NIH Public Access
Author Manuscript
Sci Transl Med. Author manuscript; available in PMC 2011 November 9.

Published in final edited form as:
Sci Transl Med. 2010 September 1; 2(47): 47ra64. doi:10.1126/scitranslmed.3001442.

Overlap and effective size of the human CD8+ T-cell receptor repertoire

Harlan S. Robins[1,*], **Santosh K. Srivastava**[1], **Paulo V. Campregher**[2], **Cameron J. Turtle**[2], **Jessica Andriesen**[1], **Stanley R. Riddell**[2], **Christopher S. Carlson**[3,$], and **Edus H. Warren**[1,$]

[1]Program in Computational Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, Washington 98109

[2]Program in Immunology, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, Washington 98109

[3]Program in Cancer Prevention, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, Washington 98109

Abstract

Diversity in T-lymphocyte antigen receptors is generated by somatic rearrangement of T-cell receptor (TCR) genes and is concentrated within the third complementarity-determining region (CDR3) of each chain of the TCR heterodimer. We sequenced the CDR3 regions from millions of rearranged TCR β chain genes in naïve and memory $CD8^+$ T-cells of seven adults. The CDR3 sequence repertoire realized in each individual is strongly biased toward specific V_β-J_β pair utilization, dominated by sequences containing few inserted nucleotides, and drawn from an effective sequence space 250-fold smaller than predicted. Surprisingly, the overlap in the naïve $CD8^+$ TCRβ CDR3 sequence repertoires of any two of the individuals is ~1000-fold larger than predicted and essentially independent of the degree of HLA matching.

Keywords

TCR repertoire; public T-cell; TCR diversity

The antigenic specificity of αβ T-lymphocytes, which recognize peptide antigens presented by class I and class II molecules of the Major Histocompatibility Complex (MHC), is in large part determined by the amino acid sequence in the hypervariable complementarity-determining region 3 (CDR3) regions of the α and β chains of the T-cell receptor (TCR). The nucleotide sequences that encode the CDR3 regions are generated by somatic rearrangement of noncontiguous variable (V), diversity (D), and joining (J) region gene segments for the β chain, and V and J segments for the α chain. The existence of multiple V, D, and J gene segments in germline DNA permits substantial combinatorial diversity in receptor composition, and receptor diversity is further augmented by the deletion of nucleotides adjacent to the recombination signal sequences (RSS) of the V, D, and J segments and template-independent insertion of nucleotides at the V_β-D_β, D_β-J_β, and V_α–J_α junctions. The diversity of distinct TCR αβ pairs has been estimated at ~$2.5*10^{18}$.[1] Using the same model[1] with updated human genome data, we calculate the diversity of the β-chain alone at ~$5*10^{11}$. Thus, the potential diversity of CDR3 sequences far exceeds the diversity that can be realized in one individual at one time, and the magnitude of this diversity has to

*corresponding author hrobins@fhcrc.org .
$contributed equally

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

date made global comparison of the αβ TCR repertoires present in different individuals virtually impossible.

Applying a high-throughput sequencing approach[2] to genomic DNA from purified naïve ($CD45RO^{-}$, $CD45RA^{hi}$, $CD62L^{+}$) and memory ($CD45RO^{+}$, $CD45RA^{low}$) $CD8^{+}$ T-cells, we assessed the realized $CD8^{+}$ TCRβ CDR3 sequence repertoire in the blood of seven healthy adults (Table S1). More than five million TCRβ CDR3 sequence reads were generated from approximately 1 million template genomes in each of the seven naïve and seven memory samples. A mean of 420,000 unique CDR3 nucleotide sequences were observed in the naïve samples, and 69,000 unique nucleotide sequences in the memory samples (Table S2). Identification of the V_β, D_β, and J_β gene segments contributing to each TCRβ CDR3 sequence was performed using a standard algorithm.[3]

The frequency with which specific V_β–J_β combinations were utilized was highly variable in each of the seven individuals (Figures 1A and 1B). Although every possible V_β–J_β combination was observed, the frequency with which specific combinations were observed varied more than 10,000-fold. V_β–J_β utilization was remarkably consistent between individuals, however, especially for the rare V_β–J_β pairs, as reflected by the fact that the variance in V_β–J_β utilization was proportional to mean utilization.

A small fraction of the TCRβ CDR3 sequences observed in the genomic DNA extracted from the naïve and memory $CD8^{+}$ T cells of each of the seven donors were predicted to generate out-of-frame TCRβ transcripts that do not encode functional TCRβ chains (Table S2). The V_β-J_β utilization in the out-of-frame CDR3 sequences was highly non-uniform and qualitatively similar to that observed in in-frame transcripts (Figure 1C), suggesting that the variability in V_β-J_β utilization in in-frame transcripts generating functional TCRβ chains expressed by naïve $CD8^{+}$ T cells is attributable, at least in part, to mechanisms that operate before the level of thymic selection.

Ordering the V_β-J_β pairs by mean utilization frequency demonstrates that 50% of the possible V_β–J_β combinations accounted for more than 98% of the sequences collectively observed in the seven donors. Several pairs (*e.g.*, $V_\beta19/J_\beta2$-2, Figures 1A and 1B) were observed at much higher frequency in the $CD8^{+}$ memory than in naïve cells of individual donors. The *TRBV30* gene segment is unique among the Vβ segments because it lies 3′ to all of the D_β, J_β, and C_β (constant region) gene segments and has an inverted transcriptional polarity; its incorporation into a V_β–J_β–D_β–C_β concatamer thus requires both inversion as well as looping out of intervening DNA.[4] Despite the increased complexity of TCRβ rearrangements involving *TRBV30*, CDR3 sequences utilizing this gene segment were frequent in all seven donors.

The observed frequencies of specific V_β–D_β–J_β combinations suggest that rearrangement between V_β and D_β gene segments is random, while that between D_β and J_β gene segments is not (Figure S1). The apparent non-random association between specific D_β and J_β gene segments is likely attributable to the organization of the TCRβ locus, in which $D_\beta1$ lies 5′ of all 13 J_β segments, while $D_\beta2$ lies 3′ of the 6 members of the $J_\beta1$ cluster but 5′ of the 7 members of the $J_\beta2$ cluster. The $D_\beta1$ segment is observed at roughly equal frequency with all 13 J_β's, while $D_\beta2$ is much more frequently paired with members of the $J_\beta2$ compared with the $J_\beta1$ family. $D_\beta2$ is observed with members of the $J_\beta1$ family about a third (.30+/−.05) as often as would be expected if the pairing were random.

Much of the predicted diversity in the TCRβ CDR3 repertoire is generated by non-templated nucleotide insertions at the V_β–D_β and D_β–J_β junctions. An estimate of $5*10^{11}$ unique TCRβ amino acid sequences is predicted by a previous model that allows for up to six insertions at each of the two junctions.[1] The cumulative distribution of TCRβ CDR3 sequences observed

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in the $CD8^+$ naïve and memory compartments, respectively, of the seven donors as a function of number of junctional insertions demonstrates that sequences with 12 or more insertions were observed, but comprised only 10% of the total (Figures 2A and 2B). In contrast, more than 10% of the observed sequences had zero, one, or two insertions, and 50% of the sequences in each donor had six or fewer total insertions at the two junctions.

To determine the effective size of the sequence space from which the $CD8^+$ TCRβ CDR3 repertoire in each individual is drawn, we explicitly enumerated the complete set of possible CDR3 sequences predicted by a model of VDJ recombination that allowed up to nine nucleotide deletions from the ends of the V_β, D_β and J_β gene segments adjacent to the RSS, followed by insertion of a total of up to seven non-templated nucleotides at the V_β–D_β and D_β–J_β junctions (Figure S2). The model allowed the total CDR3 length to range from 9 to 23 amino acids, consistent with our experimentally observed sequence data. Generation of all unique CDR3 amino acid sequences containing a total of 7 or fewer nucleotide insertions at the V_β–D_β and D_β–J_β junctions required approximately 10,000 cpu hours on a 2.3 Ghz processor (Supplement 1). Comparison of the set of sequences observed in the naïve and memory $CD8^+$ repertoires of each of the seven donors with the full set of sequences generated by the model (Figures 2C and 2D) reveals that approximately half of all the sequences observed in each donor are found in the subset of predicted sequences containing six or fewer total insertions, and 60% of the observed sequences in each donor are found in the subset of predicted sequences containing seven or fewer total insertions. The total number of sequences containing seven or fewer total insertions is 1.78×10^9 (Figure 2E), which implies that approximately two-thirds of the naïve $CD8^+$ CDR3 repertoire of each individual is drawn from a sequence space with an effective size of 2×10^9 sequences.

We previously demonstrated that at least 3×10^6 distinct TCRβ CDR3 amino acid sequences are expressed in the peripheral blood T-cell compartment of an adult,[2] which implies that any two individuals would be expected to share less than 20 CDR3 sequences if the TCRβ CDR3 repertoire in an individual were randomly chosen from a uniform distribution of $5*10^{11}$ different sequences (Supplement 2). The smaller effective size of the possible $CD8^+$ CDR3 repertoire implied by our sequence data suggested that the overlap between the $CD8^+$ CDR3 repertoires of different individuals might be significantly larger. Indeed, when we compared the naïve $CD8^+$ subsets of any two of the seven individuals who were studied, we found more than 10,000 identical TCRβ CDR3 amino acid sequences (**Figures** S3 and 3A; Supplement 3). The seven individuals include two sisters who had identical HLA-A, -B, and -C alleles, their mother, and four unrelated individuals of diverse ethnic and geographic origin who shared few HLA-A, -B, or -C alleles with each other or with the mother/daughter trio (Table S1). An overlap of over 10,000 TCRβ CDR3 sequences was even observed between the naïve $CD8^+$ repertoires of individuals 6 and 7, who shared no HLA-A, -B, or -C alleles. The overlap between the $CD8^+$ memory repertoires of any two individuals was smaller, as expected, since the composition of each individual's memory repertoire is determined by his or her cumulative history of antigenic exposures (**Figures** S4 and 3A). Nonetheless, the mean overlap between the memory $CD8^+$ TCRβ CDR3 repertoires of any two individuals exceeded 1,000 sequences. The pairwise overlap of the naïve $CD8^+$ subsets of the seven donors predicted by our model of TCRβ VDJ rearrangement is 15,400 sequences (Supplement 2), which agrees closely with the overlap observed between all 21 possible pairs of the seven individuals studied (Figure 3A). CDR3 sequences that were shared between individuals had fewer inserted nucleotides than the mean number observed across the entire repertoire (Figures 3B and 3C).

Using the TCRβ CDR3 sequence to identify clonally-derived T-cells, we compared the CDR3 sequence repertoires of the $CD8^+$ naïve and memory compartments of each of the seven donors and identified the subset of sequences that were observed in both

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

compartments to track the fate of individual T-cell clones. CDR3 sequences with high relative frequencies in the $CD8^+$ naïve compartment were more likely than sequences with low relative frequencies to be observed in the memory compartment (data not shown). Confirming previous observations from our group[2] and others,[5] we also observed that CDR3 sequence abundance is inversely correlated with total junctional insertions in both the $CD8^+$ naïve (Figure 4A) and memory compartments. The higher frequency with which sequences carrying few or no junctional insertions were observed was not due to biased amplification or sampling, because the abundance of CDR3 sequences generating out-of-frame TCRβ transcripts showed no such dependence on the number of junctional insertions (Figure 4B). Analyzing the subset of in-frame CDR3 sequences that were observed in both the naïve and memory $CD8^+$ compartments, we observed no correlation between the frequency with which individual sequences were observed in the naïve compartment and their frequency in the memory compartment (Figure S6). These results suggest that the size of $CD8^+$ clones in the naïve compartment is correlated with their probability of entering the memory compartment, but not with their size in the memory compartment. By extension, these results imply that the high prevalence of clones in the memory compartment bearing receptors with few junctional insertions is not simply attributable to their high prevalence in the naïve compartment.

The preponderance of CDR3 sequences with few non-templated insertions in the $CD8^+$ TCRβ repertoire, particularly in the memory compartment, suggests that the capacity to insert nucleotides at the V_β–D_β and D_β–J_β junctions may not be required for many $CD8^+$ immune responses. Indeed, mice deficient for Terminal deoxynucleotidyl transferase, the enzyme that catalyzes the template-independent insertion of nucleotides at the junctions, have 10-fold less diversity in their TCR CDR3 repertoires, with few insertions, yet these mice appear healthy, make efficient and specific immune responses, and display no increased susceptibility to infection.[6, 7] This, in turn, suggests the possibility that the V_β, D_β, and J_β segment sequences that contribute to recurrently generated TCRs could be subject to evolutionary pressures favoring sequences recognizing antigens from common pathogens, as these sequences are present in the germline. Indeed, components of the $CD8^+$ T cell response to ubiquitous pathogens such as Epstein-Barr virus (EBV) are characterized by highly conserved TCRβ CDR3 amino acid sequences that are found in multiple individuals and encoded by nucleotide sequences with few junctional insertions.[5, 8, 9] We looked for 12 such "public" TCRβ CDR3 sequences that have been associated with the $CD8^+$ response to EBV in individuals who express either HLA-A*0201 or HLA-B*0801, and detected 5 HLA-A*0201-associated responses in the memory $CD8^+$ compartments of donors 1 and 3, both of whom are HLA-A*0201^+, and a HLA-A*0801-associated response in the memory compartment of donor 7, who is HLA-B*0801^+. None of these responses were detected in the other four donors, all of whom were HLA-A*0201^- and HLA-B*0801^- (Table S3). The observation of the HLA-A*0201- and HLA-B*0801-associated EBV-specific CDR3 sequences only in the three donors expressing one of the associated HLA alleles was highly statistically significant ($p = 0.0002$ by Fisher exact test; Supplement 4 and Table S4).

Analysis of the crystal structure of several ternary αβTCR/peptide/MHC class I complexes (reviewed in [10]) has revealed that the CDR3 loop of the TCRβ chain primarily makes contact with bound peptide, rather than the α_1 and α_2 helices of the MHC class I heavy chain. The identification of a much larger than expected overlap in the naïve $CD8^+$ CDR3 repertoires of individuals with few, if any, shared HLA-A, -B, or -C alleles suggests that the ensemble of self peptides that participates in positive and negative selection of the repertoire may likewise share significant overlap despite the distinct peptide-binding characteristics of different HLA alleles.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature. Aug 4; 1988 334(6181):395–402. [PubMed: 3043226]
2. Robins HS, Campregher PV, Srivastava SK, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. Nov 5; 2009 114(19):4099–4107. [PubMed: 19706884]
3. Monod, M Yousfi; Giudicelli, V.; Chaume, D.; Lefranc, MP. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics. Aug 4; 2004 20(Suppl 1):i379–385. [PubMed: 15262823]
4. Malissen M, McCoy C, Blanc D, et al. Direct evidence for chromosomal inversion during T-cell receptor beta-gene rearrangements. Nature. Jan 2-8; 1986 319(6048):28–33. [PubMed: 3484541]
5. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? Nat Rev Immunol. Mar; 2008 8(3):231–238. [PubMed: 18301425]
6. Gilfillan S, Bachmann M, Trembleau S, et al. Efficient immune responses in mice lacking N-region diversity. Eur J Immunol. Nov; 1995 25(11):3115–3122. [PubMed: 7489751]
7. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. J Exp Med. Nov 5; 2001 194(9): 1385–1390. [PubMed: 11696602]
8. Argaet VP, Schmidt CW, Burrows SR, et al. Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus. J Exp Med. Dec 1; 1994 180(6): 2335–2340. [PubMed: 7964506]
9. Venturi V, Chin HY, Asher TE, et al. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. J Immunol. Dec 1; 2008 181(11):7853–7862. [PubMed: 19017975]
10. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol. 2006; 24:419–466. [PubMed: 16551255]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

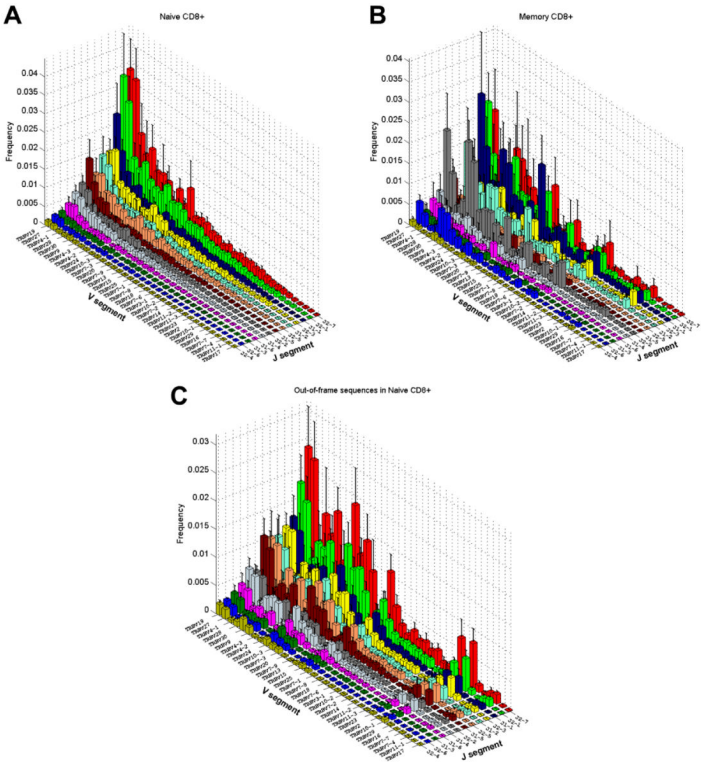

Figure 1.

Histograms depicting the mean utilization frequency of specific Vβ–Jβ gene segment combinations in the TCRβ chains expressed in naïve (A) and memory (B) $CD8^+$ T cells of seven healthy adults. All 13 Jβ segments are indicated along one axis, and the 38 of the 54 Vβ segments are indicated along the other axis. Combinations containing gene segments belonging to the Vβ5, Vβ6, and Vβ12 families are not displayed, as the segments in these families have extremely high sequence similarity at their 3′ ends and could not be unambiguously distinguished given the 60-nt sequence reads obtained in this study. (C) Histogram of the mean utilization frequency of specific Vβ–Jβ gene segment combinations in TCRβ CDR3 sequences observed in naïve $CD8^+$ T cells and predicted to generate out-of-frame TCRβ transcripts that would not encode functional TCRβ chain proteins.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

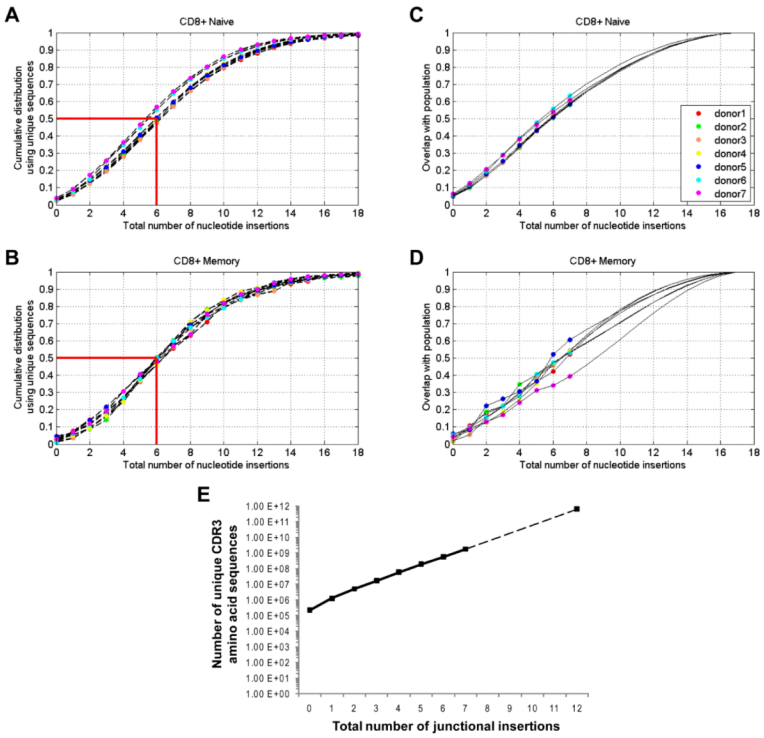

Figure 2.

Cumulative distribution of unique TCRβ CDR3 amino acid sequences in naïve (A) and memory (B) $CD8^+$ T cells in the blood of seven healthy adults, as a function of the total number of nucleotide insertions at the V_β–D_β and D_β–J_β junctions. (C),(D) The TCRβ CDR3 sequences observed in the $CD8^+$ naïve and memory compartments, respectively, of each donor were compared to the complete set of unique sequences generated by a model of TCRβ VDJ rearrangement that allows deletion of up to ten nucleotides adjacent to the RSS of the V_β, D_β, and J_β gene segments, followed by less than or equal to the indicated number of total nucleotide insertions at the two junctions. The fraction of observed sequences in the seven donors that match a sequence generated by the model is shown for naïve cells in (C) and memory cells in (D). (E) The exact number of unique TCRβ CDR3 sequences predicted by the model for 0, 1, 2, …, 7 total insertions, as well as an estimate of the number of sequences predicted by the model for 12 total insertions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

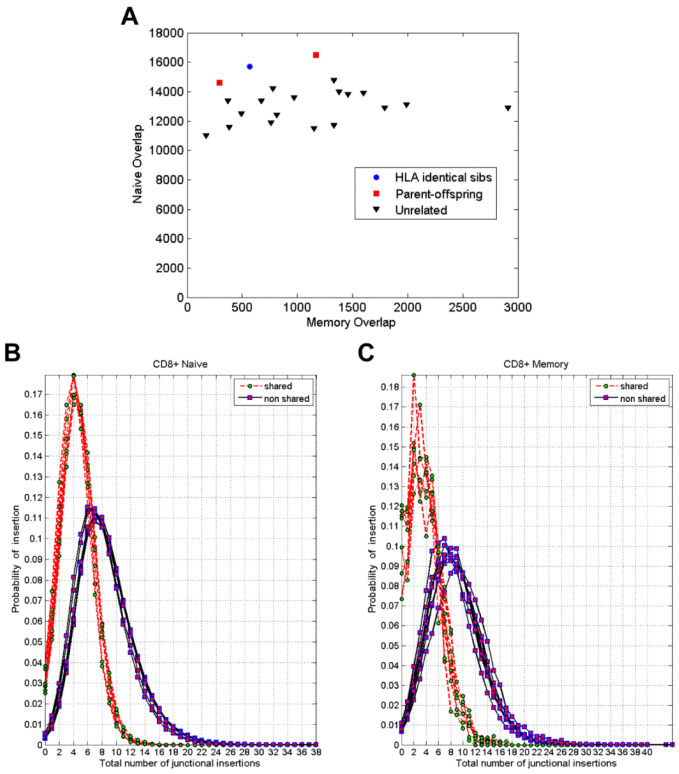

Figure 3.

Characteristics of CD8$^+$ TCRβ CDR3 amino acid sequences that were observed in at least two of the seven individuals. (A) Number of shared sequences in the naïve and memory CD8$^+$ CDR3 repertoires of every possible pair of individuals, with the HLA-identical sisters indicated by the blue diamond, each of those sisters paired with their mother indicated by the orange squares, and the remaining pairs, all of which contained two unrelated individuals, indicated by the black triangles. There are 7!/2!5! = 21 different pairs. (B), (C) Frequency distribution of shared (red curves) and non-shared (blue curves) CDR3 sequences observed in the naïve (B) and memory (C) CD8$^+$ compartments of every possible pair of individuals as a function of the total number of nucleotide insertions at the V_β–D_β and D_β–J_β junctions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

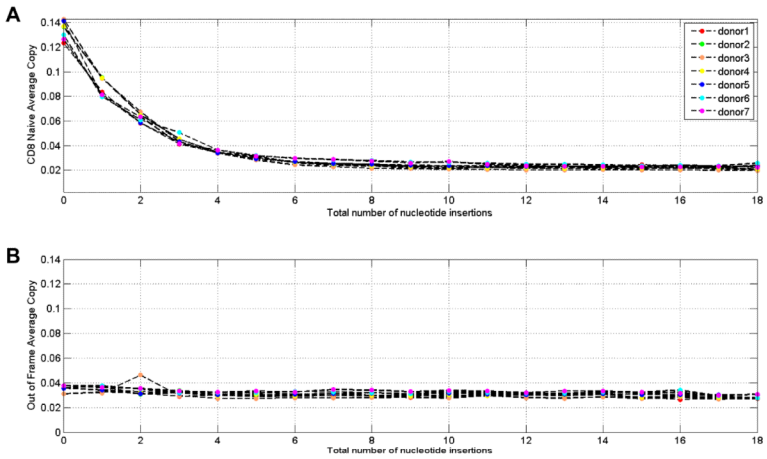

Figure 4.

Relative abundance of (A) in-frame, read-through and (B) out-of-frame TCRβ CDR3 sequences as a function of the total number of nucleotide insertions at the V_β–D_β and D_β–J_β junctions observed in naïve $CD8^+$ T cells from each of the seven donors studied.