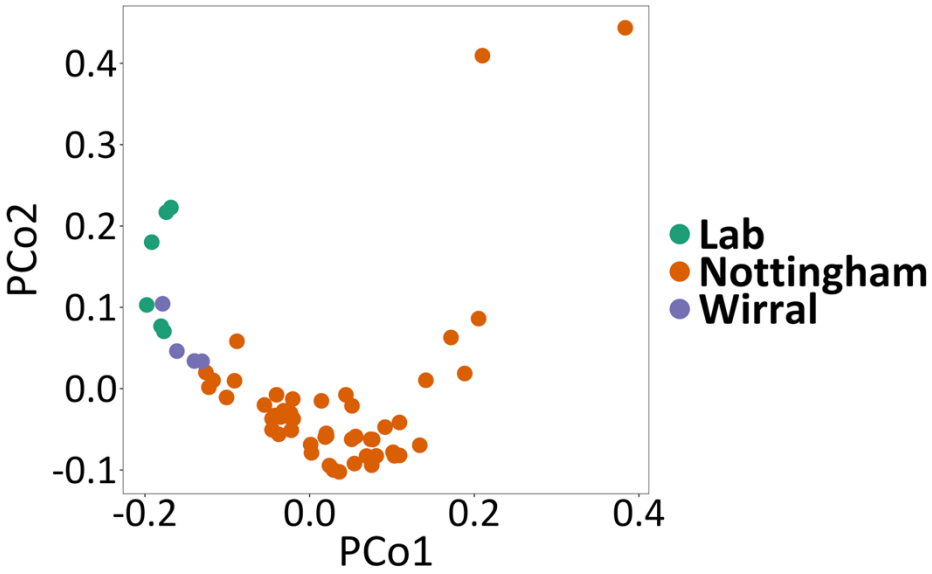
	$CD4^+$ alpha	**$CD4^+$ beta**	**$CD8^+$ alpha**	**$CD8^+$ beta**
Mean	3,232 (5.89)	1,678 (2.86)	2,561 (5.19)	1,387 (2.16)
Median	3,221 (5.87)	1601 (2.72)	2,459 (4.98)	1,251 (1.95)
Number of down-sampled sequences	54,908	58,772	49,379	64,224

B

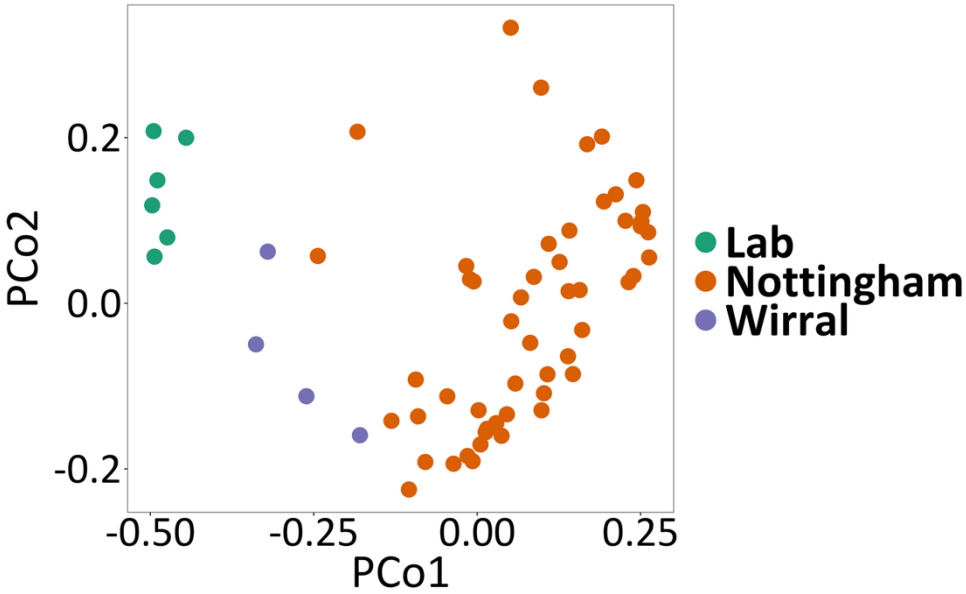
Receptor type	**Number**	**Percent**
$CD4^+$ alpha	1,113	0.199
$CD4^+$ beta	268	0.024
$CD8^+$ alpha	703	0.143
$CD8^+$ beta	157	0.017

Supplementary Material 9. Principal Co-ordinates Analysis (PCoA) showing the first two principal co-ordinates of the number of shared amino acid-defined TCR sequences between all pairwise combinations of wild and laboratory mice, with the origins of the mice colour coded for (A) $CD4^+$ alpha, (B) $CD4^+$ beta, (C) $CD8^+$ alpha and (D) $CD8^+$ beta. The percentage of variance explained by PCo1 and PCo2 is for $CD4^+$ alpha 11.1 and 9,4; $CD4^+$ beta 21.4 and 8.3; $CD8^+$ alpha 10.3 and 8.4; $CD8^+$ beta 15.5 and 7.5, respectively.

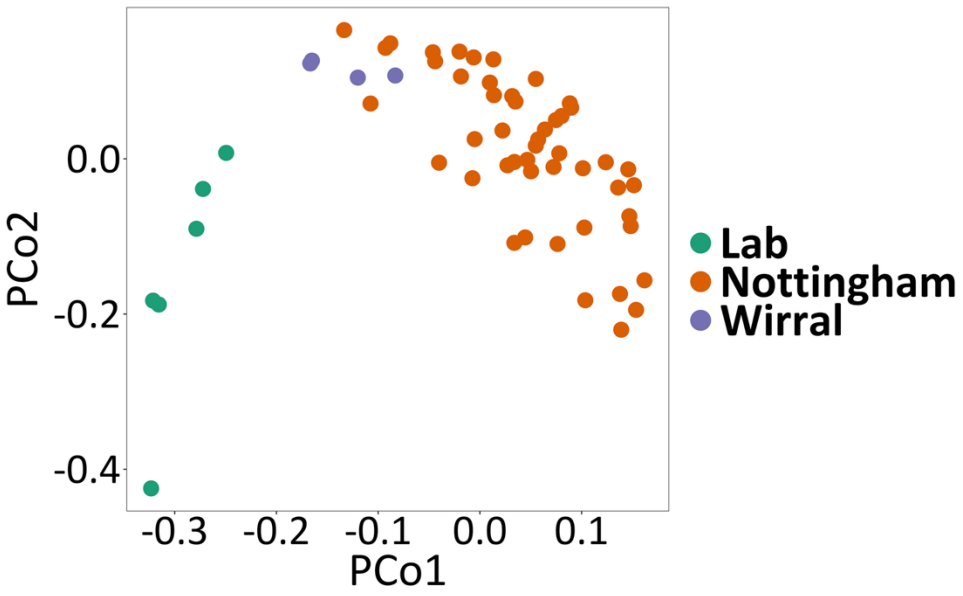
A



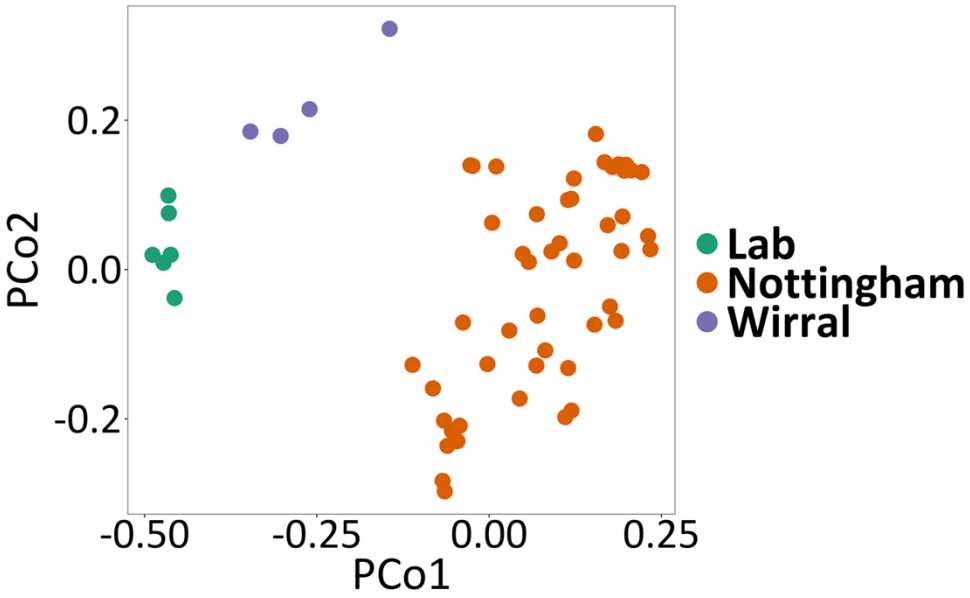
B



C

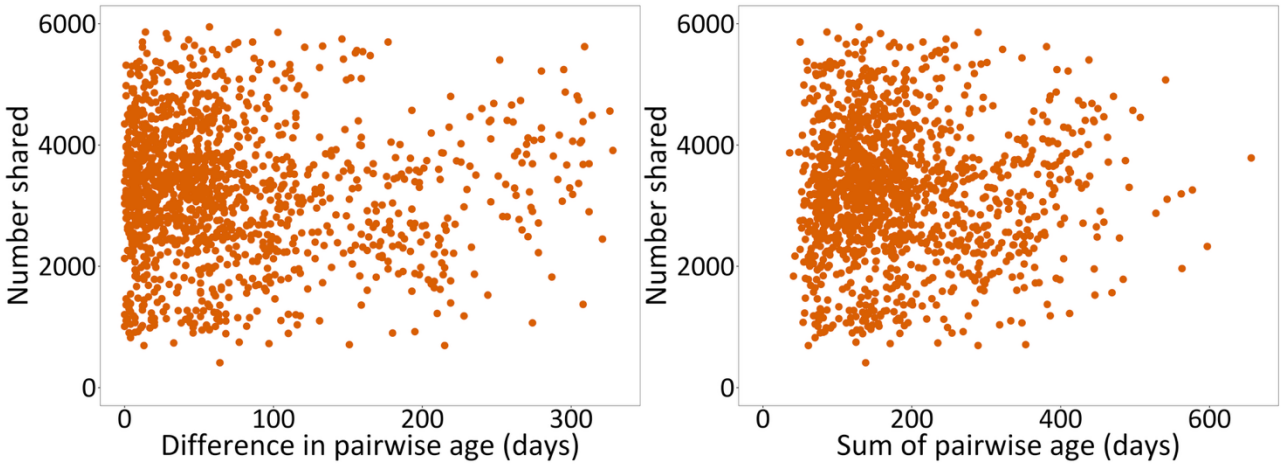

0.0
-0.2
-0.4
PCo2
-0.3
-0.2
-0.1
0.0
0.1
PCo1
Lab
Nottingham
Wirral

D

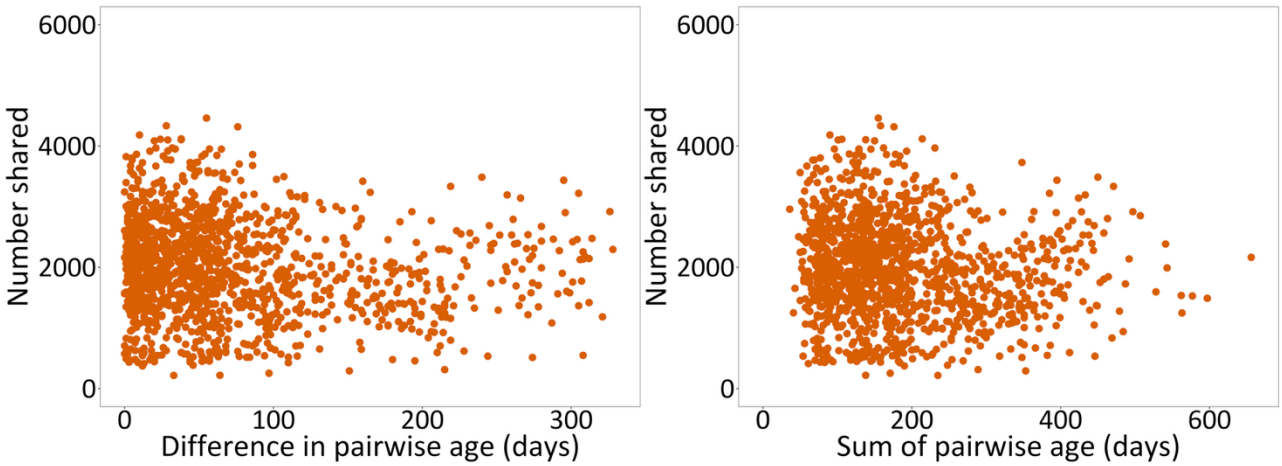

0.2
0.0
-0.2
PCo2
-0.50
-0.25
0.00
0.25
PCo1
Lab
Nottingham
Wirral

Supplementary Material 10. The number of shared amino acid-defined TCR sequences between all pairwise combinations of wild mice from the Nottingham site against the difference in pairwise age (left hand panels) and the sum of the pairwise age (right hand panels), where age is in days for (A) CD4⁺ alpha, (B) CD4⁺ beta and (C) CD8⁺ beta.

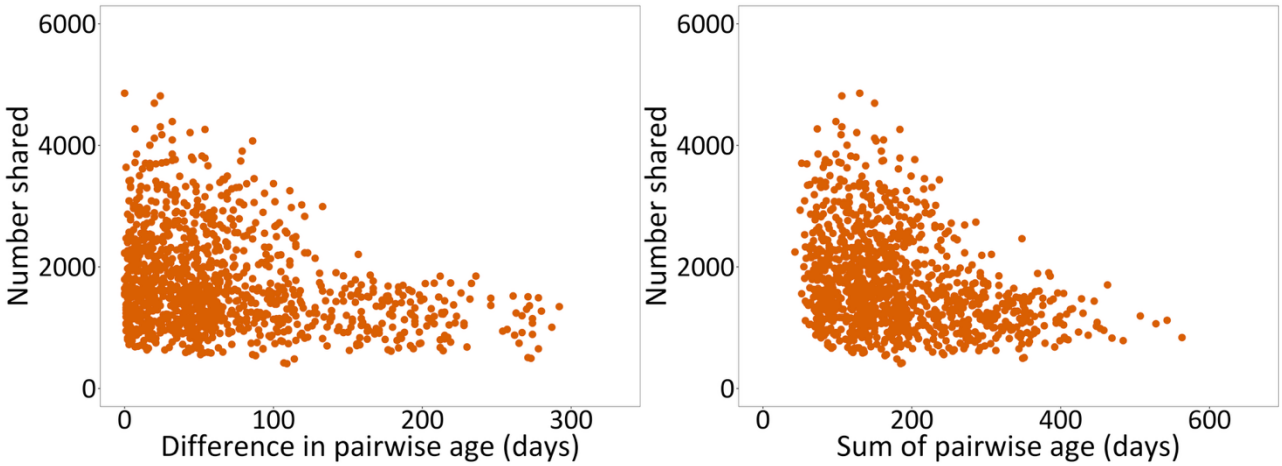
A



B



C



Supplementary Material 11. Analysis of the number of shared amino acid-defined TCR sequences and (A) pairwise difference in age and (B) pairwise sum in age of the mice, showing the slope estimates and 95 % confidence intervals (CI) for each receptor type (top table), and the P values for comparisons of the parameter estimates for the four receptor types (bottom table), with bold text showing $p < 0.05$.

A

Receptor type	Estimate	95 % CI
CD4$^+$ alpha	-0.541	-1.23, 0.145
CD4$^+$ beta	-0.859	-1.55, -0.173
CD8$^+$ alpha	-1.481	-2.30, -0.665
CD8+ beta	-1.368	-2.18, -0.551

	CD4$^+$ alpha	CD4$^+$ beta	CD8$^+$ alpha
CD4$^+$ beta	0.9182		
CD8$^+$ alpha	0.3091	0.6624	
CD8$^+$ beta	0.4258	0.7859	0.9975

B

Receptor type	Estimate	95 % CI
CD4$^+$ alpha	0.180	-0.313, 0.6723
CD4$^+$ beta	-0.434	-0.926, 0.0588
CD8$^+$ alpha	-1.392	-1.965, -0.8193
CD8+ beta	-2.110	-2.683, -1.5368

	CD4$^+$ alpha	CD4$^+$ beta	CD8$^+$ alpha
CD4$^+$ beta	0.3097		
CD8$^+$ alpha	**0.0003**	0.0620	
CD8$^+$ beta	**<0.0001**	**0.0001**	0.3049