T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity

Asaf Madi,[1,3] Eric Shifrut,[1,3] Shlomit Reich-Zeliger,[1] Hilah Gal,[1] Katharine Best,[2] Wilfred Ndifon,[1,4] Benjamin Chain,[2] Irun R. Cohen,[1] and Nir Friedman[1]

[1]*Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel;* [2]*Division of Infection and Immunity, The Cruciform Building, UCL, London WC1 6BT, United Kingdom*

The T-cell receptor (TCR) repertoire is formed by random recombinations of genomic precursor elements; the resulting combinatorial diversity renders unlikely extensive TCR sharing between individuals. Here, we studied CDR3β amino acid sequence sharing in a repertoire-wide manner, using high-throughput TCR-seq in 28 healthy mice. We uncovered hundreds of public sequences shared by most mice. Public CDR3 sequences, relative to private sequences, are two orders of magnitude more abundant on average, express restricted V/J segments, and feature high convergent nucleic acid recombination. Functionally, public sequences are enriched for MHC-diverse CDR3 sequences that were previously associated with autoimmune, allograft, and tumor-related reactions, but not with anti-pathogen-related reactions. Public CDR3 sequences are shared between mice of different MHC haplotypes, but are associated with different, MHC-dependent, V genes. Thus, despite their random generation process, TCR repertoires express a degree of uniformity in their post-genomic organization. These results, together with numerical simulations of TCR genomic rearrangements, suggest that biases and convergence in TCR recombination combine with ongoing selection to generate a restricted subset of self-associated, public CDR3 TCR sequences, and invite reexamination of the basic mechanisms of T-cell repertoire formation.

[Supplemental material is available for this article.]

The genome provides the raw material for the somatic generation of enormous T-cell receptor (TCR) diversity. These receptors are generated through a random process of DNA rearrangement, which involves the recombination of the germline V, D, and J gene segments and the deletion and insertion of nucleotides at the V(D)J junctions. The variety of the resulting TCR repertoire is required to recognize a large and unpredictable range of antigens of foreign and self origin. The potential diversity of TCR molecules synthesized during the maturation of T cells in the thymus is estimated to be $>10^{15}$ for the mouse TCRA and TCRB repertoire (Casrouge et al. 2000) and $>10^{10}$ for the β segment of the TCR. In contrast, the number of unique TCR types appearing in the peripheral lymphoid organs of an individual mouse ($\sim 10^6$) (Bousso et al. 1998) is many orders of magnitude less than this potential diversity. This excess of potential thymic TCR diversity leads to the expectation that different individuals would hardly ever share the same TCR recombination. Nevertheless, several reports have demonstrated identical TCR sequences occurring in the T-cell responses to defined antigens in different MHC-matched humans (Moss et al. 1991; Argaet et al. 1994), macaques (Venturi et al. 2008a), and mice (Venturi et al. 2008b). There have also been studies reporting substantial overlap in the naïve TCR repertoire between two mice, of ~18%–27% (Bousso et al. 1998). Shared TCR molecules can be referred to as "public"; "private" T-cell responses involve little TCR sharing. It has been suggested that an adequate sampling of individual TCR repertoires would demonstrate the true prevalence of public TCR sequences (Venturi et al. 2006, 2008b).

Here, we investigated TCR publicness using high-throughput sequencing of the TCR repertoires of a large number of mice. We obtained high-resolution repertoires of $CD4^+$ T cells in 28 individual C57BL/6 mice, naïve or immunized, based on massive parallel sequencing (TCR-seq) of T-cell mRNA (Ndifon et al. 2012). This data set allowed us to identify patterns of sequence sharing in a model system in which genetic background and MHC haplotypes are identical in all individuals. Thus, we could evaluate intrinsic factors that affect TCR sharing, such as biases in the TCR recombination process (Murugan et al. 2012; Ndifon et al. 2012), convergent recombination (Venturi et al. 2008b), and clonal selection, without confounding effects such as HLA polymorphisms and other genetic differences that occur in studies of human samples.

Results

High-throughput TCR sequencing reveals patterns of CDR3 sequence sharing among mice

To investigate TCR publicness in a repertoire-wide manner, we generated high-resolution maps of repertoires of splenic $CD4^+$ T cells in 28 individual C57BL/6 mice, based on massive parallel sequencing (TCR-seq) of T-cell mRNA (Ndifon et al. 2012). cDNA was generated using a universal Cβ primer, and was PCR amplified with

[3] **These authors contributed equally to this work.**
[4] **Present address: African Institute for Mathematical Sciences, Muizenberg 7945, South Africa**
Corresponding authors: irun.cohen@weizmann.ac.il, nir.friedman@weizmann.ac.il

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.170753.113.

a set of Vβ primers. Further processing of the resulting amplicons included cleavage followed by ligation of sequencing adaptors, to ensure complete coverage of the V-D-J junction region in each sequence read. We estimated PCR bias due to primer multiplexing by sequencing cloned TCR sequences of known V segments, and normalized read counts accordingly (Ndifon et al. 2012). The mice included 12 untreated, seven immunized with complete Freund's adjuvant (CFA), and nine immunized with CFA + ovalbumin (OVA) (see Methods for additional details). About 2.4 × 10^6 annotated nucleotide (nt) sequence reads were obtained, which corresponded to about 3.5 × 10^5 unique (nonredundant) TCR amino acid (aa) sequences. A summary of the samples is presented in Supplemental Table S1. Our analysis here is focused mainly on the aa sequences of the complementarity determining region 3 (CDR3), which is the most diverse region of the TCR molecule and is associated with antigen epitope recognition (Davis and Bjorkman 1988). Due to the degeneracy of the genetic code, the same functional CDR3 aa sequence could result from different nt recombinations—a phenomenon termed "convergent recombination" (Venturi et al. 2006, 2008b).

We found that on average, any two mice in our data set share 10.5% ± 1.8% of their expressed CDR3 aa sequences (nonredundant sequences in each mouse). There was no significant difference in pairwise sharing between the naïve and immunized groups of mice (see Fig. 1A); hence, we combined all 28 mice for further analysis. We next binned unique CDR3 aa sequences according to the number of mice in which they occurred (Fig. 1B). Most of the CDR3 aa sequences were found in only one mouse (~69% of all sequences). However, hundreds of sequences were highly shared among individual mice; 1908 sequences were shared by >75% ($n > 21$) of the mice. Notably, we found 289 CDR3 aa sequences that were shared by all 28 mice (~0.08% of all sequences). Thus, CDR3 amino acid sequences have a wide range of sharing levels, from private to highly public. Down-sampling of our data shows that the percentage of public clones grows with the number of sampled sequences (Supplemental Fig. S1), thus the observed level of public clones is likely an underestimate at our current coverage level of the TCR repertoire. We defined a sequence as "private" if it appeared in only one mouse in our data set, as "relatively private" if it was shared by two to seven mice, as "relatively public" if shared by 22–27 mice, and as "public" if shared by all 28 mice.

Public TCR sequences are much more abundant than are private sequences

We next analyzed the frequency of each CDR3 aa sequence as a function of its degree of sharing. CDR3 sequence frequency reflects two factors: the number of T cells bearing that aa sequence (which we term the "CDR3-type") and the amount of relevant mRNA produced by a cell. Thus, the frequency of a sequence reflects the numbers and the activity state of the T cells that express the specific receptor sequence. Figure 1C plots the mean frequency per mouse of each CDR3 aa sequence (averaged across all mice that share that sequence) as a function of its level of sharing. We observed a gradual increase in median frequency as a function of sharing; CDR3 sequences with higher levels of sharing tended to be more abundant (Fig. 1C; Supplemental Fig. S2). Interestingly, very frequent CDR3 aa sequences (relative frequency > 5 × 10^{-4}) appear both among private or relatively private sequences as well as among more public sequences.

Figure 1D shows the cumulative frequencies of the CDR3 aa sequences in each sharing category, averaged over all mice. The cumulative frequencies were relatively high at both extremes—the most private and the most public CDR3 aa sequences; there appeared to be a dip in the cumulative frequency curve for sequences manifesting intermediate levels of sharing. The public subgroup of 289 CDR3 sequences constitutes 10 ± 5% of the total sequences. Assuming similar per cell levels of mRNA encoding TCRB among cells of the different sharing categories, these results suggest that public CDR3 sequences represent highly abundant T-cell CDR3-types; on average, the number of T cells that bear each of the public CDR3 aa sequences is 50–100 times higher than that of an average private T-cell CDR3-type. Note, however, that many private CDR3 aa sequences were highly abundant (Fig. 1C). To confirm the abundance of public sequences, we sought their occurrence in a set of 79 TCRs that we randomly cloned and sequenced using Sanger sequencing. Twenty-two of these 79 sequences (28%) were identical to one of the 289 public CDR3 aa sequences derived by TCR-seq, validating by an independent method the relative abundance of public CDR3 sequences.

Public sequences have a very high level of convergent recombination

Previous studies reported that public TCRs manifest a higher level of convergent recombination (Venturi et al. 2006, 2011; Quigley et al. 2010; Li et al. 2012). Our analysis of a large number of individuals revealed a continuous trend; increased sharing was associated with a gradual increase in the mean degree of convergent recombination (Fig. 2A); private CDR3 aa sequences were encoded on average by one nt sequence, the public sequences were encoded by 34.5 nt sequences on average.

Figure 2B shows examples of patterns of convergent recombination of nt sequences for four CDR3 aa sequences. The two more public sequences on the left, found in 28 and 27 of the mice, showed high convergent recombination (encoded by 105 and 53 nt sequences, respectively). There was no dominant nt sequence in any mouse, nor a dominant nt sequence across mice. In contrast, the two relatively private CDR3 aa sequences shown on the right, present in seven and three mice, were encoded by only 1 or 2 nt sequences. Of note, the CDR3 aa sequence present in three mice (Fig. 3B, right panel) manifested a frequency that was markedly higher than that of the other sequences (Fig. 1C). Thus, private and public CDR3 aa segments differ markedly in their detectable degree of convergent recombination, irrespective of their relative frequency.

Public sequences differ from private sequences in gene segment usage and CDR3 length

Further analysis of the public CDR3 sequences revealed other distinct characteristics. Figure 3A shows the mean CDR3 lengths for private and relatively private sequences (shared by one to three mice) and public and relatively public ones (shared by 26–28 mice). The more public CDR3 aa sequences tended to be shorter on average by about one aa residue and, in addition, showed a significantly lower number of nt insertions in the VD and DJ junctions (Fig. 3B); this suggests that public sequences in mice are closer to germ-line configurations, as was observed in humans (Robins et al. 2010). However, the number of junctional nt deletions in both public and private sequences was similar (Fig. 3C). Public CDR3 sequences also manifested a skewed and restricted V and J segment usage compared with private sequences (Fig. 3D,E); three V genes (V2, V15, and V18) were not used in public sequences, and other V and J genes were represented at significantly different frequencies.

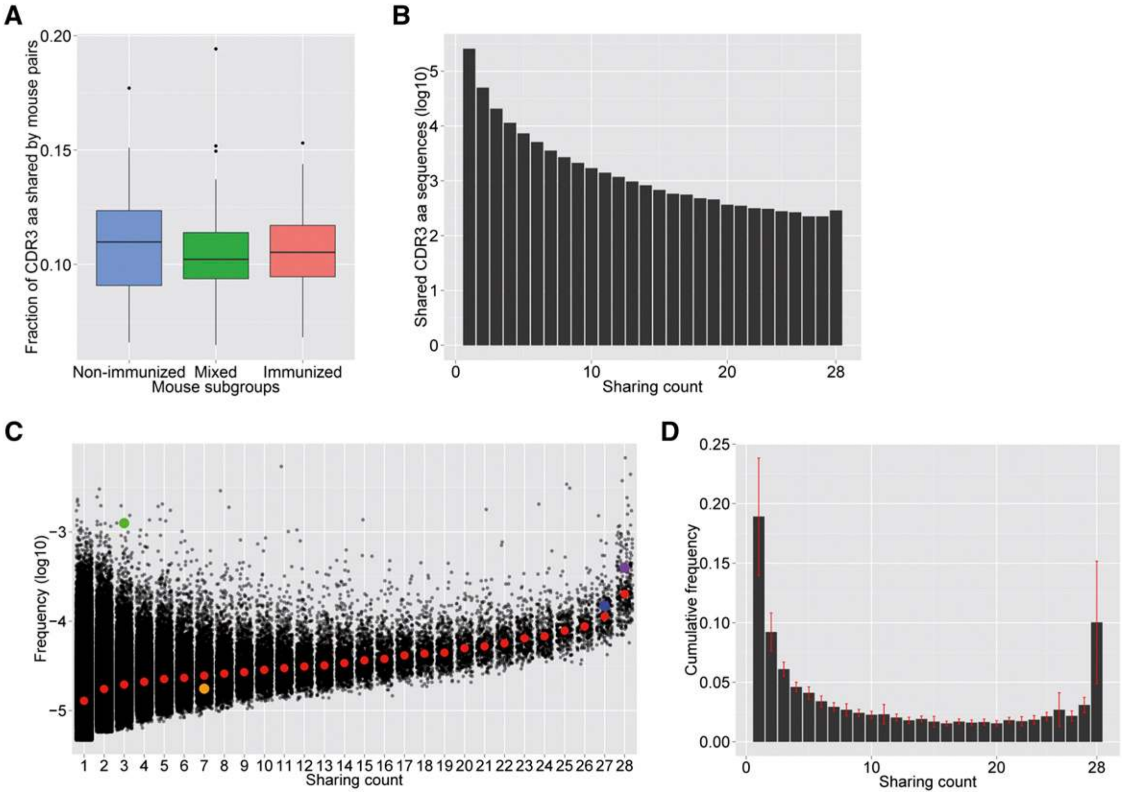

Figure 1. Analysis of the TCRβ repertoire of 28 C57BL/6 mice reveals a highly shared subset of CDR3 aa sequences present at relatively high frequencies. (*A*) Boxplot presentation of sharing between pairs of mice separated into non-immunized, immunized, and mixed (immunized/non-immunized) pairs; no significant differences are seen between the groups. (*B*) The number of CDR3 aa sequences found in each sharing category; 2.5×10^5 sequences (~69%) are private (found in only one mouse); 289 sequences (~0.08%) are public (found in all 28 mice). (*C*) Mean frequencies of all CDR3 aa sequences as a function of their sharing level. Each black dot represents the mean frequency of a single aa sequence in all the mice that share this sequence. The red dots show the median value for each sharing category. The colored dots refer to the four sequences shown in Figure 2B. (*D*) Normalized cumulative frequencies of the sequences from *C*, in each sharing category. Private CDR3 sequences account for 19% ± 5% of all sequences in each sample; the 289 public sequences account for 10% ± 5% of all sequences in each sample.

Public CDR3 sequences are enriched with self-related specificities

The marked differences between public and private CDR3 sequences raised the possibility that each class might have been driven by different classes of antigens. Interestingly, we noted that a sequence (C9: CASSLGGNQDTQYF), which was previously found to be public in NOD mice that spontaneously develop autoimmune type 1 diabetes (Tikochinski et al. 1999), was relatively public in our data set of healthy C57BL/6 mice (shared by 27 mice). The C9 CDR3 sequence, marked by the blue dot in Figure 2B, was found to recognize a peptide epitope (p277) in the mouse/human HSPD1 protein (Tikochinski et al. 1999); administration of peptide p277 to NOD mice activates anti-C9 and other regulatory T cells (Elias et al. 1999), and arrests the destruction of pancreatic beta cells both in NOD mice (Elias and Cohen 1994) and in humans with recent-onset type 1 diabetes (Raz et al. 2001, 2014). Despite the fact that the NOD and the C57BL/6 mouse strains differ in their MHC haplotypes ($H2^{g7}$ and $H2^{b}$, respectively), we found the same C9 CDR3β aa sequence to be public in both. Although we have no information about the TCRA chain that also participates in antigen recognition, the publicness of the C9 sequence prompted us to search the literature for additional annotated sequences in various disease models in different strains of mice bearing varying MHC haplotypes.

We collected from the literature 252 mouse TCRβ sequences that were previously annotated to be associated with defined immune functions, and compared them with our data set. The annotated sequences were associated with four categories of immune reactions: (1) Immunity to foreign pathogens; (2) allograft reactions; (3) tumor-related T cells; and (4) autoimmune conditions. Of the 252 annotated CDR3 sequences, we identified 124 sequences that were present in one or more of our 28 healthy C57BL/6 mice (see Supplemental Table S2 for the sequences and their references). Figure 4A shows the frequencies at which each annotated CDR3 aa sequence is expressed in the mice in our data set. The annotated sequences associated functionally with autoimmunity, allograft rejection and cancer (self or modified self) were relatively enriched with shared, public sequences compared with the sequences associated with anti-virus or anti-malarial immunity. This is evident also from Figure 4B, which shows the sharing distribution of annotated sequences of the four functional categories. We found that ~20% of the annotated sequences associated with categories of self or modified self-antigens were relatively public or public (found in >21 mice). In contrast, only ~5% of the sequences associated with reactivity to non-self-associated antigens showed similar publicness. Moreover, the CDR3 sequences associated with autoimmunity, cancer, and allograft-annotated immunity showed similar characteristics to those we identified in

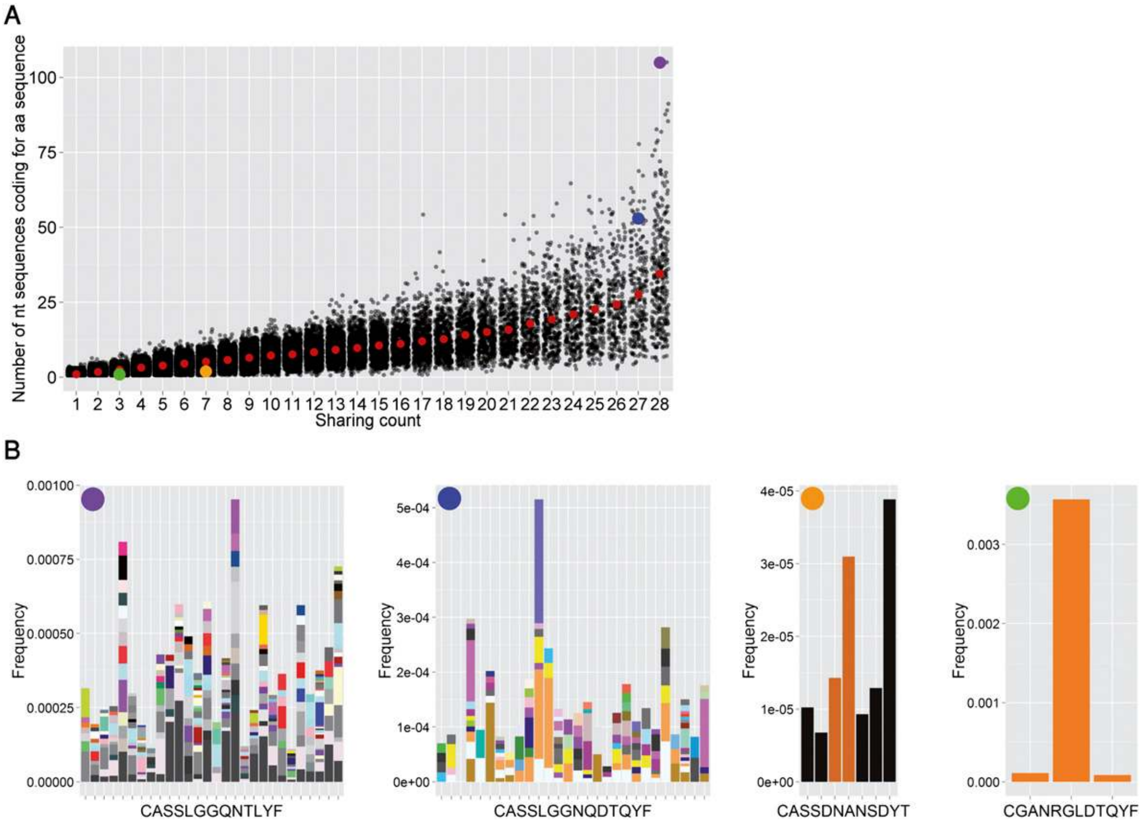

Figure 2. Convergent recombination continuously increases as a function of sharing. (*A*) Convergent recombination of all CDR3 sequences as a function of sharing level. Each black dot represents the total number of nt sequences coding for the same CDR3 aa sequence (summed across all mice in which this sequence is found). The red dots show the mean value for each sharing category. The colored dots refer to the four sequences shown in *B*. (*B*) The frequencies of nt sequences that encode four selected CDR3 aa sequences: CASSLGGQNTLYF (purple, found in $n = 28$ mice); CASSLGGNQDTQYF (blue, $n = 27$); CASSDNANSDYT (orange, $n = 7$); and CGANRGLDTQYF (green, $n = 3$). Sequences are marked also in *A* and in Figure 1C. Each color in the bars of each panel represents a different nt sequence.

the more public sequences: a higher mean frequency (Fig. 4C, $P = 1.16 \times 10^{-12}$, $P = 3.9 \times 10^{-6}$, and $P = 1.6 \times 10^{-7}$, respectively); a higher mean degree of convergent recombination ($P < 2.2 \times 10^{-16}$, $P = 4.1 \times 10^{-8}$, and $P = 4.5 \times 10^{-9}$, respectively); and fewer nt insertions (Fig. 4D, $P = 4.3 \times 10^{-10}$, $P = 1 \times 10^{-8}$, and $P = 6.3 \times 10^{-8}$, respectively), compared with the anti-pathogen-related sequences.

Simulations of TCR recombination support a combined role for recombination biases and T-cell selection in shaping TCR sharing

To further examine the forces that generate public sequences, we created in silico TCR repertoires using computer simulations. In these simulations, random TCR nt sequences are generated, where the probability for generating each sequence depended on the probability distributions of its different features that were evaluated from our actual data. The features used for the simulations are V and J gene segment usage, number of deletions in V and J, number of deletions from both sides of D, number of insertions in each junction (V-D, D-J), and nucleotide insertion bias in each junction. In a modified set of simulations, the entire junction region (between the 3′ end of the V region and the 5′ end of the J region) was simulated using position-dependent probabilities for each nucleotide as estimated from our data, without relying on D segment assignment. The simulations generated convergent recombinations emerging from two sources: The same CDR3 nt sequence can be generated by different recombination events, and a number of CDR3 nt sequences can encode the same CDR3 aa sequence (Venturi et al. 2008b). All probabilities used for the simulations were estimated either from in-frame reads or from out-of-frame reads that accounted for ~7% of the sequences in our data set. The out-of-frame sequences represent nonproductive TCR rearrangements that do not lead to receptor expression, and as a result cannot be influenced by antigen-induced selection. Simulations based on these different models and assumptions gave similar results (see Methods for full details of the assumptions used in the simulations and their realization).

We used these measured biases to create 28 repertoires of randomly generated nucleotide sequences, matching the sample size in the experimental data. We then analyzed these 28 virtual repertoires for patterns of sequence sharing in comparison with the experimental data. The simulations show a general trend that is similar to that observed experimentally, where the number of shared sequences gradually decreased with sharing level (Fig. 5A). Of note, a large number of sequences ($n = 630$) were found to be public in the simulated repertoires, somewhat higher than the number observed experimentally. In addition, simulated public CDR3 sequences manifested a higher level of convergent recombination compared with private sequences in the simulation, with a general trend similar to that observed experimentally

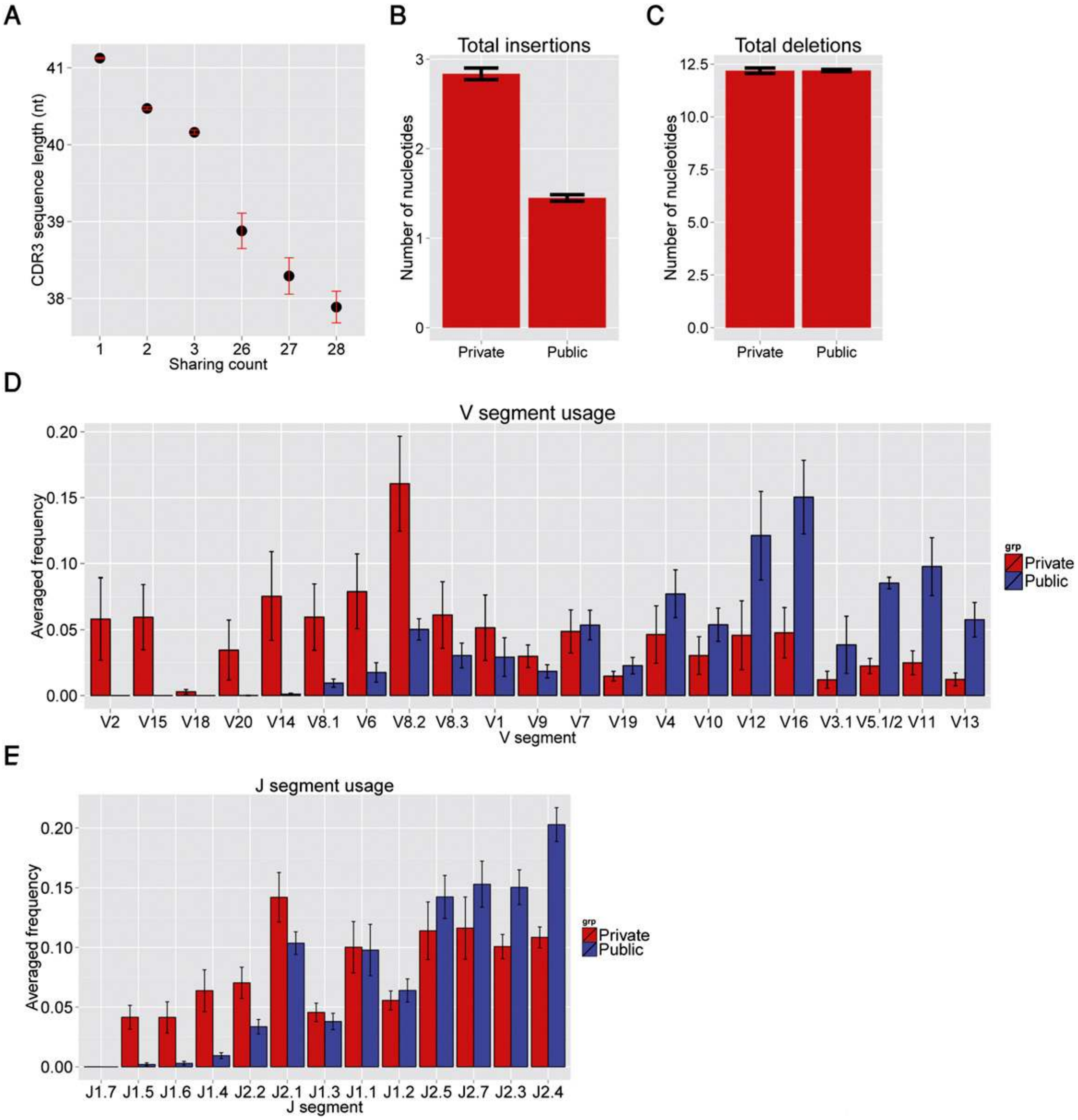

Figure 3. Public CDR3 sequences differ from private sequences in a number of characteristics. (*A*) The average length (nt) of private (shared by one to three mice) and public (shared by 26–28 mice) CDR3 sequences. Error bars, SE ($P < 8.5 \times 10^{-8}$, comparing the mean length of CDR3 nt sequences shared by $n = 3$ and by $n = 26$ mice). (*B*,*C*) The mean number of nt insertions (*B*, $P < 2.2 \times 10^{-16}$) and nt deletions (*C*, no significant difference between groups) summed over the V-D and D-J junctions, in private and public sequences. Error bars, SE. (*D*) Frequencies of V segments in private (red) and public (blue) sequences. (*E*) Frequencies of J segments in private (red) and public (blue) sequences. Segments in *D* and *E* are ordered by the ratio of their frequencies in private vs. public sequences. Error bars, SD. (All V segments except V7 and all J segments except J1.1 and J1.7 showed significant difference $P < 0.05$).

(Supplemental Fig. S3). As our simulations account only for properties of the DNA rearrangement process, and do not include T-cell selection, these results suggest that recombination biases and convergent recombination could play important roles in shaping the observed pattern of CDR3 sharing.

Further comparisons of CDR3 sharing revealed only a modest overlap between the simulated and experimentally obtained public CDR3 aa sequences; most sequences that were found to be public in the simulation were not public in the actual data and vice-versa (Fig. 5B). In contrast, most public CDR3 aa sequences found in one run of the simulation were also public in an independent random run (Fig. 5B). This observation prompted us to compare sequence sharing between simulation and experiment globally, in a sequence-specific way. Most sequences preserved very similar sharing scores in two independent runs of the simulation (Fig. 5C), which is also manifested by the high level of correlation in sharing scores ($R^2 = 0.91$, Pearson's correlation coefficient). In contrast, comparison of simulation with experimental data showed a much wider spread (Fig. 5D). In extreme cases, sequences that were private in the simulation were public in the data, while some public simulated sequences were private in the data. The correlation of sharing levels between data and simulation was significantly lower than that observed between two runs of the simulation ($R^2 = 0.55$).

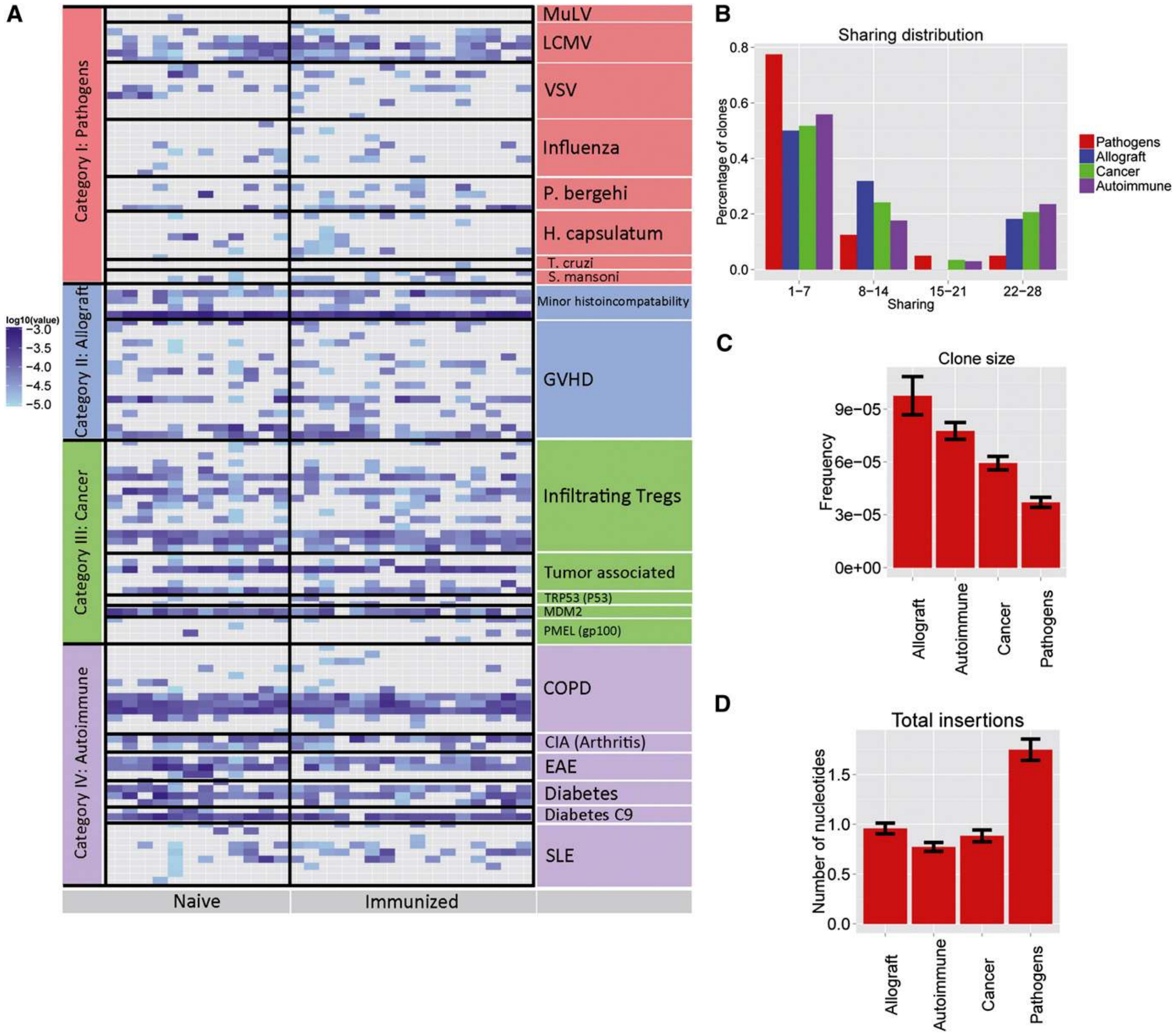

Figure 4. Annotated CDR3 sequences associated with self-related antigens feature a high level of sharing. (*A*) Frequencies of 124 annotated TCR sequences that were found in our data set. The frequency of each sequence (rows) in each mouse in our data set (columns) is shown ($\log_{10}$ scale, color bar on the *left*; gray color: sequence not found). Annotated sequences are grouped into four functional categories, according to the model in which they were detected (see Supplemental Table S2). (*B*) Sharing distributions of the annotated sequences, according to their functional category. (*C,D*) Comparison of annotated CDR3 aa sequences of the different functional categories, in terms of mean frequency in the 28 tested mice (*C*) and mean number of nt insertions in the VD and DJ junctions (*D*).

We used the consistent sharing level obtained in the simulations to define three groups of CDR3 sequences: ~93% of the sequences differed by 10 or less in their sharing level between data and simulation ($|\Delta S_{data-sim}| \leq 10$). However, 6% of the CDR3 aa sequences were shared in the data at significantly lower levels than those predicted by the simulation (red region in Fig. 5D, $\Delta S_{data-sim} < -10$), and 1.15% were shared at significantly higher levels in the data (blue region in Fig. 5D, $\Delta S_{data-sim} > 10$). Specific examples for sequences of the three groups are shown in Figure 5E, where the observed sharing level of a CDR3 aa sequence in the data is compared with its simulated sharing level in 100 runs of the simulation. These results show that some specific CDR3 sequences are shared at significantly higher (or lower) levels than expected by our null model that is based on recombination biases; this finding supports a role for clonal selection in modulating the sharing level and publicness of specific CDR3 sequences.

Public CDR3 sequences are found across MHC haplotypes but differ in their V-segment usage

As noted above, the annotated sequences were derived from various mouse strains that differ in their MHC haplotypes. To further explore the MHC restrictions of the public sequences, we used TCR-seq to map the repertoires of T cells interacting with different

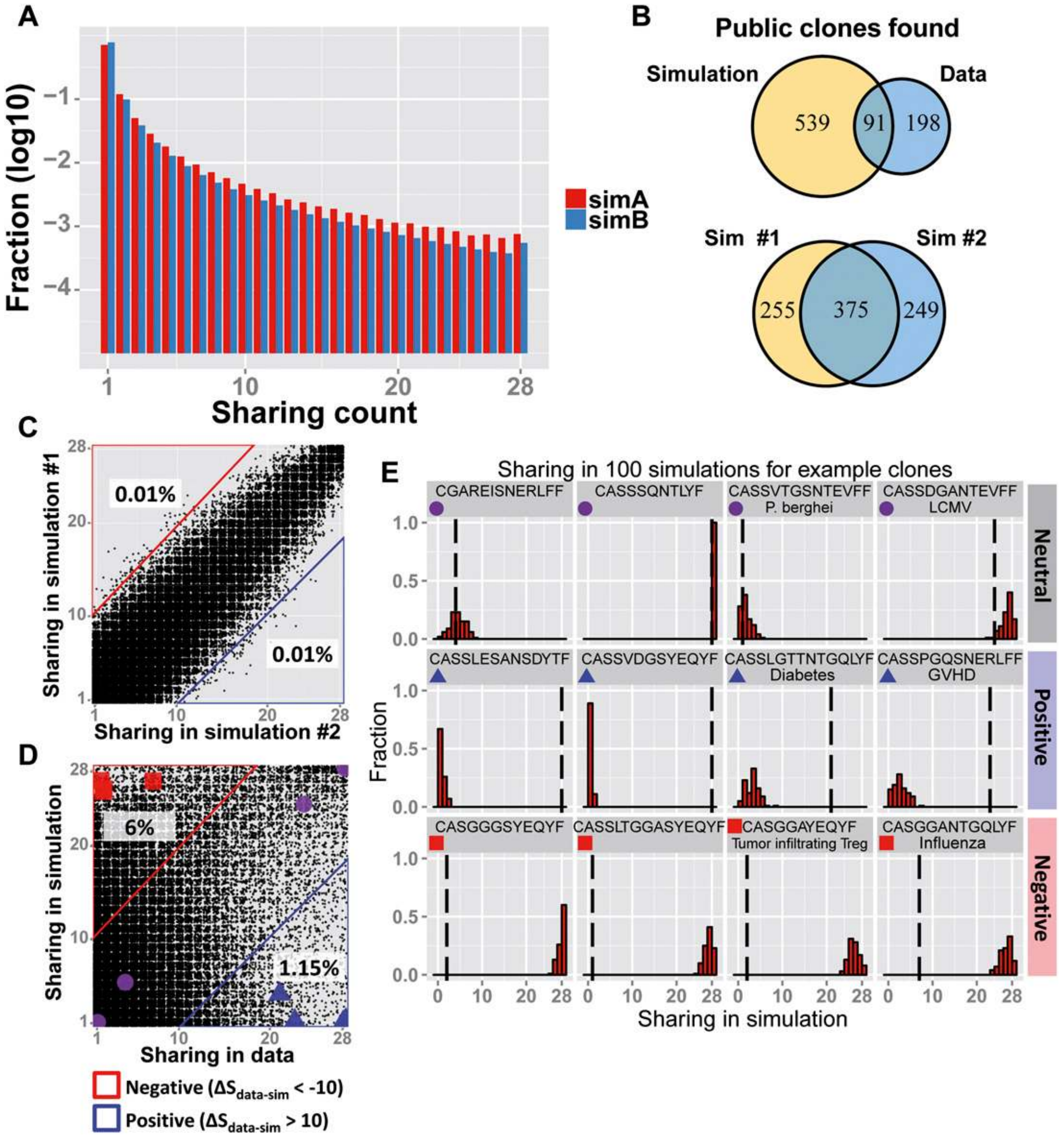

Figure 5 Simulations of the VDJ recombination process provide a measure for the impact of biases on sharing levels. (*A*) The fraction of CDR3 aa sequences found in each sharing category, in two implementations of simulation of the biased VDJ recombination process (see Methods for details). Both simulation A (red) and simulation B (blue) show a similar trend to that observed in the data (Fig. 1B), where 0.08% and 0.05% of CDR3 sequences are public, respectively. (*B*) A modest overlap of public CDR3-types is found between simulation and data, with only 91 CDR3 aa sequences found to be public in both (*top* plot). Higher overlap in public sequences exists between independent iterations of the simulation (*bottom*). (*C*) Sharing is well correlated between two independent iterations of simulation A. Only a very small number of CDR3 sequences differ by more than 10 in their sharing level between the two runs of the simulation. (*D*) A comparison of sharing level between simulation (*y*-axis) and data (*x*-axis). Each dot represents one CDR3 aa sequence. (Red region) 6% of the sequences show much lower sharing in the data ($\Delta S_{data-sim} < -10$). (Blue region) 1.15% of the sequences show much higher sharing in the data ($\Delta S_{data-sim} > 10$). (*E*) Measured (vertical dashed line) and simulated (red bars, histograms of 100 random runs of simulation A) sharing levels of four selected CDR3 aa sequences from each region in *D*. (*Top*) Neutral, showing similar sharing in simulations and data (purple circles in *D*, $|\Delta S_{data-sim}| \leq 10$); (*middle*) positive, showing much higher sharing in data vs. simulation (blue triangles in *D*, $\Delta S_{data-sim} > 10$); (*bottom*) negative (red squares in *D*, $\Delta S_{data-sim} < -10$) showing much lower sharing in the data. Two sequences in each row are taken from the annotated group of sequences shown in Figure 4.

MHC molecules: C57BL/6 CD8$^+$ T cells (which are restricted by MHC class-I H2^b); C3H.SW CD4$^+$ T cells (which have the H2^b MHC allele, but a different genetic background than C57BL/6); and C3H. HeSnJ CD4$^+$ T cells (which are congenic with the C3H.SW strain, but bear the H2^k allele). These repertories were compared with those of the MHC-II H2^b-restricted CD4$^+$ T cells of the C57BL/6 mice. We found that >82% of the 289 public CDR3 aa sequences were also present in the other T-cell repertories (Supplemental Fig. S4). Interestingly, most of these public CDR3 sequences were associated with several different V region gene segments. Moreover, the V gene segments associated with each shared CDR3 aa sequence tended to differ between the different MHC-restricted T-cell groups—see Figure 6A and Supplemental Figure S5. A global analysis of the degree of similarity in V-segment usage of the public CDR3 sequences between C57BL/6 (H2^b) CD4$^+$ T cells and the other T-cell groups showed that differences in MHC restriction were associated

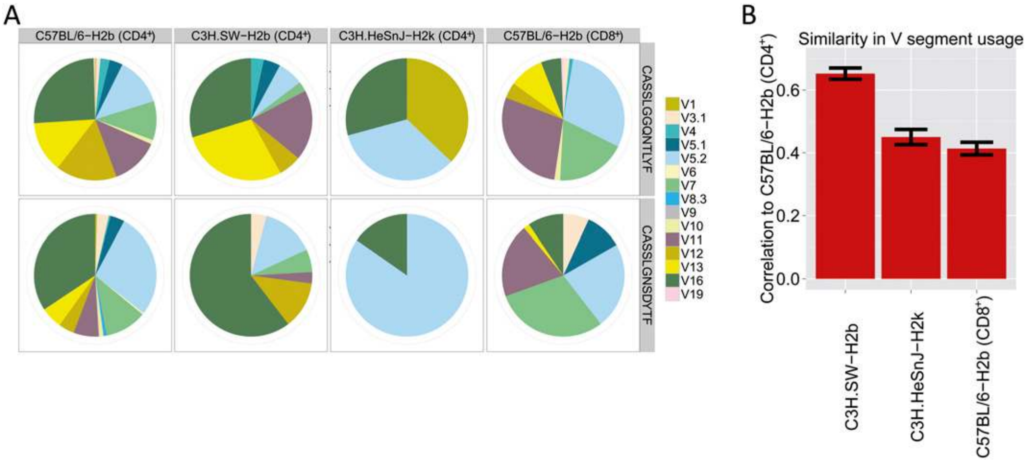

Figure 6. Differences in MHC restriction are associated with more diverse V gene usage. (*A*) V segment usage of two public CDR3 aa sequences. Weighted mean frequency of V segment usage for each sequence was calculated for different MHC-restricted T-cell groups: CD4$^+$ T cells from C57BL/6 mice (*left*, MHC-II, H2^b haplotype, $n = 28$); CD4$^+$ T cells from C3H.SW mice (second *left*, H2^b, $n = 3$); CD4$^+$ T cells of C3H.HeSnJ mice (second *right*, H2^k, $n = 2$); and CD8$^+$ T cells from C57BL/6 mice (*right*, MHC-I, H2^b, $n = 2$). V segments are indicated by the color bar on the *right*. (*B*) Average correlations in V segment usage calculated between all 289 public CDR3 aa sequences in our C57BL/6 H2^b strain data set and the three other MHC-restricted T-cell groups as in *A*. Error bars, SEM (see Supplemental Material for details).

with more diverse V gene usage (Fig. 6B; see Supplemental Material for details). For example, the C3H.SW strain bearing the H2^b haplotype showed a pattern of V segment usage that was more similar to the C57BL/6 (also a H2^b haplotype) than was the C3H. HeSnJ strain that carries the H2^k haplotype. Note that our data set does not include TCRA chain sequences, which also must be taken into account in MHC restriction. Nevertheless, these data imply that V segment usage by public T-cell CDR3 aa segments is dominated by the MHC haplotype to a greater extent than by the non-MHC genetic background.

Discussion

How might we explain the high level of convergent recombination of public sequences together with their greater abundance relative to the more private sequences? Our present knowledge of TCR repertoire development would suggest that these differences could result mainly from two mechanisms: (1) biases in the recombination process that favor the generation of certain sequences, which renders them more abundant and more public (Venturi et al. 2006; Ndifon et al. 2012); and (2) different degrees of positive selection by particular antigen epitopes, such that the more public CDR3 aa types would enjoy a selective advantage, particularly in the process of tonic stimulation needed to preserve TCR repertoires in the periphery (Ernst et al. 1999; Hochweller et al. 2010), which is where we sampled them. Negative selection could also shape the observed repertoire, leading to loss of potentially public sequences. The two mechanisms can function together: Recombination biases ensure the initial presence of certain public TCR sequences in different individuals, and positive and negative selection modulate the maintenance or loss of specific CDR3 sequences. Results of our simulations support a combination of both mechanisms. In general, the simulations suggest that recombination bias and convergent recombination can by themselves generate public sequences, and at a frequency similar or larger than that found in the data. However, we found heterogeneous behaviors when evaluating specific CDR3 sequences. Some sequences were as public in the data as predicted by the null model, whereas others were much more public or much more private than predicted by the null model. This suggests that positive and negative selection further shape patterns of sequence sharing and publicness, beyond what can be expected based on recombination biases and convergent recombination alone. The comparison with the null model can be used in future studies to focus on specific sequences that are subject to strong negative or positive selection. Involvement of selection in the generation of public clones is also supported by their enrichment for self-associated (but not pathogen-associated) antigens, and by the variable pattern of V-segment usage of public CDR3 sequences across MHC haplotypes.

The observed impact of MHC haplotype on TCR V-region usage with less effect on CDR3 segment usage (Fig. 6; Supplemental Fig. S5) is compatible with structural studies, which show that the CDR1 and CDR2 segments of the TCR, which are expressed on the V-gene segments of the TCR outside the CDR3 region, interact directly with the MHC molecule (Huseby et al. 2005; Rudolph et al. 2006). Shared CDR3 segments with variable V regions are not limited to the TCR; CDR3 regions of the immunoglobulin heavy chain associated with different V segments have also been reported in the B-cell receptor (BCR) antibody responses of humans to dengue virus (Parameswaran et al. 2013).

The novel finding of annotated CDR3 sequences in our data set of healthy mice (Fig. 4A; Supplemental Table S2) highlights a functional difference between the more private TCR sequences, which we found to be associated with all classes of antigens, and the more public sequences, which appear to be associated mainly with autoimmune conditions, allograft reactions, and tumor infiltration (Fig. 4). The standard clonal selection paradigm of adaptive immunity (Burnet 1976) would predict that T cells expressing TCRs capable of binding to self-antigens must be deleted during development, most likely in the thymus. Yet, as we see here, a set of autoimmune CDR3 aa sequences are commonly shared and even appear to be amplified with high frequency and convergent recombination. The high convergent recombination and high frequency of self-associated public CDR3-types could be explained by frequent encounters with self antigens, which are continuously present in the body, while foreign viral antigens are encountered only as a result of sporadic infection or immunization. In other words, private and public CDR3-types might express the degree and dynamics of their contact with cognate antigens subsequent to any bias in nt recombination in the thymus.

Shared autoreactivity is not limited to T-cell CDR3-types; there also appears to be an enrichment in self-reactive, autoantibodies present from birth in the healthy human antibody repertoire. We have previously found that humans, during their development in the uterus, produce a shared set of IgM and IgA autoantibody reactivities that bind to shared self-antigens (Merbl et al. 2007; Madi et al. 2009). Thus, both the TCR repertoire, shown here for CDR3 aa sequences, and the antibody repertoire, shown elsewhere for self-antigen binding, feature healthy autoimmunity that is highly shared.

Some yet to be discovered advantage must account for the high frequency of public CDR3 TCR aa sequences associated with autoimmunity; it is not likely that the prevalence of such sequences in healthy subjects has evolved to cause autoimmune disease. The example of the C9 CDR3 sequence discussed above suggests that some public TCR specificities may play a pivotal role in "controlling" autoimmunity (Tikochinski et al. 1999); hence, it is conceivable that modulating such public CDR3-types might provide a new therapeutic approach to modulating autoimmune disease. TCR diversity has been an obstacle for treatments such as T-cell vaccination based on specific TCR sequences (Cohen 2001), which might be alleviated if public TCRs can be used as effective T-cell vaccines.

The immunological homunculus theory (Cohen 1992, 2000) is based on the idea that the healthy immune system helps maintain the organism and regulate inflammation by encoding in healthy lymphocyte repertoires a functionally useful immune image of the body. The present discovery of public CDR3 aa sequences that recognize self-antigens points to a molecular basis for the immunological homunculus concept, and suggests questions for revisiting the basic processes of repertoire selection. Clinical applications could be in the offing.

Methods

Mice

Female 5- to 8-wk-old C57BL/6, C3H.HeSnJ, and C3H.SW mice were obtained from Harlan Laboratories. All mice were housed at the Weizmann Institute, in compliance with national and international regulations.

$CD4^+$ $CD8^+$ cell purification

$CD4^+$ or $CD8^+$ T cells were purified from splenocytes by magnetic bead separation. Cells were precipitated and lysed for RNA purification.

Immunization

Nine mice were injected intraperitonealy (IP) with 100 μg 1:1 ratio of Chicken Ovalbumin emulsified in CFA and seven mice with CFA only; 12 mice were not immunized. Three spleens from each immunized group were harvested on days 5 and 14 after immunization. Three of the OVA immunized mice received an additional 100 μg OVA/CFA boost injection IP on day 14 and spleens were harvested on day 60 together with one additional CFA-only immunized mouse.

Library preparation for TCR-seq and data preprocessing

Libraries were prepared and preprocessed as published (Ndifon et al. 2012). Briefly, total RNA was extracted from T cells and reverse transcribed using a TCR Cβ-specific primer linked to the 3′-end Illumina sequencing adapter. cDNA was then amplified using PCR with a Cβ-3′adp primer and a set of Vβ-specific 5′ primers. Each Vβ-specific primer included a restriction-site sequence for the AcuI restriction enzyme. PCR products were digested with AcuI enzyme, followed by ligation of 5′Illumina adaptor. Finally, products were amplified by PCR using universal primers for the 5′ and 3′ llumina adapters. The libraries were sequenced using Genome Analyzer II or HiSeq 2000 (Illumina). Sequence filtering, VDJ annotation, normalization, and translation were performed as published (Ndifon et al. 2012). See Supplemental Material for additional details.

Generating in silico TCR repertoires by simulation of the VDJ recombination process

In silico TCR repertoires were generated by computer simulations that recapitulate the biases observed in the VDJ recombination data. Simulations were implemented in two complementary approaches. In Simulation A, biases used to generate the sequences were measured from the data, using unique counts of in-frame reads, to estimate the distributions of the following properties: V and J segment usage; and, for both junctions (V-D, D-J), the number of deletions in each end of the junction, number of insertions, and the TdT nucleotide insertion bias. Biases were assumed as independent for this model. In Simulation B, biases were measured from the raw data, processed by Decombinator (Thomas et al. 2013). The following generation parameters were learned from the nonfunctional (out of frame) CDR3 sequences: V and J genes usage, conditional probabilities for deletions in each gene, and junctional nucleotide probabilities, including junction length. For both simulations, only generated nucleotide sequences determined as in-frame CDR3 were used, and all 28 virtual repertoires were simulated according to their respective experimental sample sizes.

Data access

The sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP042610.

Acknowledgments

We thank Tzachi Pilpel for helpful comments on the manuscript. This research was supported by grants from the Krenter Institute, the Minerva Foundation with funding from the Federal German Ministry for Education and Research, the Weizmann-UK program, and the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation. N.F. is the incumbent of the Pauline Recanati Career Development Chair of Immunology. A.M. was supported by the M.D. Moross Institute for Cancer Research.

References

Argaet VP, Schmidt CW, Burrows SR, Silins SL, Kurilla MG, Doolan DL, Suhrbier A, Moss DJ, Kieff E, Sculley TB, et al. 1994. Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus. *J Exp Med* **180:** 2335–2340.

Bousso P, Casrouge A, Altman JD, Haury M, Kanellopoulos J, Abastado JP, Kourilsky P. 1998. Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* **9:** 169–178.

Burnet FM. 1976. A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA Cancer J Clin* **26:** 119–121.

Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. 2000. Size estimate of the αβ TCR repertoire of naive mouse splenocytes. *J Immunol* **164:** 5782–5787.

Cohen IR. 1992. The cognitive principle challenges clonal selection. *Immunol Today* **13:** 441–444.

Cohen IR. 2000. *Tending Adam's garden: evolving the cognitive immune self.* Academic Press, London, UK.

Cohen IR. 2001. T-cell vaccination for autoimmune disease: a panorama. *Vaccine* **20:** 706–710.

Davis MM, Bjorkman PJ. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature* **334:** 395–402.

Elias D, Cohen IR. 1994. Peptide therapy for diabetes in NOD mice. *Lancet* **343:** 704–706.

Elias D, Tikochinski Y, Frankel G, Cohen IR. 1999. Regulation of NOD mouse autoimmune diabetes by T cells that recognize a TCR CDR3 peptide. *Int Immunol* **11:** 957–966.

Ernst B, Lee DS, Chang JM, Sprent J, Surh CD. 1999. The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. *Immunity* **11:** 173–181.

Hochweller K, Wabnitz GH, Samstag Y, Suffner J, Hammerling GJ, Garbi N. 2010. Dendritic cells control T cell tonic signaling required for responsiveness to foreign antigen. *Proc Natl Acad Sci* **107:** 5931–5936.

Huseby ES, White J, Crawford F, Vass T, Becker D, Pinilla C, Marrack P, Kappler JW. 2005. How the T cell repertoire becomes peptide and MHC specific. *Cell* **122:** 247–260.

Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, Cao Y, Liu X, Douek DC, Price DA, et al. 2012. Recombinatorial biases and convergent recombination determine interindividual η sharing in murine thymocytes. *J Immunol* **189:** 2404–2413.

Madi A, Hecht I, Bransburg-Zabary S, Merbl Y, Pick A, Zucker-Toledano M, Quintana FJ, Tauber AI, Cohen IR, Ben-Jacob E. 2009. Organization of the autoantibody repertoire in healthy newborns and adults revealed by system level informatics of antigen microarray data. *Proc Natl Acad Sci* **106:** 14484–14489.

Merbl Y, Zucker-Toledano M, Quintana FJ, Cohen IR. 2007. Newborn humans manifest autoantibodies to defined self molecules detected by antigen microarray informatics. *J Clin Invest* **117:** 712–718.

Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, Bodmer HC, McMichael AJ, Bell JI. 1991. Extensive conservation of α and β chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. *Proc Natl Acad Sci* **88:** 8987–8990.

Murugan A, Mora T, Walczak AM, Callan CG Jr. 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci* **109:** 16161–16166.

Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, Reich-Zeliger S, Arnon R, Friedman N. 2012. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci* **109:** 15865–15870.

Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, Artiles KL, Zompi S, Vargas MJ, Simen BB, et al. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* **13:** 691–700.

Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, Douek DC, Davenport MP, Price DA. 2010. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci* **107:** 19414–19419.

Raz I, Elias D, Avron A, Tamir M, Metzger M, Cohen IR. 2001. β-cell function in new-onset type 1 diabetes and immunomodulation with a heat-shock protein peptide (DiaPep277): a randomised, double-blind, phase II trial. *Lancet* **358:** 1749–1753.

Raz I, Ziegler AG, Linn T, Schernthaner G, Bonnici F, Distiller LA, Giordano C, Giorgino F, de Vries L, Mauricio D, et al. 2014. Treatment of recent-onset Type 1 diabetic patients with DiaPep277: results of a double-blind, placebo-controlled, randomized phase 3 trial. *Diabetes Care* **37:** 1392–1400.

Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH. 2010. Overlap and effective size of the human CD8$^+$ T cell receptor repertoire. *Sci Transl Med* **2:** 47ra64.

Rudolph MG, Stanfield RL, Wilson IA. 2006. How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* **24:** 419–466.

Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. 2013. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* **29:** 542–550.

Tikochinski Y, Elias D, Steeg C, Marcus H, Kantorowitz M, Reshef T, Ablamunits V, Cohen IR, Friedmann A. 1999. A shared TCR CDR3 sequence in NOD mouse autoimmune diabetes. *Int Immunol* **11:** 951–956.

Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, Davenport MP. 2006. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci* **103:** 18691–18696.

Venturi V, Chin HY, Price DA, Douek DC, Davenport MP. 2008a. The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8$^+$ TCRs between macaques. *J Immunol* **181:** 2597–2609.

Venturi V, Price DA, Douek DC, Davenport MP. 2008b. The molecular basis for public T-cell responses? *Nat Rev Immunol* **8:** 231–238.

Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, Asher TE, Almeida JR, Levy S, Price DA, et al. 2011. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* **186:** 4285–4294.

Received December 6, 2013; accepted in revised form July 8, 2014.

CSH PRESS

GENOME RESEARCH

T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity

Asaf Madi, Eric Shifrut, Shlomit Reich-Zeliger, et al.

Genome Res. 2014 24: 1603-1612 originally published online July 14, 2014
Access the most recent version at doi:10.1101/gr.170753.113