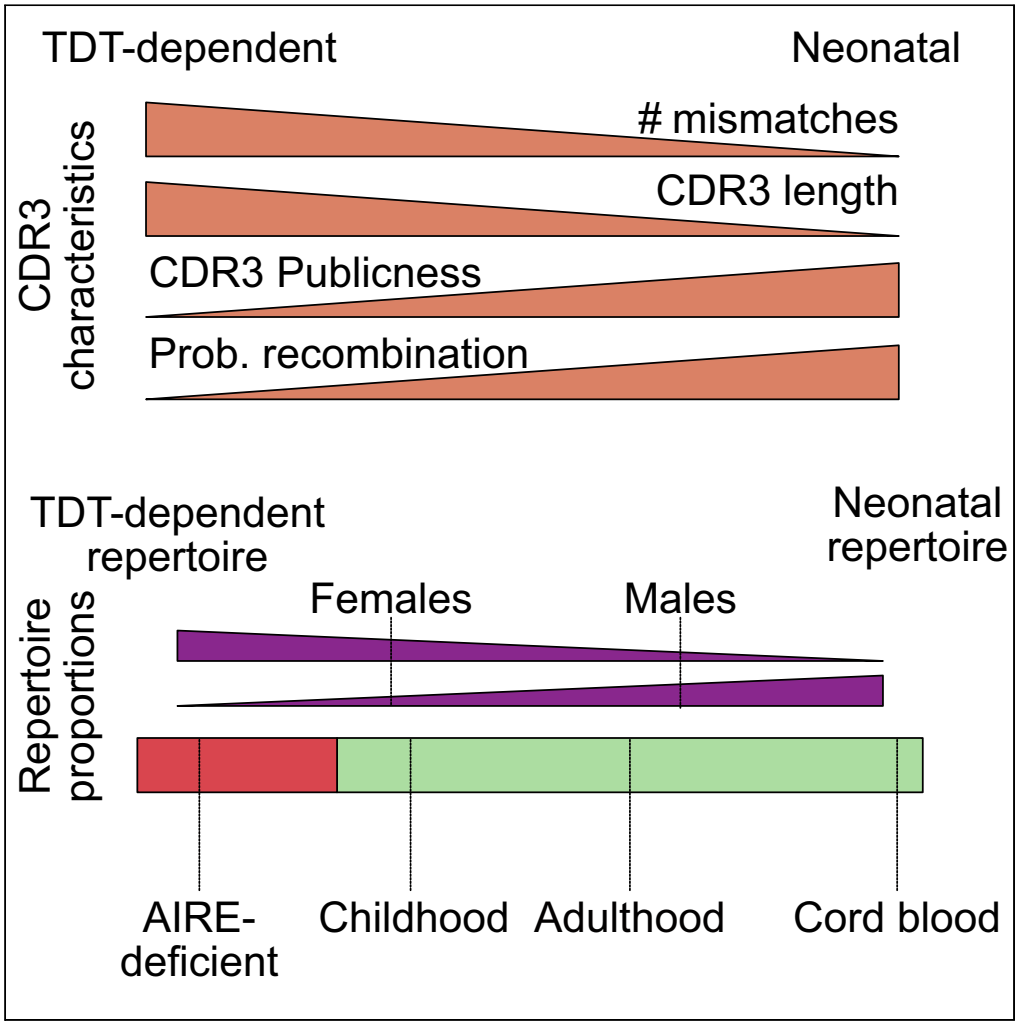
iScience

CellPress
OPEN ACCESS

Article

# Two types of human TCR differentially regulate reactivity to self and non-self antigens




Assya Trofimov, Philippe Brouillard, Jean-David Larouche, ..., Silvy Lachance, Sébastien Lemieux, Claude Perreault

s.lemieux@umontreal.ca (S.L.)
claude.perreault@umontreal.ca (C.P.)

## Highlights

Over $10^8$ TCR CDR3 sequences from ~$10^3$ individuals and 7 cohorts were analyzed

The TCR repertoire is composed of two layers: neonatal and TDT-dependent layer

~70% of frequent cord blood TCRs are associated with common pathogens

Acute graft-vs-host disease correlates with a high proportion of TDT-dependent TCRs

Trofimov et al., iScience 25, 104968
September 16, 2022 
https://doi.org/10.1016/j.isci.2022.104968

iScience

CellPress
OPEN ACCESS

Article

# Two types of human TCR differentially regulate reactivity to self and non-self antigens

Assya Trofimov,[1,2,3,7,8] Philippe Brouillard,[2,3] Jean-David Larouche,[1,4] Jonathan Séguin,[1] Jean-Philippe Laverdure,[1] Ann Brasey,[5] Gregory Ehx,[1,6] Denis-Claude Roy,[5] Lambert Busque,[5] Silvy Lachance,[4,5,10] Sébastien Lemieux,[1,2,9,10,*] and Claude Perreault[1,4,5,10,11,*]

## SUMMARY

**Based on analyses of TCR sequences from over 1,000 individuals, we report that the TCR repertoire is composed of two ontogenically and functionally distinct types of TCRs. Their production is regulated by variations in thymic output and terminal deoxynucleotidyl transferase (TDT) activity. Neonatal TCRs derived from TDT-negative progenitors persist throughout life, are highly shared among subjects, and are reported as disease-associated. Thus, 10%–30% of most frequent cord blood TCRs are associated with common pathogens and autoantigens. TDT-dependent TCRs present distinct structural features and are less shared among subjects. TDT-dependent TCRs are produced in maximal numbers during infancy when thymic output and TDT activity reach a summit, are more abundant in subjects with AIRE mutations, and seem to play a dominant role in graft-versus-host disease. Factors decreasing thymic output (age, male sex) negatively impact TCR diversity. Males compensate for their lower repertoire diversity via hyperexpansion of selected TCR clonotypes.**

## INTRODUCTION

Jawed vertebrates absolutely need a diversified TCR repertoire because classic αβT cells must respond with exquisite specificity to an enormous diversity of ligands (Mittelbrunn and Kroemer, 2021). TCR diversity is generated by somatic recombination of V(D)J gene segments and is further increased postnatally by nucleotide insertion mediated by terminal deoxynucleotidyl transferase (TDT). Notably, neonatal thymocytes, which derive from fetal hematopoietic stem cells, do not express TDT (Rudd, 2020). TDT expression reaches maximal expression in humans between 10 and 40 months (mo) of age, and then decreases progressively during adolescence and adulthood (Deibel et al., 1983; Pahwa et al., 1981). Recent estimates of the potential number of TCRs produced by V(D)J recombination range from $10^{15}$ (Mayer et al., 2019) to $10^{61}$ (deGreef et al., 2020), which vastly outnumbers the number of distinct TCRs present in a human body. Indeed, the adult human body contains approximately $4 \times 10^{11}$T cells (Jenkins et al., 2010) composed of about $10^{10}$ TCR clonotypes of various sizes (deGreef et al., 2020; Lythe et al., 2016). Initially, T cell repertoires have been presumed to be almost entirely private, and the occurrence of the same TCR in two unrelated individuals was attributed to coincidence. However, with the development of high-throughput TCR sequencing and state-of-the-art analytical algorithms, it became clear that interindividual sharing of TCR clonotypes was more common than expected (Pogorelyy et al., 2017; Sethna et al., 2019; Soto et al., 2020). Furthermore, some public clones were found to persist through an individual's life (Chu et al., 2019; Pogorelyy et al., 2017). Still, the extent of interindividual sharing of TCR clonotypes is not precisely known (Johnson et al., 2021).

Which factors influence TCR diversity? At face value, the reduced thymic output associated with aging and male sex (Clave et al., 2018) should impinge on TCR diversity. However, in contrast to mice, humans can compensate for a reduction of thymic output via minimal adjustments in homeostatic T cell proliferation (Goronzy and Weyand, 2019). Hence, the relation between thymic output and TCR diversity may not be linear. Nonetheless, analyses of TCR sequences in large cohorts have revealed a negative impact of aging on TCR diversity, while the effect of sex remains questionable (DeWitt et al., 2018; Krishna et al., 2020). Furthermore, there is an agreement that HLA polymorphism (i.e., heterozygosity for divergent alleles) positively correlates with TCR diversity and that some pathogens (e.g., CMV) can influence the composition of the TCR repertoire (DeWitt et al., 2018; Krishna et al., 2020).

[1]Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montreal, Quebec H3C 3J7, Canada

[2]Department of Computer Science and Research Operations, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

[3]Quebec Institute for Learning Algorithms (Mila), Montreal, Quebec H2S 3H1, Canada

[4]Department of Medicine, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

[5]Maisonneuve-Rosemont Hospital, Montreal, Quebec H1T 2M4, Canada

[6]Currently Interdisciplinary Cluster for Applied Genoproteomics (GIGA-I3), University of Liege, Liege 4000, Belgium

[7]Currently Fred Hutchinson Cancer Center, Seattle, WA 98109, USA

[8]Currently Department of Physics, University of Washington, Seattle, WA 98195-1560, USA

[9]Department of Biochemistry at University of Montreal, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

[10]Senior author

[11]Lead contact

*Correspondence: s.lemieux@umontreal.ca (S.L.), claude.perreault@umontreal.ca (C.P.)

https://doi.org/10.1016/j.isci.2022.104968

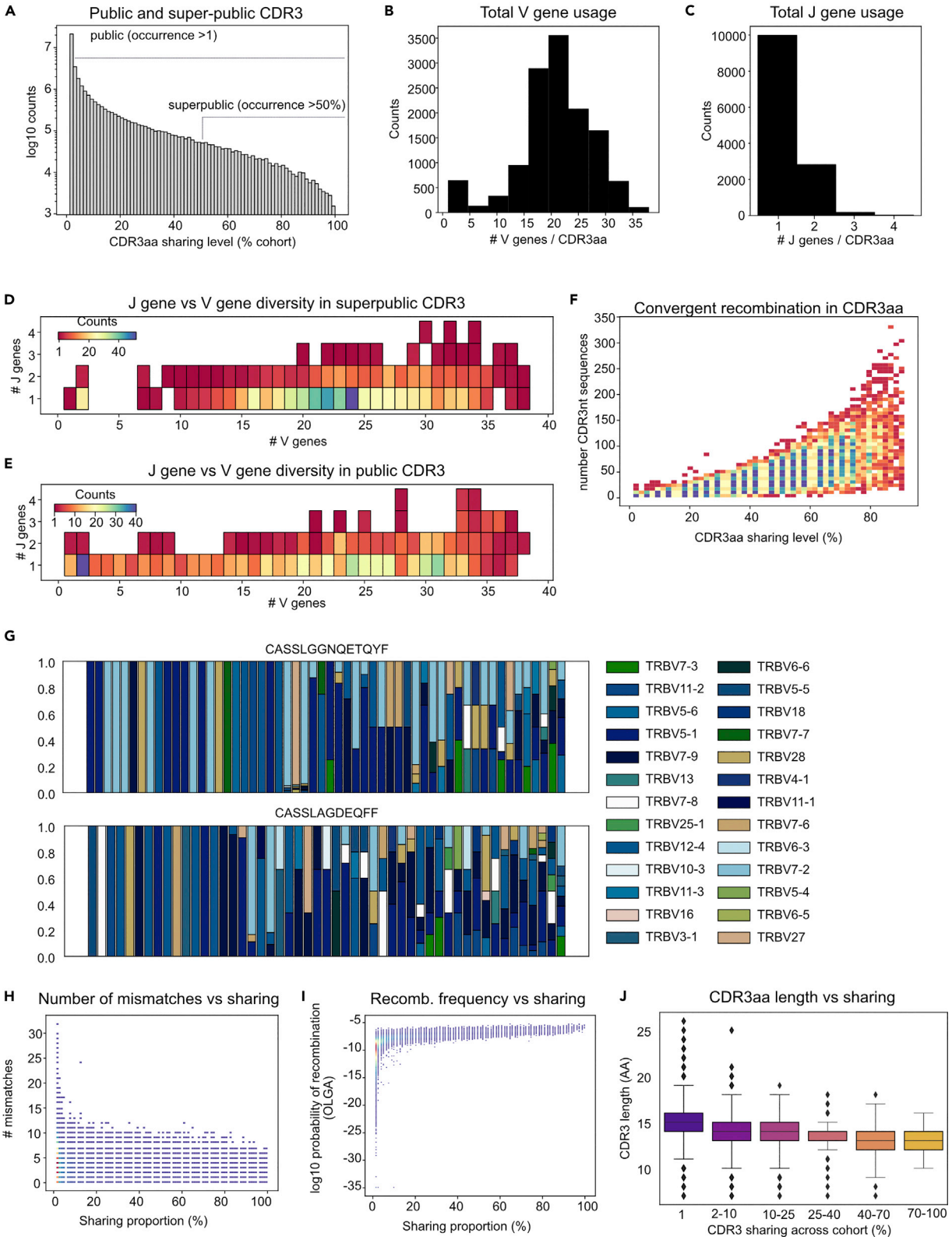

A
Public and super-public CDR3
public (occurrence >1)
superpublic (occurrence >50%)
log10 counts
CDR3aa sharing level (% cohort)
B
Total V gene usage
Counts
# V genes / CDR3aa
C
Total J gene usage
Counts
# J genes / CDR3aa
D
J gene vs V gene diversity in superpublic CDR3
Counts
# J genes
# V genes
E
J gene vs V gene diversity in public CDR3
Counts
# J genes
# V genes
F
Convergent recombination in CDR3aa
number CDR3nt sequences
CDR3aa sharing level (%)
G
CASSLGGNQETQYF
CASSLAGDEQFF
TRBV7-3
TRBV11-2
TRBV5-6
TRBV5-1
TRBV7-9
TRBV13
TRBV7-8
TRBV25-1
TRBV12-4
TRBV10-3
TRBV11-3
TRBV16
TRBV3-1
TRBV6-6
TRBV5-5
TRBV18
TRBV7-7
TRBV28
TRBV4-1
TRBV11-1
TRBV7-6
TRBV6-3
TRBV7-2
TRBV5-4
TRBV6-5
TRBV27
H
Number of mismatches vs sharing
# mismatches
Sharing proportion (%)
I
Recomb. frequency vs sharing
log10 probability of recombination (OLGA)
Sharing proportion (%)
J
CDR3aa length vs sharing
CDR3 length (AA)
1
2-10
10-25
25-40
40-70
70-100
CDR3 sharing across cohort (%)

**Figure 1. The physical characteristics of public CDR3s**

(A–E) Public CDR3s are defined as seen in at least two people in the cohort, while superpublic CDR3s are seen in at least half of the cohort. Number of unique

(B) V and (C) J genes encoding individual public CDR3aa. Relationship between the number of unique V and J genes shown by 2D histogram in

(D) superpublic and (E) public CDR3aa.

(F) Histplot showing convergent recombination of CDR3 nucleotide sequences in public CDR3aa: highly shared CDR3aa are coded by multiple synonymous nucleotides sequences.

(G) Intra-individual V gene usage diversity for two superpublic CDR3aa sequences, sorted by intra-individual entropy: CASSLAGDEQFF and CASSLGGNQETQYF.

(H) CDR3aa sharing and number of mismatches to the annotated germline.

(I) Relation between the predicted recombination frequency and CDR3aa cohort sharing percentage.

(J) CDR3aa length binned by CDR3aa cohort sharing percentage.

In this study, we analyzed the physical properties of CDR3 beta sequences in seven cohorts of individuals and their implication in immune responses against pathogens, autoantigens, and alloantigens. We found stark differences between male and female TCR repertoires, where males maintain a lower diversity but high clonality (i.e., highly abundant TCR clonotypes). In comparison, female repertoires have a higher diversity of CDR3 at low clonal frequencies. A salient finding was the identification of two non-redundant CDR3 repertoire layers based on physical characteristics, including length, number of insertions, and V/J gene usage. The neonatal layer constitutes the entire TCR repertoire of cord blood, while the TDT-dependent layer appears later in life. Unexpectedly, the cord blood TCR repertoire contains mainly public CDR3s that are associated with common pathogens.

## RESULTS

### Physical characteristics of public and superpublic CDR3s

We defined as *public* a CDR3aa (CDR3 amino acid sequence) seen in at least two individuals, while a *superpublic* CDR3aa is present in at least half of the subjects. For our first experiment, we used the Britanova cohort (Figure 1A), consisting of 79 healthy volunteers aged from 0 to 100 (see STAR Methods). We found 2,862,268 public and 15,088 superpublic CDR3aa, of which 21 were ubiquitous (present in all samples) (Figure 1A). To define the physical properties of public and superpublic CDR3aa, we first analyzed their V and J gene usage by grouping the CDR3aa sequences by the annotated V or J gene identity. As expected (DeSimone et al., 2018), while each unique CDR3aa sequence was encoded by mostly 1 or 2 J genes, many V genes can contribute to the same CDR3aa sequence. At the population level, we observed an average of 26 different V genes per public CDR3aa sequence (Figures 1B and 1C). For both public and superpublic CDR3aas, sequences encoded by a higher diversity of J genes were also encoded by numerous V genes (Figures 1D and 1E). In single individuals, up to eight different V genes could contribute to the same CDR3aa (Figure 1G). Finally, as previously reported (Gil et al., 2020; Madi et al., 2014; Venturi et al., 2008), we confirmed a positive correlation between the extent of CDR3aa sharing and the number of different nucleotide sequences encoding each CDR3aa (Figure 1F). This positive correlation points toward a trend of convergent recombination for public and superpublic CDR3aa (Quigley et al., 2010). We then used the software IgBLAST (Ye et al., 2013) to obtain the number of mismatched (i.e., not germline) nucleotides in each CDR3aa sequence in the Britanova cohort (see STAR Methods). We found that sequences shared by more individuals were also sequences with fewer mismatches (Figure 1H). This is consistent with the idea that non-templated nucleotide addition is a random process, and therefore each nucleotide mismatch lowers the likelihood of sequence sharing (Marcou et al., 2018; Sethna et al., 2019). Using the OLGA software (see STAR Methods), we calculated the probability of a CDR3 nucleotide sequence being generated during V(D)J recombination for individual CDR3aa sequences. We confirmed a positive correlation between sequence publicness and recombination probability (Figure 1I). Finally, public sequences were shorter than private ones (Figure 1J), presumably because non-templated nucleotide addition lengthens the sequence (Marcou et al., 2018; Sethna et al., 2019).

### CDR3aa sharing patterns change with age

Analysis of the Britanova cohort revealed a tight correlation between the extent of CDR3aa sharing among subjects and the frequency of the corresponding clonotypes in individual subjects. Superpublic CDR3aa were coded by high-frequency TCR clonotypes, and the most superpublic CDR3aa were found at higher-than-expected cumulative frequencies (Figure 2A). We calculated pairwise repertoire overlap distance between individuals based on the Jaccard index (see STAR Methods). Using this distance, we performed hierarchical clustering (Figures S1A and S1B) and found that individuals clustered by age and

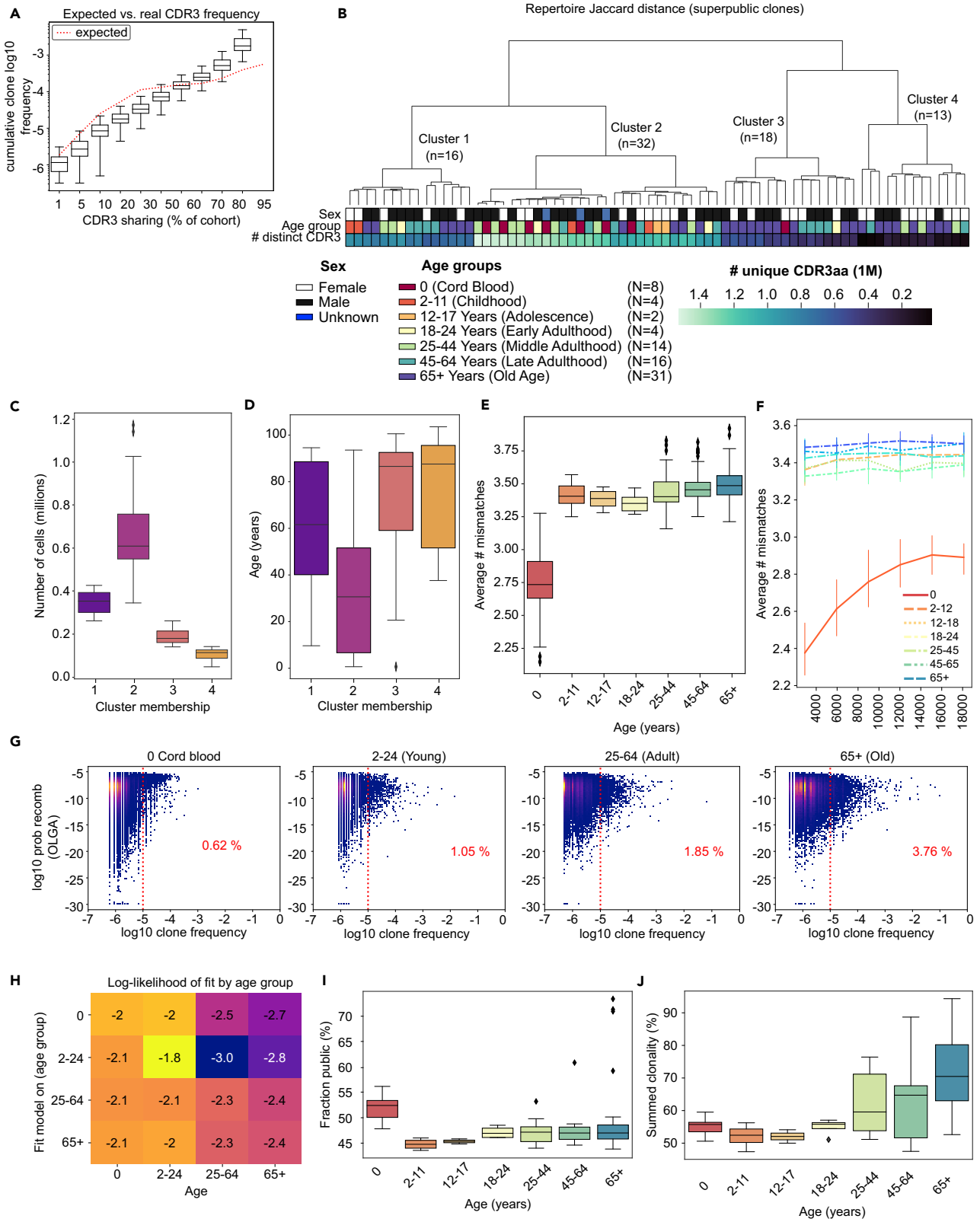



**Figure 2. CDR3 sharing between individuals as a function of age**

(A) Cumulative log10 frequency of CDR3 binned by cohort sharing percentage. The red dotted line indicates the expected frequency given the number of individuals in each sharing bin.

**Figure 2. *Continued***

(B) Hierarchical clustering of individual repertoires based on pairwise Jaccard distance. Dendrogram leaves representing individuals are colored by sex, age group, and the number of distinct CDR3aa found in individual repertoires.

(C and D) Boxplots show the distribution of the number of distinct public CDR3aa found in individuals and the age of individuals in each of the four clusters from Figure 2B.

(E) The median number of mismatches to the germline found in CDR3aa of individuals by age group.

(F) The average number of mismatches to the germline found in CDR3aa of individuals grouped by CDR3aa frequency in each repertoire. Colors represent various age groups.

(G) Aging correlates with an accumulation of high-frequency clonotypes with a high recombination frequency (as determined by OLGA score).

(H) Mean loglikelihood of fit for CDR3aa found in age groups in abscissa for models trained on age groups in ordinate. Each row represents a model trained on the age group, and each column represents the test CDR3aa; each cell contains the mean loglikelihood of fit for a Gaussian mixture model (see STAR Methods).

(I) The proportion of public CDR3aa in different age groups.

(J) Summed clonality of public CDR3aa in various age groups.

See also Figures S1 and S2.

repertoire diversity (Figure 2B), especially when looking at superpublic CDR3-sharing patterns. Indeed, upon splitting the dendrogram of superpublic CDR3s into four clusters (Figure 2B), we found that individuals in the different clusters had different repertoire sizes and age distribution (Figures 2C and 2D). Clusters #2 and #4 showed maximum divergence: individuals in cluster #2 had an average of $0.6 \times 10^6$ different CDR3 sequences and a mean age of 40, against only $0.1 \times 10^6$ sequences and a mean age of 93 in cluster #4 (Figures 2C and 2D). We wondered whether variations in TDT activity with age (Bonati et al., 1992; Deibel et al., 1983; Pahwa et al., 1981) could impact on repertoire sharing. When we aligned each CDR3 to the germline from the reference genome and counted the number of mismatches (see STAR Methods), we found that, indeed, cord blood CDR3s (TDT-negative) contained fewer mismatches than samples from other age groups (Figure 2E). Finally, when we grouped CDR3aa by descending order of frequency (see STAR Methods), we found that the most frequent CDR3aa displayed fewer mismatches than those with lower frequency, most distinctively in cord blood (Figure 2F).

Further clone size analyses showed that as individuals age, they accumulate in their repertoires more very high-frequency CDR3aa (at frequencies above 0.0001 of total repertoire), which have a recombination frequency in the higher ranges (above $10^{-10}$) (Figure 2G). Moreover, a low clonal frequency for high recombination frequency sequences could be partly due to undersampling in the repertoire (Sethna et al., 2019) since only a certain amount of CDR3s are sequences. We used a two-step strategy to evaluate the relationship between clonality and the probability of recombination at different ages. We fitted a Gaussian mixture model for each age group, and then we calculated the loglikelihood of data from other age groups under this model (see STAR Methods). We found that models fitted on repertoires of younger individuals did not fit with data from older individuals. However, since models fitted on older individuals had a similar likelihood for all age groups, we concluded that older repertoires retain characteristics of younger repertoires and outgrow them with time (Figure 2H). What distinguishes older repertoires from younger ones is a large quantity of high-frequency (presumably hyperexpanded) CDR3aa with a high recombination probability (Figures 2G and 2H).

For individual samples in the Britanova cohort, the proportion of public CDR3aa was maximum in cord blood, dropped abruptly in children, and increased progressively with age after that (Figure 2I). As a result, the proportion of public CDR3aa in subjects $\geq$65 years of age was similar to that in cord blood. The progressive increase in the fraction of public CDR3aa from childhood to old age was even more conspicuous when considering the clonality of each CDR3aa (see STAR Methods, the section on CDR3 sharing): almost 70% of repertoires in individuals $\geq$65 years of age were composed of public CDR3aa (Figure 2J). Though cord blood and samples from subjects $\geq$65 years of age contained similar proportions of public CDR3aa (Figure 2I), their clonality was very different (Figure 2J). Cord blood cells had a more uniformly distributed repertoire of public CDR3aa, without the hyperexpanded clones present in subjects $\geq$65 (Figures 2I and 2J). We validated our observation in two additional cohorts. In the Emerson cohort, containing TCR-Seq data from 666 healthy individuals (Emerson et al., 2017), we could split individuals by CMV status. We found that an age-related skew in public fraction can be observed in CMV+ and CMV– subjects (Figures S2A–S2D). The Thome cohort is smaller but contains TCR-Seq data from deceased donors' spleen and lymph nodes rather than blood (Thome et al., 2016). T cells were sorted by naive or effector memory phenotype in this study; we, therefore, analyzed those categories separately. We found the same trend of sharing by age group for the naive T cells (Figures S2E and S2F) but not for the effector memory T (TEM) cells in

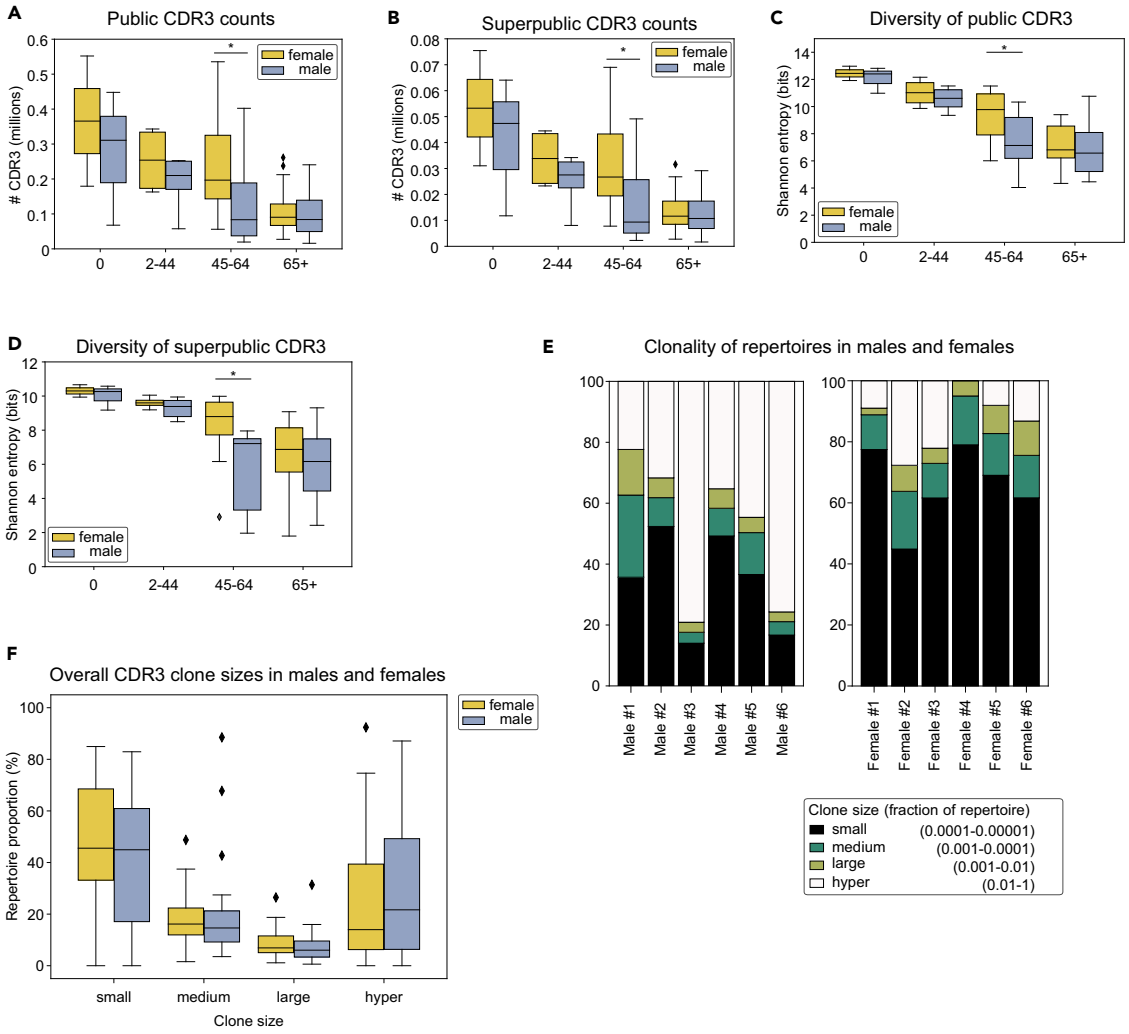



**Figure 3. CDR3 sharing between individuals as a function of sex**

(A–D) Absolute numbers of (A) public and (B) superpublic CDR3aa in male and female individuals by broad age groups. Difference statistically significant ($p < 0.05$) for young individuals (ages 2–44). Shannon entropy for (C) public and (D) superpublic CDR3aa in males and females. Statistically significant differences ($p < 0.05$, Mann-Whitney-Wilcoxon) for subjects aged 2–44 and 45–65.

(E) Clonality of CDR3aa in repertoires of adult males and females, binned by clone size.

(F) Boxplot showing the distribution of overall clone sizes of CDR3aa in males and females of all age groups.

secondary lymphoid organs (Figures S2G and S2H). The latter divergence warrants further investigation but must be considered preliminary because it is based on analyses of a small cohort of deceased donors.

These results indicate that as individuals age, their repertoire becomes preferentially populated by clones with high recombination frequencies. A high recombination frequency is likely instrumental in the abundance of highly public clones. Another possible explanation could be a preferential expansion of these T cells due to homeostatic proliferation (Murray et al., 2003) or immune activation.

## The impact of sex on the TCR repertoire

Aside from age, male sex is the factor with the most negative impact on thymic output (Clave et al., 2018). Therefore, we analyzed the potential influence of sex on CDR3 repertoire diversity and publicness by grouping individuals into broader age groups to maintain adequate comparison numbers between categories (Figure S3A). Overall, we found that males had fewer CDR3aa in their repertoires than females: this was the case for public (Figure 3A) and superpublic CDR3aa (Figure 3B). When we examined repertoire

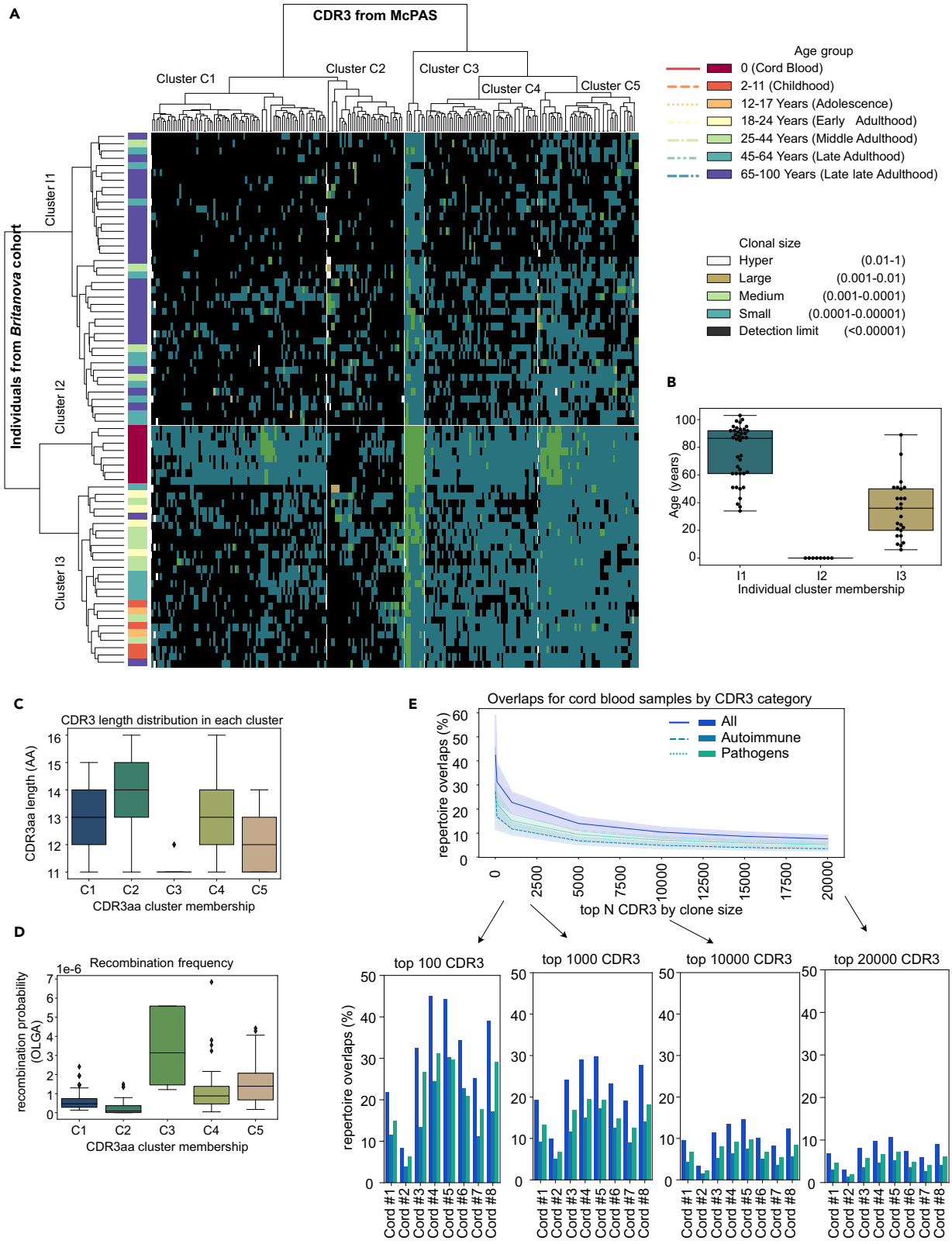



**Figure 4. Cord blood samples contain pathology-annotated CDR3s**

(A) Heatmap shows, for subjects of the Britanova cohort, the frequency of CDR3aa listed in the McPAS microbial pathogens dataset (Tickotsky et al., 2017). Rows represent individuals, columns unique CDR3aa, and cell color indicates CDR3aa clone size. Row dendrogram leaves are colored by age group.

**Figure 4. *Continued***
(B) Age distribution for individuals in three individual (Y axis) clusters from (A).
(C) Boxplot showing CDR3 lengths for CDR3 in the five X axis clusters from (A).
(D) Boxplot showing predicted recombination frequency for CDR3 in five X axis clusters from (B).
(E) Line and barplots show the percentage CDR3aa associated with autoantigens (Tickotsky et al., 2017), or pathogens (Tickotsky et al., 2017) in individual cord blood samples from the Britanova cohort, by varying top N most frequent CDR3aa.
See also Figure S4 and S5.

diversity using Shannon entropy, we found that repertoires of females were more diverse than those of males (Figures 3C and 3D). Accordingly, small-size CDR3aa clonotypes represented 70% of the repertoire in females and 50% in males (Figures 3G and 3H). In contrast, hyperexpanded CDR3 clonotypes constituted 30% of repertoire in males and 10% in females. Differences between males and females were present in all age groups and always reached statistical significance in subjects aged 2–45 but not in other groups (Figures 3A–3D). These results highlight a prominent sexual dimorphism in the TCR repertoire and suggest that it results from differences in thymic output. Female repertoires are more diverse, and males present a lower measured repertoire diversity with hyperexpansion of selected TCR clonotypes.

## Sharing of disease-specific CDR3s in different age groups

Next, we downloaded and explored the McPAS database, a manually curated catalog of pathology-associated TCR sequences (Tickotsky et al., 2017) to assess sharing in the context of pathology-specific TCR sequences. We found minimal overlap (0.1%–3%) between TCRs in two McPAS categories: microbial pathogens and autoimmune diseases (Figure S4A). To gain further insight into disease-related CDR3s, we took CDR3aa listed in the McPAS microbial pathogens dataset and analyzed their frequency in subjects from the Britanova cohort (Figure 4A). The hierarchical clustering dendrogram was separated into three clusters for individuals (I1 to I3) and five clusters for CDR3aa (C1 to C5). Age had a dramatic influence on both dimensions of this orthogonal clustering. Among clusters for individuals, cluster I2 was composed solely of cord blood samples, whereas individuals in clusters I1 and I3 had a mean age of 82 and 26 years of age, respectively (Figure 4B). The CDR3aa-based clustering adopted the following pattern: i) CDR3aa in cluster C1 were present almost exclusively in cord blood, ii) those in cluster C2 were present in few individuals without any clear pattern, and iii) CDR3aa in clusters C4 and C5 were present in young individuals (cord blood and <45 years.o.) (Figure 4A). Cluster C3 was remarkable in that it contained the most highly shared CDR3aa; they were found at high frequency in cord blood and lower frequency in almost all other individuals. CDR3aa in cluster C3 were shorter and displayed a greater recombination frequency than CDR3aa in the four other clusters (Figures 4C and 4D). Observations on microbial pathogens-related CDR3aas were replicated in autoimmune disease-associated CDR3aa (Figures S4B–S4E). First, cord blood (cluster I1 in Figure S4B) contained more autoimmunity-associated CDR3aa. Second, the most highly shared CDR3aa (cluster C1 in Figure S4B) were shorter and displayed a greater recombination frequency than CDR3aa in the four other clusters.

The key finding was that a large portion of the most frequent CDR3aas found in cord blood was disease-related CDR3aas. To evaluate this, for each top N CDR3aa by frequency, we calculated the percentage of those CDR3aas that matched disease-related CDR3aas in both McPAS CDR3aa sets. Indeed, 10% to 30% of most frequent CDR3aa in individual cord blood samples (Britanova cohort) were associated with known pathogens and autoantigens (Figure 4E). Since the Britanova cohort only contains CDR3 beta sequences, without CDR3 alpha or HLA, this result is likely an overestimation but can still be used to compare between age groups and individuals. In some cord blood samples, the summed proportions of CDR3aa associated with autoantigens and pathogens were superior to 50% (Figure 4E). We conclude that all individuals have many disease-reacting clones at a high frequency in their repertoires before birth. Are disease-related CDR3s present in older subjects? To address this question, we first filtered out low-confidence clonotypes (Figure S5A), i.e. clonotypes with a frequency below the detection limit (0.00001 of repertoire), as done for Figure 4A. Then, we compared percentages of overlaps before and after the frequency filter (Figure S5B). We found that only cord blood repertoires contained a high percentage of pathogen- and autoantigen-specific CDR3aas present at high frequencies, compared to other age groups (Figure S5B). We then sorted the clonotypes by decreasing frequency and found that cord blood but not repertoires of other age groups contained pathogen and autoantigen-associated CDR3aas among their most frequent clonotypes (Figures S5C and S5D). We validated that this was specific to pathogen and autoantigen-associated CDR3 by repeating the same comparison for a randomly sampled set of unrelated CDR3aas (Figure S5E). Two points can be made from this analysis. First, in cord blood, disease-related CDR3aa are enriched in

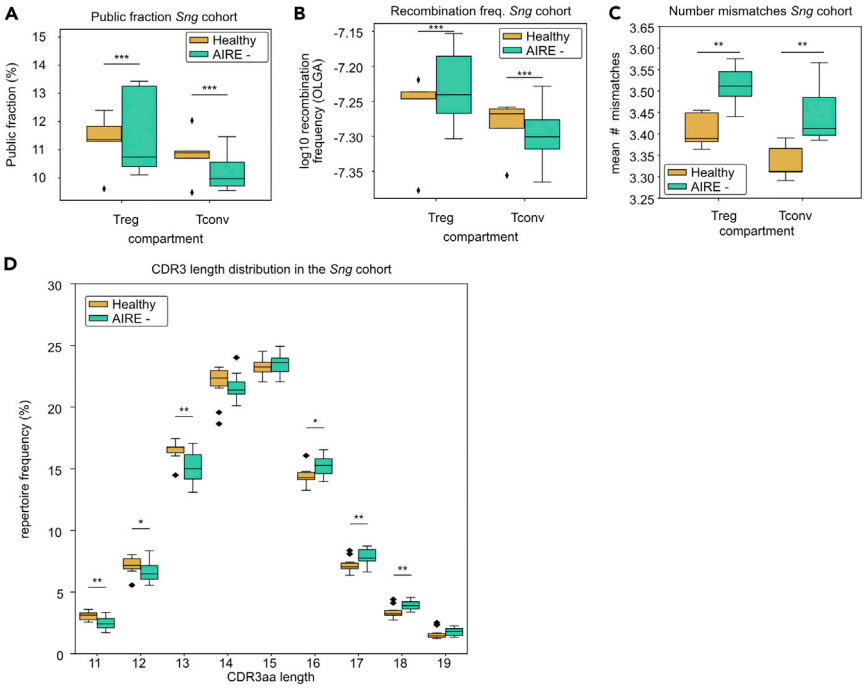



**Figure 5. CDR3aa profile in subjects with AIRE mutations**

(A–C) Boxplots showing, for the regulatory (Treg) and conventionalT cell (Tconv) compartments, (A) the public fraction, (B) the recombination frequency, and (C) the number of mismatches in subjects with AIRE mutations vs. controls.

(D) Boxplots show the CDR3aa length distribution in the Sng cohort. (*p < 0.05, **p < 0.01, ***p < 0.001, Mann-Whitney-Wilcoxon).

high-frequency clonotypes. Second, the remarkable representation of disease-related CDR3aa in the "pre-immune" repertoire of cord blood is lost in older individuals.

Our data support the notion that TCRs generated during fetal life can persist (or be continuously generated) for decades in adults (Pogorelyy et al., 2017). More importantly, they show that most of these TCRs participate in a wide variety of immune responses in adult life. Globally, our data presented so far suggest the existence of two types of CDR3: the superpublic ones, shared by many individuals and present before birth, and the private repertoire, dependent on TDT modifications. For the remainder of the study, we will refer to these two types of TCRs as *neonatal* and *TDT-dependent*.

## Negative selection targets TDT-dependent TCRs

Irrespective of their TCR type, neonatal or TDT-dependent, T cells are subjected to intrathymic positive and negative selection. Mutations in the AIRE protein are known to be associated with perturbations in negative selection and thereby causing autoimmunity (Liston et al., 2003). We, therefore, analyzed the CDR3s of the Sng cohort, which contains subjects with AIRE mutations and healthy controls (Sng et al., 2019). We found that CDR3aa repertoires of AIRE-mutated individuals had a lower public fraction than healthy repertoires (Figure 5A), with a lower recombination frequency (Figure 5B), a higher number of mismatches per CDR3aa (Figure 5C) for both regulatory and conventional T cell compartments, and longer CDR3aas (Figure 5D). These results point toward enrichment in TDT-dependent CDR3s in repertoires of AIRE-mutated individuals, which in turn suggests that thymocytes with TDT-dependent TCRs are prime subjects of negative selection.

## Effect of the TCR repertoire on graft-versus-host disease

To further evaluate the potential impact of the two types of TCRs, we reasoned that the best strategy would be to use a model in which the readout depends exclusively on T cells. Acute graft-versus-host disease (aGVHD) following allogeneic hematopoietic cell transplantation (AHCT) represents such a model. Indeed, donor T cells, particularly the $CD4^+$ subset, are necessary and sufficient for the occurrence of aGVHD (Ni

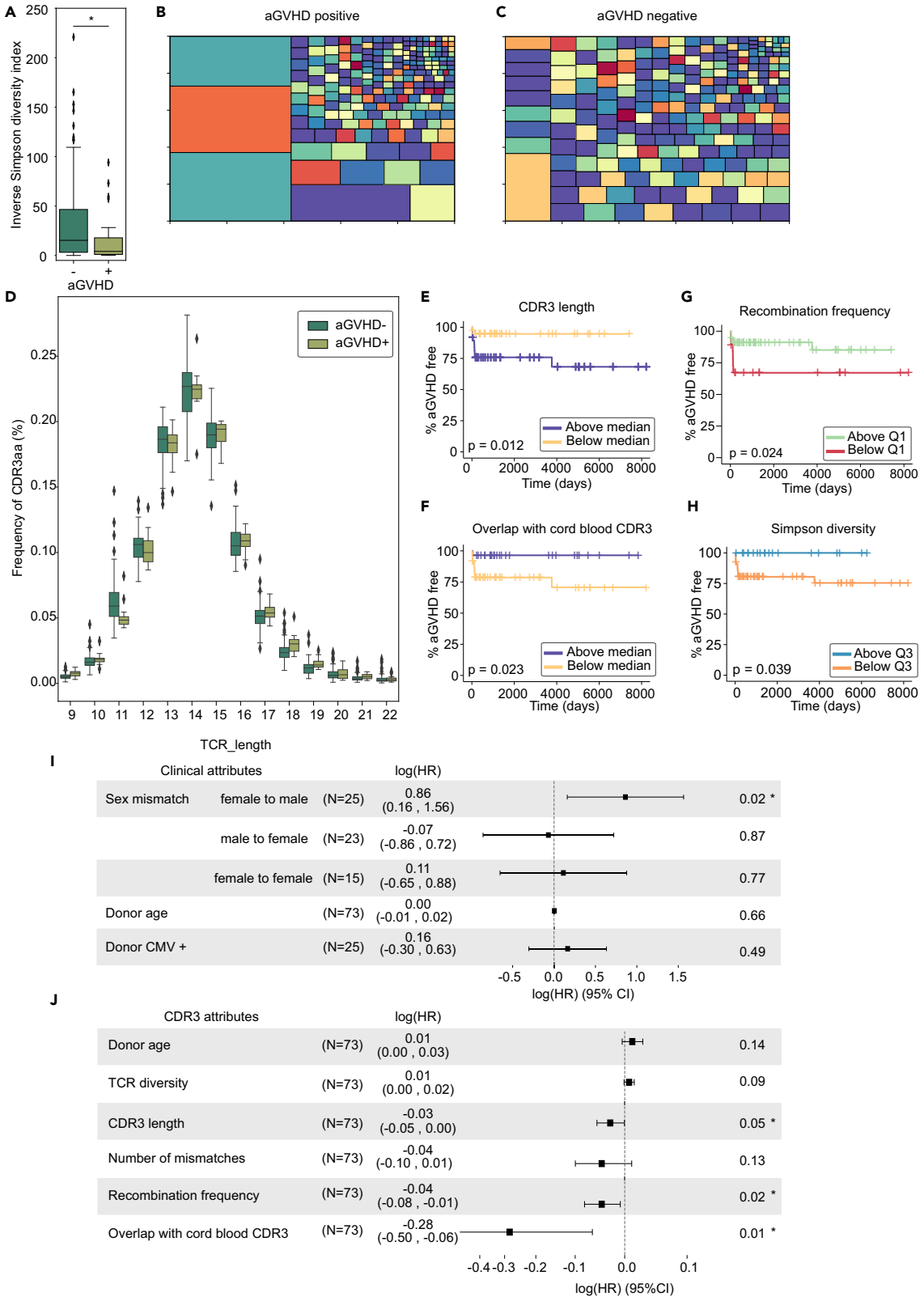



**Figure 6. CDR3aa in CD4 T cells from aGVHD+ and aGVHD- AHCT donors**

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.isci.2022.104968.

**Figure 6. *Continued***

(A–C) Inverse Simpson diversity index of CDR3 repertoires found in aGVHD+ and aGVHD-grafts. Treemaps showing CDR3aa diversity and clone sizes for two representative donors (B) aGVHD+ and (C) aGVHD-, colors selected at random for better visual distinction.

(D–J) CDR3aa length distribution in aGVHD+ and aGVHD-donors. Kaplan-Meier curves representing aGVHD onset for grafts split into two groups by median or quantiles according to (E) CDR3 length, (F) overlap with cord blood CDR3aa, (G) recombination frequency, and (H) Simpson diversity index. CoxPH models calculating hazard ratios for (I) clinical characteristics and (J) CDR3 repertoire of the donor. (*p < 0.05, **p < 0.01, ***p < 0.001).

et al., 2017; Socié and Blazar, 2009). They initiate aGVHD via recognition of host alloantigens (Martin et al., 2017; Vincent et al., 2011). Therefore, we analyzed TCRs in purified $CD4^{+}$T cells from 73 AHCT donors. Donors and recipients were HLA-matched siblings. The T cells were obtained from the peripheral blood of donors on the day of transplantation and submitted to RNA sequencing. To extract CDR3 sequences from RNA sequencing reads, we used the MIXCR software (Bolotin et al., 2015; Li et al., 2017). We classified donors as aGVHD + or aGVHD–, depending on whether their recipient presented or not severe aGVHD (see STAR Methods). Notably, aGVHD + donors had lower CDR3 diversity than aGVHD– grafts (Figure 6A). We used a treemap to display both diversity and clone size in two representative donors. Treemaps offer a visual representation of diversity at a glance, and we used these plots to compare two representative examples of aGVHD– and aGVHD + donor repertoires. In the aGVHD + donor, three hyperexpanded clones occupied almost $^{1}/_{3}$ of the repertoire (Figure 6B), while the aGVHD– donor did not have this skew (Figure 6C). The CDR3aa in aGVHD + grafts were longer (Figure 6D), had a lower recombination frequency, and more numerous mismatches than CDR3aa in aGVHD donors (Figures S5F and S5G). We then split the cohort by the median or quartiles and generated Kaplan-Meier curves to assess the impact of CDR3 features on the occurrence of aGVHD (Figures 6E–6H and S6). Overall, grafts containing a higher proportion of CDR3 with neonatal features caused less aGVHD. These features were: CDR3 length in amino acids (Figure 6E), percentage overlap with cord blood samples (Figure 6F), recombination frequency (Figure 6G), and Simpson diversity index (Figure 6H).

Finally, we used Cox proportional hazards (CoxPH) models to evaluate more accurately the impact of clinical and CDR3 features on the risk of aGVHD. For the clinical characteristics model, the sole significant correlation was a higher rate of aGVHD in male recipients of female grafts (Figure 6I). These results are concordant with previous reports (Kim et al., 2016). For the CDR3 model, we found that a shorter CDR3 length, a high number of neonatal CDR3, and a high average CDR3 recombination frequency decreased the risk of aGVHD (Figure 6J). Other characteristics and clinical traits such as donor age and CMV status had no significant impact (Figures 6I and 6J). Collectively, these results strongly suggest that donors with a higher proportion of neonatal TCRs cause less aGVHD and that aGVHD is initiated primarily by TDT-dependent TCRs.

## A stratified model of the TCR repertoire

Our final goal was to evaluate the importance of discrete features in defining neonatal and TDT-dependent TCRs. Our reasoning was based on two assumptions. First, we assumed that cord blood samples contained exclusively neonatal TCRs while all other age groups contained a mix of neonatal and TDT-dependent TCRs. Second, since thymic output and TDT activity reach their zenith during childhood, we postulated that children would generate the greatest diversity of TDT-dependent TCRs. Therefore, to get a pure and diversified population of TDT-dependent CDR3s, we selected CDR3s present in children but not in cord blood. We then confirmed that, compared to neonatal CDR3s, the TDT-dependent CDR3s were longer (Figure 7A), had more mismatches, and a lower recombination probability (Figures 7B and 7C). Notably, they also displayed a different V and J gene usage (Figures 7D and 7E).

On this dataset, we trained a logistic regression model and random forest to verify if the nonlinearity of the model could have an impact on the performance. Using all the five features (recombination frequency, # mismatches, CDR3 length, V gene, and J gene), we performed an ablation study by obtaining all possible combinations of presence/absence, totaling 31 combinations of features (Figure 7F). We trained the two models on the dataset for each combination and evaluated their performance on a held-out CDR3 repertoire of each type (the entire individual's repertoire). The performance of each model on the held-out data is represented as a single column, where black squares symbolize the absence and white squares the presence of a feature, and the performance squares are colored by the percentage of accuracy of classification (Figure 7F). The CDR3 length was crucial to the model; without the CDR3 length, the model's performance was close to the baseline of 60%, which is the proportion of neonatal CDR3s in the dataset. Adding the length improves classification accuracy by about 10% for all conditions. Numbers of mismatches and V/J

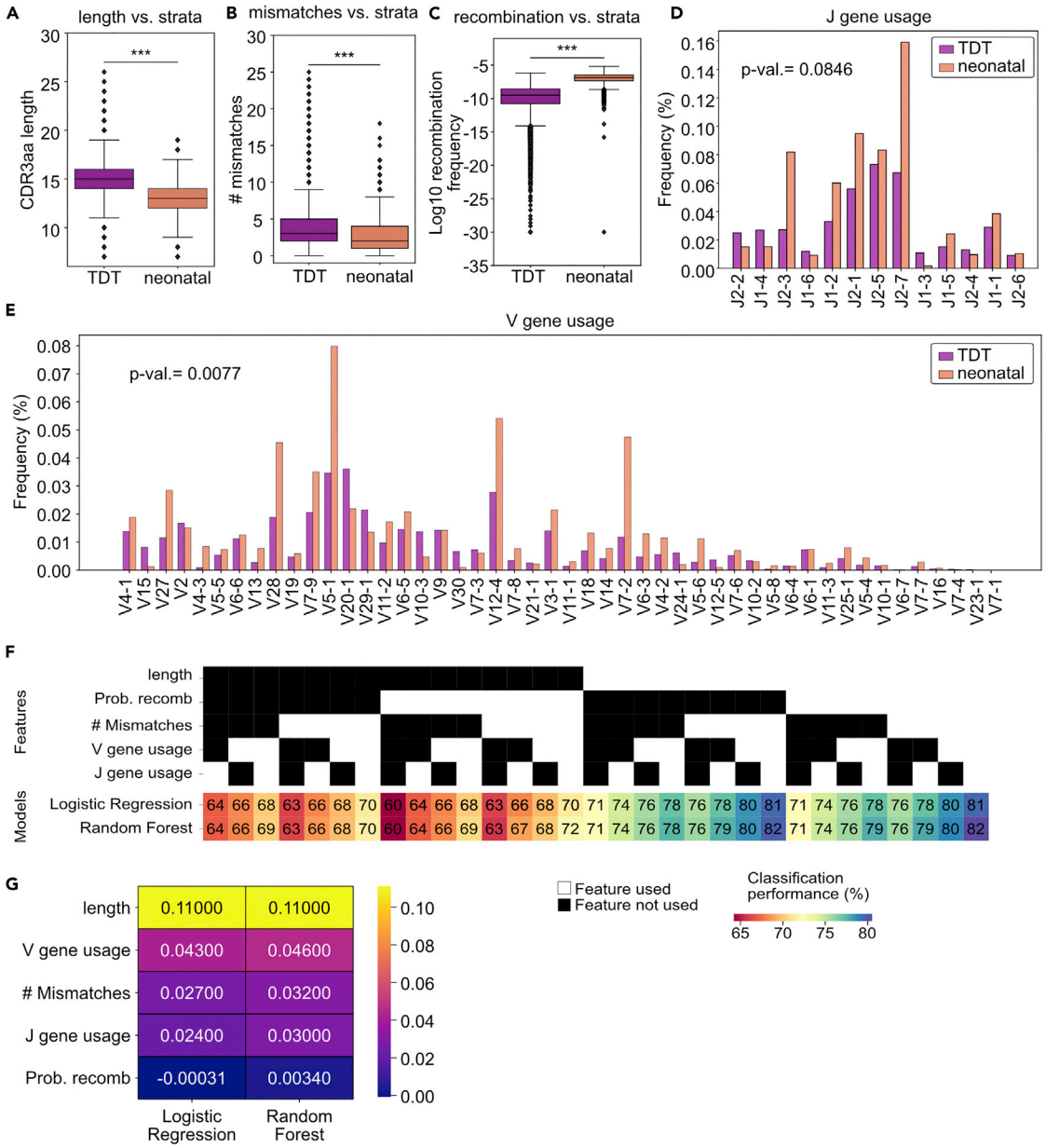



**Figure 7. Features of neonatal vs. TDT-dependent TCRs**

(A–C) Boxplots depict (A) the CDR3aa length, (B) the median number of mismatches to the germline, and (C) the median log10 recombination frequency of the TDT-dependent and neonatal strata.

(D and E) J gene and (E) V gene usage frequencies for CDR3aa in TDT-dependent and neonatal strata.

(F) Feature ablation study showing classification accuracy on held-out data for each feature combination. Black/white squares signal exclusion/inclusion of features in the dataset, and the color scheme shows classification performance.

(G) Coefficients of the linear model fitted on feature ablation study (see STAR Methods). (*p < 0.05, **p < 0.01, ***p < 0.001, Mann-Whitney-Wilcoxon).

gene usage had a more modest effect on the performance, with an accuracy gain of about 5% each. Moreover, V and J gene usage was non-redundant and having both yielded better performance than only having one or the other for both models. The inclusion of the recombination frequency did not impact the performance, most likely because it is largely redundant with CDR3 length (Figure S7A).

Finally, to validate the order of importance of the features, we fit a linear regression model on the presence/absence of features (see STAR Methods). This allowed for the direct comparison of the relative importance

of the features based on the coefficients assigned to each feature (Figure 7G). We found the following importance hierarchy: length > (# mismatches, V/J gene usage) > probability of recombination (Figure 7G). Consistency between the logistic regression and the random forest models suggests that these features robustly discriminate between neonatal and TDT-dependent TCRs. We then used the trained model to classify each CDR3 in the cohort and found that, as expected, there is a considerable dip in the proportion of neonatal TCRs after birth (Figure S7B). Then, from infancy to adulthood, there is a progressive increase in the proportion of neonatal clonotypes (Figure S7B), possibly because of their reactivity to common pathogens (Figure 4). Afterward, the proportion of neonatal CDR3 remains relatively stable with a slight trend downward with advancing age (Figure S7B).

## DISCUSSION

This report analyzed the amino acid sequence of over 100 million TCR CDR3 beta chains from over 1,000 subjects. Since both TCR chains contribute to antigen specificity, we would have wished to extend our analyses to CDR3aa alpha/beta pairs obtained through single-cell studies (Pauken et al., 2022). However, there is no available dataset that contains paired CDR3 alpha and beta chains from large human cohorts. Indeed, because of methodological constraints inherent to single-cell approaches, paired CDR3 alpha and beta chain sequencing has been limited to small numbers of subjects, typically 1 to 15 (Fischer et al., 2020; Pauken et al., 2022; Tanno et al., 2020). Focusing on CDR3 amino acid sequences instead of nucleotides and V/J gene usage allowed us to uncover more numerous public and superpublic CDR3aa than anticipated, since this analysis considers synonymous codons. In a way, our analysis examines the functional features of TCRs, which are determined by their amino acid sequence. Of note, some unconventional T cell subsets such as NK T, MAIT, and CD1-restricted cells have invariant TCRs (Godfrey et al., 2010, 2019), which are probably included in the public and superpublic CDR3aa subsets. However, TCRs from $CD4^+$ and $CD8^+$T cells certainly comprise the bulk of our analyzed TCRs. Indeed, unconventional T cell subsets typically have very low frequencies in peripheral blood (e.g., 0.01%–1.18% PBMC for NK T cells, compared to 26%–48% for $CD4^+$T cells (Autissier et al., 2010; Bernin et al., 2016)).

We found stark differences between male and female repertoires, as well as age-specific and disease-specific repertoire features. Age and sex are associated with important differences in immune responses to pathogens and self-antigens (Brodin and Davis, 2017; Liston et al., 2016). Thymic involution is instrumental in decreasing immunocompetence with age and represents a major public health issue, as illustrated by the COVID pandemic (Mittelbrunn and Kroemer, 2021; Palmer et al., 2018; Yousefzadeh et al., 2021). Aside from age, male sex is the factor with the most negative impact on thymic output (Palmer et al., 2018). We report that both aging and male sex are associated with decreased TCR diversity and hyperexpansion of public clonotypes. Female TCR repertoires are more diverse, and males compensate for their lower repertoire diversity via hyperexpansion of selected TCR clonotypes. These data argue for a strong mechanistic link between thymic output and TCR diversity.

Analyses of cord blood samples were particularly instructive. In the absence of TDT, TCRs produced before birth have short CDR3s, few mismatches (relative to germline sequences), and a biased V/J gene usage. These neonatal TCRs persist (or are continuously replenished) throughout life, are highly shared among subjects, and are likely polyreactive to self and microbial HLA-associated peptides. Three factors likely contribute to the large clone size and extensive sharing of neonatal TCRs over a lifetime. First, they have a high recombination frequency; in other words, they are easy to assemble during V(D)J recombination. Second, their reactivity to self-antigens should theoretically favor their positive selection in the thymus and their homeostatic proliferation in the periphery (Ernst et al., 1999; Hogquist and Jameson, 2014). Third, our analyses of subjects with AIRE mutations revealed that neonatal TCRs were less affected by negative selection in the thymus than TDT-dependent TCRs. Thus, neonatal TCRs may integrate all the"Goldilocks" conditions for intrathymic selection and survival in the periphery. Notably, polyreactivity to self-antigens could also favor the commitment of thymocytes bearing neonatal TCRs toward either the regulatory or alternative T cell lineages (Sood et al., 2021; Vrisekoop et al., 2014). This possibility should be explored in future studies.

While both TCR chains as well as the MHC molecule contribute to antigen specificity, in practice, most analyses of the T cell repertoire have focused on CDR3 beta, mainly for two reasons. Firstly, the sequencing of CDR3 beta is more robust than that of CDR3 alpha (Barennes et al., 2020). Secondly, CDR3 beta is the main contributor to TCR antigen specificity (Springer et al., 2020,2021). Accordingly, though the prediction of

antigen specificity is improved by paired CDR3 alpha/beta sequencing, predictions based only on CDR3 beta perform well (Fischer et al., 2020).Furthermore, analyses of CDR3 alpha and beta chain pairing in close to 1 million clonotypes from 15 individuals (Tanno et al., 2020) support our main conclusions: shared CDR3aa are relatively short with few TDT-dependent additions (Tanno et al., 2020).

The effect of the HLA genotype on the repertoire of CDR3 beta sequences is detectable (Khosravi-Maharlooei et al., 2019; Tanno et al., 2020) but remains relatively small (Emerson et al., 2017; Heikkilä et al., 2021; Pogorelyy et al., 2018; Springer et al., 2021). This is explained at least in part by the fact that most MHC-associated peptides can bind to multiple HLA alleles. Consistent with our results, studies in mice revealed that the most highly shared TCRs among mice with different MHC genotypes have shorter CDR3 sequences (Lu et al., 2019). In contrast, a single TCR has been shown to be capable to bind with as many as million different antigens (Bentzen et al., 2018; Natarajan and Krogsgaard, 2018; Wooldridge et al., 2012; Zhang et al., 2018). Therefore, while our analysis only includes beta chains and therefore overestimates polyreactivity, we found it remarkable that 10%–30% of most frequent CDR3s in cord blood samples were associated with known pathogens or autoantigens (Figure 4E). This means that humans are born with a TCR repertoire that can have a lifelong influence on their response to pathogens and the risk of autoimmunity. From an evolutionary perspective, the size of human populations has been limited by the rate of infant mortality. Hence, it would seem convenient to be born with a polyreactive T cell repertoire responsive to common pathogens.

In contrast to neonatal TCRs, TDT-dependent TCRs are longer, less shared, contain more mismatches, and display a different V/J gene profile. Their production is maximal during infancy, when thymic output and TDT activity reach a summit, and slowly decreases after that. We found that TDT-dependent TCRs were more abundant in subjects with AIRE mutations. This suggests that negative selection preferentially eliminates TDT-dependent TCRs. The ultimate role of TDT remains unclear. By ultimate role, we mean the evolutionary selected biological advantage conferred by TDT. In mice, deletion of TDT does not increase susceptibility to pathogens or the incidence of autoimmunity but decreases the breadth of antiviral responses (Haeryfar et al., 2008; Kedzierska et al., 2008). However, for the immune system, evolutionary convergence toward a higher diversity is thought to be a protection mechanism to get ahead of the arms race with pathogens (Liston et al., 2021). Therefore, a plausible hypothesis is that the presence of TDT-dependent TCRs confers an additional, more "private" layer of security against the emergence of antigen-loss variants.

aGVHD is a harbinger of chronic GVHD and has remained the nemesis of patients and physicians during the entire history of AHCT, partly because its occurrence is unpredictable. Our aGVHD cohort was composed of HLA-matched siblings. In this situation, aGVHD is caused by donor T cells that react against host minor histocompatibility antigens (Vincent et al., 2011; Warren et al., 2012). On the other hand, histoincompatibility does not always elicit fatal GVHD. Indeed, in patients that received AHCT from donors presenting multiple disparities for minor histocompatibility antigens, only 73% developed aGVHD (Martin, 1991). It has been hypothesized that some AHCT donors might be stronger alloresponders than others (Baron et al., 2007). In our cohort of 73 donor-recipient pairs, the occurrence of severe aGVHD was strongly associated with a low proportion of neonatal TCRs in the donor repertoire. Such a protective effect of neonatal TCRs would explain reports that AHCT with cord blood rather than adult hematopoietic cells may be associated with a lower risk of GVHD (Cohen et al., 2020). Moreover, while some studies report no relationship between post-transplant diversity and GVHD occurrence (Buhler et al., 2020), our results on the diversity in grafts (pre-transplant) are consistent with those of Yew and colleagues, who report that a lower TCR diversity was correlated with GVHD occurrence and relapse, while a higher percentage of cord blood cells was correlated with a higher repertoire diversity (Yew et al., 2015). If our observation is validated in further studies, it will justify the preferential selection of AHCT donors with a high proportion of neonatal TCRs in their peripheral blood.

Together, our data support an emerging model in which the T cell repertoire is composed of two strata with differential reactivity to self and non-self antigens: public neonatal TCRs and private TDT-dependent TCRs. This model is remarkably coherent with insightful theoretical predictions by Vrisekoop and colleagues who labeled the two strata the "somatic" repertoire and the "ur"-repertoire (Vrisekoop et al., 2014). Our model is also consistent with functional studies demonstrating that neonatal T cells can no longer be considered immature versions of adult cells. On the contrary, they are highly functional and respond rapidly to antigenic challenges (Davenport et al., 2020; Rudd, 2020).

### Limitations of the study

The main limitation of this study is that it was done using CDR3 beta chains only. While paired sequencing technologies exist, they are still in their infancy and paired TCR cohorts of the sizes we analyzed are unfortunately not accessible at this time. Another limitation of our study is that we only analyzed TCRs from circulating T cells. We suspect that tissue-specific T cells will likely have a different CDR3 sharing across cohorts. Finally, we included in our analysis curated sets of disease-associated CDR3s. Since both the CDR3 alpha and beta chains as well as MHC molecule play a role in the peptide recognition, our finding of disease-associated TCRs is likely an overestimation, since in our study we do not have access to the individual's HLA haplotype or the alpha chain for each TCR.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - GVHD cohort individuals
- METHOD DETAILS
  - TCR sequencing datasets
  - GVHD cohort individuals
  - Isolating public and superpublic CDR3s
  - Calculating public fraction by frequency and by sequence
  - Disease-specific CDR3 sets
  - Isolating CDR3 from bulk RNA-Seq *in silico*
  - CDR3 sharing and repertoire overlaps
  - Diversity measurements
  - Recombination probability prediction
  - Number of mismatches
  - Gaussian mixture model
  - Hierarchical clustering
  - Expected cumulative frequency
  - Overlaps by top *N* most frequent CDR3
  - Treemap
  - Survival model
  - Classification and regression models
- QUANTIFICATION AND STATISTICAL ANALYSIS

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104968.

## ACKNOWLEDGMENTS

This study was supported by grant FDN-148400 from the Canadian Institutes of Health Research (to C.P.). A.T. was supported by a studentship from the Canadian Institutes of Health Research, and J.D.L. by a studentship from the Fonds de Recherche Québec – Santé.

## AUTHOR CONTRIBUTIONS

A.T., C.P., and S.Le. designed the study. A.T. performed the main bioinformatic analyses and result interpretation. J.S., J.-P.L., S.La., L.B., S.Le., and A.B. performed RNA Sequencing experiments. P.B., J.-D.L., and G.E. contributed to bioinformatic analyses. A.T., P.B., J.-D.L., G.E., S.Le., and C.P. contributed to the analysis and interpretation of data and results. A.T. and C.P. wrote the manuscript and all authors edited and approved the final manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as a member of the LGBTQ + community. One or more of the authors of this paper received support from a program designed to increase minority representation in science.

Received: April 27, 2022
Revised: June 24, 2022
Accepted: August 12, 2022
Published: September 16, 2022

## REFERENCES

Autissier, P., Soulas, C., Burdo, T.H., and Williams, K.C. (2010). Evaluation of a 12-color flow cytometry panel to study lymphocyte, monocyte, and dendritic cell subsets in humans. Cytometry A *77*, 410–419. https://doi.org/10.1002/cyto.a.20859.

Barennes, P., Quiniou, V., Shugay, M., Egorov, E.S., Davydov, A.N., Chudakov, D.M., Uddin, I., Ismail, M., Oakes, T., Chain, B., et al. (2020). Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat. Biotechnol. *39*, 236–245. https://doi.org/10.1038/s41587-020-0656-3.

Baron, C., Somogyi, R., Greller, L.D., Rineau, V., Wilkinson, P., Cho, C.R., Cameron, M.J., Kelvin, D.J., Chagnon, P., Roy, D.-C., et al. (2007). Prediction of graft-versus-host disease in humans by donor gene-expression profiling. PLoS Med. *4*, e23. https://doi.org/10.1371/journal.pmed.0040023.

Bentzen, A.K., Such, L., Jensen, K.K., Marquard, A.M., Jessen, L.E., Miller, N.J., Church, C.D., Lyngaa, R., Koelle, D.M., Becker, J.C., et al. (2018). T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide-MHC complexes. Nat. Biotechnol. *36*, 1191–1196. https://doi.org/10.1038/nbt.4303.

Bernin, H., Fehling, H., Marggraff, C., Tannich, E., and Lotter, H. (2016). The cytokine profile of human NKT cells and PBMCs is dependent on donor sex and stimulus. Med. Microbiol. Immunol. *205*, 321–332. https://doi.org/10.1007/s00430-016-0449-y.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381. https://doi.org/10.1038/nmeth.3364.

Bonati, A., Zanelli, P., Ferrari, S., Plebani, A., Starcich, B., Savi, M., and Neri, T.M. (1992). T-cell receptor beta-chain gene rearrangement and expression during human thymic ontogenesis. Blood *79*, 1472–1483. https://doi.org/10.1182/blood.V79.6.1472.1472.

Britanova, O.V., Shugay, M., Merzlyak, E.M., Staroverov, D.B., Putintseva, E.V., Turchaninova, M.A., Mamedov, I.Z., Pogorelyy, M.V., Bolotin, D.A., Izraelson, M., et al. (2016). Dynamics of individual T cell repertoires: from cord blood to centenarians. J. Immunol. *196*, 5005–5013. https://doi.org/10.4049/jimmunol.1600005.

Brodin, P., and Davis, M.M. (2017). Human immune system variation. Nat. Rev. Immunol. *17*, 21–29. https://doi.org/10.1038/nri.2016.125.

Buhler, S., Bettens, F., Dantin, C., Ferrari-Lacraz, S., Ansari, M., Mamez, A.-C., Masouridi-Levrat, S., Chalandon, Y., and Villard, J. (2020). Genetic T-cell receptor diversity at 1 year following allogeneic hematopoietic stem cell transplantation. Leukemia *34*, 1422–1432. https://doi.org/10.1038/s41375-019-0654-y.

Chu, N.D., Bi, H.S., Emerson, R.O., Sherwood, A.M., Birnbaum, M.E., Robins, H.S., and Alm, E.J. (2019). Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. BMC Immunol. *20*, 19. https://doi.org/10.1186/s12865-019-0300-5.

Clave, E., Araujo, I.L., Alanio, C., Patin, E., Bergstedt, J., Urrutia, A., Lopez-Lastra, S., Li, Y., Charbit, B., MacPherson, C.R., et al. (2018). Human thymopoiesis is influenced by a common genetic variant within the TCRA-TCRD locus. Sci. Transl. Med. *10*. https://doi.org/10.1126/scitranslmed.aao2966.

Cohen, S., Roy, J., Lachance, S., Delisle, J.-S., Marinier, A., Busque, L., Roy, D.-C., Barabé, F., Ahmad, I., Bambace, N., et al. (2020). Hematopoietic stem cell transplantation using single UM171-expanded cord blood: a single-arm, phase 1-2 safety and feasibility study. Lancet Haematol. *7*, e134–e145. https://doi.org/10.1016/S2352-3026(19)30202-9.

Davenport, M.P., Smith, N.L., and Rudd, B.D. (2020). Building a T cell compartment: how immune cell development shapes function. Nat. Rev. Immunol. *20*, 499–506. https://doi.org/10.1038/s41577-020-0332-3.

De Simone, M., Rossetti, G., and Pagani, M. (2018). Single cell T cell receptor sequencing: techniques and future challenges. Front. Immunol. *9*, 1638. https://doi.org/10.3389/fimmu.2018.01638.

Deibel, M.R., Jr., Riley, L.K., Coleman, M.S., Cibull, M.L., Fuller, S.A., and Todd, E. (1983). Expression of terminal deoxynucleotidyl transferase in human thymus during ontogeny and development. J. Immunol. *131*, 195–200.

DeWitt, W.S., 3rd, Smith, A., Schoch, G., Hansen, J.A., Matsen, F.A., and Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. Elife *7*. https://doi.org/10.7554/eLife.38358.

Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais, C., Klinger, M., Carlson, C.S., Hansen, J.A., et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat. Genet. *49*, 659–665. https://doi.org/10.1038/ng.3822.

Ernst, B., Lee, D.S., Chang, J.M., Sprent, J., and Surh, C.D. (1999). The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. Immunity *11*, 173–181. https://doi.org/10.1016/s1074-7613(00)80092-8.

Fischer, D.S., Wu, Y., Schubert, B., and Theis, F.J. (2020). Predicting antigen specificity of single T cells based on TCR CDR3 regions. Mol. Syst. Biol. *16*, e9416. https://doi.org/10.15252/msb.20199416.

Gil, A., Kamga, L., Chirravuri-Venkata, R., Aslan, N., Clark, F., Ghersi, D., Luzuriaga, K., and Selin, L.K. (2020). Epstein-barr virus epitope-major histocompatibility complex interaction combined with convergent recombination drives selection of diverse T cell receptor α and β repertoires. mBio *11*. https://doi.org/10.1128/mBio.00250-20.

Godfrey, D.I., Stankovic, S., and Baxter, A.G. (2010). Raising the NKT cell family. Nat. Immunol. *11*, 197–206. https://doi.org/10.1038/ni.1841.

Godfrey, D.I., Koay, H.-F., McCluskey, J., and Gherardin, N.A. (2019). The biology and functional importance of MAIT cells. Nat. Immunol. *20*, 1110–1128. https://doi.org/10.1038/s41590-019-0444-8.

Goronzy, J.J., and Weyand, C.M. (2019). Mechanisms underlying T cell ageing. Nat. Rev. Immunol. *19*, 573–583. https://doi.org/10.1038/s41577-019-0180-1.

de Greef, P.C., Oakes, T., Gerritsen, B., Ismail, M., Heather, J.M., Hermsen, R., Chain, B., and de Boer, R.J. (2020). The naive T-cell receptor repertoire has an extremely broad distribution of

clone sizes. Elife *9*. https://doi.org/10.7554/eLife.49900.

Haeryfar, S.M.M., Hickman, H.D., Irvine, K.R., Tscharke, D.C., Bennink, J.R., and Yewdell, J.W. (2008). Terminal deoxynucleotidyl transferase establishes and broadens antiviral CD8+ T cell immunodominance hierarchies. J. Immunol. *181*, 649–659. https://doi.org/10.4049/jimmunol.181.1.649.

Heikkilä, N., Sormunen, S., Mattila, J., Härkönen, T., Knip, M., Ihantola, E.-L., Kinnunen, T., Mattila, I.P., Saramäki, J., and Arstila, T.P. (2021). Generation of self-reactive, shared T-cell receptor α chains in the human thymus. J. Autoimmun. *119*, 102616. https://doi.org/10.1016/j.jaut.2021.102616.

Hogquist, K.A., and Jameson, S.C. (2014). The self-obsession of T cells: how TCR signaling thresholds affect fate "decisions" and effector function. Nat. Immunol. *15*, 815–823. https://doi.org/10.1038/ni.2938.

Jenkins, M.K., Chu, H.H., McLachlan, J.B., and Moon, J.J. (2010). On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. Annu. Rev. Immunol. *28*, 275–294. https://doi.org/10.1146/annurev-immunol-030409-101253.

Johnson, S.A., Seale, S.L., Gittelman, R.M., Rytlewski, J.A., Robins, H.S., and Fields, P.A. (2021). Impact of HLA type, age and chronic viral infection on peripheral T-cell receptor sharing between unrelated individuals. PLoS One *16*, e0249484. https://doi.org/10.1371/journal.pone.0249484.

Kedzierska, K., Thomas, P.G., Venturi, V., Davenport, M.P., Doherty, P.C., Turner, S.J., and La Gruta, N.L. (2008). Terminal deoxynucleotidyltransferase is required for the establishment of private virus-specific CD8+ TCR repertoires and facilitates optimal CTL responses. J. Immunol. *181*, 2556–2562. https://doi.org/10.4049/jimmunol.181.4.2556.

Khosravi-Maharlooei, M., Obradovic, A., Misra, A., Motwani, K., Holzl, M., Seay, H.R., DeWolf, S., Nauman, G., Danzl, N., Li, H., et al. (2019). Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire. J. Clin. Invest. *129*, 2446–2462. https://doi.org/10.1172/JCI124358.

Kim, H.T., Zhang, M.-J., Woolfrey, A.E., St Martin, A., Chen, J., Saber, W., Perales, M.-A., Armand, P., and Eapen, M. (2016). Donor and recipient sex in allogeneic stem cell transplantation: what really matters. Haematologica *101*, 1260–1266. https://doi.org/10.3324/haematol.2016.147645.

Krishna, C., Chowell, D., Gönen, M., Elhanati, Y., and Chan, T.A. (2020). Genetic and environmental determinants of human TCR repertoire diversity. Immun. Ageing *17*, 26. https://doi.org/10.1186/s12979-020-00195-9.

Li, B., Li, T., Wang, B., Dou, R., Zhang, J., Liu, J.S., and Liu, X.S. (2017). Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data. Nat. Genet. *49*, 482–483. https://doi.org/10.1038/ng.3820.

Liston, A., Lesage, S., Wilson, J., Peltonen, L., and Goodnow, C.C. (2003). Aire regulates negative selection of organ-specific T cells. Nat. Immunol. *4*, 350–354. https://doi.org/10.1038/ni906.

Liston, A., Carr, E.J., and Linterman, M.A. (2016). Shaping variation in the human immune system. Trends Immunol. *37*, 637–646. https://doi.org/10.1016/j.it.2016.08.002.

Liston, A., Humblet-Baron, S., Duffy, D., and Goris, A. (2021). Human immune diversity: from evolution to modernity. Nat. Immunol. *22*, 1479–1489. https://doi.org/10.1038/s41590-021-01058-1.

Lu, J., Van Laethem, F., Bhattacharya, A., Craveiro, M., Saba, I., Chu, J., Love, N.C., Tikhonova, A., Radaev, S., Sun, X., et al. (2019). Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. Nat. Commun. *10*, 1019. https://doi.org/10.1038/s41467-019-08906-7.

Lythe, G., Callard, R.E., Hoare, R.L., and Molina-París, C. (2016). How many TCR clonotypes does a body maintain? J. Theor. Biol. *389*, 214–224. https://doi.org/10.1016/j.jtbi.2015.10.016.

Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I.R., and Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. *24*, 1603–1612. https://doi.org/10.1101/gr.170753.113.

Marcou, Q., Mora, T., and Walczak, A.M. (2018). High-throughput immune repertoire analysis with IGoR. Nat. Commun. *9*, 561. https://doi.org/10.1038/s41467-018-02832-w.

Martin, P.J. (1991). Increased disparity for minor histocompatibility antigens as a potential cause of increased GVHD risk in marrow transplantation from unrelated donors compared with related donors. Bone Marrow Transplant. *8*, 217–223.

Martin, P.J., Levine, D.M., Storer, B.E., Warren, E.H., Zheng, X., Nelson, S.C., Smith, A.G., Mortensen, B.K., and Hansen, J.A. (2017). Genome-wide minor histocompatibility matching as related to the risk of graft-versus-host disease. Blood *129*, 791–798. https://doi.org/10.1182/blood-2016-09-737700.

Mayer, A., Balasubramanian, V., Walczak, A.M., and Mora, T. (2019). How a well-adapting immune system remembers. Proc. Natl. Acad. Sci. USA *116*, 8815–8823. https://doi.org/10.1073/pnas.1812810116.

Mittelbrunn, M., and Kroemer, G. (2021). Hallmarks of T cell aging. Nat. Immunol. *22*, 687–698. https://doi.org/10.1038/s41590-021-00927-z.

Murray, J.M., Kaufmann, G.R., Hodgkin, P.D., Lewin, S.R., Kelleher, A.D., Davenport, M.P., and Zaunders, J.J. (2003). Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty. Immunol. Cell Biol. *81*, 487–495. https://doi.org/10.1046/j.1440-1711.2003.01191.x.

Natarajan, A., and Krogsgaard, M. (2018). The myriad targets of a T cell. Nat. Biotechnol. *36*, 1152–1154. https://doi.org/10.1038/nbt.4309.

Ni, X., Song, Q., Cassady, K., Deng, R., Jin, H., Zhang, M., Dong, H., Forman, S., Martin, P.J., Chen, Y.-Z., et al. (2017). PD-L1 interacts with CD80 to regulate graft-versus-leukemia activity of donor CD8+ T cells. J. Clin. Invest. *127*, 1960–1977. https://doi.org/10.1172/JCI91138.

Pahwa, R.N., Modak, M.J., McMorrow, T., Pahwa, S., Fernandes, G., and Good, R.A. (1981). Terminal deoxynucleotidyl transferase (TdT) enzyme in thymus and bone marrow. I. Age-associated decline of TdT in humans and mice. Cell. Immunol. *58*, 39–48. https://doi.org/10.1016/0008-8749(81)90147-7.

Palmer, S., Albergante, L., Blackburn, C.C., and Newman, T.J. (2018). Thymic involution and rising disease incidence with age. Proc. Natl. Acad. Sci. USA *115*, 1883–1888. https://doi.org/10.1073/pnas.1714478115.

Pauken, K.E., Lagattuta, K.A., Lu, B.Y., Lucca, L.E., Daud, A.I., Hafler, D.A., Kluger, H.M., Raychaudhuri, S., and Sharpe, A.H. (2022). TCR-sequencing in cancer and autoimmunity: barcodes and beyond. Trends Immunol. *43*, 180–194. https://doi.org/10.1016/j.it.2022.01.002.

Pogorelyy, M.V., Elhanati, Y., Marcou, Q., Sycheva, A.L., Komech, E.A., Nazarov, V.I., Britanova, O.V., Chudakov, D.M., Mamedov, I.Z., Lebedev, Y.B., et al. (2017). Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. PLoS Comput. Biol. *13*, e1005572. https://doi.org/10.1371/journal.pcbi.1005572.

Pogorelyy, M.V., Minervina, A.A., Touzel, M.P., Sycheva, A.L., Komech, E.A., Kovalenko, E.I., Karganova, G.G., Egorov, E.S., Komkov, A.Y., Chudakov, D.M., et al. (2018). Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. Proc. Natl. Acad. Sci. USA *115*, 12704–12709. https://doi.org/10.1073/pnas.1809642115.

Quigley, M.F., Greenaway, H.Y., Venturi, V., Lindsay, R., Quinn, K.M., Seder, R.A., Douek, D.C., Davenport, M.P., and Price, D.A. (2010). Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. Proc. Natl. Acad. Sci. USA *107*, 19414–19419. https://doi.org/10.1073/pnas.1010586107.

Rudd, B.D. (2020). Neonatal T cells: a reinterpretation. Annu. Rev. Immunol. *38*, 229–247. https://doi.org/10.1146/annurev-immunol-091319-083608.

Sethna, Z., Elhanati, Y., Callan, C.G., Walczak, A.M., and Mora, T. (2019). OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics *35*, 2974–2981. https://doi.org/10.1093/bioinformatics/btz035.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Simpson, E.H. (1949). Measurement of diversity. Nature *163*, 688. https://doi.org/10.1038/163688a0.

Sng, J., Ayoglu, B., Chen, J.W., Schickel, J.-N., Ferre, E.M.N., Glauzy, S., Romberg, N., Hoenig, M., Cunningham-Rundles, C., Utz, P.J., et al. (2019). AIRE expression controls the peripheral

selection of autoreactive B cells. Sci. Immunol. *4*. https://doi.org/10.1126/sciimmunol.aav6778.

Socié, G., and Blazar, B.R. (2009). Acute graft-versus-host disease: from the bench to the bedside. Blood *114*, 4327–4336. https://doi.org/10.1182/blood-2009-06-204669.

Sood, A., Lebel, M.-È., Dong, M., Fournier, M., Vobecky, S.J., Haddad, É., Delisle, J.-S., Mandl, J.N., Vrisekoop, N., and Melichar, H.J. (2021). CD5 levels define functionally heterogeneous populations of naïve human CD4+ T cells. Eur. J. Immunol. *51*, 1365–1376. https://doi.org/10.1002/eji.202048788.

Soto, C., Bombardi, R.G., Kozhevnikov, M., Sinkovits, R.S., Chen, E.C., Branchizio, A., Kose, N., Day, S.B., Pilkinton, M., Gujral, M., et al. (2020). High frequency of shared clonotypes in human T cell receptor repertoires. Cell Rep. *32*, 107882. https://doi.org/10.1016/j.celrep.2020.107882.

Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. (2020). Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. Front. Immunol. *11*, 1803. https://doi.org/10.3389/fimmu.2020.01803.

Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of T Cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. Front. Immunol. *12*, 664514. https://doi.org/10.3389/fimmu.2021.664514.

Tanno, H., Gould, T.M., McDaniel, J.R., Cao, W., Tanno, Y., Durrett, R.E., Park, D., Cate, S.J., Hildebrand, W.H., Dekker, C.L., et al. (2020). Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. Proc. Natl. Acad. Sci. USA *117*, 532–540. https://doi.org/10.1073/pnas.1915008117.

Thome, J.J.C., Grinshpun, B., Kumar, B.V., Kubota, M., Ohmura, Y., Lerner, H., Sempowski, G.D., Shen, Y., and Farber, D.L. (2016). Longterm maintenance of human naive T cells through in situ homeostasis in lymphoid tissue sites. Sci. Immunol. *1*. https://doi.org/10.1126/sciimmunol.aah6506.

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics *33*, 2924–2929. https://doi.org/10.1093/bioinformatics/btx286.

Venturi, V., Chin, H.Y., Asher, T.E., Ladell, K., Scheinberg, P., Bornstein, E., van Bockel, D., Kelleher, A.D., Douek, D.C., Price, D.A., and Davenport, M.P. (2008). TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. J. Immunol. *181*, 7853–7862. https://doi.org/10.4049/jimmunol.181.11.7853.

Vincent, K., Roy, D.-C., and Perreault, C. (2011). Next-generation leukemia immunotherapy. Blood *118*, 2951–2959. https://doi.org/10.1182/blood-2011-04-350868.

Vrisekoop, N., Monteiro, J.P., Mandl, J.N., and Germain, R.N. (2014). Revisiting thymic positive selection and the mature T cell repertoire for antigen. Immunity *41*, 181–190. https://doi.org/10.1016/j.immuni.2014.07.007.

Warren, E.H., Zhang, X.C., Li, S., Fan, W., Storer, B.E., Chien, J.W., Boeckh, M.J., Zhao, L.P., Martin, P.J., and Hansen, J.A. (2012). Effect of MHC and non-MHC donor/recipient genetic disparity on the outcome of allogeneic HCT. Blood *120*, 2796–2806. https://doi.org/10.1182/blood-2012-04-347286.

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H.A., Skowera, A., Miles, J.J., Tan, M.P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D.A., et al. (2012). A single autoimmune T cell receptor recognizes more than a million different peptides. J. Biol. Chem. *287*, 1168–1177. https://doi.org/10.1074/jbc.M111.289488.

Ye, J., Ma, N., Madden, T., L, and Ostell, J., M (2013). IgBLAST: An immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. *41* (Web Server Issue), 34–40. https://doi.org/10.1093/nar/gkt382.

Yew, P.Y., Alachkar, H., Yamaguchi, R., Kiyotani, K., Fang, H., Yap, K.L., Liu, H.T., Wickrema, A., Artz, A., van Besien, K., et al. (2015). Quantitative characterization of T-cell repertoire in allogeneic hematopoietic stem cell transplant recipients. Bone Marrow Transplant. *50*, 1227–1234. https://doi.org/10.1038/bmt.2015.133.

Yousefzadeh, M.J., Flores, R.R., Zhu, Y., Schmiechen, Z.C., Brooks, R.W., Trussoni, C.E., Cui, Y., Angelini, L., Lee, K.-A., McGowan, S.J., et al. (2021). An aged immune system drives senescence and ageing of solid organs. Nature *594*, 100–105. https://doi.org/10.1038/s41586-021-03547-7.

Zhang, S.-Q., Ma, K.-Y., Schonnesen, A.A., Zhang, M., He, C., Sun, E., Williams, C.M., Jia, W., and Jiang, N. (2018). High-throughput determination of the antigen specificities of T cell receptors in single cells. Nat. Biotechnol. *36*, 1156–1159. https://doi.org/10.1038/nbt.4282.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| $CD4^+$ T cells: RNAseq data | This study | NCBI SRA: PRJNA832136 |
| TCRseq from Emerson et al., (2017) | AdaptiveBiotech | https://clients.adaptivebiotech.com/pub/emerson-2017-natgen |
| TCRseq from Thome et al., (2016) | AdaptiveBiotech | https://adaptivebiotech.com/pub/Farber-2016-SciImmunol |
| TCRSeq from Sng et al., (2019) | AdaptiveBiotech | https://clients.adaptivebiotech.com/pub/sng-2019-sciimmunol |
| TCRSeq from (Britanova et al., 2016) | Zenodo | https://doi.org/10.5281/zenodo.826447 |
| Pathology-associated TCR database | Tickotsky et al., (2017) | http://friedmanlab.weizmann.ac.il/McPAS-TCR/ |
| Software and algorithms | | |
| MIXCR | Bolotin et al., (2015) | https://github.com/milaboratory/mixcr |
| OLGA | Sethna et al., (2019) | https://github.com/statbiophys/OLGA |
| IgBLAST | (Ye et al., 2013) | https://github.com/ncbi/igblast |
| Scipy.stats | Scipy python package | https://docs.scipy.org/doc/scipy/index.html |
| Scikit Learn | Scikit learn python package | https://scikit-learn.org/stable/ |
| Squarify | Squarify python package | https://github.com/laserson/squarify |
| Survival | Survival R package | https://cran.r-project.org/web/packages/survival/index.html |
| Lifelines | Lifelines python package | https://lifelines.readthedocs.io/en/latest/ |
| Analysis code | This study | https://github.com/TrofimovAssya/TCR_populationX_publication |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Claude Perreault (claude.perreault@umontreal.ca).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

RNA-Seq data from the GVHD cohort has been deposited in the Sequence Read Archive and are publicly available at BioSample accession PRJNA832136. DOIs are listed in the key resources table.

All original codes generated during this study are available in the form of python jupyter notebooks on Github, at the address listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### GVHD cohort individuals

The Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont approved the experiments using human materials which are reported in the present study. The *GVHD* cohort included 73 healthy

sibling-matched donors. Written informed consent was obtained from all patients or their legal guardians before sample collection or hematopoietic stem cell transplantation. For each sample, peripheral blood mononuclear cells (PBMC) were collected, and CD4 T cells were isolated by immunomagnetic positive selection (EasySep Human CD4 positive selection kit; Stemcell Technologies). Sex and age for these individuals can be found in the metadata corresponding to the RNA-Seq data available from the Sequence Read Archive under accession number PRJNA832136.

## METHOD DETAILS

### TCR sequencing datasets

We downloaded TCR sequences and additional data from four non-overlapping cohorts: (Britanova et al., 2016; Emerson et al., 2017; Sng et al., 2019; Thome et al., 2016). A total of 980 human subjects were included in these cohorts, with 401 females and 517 males; the sex of 47 subjects was unknown. For further details on numbers of sequences, see Figures S3A and S3B.

### GVHD cohort individuals

RNA was extracted from purified CD4 T cells by Trizol-Column (PureLink RNA Mini kit; Thermo Fisher Scientific). RNA was quantified by U.V. spectrophotometry (Tecan Infinite M1000), and quality was verified by Bioanalyzer (Nano RNA Chip; Agilent). Whole transcriptome libraries were prepared with the Ion Torrent Total RNA-Seq Kit v2 (Thermo Fisher Scientific) from 200 ng total poly-A enriched RNA (Dynabead mRNA direct MicroKit; Ambion). Sequencing was done on an Ion P1 chip using the Thermo Fisher Ion Proton System to a minimum of 30M reads.

### Isolating public and superpublic CDR3s

For each cohort, CDR3 beta amino acid sequences were pooled together and occurences in the cohort of each individual sequence was counted. Public CDR3aa are defined as seen in at least two individuals in the cohort, while superpublic CDR3aa are defined as seen in at least half of the individuals in the cohort.

### Calculating public fraction by frequency and by sequence

For each individual, the public fraction (Figure 2I) is calculated as the fraction of the CDR3aa that overlaps with the cohort's public CDR3aa pool. The summed clonality (Figure 2J) is calculated by summing the clonality (relative CDR3 frequency in the sample) of the CDR3aa overlapping with the cohort's public CDR3aa pool.

### Disease-specific CDR3 sets

From the *McPAS* CDR3 datasets, we downloaded the McPAS database on 2021-08-12 (Tickotsky et al., 2017). We included in our study CDR3beta amino acid sequences of human origin found in the two top categories of diseases: Pathogens and Autoimmune.

### Isolating CDR3 from bulk RNA-Seq *in silico*

From each RNA-Seq donor sample from the *GVHD* cohort, we isolated CDR3 contigs using the MIXCR software (Bolotin et al., 2015). Since the Ion Proton sequencing system generates variable-length reads, we allowed for partial alignments and performed two passes of contig assembly. To rescue as many CDR3s as possible, for incomplete TCR CDR3s, we allowed for extension via the V/J genes, since it has been shown to introduce limited errors, because of the very conserved nature of TCRs on both ends (pattern CASS————EF) (Bolotin et al., 2015).

### CDR3 sharing and repertoire overlaps

We calculated sharing of individual CDR3 sequences based solely on the amino acid sequence without matching V/J genes and nucleotide sequences. This approach was selected to assess sharing of the final protein product of CDR3s found in the body rather than to look at specific mRNA features.

Thus, for public fraction calculations based on unique CDR3aa sequences, we calculated the number of public sequences in an individual's repertoire and divided it by the total number of sequences in the repertoire (results of Figure 2I). To assess the clonality of public CDR3aa, we summed the clonal frequencies attributed to individual public sequences (results of Figure 2J).

We use the Jaccard distance $d_{J(A,B)}$ as a measure of dissimilarity between two CDR3 sets $A$ and $B$:

$$d_{J(A,B)} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Here a value of 0 means exact overlap of the CDR3 sets, and 1 means no overlap.

### Diversity measurements

We used the inverse Simpson clonality and Shannon entropy as diversity measurements. Inverse Simpson diversity index is calculated by weighted arithmetic mean of each squared clone abundance (Simpson, 1949), and we implemented this as a function in python. Shannon entropy (Shannon, 1948) is calculated using the *scipy.stats* python package.

### Recombination probability prediction

CDR3aa recombination probability was predicted using the OLGA software (Sethna et al., 2019). We downloaded the command-line tool commit version 4e0bc36 from the repository (https://github.com/statbiophys/OLGA) and selected *humanTRB* as the alignment database for the predictions.

### Number of mismatches

We used IgBLAST (Ye et al., 2013) to align each nucleotide sequence to the annotated germline V, D, and J sequences to calculate the number of mismatches. We downloaded the package from the repository (https://github.com/ncbi/igblast) commit version dfb98f8. We used the human database and specified TCRs. We obtained the aligned sequence for each result and counted the number of mismatches between the aligned sequence and the germline.

### Gaussian mixture model

We used the Gaussian mixture model from the python *GaussianMixture* function from the *sklearn* library to fit each Gaussian Mixture Model. We selected a 2 component Gaussian Mixture Model with a diagonal covariance type. We grouped individuals of the Britanova cohort by age group. For each age group, we fitted a separate Gaussian Mixture Model on the recombination frequency and clonality quantifications. Then, we evaluated the fit of this model in other age groups. This fit was calculated as a loglikelihood of fit for the data to the pre-trained model. We then reported the average loglikelihood for each age-group - model pair in the heatmap in Figure 2.

### Hierarchical clustering

We used hierarchical clustering with an unweighted pair group method with arithmetic mean agglomerative function for all hierarchical clustering experiments in this study. Visual assessment was used to split each dendrogram into clusters manually. We used the clustermap function from the seaborn python library to plot heatmaps and associated dendrograms.

### Expected cumulative frequency

For each CDR3aa sequence, we calculated the sharing percentage and grouped sequences according to the CDR3aa sharing bins. Then, we calculated for each CDR3aa the cumulative repertoire frequency by summing frequencies of all CDR3aa in each bin across all individual repertoires. The average repertoire frequency for each bin was $4.12 \times 10^{-6} \pm 1.2 \times 10^{-6}$. To draw the expected cumulative frequency line, we multiplied this overall frequency by the median number of individuals of each sharing bin. We reported this value as the mean expected cumulative frequency (red dotted line on the plot).

### Overlaps by top *N* most frequent CDR3

We ranked CDR3 in descending order of clonal frequency, and for a growing $N$, we selected the top $N$ most frequent CDR3 in each repertoire. Then, we calculated the percentage overlap with the disease-specific CDR3 set. Individual percentages were grouped by age group, and for each age group, the SD within the age group is shown on each line plot.

### Treemap

We used the treemap function from the *squarify* python library. The package was given the CDR3 clonal frequencies and a random color palette.

### Survival model

We used the *survival* package in R for plotting the Kaplan-Meier plots and the *lifelines* packages in python for the CoxPH model. For each Kaplan-Meier plot, we split the group by median as well as 25 and 75% quantiles to attempt to find the best group separation for each CDR3 characteristic. All plots can be seen in Figure S6 with associated statistical testing.

For the *CoxPH* models, we used the *lifelines* python library with the option for right-censored data. On each plot, we reported the log (hazard ratios) as well as the bottom and top 95% confidence intervals.

### Classification and regression models

The logistic regression and random forest models from the *sklearn* python library were used to classify *neonatal* and *TDT-dependent* CDR3s in Figure 7. We used the default parameters for each model: respectively, an L2 penalty with a regularization strength of 1 and the L-BFGS solver for the logistic regression and 100 estimators, a Gini impurity criterion, no max depth, and a minimum number of samples of 2 for the split for the random forest classifier.

For the ablation study, each model received the selected combination of features and learned to classify CDR3 into two classes: *neonatal* or *TDT-dependent*. During the dataset preparation, two repertories of each type (cord blood and child) were held out. These repertoires comprised the test set of new data. The performance of each iteration of the model given the combination of input features was reported in Figure 7, with performances color-coded for visual comparison.

A linear regression model was trained to determine the relative importance of the features. We used the *LinearRegression* function from the *sklearn* python library with the following default features: intercept fitting was allowed, and negative coefficients were allowed. This model received as input a binary vector of the presence/absence of the features and learned to predict the performance of either of the two models (logistic regression or random forest) obtained previously for each feature combination. The coefficients attributed to each binary feature presence/absence were used to compare relative importance. Coefficients close to zero meant there was little weight attributed to the model and vice versa. The ranking of features was obtained by ordering the absolute values of the coefficients in descending order.

## QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed using Python v3.7.6 or R v4.0.4. All statistical tests used are mentioned in the figure legends. Significance level ($p < 0.05$) results are marked with (*) in the figures. Mann-Whitney U and One-way ANOVA tests were performed using the mannwhitneyu and ANOVA functions respectively from scipy.stats python module and R. All boxes in boxplots show the first ($25^{th}$ percentile) and third quartiles ($75^{th}$ percentile) and the median while the whiskers designate the minimum (first quartile value – 1.5*interquartile range) and maximum (third quartile value +1.5*interquartile range).