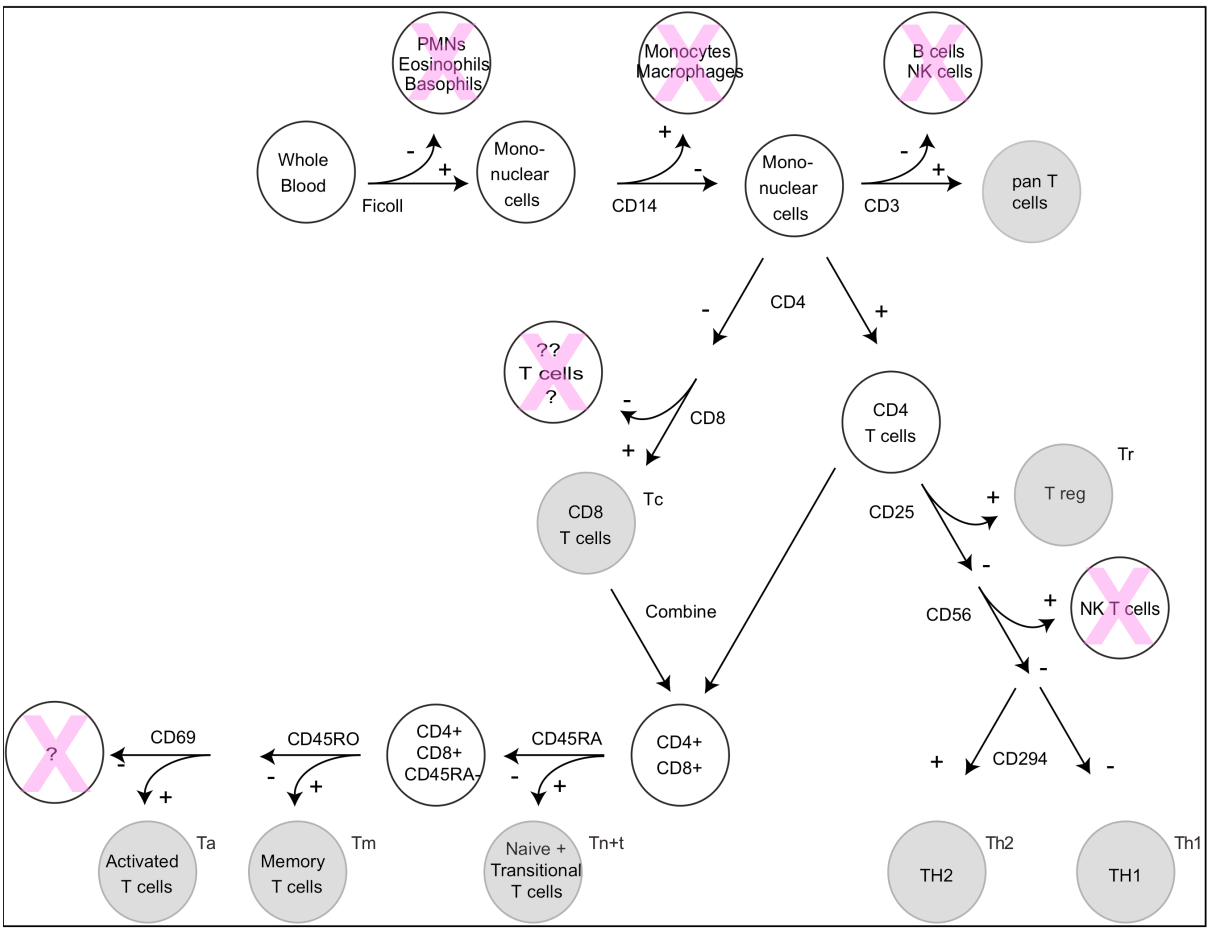

Fig. S1. Schematic diagram for isolating T cell subsets. Tr—T regulatory cell (CD4+CD25+); Th1—T helper cell 1 (CD4+CD25-CD294-); Th2—T helper cell 2 (CD4+CD25-CD294+); Tc—T cytotoxic cell (CD8+) Tn+t—naïve and transitional T cell (CD45RA+); Ta—activated T cell (CD45-RO-CD69+); Tm—memory T cell (CD45RA-RO+)

Table S1. The purity of MACS isolation protocol based on manufactory manual.

Reagent	Marker	Purity
Miltenyi Biotec	**CD25**	>90%
	CD69	>95%
	CD56	>95%
	CD294	>90%
	CD45RA	>95%
	CD45RO	>90%
	CD8 Multisort	>90%
	CD4 Multisort	>90%
Dynabeads Invitrogen	**CD3**	>98%

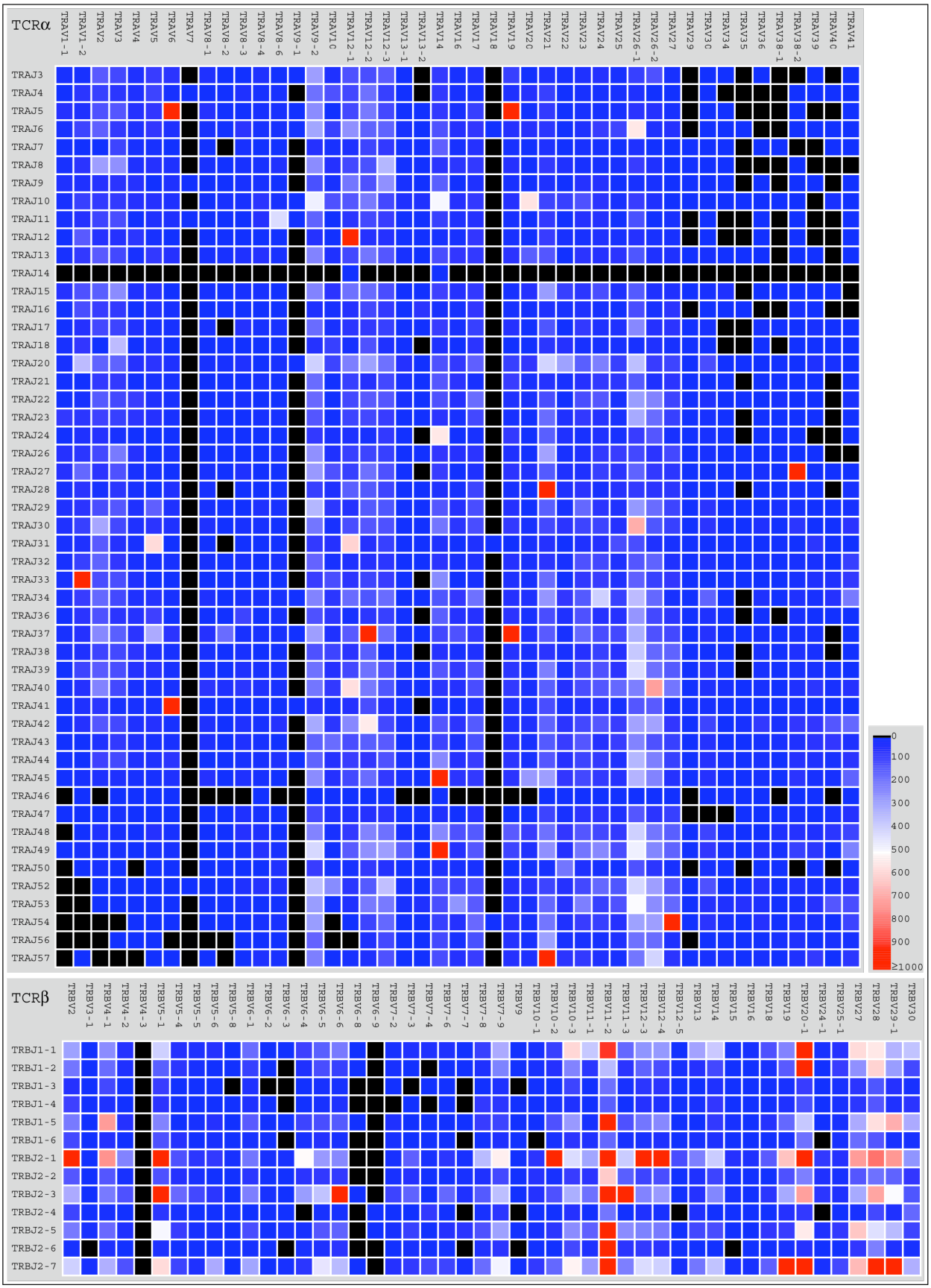
TCRα
TRAV1-1
TRAV1-2
TRAV2
TRAV3
TRAV4
TRAV5
TRAV6
TRAV7
TRAV8-1
TRAV8-2
TRAV8-3
TRAV8-4
TRAV8-6
TRAV9-1
TRAV9-2
TRAV10
TRAV12-1
TRAV12-2
TRAV12-3
TRAV13-1
TRAV13-2
TRAV14
TRAV16
TRAV17
TRAV18
TRAV19
TRAV20
TRAV21
TRAV22
TRAV23
TRAV24
TRAV25
TRAV26-1
TRAV26-2
TRAV27
TRAV29
TRAV30
TRAV34
TRAV35
TRAV36
TRAV38-1
TRAV38-2
TRAV39
TRAV40
TRAV41
TRAJ3
TRAJ4
TRAJ5
TRAJ6
TRAJ7
TRAJ8
TRAJ9
TRAJ10
TRAJ11
TRAJ12
TRAJ13
TRAJ14
TRAJ15
TRAJ16
TRAJ17
TRAJ18
TRAJ20
TRAJ21
TRAJ22
TRAJ23
TRAJ24
TRAJ26
TRAJ27
TRAJ28
TRAJ29
TRAJ30
TRAJ31
TRAJ32
TRAJ33
TRAJ34
TRAJ36
TRAJ37
TRAJ38
TRAJ39
TRAJ40
TRAJ41
TRAJ42
TRAJ43
TRAJ44
TRAJ45
TRAJ46
TRAJ47
TRAJ48
TRAJ49
TRAJ50
TRAJ52
TRAJ53
TRAJ54
TRAJ56
TRAJ57
0
100
200
300
400
500
600
700
800
900
≥1000
TCRβ
TRBV2
TRBV3-1
TRBV4-1
TRBV4-2
TRBV4-3
TRBV5-1
TRBV5-4
TRBV5-5
TRBV5-6
TRBV5-8
TRBV6-1
TRBV6-2
TRBV6-3
TRBV6-4
TRBV6-5
TRBV6-6
TRBV6-8
TRBV6-9
TRBV7-2
TRBV7-3
TRBV7-4
TRBV7-6
TRBV7-7
TRBV7-8
TRBV7-9
TRBV9
TRBV10-1
TRBV10-2
TRBV10-3
TRBV11-1
TRBV11-2
TRBV11-3
TRBV12-3
TRBV12-4
TRBV12-5
TRBV13
TRBV14
TRBV15
TRBV16
TRBV18
TRBV19
TRBV20-1
TRBV24-1
TRBV25-1
TRBV27
TRBV28
TRBV29-1
TRBV30
TRBJ1-1
TRBJ1-2
TRBJ1-3
TRBJ1-4
TRBJ1-5
TRBJ1-6
TRBJ2-1
TRBJ2-2
TRBJ2-3
TRBJ2-4
TRBJ2-5
TRBJ2-6
TRBJ2-7

Fig. S2. Germline gene segment usage for TCRα (top) and TCRβ (bottom) in the pan T sample. TCR Dβ segments are not shown. Only functional reference gene fragments according to the IMGT database are presented here. The number of sequence reads is represented by color, where red is higher than 1000, blue is the as low as 1, and black is 0. Vα and Vβ genes were listed at the X-axis on top and bottom panels; functional Jα and Jβ genes were listed at the Y-axis on top and bottom panels. All functional germline gene segments except for TRBV4-3 listed in the IMGT reference database were mapped to sequence reads in the Pan T sample. The missing germline gene segment--TRBV4-3 was observed in subset samples with extremely low abundance.

Assess the PCR and sequencing performance. The ARM-PCR technique is an improvement of the TEM-PCR technique[1]. Both TEM-PCR and ARM-PCR techniques use a pair of universal primers at the exponential phase of PCR amplification. It was proven that TEM-PCR was able to semi-quantitatively amplify targets where the relative ratio among original targets was maintained in the final products[1]. The performance of PCR and sequencing was assessed indirectly here.

First, with a few exceptions, highly random patterns of germline VJ gene segment combinations were observed (Fig. S2). All functional germline gene segments except for TRBV4-3 listed in the IMGT reference database[2] were mapped to sequence reads in the panT sample. The missing germline gene segment--TRBV4-3 was observed in subset samples with extremely low abundance, suggesting that its scarcity in the panT sample did not reflect a bias in amplification or sequencing. Together, we identified 2505 VJ combinations in the pan T sample, which account for >87% of all the possible combinations between the functional germline Vα and Jα, and Vβ and Jβ gene segments as listed in the IMGT database[2].

Secondly, with the same PCR amplification and sequencing techniques, panT cells from two samples at different time points from the same individual were sequenced. The correlation of number of sequence reads with the same V and J segments was measured. Pearson's correlation coefficient value of 0.92 was observed, indicating a high degree of correlation (Fig. S3.A). On the other hand, Pearson's correlation coefficient of number of sequence reads with the same V and J segments between Th1 and Tc was 0.09, indicating a low degree of correlation (Fig. S3. B). A high degree of correlation between the two panT samples was expected because the two blood samples were collected from the same individual, although at different time points. Also, a low degree of correlation between Tc and Th1 samples was anticipated because TCRs on the surface of Tc cells recognize antigens presented by the class I MHC molecules, while TCRs on the surface of Th1 cells recognize antigens presented by the class II MHC molecules. The observed high degree of correlation between the two panT samples indicated a very good repeatability of both the ARM-PCR amplification and 454 sequencing techniques. Results do not

reflect systematic bias in amplification or sequencing as a low degree of correlation between Tc and Th1 samples was also observed.

Fig. S3.C and Fig. S3.D illustrated the diversity of sequences with the same V (TRAV26-1) and J (TRAJ30) segment, or amplified by the same PCR primer set. It is obvious that the ARM-PCR technique was not biased to one particular sequence as many different sequences (61 in the Tc subset sample and 145 in the Th1 subset sample) were amplified with various level of abundance. The dominance of one particular sequence in the Tc samples is highly likely due to biological reason.

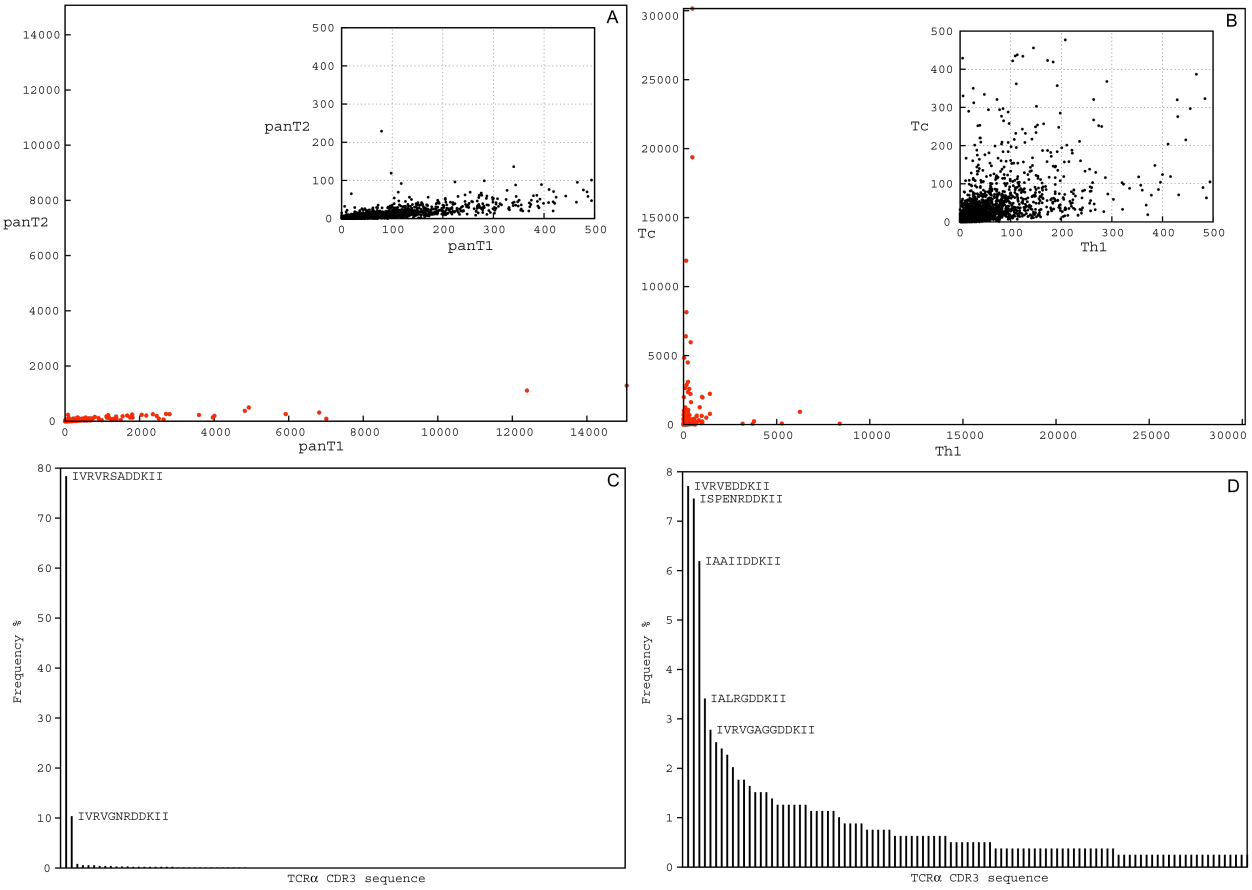

Fig. S3. Assess the PCR and sequencing performance. Correlation of number of sequence reads of the same V and J segments for two panT samples (A), and Tc versus Th1 (B). The V-J combinations for both TCR α and β chains were shown in this figure. The smaller panes within pane A and pane B show the zoom-in version of the larger panes. Each point represents the number of sequence reads of identical V and J segments in the X-axis sample versus that in the Y-axis sample. Pane A shows a high degree of correlation of 0.92 between two panT samples, whereas pane B shows a low degree of correlation of 0.09 between Tc and Th1 samples. Note that the low number of sequence reads in panT2 sample (A) were due to the fact that the panT sample was sequenced along with a panB sample in the same sequence run. The frequency of CDR3 sequences with the same V(TRAV26-1) and J(TRAJ30) segments in Tc (C) and Th1 (D) subsets. There are 61 non-redundant CDR3 sequences (1213 in total) in the Tc subset sample, and 145 non-redundant CDR3 sequences (791 in total) in the Th1 subset sample with the arrangement of TRAV26-1 and TRAJ30. The top 2 and 5 most abundant CDR3 sequences in Tc and Th1 were labeled in the pane C and D, respectively.

Taken together, ARM-PCR and the 454 sequencing technique used in this study were able to amplify and determine the sequences of a large amount of different targets. The relative ratio of the original targets was maintained as the universal PCR primer set was used at the exponential amplification phases, which was proven in a study for the TEM-PCR[1], a predecessor to the ARM-PCR technique. The repeatability of these two techniques was demonstrated with correlation coefficient value 0.92 between two panT samples from the same individual, which was unlikely due to systematic bias as an anticipated low degree of correlation was observed between two different T subset samples (Tc and Th1).

Detection of cross-contamination. One of the major concerns in this study is potential cross-contamination between different subsets of T cells using the magnetic cell sorting (MACS) system. Extra caution was used to avoid cross-contamination. The purities of cells using the MACS approach were listed in Table S1. Since some T cell subsets were isolated with concatenating MACS isolation steps (Fig. S1), the purity of a particular subset could be further improved. For instance, the CD8+ T cells were isolated by negative selection with anti-CD4 microbeads, followed by positive selection with anti-CD8 microbeads. Assuming that the purity of both anti-CD4 and anti-CD8 microbead isolation is 90%, the purity of CD8+ T cell could be $1.0-(1.0-0.9)\times(1.0-0.9)=0.99$. And more, a vigorous statistical procedure was introduced to identify and remove contaminated CDR3 sequences.

The outlined cell isolation procedure in this study produces two exclusive groups of T cell subsets: [Tc, Tr, Th1, Th2], T cells of different fates and [Tn+t, Ta, Tm], T cells of different developing stages. Common CDR3 sequences seen in two subsets of the same group are either due to cross-contamination or authentically shared between those two subsets.

First, we estimated the cross-contamination rate between any two exclusive pair of subsets based on the TCRα data where TCRα CDR3 is more likely to see in two different samples. Table S2 and S3 list the top 5 most abundant CDR3 sequences in one subset (A) and their frequency in the second subset (B). The potential contamination rate $r_{A\to B}$ is calculated as $\frac{freq_B}{freq_A}$, where $freq_A$ and $freq_B$ are the frequency of a particular CDR3 sequence in subset A and B, respectively. The contamination rate that is estimated as the first $r_{A\to B}$ smaller than 10%, where $r_{A\to B}$ is ordered descending according to the abundance of CDR3 sequences in subset A. There

are two reasons for that. First, the most abundant CDR3 in subset A is the most likely one to be observed in subset B if subset A cells contaminate subset B, so it is most reliable to estimate the contamination rate based on the most abundant CDR3. Second, increasing studies using the cell isolation procedure showed that the purity >90% and concatenating two or more than isolation steps can increase the purity and decrease the contamination. Therefore, it is unlikely that a contamination rate could be greater than 10%. Those estimated contamination rate among subsets of the same group are listed in Table S2 and S3, respectively.

For a CDR3 sequence X seen in two different exclusive subsets (A, B). Assume that CDR3 X is more abundant in subset A than in subset B. The total number of CDR3 sequences in subset A is N_A and the total number of CDR3 sequences in subset B is N_B. The contamination rate A to B is $r_{A\to B}$, which was estimated above. If CDR3 X observed in subset B are due to contamination from A cells, the expected number of CDR3 X in subset B is $N_B \times freq_A^X \times r_{A\to B}$, where $freq_A^X$ is the frequency of CDR3 X in subset A. The chance to sequence a contaminated CDR3 in subset B is relatively rare, which can be characterized with the Poisson model. For CDR3 X with n occurrences in subset B, we calculated the probability that such a CDR3 would occur n or more times if it were a contamination, using the following formula: $P = 1 - \sum_{k}^{n-1} \frac{\lambda^k \bullet e^{\lambda}}{k!}$, where $\lambda = N_B \times freq_A^X \times r_{A\to B}$, CDR3 sequences that yield $P < 0.01$ were considered unlikely to be contamination. A procedure was used to remove any contaminated CDR3 according to the Poisson test.

Table S2. Contamination rate among Tn+t, Ta and Tm subsets

Contamination type*	Order†	CDR3	Count‡	Abundance (%)§	Count¶	Abundance (%)‖	Contamination rate (%)**
Tn+t->Tm	1	AVNPNRAGSYQLT	24611	19.3	194	0.2	**1.0**
	2	APEAMGGSEKLV	21452	16.8	16	0.0	0.1
	3	AGPYLIQGAQKLV	10917	8.6	0	0	0
	4	ALSDSGVDTGRRALT	3189	2.5	0	0	0
	5	AMRDHSGGGADGLT	2511	2.0	11	0.0	0.6
Tm->Tn+t	1	AVNPPSSNTGKLI	828	0.8	42	0.0	**4.0**
	2	AVRDQGSNYQLI	555	0.5	11	0.0	1.6
	3	NKLT	522	0.5	9	0.0	1.4
	4	AMRGGNFNKFY	468	0.5	5	0.0	0.8
	5	ALSSNSGYALN	402	0.4	99	0.1	19.6
Ta->Tm	1	AVNPNRAGSYQLT	20007	15.4	194	0.2	**1.2**
	2	APEAMGGSEKLV	14950	11.5	16	0.0	0.1
	3	AGPYLIQGAQKLV	6072	4.7	0	0	0
	4	AMRDHSGGGADGLT	3533	2.7	11	0.01	0.4
	5	VVSGDSSYKLI	1571	1.2	79	0.1	6.4
Tm->Ta	1	AVNPPSSNTGKLI	828	0.8	232	0.2	21.8
	2	AVRDQGSNYQLI	555	0.5	42	0.0	**5.9**
	3	NKLT	522	0.5	175	0.1	26.1
	4	AMRGGNFNKFY	468	0.5	23	0.0	3.8
	5	ALSSNSGYALN	402	0.4	516	0.4	100.0
Ta->Tn+t	1	AVNPNRAGSYQLT	20007	15.4	24611	19.3	125.3
	2	APEAMGGSEKLV	14950	11.5	21452	16.8	146.2
	3	AGPYLIQGAQKLV	6072	4.7	10917	8.0	183.2
	4	AMRDHSGGGADGLT	3533	2.7	2511	2.0	72.4
	5	VVSGDSSYKLI	1571	1.2	1266	1.0	82.1
	30	IVKGQGAQKLV	222	0.2	6	0.0	**2.8**
Tn+t->Ta	1	AVNPNRAGSYQLT	24611	19.3	20007	15.4	79.8
	2	APEAMGGSEKLV	21452	16.83	14950	11.5	68.4
	3	AGPYLIQGAQKLV	10917	8.6	6072	4.7	54.6
	4	ALSDSGVDTGRRALT	3189	2.5	982	0.8	30.2
	5	AMRDHSGGGADGLT	2511	2.0	3533	2.7	138.1
	39	ILWIQGAQKLV	111	0.1	4	0.0	**3.5**

*the subset A (prior to ->) contaminates the subset B (after ->). †the order of abundance of CDR3 in subset A. ‡the count of CDR3 seen in subset A. §the abundance (%) of CDR3 seen in subset A. ¶the count of CDR3 seen in subset B. ‖the abundance(%) of CDR3 seen in subset B. **The potential contamination rate is calculated as the ratio of the abundance of CDR3 in subset B to that in subset A. Those bold numbers show the estimated contamination rate, which is the first rate smaller than 10%--the maximum contamination rate according to reagent manuals. All those numbers are based on TCRα data.

Table S3. Contamination rate among Tc, Tr, Th1 and Th2 subsets.

Contamination type*	Order†	CDR3	Count‡	Abundance (%)§	Count¶	Abundance (%)‖	Contamination rate (%)**
Tc->Tr	1	AVNPNRAGSYQLT	28796	21.1	1	0.0	**0.0**
	2	APEAMGGSEKLV	18410	13.5	0	0	0
	3	AGPYLIQGAQKLV	11337	8.3	0	0	0
	4	ALSDSGVDTGRRALT	3156	2.3	0	0	0
	5	AMRGPRSTGNQFY	2512	1.8	1	0.0	0
Tr->Tc	1	AVRPDKLI	198	0.2	0	0	**0**
	2	IVRVRDTGNQFY	196	0.2	0	0	0
	3	AVSNDYKLS	193	0.2	1	0.0	0.4
	4	AMSGNTGGFKTI	186	0.2	0	0	0
	5	AVRNQAGTALI	173	0.1	0	0	0
Th1->Tr	1	ALSEATGKLI	4978	5.2	15	0.0	**0.2**
	2	ALMDTGRRALT	3454	3.6	10	0.0	0.2
	3	ALSSNSGYALN	2920	3.1	9	0.0	0.2
	4	VVSGDSSYKLI	911	1	32	0.0	2.8
	5	AVNPPSSNTGKLI	695	0.7	28	0.0	3.3
Tr->Th1	1	AVRPDKLI	198	0.2	0	0	**0**
	2	IVRVRDTGNQFY	196	0.2	194	0.2	122.2
	3	AVSNDYKLS	193	0.2	60	0.1	38.4
	4	AMSGNTGGFKTI	186	0.2	2	0.0	1.3
	5	AVRNQAGTALI	173	0.1	33	0.0	23.5
Th1->Tc	1	ALSEATGKLI	4978	5.2	35	0.0	**0.5**
	2	ALMDTGRRALT	3454	3.6	40	0.0	0.8
	3	ALSSNSGYALN	2920	3.1	32	0.0	0.8
	4	VVSGDSSYKLI	911	1	1797	1.3	137.3
	5	AVNPPSSNTGKLI	695	0.7	5	0.0	0.5
Tc->Th1	1	AVNPNRAGSYQLT	28796	21.1	247	0.3	**1.2**
	2	APEAMGGSEKLV	18410	13.5	233	0.2	1.8
	3	AGPYLIQGAQKLV	11337	8.3	68	0.1	0.9
	4	ALSDSGVDTGRRALT	3156	2.3	5	0.0	0.2
	5	AMRGPRSTGNQFY	2512	1.8	6	0.0	0.3
Th2->Tr	1	AMSVNDYKLS	236	0.3	121	0.1	31
	2	IVKGQGAQKLV	177	0.2	124	0.1	42.4
	3	MRILSGSARQLT	170	0.2	95	0.1	33.8
	4	IVLMNTGFQKLV	158	0.2	54	0.0	20.7
	5	IVGDAGNNRKLI	155	0.2	86	0.1	33.6
	23	ATDTTGANNLF	93	0.1	0	0	**0**
Tr->Th2	1	AVRPDKLI	198	0.2	0	0	**0**
	2	IVRVRDTGNQFY	196	0.2	154	0.2	129.9
	3	AVSNDYKLS	193	0.2	125	0.2	107.1
	4	AMSGNTGGFKTI	186	0.2	0	0	0
	5	AVRNQAGTALI	173	0.1	96	0.1	91.8

Th2->Tc	1	AMSVNDYKLS	236	0.3	0	0	**0**
	2	IVKGQGAQKLV	177	0.2	0	0	0
	3	MRILSGSARQLT	170	0.2	0	0	0
	4	IVLMNTGFQKLV	158	0.2	0	0	0
	5	IVGDAGNNRKLI	155	0.2	0	0	0
Tc->Th2	1	AVNPNRAGSYQLT	28796	21.1	6	0.0	**0.0**
	2	APEAMGGSEKLV	18410	13.5	8	0.0	0.1
	3	AGPYLIQGAQKLV	11337	8.3	0	0	0
	4	ALSDSGVDTGRRALT	3156	2.3	0	0	0
	5	AMRGPRSTGNQFY	2512	1.8	0	0	0
Th2->Th1	1	AMSVNDYKLS	236	0.3	2	0.0	**0.6**
	2	IVKGQGAQKLV	177	0.2	19	0.0	8
	3	MRILSGSARQLT	170	0.2	26	0.0	11.4
	4	IVLMNTGFQKLV	158	0.2	72	0.1	34
	5	IVGDAGNNRKLI	155	0.2	2	0.0	1
Th1->Th2	1	ALSEATGKLI	4978	5.2	57	0.1	**1.5**
	2	ALMDTGRRALT	3454	3.6	30	0.0	1.2
	3	ALSSNSGYALN	2920	3.1	31	0.0	1.4
	4	VVSGDSSYKLI	911	1	43	0.1	6.3
	5	AVNPPSSNTGKLI	695	0.7	79	0.1	15.2

*the subset A (prior to ->) contaminates the subset B (after ->). †the order of abundance of CDR3 in subset A. ‡the count of CDR3 seen in subset A. §the abundance (%) of CDR3 seen in subset A. ¶the count of CDR3 seen in subset B. ‖the abundance(%) of CDR3 seen in subset B. **The potential contamination rate is calculated as the ratio of the abundance of CDR3 in subset B to that in subset A. Those bold numbers show the estimated contamination rate, which is the first rate smaller than 10%--the maximum contamination rate according to reagent manuals. All those numbers are based on TCRα data.

Normalizing procedure. As different subsets of T cells have quite different effective reads and functional CDR3 sequences, when calculating the overlapping CDR3, CDR3 copy number versus frequency (Fig. 2), CDR3 sequences were uniformly sampled from all CDR3 sequences in a particular subset to bring the overall number of CDR3 sequences equal to the smallest number of CDR3 in that particular group of subsets.

In order to avoid the distortion due to clonal expansion effect when calculating germline gene segment usage, CDR3 length, CDR3 amino acid content, N-nucleotide additions, and trimming at the coding ends of the germline gene segments, the entity for the same CDR3 sequence was counted only once for these comparisons irrespective of the number of copies of each individual CDR3 that was obtained.

Germline gene segment usage, CDR3 length, N-addition, nibbling at gene segments and amino acids usage. Statistical comparisons of germline gene segment usages distributions were accomplished by two methods: determination of Pearson correlation coefficient, r, and chi-

square analysis. Both were done pairwise if a comparison between more than two parties were carried out. Results from Pearson correlation coefficient analysis were provided as the coefficient r and P-value. The P-value is the probability that one would have found the current result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional 5% ($P<0.05$), the correlation coefficient was termed statistically significant. To overcome the influence of sample size on $\chi 2$ value, Monte Carlo simulation with 500,000 replications, were used to compare proportions on gene segments usage and to estimate the P-value.

Table S4. CDR3 peptide length, N-nucleotide addition, nibbling at 3' end of V and D segments, and at 5' end of D and J segments.

Subset	TCRα				TCRβ					
	CDR3	N-add.	V'3 del	5'J del	CDR3	N-add.	V'3 del	5'D del	D'3 del	5'J del
Tr	11.7(1.6)	5.0(3.0)	3.5(3.4)	4.9(3.3)	12.4(1.7)	9.0(4.9)	3.5(2.4)	3.3(2.8)	2.6(2.1)	4.1(2.9)
Th1	11.7(1.6)	4.9(3.0)	3.5(3.3)	4.8(3.3)	12.3(1.6)	8.7(4.8)	3.5(2.4)	3.3(2.9)	2.7(2.2)	4.0(2.9)
Th2	11.7(1.6)	4.9(3.0)	3.6(3.4)	4.8(3.3)	12.3(1.7)	8.9(4.9)	3.5(2.4)	3.3(2.9)	2.6(2.1)	4.0(2.9)
Tc	11.6(1.8)	5.0(3.1)	3.6(3.6)	4.7(3.4)	12.3(1.6)	8.7(4.9)	3.3(2.4)	3.4(2.9)	2.7(2.2)	3.9(2.7)
Tn+t	11.6(1.7)	5.0(3.1)	3.4(3.4)	4.7(3.4)	12.3(1.6)	8.7(4.9)	3.4(2.5)	3.3(2.9)	2.7(2.2)	4.0(2.8)
Ta	11.7(1.7)	5.0(3.2)	3.5(3.3)	4.7(3.3)	12.3(1.7)	9.1(5.0)	3.5(2.4)	3.3(2.9)	2.6(2.2)	4.0(2.8)
Tm	11.7(1.6)	4.9(3.0)	3.5(3.3)	4.8(3.3)	12.3(1.6)	8.7(4.8)	3.5(2.4)	3.3(2.9)	2.6(2.1)	4.1(2.9)

Note that clonal expansion effects are eliminated by counting the entity only once for these comparisons irrespective of the number of copies of each individual CDR3 that was obtained.

TCR α and β chains are generated by combinatorial joining of rearranged gene segments of variable (V), diversity (D) and joining (J) regions. The usage of those domains are nevertheless random. Immune responses usually show some level of bias in the usage of V, D and J gene segments[3-7]. Several methods have been developed to analyze the repertoire of T cells under physiologic conditions as well as in various pathological situations. The majority of these studies examine the extent of skewing of either Vβ usage or CDR3 lengths (Vβ mAb staining[8], spectratyping[9], CDR3 length polymorphism analysis[10]). However, it is difficult for these methods to quantitate T cell diversity at a clonal level. With the availability of a large amount of sequence data, the germline gene segments usage, CDR3 lengths distribution, deletion of nucleotides at ends of germline gene segments, and the number of non-templated nucleotides addition were examined here. In order to avoid the distortion of these parameters by expanded clones, the entity for each unique CDR3 sequence was counted once irrespective of how many copies were observed.

In general, the usage patterns of germline gene segments are similar among different subsets of T cells. Pairwise Pearson's correlation coefficient for the usage of Vα, Jα, Vβ, Dβ, Jβ, respectively between subsets of T cells were computed. Most of those correlation coefficient values indicated a high level of correlation, ranging from 0.70 to 1.00, suggesting the similarity on the germline gene segment usage among different subsets of T cells. χ^2-test with Monte Carlo simulation on 500,000 replications shows no statistically significant difference in terms of germline gene segments usage after the clonal expansion effects were ruled out.

Fig. S4 displays the amino acids distribution of the CDR3 regions for TCRα (A) and TCRβ (B). Basically, the patterns of amino acids were similar to each other among different subsets of T cells. Table S4 listed CDR3 lengths distribution, number of non-templatedd nucleotides (N) addition, and deletion of nucleotides at ends of germline gene segment for different subsets of T cells. No statistically significant differences were found among different subsets of T cells.

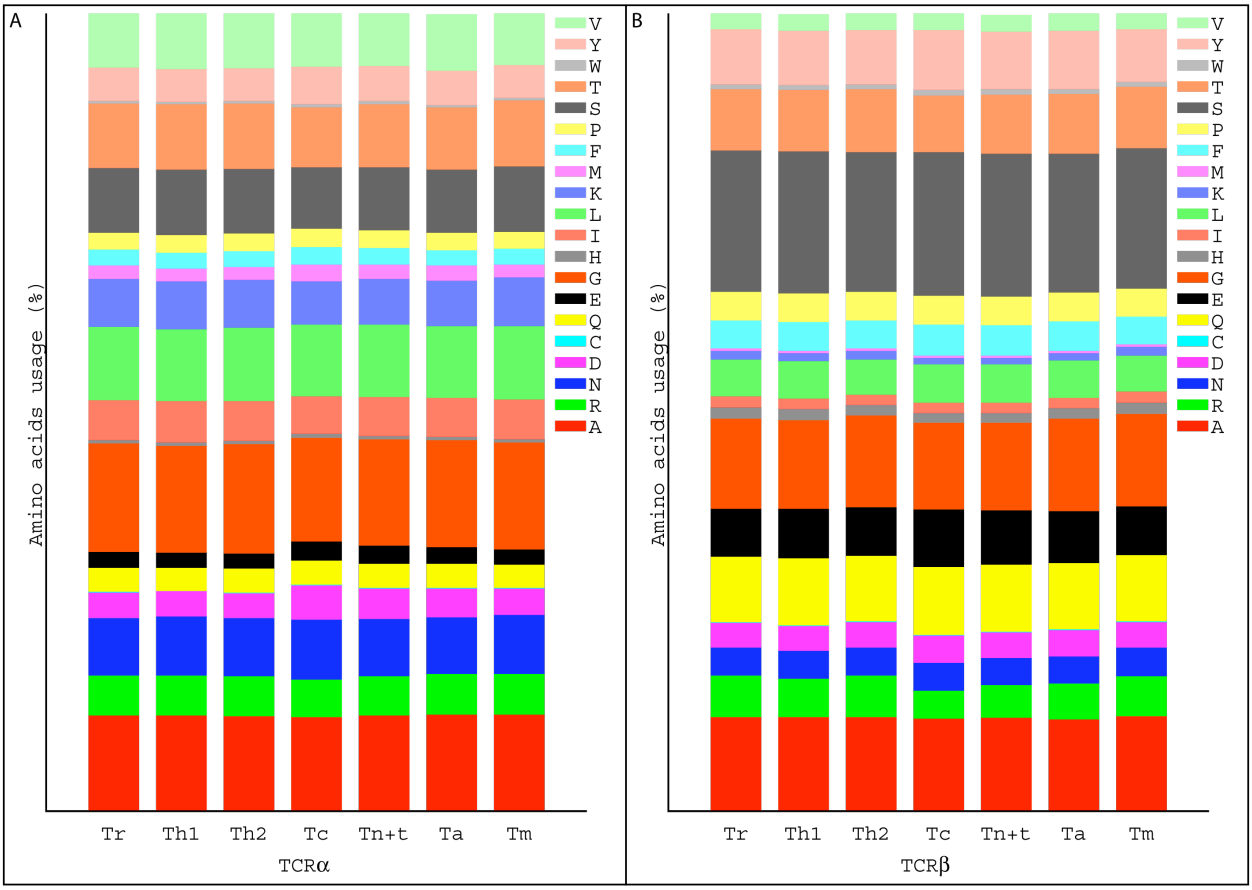

Fig. S4. Amino acids usage in the TCRα (A) and TCRβ(B) CDR3 region of different subsets of T cells. Clonal expansion effects are eliminated by counting CDR3 only once irrespective of the number of copies of each individual CDR3 that was obtained.

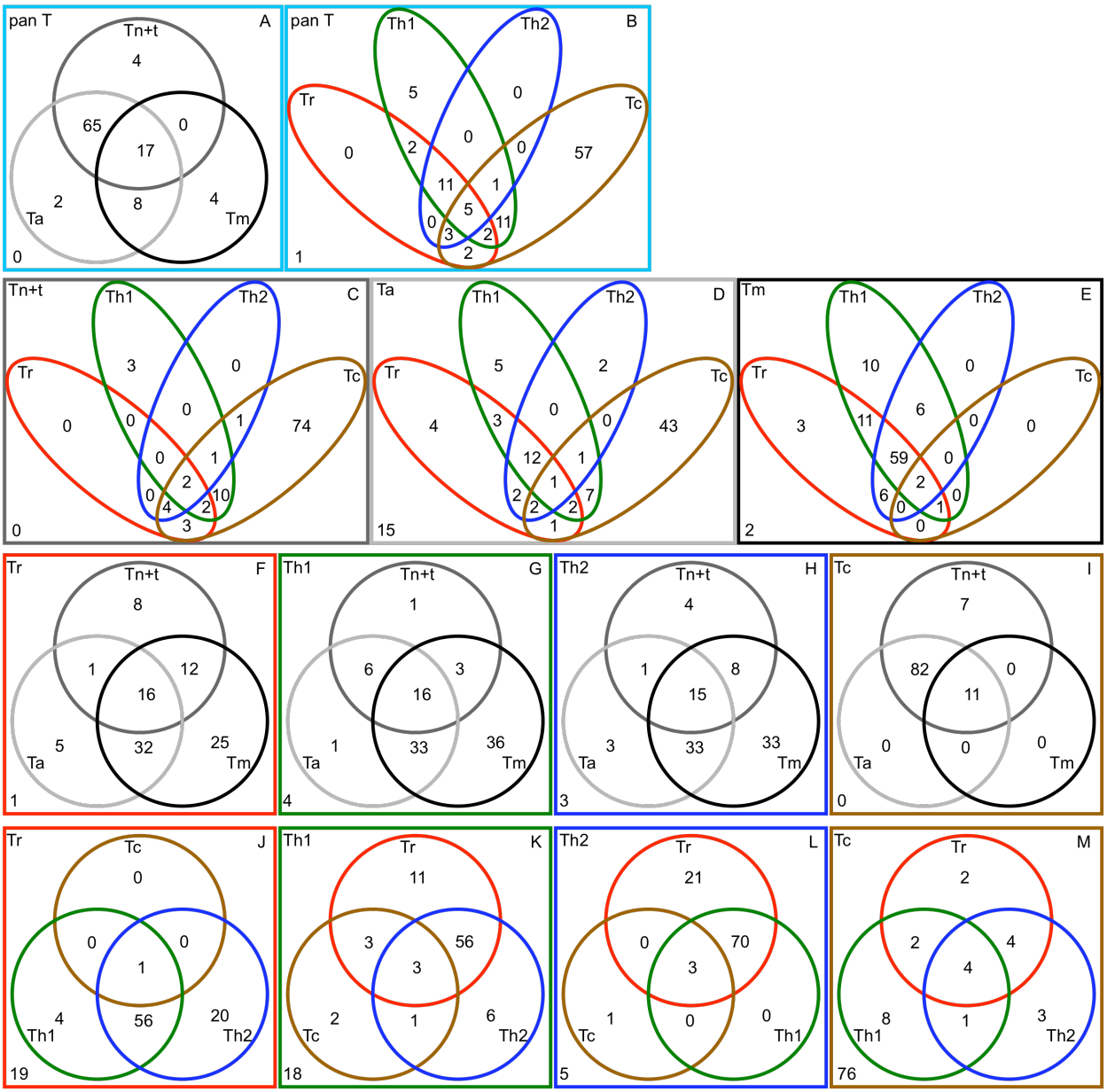

Fig S5. The first 100 most abundant TCRα CDR3 sequences of a particular subset (labeled at the top left corner of each box) common to those in subsets of T cells of either different developing stages(grey:Tn+t; darkgrey: Ta and Tm: black) or different fates (red: Tr; green: Th1; blue: Th2; and brown: Tc). The number of CDR3 sequences that are unique to each subset are shown in the non-overlapping sections. The number of CDR3 sequences that are common to any two, three and four of these subsets are indicated in the relevant overlapping areas. The number of CDR3 sequences that are not found in those examined subsets is labeled at the bottom left corner of each box.

Table S5. The top 10 most frequent CDR3 sequences for TCR α and β listed for T cell subsets.

TCRα							
Tr	Th2	Tm	Th1	Tc	Tn+t	Ta	panT
IVRVRDTGNQFY	AMSVNDYKLS	AVNPPSSNTGKLI	ALSEATGKLI	AVNPNRAGSYQLT	AVNPNRAGSYQLT	AVNPNRAGSYQLT	AVNPNRAGSYQLT
AVRPDKLI	IVKGQGAQKLV	AVRDQGSNYQLI	ALMDTGRRALT	APEAMGGSEKLV	APEAMGGSEKLV	APEAMGGSEKLV	APEAMGGSEKLV
AMSGNTGGFKTI	MRILSGSARQLT	NKLT	ALSSNSGYALN	AGPYLIQGAQKLV	AGPYLIQGAQKLV	AGPYLIQGAQKLV	AGPYLIQGAQKLV
APHGGGGADGLT	IVLMNTGFQKLV	AMRGGNFNKFY	VVSGDSSYKLI	ALSDSGVDTGRRALT	ALSDSGVDTGRRALT	AMRDHSGGGADGLT	AMRDHSGGGADGLT
AVRNQAGTALI	IVRVRDTGNQFY	ALSSNSGYALN	AVNPPSSNTGKLI	AMRGPRSTGNQFY	AMRDHSGGGADGLT	VVSGDSSYKLI	ALMDTGRRALT
IVIWDDKII	IVGDAGNNRKLI	IVRVRDTGNQFY	NKLT	AMRDHSGGGADGLT	AMRGPRSTGNQFY	AYRSGFDAGKST	ALSDSGVDTGRRALT
AAAYNTDKLI	AAAYNTDKLI	AVNIAGGGNKLT	ASPSYNTDKLI	AYRSGFDAGKST	AYRSGFDAGKST	AMRGPRSTGNQFY	AMRGPRSTGNQFY
IVRAGIGGATNKLI	IVRVKRNNDMR	IVRVAEKLT	APEAMGGSEKLV	VVSGDSSYKLI	VVSGDSSYKLI	ALSEATGKLI	ALSEATGKLI
AVVSGTYKYI	IVRYGGGADGLT	LLTGTASKLT	IVRVRDTGNQFY	VVNLNNNARLM	ALSELEQDTGRRALT	ALMDTGRRALT	VVSGDSSYKLI
AMSGANAGKST	IALRGDDKII	ALLNAGKST	ALLNAGKST	ALSELEQDTGRRALT	VVNLNNNARLM	ALSDSGVDTGRRALT	ALSSNSGYALN
TCRβ							
Tr	Th2	Tm	Th1	Tc	Tn+t	Ta	panT
ASSQKREWMYEQY	ASSLGDRMQY	SANKDRVPEAF	ASSFGATTDTQY	ASSFGQGPNPGANVLT	ASSFGQGPNPGANVLT	ASSFGQGPNPGANVLT	ASSFGATTDTQY
AHGSGSNQPQH	ASSLQKGYT	ASTSRGGRQETQY	SANKDRVPEAF	ASRLAGGTDTQY	ASRLAGGTDTQY	ASSTGTNQPQH	SIPKERGPIRYEQY
ASSPRETQY	ASSLGQAFTEAF	ASSFGATTDTQY	SAGQGANYGYT	SIPKERGPIRYEQY	SIPKERGPIRYEQY	ASSPGATTDTQY	ASRLAGGTDTQY
ASSSTGTGFQPQH	ASSPGQGPYEQY	ASSATGLPQPQH	ASSDPGGSNEQF	ASSIAGTAYEQY	ASSIAGTAYEQY	ASRLAGGTDTQY	ASSFGQGPNPGANVLT
ASSLGGLLAGNYTDTQY	ASSPARSSSQETQY	SASQGADTQY	ASTSRGGRQETQY	ASSTGTNQPQH	ASSTGTNQPQH	SIPKERGPIRYEQY	ASSTGTNQPQH
ASSYRDRRNNEQF	ASRGWGGLSYNEQF	ATGTRILGANVLT	ATGTRILGANVLT	ASTLGGLIYNEQF	ASTLGGLIYNEQF	SANKDRVPEAF	SANKDRVPEAF
AVGGLGGYT	ASSSTGTGFQPQH	ASSKSGESYNEQF	ASSLEAGPSYEQY	SVPEGTNYGYT	ASSLEQGVRSDEQF	SVPEGTNYGYT	ASSIAGTAYEQY
ASSTGQGADTQY	ASSLAGANVLT	ASSQDRDLSAF	ASSLASPRQNYGYT	ASSDPGGSNEQF	SVPEGTNYGYT	ASTLGGLIYNEQF	ASSDPGGSNEQF
ASSLARYTEAF	ASSLGGLLAGNYTDTQY	ASSDPGGSNEQF	ASSATGLPQPQH	ASSLEQGVRSDEQF	ASSDPGGSNEQF	SAGQGANYGYT	ASSLEQGVRSDEQF
ASRRDNYGYT	AIRQRQGPQH	ASSQDRVGNTEAF	ASSLEPLAKNIQY	ASSPEQGARADTQY	ASSPEQGARADTQY	SVSGTGYLLNTEAF	SVPEGTNYGYT

Tr—T regulatory cell (CD4+CD25+); Th1—T helper cell 1 (CD4+CD25-CD294-); Th2—T helper cell 2 (CD4+CD25-CD294+); Tc—T cytotoxic cell (CD8+) Tn+t—naïve and transitional T cell (CD45RA+); Ta—activated T cell (CD45-RO-CD69+); Tm—memory T cell (CD45RA-RO+); aa—amino acids; na–-nucleic acids. Sequences with a white background are unique to the subset. Sequences with the same color background are shared between subsets.

Calculation of the probability of appearance of either two identical TCRα or two identical TCRβ sequences. We estimated the probability that two exactly identical TCRα or TCRβ genes with the same CDR3 sequences were generated via two independent rearrangements following a similar strategy by Saada et al[(11)].

For TCRβ genes, the CDR3 length ranges from 9 to 17 amino acids. The number of non-templated nucleotide addition ranges from 0 to 25. The number of nucleotides trimmed at the germline gene segments ranges from 0 to 10 for 3' end of Vβ, from 0 to 10 for 5' end of Dβ, from 0 to 10 for 3' end of Dβ, and from 0 to 13 for 5' end of Jβ, respectively. The calculation is based on these parameters and on the following assumptions:

1. The number of functional gene segments for the germline gene segments of human TCRβ chain are: 48Vβ, 2Dβ and 13Jβ according to the IMGT database[(2)].
2. Deletions are between 0 and 10 for 3' end of Vβ, hence there are 11 options for deletion at the Vβ. Deletions are between 0 and 13 for 5' end of Jβ, hence there are 14 options for deletion at the Jβ.
3. Since the locations of Dβ segment in the CDR3 region are random with 26 options as the maximum length of N-addition is 25. According to the methods described in the main text, the shortest length of recognizable Dβ segment is 6. The length Dβ segments could range from 6 to its full length where the length of Dβ1 is 12 and that of Dβ2 is 16. Thus, the contribution of

Dβ segments to the CDR3 diversity is $26\times\left(\sum_{k1=6}^{l1=12}(l1-k1+1)+\sum_{k2=6}^{l2=16}(l2-k2+1)\right)$ where the first and second terms in parenthesis show the possible number of options for using different length of Dβ1 and Dβ2 segments.

4. N addition is up to 25 nucleotides. There are four options for each nucleotide (A, G, C, or T).

   Hence the number of possibilities for N addition one overhanging strand is $\sum_{k=0}^{25}4^k$

Taken all those factors together, the total number of possible CDR3 sequences for TCRβ chain is:

$$[11\times 48]_V\times\left[26\times\left(\sum_{k1=6}^{l1=12}(l1-k1+1)+\sum_{k2=6}^{l2=16}(l2-k2+1)\right)\right]_D\times[14\times 13]_J\times\left[\sum_{k=0}^{25}4^k\right]_{addition}=3.53\times 10^{23}$$

The calculation for total number of possible CDR3 sequences for TCRα chain is much simpler due to its lack of D segments. According to the IMGT database(2), there are 45 functional Vα and 50 functional Jα segments in a human genome. The number of non-templated nucleotide addition ranges from 0 to 15. The number of nucleotides deleted at the germline gene segments ranges from 0 to 15 for 3' end of Vα, and from 0 to 17 for 5' end of Jα, respectively. The total number of possible CDR3 sequences for TCRα chain is:

$$[16\times 45]_V\times[18\times 50]_J\times\left[\sum_{k=0}^{15}4^k\right]_{addition}=9.28\times 10^{14}$$

These estimates do not take into account the fact that two different rearrangements can result in the same sequences.

In mice, the number of mature T cells that leave the thymus each day was estimated at 2×10^6 *(16)*. As did in the study(11), the number of mature T cells from thymus made daily by a human was estimated by scaling up the mice data according to the assumption that the number of T cells created is proportional to body weight. The average weight of a mouse is 20g and the average weight of a human in 60kg. Thus, the number of mature T cells leaving a human thymus daily is $2\times 10^6\times(60\times 10^3/20)=6\times 10^9$. Assuming the average lifespan is 100 years (36500 days), we got an estimate of $36500\times 6\times 10^9=2.19\times 10^{14}$ T cells created in the lifetime of a long-lived human.

According to the 'birthday paradox'[12], the probability that two T cells have the same CDR3 sequences with an individual's lifetime can be estimated by $p = 1.0 - \frac{y!}{(y-x)!y^x}$, where y is the number of possible CDR3 sequences, and x is the total number of T cells generated in an individuals lifetime. According to the Stirling's approximation, $p \approx 1.0 - \frac{y^{y-x+0.5}e^{-x}}{(y-x)^{y-x+0.5}}$. For TCRβ chain where $y = 3.53 \times 10^{23}$, $p \simeq 0.0$, implying that clonality can be established on the basis of TCRβ chain. For TCRα chain where $y = 9.28 \times 10^{14}$, $p \simeq 1.0$, suggesting that there are chances that identical TCRα chain CDR3 sequences can be generated from independent rearrangements.

References

1. Han J*, et al.* (2006) Simultaneous amplification and identification of 25 human papillomavirus types with Templex technology. (Translated from eng) *J Clin Microbiol* 44(11):4157-4162 (in eng).
2. Lefranc MP (2003) IMGT, the international ImMunoGeneTics database. (Translated from eng) *Nucleic Acids Res* 31(1):307-310 (in eng).
3. Kjer-Nielsen L*, et al.* (2003) A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity. (Translated from eng) *Immunity* 18(1):53-64 (in eng).
4. Maryanski JL, Jongeneel CV, Bucher P, Casanova JL, & Walker PR (1996) Single-cell PCR analysis of TCR repertoires selected by antigen in vivo: a high magnitude CD8 response is comprised of very few clones. (Translated from eng) *Immunity* 4(1):47-55 (in eng).
5. Matsumoto Y*, et al.* (2006) CDR3 spectratyping analysis of the TCR repertoire in myasthenia gravis. (Translated from eng) *J Immunol* 176(8):5100-5107 (in eng).
6. Koga M, Yuki N, Tsukada Y, Hirata K, & Matsumoto Y (2003) CDR3 spectratyping analysis of the T cell receptor repertoire in Guillain-Barre and Fisher syndromes. (Translated from eng) *J Neuroimmunol* 141(1-2):112-117 (in eng).

7. Stamatopoulos K*, et al.* (2005) Immunoglobulin light chain repertoire in chronic lymphocytic leukemia. (Translated from eng) *Blood* 106(10):3575-3583 (in eng).
8. MacDonald HR, Casanova JL, Maryanski JL, & Cerottini JC (1993) Oligoclonal expansion of major histocompatibility complex class I-restricted cytolytic T lymphocytes during a primary immune response in vivo: direct monitoring by flow cytometry and polymerase chain reaction. (Translated from eng) *J Exp Med* 177(5):1487-1492 (in eng).
9. Even J (1995) T cell repertoires in healthy and diseased human tissues analyzed by T cell receptor [beta]-chain CDR3 size determination: evidence for oligoclonal expansions in tumours and inflammatory diseases. *Res. Immunol.* 146:65-80.
10. Cochet M*, et al.* (1992) Molecular detection and in vivo analysis of the specific T cell response to a protein antigen. (Translated from eng) *Eur J Immunol* 22(10):2639-2647 (in eng).
11. Saada R, Weinberger M, Shahaf G, & Mehr R (2007) Models for antigen receptor gene rearrangement: CDR3 length. (Translated from eng) *Immunol Cell Biol* 85(4):323-332 (in eng).
12. H. ME (1966) Generalized Birthday Problem. *American Mathematical Monthly* 73:385-387.