BMC Bioinformatics

RESEARCH ARTICLE Open Access

Balancing sensitivity and specificity in distinguishing TCR groups by CDR sequence similarity

Neerja Thakkar and Chris Bailey-Kellogg*

Abstract

Background: Repertoire sequencing is enabling deep explorations into the cellular immune response, including the characterization of commonalities and differences among T cell receptor (TCR) repertoires from different individuals, pathologies, and antigen specificities. In seeking to understand the generality of patterns observed in different groups of TCRs, it is necessary to balance how well each pattern represents the diversity among TCRs from one group (sensitivity) vs. how many TCRs from other groups it also represents (specificity). The variable complementarity determining regions (CDRs), particularly the third CDRs (CDR3s) interact with major histocompatibility complex (MHC)-presented epitopes from putative antigens, and thus encode the determinants of recognition.

Results: We here systematically characterize the predictive power that can be obtained from CDR3 sequences, using representative, readily interpretable methods for evaluating CDR sequence similarity and then clustering and classifying sequences based on similarity. An initial analysis of CDR3s of known structure, clustered by structural similarity, helps calibrate the limits of sequence diversity among CDRs that might have a common mode of interaction with presented epitopes. Subsequent analyses demonstrate that this same range of sequence similarity strikes a favorable specificity/sensitivity balance in distinguishing twins from non-twins based on overall CDR3 repertoires, classifying CDR3 repertoires by antigen specificity, and distinguishing general pathologies.

Conclusion: We conclude that within a fairly broad range of sequence similarity, matching CDR3 sequences are likely to share specificities.

Keywords: Antigen-specific recognition, CDR classification, Immune repertoire, Sequence similarity, T cell receptor

Background

The recognition by T cell receptors (TCRs) of non-self peptide epitopes presented by major histocompatibility complex (MHC) proteins drives the cellular immune response against the non-self offender. In the case of intracellular non-self peptides, e.g., infected or cancerous cells, the ternary MHC:peptide: TCR recognition can lead to the killing of abnormal cells presenting these peptides; in the case of extracellular non-self peptides, e.g., pathogens or biotherapeutics, it can lead to the development of a humoral response to neutralize or clear the antigens containing these peptides. Consequently, modeling and predicting MHC and TCR recognition propensities supports wide-ranging applications, for example: developing vaccines against infectious diseases [1–6] as well as understanding escape mechanisms [7–10], identifying cancer neoantigens and developing specifically targeted vaccines [11–14], discovering potential drivers of allergy, autoimmunity, and tolerance [15–18], and understanding and mitigating anti-biotherapeutic immune responses [19–27].

In the MHC:peptide: TCR recognition process (Fig. 1) the TCR represents the main source of variability and training in distinguishing of self vs. non-self peptides [28]. MHC is genetically encoded and even the effective diversity across global populations due to allelic variation is limited, since there is degeneracy in the MHC binding groove pockets that hold the peptide side-chains [29, 30]. In contrast, TCRs are much more diverse, with hypervariable complementarity determining regions (CDRs), particularly the third CDRs (CDR3s), which are

* Correspondence: cbk@cs.dartmouth.edu
Department of Computer Science, Dartmouth, Hanover, NH, USA

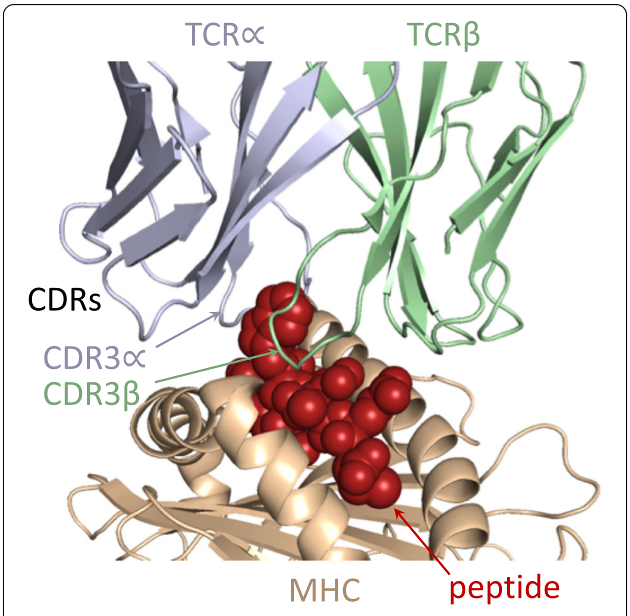

Fig. 1 MHC:peptide: TCR recognition. An MHC (brown) presents a peptide (red spheres) to a TCR (light blue: α chain; light green: β chain) (PDB id 2bnr [67], rendered via PyMOL [68]). The TCR interacts with the MHC and peptide through its CDR loops, with the first two CDRs of each TCR chain generally recognizing mainly the MHC and the third the peptide; here the CDR3α and CDR3β are "grabbing" a methionine and tryptophan in the middle of the peptide (inspired by Chen et al. [67])

derived from variable-diversity-joining (VDJ) recombination. Overall TCR theoretical diversity is estimated to be perhaps 10^{15} [31] and practical diversity in any individual roughly 10^6 [32–34]. The CDR3s, which contribute the bulk of the diversity, are thereby able to specifically recognize a wide array of MHC-presented antigen peptides, while the other CDRs, which are largely genetically encoded, are primarily responsible for recognizing the MHC itself (see again Fig. 1) [28, 35].

An individual's set of TCRs, or TCR repertoire, is shaped by thymic training against self along with a lifetime of exposure to different antigens. It is presumably much smaller than that of all possible antigens, and there is substantial degeneracy, with different TCRs able to recognize the same antigenic region and the same TCR able to accommodate different antigenic regions [36–39]. Thus numerous interesting and important questions center on the relationship between TCR repertoire and recognition propensities, including the impacts of genetics vs. training and exposure, the ability of CDRs to accommodate antigenic diversity, and commonalities across pathology- or antigen-specific populations. The advent of large-scale repertoire sequencing [40, 41], initially for CDRs alone [32, 42–44], and more recently even for paired α/β chains [45, 46], provides opportunities to gain insights into patterns of TCR diversity and recognition. An early study tackled the importance of genetics by studying pairs of monozygous twins, and, among other analyses, found that twin pairs had more identical CDR3 sequences than non-twins [47]. More recent publications have shifted the focus from identical sequences to similar sequences. Paired α/β TCRs, resulting from single-cell sequencing and antigen-specific selection, could be classified very well according to their epitope specificities, and the sequences elucidated patterns conferring those specificities [48]. Similarly, epitope-specific repertoires from a variety of viral infection contexts pooled across many subjects revealed distinctive motifs, and furthermore the CDRs from a set of *M. tuberculosis* subjects clustered into groups with strong MHC associations enabling design of specific MHC-peptide-TCR interactions [49]. An extensive analysis of available TCR sequence data from a wide range of subjects revealed pathogen-specific MHC-TCR associations along with structural insights into MHC and TCR covariation [50].

TCR repertoire sequencing thus provides the opportunity to generalize from a set of samples to patterns than are predictive of relationships among subjects, pathologies, antigens, MHC restrictions, and so forth [48, 50–52]. Here, we systematically investigate, over a diverse group of repertoire datasets, the extent to which sequence enables prediction of such relationships. As always in statistical / machine learning approaches, one must thread the needle between under-generalization, missing out on predictions that would be true (i.e., lacking sensitivity), and over-generalization, making predictions that end up not being true (i.e., lacking specificity). Thus we characterize the trade offs between specificity and sensitivity in these various studies. We show that in general it is possible to obtain a good specificity-sensitivity balance and make a large fraction of high-confidence predictions of TCR function from sequence.

Results

We study a diverse group of CDR datasets in order to evaluate in general how predictive TCR sequence is of relationships among groups (subjects, epitopes, pathologies, etc.). Since single cell methods are only now becoming available, many repertoire analysis efforts use standard sequencing approaches to characterize CDR3, which, as discussed above, is the main source of variability and antigen-specific recognition. So as to provide a consistent and interpretable basis for drawing conclusions, as well as to evaluate the information content provided by CDR3 alone, we use only CDR3 for all datasets, and separately analyze CDR3α and CDR3β.

The approach we take is representative of the key principles underlying the many possible sequence-based prediction methods and parameterizations, with advantages of being readily interpretable and able to function well even with limited data. In short, we evaluate CDR3 sequence similarity (technically dissimilarity, as lower is better), with a score which we term *CDRdist*, with 0 indicating an exact match and 1 no identifiable similarity. We assess the predictive power of CDR3 sequences by using this score to perform 1-nearest-neighbor classification (Fig. 2), predicting the group (e.g., associated antigen or disease) to which one CDR belongs based on the known group of the most similar CDR. By checking whether the predicted group is correct or not, we can evaluate both sensitivity (fraction of CDRs from a group that are correctly predicted to be in that group) and specificity (fraction of CDRs predicted to be in a group that actually are in that group). In order to evaluate only the impacts of similarity, without being confounded by duplicates which can render classification trivial, we do not consider exact matches (score of 0). In order to gain deeper insights into how the degree of similarity impacts classification performance, we slide a threshold from > 0 to 1, making a prediction only if the nearest neighbor is sufficiently similar, and assessing how relaxing the required score manifests in specificity vs. sensitivity trade-offs. The score also provides the basis for clustering, elucidating sequence similarity-based groupings for CDRs and revealing sequence patterns conferring the observed performance trade-offs.

In the following sections, we apply this general framework to characterize several datasets: CDR groups defined by structure, evaluating sequence variation within and between clusters; repertoires from twins, studying the relative similarity between an individual and their twin vs. others; a number of different human and murine repertoires, assessing CDR distinctiveness as it relates to antigen specificity as well as underlying pathology.

Extent of sequence similarity within/between structural clusters

The Structural T-Cell Receptor Database (STCRDab) [53] structurally clusters CDR3s into canonical classes, separately for α and β, and separately by length(s). A set of CDR3 sequences and associated structural classes were downloaded and investigated for relative intra- vs. inter-class sequence similarity. After removing duplicates, there were 142 unique sequences across four α groups and six β groups (Table 1). While it is common for structural clusters to be characterized by their individual sequence profiles, we also sought to understand the extent to which the sequence pattern from one cluster could be generalized before encroaching on another cluster.

The distance from each unique CDR to the most similar (but distinct) CDR within its structural class tends to be smaller than the distance to the most similar (but distinct) CDR from another class (Fig. 3 (a, b)). When the distance is less than about 0.2 or 0.3, the closest CDR tends to be within the same structural class (below 0.2: 97% in the same structural class; below 0.3: 96%), while when it is above about 0.6 or 0.7, the CDRs tends to be within different classes (above 0.6: 69% in different

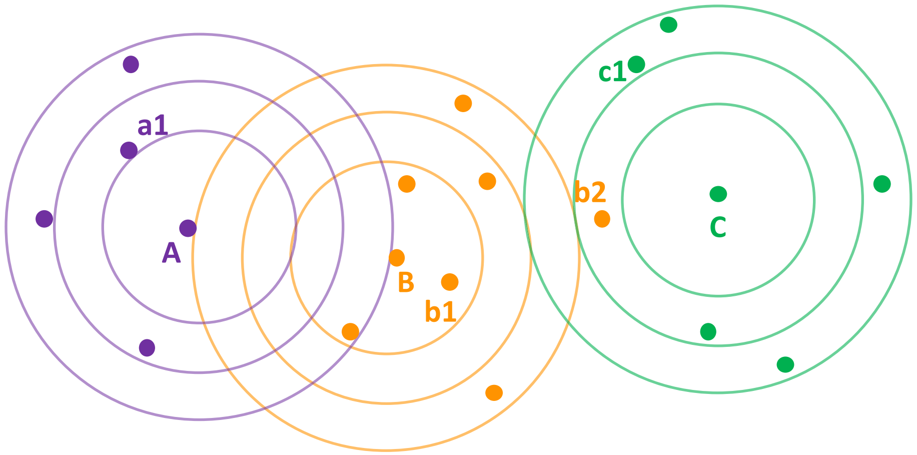

Fig. 2 Classifying CDRs by sequence similarity. This illustration plots in a schematic low-dimensional space the locations (dots) of CDRs from three different classes (colors). 1-nearest-neighbor classification predicts the class of one CDR from that of the most similar one; we here refine that to require the nearest neighbor to be close enough, within a specific distance threshold. Contour rings show sequence distances of 0.2, 0.3, and 0.4 from three query CDRs ("**A**", "**B**", and "**C**") from those classes. At a threshold of 0.2, only "**B**" has a close-enough nearest neighbor, "b1", which is of the same class, so 1-nearest-neighbor classification is correct. At this threshold, "**A**" and "**C**" are not predicted. When the threshold is relaxed to 0.3, "**A**" now has a close-enough nearest neighbor, "a1", of the same class, so it is also correctly predicted. However, "**C**" has "b2", of a different class, as its close-enough nearest neighbor, so it is incorrectly predicted. In this manner, we study trends trading off correct, incorrect, and unidentified, as the threshold is varied

Table 1 Unique CDRs from different structural classes according to STCRDab [53]

Class name	Length(s)	Number of unique sequences
A3–10-A	10	14
A3–13-B	13	3
A3–13-A	13	6
A3–10_11_12-A	10	10
B3–10_11_12_13-A	10, 11, 12, 13	56
B3–10_11-A	10, 11	20
B3–12-A	12	6
B3–12-B	12	10
B3–13_14-A	13, 14	14
B3–14-A	14	3

structural classes; above 0.7: 71%). This observation supports the use of nearest-neighbor classification, predicting structural class from sequence matches, which we elaborate to study specificity-sensitivity trade-offs by subjecting it to a distance threshold; i.e., only make a prediction if the nearest neighbor is closer than a given threshold. When the threshold is less than about 0.2, many CDRs are unclassified but those that are tend to be correct; between 0.2 and 0.4, the number of unclassified CDRs drops substantially while still retaining high specificity; and above that range the specificity drops in order to obtain further sensitivity (Fig. 3 (c, d)).

In order to more directly explore the relationship between sequence similarity and structural similarity, structures were downloaded from STCRDab and the CDR3β loops extracted for those in which electron density was present. When the same sequence was present in multiple structures, a representative was chosen as that with minimal sum of main-chain root mean squared deviation (RMSD) to the others. For each such CDR, the most similar sequence in its STCRDab CDR3β structural class and the most similar sequence from another CDR3β class were compared, in terms of both *CDRdist* and RMSD. This comparison (Fig. 4) thus elaborates the implications of Fig. 3, characterizing when a closer sequence implies a closer structure. Some examples are illustrated, limited to cases with good sequence score, such that a classification decision would be made under the thresholding approach above.

This analysis helps calibrate the general level of confidence one can have that two CDRs of a given degree of

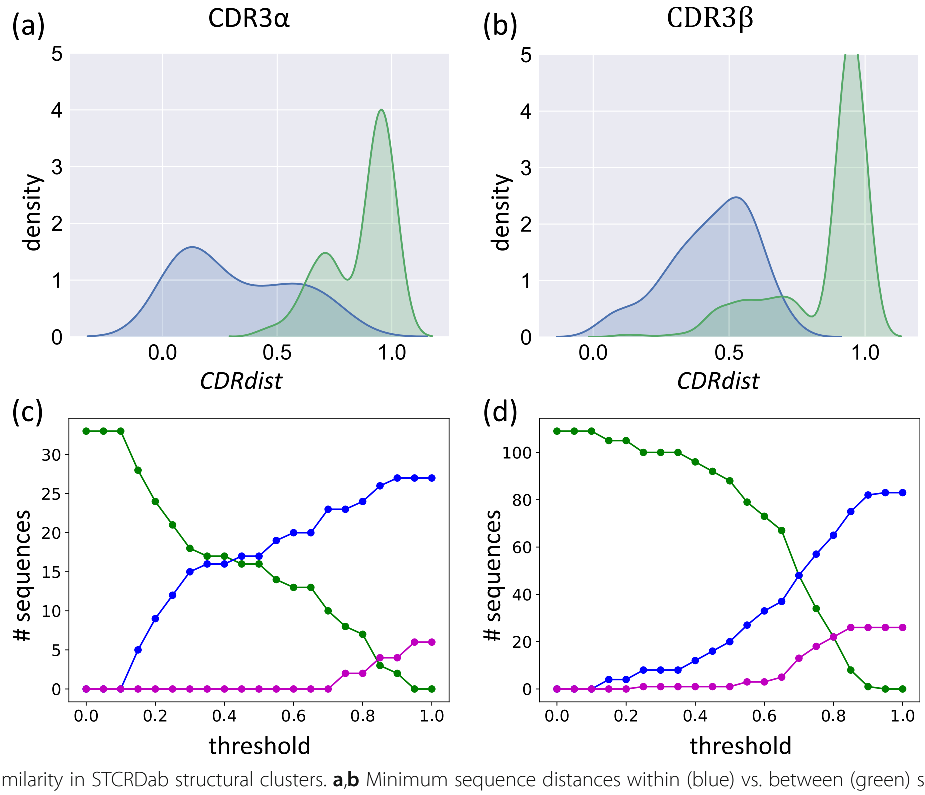

Fig. 3 Sequence similarity in STCRDab structural clusters. **a**,**b** Minimum sequence distances within (blue) vs. between (green) structural classes for (**a**) CDR3α sequences and (**b**) CDR3β sequences, plotted as a density estimate. **c**, **d** Accuracy of nearest-neighbor classification using sequence similarity to predict structural cluster for (**c**) CDR3α sequences and (**d**) CDR3β sequences. As the threshold required to make a classification is increased (*x*-axis), the number of sequences (*y*-axis) that are unclassified (green line) decreases, trading off how many are correctly (blue line) vs. incorrectly (magenta) classified

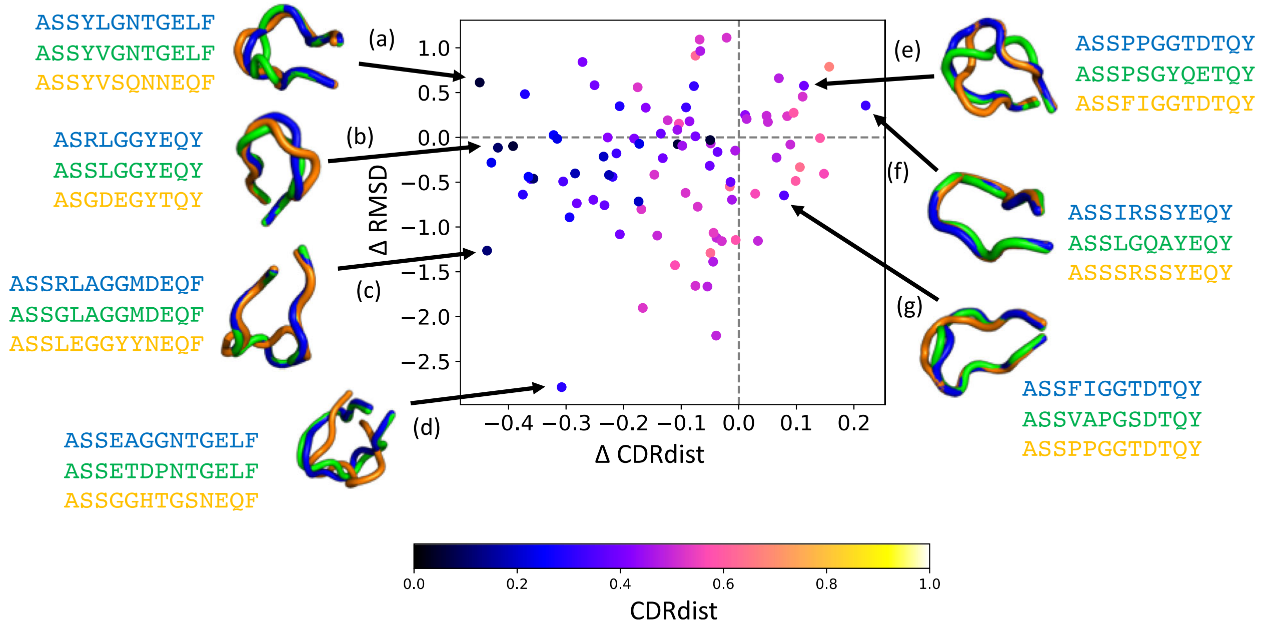

Fig. 4 Sequence similarity differences vs. structural similarity differences. Each point represents a CDR3β, and its coordinate represents the difference in sequence similarity, *CDRdist,* (*x*-axis) and structural similarity, RMSD, (*y*-axis) between the most similar sequence in its STCRDab structural class and the most similar sequence in any other STCRDab CDR3β structural class. A negative number indicates that the CDR is more similar to the neighbor in its structural class than to the other neighbor; i.e., left indicates more similarity in sequence and bottom more similarity in structure to an in-class CDR than to an out-of-class one. Color represents the distance between the CDR and the in-class neighbor. Structural superpositions highlight example relationships between a target CDR (blue), its in-class neighbor (green), and its other neighbor (orange). **a** Sequence-based classification is correct, but actually does not yield the best structural match. **b-d** Classification is correct, and yields (**b**) similar, (**c**) better, and (**d**) much better structural matches. **e-f** Classification according to STCRDab classes is incorrect but is actually consistent with the relative structural similarity: the other structure is (**e**) more similar or (**f**) about the same, though in (**e**) neither structure is particularly similar. **g** Classification is incorrect, and the closest sequence in the same structural class is more similar to that in the other class

local sequence similarity are likely to adopt similar structures, under a readily interpretable classification approach.

CDR similarity across repertoires from twins

An early landmark study in TCR repertoire analysis characterized three pairs of monozygous twins (“A”, “B”, and “D”), evaluating general characteristics of the repertoires (e.g., diversity) as well as the extent of identity across subjects [47]. The analysis published with that study revealed that the number of identical CDR3 sequences between two individuals was significantly increased if they were twins. We sought to relax the identification of identical sequences across individuals to allow for different degrees of similarity, in particular to test whether twins had more similar sequences than non-twins. As throughout this paper, we explicitly did not consider exact matches, in this case thereby evaluating the “residual” information beyond the previously studied identity. In order to focus the analysis on the strongest signal, we considered only the 1000 most abundant CDR3 sequences from each repertoire, with read counts ranging from over 100,000 down to around 100. Zvyagin et al. [47] observed that shared clonotypes were significantly higher among the most abundant CDR3β sequences for any pair of individuals, and this was even more evident for twins, so we focused on these more significant sequences.

For each unique CDR from each subject, the closest non-identical CDR in each other subject was identified. These matches served as the basis for evaluating nearest-neighbor classification, assessing whether the closest CDR was from a twin (correct) or a non-twin (incorrect). Over the range of required distance thresholds, the number of correct classifications outpaces that of incorrect ones (Fig. 5 and Additional file 1: Table S1), though not as strongly as for the structural clusters. For CDR3α, at a threshold of 0.2 about 46% of each twin pair’s CDRs are correctly classified and about 20% incorrectly classified, with the remaining 34% unidentified. Raising the threshold to 0.3 yields roughly 57% correct, but at the cost of roughly 28% incorrect, leaving only 15% unidentified. Further increases in the threshold continue this trend, with 0.4 resulting in 63% correct but 31% incorrect, and 6% unidentified. Results for CDR3β follow the same sensitivity-specificity trend over this range, but at a lower accuracy: at a threshold of 0.2, there are about 26% correct vs. 14% incorrect; at 0.3 the

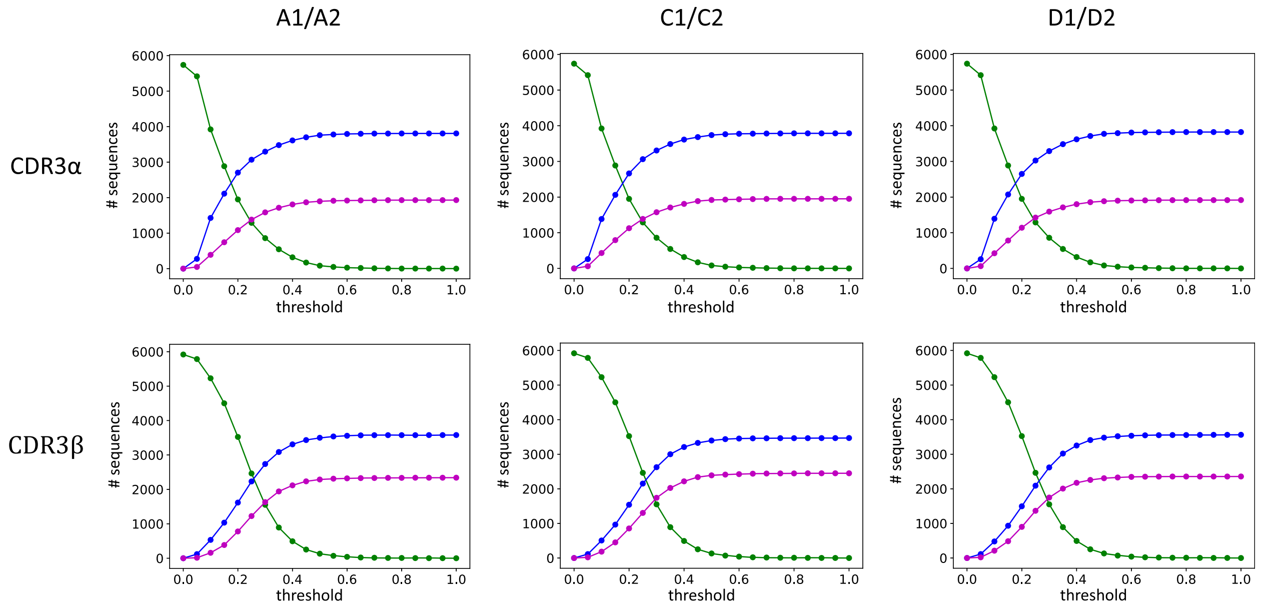

Fig. 5 CDR3 similarity in twins vs. non-twins. The plots track the performance of nearest-neighbor classification, using sequence similarity to predict whether a CDR is from a twin or an unrelated individual. As the threshold required to make a classification is increased (*x*-axis), the fraction of sequences (*y*-axis) that are unclassified (green line) decreases, trading off the fraction that are correctly (blue line) vs. incorrectly (magenta) classified

trade-off is 45% correct vs. 29% incorrect, and at 0.4, 55% correct vs. 37% incorrect. Overall, since identical sequences were explicitly excluded from the classification, this result demonstrates that twin repertoires share not only more identical sequences, but also more similar sequences.

To quantify the significance of the difference in twin vs. non-twin nearest neighbors, the distribution of the closest twin distance was compared to that of the closest non-twin match with a Mann-Whitney U test. CDR3β matches between twins in pair A are closer to each other than to their closest non-twin matches (*p*-value 7×10^{-4}) as are those in twin pair D (1×10^{-9}); however, the distinction does not hold for twin pair C (0.24). For CDR3α, however, only twins A are significantly different from others (0.017), with C marginally above a 5% cutoff (0.07) and D insignificant (0.198). We conclude that two of the three twin pairs have more similar CDR3β sequence, but only one pair has more similar CDR3α sequences. In contrast, the CDR3α classification performance is better than that for CDR3β, even though the overall distributions are more similar, suggesting that focusing specifically on close-enough pairs reveals additional valuable information distinctive of twin pairs.

Analyzing patterns of CDR sequence similarity within and between repertoires can yield biological insights into the basis for specificity (or lack thereof) and suggest directions for further investigation. To explore such patterns for the twin pairs, CDRs in each individual's repertoire were clustered for illustration according to *CDRdist* at a maximum distance of 0.3, which as shown above yields a good specificity-sensitivity trade-off. Clusters from the different individuals were then compared according to an aggregate sequence score, *clusterdist*, computed as the average cluster member distances. Figure 6 illustrates some of the patterns of cluster specificities, with motifs revealing the determinants of specificity (or not). In these examples, a C-terminal DSNYQLIW appears to be common to some CDRs in all of the individuals, while AGNNRKLIW is more specific to individual A1, and VV**GREYGNKLVF distinguishes the twin pair A1/A2 from the other individuals.

Classification of epitope specificity by CDR similarity

As discussed in the introduction, a pair of seminal studies published in 2017 studied epitope-specific TCR repertoires across many individuals. This section characterizes sensitivity-specificity trade-offs in predicting within these datasets which epitope each CDR3 recognizes based on the epitopes for similar CDR3s.

Dash et al. [48] used peptide-MHC (pMHC) tetramer selection and single-cell amplification to collect 4635 paired α/β TCR sequences from 10 epitope-specific repertoires. The 1211 unique CDR3α and 1244 unique CDR3β mouse sequences came from 78 mice and were associated with epitopes labeled NP, PA, F2, and PB1 (presented during influenza infection) and M38, m139,

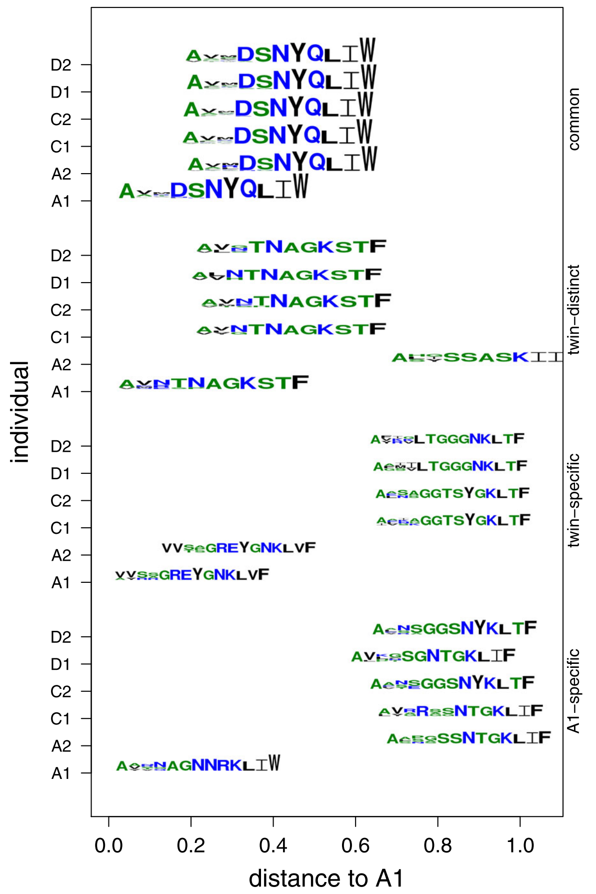

Fig. 6 Example twin motifs with different specificity patterns. Motifs represent sequences for four different clusters from subject A1, and the closest clusters to each of those in each of the five other individuals. The *x*-axis indicates the distance from the A1 cluster to a cluster from another individual, assessed as the average over pairs of CDRs. Sets of clusters illustrate different types of patterns: (common) all of the clusters are similar to the cluster from A1 (within 0.3); (twin-distinct) clusters from unrelated individuals are closer than that from the twin; (twin-specific) the twin's cluster is closer than those from other individuals; (A1-specific) the cluster is far from all clusters from all other individuals

and M45 (presented during murine cytomegalovirus infection). The 276 unique CDR3β and 294 unique CDR3α human sequences came from 32 humans and were associated with epitopes labeled M1 from influenza virus, pp65 from human cytomegalovirus, and BMLF1 from Epstein-Barr virus. Among other analyses, Dash et al. performed nearest-neighbor classification based on a custom sequence similarity score using entire paired TCR sequences, and were able to assign 78% (mouse) and 81% (human) of the TCRs to their correct epitope group. We again sought to characterize how specificity and sensitivity vary, and as throughout the paper directly focused on unpaired CDR3 only and the single most similar sequence to a given one.

Figure 7 illustrates performance over the range of allowed distance thresholds and Additional file 1: Table S2 gives details. At 0.2, 53% of the mouse CDR3βs are classified correctly, as are 43% of the human ones, with 16% mouse and 4% human incorrect. Relative to the CDRs that are actually identified (69% mouse and 47% human), these fractions are 77% mouse correct and 91% human correct, comparable to the previous results (78 and 81%, respectively). Thus even using just the CDR3β alone, *if the single closest neighbor is close enough* then it is highly predictive of the epitope group. Relaxing the threshold to 0.3 yields more correct classifications, 59% mouse and 49% human, at the cost of more incorrect, 29% mouse and 9% human. Consequently the accuracy among those identified is somewhat lower but still comparable, 67% mouse and 87% human. Further relaxing the threshold continues to yield further increases for both correct and incorrect classifications, e.g., at 0.4, 62% of the mouse sequences are correctly classified vs. 36% incorrectly, and 58% correct vs. 16% incorrect for human, which translates to 63% mouse and 79% human correct among those identified. Thus the suitable balance between accuracy and number of predictions again appears to fall in the 0.2 to 0.4 range as we observed first in the structural clusters as evidence of similar conformation. Similar trends hold for analysis of the CDR3α sequences, but CDR3β sequences had greater predictive power, particularly for murine sequences. Other studies, e.g., [47, 49], have likewise found CDR3β sequences to be more informative than CDR3α sequences.

In order to gain some insights into the factors conferring epitope specificity, each of the repertoires was clustered, and as with the twins dataset, the clusters were evaluated for their relative specificity. Within the murine and human groups of repertoires, the cluster similarity score *clusterdist* was computed for each pair of clusters. Each cluster's specificity to its repertoire was characterized by the smallest *clusterdist* to a cluster from a different repertoire, since a relatively small *clusterdist* indicates that the sequence pattern defining a cluster common to epitopes in other repertoires while a relatively large score indicates that no cluster in another repertoire has similar epitopes. Figure 8 illustrates some examples of relatively specific and relatively non-specific clusters at a threshold of 0.3, shown above to yield a good specificity-sensitivity balance. For example, TCS*GTGG*NYAEQFF is common to both PB1 and M38 from mice, with a distance of only 0.12 between the two clusters. On the other hand, ASGLVP*G*VYEQYF is distinct to the human M1 repertoire. Its closest motif is ASS**TGTG*YGYTF in p65 at a distance of 0.79, which is relatively large, indicating that ASGLVP*G*VYEQYF is unique to M1.

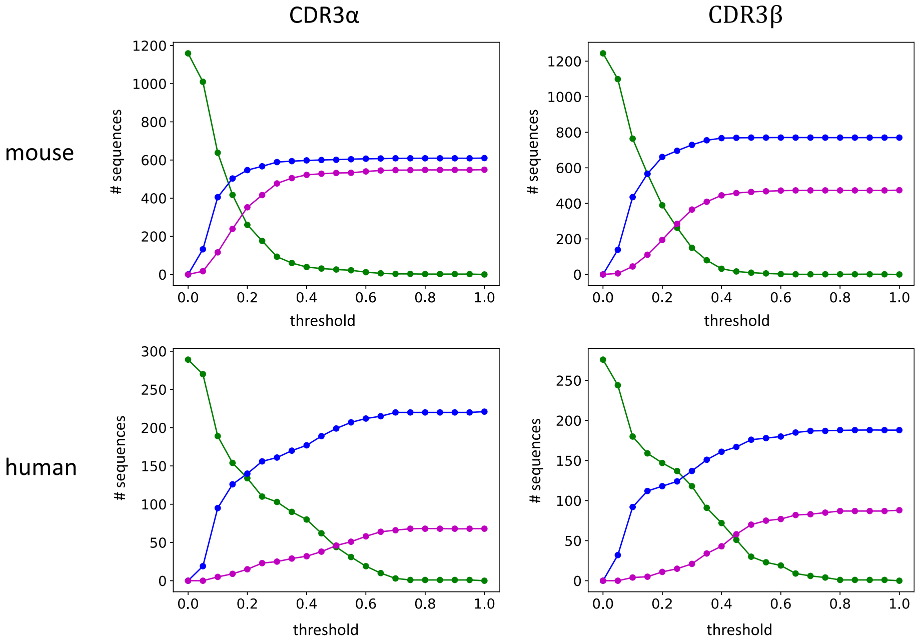

Fig. 7 Epitope specificity in Dash et al. repertoires. The plots track the performance of nearest-neighbor classification, using sequence similarity to predict a CDR's epitope from a set of seven different murine epitopes or from a set of three different human ones. As the threshold required to make a classification is increased (*x*-axis), the fraction of sequences (*y*-axis) that are unclassified (green line) decreases, trading off the fraction that are correctly (blue line) vs. incorrectly (magenta) classified

As shown in Fig. 7, the specificity-sensitivity balance shifts as the threshold varies, so in order to understand the sequence patterns driving that shift, we evaluated the evolution of the underlying clusters (illustrative examples in Fig. 9). As the threshold increases, clusters tend to become larger, containing more unique sequences, and are therefore more diverse and less specific. In particular, for the NP cluster, the nearest cluster in a different repertoire at 0.2 is at a *clusterdist* of 0.43, but that falls to 0.36 for the 0.3 NP cluster; similarly, for PB1-F2 the nearest other-repertoire cluster at 0.2 is 0.49 away, down to 0.38 at 0.4.

Turning to the other recent large TCR repertoire study, Glanville et al. [49] collected 2068 unique sequences using the pMHC tetramers to isolate antigen-specific T cells spanning eight tetramer antigen-MHC (more specifically

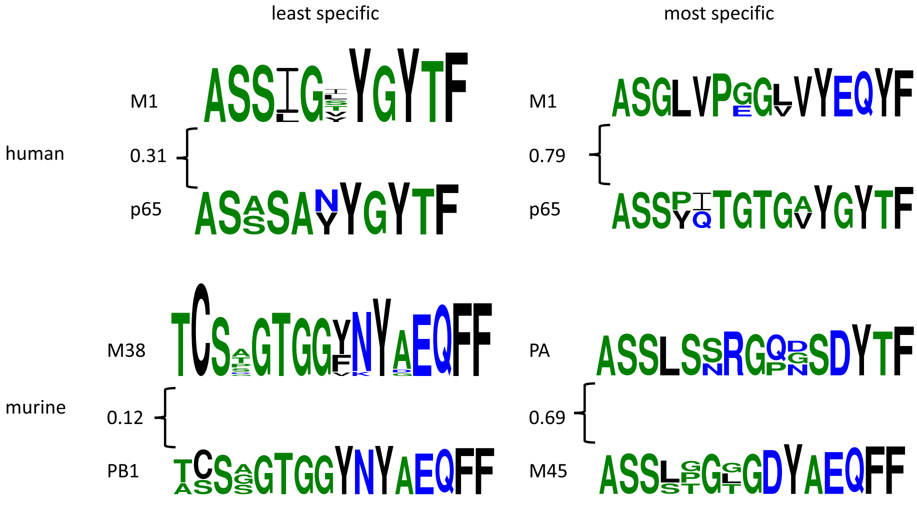

Fig. 8 Example relatively specific and relatively non-specific epitope clusters in Dash et al. repertoires. Each motif represents sequences from a cluster of epitopes at a 0.3 threshold. The distance between an example cluster from one cluster and the nearest from another, computed as the average over pairs of their CDRs, is annotated at the curly brace

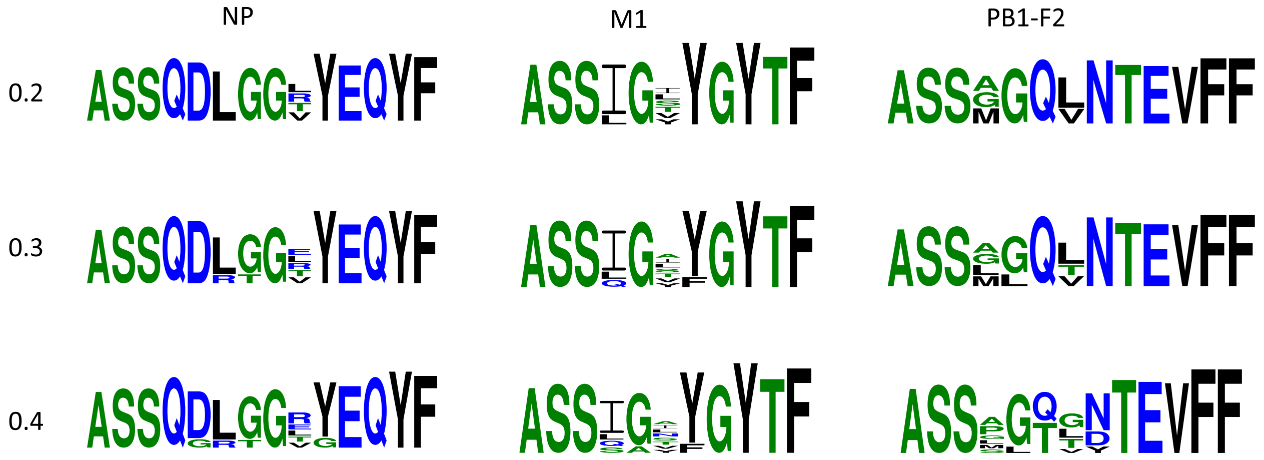

Fig. 9 Example progression of clusters in Dash et al. repertoires as the threshold is relaxed. Motifs represent clusters from three different repertoires (columns: NP, M1, PB1-F2) at three different thresholds (rows: 0.2, 0.3, 0.4)

human MHC, called human leukocyte antigen or HLA) specificities: pp50 associated with HLA-A1 (271 sequences), NP177 associated with HLA-B7 (213), pp65 + HLA-A2 (155), pp65 + HLA-B7 (56), BMLF1 + HLA-A2 (700), M1 + HLA- A2 (56 sequences), and NP44 + HLA-A1 (24 sequences). Their analysis of this data showed that the antigen-specific repertoires tended to share more similar sequences, and revealed some 2-, 3-, and 4-mer motifs enriched in different repertoires. We sought to build on this analysis by once again systematically evaluating the extent of generalization and predictiveness supported by sequence patterns.

Nearest-neighbor classification was performed for each pair of specificities at thresholds of 0.2, 0.3, and 0.4, predicting the epitope+HLA of one CDR based on that of the most similar one (Fig. 10). NP44 and pp65 associated with HLA-B7 generally do very well, and BMLF1 also does relatively well. NP177 and pp65 associated with HLA-A2 are more easily confused with other epitopes. On average, over all the pairwise comparisons, a threshold of 0.2 yields about 19% correctly classified vs. 2% incorrectly classified, with 74% unidentified. Raising the threshold to 0.3 yields 38% correct vs. 7% incorrect with 56% unidentified sequences, and 0.4 continues the trend to 55% vs. 15% with 30% unidentified. The standard deviations on these correct percentages are around 11–13%, as some pairs are clearly quite better than others. Overall, these tests confirmed that the 0.2 to 0.4 range balances accuracy and a sufficient number of predictions for the epitopes in this dataset.

Classification of pathology by CDR similarity

McPAS-TCR catalogues TCR sequences from T cells associated with various pathological conditions in humans and mice [54]. This repository allowed us to move up from epitope specificity to pathology specificity, evaluating how well CDR3β similarity supports classification of the general pathology from which it was derived. The McPAS-TCR human TCR sets with at least 50 unique CDR sequences were downloaded and split into two groups: “small”, with fewer than 400, and “large” with more than 400, yielding a relatively equal number of repertoires per group with relatively balanced number of sequences per repertoire (Table 2). All pairs of pathologies within the same size group were then subjected to nearest-neighbor classification, predicting the pathology of one CDR based on that of the most similar one.

The pairwise classification performance (numbers of correct, incorrect, and unidentified) was calculated for thresholds of 0.2, 0.3, and 0.4, in the common range demonstrating favorable specificity-sensitivity performance across all studies (Fig. 11). On average over all pairwise classifications in the small repertoire group, the percentage of correct classifications increases from 19% at 0.2 to 34% at 0.3 and 49% at 0.4, traded off against 2, 6, and 14% incorrect, respectively, as the fraction of identified sequences goes from 24% up to 40% and finally 63%. Likewise, for the pairwise classifications in the large repertoire group, the fraction of correct classifications averages 24% at 0.2, 48% at 0.3, and 64% at 0.4, with corresponding incorrect classification averages of 4, 12, and 20% and identified averages of 34, 56%, and finally 84%. Over all these comparisons, the standard deviations for correct fractions ranges from 8 to 10%, with some pairs clearly much better than others. In general, larger repertoires are able to classify more sequences than small ones, and do so at a higher accuracy, presumably due to simply having a higher probability of containing a sufficiently close sequence. Some infectious diseases such as influenza, yellow fever, HIV, and hepatitis C all do particularly well in the classification task,

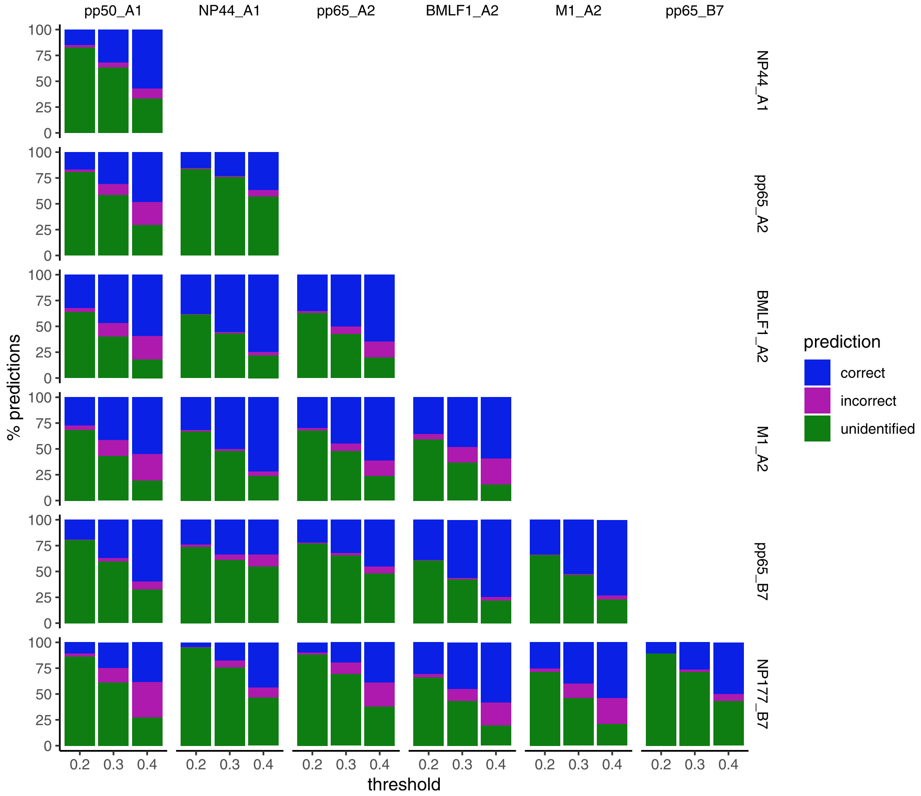

Fig. 10 Epitope specificity in Glanville et al. repertoires. For each pair of epitope+HLA specificities, 1-nearest-neighbor classification was performed to predict whether a CDR belongs to one or the other. Each sub-plot summarizes the performance of such a classification test for one pair of specificities (row vs. column). The bars indicate the percentage of correct (blue), incorrect (magenta), and unidentified (green) predictions between a pair of epitope+HLA specificities (row and column) at thresholds of 0.2, 0.3, and 0.4

even against each other. With these datasets, cancers and autoimmune diseases are confused, and diabetes performs poorly. Further studies are required to ascertain whether differences in classification performance across specific pathologies and general pathology types reflect inherent immunological differences or reveal artifacts in experimental procedures that can perhaps be mitigated with systematic integration methodologies.

Discussion

This study centers on the use of a representative approach to assessing TCR similarity, with local alignments between size-matched CDR3s for each chain separately, and a relaxed substitution scoring matrix combined with a relatively large gap penalty. This set of choices strikes a balance between looking for the local "hot spots" that mediate binding, and accounting for the overall structural context in which those hot spots are situated. While the detailed outcomes would surely be different if the score moved in one direction toward global alignment or in the other direction toward alignment-free motifs, the same general specificity-sensitivity trends would likely hold. In contrast, integrating information across all six CDRs (and even framework regions) [48–50], rather than considering only CDR3α or CDR3β independently, would likely yield higher overall performance. However, we felt it worthwhile to explore how much information was encoded just in the CDR3 regions, and found them to be strikingly informative.

The 1-nearest-neighbor classifier employed here for specificity-sensitivity assessments is one of the simplest approaches possible, but makes it straightforward to understand and analyze the basis for predictions. A model-based approach, e.g., a linear classifier or even a nonlinear model [51, 52, 55, 56], could give better predictive performance, but would also confound some of the analyses due to the differences in sizes and diversity in different groups. At the same time, a statistical learning approach could provide insights into the importance to different groups of particular CDRs and particular residue positions, directly reveal amino acid motifs conferring specificity, and so forth. In order to focus on the information provided by CDR similarity alone, the analyses presented here did not allow for identity in the 1-nearest-neighbor classification and did

Table 2 Pathology-associated repertoires [54] for small (a) and large (b) size-matched groups

Pathology	Category	Size
(a) small		
Allergy	Allergy	259
Celiac disease (celiac)	Autoimmune	70
Multiple sclerosis (MS)	Autoimmune	116
Rheumatoid Arthritis (RA)	Autoimmune	270
Clear cell renal carcinoma (clearcell)	Cancer	68
Hepatitis C virus (HepC)	Pathogens	85
Yellow fever virus (YF)	Pathogens	179
(b) large		
Diabetes Type 1 (diabetes)	Autoimmune	724
Melanoma	Cancer	475
Cytomegalovirus (CMV)	Pathogens	921
Epstein Barr virus (EBV)	Pathogens	1061
Human immunodeficiency virus (HIV)	Pathogens	649
Influenza	Pathogens	2939

not consider abundance. This made the classification task somewhat harder, but also provided a uniform basis across the different studies. In a specific practical application, leveraging identity and abundance would be advantageous; statistical learning approaches could offer a natural means to incorporate this information.

Our analysis of structural clusters provided some intriguing insights into sequence-structure relationships, while the rest of the paper explored a range of sequence-function relationships. The structural analysis was somewhat limited due to limited available structures, but a more complete loop modeling approach [57–59] might provide additional power. The functional analysis could benefit from incorporation of MHC restriction information [50], in order to reveal associations among the presented antigens, the presenting MHCs, and the CDRs. And ultimately, combining sequence, structure, and function in an integrated model could provide much deeper insights into the basis for specific recognition.

We individually analyzed repertoires for pairs of twins [47] and for particular antigen specificities [48, 49], along with aggregated collections of pathology-related repertoires [54], but we did not seek to combine information across these different studies. An integrative analysis could provide insights into common modalities of recognition, as has been shown for MHC restrictions [50], but which could also span broader functional associations across antigens from different pathogens as well as from different "self"s. An integrative analysis could thus seek to account for private and public aspects of recognition, gain insights into genetics vs. exposure, and support modeling of the development of immunity.

Conclusion

This paper has systematically explored the utility of using CDR3 sequence similarity to predict structural class and functional group (namely antigen specificity or pathology association). Based on a representative measure of similarity and an interpretable classification method, the information content in CDR3 alone was shown to support highly specific predictions at a

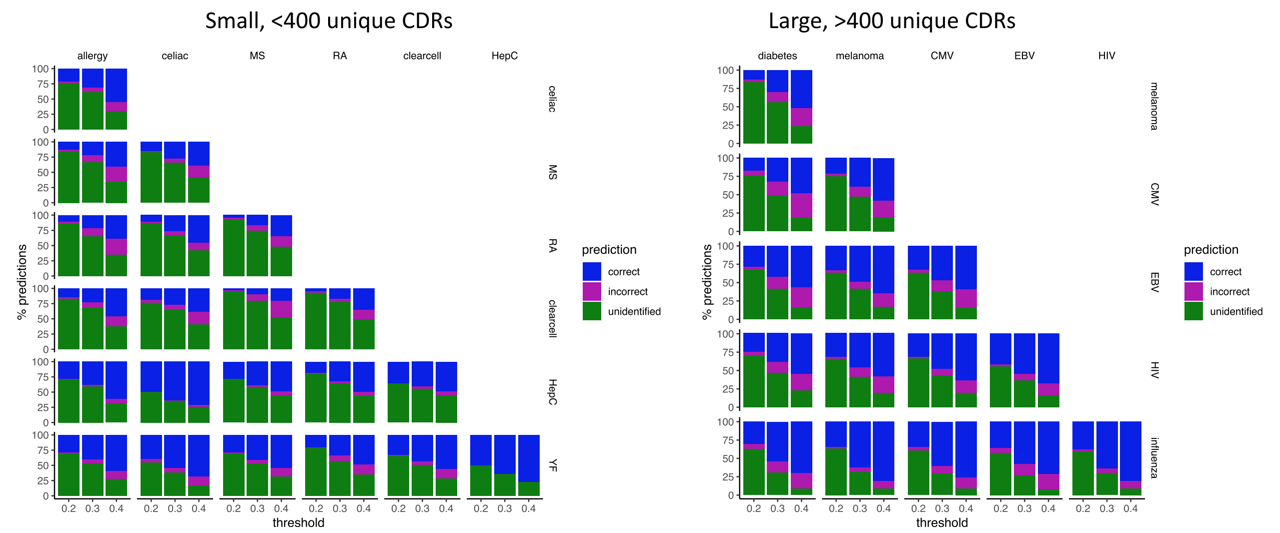

Fig. 11 Pathology specificity of McPAS-TCR repertoires. For each pair of pathologies in a set of size-matched repertroies (7 small, 6 large), 1-nearest-neighbor classification was performed to predict whether a CDR was correlated with one pathology or the other. Each sub-plot summarizes the performance of such as classification test for one such pair of pathologies (row vs. column). The bars indicate the percentage of correct (blue), incorrect (magenta), and unidentified (green) predictions between a pair of pathologies (row and column) at thresholds of 0.2, 0.3, and 0.4

sufficiently stringent similarity threshold, and to maintain good specificity even while increasing sensitivity by substantially relaxing the threshold. Furthermore, patterns supporting predictions within and between groups were shown to provide insights into the structural and functional bases for recognition. We conclude that, if suitably controlled as demonstrated here, predictive frameworks can productively leverage sequence patterns in characterizing and predicting TCR sequence-structure-function relationships.

Methods

Data processing

When data was processed for each repertoire, the first "C" from each sequence was removed as it is uninformative for scoring, uniformly inflating the scores. Duplicate sequences were combined within each repertoire.

Sequence similarity score

Given a pair of CDRs to evaluate for similarity, local alignment was performed using the Smith-Waterman (SW) algorithm [60], implemented in the Python package swalign [61]. SW was applied with the BLOSUM45 substitution matrix [62] to allow for biochemical diversity, and a gap penalty of − 10 to focus on matching largely gap-less substrings. Manual inspection of some CDR alignments suggested that these parameters accomplished the intended goals. So as to generate a score that is universally comparable across different CDR sets, the alignment score was normalized by dividing by the self-scores of the two sequences. To provide a dissimilarity measure suitable for clustering, the normalized score was then subtracted from 1. Since SW scores are non-negative and self-scores are maximal, the final score is between 0 (identical) and 1 (no discernable similarity). Formally, for two sequences A and B, the distance is:

$$CDRdist(A,B) = 1 - \sqrt{\frac{SW(A,B)^2}{SW(A,A) * SW(B,B)}}$$

Nearest neighbor classification

Given a set of CDRs that are labeled as belonging to one of two or more different groups, a nearest neighbor classifier predicts the label of another CDR based on "nearby" labeled CDRs, in terms of *CDRdist*. A 1-nearest-neighbor classifier was used for the results here, making the assignment on the single closest (but not identical) CDR rather than taking a vote among several. Furthermore, the allowed distance was thresholded, such that if no neighbor was sufficiently close, then no prediction would be made.

Clustering

CDRs were clustered using hierarchical agglomerative clustering via the linkage function in the scipy.cluster.hierarchy package of scipy [63], with the *CDRdist* as a comparison function. The resulting dendogram was cut at the specified *CDRdist* threshold (i.e., 0.2, 0.3, or 0.4 in the example results) in order to define clusters.

Cluster similarity score

In order to characterize how specific or non-specific clusters were to the repertoires they came from, the distance between a pair of clusters was computed as the average pairwise distance between their members:

$$clusterdist(C1, C2) = \frac{\sum_{c \in C1} \sum_{c' \in C2} CDRdist(c, c')}{|C1||C2|}$$

Then the relative specificity of a cluster to its repertoire was characterized in terms of its distance to clusters from other repertoires, with a small score indicating non-specificity (i.e., similarity to a cluster from another repertoire).

Logos

Sequence logos were generated by Weblogo version 2.8.2 [64, 65] in conjunction with the biopython [66] motif library Bio.motifs.Motif.

Additional file

Additional file 1: Classification Results Details. (DOCX 18 kb)

Abbreviations

CDR: Complementarity determining region; HLA: Human leukocyte antigen; MHC: Major histocompatibility complex; pMHC: Peptide – MHC complex; RMSD: Root mean squared deviation; SW: Smith-Waterman; TCR: T cell receptor; VDJ: Variable-diversity-joining

Acknowledgements

We thank Jennifer Franks and Yoonjoo Choi for valuable comments on this work.

Funding

Salaries supporting the development and analysis described here were funded in part by NIH grant 2R01GM098977.

Availability of data and materials

The datasets analyzed during the current study are publicly available as follows:

- structural data from SCTRDab http://opig.stats.ox.ac.uk/webapps/stcrdab/ [53]
- twins data from http://labcfg.ibch.ru/tcr.html#MZTwins, derived from Zvyagin et al. [47]
- antigen-specific data for Dash et al. [48] from the Additional file 1: Tables S1-S2 (Table 1)
- antigen-specific data for Glanville et al. [49] from https://vdjdb.cdr3.net/search searching for PMID 28636592
- pathology-associated data from McPAS-TCR http://friedmanlab.weizmann.ac.il/McPAS-TCR/ [54]

All of our analysis code is freely available, under an MIT license, on github, at https://github.com/neerjathakkar/Distinguishing-TCR-Groups. This includes directions for obtaining and preprocessing these datasets that have been provided by others. The code is in platform-independent Python.

Authors' contributions

NT was primarily responsible for developing and applying the analysis methods. CBK developed additional pre- and post-processing methods, performed auxiliary analyses, and discussed NT's analyses. Both authors wrote the manuscript together. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 30 January 2019 Accepted: 29 April 2019
Published online: 15 May 2019

References

1. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, et al. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med. 2006;13:100–6.
2. Elliott SL, Suhrbier A, Miles JJ, Lawrence G, Pye SJ, Le TT, et al. Phase I trial of a CD8+ T-cell peptide epitope-based vaccine for infectious mononucleosis. J Virol. 2008;82:1448–57.
3. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. Open Biol. 2013;3:120139.
4. Barouch DH, O'Brien KL, Simmons NL, King SL, Abbink P, Maxfield LF, et al. Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat Med. 2010;16:319–23.
5. Tian Y, da Silva Antunes R, Sidney J, Lindestam Arlehamn CS, Grifoni A, Dhanda SK, et al. A review on T cell epitopes identified using prediction and cell-mediated immune models for mycobacterium tuberculosis and Bordetella pertussis. Front Immunol. 2018;9:2778.
6. Theiler J, Korber B. Graph-based optimization of epitope coverage for vaccine antigen design. Stat Med. 2018;37:181–94.
7. Hilleman MR. Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. Proc Natl Acad Sci. 2004; 101(Supplement 2):14560–6.
8. Calis JJA, de Boer RJ, Keşmir C. Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. PLoS Comput Biol. 2012;8:e1002412.
9. He L, De Groot AS, Gutierrez AH, Martin WD, Moise L, Bailey-Kellogg C. Integrated assessment of predicted MHC binding and cross-conservation with self reveals patterns of viral camouflage. BMC Bioinformatics. 2014; 15(Suppl 4):S1.
10. He L, De Groot AS, Bailey-Kellogg C. Hit-and-run, hit-and-stay, and commensal bacteria present different peptide content when viewed from the perspective of the T cell. Vaccine. 2015;33:6922–9.
11. Robbins PF, Lu Y-C, El-Gamil M, Li YF, Gross C, Gartner J, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat Med. 2013;19:747–52.
12. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2014;371:2189–99.
13. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.
14. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547: 217–21.
15. Oseroff C, Sidney J, Kotturi MF, Kolla R, Alam R, Broide DH, et al. Molecular determinants of T cell epitope recognition to the common Timothy grass allergen. J Immunol. 2010;185:943–55.
16. Oseroff C, Sidney J, Vita R, Tripple V, McKinney DM, Southwood S, et al. T cell responses to known allergen proteins are differently polarized and account for a variable fraction of total response to allergen extracts. J Immunol. 2012;189:1800–11.
17. Gross DM, Forsthuber T, Tary-Lehmann M, Etling C, Ito K, Nagy ZA, et al. Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. Science. 1998;281:703–6.
18. Losikoff PT, Mishra S, Terry F, Gutierrez A, Ardito MT, Fast L, et al. HCV epitope, homologous to multiple human protein sequences, induces a regulatory T cell response in infected patients. J Hepatol. 2015;62:48–55.
19. Salvat RS, Choi Y, Bishop A, Bailey-Kellogg C, Griswold KE. Protein deimmunization via structure-based design enables efficient epitope deletion at high mutational loads. Biotechnol Bioeng. 2015;112:1306–18.
20. Salvat RS, Verma D, Parker AS, Kirsch JR, Brooks SA, Bailey-Kellogg C, et al. Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. Proc Natl Acad Sci. 2017;114:201621233.
21. Blazanovic K, Zhao H, Choi Y, Li W, Salvat RS, Osipovitch DC, et al. Structure-based redesign of lysostaphin yields potent antistaphylococcal enzymes that evade immune cell surveillance. Mol Ther Methods Clin Dev. 2015;2:15021.
22. Zhao H, Verma D, Li W, Choi Y, Ndong C, Fiering SN, et al. Depletion of T cell epitopes in Lysostaphin mitigates anti-drug antibody response and enhances antibacterial efficacy in vivo. Chem Biol. 2015;22:629–39.
23. King C, Garza EN, Mazor R, Linehan JL, Pastan I, Pepper M, et al. Removing T-cell epitopes with computational protein design. Proc Natl Acad Sci. 2014;111:8577–82.
24. Griswold KE, Bailey-Kellogg C. Design and engineering of deimmunized biotherapeutics. Curr Opin Struct Biol. 2016;39:79–88.
25. Cantor JR, Yoo TH, Dixit A, Iverson BL, Forsthuber TG, Georgiou G. Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. Proc Natl Acad Sci. 2011;108:1272–7.
26. Lamberth K, Reedtz-Runge SL, Simon J, Klementyeva K, Pandey GS, Padkjær SB, et al. Post hoc assessment of the immunogenicity of bioengineered factor VIIa demonstrates the use of preclinical tools. Sci Transl Med. 2017;9: eaag1286.
27. Salvat RS, Parker AS, Choi Y, Bailey-Kellogg C, Griswold KE. Mapping the Pareto optimal design space for a functionally deimmunized biotherapeutic candidate. PLoS Comput Biol. 2015;11:e1003988.
28. Rudolph MG, Stanfield RL, Wilson IA. How TCRS bind MHCS, peptides, and CORECEPTORS. Annu Rev Immunol. 2006;24:419–66.
29. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics. 1999;50:201–12.
30. Lund O, Nielsen M, Kesmir C, Petersen AG, Roder G, Justesen S, et al. Definition of supertypes for HLA molecules using clustering of specificity matrices. Immunogenetics. 2004;55:797–810.
31. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature. 1988;334:395–402.
32. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009;114:4099–107.
33. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee J-Y, et al. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci. 2014;111: 13139–44.
34. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. Science. 1999; 286:958–61.
35. Hennecke J, Wiley DC. T cell receptor-MHC interactions up close. Cell. 2001; 104:1–4.
36. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol. 2014;35:581–90.
37. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. Cell. 2014;157:1073–87.
38. Wooldridge L, Ekeruche-Makinde J, van den Berg HA, Skowera A, Miles JJ, Tan MP, et al. A single autoimmune T cell receptor recognizes more than a million different peptides. J Biol Chem. 2012;287:1168–77.

39. Hemmer B, Vergelli M, Gran B, Ling N, Conlon P, Pinilla C, et al. Predictable TCR antigen recognition based on peptide scans leads to the identification of agonist ligands with no sequence homology. J Immunol. 1998;160:3631–6.
40. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012;135:183–91.
41. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. Trends Biotechnol. 2017;35:203–14.
42. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009;19:1817–24.
43. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 2011;21:790–7.
44. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur J Immunol. 2012;42:3073–83.
45. Dash P, McClaren JL, Oguin TH, Rothwell W, Todd B, Morris MY, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. J Clin Invest. 2011;121:288–95.
46. 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.
47. Zvyagin IV, Pogorelyy MV, Ivanova ME, Komech EA, Shugay M, Bolotin DA, et al. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. Proc Natl Acad Sci. 2014;111:5980–5.
48. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017;547:89–93.
49. 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.
50. DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. Elife. 2018;7.
51. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Front Immunol. 2018;9:224.
52. Cinelli M, Sun Y, Best K, Heather JM, Reich-Zeliger S, Shifrut E, et al. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. Bioinformatics. 2017;33:btw771.
53. Leem J, de Oliveira SHP, Krawczyk K, Deane CM. STCRDab: the structural T-cell receptor database. Nucleic Acids Res. 2018;46:D406–12.
54. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics. 2017;33:2924–9.
55. Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, et al. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. Front Immunol. 2017;8:430.
56. Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, et al. Learning the high-dimensional Immunogenomic features that predict public and private antibody repertoires. J Immunol. 2017;199:2985–97.
57. Leimgruber A, Ferber M, Irving M, Hussain-Kahn H, Wieckowski S, Derré L, et al. TCRep 3D: an automated in silico approach to study the structural properties of TCR repertoires. PLoS One. 2011;6:e26301.
58. Stein A, Kortemme T. Improvements to robotics-inspired conformational sampling in Rosetta. PLoS One. 2013;8:e63090.
59. Gowthaman R, Pierce BG. TCRmodel: high resolution modeling of T cell receptors from sequence. Nucleic Acids Res. 2018;46:W396–401.
60. Waterman M, Smith T, Beyer W. Some biological sequence metrics. Adv Math (N Y). 1976;20:367–87.
61. Breese MR. swalign. 2015. https://github.com/mbreese/swalign.
62. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89:10915–9.
63. Jones E, Oliphant E, Peterson P, SciPy: open source scientific tools for Python. 2001. http://www.scipy.org/.
64. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14:1188–90.
65. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18:6097–100.
66. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
67. Chen J-L, Stewart-Jones G, Bossi G, Lissin NM, Wooldridge L, Choi EML, et al. Structural and kinetic basis for heightened immunogenicity of T cell vaccines. J Exp Med. 2005;201:1243–55.
68. Schrodinger LLC. The PyMOL molecular graphics system, version 1.8; 2015.