Chromatin conformation governs T-cell receptor Jβ gene segment usage

Wilfred Ndifon[1], Hilah Gal[1], Eric Shifrut, Rina Aharoni, Nissan Yissachar, Nir Waysbort, Shlomit Reich-Zeliger, Ruth Arnon, and Nir Friedman[2]

Department of Immunology, Weizmann Institute of Science, Rehovot, 76100 Israel

Edited* by Michael Sela, Weizmann Institute of Science, Rehovot, Israel, and approved August 13, 2012 (received for review March 6, 2012)

T cells play fundamental roles in adaptive immunity, relying on a diverse repertoire of T-cell receptor (TCR) α and β chains. Diversity of the TCR β chain is generated in part by a random yet intrinsically biased combinatorial rearrangement of variable (Vβ), diversity (Dβ), and joining (Jβ) gene segments. The mechanisms that determine biases in gene segment use remain unclear. Here we show, using a high-throughput TCR sequencing approach, that a physical model of chromatin conformation at the DJβ genomic locus explains more than 80% of the biases in Jβ use that we measured in murine T cells. This model also predicts correctly how differences in intersegment genomic distances between humans and mice translate into differences in Jβ bias between TCR repertoires of these two species. As a consequence of these structural and other biases, TCR sequences are produced with different a priori frequencies, thus affecting their probability of becoming public TCRs that are shared among individuals. Surprisingly, we find that many more TCR sequences are shared among all five mice we studied than among only subgroups of three or four mice. We derive a necessary mathematical condition explaining this finding, which indicates that the TCR repertoire contains a core set of receptor sequences that are highly abundant among individuals, if their a priori probability of being produced by the recombination process is higher than a defined threshold. Our results provide evidence for an expanded role of chromatin conformation in VDJ rearrangement, from control of gene accessibility to precise determination of gene segment use.

lymphocyte receptor repertoires | public T-cell clones | VDJ recombination | epigenetics | next generation sequencing

A large diversity of T-cell receptor (TCR) αβ chains is essential for reliable antigen recognition and for proper functioning of the adaptive immune system (1, 2). The TCR interacts with a wide array of antigens bound to major histocompatibility complex (MHC) molecules displayed on the surface of cells (2). TCR diversity is generated by random rearrangement of the variable (Vα) and joining (Jα), and the variable (Vβ), diversity (Dβ), and joining (Jβ), gene segments of the TCR α and β chains, respectively (1) (*SI Appendix*, Fig. S1*A*). Diversity is further increased by nucleotide insertions and deletions at the junctions between pairs of rearranging genes, forming the highly variable complementarity determining region 3 (CDR3) that is directly implicated in antigen recognition. These processes result in a huge potential TCR diversity, with as many as 10^{15} distinct clonotypes estimated to be realizable in the mouse TCR αβ repertoire (2).

Many studies published over the past 20 y indicate that the TCR repertoire is biased—i.e., not all potential sequences are found with the same probability. These biases can stem from properties of the gene rearrangement process, as well as from thymic selection and the expansion of T-cell clones. Here, we focus on biases that result from the gene rearrangement process, in which different TCR sequences are produced with different a priori probabilities. In general, bias in TCR sequences is evident in the unequal frequencies of gene segments observed, and also in the biased number of nucleotides inserted/deleted at junction regions. This bias leads to characteristic Gaussian distributions of CDR3 region lengths (3). In addition, many TCR sequences can be produced in different ways, both through different recombination events and through convergence of different rearranged nucleotide sequences to the same encoded amino acid sequence (4). These processes, collectively termed "convergent recombination" (4), further contribute to increased a priori probabilities of producing specific TCR sequences. The mechanistic rules controlling these biases in TCR production frequency are still largely unknown (5). Understanding of such rules can be advanced by analyzing larger datasets to obtain better statistics on the structure of repertoires, and also by comparing repertoires of different individuals. Such interindividual comparisons can help to distinguish between general features of repertoire structure, such as those generated by biases in gene rearrangement, and those features that are specific to individuals due to their own immune history. This reasoning led us to use high-throughput sequencing to map the TCR β chain (TCRB) repertoire of a number of individual inbred mice, to better understand biases in gene rearrangement and investigate the potential mechanisms that generate these biases.

Inbred mouse strains can serve as useful models for studying biases in the gene rearrangement process of lymphocyte receptors, while controlling for effects of other factors that shape the repertoire. First, identical genetic backgrounds ensure that all individual mice are homozygous for identical VDJ gene alleles and are thus easier to compare. Moreover, because all individuals have the same MHC alleles, thymic selection is also expected to have similar effects on the repertoires of different individuals. Second, using young mice that grow in clean conditions reduces effects of exposure to infections on the structure of the repertoire. Finally, it is possible to obtain a large number of lymphocytes, which allows for sorting of subpopulations such as $CD4^+$ or $CD8^+$ T cells, thus obtaining results that are less affected by potential differences in the repertoires of those different cell subpopulations.

The recent emergence of high-throughput sequencing technologies has revolutionized the analysis of lymphocyte repertoires. While traditional methods for assessment of the T-cell repertoire, such as spectratyping and Sanger sequencing (3, 6), provide only limited information about variable CDR3 length and diversity, these new high-throughput methods enable parallel sequencing of millions of short DNA sequences (7), providing a unique opportunity to map immune repertoires at a very high resolution (8). Deep sequencing technologies have already been applied to study antibody repertoires (8, 9) and, more recently, to define the full spectrum of β chains found in both the naïve and the antigen-experienced human T-cell repertoires (10–13).

Author contributions: W.N., H.G., R. Aharoni, R. Arnon, and N.F. designed research; W.N., H.G., R. Aharoni, and S.R.-Z. performed research; W.N., N.Y. and N.W. contributed new reagents/analytic tools; W.N., H.G., and E.S. analyzed data; and W.N. and N.F. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequence data reported in this paper has been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra (accession no. SRA057715).

[1]W.N. and H.G. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: nir.friedman@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1203916109/-/DCSupplemental.

Here, we describe our development and application of an experimental and computational approach for characterizing the TCR repertoire based on massively parallel sequencing (TCR-seq). Using this approach, we mapped with high resolution the TCRB repertoires of individual C57BL/6 mice, aiming to reveal general organizing principles that affect repertoire biases.

Results

Biases in Gene Segment Use Are Similar Between Individual Inbred Mice. We developed an affordable experimental and computational TCR-seq approach (*SI Appendix*, Fig. S1 *B* and *C*; *Methods*) for characterizing the TCR repertoire. We applied this approach to sequence the rearranged TCRB CDR3 region of splenic $CD4^+$ and $CD8^+$ T cells of seven individual C57BL/6 mice, and obtained $\sim 10^7$ total sequence reads, $\sim 4 \times 10^5$ of which we defined as unique T-cell clonotypes (*SI Appendix*, Table S1 and *Methods*). As we look for biases in the rearrangement process itself, we analyzed both unique in-frame ("selected") and unique out-of-frame ("unselected") clonotype sequences. The latter sequences contain one or more projected stop codons within their Vβ/Dβ/Jβ sequence, and are likely to represent failed rearrangements in a cell that had successfully rearranged the second TCRB allele (14). The frequency of clonotypes containing a stop codon is ~2% in our data (excluding sequences with Vβ17, which has a stop codon in its germline sequence in this mouse strain) (15, 16). This value is within the range of stop codon-containing TCRB transcripts measured in recent studies of human TCRB repertoires by means of high-throughput sequencing (11, 13). Thus, these out-of-frame sequences are assumed to represent the original landscape of rearranged CDR3 regions, without biases due to thymic selection (15, 16). Analysis of unique sequences (regardless of their copy number) minimizes effects of clonal expansion on repertoire statistics.

We start by analyzing the frequencies of individual Vβ and Jβ genes measured in $CD4^+$ T cells. We find that the frequencies measured in selected clonotypes from different mice are very similar (Fig. 1 *A*–*C*), suggesting that these frequencies might be determined by common organizing principles. Similarity in Vβ and Jβ gene frequencies between individuals was also observed in previous analyses of human T cells (12, 13). We further characterize common biases in gene use found in our data. We find that the measured gene frequencies vary widely. For example, several genes (e.g., Vβ18) appear in <0.5% of unique clonotypes, whereas others (e.g., Vβ10) appear in >5% (Fig. 1 *A* and *B*). Focusing on the use of Jβ genes, we find that segments belonging to the first DJCβ cluster (Jβ 1.1–1.7) rearrange with Dβ1 34 ± 8 times more frequently than with Dβ2, whereas Jβ segments from the second DJCβ cluster (Jβ 2.1–2.7) rearrange only slightly less often with Dβ1 than with Dβ2. A similar pattern of Jβ use is also observed in unselected clonotypes. The "unconventional" pairings between Jβ 1.1–1.7 and Dβ2 were previously observed using both standard Sanger sequencing (17) and high-throughput sequencing of human T cells (12) (*SI Appendix, SI Text*, section 1). Thus, the frequencies of individual Vβ and Jβ genes are highly biased, as are the frequencies of Dβ-Jβ pairs.

Because the gene segment frequencies we measured were mostly based on short reads (of lengths 40 nt for datasets M1–M4 and M6–M8, and 80 nt for M5), we evaluated their accuracy using a simulated dataset of 10^5 TCRB sequences, with characteristics similar to our experimental data (*SI Appendix, SI Text*, section 4). The simulation results indicate that our measured Jβ frequencies are very accurate for both read lengths, whereas the

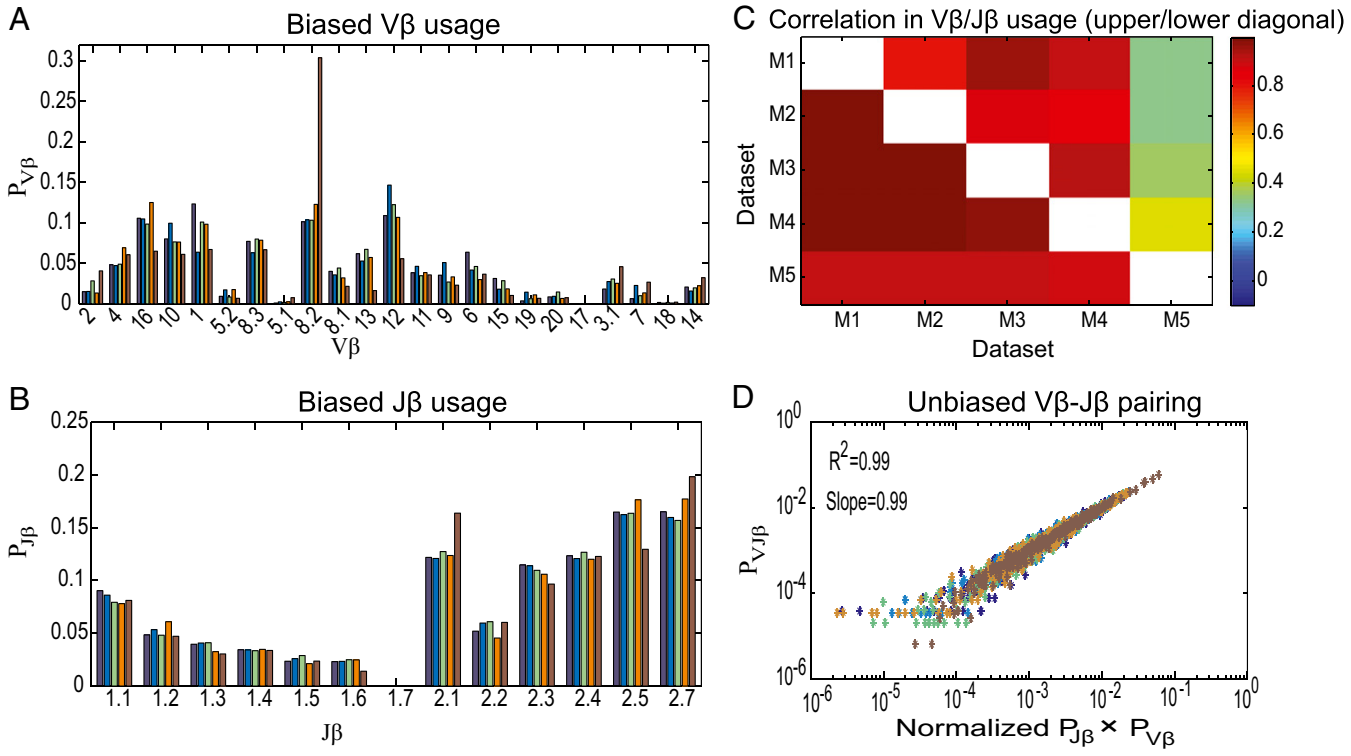

Fig. 1. TCRB repertoire sequencing reveals a biased Vβ and Jβ gene segment use that is similar between individual mice. (*A* and *B*) Measured frequencies of Vβ gene segments, $P_{Vβ}$ (*A*), and Jβ gene segments, $P_{Jβ}$ (*B*), in selected (in-frame) clonotypes from mice M1–M5. Vβ and Jβ gene segments are ordered according to their relative genomic positions in the mouse genome (18). (*C*) Squared correlation coefficients between measured frequencies of Vβ/Jβ gene segments in selected clonotypes, calculated for all possible pairs of mice M1–M5. (*D*) Vβ and Jβ gene frequencies are independent. The probability that a selected clonotype carries a particular pair of Vβ and Jβ ($P_{VJβ}$) is plotted vs. the normalized product of $P_{Vβ}$ and $P_{Jβ}$. Data from mice M1–M5. The average of $P_{VJβ}$ from different mice is not significantly different from the average of normalized $P_{Vβ} \times P_{Jβ}$ ($P = 0.62$, Wilcoxon signed-rank test), consistent with statistical independence of Vβ and Jβ frequencies. A linear fit to the data has a slope of 0.99, further supporting statistical independence.

Vβ frequencies have some inaccuracies, especially for a few Vβ segments that have a similar coding sequence at their 3′ end (*SI Appendix*, Figs. S2–S5). In particular, for 80-nt reads, the frequencies of Vβ 5.1/5.2, 8.1, and 16 differ from their expected values by >20% (*SI Appendix*, Fig. S3*A*). These results are supported by the high correlation we find between our measured Jβ frequencies and measurements made previously (15) using other methods ($R^2 = 0.96$), and also between the Vβ frequencies we measured by 80-nt TCR-seq vs. staining mouse splenic T cells using a panel of 15 Vβ-specific antibodies ($R^2 = 0.81$; *SI Appendix*, *SI Text*, section 1 and Table S2). Overall, these results suggest that our TCR-seq analysis quantifies very accurately Jβ frequencies, and it also provides good estimates of Vβ frequencies, with few exceptions.

The unprecedented resolution offered by TCR-seq allows for characterization of additional, less well understood features of the murine TCR repertoire. In particular, we analyzed the codistribution of Vβ and Jβ genes in unique clonotypes. We find that the frequencies of Vβ-Jβ pairs vary widely, with some pairs appearing ∼1,000× more frequently than others (Fig. 1*D*). Interestingly, the probability of finding a particular Vβ-Jβ pair is not significantly different from the probability calculated for this pair by assuming that Vβ and Jβ frequencies are independent; this holds for both selected (Fig. 1*D*; $P = 0.62$, Wilcoxon signed-rank test) and unselected (*SI Appendix*, Fig. S6*A*; $P = 0.67$, Wilcoxon signed-rank test) clonotypes. This observation suggests strongly that Vβ and Jβ frequencies are statistically independent, and is consistent with previous results showing similar Jβ frequencies in murine splenic T cells carrying a subset of different Vβ genes (19). The observed independence requires that the frequency of a particular Vβ paired with Dβ1 is not significantly different from the frequency of the same Vβ paired with Dβ2, which is supported by our data (*SI Appendix*, Fig. S6*B*; $P = 0.06$, Wilcoxon signed-rank test). Importantly, the fact that we can predict accurately the frequencies of all 299 possible Vβ-Jβ pairs using only the 36 individual Vβ and Jβ frequencies (Fig. 1*D*) indicates that the TCR Vβ-Jβ repertoire is much less complex than expected.

Mechanical Model of Chromatin Explains Observed Biases in Jβ-Dβ Pairing. Biases in Vβ/Jβ gene use have important implications for the effectiveness of T-cell–mediated immunity (5, 20, 21). Our finding that gene use biases measured in different mice are very similar (Fig. 1*A*–*C*) prompted us to investigate common organizing principles of those biases. Previous work suggested that the degree of sequence conservation of recombination signal sequences (RSSs) flanking individual TCR gene segments is correlated qualitatively with frequencies of murine Jβ genes (22), and quantitatively with rearrangement frequencies for extrachromosomal recombination substrates (23, 24). However, accurate quantitative prediction of rearrangement frequencies for chromosomal receptor genes, as observed in vivo, has not previously been possible. Examination of our sequencing data revealed a regular pattern relating frequencies of Dβ-Jβ gene pairs and the genomic distance between them. We hypothesized that mechanical properties of chromatin can generate the observed pattern by modulating the frequencies of random encounters between pairs of rearranging Dβ and Jβ genes. We tested this hypothesis by fitting to the measured Dβ-Jβ frequencies a biophysical model that was previously used to describe chromatin conformation of a yeast chromosome (25) (*Methods*). The model quantifies the expected frequency of interactions between a given pair of Dβ and Jβ genes based on the genomic distance between those genes. The model describes chromatin as a self-avoiding polymer that may be constrained in space into a curved shape. It contains two free parameters corresponding to the flexibility (or persistence length) of the chromatin and the radius of its constrained curvature. We applied this model to the genomic region spanning Dβ1 and Jβ2.7, and evaluated the best-fit flexibility and curvature parameters to the measured Dβ-Jβ pairing frequencies. Strikingly, we find that the model explains 83% ($P = 0.01$, permutation test) of the biases in average Jβ frequencies found in unselected clonotypes (Fig. 2*A*). The predictive accuracy of the model

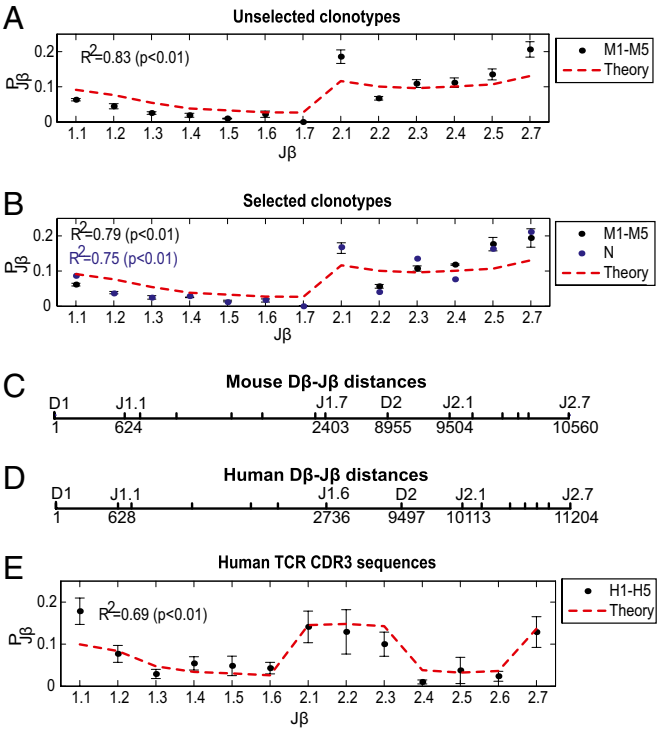

Fig. 2. Chromosome conformation determines a substantial fraction of the variation in Jβ frequencies. (*A*) Average of measured Jβ frequencies in unselected clonotypes from mice M1–M5 (circles) and corresponding theoretical frequencies (red dashed line) calculated using a biophysical model for chromatin conformation of the DJβ locus (model adapted from ref. 25; see *Methods* for details). Error bars indicate SD of the measured frequencies. We fit the model to the average Dβ–Jβ frequencies found in mice M1–M5 (*SI Appendix*, Table S3) using mouse RSS distances (*SI Appendix*, Table S4) as the independent variables, yielding the best-fit parameter estimates b_{est} = 68.77 nm and c_{est} = 10.86 kbp (*SI Appendix*, Table S5). (*B*) Average of measured Jβ frequencies in selected clonotypes from mice M1–M5 and in a published dataset (15) (*SI Appendix*, Table S6) and the theoretical frequencies computed in *A*. (*C* and *D*) Positions of Dβ and Jβ genes, relative to Dβ1, in (*C*) mice and (*D*) humans, based on the positions (18) of corresponding 12-RSSs for Jβ and 23-RSSs for Dβ. (*E*) Average of measured Jβ frequencies in human T cells (10, 11) (*SI Appendix*, Table S7) and theoretical frequencies computed by applying the biophysical model, with parameter values given in *A*, to human Dβ–Jβ RSS distances (*SI Appendix*, Table S8).

is not sensitive to the particular copy-number cutoffs used for defining clonotypes (*SI Appendix*, Fig. S7). The mechanical model fits the data much better than a genetic model (23, 24, 26) that is based on sequence conservation of RSSs flanking individual Jβ genes (*SI Appendix*, Fig. S8*A*).

The model fit suggests that during rearrangement, the chromatin found at the Dβ1–Jβ2.7 genomic region is highly flexible, with an apparent persistence length of ∼20 nm or below (*SI Appendix*, *SI Text*, section 2). A highly flexible structure is consistent with reported extensive remodeling of chromatin at this genomic region during rearrangement, by protein factors such as switch/sucrose nonfermentable protein complex (SWI/SNF) (27) and high-mobility group proteins (HMG) (28). The latter protein had been shown to decrease greatly the persistence length of naked DNA, from ∼50 nm to only ∼5 nm (29). Additionally, the model predicts that the Dβ1–Jβ2.7 genomic region is constrained in a curved conformation during rearrangement such that the frequency of random encounters between pairs of Dβ-Jβ genes is maximal at both small and large genomic distances (∼340 bp and ∼10,500 bp, respectively, for the fitted parameters; *SI Appendix*, *SI Text*, section 2). The model also predicts correctly the distinct

IMMUNOLOGY

pattern of pairing of the first and second DJCβ clusters that was described previously.

Applying this model to the Vβ region does not provide a good fit. We assume that this is due to the much longer genomic region spanned by Vβ genes (*SI Appendix*, Fig. S1*A*), which may be constrained into a more complex structure, potentially containing several loops (30, 31). Prediction of such a multiloop structure requires an extension of the biophysical model, and potentially more data to constrain it. However, the general assumptions of the model suggest that the rates of primary rearrangement between the same Vβ and the two different Dβs will be similar: because the Vβ–Dβ genomic distance is substantially larger than the Dβ1–Dβ2 distance, the distance of any Vβ segment to Dβ1 is not substantially different from its distance to Dβ2, and the model thus predicts a similar rearrangement rate. This provides a plausible mechanistic basis for the observed independence between Vβ and Jβ frequencies (Fig. 1*D*).

Interestingly, the Jβ frequencies that we measured are highly similar between selected and unselected $CD4^+$ unique clonotypes (*SI Appendix*, Fig. S9*A*), and also between selected $CD4^+$ and $CD8^+$ unique clonotypes (*SI Appendix*, Fig. S9*B*). Consequently, the biophysical model provides an explanation also for Jβ frequencies measured in selected $CD4^+$ (Fig. 2*B*) and $CD8^+$ clonotypes (*SI Appendix*, Table S5). These findings suggest that biases in Jβ gene use that occur during genomic rearrangement are preserved in the peripheral foreign antigen-inexperienced T-cell repertoire, and are largely unaffected by thymic selection and homeostatic clonal expansion.

Differences in TCR Jβ Gene Frequencies Between Humans and Mice Can Be Explained by the Chromatin Conformation Model. Distances between Dβ and Jβ gene segments generally differ between species (Fig. 2 *C* and *D*). Our model suggests that such differences will translate into variation in Jβ frequencies. We tested this prediction using previously published frequencies of TCR Jβ genes measured in human blood (10, 11). As in our data, the human data also shows similarity in Jβ biases between individuals and between T-cell subsets. However, there are differences in Jβ frequencies between humans and mice, especially in the second DJCβ cluster. To minimize overfitting, we applied the model to the human Dβ1–Jβ2.7 genomic region using estimates for the two model parameters, chromatin flexibility and curvature, obtained for the mouse data. Remarkably, the model explains 69% ($P = 0.01$, permutation test) of the variation in the human data (Fig. 2*E*), despite the fact that it was parameterized using the mouse data. In particular, the model provides a mechanistic explanation for the different pattern of Jβ 2.1–2.7 frequencies found between the two species. In contrast, a genetic model based on sequence conservation of RSSs (23, 24, 26) for individual human Jβ genes cannot explain the human data (*SI Appendix*, Fig. S8*B*). Together, the results presented above provide strong evidence that chromatin conformation determines biases in Jβ gene use, in both mice and humans.

Model for TCR Sharing Based on Biased Production Frequencies Predicts a Threshold for Public TCR Sequences. Our data allows for analysis of the level of sequence sharing among the five individual mice for which we obtained $CD4^+$ TCR sequences. Surprisingly, we find that many more TCRB amino acid sequences are shared among all five mice than among only subsets of either three or four mice (Fig. 3 *A* and *B*, *SI Appendix*, Fig. S10, and *Methods*). This trend is evident for both selected and unselected TCR amino acid sequences. We also observe that the highly shared sequences use preferentially the Jβ (Fig. 3*C*) and Vβ (*SI Appendix*, Fig. S11) gene segments that are most frequent in the repertoire.

Next, we checked whether the observed pattern of sharing can be explained as a result of biases in the frequencies at which the highly shared sequences are produced. Thus, we applied a probabilistic model to link those frequencies to patterns of TCR sequence sharing (*SI Appendix*, *SI Text*, section 3). Assuming that each sequence has an a priori probability, f, of being made, the model predicts that there is a threshold probability above which a sequence is more likely to be shared among all individuals in a group than being exclusive to any particular subgroup (Fig. 3*D*). This threshold probability is given by

$$f_T = 1 - 2^{-1/N},$$

where N is the total number of sequences found in each individual. For a large value of N, the threshold level is well approximated by

$$f_T \cong \frac{1}{N}\log_e 2.$$

For a hypothetical unbiased repertoire of size much larger than N, the a priori production probability of all sequences will be below threshold, resulting in a vanishingly small probability of sharing. Biases, such as those introduced by chromatin conformation, and also by other factors discussed above, can cumulatively increase f of specific sequences above threshold, making those particular sequences more likely to be public. We show this predicted trend by comparing the calculated sharing probability of sequences that are found below and above the publicness threshold, f_T (Fig. 3*D*). For a sequence with a priori frequency f that is below threshold (Fig. 3*D*, black line), the probability of sharing declines with increasing subgroup size. This sequence is more likely to be private to one mouse than to be shared by two or more mice. However, a sequence with an a priori frequency that is above threshold (Fig. 3*D*, blue line) is more likely to be shared by the entire group of five individuals than by any smaller subset. Hence, the pattern of amino acid TCR sequence sharing (both selected and unselected) that we observe in our data (Fig. 3 *A* and *B* and *SI Appendix*, Fig. S10) is consistent with the existence of a significant number of sequences whose a priori production frequencies are above the defined threshold.

Discussion

An earlier perspective on the structure of lymphocyte receptor repertoires held that the repertoires are primarily unbiased with respect to the frequencies of receptor genes (reviewed in ref. 32), and that significant biases in those frequencies are mainly due to lymphocyte selection and clonal expansion. However, recent work has shown that TCRB gene segment frequencies are substantially biased even in the primary TCRB repertoire, before T-cell selection (32, 33). Focusing on frequencies of Dβ-Jβ pairing, our results indicate a general mechanism that shapes biases in gene segment use. The model shows how biases in Jβ gene use can emerge naturally from differences in the degree to which chromatin conformation constrains rearrangement rates between different pairs of genes, according to the genomic distances between them. Previous work (23, 24) showed that rearrangement frequencies for synthetic sequence constructs representing simplified models of lymphocyte receptor loci can be predicted accurately based on the degree of sequence similarity between individual RSSs found in those constructs and physiological RSSs found in mice. However, this approach cannot explain rearrangement frequencies for receptor genes measured in vivo, as we demonstrated in this work (*SI Appendix*, Fig. S8). By introducing a very different approach for explaining gene rearrangement frequencies, we showed that chromatin conformation determines a substantial proportion of the biases in Jβ gene use measured in both mice and humans (Fig. 2 *A*, *B*, and *E*). Our approach is general, relying only on chromatin conformation at genomic loci for receptor genes of interest, and it is therefore applicable also to other loci. It will be interesting to assess the extent to which chromatin conformation explains biases in gene segment use in other rearranged receptors such as the TCR α, or the Ig heavy and light chains. The ability of chromatin conformation to explain rearrangement frequencies at these other receptor loci will depend on the magnitude of the contributions

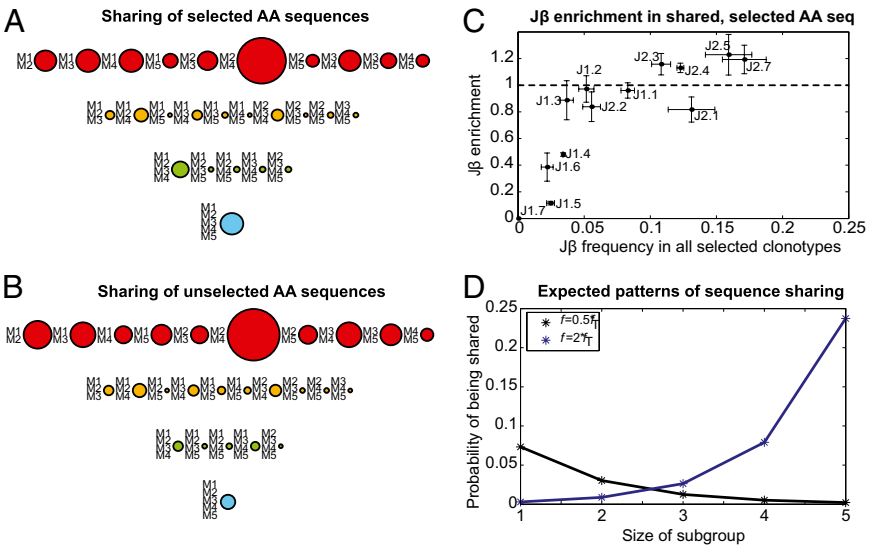

Fig. 3. Biases in the repertoire affect patterns of sequence sharing. (*A* and *B*) Measured number of (*A*) selected and (*B*) unselected amino acid (AA) sequences shared among different subgroups of five mice (M1–M5). In each subplot, circle area is proportional to the number of shared sequences of 1.5×10^4 (*A*) and 2×10^3 (*B*) unique sequences sampled randomly of the total clonotypes obtained for each mouse. Notably, the number of selected sequences shared among all five mice (public) is larger than that shared by any subset of three or four mice. (*C*) The enrichment for each Jβ gene (its frequency in the selected public clonotype subset divided by its frequency in all selected clonotypes) is plotted against the corresponding Jβ frequency in all selected clonotypes. (*D*) The probability that in a group of five individuals a sequence will be shared among only a particular subgroup of the indicated size is shown for two sequences: one (black) with a priori production frequency f that is lower than the threshold frequency determining sequence publicness f_T (defined in the text), and the other (blue) with $f > f_T$. The probability was calculated using *SI Appendix*, Eq. **S28**, with $n = 10^5$. Sharing probability decreases with group size for $f < f_T$, but increases with group size for $f > f_T$.

from other factors (1, 23, 24, 27, 28, 34–37) that influence the rearrangement process.

In particular, the demonstrated effect of chromatin conformation on gene segment use is potentially further modulated by *cis*-acting elements such as RSSs (1, 23, 24, 34), coding ends of genes (35), Vβ/Dβ promoters (34), and other molecular factors that regulate the accessibility of individual genes to enzymes involved in VDJ rearrangement (27, 28, 36, 37). Additionally, gene segment use may be further modulated by thymic selection, homeostatic T-cell expansion in peripheral lymphoid organs, and clonal expansion during specific immune responses (38). However, thymic selection has only a weak effect on Jβ gene use, as evidenced by the high correlation in Jβ frequencies that we found between selected and unselected T-cell clonotypes (*SI Appendix*, Fig. S9*A*). Moreover, we found that Jβ frequencies are highly similar between selected $CD4^+$ and $CD8^+$ T-cell clonotypes (*SI Appendix*, Fig. S9*B*), despite the fact that thymic selection of these clonotypes depends on interactions with class II vs. class I MHC molecules, respectively. Together, these findings support a picture in which biases in Jβ gene use are mostly determined during VDJ rearrangement, by chromatin conformation at rearranging genomic loci.

The mechanical model could not predict measured Vβ frequencies based on chromatin conformation; this could be due to stronger modifying effects of thymic selection and other factors, as discussed above, on Vβ gene use. However, the longer Vβ region may be constrained during rearrangement into several loops, and may still be explained by an extended version of the mechanical model for chromatin. Support for this view comes from chromosome conformation capture experiments that show contraction of the TCRB locus in double-negative thymocytes (30). The results of these experiments suggest that the two long regions in the locus devoid of Vβ genes are "looped out" and thus located away from the DJβ domain, whereas areas that contain Vβ genes tend to be closer to this domain.

A recent review summarizes a large number of studies in which public TCR clones were identified among individuals, in humans, other primates, and mice (32). Most of these studies were based on identification and sequencing of antigen-specific clones. Recently, a high-throughput study of TCR sequences of a specific Vβ-Jβ pair in humans provided evidence for roles of convergent recombination in enhancing sequence publicness in an unbiased way (39). However, because a single Vβ-Jβ combination was studied, effects of biases in gene use could not be investigated in that study (39). Mapping of the entire spectrum of Vβ-Dβ-Jβ combinations allowed us to gain a wider view of the effects of biases in the rearrangement process on TCR sequence sharing. Biases in the rearrangement process and in the convergence of different TCR nucleotide sequences to the same amino acid sequence (4, 32) determine the a priori probability f that each TCR amino acid sequence will be produced. This a priori probability in turn largely determines which sequences are intrinsically "public," meaning that they are more likely to be produced in multiple individuals than in fewer individuals. However, the precise relationship between f and sequence publicness had not previously been determined. Our data for sequence sharing motivated us to derive a mathematical expression for a threshold value for f, above which a particular sequence will be public, and below which it will be private (*SI Appendix*, *SI Text*, section 3). This threshold value is inversely proportional to the total number of sequences sampled from each individual. Indeed, the fraction of sequences shared among all five mice grows with sample size as predicted by the model, whereas the fraction of sequences shared by subgroups becomes saturated (*SI Appendix*, Fig. S12). The existence of this well-defined threshold governing the distinction between private and public sequences means that even a small change in f, e.g., due to a corresponding change in the relative rearrangement frequency of a particular gene, can alter systematically sharing patterns for sequences whose f values are close to the threshold.

Finally, we would like to suggest that the identified mechanism by which the genomic distance between gene segments affects their probability for recombination could be harnessed in productive ways, linking genetic changes to beneficial variation of repertoire structure on evolutionary time scales. An intriguing possibility is that genetic changes such as insertions or deletions in noncoding regions, as well as deletions or duplications of genes (40), could change distances between VDJ gene segments, thus altering their frequencies in the repertoire. Such changes can in turn tune the composition of the set of public clonotypes in accordance with threats posed by common pathogens, and potentially also with evolving needs for self-maintenance (41, 42).

Methods

Additional details can be found in *SI Appendix*.

Library Construction and Sequencing. Library construction protocol is schematically described in *SI Appendix*, Fig. S1*B*. Total RNA was extracted from splenic T cells of C57BL/6 mice and reverse transcribed using a TCR Cβ-specific primer linked to the 3′-end Illumina sequencing adapter. cDNA was used as template for high-fidelity PCR amplification (18 cycles) using the Cβ primer and a set of 23 Vβ primers. Each Vβ-specific primer was anchored to

a restriction site sequence for a restriction enzyme (AcuI) that we used to cleave part of the primer sequence, such that sequencing starts closer to the Vβ–Dβ junction region; this allows for good coverage of CDR3 with a single Illumina read. This step was followed by ligation of the Illumina 5′ adapter, which was linked to a 3-bp barcode sequence in its 3′ end, and a second round of PCR amplification (24 cycles) using primers for the 5′ and 3′ Illumina adapters. Final PCR products were gel purified and sequenced using Genome Analyzer II (Illumina).

Processing and Characterization of TCR Sequences. The analysis pipeline is schematically described in *SI Appendix*, Fig. S1C. Sequencing reads were quality filtered (Q value ≥20) and assigned germline Vβ/Jβ gene segments (18), using the following threshold alignment lengths (determined by a permutation analysis; *SI Appendix*, Fig. S13): for datasets M1–M4 and M7–M8 (40-nt reads): 11 nt for Vβ, 9 nt for Jβ. For M5 (80-nt reads): 12 nt for Vβ, 11 nt for Jβ. Assigned reads were clustered to reduce effects of sequencing errors. Cluster sequences were translated, and those containing a stop codon were designated as "unselected" and the rest as "selected." For some clusters, we could also assign a Dβ gene, requiring a perfect match of length >6 nt, because of the high similarity between the two germline Dβ genes (18). Cluster sizes were corrected for PCR amplification bias using a new probabilistic method applied to a synthetic library of 79 cloned TCRs (representing all 23 Vβ segments), which was sequenced and processed in parallel with our experimental libraries (*SI Appendix, SI Methods*). To increase the signal-to-noise ratio of our data, we analyzed cluster sequences (called clonotypes) that have an enzymatic cleavage error of ≤2 nt, and a bias-corrected cluster size of ≥5.

Biophysical Model for Gene Rearrangement Frequency. We adapted a previously published model (25) to calculate gene frequencies. The biophysical model gives the theoretical frequency of the *i*th Jβ gene as $P(J\beta_i) = K[\alpha_1{}^{-3/2} \exp(-2\alpha_1{}^{-2}) + \alpha_2{}^{-3/2}\exp(-2\alpha_2{}^{-2})]$, where $\alpha_j = (d_j/b)(1 - d_j/c)$, $j = 1,2$. $d_{i,j}$ (in bp) is the genomic distance between the start position of the 12-bp spacer RSS (12-RSS) of $J\beta_i$ and the start position of the 23-bp spacer RSS (23-RSS) of $D\beta_j$ (*SI Appendix*, Table S4), K is a normalization constant, and both b (in nm) and c (in bp) are free parameters. We fit the model to measured Dβ–Jβ frequencies by means of simulated annealing (43) followed by gradient descent. See *SI Appendix, SI Text*, section 2, for additional information.

Sequence Sharing Analysis. For analysis of selected (unselected) sequences, we sampled randomly 15,000 (2,000) unique amino acid sequences from each dataset (M1–M5), where the chance of selection is proportional to the number of times each amino acid sequence appears in the dataset. This method allows for analysis of sharing between datasets of different sizes, providing a direct comparison of data and model, which is based on the a priori frequency of generating unique sequences (*SI Appendix, SI Text*, section 3).

ACKNOWLEDGMENTS. We thank I. Cohen, T. Pilpel, and R. Sorek for helpful discussions and comments on the manuscript; S. Horn-Saban, D. Zalcenstein, and D. Leshkowitz for technical support with Illumina sequencing and helpful discussions; and the anonymous reviewers for their insightful comments. This research was supported by the International Human Frontier Science Program Organization and the Benoziyo Center for Neurological Diseases. W.N. was supported by a postdoctoral fellowship from the Weizmann Institute of Science. N.F. is the incumbent of the Pauline Recanati Career Development Chair of Immunology.

1. Bassing CH, Swat W, Alt FW (2002) The mechanism and regulation of chromosomal V(D)J recombination. *Cell* 109(Suppl):S45–S55.
2. Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. *Nature* 334:395–402.
3. Gorski J, et al. (1994) Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. *J Immunol* 152:5109–5119.
4. Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8:231–238.
5. Turner SJ, Doherty PC, McCluskey J, Rossjohn J (2006) Structural determinants of T-cell receptor bias in immunity. *Nat Rev Immunol* 6:883–894.
6. Pannetier C, et al. (1993) The sizes of the CDR3 hypervariable regions of the murine T-cell receptor β chains vary as a function of the recombined germ-line segments. *Proc Natl Acad Sci USA* 90:4319–4323.
7. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135–1145.
8. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810.
9. Wu YC, et al. (2010) High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116:1070–1078.
10. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.
11. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA (2009) Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* 19: 1817–1824.
12. Robins HS, et al. (2010) Overlap and effective size of the human $CD8^+$ T cell receptor repertoire. *Sci Transl Med* 2:47ra64.
13. Warren RL, et al. (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* 21:790–797.
14. Jung D, Giallourakis C, Mostoslavsky R, Alt FW (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* 24: 541–570.
15. Candéias S, Waltzinger C, Benoist C, Mathis D (1991) The V β 17+ T cell repertoire: Skewed J β usage after thymic selection; dissimilar CDR3s in CD4+ versus CD8+ cells. *J Exp Med* 174:989–1000.
16. Wade T, Bill J, Marrack PC, Palmer E, Kappler JW (1988) Molecular basis for the nonexpression of V β 17 in some strains of mice. *J Immunol* 141:2165–2167.
17. Manfras BJ, Terjung D, Boehm BO (1999) Non-productive human TCR beta chain genes represent V-D-J diversity before selection upon function: Insight into biased usage of TCRBD and TCRBJ genes and diversity of CDR3 region length. *Hum Immunol* 60:1090–1100.
18. Lefranc MP, et al. (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* 37(Database issue):D1006–D1012.
19. Kato T, et al. (1994) Comparison of the J beta gene usage among different T cell receptor V beta families in spleens of C57BL/6 mice. *Eur J Immunol* 24:2410–2414.
20. Bousso P, et al. (1998) Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* 9:169–178.
21. Menezes JS, et al. (2007) A public T cell clonotype within a heterogeneous autoreactive repertoire is dominant in driving EAE. *J Clin Invest* 117:2176–2185.
22. Livak F, Burtrum DB, Rowen L, Schatz DG, Petrie HT (2000) Genetic modulation of T cell receptor gene segment usage during somatic recombination. *J Exp Med* 192: 1191–1196.
23. Lee AI, et al. (2003) A functional analysis of the spacer of V(D)J recombination signal sequences. *PLoS Biol* 1:E1.
24. Cowell LG, Davila M, Yang K, Kepler TB, Kelsoe G (2003) Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. *J Exp Med* 197:207–220.
25. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295:1306–1311.
26. Merelli I, et al. (2010) RSSsite: A reference database and prediction tool for the identification of cryptic recombination signal sequences in human and murine genomes. *Nucleic Acids Res* 38(Web Server issue):W262–W267.
27. Osipovich O, et al. (2007) Essential function for SWI-SNF chromatin-remodeling complexes in the promoter-directed assembly of *Tcrb* genes. *Nat Immunol* 8:809–816.
28. van Gent DC, Hiom K, Paull TT, Gellert M (1997) Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J* 16:2665–2670.
29. McCauley M, Hardwidge PR, Maher LJ, 3rd, Williams MC (2005) Dual binding modes for an HMG domain from human HMGB2 on DNA. *Biophys J* 89:353–364.
30. Skok JA, et al. (2007) Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes. *Nat Immunol* 8:378–387.
31. Bossen C, Mansson R, Murre C (2012) Chromatin topology and the regulation of antigen receptor assembly. *Annu Rev Immunol* 30:337–356.
32. Miles JJ, Douek DC, Price DA (2011) Bias in the αβ T-cell repertoire: Implications for disease pathogenesis and vaccination. *Immunol Cell Biol* 89:375–387.
33. Wilson A, Maréchal C, MacDonald HR (2001) Biased V β usage in immature thymocytes is independent of DJ β proximity and pT α pairing. *J Immunol* 166:51–57.
34. Sleckman BP, Gorman JR, Alt FW (1996) Accessibility control of antigen-receptor variable-region gene assembly: Role of *cis*-acting elements. *Annu Rev Immunol* 14: 459–481.
35. Gerstein RM, Lieber MR (1993) Coding end sequence can markedly affect the initiation of V(D)J recombination. *Genes Dev* 7(7B):1459–1469.
36. McMurry MT, Krangel MS (2000) A role for histone acetylation in the developmental regulation of VDJ recombination. *Science* 287:495–498.
37. Yancopoulos GD, Alt FW (1985) Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. *Cell* 40:271–281.
38. Correia-Neves M, Waltzinger C, Mathis D, Benoist C (2001) The shaping of the T cell repertoire. *Immunity* 14:21–32.
39. Venturi V, et al. (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* 186:4285–4294.
40. Kidd MJ, et al. (2012) The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* 188: 1333–1340.
41. Cohen IR (2004) *Tending Adam's Garden* (Elsevier, London).
42. Yoles E, et al. (2001) Protective autoimmunity is a physiological response to CNS trauma. *J Neurosci* 21:3740–3748.
43. Styblinski MA, Tang TS (1990) Experiments in non-convex optimization, stochastic approximation with function smoothing and simulated annealing. *Neural Netw* 3: 467–483.