Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing

Ivan V. Zvyagin[a], Mikhail V. Pogorelyy[a], Marina E. Ivanova[a], Ekaterina A. Komech[a], Mikhail Shugay[a,b], Dmitry A. Bolotin[a], Andrey A. Shelenkov[a,c], Alexey A. Kurnosov[a], Dmitriy B. Staroverov[a], Dmitriy M. Chudakov[a,b], Yuri B. Lebedev[a], and Ilgar Z. Mamedov[a,1]

[a]Department of Genomics and Postgenomic Technologies, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow 117997, Russia; [b]Department of Biomedical Technologies, Pirogov Russian National Research Medical University, Moscow 117997, Russia; and [c]Department of Genetics and Biotechnology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russia

Edited by Philippa Marrack, Howard Hughes Medical Institute, National Jewish Health, Denver, CO, and approved March 11, 2014 (received for review October 15, 2013)

Adaptive immunity in humans is provided by hypervariable Ig-like molecules on the surface of B and T cells. The final set of these molecules in each organism is formed under the influence of two forces: individual genetic traits and the environment, which includes the diverse spectra of alien and self-antigens. Here we assess the impact of individual genetic factors on the formation of the adaptive immunity by analyzing the T-cell receptor (TCR) repertoires of three pairs of monozygous twins by next-generation sequencing. Surprisingly, we found that an overlap between the TCR repertoires of monozygous twins is similar to an overlap between the TCR repertoires of nonrelated individuals. However, the number of identical complementary determining region 3 sequences in two individuals is significantly increased for twin pairs in the fraction of highly abundant TCR molecules, which is enriched by the antigen-experienced T cells. We found that the initial recruitment of particular TCR V genes for recombination and subsequent selection in the thymus is strictly determined by individual genetic factors. J genes of TCRs are selected randomly for recombination; however, the subsequent selection in the thymus gives preference to some α but not β J segments. These findings provide a deeper insight into the mechanism of TCR repertoire generation.

immunogenetics | TCR repertoire analysis | twin studies

Adaptive immunity is provided by B and T cells bearing B-cell receptors (BCRs) and Ig-like T-cell receptors (TCRs), respectively. These hypervariable molecules are the key part of the adaptive immune system as they can potentially recognize any alien agent and drive specific immune responses. The α/β TCRs recognize short peptides in the complex with major histocompatibility complex (MHC) molecules and play the key role in the targeted immune response. The total diversity of TCR molecules in an individual human organism is initially formed via genomic recombination with subsequent positive and negative selection at several stages of maturation and activation. The maximal theoretical diversity of TCRβ chain's amino acid sequences in humans is estimated between 5×10^{11} (1) and 10^{14} (2), whereas the maximal number of α/β pairs reaches 10^{18} (3). This huge number of variants is probably never achieved: the whole TCRβ chain repertoire size in a single human organism is estimated at $1–5 \times 10^6$ (1, 4–6), although this is only a lower bound estimate. Two driving forces shape the final face of individual TCR repertoire: the individual genetics and the complexity of environmental factors. The genes coding for proteins involved in VDJ recombination, antigen processing and presentation, and products of genes participating in the immune response signaling belong to the first type of the repertoire-forming factors. The spectrum of the organism's self-peptides presented in the thymus also depends on the individual's set of the MHC molecules. Moreover, this spectrum of peptides is determined by the amino acid sequences of the organism's proteins, which thus can also be considered a genetic factor. Furthermore, TCRs arising to the same alien antigenic peptides are known to be MHC restricted (7). The environmental factors include the whole range of pathogens met by the individual including disease-causing bacteria and viruses, as well as vaccines, symbionts, etc. The genetic component can potentially have a major impact on the initial recombination and selection in the thymus forming the naïve TCR repertoire, whereas the subsequent interference with antigens provides the selective expansion of some TCRs and forms the final repertoire structure. However, the particular impact of genetic factors on TCR repertoire structure and diversity is unknown.

All genes of monozygous (MZ) twins are identical (including those responsible for the TCR repertoire formation), and therefore, MZ twins are widely used in the studies where the genetic impact is evaluated. Several studies of TCR repertoires were performed mainly focusing on diseases concordant and discordant MZ twins and using complementarity determining region 3 (CDR3) spectratyping and/or low depth sequencing (8–11). Some of these studies reported the common use of particular V genes and common clonotypes. In recent years, the high-throughput sequencing technologies paved the way to whole-repertoire studies of individual TCRs that led to new findings in the field of adaptive immunity (1, 5, 6, 12–22). In this study, for the first time to the best of our knowledge, we obtain and compare the α and β chain TCR repertoires of three pairs of MZ twins using next-generation sequencing (NGS).

Significance

The power of adaptive immunity in humans is realized through the hypervariable molecules: the T-cell receptors (TCRs). Each of those is built from genetically encoded parts with the addition of random nucleotides finally forming individual TCR repertoire. Despite that the individual TCR repertoire potentially can include 10^{11}–10^{14} different variants, substantially less molecules are found in a single individual. The particular genetic impact on the final set of TCR molecules is still poorly understood. In this study, for the first time to the best of our knowledge, we compare deep TCR repertoires of genetically identical twins. We found that, although TCR repertoires of any pair of individuals have the same amount of identical receptors, twin repertoires share certain specific features.

Author contributions: I.V.Z., Y.B.L., and I.Z.M. designed research; I.V.Z., M.E.I., E.A.K., and D.B.S. performed research; I.V.Z., M.V.P., M.S., D.A.B., and A.A.S. analyzed data; and A.A.K., D.M.C., Y.B.L., and I.Z.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the NCBI sequence read archive database (accession no. SRP028752).

[1]To whom correspondence should be addressed. E-mail: Imamedov@mx.ibch.ru.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319389111/-/DCSupplemental.

Results

Library Preparation and Sequencing. RNA samples from three female pairs of monozygous twins (A, C, and D) were studied in this work. All individuals were HLA (MHC I) typed, and we found that each of the three MZ pairs had the same MHC I alleles profile (Table S1). TCR repertoires (for α and β chains) were obtained for all three pairs of monozygous twins. Sequencing libraries were prepared as described previously (12, 15, 23) with several modifications (Fig. S1 and *SI Materials and Methods*). To avoid any potential cross-contamination during library preparation and sequencing, we introduced sample-specific barcodes on both ends of each TCR cDNA library in the course of cDNA synthesis and amplification. In total, we obtained around 50,000,000 sequencing reads for 12 samples (6 α and 6 β) using the Illumina sequencing platform. After removing the reads with improper index combinations, TCR clonesets were generated for each library using our MiTCR software package (24) (see Table S2 for reads and clones distribution). Briefly, MiTCR extracts the CDR3 sequence from each read, identifies V, D, and J gene segments, merges and counts sequences with the identical CDR3 to form clonotypes, filters out or rescues low-quality reads, and provides advanced correction of PCR and sequencing errors. MiTCR CDR3 extraction yielded 2.2–3.6 million reads for each of the 12 libraries (Table S2). The identified number of clonotypes was different for each individual, varying from ∼100,000 to nearly 500,000. The resulting clonotypes are available at http://labcfg.ibch.ru/tcr.html#MZTwins and http://mitcr.milaboratory.com/datasets/twins2013/ in MiTCR format (24). A portion of each library was comprised by the out-of-frame clonotypes representing the nonfunctional TCR sequences formed during the recombination step. The percentage of such sequences was different for α and β TCR libraries, varying in most cases from 12.5% to 14.5% and from 2.9% to 4.1% for α and β libraries, respectively (Table S2).

V Gene Usage but Not J Gene Usage Is Similar in Twins Before the T-Cell Selection in Thymus. Proteins involved in TCR gene recombination could affect the initial TCR repertoire formed at the very early stage before selection in the thymus. MZ twins have identical genes coding for the TCR recombination machinery; thus, their TCR repertoires are expected to be more similar at this step. In course of T-cell maturation, TCR locus recombination events can produce nonfunctional TCRs with frameshifts or stop codons (25). In this case, the T cell tries to arrange the second allele, and if the successful (in-frame) TCR formation occurs, the T cell carries both functional and nonfunctional TCR genes (26). The latter allele expression is usually down-regulated by the nonsense-mediated mRNA decay (NMD). However, the corresponding RNA is still present in the T cell at some concentration.

To analyze the nonfunctional TCR repertoire, we first removed artificial (carrying PCR or sequencing errors) out-of-frame clonotypes by mapping them to the in-frame sequences with one indel nucleotide allowed in the CDR3. The remaining genuine out-of-frame sequences were not affected by any type of selection (either in the thymus or on the periphery) and thus were used to characterize the initial repertoire formed by recombination itself (1, 2, 20). We then calculated the Jensen–Shannon (JS) divergence divided by mean entropy (*SI Materials and Methods*) for V and J gene usage in all possible pairs formed by six individuals for α and β out-of-frame TCR clonotypes, irrespective of each clonotype frequency (Fig. 1). The more similar distribution of V or J segments between a pair of individuals is characterized by the lower JS divergence (distance) value. V segment usage for TCRβ out-of-frame clonotypes was more similar (4–10 times lower JS divergence) in each twin pair compared with other individuals (Fig. 1, Vβ). For TCRα out-of-frame clonotypes the lowest JS distance was also observed between the individual and her twin except for C1 (Fig. 1, Vα). In contrast, JS divergence for J gene usage was distributed in a nearly stochastic manner, indicating no significant skew in J segment use in twin pairs for both α and β TCR chains (Fig. 1, Jα and Jβ).

J Gene Usage Is Affected by Selection in the Thymus. The next step of TCR repertoire formation includes the selection of T cells with functional CDR3 sequences against the individual's specific repertoire of MHC molecules in the thymus. After maturation in the thymus, a pool of naïve T cells able to recognize MHC and at the same time not to recognize auto-antigens is formed. At this step, individual genetics is the main driving force as the MHC molecules and self-antigenic peptides are both genetically encoded. As the high-throughput sequencing provides deep TCR repertoire profiling, the main diversity of the obtained repertoire is generated by the naïve T cells. At the same time, antigen-experienced expanded clones represented by a large number of sequencing reads have a minor influence on the total clonotype diversity [for example, Robins et al. (1) report a mean of 420,000 unique CDR3 sequences in FACS-sorted naïve samples compared with 69,000 unique sequences in a memory pool]. Thus, we suggested that the entire set of the in-frame TCR clonotypes sequenced is dominated by a naïve T-cell repertoire. We calculated the JS divergence for V and J gene usage for α and β in-frame TCR clonotypes of all possible pairs of individuals (Fig. 2).

V gene usage for both α and β in-frame TCR clonotypes was obviously closer for all three twin pairs compared with any pair of unrelated individuals (Fig. 2, Vβ and Vα). J gene usage for the TCRβ in-frame clonotypes demonstrated diverse patterns of JS divergence: for individuals A1 and A2, the shortest JS distance was to the twin individual, whereas for individuals D1/D2 and C1/C2, the distribution of JS distance was stochastic (Fig. 2, Jβ).

A significant change in the pattern of JS divergence of Jα usage was observed for the in-frame clonotypes compared with the out-of-frame clonotypes. Jα usage for the in-frame clonotypes was much more similar for all three twin pairs than for random pairs of individuals, whereas for the out-of-frame TCRα repertoires, J genes demonstrated stochastic distribution independent of kinship (Figs. 1, Jα, and 2, Jα). Additionally average JS distance for Jα in-frames was considerably (5–10 times) shorter compared with out-of-frames between any pair of individuals. This finding suggests that TCRs' J segments are initially selected for recombination in a relatively random manner in all humans, but during the subsequent selection, the usage of Jα segments becomes more universal for all individuals. Twin pairs

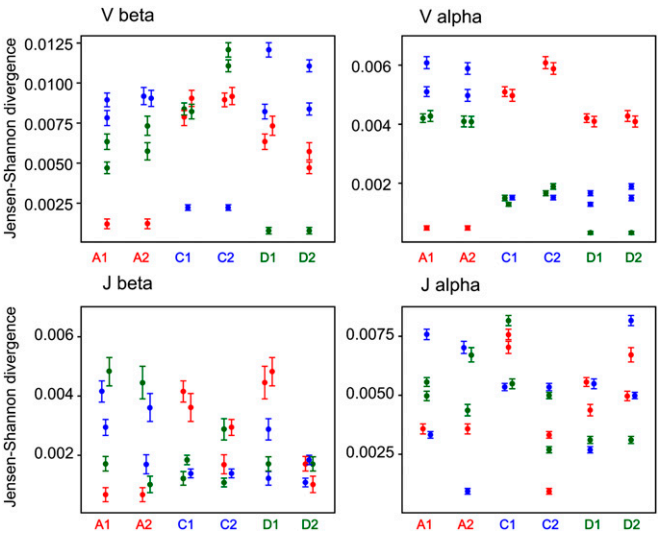

Fig. 1. JS divergence for Vβ, Jβ and Vα, Jα segment usage of out-of-frame clonotypes. Each column indicates the JS divergence divided by the mean entropy (for each pair) between an individual and five other individuals. Error bars indicate SD calculated using the bootstrap analysis. Red circles, individuals A1 or A2; blue circles, individuals C1 or C2; green circles, individuals D1 or D2.

IMMUNOLOGY

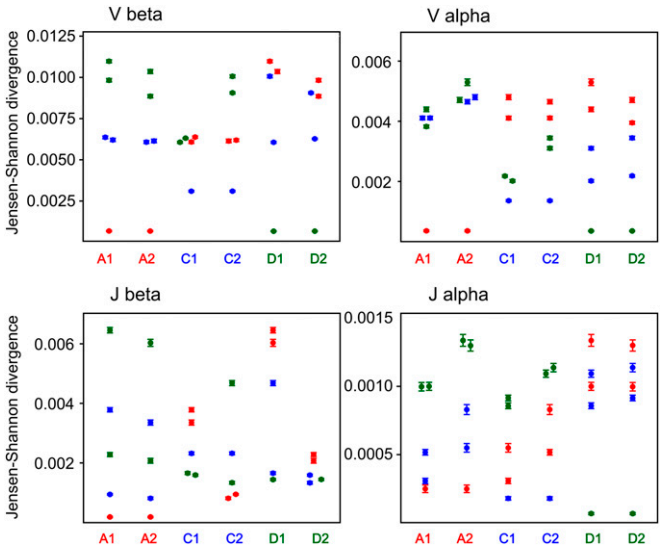

Fig. 2. JS divergence for Vβ, Jβ and Vα, Jα segment usage of in-frame clonotypes. Each column indicates the JS divergence divided by the mean entropy (for each pair) between an individual and five other individuals. Error bars indicate SD calculated using the bootstrap analysis. Red circles, individuals A1 or A2; blue circles, individuals C1 or C2; green circles, individuals D1 or D2.

use even a more similar pattern of Jα segments after selection compared with unrelated individuals.

MZ Twins Have More Shared Clonotypes Among the Most Abundant T-Cell Clones. The final individual repertoire of TCRs results from three sequential processes: the recombination of the germ-line TCR gene locus, selection in the thymus, and interaction with different antigens (including self-antigens). Therefore, as MZ twins bear identical genes coding for proteins involved in TCR recombination and identical sets of genes involved in antigen presentation, they are expected to have more similar sets of TCR molecules. Indeed, this fact is clearly observed from the similar usage of V gene segments (see Fig. 2) coding for CDR1 and CDR2 responsible for interaction with MHC. Moreover, twins reared together are exposed to a similar set of antigens processed in the same way for a long period of their life. Thus, we expected the set of TCR CDR3 sequences to have more similarity between twins than between unrelated individuals.

We analyzed the number of amino acid CDR3 clonotypes (i.e., having an identical amino acid sequence of CDR3 independently of their nucleotide CDR3 sequence) shared between each pair of individuals (Table S3) for α and β TCR repertoires. The minimal number of shared CDR3β clonotypes (6,906) was observed for individuals A1 and A2, both having the lowest number of identified clonotypes (hereafter "cloneset size"; 228,772 and 102,989 individual clonotypes, respectively), whereas the maximal number (39,479) was determined for C1 and D2 having the largest clonotype sets (around 527,000 and 500,000, respectively). The same pattern was observed for overlap of the α chain CDR3 repertoires. The number of shared clonotypes for all possible pairs of individuals and the product of the intersected clonesets sizes is shown in Fig. 3. We observed a linear correlation between the number of shared CDR3 clonotypes and the product of the cloneset sizes for both α and β TCR chains. R^2 was 0.9752 and 0.9814 (for α and β regression lines, respectively). Thus, the number of shared clonotypes between TCR repertoires of two individuals is mainly determined by the intersected cloneset sizes. The impact of the identical TCR recombination machinery and MHC turns out not to be substantial as at least two of the three twin pairs exactly fit the regression line (Fig. 3, red circles). Using the data for all pairs, we calculated 98% prediction intervals for both lines. We also added the data on the number of shared CDR3 clonotypes obtained by Warren et al. (6) for two pairs of individuals (Fig. 3, empty triangles). These clonesets were generated using the similar RNA-based TCR library preparation method. Data on them fell into our model's prediction interval despite that the TCR extraction and cloneset generation methods were different. Thus, the number of CDR3 clonotypes shared between TCR repertoires of two individuals can be predicted from the intersected clonesets sizes quite precisely. The normalized number of shared CDR3 clonotypes (i.e., the number of shared CDR3 clonotypes divided by the product of the intersected cloneset sizes) can be approximated as a cloneset size-independent measure of CDR3 repertoire similarity between two individuals.

The abundant clonotypes (i.e., clonotypes with a higher number of sequencing reads) are enriched by antigen experienced T-cell clones, whereas the clonotypes with a low number of reads are to a great extent represented by the naïve T cells [it should be mentioned, however, that some naïve T lymphocytes can be abundant (27) and some antigen experienced T-cell clones are quite rare]. Fig. 4 shows the relationship between the number of shared CDR3 clonotypes and the abundance of the selected clonotypes in the datasets for each possible pair of individuals (1,000 most abundant clonotypes in individual 1 vs. 1,000 most abundant clonotypes in individual 2; 2,000 most abundant clonotypes in individual 1 vs. 2,000 most abundant clonotypes in individual 2, etc.). The number of shared CDR3 clonotypes was divided by the product of the intersected cloneset sizes (1,000 × 1,000, 2,000 × 2,000, etc.) for normalization. Previous observations suggest that if we randomly pick a limited number (1,000, 2,000, etc.) of clonotypes from each individual cloneset (independent of each clonotype read count), the normalized number of shared clonotypes remains constant.

We found that the normalized number of shared clonotypes is significantly higher among the most abundant TCRβ CDR3 clonotypes for any pair of individuals. Notably, each of the six individuals had an even more increased number of the most abundant CDR3β clonotypes shared with her twin (Fig. 4). Similarly, more shared CDR3 amino acid sequences were observed among abundant α clonotypes. However, the difference between the twin and unrelated pairs was not obvious. Although individuals C1, C2, D1, and D2 shared the highest number of common clonotypes with their twins, the difference with the unrelated individuals was minimal, whereas no difference was observed for the individuals A1 and A2 (Fig. S2).

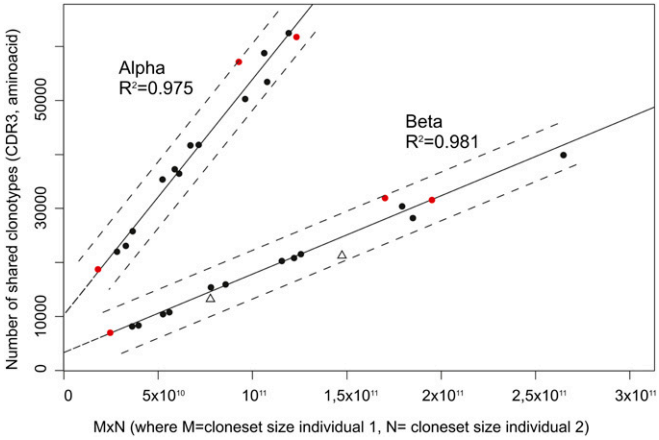

Fig. 3. Correlation between the number of shared clonotypes in each pair of individuals and the product of intersected cloneset sizes. Red circles correspond to twin pairs, and black circles correspond to unrelated pairs. Empty triangles correspond to the number of shared clonotypes obtained for two pairs of individuals by Warren et al. (6). Dash lines indicate the 98% prediction intervals for both regression lines.

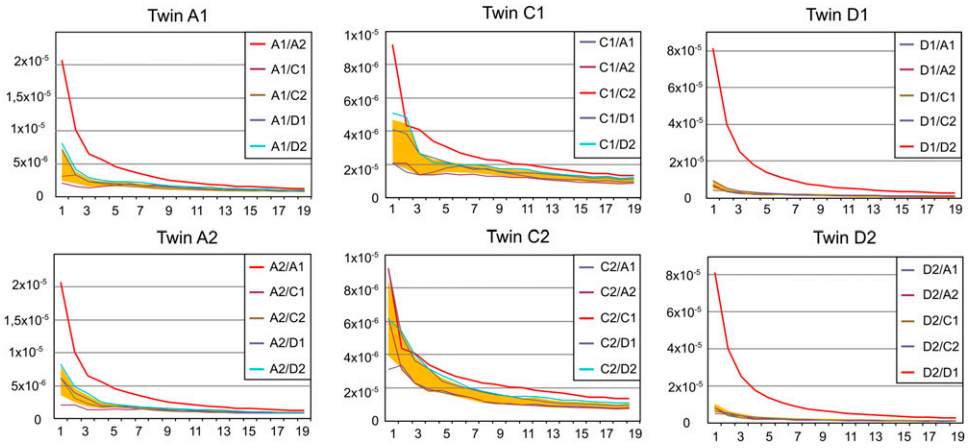

Fig. 4. The normalized number of identical TCRβ CDR3 amino acid sequences for each possible pair of individuals among 1,000 most abundant clonotypes, 2,000 most abundant clonotypes, etc. *x* axis, the number of the most abundant clonotypes (×1,000) intersected; *y* axis, the normalized number of shared clonotypes. The shaded area (orange) indicates the mean ± SD calculated for pairs formed by twin with each unrelated individual.

Shared Clonotypes Features Are Different for Twins and Unrelated Individuals. The antigen specificity of the T-cell receptors to the peptide-MHC complex is provided by three complement determining regions (CDRs). One of them (CDR3) is hypervariable (i.e., can consist of potentially any combination of amino acids), whereas CDR1 and CDR2 are selected from a set of about 50 possible variants (genetically encoded V genes). Thus, T-cell receptors having identical amino acid sequences of all three CDRs most probably recognize the same antigen-MHC complex in different individuals.

We analyzed the percentage of TCR clonotypes having the same V gene among clonotypes with identical amino acid CDR3s for each possible pair of six individuals for both α and β chains (Fig. 5). For the whole set of identified clonotypes the percentage of identical V genes among TCRs with identical CDR3 was similar for any pair of individuals for both α and β chains. Twenty-eight percent to 42% of clonotypes with identical CDR3β additionally have the same V gene, wherein three monozygous twin pairs have the highest percent ($P = 0.0088$, Mann–Whitney U test; Fig. 5, β all). For the TCRα clonotypes, this percent lied between 62% and 69% with no significant deviation for the twin pairs ($P = 0.233$, Mann–Whitney U test; Fig. 5, α all). We also calculated the percent of clonotypes with identical V genes among the carriers of identical CDR3s taken from the top 10,000 most abundant clonotypes for each possible pair of individuals for α and β TCRs (Fig. 5, β top 10K and α top 10K). We observed a significant increase of this value among the most abundant TCRβ sequences for all pairs (mean, 47%) compared with all clonotypes (mean, 33%). Interestingly, this percent was significantly ($P = 0.0044$, Mann–Whitney U test) higher for all three twin pairs (from 57% for C1/C2 to 81% for D1/D2). In contrast, there was no significant difference between the percentage of identical V genes among all and the top 10,000 most abundant TCRα clonotypes with identical CDR3 (means were 65% and 64%, respectively). This percentage for twin pairs among the 10,000 most abundant TCRα was slightly higher ($P = 0.048$, Mann–Whitney U test).

To further characterize the shared TCRs pool, we analyzed the number of random nucleotides added at the junction between the V-D and D-J segments for the β chain and V-J for the α chain. The average number of added nucleotides in the individual repertoire was 7.4 (sum for both junctions) for the β chain and 6.2 (for the single junction) for the α chain. In the shared TCR pools, we observed a significant decrease in the number of added nucleotides (3.3 and 3.8 for β and α, respectively), with no difference between twin pairs and unrelated individual pairs. This decrease is not surprising as TCRs with a lower number of inserted nucleotides have a higher chance for independent convergent generation in different individuals [as indicated previously (18, 20, 21)] and thus a higher chance to match in two or more humans. In contrast, when we intersected the most abundant clonotypes (top 10,000), the number of inserted nucleotides for matched TCRs differed between twin pairs and unrelated pairs (2.9 vs. 1.9, respectively, $P < 0.00001$, two-sided t test) for β chains. For α chains, the number of inserted nucleotides was very close for twin and unrelated pairs for the most abundant shared clonotypes (2.07 and 1.90, respectively, $P = 0.22$, two-sided t test).

Clonotypes shared by twins and by unrelated individuals can represent T cells fighting common viruses like CMV and Epstein–Barr virus (EBV). We stained peripheral blood mononuclear cells (PBMCs) from donor A2 with the HLA-A*02 multimer with CMV peptide NLVPMVATV(NLV), FACS sorted the stained cells, generated the TCR libraries, and sequenced them by Illumina. We identified 25 β and 21 α distinct clonotypes and searched for their exact CDR3 amino acid matches in other individual datasets. Nine β and 12 α matched clonotypes were found in at least one other individual (Table S4). Interestingly, the matched clonotypes were found in the datasets of C1/C2 twins despite that they do not have the HLA-A*02 alleles. When comparing the V segments of matched CMV-NLV-specific donor A2 clonotypes, we found that individuals A1, D1, and D2 in many cases have the same V gene in addition to the same CDR3β (5/6 for A1, 4/6 for D1, and 3/5 for D2). In contrast, none of the matched CDR3β sequences from C1 and C2 had the same V segment. Thus,

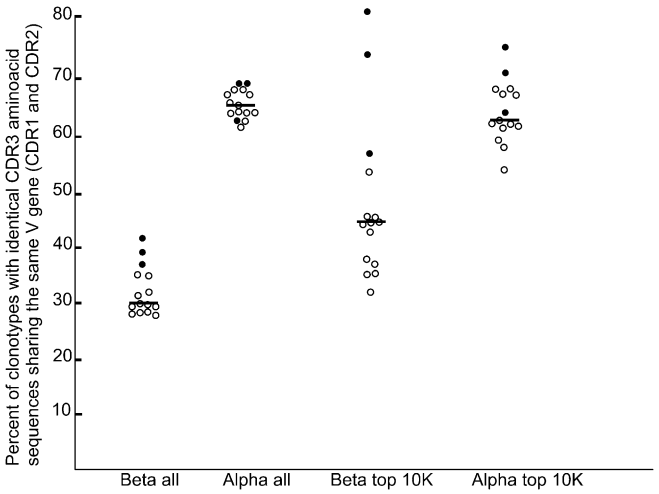

Fig. 5. The percentage of clonotypes with identical CDR3 amino acid sequences sharing the same V gene for all possible pairs of individuals. Black circles indicate twin pairs, empty circles indicate all other pairs, and lines indicate median values. β all, among all identified TCRβ clonotypes; α all, among all identified TCRα clonotypes, β top 10K, among 10,000 most abundant TCRβ clonotypes; α top 10K, among 10,000 most abundant TCRα clonotypes.

we can suggest that these clonotypes recognize the same CMV peptide (as they have identical CDR3) but presented by another MHC molecule (as they have different V segments coding for the CDR1/CDR2 that recognize MHC) (28). No such observation was made for the α chain where identical/different V segments among matched CMV-positive CDR3s are distributed quite randomly between individuals. For donor A2, we calculated the portion of HLA-A02-NLV-specific CDR3 clonotypes among all of the clonotypes shared with each other individual for several abundance categories (top 10,000 most abundant clonotypes in both donors, top 30,000, etc.; Fig. S3*A*). These data suggest that more identical CMV-specific T cells are present among abundant clonotypes. However, the absolute numbers of such clonotypes are too low to make confident conclusions. To get deeper insight into this issue, we generated a list of 257 TCRβ CDR3 sequences known to recognize different CMV or EBV peptides in complex with a number of MHC I variants (HLA-A*02; HLA-B*07, B*08; at least one allele is present in every individual; Table S1) from the previously published works (*SI Materials and Methods*). Then we compared the list with the clonesets of all six individuals from this work. Sixty-nine of 257 CMV/EBV-specific amino acid CDR3 sequences were found in at least one individual, and 6 sequences were found in all six individuals. Next, we calculated the percent of CMV/EBV-specific CDR3s among clonotypes shared by each pair of six individuals (Fig. S3*B*). The portion of such CDR3 sequences is again higher in the fraction of abundant clonotypes. The higher percentage of CMV/EBV-specific clonotypes suggests that a significant portion of abundant clonotypes in the human organism may be represented by common virus-specific TCRs. More generally, the higher proportion of shared clones among abundant clonotypes could be a consequence of peripheral selection and expansion of antigen-specific clonotypes driven by persistent pathogens. We observed no significant difference between twins and nonrelated individuals for this parameter.

Discussion

To summarize the results of the current study, we conclude the following.

V gene segment selection for TCR β and α chains is very similar in twin pairs. This result could be expected because twins have identical sets of MHC molecules and V segments coding for MHC recognizing motifs (29, 30). However, we found that not only the TCRs of cells affected by thymic selection (in-frame TCRs), but also the TCRs formed before selection (out-of-frame TCRs) are strictly dependent on individual genetic factors. Each of the analyzed twin pairs has a much closer distribution of V gene usage in the out-of-frame clonesets than any pair of random individuals, as evaluated by JS divergence. The same conclusion is valid for V gene segments usage in α chain TCRs. These findings indicate that the selection of variable segments for the mature TCRs during recombination is determined by the genes coding for the TCR recombination machinery, as it was suggested previously (31).

In contrast, the selection of J gene segments for both α and β TCR chains at the initial recombination step (as evaluated from the out-of-frame clonesets) is not different in twin pairs and random pairs of individuals. This finding indicates that the selection of J gene segments for recombination is not strictly determined by genetic factors. The pattern of JS divergence of J segment distributions is dramatically changed for the in-frame α TCRs after thymic selection. For each of the six individuals, her twin becomes the closest one in α segment usage leaving behind all unrelated individuals. Moreover, we observe that JS divergences fall down for all pairs of individuals when moving from the out-of-frame to the in-frame Jα segment usage. At the same time, Jβ segment usage remains generally independent on the individuals' kinship. This finding most probably suggests that the part of the TCRα chain receptor encoded by the J gene segment is involved in the recognition of MHC molecules at the selection stage in the thymus, whereas the same part of the β chain molecule is not. However, this suggestion has to be validated by more independent studies.

An accurate estimation of the shared clonotypes amount when comparing large individual TCR repertoires appears to be quite a challenging task: missing particular points when performing such calculations can lead to incorrect conclusions and false discoveries. First, the compared datasets should be generated using the same protocol of sample preparation, comparable depth of sequencing, and TCR sequence extraction algorithms. Additionally, many PCR cycles and millions of sequencing reads inevitably generate a huge artificial TCR diversity and lead to an incorrect evaluation of the cloneset size. This artificial diversity can be partially avoided by using proper error correction algorithms (12, 24). Second, possible cross-contamination between individual TCR datasets should be avoided by standard precautions (PCR boxes, clean rooms, etc.), as well as by the use of additional platform-specific strategies preventing molecular exchange in course of library preparation and sequencing (e.g., double sample barcoding; *SI Materials and Methods* and Fig. S1). Finally, the number of shared clonotypes between a pair of individuals can be compared correctly only if the comparable amount of starting material (T cells and isolated RNA or DNA) and the similar sequencing depth are achieved. For example, if a total of 10,000 clonotypes is identified for samples 1 and 2, they are mainly comprised by the expanded T-cell clones and have a higher proportion of public (common) clonotypes, whereas 200,000 sample 3 clonotypes would to a great extent consist of the naïve T cells and contain a lower portion of public TCRs. In this example the normalized (divided by the product of cloneset sizes) number of shared clonotypes will be higher for the sample pair 1/2 than for the pairs 1/3 and 2/3.

In this study, we clearly demonstrate that the number of CDR3 clonotypes shared between any pair of individuals is strictly dependent on the product of the intersected cloneset sizes for both α and β TCR chains. This finding is in agreement with the observation made for the out-of-frame (generated) repertoire by Murugan et al. (2). Surprisingly, MZ twins do not deviate from this trend. The lines depicting the linear relationship are approximated with $p = b_1 \times MN + b_0$, where p is the number of identical CDR3s in a pair of individuals; M and N are the sizes of the two intersected clonesets; and b_1 and b_0 are the slope and the intercept, respectively. Different values of b_1 specifying the slope of lines for β (1.472×10^{-7}) and for α chains (4.351×10^{-7}) indicate different potential diversity and thus different numbers of shared clonotypes for α and β chains. The more diverse β chain potentially has less shared clonotypes in a pair of individual repertoires because amino acid sequence convergence is less probable. An extrapolation of the line to the zero intersected cloneset sizes indicates that every pair of individuals will inevitably have some shared clonotypes and their number is reflected by the b_0 coefficient. Presence of such shared clonotypes can be partially explained by the existence of natural killer T (NKT) and mucosa-associated invariant T (MAIT) cells having the identical TCRα chain sequences in all humans (32, 33). Indeed, we found all of the CDR3 sequences originated from MAIT cells and 8 of 12 CDR3 sequences originated from NKT cells (32) in our datasets. Moreover, the majority of MAIT CDR3 sequences were present in all individuals within 10,000 most abundant clonotypes. The frequently formed public clonotypes against common persisting viruses like CMV and EBV can also contribute to the preexisting shared sequences. Thus, the real behavior of the line near zero is more complex: when a small number of clonotypes (less than ~10,000–20,000) is detected the chance to pick public clonotypes is higher, thus increasing the actual portion of shared clonotypes in a pair of individuals (as reflected in Fig. 4 and Fig. S2).

In general, deep TCR repertoires (mainly composed by low abundant T cells) of MZ twins have the same amount of shared CDR3 clonotypes compared with unrelated individuals. The shared clonotypes in all people are characterized by a lower number of nucleotides added at the germ-line segment junctions compared with individual repertoires, as shown previously

(20, 34, 35). Interestingly, individual TCR repertoires of mice (nearly genetically identical within the same strain) share at least 10 times more CDR3 clonotypes (36) than human pairs. This fact is probably explained by the overall lower number of added nucleotides in mice TCRs [around 4.0 (34)] providing a higher chance to generate identical CDR3 independently. The observed percentage of identical V segments among TCRs with identical CDR3 shared by twins is also similar to unrelated pairs.

In contrast, the most abundant clonotypes (supposed to mainly reflect the antigen experienced T cells) shared by identical twins show some characteristics of β chain that clearly distinguish them from the ones shared by random individuals. First, the proportion of shared CDR3 sequences among abundant β clonotypes is significantly higher in twin pairs. Second, abundant clonotypes shared by twins have a higher number of added nucleotides compared with unrelated humans. Finally, the most abundant clonotypes with identical CDR3s share the same Vβ segments in twin pairs for β TCR chains more often. Most probably, these clonotypes in twins have an initial higher probability to rearrange and then are activated by the same antigens presented in the same MHC context. The results of the experiments with CMV-specific T cells sorting also favor this suggestion. However, it is not possible to exclude the influence of environmental factors: all three pairs of twins in this study were reared together and most probably met the same infectious agents, vaccines, etc. To further clarify this point deep TCR profiling of twins reared apart is needed. At the same time, we did not observe any significant difference in characteristics of most abundant TCRα chain clonotypes shared between the twins and between the unrelated individuals. This finding suggests that α chain CDR3 is less involved in specific antigenic peptides recognition.

Materials and Methods

A detailed description of experimental procedures is given in *SI Materials and Methods* and Fig. S4. Briefly, RNA was isolated from PBMCs of three female pairs of healthy MZ twins and used for TCR cDNA library preparation (see Table S5 for oligonucleotides used). cDNA libraries were sequences by Illumina. TCR clonotypes were generated using MiTCR software.

ACKNOWLEDGMENTS. This work was supported by the Molecular and Cellular Biology Program Russian Academy of Sciences, Russian Foundation for Basic Research (14-04-01062, 14-04-01823, 13-04-01124, and 14-04-01247), European Regional Development Fund (CZ.1.05/1.1.00/02.0068), Council of the President of the Russian Federation for Young Scientists (MD-3044.2014.4 and SP-2039.2012.4), and the Dmitry Zimin Dynasty Foundation.

1. Robins HS, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* 2(47):47ra64.
2. 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 USA* 109(40):16161–16166.
3. Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. *Nature* 334(6181):395–402.
4. Arstila TP, et al. (1999) A direct estimate of the human alphabeta T cell receptor diversity. *Science* 286(5441):958–961.
5. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114(19):4099–4107.
6. 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(5):790–797.
7. Miles JJ, Douek DC, Price DA (2011) Bias in the αβ T-cell repertoire: Implications for disease pathogenesis and vaccination. *Immunol Cell Biol* 89(3):375–387.
8. Fozza C, et al. (2012) T-cell receptor repertoire analysis in monozygotic twins concordant and discordant for type 1 diabetes. *Immunobiology* 217(9):920–925.
9. Gulwani-Akolkar B, Shalon L, Akolkar PN, Fisher SE, Silver J (1994) Analysis of the peripheral blood T-cell receptor (TCR) repertoire in monozygotic twins discordant for Crohn's disease. *Autoimmunity* 17(3):241–248.
10. Somma P, et al. (2007) Characterization of CD8+ T cell repertoire in identical twins discordant and concordant for multiple sclerosis. *J Leukoc Biol* 81(3):696–710.
11. Utz U, et al. (1993) Skewed T-cell receptor repertoire in genetically identical twins correlates with multiple sclerosis. *Nature* 364(6434):243–247.
12. Bolotin DA, et al. (2012) Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur J Immunol* 42(11): 3073–3083.
13. Britanova OV, et al. (2012) First autologous hematopoietic SCT for ankylosing spondylitis: A case report and clues to understanding the therapy. *Bone Marrow Transplant* 47(11):1479–1481.
14. 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(10): 1817–1824.
15. Mamedov IZ, et al. (2011) Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol Med* 3(4):201–207.
16. Nguyen P, et al. (2011) Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 12:106.
17. Sherwood AM, et al. (2011) Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after αβ and γδ T cell commitment. *Sci Transl Med* 3(90):90ra61.
18. Venturi V, et al. (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* 186(7):4285–4294.
19. Wang C, et al. (2010) High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci USA* 107(4): 1518–1523.
20. Putintseva EV, et al. (2013) Mother and child T cell receptor repertoires: Deep profiling study. *Front Immunol* 4:463.
21. Shugay M, et al. (2013) Huge Overlap of Individual TCR Beta Repertoires. *Front Immunol* 4:466.
22. Turchaninova MA, et al. (2013) Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* 43(9):2507–2515.
23. Mamedov IZ, et al. (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol* 4:456.
24. Bolotin DA, et al. (2013) MiTCR: Software for T-cell receptor sequencing data analysis. *Nat Methods* 10(9):813–814.
25. Krangel MS (2009) Mechanics of T cell receptor gene rearrangement. *Curr Opin Immunol* 21(2):133–139.
26. Weischenfeldt J, et al. (2008) NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev* 22(10):1381–1396.
27. Quigley MF, et al. (2010) Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci USA* 107(45):19414–19419.
28. Koning D, et al. (2013) CD8+ TCR repertoire formation is guided primarily by the peptide component of the antigenic complex. *J Immunol* 190(3):931–939.
29. Garcia KC, Adams JJ, Feng D, Ely LK (2009) The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* 10(2):143–147.
30. Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* 24:419–466.
31. Melenhorst JJ, et al. (2008) Contribution of TCR-beta locus and HLA to the shape of the mature human Vbeta repertoire. *J Immunol* 180(10):6484–6489.
32. Greenaway HY, et al. (2013) NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination. *Immunobiology* 218(2):213–224.
33. Venturi V, Rudd BD, Davenport MP (2013) Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr Opin Immunol* 25(5):639–645.
34. Li H, et al. (2012) Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. *J Immunol* 189(5):2404–2413.
35. Venturi V, et al. (2006) Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci USA* 103(49):18691–18696.
36. Ndifon W, et al. (2012) Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci USA* 109(39):15865–15870.