Generalizable features for the diagnosis of infectious disease, autoimmunity and cancer from adaptive immune receptor repertoires

Zichang Xu[1], Hendra S Ismanto[1,2], Dianita S Saputri[1,2], Soichiro Haruna[2], Guanqun Sun[3], Jan Wilamowski[2], Shunsuke Teraguchi[2,4], Ayan Sengupta[5], Songling Li[1,2] and Daron M Standley[1,2]

[1]Department of Systems Immunology, Immunology Frontier Research Institute (IFReC), Osaka University, Suita 565-0871, Japan

[2]Department of Genome Informatics, Research Institute for Microbial Diseases (RIMD), Osaka University, Suita 565-0871, Japan

[3]School of Information Science, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Japan

[4]Faculty Data Science, Shiga University, Hikone 522-8522, Japan

[5]Cogent Labs, Tokyo 106-032, Japan

Abstract

Classifiers trained on adaptive immune receptors (AIRs) have been reported to accurately identify a broad range of diseases. However, when using the conventional "clonotype" representation of AIRs, different donors exhibit vastly different features, limiting the generalizability of these classifiers. To address this problem, we developed a representation of AIRs based on paratope adjacency that significantly improved donor sharing and classification performance on new donors across a wide range of diseases (mean ROC AUC 0.893) compared to classifiers trained on clonotypes (0.714). Surprisingly, in cancer data, we observed that some of the AIRs that were important for the classification were significantly more abundant in healthy controls than in patients. Such healthy-biased, but not disease-biased, AIRs were predicted to target known cancer-associated antigens at dramatically higher rates than background values (Z-scores > 75), constituting an overlooked reservoir of cancer-targeting immune cells that are diagnostic and identifiable from a routine blood test.

Introduction

Adaptive immune receptors (AIRs), which consist of B cell receptors (BCRs) and T cell receptors (TCRs), play a direct role in the detection and prevention of various diseases by recognizing specific endogenous and exogenous antigens. Their function depends on the ability of each B or T cell to express one of many possible AIR variable region coding sequences, which are assembled by recombination of V, D and J gene segments (**Fig. S1A**). The topology of a folded variable region brings together three distinct segments called complementarity determining regions (CDR1, CDR2, add CDR3) (**Fig. S1B**). Most physical contacts between AIRs and antigens occur within or near the CDRs (**Fig. S1C**). Because B and T cells are selected to preferably bind to "non-self" antigens, and binding can stimulate their further proliferation, the frequency distribution of distinct AIR sequences is highly skewed toward a small fraction of high-frequency proliferating sequences with a very long tail of lower frequency sequences (**Fig. S1D**). Both infectious disease and cancer will produce non-self antigens. Conversely, if self/non-self recognition fails, B or T cells targeting endogenous molecules can proliferate, leading to autoimmune disease. As a result, the frequency distribution of distinct AIR sequences is very sensitive to the presence of a wide range of diseases.

Since AIRs are generated by a stochastic process, and the number of B and T cells in an individual is limited, the repertoires of AIRs in different donors differ to the extent that, if a specific AIR is detected in a blood sample from one individual, the identical AIR is unlikely to be observed in a second individual. Thus, despite their exquisite sensitivity to disease, the lack of "donor sharing" presents a challenge for the use of AIRs as disease biomarkers.

Several recent studies have described machine learning classifiers to tackle this problem[1-10], two which (DeepRC[10] and immuneML[9]) are available as stand-alone tools that can be easily trained on one's own AIR data. A common way of improving donor sharing is to cluster or embed similar but non-identical AIRs in a lower dimensional coordinate system[6-8]. The most common way of describing AIRs for clustering is "clonotyping". Clonotype nomenclature derives from V(D)J

recombination, and consists of the V and J gene names, along with the amino acid sequence of the CDR3 region, where the V and J coding regions join. Because of its use of gene names, clonotyping is useful as a qualitative description, but it is not convenient for computational analysis of AIRs for at least two reasons. First, clonotypes mix categorical variables (gene names) and quasi-continuous variables (CDR3 sequence) making it difficult to define a continuous measure of their pairwise similarity or difference. Secondly, because of their categorical nature, clonotypes tend to divide AIRs into immutable "clonotype clusters". Such clusters typically consist of AIRs with the same V and J gene names, same CDR3 length, and highly similar CDR3 sequences.

Clusters are important and probably necessary for utilizing AIRs as features. However, as we show below, equating AIRs with their clusters leads to a very sparse representation of donors, hindering their use in disease classification. If we instead represent AIRs by their paratopes by concatenating the sequences of the three CDRs, then the similarity or difference between any two AIRs can be defined by conventional protein sequence comparison methods. From such a perspective, clusters are simply dense regions in a network of AIRs with similar paratopes. This similarity network is probably required for robust defense against pathogens and tumor cells that evolve rapidly in order to evade immune surveillance. Here we utilize this similarity network to extend the cluster representation of AIRs in order improve donor feature generalizability.

We describe the details of our approach in *Methods* and illustrate the main idea in **Fig. 1**. Briefly, for disease diagnosis, we need a single feature vector to describe each donor. When using clusters to represent AIRs, an individual donor can be represented by the frequency of clusters in the donor's repertoire. We denote such features "cluster frequencies" (CFs) (**Fig. 1A**). We next extend CFs by allowing a given AIR to "occupy" multiple clusters. This is accomplished through an adjacency matrix A that describes the network of pairwise paratope similarities (**Fig. 1B**). Here, CFs consist of "clonotype cluster frequencies" (CCFs) or "paratope cluster frequencies" (PCFs), while cluster occupancies (COs), consist of "clonotype cluster occupancies" (CCOs) or "paratope cluster occupancies" (PCOs), depending on whether we use clonotype or paratope similarity to

define the original clusters (**Fig. 1C**). All features are indexed by a specific cluster, allowing feature importance to be used to identify biologically important AIRs.

Because AIR frequencies follow a longtail distribution (**Fig. S1D**), cluster frequencies do as well, which results in very long feature vectors. We thus require a way of dropping clusters that are not populated by many donors. We introduce a parameter d_{min} that sets a threshold for the required donor diversity. By discarding features with low donor diversity, we can reduce the sparsity of the features, reduce the dimensionality of the feature vectors, and reduce the number of AIRs in the paratope adjacency matrix. The parameter d_{min} is thus tuned in the training process, as described in *Results*.

To evaluate the classifiers for a wide range of diseases, we collected AIR sequences for 6 diseases covering 3 general categories (**Table S1**): infectious disease (COVID-19, HIV), autoimmune disease (autoimmune hepatitis, type 1 diabetes), and cancer (colorectal cancer, non-small cell lung cancer). Wherever possible, we utilized disease and healthy control data from the same study in order to minimize batch effects (i.e. any bias other than the disease of interest), or used different studies for training and testing.

Results

We randomly split donors into training (70%) and test (30%) groups. For the classifiers described here, we selected the optimal hyperparameter d_{min} by leave-one-out cross validation (LOOCV) using the training donors. The previously-published methods DeepRC[10] and immuneML[9] were trained as described in methods. We then evaluated all trained classifiers on the test set (**Fig. 1C**). For the classifiers described here, with the exception of feature generation (CCF, CCO, PCF, or PCO), all steps in the pipeline were identical.

Table 1 contains a summary of the performance of the classifiers on the independent test donors using each of the four features. In addition to the classifiers described here, we also included the previously-published methods DeepRC[10] and immuneML[9] using various settings. The PCO-based

classifier achieved a higher area under the receiver operating characteristic curve (ROC AUC) than all other classifiers for all six diseases. The mean ROC AUC for the PCO-based classifier (0.893) was substantially higher than that for the CCF-based classifier (0.714), and far above the performance of a random classifier (0.5). Three of the PCO ROC AUCs—HIV (0.985), AIH (0.947) and NSCLC (0.985)—were close to perfect. The remaining three—COVID-19 (0.896), T1D (0.725), CRC (0.821)—had the fewest number of disease donors suggesting that at least 50 donors may be needed for robust classification. In contrast, the CCF-based classier ROC AUCs only exceeded 0.9 in one case (HIV) for which the sequencing depth was much higher than the rest of the diseases, and completely failed for another case (T1D). In the following sections we discuss two diseases—COVID-19 and NSCLC—in depth. The remaining cases are described in (**Figs S3-6 and Suppl text**).

		COVID19	**HIV**	**AIH**	**T1D**	**CRC**	**NSCLC**	**Average AUC**
Donors	**Healthy**	58	128	59	11	88	294	
	Disease	44	94	59	34	20	204	
ROC AUC	**DeepRC (CNN)**	0.407	0.888	0.66	0.5	0.541	0.715	0.619
	DeepRC (LSTM)	0.813	0.861	0.5	0.5	0.458	0.741	0.646
	immuneML (LR)	0.717	0.946	0.688	0.65	0.78	0.802	0.764
	immuneML (SVM)	0.781	0.744	0.5	0.5	0.75	0.836	0.685
	immuneML (KNN)	0.565	0.851	0.656	0.525	0.76	0.795	0.692
	immuneML (RF)	0.838	0.876	0.706	0.5	0.85	0.895	0.777
	immuneML (PBC)	0.533	0.5	0.713	0.5	0.66	0.317	0.537
	CCF	0.648	0.941	0.653	0.500	0.779	0.762	0.714
	PCF	0.867	0.968	0.787	0.500	0.800	0.967	0.815
	CCO	0.863	0.975	0.828	0.625	0.700	0.409	0.733
	PCO	**0.896**	**0.985**	**0.947**	**0.725**	**0.821**	**0.985**	**0.893**
PR AUC	**DeepRC (CNN)**	0.706	0.796	0.636	0.714	0.451	0.651	0.659

	DeepRC (LSTM)	0.928	0.856	0.444	0.714	0.416	0.696	0.676
	immuneML (LR)	0.796	0.914	0.729	0.871	0.718	0.802	0.805
	immuneML (SVM)	0.894	0.774	0.722	0.857	0.719	0.905	0.812
	immuneML (KNN)	0.71	0.785	0.7	0.835	0.782	0.827	0.773
	immuneML (RF)	0.878	0.831	0.757	0.857	**0.810**	0.851	0.831
	immuneML (PBC)	0.767	0.723	0.722	0.857	0.560	0.31	0.657
	CCF	0.818	0.936	0.544	0.857	0.332	0.605	0.682
	PCF	0.893	0.930	0.740	0.857	0.565	0.965	0.825
	CCO	0.890	0.947	0.808	0.829	0.390	0.371	0.706
	PCO	**0.932**	**0.966**	**0.953**	**0.902**	0.468	**0.982**	**0.867**
	Rand	0.516	0.236	0.444	0.714	0.152	0.421	0.413

Table 1. Summary of performance metrics. Performance of previously-published methods (DeepRC, immuneML) using various settings, along with the classifiers described here trained on each of the four features (CCF, PCF, CCO, and PCO) applied to six diseases (COVID-19, HIV, AIH, T1D, CRC, and NSCLC). The numbers of donors are listed under each disease. Areas under the curve (AUC) for both receiver operating characteristic (ROC) and precision-recall (PR) curves are given. For the PR AUC values, the expected performance of a random predictor (Rand), is given by the ratio of positive to negative donors. The abbreviations in column two are as follows: Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Probabilistic Binary Classifier (PBC).

COVID-19

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which initially targets the epithelial tissue of the nasal cavity and spreads to the upper respiratory tract, lung, and other organs[11]. Diagnosis is usually confirmed by polymerase chain reaction (PCR), which can detect the presence of viral nucleic acids[12]. In order to evaluate the performance of an AIR-based diagnostic, we utilized BCR heavy chain data from a single cell sequencing study of peripheral blood mononuclear cells (PBMCs) of 44 COVID-19 patients (106,640 BCRs) and 58 healthy donors (174,139 BCRs), all of whom were unvaccinated[13]. After splitting randomly, 71 donors were used for training and 31 were held out for testing. **Fig. 2A** shows the effect of d_{min} on the sparsity of the features as well as the LOOCV target function, which reached a maximum at d_{min}=0.7 (pink line). **Fig. 2B** shows ROC curves for the test data

using the PCO features, which achieved an AUC of 0.896. From the qualitative agreement between the LOOCV and test ROC curves, it appears that the classifier was not overfit and generalized well to new donors. In **Fig. 2C** the test ROC AUCs for classifiers using all four features are shown, which indicate that only the model trained on CCF features was unable to classify test donors, presumably due to the poor sharing of clonotypes across donors. In **Fig. 2D** we show precision-recall (PR) AUCs, which indicate that the PCO-based classifier achieved the highest value (0.932) among all trained models.

The stark difference in the performance of CCF-based features compared to the rest suggests that the computation of paratope occupancies made a critical difference in this case. In order understand this effect, we examined differences between CCF and CCO features, which utilized the same clonotype clusters but were transformed to occupancies using paratope similarities. To this end, we first identified the feature with the highest CCO importance. We then examined the clonotype cluster (Cluster 1) corresponding to this feature. **Fig. 2E** shows that this cluster consisted of only three AIRs from two donors (donor 10 and donor 32). By examining the CCO calculation, we found that AIRs from many donors shared paratope-level similarity to one or more of the three members of cluster 1 (**Fig. 2F**). Thus, while the CCF feature corresponding to this cluster was zero for all but the two of the donors (10 and 32), the corresponding CCO feature was non-zero for many donors (**Fig. 2G**). This example highlights the fact that an AIR that belongs to a given clonotype cluster can nevertheless have significant paratope similarity to AIRs in different clusters. Including such paratope-level similarities was critical to the performance of the CCO and PCO COVID-19 classifier performance.

Non-small Cell Lung Cancer

Non-small Cell Lung Cancer (NSCLC) is the most common type of lung cancer and includes squamous-cell carcinoma, large-cell carcinoma, adenocarcinoma, as well as several rarer subtypes[14]. Diagnosis traditionally utilizes a wide range of evidence, including symptoms (e.g. persistent cough), imaging data, and tissue biopsy; however, blood-based biomarkers are recently becoming more common[15]. As with many cancers, it is known that disease-specific T cells often

infiltrate tumors, but it is less understood whether or not these T cells also circulate in the blood. Thus, TCR-based diagnosis potentially represents a less invasive approach than biopsy. Unlike the previous examples, we were unable to find a single study that provided both disease and healthy PBMCs for NSCLC. As an alternative we instead constructed a diverse dataset from seven studies: Three for healthy training[16-18], two for NSCLC training [19,20], one for healthy testing[21] and one for NSCLC testing[22]. This resulted in a total of 204 NSCLC donors (6,734,867 TCRs) and 294 healthy donors (4,120,597 TCRs). After splitting the data, there were 344 training donors and 152 held out for testing. **Fig. 3A** shows the sparsity of the resulting features, indicating that occupancy-based features (CCO, PCO) were about 20% less sparse than the frequency-based features at small values of d_{min}, but converged at larger values, while the LOOCV target function remained perfect at all values. **Fig. 3B** shows that the PCO-based classifier trained using the highest d_{min} (0.9) achieved a ROC AUC of 0.985 for the test donors. **Figs. 3C-D** indicate that only the classifiers trained on occupancy-based features (CCO, PCO) performed well on test donors, and that the performance was close to perfect (ROC AUC 0.985; PR AUC 0.982).

NSCLC classifiers trained on PCOs were robust against batch effects

Use of a large NSCLC dataset composed of multiple studies provided the opportunity to systematically explore the robustness of the different features against batch effects. To this end, we sampled all possible splits of the datasets where the test set consisted of one healthy study and one NSCLC study, with the remaining studies used for training. This resulted in twelve splits. These results indicated that, in terms of ROC (**Fig. 3E**) and PR AUC (**Fig. 3F**), both PCF and PCO features performed well overall, but that CCF and CCO features performed inconsistently, in agreement with **Figs. 3C-D**. It was possible to combine the NSCLC studies in such a way that the PCO model failed to generalize to the test data. Out of a total of twelve combinations, there were two such splits. Interestingly, these two cases where ones where the size of the training data was much smaller than that of the test data, consistent with other diseases, again emphasizing the need for sufficiently large training data. Overall, however, the performance of the PCO-based classifier was more robust than that of the alternatives.

Surprising relationship between healthy-biased clusters and cancer antigen specificity

Because each feature corresponds to an AIR cluster, we hypothesized that AIRs within clusters corresponding to important features might be more likely to target antigens that are specific to the disease in question. To test this hypothesis, we again examined the NSCLC dataset, for which we had the largest amount of data. First, we ranked the PCO clusters by their feature importance (**Fig. 4A**) and examined the proportion of healthy and NSCLC-derived TCRs in each. We observed clusters with significantly more TCRs from healthy ("healthy-biased") and with significantly for TCRs from disease ("NSCLC-biased") using a t-test p-value cutoff of 0.001 (**Fig. 4B**). The features of these class-imbalanced clusters are shown as a heatmap in **Fig. 4C**. Next, using these class-imbalanced clusters, we searched two databases (McPAS[23] and VDJdb[24]), whose TCRs have been annotated by their targeted antigen and associated disease. Interestingly, we identified strong hits to cancer-associated antigens using TCRs from either the healthy- or NSCLC-biased clusters. Surprisingly, there were more hits from the healthy-biased clusters than NSCLC-biased clusters (**Fig. 5A)**. **Fig. 5B** illustrates some of these hits, which included antigens such as MLANA[25], EphA2[26], BST2[27], TKT[28] and IGF2BP2[29], that have been reported as prognostic markers for NSCLC.

Identification of TCRs targeting cancer-associated antigens could be of interest not only for diagnostic purposes but also for the discovery of cancer therapeutics. While it is easy to imagine why TCRs from NSCLC-biased clusters would target cancer-associated antigens, the even higher rate of hits from healthy-biased clusters is more difficult to rationalize. We speculated that T cells that target cancer-associated antigens might appear high in healthy donors if, in NSCLC patients, they migrated out of peripheral blood toward tissues where tumors were present (i.e. lungs). Indeed, it is known that such tumor infiltrating lymphocytes (TILs) are observed in almost all cancers[30]. To test this idea, we repeated the database queries using randomly selected TCRs from healthy donors. Compared to TCRs from healthy-biased clusters, we found dramatically fewer hits using randomly selected healthy TCRs (Z-score = 75), as shown in **Fig. 5C**. In contrast, TCRs from NSCLC-biased clusters exhibited a similar level of cancer-associated hits to randomly selected TCRs (Z-score = 1.41), as shown in **Fig. 5D**. These observations were not qualitatively sensitive to the similarity threshold used to define a hit (**Fig. S4**). We next performed an analogous analysis

of Colorectal Cancer (CRC) cases (Fig. **S5-6**). Consistently, TCRs from healthy-biased clusters had significantly more cancer-related hits than randomly selected healthy TCRs (Z-score = 39.5) (**Fig. S6C**), regardless of the similarity threshold (**Fig. S7**). A recent report showed that several cancer-specific TCRs could target multiple tumor types via HLA A*02:01-restricted epitopes EAAGIGILTV, LLLGIGILVL, and NLSALGIFST from Melan A (MLANA), BST2, and IMP2 (IGF2BP2) and that PBMCs of healthy donors express such TCRs[31]. Consistently, many TCRs from healthy-biased important clusters were predicted to target MLANA and BST2 or MLANA and IGF2BP2 via highly similar epitopes in both the NSCLS (**Fig. S8**) and CRC (**Fig. S9**) datasets. Taken together, these results strongly suggest that TCRs in the healthy-biased clusters have migrated out of the peripheral blood to infiltrate tumors. This is an important finding, as it has become common practice in repertoire analysis to focus on clusters that are over-represented in patients[32,33]. Our results suggest that, in contrast, clusters that are over-represented in healthy donors contain many tumor-specific TCRs, potentially providing a simple way of identifying therapeutic T cells.

Discussion

Medical diagnosis has been an enduring challenge throughout history. As early as 16th century BCE, Egyptians documented methods for attributing an established disease as the primary cause for a set of symptoms, and this approach forms the foundation of modern differential diagnosis algorithms[34]. However, the landscape of medical diagnosis is rapidly evolving with the increasing prevalence of routine health checks that generate extensive standardized datasets from large populations[35]. This proliferation of data-driven approaches is poised to reshape the future of medical diagnosis, offering significant benefits for both individuals and healthcare providers. These include enhanced precision and accuracy[36], early detection of abnormalities even in the absence of symptoms[37], and personalized diagnoses tailored to an individual's medical history[38]. As a result, classical medical diagnosis methods are undergoing a disruptive transformation by leveraging recent breakthroughs in computational data science along with innovations in biotechnology.

AIRs are a particularly rich and emerging source of information that have not traditionally been used in medical diagnosis. A significant challenge in their diagnostic use is the extensive diversity and limited sharing of AIRs among donors. This stands in stark contrast to classical blood biomarkers where the chemical species in question can be assumed to exist in all donors and only its level requires quantification. In the case of AIRs, the natural inclination to gather more extensive and larger datasets for machine learning purposes is at odds with the goal of achieving widespread mutual sharing among all donors. This problem is exacerbated by the use of clonotype nomenclature, which leads to separation of AIRs into discrete clusters defined by their V and J gene usage and CDR3 length.

Our results demonstrate that an AIR representation that allows continuous similarity to be evaluated enables clusters to be framed within a vast network of similarities, resulting in a given AIR to have occupancy adjacency to multiple clusters. This, in turn, provides a solution to the donor sharing problem by providing non-zero occupancy for most clusters by most donors. This view of adaptive immunity makes sense from the point of view of host defense, as pathogens represent a diverse an unpredictable set of threats. Our findings show that, as features, dense features based on paratope clusters and occupancies generalize much better to new donors than sparse features based on clonotype cluster frequencies, and that these observations were consistent across three broad categories of disease. We found that there was obvious improvement with added training data, which is encouraging given the rapid growth of publicly-available AIR data.

In addition to these general observations, we found that, in the two cancer datasets clusters that corresponded to important features included both healthy- and disease-biased clusters. One possible explanation for the importance of healthy-biased clusters is that they are not relevant to cancer, and thus down-regulated in the patients. However, the opposite interpretation—that they are relevant to cancer but have migrated out of the peripheral blood (e.g. to the site of the tumor) is also conceivable. The latter explanation would be consistent with the observation that infiltrating lymphocytes are common in tumors of both of the cancers examined here. Although

we cannot know what happened to the T cells that are missing in the cancer patients, the extremely high hit rate to cancer-associated TCRs in healthy-biased compared to background healthy TCRs supports idea that these T cells left the periphery of cancer patients as a result of targeting the tumors. Moreover, the close resemblance of the TCRs identified here to those found to be reactive to multiple types of cancer[31] further strengthens this argument. The notion that tumors can affect global migration from peripheral blood has been demonstrated in brain cancer and metastatic melanomas[39]. We thus believe that TCR-based classification of cancer from PBMCs will have applications not only for diagnosis but also for discovery of broadly protective anti-cancer T cells.

Methods

AIR datasets

AIRR data were collected from published studies covering a wide range of diseases and included both BCRs and TCRs (**Table S1**). In order to minimize batch effects (i.e., any bias other than the one of interest), we sought studies that included both healthy and disease cohorts. Training (70%) and test (30%) splitting was performed at the level of donors. We note that, among the selected studies, the HIV dataset was significantly larger than the others in terms of the number of donors and depth of sequencing per donor. For non-small cell lung cancer (NSCLC), we were unable to obtain both healthy and disease data from the same studies, so we instead utilized several studies for healthy and disease data. This allowed us to perform training and test splitting at the study level for NSCLC.

AIR representation

Clonotyping is by far the most common way of describing AIRs[1,5,40-42]. In the clonotype representation, each AIR is represented by its V gene, J gene, and CDR3 sequence. Alternatively, AIRs can be described by their three CDR sequences, which we refer to here as the 'paratope' representation. Paratope sequences have been used to functionally cluster both BCRs[43] and TCRs[7].

Here, a single pseudo sequence was constructed by concatenating the three CDR segments of the amino acid sequence.

Comparison with other repertoire classification methods

To systematically evaluate the performance of repertoire classification and disease diagnosis, two previously-published methods, DeepRC[10] and immuneML[9], were evaluated. DeepRC and immuneML were applied to either BCR or TCR repertoire classification tasks using amino acid sequences of the CDR3 region with or without V/J gene information as input.

DeepRC utilizes a Deep Learning framework that integrates transformer-like attention, or, equivalently, modern Hopfield networks, into the learning process. The DeepRC software was obtained from https://github.com/ml-jku/DeepRC. Sequence embedding methods, including Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), were performed independently in the DeepRC runs. Two important parameters "n_updates" and "evaluate_at" were set to 10^5 and 5^3, respectively, following author recommendations. Other parameters were set to default values.

immuneML is well-established software suite for immune repertoire analysis powered by machine learning. immuneML (v 2.2.6) was obtained from https://github.com/uio-bmi/immuneML. Three repertoire encoding methods including Word2Vec, Kmer Frequency, and Evenness Profile, were utilized in combination with four machine learning methods, including Logistic Regression (LR), Random Forest (RF), support vector machine (SVM), and K-nearest neighbors (KNN). The Sequence Abundance for repertoire encoding and Probabilistic Binary Classifier (PBC) were also independently assessed in the context of immuneML. Parameters for encoding and machine learning were optimized using a grid search following the provided documentation.

Clustering

We used two different ways of clustering AIRs and compared the performance of both approaches.

Clonotype-based clustering: To cluster clonotypes, AIRs with identical V gene, J gene, and CDR3 length were first combined; then, the AIRs within these groups were clustered by their CDR3 amino acid sequences. Previous clonotype clustering studies have used different methods for such CDR3 clustering. For example, a sequence identity threshold of 90% was used in a previous study[41] to identify BCRs enriched in HIV-positive donors. Alternatively, another report[5] used a Hamming distance cutoff of 0.15 in the CDR3 sequences to stratify celiac disease patients. In yet another study[42], a Levenshtein distance cutoff of 2 was used and the CDR3 length constraint was removed in order to analyze differences in the sharing of clonotypes between COVID-19 and healthy donors. Here, we used an 80% sequence identity and 100% alignment coverage as implemented by CD-HIT[44].

Paratope-based clustering: Paratope pseudo sequence clustering was carried out using the MMseqs2 (version 13.45111) cluster method[45] with a sequence identity threshold of 80% and a coverage threshold of 90%. The antigen and epitope specificity of this method of clustering BCRs was experimentally shown to be >95%[46]. A cluster of CDR3 sequences defined by an 80% sequence identity and an identical length criterion is significantly more permissive compared to a paratope alignment with an identical identity threshold and 90% coverage. This discrepancy arises from the extended length characteristic of paratope sequences. Within the clonotype methodology, the leniency afforded by the CDR3 alignment is counterbalanced by the stringent necessity for exact matches in V and J gene names. Conversely, in the paratope-based strategy, where gene names are not taken into consideration, the responsibility for specificity falls entirely upon the quality of the paratope sequence alignment.

AIR adjacency matrices

AIR adjacency matrices were computed from paratope pseudo sequences using the MMseqs2 search method[45], with a sequence identity threshold of 80% and a coverage threshold of 90%.

Let us assume there are a total of M distinct AIR sequences for D donors. For any given pair of AIRs denoted by a_i and a_j let us denote the sequence identity by p_{ij} and the coverage by q_{ij}. Now, elements of the adjacency matrix $A^{M\times M}$ can be defined by equation (1):

$$A_{ij} = \begin{cases} 1 \; if \; p_{ij} > 0.8 \; and \; q_{ij} > 0.9 \\ 0 \qquad\qquad\qquad otherwise \end{cases} \qquad \text{where } i \in [1, M] \; and \; j \in [1, M] \qquad (1)$$

Where A_{ij} denotes the element in row i and column j. By definition, the diagonal elements of A are equal to unity.

Cluster Frequencies

In order to construct a feature vector for each donor we first clustered all of the AIR data for a given disease dataset and described each donor by the frequency of AIRs in the resulting clusters. Hence the feature vector for each donor is essentially a numerical representation that captures the distribution of their AIRs across the different clusters in the dataset.

Let us denote an AIR dataset by X_{AIR} and let us assume there are N clusters defined as $\{S_1 \cdots S_N\}$ after performing the clustering as defined above. Then, each of the M AIRs belongs to one of N clusters. Formally, we can define a set of distinct AIRs as $X_{AIR} = \{a_1 \cdots a_M\}$ where:

1. For all $a_i \in X_{AIR}$ there exists a $k \in [1, N]$ such that $a_i \in S_k$.
2. For all $a_i, a_j \in X_{AIR}$, $a_i = a_j$ if and only if $i = j$.

Let $C^{M\times N}$ be a $M \times N$ matrix that contains the one-hot encoding of the cluster assignment for each AIR, such that each element $C_{ik} \in C^{M\times N}$ is defined as follows:

$$C_{ik} = \begin{cases} 1 & if \; a_i \in S_{k,} \\ 0 & otherwise \end{cases} \qquad \text{where } i \in [1, M] \; and \; k \in [1, N].$$

That is, if AIR a_i belongs to cluster S_k, C_{ik} is equal to one, and zero for all other values of k. If the set of rows corresponding to donor D is defined as X_{AIR}^D, the cluster frequency CF of a given cluster S_k for donor D can be defined as:

$$CF_k^D = \sum_{i \in X_{AIR}^D} C_{ik} \quad \text{where } k \in [1, N] \qquad (2)$$

Now, the feature vector for a donor D is defined as $\{CF_1^D, \cdots, CF_k^D\}$. In practice, the frequency distribution of AIR sequences across different clusters follows a longtail distribution (**Fig. S1D**). Furthermore, each AIR is usually accompanied by a "clone count" that adds a scalar multiplicative factor to each row, which we ignore here for simplicity. The procedure above is illustrated in **Fig. 1C**.

As mentioned above, we can perform the clustering in two different ways and hence can produce two different representations of feature vectors for each donor. We refer to features based on clonotype clusters as clonotype cluster frequencies (CCFs) and those based on paratope clusters and paratope cluster frequencies (PCFs).

Cluster occupancies

Because of the longtail distribution of AIR sequences, the size distributions of clonotype or paratope clusters also follow a longtail distribution, resulting in a high-dimensional feature space. On top of this, the use of one-hot encoding results in features that are very sparse. We hypothesized that reducing sparsity would be an effective strategy for downsizing the features and reducing sparsity, as discussed in the next section. To this end, we utilized the adjacency matrix to identify the 'occupancy' of a given AIR in a given cluster. The occupancy matrix $O^{M\times N}$ is defined as the matrix multiplication of the adjacency matrix $A^{M\times M}$ and cluster matrix $C^{M\times N}$ containing one-hot cluster encoding. Formally, it can be written as $O^{M\times N} = A^{M\times M}\, C^{M\times N}$, where each element $O_{ik} \in O^{M\times N}$ can be defined as:

$$O_{ik} = \sum_j^M A_{ij}\, C_{jk} \qquad (3)$$

and A_{ij} is as defined by equation (1). If the adjacency matrix A contains nonzero elements that connect AIRs from different clusters, $O^{M\times N}$ will be less sparse than $C^{M\times N}$. Moreover, similar to equation (2), we can compute the overall cluster occupancy CO_k^D of cluster S_k for a given donor D as follows:

$$CO_k^D = \sum_{i \in X_{AIR}^D} O_{ik} \qquad (4)$$

Now, the cluster occupancy-based feature vector for a donor D is defined as $\{CO_1^D, \cdots, CO_k^D\}$. We refer to cluster occupancy-based features derived from clonotype clusters and paratope clusters as clonotype cluster occupancies (CCOs) and paratope cluster occupancies (PCOs), respectively. The occupancy calculation is illustrated in **Fig. 1B**.

Cluster Downsizing

Using all the cluster information for creating the feature vectors will often result in a high-dimensional feature space. This is mainly because the cluster sizes follow a longtail distribution resulting in a large number of clusters with very few members. Additionally, since the time complexity of calculating the adjacency matrix is quadratic in nature, it becomes increasingly difficult to calculate the full matrix when there are a large number of AIRs. To tackle both problems, we propose an algorithm for downsizing the number of clusters and thus reducing the number of AIRs and features for each donor. That is, downsizing the number of clusters is intended to retain potentially meaningful AIRs for feature generation. To this end, we retain only those clusters that meet a donor diversity criterion, as described below.

We first define the set of binary classes as $Y = \{y_1, y_2\}$ where y_1 denotes the class of healthy donors and y_2 denotes the class of donors with a disease. Next, we let the number of donors that belongs to class y_p be defined by N_p where $p \in [1,2]$. The number of donors that belong to class y_p and who have AIRs in cluster S_k can be defined by $n_p(S_k)$. Then, we define the cluster ratio tuple of a cluster S_k as ($r_k^{y_1}, r_k^{y_2}$) where $r_k^{y_p} = \frac{n_p(S_k)}{N_p}$. In other words, $r_k^{y_p}$ is the ratio of the

number of donors of class y_p in cluster S_k to the total number of donors in class y_p. We next introduce a threshold parameter d_{min}, that defines a minimum ratio, below which all clusters will be dropped and the respective AIRs ignored. For a cluster to be selected at least one value of the cluster ratio tuple should be more than d_{min} . More concretely, for a cluster S_k if $max\left(r_k^{y_1}, r_k^{y_2}\right) < d_{min}$ then it will be dropped. This downsizing is applied to all datasets in the context of training with d_{min} used as a hyperparameter. In addition, for the very large HIV dataset, we used this downsizing as a pre-processing step, since the adjacency calculation was prohibitively time-consuming for the entire dataset. It is important to note that such downsizing was performed after splitting the data into training and test groups; that is, only the labels of training AIRs were used to drop clusters.

Classification

For the binary classification task, we used the XGBoost algorithm via the XGBClassifier (version 1.6.2) Python API. The training was performed on a randomly selected 70% of donors, and, with the exception of weighting positive classes by the ratio of true to false training donors, all parameters were set to default values. Classification performance was evaluated on the remaining 30% split test data using the area under the receiver operating characteristic curve (ROC AUC) or the area under the precision-recall curve (PR AUC) as implemented in the Scikit-Learn Python package (version 1.0.2). Four workflows were constructed that differed only in the clustering method (clonotype or paratope) and feature construction (cluster frequencies or cluster occupancies), as shown in **Fig. 1C**.

Statistical analysis and Visualization

T-test and Z-score calculations were performed using the Python package Scipy. The Python packages Seaborn and Matplotlib were used for preparing graphs. The Chord diagram was made using the Python package Pycirclize. Sequence logos were depicted by WebLogo (https://weblogo.berkeley.edu/logo.cgi).

Healthy and disease-biased clusters

After picking the best d_{min} value, we divided important clusters into those with significantly more TCRs from one class (healthy or disease) using a t-test p-value cutoff of 0.001. We then ranked the healthy or disease-biased clusters based on their p-values to identify the 100 most-significant healthy- or disease-biased clusters for downstream cancer-associated sequence annotation.

Cancer-associated TCR datasets

We collected TCR sequences that were annotated to target cancer-associated antigens in VDJdb (https://vdjdb.cdr3.net/) or McPAS (http://friedmanlab.weizmann.ac.il/McPAS-TCR/). After preprocessing, 3654 and 951 cancer-associated TCR beta chains were independently prepared from VDJdb and McPAS, respectively. These 4605 TCR beta chains were used for searching and annotating the query input by the following criteria: sequence identity of CDR1 >= 90%, CDR2 >= 90%, CDR3 >= 80%; alignment coverage >= 90%. A hit was defined as possessing at least one matching cancer-associated TCR from VDJdb or McPAS.

Cancer-associated TCR background comparison

To fairly compare the hit ratio of cancer-associated TCRs from a given set of clusters with the background of the same class of donors (healthy or disease), we constructed a set of queries by randomly selecting an identical number of sequences from the corresponding donors as those in the cluster. The selection was repeated 100 times with different random seeds. We calculated the Z-score based on the mean and standard deviation of the hit ratio of these 100 background database queries and the hit ratio of TCRs from the clusters of interest.

Figure Captions

Figure 1. **Strategy of cluster frequencies and cluster occupancies in classification algorithm. A,** Cluster frequencies are illustrated by a simple example of two donors (A and B) whose AIRs form 6 clusters, which are one-hot encoded in the matrix *C*. Separately summing the rows of this matrix that belong to each donor yields a cluster frequency feature. **B,** Cluster occupancies are derived

similarly to cluster frequencies except we introduce an adjacency matrix that describes the pairwise similarities between each AIR. If there are similarities between AIRs belonging to different clusters, the occupancy matrix *O* will differ from *C.* Summing the rows for each donor yields a feature vector that, in general, will be less sparse than that derived from cluster frequencies. **C,** The flowchart of classification pipeline. Starting from a set of training and test donors, the first step is feature generation using one of four methods; thereafter, the remaining steps are identical and consist of training/hyperparameter selection and testing the classifier on the test features. The four methods of feature generation differ in the type of clustering (clonotype or paratope) and whether the clusters are transformed into features as frequencies or occupancies.

Figure 2. **Assessment of COVID-19 diagnosis based on BCRs. A,** The sparsity of each of the four feature matrices is shown as a function of d_{min}, with the vertical cyan line showing the value from training LOOCV hyperparameter optimization, with the pink line indicating the target function (mean of training ROC AUC and PR AUC). **B,** ROC curves for the LOOCV and test predictions using the PCO classifier. **C,** Test ROC AUCs for all four classifiers. **D,** Test PR AUCs for all four classifiers, with a random classifier indicated as a grey bar. **E,** Impact of occupancy on important features. The clonotype cluster corresponding to the most important feature in the COVID-19 CCO training (cluster 1) has three members representing two donors (Donor 10 and Donor 32). The small size, and low sharing of cluster 1 will result in a sparse CCF feature. **F,** AIRs with similar paratopes from other clonotype clusters are visualized as a chord diagram, where each color denotes a different donor. The original two donors are colored purple and cyan as in the clonotype cluster. **G,** The high degree of paratope-level similarity across multiple BCRs from multiple donors results in a dense CCO feature vector.

Figure 3. **Assessment of NSCLC diagnosis based on TCRs.** A, The sparsity of each of the four feature matrices is shown as a function of d_{min}, with the vertical cyan line showing the value from training LOOCV hyperparameter optimization, with the pink line indicating the target function (mean of training ROC AUC and PR AUC). **B,** ROC curves for the LOOCV and test predictions using

the PCO classifier. **C,** Test ROC AUCs for all four classifiers. **D,** Test PR AUCs for all four classifiers, with a random classifier indicated as a grey bar. **E,** Summary of 12 ROC AUC values on test sets of NSCLC using different combinations of studies for training and testing. **F,** PR AUC values on the same 12 test sets.

Figure 4. **Assessment of cluster features in NSCLC patient classification**. **A**, Feature Importance. The Y-axis shows XGBoost feature importance in descending order. **B**, Top-six most important features. Each dot represents the ratio of healthy (HD) and disease (NSCLC) donors. The t-test was performed to compare the ratios between HD and NSCLC cohorts (***$p \leq 0.001$). **C**, The heatmap of top 100 diseases and healthy (HD) significant cluster features. The *p*-value was calculated by the t-test. The rank of the feature was set by the *p-value* in ascending order. The value of each cell is the normalized ratio from 0 (white) to 1 (blue). The left Y-axis showed each donor from healthy and NSCLC cohorts grouped by the corresponding study.

Figure 5. **Functional annotation of healthy- and NSCLC-biased clusters**. **A**, Histogram of hits to TCRs targeting the indicated antigens from healthy (blue bars) and NSCLC-biased (pink bars) clusters after querying the McPAS and VDJdb cancer databases. Each count represents one database hit. **B**, Sequences from important clusters matching known cancer-targeting antigens, as shown by their alignments and sequence logos based on all members of the same cluster. **C**, Hit rates of TCRs from healthy-biased clusters to cancer-targeting TCRs (vertical read line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical line representing mean). The horizontal arrow indicates the Z-score. **D,** Hit rates of TCRs from NSCLC-biased clusters to cancer-targeting TCRs (vertical read line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical line representing mean). The horizontal arrow indicates the Z-score.

Figure S1. **Adaptive immune receptors. A,** B cell receptors (also known as antibodies in their soluble form) consist of two heavy and two light chains, while T cell receptors (TCRs) consist of a single beta and a single alpha chain. BCR and TCR coding sequences are generated by

combinatorial rearrangement of V D and J genes, which result in diverse complementarity determining regions (CDRs 1-3). **B,** CDRs 1-3 are arranged to form a continuous molecular surface called a "paratope" (in blue). **C,** Contacts with antigen (light blue) are primarily restricted to CDRs compared with the background assessable surface (light brown). **D,** The distribution of BCRs and TCRs in each donor follow a longtail distribution, with the vast number of clones having very low frequencies (clones with counts less than 10 not shown).

Figure S2. **Assessment of HIV diagnosis based on BCRs and AIH diagnosis based on TCRs**. **A/E,** The sparsity of each of the four feature matrices is shown as a function of d_{min}, with the vertical cyan line showing the value from training LOOCV hyperparameter optimization, with the pink line indicating the target function (mean of training ROC AUC and PR AUC). **B/F,** ROC curves for the LOOCV and test predictions using the PCO classifier. **C/G,** Test ROC AUCs for all four classifiers. **D/H,** Test PR AUCs for all four classifiers, with a random classifier indicated as a grey bar. Impact of occupancy on important features.

Figure S3. **Assessment of T1D and CRC diagnosis based on TCRs**. **A/E,** The sparsity of each of the four feature matrices is shown as a function of d_{min}, with the vertical cyan line showing the value from training LOOCV hyperparameter optimization, with the pink line indicating the target function (mean of training ROC AUC and PR AUC). **B/F,** ROC curves for the LOOCV and test predictions using the PCO classifier. **C/G,** Test ROC AUCs for all four classifiers. **D/H,** Test PR AUCs for all four classifiers, with a random classifier indicated as a grey bar. Impact of occupancy on important features.

Figure S4**. Cancer-associated hit rates from healthy-biased clusters and NSCLC-biased clusters using various similarity thresholds.** TCRs from healthy-biased clusters had significantly more cancer-related hits than randomly selected background TCRs under three different CDR3 sequence identity thresholds (80%, 90% and 100%).

Figure S5. **Assessment of cluster features in CRC patient classification**. **A**, Feature Importance. The Y-axis shows XGBoost feature importance in descending order. **B**, Top-six most important features. Each dot represents the ratio of healthy (HD) and disease (CRC) donors. The t-test was performed to compare the ratios between HD and CRC cohorts (***$p \leq 0.001$). **C**, The heatmap of top 100 CRC and healthy (HD) biased cluster features. The *p*-value was calculated by the t-test. The rank of the feature was set by the *p-value* in ascending order. The value of each cell is the normalized ratio from 0 (white) to 1 (blue). The left Y-axis shows each donor from healthy and CRC cohorts grouped by the corresponding study.

Figure S6. **Functional annotation of healthy- and CRC-biased clusters**. Histogram of hits to TCRs targeting the indicated antigens from healthy (blue bars) and CRC-biased (pink bars) clusters after querying the McPAS and VDJdb cancer databases. Each count represents one database hit. **B**, Sequences from important clusters matching known cancer-targeting antigens, as shown by their alignments and sequence logos based on all members of the same cluster. **C**, Hit rates of TCRs from healthy-biased clusters to cancer-targeting TCRs (vertical read line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical line representing mean). The horizontal arrow indicates the Z-score. **D,** Hit rates of TCRs from CRC-biased clusters to cancer-targeting TCRs (vertical read line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical line representing mean). The horizontal arrow indicates the Z-score.

Figure S7**. Cancer-associated hit rates from healthy-biased clusters and CRC-biased clusters using various similarity thresholds.** TCRs from healthy-biased clusters had significantly more cancer-related hits than randomly selected background TCRs under three different CDR3 sequence identity thresholds (80%, 90% and 100%).

Figure S8. **Predicted MLANA, BST2, and IGF2BP2-targeting TCRs**. **A,** the upset plot shows the TCR hit number of predicted MLANA, BST2, and IGF2BP2-targeting TCRs from healthy- (left) and

NSCLC-biased clusters (right) under CDR3 sequence identities 80 or 90 %. Some TCRs are predicted to target two antigens. **B,** Sequence alignments of query-template hits indicate that multiple healthy donors harbor TCRs predicted to target two cancer-related antigens.

Figure S9. **Predicted MLANA, BST2, and IGF2BP2-targeting TCRs**. **A,** the upset plot shows the TCR hit number of predicted MLANA, BST2, and IGF2BP2-targeting TCRs from healthy- (left) and CRC-biased clusters (right) under CDR3 sequence identities 80 or 90 %. Some TCRs are predicted to target two antigens. **B,** Sequence alignments of query-template hits indicate that multiple healthy donors harbor TCRs predicted to target two cancer-related antigens.

References

1 Chen, Y. *et al.* A Deep Learning Model for Accurate Diagnosis of Infection Using Antibody Repertoires. *J Immunol* **208**, 2675-2685 (2022). https://doi.org:10.4049/jimmunol.2200063

2 Foers, A. D. *et al.* Classification of intestinal T-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. *J Pathol* **253**, 279-291 (2021). https://doi.org:10.1002/path.5592

3 Ostrovsky-Berman, M., Frankel, B., Polak, P. & Yaari, G. Immune2vec: Embedding B/T Cell Receptor Sequences in R (N) Using Natural Language Processing. *Front Immunol* **12**, 680687 (2021). https://doi.org:10.3389/fimmu.2021.680687

4 Park, J. J. *et al.* Machine learning identifies T cell receptor repertoire signatures associated with COVID-19 severity. *Commun Biol* **6**, 76 (2023). https://doi.org:10.1038/s42003-023-04447-4

5 Shemesh, O., Polak, P., Lundin, K. E. A., Sollid, L. M. & Yaari, G. Machine Learning Analysis of Naive B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. *Front Immunol* **12**, 627813 (2021). https://doi.org:10.3389/fimmu.2021.627813

6 Cinelli, M. *et al.* Feature selection using a one dimensional naive Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* **33**, 951-955 (2017). https://doi.org:10.1093/bioinformatics/btw771

7 Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89-93 (2017). https://doi.org:10.1038/nature22383

8 Eliyahu, S. *et al.* Antibody Repertoire Analysis of Hepatitis C Virus Infections Identifies Immune Signatures Associated With Spontaneous Clearance. *Front Immunol* **9**, 3004 (2018). https://doi.org:10.3389/fimmu.2018.03004

9 Pavlovic, M. *et al.* The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell* **3**, 936-944 (2021). https://doi.org:10.1038/s42256-021-00413-z

10 Widrich, M. *et al.* DeepRC: Immune repertoire classification with attention-based deep massive multiple instance learning. *bioRxiv*, 2020.2004.2012.038158 (2020). https://doi.org:10.1101/2020.04.12.038158

11 Khan, M. *et al.* Visualizing in deceased COVID-19 patients how SARS-CoV-2 attacks the respiratory and olfactory mucosae but spares the olfactory bulb. *Cell* **184**, 5932-5949 e5915 (2021). https://doi.org:10.1016/j.cell.2021.10.027

12 Hayden, M. K. *et al.* The Infectious Diseases Society of America Guidelines on the Diagnosis of COVID-19: Antigen Testing. *Clin Infect Dis* (2023). https://doi.org:10.1093/cid/ciad032

13 Edahiro, R. *et al.* Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nat Genet* (2023). https://doi.org:10.1038/s41588-023-01375-1

14 Gridelli, C. *et al.* Non-small-cell lung cancer. *Nature Reviews Disease Primers* **1**, 15009 (2015). https://doi.org:10.1038/nrdp.2015.9

15 Thunnissen, E. *et al.* The challenge of NSCLC diagnosis and predictive analysis on small samples. Practical approach of a working group. *Lung Cancer* **76**, 1-18 (2012). https://doi.org:10.1016/j.lungcan.2011.10.017

16 de Greef, P. C. *et al.* The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife* **9** (2020). https://doi.org:10.7554/eLife.49900

17 Schultheiss, C. *et al.* Next-Generation Immunosequencing Reveals Pathological T-Cell Architecture in Autoimmune Hepatitis. *Hepatology* **73**, 1436-1448 (2021). https://doi.org:10.1002/hep.31473

18 Swanson, P. A., 2nd *et al.* AZD1222/ChAdOx1 nCoV-19 vaccination induces a polyfunctional spike protein-specific T(H)1 response with a diverse TCR repertoire. *Sci Transl Med* **13**, eabj7211 (2021). https://doi.org:10.1126/scitranslmed.abj7211

19 Reuben, A. *et al.* Comprehensive T cell repertoire characterization of non-small cell lung cancer. *Nat Commun* **11**, 603 (2020). https://doi.org:10.1038/s41467-019-14273-0

20 Wu, L. *et al.* T-Cell Receptor Profiling and Prognosis After Stereotactic Body Radiation Therapy For Stage I Non-Small-Cell Lung Cancer. *Front Immunol* **12**, 719285 (2021). https://doi.org:10.3389/fimmu.2021.719285

21 Hamm, D. E. ImmuneAccess immunoSEQ hsTCRB-V4b Control Data. (2020). https://doi.org:https://doi.org/10.21417/ADPT2020V4CD

22 Joshi, K. *et al.* Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nat Med* **25**, 1549-1559 (2019). https://doi.org:10.1038/s41591-019-0592-2

23 Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924-2929 (2017). https://doi.org:10.1093/bioinformatics/btx286

24 Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* **46**, D419-D427 (2018). https://doi.org:10.1093/nar/gkx760
25 Der, S. D. *et al.* Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol* **9**, 59-64 (2014). https://doi.org:10.1097/JTO.0000000000000042
26 Brannan, J. M. *et al.* EphA2 in the early pathogenesis and progression of non-small cell lung cancer. *Cancer Prev Res (Phila)* **2**, 1039-1049 (2009). https://doi.org:10.1158/1940-6207.CAPR-09-0212
27 Suzuki, K. *et al.* Prognostic immune markers in non-small cell lung cancer. *Clin Cancer Res* **17**, 5247-5256 (2011). https://doi.org:10.1158/1078-0432.CCR-10-2805
28 Niu, C. *et al.* Transketolase Serves as a Biomarker for Poor Prognosis in Human Lung Adenocarcinoma. *J Cancer* **13**, 2584-2593 (2022). https://doi.org:10.7150/jca.69583
29 Han, L. *et al.* IGF2BP2 Regulates MALAT1 by Serving as an N6-Methyladenosine Reader to Promote NSCLC Proliferation. *Front Mol Biosci* **8**, 780089 (2021). https://doi.org:10.3389/fmolb.2021.780089
30 Brummel, K., Eerkens, A. L., de Bruyn, M. & Nijman, H. W. Tumour-infiltrating lymphocytes: from prognosis to treatment selection. *Br J Cancer* **128**, 451-458 (2023). https://doi.org:10.1038/s41416-022-02119-4
31 Dolton, G. *et al.* Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. *Cell* **186**, 3333-3349 e3327 (2023). https://doi.org:10.1016/j.cell.2023.06.020
32 Huang, C. *et al.* The landscape and diagnostic potential of T and B cell repertoire in Immunoglobulin A Nephropathy. *J Autoimmun* **97**, 100-107 (2019). https://doi.org:10.1016/j.jaut.2018.10.018
33 Liu, X. *et al.* T cell receptor beta repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Ann Rheum Dis* **78**, 1070-1078 (2019). https://doi.org:10.1136/annrheumdis-2019-215442
34 van Middendorp, J. J., Sanchez, G. M. & Burridge, A. L. The Edwin Smith papyrus: a clinical reappraisal of the oldest known document on spinal injuries. *Eur Spine J* **19**, 1815-1823 (2010). https://doi.org:10.1007/s00586-010-1523-6
35 Chen, P. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nat Mater* **18**, 410-414 (2019). https://doi.org:10.1038/s41563-019-0345-0
36 Song, Y. *et al.* Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **18**, 2775-2780 (2021). https://doi.org:10.1109/TCBB.2021.3065361
37 Liu, S. *et al.* in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI).* 1015-1018.
38 Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* **1**, 39 (2018). https://doi.org:10.1038/s41746-018-0040-6
39 Bochem, J. *et al.* Early disappearance of tumor antigen-reactive T cells from peripheral blood correlates with superior clinical outcomes in melanoma under anti-PD-1 therapy. *J Immunother Cancer* **9** (2021). https://doi.org:10.1136/jitc-2021-003439

40 Miho, E., Roskar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat Commun* **10**, 1321 (2019). https://doi.org:10.1038/s41467-019-09278-8

41 Roskin, K. M. *et al.* Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nat Immunol* **21**, 199-209 (2020). https://doi.org:10.1038/s41590-019-0581-0

42 Ruiz Ortega, M., Spisak, N., Mora, T. & Walczak, A. M. Modeling and predicting the overlap of B- and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals. *PLoS Genet* **19**, e1010652 (2023). https://doi.org:10.1371/journal.pgen.1010652

43 Richardson, E. *et al.* A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *MAbs* **13**, 1869406 (2021). https://doi.org:10.1080/19420862.2020.1869406

44 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012). https://doi.org:10.1093/bioinformatics/bts565

45 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017). https://doi.org:10.1038/nbt.3988

46 Saputri, D. S. *et al.* Deciphering the antigen specificities of antibodies by clustering their complementarity determining region sequences. *mSystems*, e0072223 (2023). https://doi.org:10.1128/msystems.00722-23

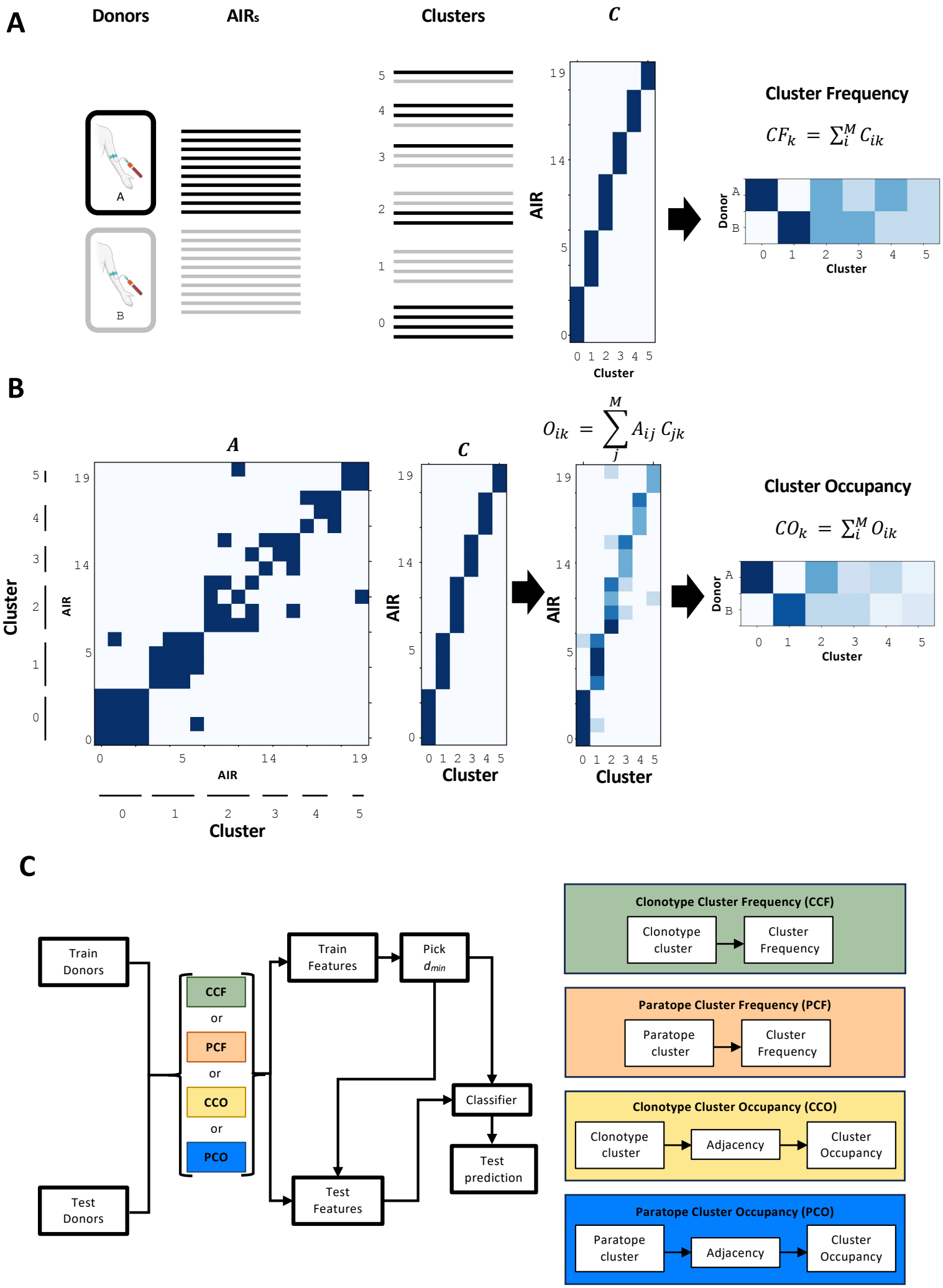

A
Donors
AIRs
Clusters
C
A
B
AIR
Cluster
Cluster Frequency
$CF_k = \sum_i^M C_{ik}$
Donor
Cluster
B
A
$O_{ik} = \sum_j^M A_{ij} C_{jk}$
Cluster Occupancy
$CO_k = \sum_i^M O_{ik}$
C
Train Donors
Test Donors
CCF
or
PCF
or
CCO
or
PCO
Train Features
Pick d_{min}
Test Features
Classifier
Test prediction
Clonotype Cluster Frequency (CCF)
Clonotype cluster
Cluster Frequency
Paratope Cluster Frequency (PCF)
Paratope cluster
Cluster Frequency
Clonotype Cluster Occupancy (CCO)
Clonotype cluster
Adjacency
Cluster Occupancy
Paratope Cluster Occupancy (PCO)
Paratope cluster
Adjacency
Cluster Occupancy

Figure 1

COVID-19

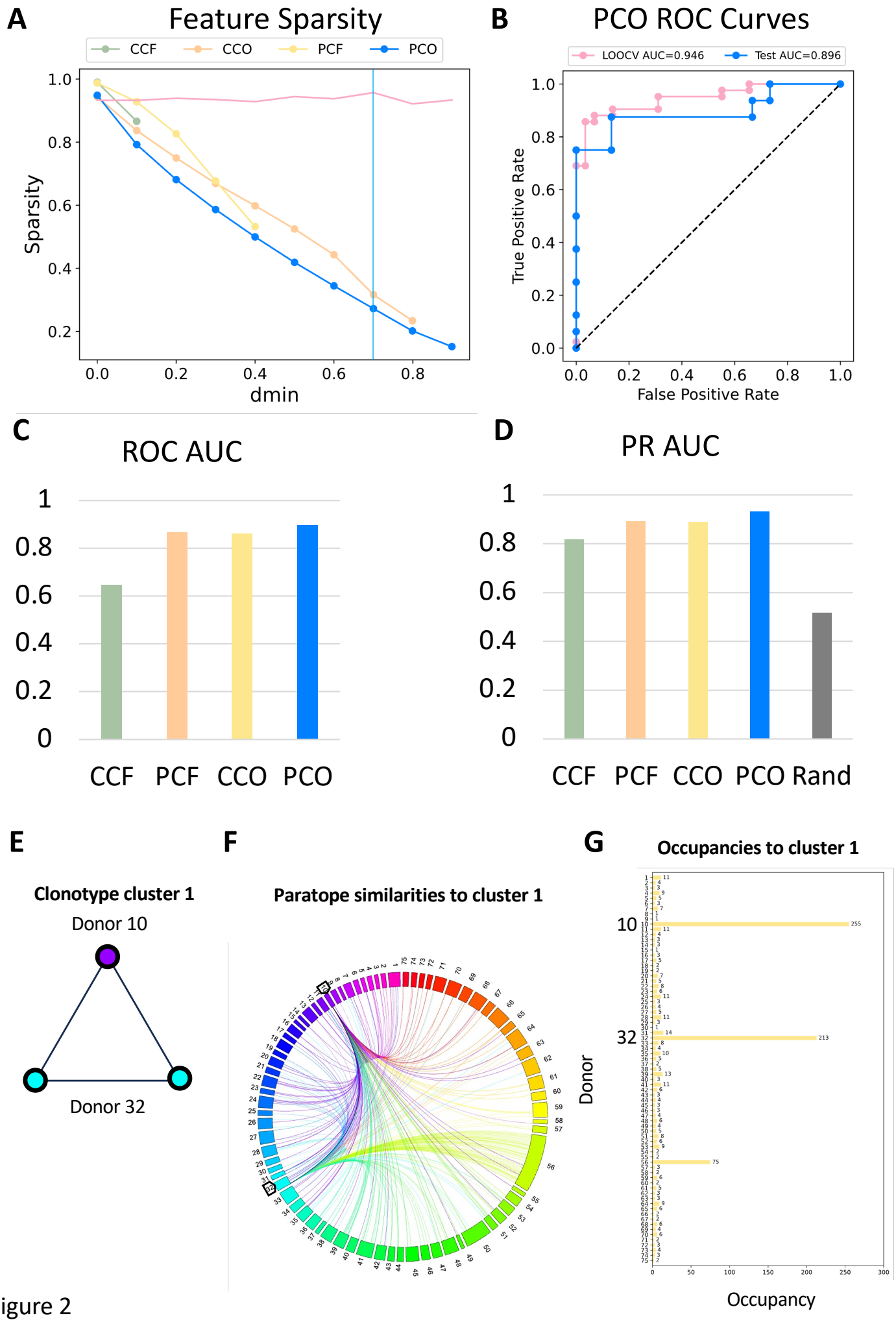


Figure 2

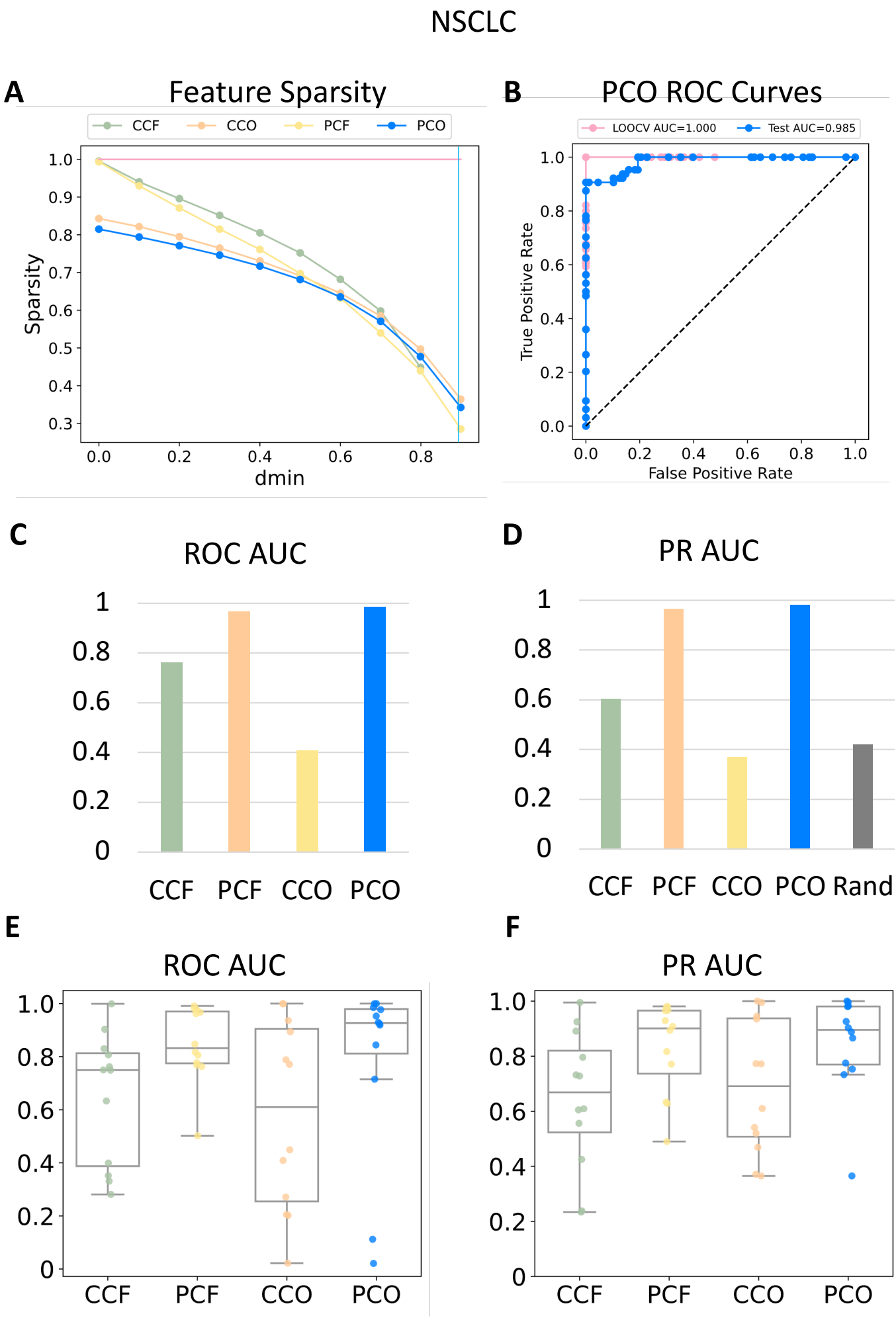

NSCLC
A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC Curves
LOOCV AUC=1.000
Test AUC=0.985
True Positive Rate
False Positive Rate
C
ROC AUC
CCF
PCF
CCO
PCO
D
PR AUC
CCF
PCF
CCO
PCO
Rand
E
ROC AUC
CCF
PCF
CCO
PCO
F
PR AUC
CCF
PCF
CCO
PCO

Figure 3

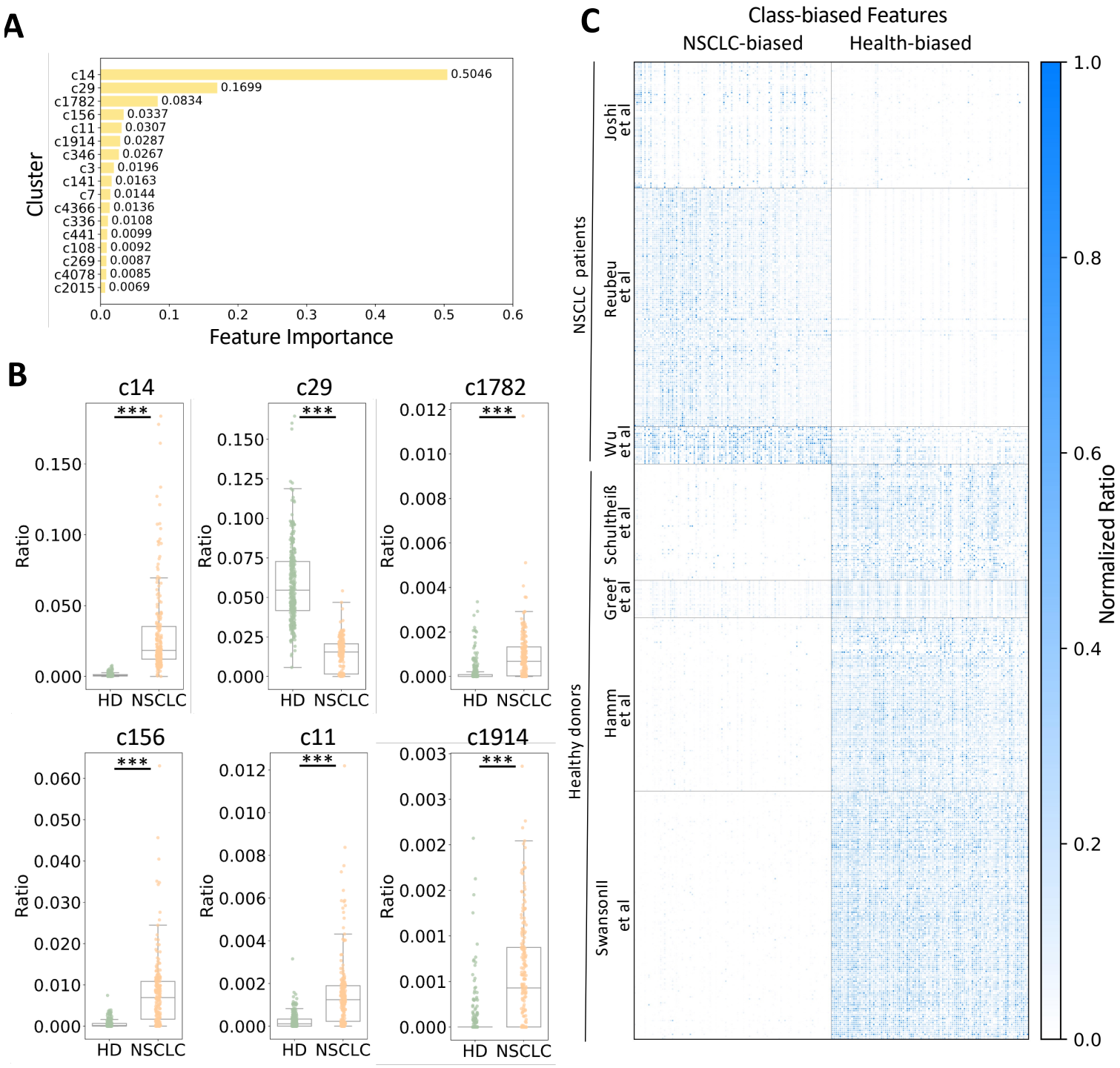
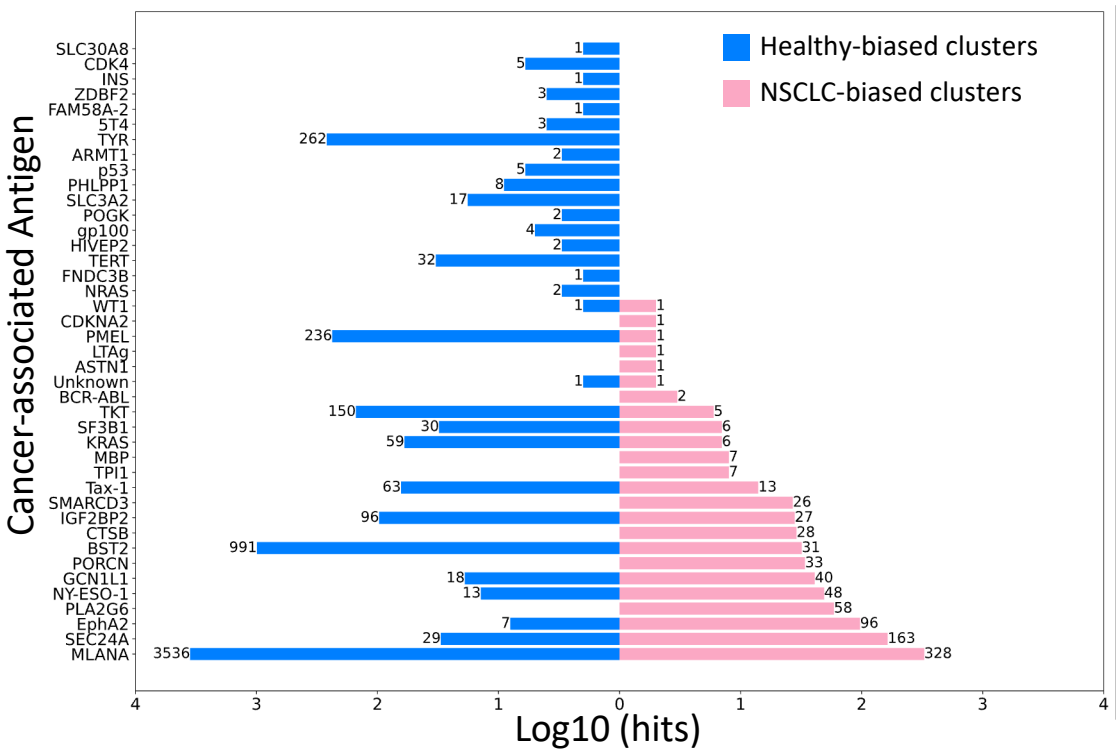
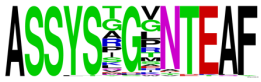

A
Cluster
c14 0.5046
c29 0.1699
c1782 0.0834
c156 0.0337
c11 0.0307
c1914 0.0287
c346 0.0267
c3 0.0196
c141 0.0163
c7 0.0144
c4366 0.0136
c336 0.0108
c441 0.0099
c108 0.0092
c269 0.0087
c4078 0.0085
c2015 0.0069
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Feature Importance
B
c14
c29
c1782
c156
c11
c1914
Ratio
***
HD NSCLC
C
Class-biased Features
NSCLC-biased
Health-biased
NSCLC patients
Healthy donors
Joshi et al
Reubeu et al
Wu et al
Schultheiß et al
Greef et al
Hamm et al
SwansonII et al
Normalized Ratio
1.0
0.8
0.6
0.4
0.2
0.0

Figure 4

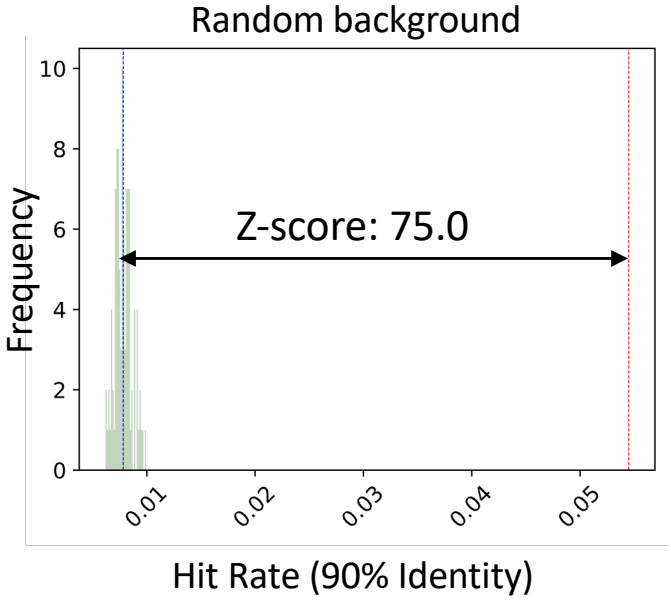
A



B

Cluster		V/ J gene	CDR3 (length)	Identity	Template Antigen	Cluster CDR3 Logo
c14	Query Template	TRBV9/TRBJ1-1 TRBV9/TRBJ1-1	ASSVD-GGGTEAF (12) ASSVDSGAGTEAF (13)	84%	MLANA	
c29	Query Template	TRBV6-5/TRBJ1-1 TRBV6-5/TRBJ1-1	ASSYSTGVNTEAF (13) ASSYSVGVNTEAF (13)	92%	SEC24A	
c1782	Query Template	TRBV6-1/TRBJ1-2 TRBV9/TRBJ1-1	ASRP-RTTNYGYT (12) ASRPSRGTNYGYT (13)	84%	ASTN1	
c156	Query Template	TRBV2/TRBJ2-7 TRBV2/TRBJ2-5	ASSGQGAGEQY (11) ASSGQGAGTQY (11)	90%	NY-ESO-1	
c11	Query Template	TRBV2/TRBJ2-7 TRBV2/TRBJ2-7	ASSLRTGFYEQY (12) ASSLRTGSYEQY (12)	91%	MLANA	
c1914	Query Template	TRBV6-1/TRBJ1-2 TRBV6-1/TRBJ1-2	ASSYSSVNYGYT (12) ASSY-SVNYGYT (11)	91%	PMEL	

C



D

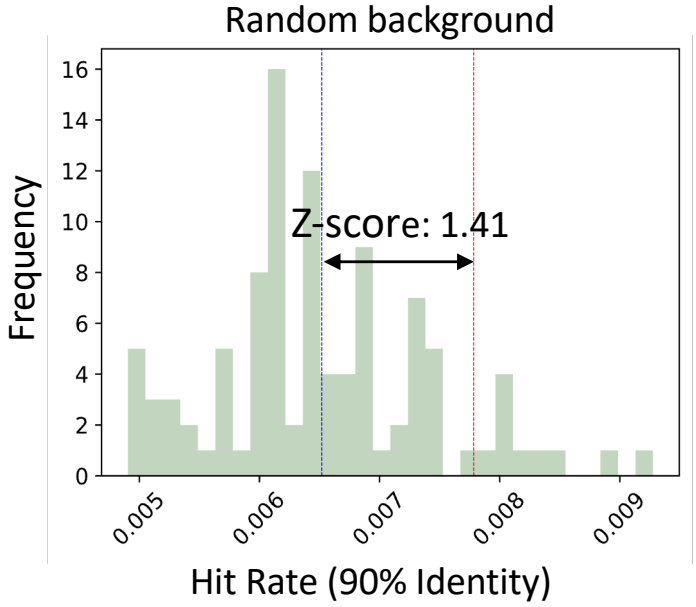
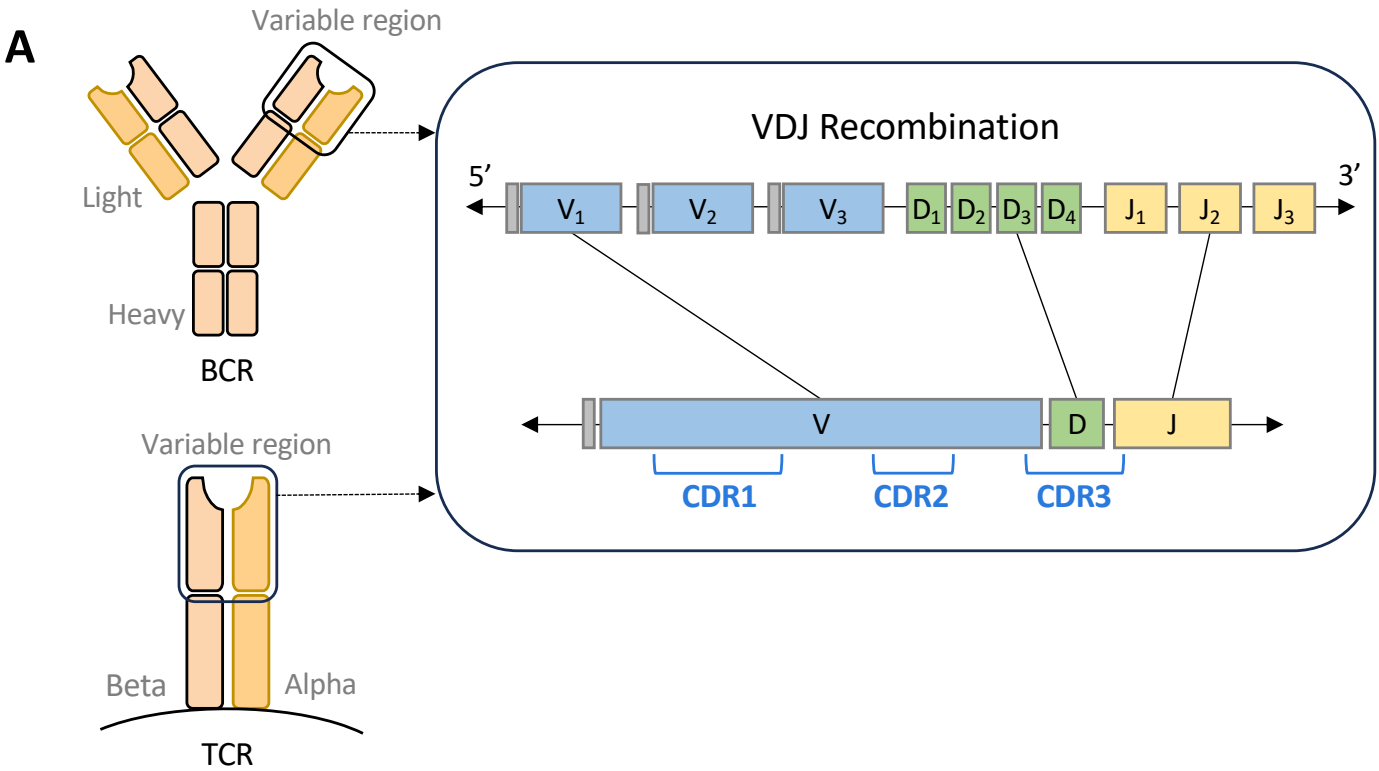
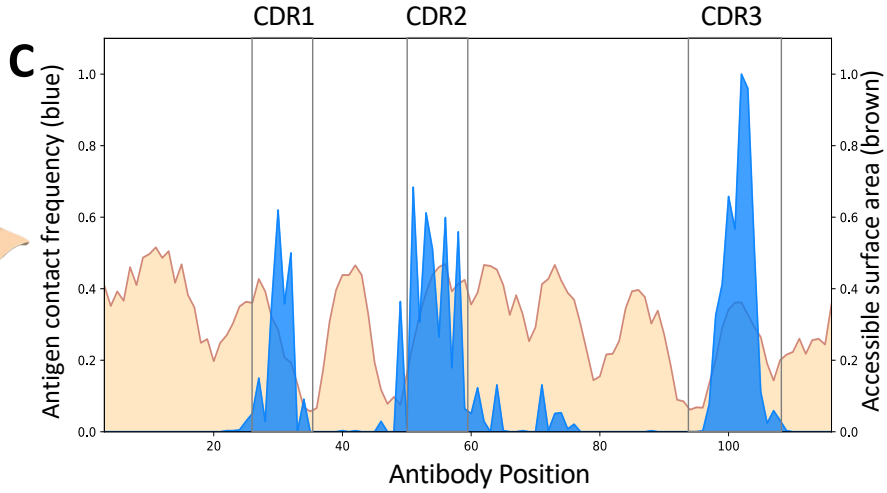


Figure 5


A
Variable region
Light
Heavy
BCR
Variable region
Beta
Alpha
TCR
VDJ Recombination
5'
3'
V1
V2
V3
D1
D2
D3
D4
J1
J2
J3
V
D
J
CDR1
CDR2
CDR3

B


C
CDR1
CDR2
CDR3
Antigen contact frequency (blue)
Accessible surface area (brown)
Antibody Position
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

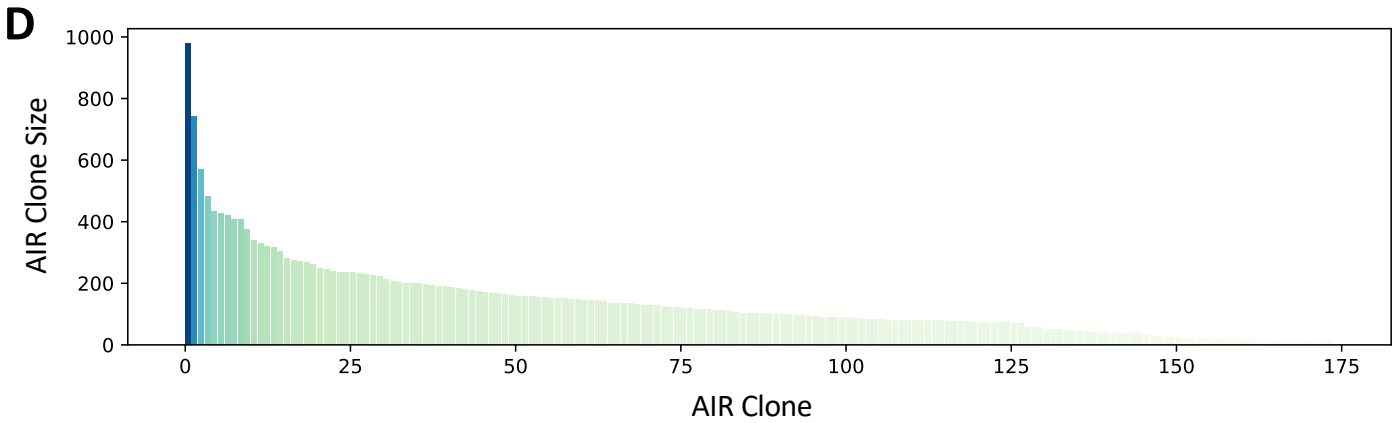

D
AIR Clone Size
AIR Clone
0
200
400
600
800
1000
0
25
50
75
100
125
150
175

Figure S1

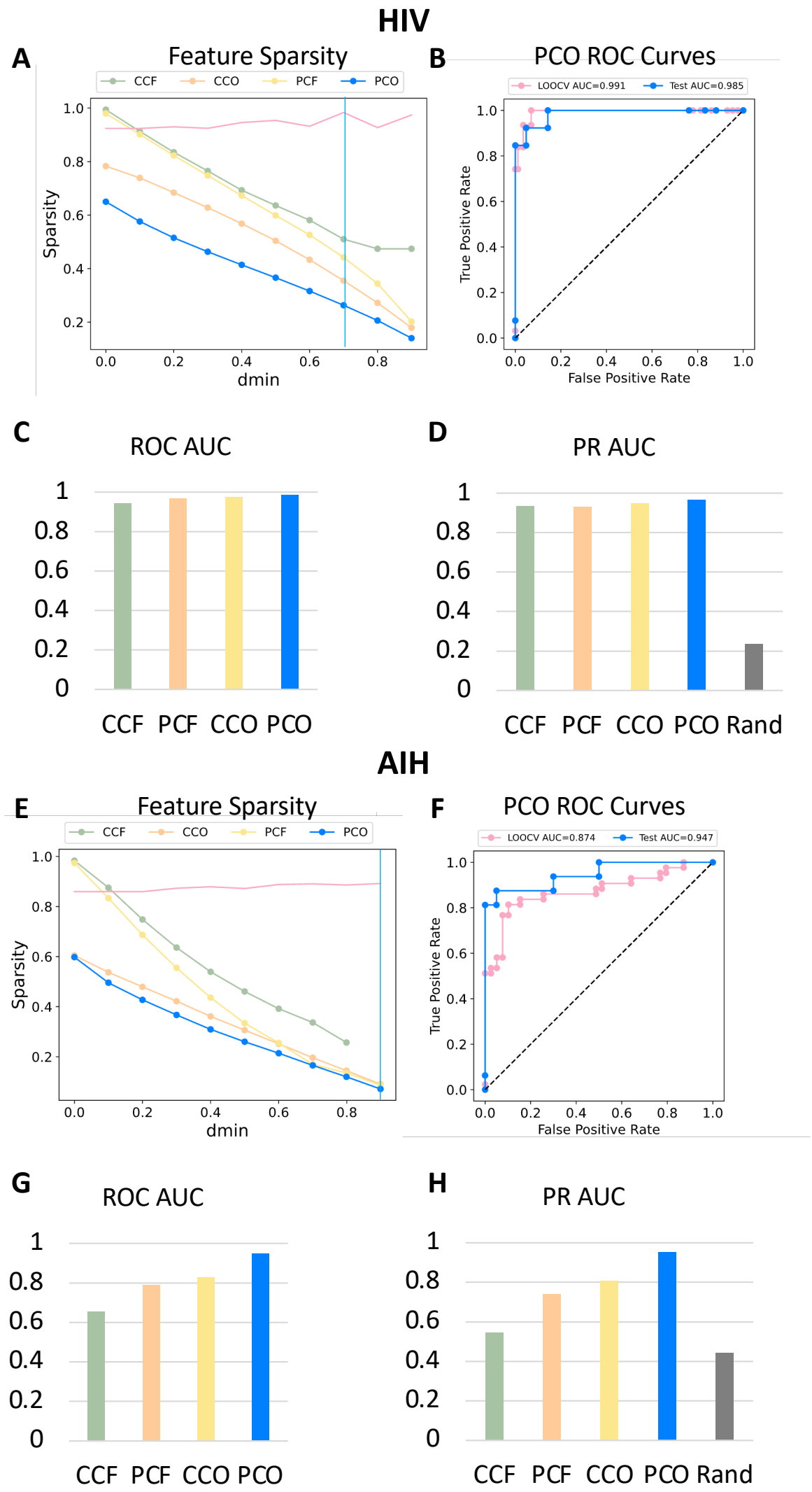

HIV
A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC Curves
LOOCV AUC=0.991
Test AUC=0.985
True Positive Rate
False Positive Rate
C
ROC AUC
CCF PCF CCO PCO
D
PR AUC
CCF PCF CCO PCO Rand
AIH
E
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
F
PCO ROC Curves
LOOCV AUC=0.874
Test AUC=0.947
True Positive Rate
False Positive Rate
G
ROC AUC
CCF PCF CCO PCO
H
PR AUC
CCF PCF CCO PCO Rand

Figure S2

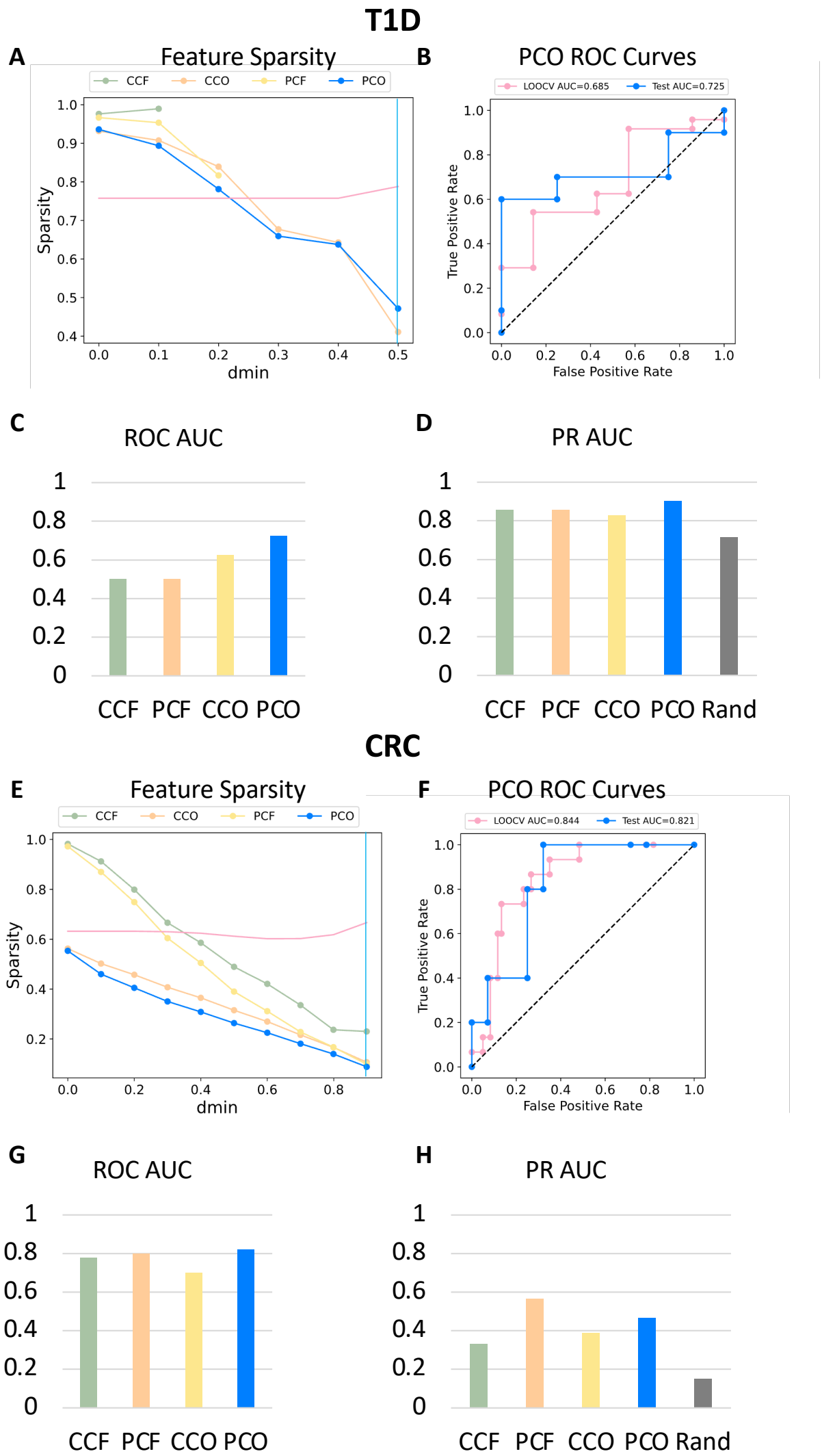

T1D
A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC Curves
LOOCV AUC=0.685
Test AUC=0.725
True Positive Rate
False Positive Rate
C
ROC AUC
CCF PCF CCO PCO
D
PR AUC
CCF PCF CCO PCO Rand
CRC
E
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
F
PCO ROC Curves
LOOCV AUC=0.844
Test AUC=0.821
True Positive Rate
False Positive Rate
G
ROC AUC
CCF PCF CCO PCO
H
PR AUC
CCF PCF CCO PCO Rand

Figure S3

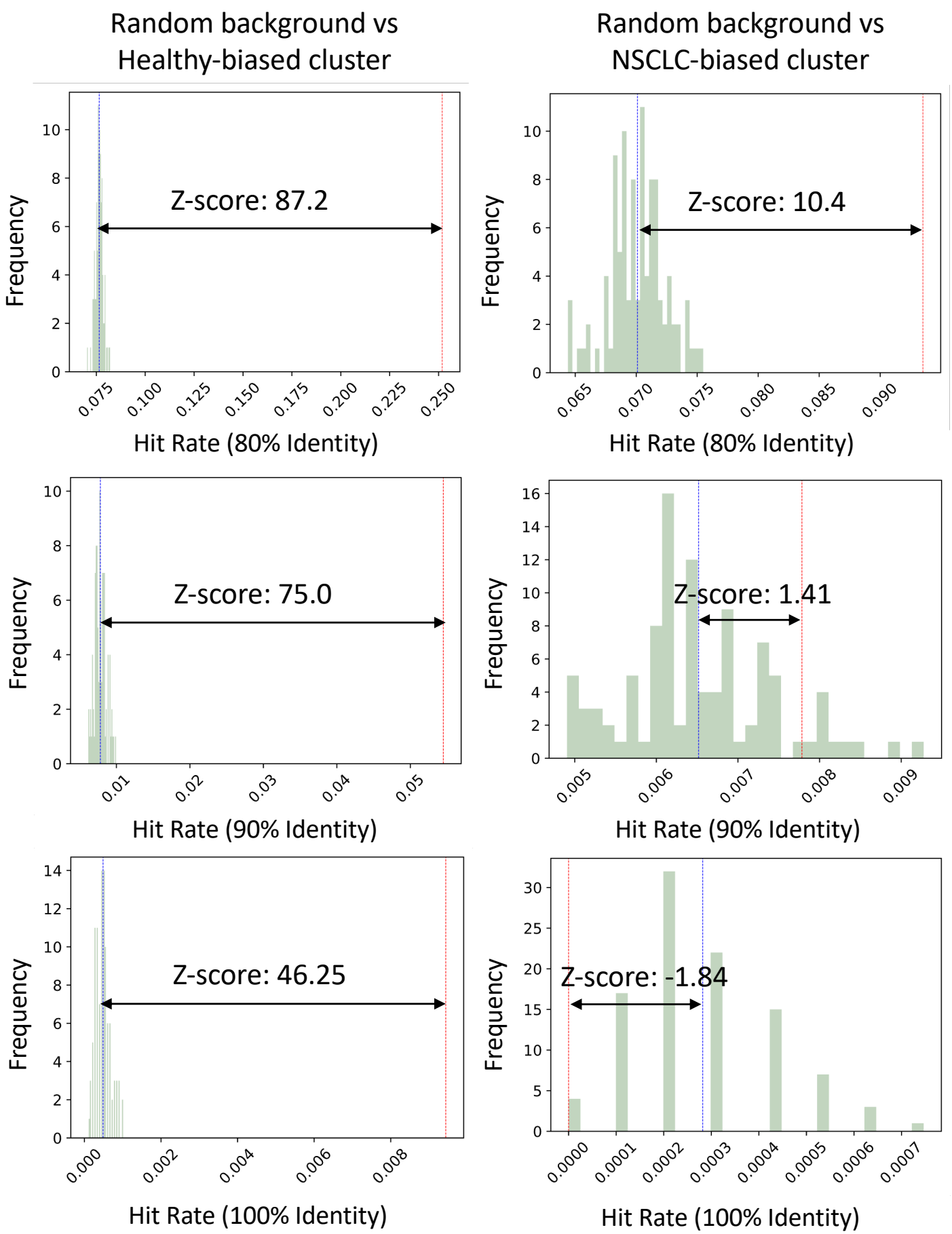

Random background vs
Healthy-biased cluster
Random background vs
NSCLC-biased cluster
Frequency
Z-score: 87.2
Hit Rate (80% Identity)
Z-score: 10.4
Hit Rate (80% Identity)
Z-score: 75.0
Hit Rate (90% Identity)
Z-score: 1.41
Hit Rate (90% Identity)
Z-score: 46.25
Hit Rate (100% Identity)
Z-score: -1.84
Hit Rate (100% Identity)

Figure S4

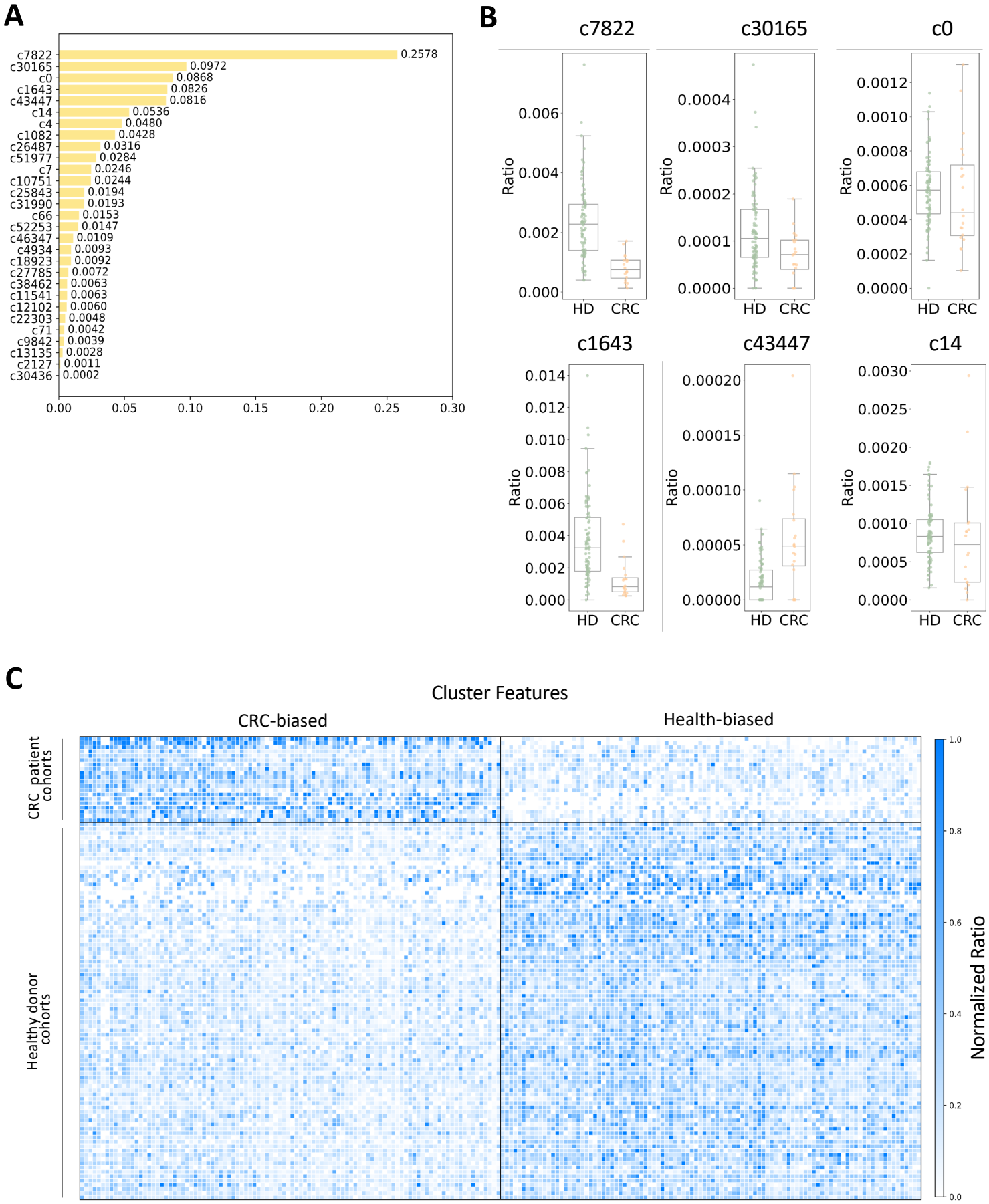
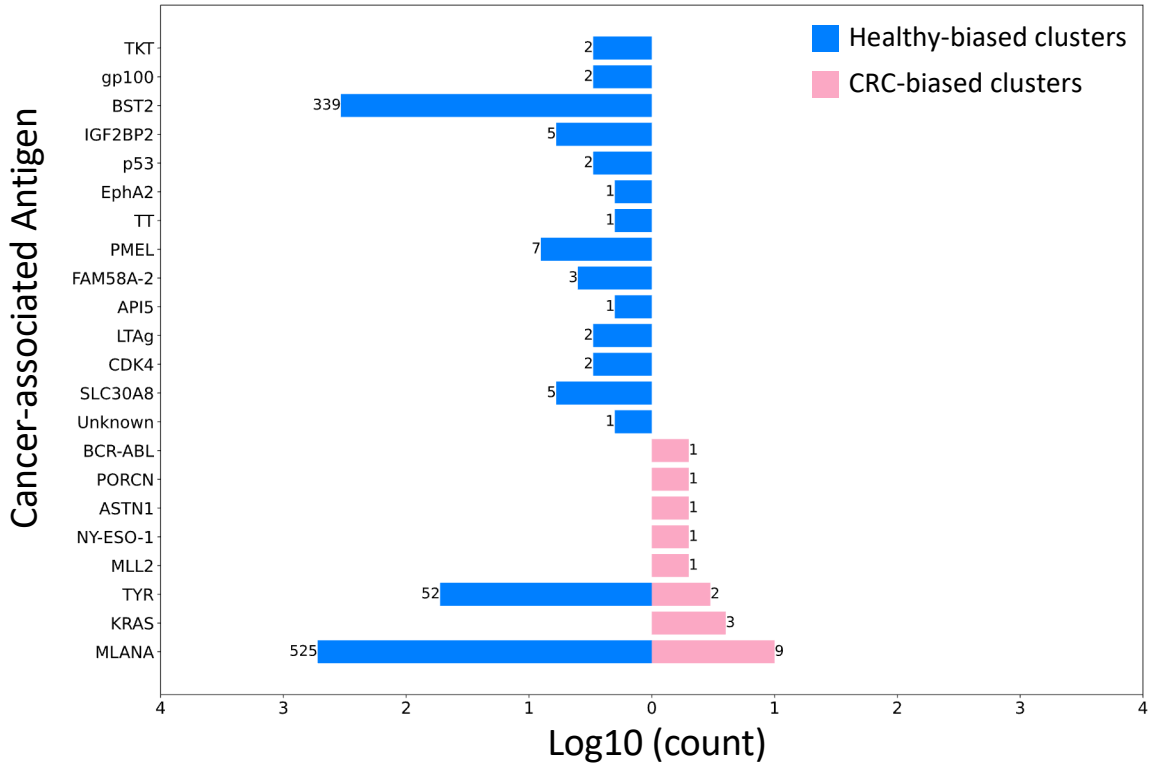

A
c7822 0.2578
c30165 0.0972
c0 0.0868
c1643 0.0826
c43447 0.0816
c14 0.0536
c4 0.0480
c1082 0.0428
c26487 0.0316
c51977 0.0284
c7 0.0246
c10751 0.0244
c25843 0.0194
c31990 0.0193
c66 0.0153
c52253 0.0147
c46347 0.0109
c4934 0.0093
c18923 0.0092
c27785 0.0072
c38462 0.0063
c11541 0.0063
c12102 0.0060
c22303 0.0048
c71 0.0042
c9842 0.0039
c13135 0.0028
c2127 0.0011
c30436 0.0002
0.00 0.05 0.10 0.15 0.20 0.25 0.30
B
c7822
c30165
c0
c1643
c43447
c14
Ratio
HD CRC
C
Cluster Features
CRC-biased
Health-biased
CRC patient cohorts
Healthy donor cohorts
Normalized Ratio
0.0 0.2 0.4 0.6 0.8 1.0

Figure S5

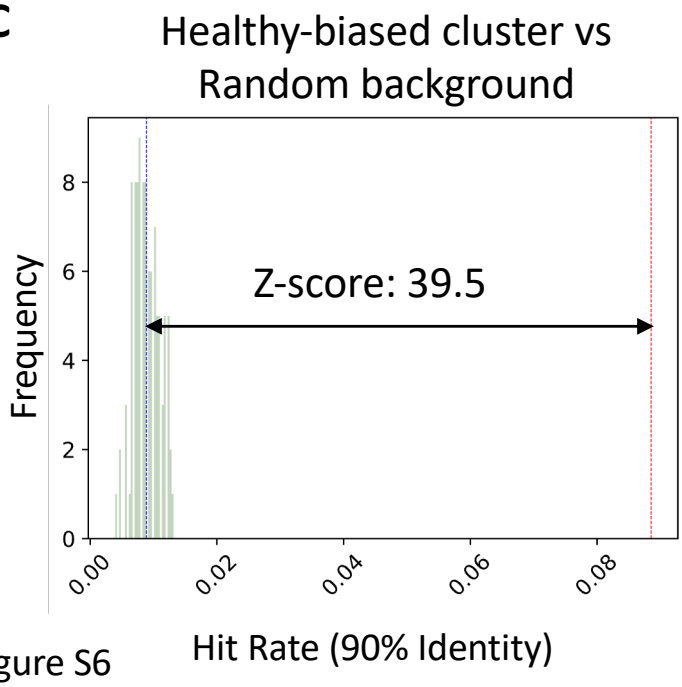
A



B

Cluster		V/ J gene	CDR3 (length)	Identity	Template Antigen	Cluster CDR3 Logo
c7822	Query Template	TRBV6-1/TRBJ2-7 TRBV6-1/TRBJ2-7	ASSETGTGPYEQY (13) ASS-TGTGSYEQY (12)	84%	TKT	
c22303	Query Template	TRBV6-4/TRBJ2-3 TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13) ASSDSSGSTDTQY (13)	100%	BST2	
c13360	Query Template	TRBV6-1/TRBJ2-3 TRBV6-1/TRBJ2-3	ASSELGGGADTQY (13) ASSELSGGADTQY (13)	92%	MLANA	
c26548	Query Template	TRBV6-4/TRBJ2-3 TRBV6-4/TRBJ2-3	ASSPRGGT-DTQY (12) ASSPRGGTADTQY (13)	92%	PMEL	
c14839	Query Template	TRBV9/TRBJ2-5 TRBV9/TRBJ2-5	ASSVATGGGETQY (13) ASSVAGGGQETQY (13)	84%	KRAS	
c29566	Query Template	TRBV4-2/TRBJ1-2 TRBV4-2/TRBJ1-2	ASSQEGWDGYT (11) ASSQEGGDGYT (11)	91%	MLANA	

C



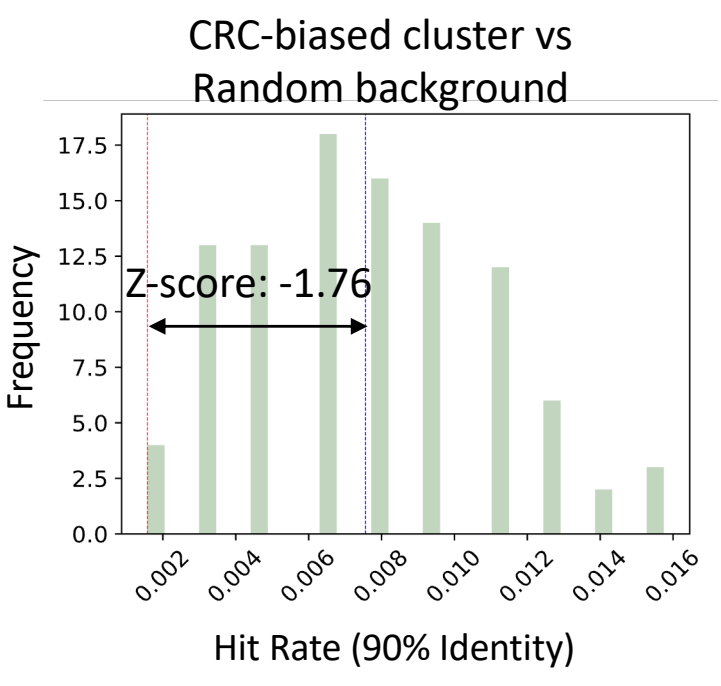


Figure S6

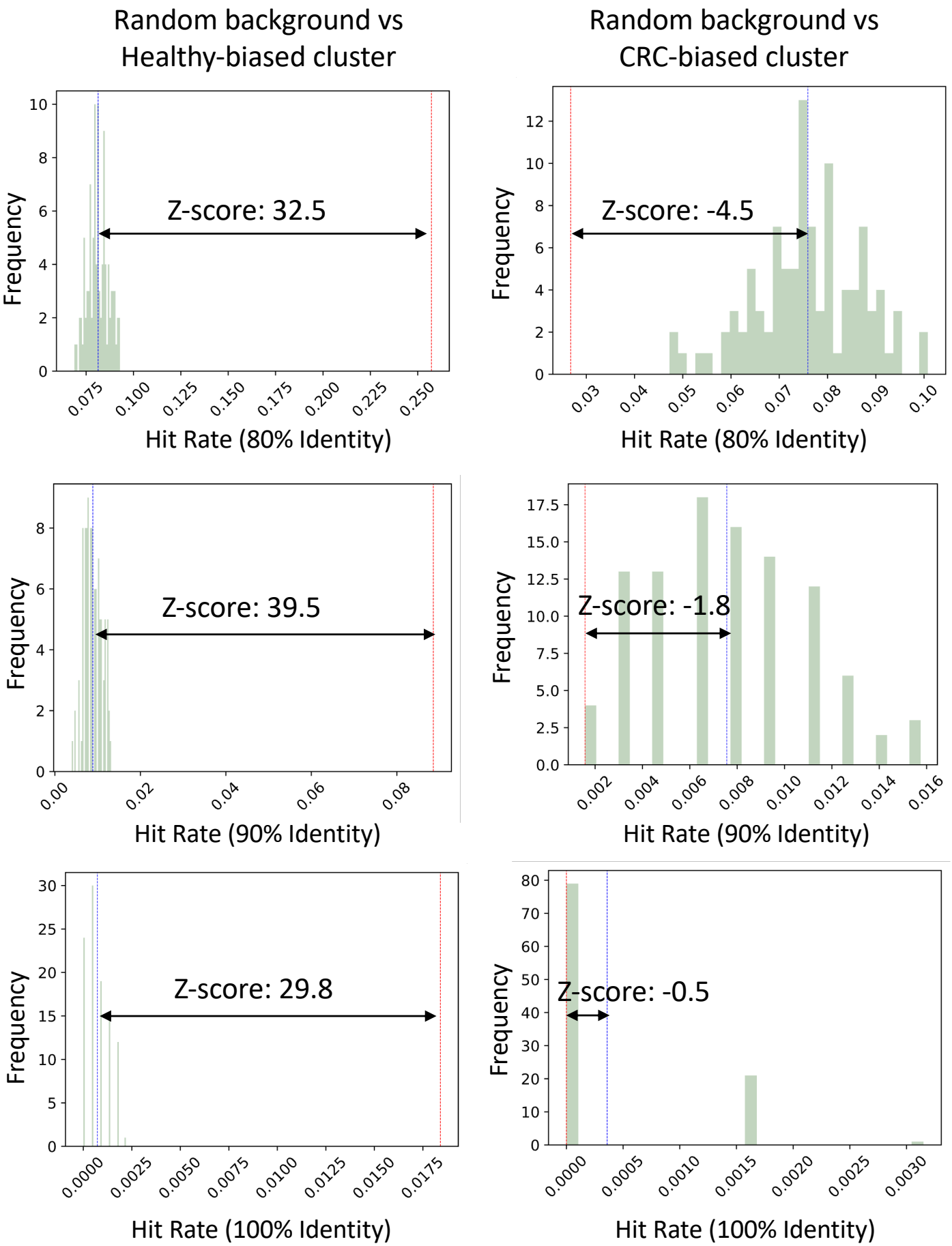
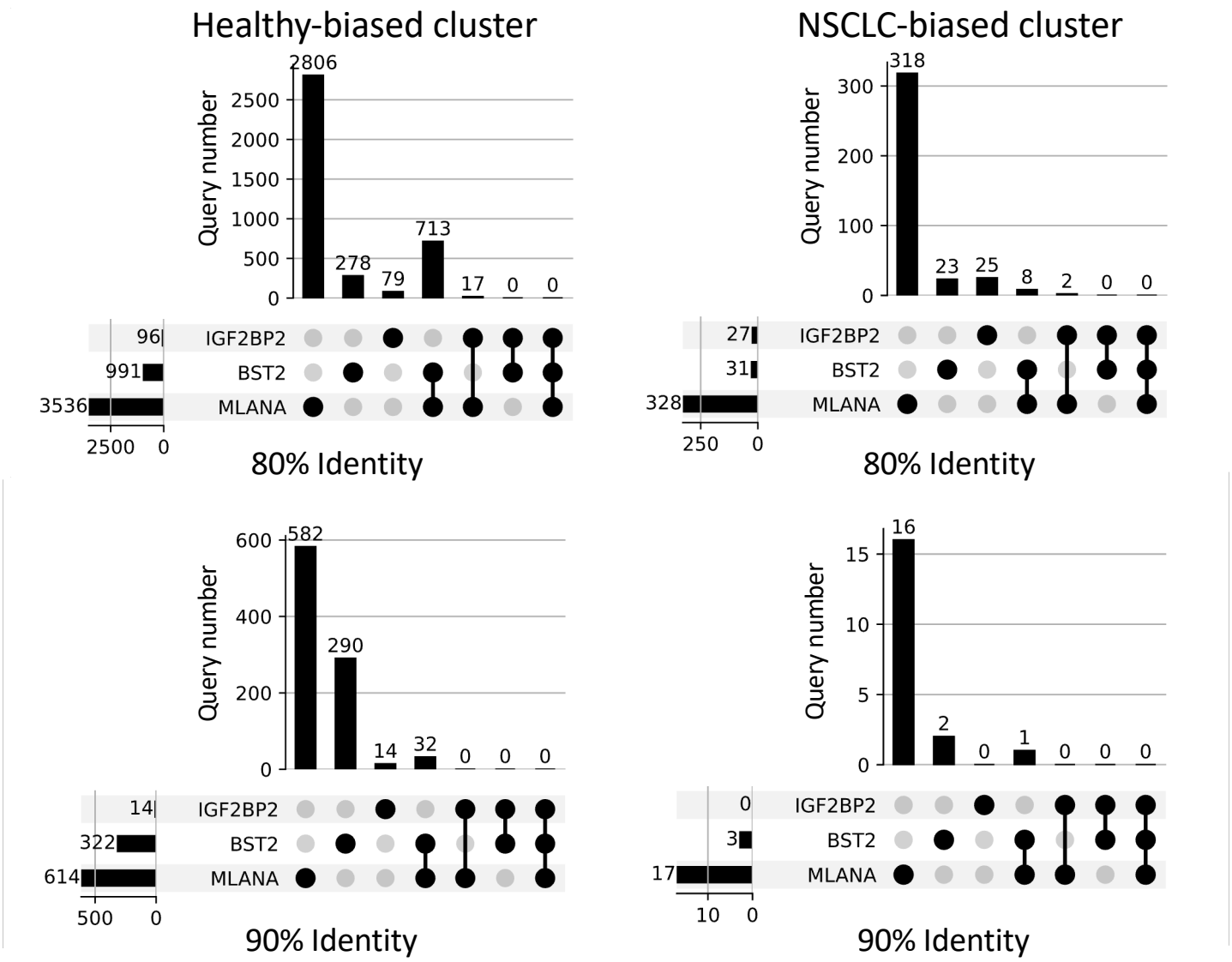

Random background vs
Healthy-biased cluster
Random background vs
CRC-biased cluster
Z-score: 32.5
Z-score: -4.5
Z-score: 39.5
Z-score: -1.8
Z-score: 29.8
Z-score: -0.5
Frequency
Hit Rate (80% Identity)
Hit Rate (90% Identity)
Hit Rate (100% Identity)

Figure S7

A

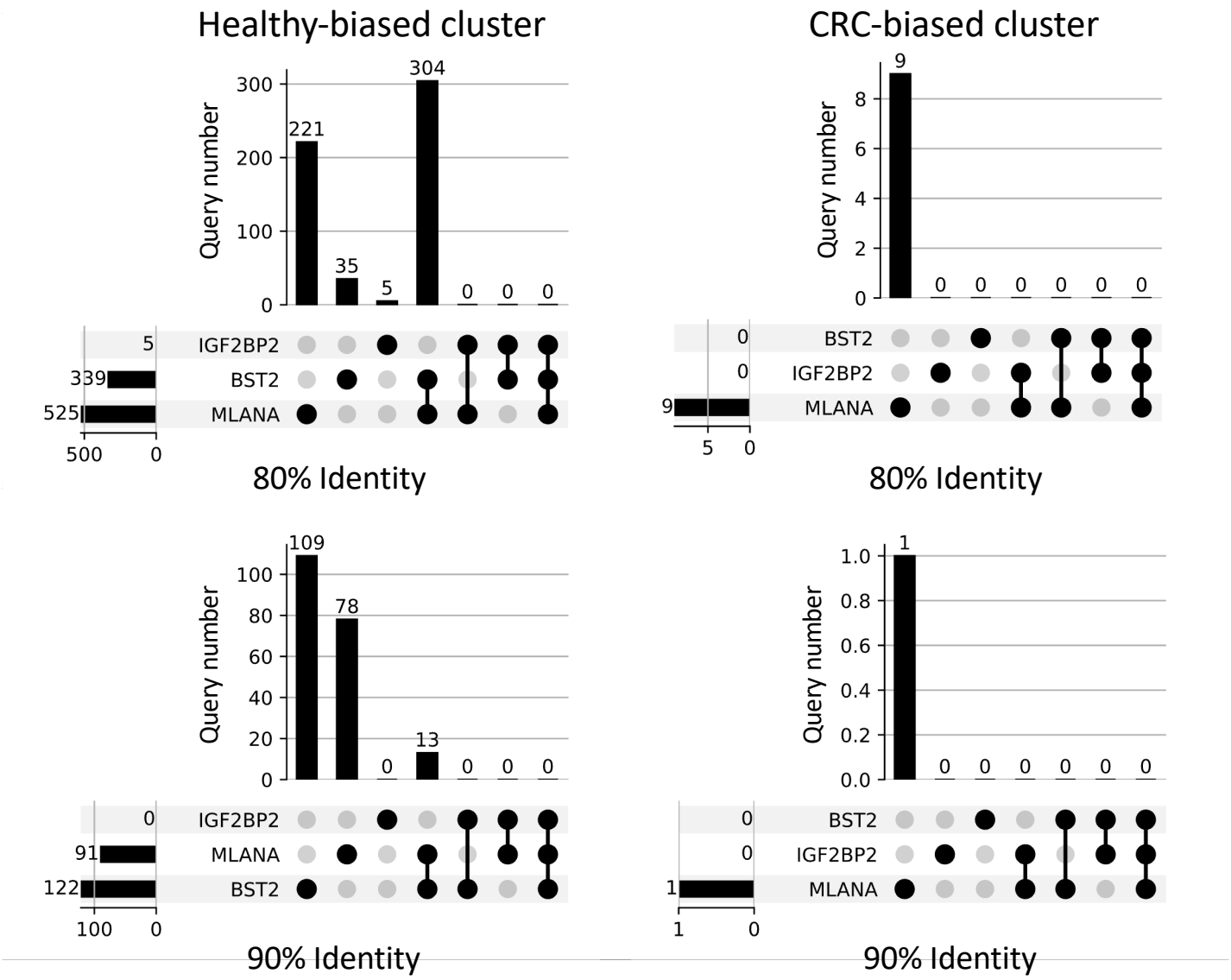


B

Query-template pair	V/ J gene	CDR3 (length)	Identity	Template Antigen (epitope)
ADPT2020V4CD_66_1990	TRBV6-4/TRBJ2-3	ASSDSGGSTDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
HD_46_2316	TRBV6-4/TRBJ2-3	ASSDSSGATDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
D0129d_58_10368	TRBV6-4/TRBJ2-3	ASSDSSGATDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
GreefHD_12_4916	TRBV19/TRBJ1-5	ASSIGLYSNQPQH (13)		
CancerTemplate_1_3616	TRBV19/TRBJ1-5	ASSIGGGSNQPQH (13)	84%	IGF2BP2 (NLSALGIFST)
CancerTemplate_1_3638	TRBV19/TRBJ1-5	ASSWGLLSNQPQH (13)	84%	MLANA (ELAGIGILTV)
D0129d_25_5918	TRBV19/TRBJ1-5	ASSIGLGSNQPQH (13)		
CancerTemplate_1_3616	TRBV19/TRBJ1-5	ASSIGGGSNQPQH (13)	92%	IGF2BP2 (NLSALGIFST)
CancerTemplate_1_3638	TRBV19/TRBJ1-5	ASSWGLLSNQPQH (13)	84%	MLANA (ELAGIGILTV)

Figure S8

A



B

Query-template pair	V/ J gene	CDR3 (length)	Identity	Template Antigen (epitope)
ADPT2020V4CD_59_8715	TRBV6-4/TRBJ2-3	ASSDSSGATDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
ADPT2020V4CD_6_9390	TRBV6-4/TRBJ2-3	ASSDSGGSTDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
ADPT2020V4CD_66_1990	TRBV6-4/TRBJ2-3	ASSDSGGSTDTQY (13)		
CancerTemplate_1_2275	TRBV6-4/TRBJ2-3	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_393	TRBV6-4/TRBJ2-3	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)
ADPT2020V4CD_52_5671	TRBV19/TRBJ1-5	ASSDSSGATDTQY (13)		
CancerTemplate_1_3616	TRBV19/TRBJ1-5	ASSDSSGSTDTQY (13)	92%	BST2 (LLLGIGILV)
CancerTemplate_1_3638	TRBV19/TRBJ1-5	ASSDSGGATDTQY (13)	92%	MLANA (ELAGIGILTV)

Figure S9