Thymic selection of the T cell receptor repertoire is biased toward autoimmunity in females

Hélène Vantomme[1,2], Valentin Quiniou[1], Leslie Adda[1§], Charline Jouannet[1§], Vanessa Mhanna[1§], Céline Albalaa[1$], Pierre Barennes[1$], Nicolas Coatnoan[1,2$], Vimala Diderot[1,2$], Johanna Dubois[1$], Gwladys Fourcade[1$], Kenz Le Gouge[1$], Otriv Frédéric Nguekap Tchoumba[1$], Martin Pezous[1$], Paul Stys[1$], Adrien Six[1], Encarnita Mariotti-Ferrandiz[1,3*] and David Klatzmann[1,2*#].

[§]contributed equally, alphabetical order

[$]contributed equally, alphabetical order

*co-senior authors

[#]corresponding author

1. Immunology, Immunopathology, Immunotherapy i3 Lab, INSERM UMRS-959, Sorbonne Université, Paris, France
2. Biotherapy (CIC-BTi) and Inflammation-Immunolopathology-Biotherapy Department (i2B), Assistance Publique -Hôpitaux de Paris (AP-HP), Paris France
3. Institut Universitaire de France

Corresponding author :
Prof. David KLATZMANN MD, PhD
I3 Lab (Immunology Immunopathology Immunotherapy) INSERM UMRS 959 – Sorbonne Université
Bâtiment CERVI, GHPS Pitié-Salpêtrière, 83 boulevard de l'Hôpital, 75013 Paris, France
Tel : +33 1 42 17 74 61
Email : david.klatzmann@sorbonne-universite.fr

The authors have declared that no conflict of interest exists.

Abstract: (200 words)

Women represent about 80% of patients with autoimmune diseases. This may partly result from sex-based differences in T cell receptor (TCR) selection during thymocyte development, potentially influenced by hormones and the lower expression of the Autoimmune Regulator (AIRE) transcription factor in females.

To investigate this, we analyzed sex-specific differences in TCR generation and selection. We examined TCR repertoires in double-positive thymocytes and single-positive thymic cells, including $CD8^+$ and $CD4^+$ effector T cells and regulatory T cells (Tregs), derived from male and female organ donors. Minimal sex-based differences were observed in V and J gene usage, and there were no notable differences in TCR repertoire diversity, complementarity-determining region 3 (CDR3) length, amino acid composition, or network structure. No TCR sequences were exclusive to either sex.

However, female effector T cells exhibited a significantly higher prevalence of TCRs specific to self-antigens implicated in autoimmunity compared to males, while female Tregs showed a reduced frequency of such TCRs. These differences were not observed for TCRs targeting self-antigens unrelated to autoimmunity or antigens associated with cancer or viruses.

Our findings highlight a sex-specific imbalance in thymic selection of TCRs with autoimmunity-associated specificities, providing mechanistic insight into the increased susceptibility of women to autoimmune diseases.

Introduction

Several studies have highlighted a sex imbalance in diseases involving the immune system. This is particularly the case for autoimmune diseases (AIDs), where 80% of those affected are women and the severity between the sexes vary depending on the pathology (1–3). In addition, men are more severely affected by infectious disease such as COVID-19 and tuberculosis (4, 5). The response to preventive immunotherapies such as vaccination, and to curative treatments, like immune checkpoint inhibitors or anti-TNF alpha, is also sex dependent (6). All of these observations suggest intrinsic biological differences in immune responses between males and females.

A variety of factors have been identified that could help explain these observed sex-based differences. Among these, sex hormones have been shown to influence immune responses by acting on immune cells that express specific receptors to these hormones (7–9). Sex hormones could directly influence the adaptive immune response at the level of thymocyte differentiation and selection by acting on the expression of Autoimmune Regulator (AIRE) in thymic epithelial cells. AIRE drives the thymic expression of otherwise tissue-specific antigens, contributing to both the negative selection of effector cells and the positive selection of Tregs that recognize such antigens (10). Testosterone increases AIRE expression in medullary thymic epithelial cells, whereas estradiol decreases it (11, 12). Furthermore, a recent study has shown that the transcriptomic profiles of these specialized antigen-presenting cells in the thymus differ between males and females (13). This clearly suggests that there could be differences in the selection of the T cell receptor (TCR) repertoire between males and females.

Despite these findings, little is known about potential differences in the TCR repertoire between the sexes. The generation of the TCR is a key process in T-cell development that

takes place in the thymus. The TCR consists of two chains, alpha (TRA) and beta (TRB), each of which resulting from a random genetic rearrangement involving the V (variable), D (diversity) [for the TRB], and J (joining) genes. The combination of these gene segments is accompanied by insertions and deletions, creating a highly diverse region, the Complementarity Determining Region 3 (CDR3) (14). It is this region of the TCR that mostly interacts with the peptide, a critical interaction for antigen recognition and subsequent T cell activation. The ability to specifically recognize diverse antigens via the CDR3 is central to the efficiency and specificity of the adaptive immune response.

Because TCRs are randomly generated, many could be useless while some could be harmful by recognizing self-antigens and causing autoimmunity. The current paradigm is that during thymocyte development, thymocytes are selected based on the avidity of their TCRs for antigens presented by cortical and medullary thymic epithelial cells. Thymocytes expressing TCRs with too low or too high avidity are eliminated (negative selection), while those with an intermediate avidity are positively selected and mature into single-positive cells, either CD4+ or CD8+. Thymocytes with the highest of these intermediate affinities develop in Tregs. Since AIRE plays a major role in these processes, sex differences in AIRE expression could have a major impact on thymic selection.

The TCR repertoire of female and male peripheral T cells have shown that their diversity, particularly that of the TRB repertoire, is influenced not only by the age of the individuals but also by their sex (15–18). However, another study showed no association between the sex of individuals and the diversity of their TCR repertoire (19). Nevertheless, the study of peripheral blood TCRs reflects that of an immune system that has gone through multiple immune responses. For example, differences observed between males and females for responses to

self-antigens associated with autoimmunity could reflect either the cause (the TCRs drive the disease) or the consequence (the TCRs are expanded because of the disease).

To assess whether there are significant differences in the generation and thymic selection of the TCR repertoire between males and females, we therefore focused on analyzing the TCR repertoires of thymocytes. These cells are the most recent product of T cell development and have not yet been involved in immune responses. Their repertoires thus represent the mechanisms of TCR generation for the analysis of CD4/CD8 double positive (DP) cells, and of TCR selection for CD4 and CD8 simple positive (SP) T cells. We generated a unique data set comprising the sequences of the TRA and TRB chains of DP cells, CD8 and CD4 SP effector cells, and CD4 SP Tregs from organ donors, both infants and adults. The analysis of TRA and TRB repertoires using state-of-the-art strategies revealed no sex differences in repertoire generation (DP cell repertoire), but a bias towards autoimmunity in the selection of the CD8 SP T cells in females.

Results

Minimal sex-based differences in TCR V and J gene usage across thymic cell subsets.

The first level of diversity in generating TCRs is the usage of V and J genes (20). We analyzed the usage of TRA and TRB V and J genes and their combinations in each thymocyte population (Figure 1).

Males and females had a relatively similar V gene usage as attested by the absence of a clear separation in the Principal Component Analysis (PCA), for both TRA and TRB, and across all studied cell subtypes (Figure 2A). However, we observed some individual differences in V and J gene usage between males and females, with some being specific to certain cell subtypes (Supplemental Figures 2-5), and others observed across multiple cell subtypes (Supplemental Figures 2-5). For example, TRBV6-5 was more highly expressed in females compared to males in DP cells (Supplemental Figure 2C) and CD8 SP cells (Supplemental Figure 3C).

Next, we compared the usage of TRAV-TRAJ and TRBV-TRBJ gene combinations between all individuals using the Jensen-Shannon Divergence (JSD) score (Figure 2B). The hierarchical clustering showed no clear separation between male and female individuals for any of the chain or cell subtype, indicating no significant differences in the gene combination usage between males and females (Figures 2B). A similar observation was made when directly analyzing the frequency of these V-J gene combinations (Supplemental Figures 2E-5E, 2F-5F).

Comparable TCR repertoire diversity between males and females

A high diversity of the TCR repertoire is essential for generating a broad potential reactivity against a variety of antigens. We thus compared the TCR repertoire diversity between males and females. To achieve this, we analyzed Rényi diversity curves (from zero to infinite α parameters) that measure various aspects of TCR repertoire diversity. At different α values,

the Rényi index places more or less weight on the contribution of the most frequent clonotypes. Higher α values emphasize on the most abundant clonotypes, thereby revealing the impact of expanded clones on the overall diversity.

Rényi curves showed comparable diversities between sexes in DP and CD4 Teff cells, for both TRA and TRB, and a difference in CD8 and CD4 Treg cells (Supplemental Figure 6). In CD8 cells, the curves begin to diverge at low α values (Supplemental Figure 6), with an earlier inflection in males, indicating a less balanced distribution of TCR clones. This suggests that in males, fewer clones dominate the repertoire at lower diversity thresholds compared to females. In Tregs, while the inflection point is similar, the difference seems to be characterized by higher diversity and richness in females compared to males (Supplemental Figure 6). However, no significant differences were observed for the Shannon diversity index (α = 1), the Simpson index (α = 2), and the Berger-Parker index (α = infinite), for both chains and all cell subtypes (Figure 3B). These results indicate that the thymic TCR repertoire diversity is similar between males and females.

Comparable CDR3aa length distribution in male and female TCR repertoires

Most of the repertoire diversity is generated by the random addition of nucleotides within the VDJ junctions, generating the hypervariable CDR3 region that is interacting with the peptide-MHC complex, and is mostly driving the TCR specificity. The CDR3 length directly influences the TCR's ability to interact with the peptide (21). We thus studied the distribution of CDR3 aa (CDR3aa) sequence lengths between males and females to identify potential TCR generation and/or selection biases (Figure 1).

Our results showed that males and females have a comparable distribution of CDR3aa lengths for both TRA (Supplemental Figure 7) and TRB (Figure 4A). CDR3aa sequences in TRA were

shorter than those in TRB, as previously described in peripheral blood T cells (22, 23). Although there were subtle differences in certain CDR3aa lengths between males and females, these variations were minimal, involving sequences that represent less than 2% of the total TCR repertoire (Figure 4A, Supplemental Figure 7). Altogether, the overall distribution of CDR3aa length is similar between males and females.

Minimal differences in CDR3aa usage in thymic TCR repertoires of males and females

We analyzed the aa usage in the FG loop of the CDR3, which is in contact with the peptide-MHC complex (24). We classified the aa according to their hydropathy property: hydrophilic, hydrophobic, or neutral (25). We observed significant differences in the usage of some aa between males and females for both TRA (Supplemental Figure 8) and TRB (Figure 4B). In DP cells, the differences preferentially involve hydrophobic aa: females exhibit higher alanine (A) usage in TRA and lower in TRB compared to males (Figure 4B, Supplemental Figure 8). Additionally, females show increased usage of phenylalanine (F) and isoleucine (I) in TRB (Figure 4B). Furthermore, other aa with varying hydropathy characteristics are differentially used between males and females in CD4 Teff and Treg populations (Figure 4B, Supplemental Figure 8).

We then analyzed in more detail the usage of aa at positions p109 and p110 as the presence of hydrophobic aa at these two positions in TRB has been associated with a stronger recognition of self-antigens (26–28). We did not observe any significant differences between males and females in the usage of hydrophobic aa at these two positions, across all cell subtypes in TRB (Figure 4C).

Slight sex-based variations in thymic TCR generation probability distribution.

Another approach to evaluate biases in the generation and selection of the TCR repertoire is to compare the probability of generation (Pgen) of CDR3 nucleotide sequences (Figure 1). Pgen represents the likelihood of a sequence being produced, according to models of V(D)J recombination (14, 29). For DP and CD8 cells only and for each TCR chain, a sequence generation model was created using 100,000 non-productive nucleotide sequences. Based on this model, the Pgen of all productive TCR sequences composing each repertoire was calculated.

When comparing the distribution of Pgen between males and females, we observe a very slight difference of distribution characterized by a nearly null Kolmogorov-Smirnov score ($D_{K-S} < 0.05$) (Figure 5). This difference is more pronounced in the TRB compared to the TRA chain (Figure 5). As this small difference could reflect differences linked to variation among individuals irrespective of their sex, we generated 20,000 permuted mixed-sex groups with our dataset. We found the same significant distribution difference of Pgen sequences between these mixed groups population (DP – TRA: $D_{K-S} = 0.0183 \pm 0.0145$ ($\mu \pm sd$), DP – TRB: $D_{K-S} = 0.0190 \pm 0.0161$, CD8 – TRA: $D_{K-S} = 0.0227 \pm 0.0153$, and CD8 – TRB: $D_{K-S} = 0.0259 \pm 0.0140$). However, the difference observed in the Pgen distribution between males and females is higher than that between the mixed groups in the DP TRB.

Similar network structure of thymic TCR repertoires in males and females

Next, we evaluated whether thymic selection in males and females is directed towards similar or distinct sequences. For each TCR repertoire, CDR3aa sequences are connected if they differ by one aa, i.e. have a Levenshtein distance of =1 (Figure 1). Such linked sequences have high probability to share the same specificities (30). A random downsampling to the smallest

sample was performed to compare sample of similar sizes, with 100 iterations for each, detailed in Supplemental Table. The shapes of the network are depicted in Figure 6A.

To analyze more precisely these clustering we used various metrics. When comparing the number of clustered sequences in each individual network between males and females, no significant differences were observed for any cell type in either chain (Figure 6B). Similarly, the degree of sequence similarity, reflected by the density of edges between sequences within an individual's TCR repertoire, was comparable between males and females (Figure 6C). These findings indicate that there are no significant sex-specific differences in the similarity of CDR3aa sequences, suggesting that males and females produce CDR3aa sequences with the same degree of variability within their TCR repertoires.

Absence of sex-specific CDR3aa motifs in thymic TCR repertoires of males and females

We then sought whether we could detect CDR3aa sequence motifs that were more present in one group compared to the other, in order to evaluate whether the selection of CDR3aa was biased towards particular sequences differently between males and females (Figure 1). To identify sequences with shared characteristics that might indicate a probable common specificity among individuals of the same group, the analysis focused exclusively on CDR3aa sequences of the TRB, as this region has a greater influence on TCR specificity (31, 32). Local motifs were defined using Gliph2 (33), which are a sequences of three to five aa, as well as global motifs, which are a sequence of more than three aa where, at one position, an aa can be substituted by another aa if it has a positive BLOSUM62 score (34).

We found several hundred motifs differentially expressed between males and females for each cell subpopulation. Larger number of motifs were specific to females, ranging from 426 motifs for DP to 278 motifs for CD4 Tregs, compared to males, ranging from 328 motifs for

CD4 Teffs to 34 motifs for CD4 Tregs (Figure 7A). The identified motifs are mainly local motifs and mostly restricted to their respective sex group, meaning that motifs identified in females are found almost exclusively in females, and vice versa (Supplemental Figure 9).

We observed minimal motif overlap between different cell subtypes (Figure 7B). Only two female associated motifs overlapped: one between CD4 Teffs and CD4 Tregs, and another between CD4 Teffs and CD8 cells (Figure 7B). In contrast, no overlap was found between male-associated motifs (Figure 7B). Additionally, we noted that one motif associated with female CD8 cells motif overlapped with a motif associated with male CD4 Teff cells (Figure 7B).

We next evaluated the usage of the differentially expressed motifs in external datasets. No public thymic TCR dataset reports repertoires according to cell subsets. We thus tested these motifs on TCRs from pediatric bulk thymocytes, which contains TCR data from children aged two to eight months with a male-to-female ratio of 5:3 (35, 36). We calculated the usage of these differentially expressed motifs in each sample of this dataset. We were unable to separate individuals by sex using these motif usages (Figure 7C).

We further evaluated these motifs using a peripheral blood TCR dataset from healthy individuals, where CD8, CD4 Teff, and CD4 Treg cells had been sorted. The usage of these motifs could not distinguish males from females across all cell subsets (Figure 7D, Supplemental Figure 10).

Sex-biased enrichment of TCRs associated with autoimmune diseases and bacterial antigens.

We looked for the usage of TCRs with known specificity as represented in IEDB, Mc-PAS and VDJdb public databases (37–39). We compiled the TCRs from these three databases, retaining only those with high sequence reliability scores (see Materials and Methods). We focused on TRB sequences only, due to their greater representation in the databases (31, 40). We

identified 55 368 unique TRB CDR3aa sequences with high reliability and specificity assignment scores for at least one specificity.

We classified the TCR specificities based on the antigen they recognize: viral or bacterial peptides, or human peptides overexpressed in cancers, autoimmune diseases (AIDs), or neither of these diseases (Supplemental Figure 11A). We looked for these TRB CDR3aa sequences in our thymocyte dataset. Depending on the individuals and cell subtypes, they represented around 0.82% to 3.58% of the TRB repertoires.

We then compared the distribution of these CDR3aa between males and females for each cell subtype. We observed a significant enrichment of unique TRB CDR3aa sequences specific to self-antigens not associated with pathologies and those associated with cancers in our thymic TCR dataset, for both females and males (Supplemental Figure 11B). This enrichment is observed across all cell subtypes and is approximately 1.25 to 1.5 times higher than the baseline distribution of specificities in the pooled database (Supplemental Figure 11B). Strikingly, in the TCR repertoires with one specificity only, we observed in CD8 SP cells, a significantly higher proportion of unique TRB CDR3aa sequences associated with AIDs of females compared to males (Figure 8A, Supplemental Figure 12). Additionally, in the specific TCR repertoires only, there was a trend towards a higher proportion of thymic sequences specific to bacterial compounds in female CD8 SP cells compared to males and an opposite trend observed for CD4 Tregs SP (Figure 8A, Supplemental Figure 12). These differences observed in CD8 SP cells persist when examining the usage of these sequences (i.e. the cumulative usage frequency), with significantly higher usage of sequences specific to bacterial compounds and those associated with AIDs in females compared to males (Figure 8B, Supplemental Figure 12).

Donor age had no significant effect on the sex-biased enrichment of self- and bacteria-specific TRB CDR3 sequences in CD8 SP and CD4 Treg SP cells ($p \geq 0.52$), indicating that the observed differences are age-independent (Supplemental Figure 13). We also investigated for the presence of polyspecific TCRs, which are capable of recognizing multiple antigens from different organisms (41–43). These sequences were enriched across all cell subtypes, compared to their representation in the reference database (Supplemental Figure 11C). However, no significant differences were observed in the proportion or usage of these polyspecific sequences between males and females. Interestingly, an inverse usage trend was noted between CD8 and CD4 Treg SP cells: females exhibited higher usage of these sequences in the CD8 SP repertoire but a lower usage in the CD4 Treg SP repertoire compared to males (Figure 8C-8D).

In conclusion, our analysis of TCR specificity reveals significant sex-specific differences in the thymic selection of CD8 effector T cells specific for autoimmune and bacterial antigens that are overrepresented in female versus males.

Discussion

It is striking that so little is known about possible differences in the TCR repertoire between sexes given the yet unexplained much higher prevalence of autoimmune diseases in females. Exploring the hypothesis that the generation and/or the selection of the TCR repertoires may be different between sexes requires studying the TCR repertoire where they are formed and selected, in the thymus. We thus generated a unique and valuable dataset comprising thymic samples collected from deceased organ donors of various ages, and young children with cardiac surgery. Most importantly, we separated the various thymocytes population to be able to study both TCR repertoire generation at the level of DP cells, and repertoire selection

at the level of CD8, CD4 Teff and Treg SP cells. We performed bulk sequencing of the TCRs from these cells, the only method that yet generates enough sequences for sensitive comparison. Indeed, single cell sequencing that analyzes on a few thousand cells would not have allowed making the analyses that can be performed with hundreds of thousand cells.

TCR generation

Our analyses first revealed similarities rather than dissimilarities in TCR generation. There were no significant sex differences for most of the variables that we analyzed comparing males and females DP cells' repertoires, except for differences in the usage of few V and J genes from both TCR chains. However, these few differences are likely a result of random variation, considering that we performed hundreds of comparisons. This is supported by the fact that there are no significant differences in the usage of TRBV-TRBJ and TRAV-TRAJ gene associations between males and females. We also observed no differences in diversity indexes in DP cells' repertories between males and females. Altogether, the overall combinatorial diversity of the TCR repertoire remains comparable between sexes.

A detailed analysis of the CDR3aa region in DP cells showed a differential usage of some hydrophobic aa in the FG loop region for both TCR chains. However, no significant sex-based differences were found in the usage of hydrophobic, hydrophilic, or neutral aa at the critical p109 and p110 positions in TRB that have been described as influencing self-antigen recognition due to the impact of hydrophobic aa on TCR self-reactivity (26–28). These findings suggest that the subtle sex-based differences in the FG loop's aa usage are minor and do not significantly impact the overall characteristics of the CDR3 region in a sex-dependent manner.

The Pgen analysis of DP cells revealed only minimal differences in the distribution of Pgen between males and females, which were similar to those observed in control groups. This

indicates that the mechanisms governing TCR generation are largely conserved across sexes, with only minor, likely random, deviations.

Additionally, although we identified TRB CDR3aa sequence motifs differentially expressed between males and females at the DP cell stage in our dataset, these motifs could not reliably differentiate individuals by sex in external datasets. This indicates that they are more dataset specific than sex-specific. This might also be explained by the different characteristics of our data and other dataset. For example, in the Arstila dataset, the analyses were performed on unsorted cells from young subjects. Younger individuals typically have a more diverse and richer TCR repertoire, with greater repertoire sharing in the periphery, which could obscure sex-specific patterns (16, 44). This suggests that these motifs might be context-specific or influenced by other factors such as age, genetic background, or environmental exposures. Moreover, we did not observe any sex-specific differences in CDR3 TRB specificity usage in DP cells' repertoires. In summary, our results from DP cells analyses could not identify any relevant sex differences in TCR generation.

TCR selection

We identified three genes with differential usage between males and females. While some of these differences could be random, resulting from multiple comparisons, the preferential use of TRBV6-5 in females has previously been observed in the peripheral TCR repertoire (45). Notably, this gene is overrepresented in female lupus patients compared to their controls (51,52). Further research is needed to validate this observation and then elucidate its significance.

Our most striking observation relates to significant sex-specific differences in specificities of CD8 T cells, with a biased CD8 selection process in the thymus favoring sequences associated with autoimmune diseases and bacterial antigens in females. The fact that this bias does not

affect DP cells indicates that it is related to TCR selection rather than generation. Its relevance is strongly supported by the fact that (i) there is no such bias for other self-antigens that are not associated with autoimmunity and (ii) the bias affects both the effectors and regulators of autoimmunity with a directionality compatible with the increased prevalence of autoimmune disease in females. TCRs specific for autoimmunity related self-antigen are increased in the CD8 effector T cells while decreased in the CD4 Tregs. It is worth nothing that some bacterial antigens, but not viral antigens, are also affected by this bias. This could potentially be explained by the fact that bacterial antigens are expressing numerous mimotopes of self-antigens and are known to contribute to the shaping of the TCR repertoire. The observed bias towards bacterial antigens in female CD8 T cells could likewise provide more effector cells specific for bacterial antigens that are mimics of self-antigens, contributing to autoimmune disease development by molecular mimicry (48). In this line, antigens linked to celiac disease (CeD) and type 1 diabetes (T1D) are highly represented in the database and these two AIDs have been linked to bacterial components that may mimic self-antigens (49–53). Similarly, it is well-established that the gut microbiota influences thymic development of T cells (54). Sexual hormones could influence this biased selection by affecting the composition of the gut microbiota (55) and/or by affecting the selection of the TCRs specific for these antigens.

Finally, we have recently described polyspecific TCRs that can respond to multiple unrelated viral antigens and hypothesized their possible involvement in autoimmunity (43). We observed a tendency towards higher usage of polyspecific sequences in the CD8 SP repertoire of females, although not statistically significant. This may indicate a preferential cross-activation of CD8 T cells in females, potentially enhancing protection against pathogens as well as increasing the risk of heightened inflammatory reactivity and autoimmunity.

Conversely, the frequency of polyspecific TCR sequence usage was lower in female CD4 Tregs SP cells. This could limit CD4 Treg antigen coverage and be detrimental to prevent autoimmunity. However, little is known about polyspecific TCRs and these observations warrant further investigation.

To further evaluate the robustness of our findings, we performed additional analyses to assess whether donor age could confound the observed sex-specific differences. A potential limitation of our study is the relatively wide age range of donors. However, the absence of age-related clustering in repertoire features (data not shown), and the lack of age impact on CD8 and CD4 Treg bacterial-specific TRB usage and CD8 and CD4 Treg self-associated to autoimmune disease-specific TRB usage patterns (Supplemental Figure 13), argue against age as a confounding factor. These results support an intrinsic sex bias in thymic TCR selection rather than an age-related effect.

While our findings provide valuable insights into sex-specific differences in thymic selection of TCR repertoires, our TCR specificity analyses were limited. Indeed, we only studied the TRB CDR3aa region of the TCR, which is the most critical for defining TCR specificity (31, 32) while both the TRA and the TRB chains contribute to TCR specificity (31, 32). This limitation is intrinsic to the limited number of TRA with assigned specificity in databases and to the fact that, given the nature of our studies, there is a need to study large repertoires, which is currently not feasible by single cell TCR sequencing. However, as methods for inferring specificity continue to evolve, they will likely allow more refined analyses of our dataset in the future (56). In addition, the continuous development of the TCR database should enrich these analyses.

Altogether, our analyses of TCR specificity revealed a biased selection in females towards sequences associated with autoimmune diseases, potentially underlying the higher

prevalence of autoimmune diseases in females. Future research should focus on validating these observations in larger datasets.

Methods

Samples

Thymus samples were obtained from twenty-two organ donors aged between 73 days and 64 years, without any specific pathologies (Figure 1). Adult samples were collected post-surgery at the Cardiac Surgery Departments of many hospitals in France, after approval by the Agence de Biomédecine and the Ministry of Research (#PFS14-009). Pediatric samples were obtained after total thymectomy during cardiac surgeries at Necker Hospital. The male-to-female across donors is balanced to 1:1, although it ranges from 1.1:1 and 1.33:1 depending on the cell subtypes analyzed. Age distributions of male and female donors were compared using two-sample Kolmogorov–Smirnov tests, with no significant differences observed across thymic subsets ($p > 0.46$), indicating no age-related confounding (Supplementary Figure 1).

Thymocyte isolation and RNA extraction

Thymic cell suspensions were prepared through automated tissue dissociation using gentleMACS dissociators (Miltenyi®) or by mechanical disruption, followed by filtration through a 70-µm nylon mesh. Cells were stained with anti-CD3 (AF700), anti-CD4 (APC), anti-CD8 (FITC), and for some samples, anti-CD25 (PE) antibodies. Cell sorting was performed by fluorescent activated cell sorting (FACS) on a Becton Dickinson FACSAriaII, achieving a purity of >85%: Populations sorted included DP CD3+ (CD3+CD4+CD8+), SP CD8+ (CD3+CD4-CD8+), SP CD4+ (CD3+CD4+CD8-), with further separation into SP CD4 Teff (CD3+CD4+CD8-CD25-)

and SP CD4 Treg (CD3+CD4+CD8-CD25+) for sixteen of samples (Figure 1). For the sorted SP CD4, the Treg percentage varied between 5.8% and 16% across samples, with an average of 9.63%. Six SP CD4 samples were not sorted into Teffs and Tregs, we thus considered them as SP CD4 Teff by assuming a purity of Teff within the total SP CD4 around 90%.

RNA was isolated using the RNAqueous extraction kit (Invitrogen®), according to the manufacturer's protocol. The RNA were quantified and sample integrity were determined on Nanodrop (ThermoFisher) or on Tapestation 4200 (Agilent, RNA screentape).

TCR library preparation and sequencing

TCR library preparation and sequencing were performed as described by Barennes et al. (57). Briefly, RNA was processed using the SMARTer Human TCR a/b profiling v1 kit (Takarabio) according to the manufacturer's protocol. Amplicons were purified with AMPure XP beads (Beckman Coulter), quantified, and their integrity assessed using 2100 Bioanalyzer System (Agilent, DNA 1000 kit) or Tapestation 4200 (Agilent, D1000 screentape). Bulk Next-Generation Sequencing was performed using either a Hiseq 2500 (Illumina) with SR-300 protocol plus 10% PhiX on the LIGAN-PM Genomics platform (Lille, France), or a Novaseq 6000 (Illumina) with PE-250 protocol plus 10% PhiX on the LIGAN-PM Genomics platform or ICM platform (Paris, France).

Data processing

Raw FASTQ/FASTA files were aligned to TRA and TRB using MiXCR (v3.0.13, RNA-seq parameters), which corrects for sequencing and PCR errors (58). Samples with fewer than 1,000 CDR3aa sequences, an imbalanced TRA/TRB ratio (<1/9), or a low sequence count relative to sorted cell counts (< 0.9 ratio) were removed. Sequences with CDR3aa lengths outside 6-23 amino acids were also discarded. To standardize clonotype counts and unique

number of clonotypes across sequencing techniques, initial counts for each clonotype were reduced by 1 to exclude those with counts of 0 (singletons).

Additionally, due to high similarity between certain TRBV sequences, which could bias the analysis depending on the sequencing method used, TRBV06-2 and TRBV06-3 were merged as TRBV06-2/3, and TRBV12-3 and TRBV12-4 were merged as TRBV12-3/4 (59).

Gene usage analysis

For both TRA and TRB chain across all cell subtypes, we analyzed the V and J gene usage and those of the VJ gene associations (Figure 1).

The usage of gene is calculated as follow: $\frac{\Sigma i \epsilon G^{ci}}{\Sigma j \epsilon S^{cj}}$

Where:

- $\Sigma i \epsilon G^{ci}$ represents the sum of the counts of the clonotypes using the interested gene or gene association
- $\Sigma j \epsilon S^{cj}$ represents the total sum of the counts of all clonotypes in the sample

In this notation, G is the set of clonotypes using the interested gene, S is the set of all clonotypes in the sample, and c_i and c_j are the counts of the clonotypes *i* and *j*, respectively.

The usage of the V gene and the VJ gene combinations was investigated through dimensional reduction using Principal Component Analysis (PCA) (MASS package v7.3-60, factoMineR package v2.8, factoextra package v1.0.7). Confidence ellipses were represented with a 5% alpha risk. Additionally, for VJ gene combinations usage, a heatmap was constructed using the Jensen-Shannon Divergence (JSD) distance to measure the distance between samples based on their gene usage distribution (complexHeatmap package v2.16.0, philentropy package v0.7.0). Hierarchical clustering was performed using the Euclidean distance and the complete linkage method.

Diversity analysis

For the TCR repertoire diversity analysis, to correct for sequencing depth bias and differences in cellular richness between individuals, rarefactions were performed fifty times for each sample, using X clonotypes, where X represents the effective diversity number (60).

Effective diversity was calculated as the exponential of the Shannon index (e^H) (61).

Rényi indices (11 indices from 0 to Infinity) were calculated for each rarefaction (62).

The median Shannon index, Simpson index and Berger-Parker index were specifically compared between sex groups.

Distribution of CDR3aa length analysis

The usage of all CDR3 by the amino acid length is calculated as follow: $\frac{\Sigma i \epsilon L^{ci}}{\Sigma j \epsilon S^{cj}}$

Where

- $\Sigma i \epsilon L^{ci}$ represents the sum of the counts of the CDR3 characterized by the amino acid length of interest.
- $\Sigma j \epsilon S^{cj}$ represents the total sum of counts of all CDR3s per sample.

In this notation, L is the set of CDR3s characterized by the amino acid length of interest, S is the set of all CDR3s in the sample, and c_i and c_j are the counts of the CDR3s *i* and *j*, respectively.

Usage of amino acid in CDR3 region

For each CDR3aa sequence, alignment was performed following the IMGT convention (63).

The aa usage is defined at different scales. Firstly, usage at the sequence level, within the CDR3 loop region (positions p108 to p114), involved in MHC-peptide complex interactions, assessed for CDR3aa sequences 9-21 aa long in both TRA and TRB chains (63). This aa usage was defined as the percentage of their representation within the sequence as follow:

(1) CDR3aa loop usage (%) = $\frac{n_{aa}}{L}$ x 100

Where n_{aa} is the count of the specific amino acid within the sequence and *L* is the total length of the sequence.

Secondly, the CDR3aa usage is calculated at particular positions, the p109 and p110 in the repertoire of each individual. The usage of each aa a at position p is calculated as the proportion of CDR3 sequences in which a appears at p and can be defined as follow:

(2) CDR3aa usage in position (%) = $\frac{N_{a,p}}{n}$ x 100

Where n is the total number of CDR3 sequences and $N_{a,p}$ the number of sequences.

Additionally, aa were classified into three hydropathy categories (neutral, hydrophobic, and hydrophilic, using the Kyte-Doolitle scale (25, 64).

Probability of generation analysis

The probability of generation analysis of DP and CD8 cell subtypes consisted initially in the creation of a generative model of V(D)J recombination using 400,000 randomly selected non-productive sequences from all samples for both TRA and TRB chains, with the IGOR tool (65). Generation probabilities (Pgen) for all sequences in the dataset were then calculated using the OLGA tool, based on the respective generative models for each cell subtype and TCR chain (66). Comparisons of Pgen values at the nucleotide level between males and females were performed. To ensure statistical robustness, control groups with equivalent male and female sample numbers were created by permuting the samples into two groups 20,000 times.

TCR network structure by sequence similarity

For each sample, 100 random sub-samplings of CDR3aa sequences were performed, using the minimum CDR3aa count per cell subtype. Two CDR3aa were connected if their Levenshtein distance was = 1 (Figure 1) (67). Network analysis focused on two metrics: the proportion of connected sequences and network density, defined as the ratio of actual connections to all possible connections. Median values across 100 subsampling iterations were used for

analysis. Levenshtein distances were computed with the stringdist package (v0.9.10), networks were visualized using Cytoscape (v3.8.2), and network density was calculated with igraph package (v1.5.1).

TRB CDR3aa motif search

Two types of structural motifs were searched: local motifs (strict sequential motifs) and global motifs (sequences with substitutions preserving a positive BLOSUM62 score) (Figure 1). Enriched motifs were selected using gliph2 function with turboGliph package (v0.99.2), with specific parameters defined as follow:

- Input sequences were from the sex group of interest (male or female), with reference sequences matched for TRBV gene usage
- CDR3aa sequences were required to be longer than eight aa
- Local motif lengths were restricted to 3-5 aa
- Only motifs constituted with strictly more than two unique CDR3aa are conserved
- No boost of local motif significance of Fisher test

Additional filters were applied so that: (i) a motif includes public CDR3aa sequences (shared by at least two individuals), (ii) a significant enrichment (Fisher's test, $p < 0.01$) and (iii) a usage difference between groups of at least twofold (Wilcoxon test, $p < 0.05$).

Motif validation was performed on an publicly published pediatric thymic dataset from Arstila team (35, 36) consisting in a bulk TCR dataset from infants aged from seven days to eight months. Here, we removed one boy named “Thymus A” corresponding to a twin of the “Thymus B” and the “Thymus 1” neonate sample so that we analyze height samples with a male-to-female ratio of 5:3. Additional validation was carried out on sorted peripheral blood TCR repertoires (CD8, CD4 Teff and CD4 Treg) from seventy-five healthy volunteers aged from eighteen to eighty-four years (male-to-female ratio of 1.03:1), whose libraries were

generated under the same conditions as those of our thymic dataset. They were recruited within the Transimmunom observational trial (NCT02466217) and HEALTHIL-2 (NCT03837093) (68). PBMCs from healthy individuals were isolated using Ficoll density gradient centrifugation and enriched for CD4+ T cells via EasySep™ Human CD4+ T Cell magnetic beads (Stemcell). Further separation into effector T cells (CD4+ CD25-) and Tregs (CD4+ CD25+) was performed using EasySep™ Human Pan-CD25 magnetic beads (Stemcell). Purity was assessed by flow cytometry based on the expression of CD4 and FoxP3, with a purity threshold of >80%. From the frozen PBMC aliquots, CD8+ T cells (CD3+ CD8+) were sorted using a FACS ARIA II cell sorter.

TRB CDR3aa specificity analysis

We established a curated and harmonized TCR specificity database by integrating data from three public TCR repositories: Mc-PAS, IEDB, and VDJdb (38, 39, 69) (Figure 1). Reliability levels for sequences and specificity assignments were added based on the identification methods used in the original studies, ensuring a robust dataset for specificity analysis (70). Specificity groups were categorized by antigen class, including viral, bacterial, yeast, parasitic and self-antigens linked to diseases (e.g. autoimmune conditions, cancer) or unrelated to diseases, as well as antigens from wheat, plants or animals.

This harmonized database was used to identify exact matches with the CDR3aa sequences of our TRB sequences. Enrichment of CDR3aa sequences associated with a given specificity group in male and female thymic repertoires for each cell subtype compared to the specificity distribution in the database was tested. Specificity group distributions between male and female samples were also performed, as was the overall usage of these sequences across the entire TCR repertoire.

Additionally, we focused on polyspecific sequences defined as CDR3aa TRB sequences recognizing at least two different species in case of microorganisms, or belonging to multiple specificity groups for other antigens. The same analyses as described above were conducted for these polyspecific sequences.

Statistical analysis

For V-J gene usage, CDR3aa length usage, CDR3aa usage, network similarity and specificity analyses, the comparison between males and females were statistically analysed using the Wilcoxon test. The statistical tests were performed using ggpubr library (v0.6.0) in R. Significance was denoted by the number of asterisks from one to four, each representing respectively $p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$, and $p \leq 0.0001$. The comparisons of the Rényi curves and Pgen were evaluated using the Kolmogorov-Smirnov test. The comparisons of the proportions of specific sequences in male and female relative to those in the specificity database were tested using the Wilcoxon test.

Data availability

Scripts an data supporting the findings of this study can be available upon request.

Author contributions

VQ and HV conducted all cell sorting. RNA extraction and TCR library preparation were performed by HV, VQ, LA, PB, NC, VD, JD, GF, KL and PS. HV, VQ, VM, LA, KL, ONT, MP, PS, AS, EMF and DK contributed valuable suggestions for experimental design and participated in discussions of the results. HV and CJ pooled and curated the reference specificity database. LA, CJ and VM contributed equally in this work. HV authored the manuscript, with VQ, VM, CA, JD, EMF and DK providing comments and critical reviews. DK conceptualized, supervised and provided funding for the study.

Acknowledgment

We would like to thank Bruno Gouritin for his assistance with cell sorting and to Marie Surroque for her support with RNA quantification.

References

1. C. Libert, L. Dejager, I. Pinheiro, The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol* **10**, 594–604 (2010).

2. L. C. Golden, R. Voskuhl, The Importance of Studying Sex Differences in Disease: The Example of Multiple Sclerosis. *J Neurosci Res* **95**, 633–643 (2017).

3. T. Sokka, *et al.*, Women, men, and rheumatoid arthritis: analyses of disease activity, disease characteristics, and treatments in the QUEST-RA Study. *Arthritis Res Ther* **11**, R7 (2009).

4. Y. Li, M. Jerkic, A. S. Slutsky, H. Zhang, Molecular mechanisms of sex bias differences in COVID-19 mortality. *Crit Care* **24**, 405 (2020).

5. D. Hertz, B. Schneider, Sex differences in tuberculosis. *Semin Immunopathol* **41**, 225–237 (2019).

6. S. L. Klein, R. Morgan, The impact of sex and gender on immunotherapy outcomes. *Biology of Sex Differences* **11**, 24 (2020).

7. H.-J. Lee, C. Chang, Recent advances in androgen receptor action. *CMLS, Cell. Mol. Life Sci.* **60**, 1613–1622 (2003).

8. P. Chen, B. Li, L. Ou-Yang, Role of estrogen receptors in health and disease. *Front Endocrinol (Lausanne)* **13**, 839005 (2022).

9. K. M. Scarpin, J. D. Graham, P. A. Mote, C. L. Clarke, Progesterone action in human tissues: regulation by progesterone receptor (PR) isoform expression, nuclear positioning and coregulator expression. *Nucl Recept Signal* **7**, e009 (2009).

10. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol* **14**, 377–391 (2014).

11. M.-L. Zhu, *et al.*, Sex bias in CNS autoimmune disease mediated by androgen control of autoimmune regulator. *Nat Commun* **7**, 11350 (2016).

12. N. Dragin, *et al.*, Estrogen-mediated downregulation of AIRE influences sexual dimorphism in autoimmune diseases. *J Clin Invest* **126**, 1525–1537.

13. L. N. Stankiewicz, *et al.*, Sex-biased human thymic architecture guides T cell development through spatially defined niches. *Dev Cell* S1534-5807(24)00539–2 (2024). https://doi.org/10.1016/j.devcel.2024.09.011.

14. A. Murugan, T. Mora, A. M. Walczak, C. G. Callan, Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16161–16166 (2012).

15. K. Yoshida, *et al.*, Aging-related changes in human T-cell repertoire over 20 years delineated by deep sequencing of peripheral T-cell receptors. *Experimental Gerontology* **96**, 29–37 (2017).

16. A. Trofimov, *et al.*, Two types of human TCR differentially regulate reactivity to self and non-self antigens. *iScience* **25**, 104968 (2022).

17. J. Mika, K. Yoshida, Y. Kusunoki, S. M. Candéias, J. Polanska, Sex- and age-specific aspects of human peripheral T-cell dynamics. *Front Immunol* **14**, 1224304 (2023).

18. M. Gong, *et al.*, Age-related changes in the TRB and IGH repertoires in healthy adult males and females. *Immunology Letters* **240**, 71–76 (2021).

19. Q. Qi, *et al.*, Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U S A* **111**, 13139–13144 (2014).

20. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).

21. F. Wang, *et al.*, On defining the rules for interactions between the T cell receptor and its ligand: a critical role for a specific amino acid residue of the T cell receptor beta chain. *Proc Natl Acad Sci U S A* **95**, 5217–5222 (1998).

22. K. Yu, J. Shi, D. Lu, Q. Yang, Comparative analysis of CDR3 regions in paired human αβ CD8 T cells. *FEBS Open Bio* **9**, 1450–1459 (2019).

23. J. A. Carter, *et al.*, Single T Cell Sequencing Demonstrates the Functional Role of αβ TCR Pairing in Cell Lineage and Antigen Specificity. *Front Immunol* **10**, 1516 (2019).

24. K. C. Garcia, *et al.*, Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science* **279**, 1166–1172 (1998).

25. C. Pommié, S. Levadoux, R. Sabatier, G. Lefranc, M.-P. Lefranc, IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of Molecular Recognition* **17**, 17–32 (2004).

26. M. Khosravi-Maharlooei, *et al.*, Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire. *J Clin Invest* **129**, 2446–2462 (2019).

27. J. Lu, *et al.*, Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nat Commun* **10**, 1019 (2019).

28. B. D. Stadinski, *et al.*, Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat Immunol* **17**, 946–955 (2016).

29. Y. Elhanati, A. Murugan, C. G. Callan, T. Mora, A. M. Walczak, Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci U S A* **111**, 9875–9880 (2014).

30. P. Meysman, *et al.*, On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics* **35**, 1461–1468 (2019).

31. I. Springer, N. Tickotsky, Y. Louzoun, Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Front Immunol* **12**, 664514 (2021).

32. D. Korpela, *et al.*, EPIC-TRACE: predicting TCR binding to unseen epitopes using attention and contextualized embeddings. *Bioinformatics* **39**, btad743 (2023).

33. H. Huang, C. Wang, F. Rubelt, T. J. Scriba, M. M. Davis, Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol* **38**, 1194–1202 (2020).

34. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–10919 (1992).

35. N. Heikkilä, *et al.*, Characterization of human T cell receptor repertoire data in eight thymus samples and four related blood samples. *Data in Brief* **35**, 106751 (2021).

36. J. Mattila, *et al.*, Analysis of thymic generation of shared T-cell receptor α repertoire associated with recognition of tumor antigens shows no preference for neoantigens over wild-type antigens. *Cancer Med* **12**, 13486–13496 (2023).

37. R. Vita, *et al.*, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* **47**, D339 (2018).

38. N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).

39. M. Shugay, *et al.*, VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* **46**, D419–D427 (2018).

40. E. Rosati, *et al.*, Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology* **17**, 61 (2017).

41. L. Wooldridge, *et al.*, A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* **287**, 1168–1177 (2012).

42. S. R. Hadrup, *et al.*, Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat Methods* **6**, 520–526 (2009).

43. V. Quiniou, *et al.*, Human thymopoiesis produces polyspecific CD8+ α/β T cells responding to multiple viral antigens. *Elife* **12**, e81274 (2023).

44. C. E. van de Sandt, *et al.*, Newborn and child-like molecular signatures in older adults stem from TCR shifts across human lifespan. *Nat Immunol* **24**, 1890–1907 (2023).

45. T. Schneider-Hohendorf, *et al.*, Sex bias in MHC I-associated shaping of the adaptive immune system. *Proc Natl Acad Sci U S A* **115**, 2168–2173 (2018).

46. A. M. Roca, B. I. Chobrutskiy, B. M. Callahan, G. Blanck, T-cell receptor V and J usage paired with specific HLA alleles associates with distinct cervical cancer survival rates. *Hum Immunol* **80**, 237–242 (2019).

47. X. Hou, *et al.*, Characterisation of T and B cell receptor repertoire in patients with systemic lupus erythematosus. *Clin Exp Rheumatol* **41**, 2216–2223 (2023).

48. M. Rojas, *et al.*, Molecular mimicry and autoimmunity. *Journal of Autoimmunity* **95**, 100–123 (2018).

49. J. Repac, B. Božić, B. Božić Nedeljković, Microbes as triggers and boosters of Type 1 Diabetes - Mediation by molecular mimicry. *Diabetes Res Clin Pract* **202**, 110824 (2023).

50. K. Girdhar, *et al.*, A gut microbial peptide and molecular mimicry in the pathogenesis of type 1 diabetes. *Proc Natl Acad Sci U S A* **119**, e2120028119 (2022).

51. J. Petersen, *et al.*, T cell receptor cross-reactivity between gliadin and bacterial peptides in celiac disease. *Nat Struct Mol Biol* **27**, 49–61 (2020).

52. D. S. Vazquez, H. M. Schilbert, V. I. Dodero, Molecular and Structural Parallels between Gluten Pathogenic Peptides and Bacterial-Derived Proteins by Bioinformatics Analysis. *Int J Mol Sci* **22**, 9278 (2021).

53. N. Tai, *et al.*, Microbial antigen mimics activate diabetogenic CD8 T cells in NOD mice. *J Exp Med* **213**, 2129–2146 (2016).

54. R. Hebbandi Nanjundappa, C. Sokke Umeshappa, M. B. Geuking, The impact of the gut microbiota on T cell ontogeny in the thymus. *Cell. Mol. Life Sci.* **79**, 221 (2022).

55. J. G. M. Markle, *et al.*, Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* **339**, 1084–1088 (2013).

56. Z. Ghoreyshi, J. George, Quantitative approaches for decoding the specificity of the human T cell repertoire. *Frontiers in Immunology* **14** (2023).

57. P. Barennes, *et al.*, Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol* **39**, 236–245 (2021).

58. D. A. Bolotin, *et al.*, MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* **12**, 380–381 (2015).

59. A. Omer, *et al.*, T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Medicine* **14**, 2 (2022).

60. W. Chaara, *et al.*, RepSeq Data Representativeness and Robustness Assessment by Shannon Entropy. *Front Immunol* **9**, 1038 (2018).

61. C. E. Shannon, The mathematical theory of communication. 1963. *MD Comput* **14**, 306–317 (1997).

62. A. Rényi, "On Measures of Entropy and Information" in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, (University of California Press, 1961), pp. 547–562.

63. M.-P. Lefranc, *et al.*, IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology* **27**, 55–77 (2003).

64. J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132 (1982).

65. Q. Marcou, T. Mora, A. M. Walczak, High-throughput immune repertoire analysis with IGoR. *Nat Commun* **9**, 561 (2018).

66. Z. Sethna, Y. Elhanati, C. G. Callan, A. M. Walczak, T. Mora, OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).

67. A. Madi, *et al.*, T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife* **6**, e22057 (2017).

68. R. Lorenzon, *et al.*, Clinical and multi-omics cross-phenotyping of patients with autoimmune and autoinflammatory diseases: the observational TRANSIMMUNOM protocol. *BMJ Open* **8**, e021037 (2018).

69. R. Vita, *et al.*, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **47**, D339–D343 (2019).

70. C. Jouannet, H. Vantomme, D. Klatzmann, E. Mariotti-Ferrandiz, Benchmarking unsupervised methods for inferring TCR specificity. [Preprint] (2025). Available at: https://www.biorxiv.org/content/10.1101/2024.10.26.620398v2 [Accessed 20 January 2025].

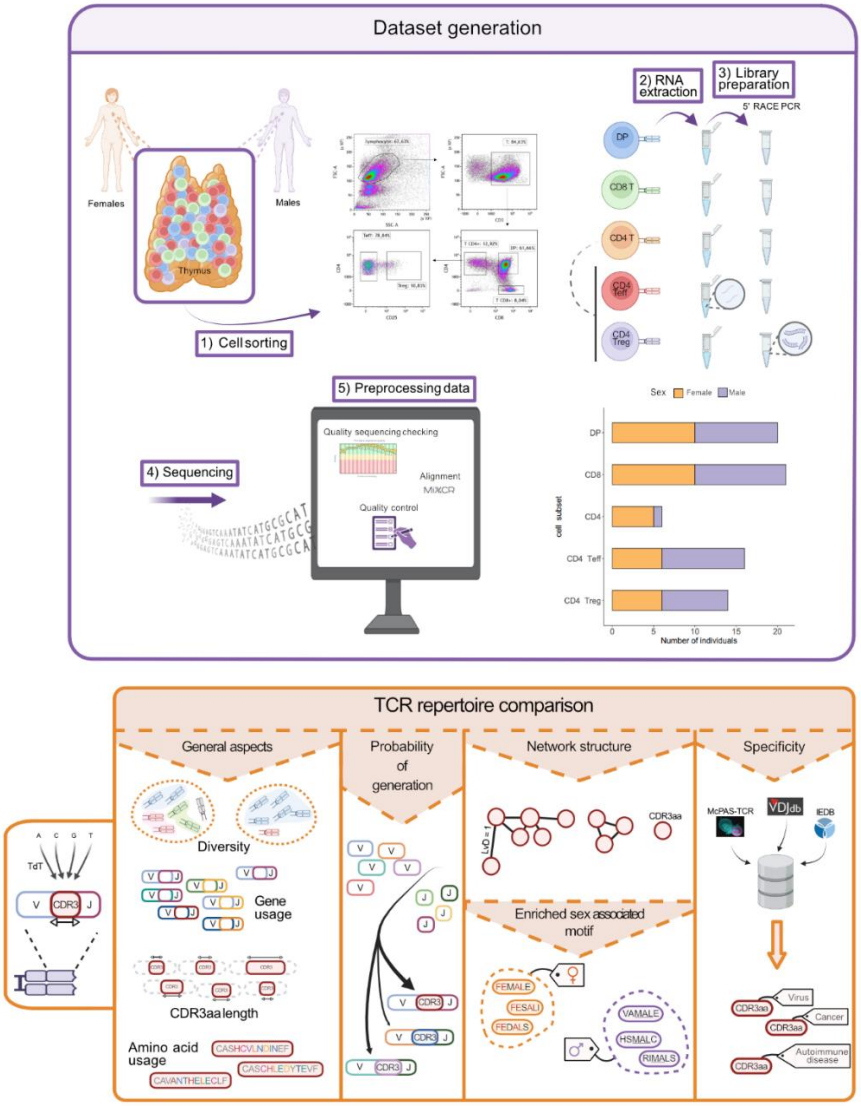

Figure 1: Schematic overview of the generation of thymic TCR dataset and analytical pipeline.

Top Panel: Generation of the thymic TCR dataset. From deceased human thymuses of males and females, we isolated key T-cell subtypes through cell sorting. These subtypes included double-positive (DP) cells (CD3+ CD4+ CD8+), single-positive (SP) CD8+ cells (CD3+ CD4- CD8+), SP CD4+ cells (CD3+ CD4+ CD8-), and further separated SP CD4+ cells into T effector (Teff, CD3+ CD4+ CD8- CD25-) and Treg (CD3+ CD4+ CD8- CD25+) cells. TCR libraries were generated from the RNA of each cell population using rapid amplification of cDNA-ends by PCR (5'RACE PCR). Following sequencing, data preprocessing involved quality sequencing checks, contig alignment, and quality control. The final dataset comprised 20 DP samples (male-to-female ratio of 1:1), 21 SP CD8+ samples (1.1:1), 6 SP CD4+ samples (1:5), 16 SP CD4 Teff samples (1.67:1), and 14 SP CD4 Treg samples (1.33:1). Males are depicted in violet and females in orange.

Bottom Panel: Analytical pipeline. We compared the TCR repertoires of males and females across various dimensions. General aspects of the TCR repertoire were evaluated, including diversity, gene usage, CDR3aa length distribution, and aa usage within the CDR3 region. Additionally, we analyzed the probability of sequence generation, the TCR repertoire structure based on CDR3aa sequence similarity, identification of differentially expressed TRB CDR3aa motifs between sex, and TRB CDR3aa sequence specificity.

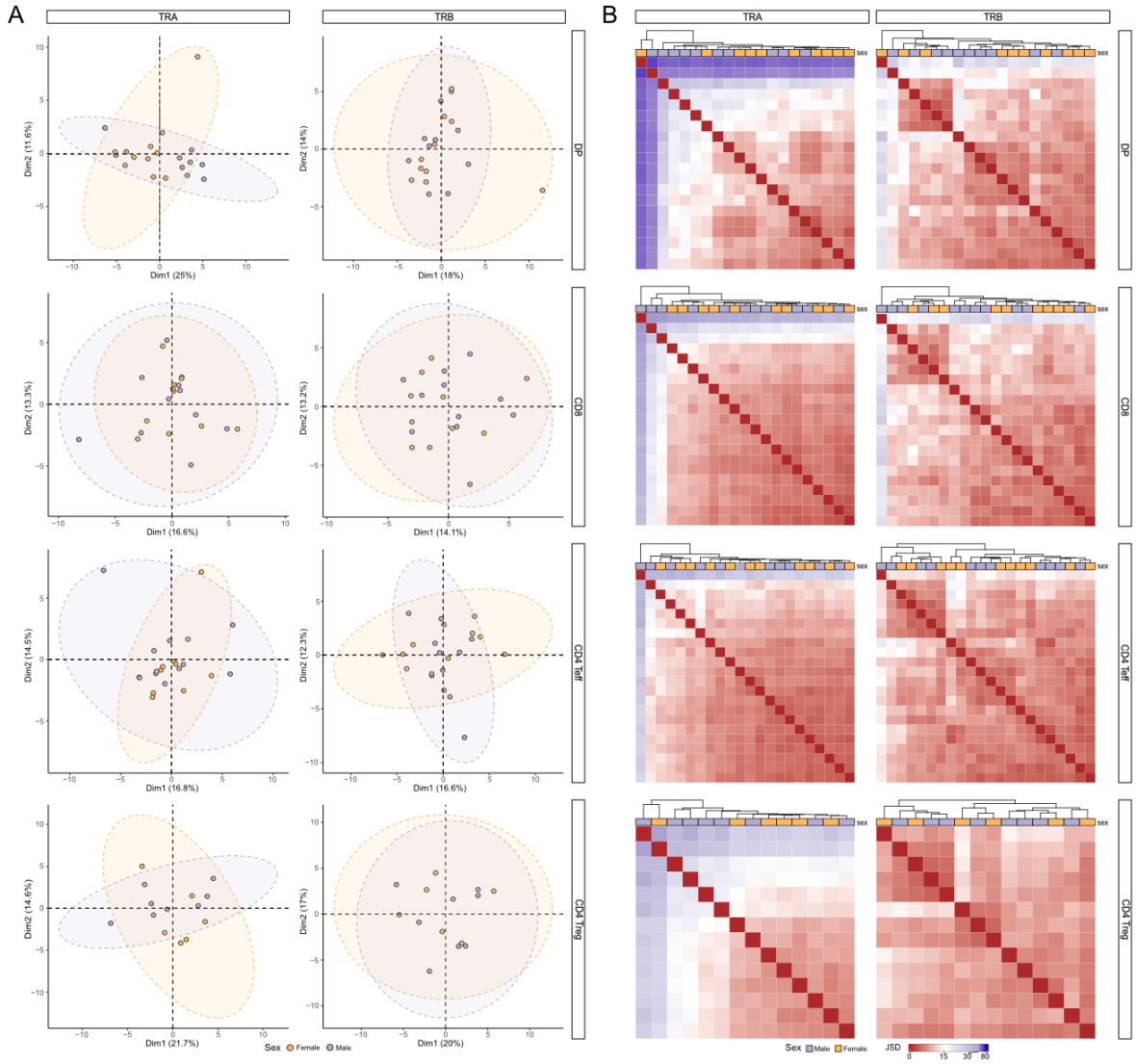

Figure 2: Comparable overall TCR gene usage between males and females.

(A) Principal Component Analysis (PCA) derived from the distribution of TRAV (left) and TRBV (right) gene usage frequencies across sex groups (males vs females), showing results for DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14) cells (displayed from top to bottom). Each point represents an individual. Ellipses indicate 95% confidence intervals. **(B)** Heatmap showing the Jensen-Shannon Divergence (JSD) score between samples, derived from the distributional usage of TRAV-TRAJ (left) and TRBV-TRBJ (right) gene associations in DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14) cells (displayed from top to bottom). Hierarchical clustering was performed using the Euclidean distance and the complete linkage method. Males are shown in violet; females in orange.

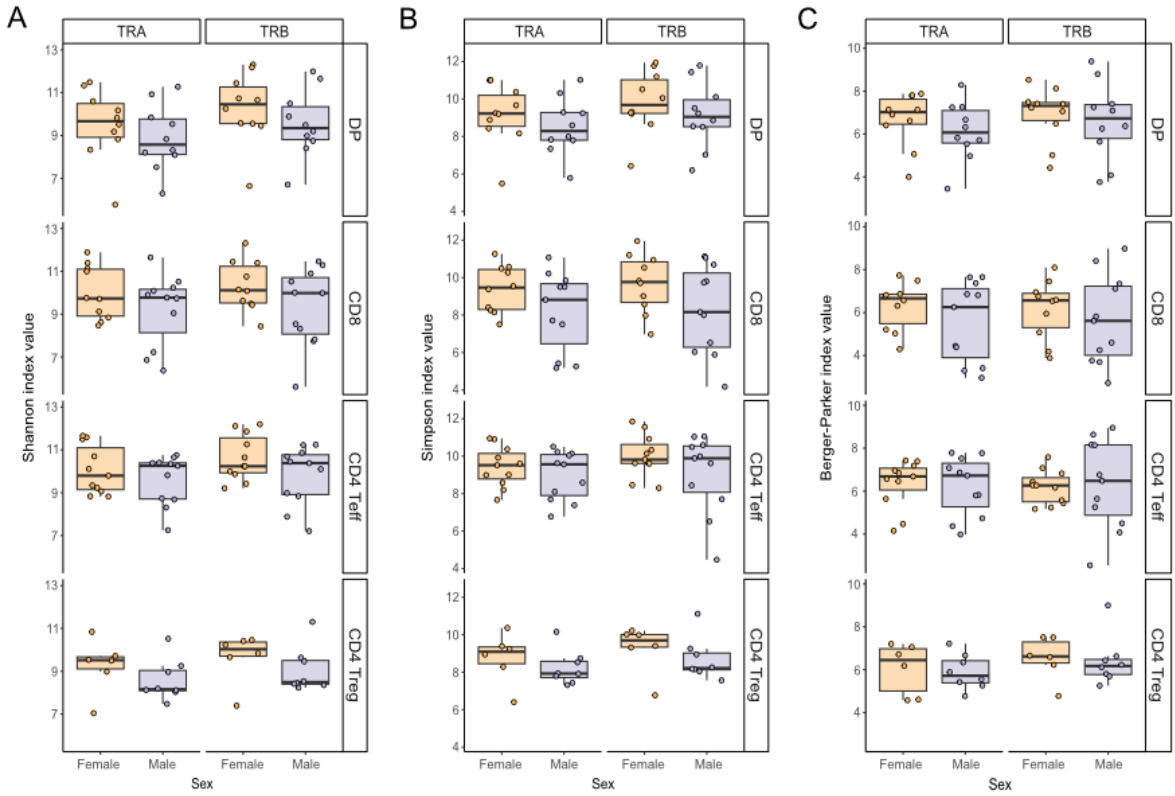

Figure 3: Comparable thymic TCR repertoire diversity between males and females.

Boxplots display Shannon **(A)**, Simpson **(B)**, and Berger-Parker **(C)** index values for TRA (left) and TRB (right), across thymic T cell subtypes, displayed from top to bottom, in DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14). Each point represents the median value from 50 rarefactions per sample. Statistical analysis (Wilcoxon test) showed no significant sex bias in TCR repertoire diversity ($p > 0.05$). Males are shown in violet; females in orange.

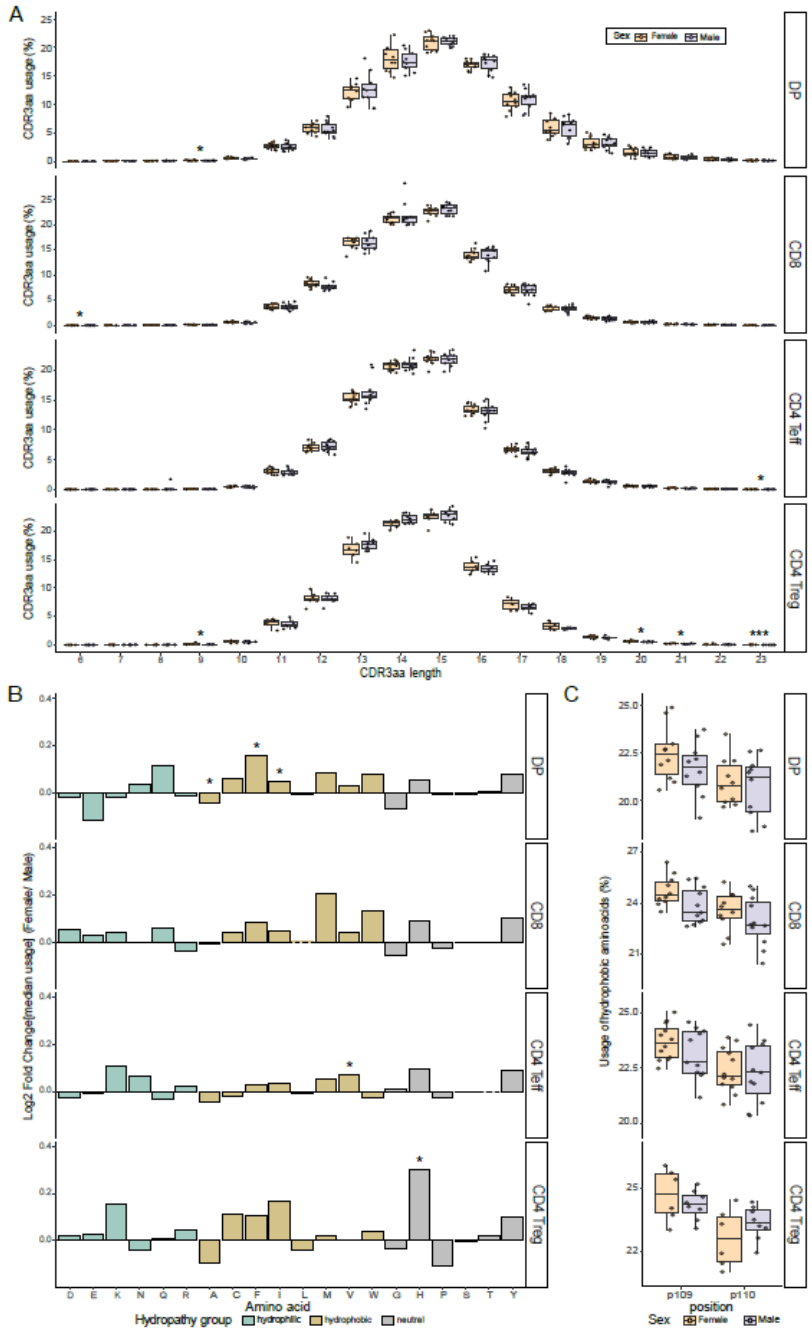

Figure 4: CDR3aa length and amino acid composition of TRB CDR3s in males and females.

(A) Distribution of TRB CDR3aa length usage in DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and Treg (N = 14) SP cells (displayed from top to bottom). Stars indicate statistical differences between males and females based on the p-value of the Wilcoxon test (*: $p < 0.05$, **: $p < 0.01$). **(B)** Amino acid (aa) usage between males and females represented as the log2 fold change of the median usage of females over males for each aa in the p108 to p114 CDR3aa region for TRB. A line at log2 fold change = 0 indicates the direction of the difference in usage frequency. Bars are colored according to the hydropathy class of the aa (neutral in gray, hydrophilic in blue-green and hydrophobic in gold). Stars indicate statistical differences of usage between males and females based on the p-value of the Wilcoxon test (*: $p < 0.05$, **: $p < 0.01$). **(C)** Cumulative hydrophobic aa usage for each individual at p109 and p110 positions in TRB across thymic T cell subtypes, including

DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14). Statistical analysis showed no significant sex bias (Wilcoxon test: $p > 0.05$ for the two positions). Males are depicted in violet; females in orange.

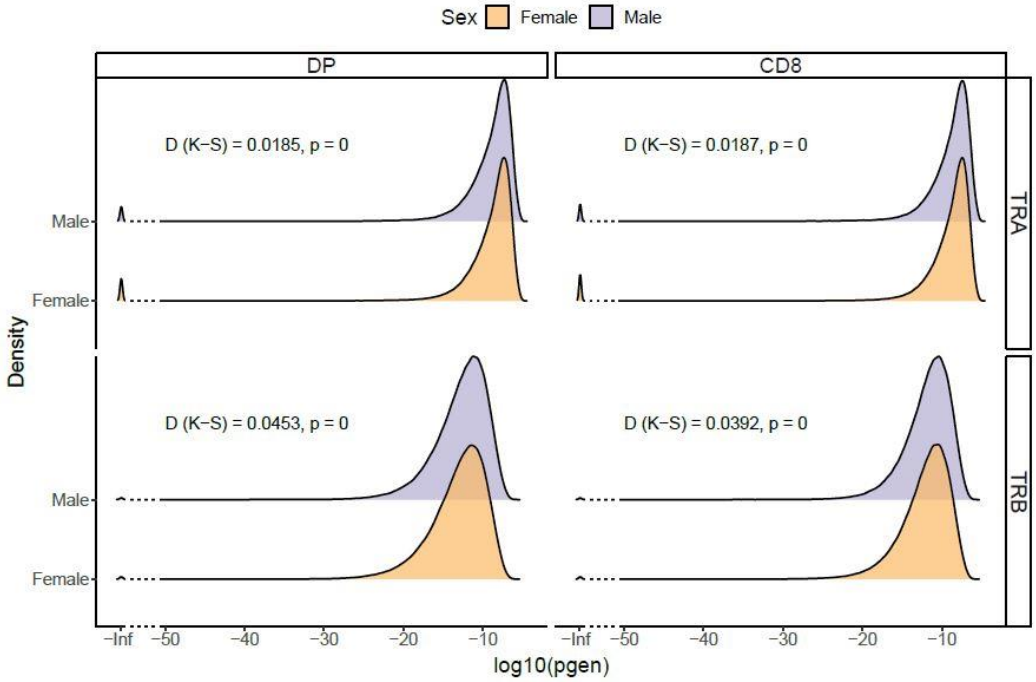

Figure 5: Probabilities of generation of TCRs in males and females.

The figure shows the distribution of Pgen (probability of generation) sequences between males and females for TRA and TRB in DP and CD8 cells. A V(D)J recombination model was created using 400,000 non-productive random sequences derived from the nonproductive sequences of all individuals for each chain, both for DP and CD8 cells (65). This model was then used to calculate the Pgen of sequences for each individual (66). The overall distribution comparison between males and females was tested using the Kolmogorov-Smirnov test, with the D value and associated p-value indicated. Males are depicted in violet and females in orange.

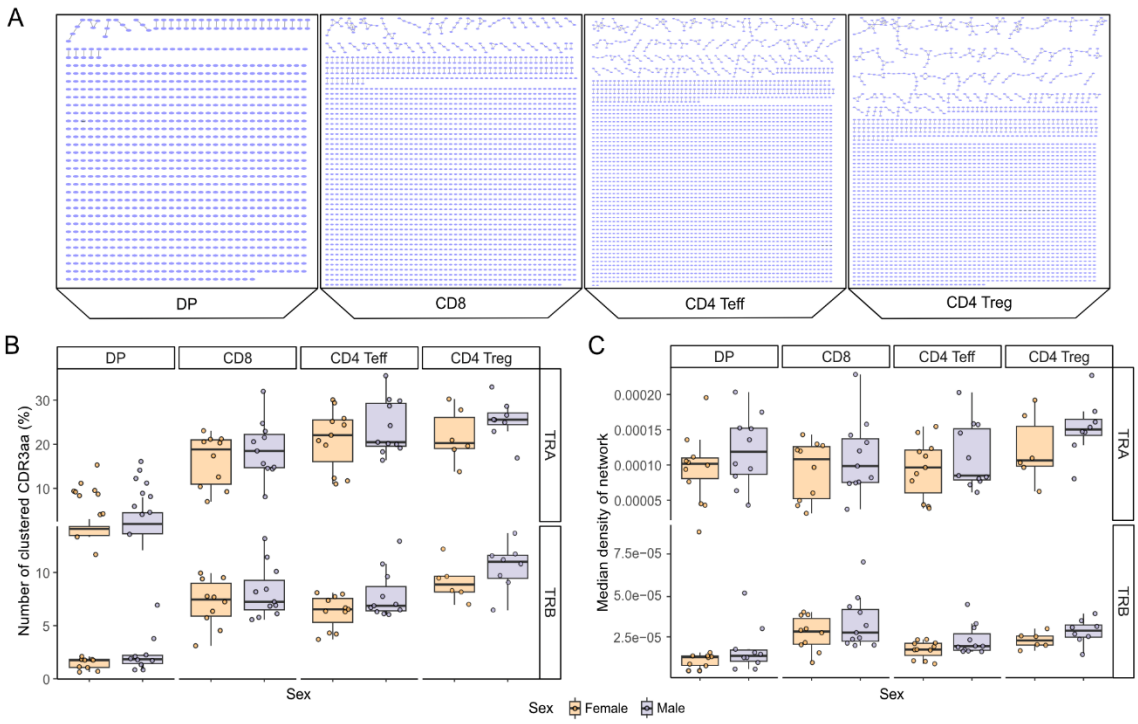

Figure 6: Comparable TCR repertoire network structure based on CDR3 amino acid sequence similarity between males and females. For each sample, 100 random subsamplings were performed on the minimum number of CDR3aa per cell subtype. Two CDR3aa are linked if they have a Levenshtein distance of one. **(A)** TRA Network structure of subsampled TCR repertoire of a male subject for DP, CD8, CD4 Teff and CD4 Treg SP (from left to right). Each point represents a CDR3aa. **(B-C)** Comparison of the proportion of linked sequences **(B)** and network density **(C)** between males and females samples, in DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14). Each point represents the median value from 100 subsampling iterations for each sample. Statistical analysis using the Wilcoxon test revealed no significant sex differences for these two metrics ($p > 0.05$). Males are depicted in violet; females in orange.

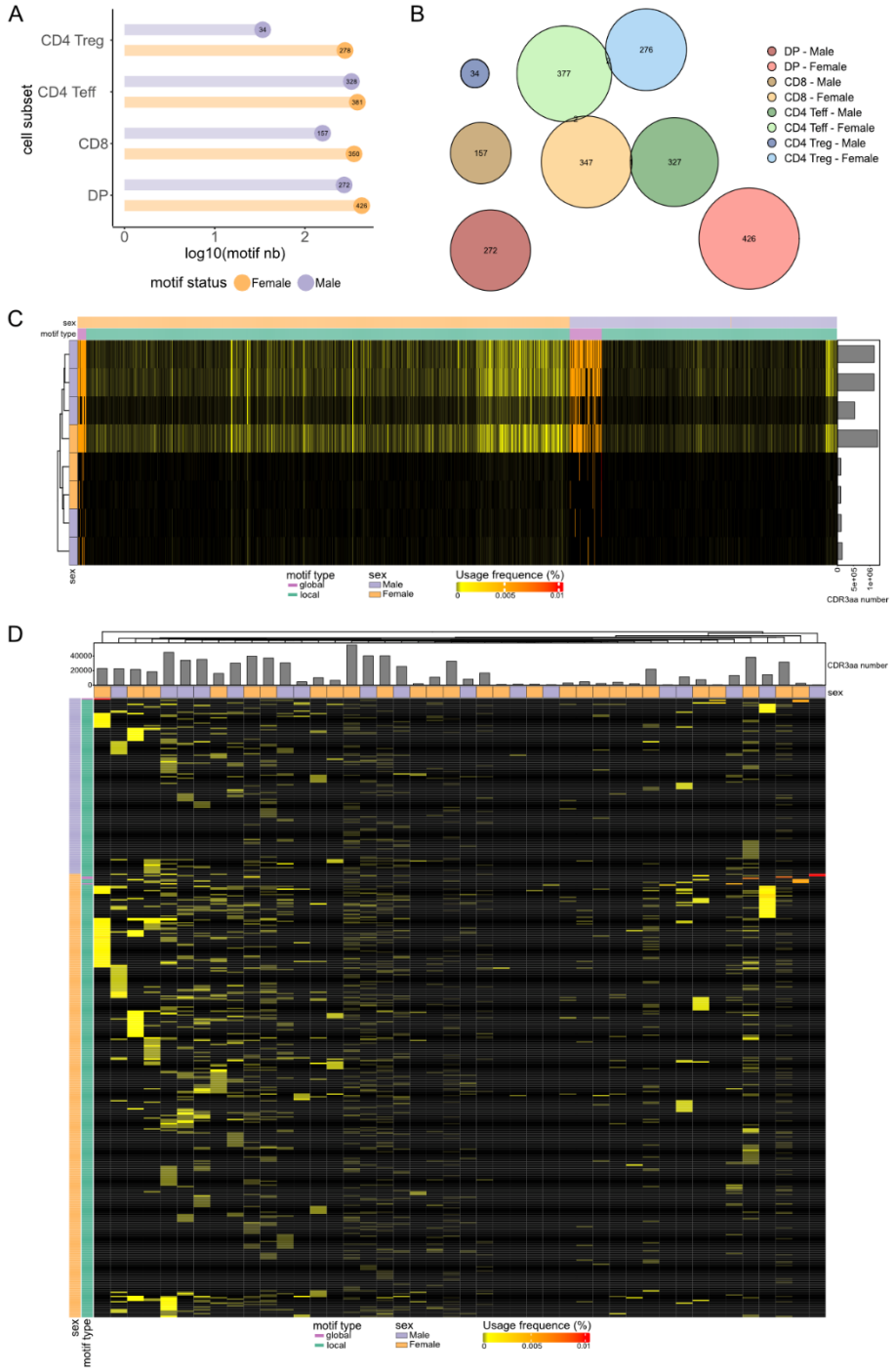

Figure 7: Thymic TRB TCR sex-associated motifs.

Different structural motifs found differentially expressed between males and females in our dataset. We distinguish local motifs as distinct aa sequences and global motifs as motif region with one variable aa position maintaining a BLOSUM62 score ≥ 0. **(A)** Number of male and female associated motifs by cell subset. **(B)** Euler diagram illustrating the distribution and overlap between all sex associated motifs. Numbers indicate the number of motifs in overlap zones. **(C-D)** Validation of these sex associated TRB CDR3aa motifs. Heatmap of all these TRB CDR3aa motif usage in the external thymic pediatric dataset (35, 36) **(C)** and those of TRB CD8 in the peripheral dataset **(D)**. Sex and total CDR3aa number are depicted by sample. Males are depicted in violet and females in orange then local motifs in blue-green and global motifs in magenta.

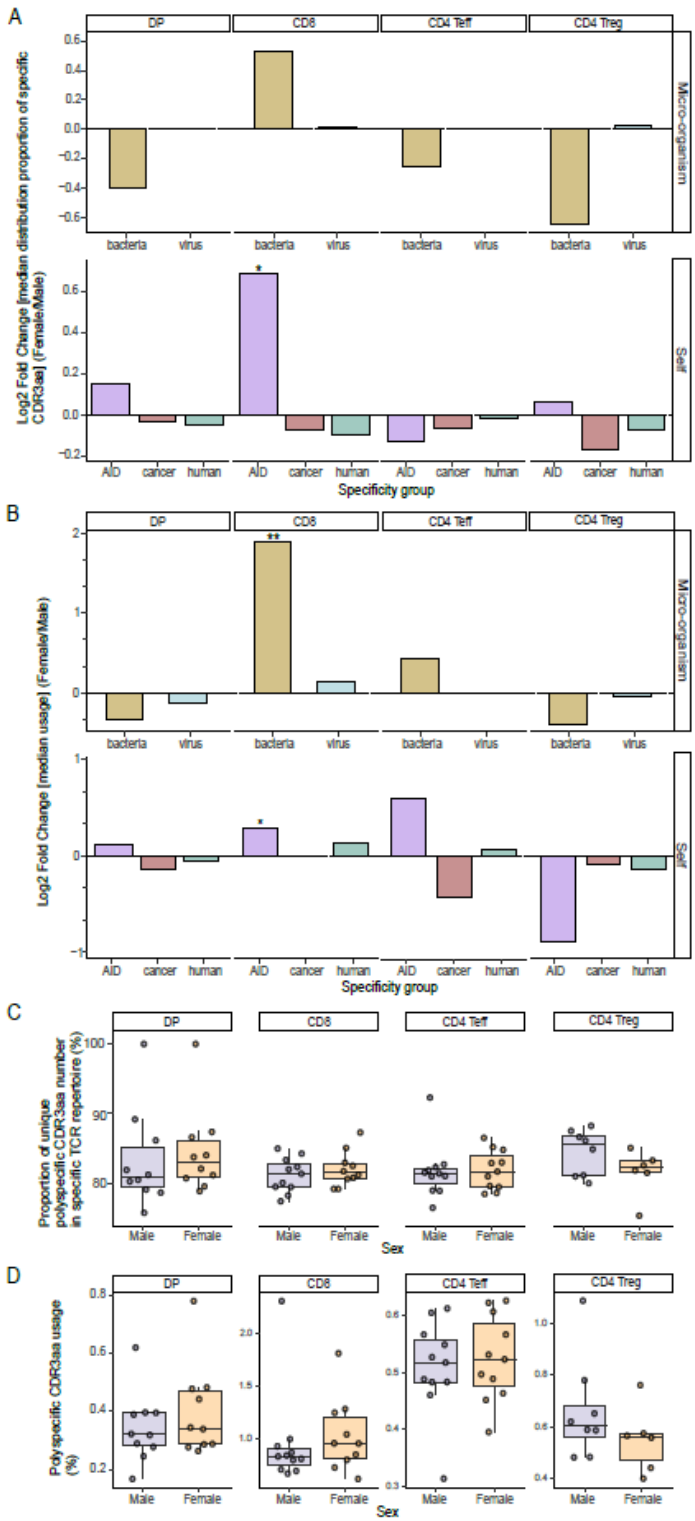

Figure 8: Sex-biased enrichment of TCRs specific for known antigens.
From a pooled and curated database, an exact match with this database infers the specificity of TRB CDR3aa of our thymic dataset. Many specificity groups are defined with the nature of the antigen peptide targeted (bacteria, virus, autoimmune disease [AID], cancer and self-peptide no associated to disease [human]). This analysis compares the distributions of the proportion of unique TRB CDR3aa sequences with a specific specificity **(A)** and their usage **(B)** between females and males across cell subtypes, using the log2 fold change of the median values (females over males), following each specificity group. These groups of specificity are additionally classified as microorganism in top (bacteria in gold and virus in light blue) and self at bottom (AID in magenta, cancer in red, human in blue-green). Polyspecific CDR3aa are defined here as CDR3aa capable of recognizing multiple antigens from different organisms (for no self-antigens) or from different specificity groups (e.g. categorized microorganisms, categorized self-antigens, allergens…). Proportion of unique polyspecific CDR3aa in the specific repertoire **(C)** and their usage **(D)** is compared between males and females, in DP (N = 22), CD8 (N = 23), CD4 Teff (N = 24) and CD4 Treg (N = 14) cells. Stars indicate statistical differences between males and females based on the p-value of the Wilcoxon test (*: $p < 0.05$, **: $p < 0.01$). Males are depicted in violet and females in orange.