Supporting Information

Zvyagin et al. 10.1073/pnas.1319389111

SI Materials and Methods

Blood Samples, RNA/DNA Isolation, and HLA Typing. This study was approved by the local ethics committee and conducted in accordance with the Declaration of Helsinki. All donors were informed of the final use of their blood and signed an informed consent. Ten milliliters of peripheral blood was obtained from each of six systemically healthy donors including three pairs of monozygous twins (Caucasian population, Russia). Three female pairs aged 21 (D1-D2), 27 (A1-A2), and 46 (C1-C2) years were used for this study. Peripheral blood was collected into EDTA-treated Vacutainer tubes (BD Biosciences). Peripheral blood mononuclear cells (PBMCs) (at least 10^7 per each sample) were isolated by Ficoll-Paque (Paneco) density gradient centrifugation. Total RNA was isolated using TRIzol reagent (Invitrogen) according to the manufacturer's protocol. A small aliquot of cells was used for genomic DNA isolation using the standard phenol-chlorophorm method. HLA typing was performed using the AllSet Gold SSP HLA-ABDRDQ Low Res Kit and HLA-Cw Low Res Kit (Invitrogen).

Sample Preparation. All of the RNA (on average 10 mkg for each of six samples) obtained from 10 mL of whole blood was used for cDNA synthesis. First-strand cDNA was synthesized for 2 h using the Mint cDNA synthesis kit (Evrogen) with primers bc1R (specific to both constant regions of the human TCRβ; for primers sequences, see Table S5) and ac1R (specific to the constant region of the human TCRα) according to the manufacturer's protocol. Each tube contained 2.5 mkg of total RNA. Plug oligo (Evrogen) was added after 30 min of synthesis.

All the first-strand cDNA obtained from the synthesis was used as template for the first PCR amplification. Each tube (PCR mixture, volume 15 mkL) contained 1× Encyclo polymerase buffer (Evrogen), 0.125 mM of each dNTP, 5 pmol of primers bc2R and ac2R, and one of the primers Na21-SB-M1 with the 5-nt unique sample barcode (SB) for each sample, as well as 0.3 mkL of Encyclo polymerase mix, and 1 mkL of undiluted first-strand cDNA. The PCR amplification protocol was as follows: 94 °C for 20 s; 65 °C for 20 s; and 72 °C for 50 s for 18 cycles. After the first PCR step, the whole PCR product for each sample was mixed and purified using the QIAquick PCR purification kit (Qiagen). One microliter of the purified PCR product was used in each of the 10 second step PCR reactions (25 mkL each) to generate the library of TCRβ or TCRα. The PCR mixture (each tube) contained 1× Encyclo polymerase buffer (Evrogen), 0.125 mM of each dNTP, 10 pmol of primer NNa, and either primer ac3R-SB (for TCRα chain library) or primer bc3R-SB (for TCRβ chain library), as well as 0.3 mkL of Encyclo polymerase mix and 1 mkL of undiluted first-strand cDNA. The same sample barcode as in first PCR was used for each sample at this stage.

Sorting for CMV NV9-Specific T Cells. PBMCs were obtained from donor A2 as described above. Cells were stained with CD8-PC7 (BeckmanCoulter) and PeliMer A2/CMV (pp65) PE [HLA-A*0201 NLVPMVATV (NLV); Sanquin] according to the manufacturer's protocol and sorted using FACSAria III (BD Bioscience) directly in TRIzol reagent. RNA was isolated from CD8$^+$/HLA-A*02-NLV$^+$ and CD8$^+$/HLA-A*02-NLV$^-$ fractions as described above. T-cell receptor (TCR) cDNA libraries were generated as described above and sequenced by Illumina (MiSeq 2x150). Data analysis was performed as described below. TCR clonotypes from the CD8$^+$ /HLA-A*02-NLV$^+$ fraction that were identical to highly abundant sequences from the CD8$^+$/HLA-A*02-NLV$^-$ fraction (negative control) were discarded as potential false positives.

Search for Other CMV- or EBV-Specific Complementarity Determining Region 3 Sequences. The list of 257 CMV or EBV recognizing complementarity determining region 3 (CDR3) sequences was obtained from refs. 1–9.

NGS and Data Analysis. The libraries were mixed, and TruSeq adaptors were ligated and paired end (2 × 125) sequenced on tree lanes of Illumina GA IIx. The raw sequencing data are available at NCBI sequence read archive (SRP028752). Sequences were separated into 12 datasets (6 α and 6 β) based on the sample barcodes introduced in the course of library preparation. Sequencing reads having nonidentical SBs were removed. This system of double sample barcoding allows protection of data from sequence exchange between datasets. The double barcoding is crucial for the studies where the exact number of shared clonotypes between samples is evaluated. Interestingly, the percentage of sequencing reads having distinct barcodes was significantly higher for samples sequenced in the same Illumina lane, probably indicating some ends exchange during bridge amplification. Clonotypes were assembled using MiTCR software (10) with default parameters (quality threshold for each nucleotide within the CDR3 was set as Phred >25; the strictest "eliminate these errors" correction algorithm was used).

Statistical Analysis. ***Jensen–Shannon divergence.*** To compare the clonotype V and J gene usage distribution in pairs of individuals, we used the Jensen–Shannon divergence (JS), which is a symmetrized and smoothed version of the Kullback–Leibler divergence, used to quantify the similarity between two probability distributions. JS is defined as follows (11):

$$JS(P,Q)=\frac{1}{2}\sum_{i}^{n}\log_2\left[\left(\frac{P_i}{\frac{1}{2}(P_i+Q_i)}\right)\right]\cdot P_i + \frac{1}{2}\sum_{i}^{n}\log_2\left(\frac{Q_i}{\frac{1}{2}(P_i+Q_i)}\right)\cdot Q_i.$$

Entropy was calculated using the formula

$$H(P)=-\sum_{i}^{n}P_i\cdot\log_2 P_i.$$

JS distances were divided by mean entropy of two distributions. The calculated entropy values for different data categories were as follows: 4.70–4.98 for Vβ out-of-frames, 4.79–4.96 for Vβ in-frames, 4.99–5.11 for Vα out-of-frames, 5.10–5.13 for Vα in-frames; 3.34–3.39 for Jβ out-of-frames, 3.31–3.40 for Jβ in-frames, 5.46–5.49 for Jα out-of-frames; and 5.49–5.52 for Jα in-frames.

Bootstrap estimates for JS divergence SD were computed by performing random sampling of half of clonotypes from both clonesets, calculating the JS divergence between these subsamples, and repeating the procedure 100 times.

To evaluate the minimal number of clonotypes for a reliable JS estimate, we performed the rarefaction analysis. Random clonotype samples of increasing size were selected from clonesets of individual A1 and JS distance to the cloneset of individual C1 was calculated (Fig. S4). We concluded that 4,000 clonotypes are enough to evaluate the JS distance correctly.

Linear regression. Linear regression analysis was performed using R programming language (12). Linear models $p = b_1 \times MN + b_0$ (where p is the number of identical CDR3s between a pair of individuals, M and N are the sizes of two intersected clonesets, b_1 is the slope, and b_0 is intercept) were fit using the least-squares method. Using regression analysis, we estimated $b_1 = 1.472 \times 10^{-7}$ (98% CI: 1.411×10^{-7}, 1.534×10^{-7}) and $b_0 = 3{,}072$ (98% CI: 2,563, 3,581) for β chains and $b_1 = 4.351 \times 10^{-7}$ (98% CI: 3.860×10^{-7}, 4.842×10^{-7}), $b_0 = 10{,}790$ (98% CI: 7,066, 14,520) for α chains, respectively. Adjusted R^2 for α and β chain intersection linear models was 0.9752 and 0.9814, respectively. Predictive intervals and confidence intervals for b_0 and b_1 were computed using R programming language.

Data from Warren et al. (13) was not used for model fit and predictive interval construction.

1. Cohen GB, et al. (2002) Clonotype tracking of TCR repertoires during chronic virus infections. *Virology* 304(2):474–484.
2. Dong L, Li P, Oenema T, McClurkan CL, Koelle DM (2010) Public TCR use by herpes simplex virus-2-specific human CD8 CTLs. *J Immunol* 184(6):3063–3071.
3. Iancu EM, et al. (2009) Clonotype selection and composition of human CD8 T cells specific for persistent herpes viruses varies with differentiation but is stable over time. *J Immunol* 183(1):319–331.
4. Lehner PJ, et al. (1995) Human HLA-A0201-restricted cytotoxic T lymphocyte recognition of influenza A is dominated by T cells bearing the V beta 17 gene segment. *J Exp Med* 181(1):79–91.
5. Miconnet I, et al. (2011) Large TCR diversity of virus-specific CD8 T cells provides the mechanistic basis for massive TCR renewal after antigen exposure. *J Immunol* 186(12): 7039–7049.
6. Miles JJ, et al. (2005) T-cell grit: Large clonal expansions of virus-specific CD8+ T cells can dominate in the peripheral circulation for at least 18 years. *Blood* 106(13): 4412–4413.
7. Price DA, et al. (2005) Avidity for antigen shapes clonal dominance in CD8+ T cell populations specific for persistent DNA viruses. *J Exp Med* 202(10):1349–1361.
8. Schwanninger A, et al. (2008) Age-related appearance of a CMV-specific high-avidity CD8+ T cell clonotype which does not occur in young adults. *Immun Ageing* 5:14.
9. Venturi V, et al. (2008) TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J Immunol* 181(11):7853–7862.
10. Bolotin DA, et al. (2013) MiTCR: Software for T-cell receptor sequencing data analysis. *Nat Methods* 10(9):813–814.
11. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37(1):145–151.
12. R Core Team (2012) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna). Available at www.R-project.org.
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(5):790–797.

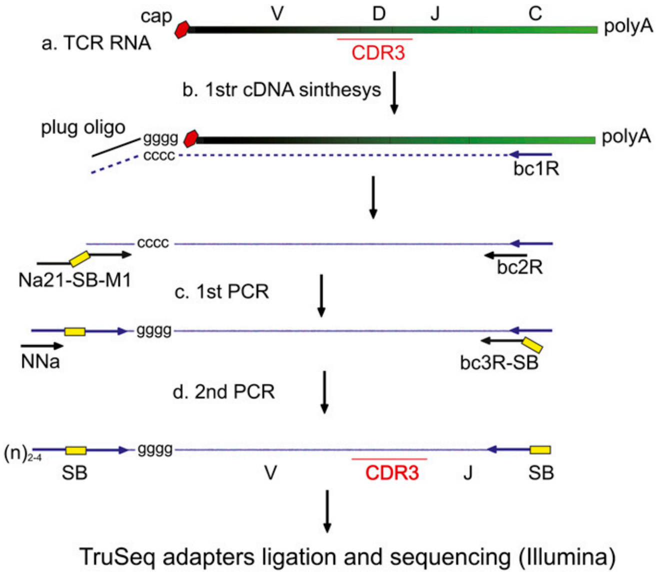

Fig. S1. Scheme of TCR library preparation (detailed for TCRβ). (*A*) Structure of human TCRβ RNA. V, D, J, and C genomically encoded segments. (*B*) First-strand cDNA synthesis is performed with the primer (bc1R) specific to TCR constant (C) region. Plug oligo is used for template switch and generates a universal primer annealing site for each molecule. (*C*) First PCR amplification step is performed with nested primer (bc2R) corresponding to the TCR constant region and Na21-SB-M1 primer with a 5nt sample-specific barcode. (*D*) Second PCR amplification step is performed with nested primer (bc3R-SB) corresponding to the junction of TCR C and J regions and NNa primer. bc3R-SB primer introduce the second sample barcode identical to the one introduced in the first amplification step for each individual sample. Finally each TCR library obtained from each donor is marked by the same individual sample barcode at both ends. This double barcoding system allows elimination of sequence reads resulting from intersample exchange occurring during the course of amplification after Illumina adapter ligation or bridge amplification on the solid phase. NNa primer is composed of a mixture of three almost identical primers having two, three, and four N on their 5′ end. This approach allows to reduce low complexity on the end of the library that is critical for Illumina sequencing.

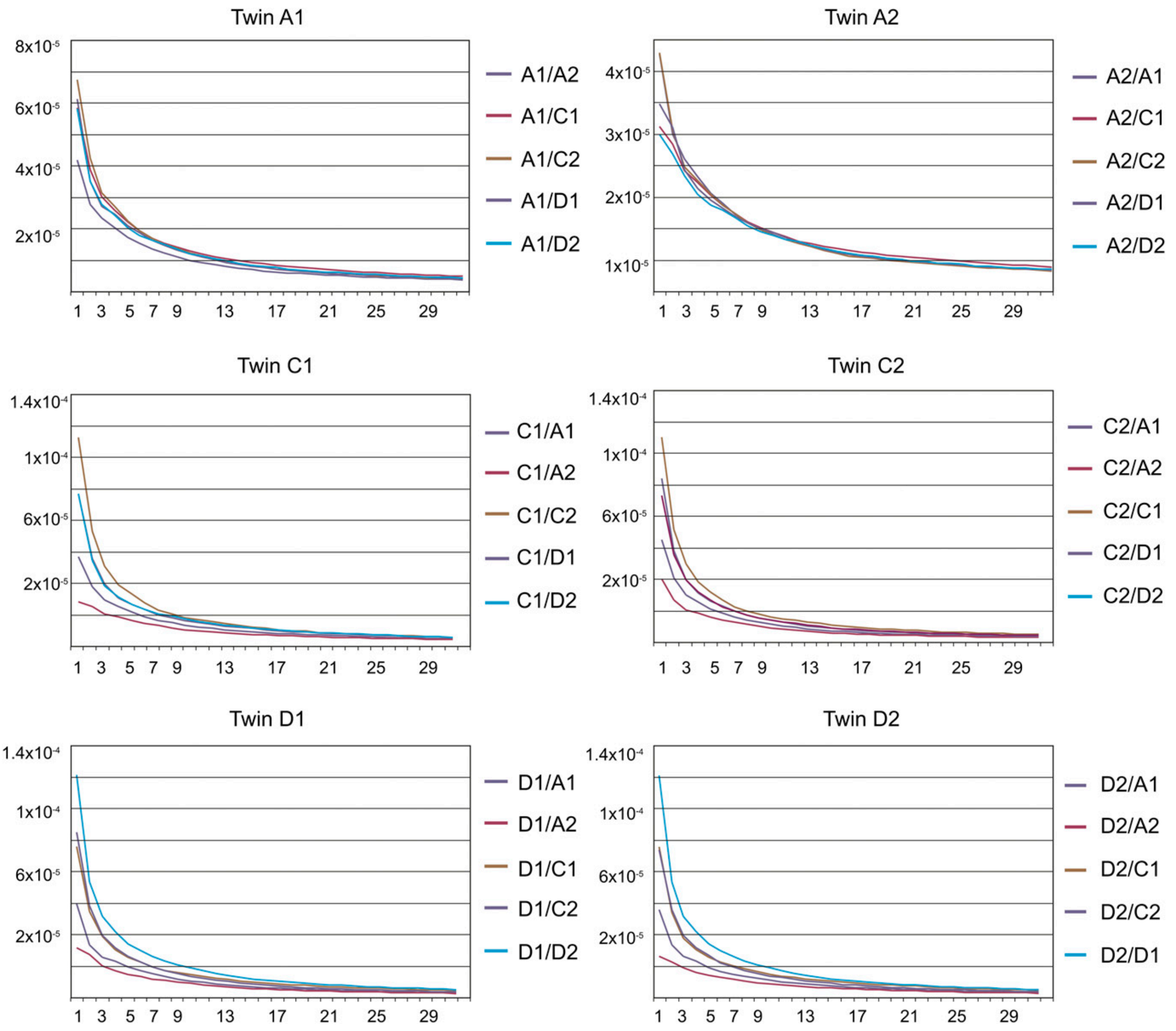

Fig. S2. Normalized number of identical TCRα CDR3 amino acid sequences for each possible pair of individuals among the 1,000 most abundant clonotypes, 2,000 most abundant clonotypes, etc. *x* axis, the number of most abundant clonotypes (×1,000) intersected for each of two individuals; *y* axis, normalized number of shared clonotypes.

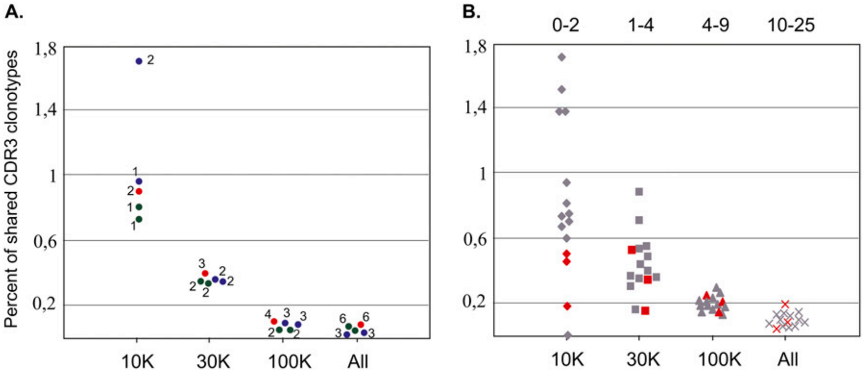

Fig. S3. Percent of shared clonotypes pools composed by cytomegalovirus (CMV) and Epstein-Barr virus (EBV)-specific sequences. (*A*) Each dot indicates the percent (*y* axis) of CDR3β clonotypes shared between individual A2 and each of the other five individuals comprised by HLA-A*02/CMV-NV9 specific clonotypes (identified by FACS experiment; this work). Absolute numbers of HLA-A*02/CMV-NV9 specific clonotypes shared between each pair are indicated near each dot. Red dot, A1; blue dots, C1 and C2; green dots, D1 and D2. Results are shown for different cohorts intersection: 10K, 10,000 most abundant (top) clonotypes from one individual are intersected with 10,000 most abundant clonotypes from another individual; 30K, top 30,000 from each; 100K, top 100,000 from each; all, the entire repertoires are compared. (*B*) The percent (*y* axis) of clonotypes shared between any possible pair of individuals occupied by sequences identified as CMV or EBV specific in nine previously published articles. Intervals of absolute numbers of CMV- or EBV-specific specific clonotypes shared between pairs for different cohorts are indicated at the top of each column. Red dots, triangles, and crosses indicate twin pairs.

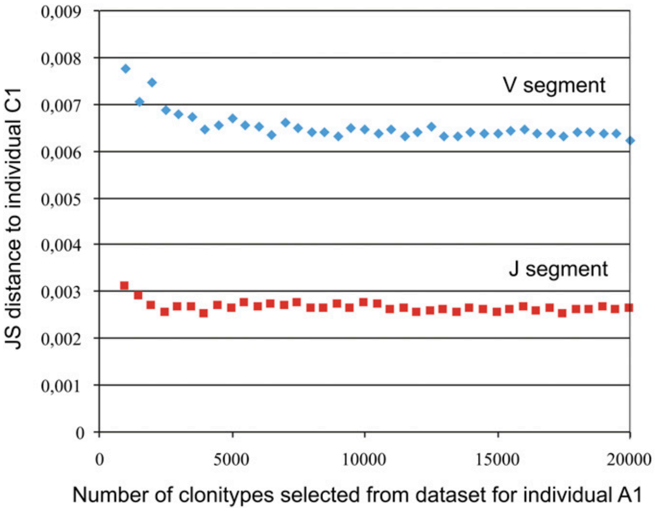

Fig. S4. Rarefaction analysis for JS distance between individuals A1 and C1 (β chain). *x* axis, number of clonotypes selected randomly from the dataset for individual A1; *y* axis, JS distance between individuals A1 and C1 (blue, Vβ segment; red, Jβ segment).

Table S1. Results of MHC I HLA typing

Twin	HLA-A(1)	HLA-A(2)	HLA-B(1)	HLA-B(2)	HLA-C(1)	HLA-C(2)
A1	A*02	A*02	B*14	B*15	C*03	C*08:25
A2	A*02	A*02	B*14	B*15	C*03	C*08:25
C1	A*01	A*32	B*08/B*15:180	B*44	C*05	C*05:43/C*07
C2	A*01	A*32	B*08/B*15:180	B*44	C*05	C*05:43/C*07
D1	A*02/ A*68	A*68	B*07	B*41	C*07	C*17
D2	A*02/ A*68	A*68	B*07	B*41	C*07	C*17

Table S2. Reads and clones distribution per sample

Twin	*N* of reads	Good sequences*	Number of clonotypes[†]	Out-of-frame clonotypes number[‡] (%)
A1α	3,522,985	2,549,023	211,876	25,519 (12.0)
A2α	5,193,253	2,769,645	110,504	14,622 (13.2)
C1α	3,586,583	2,933,398	457,361	52,804 (11.5)
C2α	4,331,900	3,680,532	396,306	81,063 (20.4)
D1α	3,821,093	2,284,140	350,622	38,162 (10.9)
D2α	5,058,737	3,109,265	397,690	41,814 (10.5)
A1β	3,406,486	2,549,023	246,499	8,747 (3.5)
A2β	3,512,006	2,769,645	109,685	4,179 (3.8)
C1β	4,003,193	2,933,398	587,935	22,276 (3.8)
C2β	5,017,740	3,680,532	401,612	31,537 (7.8)
D1β	3,068,713	2,284,140	371,390	10,191 (2.7)
D2β	4,195,262	3,109,265	556,509	15,015 (2.7)

Interestingly, the individual C2 has a significantly greater than average number of out-of-frame sequences for both α (23.4%) and β (8.2%) TCR chains. The increased number of out-of-frame sequences probably results from some individual characteristic of the nonsense-mediated mRNA decay (NMD), which down-regulates the out-of frame TCR mRNA.

*Sequences with CDR3 and V, J segments identified.

[†] Here "clonotype" is the cohort of sequences with identical nucleotide sequence of CDR3 V and J segments.

[‡] True out-of-frame clonotypes: after filtering out of erroneous sequences comprising in-frame clonotypes with insertion/deletion.

Table S3. Number of amino acid CDR3 sequences shared between each pair of individuals

Twin (number of clonotypes)*	A1 (228,772)	A2 (102,989)	C1 (527,428)	C2 (367,959)	D1 (337,788)	D2 (499,671)
β Chain TCR						
A1 (228,772)	NA	6,906	20,451	15,654	15,109	19,915
A2 (102,989)	6,906	NA	10,530	8,030	7,935	10,208
C1 (527,428)	20,451	10,530	NA	31,224	30,076	39,479
C2 (367,959)	15,654	8,030	31,224	NA	21,259	28,050
D1 (337,788)	15,109	7,935	30,076	21,259	NA	31,683
D2 (499,671)	19,915	10,208	39,479	28,050	31,683	NA
Twin (number of clonotypes)*	A1 (178,958)	A2 (95,922)	C1 (368,081)	C2 (330,996)	D1 (285,550)	D2 (320,701)
α Chain TCR						
A1 (178,958)	NA	18,434	41,413	36,423	34,943	36,975
A2 (95,922)	18,434	NA	25,440	22,748	21,746	22,954
C1 (368,081)	41,413	25,440	NA	61,472	58,402	62,082
C2 (330,996)	36,423	22,748	61,472	NA	49,917	53,077
D1 (285,550)	34,943	21,746	58,402	49,917	NA	56,682
D2 (320,701)	36,975	22,954	62,082	53,077	56,682	NA

*Here "clonotype" is the cohort of sequences with identical amino acid sequence of CDR3. NA, not applicable.

Table S4. Sequences of TCRs (CDR3) from T cells sorted for HLA-A*02 CMV-NV9 multimer from individual A2 and found in other individuals

*CDR3 sequence	A1	A2	C1	C2	D1	D2
β Chain CDR3						
CASSSANYGYTF	*241,444 (v12-3)*[†]	*26 (v12-3)*	1,325 (v6-2)	14,543 (v27)	8,027 (v28)	31,132 (v7-9)
CAWVPGTGGTEAFF		*28 (v30)*			*313,029 (v30)*	*163,040 (v30)*
CASSPGLPYEQYF	*20,412 (v12-3)*	*248 (v12-3)*		211,319 (v13)		
CASSLRGGGDTQYF		*99 (v7-3)*	245,440 (v2)			462,474 (v28)
CASSRVPSEQFF	40,995 (v7-2)	*4 (v7-2)*			*60,761 (v7-2)*	*21,954 (v7-2)*
CATDAGQGLFYGYTF		*116 (v6-5)*			*293,725 (v6-5)*	
CASSALGGAGTGELFF	*890 (v9)*	*455 (v9)*				
CASSLTGNTEAFF	4,498 (v12-3)	*576 (v5-1)*	4,272 (v7-9)	5,748 (v27)	15,975 (v7-3)	*8,625 (v5-1)*
CSVGRAQNEQFF	*187,056 (29-1)*	*14 (29-1)*			270,889 (29-1)	
α Chain CDR3						
CAVRSNFGNEKLTF	2,646 (v21)	37 (v41)	23,206 (v21)	15,779 (v21)	15,132 (v12-2)	11,770 (v21)
CAGPMKTSYDKVIF	*563 (v35)*	*66 (v35)*	72,993 (v27)			*67,449 (v35)*
CAFNDYKLSF	*2,957 (v24)*	*7 (v24)*	*8,808 (v24)*	*4,701 (v24)*	*5,110 (v24)*	3,468 (38-1)
CASFNTGNQFYF	57,527 (v24)	169 (v12-3)				
CAPPEGGATNKLIF		*3,617 (1-2)*				*9,502 (v21)*
CAVRDIRLMF		*245 (v3)*		*182,339 (v3)*		
CAVDIETSGSRLTF		*158 (v39)*			*23,414 (v39)*	*44,478 (v39)*
CAASRDQGAQKLVF	*88,617 (13-1)*	*11 (13-1)*	*72,751 (13-1)*	*269,897 (13-1)*	*306,737 (13-1)*	81,561 (v23)
CAASKDGGFKTIF	*22,792 (13-1)*	*50 (13-1)*	*100,647 (13-1)*	*64,777 (13-1)*	*10,555 (13-1)*	292,094 (v23)
CAVRDTDARLMF	*204,606 (v3)*	*189 (v3)*				
CAASILTGGGNKLTF	*3,043 (13-1)*	*297 (13-1)*	*1,798 (13-1)*	*654 (13-1)*	*943 (13-1)*	*22,896 (13-1)*
CAVRDTRLMF		2,693 (v3)	431,941 (1-2)	22,344 (1-2)		

The complete set of clonotypes from T cells sorted with HLA-A*02 CMV-NV9 multimer from individual A2 is available at http://labcfg.ibch.ru/tcr.html#MZTwins.

*CDR3 sequence is given from conservative cysteine (C) to conservative phenylalanine (F).

[†]For each CDR3, rank (i.e., the position in the list of all individual clonotypes arrange by number of sequencing reads from abundant to rare) and V gene (in parentheses) are given. Cells with V gene matched between A2 and any other in individual are italic.

Table S5. Oligonucleotides used for library preparation

Name	Sequence	TCR chain	Application
ac1R	ACACATCAGAATCCTTACTTTG	α	cDNA synthesis
bc1R	CAGTATCTGGAGTCATTGA	β	cDNA synthesis
ac2R	TACACGGCAGGGTCAGGGT	α	First PCR
bc2R	TGCTTCTGATGGCTCAAACAC	β	First PCR
Na-SB2-M1	CGAGCGTGACGACGACAGTAGTCGTGGTATCAACGCAGAGTAC	α + β	First PCR
Na-SB3-M1	CGAGCGTGACGACGACAGACTTCGTGGTATCAACGCAGAGTAC	α + β	First PCR
Na-SB4-M1	CGAGCGTGACGACGACAGTCACTGTGGTATCAACGCAGAGTAC	α + β	First PCR
Na-SB5-M1	CGAGCGTGACGACGACAGGATTCGTGGTATCAACGCAGAGTAC	α + β	First PCR
Na-SB6-M1	CGAGCGTGACGACGACAGGTCTTGTGGTATCAACGCAGAGTAC	α + β	First PCR
Na-SB7-M1	CGAGCGTGACGACGACAGAGTCTGTGGTATCAACGCAGAGTAC	α + β	First PCR
NNa	$(N)_{2-4}$CGAGCGTGACGACGACAG	α + β	Second PCR
ac3R-SB2	TAGTCGGGTCAGGGTTCTGGATAT	α	Second PCR
ac3R-SB3	ACTTCGGGTCAGGGTTCTGGATAT	α	Second PCR
ac3R-SB4	TCACTGGGTCAGGGTTCTGGATAT	α	Second PCR
ac3R-SB5	GATTCGGGTCAGGGTTCTGGATAT	α	Second PCR
ac3R-SB6	GTCTTGGGTCAGGGTTCTGGATAT	α	Second PCR
ac3R-SB7	AGTCTGGGTCAGGGTTCTGGATAT	α	Second PCR
bc3R-SB2	TAGTCACACRTTKTTCAGGTCCTC	β	Second PCR
bc3R-SB3	ACTTCACACRTTKTTCAGGTCCTC	β	Second PCR
bc3R-SB4	TCACTACACRTTKTTCAGGTCCTC	β	Second PCR
bc3R-SB5	GATTCACACRTTKTTCAGGTCCTC	β	Second PCR
bc3R-SB6	GTCTTACACRTTKTTCAGGTCCTC	β	Second PCR
bc3R-SB7	AGTCTACACRTTKTTCAGGTCCTC	β	Second PCR

SBs are marked in red. Each sample barcode pair differs by a minimum of two nucleotides.