Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity

William S DeWitt III[1,2], Anajane Smith[3], Gary Schoch[3], John A Hansen[3,4], Frederick A Matsen IV[1,2], Philip Bradley[1,5*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [2]Department of Genome Sciences, University of Washington, Seattle, WA; [3]Clinical Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [4]Department of Medicine, University of Washington, Seattle, WA; [5]Institute for Protein Design, University of Washington, Seattle, WA, USA.

*For correspondence: pbradley@fredhutch.org

Abstract

The T cell receptor (TCR) repertoire encodes immune exposure history through the dynamic formation of immunological memory. Statistical analysis of repertoire sequencing data has the potential to decode disease associations from large cohorts with measured phenotypes. However, the repertoire perturbation induced by a given immunological challenge is conditioned on genetic background via major histocompatibility complex (MHC) polymorphism. We explore associations between MHC alleles, immune exposures, and shared TCRs in a large human cohort. Using a previously published repertoire sequencing dataset augmented with high-resolution MHC genotyping, our analysis reveals rich structure: striking imprints of common pathogens, clusters of co-occurring TCRs that may represent markers of shared immune exposures, and substantial variations in TCR-MHC association strength across MHC loci. Guided by atomic contacts in solved TCR:peptide-MHC structures, we identify sequence covariation between TCR and MHC. These insights and our analysis framework lay the groundwork for further explorations into TCR diversity.

1 Introduction

T cells are the effectors of cell-mediated adaptive immunity in jawed vertebrates. To control a broad array of pathogens, massive genetic diversity in loci encoding the T cell receptor (TCR) is generated somatically throughout an individual's life via a process called V(D)J recombination. All nucleated cells regularly process and present internal peptide antigens on cell surface molecules called major histocompatibility complex (MHC). Through the interface of TCR and MHC, a rare T cell with a TCR having affinity for a peptide antigen complexed with MHC (pMHC) is stimulated to initiate an immune response to an infected (or cancerous) cell. The responding T cell proliferates clonally, and its progeny

inherit the same antigen-specific TCR, constituting long-term immunological memory of the antigen. The diverse population of TCR clones in an individual (the TCR repertoire) thus dynamically encodes a history of immunological challenges.

Advances in high-throughput TCR sequencing have shown the potential of the TCR repertoire as a personalized diagnostic of pathogen exposure history, cancer, and autoimmunity (Kirsch et al., 2015; Friedensohn et al., 2017). Public TCRs—defined as TCR sequences seen in multiple individuals and perhaps associated with a shared disease phenotype—have been found in a range of infectious and autoimmune diseases and cancers including influenza, Epstein-Barr virus, and cytomegalovirus infections, type I diabetes, rheumatoid arthritis, and melanoma (Venturi et al., 2008; Li et al., 2012; Madi et al., 2017; Pogorelyy et al., 2017; Dash et al., 2017; Glanville et al., 2017; Chu et al., 2018; Pogorelyy et al., 2018). By correlating occurrence patterns of public TCRβ chains with cytomegalovirus (CMV) serostatus across a large cohort of healthy individuals, Emerson et al. identified a set of CMV-associated TCR chains whose aggregate occurrence was highly predictive of CMV seropositivity (Emerson et al., 2017). Staining with multimerized pMHC followed by flow cytometry has been used to isolate and characterize large populations of T cells that bind to defined pMHC epitopes (Dash et al., 2017; Glanville et al., 2017), providing valuable data on the mapping between TCR sequence and epitope specificity. We and others have leveraged these data to develop learning-based models of TCR:pMHC interactions, using TCR distance measures (Dash et al., 2017), CDR3 sequence motifs (Glanville et al., 2017) and k-mer frequencies (Cinelli et al., 2017), and other techniques.

MHC proteins in humans are encoded by the human leukocyte antigen (HLA) loci, among the most polymorphic in the human genome (Robinson et al., 2014). Within an individual, six major antigen-presenting proteins are each encoded by polymorphic alleles. The set of these alleles comprise the individual's HLA type, which is unlikely to be shared with an unrelated individual and which determines the subset of peptide epitopes presented to T cells for immune surveillance. Specificity of a given TCR for a given antigen is biophysically modulated by MHC structure: MHC binding specificity determines the specific antigenic peptide that is presented, and the TCR binds to a hybrid molecular surface composed of peptide- and MHC-derived residues. Thus, population-level studies of TCR-disease association are severely complicated by a dependence on individual HLA type.

Here we report an analysis of the occurrence patterns of public TCRs in a cohort of 666 healthy volunteer donors, in which information on only TCR sequence and HLA association guide us to inferences concerning disease history. To complement deep TCRβ repertoire sequencing available from a previous study (Emerson et al., 2017), we have assembled high-resolution HLA typing data at the major class I and class II HLA loci on the same cohort, as well as information on age, sex, ethnicity, and CMV serostatus. We focus on statistical association of TCR occurrence with HLA type, and show that many of the most highly HLA-associated TCRs are likely responsive to common pathogens:

for example, eight of the ten TCRβ chains most highly associated with the HLA-A*02:01 allele are likely responsive to one of two viral epitopes (influenza $M1_{58}$ and Epstein-Barr virus $BMLF1_{280}$). We introduce new approaches to cluster TCRs by primary sequence and by the pattern of occurrences among individuals in the cohort, and we identify highly significant TCR clusters that may indicate markers of immunological memory. Four of the top five most significant clusters appear linked with common pathogens (parvovirus B19, influenza virus, CMV, and Epstein-Barr virus), again highlighting the impact of viral pathogens on the public repertoire. We also find HLA-unrestricted TCR clusters, some likely to be mucosal-associated invariant T (MAIT) cells, which recognize bacterial metabolites presented by non-polymorphic MR1 proteins, rather than pMHC (Kjer-Nielsen et al., 2012). Our global, unbiased analysis of TCR-HLA association identifies striking variation in association strength across HLA loci and highlights trends in V(D)J generation probability and degree of clonal expansion that illuminate selection processes in cellular immunity. Guided by structural analysis, we used our large dataset of HLA-associated TCRβ chains to identify statistically significant sequence covaration between the TCR CDR3 loop and the DRB1 allele sequence that preserves charge complementarity at the TCR:pMHC interface. These analyses help elucidate the complex dependence of TCR sharing on HLA type and immune exposure, and will inform the growing number of studies seeking to identify TCR-based disease diagnostics.

2 Results

2.1 The matrix of public TCRs

Of the around 80 million unique TCRβ chains (defined by V-gene family and CDR3 sequence) in the 666 cohort repertoires, about 11 million chains are found in at least two individuals and referred to here as *public* chains (for a more nuanced examination of TCR chain sharing see Elhanati et al., 2018). The occurrence patterns of these public TCRβs—the subset of subjects in which each distinct chain occurs—can be thought of as forming a very large binary matrix M with about 11 million rows and 666 columns. Entry $M_{i,j}$ contains a one or a zero indicating presence or absence, respectively, of TCR i in the repertoire of subject j (ignoring for the moment the abundance of TCR i in repertoire j). Emerson et al. (2017) demonstrated that this binary occurrence matrix M encodes information on subject genotype and immune history: they were able to successfully predict HLA-A and HLA-B allele type and CMV serostatus by learning sets of public TCRβ chains with occurrence patterns that were predictive of these features. Specifically, each feature—such as the presence of a given HLA allele (e.g. HLA-A*02) or CMV seropositivity—defines a subset of the cohort members positive for that feature, and can be encoded as a vector of 666 binary digits. This phenotype occurrence pattern of zeros and ones can be compared to the occurrence patterns of all the public TCRβ chains to identify similar

patterns, as quantified by a p-value for significance of co-occurrence across the 666 subjects; thresholding on this p-value produces a subset of significantly associated TCRβ chains whose collective occurrence in a repertoire was found by Emerson et al. to be predictive of the feature of interest (in cross-validation and, for CMV, on an independent cohort). Generalizing from these results, it is reasonable to expect that other common immune exposures may be encoded in the occurrence matrix M, and that these encodings could be discovered if we had additional phenotypic data to correlate with TCR occurrence patterns. In this study, we set out to discover these encoded exposures *de novo*, without additional phenotypic correlates, by learning directly from the structure of the occurrence matrix M and using as well the sequences of the TCRβ chains (both their similarities to one another and to TCR sequences characterized in the literature). To support this effort we assembled additional HLA typing data for the subjects, now at 4-digit resolution and including MHC class II alleles, and we compiled a dataset of annotated TCRβ chains by combining online TCR sequence databases, structurally characterized TCRs, and published studies (see Methods; Shugay et al., 2017; Tickotsky et al., 2017; Berman et al., 2000; Dash et al., 2017; Glanville et al., 2017; Song et al., 2017; Kasprowicz et al., 2006). Here we describe the outcome of this discovery process, and we report a number of intriguing general observations about the role of HLA in shaping the T cell repertoire.

2.2 Globally co-occurring TCR pairs form clusters defined by shared associations

We hypothesized that we could identify unknown immune exposures encoded in the public repertoire by comparing the occurrence patterns of individual TCRβ chains to one another. A subset of TCRβ chains that strongly co-occur among the 666 subjects might correspond to an unmeasured immune exposure that is common to a subset of subjects. Since shared HLA restriction could represent an alternative explanation for significant TCR co-occurrence, we also compared the TCR occurrence patterns to the occurrence patterns for class I and class II HLA alleles. We began by analyzing TCR occurrence patterns over the full set of cohort members. For each pair of public TCRβ chains t_1 and t_2 we computed a co-occurrence p-value $P_{\mathrm{CO}}(t_1, t_2)$ that reflects the probability of seeing an equal or greater overlap of shared subjects (i.e., subjects in whose repertoires both t_1 and t_2 are found) if the occurrence patterns of the two TCRs had been chosen randomly (for details, see Methods). In a similar manner we computed, for each HLA allele a and TCR t, an association p-value $P_{\mathrm{HLA}}(a, t)$ that measures the degree to which TCR t tends to occur in subjects positive for allele a. Finally, for each pair of strongly co-occurring ($P_{\mathrm{CO}} < 1 \times 10^{-8}$) TCRβ chains t_1 and t_2, we looked for a mutual HLA association that might explain their co-occurrence, by finding the allele having the strongest association with both t_1 and t_2, and noting its association p-value:

$$P_{\mathrm{HLA}}(t_1, t_2) = \min_{a \in \mathcal{A}} \max_{t \in \{t_1, t_2\}} P_{\mathrm{HLA}}(a, t),$$

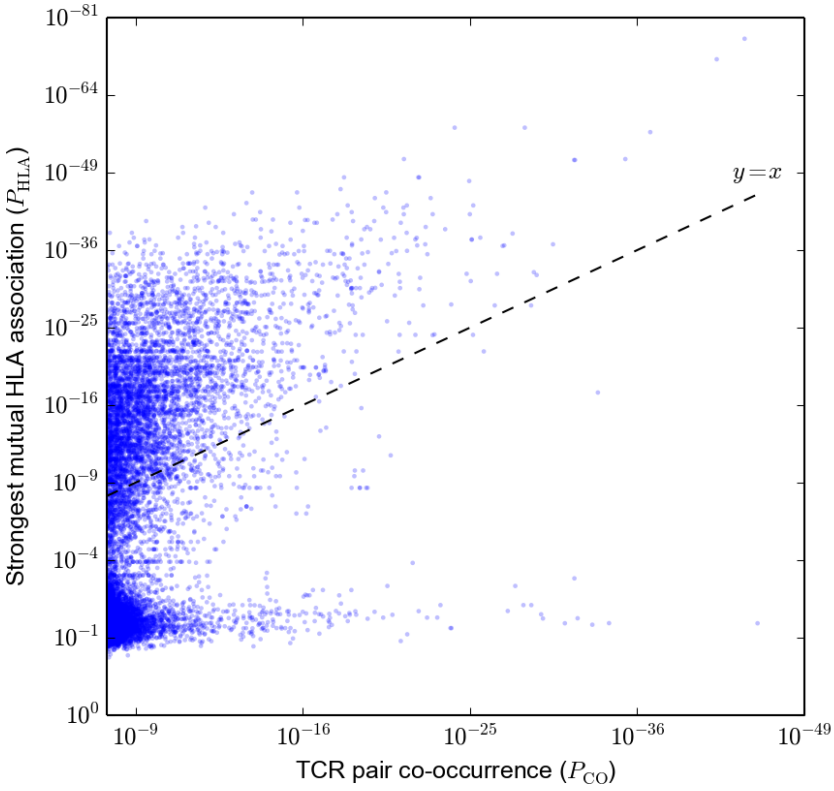


Figure 1. Strongly co-occurring TCR pairs form two broad classes distinguished by HLA-association strength. The co-occurrence p-value P_{CO} for each pair of public TCRs is plotted (x-axis) against the HLA-association p-value P_{HLA} for the HLA allele with the strongest mutual association with that TCR pair (y-axis).

where $\mathcal{A}$ denotes the set of all HLA alleles. In words, we take the p-value of the strongest HLA allele association with the TCR pair, where the association of an HLA allele with a TCR pair is defined by the weakest association of the allele among the individual TCRs.

Based on this analysis, we identified two broad classes of strongly co-occurring TCR pairs (Figure 1): those with a highly significant shared HLA association, where the co-occurrence of the two TCRs can be explained by a shared HLA allele association (i.e. a common HLA restriction), and those with only modest shared HLA-association p-value, for which another explanation of co-occurrence must be sought. Points above the dashed $y = x$ line correspond to pairs of TCRs for which there exists an HLA allele whose co-occurrence with each of the TCRs is stronger than their mutual co-occurrence, while for points below the line no such HLA allele was present in the dataset.

We used a neighbor-based clustering algorithm, DBSCAN (Ester et al., 1996), to link strongly co-occurring TCR pairs together to form larger correlated clusters (see Methods), and then investigated phenotype associations with these clusters. At an approximate family-wise error rate of 0.05 (see Methods), we identified 28 clusters of co-occurring TCRs, with sizes ranging from 7 to 386 TCRs (Figure 2). Given one of these clusters of co-occurring TCRs, we can count the number of cluster member TCRs found in each subject's repertoire. The aggregate occurrence pattern of the cluster can be visualized as a rank plot of this cluster TCR count over the subjects (the black curves in Figure 2B-C). This ranking can also be compared with other phenotypic or genotypic features of the same subjects. In particular, by comparing this aggregate occurrence pattern to a control pattern generated by repeatedly choosing equal numbers of subjects independently at random (dotted green lines in Figure 2B-C), we can identify a subset of the cohort with an apparent enrichment of cluster member TCRs and look for overlap between this subset and other defined cohort features. Performing this comparison against the occurrence patterns of class I and class II HLA alleles revealed that the majority of the TCR clusters were strongly associated with at least one HLA allele (as depicted for a DRB1*15:01-associated cluster in Figure 2B and summarized in Figure 2A).

In addition, there were two large clusters of TCRs which were not strongly associated with any of the typed HLA alleles. Visual inspection of the CDR3 regions of TCRs in one of these clusters revealed a distinctive 'YV' C-terminal motif that is characteristic of the TRBJ2-7*02 allele (Figure 2–Figure Supplement 1), and indeed the 41 subjects whose repertoires indicated the presence of this genetic variant were exactly the 41 subjects enriched for members of this TCR cluster (Figure 2C). This demonstrated that population diversity in germline allele sets manifests as occurrence pattern clustering. The other large, non-HLA associated TCR cluster had a number of distinctive features as well: strong preference for the TRBV06 family, followed by TRBV20 and TRBV04 (Figure 2–Figure Supplement 2); low numbers of inserted 'N' nucleotides; and a skewed age distribution biased toward younger subjects (Figure 2–Figure Supplement 3). These features, together with the lack of apparent HLA restriction, suggested that this cluster represented an invariant T cell subset, specifically

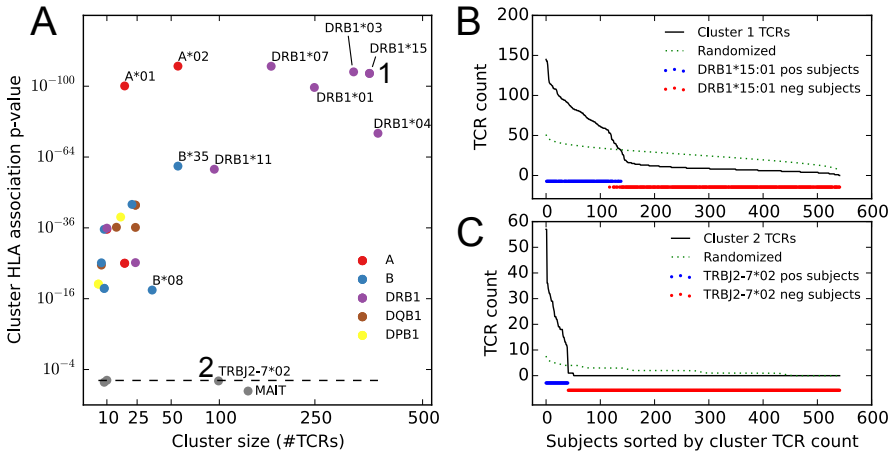


Figure 2. Clustering public TCRβ chains by co-occurrence over the full cohort identifies associations with HLA and TRBJ alleles as well as an invariant T cell subset. **(A)** Cluster size (x-axis) versus the p-value of the most significant HLA allele association (y-axis), with markers colored according to the locus of the associated allele. Dashed line indicates random expectation based on the total number of alleles, assuming independence. **(B)** Count of cluster member TCRs found in each subject for the cluster labeled '1' in (A). The dotted line represents an averaged curve based on randomly and independently selecting subject sets for each member TCR. Red and blue dots indicate the occurrence of the DRB1*15:01 allele in the cohort. **(C)** Count of cluster member TCRs found in each subject for the cluster labeled '2' in (A). The dotted line again represents a control pattern, and the red and blue dots indicate the occurrence of the TRBJ2-7*02 allele.

Figure 2–Figure supplement 1. TCRdist tree of the members of the TRBJ2-7*02-associated cluster.

Figure 2–Figure supplement 2. TCRdist tree of the members of the putative MAIT cell cluster.

Figure 2–Figure supplement 3. More details on the MAIT cell cluster: subject age and N-nucleotide insertion distributions; TCRα chains paired with cluster member TCRβ chains in the pairSEQ dataset of Howie et al. (2015).

Table 1. The top 50 most significant HLA-associated public TCRβ chains and the top 10 for A*02:01 (indicated in bold).

Association *p*-value	Overlap[a]	TCR subjects[b]	HLA subjects[c]	Total subjects[d]	V-family	CDR3	HLA allele[e]	Epitope annotation
3.7×10^{-90}	231	267	268	629	TRBV19	CASSIRSSYEQYF	**A*02:01**	Influenza virus
2.4×10^{-72}	179	191	268	629	TRBV29	CSVGTGGTNEKLFF	**A*02:01**	Epstein-Barr virus
3.8×10^{-66}	107	124	134	522	TRBV20	CSARNRDYGYTF	DRB1*03:01-DQ	
1.9×10^{-65}	92	95	151	630	TRBV05	CASSLVVSPYEQYF	DRB1*07:01	
6.7×10^{-64}	91	94	134	522	TRBV30	CAWSRDSGSGNTIYF	DRB1*15:01-DQ	
7.5×10^{-59}	51	53	66	630	TRBV15	CATSREEGDGYTF	B*35:01	
3.6×10^{-57}	89	96	134	522	TRBV11	CASSPGQGPGNTIYF	DRB1*15:01-DQ	
7.4×10^{-56}	57	57	95	630	TRBV02	CASSENQGSQPQHF	DRB1*04:01	
1.5×10^{-52}	86	87	184	629	TRBV06	CASSYDSGTGELFF	C*07:01	
3.3×10^{-52}	136	143	268	629	TRBV19	CASSIRSAYEQYF	**A*02:01**	Influenza virus
1.2×10^{-51}	71	96	94	630	TRBV27	CASSLGGQNYGYTF	B*44:02	
1.8×10^{-50}	52	52	94	630	TRBV28	CASSSSPLNYGYTF	DRB1*01:01	
3.8×10^{-49}	69	71	142	630	TRBV04	CASSPGQGEGYEQYF	B*08:01	Epstein-Barr virus
6.3×10^{-49}	92	98	189	629	TRBV11	CASSFGQMNTEAFF	A*01:01	
1.3×10^{-48}	73	75	156	630	TRBV18	CASSPPTESYGYTF	B*07:02	
3.2×10^{-48}	79	87	151	630	TRBV14	CASSQAGMNTEAFF	DRB1*07:01	
8.7×10^{-47}	49	49	95	630	TRBV11	CASSLDQGGSSSYNEQFF	DRB1*04:01	
3.2×10^{-46}	50	51	95	630	TRBV20	CSAQREYNEQFF	DRB1*04:01	
3.3×10^{-46}	68	69	134	522	TRBV05	CASSFWGRDTQYF	DRB1*03:01-DQ	
3.3×10^{-46}	54	59	94	630	TRBV05	CASSWTGGGGANVLTF	DRB1*01:01	
3.1×10^{-45}	54	60	94	630	TRBV02	CASSEARGAGQPQHF	DRB1*01:01	
1.4×10^{-44}	41	42	69	630	TRBV14	CASSPLGPGNTIYF	DRB1*11:01	
2.4×10^{-43}	92	121	134	522	TRBV07	CASSPTGLQETQYF	DRB1*03:01-DQ	
4.1×10^{-43}	43	52	61	630	TRBV19	CASSPTGGIYEQYF	B*44:03	Multiple sclerosis
4.5×10^{-43}	39	40	66	629	TRBV10	CASSESPGNSNQPQHF	C*12:03	
6.7×10^{-43}	76	86	134	522	TRBV28	CASRGRPEAFF	DRB1*15:01-DQ	
7.5×10^{-43}	50	54	94	630	TRBV19	CASSPTQNTEAFF	DRB1*01:01	
1.7×10^{-42}	84	110	142	630	TRBV07	CASSSGPNYEQYF	B*08:01	
1.7×10^{-42}	61	81	95	630	TRBV05	CASSFPGEDTQYF	DRB1*04:01	
1.3×10^{-41}	47	49	95	630	TRBV18	CASSPPAGAAYEQYF	DRB1*04:01	
1.5×10^{-41}	75	87	151	630	TRBV28	CASSLTSGGQETQYF	DRB1*07:01	
2.3×10^{-41}	64	67	151	630	TRBV07	CASSLGQGFYNSPLHF	DRB1*07:01	
8.2×10^{-40}	77	92	134	522	TRBV19	CASSISVYGYTF	DRB1*15:01-DQ	
2.4×10^{-39}	43	54	66	630	TRBV10	CAISTGDSNQPQHF	B*35:01	Epstein-Barr virus
3.4×10^{-39}	115	193	156	630	TRBV09	CASSGNEQFF	B*07:02	
9.5×10^{-39}	151	260	189	629	TRBV19	CASSIRDSNQPQHF	A*01:01	
1.2×10^{-38}	100	103	268	629	TRBV20	CSARDGTGNGYTF	**A*02:01**	Epstein-Barr virus
1.3×10^{-38}	56	60	130	629	TRBV25	CASSEYSLTDTQYF	C*04:01	
2.1×10^{-38}	109	116	268	629	TRBV20	CSARDRTGNGYTF	**A*02:01**	Epstein-Barr virus
2.3×10^{-38}	102	106	268	629	TRBV19	CASSVRSSYEQYF	**A*02:01**	Influenza virus
6.4×10^{-38}	54	54	151	630	TRBV10	CAISESQDLNTEAFF	DRB1*07:01	
1.1×10^{-37}	43	45	94	630	TRBV07	CASSLAGPPNSPLHF	DRB1*01:01	
1.2×10^{-37}	44	60	66	630	TRBV09	CASSARTGELFF	B*35:01	Epstein-Barr virus
3.3×10^{-37}	79	88	189	629	TRBV19	CASSIDGEETQYF	A*01:01	
5.4×10^{-37}	64	70	134	522	TRBV05	CASSLESPNYGYTF	DRB1*03:01-DQ	
2.0×10^{-36}	38	43	69	630	TRBV06	CASGAGHTDTQYF	DRB1*11:01	
2.9×10^{-36}	54	55	151	630	TRBV05	CASSLVVQPYEQYF	DRB1*07:01	
3.3×10^{-36}	57	81	95	630	TRBV11	CASSPGQDYGYTF	DRB1*04:01	
2.4×10^{-35}	50	53	109	522	TRBV27	CASNRQGPNTEAFF	DQB1*03:01-DQA1*05:05	
5.7×10^{-35}	75	95	134	522	TRBV18	CASSGQANTEAFF	DRB1*03:01-DQ	
2.2×10^{-33}	86	88	268	629	TRBV14	CASSQSPGGTQYF	**A*02:01**	Epstein-Barr virus
1.8×10^{-32}	84	86	268	629	TRBV10	CASSEDGMNTEAFF	**A*02:01**	
4.3×10^{-32}	86	89	268	629	TRBV05	CASSLEGQASSYEQYF	**A*02:01**	Melanoma
4.3×10^{-32}	86	89	268	629	TRBV29	CSVGSGGTNEKLFF	**A*02:01**	Epstein-Barr virus

[a] Number of subjects positive for both the TCRβ chain and the indicated HLA allele.
[b] Number of subjects positive for the TCRβ chain with available HLA typing at the corresponding locus.
[c] Number of subjects positive for the indicated HLA allele.
[d] Total number of subjects with available HLA typing at the corresponding locus.
[e] The following DR-DQ haplotype abbreviations are used: DRB1*03:01-DQ (DRB1*03:01-DQA1*05:01-DQB1*02:01) and DRB1*15:01-DQ (DRB1*15:01-DQA1*01:02-DQB1*06:02).

MAIT (mucosal-associated invariant T) cells (Kjer-Nielsen et al., 2012; Venturi et al., 2013; Pogorelyy et al., 2017). Since MAIT cells are defined primarily by their alpha chain sequences, we searched in a recently published paired dataset (Howie et al., 2015) for partner chains of the clustered TCRβ chain sequences, and found a striking number that matched the MAIT consensus (TRAV1-2 paired with TRAJ20/TRAJ33 and a 12 residue CDR3, Figure 2–Figure Supplement 3D). We also looked for these clustered TCRs in a recently published MAIT cell sequence dataset (Howson et al., 2018) and found that 93 of the 138 cluster member TCRs occurred among the $31{,}654$ unique TCRs from this dataset; of these 93 TCRβ chains, 27 were found among the 78 most commonly occurring TCRs in the dataset (the TCRs occurring in at least 7 of the 24 sequenced repertoires), a highly significant overlap ($P < 2 \times 10^{-52}$ in a one-sided hypergeometric test). These concordances indicate that our untargeted approach has detected a well-studied T cell subset *de novo* through analysis of occurrence patterns.

2.3 HLA-associated TCRs

These analyses suggested to us that TCR co-occurrence patterns across the full cohort of subjects are strongly influenced by the distribution of the HLA alleles, in accordance with the expectation that the majority of αβ TCRs are HLA-restricted. Covariation between TCRs responding to the same HLA-restricted epitopes would only be expected in subjects positive for the restricting alleles, with TCR presence and absence outside these subjects likely introducing noise into the co-occurrence analysis. Thus we next analyzed patterns of TCR co-occurrence within subsets of the cohort positive for specific HLA alleles, and we restricted our co-occurrence analysis to TCRs having a statistically significant association with the specific allele defining the cohort subset. At a false discovery rate of 0.05 (estimated from shuffling experiments; see Methods), we were able to assign $16{,}951$ TCRβ sequences to an HLA allele (or alleles: DQ and DP alleles were analyzed as αβ pairs, and there were 5 DR/DQ haplotypes whose component alleles were so highly correlated across our cohort that we could not assign TCR associations to individual DR or DQ components; see Methods). Table 1 lists the top 50 HLA-associated TCR sequences by association p-value and top 10 associated TCRs for the well-studied A*02:01 allele.

We find that 8 of the top 10 A*02:01-associated TCRs have been previously reported and annotated as being responsive to viral epitopes, specifically influenza $M1_{58}$ and Epstein-Barr virus (EBV) $BMLF1_{280}$ (Shugay et al., 2017; Tickotsky et al., 2017). Moreover, each of these 8 TCRβ chains is present in a recent experimental dataset (Dash et al., 2017) that included tetramer-sorted TCRs positive for these two epitopes; each TCR has a clear similarity to one of the consensus epitope-specific repertoire clusters identified in that work, with the EBV TRBV20, TRBV29, and TRBV14 TCRs, respectively, matching the three largest branches of the $BMLF1_{280}$ TCR tree, and the three influenza $M1_{58}$ TCRs all matching the dominant TRBV19 'RS' motif consensus (Figure 10). TCRs with annotation matches are sparser in the top 50 across all alleles, which is

likely due in part to a paucity of experimentally characterized non-A*02 TCRs, however we again see EBV-epitope responsive TCRs (with B*08:01 and B*35:01 restriction).

A global comparison of TCR feature distributions for HLA-associated versus non-HLA-associated TCRs provides further evidence of functional selection. As shown in Figure 3A, HLA-associated TCRs are on average more clonally expanded than a set of background, non-HLA associated TCRs with matching frequencies in the cohort. They also have lower generation probabilities—are harder to make under a simple random model of the VDJ rearrangement process—which suggests that their observed cohort frequencies may be elevated by selection (Figure 3B, see Methods for further details on the calculation of clonal expansion indices and generation probabilities; also see Pogorelyy et al., 2018). Examination of two-dimensional feature distributions suggests that these shifts are correlated, with HLA-associated TCRs showing an excess of lower-probability, clonally expanded TCRs (Figure 3C); this trend appears stronger for class-I associated TCRs than for class II-associated TCRs (Figure 3–Figure Supplement 1).

To give a global picture of TCR-HLA association, we counted the number of significant TCR associations found for each HLA allele in the dataset, and plotted this number against the number of subjects in the cohort with that allele (Figure 4). As expected, the more common HLA alleles have on average greater numbers of associated TCRs (since greater numbers of subjects permit the identification of more public TCRs, and the statistical significance assigned to an observed association of fixed strength grows as the number of subjects increases). What was somewhat more surprising is that the slope of the correlation between cohort frequency and number of associated TCRs varied dramatically among the HLA loci, with HLA-DRB1 alleles having the largest number of associated TCRs for a given allele frequency and HLA-C alleles having the smallest. The best-fit slope for the five DR/DQ haplotypes (12.2) was roughly the sum of the DR (7.99) and DQ (3.39) slopes, suggesting as expected that these haplotypes were capturing TCRs associated with both the DR and DQ component alleles.

2.4 HLA-restricted TCR clusters

We next sought to identify TCR clusters that might represent HLA-restricted responses to shared immune exposures. We performed this analysis for each HLA allele individually, restricting our clustering to the set of TCR chains significantly-associated with that allele and comparing occurrence patterns only over the subset of subjects positive for that allele. The smaller size of many of these allele-positive cohort subsets reduces our statistical power to detect significant clusters using co-occurrence information. To counter this effect, we used TCRdist (Dash et al., 2017) to leverage the TCR sequence similarity which is often present within epitope-specific responses (Dash et al., 2017; Glanville et al., 2017) (e.g., A*02:01 TCRs in Table 1 and Figure 10). We augmented the probabilistic similarity measure used to define neighbors for DBSCAN clustering to incorporate information about TCR sequence similarity, in addition

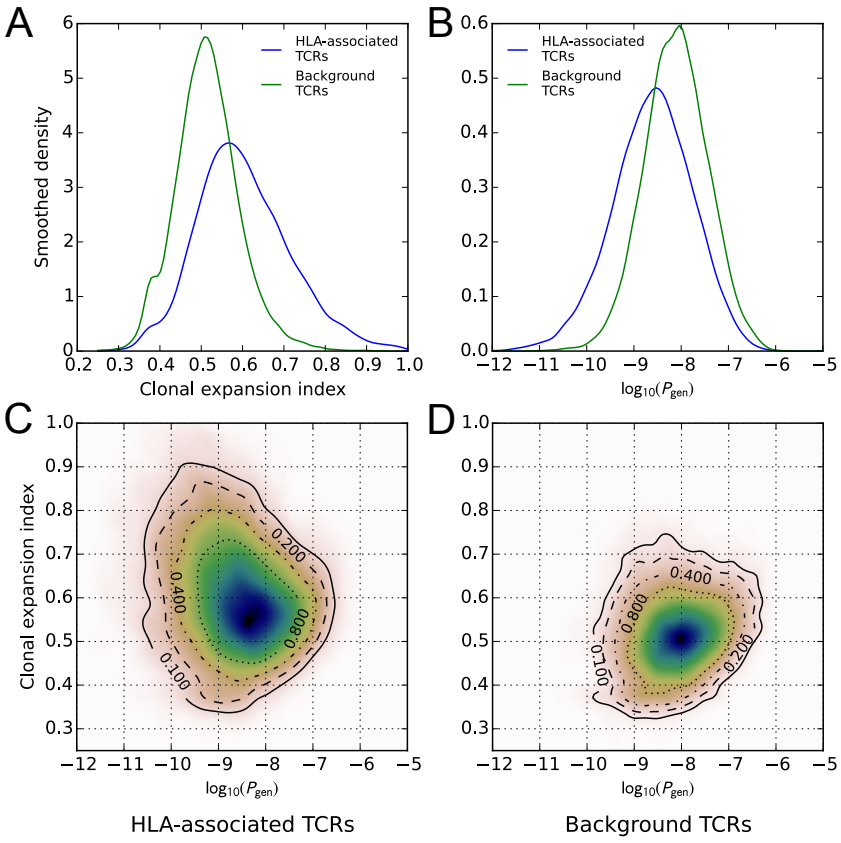


Figure 3. HLA-associated TCRs are more clonally expanded and have lower generation probabilities than equally common, non-HLA associated TCRs. **(A)** Comparison of clonal expansion index distributions for the set of HLA-associated TCRs (blue) and a cohort-frequency matched set of non HLA-associated TCRs (green). **(B)** Comparison of VDJ-rearrangement TCR generation probability (P_{gen}) distributions for the set of HLA-associated TCRs (blue) and a cohort-frequency matched set of non HLA-associated TCRs (green). **(C)** Two-dimensional distribution (P_{gen} versus clonal expansion index) for HLA-associated TCRs. **(D)** Two-dimensional distribution (P_{gen} versus clonal expansion index) for frequency-matched background TCRs.

Figure 3–Figure supplement 1. Two-dimensional feature distributions for HLA-associated TCR subsets defined by HLA locus.

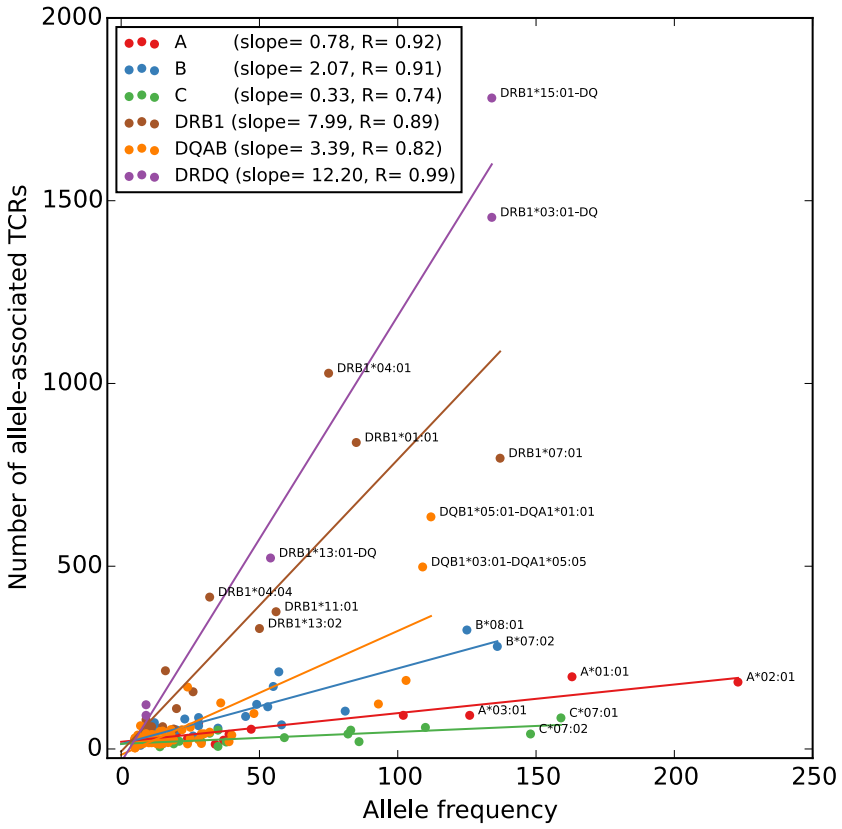


Figure 4. Rates of TCR association vary substantially across HLA loci. The number of HLA-associated TCRs (y-axis) is plotted as a function of allele frequency in the cohort (x-axis). Best fit lines are shown for each locus and also for the set of five DR/DQ haplotypes ('DRDQ') which could not be separated into component alleles in this cohort. The following DR-DQ haplotype abbreviations are used: DRB1*03:01-DQ (DRB1*03:01-DQA1*05:01-DQB1*02:01), DRB1*15:01-DQ (DRB1*15:01-DQA1*01:02-DQB1*06:02), and DRB1*13:01-DQ (DRB1*13:01-DQA1*01:03-DQB1*06:03).

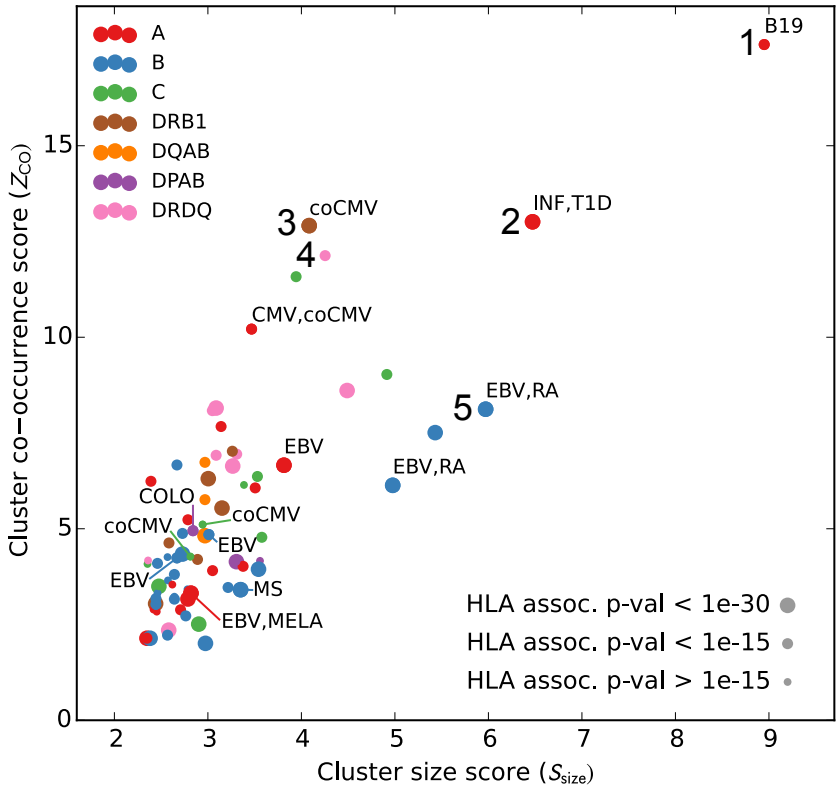


Figure 5. Many HLA-restricted TCR clusters contain TCRβ chains annotated as pathogen-responsive. Each point represents one of the 78 significant HLA-restricted TCR clusters, plotted based on a normalized cluster size score (S_{size}, x-axis) and an aggregate TCR co-occurrence score for the member TCRs (Z_{CO}, y-axis). Markers are colored by the locus of the restricting HLA allele and sized based on the strength of the association between cluster member TCRs and the HLA allele. The database annotations associated to TCRs in each cluster are summarized with text labels using the following abbreviations: B19=parvovirus B19, INF=influenza, EBV=Epstein-Barr Virus, RA=rheumatoid arthritis, MS=multiple sclerosis, MELA=melanoma, T1D=type 1 diabetes, CMV=cytomegalovirus. Clusters labeled 'coCMV' are significantly associated ($P < 1 \times 10^{-5}$) with CMV seropositivity (see main text discussion of cluster #3). Clusters labeled 1–5 are discussed in the text and examined in greater detail in Figure 6.

Figure 5–Figure supplement 1. Distributions of cluster co-occurrence scores on the two validation cohorts.

to cohort co-occurrence (see Methods). We independently clustered each allele's associated TCRs and merged the clustering results from all alleles; using the Holm multiple testing criterion (Holm, 1979) to limit the approximate family-wise error rate to 0.05, we found a total of 78 significant TCR clusters.

We analyzed the sequences and occurrence patterns of the TCRs belonging to these 78 clusters in order to assess their potential biological significance and prioritize them for further study (Table 3). Each cluster was assigned two scores (Figure 5): a size score (S_{size}, x-axis), reflecting the significance of seeing a cluster of that size given the total number of TCRs clustered for its associated allele, and a co-occurrence score (Z_{CO}, y-axis), reflecting the degree to which the TCRs in that cluster co-occur within its allele-positive cohort subset (see Methods). In computing the co-occurrence score, we defined a subset of individuals with an apparent enrichment for the member TCRs in each cluster; the size of this enriched subset of subjects is given in the 'Subjects' column in Table 3. We rank ordered the 78 clusters based on the sum of their size and co-occurrence scores (weighted to equalize dynamic range); the top 5 clusters are presented in greater detail in Figure 6. HLA associations, member TCR and enriched subject counts, cluster center TCR sequences, scores, and annotations for all 78 clusters are given in Table 3.

We found that a surprising number of the most significant HLA-restricted clusters had links to common viral pathogens. For example, the top cluster by both size and co-occurrence (Figure 6, upper panels) is an A*24:02-associated group of highly similar TCRβ chains, five of which can be found in a set of 12 TCRβ sequences reported to respond to the parvovirus B19 epitope `FYTPLADQF` as part of a highly focused CD8+ response to acute B19 infection (Kasprowicz et al., 2006). The subject TCR-counts curve for this cluster (Figure 6, top right panel) shows a strong enrichment of member TCRs in roughly 30% of the A*24:02 repertoires, which is on the low end of prevalence estimates for this pathogen (Heegaard and Brown, 2002) and may suggest that, if cluster enrichment does correlate with B19 exposure, there are likely to be other genetic or epidemiologic factors that determine which B19-exposed individuals show enrichment. The second most significant cluster by both measures is an A*02:01-associated group of TRBV19 TCRs with a high frequency of matches to the influenza M1_{58} response (41/43 TCRs, labeled 'INF-pGIL' for the first three letters of the GILGFVFTL epitope). Notably, the cluster member sequences recapitulate many of the core features of the tree of experimentally identified M1_{58} TCRs (Figure 10): a dominant group of length 13 CDR3 sequences with an 'RS' sequence motif together with a smaller group of length 12 CDR3s with the consensus `CASSIG.YGYTF`.

Rounding out the top five, the third and fifth most significant clusters also appear to be pathogen-associated. Cluster #3 brings together a diverse set of DRB1*07:01-associated TCRβ chains (Figure 6, second page, middle dendrogram), none of which matched our annotation database. However, it was strongly associated with CMV serostatus: As is evident in the subject TCR-counts panel for this cluster (Figure 6, second page, middle right), there is a highly significant ($P < 3 \times 10^{-19}$) association between CMV seropositivity (blue

dots at the bottom of the panel) and cluster enrichment (here defined as a subject TCR count ≥ 3). Finally, the B*08:01-associated cluster #5 (bottom panels in second page Figure 6) appears to be EBV-associated: four of the TCRβ chains in this cluster match TCRs annotated as binding to EBV epitopes (two matches for the B*08:01-restricted FLRGRAYGL epitope and two for the B*08:01-restricted RAKFKQLL epitope). The fact that this cluster brings together sequence-dissimilar TCRs that recognize different epitopes from the same pathogen supports the hypothesis that at least some of the observed co-occurrence may be driven by a shared exposure.

As a preliminary validation of the clusters identified here, we examined the occurrence patterns of cluster member TCRs in two independent cohorts: a set of 120 individuals ("Keck120") that formed the validation cohort for the original Emerson et al. study, and a set of 86 individuals ("Brit86") taken from the aging study of Britanova et al. (2016). Whereas the Keck120 repertoires were generated using the same platform as our 666-member discovery cohort, the Brit86 repertoires were sequenced from cDNA libraries using 5'-template switching and unique molecular identifiers. In the absence of HLA typing information for these subjects, we simply evaluated the degree to which each cluster's member TCRs co-occurred over the entirety of each of these validation cohorts, using the co-occurrence score described above ($Z_{\mathrm{CO}}^{\mathrm{Keck120}}$ and $Z_{\mathrm{CO}}^{\mathrm{Brit86}}$ columns in Table 3). Although rare alleles and cluster-associated exposures may not occur with sufficient frequency in these smaller cohorts to generate co-occurrence signal, co-occurrence scores support the validity of the clusterings identified on the discovery cohort: 94% of the Keck120 scores and 92% of the Brit86 scores are greater than 0, indicating a tendency of the clustered TCRs to co-occur (smoothed score distributions are shown in Figure 5–Figure Supplement 1).

2.5 Covariation between CDR3 sequence and HLA allele

Given our large dataset of HLA-associated TCRβ sequences, we set out to look for correlations between CDR3 sequence and HLA allele sequence. Previous studies have identified correlations between TCR V-gene usage and HLA alleles (Sharon et al., 2016; Blevins et al., 2016). In our previous work on epitope-specific TCRs (Dash et al., 2017), we identified a significant negative correlation between CDR3 charge and peptide charge, suggesting a tendency toward preserving charge complementarity across the TCR:pMHC interface. Although the CDR3 loop primarily contacts the MHC-bound peptide, computational analysis of solved TCR:peptide:MHC structures in the Protein Data Bank (Berman et al., 2000) (see Methods) identified a number of HLA sequence positions that are frequently contacted by CDR3 amino acids (Table 2). For each frequently-contacted HLA position with charge variability among alleles we computed the covariation between HLA allele charge at that position and average CDR3 charge for allele-associated TCRs. Since portions of the CDR3 sequence are contributed by the V- and J-gene germline sequences, and covariations are known to exist between HLA and V-gene usage, we also performed a covariation

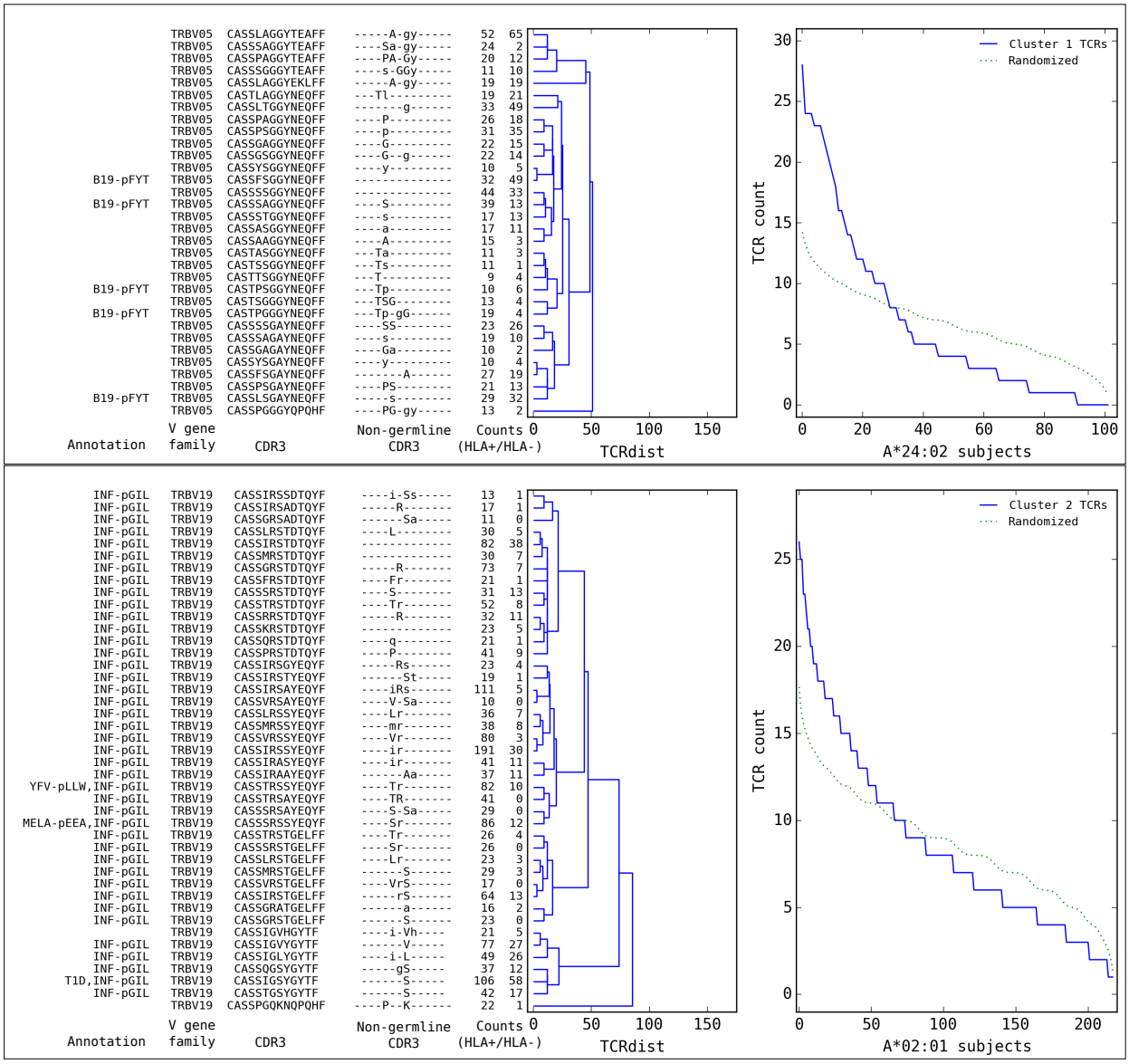
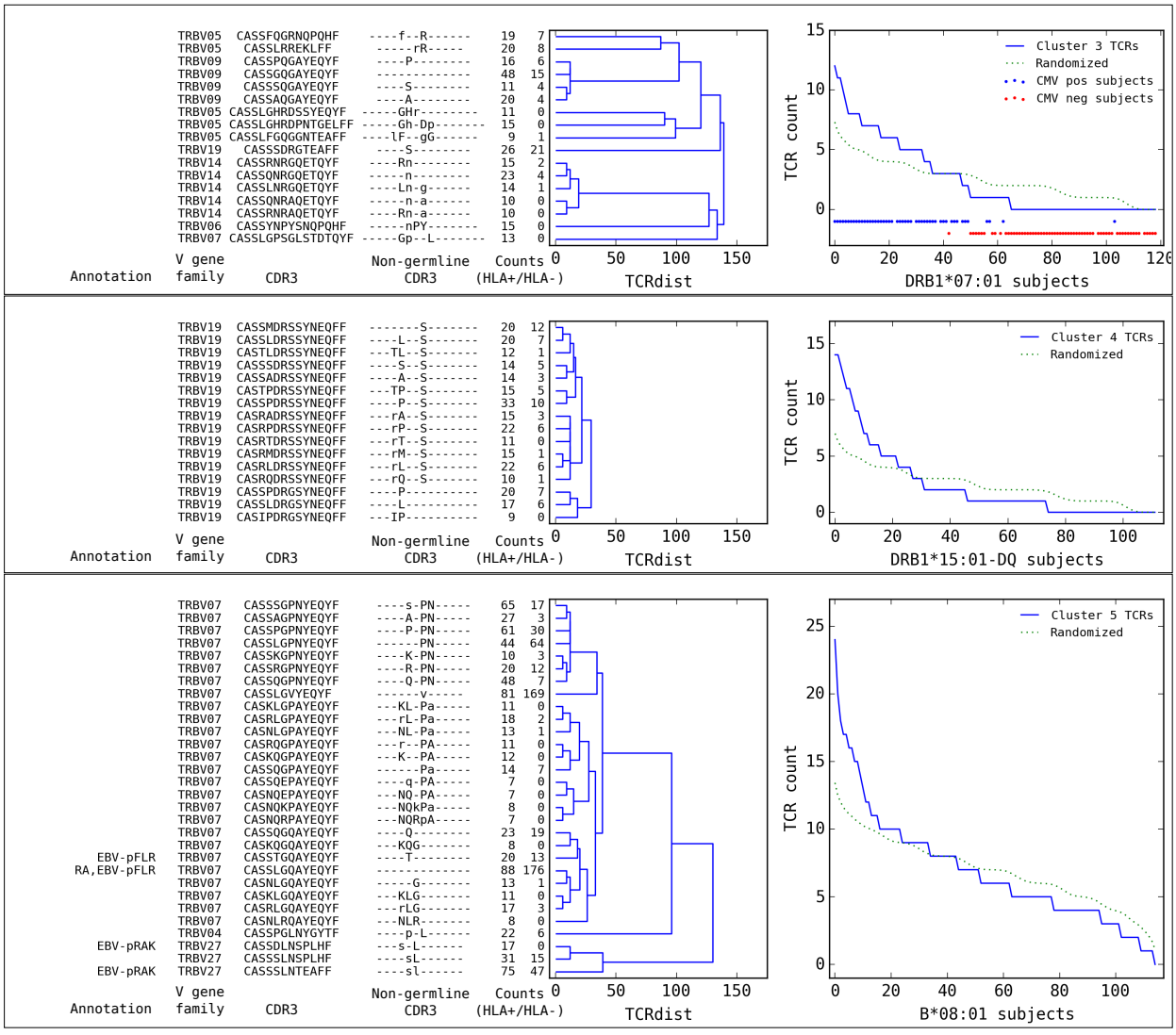


Figure 6. Top five HLA-restricted clusters (continued on following page). Details on the TCR sequences, occurrence patterns, and annotations for the five most significant clusters (labeled 1–5 in Figure 5) based on size and TCR co-occurrence scores. Each panel consists of a TCRdist dendrogram (left side, labeled with annotation, CDR3 sequence, and occurrence counts for the member TCRs) and a per-subject TCR count profile (right side) showing the aggregate occurrence pattern of the member TCRs (blue curve) and a control pattern (green curve) produced by averaging occurrence counts from multiple independent randomizations of the subject set for each TCR. The numbers in the two 'Counts' columns represent the number of HLA+ (left) and HLA- (right) subjects whose repertoire contained the corresponding TCR, where HLA+/- means positive/negative for the restricting allele (for example, A*24:02 in the case of cluster 1). Annotations use the following abbreviations: B19 (parvovirus B19), INF (influenza virus), YFV (yellow fever virus), MELA (melanoma), T1D (type 1 diabetes), EBV (Epstein-Barr virus), RA (rheumatoid arthritis). In cases where the peptide epitope for the annotation match is known, the first three peptide amino acids are given after '-p'. Non-germline CDR3 amino acids with 2 or 3 non-templated nucleotides in their codon are shown in uppercase, while amino acids with only a single non-templated coding nucleotide are shown in lowercase.



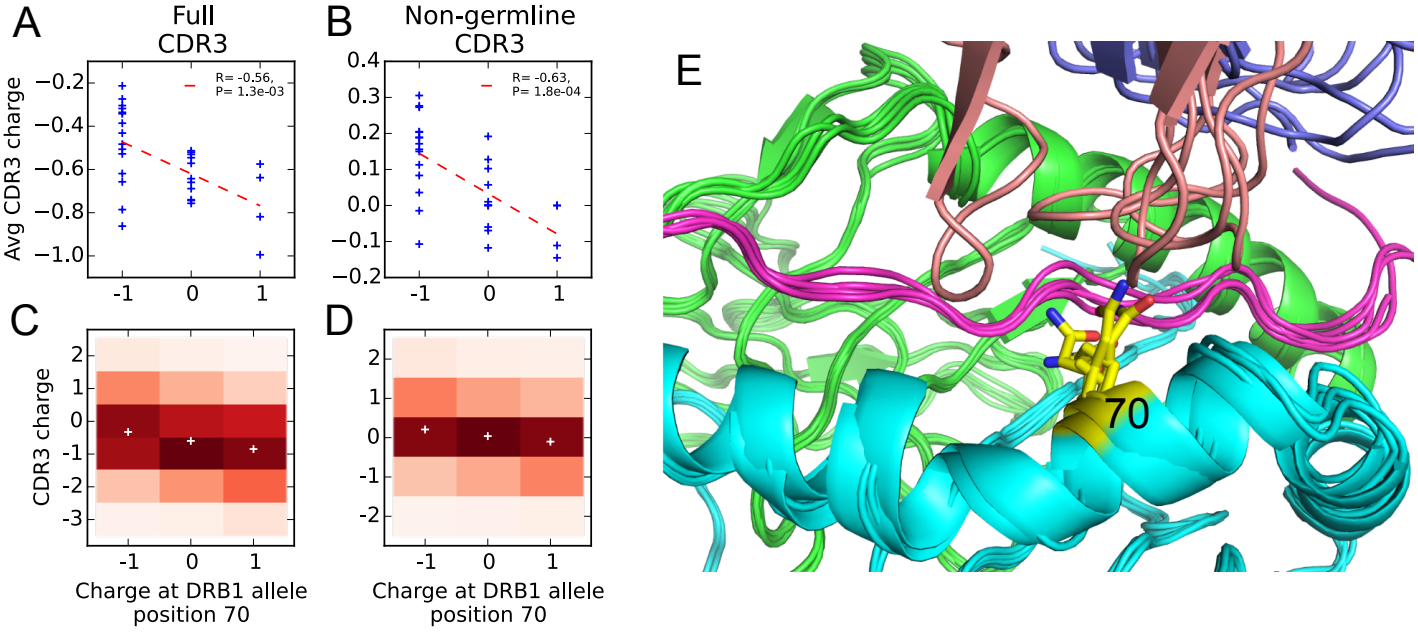


Figure 7. Negative correlation between HLA allele charge at DRB1 position 70 and CDR3 charge of HLA-associated TCRs. **(A-B)** Allele charge (x-axis) versus average CDR3 charge of allele-associated TCRβ chains (y-axis) for 30 HLA-DRB1 alleles. Charge of the CDR3 loop was calculated over the full CDR3 sequence (A) or over the subset of CDR3 amino acids with at least one non-germline coding nucleotide (B). **(C-D)** CDR3 charge distributions for TCRs associated with alleles having defined charge at position 70 (x-axis) using the full (C) or non-germline (D) CDR3 sequence (mean values shown as white pluses). **(E)** Superposition of five TCR:peptide:HLA-DR crystal structures (PDB IDs 1j8h, 2iam, 2wbj, 3o6f, and 4e41; Hennecke and Wiley, 2002; Deng et al., 2007; Harkiolaki et al., 2009; Yin et al., 2011; Deng et al., 2012) showing the DRα chain in green, the DRβ chain in cyan, the peptide in magenta, the TCRβ chain in blue with the CDR3 loop colored reddish brown. The TCRα chain is omitted for clarity, and position 70 is highlighted in yellow.

Table 2. Covariation between HLA allele charge and average CDR3 charge of HLA-associated TCRs for HLA positions frequently contacted by CDR3 amino acids in solved TCR:pMHC crystal structures.

MHC class	Position[a]	Contact frequency[b]	Full CDR3 R-value	Full CDR3 p-value	Non-germline CDR3[c] R-value	Non-germline CDR3[c] p-value	AAs[d]
II-β	70	1.48	−0.47	3.3×10^{-4}	−0.52	6.1×10^{-5}	DEGQR
II-α	64	1.09	−0.15	0.33	−0.07	0.64	ART
I	152	0.47	0.00	0.99	−0.04	0.72	AERTVW
I	151	0.46	0.08	0.50	0.06	0.59	HR
I	69	0.26	−0.13	0.28	−0.14	0.24	ART
I	76	0.21	−0.08	0.49	−0.14	0.25	AEV
I	70	0.12	0.02	0.86	0.08	0.50	HKNQS

[a] Only positions whose charge varies across alleles are included.
[b] Total number of CDR3 residues contacted (using a sidechain heavyatom distance threshold of 4.5Å) divided by number of structures analyzed.
[c] CDR3 charge is calculated over amino acids with at least non-germline coding nucleotide.
[d] Amino acids present at this HLA position.

analysis restricting to 'non-germline' CDR3 sequence positions whose coding sequence is determined by at least one non-templated insertion base (based on the most parsimonious VDJ reconstruction; see Methods). We found a significant negative correlation ($R = -0.47, P < 4 \times 10^{-4}$ for the full CDR3 sequence; $R = -0.52, P < 7 \times 10^{-5}$ for the non-germline CDR3 sequence) between CDR3 charge and the charge at position 70 of the class II beta chain. We did not see a significant correlation for the frequently contacted position on the class II alpha chain, perhaps due to the lack of sequence variation at the DRα locus and/or the more limited number of DQα and DPα alleles. None of the five class I positions showed significant correlations, which could be due to their lower contact frequencies, a smaller average number of associated TCRs (51 for class I versus 309 for class II), bias toward A*02 in the structural database, or noise introduced from multiple contacted positions varying simultaneously. Further analysis of the class II correlation suggested that it was driven largely by HLA-DRB1 alleles: position 70 correlations were -0.56 versus -0.10 for DR and DQ, respectively, over the full CDR3 and -0.64 vs -0.38 for the non-germline CDR3. Figure 7 provides further detail on this DRB1-TCR charge anti-correlation, including a structural superposition showing the proximity of position 70 to the TCRβ CDR3 loop.

2.6 CMV-associated TCRβ chains are largely HLA-restricted

We analyzed the HLA associations of strongly CMV-associated TCRβ chains to gain insight into their predictive power across genetically diverse individuals. Here we change perspective somewhat from earlier sections, in that we select TCRs based on their CMV association and then evaluate HLA associ-

ation, rather than the other way around. In their original study, Emerson et al. identified a set of TCRβ chains that were enriched in CMV seropositive individuals and showed that by counting these CMV-associated TCRβ chains in a query repertoire they could successfully predict CMV serostatus both in cross-validation and on an independent test cohort. The success of this prediction strategy across a diverse cohort of individuals raises the intriguing question of whether these TCRβs are primarily HLA-restricted in their occurrence and in their association with CMV, or whether they span multiple HLA types. To shed light on this question we focused on a set of 68 CMV-associated TCRβ chains whose co-occurrence with CMV seropositivity was significant at a p-value threshold of 1.5×10^{-5} (corresponding to an FDR of 0.05; see Methods). For each CMV-associated TCRβ chain, we identified its most strongly associated HLA allele and compared the p-value of this association to the p-value of its association with CMV (Figure 8A). From this plot we can see that the majority of the CMV-associated chains do appear to be HLA-associated, having p-values that exceed the FDR 0.05 threshold for HLA association. The excess of highly significant HLA-association p-values for these CMV-associated TCRβs can be seen in Figure 8B, which compares the observed p-value distribution to a background distribution of HLA association p-values for randomly selected frequency-matched public TCRβs.

As a next step we looked to see whether these HLA associations fully explained the CMV association, in the sense that the CMV association was only present in subjects positive for the associated allele. For each of the 68 CMV-associated TCRs, we divided the cohort into subjects positive for its most strongly associated HLA allele and subjects negative for that allele. Here we considered both 2- and 4-digit resolution alleles when defining the most strongly associated allele, to allow for TCRs whose association extends beyond a single 4-digit allele. We computed association p-values between TCR occurrence and CMV seropositivity over these two cohort subsets independently and compared them (Figure 8C). We see that the majority of the points lie below the $y = x$ line—indicating a stronger CMV-association on the subset of the cohort positive for the associated allele—and also below the line corresponding to the expected minimum of 68 uniform random variables (i.e. the expected upper significance limit in the absence of CMV association on the allele-negative cohort subsets). There are however a few TCRβs which do not appear strongly HLA-associated and for which the CMV-association remains strong in the absence of their associated allele (the points above the line $y = x$ in Figure 8C). For example, the public TCRβ chain defined by `TRBV07` and the CDR3 sequence `CASSSDSGGTDTQYF` (which corresponds to the highest point in Figure 8C) is strongly CMV-associated ($22/23$ subjects with this chain are CMV positive; $P < 3 \times 10^{-7}$) but does not show evidence of HLA association in our dataset. TCRs with HLA promiscuity may be especially interesting from a diagnostic perspective, since their phenotype associations may be more robust to differences in genetic background.

Finally, we looked to see whether CMV assocation completely explained the observed HLA associations, in the sense that a response to one or more CMV epitopes was likely the only driver of HLA association, or whether there

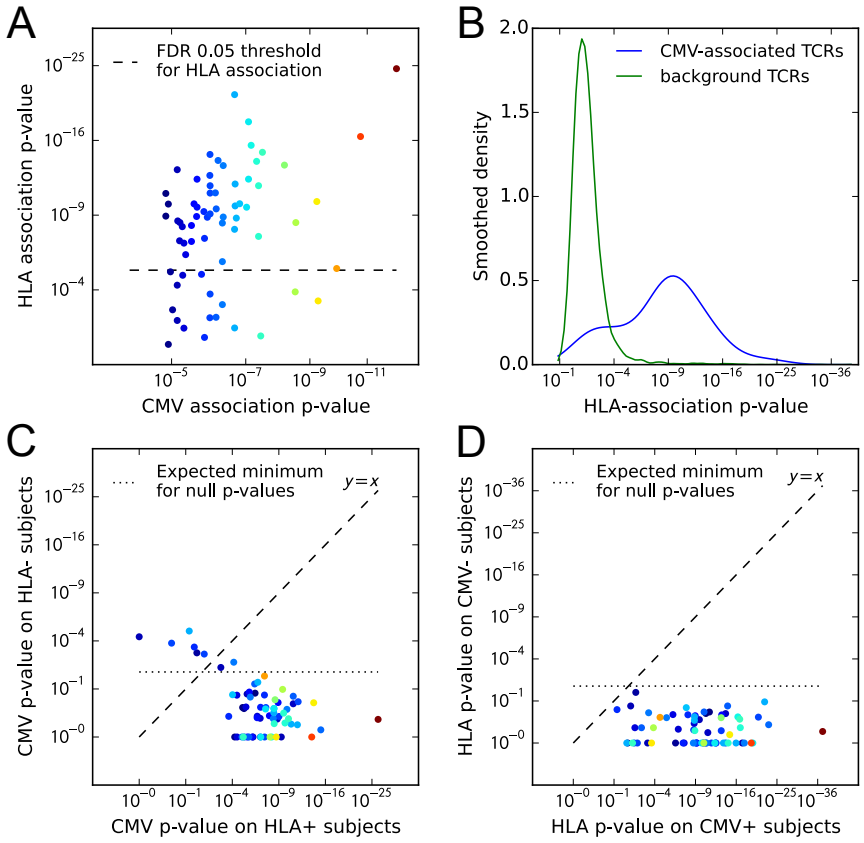


Figure 8. CMV-associated TCRβ chains are largely HLA-restricted. **(A)** Comparison of CMV-association (x-axis) and HLA-association (y-axis) p-values for 68 CMV-associated TCRβ chains shows that the majority are also HLA associated. **(B)** Smoothed densities comparing HLA-association p-value distributions for the 68 CMV-associated chains (blue) and a cohort-frequency matched set of 6800 randomly selected public TCRβ chains. CMV-associated TCRs are much more strongly HLA-associated than would be expected based solely on their cohort frequency. **(C)** CMV-association p-values computed over subsets of the cohort positive (x-axis) or negative (y-axis) for the HLA allele most strongly associated with each TCR. For most of the TCR chains, CMV association is restricted to the subset of the cohort positive for their associated HLA allele. **(D)** HLA-association p-values computed over CMV-positive (x-axis) or CMV-negative (y-axis) subsets of the cohort suggest that for these 68 CMV-associated TCRβ chains, HLA-association is driven solely by response to CMV (rather than generic affinity for their associated allele, for example, or additional self or viral epitopes). In panels (A), (C), and (D), points are colored by CMV-association p-value; in all panels we use a modified logarithmic scale based on the square root of the exponent when plotting p-values in order to avoid compression due to a few highly significant associations.

might be evidence for other epitope-specific responses by these TCRβ chains or a more general affinity for the associated allele, perhaps driven by common self antigens. Put another way, do we see evidence for pre-existing enrichment of any of these TCRβ chains when their associated allele is present, even in the absence of CMV, which might suggest that the CMV response recruits from a pre-selected pool enriched for TCRs with intrinsic affinity for the restricting allele? To approach this question we split the cohort into CMV seropositive and seronegative subjects and computed, for each of the 68 CMV-associated TCRs, the strength of its association with its preferred allele over these two subsets separately. Figure 8D compares these HLA-association p-values computed over the subsets of the cohort positive (289 individuals, x-axis) and negative (352 individuals, y-axis) for CMV. We can see in this case that all of the associations on the CMV-positive subset are stronger than those on the CMV-negative subset, and indeed the CMV-negative p-values do not appear to exceed random expectation given the number of comparisons performed. Thus, the apparent lack of any significant HLA-association on the CMV-negative cohort subset suggests that the HLA associations of these CMV-predictive chains are largely driven by CMV exposure. A limitation of this analysis is that, although the CMV-negative subset of the cohort is larger than the CMV-positive subset, the number of TCR occurrences in the CMV-negative subset is likely lower than in the CMV-positive subset for these CMV-associated chains, which will limit the strength of the HLA associations that can be detected.

3 Discussion

Each individual's repertoire of circulating immune receptors encodes information on their past and present exposures to infectious and autoimmune diseases, to antigenic stimuli in the environment, and to tumor-derived epitopes. Decoding this exposure information requires an ability to map from amino acid sequences of rearranged receptors to their eliciting antigens, either individually or collectively. One approach to developing such an antigen-mapping capability would involve collecting deep repertoire datasets and detailed phenotypic information on immune exposures for large cohorts of genetically diverse individuals. Correlation between immune exposure and receptor occurrence across such datasets could then be used to train statistical predictors of exposure, as demonstrated by Emerson et al. for CMV serostatus. The main difficulty with such an approach, beyond the cost of repertoire sequencing, is likely to be the challenge of assembling accurate and complete immune exposure information.

For this reason, we set out to discover potential signatures of immune exposures *de novo*, in the absence of phenotypic information, using only the structure of the public repertoire—its receptor sequences and their occurrence patterns. By analyzing co-occurrence between pairs of public TCRβ chains and between individual TCRβ chains and HLA alleles, we were able to identify statistically significant clusters of co-occurring TCRs across a large cohort of individuals and in a variety of HLA backgrounds. Indirect evidence from

sequence matches to experimentally-characterized receptors suggests that some of these TCR clusters may reflect hidden immune exposures shared among subsets of the cohort members; indeed, several of the most significant clusters appear linked to common viral pathogens (parvovirus B19, influenza, CMV, and EBV).

The results of this paper demonstrate the potential for a productive dialog between statistical analysis of TCR repertoires and immune exposure analysis. Specifically, sequences from the statistically-inferred clusters defined here could be tested for antigen reactivity or combined with immune exposure data to infer the driver of TCR expansion, as was done here for the handful of CMV-associated clusters based on CMV serostatus information. In either case our clustering approach will reduce the amount of independent data required, since the immune phenotype data is used for annotation of a modest number of defined TCR groupings rather than direct discovery of predictive TCRs from the entire public repertoire. We can also look for the presence of specific TCRs and TCR clusters identified here in other repertoire datasets, for example from studies of specific autoimmune diseases or pathogens, as a means of assigning putative functions. However the answer may not be entirely straightforward: it remains possible that enrichment for other cluster TCRs, rather than being associated with an exposure *per se*, is instead associated with some subject-specific genetic or epigenetic factor that determines whether a specific TCR response will be elicited by a given exposure.

The finding by Emerson et al.—now replicated and extended in this work—that there are large numbers of TCRβ chains whose occurrence patterns (independent of potential TCRα partners) are strongly associated with specific HLA alleles, raises the question of what selective forces drive these biased occurrence patterns. Our observations point to a potential role for responses to common pathogens in selecting some of these chains in an HLA-restricted manner. Self-antigens (presented in the thymus and/or the periphery) may also play a role in enriching for specific chains, as suggested by Madi et al. (2017) in their work on TCR similarity networks formed by the most frequent CDR3 sequences. Our conclusions diverge somewhat from this previous work, which may be explained by the following factors: our use of HLA-association rather than intra-individual frequency as a filter for selecting TCRs, our inclusion of information on the V-gene family in addition to the CDR3 sequence when defining TCR sharing and computing TCR similarity, and our use of TCR occurrence patterns, rather than CDR3 edit distance, to discover TCR clusters. We also find it interesting that class II loci appear on average to have greater numbers of associated TCRβ chains than class I loci (Figure 4): presumably this reflects differences in selection and/or abundance between the CD4+ and CD8+ T cell compartments, but the underlying explanation for this trend is unclear. It is also worth pointing out that our primary focus on presence/absence of TCRβ chains (rather than abundance) assumes relatively uniform sampling depths across the cohort; in the limit of very deep repertoire sequencing, pathogen-associated chains may be found (presumably in the naive pool) even in the absence of the associated immune challenge, while shallow sampling reliably picks out only

the most expanded T cell clones. Here the use of clusters of responsive TCRs rather than individual chains lessens stochastic fluctuations in TCR occurrence patterns, providing some measure of robustness.

We look forward to the accumulation of new data sets, which will enable future researchers to move beyond the limitations of the study presented here. An ideal study would perform discovery on repertoire data from multiple large cohorts, rather than the single large cohort generated with a single sequencing platform. Although we do validate TCR clusters on two independent datasets, with one from a different immune profiling technology, performing discovery on multiple large cohorts would presumably give more robust results. Future analyses of independent, HLA-typed cohorts will provide additional validation of trends seen here. We also hope that future studies will have rich immune exposure data beyond CMV serostatus: although the cohort members were all nominally healthy at the time of sampling, it is likely that there are a variety of immune exposures, some presaging future pathologies, that can be observed in a diverse collection of 650+ individuals. As an example, two of our EBV-annotated clusters contain TCRβ chains also seen in the context of rheumatoid arthritis: cross-reactivity between pathogen and autoimmune epitopes may mean that TCR clusters discovered on the basis of common infections also provide information relevant in the context of autoimmunity.

4 Materials and Methods

4.1 Datasets

TCRβ repertoire sequence data for the 666 members of the discovery cohort was downloaded from the Adaptive biotechnologies website using the link provided in the original Emerson et al. (2017) publication (`https://clients.adaptivebiotech.com/pub/Emerson-2017-NatGen`). The repertoire sequence data for the 120 individuals in the "Keck120" validation set was included in the same download. Repertoire sequence data for the 86 individuals in the "Brit86" validation set was downloaded from the NCBI SRA archive using the Bioproject accession PRJNA316572 (Britanova et al., 2016) and processed using scripts and data supplied by the authors (`https://github.com/mikessh/aging-study`) in order to demultiplex the samples and remove technical replicates. Repertoire sequence data for TCRβ chains from MAIT cells was downloaded from the NCBI SRA archive using the Bioproject accession PRJNA412739 (Howson et al., 2018).

V and J genes were assigned by comparing the TCR nucleotide sequences to the IMGT/GENE-DB (Giudicelli et al., 2005) nucleotide sequences of the human TR genes (sequence data downloaded on 9/6/2017 from `http://www.imgt.org/genedb/`). CDR3 nucleotide and amino acid sequences and most-parsimonious VDJ recombination scenarios were assigned by the TCRdist pipeline (Dash et al., 2017) (the most parsimonious recombination scenario, used for identifying non-germline CDR3 amino acids, is the one requiring the fewest non-templated

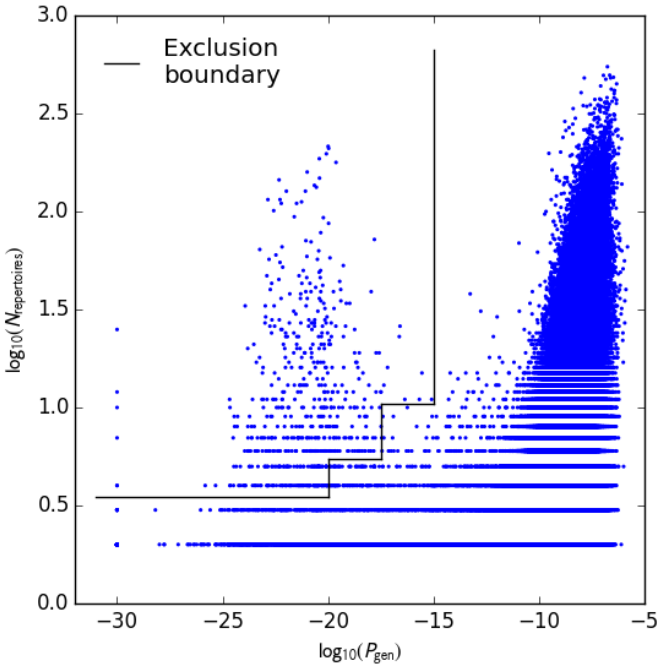


Figure 9. Analysis of TCR sharing at the nucleotide level and VDJ recombination probabilities helps to identify potential contamination. Each point represents a TCRβ nucleotide sequence that occurs in more than one repertoire, plotted according to its generation probability (P_{gen}, x-axis) and the number of repertoires in which it was seen ($N_{\mathrm{repertoires}}$, y-axis). Very low probability nucleotide sequences that are shared across many repertoires represent potential cross-contamination, as confirmed for one large cluster of artifactual sequences (see the main text). We excluded all TCRβ nucleotide sequences lying above the boundary indicated by the black line.

nucleotide insertions). To define the occurrence matrix of public TCRs and assess TCR-TCR, TCR-HLA and TCR-CMV association, a TCRβ chain was identified by its CDR3 amino acid sequence and its V-gene family (e.g., TRBV6-4*01 was reduced to TRBV06). TCR sequence reads for which a unique V-gene family could not be determined (due to equally well-matched V genes from different families, a rare occurrence in this dataset) were excluded from the analysis.

4.2 Eliminating potential cross-contamination

A preliminary analysis of TCR sharing at the nucleotide level was conducted to identify potential cross-contamination in the discovery cohort repertoires. Each TCRβ nucleotide sequence that was found in multiple repertoires was as-

signed a generation probability (P_{gen}, see below) in order to identify nucleotide sequences with suspiciously high sharing rates among repertoires. Visual comparison of the sharing rate (the number of repertoires in which each TCRβ nucleotide sequence was found) to the generation probability (Figure 9) showed that the majority of highly-shared TCRs had correspondingly high generation probabilities; it also revealed a cluster of TCR chains with unexpectedly high sharing rates. Examination of the sequences of these highly-shared TCRs revealed them to be variants of the consensus sequence `CFFKQKTAYEQYF` (coding sequence: `tgttttttcaagcagaagacggcatacgagcagtacttc`). Consultation with scientists at Adaptive Biotechnologies confirmed that these sequences were likely to represent a technical artifact. We elected to remove all TCRβ nucleotide sequences whose sharing rates put them outside the decision boundary indicated by the black line in Figure 9, which eliminated the vast majority of the artifactual variants as well as a handful of other highly shared, low-probability sequences.

4.3 Measuring clonal expansion

Each public TCRβ chain was assigned a clonal expansion index (I_{exp}) determined by its frequencies in the repertoires in which it was found. First, the unique TCRβ chains present in each repertoire were ordered based on their inferred nucleic acid template count (Carlson et al., 2013), and assigned a rank ranging from 0 (lowest template count) to $S-1$ (highest template count), where S is the total number of chains present in the repertoire. TCRs with the same template count were assigned the same tied rank equal to the midpoint of the tied group. In order to compare across repertoires, the ranks for each repertoire were then normalized by dividing by the number of unique sequences in the repertoire. The clonal expansion index for a given public TCR t was taken to be its average normalized rank for the repertoires in which it occurred:

$$I_{\mathrm{exp}}(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{r_i}{S_i - 1},$$

where the sum is taken over the m repertoires in which t is found, r_i is the template-count rank of TCR t in repertoire i, and S_i is the total size of repertoire i.

4.4 HLA typing

HLA genotyping was performed and confirmed by molecular means (either Sanger sequencing or next-generation sequencing) and independently by imputation of HLA alleles using data generated by high density single-nucleotide polymorphism arrays. HLA typing data availability varied across loci as follows: HLA-A (629 subjects), HLA-B (630 subjects), HLA-C (629 subjects), HLA-DRB1 (630 subjects), HLA-DQA1 (522 subjects), HLA-DQB1 (630 subjects), HLA-DPA1 (606 subjects), and HLA-DPB1 (472 subjects). When calculating the association p-values between TCRβ chains and HLA alleles reported in Table 1, the cohort

was restricted to the subset of subjects with available HLA typing at the relevant locus. For comparing TCR association rates across loci in Figure 4, associations were calculated over the cohort subset (522 subjects) with typing data at all compared loci (A, B, C, DRB1, DQA1, and DQB1) in order to avoid spurious differences in association strengths arising from differential data availability among the loci. Due to their very strong linkage on our cohort, five DR-DQ haplotypes were treated as single allele units for association calculations and clustering: DRB1*03:01-DQA1*05:01-DQB1*02:01, DRB1*15:01-DQA1*01:02-DQB1*06:02, DRB1*13:01-DQA1*01:03-DQB1*06:03, DRB1*10:01-DQA1*01:05-DQB1*05:01, and DRB1*09:01-DQA1*03:02-DQB1*03:03.

4.5 TCR generation probability

We implemented a version of the probabilistic model proposed by Walczak and co-workers (Murugan et al., 2012) in order to assign to each public TCRβ chain (defined by a V-gene family and a CDR3 amino acid sequence) a generation probability, P_{gen}, which captures the probability of seeing that TCRβ in the preselection repertoire. P_{gen} is calculated by summing the probabilities of the possible VDJ rearrangements that could have produced the observed TCR:

$$P_{\text{gen}}(\text{V}_{\text{family}}, \text{CDR3}_{\text{aa}}) = \sum_{s \in \mathcal{S}} P(s)$$

where $\mathcal{S}$ represents the set of possible VDJ recombination scenarios capable of producing the observed TCR V family and CDR3 amino acid sequence. To compute the probability of a given recombination scenario s, we use the factorization proposed by Marcou et al. (2018), which captures observed dependencies of V-, D-, and J-gene trimming on the identity of the trimmed gene and of inserted nucleotide identity on the identity of the preceding nucleotide:

$$\begin{aligned} P(s) =& P(V_s)P(D_s|J_s)P(J_s) \\ & \times P(\text{del}_s V|V_s)P(\text{del}_s D5', \text{del}_s D3'|D_s)P(\text{del}_s J|J_s) \\ & \times P(\text{Ins}_s VD) \prod_i^{\text{Ins}_s VD} P(n_i|n_{i-1}) \\ & \times P(\text{Ins}_s DJ) \prod_i^{\text{Ins}_s DJ} P(m_i|m_{i-1}) \end{aligned}$$

Here the recombination scenario s consists of a choice of V gene (V_s), D gene (D_s), J gene (J_s), number of nucleotides trimmed back from the end of the V gene ($\text{del}_s V$) or J gene ($\text{del}_s J$) or D gene ($\text{del}_s D5'$ and $\text{del}_s D3'$), number of nucleotides inserted between the V and D genes ($\text{Ins}_s VD$) and between the D and J genes ($\text{Ins}_s DJ$) and the identities of the inserted nucleotides ($\{n_i\}$ and $\{m_i\}$ respectively). At the start of the calculation, the CDR3 amino acid sequence is converted to a list of potential degenerate coding nucleotide sequences. Since

each amino acid other than Leucine, Serine, and Arginine has a single degenerate codon (and these three amino acids have two such codons), this list of nucleotide sequences is generally not too long. The generation probability is then taken to be the sum of the probabilities of these degenerate nucleotide sequences. Since the total number of possible recombination scenarios is in principle quite large, we make a number of approximations to speed the calculation: we limit *excess trimming* of genes to at most three nucleotides, where excess trimming is defined to be trimming back a nucleotide which matches the target CDR3 nucleotide (therefore requiring non-templated reinsertion of the same nucleotide); at most 2 palindromic nucleotides are allowed; sub-optimal D gene alignments are only considered up to a score gap of 2 matched nucleotides relative to the best match. The parameters of the probability model are fit by a simple iterative procedure in which we generate rearrangements using an initial model, compare the statistics of those rearrangements to statistics derived from observed out-of-frame rearrangements in the dataset, and adjust the probability model parameters to iteratively improve agreement.

4.6 Co-occurrence calculations

We used the hypergeometric distribution to assess the significance of an observed overlap between two subsets of the cohort, taking our significance p-value to be the probability of seeing an equal or greater overlap if the two subsets had been chosen at random:

$$P_{\text{overlap}}(k, N_1, N_2, N) = \sum_{j \geq k} \frac{\binom{N_1}{j}\binom{N-N_1}{N_2-j}}{\binom{N}{N_2}}$$

where k is the size of the overlap, N_1 and N_2 are the sizes of the two subsets, and N is the total cohort size. A complication arises when assessing TCR-TCR co-occurrence in the presence of variable-sized repertoires: TCRs are more likely to come from the larger repertoires than the smaller ones, which violates the assumptions of the hypergeometric distribution and leads to inflated significance scores. In particular, when we use the hypergeometric distribution to model the overlap between the sets of subjects in which two TCR chains are found, we implicitly assume that all subjects are equally likely to belong to a TCR chain's subject set. If the subject repertoires vary in size, this assumption will not hold. For example, in the limit of a subject with an empty repertoire, no TCR subject sets will contain that subject, which will inflate all the overlap p-values since we are effectively overstating the size N of the cohort by 1. On the other hand, if one of the subject repertoires contains all the public TCR chains, then each TCR-TCR overlap will automatically contain that subject, again inflating the p-values since we are artificially adding 1 to each of k, N_1, N_2, and N. We developed a simple heuristic to correct for this effect using a per-subject bias factor by defining

$$b_i = \frac{S_i N}{\sum_{j=1}^{N} S_j},$$

where S_i is the size of repertoire i and N is the cohort size. To score an overlap $\mathcal{O}$ of size k involving subjects $s_1, ..., s_k$, we adjust the overlap p-value by the product of the bias factors of the subjects in the overlap:

$$P_{\mathrm{CO}}(\mathcal{O}, N_1, N_2, N) = \Big(\prod_{j=1}^{k} b_{s_j} \Big) P_{\mathrm{overlap}}(k, N_1, N_2, N).$$

This has the effect of decreasing the significance assigned to overlaps involving larger repertoires, yet remains fast to evaluate, an important consideration given that the all-vs-all TCR co-occurrence calculation involves about 10^{14} pairwise comparisons (and this calculation is repeated multiple times with shuffled occurrence patterns to estimate false-discovery rates). When clustering by co-occurrence, we augmented this heuristic p-value correction by also eliminating repertoires with very low (fewer than $30{,}000$) or very high (more than $120{,}000$) numbers of public TCRβ chains (nonzero entries in the occurrence matrix M), as well as five additional repertoires which showed anomolously high levels of TCR nucleotide sharing with another repertoire—all with the goal of reducing potential sources of spurious TCR-TCR co-occurrence signal.

4.7 Estimating false-discovery rates

We used the approach of Storey and Tibshirani (2003) to estimate false-discovery rates for detecting associations between TCRs and HLA alleles and between TCRs and CMV seropositivity. Briefly, for a fixed significance threshold P we estimate the false-discovery rate (FDR) by randomly permuting the HLA allele or CMV seropositivity assignments 20 times and computing the average number of significant associations discovered at the threshold P in these shuffled datasets. The estimated FDR is then the ratio of this average shuffled association number to the number of significant associations discovered in the true dataset at the same threshold. In order to estimate a false-discovery rate for TCR-TCR co-occurrence over the full cohort, we performed 20 co-occurrence calculations on shuffled occurrence matrices, preserving the per-subject bias factors during shuffling by resampling each TCR's occurrence pattern with the bias distribution $\{b_i\}$ determined by the subject repertoire sizes.

4.8 TCR clustering

We used the DBSCAN (Ester et al., 1996) algorithm to cluster public TCRβ chains by their occurrence patterns. DBSCAN is a simple and robust clustering procedure that requires two input parameters: a similarity/distance threshold (T_{sim}) at which two points in the dataset are considered to be neighbors, and a minimum number of neighbors (N_{core}) for a point to be considered a *core*, as opposed to a *border*, point. DBSCAN clusters consist of the connected components of the neighbor-graph over the core points, together with any border point neighbors the core cluster members have. To prevent the discovery of

fictitious clusters, T_{sim} and N_{core} can be selected so that core points (points with at least N_{core} neighbors) are unlikely to occur by chance. There is a trade-off between the two parameter settings: as T_{sim} is relaxed, points will tend to have more neighbors on average and thus N_{core} should be increased, which biases toward discovery of larger clusters; conversely, more stringent settings of T_{sim} are compatible with smaller values for N_{core} which permits the discovery of smaller, more tightly linked clusters.

For clustering TCRs by co-occurrence over the full cohort, we used a threshold of $T_{\text{sim}}= 10^{-8}$ and chose a value for N_{core} (6) such that no core points were found in any of the 20 shuffled datasets. In other words, two TCRs t_1 and t_2 were considered to be neighbors for DBSCAN clustering if $P_{\text{CO}}(t_1, t_2) < 10^{-8}$; a TCR was considered a core point if it had at least 6 neighbors. Choosing parameters for HLA-restricted TCR clustering was slightly more involved due to the variable number of clustered TCRs for different alleles, and the more complex nature of the similarity metric, whose dependence on TCR sequence makes shuffling-based approaches more challenging. To begin, we transformed the TCRdist sequence-similarity measure into a significance score P_{TCRdist} which captures the probability of seeing an observed or smaller TCRdist score for two randomly selected TCRβ chains. Since public TCRβ chains are on average shorter and closer to germline than private TCRs, we derived the P_{TCRdist} CDF by performing TCRdist calculations on randomly selected public TCRs seen in at least 5 repertoires. We identified neighbors for DBSCAN clustering using a similarity score P_{sim} that combines co-occurrence and TCR sequence similarity:

$$P_{\text{sim}}(t_1, t_2) = f\big(P_{\text{TCRdist}}(t_1, t_2) \cdot P_{\text{CO}}(t_1, t_2)\big)$$

where the transformation by $f(x) = x - x\log(x)$ corrects for taking the product of two p-values because $f(x)$ is the cumulative distribution function of the product of two uniform random variables. Thus, if P_{TCRdist} and P_{CO} are independent and uniformly distributed, the same will be true of P_{sim}.

For HLA-restricted clustering using this combined similarity measure we set a fixed value of $T_{\text{sim}}= 10^{-4}$ and adjusted the N_{core} parameter as a function of the total number of TCRs clustered for each allele. As in global clustering, our goal was to choose N_{core} such that core points were unlikely to occur by chance (more precisely, had a per-allele probability less than 0.05). We estimated the probability of seeing core points by modeling neighbor number using the binomial distribution, assuming that the observed neighbor number of a given TCR during clustering is determined by $M-1$ independent Bernoulli-distributed neighborness tests with rate r, where M is the number of clustered TCRs. Rather than assuming a fixed neighbor-rate r across TCRs, we captured the observed variability in neighbor-rate (due, for example, to unequal V-gene frequencies and variable CDR3 lengths) by using a mixture of 20 rates estimated from similarity comparisons on randomly chosen public TCRs.

We also used this neighbor-number model to assign a p-value (P_{size}) to each cluster reflecting the likelihood of seeing the observed degree of clustering by chance. Since DBSCAN clusters are effectively single-linkage-style partitionings

of the core points (together with any neighboring border points), they can have a variety of shapes, ranging from densely interconnected graphs, to extended clusters held together by local neighbor relationships (Ester et al., 1996). Modeling the total size of these arbitrary groupings is challenging, so we took the simpler and more conservative approach of assigning *p*-values based on the size of the largest TCR neighborhood (set of neighbors for a single TCR) contained within each cluster. We identified the member TCR with the greatest number of neighbors in each cluster (the *cluster center*) and computed the likelihood of seeing an equal or greater neighbor-number under the mixture model described above. This significance estimate is conservative in that it neglects clustering contributions from TCRs outside the neighborhood of the cluster center, however in practice we observed that the majority of TCR clusters were dominated by a single dense region of repertoire space and therefore reasonably well-captured by a single neighborhood. To control false discovery when combining DBSCAN clusters from independent clustering runs for different HLA alleles, we used the Holm method (Holm, 1979) applied to the sorted list of cluster P_{size} values, with a target family-wise error rate (FWER) of 0.05 (i.e., we attempted to limit the overall probability of seeing a false cluster to 0.05). In the Holm FWER calculation we set the total number of hypotheses equal to the total number of TCRs clustered across all alleles minus the cumulative neighbor-count of the cluster centers (we exclude cluster center neighbors since their neighbor counts are not independent of the neighbor count of the cluster center).

4.9 Analyzing TCR clusters

For each (global or HLA-restricted) TCR cluster, we analyzed the occurrence patterns of the member TCRs in order to identify a subset of the (full or allele-positive) cohort enriched for those TCRs. We counted the number of cluster member TCRs found in each subject's repertoire and sorted the subjects by this TCR count (rank plots in Figure 2B-C and in the right panels of Figure 6). For comparison, we generated control TCR count plots by independently resampling the subjects for each member TCR, preserving the frequency of each TCR and biasing by subject repertoire size. Each complete resampling of the cluster member TCR occurrence patterns produced a subject TCR rank plot; we repeated this resampling process 1000 times and averaged the rank plots to yield the green ('randomized') curves in Figure 2B-C and Figure 6. To compare the observed and randomized curves, we took a signed difference

$$D_{\text{CO}} = \max_{1 \le i \le N} \left(\sum_{j \le i} (C_j - R_j) + \sum_{j > i} (R_j - C_j) \right)$$

between the observed counts C_j and the randomized counts R_j, where the value of the subject index $i = i_{\max}$ that maximizes the right-hand side in the equation above represents a switchpoint below which the observed counts generally exceed the randomized counts and above which the reverse is true (both sets of counts are sorted in decreasing order). We take this switchpoint

$i_{\max}$ as an estimate of the number of enriched subjects for the given cluster (this is the value given in the 'Subjects' column in Table 3).

Since the raw D_{CO} values are not comparable between clusters of different sizes and for different alleles, we transformed these values to a Z-score (Z_{CO}) by generating, for each cluster, 1000 additional random TCR count curves and computing the mean (μ_D) and standard deviation (σ_D) of their $D_{\mathrm{CO}}^{\mathrm{rand}}$ score distribution:

$$Z_{\mathrm{CO}} = \frac{D_{\mathrm{CO}} - \mu_D}{\sigma_D}$$

We used this co-occurrence score Z_{CO} together with a log-transformed version of the cluster size p-value,

$$S_{\mathrm{size}} = \sqrt{-\log_{10}(P_{\mathrm{size}})}$$

for visualizing clustering results in Figure 5 (S_{size} on the x-axis and Z_{CO} on the y-axis) and prioritizing individual clusters for detailed follow-up.

4.10 TCR annotations

We annotated public TCRs in our dataset by matching their sequences against two publicly available datasets: VDJdb (Shugay et al., 2017), a curated database of TCR sequences with known antigen specificities (downloaded on 3/29/18; about $17,000$ human TCRβ entries) and McPAS-TCR (Tickotsky et al., 2017), a curated database of pathogen-associated TCR sequences (downloaded on 3/29/18; about $9,000$ human TCRβ entries). VDJdb entries are associated with a specific MHC-presented epitope, whereas McPAS-TCR also includes sequences of TCRs isolated from diseased tissues whose epitope specificity is not defined. We added to this merged annotation database the sequences of structurally characterized TCRs of known specificity (see below), as well as literature-derived TCRs from a handful of primary studies (Dash et al., 2017; Glanville et al., 2017; Song et al., 2017; Kasprowicz et al., 2006). For matches between HLA-associated TCRs and database TCRs of known specificity, we filtered for agreement (at 2-digit resolution) between the associated HLA allele in our dataset and the presenting allele from the database. In other words, TCRs belonging to B*08:01-restricted clusters were not annotated with matches to database TCRs that bind to A*02:01-presented peptides.

4.11 Structural analysis

We analyzed a set of experimentally determined TCR:peptide-MHC structures to find MHC positions frequently contacted by the CDR3β loop. Crystal structures of complexes involving human TCRs and human class I or class II HLA alleles were identified using BLAST (Altschul et al., 1997) searches against the RCSB PDB (Berman et al., 2000) sequence database (`ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt`). Structural coverage of HLA loci and alleles is sparse and highly biased toward well studied alleles such as HLA-A*02.

Given the high degree of structural similarity among class I and among class II MHC structures solved to date, we elected to share contact information across loci using trans-locus sequence alignments. For class I we used the merged alignment (`ClassI_prot.txt`) available from the IPD-IMGT/HLA (Robinson et al., 2014) database. Starting with multiple sequence alignments for individual class II loci from the IPD-IMGT/HLA database, we inserted gaps as needed in order to created merged alignments for the class II α and β chains. These alignments provided a common reference frame in which to combine residue-residue contacts from the TCR:peptide-MHC structures. We considered two amino acid residues to be in contact if they had a side chain heavyatom contact distance less than or equal to 4.5Å. The CDR3β contact frequency for an alignment position (class I, class II-α, or class II-β) was defined to be the total number of contacted CDR3β amino acids observed for that position, divided by the total number of structures analyzed. Redundancy in the structural database was assessed at the level of TCR and HLA sequence, ignoring the sequence of the peptide. Contacts from a set of n structures all containing the same TCR and HLA were given a weight of $1/n$ when computing the residue contact frequencies.

5 Acknowledgments

This work was supported in part through the NIH/NCI Cancer Center Support Grant P30 CA015704 and by NIH NHLBI grant R01-HL105914 to JH, as well as R01 GM113246 and U19 AI117891. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation. We gratefully acknowledge superlative computing support from Fred Hutch scientific computing and thank Paul Thomas and Jeremy Crawford for helpful comments on a preliminary version of this manuscript.

References

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL `http://dx.doi.org/10.1093/nar/28.1.235`.

Sydney J Blevins, Brian G Pierce, Nishant K Singh, Timothy P Riley, Yuan Wang, Timothy T Spear, Michael I Nishimura, Zhiping Weng, and Brian M Baker. How structural adaptability exists alongside HLA-A2 bias in the human $\alpha\beta$ TCR repertoire. *Proc. Natl. Acad. Sci. U. S. A.*, 113(9):E1276–85, March 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1522069113. URL `http://dx.doi.org/10.1073/pnas.1522069113`.

Olga V Britanova, Mikhail Shugay, Ekaterina M Merzlyak, Dmitriy B Staroverov, Ekaterina V Putintseva, Maria A Turchaninova, Ilgar Z Mamedov, Mikhail V Pogorelyy, Dmitriy A Bolotin, Mark Izraelson, Alexey N Davydov, Evgeny S Egorov, Sofya A Kasatskaya, Denis V Rebrikov, Sergey Lukyanov, and Dmitriy M Chudakov. Dynamics of individual T cell repertoires: From cord blood to centenarians. *J. Immunol.*, 196(12): 5005–5013, June 2016. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1600005. URL http://dx.doi.org/10.4049/jimmunol.1600005.

Christopher S Carlson, Ryan O Emerson, Anna M Sherwood, Cindy Desmarais, Moon-Wook Chung, Joseph M Parsons, Michelle S Steen, Marissa A LaMadrid-Herrmannsfeldt, David W Williamson, Robert J Livingston, David Wu, Brent L Wood, Mark J Rieder, and Harlan Robins. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.*, 4:2680, 2013. ISSN 2041-1723. doi: 10.1038/ncomms3680. URL http://dx.doi.org/10.1038/ncomms3680.

Nathaniel D Chu, Haixin Sarah Bi, Ryan O Emerson, Anna M Sherwood, Michael E Birnbaum, Harlan S Robins, and Eric J Alm. Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in public receptors. February 2018. URL https://www.biorxiv.org/content/early/2018/02/08/262667.

Mattia Cinelli, Yuxin Sun, Katharine Best, James M Heather, Shlomit Reich-Zeliger, Eric Shifrut, Nir Friedman, John Shawe-Taylor, and Benny Chain. Feature selection using a one dimensional naïve bayes classifier increases the accuracy of support vector machine classification of cdr3 repertoires. *Bioinformatics*, 33(7):951–955, 2017.

Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi H O Nguyen, Katherine Kedzierska, Nicole L La Gruta, Philip Bradley, and Paul G Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, June 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature22383. URL http://dx.doi.org/10.1038/nature22383.

Lu Deng, Ries J Langley, Patrick H Brown, Gang Xu, Leslie Teng, Qian Wang, Monica I Gonzales, Glenda G Callender, Michael I Nishimura, Suzanne L Topalian, et al. Structural basis for the recognition of mutant self by a tumor-specific, mhc class ii–restricted t cell receptor. *Nature immunology*, 8(4):398, 2007.

Lu Deng, Ries J Langley, Qian Wang, Suzanne L Topalian, and Roy A Mariuzza. Structural insights into the editing of germ-line–encoded interactions between t-cell receptor and mhc class ii by vα cdr3. *Proceedings of the National Academy of Sciences*, 109(37): 14960–14965, 2012.

Yuval Elhanati, Zachary Sethna, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. March 2018. URL https://www.biorxiv.org/content/early/2018/03/02/275602.

Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder, and Harlan S Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, April 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3822. URL http://dx.doi.org/10.1038/ng.3822.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231, Portland, Oregon, 1996. AAAI Press. URL `http://dl.acm.org/citation.cfm?id=3001460.3001507`.

Simon Friedensohn, Tarik A Khan, and Sai T Reddy. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends in biotechnology*, 35(3):203–214, 2017.

Veronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. Imgt/gene-db: a comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic acids research*, 33(suppl_1):D256–D261, 2005.

Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M Krams, Christina Pettus, Nikhil Haas, Cecilia S Lindestam Arlehamn, Alessandro Sette, Scott D Boyd, Thomas J Scriba, Olivia M Martinez, and Mark M Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, June 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature22976. URL `http://dx.doi.org/10.1038/nature22976`.

Maria Harkiolaki, Samantha L Holmes, Pia Svendsen, Jon W Gregersen, Lise T Jensen, Roisin McMahon, Manuel A Friese, Gijs Van Boxel, Ruth Etzensperger, John S Tzartos, et al. T cell-mediated autoimmune disease due to low-affinity crossreactivity to common microbial peptides. *Immunity*, 30(3):348–357, 2009.

Erik D Heegaard and Kevin E Brown. Human parvovirus b19. *Clinical microbiology reviews*, 15(3):485–505, 2002.

Jens Hennecke and Don C Wiley. Structure of a complex of the human α/β t cell receptor (tcr) ha1. 7, influenza hemagglutinin peptide, and major histocompatibility complex class ii molecule, hla-dr4 (dra0101 and drb10401): insight into tcr cross-restriction and alloreactivity. *Journal of Experimental Medicine*, 195(5):571–581, 2002.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Bryan Howie, Anna M Sherwood, Ashley D Berkebile, Jan Berka, Ryan O Emerson, David W Williamson, Ilan Kirsch, Marissa Vignali, Mark J Rieder, Christopher S Carlson, and Harlan S Robins. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.*, 7(301):301ra131, August 2015.

Lauren J Howson, Giorgio Napolitani, Dawn Shepherd, Hemza Ghadbane, Prathiba Kurupati, Lorena Preciado-Llanes, Margarida Rei, Hazel C Dobinson, Malick M Gibani, Karen Wei Weng Teng, et al. Mait cell clonal expansion and tcr repertoire shaping in human volunteers challenged with salmonella paratyphi a. *Nature communications*, 9 (1):253, 2018.

V Kasprowicz, A Isa, K Jeffery, K Broliden, T Tolfvenstam, P Klenerman, and P Bowness. A highly restricted t-cell receptor dominates the CD8+ t-cell response to parvovirus B19 infection in HLA-A*2402-positive individuals. *J. Virol.*, 80(13):6697–6701, July 2006.

Ilan Kirsch, Marissa Vignali, and Harlan Robins. T-cell receptor profiling in cancer. *Molecular oncology*, 9(10):2063–2070, 2015.

Lars Kjer-Nielsen, Onisha Patel, Alexandra J Corbett, Jérôme Le Nours, Bronwyn Meehan, Ligong Liu, Mugdha Bhati, Zhenjun Chen, Lyudmila Kostenko, Rangsima Reantragoon, et al. Mr1 presents microbial vitamin b metabolites to mait cells. *Nature*, 491 (7426):717, 2012.

Hanjie Li, Congting Ye, Guoli Ji, and Jiahuai Han. Determinants of public T cell responses. *Cell Res.*, 22(1):33–42, January 2012.

Asaf Madi, Asaf Poran, Eric Shifrut, Shlomit Reich-Zeliger, Erez Greenstein, Irena Zaretsky, Tomer Arnon, Francois Van Laethem, Alfred Singer, Jinghua Lu, Peter D Sun, Irun R Cohen, and Nir Friedman. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife*, 6, July 2017. ISSN 2050-084X. doi: 10.7554/eLife.22057. URL `http://dx.doi.org/10.7554/eLife.22057`.

Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with igor. *Nature communications*, 9(1):561, 2018.

Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.

Mikhail V Pogorelyy, Yuval Elhanati, Quentin Marcou, Anastasiia L Sycheva, Ekaterina A Komech, Vadim I Nazarov, Olga V Britanova, Dmitriy M Chudakov, Ilgar Z Mamedov, Yury B Lebedev, Thierry Mora, and Aleksandra M Walczak. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput. Biol.*, 13(7):e1005572, July 2017. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1005572. URL `http://dx.doi.org/10.1371/journal.pcbi.1005572`.

Mikhail V Pogorelyy, Anastasia A Minervina, Dmitriy M Chudakov, Ilgar Z Mamedov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Method for identification of condition-associated public antigen receptor sequences. *Elife*, 7, March 2018. ISSN 2050-084X. doi: 10.7554/eLife.33050. URL `http://dx.doi.org/10.7554/eLife.33050`.

James Robinson, Jason A Halliwell, James D Hayhurst, Paul Flicek, Peter Parham, and Steven GE Marsh. The ipd and imgt/hla database: allele variant databases. *Nucleic acids research*, 43(D1):D423–D431, 2014.

Eilon Sharon, Leah V Sibener, Alexis Battle, Hunter B Fraser, K Christopher Garcia, and Jonathan K Pritchard. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.*, August 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3625. URL `http://dx.doi.org/10.1038/ng.3625`.

Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, Alexey V Eliseev, Ewald Van Dyk, Pradyot Dash, Meriem Attaf, Cristina Rius, Kristin Ladell, James E McLaren, Katherine K Matthews, E Bridie Clemens, Daniel C Douek, Fabio Luciani, Debbie van Baarle, Katherine Kedzierska, Can Kesmir, Paul G Thomas, David A Price, Andrew K Sewell, and Dmitriy M Chudakov. VDJdb: a curated database of t-cell receptor sequences with known

antigen specificity. *Nucleic Acids Res.*, September 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx760. URL https://academic.oup.com/nar/article/doi/10.1093/nar/gkx760/4101254/VDJdb-a-curated-database-of-T-cell-receptor.

Inyoung Song, Anna Gil, Rabinarayan Mishra, Dario Ghersi, Liisa K Selin, and Lawrence J Stern. Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8+ T cell epitope. *Nat. Struct. Mol. Biol.*, 24(4):395–406, April 2017.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. Mcpas-tcr: a manually curated catalogue of pathology-associated t cell receptor sequences. *Bioinformatics*, 33(18):2924–2929, 2017.

Vanessa Venturi, David A Price, Daniel C Douek, and Miles P Davenport. The molecular basis for public t-cell responses? *Nat. Rev. Immunol.*, 8(3):231–238, March 2008.

Vanessa Venturi, Brian D Rudd, and Miles P Davenport. Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr. Opin. Immunol.*, 25(5):639–645, October 2013.

Yiyuan Yin, Yili Li, Melissa C Kerzic, Roland Martin, and Roy A Mariuzza. Structure of a tcr with high affinity for self-antigen reveals basis for escape from negative selection. *The EMBO Journal*, 30(6):1137–1148, 2011.

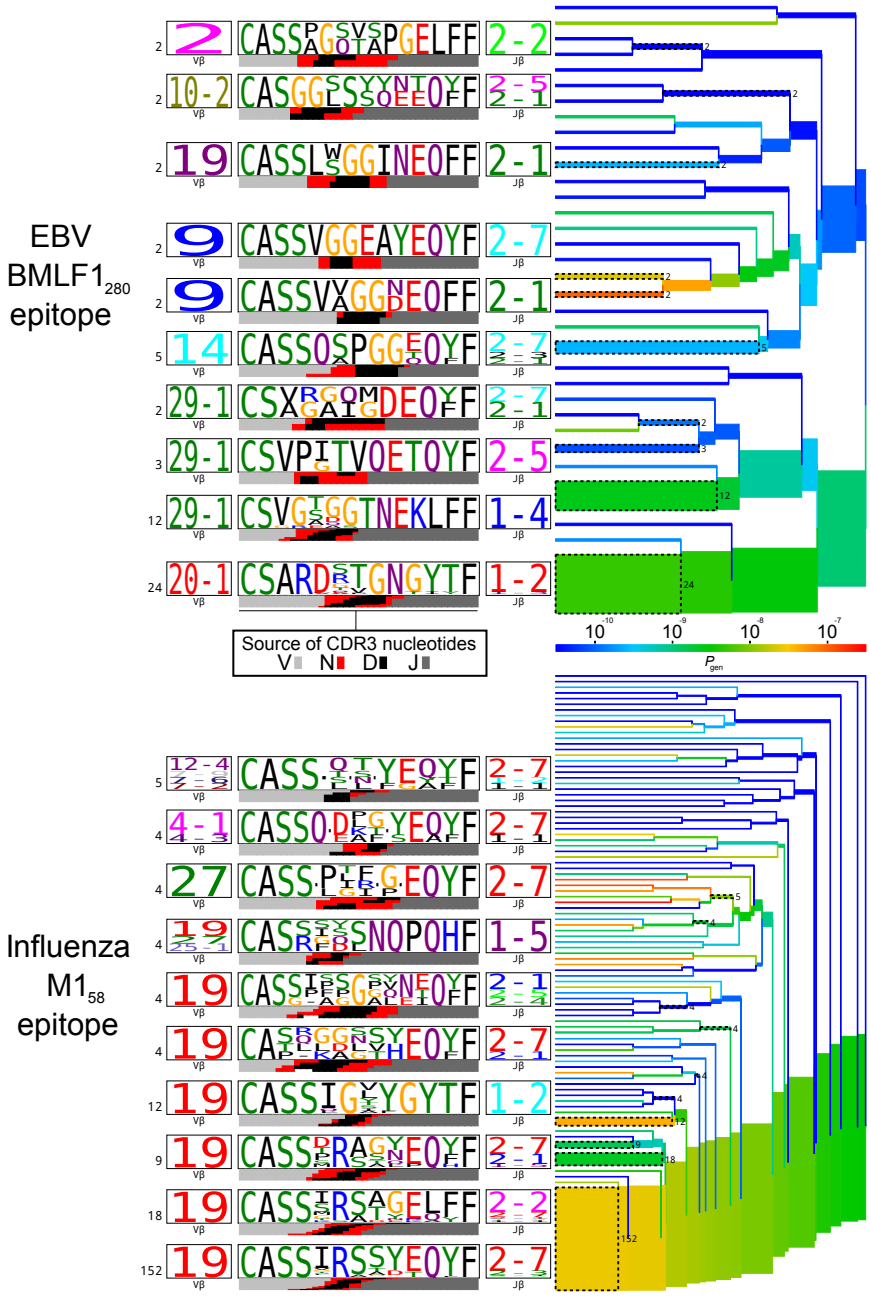


Figure 10. TCRdist trees of experimentally determined pathogen-responsive TCRβ chains for two immunodominant epitopes, EBV $\mathrm{BMLF1}_{280}$ and influenza $\mathrm{M1}_{58}$. TCR beta chain sequences were taken from the dataset of Dash et al. (2017). On the right-hand side are average-linkage dendrograms of TCRdist receptor clusters colored by generation probability (P_{gen}). TCR logos for selected receptor subsets (the branches of the tree enclosed in dashed boxes labelled with size of the TCR clusters) are shown on the left. Each logo depicts the V- (left side) and J- (right side) gene frequencies, CDR3 amino acid sequences (middle), and inferred rearrangement structure (bottom bars coloured by source region, light grey for the V-region, dark grey for J, black for D, and red for N-insertions) of the grouped receptors.

Table 3. HLA-restricted TCR clusters with size (S_{size}) and co-occurrence (Z_{CO}) scores, annotations (abbreviated as in Figure 5), and validation scores.

Rank	HLA allele	Allele frequency	TCRs	Subjects	Cluster center	S_{size}	Z_{CO}	Annotations	$Z_{\text{CO}}^{\text{Keck120}}$	$Z_{\text{CO}}^{\text{Brit86}}$
1	A*24:02	102	32	29	TRBV05,CASSGSGGYNEQFF	8.95	17.64	B19	10.38	6.74
2	A*02:01	218	43	66	TRBV19,CASSGRSTDTQYF	6.47	13.01	INF, T1D	12.28	4.28
3	DRB1*07:01	119	17	36	TRBV09,CASSGQGAYEQYF	4.08	12.91	coCMV	9.46	6.40
4	DRB1*15:01-DQ	112	16	27	TRBV19,CASSPDRSSYNEQFF	4.25	12.13		1.65	1.72
5	B*08:01	115	30	34	TRBV07,CASSQGPAYEQYF	5.97	8.12	EBV, RA	3.83	1.83
6	C*04:01	104	7	24	TRBV19,CASSPGGDYNEQFF	3.94	11.58		4.48	2.01
7	C*04:01	104	11	20	TRBV04,CASSHSGTGETYEQYF	4.91	9.03		7.52	1.66
8	B*15:01	55	23	27	TRBV19,CASSTTSGSYNEQFF	5.43	7.51		10.31	4.01
9	DRB1*03:01-DQ	108	26	39	TRBV29,CSVAPGWGMNTEAFF	4.49	8.61		10.96	7.09
10	A*01:01	154	8	44	TRBV24,CATSDGDTQYF	3.47	10.21	CMV, coCMV	3.80	2.42
11	B*35:01	56	18	24	TRBV10,CATGTGDSNQPQHF	4.98	6.13	EBV, RA	4.50	5.42
12	DRB1*03:01-DQ	108	11	35	TRBV07,CASSLSLAGSYNEQFF	3.09	8.15		5.35	1.40
13	A*02:01	218	10	84	TRBV20,CSARDRTGNGYTF	3.81	6.66	EBV	7.14	3.50
14	DRB1*15:01-DQ	112	15	38	TRBV05,CASSLRGVRTDTQYF	3.05	8.08		8.73	3.31
15	A*01:01	154	6	30	TRBV10,CAISESRASGDYNEQFF	3.14	7.67		11.31	2.99
16	DRB1*13:01-DQ	43	7	7	TRBV20,CSASAGESNQPQHF	3.14	7.64		-0.55	-0.35
17	DRB1*03:01-DQ	108	16	32	TRBV20,CSARGGGRSYEQYF	3.31	6.95		2.57	3.09
18	DRB1*11:01	58	14	20	TRBV06,CASSYSVRGRYSNQPQHF	3.26	7.02		8.72	3.44
19	C*08:02	37	6	15	TRBV28,CASSLGIHYEQYF	3.53	6.37		1.82	4.37
20	DRB1*15:01-DQ	112	13	51	TRBV12,CASSLAGTEKLFF	3.27	6.64		4.61	3.01
21	DRB1*03:01-DQ	108	11	23	TRBV05,CASSSTGLRSYEQYF	3.09	6.92		4.73	5.81
22	A*02:01	218	7	64	TRBV04,CASSQGTGRYEQYF	3.51	6.07		2.79	3.23
23	C*03:04	72	5	13	TRBV09,CASSVAYRGNEQFF	3.39	6.14		6.26	3.23
24	DQB1*03:01-DQA1*05:05	84	10	39	TRBV09,CASSVGTVQETQYF	2.97	6.73		3.02	3.54
25	DRB1*04:01	78	25	35	TRBV05,CASSRQGAGETQYF	3.00	6.31		5.82	1.55
26	B*08:01	115	7	30	TRBV12,CASSFEGLHGYTF	2.67	6.67		3.77	2.95
27	C*04:01	104	6	25	TRBV06,CASRTGLAGTDTQYF	3.58	4.78		3.53	3.76
28	DRB1*07:01	119	9	42	TRBV14,CASSLAGMNTEAFF	3.15	5.54		6.99	5.58
29	DQB1*03:01-DQA1*05:05	84	7	36	TRBV02,CASSELENTEAFF	2.97	5.76		5.25	3.24
30	DPB1*03:01-DPA1*01:03	42	7	16	TRBV30,CAWSADSNQPQHF	3.56	4.16		2.42	1.73
31	B*15:01	55	18	27	TRBV29,CSVETRDYEQYF	3.54	3.94		13.81	4.29
32	A*01:01	154	4	26	TRBV09,CASSVGVDSTDTQYF	2.39	6.24		-0.31	2.17
33	C*07:02	142	4	14	TRBV25,CASSPGDEQYF	2.94	5.11	coCMV	6.37	3.69
34	B*08:01	115	6	38	TRBV29,CSVGSGDYEQYF	3.01	4.85	EBV	2.73	0.75
35	A*01:01	154	6	37	TRBV20,CSAPGQGAVEQYF	2.79	5.24		2.42	3.00
36	A*23:01	22	5	7	TRBV06,CASSDGNSGNTIYF	3.38	4.02		1.91	4.11
37	DQB1*03:01-DQA1*05:05	84	7	29	TRBV15,CATSRDPGGNQPQHF	2.97	4.82		5.00	2.67
38	DPB1*04:01-DPA1*01:03	274	5	65	TRBV19,CASSIKGDTEAFF	3.31	4.14		4.89	3.42
39	DPB1*04:01-DPA1*01:03	274	4	55	TRBV19,CASRLSGDTQYF	2.84	4.95	COLO	3.80	1.25
40	B*07:02	125	7	37	TRBV02,CASRGETQYF	2.73	4.88		3.20	2.11
41	B*44:03	41	9	20	TRBV19,CASSATGGIYEQYF	3.35	3.41	MS	6.61	8.76
42	A*24:02	102	6	31	TRBV30,CAWSPGTGDYEQYF	3.05	3.91		3.56	2.99
43	DRB1*07:01	119	13	31	TRBV18,CASSPSVRNTEAFF	2.89	4.20		5.32	0.96
44	B*57:01	27	5	14	TRBV12,CASSPPEGETQYF	3.22	3.47		6.31	1.94
45	C*06:02	74	4	14	TRBV02,CASSAGTASTDTQYF	2.81	4.27	coCMV	4.76	3.06
46	A*11:01	47	5	7	TRBV09,CASSPKGVGYEQYF	2.75	4.31		2.43	3.32
47	DRB1*01:01	82	9	21	TRBV19,CASSIPGLAYEQYF	2.58	4.63		0.96	-0.49
48	B*07:02	125	7	21	TRBV09,CASSDRRGYTF	2.73	4.34		4.57	0.45
49	B*08:01	115	6	22	TRBV07,CASSSTGAGNQPQHF	2.67	4.24	EBV	1.00	2.85
50	B*18:01	46	5	6	TRBV27,CASSPTSEDTQYF	2.57	4.26		5.79	-0.23
51	B*27:05	36	7	13	TRBV06,CASSLRLAGLYEQYF	2.64	3.81		9.25	1.08
52	B*35:01	56	4	7	TRBV07,CASSQGPGRTYEQYF	2.46	4.10		-	-
53	B*35:03	16	4	7	TRBV10,CAISVGNEQFF	2.78	3.42		1.50	0.73
54	A*02:01	218	5	126	TRBV29,CSVGTGGTNEKLFF	2.82	3.32	EBV, MELA	5.65	2.37
55	DRB1*03:01-DQ	108	6	18	TRBV02,CASSAGAGTEAFF	2.36	4.17		0.98	2.79
56	B*44:02	79	4	18	TRBV02,CASSADSSYNEQFF	2.57	3.65		2.09	2.12
57	C*03:04	72	3	8	TRBV27,CASSPRPYNEQFF	2.35	4.08		1.36	3.22
58	A*24:02	102	4	12	TRBV20,CSAREDGHEQYF	2.62	3.54		0.83	2.94
59	A*01:01	154	12	65	TRBV19,CASSIRDHNQPQHF	2.79	3.17		8.44	2.33
60	B*27:05	36	4	12	TRBV07,CASSPPGGSAYNEQFF	2.64	3.23		1.13	2.12
61	C*14:02	23	4	9	TRBV02,CASSGDTSTNEKLFF	2.48	3.50		6.23	-
62	B*27:05	36	9	12	TRBV27,CASSSGTSGNNEQFF	2.64	3.16		4.32	3.24
63	C*12:03	53	6	25	TRBV15,CATSRENEKLFF	2.90	2.51		1.88	3.08
64	A*68:01	29	4	16	TRBV05,CASSLIATNEKLFF	2.71	2.88		3.67	1.23
65	B*51:01	53	6	20	TRBV04,CASSQDYPGGSYEQYF	2.76	2.73		6.43	5.18
66	B*35:01	56	4	8	TRBV27,CASSLGAATGELFF	2.46	3.32		4.52	3.01
67	B*15:01	55	4	20	TRBV06,CASSAGTGRYEQYF	2.44	3.18		2.40	2.23
68	B*44:03	41	7	14	TRBV07,CASSSGESGANVLTF	2.97	2.01		3.92	4.81
69	DRB1*04:02	14	4	6	TRBV03,CASSQASGGANEQFF	2.44	3.04		2.04	2.22
70	B*15:01	55	4	10	TRBV19,CASSHRGGNEQFF	2.44	3.03		0.92	3.58
71	B*15:01	55	5	7	TRBV05,CASSLGVSAGELFF	2.44	2.98		-0.32	-0.12
72	A*32:01	34	3	5	TRBV12,CASSYGPGNQPQHF	2.45	2.84		5.76	3.18
73	A*02:01	218	4	23	TRBV19,CASSTGTATNEKLFF	2.42	2.89		0.84	-
74	DRB1*15:01-DQ	112	7	51	TRBV28,CASSLLGGQPQHF	2.58	2.35		0.66	1.89
75	B*18:01	46	5	15	TRBV27,CASSFPGKEQYF	2.57	2.22		-0.35	5.62
76	B*49:01	16	3	8	TRBV29,CSVERGYNEQFF	2.38	2.14		1.03	0.43
77	A*23:01	22	3	6	TRBV20,CSARDREGAGYGYTF	2.35	2.14		-0.16	-0.12
78	B*55:01	13	3	10	TRBV19,CASRGGNQPQHF	2.36	2.09		0.95	-0.28

Table 4. PDB structures analyzed.

PDB ID[a]	HLA allele	Vα	Jα	CDR3α	Vβ	Jβ	CDR3β	Peptide
5bs0	A*01	TRAV21*01	TRAJ28*01	CAVRPGGAGPFFVVF	TRBV5-1*01	TRBJ2-7*01	CASSFNMATGQYF	ESDPIVAQY
3qdj	A*02	TRAV12-2*01	TRAJ23*01	CAVNFGGGKLIF	TRBV6-4*01	TRBJ1-1*01	CASSLSFGTEAFF	AAGIGILTV
4l3e	A*02	TRAV12-2*01	TRAJ23*01	CAVNFGGGKLIF	TRBV6-4*01	TRBJ1-1*01	CASSWSFGTEAFF	ELAGIGILTV
5e9d	A*02	TRAV12-2*01	TRAJ24*02	CAVTKYSWGKLQF	TRBV6-5*01	TRBJ2-7*01	CASRPGWMAGGVELYF	ELAGIGILTV
3qfj	A*02	TRAV12-2*01	TRAJ24*02	CAVTTDSWGKLQF	TRBV6-5*01	TRBJ2-7*01	CASRPGLAGGRPEQYF	LLFGFPVYV
4ftv	A*02	TRAV12-2*01	TRAJ24*02	CAVTTDSWGKLQF	TRBV6-5*01	TRBJ2-7*01	CASRPGLMSAQPEQYF	LLFGYPVYV
3hg1	A*02	TRAV12-2*01	TRAJ27*01	CAVNVAGKSTF	TRBV30*01	TRBJ2-2*01	CAWSETGLGTGELFF	ELAGIGILTV
4eup	A*02	TRAV12-2*01	TRAJ45*01	CAVSGGGADGLTF	TRBV28*01	TRBJ2-1*01	CASSFLGTGVEQYF	ALGIGILTV
5c0c	A*02	TRAV12-3*01	TRAJ12*01	CAMRGDSSYKLIF	TRBV12-4*01	TRBJ2-4*01	CASSLWEKLAKNIQYF	RQFGPDWIVA
5eu6	A*02	TRAV21*01	TRAJ53*01	CAVLSSGGSNYKLTF	TRBV7-3*01	TRBJ2-3*01	CASSFIGGTDTQYF	YLEPGPVTV
2p5e	A*02	TRAV21*01	TRAJ6*01	CAVRPLLDGTYIPTF	TRBV6-5*01	TRBJ2-2*01	CASSYLGNTGELFF	SLLMWITQC
2bnq	A*02	TRAV21*01	TRAJ6*01	CAVRPTSGGSYIPTF	TRBV6-5*01	TRBJ2-2*01	CASSYVGNTGELFF	SLLMWITQV
4mnq	A*02	TRAV22*01	TRAJ40*01	CAVDSATALPYGYIF	TRBV6-5*01	TRBJ1-1*01	CASSYQGTEAFF	ILAKFLHWL
5men	A*02	TRAV22*01	TRAJ40*01	CAVDSATSGTYKYIF	TRBV6-5*01	TRBJ1-1*01	CASSYQGTEAFF	ILAKFLHWL
5isz	A*02	TRAV24*01	TRAJ27*01	CAFDTNAGKSTF	TRBV19*01	TRBJ2-7*01	CASSIFGQREQYF	GILGFVFTL
5d2l	A*02	TRAV24*01	TRAJ49*01	CAFITGNQFYF	TRBV7-2*02	TRBJ2-5*01	CASSQTQLWETQYF	NLVPMVATV
3gsn	A*02	TRAV24*01	TRAJ49*01	CARNTGNQFYF	TRBV6-5*01	TRBJ1-2*01	CASSPVTGGIYGYTF	NLVPMVATV
5d2n	A*02	TRAV26-2*01	TRAJ43*01	CILDNNNDMRF	TRBV7-6*01	TRBJ1-4*01	CASSLAPGTTNEKLFF	NLVPMVATV
5euo	A*02	TRAV27*01	TRAJ37*02	CAGAIGPSNTGKLIF	TRBV19*01	TRBJ2-7*01	CASSIRSSYEQYF	GILGFVFTL
5hho	A*02	TRAV27*01	TRAJ42*01	CAGAGSQGNLIF	TRBV19*01	TRBJ2-7*01	CASSIRSSYEQYF	GILEFVFTL
2vlr	A*02	TRAV27*01	TRAJ42*01	CAGAGSQGNLIF	TRBV19*01	TRBJ2-7*01	CASSSRASYEQYF	GILGFVFTL
1oga	A*02	TRAV27*01	TRAJ42*01	CAGAGSQGNLIF	TRBV19*01	TRBJ2-7*01	CASSSRSSYEQYF	GILGFVFTL
1bd2	A*02	TRAV29/DV5*01	TRAJ54*01	CAAMEGAQKLVF	TRBV6-5*01	TRBJ2-7*01	CASSYPGGGFYEQYF	LLFGYPVYV
5e6i	A*02	TRAV35*01	TRAJ37*02	CAGPGGSSNTGKLIF	TRBV27*01	TRBJ2-2*01	CASSLIYPGELFF	GILGFVFTL
3qeq	A*02	TRAV35*01	TRAJ49*01	CAGGTGNQFYF	TRBV10-3*01	TRBJ1-5*01	CAISEVGVGQPQHF	AAGIGILTV
4zez	A*02	TRAV38-2/DV8*01	TRAJ30*01	CAYGEDDKIIF	TRBV25-1*01	TRBJ2-7*01	CASRRGPYEQYF	KLVALVINAV
5jhd	A*02	TRAV38-2/DV8*01	TRAJ52*01	CAWGVNAGGTSYGKLTF	TRBV19*01	TRBJ1-2*01	CASSIGVYGYTF	GILGFVFTL
3o4l	A*02	TRAV5*01	TRAJ31*01	CAEDNNARLMF	TRBV20-1*01	TRBJ1-2*01	CSARDGTGNGYTF	GLCTLVAML
3vxs	A*24	TRAV21*01	TRAJ12*01	CAVRMDSSYKLIF	TRBV7-9*01	TRBJ2-2*01	CASSSWDTGELFF	RYPLTLGWCF
3vxm	A*24	TRAV8-3*01	TRAJ28*01	CAVGAPSGAGSYQLTF	TRBV4-1*01	TRBJ2-7*01	CASSPTSGIYEQYF	RFPLTFGWCF
3sjv	B*08	TRAV12-1*01	TRAJ23*01	CVVRAGKLIF	TRBV6-2*01	TRBJ2-4*01	CASGQGNFDIQYF	FLRGRAYGL
3ffc	B*08	TRAV14/DV4*01	TRAJ49*01	CAMREDTGNQFYF	TRBV11-2*01	TRBJ2-3*01	CASSFTWTSGGATDTQYF	FLRGRAYGL
1mi5	B*08	TRAV26-2*01	TRAJ52*01	CILPLAGGTSYGKLTF	TRBV7-8*01	TRBJ2-7*01	CASSLGQAYEQYF	FLRGRAYGL
4qrp	B*08	TRAV9-2*01	TRAJ43*01	CALSDPVNDMRF	TRBV11-2*01	TRBJ1-5*01	CASSLRGRGDQPQHF	HSKKKCDEL
4g9f	B*27	TRAV14/DV4*02	TRAJ21*01	CAMRDLRDNFNKFYF	TRBV6-5*01	TRBJ1-1*01	CASREGLGGTEAFF	KRWIIMGLNK
4jrx	B*35	TRAV19*01	TRAJ34*01	CALSGFYNTDKLIF	TRBV6-1*01	TRBJ1-1*01	CASPGETEAFF	LPEPLPQGQLTAY
2ak4	B*35	TRAV19*01	TRAJ34*01	CALSGFYNTDKLIF	TRBV6-1*01	TRBJ2-7*01	CASPGLAGEYEQYF	LPEPLPQGQLTAY
3mv7	B*35	TRAV20*01	TRAJ58*01	CAVQDLGTSGSRLTF	TRBV9*01	TRBJ2-2*01	CASSARSGELFF	HPVGEADYFEY
4jry	B*35	TRAV39*01	TRAJ33*01	CAVGGGSNYQLIW	TRBV5-6*01	TRBJ2-7*01	CASSRTGSTYEQYF	LPEPLPQGQLTAY
3dxa	B*44	TRAV26-1*01	TRAJ13*02	CIVWGGYQKVTF	TRBV7-9*01	TRBJ2-1*01	CASRYRDDSYNEQFF	EENLLDFVRF
3kpr	B*44	TRAV26-2*01	TRAJ52*01	CILPLAGGTSYGKLTF	TRBV7-8*01	TRBJ2-7*01	CASSLGQAYEQYF	EEYLKAWTF
4mji	B*51	TRAV17*01	TRAJ22*01	CATDDDSARQLTF	TRBV7-3*01	TRBJ2-2*01	CASSLTGGGELFF	TAFTIPSI
2ypl	B*57	TRAV5*01	TRAJ13*01	CAVSGGYQKVTF	TRBV19*01	TRBJ1-2*01	CASTGSYGYTF	KAFSPEVIPMF
4p4k	DPA1*01/DPB1*352	TRAV9-2*01	TRAJ28*01	CALSLYSGAGSYQLTF	TRBV5-1*01	TRBJ2-5*01	CASSLAQGGETQYF	QAFWIDLFETIG
4may	DQA1*01/DQB1*05	TRAV13-1*01	TRAJ48*01	CAASSFGNEKLTF	TRBV7-3*01	TRBJ2-3*01	CATSALGDTQYF	QLVHFVRDFAQL
5ks9	DQA1*03/DQB1*03	TRAV20*01	TRAJ39*01	CAVALNNNAGNMLTF	TRBV9*01	TRBJ2-3*01	CASSVAPGSDTQYF	APSGEGSFQPSQENPQ
4gg6	DQA1*03/DQB1*03	TRAV26-2*01	TRAJ45*01	CILRDGRGGADGLTF	TRBV9*01	TRBJ2-7*01	CASSVAVSAGTYEQYF	QQYPSGEGSFQPSQENPQ
4z7u	DQA1*03/DQB1*03	TRAV26-2*01	TRAJ49*01	CILRDRSNQFYF	TRBV9*01	TRBJ2-5*01	CASSTTPGTGTETQYF	APSGEGSFQPSQENPQGS
4z7v	DQA1*03/DQB1*03	TRAV26-2*01	TRAJ54*01	CILRDSRAQKLVF	TRBV9*01	TRBJ2-7*01	CASSAGTSGEYEQYF	APSGEGSFQPSQENPQGS
4z7w	DQA1*03/DQB1*03	TRAV8-3*01	TRAJ36*01	CAVGETGANNLFF	TRBV6-1*01	TRBJ2-1*01	CASSEARRYNEQFF	APSGEGSFQPSQENPQGS
4ozh	DQA1*05/DQB1*02	TRAV26-1*01	TRAJ32*01	CIVWGGATNKLIF	TRBV7-2*01	TRBJ2-3*01	CASSVRSTDTQYF	APQPELPYPQPGS
4ozg	DQA1*05/DQB1*02	TRAV26-1*01	TRAJ45*01	CIVLGGADGLTF	TRBV7-2*01	TRBJ2-3*01	CASSFRFTDTQYF	APQPELPYPQPGS
4ozf	DQA1*05/DQB1*02	TRAV26-1*01	TRAJ54*01	CIAFQGAQKLVF	TRBV7-2*01	TRBJ2-3*01	CASSFRALAADTQYF	APQPELPYPQPGS
4ozi	DQA1*05/DQB1*02	TRAV4*01	TRAJ4*01	CLVGDGGSFSGGYNKLIF	TRBV20-1*01	TRBJ2-5*01	CSAGVGGQETQYF	QPFPQPELPYPGS
5ksa	DQA1*05/DQB1*03	TRAV20*01	TRAJ33*01	CAVQFMDSNYQLIW	TRBV9*01	TRBJ2-7*01	CASSVAGTPSYEQYF	QPQQSFPEQEA
5ksb	DQA1*05/DQB1*03	TRAV20*01	TRAJ6*01	CAVQASGGSYIPTF	TRBV9*01	TRBJ2-3*01	CASSNRGLGTDTQYF	GPQQSFPEQEA
4e41	DRA*01/DRB1*01	TRAV22*01	TRAJ18*01	CAVDRGSTLGRLYF	TRBV5-8*01	TRBJ2-5*01	CASSQIRETQYF	GELIGILNAAKVPAD
2iam	DRA*01/DRB1*01	TRAV22*01	TRAJ54*01	CAALIQGAQKLVF	TRBV6-6*01	TRBJ1-3*01	CASTYHGTGYF	GELIGILNAAKVPAD
1fyt	DRA*01/DRB1*01	TRAV8-4*01	TRAJ48*01	CAVSESPFGNEKLTF	TRBV28*01	TRBJ1-2*01	CASSSTGLPYGYTF	PKYVKQNTLKLAT
3o6f	DRA*01/DRB1*04	TRAV26-2*01	TRAJ32*01	CTVYGGATNKLIF	TRBV20-1*01	TRBJ1-6*01	CSARGGSYNSPLHF	FSWGAEGQRPGFGSGG
1j8h	DRA*01/DRB1*04	TRAV8-4*01	TRAJ48*01	CAVSESPFGNEKLTF	TRBV28*01	TRBJ1-2*01	CASSSTGLPYGYTF	PKYVKQNTLKLAT
2wbj	DRA*01/DRB1*15	TRAV17*01	TRAJ40*01	CATDTTSGTYKYIF	TRBV20-1*01	TRBJ2-1*01	CSARDLTSGANNEQFF	MDFARVHFISALHGSGG
4h1l	DRA*01/DRB3*03	TRAV8-3*01	TRAJ37*01	CAVGASGNTGKLIF	TRBV19*01	TRBJ2-2*01	CASSLRDGYTGELFF	QHIRCNIPKRISA
1zgl	DRA*01/DRB5*01	TRAV9-2*01	TRAJ12*01	CALSGGDSSYKLIF	TRBV5-1*01	TRBJ1-1*01	CASSLADRVNTEAFF	VHFFKNIVTPRTPGG

[a] If there are multiple structures with the same TCR and HLA allele, only the ID of the highest-resolution structure is given. During CDR3β contact analysis, however, we combined the contacts from all redundant structures, downweighting so as to equalize the contribution from all TCR/HLA pairs.

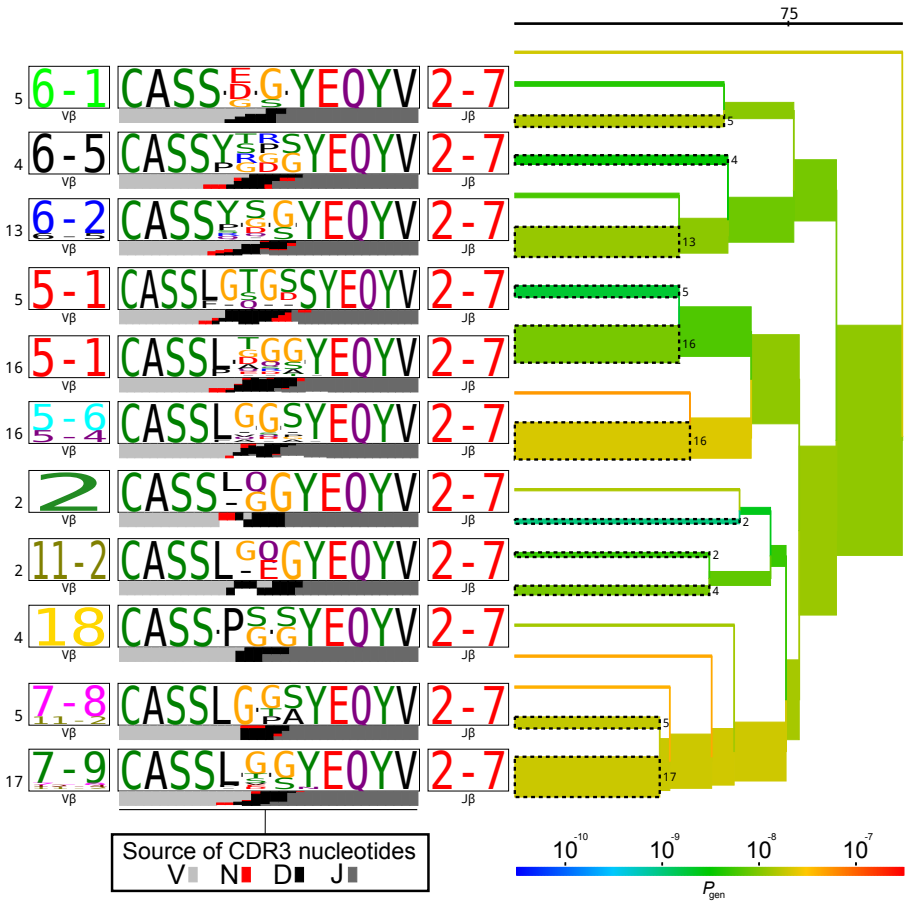


Figure 2–Figure supplement 1. TCRdist tree of the members of the TRBJ2-7*02-associated cluster. Average-linkage dendrogram of TCRdist receptor clusters colored by generation probability (P_{gen}), with TCR logos for selected receptor subsets (the branches of the tree enclosed in dashed boxes labelled with size of the TCR clusters). Each logo depicts the V- (left side) and J- (right side) gene frequencies, CDR3 amino acid sequences (middle), and inferred rearrangement structure (bottom bars coloured by source region, light grey for the V-region, dark grey for J, black for D, and red for N-insertions) of the grouped receptors.

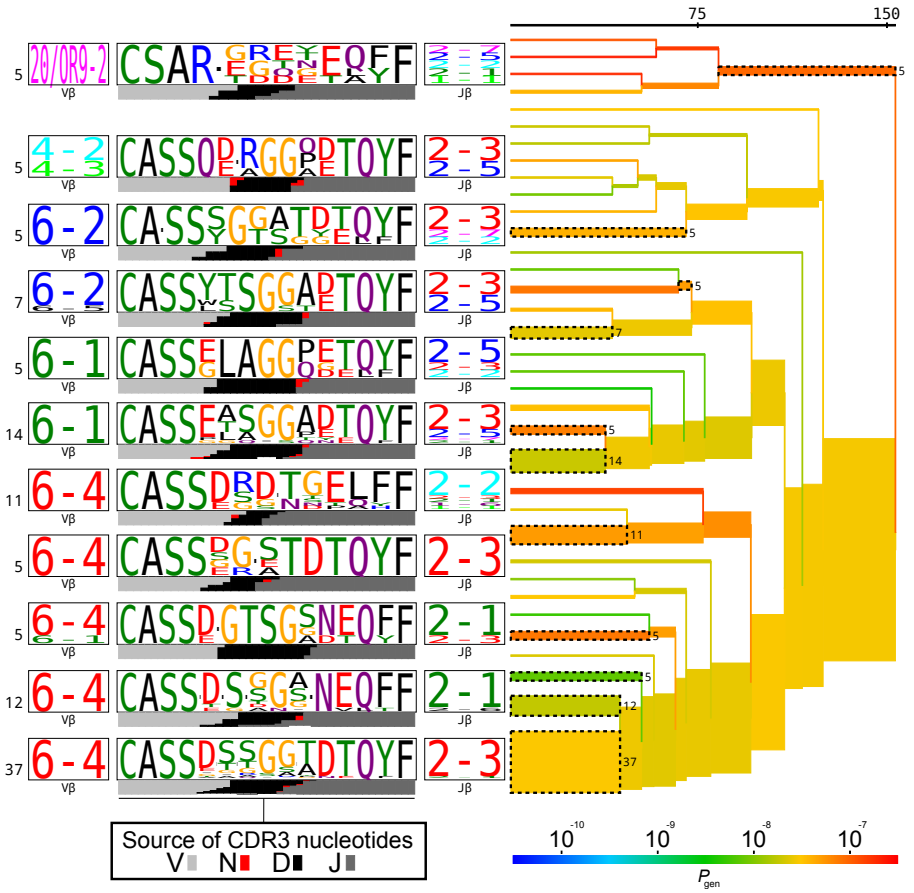
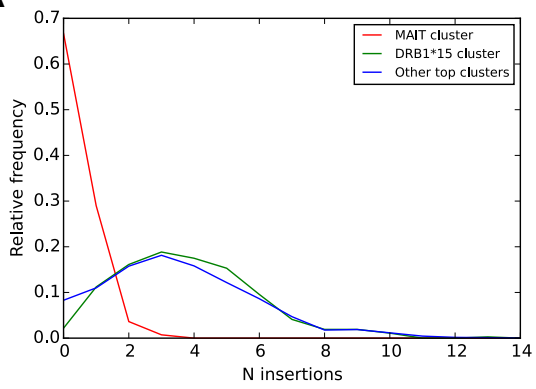
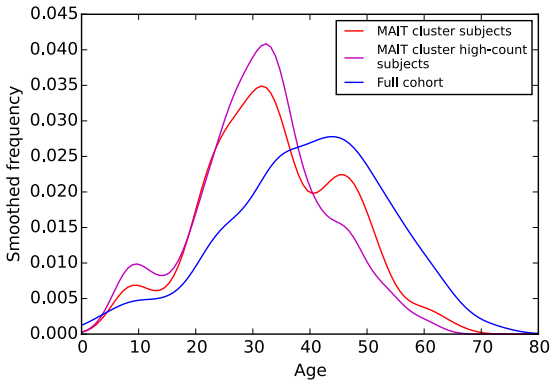


Figure 2–Figure supplement 2. TCRdist tree of the members of the putative MAIT cell cluster. Average-linkage dendrogram of TCRdist receptor clusters colored by generation probability (P_{gen}), with TCR logos for selected receptor subsets (the branches of the tree enclosed in dashed boxes labelled with size of the TCR clusters). Each logo depicts the V- (left side) and J- (right side) gene frequencies, CDR3 amino acid sequences (middle), and inferred rearrangement structure (bottom bars coloured by source region, light grey for the V-region, dark grey for J, black for D, and red for N-insertions) of the grouped receptors.

A



B



C

V-alpha	J-alpha	CDR3-alpha	TCR-beta
TRAV1-1*01	TRAJ3*01	CAVRERRSSASKIIF	V06,CASSDSAGGPDTQYF
TRAV1-2*01	TRAJ11*01	CAVPFYSGYSTLTF	V06,CASSERDTGELFF
TRAV1-2*01	TRAJ12*01	CAALDSSYKLIF	V06,CASSDSGGANEQFF
TRAV1-2*01	TRAJ16*01	CAKDGQKLLF	V06,CASSDSSTDTQYF
TRAV1-2*01	TRAJ20*01	CAALGNDYKLSF	V06,CASSDSGGSYNEQFF
TRAV1-2*01	TRAJ20*01	CAGLSNDYKLSF	V06,CASSDSAGGPDTQYF
TRAV1-2*01	TRAJ20*01	CAVIRNDYKLSF	V06,CASSDSAGGPDTQYF
TRAV1-2*01	TRAJ20*01	CAVKSGDYKLSF	V06,CASSELAGGPDTQYF
TRAV1-2*01	TRAJ20*01	CAVNGDDYKLSF	V06,CASSDSSGGADTQYF
TRAV1-2*01	TRAJ20*01	CAVRDGDYKLSF	V06,CASSERGGTDTQYF
TRAV1-2*01	TRAJ33*01	CAVGESNYQLIW	V06,CASSDSGGSYNEQFF
TRAV1-2*01	TRAJ33*01	CAVRDRNYQLIW	V06,CASSDSSTGELFF
TRAV1-2*01	TRAJ33*01	CAVRDSNYQLIW	V06,CASSDSAGGTDTQYF
TRAV1-2*01	TRAJ33*01	CTPLDSNYQLIW	V04,CASSQDRGGQETQYF
TRAV1-2*01	TRAJ4*01	CAAPRSPGGYNKLIF	V06,CASSDSSTDTQYF
TRAV10*01	TRAJ27*01	CVVSVGTNAGKSTF	V06,CASSERDTGELFF
TRAV10*01	TRAJ6*01	CVVGQRGGSYIPTF	V04,CASSQDRGGADTQYF
TRAV12-1*01	TRAJ5*01	CVEDTGRRALTF	V06,CASSERGGTDTQYF
TRAV13-1*01	TRAJ5*01	CAANRTGRRALTF	V06,CASSEDSSGANVLTF
TRAV13-1*01	TRAJ7*01	CAAVFYGNNRLAF	V06,CASSDSAGGTDTQYF
TRAV13-2*01	TRAJ34*01	CAYNTDKLIF	V06,CASSDSTSGSNEQFF
TRAV14/DV4*02	TRAJ3*01	CAMREGLTLGSSASKIIF	V04,CASSQDRGGQETQYF
TRAV14/DV4*02	TRAJ6*01	CAMSKGEGGSYIPTF	V06,CASSDSGGGTDTQYF
TRAV19*01	TRAJ8*01	CALSEALGFQKLVF	V06,CASSDSSTGELFF
TRAV2*01	TRAJ21*01	CAVVGYNFNKFYF	V06,CASSDSTSGADTQYF
TRAV3*01	TRAJ13*02	CAGLRGGYQKVTF	V06,CASSDSGGSYNEQFF
TRAV4*01	TRAJ31*01	CLVGASNARLMF	V06,CASSGTSGSYNEQFF
TRAV41*01	TRAJ54*01	CAVRRVAQKLVF	V06,CASSASGGADTQYF
TRAV5*01	TRAJ22*01	CAERPPSGSARQLTF	V04,CASSQERGGQETQYF
TRAV6*01	TRAJ5*01	CALDSGGGRRALTF	V06,CASSGTSGGTDTQYF
TRAV8-1*01	TRAJ13*02	CAVILQRGYQKVTF	V06,CASSDSGGSYNEQFF
TRAV8-2*01	TRAJ24*02	CVVSATDSWGKLQF	V06,CASSDRDTGELFF
TRAV8-3*01	TRAJ45*01	CAVGAAMYSGGGADGLTF	V06,CASSETSGSYNEQFF
TRAV8-4*01	TRAJ26*01	CAVSGNYGQNFVF	V06,CASSETSGGADTQYF
TRAV9-2*01	TRAJ10*01	CALSGGGNKLTF	V06,CASSDGTGGNEQFF
TRAV9-2*01	TRAJ8*01	CGGFQKLVF	V06,CASSYSTSGGNEQFF

Figure 2–Figure supplement 3. Further details on the putative MAIT cell TCR cluster. **(A)** Distribution of N-nucleotide insertions for TCRβ chains in the MAIT cluster (red), in the DRB1*15-associated cluster (green), and in the union of the members of the top 10 clusters (excluding the members of the MAIT cluster, blue). MAIT cell cluster members have very few N-insertions relative to the members of the other clusters. **(B)** Subjects enriched for MAIT cluster TCRs (red curve) are younger than the cohort as a whole (blue curve), a trend that is further strengthened in the top half of the enriched subjects by member-TCR count (the 'high-count subjects', magenta curve). **(C)** TCRα chains paired with MAIT cluster TCRβ chains in the pairSEQ dataset of Howie Howie et al. (2015). Ten of the 36 paired TCRα chains match the MAIT sequence consensus (TRAV1-2, TRAJ20 or TRAJ33, and a 12 residue CDR3, enclosed in the blue box).

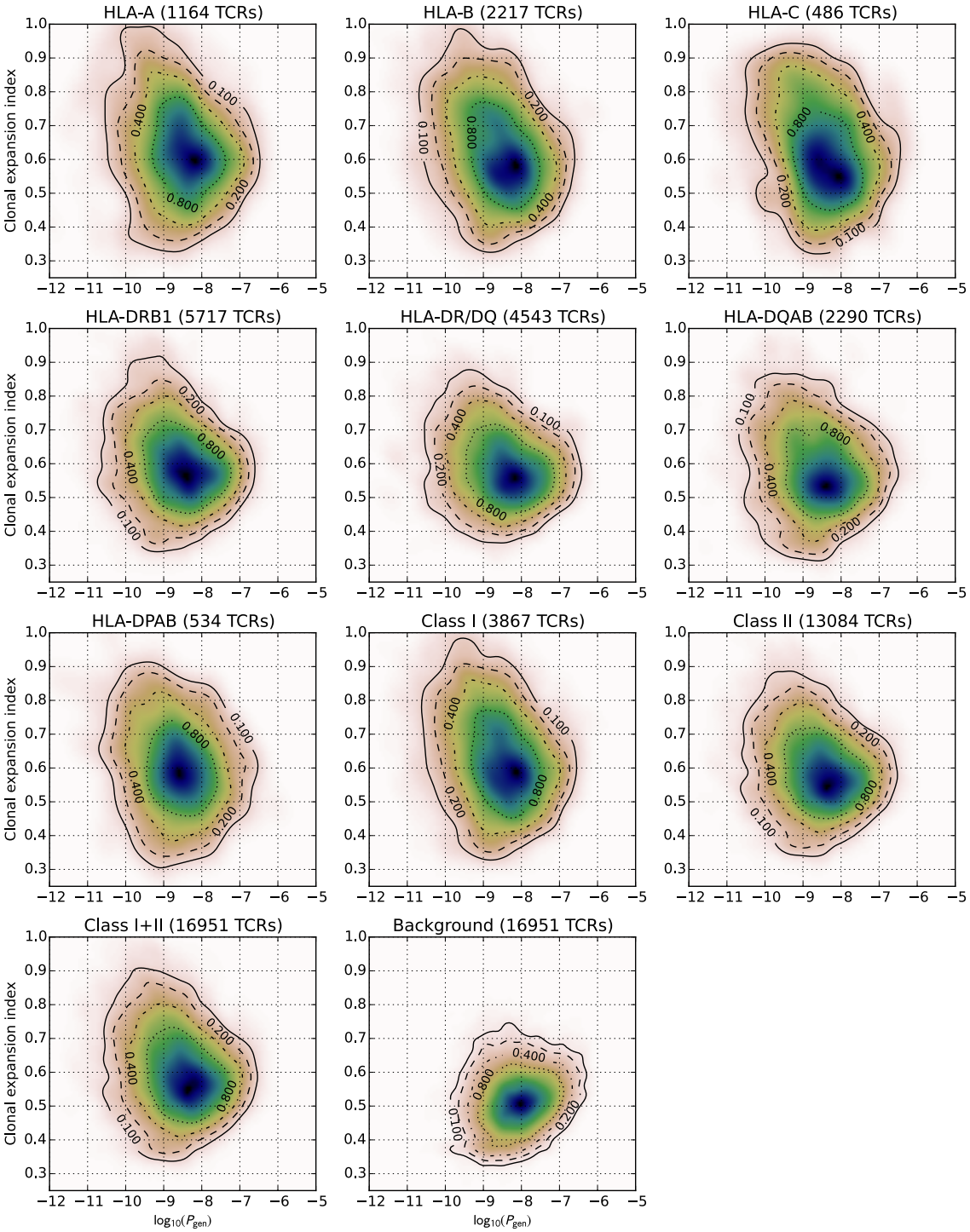


Figure 3–Figure supplement 1. Two-dimensional distributions of TCR generation probability (x-axis, P_{gen}) and clonal expansion index (y-axis) for TCRs with the indicated HLA associations (panel headers), and for a background set of non-HLA associated, cohort-frequency matched TCRs.

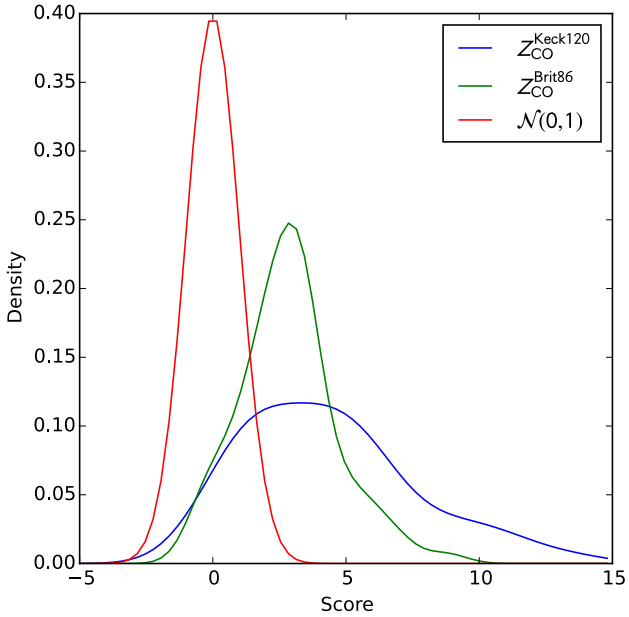


Figure 5–Figure supplement 1. Smoothed distributions of cluster co-occurrence scores on the two validation cohorts. Gaussian kernel density estimation (KDE)-smoothed distributions of the cluster member TCR co-occurrence scores (Z_{CO}) for the two validation cohorts. A standard normal distribution is shown as an approximate null expectation for these Z-scores.