ORIGINAL RESEARCH
published: 31 July 2019
doi: 10.3389/fimmu.2019.01516

Single T Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage and Antigen Specificity

Jason A. Carter[1,2], *Jonathan B. Preall*[2], *Kristina Grigaityte*[2,3], *Stephen J. Goldfless*[4], *Eric Jeffery*[4], *Adrian W. Briggs*[4], *Francois Vigneault*[4] and *Gurinder S. Atwal*[2,3]*

[1] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, United States, [2] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, United States, [3] Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, United States, [4] Juno Therapeutics, Seattle, WA, United States

Although structural studies of individual T cell receptors (TCRs) have revealed important roles for both the α and β chain in directing MHC and antigen recognition, repertoire-level immunogenomic analyses have historically examined the β chain alone. To determine the amount of useful information about TCR repertoire function encoded within $\alpha\beta$ pairings, we analyzed paired TCR sequences from nearly 100,000 unique CD4$^+$ and CD8$^+$ T cells captured using two different high-throughput, single-cell sequencing approaches. Our results demonstrate little overlap in the healthy CD4$^+$ and CD8$^+$ repertoires, with shared TCR sequences possessing significantly shorter CDR3 sequences corresponding to higher generation probabilities. We further utilized tools from information theory and machine learning to show that while α and β chains are only weakly associated with lineage, $\alpha\beta$ pairings appear to synergistically drive TCR-MHC interactions. V$\alpha\beta$ gene pairings were found to be the TCR feature most informative of T cell lineage, supporting the existence of germline-encoded paired $\alpha\beta$ TCR-MHC interaction motifs. Finally, annotating our TCR pairs using a database of sequences with known antigen specificities, we demonstrate that approximately a third of the T cells possess α and β chains that each recognize different known antigens, suggesting that $\alpha\beta$ pairing is critical for the accurate inference of repertoire functionality. Together, these findings provide biological insight into the functional implications of $\alpha\beta$ pairing and highlight the utility of single-cell sequencing in immunogenomics.

Keywords: TCR–T cell receptor, CD4 and CD8 T cell repertoires, TCR repertoire diversity, single-cell sequencing, machine learning

OPEN ACCESS

Edited by:
Remy Bosselut,
National Cancer Institute (NCI),
United States

Reviewed by:
Paolo Dellabona,
San Raffaele Scientific Institute (IRCCS), Italy
Nicole L. La Gruta,
Monash University, Australia

***Correspondence:**
Gurinder S. Atwal
atwal@cshl.edu

Specialty section:
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology

Received: *05 April 2019*
Accepted: *18 June 2019*
Published: *31 July 2019*

Citation:
Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Jeffery E, Briggs AW, Vigneault F and Atwal GS (2019) Single T Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage and Antigen Specificity. Front. Immunol. 10:1516. doi: 10.3389/fimmu.2019.01516

INTRODUCTION

With potentially up to 10^{15} unique $\alpha\beta$ T cell receptor (TCR) pairs, a wealth of clinically-relevant information pertaining to infectious disease, autoimmunity, and cancer immunotherapy is encoded within the remarkable diversity of the TCR repertoire (1–3). As limitations in technology have historically precluded meaningful single-cell sequencing experiments, our current understanding of the TCR repertoires' diversity, structure, and function is almost entirely based on bulk-sequencing of the β chain repertoire alone (4–6). While such approaches have yielded impressive insights into adaptive immunity, they, *de facto*, are forced to make use of the assumption that the pairing of $\alpha\beta$ TCR chains contains little useful information. In contrast, structural insights gleaned from a relatively small number of TCR-peptide-MHC structures have clearly defined important roles for both the α and β TCR chains in driving alloreactivity and antigen specificity (7–10). While our understanding of the underlying biology suggests that $\alpha\beta$ pairings may themselves contain useful information on TCR function and repertoire diversity, whether this theoretical information can be approximated from bulk-sequencing, and if not, whether it can be utilized to meaningfully improve our understanding of the TCR repertoire remains largely a matter of conjecture.

While previous methods for paired $\alpha\beta$ TCR sequencing have been developed (11–15), only recently have technological advances enabled high-throughput capture of paired $\alpha\beta$ TCR sequences (16–18). We recently took advantage of one such single-cell sequencing method to capture more than 200,000 paired $\alpha\beta$ TCR sequences from the peripheral blood of five healthy individuals, finding that the use of bulk and single-cell sequencing often resulted in significantly different diversity estimates (19). In the present study, we asked whether we could infer additional information about TCR repertoire function when examining paired $\alpha\beta$ sequences relative to either of the single chain repertoires. Toward this, we used 10× Genomics single-cell platform (17) to add ~11,000 new $\alpha\beta$ paired sequences to the ~86,000 $CD4^+$ and $CD8^+$ TCR sequences we previously obtained using the AbVitro method (18, 19). In addition to providing the most comprehensive comparison of the human $CD4^+$ and $CD8^+$ $\alpha\beta$ TCR repertoires to date, we examined the ability of $\alpha\beta$ pairings to provide information about T cell lineage and antigen specificity beyond that contained in the single-chain repertoires. At similar repertoire depths, we find that the paired $\alpha\beta$ repertoire contains useful information about TCR function, both in terms of MHC recognition and antigen specificity, that is not accessible through conventional bulk-sequencing. Consequently, our study demonstrates the utility of using new single-cell sequencing approaches, in addition to conventional high-throughput bulk-sequencing, to capture a more accurate picture of TCR repertoire function.

RESULTS

Overlap Between the $CD4^+$ and $CD8^+$ Repertoires

During thymic positive selection, bipotent T cell precursors differentiate into either the $CD4^+$ helper T cell or the $CD8^+$ cytotoxic T cell lineage. Although this lineage selection process is contingent upon the interaction of the heterodimeric $\alpha\beta$ TCR with either MHC class II or class I, respectively, understanding the general TCR features that mediate the TCR-pMHC interaction remains an area of active interest (20, 21). Potentially, the required ability to recognize structurally diverging MHC classes creates systematic differences in the $CD4^+$ and $CD8^+$ TCR repertoires. In support of this idea, previous studies have identified certain germline regions and CDR3 features in the single chain repertoires that are associated with up to ~5 times increase in likelihood for either $CD4^+$ or $CD8^+$ status (22–24). If $\alpha\beta$ TCR pairing is an important component for understanding the differences between two TCR repertoires, we hypothesized that $\alpha\beta$ pairs should be much less commonly shared between the $CD4^+$ or $CD8^+$ populations. That is, the information about $\alpha\beta$ pairing should correlate with increased functional specificity for one of the two MHC classes.

With this goal in mind, we first addressed how the paired $\alpha\beta$ TCR repertoires differ between the $CD4^+$ and $CD8^+$ T cell populations, independent from an individual's HLA type (**Supplemental Table 1**). Toward this, we obtained paired $\alpha\beta$ TCR sequences delineated by $CD4^+$ and $CD8^+$ lineage from our recently published work (19). In addition to these sequences captured using the AbVitro microfluidic platform (18), we resequenced samples from two individuals using the independent 10× Genomics single-cell sequencing platform (17). While we obtained only a small number of TCR sequences during resequencing, potentially due to RNA degradation secondary to prolonged storage times, a large fraction of these new TCR sequences were also found in the original dataset (**Figure 1A**). These findings strongly suggest the ability of both of these methods to accurately obtain TCR sequences in a high-throughput fashion and allowed us to confidently generate new single-cell datasets for two additional individuals. Combining results from the two methods allowed us to analyze nearly 100,000 unique paired $\alpha\beta$ TCR sequences drawn from the $CD4^+$ and $CD8^+$ TCR repertoire of seven healthy individuals. In order to avoid introducing biases stemming from large clonal expansions, we will consider only the unique set of TCR sequences for each repertoire. We additionally note that the $CD4^+$ and $CD8^+$ repertoires may still be biased by the presence of many similar, but not identical, clones responding to the same viral epitope (26, 27). However, as each of these similar clones still must maintain its ability to recognize a particular MHC class and should represent a relatively small fraction of the repertoire in healthy individuals, the impact of these sequences on the observed repertoires is expected to be minimal.

Considering the unique set of TCR clonotypes (V$\alpha\beta$ and amino acid CDR3$\alpha\beta$) across all individuals, we found that the paired $CD4^+$ and $CD8^+$ repertoires were largely distinct from one another ($\alpha\beta_{overlap}$ =0.65% of total $\alpha\beta$ sequences). Splitting the paired repertoire into the constituent α chain ($\alpha_{overlap}$=7.8%) and β chain ($\beta_{overlap}$=4.7%) repertoires resulted in considerably higher overlap between the two lineages (**Figure 1B**). We note that $\alpha_{overlap} \cdot \beta_{overlap} \approx \alpha\beta_{overlap}$, potentially reflecting roughly independent contributions of the α and β chains. Quantifying the overlap between the $CD4^+$ and $CD8^+$ TCR repertoires within each individual, we observed greater similarity between the $CD4^+$ and $CD8^+$ single chain repertoires than between the paired $\alpha\beta$ repertoires (**Figure 1C**). The decreased similarity of the paired TCR repertoires relative to the single chain repertoires, however, is not unique to the comparison of the $CD4^+$ and $CD8^+$ repertoires. For example, comparison of the single chain and paired $CD4^+$ or $CD8^+$ repertoires between individuals produces similar decreases in repertoire overlap and is likely reflective of

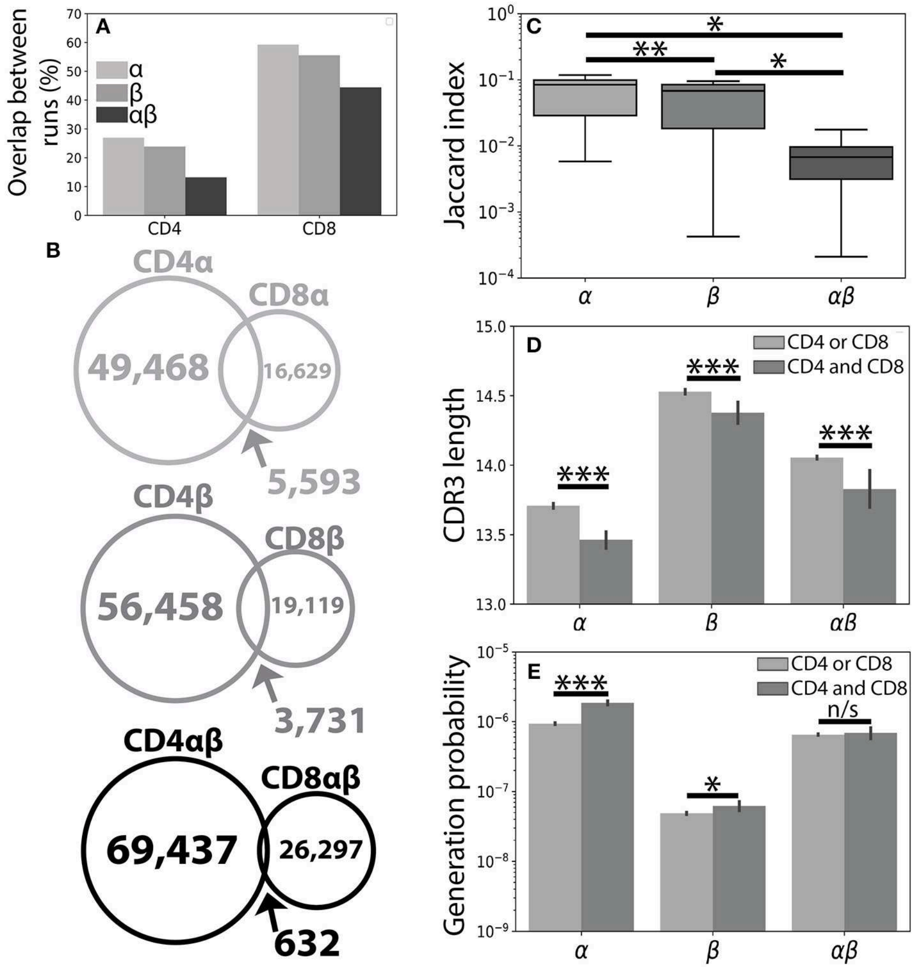


FIGURE 1 | The CD4$^+$ and CD8$^+$ TCR repertoires are largely distinct. **(A)** A single peripheral sample from Subject 1 was sequenced using both the Abvitro (18) and 10× Genomics (17) platforms. The number of TCR clones found using both methods is reported as a fraction of the total number of TCRs observed in the considerably smaller 10× run. **(B)** Population unique TCRs, defined as V$\alpha\beta$-CDR3$\alpha\beta$ amino acid clonotypes, were calculated by combining CD4$^+$ and CD8$^+$ sequences from 7 healthy individuals. The overlap between CD4$^+$ and CD8$^+$ repertoires were calculated for the α (top), β (middle), and paired $\alpha\beta$ (bottom) repertoires. **(C)** The Jaccard index quantitatively measures the overlap between the CD4$^+$ and CD8$^+$ repertoires separately for each individual. Significance between samples was assessed using a paired sample Student's *t*-test. **(D)** Bar plots show mean CDR3 lengths for α, β, and $\alpha\beta$ TCR sequences found exclusively in (light gray) or shared between (dark gray) the CD4$^+$ and CD8$^+$ lineages. Error bars represent bootstrapped 99% confidence intervals of the mean with significance assessed using the Mann-Whitney *U* test. **(E)** The probability of generating a given CDR3 sequence was estimated using OLGA (25). $\alpha\beta$ probabilities were calculated by assuming complete independence between the α and β chain. Shared sequences were more likely to be randomly generated than those found in only one of the repertoires. For all panels, n/s-not significant, $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

the lower generation probability associated with a given $\alpha\beta$ TCR pair relative to either of its constituent single chains.

Previous findings have suggested that TCRs shared between individuals may have shorter CDR3β sequences and may be closer to germline recombination sequences than clonotypes found only in a single individual (28). Accordingly, TCR sequences shared between the CD4$^+$ and CD8$^+$ lineages were, on average, shorter than those found only in one of the two lineages with respect to the α, β. and $\alpha\beta$ repertoires (**Figure 1D** and **Supplemental Figure 1**). We further confirmed this finding using the OLGA software package to calculate the probability of randomly generating a given CDR3 sequence (25). As expected, TCR clones found in both repertoires additionally had higher generation probabilities than those found in a single repertoire (**Figure 1E**). Given the relative uniqueness of $\alpha\beta$ TCRs for the CD4$^+$ and CD8$^+$ repertoires and previous structural findings implicating both chains in determining TCR-pMHC binding (7–10), we next asked whether $\alpha\beta$ pairings could provide more information about T cell lineage than either chain alone.

Association of VJ Germline Segment Usage With CD4$^+$-CD8$^+$ Status

High-throughput sequencing of the β chain repertoire has revealed an association between the expression levels of specific TCR V-regions and MHC polymorphisms (29) and identified HLA-associated TCRβ sequences (30, 31). Furthermore, significant biases in V and J germline segment use between the single-chain CD4$^+$ and CD8$^+$ repertoires have been previously identified (22–24). One possible explanation for these observations, and generally a mechanism that enables MHC restriction, posits the existence of germline-encoded sequences that have been evolutionarily hard-wired into the Variable (V) region's CDR1 and CDR2 loops (32, 33). Evidence for such hard-wired regions biasing, but not completely determining, the interactions between TCRs and pMHC complexes is primarily drawn from a multitude of structural studies, which have identified a widely conserved TCR-MHC docking orientation (20), as well as other conserved TCR-MHC interaction motifs (34–38). The preference of specific germline regions for a particular MHC class is thought to create systematic biases in the CD4$^+$ or CD8$^+$ repertoires. We thus next hypothesized that if the paired $\alpha\beta$ repertoire contained additional information about the function of the TCR repertoire, paired germline features should be more informative of T cell lineage than either of the single-chain repertoire alone. Specifically, information about $\alpha\beta$ pairing should allow us to better understand the factors that influence TCR interaction with MHC and ultimately the factors at play in T cell differentiation.

To further explore this possibility, we split the CD4$^+$ and CD8$^+$ repertoires into unique α, β, and $\alpha\beta$ subsets, which allows us to directly compare each single-chain repertoire with that of pairs at similar sample sizes. We then calculated the odds ratio (OR) of observing a given Vα or Vβ in the CD4$^+$ repertoire relative to the CD8$^+$ repertoire. In this sense, the OR compares the odds of a given TCR feature being used in a given CD4$^+$ TCR to the odds of it being used in a cell from the CD8$^+$ population. Thus, an odds ratio that is >1 indicates a CD4$^+$ bias, while an OR <1 is reflective of preferential use in the CD8$^+$ repertoire. Calculating Bonferroni-corrected p values using the Fisher's Exact test, we identified weak, but statistically significant associations in both the Vα and Vβ single-chain repertoires (**Figures 2A,B** and **Supplemental Figures 3A–D**). Interestingly, these associations are significantly weaker than previously reported, potentially due to the more rigorous correction for PCR biases enabled by unique molecular identifiers (UMIs) available in single-cell sequencing (22). We further note weaker associations between T cell lineage and single chain Jα and Jβ usage (**Supplemental Figures 2A,B**, **3D–H**).

The role of $\alpha\beta$ germline segment pairing in biasing T cell differentiation was similarly examined by comparing the odds of observing a given V$\alpha\beta$ or J$\alpha\beta$ pair in each of the two repertoires. We show the statistically significant ($q \leq 0.05$) CD4$^+$:CD8$^+$ odds ratios for 349 V$\alpha\beta$ and 79 J$\alpha\beta$ pairs associated with a significant lineage specification bias (**Figure 2C**

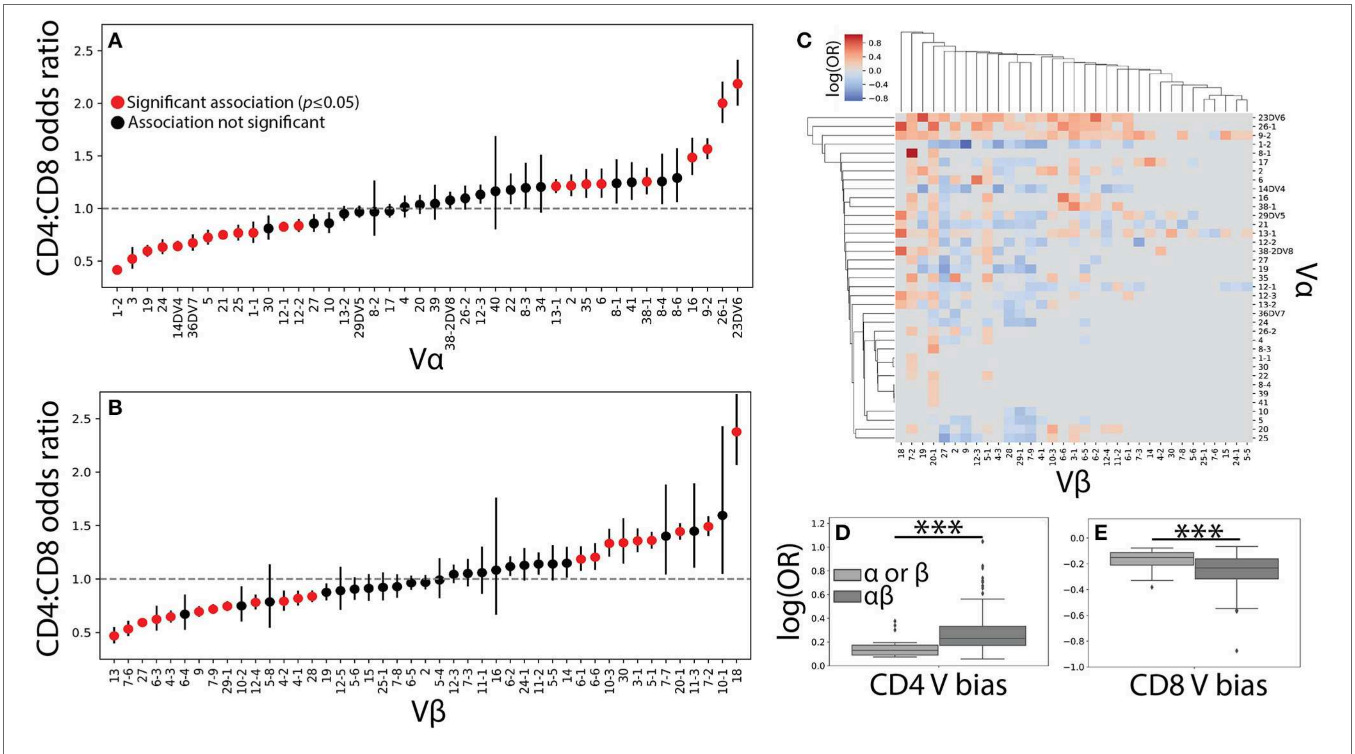


FIGURE 2 | V$\alpha\beta$ pairings encode strong associations with T cell lineage. **(A)** The CD4$^+$ and CD8$^+$ TCR repertoires were then pooled across individuals and the CD4$^+$:CD8$^+$ odds ratio (OR) was calculated for each Vα and **(B)** Vβ germline region. An OR> 1 represents a CD4$^+$ bias, while an OR< 1 represents a CD8$^+$ bias with error bars representing the 95% confidence interval. The mean is represented by a red or black dot, with red representing statistical significance at p<0.05 by Fisher's exact test after Bonferroni correction. **(C)** Significant (q < 0.05 by Fisher's exact test) log odds ratios reveals strong CD4$^+$:CD8$^+$ biases for 349 V$\alpha\beta$ pairs. **(D)** Boxplots were calculated for the set of all significant odds ratios associated with single chains (Vα or Vβ) and compared with those associated with $\alpha\beta$ pairs. Paired associations for both CD4$^+$ and **(E)** CD8$^+$ status were significantly stronger (***p < 0.001 by Mann-Whitney U test) than those for the single chain alone.

and **Supplemental Figure 2C**). The strength of association with T cell lineage was significantly stronger for V$\alpha\beta$ pairs than for J$\alpha\beta$ pairs, likely reflecting the contribution of the CDR1 and CDR2 loops present in each V region to MHC binding (**Supplemental Figures 2D–E**). This finding supports the existence of germline-encoded TCR-MHC interaction motifs and raises the possibility that such motifs in both the α and β chains act in concert with one another.

Unsurprisingly, paired V$\alpha\beta$ provides a more nuanced view of germline associations when compared with the single-chain repertoires alone, with associations confined too specific pairs (**Figure 2C**). Qualitatively, our data reveals several associations in the paired data that would have otherwise been missed in the single chain results. For example, TRBV20-1 is strongly associated with CD4$^+$ status in the single chain dataset, but paired analysis reveals several α chains for which TRBV20-1 has significant CD8$^+$ associations (e.g., TRAV1-2, TRAV19, TRAV36DV7). Similarly, TRAV4 has no association in the single chain data, but several associations with specific β chains (e.g., TRBV6-5, TRBV5-1, TRBV2). The observed associations between paired V$\alpha\beta$ germline regions and T cell lineage were additionally, on average, significantly stronger than those associations found for either of the single chain repertoires individually (**Figures 2D–E**). Biologically, this finding is consistent with the notion that both the α and β chain contribute substantially to TCR-pMHC binding (7–10). Furthermore, these germline region biases are observed across individuals of differing HLA types and are consequently likely to be representative of differences between MHC classes rather than from individual MHC polymorphisms.

CDR3 Features Alone Are Weakly Associated With T Cell Lineage

Conventionally, the CDR1 and CDR2 loops encoded entirely within the germline Vα and Vβ regions have been thought to predominate the TCRs interaction with MHC. However, recent structural evidence has additionally noted interactions between the CDR3 region, which predominantly drives antigen specificity, and MHC (20, 21). As such, we additionally investigated the role of CDR3$\alpha\beta$ pairing in driving MHC class I or II recognition. To gain a better understanding of CDR3 composition, we first calculated the frequency with which each amino acid was used across all α or β CDR3 regions. We observed strong differences in amino acid usage between the α and β chains, likely due to differences in the germline composition of α and β V(D)J segments (**Figures 3A,B**). However, we observed only small, insignificant differences in amino acid use between the CD4$^+$ and CD8$^+$ repertoires (**Figures 3C,D**). Although the overall effect size remained small, we did note increased use of negatively charged amino acids in the CD8$^+$ T cell population in both the α and β repertoires. Similarly, our data suggested an increased use of positively charged amino acids in the CD4$^+$ population.

In order to gain a better understanding of how CDR3 net charge may effect MHC recognition, we calculated the odds ratio for CDR3 net charge between the two T cell populations. As expected (22, 23), we found that net positive charges were significantly associated with CD4$^+$ status and net negative charges were associated with the CD8$^+$ population (**Figure 3E** and **Supplemental Figures 4A–D**). We next calculated the odds ratio for joint CDR3$\alpha\beta$ charge pairs, finding a similar pattern to the single chain data (**Figure 3F**). Associations between T cell lineage and CDR3 charge were stronger for the paired chains, though these differences not statistically significant for CDR3 charge (**Figures 3G,H**).

To further explore the relationship between the CDR3 region and T cell lineage, we next examined CDR3 length. As found in previous studies, we identified only very weak relationships between lineage and CDR3α and CDR3β lengths (**Figure 3I** and **Supplemental Figures 4E–H**) (22, 23). Interestingly, however, we do observe a small number of CDR3$\alpha\beta$ length pairs that have significant associations with CD4$^+$ and CD8$^+$ status (**Figure 3J**). Again, we find that these paired interactions are substantially stronger than those found in the single-chain repertoire (**Figures 3K,L**). We note that associations for CDR3 charge and length were substantially weaker than those identified for V$\alpha\beta$ pairs, consistent with CDR3 sequence playing a smaller role in the TCR-MHC interaction than germline regions.

Paired Chain Sequences Are More Informative of CD4$^+$-CD8$^+$ Status Than Single Chains

To quantify the amount of information about CD4$^+$ and CD8$^+$ status encoded in the α, β, and $\alpha\beta$ TCR sequences, we next calculated the mutual information (39), corrected for finite sample sizes, between several TCR features and T cell lineage (**Figure 4A**). In brief, mutual information allows us to quantify the dependence of two random variables (e.g., the dependence of CD4$^+$/CD8$^+$ status on Vα gene usage), with a mutual information value of zero corresponding to statistical independence. Examining single chain V and J germline region usage frequencies, as well as CDR3 length and charge distributions, we find a small but non-negligible amount of information about T cell lineage. If the α and β chains encode information about T cell lineage in a conditionally independent manner, the expected information content of $\alpha\beta$ pairs can be found by summing the information contained by each chain individually ($\alpha + \beta$). Alternatively, the α and β could encode redundant information (as would be the case if the β chain was the predominate determinant of TCR-pMHC interactions) and would result in the information contained in $\alpha\beta$ pairs being less than the sum of the two chains ($\alpha\beta < \alpha + \beta$). Surprisingly, we observe synergistic information (41) in which the paired chains carry more information than the individual chains summed together ($\alpha\beta > \alpha + \beta$), which would suggest that the germline encoded interactions may act in a synergistic manner and highlights the importance of both chains in determining TCR function. Despite this observed synergy, the overall amount of information encoded in these general TCR features about T cell lineage remains relatively low.

Building from this observed synergistic information built into $\alpha\beta$ pairings, we next asked whether the use of paired sequences would better allow us to predict T cell lineage from

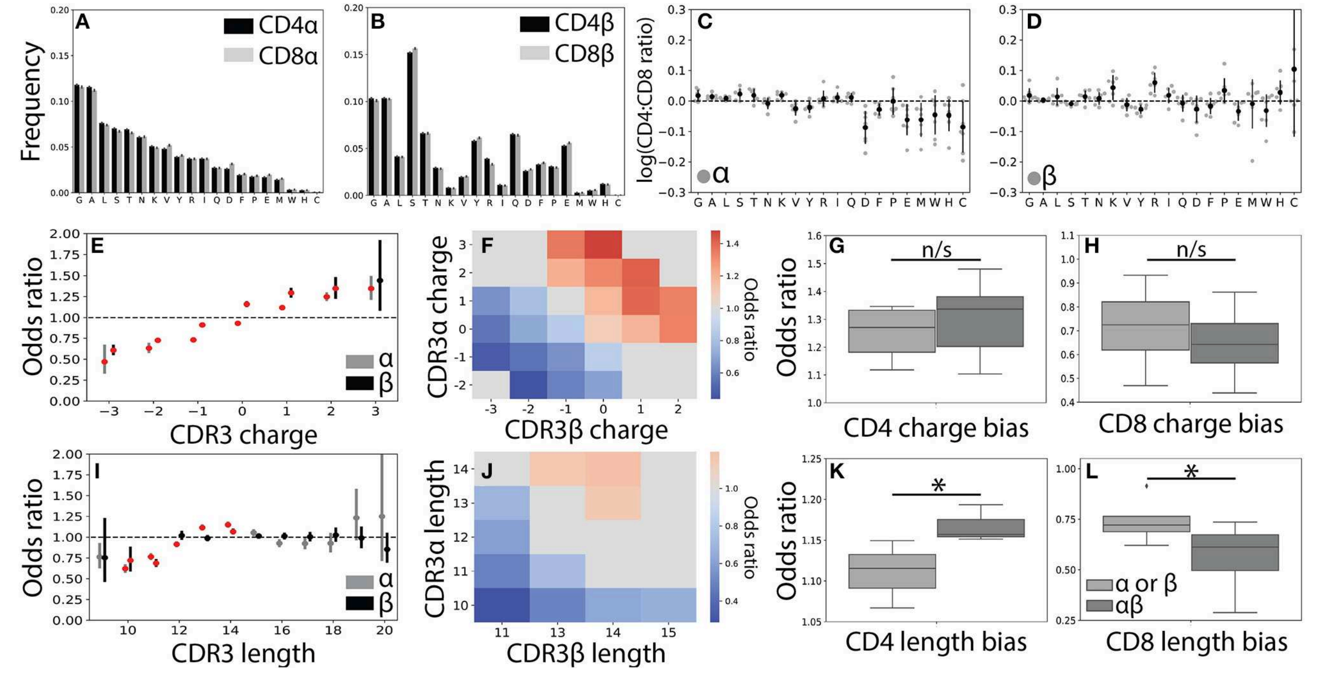


FIGURE 3 | CDR3 features correlate weakly with T cell lineage. **(A)** Usage frequencies for all 20 amino acids, rank ordered by prevalence in CDR3α, are shown for CDR3α and **(B)** CDR3β sequences across the CD4$^+$ and CD8$^+$ repertoires. **(C)** The CD4$^+$:CD8$^+$ usage ratio for all amino acids are shown for the α and **(D)** β chains. The frequency with which each amino acid is used is shown for each individual (gray circles) with the population mean and standard deviation shown in black. Amino acid usage was found to not significantly differ across the CD4$^+$ and CD8$^+$ repertoires using a one sample t-test after Bonferroni correction. **(E)** CD4$^+$ odds ratios (OR) quantify the strength of association of CDR3 net charge with lineage for both the α (gray, left) and β (black, right) chains. Red markers indicate statistical significance ($p < 0.05$ after Bonferroni correction). **(F)** Significant ($p < 0.05$ after Bonferroni correction) log odds ratios reveals strong CD4$^+$:CD8$^+$ bias for 23 CDR3$\alpha\beta$ charge pairs. Values that are not statistically significant are shown in gray (OR defined as being equal to one). **(G)** Boxplots compare the strength of association between T cell lineage and either single-chain features (α or β) and paired ($\alpha\beta$) CDR3 charges. Paired charges show stronger associations when compared with those of the single-chain for both the CD4$^+$ and **(H)** CD8$^+$ populations. **(I)** Single-chain odds ratio associations for CDR3 length. **(J)** Significant paired CDR3$\alpha\beta$ length association with T cell lineage. As in **(F)**, only statistically significant CD4$^+$:CD8$^+$ odds ratios are shown. **(K)** Boxplot compares length association strength with CD4$^+$ or **(L)** CD8$^+$ status for paired and single-chain features. *$p < 0.05$ by Mann Whitney U.

TCR features using machine learning classification. We obtained the highest accuracy using a gradient boosted decision tree classifier, specifically the XGBoost (40) algorithm (see Methods). Although the α (AUC $\approx$ 0.59 $\pm$ 0.006) and β (AUC $\approx$ 0.60 $\pm$ 0.006) chains were both only weakly informative of lineage, we found that the information encoded by paired TCR sequences ($\alpha\beta$ AUC $\approx$ 0.64 $\pm$ 0.005) allowed for a significant increase in model performance (**Figure 4B**). As our mutual information calculations demonstrated the presence of synergistic information within $\alpha\beta$ pairings, we reasoned that our machine learning classifier should reflect this additional information. To address this question, we independently trained classifiers on both the α and β chains and created a soft-voting ensemble ($\alpha + \beta$) to predict CD4$^+$ and CD8$^+$ lineage. Classifiers trained on $\alpha\beta$ pairs together significantly outperformed those trained on the additive model (additive AUC$\approx$ 0.63 $\pm$ 0.004, $p \leq 3 \times 10^{-15}$ against $\alpha\beta$ using a Mann Whitney U test), again suggesting a synergistic relationship between α and β pairs with respect to T cell lineage specification (**Figure 4B** inset).

Of note is a previous report using a SVM classifier and CDR3 length-dependent parametrization to predict T cell lineage from TCR sequences with >90% accuracy (23). This approach, however, failed to achieve the same degree of predictive accuracy when using our dataset (**Supplemental Figure 5**). To better understand this finding, we compared the TCR sequences from this previous study (23) with those reported here and an additional bulk-sequencing TCRβ dataset (24). We find that the aforementioned increased predictive accuracy is driven by anomalous Vβ and Jβ gene frequencies in the Li et al. dataset, possibly due to a lack of rigorous PCR correction, as compared with the other two datasets (**Supplemental Figure 6**).

Association of Paired $\alpha\beta$ Sequences With Known Peptide Specificity

Given the increased information contained within paired $\alpha\beta$ TCR sequences about T cell lineage, we next asked whether these paired sequences could provide us with additional information about peptide specificity. More specifically, we wondered whether information from $\alpha\beta$ pairing could be used to significantly improve our ability to understand the functional aspects of the TCR repertoire. In order to address this question, we downloaded more than 20,000 CDR3 sequences with known antigen specificities from a previously published repository [VDJdb (42)]. Of these known TCRs, more than ~96%

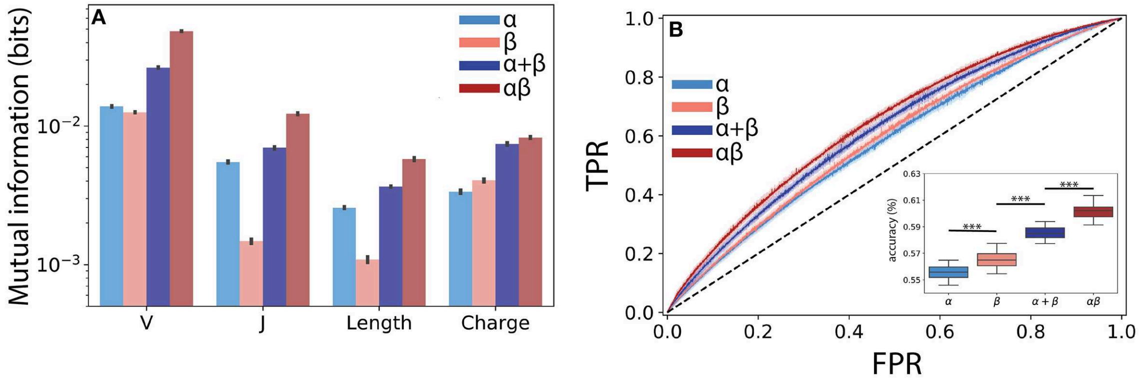


FIGURE 4 | Paired $\alpha\beta$ sequences are more informative of T cell lineage than single chain sequences alone. **(A)** Mutual information estimates (bits) were calculated using a finite-sampling correction to quantify the amount of information about T cell lineage by various TCR features drawn from the α, β, α and β summed together ($\alpha + \beta$), and paired $\alpha\beta$ repertoires. $\alpha + \beta$ sum represents expected mutual information if contributions from each chain are conditionally independent of T cell lineage. **(B)** A boosted tree classifier [XGBoost (40)] was trained using a constant length vector encoding V and J region usage, CDR3 length, and CDR3 amino acid usage frequencies. Receiver operating characteristic (ROC) curves showing true positive rate (TPR) vs. false positive rate (FPR) for the classifiers trained using the α and β single chain and the $\alpha\beta$ paired repertoire. To determine the expected accuracy of a classifier trained using independent α and β chains, we created a soft-voting classifier trained on each chain individually ($\alpha + \beta$). Inset shows prediction accuracy for classifiers trained using data from ECH repertoire. Shaded region represents one standard deviation. Boxplot inset show maximum accuracy in predicting CD4$^+$ or CD8$^+$ from each repertoire through 5-fold cross-validation repeated 10 times. Statistical significance between groups was assessed by Mann-Whitney *U* test (***$p < 0.001$).

were known to recognize peptides presented by MHC class I, with the remaining ~4% recognizing peptides presented in the context of an MHC class II molecule. We note that the antigen annotations provided for the curated VDJdb TCR sequences were obtained experimentally, most frequently through tetramer sorting assays (42).

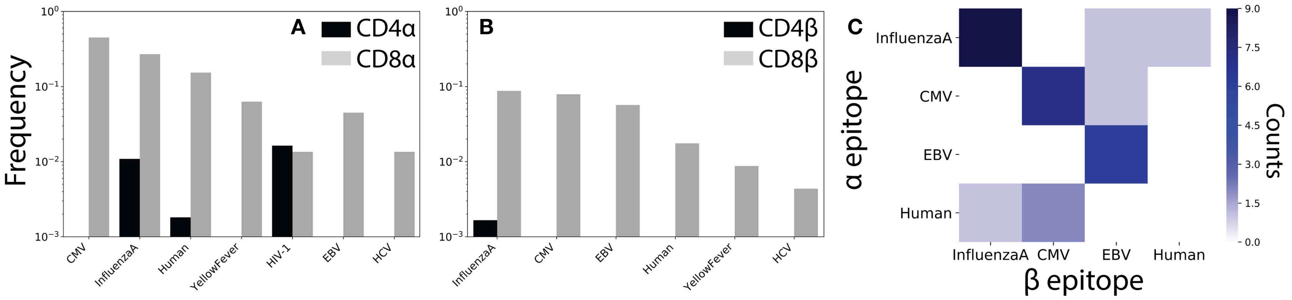
We first compared our single chain CD4$^+$ and CD8$^+$ TCR repertoires against these known sequences, using clonotypes composed of the V region plus amino acid CDR3 sequence, reporting the fraction of each repertoire with known antigen annotations (**Figures 5A,B**). In total, we identified 287 α and β TCR sequences with experimental antigen specificity annotations, of which 17 (~5%) were found in the CD4$^+$ repertoire (in line with the 4% of VDJdb annotations corresponding to MHC II restricted epitopes). Of these sequences, ~80% of α and β chains were associated with highly prevalent viral infections (Cytomegalovirus, Epstein-Barr virus, Influenza A) to which public TCR clones have previously been observed in otherwise healthy individuals (43). Interestingly, the remaining 20% of annotated sequences recognized epitopes that should not be present in our healthy cohort (e.g., Yellow Fever, HIV, and Hepatitis C). Of note, this result further demonstrates the ability of single-cell sequencing (17, 18) approaches to capture large numbers of TCRs which have previously been observed using bulk-sequencing methodologies.

To better understand how analysis of $\alpha\beta$ paired TCR sequences would influence our ability to understand TCR antigen specificity and repertoire-level function, we next asked which of our $\alpha\beta$ pairs had known peptide specificities for both the α and β chains individually. We observed 1 CD4$^+$ and 28 CD8$^+$ TCR pairs for which for which both chains had known antigen specificities. Of these, 6 (~21%) TCR pairs recognized epitopes from different species and an additional 2 (~7%) pairs recognized different epitopes from the same species (**Figures 5C,D**). In contrast to the single-chain repertoires, all TCR pairs with matching antigen specificities recognized a viral antigen expected to be found in healthy individuals (**Figure 5C**). We additionally note several non-monogamous $\alpha\beta$ pairings in which the same α chain is paired with β chains recognizing different antigens. For example, the α sequence Vα1-2 CAVMDSSYKLIF has previously been shown to recognize a human Bone Marrow Stromal Cell Antigen 2 (BST2) epitope and is here shown to pair with β sequences that have been shown to interact with both Influenza A and CMV epitopes (**Figure 5D**).

While promiscuity in TCR pairing has been widely reported (19, 26, 44, 45), these results serve to further emphasize the functional importance of $\alpha\beta$ pairing in determining antigen specificity. That is, our findings clearly demonstrate the ability of an identical TCR sequence to recognize differing pMHC complexes depending on its pair. Given recent efforts to infer antigen exposure history from the β chain repertoire (24), the prevalence of antigen false-positives observed in the single chain repertoires (i.e., TCRs recognizing antigens not present in healthy donors) may be of particular relevance. Further, these findings demonstrate that even limited sequencing of the paired $\alpha\beta$ repertoire may be able to provide accurate information about previous antigen exposure and repertoire function.

DISCUSSION

Although the theoretical importance of $\alpha\beta$ pairing is not debated, the actual amount of functional information which can be extracted from repertoire level analyses of $\alpha\beta$ TCR pairs remains uncertain. In this study, we have contributed more than 11,000 unique $\alpha\beta$ paired sequences to our previously published database, providing us with nearly 100,000 unique TCR pairs



D

Lineage	(Vα) CDR3α	(Vβ) CDR3β	α/β epitope species	α/β epitope gene
CD4	(20) CAVQYGGSQGNLIF	(13) CASTRQNSNQPQHF	InfluenzaA / InfluenzaA	HA/HA
CD8	(26-2) CILSNNNDMRF	(30) CAWSVSDVANTEAFF	CMV / CMV	p65 / p65
CD8	(26-2) CILSNNNDMRF	(30) CAWSISDLAKNIQYF	CMV / CMV	p65 / p65
CD8	(3) CAVRGTNARLMF	(12-4) CASSSVTEAFF	CMV / CMV	p65 / p65
CD8	*(17) CATGTGANNLFF*	*(5-1) CASSLDSQSSGNTIYF*	*CMV / CMV*	*p65 / IE1*
CD8	(24) CAFPYNNNDMRF	(27) CASSLEGFTEAFF	CMV / CMV	p65 / p65
CD8	*(24) CARNTGNQFYF*	*(5-1) CASSLDSQSSGNTIYF*	*CMV / CMV*	*p65 / IE1*
CD8	(24) CAFPYNNNDMRF	(27) CASSLEGYTEAFF	CMV / CMV	p65 / p65
CD8	**(24) CARNTGNQFYF**	**(20-1) CSARDRTGNGYTF**	**CMV / EBV**	**p65 / BMLF1**
CD8	(20) CAVQDLGTSGSRLTF	(9) CASSARSGELFF	EBV / EBV	EBNA1 / EBNA1
CD8	(5) CAEDNNARLMF	(20-1) CSARDRTGNGYTF	EBV / EBV	BMLF1 / BMLF1
CD8	(5) CAEDNNARLMF	(20-1) CSARDGTGNGYTF	EBV / EBV	BMLF1 / BMLF1
CD8	(5) CAEYSSASKIIF	(14) CASSQSPGGTQYF	EBV / EBV	BMLF1 / BMLF1
CD8	(5) CADDNNARLMF	(20-1) CSARDGTGNGYTF	EBV / EBV	BMLF1 / BMLF1
CD8	(5) CAESIGKLIF	(29-1) CSVGTGGTNEKLFF	EBV / EBV	BMLF1 / BMLF1
CD8	**(1-2) CAVMDSSYKLIF**	**(6-1) CASSELAGSYNEQFF**	**HomoSapiens / CMV**	**BST2 / p65**
CD8	**(1-2) CAVMDSSYKLIF**	**(5-1) CASSLDSQSSGNTIYF**	**HomoSapiens / CMV**	**BST2 / IE1**
CD8	**(1-2) CAVMDSSYKLIF**	**(6-4) CASSDSAGGTDTQYF**	**HomoSapiens / InfluenzaA**	**BST2 / M1**
CD8	**(21) CAVNSGGGADGLTF**	**(9) CASSARTGELFF**	**InfluenzaA / EBV**	**M1 / EBNA1**
CD8	**(1-2) CAVLDSSYKLIF**	**(6-1) CASSELAGGQETQYF**	**InfluenzaA / HomoSapiens**	**M1 / MLANA**
CD8	(27) CAGGDGGSQGNLIF	(19) CASSSRSTGELFF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(38-1) CAFMMNAGGTSYGKLTF	(19) CASSIGVYGYTF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(25) CAGNYGGSQGNLIF	(19) CASSTRSTGELFF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(25) CAGTYGGSQGNLIF	(19) CASSPRSTDTQYF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(27) CAGASGSSNTGKLIF	(19) CASSIRSAYEQYF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(27) CAGAVGSSNTGKLIF	(19) CASSIRSSYEQYF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(27) CAGGGSQGNLIF	(19) CASSIRSSYEQYF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(27) CAGGGSQGNLIF	(19) CASSVRSSYEQYF	InfluenzaA / InfluenzaA	M1 / M1
CD8	(13-2) CAENWGGGSQGNLIF	(19) CASSSRATGELFF	InfluenzaA / InfluenzaA	M1 / M1

FIGURE 5 | $\alpha\beta$ pairing provides additional information about antigen specificity. **(A)** The VDJdb (42) database of TCRs with known peptide specificity was compared to the $CD4^+$ and $CD8^+$ single-chain repertoires for the α and **(B)** β chains. **(C)** Antigen specificities for paired $\alpha\beta$ TCR sequences. Counts along the diagonal represent pairs with matching specificity while those off the diagonal represent mismatched pairs. **(D)** Table shows CDR3 sequence and antigen specificity for all $\alpha\beta$ pairs for which annotations were available for both chains. TCR pairs for which the α and β chains recognize epitopes from different species are shown in bold, while those recognizing different epitopes from the same species are shown in italics. CMV, cytomegalovirus; EBV, Epstein-Barr virus; HIV, human immunodeficiency virus; HCV, hepatitis C virus; HTLV, human T-lymphotropic virus; DENV, Dengue virus.

split between the $CD4^+$ and $CD8^+$ T cell lineages. To better understand how high-throughput examination of $\alpha\beta$ pairing can inform on repertoire function, we chose to focus on (i) how TCR pairing might influence MHC recognition and subsequently inform on biases between the $CD4^+$ and $CD8^+$ repertoires, and (ii) how TCR pairing might provide additional information on the antigen specificity of the TCR repertoire.

A growing number of studies have begun to elucidate a number of molecular interactions conserved between multiple structures leading to the hypothesis that such interaction motifs have been evolutionarily incorporated into the germline $V\alpha\beta$ sequences (32, 33). Although these conclusions are primarily based on a limited number of solved TCR-pMHC structures, bulk-sequencing of the β chain has revealed statistical associations between features of the TCR repertoire and individual MHC polymorphisms (29). However, previous studies have not differentiated between the $CD4^+$ and $CD8^+$ paired TCR repertoires. It was therefore unknown whether $\alpha\beta$ pairing could influence the effects of these germline biases. Our analysis of the healthy $CD4^+$ and $CD8^+$ TCR repertoires revealed that while individual α and β chains were more commonly found in both repertoires, paired $\alpha\beta$ sequences tended to be specific for one lineage. As has been previously suggested for β chains (28), we found that sequences shared between the two cell lineages tended to be shorter and have higher generation probabilities (i.e., are closer to the germline sequences) than those found only in one repertoire. Together, these results suggested that a large portion of paired $\alpha\beta$ sequences were relatively specific for one MHC class and supported previous findings of systematic differences between the $CD4^+$ and $CD8^+$ repertoires (22–24).

Comparing the α and β single-chain repertoires between the $CD4^+$ and $CD8^+$ expectedly revealed that V and J germline region usage, as well as CDR3 charge and length distributions, differed between the two repertoires. Consistent with previous small-scale structural findings (7–10), our results demonstrate that $\alpha\beta$ pairings encode substantially stronger associations with T cell lineage than either of the single-chain repertoires alone. Rigorously quantifying the strength of these associations using mutual information and machine learning classifiers, our results showed that the majority of information about T cell lineage carried by TCRs is encoded by the V germline region, with significantly less information present in the J region, CDR3 charge, and CDR3 length. Though the total amount of information remained relatively low, these methods revealed substantial synergy between the α and β chains with respect to lineage association. To the best of our knowledge, such synergy between TCR chains, particularly for V$\alpha\beta$ pairs, has not been previously demonstrated at the repertoire level. Biologically, one possible explanation of these findings is a model in which Vα and Vβ chains evolved to, in concert, bias TCRs toward interaction with either of the MHC classes. Future studies employing a substantially larger cohort will be necessary to further unravel the relationship between specific TCR features and HLA-types, as well as specific MHC polymorphisms.

Finally, given the observed importance of $\alpha\beta$ pairing in driving MHC specificity, we asked whether TCR pairings could similarly influence peptide specificity. Toward this, we annotated our TCR sequences using the antigen specificity information contained within the VDJdb sequence repository (42). We found that approximately one third of our TCR pairs for which both chains had known antigen specificities were mismatched (i.e., had different known antigen specificities for the α and β chain). Intriguingly, TCR pairs recognizing the same antigen individually were always associated with common viral peptides that would be expected to be present in otherwise healthy individuals. Conversely, TCR pairs with different antigen specificities tended to recognize viral peptides that are not found in healthy individuals, suggesting that the antigen specificity of single chains is largely dependent upon its pair. This result is consistent with previous findings from bulk-sequencing of the β chain in CMV patients, in which even CMV seronegative patients were found to have low levels of CMV-associated TCRs (24). While this study was largely successful in predicting whether an individual was infected with CMV from the TCRβ repertoire, it required the use of a large number of TCR sequences with known CMV associations. Given the demonstrated increase in antigen specificity information contained within $\alpha\beta$ pairing, we hypothesize that the increased availability of such single-cell approaches may ultimately increase diagnostic efficiency and accuracy.

In summary, we have generated and comprehensively analyzed the largest database of CD4$^+$ and CD8$^+$ paired $\alpha\beta$ TCR sequences to date using recently developed high-throughput single-cell technologies. While such single-cell methods remain cost-prohibitive for large cohort studies, we have demonstrated the ability of current paired $\alpha\beta$ sequencing to provide useful insights into TCR repertoire function beyond those available from conventional bulk-sequencing. Biologically, our results have shown substantial synergistic information about T cell lineage encoded within TCR pairings and suggested the utility of $\alpha\beta$ pairings when determining antigen specificities for an individual's TCR repertoire. Together, our results demonstrate the power of paired $\alpha\beta$ sequencing to inform on repertoire function and suggest that current paired $\alpha\beta$ repertoire sequencing are capable of opening new avenues of research when use in conjunction with TCRβ sequencing. We further believe that the rigorous examination of the normal $\alpha\beta$ TCR repertoires presented in this study will prove to be valuable in understanding the perturbations caused by infectious, oncological and autoimmune disease states.

MATERIALS AND METHODS

Single-Cell Barcoding and Sequencing

TCR sequences for Subjects 1-5, along with each patient's HLA type, were obtained from Grigaityte et al. (19). As described previously, peripheral blood mononuclear cells (PBMCs) were obtained from five healthy donors after obtaining appropriate informed consent. Blood samples underwent pan T cell enrichment before single-cell barcoding-in-emulsion using the AbVitro microfluidic platform (18). In brief, single-cell sequencing was performed by probabilistically loading individual T cells into ∼65 picoliter oil-emulsion droplets and TCR-targeted reverse-transcriptase PCR is performed. Unique droplet barcodes, along with unique molecular identifier (UMI) barcodes, are similarly loaded into droplets and attached to TCR cDNA within each droplet. Droplets are then lysed and next-generation sequencing performed on the pooled product using the Illumina MiSeq platform (18). Raw sequences were processed using a custom pipeline (19) to identify $\alpha\beta$ pairs utilizing MiXCR 2.2.1 (46) to identify V(D)J segments and annotate the CDR3 region of each TCR. As described in detail previously (19), the quality of TCR pairs were ensured by setting a minimum read depth for including a given TCR sequence and collapsing reads from a single droplet with a nucleotide CDR3 Hamming distance of 1. We excluded droplets with more than one unique α or β chain given that we cannot readily differentiate droplets with an allelic inclusion T cell from those containing two different T cells.

All TCR sequences for Subject 6 and 7, as well as for a subset of Subjects 1 and 3, were obtained using the 10× Genomics commercial single-cell sequencing platform (17). PBMCs for Subjects 6 and 7 were purchased from ATCC (PCS-800-011TM). CD4$^+$ and CD8$^+$ T cell populations were separated using either magnetic bead enrichment according to the manufacturer protocol (EasySep Human T Cell Enrichment Kit, StemCell Technologies) or fluorescence activated cell sorting (Becton Dickinson FACSARIA SORP). Following the manufacturer's instructions, ∼5,000 cells per lane were loaded into the Chromium Controller using the Single Cell V(D)J reagent kit for emulsion-barcoding (17) and sequenced using an Illumina HiSeq 2500 sequencer. Raw sequencing reads were processed as described above (19). PBMC samples for Subjects 1 and 3 used for sequencing on the 10× platform were frozen and stored for several months, potentially leading to RNA degradation and resulting in the low number of captured sequences.

The Li et al. dataset (23) was provided by N. P. Weng as a processed datafile containing VJ segments and CDR3 amino acid sequences. The Emerson et al. dataset (24) was downloaded from Adaptive Biotechnologies open-access immuneACCESS database (https://clients.adaptivebiotech.com/immuneaccess). While healthy and diseased TCR repertoires were obtained, only the 17 healthy patients were studied here.

Data Analysis

Paired $\alpha\beta$ TCR sequences, along with clonotype information about V(D)J segment use and CDR3 amino acid sequences, were divided into CD4$^+$ and CD8$^+$ repertoires. T cells lacking a lineage designation or expressing two unique TCRs (i.e., dual receptor T cells) were excluded from subsequent analysis. As we care about identifying features of the TCR repertoires between the CD4$^+$ and CD8$^+$ populations, we count each unique TCR clonotype only once. That is, clonal expansion in the CD4$^+$ and CD8$^+$ populations would bias our analysis of the factors that affect differentiation. As such, we include each TCR clonotype only once into our final dataset. We then identified TCR clonotypes that were shared between the CD4$^+$ and CD8$^+$ compartments and the degree of overlap between the two TCR repertoires was quantified using the Jaccard Index (J):

$$J(CD4, CD8) = \frac{|CD4 \cap CD8|}{|CD4 \cup CD8|} \quad (1)$$

Here $|CD4 \cap CD8|$ refers to the cardinality of the intersection between the CD4$^+$ and CD8$^+$ TCR repertoires (i.e., the number of TCRs found in both repertoires). $|CD4 \cup CD8|$ refers to the union of the two repertoires (i.e., the number of TCRs found in either of the two repertoires). The Jaccard Index was calculated independently for the α ($J(CD4_\alpha, CD8_\alpha)$), β ($J(CD4_\beta, CD8_\beta)$), and $\alpha\beta$ ($J(CD4_{\alpha\beta}, CD8_{\alpha\beta})$) TCR repertoires.

Furthermore, as done previously (19), the paired $\alpha\beta$ repertoire consists of all unique, paired TCR sequences and the α and β individual chain repertoires were derived directly from the paired repertoire. That is, the individual α repertoire consists of all the α chains present in the paired dataset. Thus, the α, β, and $\alpha\beta$ datasets are all of the same size and differences in sample size do not drive the observed differences. Furthermore, all boxplots represent median and inter-quartile range.

VJ Segment Usage

V(D)J segments were identified from raw sequences by MiXCR and annotated according to the International ImMunoGeneTics (IMGT) V(D)J gene definitions (47). The odds ratio (OR) for a given TCR characteristic and T cell lineage was calculated by counting the number of TCRs with (C^+) and without (C^-) that characteristic within the CD4$^+$ (T^4) and CD8$^+$ (T^8) repertoires. The OR is then given as:

$$OR = \frac{|C^+ \in T^4| * |C^- \in T^8|}{|C^- \in T^4| * |C^+ \in T^8|} \quad (2)$$

The numerator is the number of CD4$^+$ TCRs with a given feature multiplied by the number of CD8$^+$ TCRs without that feature. The denominator is given by the number of CD4$^+$ cells without that feature multiplied by the number of CD8$^+$ with that feature. 95% confidence intervals and a p-value were then calculated for each OR using Fisher's exact test implemented using the SciPy library (www.scipy.org). Multiple hypothesis testing correction was applied to single chain p-values using a Bonferroni correction and paired chains p-values, given the larger number of tested hypotheses, were converted to q-values (48). Significance was assessed at the $p < 0.05$ or $q < 0.05$ level.

CDR3 Features

Sequence logos showing the amino acid frequency for a given position in the sequence were generated using all α and β CDR3 sequences of length 14 using WebLogo (49). Of note, we defined the CDR3 length to be inclusive of the proximal cysteine and terminal phenylalanine that define the CDR3 region. The ratio of each amino acid in CDR3 between each population was calculated by dividing the frequency of a given amino acid across all CD4$^+$ CDR3 sequences for a given chain by the frequency with which that amino acid occurred across all CD8$^+$ CDR3 sequences. CDR3 charge was calculated as difference between the number of positively charged amino acids (R and K) and negatively charged amino acids (D and E) present in the CDR3 region.

Mutual Information

The mutual information (I, bits), between a given feature, X, and T cell lineage (L) was calculated as:

$$I(X; L) = \sum_{x \epsilon X} \sum_{l \epsilon L} p(x, l) \log_2 \left(\frac{p(x, l)}{p(x)p(l)} \right) \quad (3)$$

In order to correct for biases in our MI estimate arising from our limited sample sizes, we applied a bootstrapping based finite-sampling correction previously described (19, 50). We additionally calculate the synergistic information (41) (S) according to:

$$S(X_\alpha, X_\beta, L) = I(X_\alpha, X_\beta; L) - I(X_\alpha; L) - I(X_\beta; L) \quad (4)$$

where X_α and X_β refer to TCRα and TCRβ features, respectively.

Machine Learning

Extreme Gradient Boosted decision tree classifiers were trained using the Python XGBoost implementation (40). α and β chain TCR sequences were converted into length-independent vectors encompassing V and J regions (categorically encoded), CDR3 length, CDR3 charge, and CDR3 amino acid usage frequencies. Only the unique set of TCRs were used for training and testing, and TCR sequences found in both the CD4$^+$ and CD8$^+$ repertoires were removed. Classifiers were trained and tested using 5-fold cross-validation, which was repeated 10 times, for the each of the α, β, or $\alpha\beta$ repertoires. The independent $\alpha + \beta$ classifier was an ensemble classifier created by training XGBoost classifiers on each chain independently. Final predictions for this ensemble were made using a soft-voting approach. Receiver operating characteristic curves (ROC), as well as the area under the ROC curve (AUC) and classifier accuracy, were calculate using the sklearn metrics package (51). ROC curves and AUC values were calculated using the predicted probability of a given TCR chain belonging to the CD4$^+$ population. Accuracy was calculated as the percentage of correct predictions divided by the total number of predictions made, where a TCR was predicted to be CD4$^+$ sequence if the predicted CD4$^+$ probability was >50%. For SVM's trained on the Li et al. and Emerson et al. dataset, CDR3β amino acid sequences were first converted in numeric vectors using Atchley factors (23, 52).

DATA AVAILABILITY

Sequencing data and custom Python scripts used for data analysis are freely available at our Github Repository (https://github.com/JasonACarter/CD4_CD8-Manuscript).

AUTHOR CONTRIBUTIONS

JC and GA contributed to the conception and design of the study. JC, JP, KG, SG, and EJ performed research with supervision from AB, FV, and GA. JC and GA analyzed data and wrote the paper. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

JC was partially supported by NIHGM MSTP Training award T32-GM008444. KG was funded by the Ferish-Gerry fellowship from the Watson School of Biological Sciences. GA was funded by the Simons Foundation, a LIBH grant, and the Stand Up To Cancer-Breast Cancer Research Foundation Convergence Team Translational Cancer Research Grant, Grant Number SU2C-BCRF 2015-001.

ACKNOWLEDGMENTS

The authors thank Doug Fearon and Ronald Hause for comments on the manuscript, Pamela Moody and the CSHL Flow Cytometry Shared Resource for help with FACS experiments, and the CSHL DNA Sequencing Core for next-generation sequencing. We additionally thank N. P. Weng for providing the β chain bulk sequencing dataset from Li et al. (23).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2019.01516/full#supplementary-material

Supplemental Figure 1 | CDR3 sequences shared between the CD4$^+$ and CD8$^+$ repertoires tend to be shorter than those found in only one repertoire. CDR3 length distributions show sequences found in both the CD4$^+$ and CD8$^+$ repertoires (∩) are shorter than those found in only one of the two repertoires (⊕) for the **(A)** α, **(B)** β, and **(C)** paired $\alpha\beta$ repertoires. For paired sequences, we report the average length of the α and β chains. **(D)** Heatmaps showing frequency with which each α and β CDR3 length pair is present in the TCR repertoire shared between the CD4$^+$ and CD8$^+$ lineages and for the **(E)** TCR repertoire present in only one of the two lineages. Dashed red lines indicate the average length for the α (14 amino acids) and β chains (15 amino acids).

Supplemental Figure 2 | J germline region bias for the α, β, and $\alpha\beta$ repertoires. **(A)** The CD4$^+$ and CD8$^+$ TCR repertoires were then pooled across individuals and the CD4$^+$:CD8$^+$ odds ratio (OR) was calculated for each Jα and **(B)** Jβ single-chain germline region. An OR> 1 represents a CD4$^+$ bias, while an OR< 1 represents a CD8$^+$ bias with error bars representing the 95% confidence interval. The mean is represented by a red or black dot, with red representing statistical significance at the p<0.05 by Fisher's exact test level after applying Bonferroni correction. **(C)** Significant ($q < 0.05$ by Fisher's exact test) log odds ratios reveals strong CD4$^+$:CD8$^+$ biases for 79 J$\alpha\beta$ pairs. **(D)** Boxplots were calculated for the set of all significant odds ratios associated with single chains (Jα or Jβ) and compared with those associated with J$\alpha\beta$ pairs. Paired associations for both CD4$^+$ and **(E)** CD8$^+$ status were significantly stronger (***$p < 0.001$ by Mann-Whitney U test) than those associated with a single chain alone. Associations for the J region were, overall, substantially weaker than those observed for the V chain.

Supplemental Figure 3 | V and J germline region usage. **(A)** Single-chain V region distributions for the α and **(B)** β chains. **(C)** Paired V$\alpha\beta$ usage for the CD4$^+$ and **(D)** CD8$^+$ T cell populations. **(E)** Single-chain J region distributions for the α and **(F)** β chains. **(G)** Paired V$\alpha\beta$ usage for the CD4$^+$ and **(H)** CD8$^+$ T cell populations.

Supplemental Figure 4 | CDR3 charge and length distributions. **(A)** Single-chain CDR3 charge for the α and **(B)** β chains, separated by CD4$^+$ and CD8$^+$ populations. **(C)** Paired CDR3$\alpha\beta$ charge usage for the CD4$^+$ and **(D)** CD8$^+$ T cell populations. **(E)** Single-chain CDR3 length for the α and **(F)** β chains, separated by CD4$^+$ and CD8$^+$ populations. **(G)** Paired CDR3$\alpha\beta$ length distributions for the CD4$^+$ and **(H)** CD8$^+$ T cell populations.

Supplemental Figure 5 | SVM trained on CDR3β sequences converted to Atchley factors. A support vector machine (SVM) was trained on vectors composed of CDR3β sequences converted into numerical array according to their Atchley factors. As these vectors are dependent on the length of the CDR3 sequence, SVMs were trained separately for CDR3 sequences of lengths between 10 and 15, as previously done (23). For comparison, SVM accuracy for classifiers trained on CDR3β sequences converted to our constant length vector are also shown (Constant). **(A)** Accuracy for each model is reported as the percentage of correctly predicted CDR3 sequences using an independent testing set (25% of dataset). The Li et al. dataset is well-described by this SVM model, with accuracy as high as 96%. However, this model fails to accurately describe either the dataset used in this study or that of Emerson et al. **(B)** Receiver operator curves (ROC) for the current dataset, **(C)** the Emerson et al. dataset, and **(D)** the Li et al. dataset show length-dependent SVMs accurately predict the Li et al. dataset, but fail to do so for the other two datasets.

Supplemental Figure 6 | V and J region usage patterns vary substantially between the Li et al. and Emerson et al. datasets. **(A)** β TCR sequences were obtained from 621,085 CD4$^+$ and 64,725 CD8$^+$ cells previously by Li et al. (23). Comparison of V-usage frequencies for each germline region reveals large differences between the CD4$^+$ and CD8$^+$ repertoires in this dataset. **(B)** V-usage frequencies observed by comparing 3,212,682 CD4$^+$ and 1,774,260 CD8$^+$ TCR sequences taken from Emerson et al. reveal less variation between the two cell types (24) and more closely resemble the results obtained in the present study (**Supplemental Figure 3**). **(C)** Similar results were obtained for Jβ region usage in the Li et al. and **(D)** Emerson et al. datasets. **(E)** We quantified the difference in V segment use in the CD4$^+$ and CD8$^+$ populations by calculating the odds ratio (OR) for each V region in the Li et al. dataset and **(F)** the Emerson et al. dataset independently. **(G)** Jβ usage between Li et al. dataset and **(H)** Emerson et al. datasets. **(I)** Mutual information with finite sampling correction was calculated for the association between β chain features (Vβ, Jβ, CDR3β length and charge) and lineage for the dataset used in this study (from Table), by Li et al. (23) and Emerson et al. (24). Substantially higher mutual information values, indicating stronger associations, were found for the Li et al. dataset as compared to the other two datasets.

Supplemental Table 1 | Demographic information for each subject. Peripheral blood mononuclear cells (PBMCs) were previously obtained from 5 healthy individuals (S1–S5) and sequenced using single-cell barcoding in emulsion (18, 19). The original PBMC samples from S1 and S3, as well as new samples from additional healthy individuals (S6, S7), were sequenced using a commercially available single-cell system (10× Genomics) (17). In all, we obtain 70,108 and 26,946 unique TCR pairs from CD4$^+$ CD8$^+$ T cells, respectively. Demographic information, as well as HLA types, are provided as available (19).

REFERENCES

1. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science* (1999) 2886:958–61. doi: 10.1126/science.286.5441.958
2. Miles JJ, Douek DC, Price DA. Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol.* (2011) 89:375–87. doi: 10.1038/icb.2010.139
3. Davis MM, Tato CM, Furman D. Systems immunology: just getting started. *Nat Immunol.* (2017) 18:725–32. doi: 10.1038/ni.3768
4. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood.* (2009) 114:4099–107. doi: 10.1182/blood-2009-04-217604
5. Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief Bioinform.* (2018) 19:554–65. doi: 10.1093/bib/bbw138
6. Hou D, Chen C, Seely EJ, Chen S, Song Y. High-throughput sequencing-based immune repertoire study during infectious disease. *Front Immunol.* (2016) 7:336. doi: 10.3389/fimmu.2016.00336
7. Marrack P, Krovi SH, Silberman D, White J, Kushnir E, Nakayama M, et al. The somatically generated portion of T cell receptor CDR3α contributes to the MHC allele specificity of the T cell receptor. *eLife.* (2017) 6:e30918. doi: 10.7554/eLife.30918
8. Stadinski BD, Trenh P, Smith RL, Bautista B, Huseby PG, Li G, et al. A role for differential variable gene pairing in creating T cell receptors specific for unique major histocompatibility ligands. *Immunity.* (2011) 35:694–704. doi: 10.1016/j.immuni.2011.10.012
9. Stadinski BD, Trenh P, Duke B, Huseby PG, Li G, Stern LJ, et al. Effect of CDR3 sequences and distal V gene residues in regulating TCR-MHC contacts and ligand specificity. *J Immunol.* (2014) 192:6071–82. doi: 10.4049/jimmunol.1303209
10. Yin L, Huseby E, Scott-Browne J, Rubtsova K, Pinilla C, Crawford F, et al. A single T cell receptor bound to major histocompatibility complex class I and class II glycoproteins reveals switchable TCR conformers. *Immunity.* (2011) 35:23–33. doi: 10.1016/j.immuni.2011.04.017
11. Simon MD, Rossetti G, Pagani M. Single cell T cell receptor sequencing: techniques and future challenges. *Front Immunol.* (2018) 9:1638. doi: 10.3389/fimmu.2018.01638
12. Dash P, McClaren JL, Oguin III TH, Rothwell W, Todd B, Morris MY, et al. Paired analysis of the TCRα and TCRβ chains at the single-cell level in mice. *J Clin Invest.* (2010) 121:288–95. doi: 10.1172/JCI44752
13. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol.* (2014) 32:684–92. doi: 10.1038/nbt.2938
14. Munson DJ, Egelston CA, Chiotti KE, Parra ZE, Bruno TC, Moore BL, et al. Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR. *Proc Natl Acad Sci USA.* (2016) 113:8272–7. doi: 10.1073/pnas.1606994113
15. Stubbington MJT, Lonnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods.* (2016) 13:329–32. doi: 10.1038/nmeth.3800
16. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med.* (2015) 7:301ra131. doi: 10.1126/scitranslmed.aac5624
17. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun.* (2017) 8:14049. doi: 10.1038/ncomms14049
18. Briggs AW, Goldfless SJ, Timberlake S, Belmont BJ, Clouser CR, Koppstein D, et al. Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion. *bioRxiv Preprint.* (2017). doi: 10.1101/134841
19. Grigaityte K, Carter JA, Goldfless SJ, Jeffery EW, Hause RJ, Jiang Y, et al. Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire. *bioRxiv.* (2017) 213462. doi: 10.1101/213462
20. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T cell antigen receptor recognition of antigen-presenting molecules. *Annu Rev Immunol.* (2015) 33:169–200. doi: 10.1146/annurev-immunol-032414-112334
21. La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. Understanding the drivers of MHC restriction of T cell receptors. *Nat Rev Immunol.* (2018) 18:467–78. doi: 10.1038/s41577-018-0007-5
22. Klarenbeek PL, Doorenspleet ME, Esveldt RE, van Schaik BDC, Lardy N, van Kampen AHC, et al. Somatic variation of T-cell receptor genes strongly associate with HLA class restriction. *PLoS ONE.* (2015) 10:e1040815. doi: 10.1371/journal.pone.0140815
23. Li HM, Hiroi T, Zhang Y, Shi A, Chen G, De S, et al. TCRβ repertoire of CD4$^+$ and CD8$^+$ T cells is distinct in richness, distribution and CDR3 amino acid composition. *J Leuk Biol.* (2016) 99:505–13. doi: 10.1189/jlb.6A0215-071RR
24. Emerson R, Sherwood A, Desmarais C, Malhotra S, Phippard D, Robins H. Estimating the ratio of CD4$^+$ to CD8$^+$ T cells using high-throughput sequence data. *J Immunol Methods.* (2013) 391:14–21. doi: 10.1016/j.jim.2013.02.002
25. Sethna Z, Elhanati Y, Callan CG Jr, Mora T, Walczak AM. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics.* (2019) btz035. doi: 10.1093/bioinformatics/btz035
26. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souqette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature.* (2017) 547:89–93. doi: 10.1038/nature22383
27. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature.* (2017) 547:94–8. doi: 10.1038/nature22976
28. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol.* (2011) 186:4285–94. doi: 10.4049/jimmunol.1003898
29. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet.* (2016) 48:995–1002. doi: 10.1038/ng.3625
30. DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen IV FA, Bradley P. Human T cell receptor occurence patterns encode immune history, genetic background, and receptor specificity. *eLife.* (2018) 7:e38358. doi: 10.7554/eLife.38358
31. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet.* (2017) 49:659–65. doi: 10.1038/ng.3822
32. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW. Evolutionarily conserved amino acids that control TCR-MHC interaction. *Annu Rev Immunol.* (2008) 26:171–203. doi: 10.1146/annurev.immunol.26.021607.090421
33. Garcia KC, Adams JJ, Feng D, Ely LK. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol.* (2009) 10:143–7. doi: 10.1038/ni.f.219
34. Huseby ES, White J, Crawford F, Vass T, Becker D, Pinilla C, et al. How the T cell repertoire becomes peptide and MHC specific. *Cell.* (2005) 122:247–60. doi: 10.1016/j.cell.2005.05.013
35. Feng D, Bond CJ, Ely LK, Maynard J, Garcia KC. Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'. *Nat Immunol.* (2007) 8:975–83. doi: 10.1038/ni1502
36. Dai S, Huseby ES, Rubtsova K, Scott-Browne J, Crawford F, Macdonald WA, et al. Crossreactive T cells spotlight the germline rules for $\alpha\beta$ T cell receptor interactions with MHC molecules. *Immunity.* (2008) 28:324–34. doi: 10.1016/j.immuni.2008.01.008
37. Scott-Browne JP, White J, Kappler JW, Gapin L, Marrack P. Germline-encoded amino acids in the $\alpha\beta$ T-cell receptor control thymic selection. *Nature.* (2009) 458:1043–6. doi: 10.1038/nature07812
38. Adams JJ, Narayanan S, Birnbaum ME, Sidhu SS, Blevins SJ, Gee MH, et al. Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nat Immunol.* (2016) 17:87–94. doi: 10.1038/ni.3310
39. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci USA.* (2014) 111:3354–9. doi: 10.1073/pnas.1309933111
40. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *arXiv.* (2016) 1603.02754. doi: 10.1145/2939672.2939785

41. Brenner N, Strong SP, Koberle R, Bialek W, de Ruyter van Steveninck RR. Synergy in a neural code. *Neural Comput.* (2000) 12:1531–52. doi: 10.1162/089976600300015259
42. Shugay M, Bagaev D, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specifcity. *Nucleic Acids Res.* (2018) 46:D419–27. doi: 10.1093/nar/gkx760
43. Chen G, Yang X, Ko A, Sun X, Gao M, Zhang Y, et al. Sequence and structural analyses reveal distinct and highly diverse human CD8$^+$ TCR repertoires to immunodominant viral antigens. *Cell Rep.* (2017) 19:569–83. doi: 10.1016/j.celrep.2017.03.072
44. Cukalac T, Kan WT, Dash P, Guan J, Quinn KM, Gras S, et al. Paired TCR$\alpha\beta$ analysis of virus-specific CD8$^+$ T cells exposes diveristy in a previously defined 'narrow' repertoire. *Immunol Cell Biol.* (2015) 93:804–14. doi: 10.1038/icb.2015.44
45. Lee ES, Thomas PG, Mold JE, Yates AJ. Identifying T cell receptors from high-throughput sequencing: dealing with promiscuity in TCRα and TCRβ pairing. *PLoS Comput Biol.* (2017) 13:e1005313. doi: 10.1371/journal.pcbi.1005313
46. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods.* (2015) 12:380–1. doi: 10.1038/nmeth.3364
47. Monod MY, Giudicelli V, Chaume D, Lefranc MP. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics.* (2004) 20:i379–85. doi: 10.1093/bioinformatics/bth945
48. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA.* (2003) 100:9440–5. doi: 10.1073/pnas.1530509100
49. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* (2004) 14:1188–90. doi: 10.1101/gr.849004
50. Strong SP, Koberle R, Ruyter van Steveninck RR, Bialek W. Entropy and information in neural spike trains. *Phys Rev Lett.* (1998) 80:197. doi: 10.1103/PhysRevLett.80.197
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* (2011) 12:2825–30. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
52. Atchley WR, Zhao J, Fernandes AD, Druke T. Solving the protein sequence metric problem. *Proc Natl Acad Sci USA.* (2005) 102:6395–400. doi: 10.1073/pnas.0408677102

Conflict of Interest Statement: SG, EJ, and AB are employed by Juno Therapeutics and hold equity in its parent company, Celgene. FV was formerly employed by Juno Therapeutics and is currently employed by and holds equity in Shape Therapeutics.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.