Robust diagnosis of infectious disease, autoimmunity and cancer from the paratope networks of adaptive immune receptors

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 of Data Science, Shiga University, Hikone 522-8522, Japan

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

Abstract

Liquid biopsies based on peripheral blood offer a minimally invasive alternative to solid tissue biopsies for the detection of diseases, primarily cancers. However, such tests currently consider only the serum component of blood, overlooking a potentially rich source of biomarkers: adaptive immune receptors (AIRs) expressed on circulating B and T cells. Machine learning-based classifiers trained on AIRs have been reported to accurately identify not only cancers, but also autoimmune and infectious diseases as well. However, when using the conventional "clonotype cluster" representation of AIRs, donors within a disease or healthy cohort exhibit vastly different features, limiting the generalizability of these classifiers. This paper addresses the challenge of classifying specific diseases from circulating B or T cells by developing a novel representation of AIRs based on similarity networks constructed from their antigen-binding regions (paratopes). Features based on this novel representation, paratope cluster occupancies (PCOs), significantly improved disease classification performance for infectious disease, autoimmunity and cancer. Under

identical methodological conditions, classifiers trained on PCOs achieved a mean ROC AUC of 0.893 when applied to new donors, compared to clonotype cluster-based classifiers (0.714) or the best-performing published classifier (0.777). Surprisingly, for cancer patients, we observed that some of the AIRs that were important for classification were significantly more abundant in healthy controls than in individuals with disease. These "healthy-biased" AIRs were predicted to target known cancer-associated antigens at dramatically higher rates than healthy AIRs as a whole (Z scores > 75), suggesting the existence of an overlooked reservoir of cancer-targeting immune cells that are diagnostic and identifiable from a routine blood test. Consequently, PCOs not only enhance classification of a broad range of diseases but also identify immune cells with therapeutic potential.

Introduction

Liquid biopsies that extract circulating tumor DNA, extracellular vesicles, or circulating tumor cells from peripheral blood offer a range of advantages over traditional methods of medical diagnosis[1]. Compared to solid tissue biopsies, these test are minimally invasive, which facilitates routine monitoring, which offers the promise of disease detection before clinical symptoms manifest[2]. More broadly, such approaches have the potential to empower individuals to manage their own health and thus to reduce the cost of and accessibility to healthcare. Furthermore, liquid biopsies may provide molecular profiles of heterogeneous diseases, which can inform personalized treatment interventions[3]. Finally, technological improvement of next-generation sequencing (NGS) and machine learning (ML) are expected to further accelerate improvements in the sensitivity and specificity of these tests[4].

Despite their great promise, current blood-based liquid biopsies have limitations as well. The sensitivity of circulating tumor DNA detection can be low, especially in early-stage cancers or diseases with low tumor burden[5]. Another shortcoming is the exclusive focus on serum. Blood is a rich source of immune cells, which play a direct role in responses to many

diseases. Incorporating the sequences of adaptive immune receptors (AIRs) would greatly enrich the information content of liquid biopsies, potentially improving diagnostic sensitivity[6]. As AIR sequencing and computational analysis technologies have continued to improve, application of AIR data has grown from basic research to application to biomarkers for disease and for guiding immunotherapies[7]. Taken together, expanding current liquid biopsies to include AIR sequence information warrants further exploration.

Each adaptive immune cell expresses a unique AIR, whose coding sequence is generated by rearrangement of germline V D and J genes[8] (**Fig. S1A**). The number of possible combinations far exceeds the number of B or T cells in any one individual[9, 10]. The resulting extraordinary diversity allows adaptive immune cells to engage with and remember nearly any disease-associated antigen. Upon antigen engagement, adaptive immune cells proliferate, resulting in dramatic differences in the levels of specific AIR clones in peripheral blood. For use as a liquid biopsy, the diverse AIR signals in a given donor must be formulated as a single feature vector that can be compared to those of other donors. In recent years, statistics- and ML-based approaches have been actively explored in order to construct such features in order to classify donors according to their disease status[11-20].

The main obstacle to the use of AIRs features of disease status is their diversity. The pairwise sharing of AIRs between different donors from typical blood samples ranges from 1-6% and decreases rapidly with an increase in the number of donors[21, 22]. This “donor sharing problem” has severely hindered the use of AIRs as traditional biomarkers as it prevents ML classifiers from being able to identify general features associated with particular disease. The extent of donor sharing depends, however, on the way of constructing the AIR features. The traditional representation of an AIR is as “clonotype”: its V and J gene names, along with its CDR3 amino acid sequence. A clonotype is a useful qualitative nomenclature because it describes the receptor’s gene rearrangement history. However, because clonotypes mix categorical (gene

names) and continuous (CDR3 amino acid sequence) variables, it is not ideal for quantifying the similarities or differences between AIRs.

An alternative approach is to represent AIRs by a single amino acid sequence containing the residues near the antigen binding interface, also known as the "paratope". Paratopes bring 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**). By concatenating the three complementarity regions[23] or by predicting the paratope[24] a single paratope sequence can be constructed that is readily handled by standard protein sequence analysis methods for alignment, clustering, or searching[25, 26]. This approach has thus been used for grouping BCRs that target a common antigen[23, 24, 27, 28] or TCRs that target a given peptide-MHC complex[17, 29]. In the context of disease diagnosis, an important strength of the paratope representation is that AIRs form extended networks that join together different clonotypes and, importantly, different donors.

Results

Paratope adjacencies connect different clonotypes and donors

The impact of the paratope representation is best seen from an example. For this purpose, we constructed clonotype (**Fig. 1A**) and paratope (**Fig. 1B**) networks using the BCRs of ten COVID-19 from a previous study[30]. Notably, none of the clonotype networks connected different donors, consistent with the previously reported low sharing of BCR clones[9, 22]. The frequency of cluster sizes was systematically lower for clonotypes, than for paratopes, which formed a number of very large networks (**Fig. 1C**). Importantly, the larger paratope networks often contained many different clonotypes and connected many donors (**Fig. 1D**). In this context, we refer to the AIRs that are connected in the paratope networks as "adjacent". These observations on clonotype and paratope networks form the basis of the features introduced here. The key observation is that AIRS can be "paratope adjacent" even if they come from different donors. Therefore, features based on such adjacency are likely to

have elements that are shared among different donors. This is essential to group donors belonging to a common disease class.

Feature engineering

We describe the details of our approach in the *Methods* section. Briefly, our goal is to derive a single feature vector for each donor. We assume each AIR belongs to a cluster. Such clusters can be defined in terms of clonotypes or paratopes. An individual donor can be represented by the frequency of clusters. We denote such features as "cluster frequencies" (CFs) (**Fig. S2A**). In this way, CFs consist of "clonotype cluster frequencies" (CCFs) or "paratope cluster frequencies" (PCFs)), depending on whether we use clonotype or paratope clustering. We next extend the CFs by considering the networks of paratopes (**Fig. 1B**). The networks are represented mathematically by an adjacency matrix A of pairwise paratope edges (**Fig. S2B**). Importantly, a given AIR can have edges with AIRs in the same cluster as well to AIRs belonging to different clusters. We denote the frequency of edges to all clusters as "occupancies" to emphasize the notion that a given AIR can "occupy" multiple clusters. Cluster occupancies (COs) consist of "clonotype cluster occupancies" (CCOs) or "paratope cluster occupancies" (PCOs), depending on whether we use clonotype or paratope clustering (**Fig. S2C**). Because all features are indexed by a specific cluster, the feature importance, as calculated by XGBoost, can be used to identify biologically important AIRs.

AIR data downsizing

Because AIR frequencies follow a long-tailed distribution (**Fig. S1D**), cluster frequencies also have a long-tailed distribution, which results in very long feature vectors. We thus require a way to drop 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 the *Results* section.

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 to minimize batch effects (i.e., any bias other than that of the disease of interest) or used different studies for training and testing.

Classifier training

We randomly split the donors into training (70%) and test (30%) groups. For the classifiers described here, we separately selected the optimal hyperparameter d_{min} by leave-one-out cross validation (LOOCV) using the training donors. Two previously published methods, DeepRC[20] and immuneML[19], were identically trained, as described in the *Methods* section. We then evaluated all the trained classifiers on the test set (**Fig. S2C**). Aside from the feature generation steps (CCF, CCO, PCF, and PCO), all steps in the pipeline were identical for all of the newly developed classifiers. In the following sections, we discuss two representative diseases—COVID-19 (using single cell BCR sequence data) and NSCLC (using bulk TCR sequence data)—in depth. Our findings for the remaining diseases are described in **Figs S3-6 and the Suppl text**. XXX

Classification of COVID-19 patients and healthy donors from BCR data

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[31]. Diagnosis is usually confirmed by polymerase chain reaction (PCR), which can detect the presence of viral nucleic acids[32]. To evaluate the performance of a theoretical AIR-based diagnostic system, we utilized BCR heavy chain data from a single-cell sequencing study of peripheral blood mononuclear cells (PBMCs) from 44 COVID-19 patients (106,640 BCRs) and 58 healthy donors (174,139 BCRs), all of whom were

unvaccinated[30]. After splitting the data randomly, 71 donor datasets were used for training, and 31 were used for testing. The sparsity of the features decreases roughly linearly with d_{min}, while the LOOCV target function reached a maximum at a d_{min}value of 0.7 (**Fig. 2A**). Using this value of d_{min},the classifier trained on the PCO features achieved an area under the receiver operating characteristic curve (ROC AUC) of 0.896 for test data, which was close to the LOOCV result on training data (0.946), indicating that the classifier was not overfitted and generalized well to new donors (**Fig. 2B**). When compared with the other new classifiers, only the model trained on CCF features was unable to classify the test donors (**Fig. 2C)**, presumably due to the poor sharing of clonotypes across donors, as described above. The performance of the previously published classifiers ranged from 0.407 to 0.838, highlighting the importance of the technology in achieving robust results. The precision–recall (PR) AUCs indicate that the PCO-based classifier achieved the highest value (0.932) among all the trained models (**Fig. 2D)**. The above results demonstrate the improved generalizability of the PCO-based classifier, in particular in comparison to the CCF-based classifier. XXX

The stark difference in performance between the CCF-based classifiers and the remaining types classifiers suggests that the computation of paratope occupancies had a critical impact on differentiating healthy individuals from those with COVID-19. To understand this effect, we examined differences between CCF and CCO features, the latter of which utilized the same clonotype clusters as the former but were transformed to occupancies using paratope similarities. To this end, we first identified the feature with the highest CCO importance. We subsequently examined the clonotype cluster (Cluster 1) corresponding to this feature. This cluster consisted of only three AIRs from 2 of 104 possible donors: donor 10 and donor 32 (**Fig. 2E)**. By examining the CCO calculations, we found that AIRs from many donors shared paratope-level similarity with 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 donors (10 and 32), the corresponding CCO feature was nonzero for many donors (898 AIRs from 75 of 104 donors) (**Fig. 2G**). This example highlights the fact that an AIR that belongs to a given clonotype cluster can also have significant paratope similarity to AIRs in

different clusters and that including such paratope-level similarities was therefore critical to the good performance of the CCO- and PCO-based COVID-19 classifiers.

Classification of Non-small cell lung cancer patients from healthy donors using TCR data

Non-small cell lung cancer (NSCLC) is the most common type of lung cancer; it consist primarily of squamous cell carcinoma, large cell carcinoma, and adenocarcinoma, but several rarer subtypes have also been identified[33]. Diagnosis traditionally relies upon a wide variety of evidence, including symptoms (e.g., persistent cough), imaging data, and tissue biopsy; however, diagnoses based on blood-based biomarkers (*EGFR, HER2, BRAF, KRAS, MET etc*)[34] have recently become more common in the application of liquid biopsy in NSCLC [35]. Like in many other cancers, in NSCLC, disease-specific T cells often infiltrate tumors, but whether these T cells also circulate in the blood remains poorly understood. A TCR-based diagnosis would be beneficial as a less invasive approach than biopsy. Unlike for the other diseases, we were unable to find a single study that included PBMCs from both NSCLC patients and healthy controls. Therefore, we constructed a diverse dataset from seven studies: three with data from healthy individuals[36-38] and two with data from individuals with NSCLC[39, 40] to serve as the training set and one each with data from healthy individuals[41] and from individuals with NSCLC[42] to serve as the testing set. Across the studies, a total of 204 NSCLC donors (6,734,867 TCRs) and 294 healthy donors (4,120,597 TCRs) were ultimately included. After the data were split, 344 donors were used for classifier training, and 152 were retained for testing. **Fig. 3A** shows the sparsity of the resulting features, indicating that the occupancy-based features (CCO, PCO) were approximately 20% less sparse than the frequency-based features at small values of d_{min} and 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 an ROC AUC of 0.985 for the test donors. **Figs. 3C-D** indicate that, among the new classifiers, only the those trained on occupancy-based features (CCO, PCO) performed well on test donors and that their performance was close to perfect (ROC AUC 0.985; PR AUC 0.982). The ROC AUCs of the

previously published methods ranged from 0.317 to 0.895. Taken together, PCO-based classifiers demonstrated the potential for nearly perfect diagnosis on a large cohort of cancer patients and healthy controls. The less than perfect performance of the next-best classifier--0.895 by immuneML(RF)—indicated that merely increasing the size of the training data was not sufficient for robust classification.

Assessment across infectious disease, autoimmunity and cancer

In addition to COVID-19 and NSCLC, we extended our benchmark to include an additional infectious disease, Human immunodeficiency virus (HIV); autoimmunity (Autoimmune hepatitis (AIH), Type 1 Diabetes (T1D); and, an additional cancer, Colorectal cancer (CRC) (**Figs. S3-4**). When we assessed all classifiers on all six independent test donor sets, we found that the PCO-based classifier exhibited the best overall performance (**Table 1**). The PCO-based classifier achieved the greatest ROC AUC among all the classifiers for all six diseases except for CRC, the disease with the fewest patients (20), and for which the difference (0.850 vs 0.821) was rather small. The mean ROC AUC of the PCO-based classifier (0.893) was substantially greater than that of the CCF-based classifier (0.714) and the previously published methods (0.537-0.777). Three of the PCO ROC AUCs—those for HIV (0.985), autoimmune hepatitis (AIH, 0.947) and NSCLC (0.985)—were close to perfect. The remaining three diseases—COVID-19 (0.896), type 1 diabetes (T1D, 0.725), and CRC (0.821)—were represented by fewer disease donors, suggesting that at least 50 disease donors may be needed for robust classifier performance. In contrast, the CCF-based classier ROC AUCs exceeded 0.9 for only one disease (HIV), for which the sequencing depth was much greater than that of the remaining diseases and performed no better than chance for T1D. It is noteworthy that, removal of the paratope features reduced performance (CCF, 0.714) to a value similar to that of immuneML(RF) (0.777), suggesting that the improvement was likely due to this innovation. The above results demonstrate that PCO-based classifiers can successfully distinguish disease patients from healthy donors in infectious disease, autoimmunity and cancer.

		COVID-19	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	0.780	0.802	0.764
	immuneML (SVM)	0.781	0.744	0.5	0.5	0.750	0.836	0.685
	immuneML (KNN)	0.565	0.851	0.656	0.525	0.760	0.795	0.692
	immuneML (RF)	0.838	0.876	0.706	0.5	**0.850**	0.895	0.777
	immuneML (PBC)	0.533	0.5	0.713	0.5	0.660	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 developed in this study, each trained on one of the four features (CCF, PCF, CCO, and PCO) and applied to six diseases (COVID-19, HIV, AIH, T1D, CRC, and NSCLC). The numbers of donors are listed under each disease. Areas under both the receiver operating characteristic (ROC AUC) and precision-recall (PR AUC) 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 the second column 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), and probabilistic binary classifier (PBC).

Classifiers trained on PCOs were robust against batch effects

Because our large NSCLC dataset was composed of data from multiple studies, it provided the opportunity to systematically explore the robustness of the different features against batch effects: effects that arise from differences between samples that are not rooted in the experimental design and can have various sources[43]. To this end, we sampled all possible splits of the datasets where the test set consisted of the data from one healthy study and

one NSCLC study, while the data from the remaining studies were used for training. This process resulted in twelve splits with various numbers of AIRs in the training and test sets. These results indicated that, in terms of the ROC (**Fig. 3E**) and PR AUCs (**Fig. 3F**), the classifiers based on PCF and PCO features performed well overall, but those based on the CCF and CCO features performed inconsistently, in agreement with the data shown in **Figs. 3C-D**. Furthermore, two of the twelve splits showed that it was possible to combine the NSCLC studies in such a way that the PCO model failed to generalize to the test data. Interestingly, in these two splits, the size of the training dataset was much smaller than that of the test dataset, which was consistent with the findings for the other diseases like T1D and CRC, again emphasizing the need for sufficient training data. Overall, however, the performance of the PCO-based classifier was more robust than that of the alternative new classifiers.

Cancer data reveals a surprising relationship between healthy-biased clusters and cancer antigen specificity

The underlying AIRs can be investigated using the XGBoost importance to identify clusters of interest. Because each feature corresponds to an AIR cluster, we hypothesized that AIRs within clusters corresponding to important features with high importance values 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 proportions of healthy- and NSCLC-derived TCRs in each. We observed clusters both with significantly more TCRs from healthy donors (“healthy-biased” clusters) and with significantly more TCRs from disease donors (“NSCLC-biased” clusters) 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[44] and VDJdb[45]), whose TCRs are annotated by their targeted antigen and associated disease. Interestingly, we identified strong hits to cancer-associated antigens using the TCRs from both the healthy- and NSCLC-biased clusters, but surprisingly, there

were more hits from the healthy-biased clusters than from the NSCLC-biased clusters (**Fig. 5A)**. **Fig. 5B** illustrates some of these hits, which included antigens such as MLANA[46], EphA2[47], BST2[48], TKT[49] and IGF2BP2[50], which have been reported as prognostic markers for NSCLC. These findings demonstrate the further application of PCO-based features as a means of identifying potential cancer-fighting T cells.

We speculated that T cells that target cancer-associated antigens might be more common in healthy donors if, in NSCLC patients, they migrate from the peripheral blood toward tumor-presenting tissues (i.e., the lungs). To test this idea, we repeated the database queries using randomly selected TCRs from healthy donors. Compared to TCRs from healthy biased clusters, randomly selected TCRs from healthy donors resulted in dramatically fewer hits (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 from donors with NSCLC (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**). Taken together, they show that the healthy-biased TCRs are indeed distinct from typical healthy donor-derived TCRs, supporting the notion that they would have the potential to migrate out of the periphery in cancer patients.

We next performed an analogous analysis for colorectal cancer (CRC) cases (Fig. **S5-6**). Consistently, TCRs from healthy-biased clusters had significantly more cancer-related hits than did randomly selected TCRs from healthy donors (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 the HLA A*02:01-restricted epitopes EAAGIGILTV, LLLGIGILVL, and NLSALGIFST from Melan A (MLANA), BST2, and IMP2 (IGF2BP2) and that PBMCs from healthy donors expressed such TCRs[51]. 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 NSCLC (**Fig. S8**) and CRC (**Fig. S9**) datasets. Taken

together, these results strongly suggest that the TCRs in the healthy-biased clusters migrate out of the peripheral blood to infiltrate tumors in individuals with NSCLC and CRC. These findings highlight the interpretability of our ML classifiers to identify underlying mechanisms and potentially therapeutic immune cells.

Discussion

Liquid biopsies that can detect specific diseases from blood have the potential to reshape the future of medical diagnosis. Adaptive immune cells, which circulate in peripheral blood, are highly sensitive to a broad range of diseases. However, harnessing this sensitivity in a reproducible and general manner has presented a challenge, due to the high diversity and low inter-donor sharing of AIRs. Moreover, 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 the clonotype nomenclature, which separates AIRs into discrete clusters defined by their V and J gene usage and CDR3 length. Our results demonstrate that an AIR representation that incorporates the extended networks of paratopes allows donors to be connected and improves classifier performance compared to clonotype-based classifiers.

We found that there was obvious improvement with additional training data, which is encouraging given the rapid growth of publicly available AIR data. Beyond a threshold value of 200 donors, the PCO-based classifiers were very robust, achieving ROC AUC values of 0.985 (HIV and NSCLC). This suggests that initial clinical validation of specific diseases could be performed with smaller numbers of donors (e.g. 20 patients) and then scaled up to larger numbers (e.g. 50 patients) based on initial results.

The network view of adaptive immunity makes sense from the perspective of host defense, as pathogens represent a diverse and unpredictable set of threats. Indeed, Immune Network

Theory, originally proposed by Jerne and Hoffmann helped to explain the immune system's ability to distinguish self from non-self[52,53]. Here, we see from the perspective of disease classification that dense features based on paratope networks generalize much better to new donors than sparse features based on less connected clonotypes and that these observations were consistent across three broad categories of disease.

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. Typically, in repertoire analysis, there is a tendency to focus on clusters that are overrepresented in patients[54, 55]. However, we observed the reverse: healthy-biased clusters were important for disease classification as well. One possible explanation for the importance of healthy-biased clusters is that they are not relevant to cancer and are thus downregulated in individuals with cancer. However, the opposite interpretation—that these clusters 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 is consistent with the observation that tumor-infiltrating lymphocytes (TILs) have been observed in almost all cancers[56]. Although we cannot determine what happens to the T cells missing from cancer patients, the extremely high hit rate of cancer-associated TCRs in healthy-biased clusters with respect background healthy TCRs supports the idea that these T cells leave the periphery of cancer patients as a result of targeting tumors. Moreover, the close resemblance of the TCRs identified here to those found to be reactive to multiple types of cancer[51] further strengthens this argument. The notion that tumors can affect global migration from peripheral blood has been demonstrated in brain cancer and metastatic melanoma[57]. In one NSCLC source report[42], the authors discovered that many expanded intratumoral TCRs were detectable in blood samples at the time of lung tumor resection from NSCLC patient, consistent with the notion that T cells in peripheral blood, and TCR-based classification of cancer cells from PBMCs, can be used for disease monitoring and personalized immunotherapy development.

XXX

Undoubtedly, there are limitations to this study that suggest future directions. One concerns the focus on specific disease and limited sample sizes. In this study, we focused on classifying single disease patients from a healthy donor cohort. However, future work must also explore the classification of different cancer types or multiple diseases simultaneously. To achieve this, larger and higher-quality AIR datasets with accompanying clinical information, such as tumor stage, treatment history, and survival outcomes, will be necessary. Expanding the scope of the study to include more diverse disease types and larger cohorts will also help validate the generalizability of our findings and provide a more comprehensive understanding of the TCR repertoire's role in disease diagnosis and monitoring. Another area we have not yet explored is the combination of BCR and TCR information. Our study focused on BCRs for infectious diseases and TCRs for autoimmune diseases and cancer; however, it would be valuable to investigate the potential of combining both BCR and TCR information, including paired (light-heavy, alpha-beta) chains, to train the classifier. By integrating data from both arms of the adaptive immune system, future studies may potentially improve classification performance and provide a more holistic view of the immune response in various disease states. In addition, incorporating more advanced ML model is of interest. With the advancement of large language models, it will be interesting to explore whether paratope networks emerge naturally from training rather than through explicit calculation of the adjacency matrix. By leveraging these advanced models, we may capture more complex patterns and interactions within TCR and BCR repertoires. In parallel, extending our approach through use of deep learning models, such as graph neural networks, may further improve diagnostic performance.

XXX In conclusion, our study represents a proof of concept showing the use of paratope networks improves the robustness of disease classifiers with the potential to diagnose infectious diseases, autoimmune disorders, and cancer from adaptive immune receptor repertoire data. These findings highlight the potential of our method to provide a deeper understanding of the role of the adaptive immune system in three major disease states and

contributes to the development of more accurate and widely applicable liquid biopsies. Moreover, since most liquid biopsies utilizing peripheral blood utilize only the serum component, our technology is entirely compatible with such tests and may well add breadth and sensitivity to existing tests. As the field of AIR repertoire analysis continues to evolve, we anticipate that further refinement will enhance the performance and generalizability of this approach, ultimately leading to improved patient care and outcomes.

Figure Captions

Figure 1. **Paratope networks enable donor sharing.** BCR heavy chain networks were constructed for 10 COVID-19 patients, where nodes are colored by donor and edges represent BCR sharing the same clonotype or similar paratope. Only networks larger than 10 are shown for simplicity. **A,** Clonotype networks are typically small and do not connect different donors. **B,** Paratope networks are much larger and generally connect many donors. **C,** The network sizes and frequencies were distinct even on a log scale. **D,** A close-up view of one of the larger paratope networks (circled in B), whch is made up of many different clonotypes, includes all 10 donors.

Figure 2. **Assessment of COVID-19 diagnosis based on BCRs A,** The sparsity of each of the four feature matrices as a function of d_{min}. The vertical cyan line shows the value from LOOCV hyperparameter optimization, and the pink line indicates 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,** ROC AUCs for all four classifiers in the test set. **D,** PR AUCs for all four classifiers in the test set; the performance of a random classifier is indicated as a gray 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 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,** A 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 as a function of d_{min}, with the vertical cyan line showing the value from LOOCV hyperparameter optimization and 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,** ROC AUCs for all four classifiers in the test set. **D,** PR AUCs for all four classifiers in the test set; the AUC of a random classifier is indicated as a gray bar. **E,** Summary of 12 ROC AUC values in the test sets of NSCLC using different combinations of studies for training and testing. **F,** PR AUC values in the same 12 test sets.

Figure 4. **Assessment of cluster features in NSCLC patient classification A**, Feature importance. The Y-axis shows the importance of the XGBoost features in descending order. **B**, The six most important features. Each dot represents the proportions of healthy (HD) and disease (NSCLC) donors exhibiting those features. A t test was performed to compare the proportions between the HD and NSCLC cohorts (***$p \leq 0.001$). **C**, Heatmap of the top 100 significant disease and healthy (HD) cluster features. *p* values were calculated with the t test; by sorting the *p* values in ascending order, the ranks of the corresponding features are determined. The value of each cell is the normalized ratio from 0 (white) to 1 (blue). The left Y-axis shows each donor from the 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 red line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical lines representing the means). The horizontal arrow indicates the Z score. **D,** Hit rates of TCRs from NSCLC-biased clusters to cancer-targeting TCRs (vertical red line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical lines representing the means). 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 results 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,** Contact with an antigen (light blue) most often occurs with the CDRs than with the background assessable surface (light brown). **D,** The distribution of BCRs and TCRs for each donor follows a long-tailed distribution, with the large majority of clones having very low frequencies (clones with counts less than 10 are not shown).

Figure S2. **Cluster frequency and occupancy strategy in the classification algorithm. A,** Cluster frequencies are illustrated with two example donors (A and B) whose AIRs form 6 clusters, which are one-hot encoded in matrix *C*. The cluster frequency feature of a donor is obtained by summing the rows of this matrix that belong to that donor. **B,** Cluster occupancies are derived similarly to cluster frequencies except that we introduce an adjacency matrix describing 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 is generally less sparse than that derived from cluster frequencies. **C,** Flowchart of the classification pipeline. Starting from a set of training and test donors, the first step is feature generation using one of four methods, which differ in the type of clustering (clonotype or paratope) and whether the clusters are transformed into features based on frequencies or occupancies. Thereafter, the remaining steps are identical and consist of training/hyperparameter selection and testing the classifier on the test features.

Figure S3. **Assessment of HIV diagnosis based on BCRs and AIH diagnosis based on TCRs**. **A/E,** The sparsity of each of the four feature matrices as a function of d_{min}, with the vertical cyan line showing the value from LOOCV hyperparameter optimization and 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,** ROC AUCs for all four classifiers in the test set. **D/H,** PR AUCs for all four classifiers in the test set; the AUC of a random classifier is indicated as a gray bar. Impact of occupancy on important features.

Figure S4. **Assessment of T1D and CRC diagnosis based on TCRs**. **A/E,** The sparsity of each of the four feature matrices as a function of d_{min}, with the vertical cyan line showing the value from LOOCV hyperparameter optimization and 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,** ROC AUCs for all four classifiers in the test set. **D/H,** PR AUCs for all four classifiers in the test set; the AUC of a random classifier is indicated as a gray bar. Impact of occupancy on important features.

Figure S5**. 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 did those from randomly selected background TCRs under three different CDR3 sequence identity thresholds (80%, 90% and 100%).

Figure S6. **Assessment of cluster features in CRC patient classification**. **A**, Feature importance. The Y-axis shows the importance of the XGBoost features in descending order. **B**, The top six most important features. Each dot represents the proportion of healthy (HD) and disease (CRC) donors with that feature. A t test was performed to compare the proportions between the HD and CRC cohorts (***$p \leq 0.001$). **C**, Heatmap of the top 100 CRC- and healthy (HD)-biased cluster features. *p* values were calculated with the t test; by sorting the p values in ascending order, the ranks of the corresponding features was determined. The value of each cell is the normalized ratio from 0 (white) to 1 (blue). The left Y-axis shows each donor from the healthy and CRC cohorts grouped by the corresponding study.

Figure S7. **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 red line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical lines representing the means). The horizontal arrow indicates the Z score. **D,** Hit rates of TCRs from CRC-biased clusters to cancer-targeting TCRs (vertical red line) and the corresponding hit rates of 100 randomly selected sets of queries from the same donors (green bars with blue vertical lines representing the means). The horizontal arrow indicates the Z score.

Figure S8**. 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 did those from randomly selected background TCRs under three different CDR3 sequence identity thresholds (80%, 90% and 100%).

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

Figure S10. **Predicted MLANA, BST2, and IGF2BP2-targeting TCRs**. **A,** The UpSet plot shows the TCR hit numbers for the predicted MLANA, BST2, and IGF2BP2-targeting TCRs from healthy (left) and CRC-biased (right) clusters with CDR3 sequence identities of 80 or 90%. Some TCRs are predicted to target two antigens. **B,** Sequence alignments of query-template hits indicating that multiple healthy donors harbor TCRs predicted to target two cancer-related antigens.

References

1. Lone SN, Nisar S, Masoodi T, Singh M, Rizwan A, Hashem S, El-Rifai W, Bedognetti D, Batra SK, Haris M, Bhat AA, Macha MA. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Mol Cancer. 2022;21(1):79. Epub 20220318. doi: 10.1186/s12943-022-01543-7. PubMed PMID: 35303879; PMCID: PMC8932066.

2. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic - implementation issues and future challenges. Nat Rev Clin Oncol. 2021;18(5):297-312. Epub 20210120. doi: 10.1038/s41571-020-00457-x. PubMed PMID: 33473219.
3. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol. 2017;14(9):531-48. Epub 20170302. doi: 10.1038/nrclinonc.2017.14. PubMed PMID: 28252003.
4. Im YR, Tsui DWY, Diaz LA, Jr., Wan JCM. Next-Generation Liquid Biopsies: Embracing Data Science in Oncology. Trends Cancer. 2021;7(4):283-92. Epub 20201213. doi: 10.1016/j.trecan.2020.11.001. PubMed PMID: 33317961; PMCID: PMC8408348.
5. Cescon DW, Bratman SV, Chan SM, Siu LL. Circulating tumor DNA and liquid biopsy in oncology. Nat Cancer. 2020;1(3):276-90. Epub 20200320. doi: 10.1038/s43018-020-0043-5. PubMed PMID: 35122035.
6. Snir T, Efroni S. T cell repertoire sequencing as a cancer's liquid biopsy—can we decode what the immune system is coding? Current Opinion in Systems Biology. 2020;24:135-41. doi: https://doi.org/10.1016/j.coisb.2020.10.009.
7. Frank ML, Lu K, Erdogan C, Han Y, Hu J, Wang T, Heymach JV, Zhang J, Reuben A. T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. Clin Cancer Res. 2023;29(6):994-1008. doi: 10.1158/1078-0432.CCR-22-2469. PubMed PMID: 36413126; PMCID: PMC10011887.
8. Murphy K, Weaver C. Janeway's immunobiology. 9th edition. ed. New York, NY: Garland Science,Taylor & Francis Group; 2017.
9. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566(7744):393-7. Epub 20190121. doi: 10.1038/s41586-019-0879-y. PubMed PMID: 30664748; PMCID: PMC6411386.
10. Elhanati Y, Sethna Z, Callan CG, Jr., Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev. 2018;284(1):167-79. doi: 10.1111/imr.12665. PubMed PMID: 29944757; PMCID: PMC6033145.
11. Chen Y, Ye Z, Zhang Y, Xie W, Chen Q, Lan C, Yang X, Zeng H, Zhu Y, Ma C, Tang H, Wang Q, Guan J, Chen S, Li F, Yang W, Yan H, Yu X, Zhang Z. A Deep Learning Model for Accurate Diagnosis of Infection Using Antibody Repertoires. J Immunol. 2022;208(12):2675-85. Epub 20220523. doi: 10.4049/jimmunol.2200063. PubMed PMID: 35606050.
12. Foers AD, Shoukat MS, Welsh OE, Donovan K, Petry R, Evans SC, FitzPatrick ME, Collins N, Klenerman P, Fowler A, Soilleux EJ. Classification of intestinal T-cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status. J Pathol. 2021;253(3):279-91. Epub 20210106. doi: 10.1002/path.5592. PubMed PMID: 33225446; PMCID: PMC7898595.
13. 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. 2021;12:680687. Epub 20210722. doi: 10.3389/fimmu.2021.680687. PubMed PMID: 34367141; PMCID: PMC8340020.

14. Park JJ, Lee KAV, Lam SZ, Moon KS, Fang Z, Chen S. Machine learning identifies T cell receptor repertoire signatures associated with COVID-19 severity. Commun Biol. 2023;6(1):76. Epub 20230120. doi: 10.1038/s42003-023-04447-4. PubMed PMID: 36670287; PMCID: PMC9853487.
15. Shemesh O, Polak P, Lundin KEA, Sollid LM, Yaari G. Machine Learning Analysis of Naive B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. Front Immunol. 2021;12:627813. Epub 20210310. doi: 10.3389/fimmu.2021.627813. PubMed PMID: 33790900; PMCID: PMC8006302.
16. Cinelli M, Sun Y, Best K, Heather JM, Reich-Zeliger S, Shifrut E, Friedman N, Shawe-Taylor J, Chain B. Feature selection using a one dimensional naive Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. Bioinformatics. 2017;33(7):951-5. doi: 10.1093/bioinformatics/btw771. PubMed PMID: 28073756; PMCID: PMC5860388.
17. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017;547(7661):89-93. Epub 20170621. doi: 10.1038/nature22383. PubMed PMID: 28636592; PMCID: PMC5616171.
18. Eliyahu S, Sharabi O, Elmedvi S, Timor R, Davidovich A, Vigneault F, Clouser C, Hope R, Nimer A, Braun M, Weiss YY, Polak P, Yaari G, Gal-Tanamy M. Antibody Repertoire Analysis of Hepatitis C Virus Infections Identifies Immune Signatures Associated With Spontaneous Clearance. Front Immunol. 2018;9:3004. Epub 20181221. doi: 10.3389/fimmu.2018.03004. PubMed PMID: 30622532; PMCID: PMC6308210.
19. Pavlovic M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, Waagan K, Bernal FLM, Costa AA, Corrie B, Akbar R, Al Hajj GS, Balaban G, Brusko TM, Chernigovskaya M, Christley S, Cowell LG, Frank R, Grytten I, Gundersen S, Haff IH, Hovig E, Hsieh PH, Klambauer G, Kuijjer ML, Lund-Andersen C, Martini A, Minotto T, Pensar J, Rand K, Riccardi E, Robert PA, Rocha A, Slabodkin A, Snapkov I, Sollid LM, Titov D, Weber CR, Widrich M, Yaari G, Greiff V, Sandve GK. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. Nat Mach Intell. 2021;3(11):936-44. Epub 20211116. doi: 10.1038/s42256-021-00413-z. PubMed PMID: 37396030; PMCID: PMC10312379.
20. Widrich M, Schäfl B, Pavlović M, Sandve GK, Hochreiter S, Greiff V, Klambauer G. DeepRC: Immune repertoire classification with attention-based deep massive multiple instance learning. bioRxiv. 2020:2020.04.12.038158. doi: 10.1101/2020.04.12.038158.
21. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010;2(47):47ra64. doi: 10.1126/scitranslmed.3001442. PubMed PMID: 20811043; PMCID: PMC3212437.
22. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, Crowe JE, Jr. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019;566(7744):398-402. Epub 20190213. doi: 10.1038/s41586-019-0934-8. PubMed PMID: 30760926; PMCID: PMC6949180.

23. Xu Z, Li S, Rozewicki J, Yamashita K, Teraguchi S, Inoue T, Shinnakasu R, Leach S, Kurosaki T, Standley DM. Functional clustering of B cell receptors using sequence and structural features. Molecular Systems Design & Engineering. 2019;4(4):769-78. doi: 10.1039/C9ME00021F.
24. Richardson E, Galson JD, Kellam P, Kelly DF, Smith SE, Palser A, Watson S, Deane CM. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. MAbs. 2021;13(1):1869406. doi: 10.1080/19420862.2020.1869406. PubMed PMID: 33427589; PMCID: PMC7808390.
25. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2. Epub 20121011. doi: 10.1093/bioinformatics/bts565. PubMed PMID: 23060610; PMCID: PMC3516142.
26. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-8. Epub 20171016. doi: 10.1038/nbt.3988. PubMed PMID: 29035372.
27. Ismanto HS, Xu Z, Saputri DS, Wilamowski J, Li S, Nugraha DK, Horiguchi Y, Okada M, Arase H, Standley DM. Landscape of infection enhancing antibodies in COVID-19 and healthy donors. Comput Struct Biotechnol J. 2022;20:6033-40. Epub 20221104. doi: 10.1016/j.csbj.2022.11.001. PubMed PMID: 36348766; PMCID: PMC9635252.
28. Saputri DS, Ismanto HS, Nugraha DK, Xu Z, Horiguchi Y, Sakakibara S, Standley DM. Deciphering the antigen specificities of antibodies by clustering their complementarity determining region sequences. mSystems. 2023;8(6):e0072223. Epub 20231117. doi: 10.1128/msystems.00722-23. PubMed PMID: 37975681; PMCID: PMC10734444.
29. Mayer-Blackwell K, Schattgen S, Cohen-Lavi L, Crawford JC, Souquette A, Gaevert JA, Hertz T, Thomas PG, Bradley P, Fiore-Gartland A. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. eLife. 2021;10:e68605. doi: 10.7554/eLife.68605.
30. Edahiro R, Shirai Y, Takeshima Y, Sakakibara S, Yamaguchi Y, Murakami T, Morita T, Kato Y, Liu YC, Motooka D, Naito Y, Takuwa A, Sugihara F, Tanaka K, Wing JB, Sonehara K, Tomofuji Y, Japan C-TF, Namkoong H, Tanaka H, Lee H, Fukunaga K, Hirata H, Takeda Y, Okuzaki D, Kumanogoh A, Okada Y. Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. Nat Genet. 2023. Epub 20230424. doi: 10.1038/s41588-023-01375-1. PubMed PMID: 37095364.
31. Khan M, Yoo SJ, Clijsters M, Backaert W, Vanstapel A, Speleman K, Lietaer C, Choi S, Hether TD, Marcelis L, Nam A, Pan L, Reeves JW, Van Bulck P, Zhou H, Bourgeois M, Debaveye Y, De Munter P, Gunst J, Jorissen M, Lagrou K, Lorent N, Neyrinck A, Peetermans M, Thal DR, Vandenbriele C, Wauters J, Mombaerts P, Van Gerven L. Visualizing in deceased COVID-19 patients how SARS-CoV-2 attacks the respiratory and olfactory mucosae but spares the olfactory bulb. Cell. 2021;184(24):5932-49 e15. Epub 20211103. doi: 10.1016/j.cell.2021.10.027. PubMed PMID: 34798069; PMCID: PMC8564600.
32. Hayden MK, Hanson KE, Englund JA, Lee F, Lee MJ, Loeb M, Morgan DJ, Patel R, El Alayli A, El Mikati IK, Sultan S, Falck-Ytter Y, Mansour R, Amarin JZ, Morgan RL, Murad MH, Patel P, Bhimraj A, Mustafa RA. The Infectious Diseases Society of America Guidelines on

the Diagnosis of COVID-19: Antigen Testing. Clin Infect Dis. 2023. Epub 20230126. doi: 10.1093/cid/ciad032. PubMed PMID: 36702617.
33. Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, Petrella F, Spaggiari L, Rosell R. Non-small-cell lung cancer. Nature Reviews Disease Primers. 2015;1(1):15009. doi: 10.1038/nrdp.2015.9.
34. Tomasik B, Skrzypski M, Bienkowski M, Dziadziuszko R, Jassem J. Current and future applications of liquid biopsy in non-small-cell lung cancer-a narrative review. Transl Lung Cancer Res. 2023;12(3):594-614. Epub 20230309. doi: 10.21037/tlcr-22-742. PubMed PMID: 37057121; PMCID: PMC10087994.
35. Thunnissen E, Kerr KM, Herth FJ, Lantuejoul S, Papotti M, Rintoul RC, Rossi G, Skov BG, Weynand B, Bubendorf L, Katrien G, Johansson L, Lopez-Rios F, Ninane V, Olszewski W, Popper H, Jaume S, Schnabel P, Thiberville L, Laenger F. The challenge of NSCLC diagnosis and predictive analysis on small samples. Practical approach of a working group. Lung Cancer. 2012;76(1):1-18. Epub 20111203. doi: 10.1016/j.lungcan.2011.10.017. PubMed PMID: 22138001.
36. de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermsen R, Chain B, de Boer RJ. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. Elife. 2020;9. Epub 20200318. doi: 10.7554/eLife.49900. PubMed PMID: 32187010; PMCID: PMC7080410.
37. Schultheiss C, Simnica D, Willscher E, Oberle A, Fanchi L, Bonzanni N, Wildner NH, Schulze Zur Wiesch J, Weiler-Normann C, Lohse AW, Binder M. Next-Generation Immunosequencing Reveals Pathological T-Cell Architecture in Autoimmune Hepatitis. Hepatology. 2021;73(4):1436-48. Epub 20210208. doi: 10.1002/hep.31473. PubMed PMID: 32692457.
38. Swanson PA, 2nd, Padilla M, Hoyland W, McGlinchey K, Fields PA, Bibi S, Faust SN, McDermott AB, Lambe T, Pollard AJ, Durham NM, Kelly EJ, AstraZeneca/Oxford VRCSG. AZD1222/ChAdOx1 nCoV-19 vaccination induces a polyfunctional spike protein-specific T(H)1 response with a diverse TCR repertoire. Sci Transl Med. 2021;13(620):eabj7211. Epub 20211117. doi: 10.1126/scitranslmed.abj7211. PubMed PMID: 34591596; PMCID: PMC9924073.
39. Reuben A, Zhang J, Chiou SH, Gittelman RM, Li J, Lee WC, Fujimoto J, Behrens C, Liu X, Wang F, Quek K, Wang C, Kheradmand F, Chen R, Chow CW, Lin H, Bernatchez C, Jalali A, Hu X, Wu CJ, Eterovic AK, Parra ER, Yusko E, Emerson R, Benzeno S, Vignali M, Wu X, Ye Y, Little LD, Gumbs C, Mao X, Song X, Tippen S, Thornton RL, Cascone T, Snyder A, Wargo JA, Herbst R, Swisher S, Kadara H, Moran C, Kalhor N, Zhang J, Scheet P, Vaporciyan AA, Sepesi B, Gibbons DL, Robins H, Hwu P, Heymach JV, Sharma P, Allison JP, Baladandayuthapani V, Lee JJ, Davis MM, Wistuba, II, Futreal PA, Zhang J. Comprehensive T cell repertoire characterization of non-small cell lung cancer. Nat Commun. 2020;11(1):603. Epub 20200130. doi: 10.1038/s41467-019-14273-0. PubMed PMID: 32001676; PMCID: PMC6992630.
40. Wu L, Zhu J, Rudqvist NP, Welsh J, Lee P, Liao Z, Xu T, Jiang M, Zhu X, Pan X, Li P, Zhou Z, He X, Yin R, Feng J. T-Cell Receptor Profiling and Prognosis After Stereotactic Body Radiation Therapy For Stage I Non-Small-Cell Lung Cancer. Front Immunol.

2021;12:719285. Epub 20211018. doi: 10.3389/fimmu.2021.719285. PubMed PMID: 34733273; PMCID: PMC8559517.
41. Hamm DE. ImmuneAccess immunoSEQ hsTCRB-V4b Control Data2020. doi: https://doi.org/10.21417/ADPT2020V4CD.
42. Joshi K, de Massy MR, Ismail M, Reading JL, Uddin I, Woolston A, Hatipoglu E, Oakes T, Rosenthal R, Peacock T, Ronel T, Noursadeghi M, Turati V, Furness AJS, Georgiou A, Wong YNS, Ben Aissa A, Sunderland MW, Jamal-Hanjani M, Veeriah S, Birkbak NJ, Wilson GA, Hiley CT, Ghorani E, Guerra-Assuncao JA, Herrero J, Enver T, Hadrup SR, Hackshaw A, Peggs KS, McGranahan N, Swanton C, consortium TR, Quezada SA, Chain B. Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. Nat Med. 2019;25(10):1549-59. Epub 20191007. doi: 10.1038/s41591-019-0592-2. PubMed PMID: 31591606; PMCID: PMC6890490.
43. Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. BMC Bioinformatics. 2022;23(Suppl 6):279. Epub 20220714. doi: 10.1186/s12859-022-04775-y. PubMed PMID: 35836114; PMCID: PMC9284682.
44. 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(18):2924-9. doi: 10.1093/bioinformatics/btx286. PubMed PMID: 28481982.
45. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, Komech EA, Sycheva AL, Koneva AE, Egorov ES, Eliseev AV, Van Dyk E, Dash P, Attaf M, Rius C, Ladell K, McLaren JE, Matthews KK, Clemens EB, Douek DC, Luciani F, van Baarle D, Kedzierska K, Kesmir C, Thomas PG, Price DA, Sewell AK, Chudakov DM. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. 2018;46(D1):D419-D27. doi: 10.1093/nar/gkx760. PubMed PMID: 28977646; PMCID: PMC5753233.
46. Der SD, Sykes J, Pintilie M, Zhu CQ, Strumpf D, Liu N, Jurisica I, Shepherd FA, Tsao MS. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol. 2014;9(1):59-64. doi: 10.1097/JTO.0000000000000042. PubMed PMID: 24305008.
47. Brannan JM, Sen B, Saigal B, Prudkin L, Behrens C, Solis L, Dong W, Bekele BN, Wistuba I, Johnson FM. EphA2 in the early pathogenesis and progression of non-small cell lung cancer. Cancer Prev Res (Phila). 2009;2(12):1039-49. Epub 20091124. doi: 10.1158/1940-6207.CAPR-09-0212. PubMed PMID: 19934338.
48. Suzuki K, Kachala SS, Kadota K, Shen R, Mo Q, Beer DG, Rusch VW, Travis WD, Adusumilli PS. Prognostic immune markers in non-small cell lung cancer. Clin Cancer Res. 2011;17(16):5247-56. Epub 20110609. doi: 10.1158/1078-0432.CCR-10-2805. PubMed PMID: 21659461.
49. Niu C, Qiu W, Li X, Li H, Zhou J, Zhu H. Transketolase Serves as a Biomarker for Poor Prognosis in Human Lung Adenocarcinoma. J Cancer. 2022;13(8):2584-93. Epub 20220513. doi: 10.7150/jca.69583. PubMed PMID: 35711845; PMCID: PMC9174852.
50. Han L, Lei G, Chen Z, Zhang Y, Huang C, Chen W. IGF2BP2 Regulates MALAT1 by Serving as an N6-Methyladenosine Reader to Promote NSCLC Proliferation. Front Mol

Biosci. 2021;8:780089. Epub 20220117. doi: 10.3389/fmolb.2021.780089. PubMed PMID: 35111811; PMCID: PMC8802805.

51. Dolton G, Rius C, Wall A, Szomolay B, Bianchi V, Galloway SAE, Hasan MS, Morin T, Caillaud ME, Thomas HL, Theaker S, Tan LR, Fuller A, Topley K, Legut M, Attaf M, Hopkins JR, Behiry E, Zabkiewicz J, Alvares C, Lloyd A, Rogers A, Henley P, Fegan C, Ottmann O, Man S, Crowther MD, Donia M, Svane IM, Cole DK, Brown PE, Rizkallah P, Sewell AK. Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. Cell. 2023;186(16):3333-49 e27. Epub 20230724. doi: 10.1016/j.cell.2023.06.020. PubMed PMID: 37490916.

52. Jerne NK. Towards a network theory of the immune system. Ann Immunol (Paris). 1974;125C(1-2):373-89. PubMed PMID: 4142565.

53. Hoffmann GW. A theory of regulation and self-nonself discrimination in an immune network. Eur J Immunol. 1975;5(9):638-47. doi: 10.1002/eji.1830050912. PubMed PMID: 11993326.

54. Huang C, Li X, Wu J, Zhang W, Sun S, Lin L, Wang X, Li H, Wu X, Zhang P, Xu G, Wang H, Liu H, Liu Y, Chen D, Zhuo L, Li W, Yang H, Wang J, Wang L, Liu X. The landscape and diagnostic potential of T and B cell repertoire in Immunoglobulin A Nephropathy. J Autoimmun. 2019;97:100-7. Epub 20181029. doi: 10.1016/j.jaut.2018.10.018. PubMed PMID: 30385082.

55. Liu X, Zhang W, Zhao M, Fu L, Liu L, Wu J, Luo S, Wang L, Wang Z, Lin L, Liu Y, Wang S, Yang Y, Luo L, Jiang J, Wang X, Tan Y, Li T, Zhu B, Zhao Y, Gao X, Wan Z, Huang C, Fang M, Li Q, Peng H, Liao X, Chen J, Li F, Ling G, Zhao H, Luo H, Xiang Z, Liao J, Liu Y, Yin H, Long H, Wu H, Yang H, Wang J, Lu Q. T cell receptor beta repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. Ann Rheum Dis. 2019;78(8):1070-8. Epub 20190517. doi: 10.1136/annrheumdis-2019-215442. PubMed PMID: 31101603.

56. Brummel K, Eerkens AL, de Bruyn M, Nijman HW. Tumour-infiltrating lymphocytes: from prognosis to treatment selection. Br J Cancer. 2023;128(3):451-8. Epub 20221223. doi: 10.1038/s41416-022-02119-4. PubMed PMID: 36564565; PMCID: PMC9938191.

57. Bochem J, Zelba H, Spreuer J, Amaral T, Wagner NB, Gaissler A, Pop OT, Thiel K, Yurttas C, Soffel D, Forchhammer S, Sinnberg T, Niessner H, Meier F, Terheyden P, Konigsrainer A, Garbe C, Flatz L, Pawelec G, Eigentler TK, Loffler MW, Weide B, Wistuba-Hamprecht K. 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. 2021;9(12). doi: 10.1136/jitc-2021-003439. PubMed PMID: 34933966; PMCID: PMC8693089.

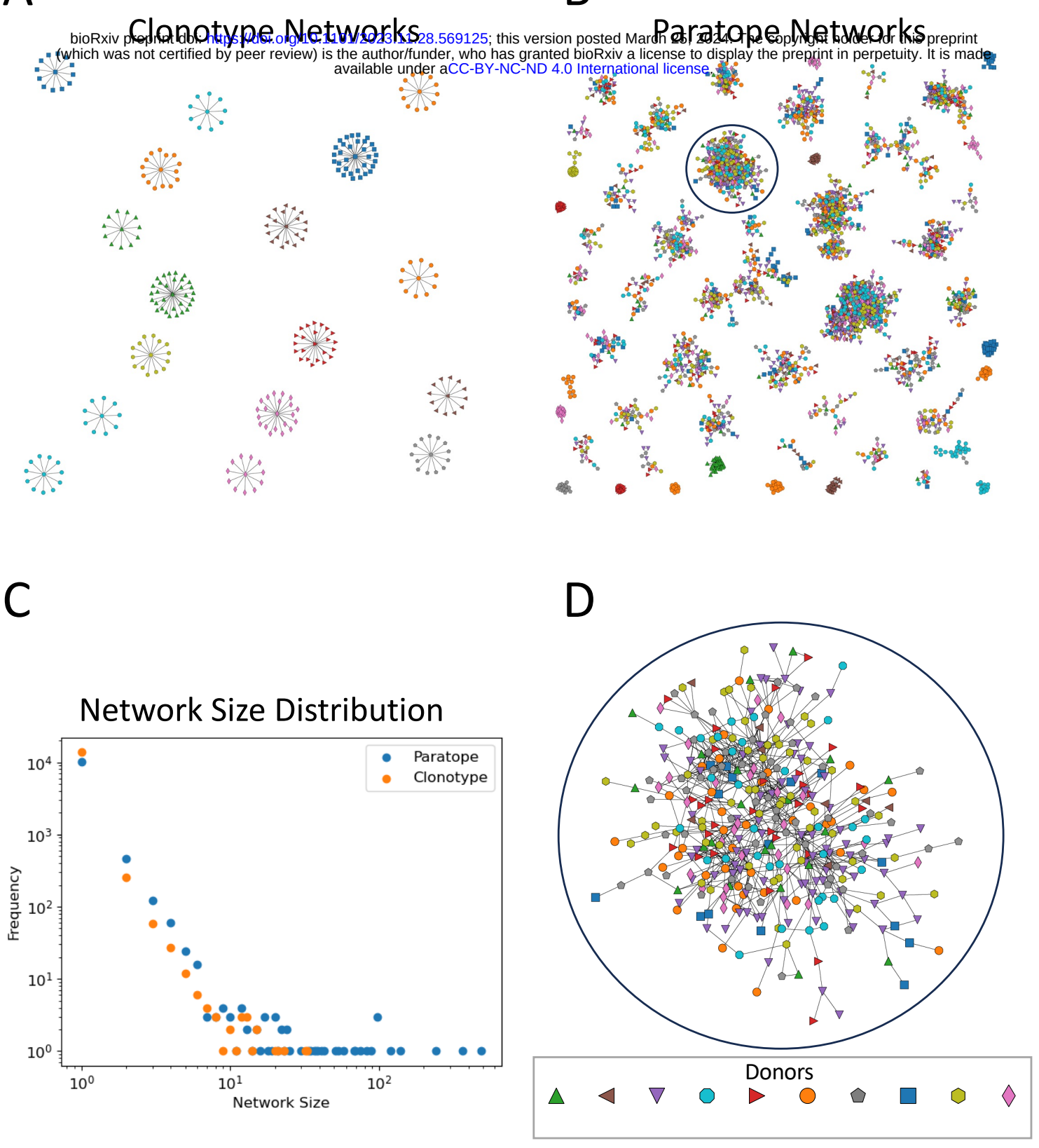


Figure 1

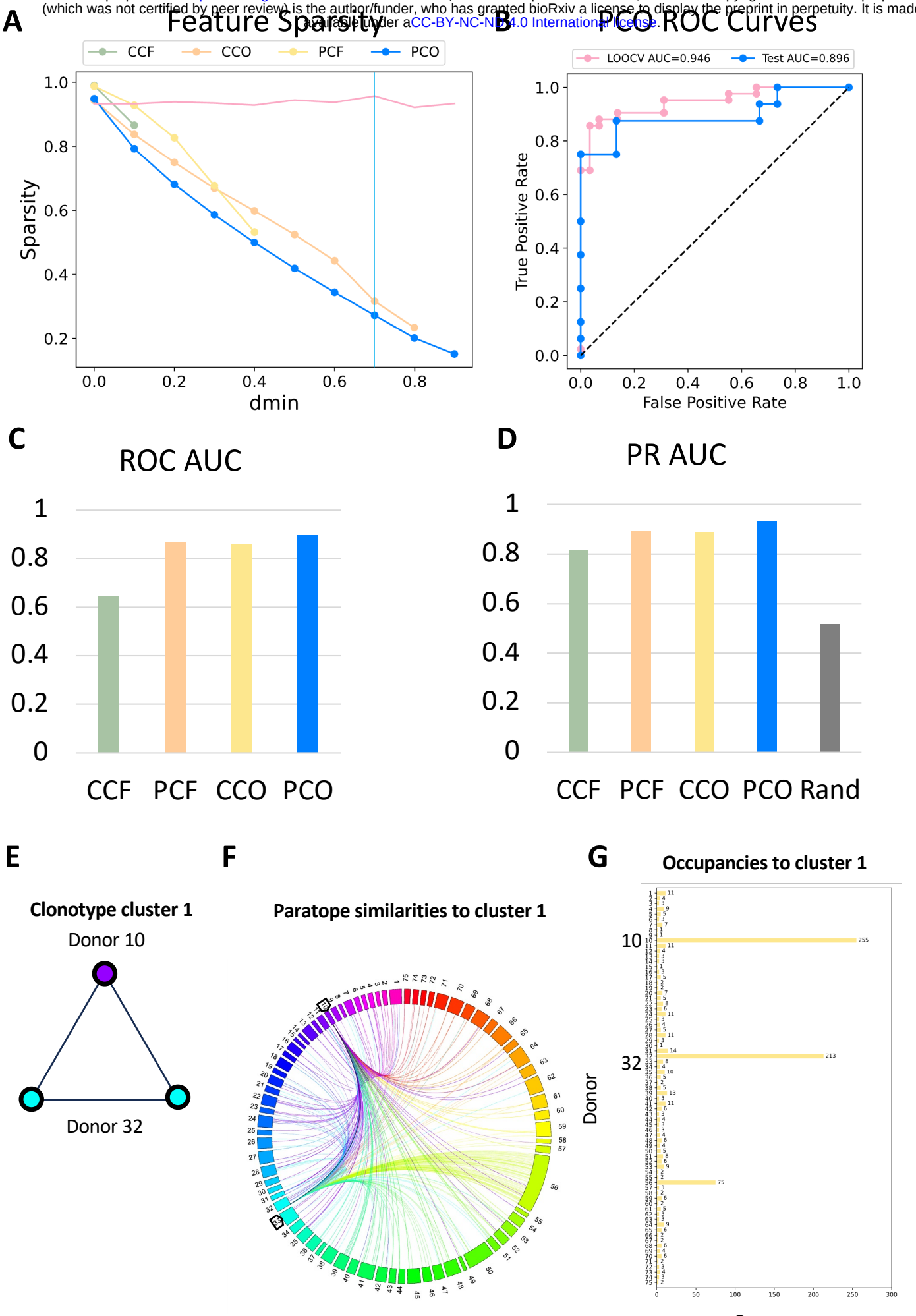


Figure 2

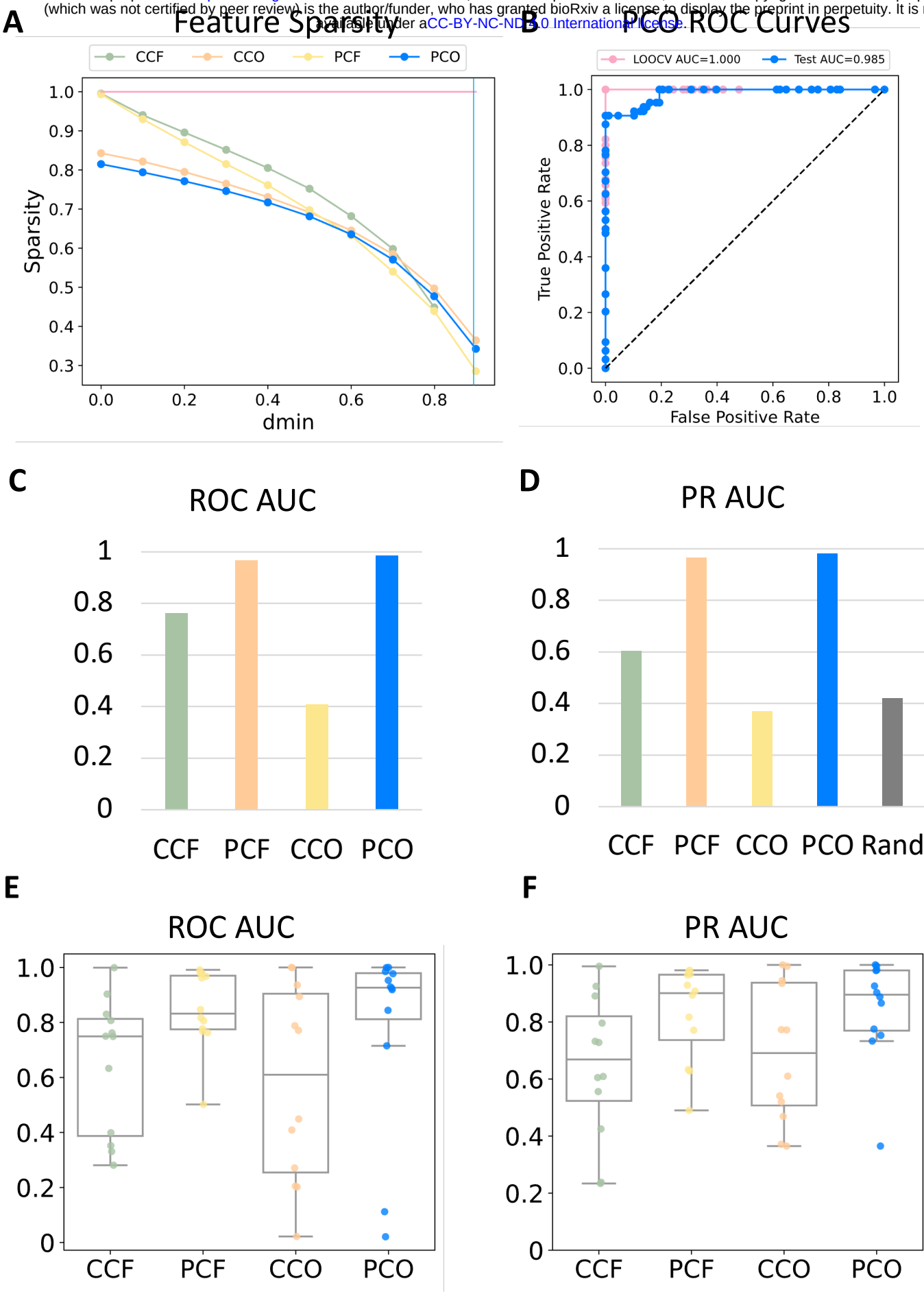


Figure 3

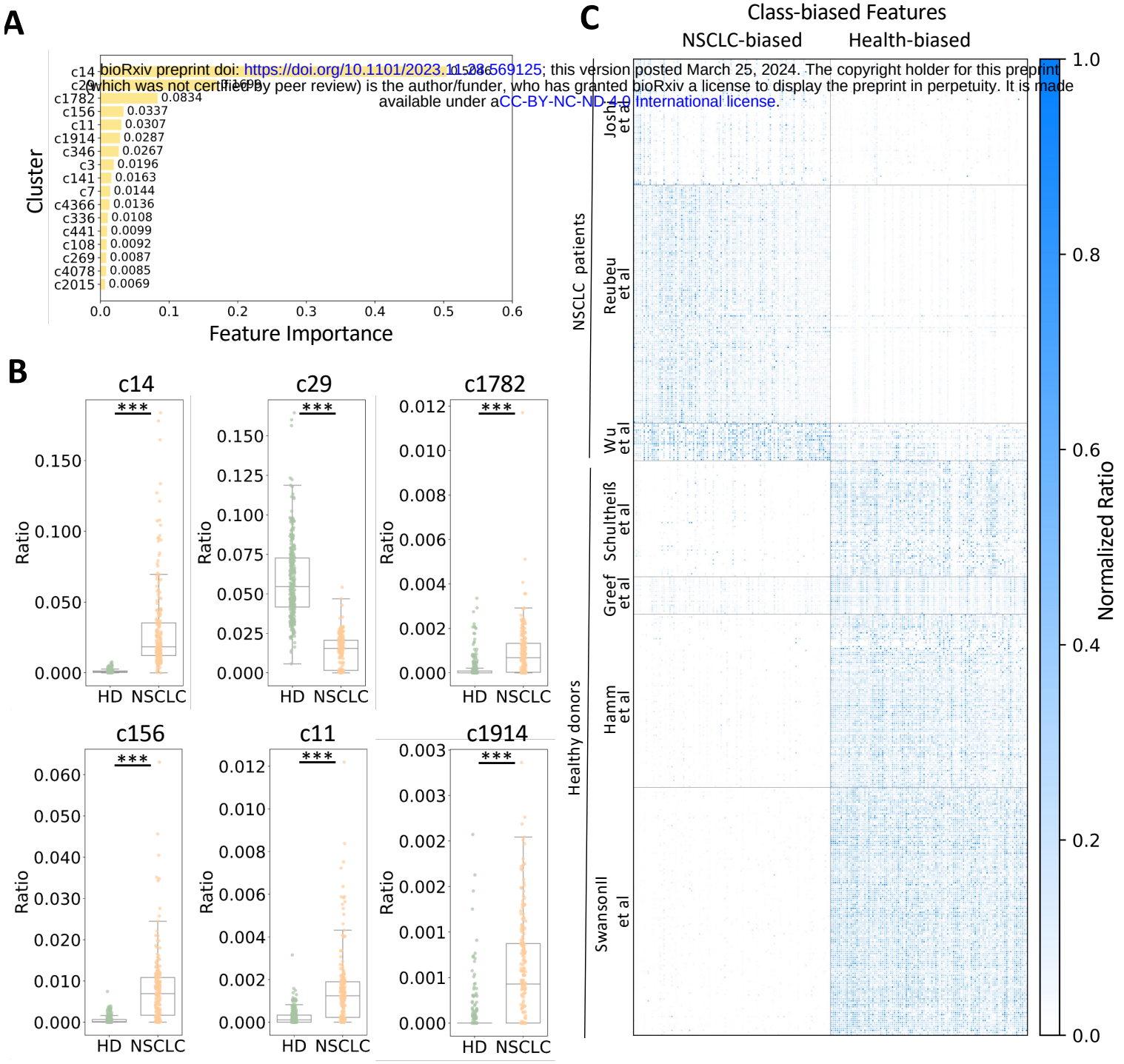
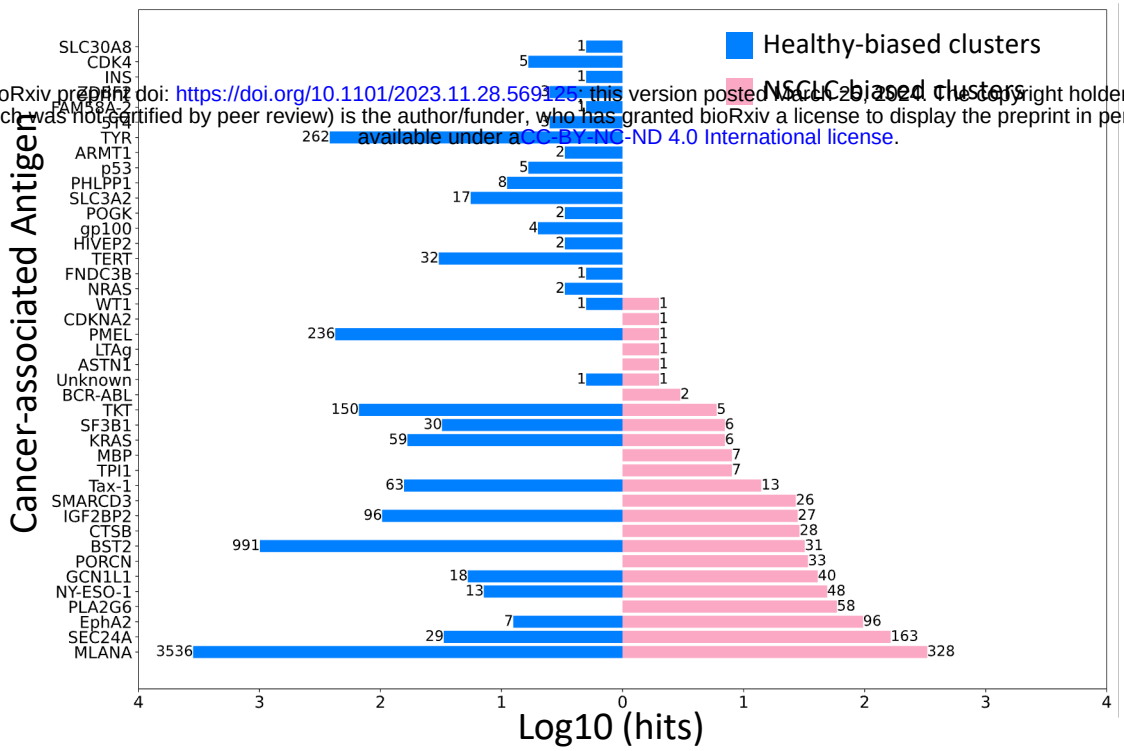


Figure 4

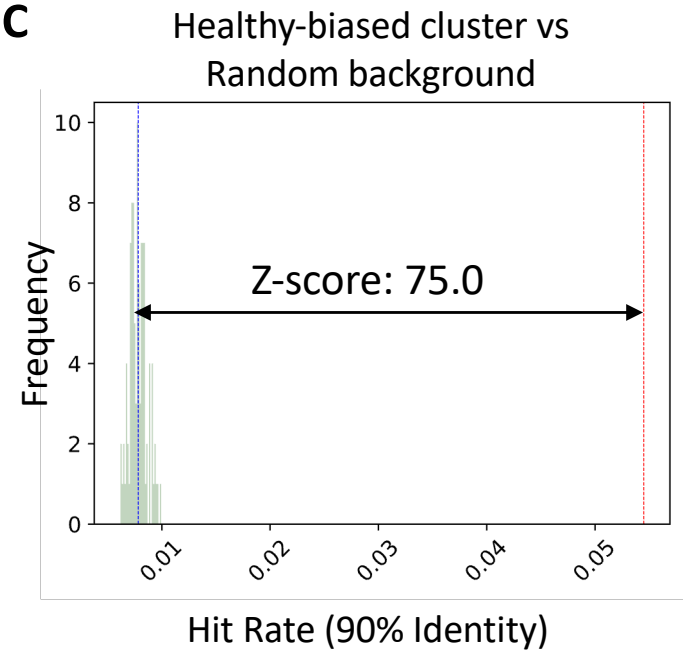
A



B

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

C



D

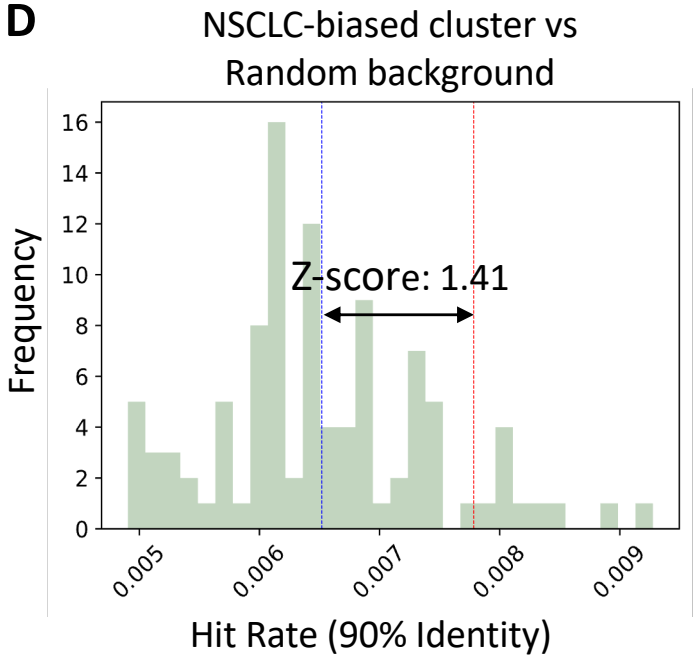
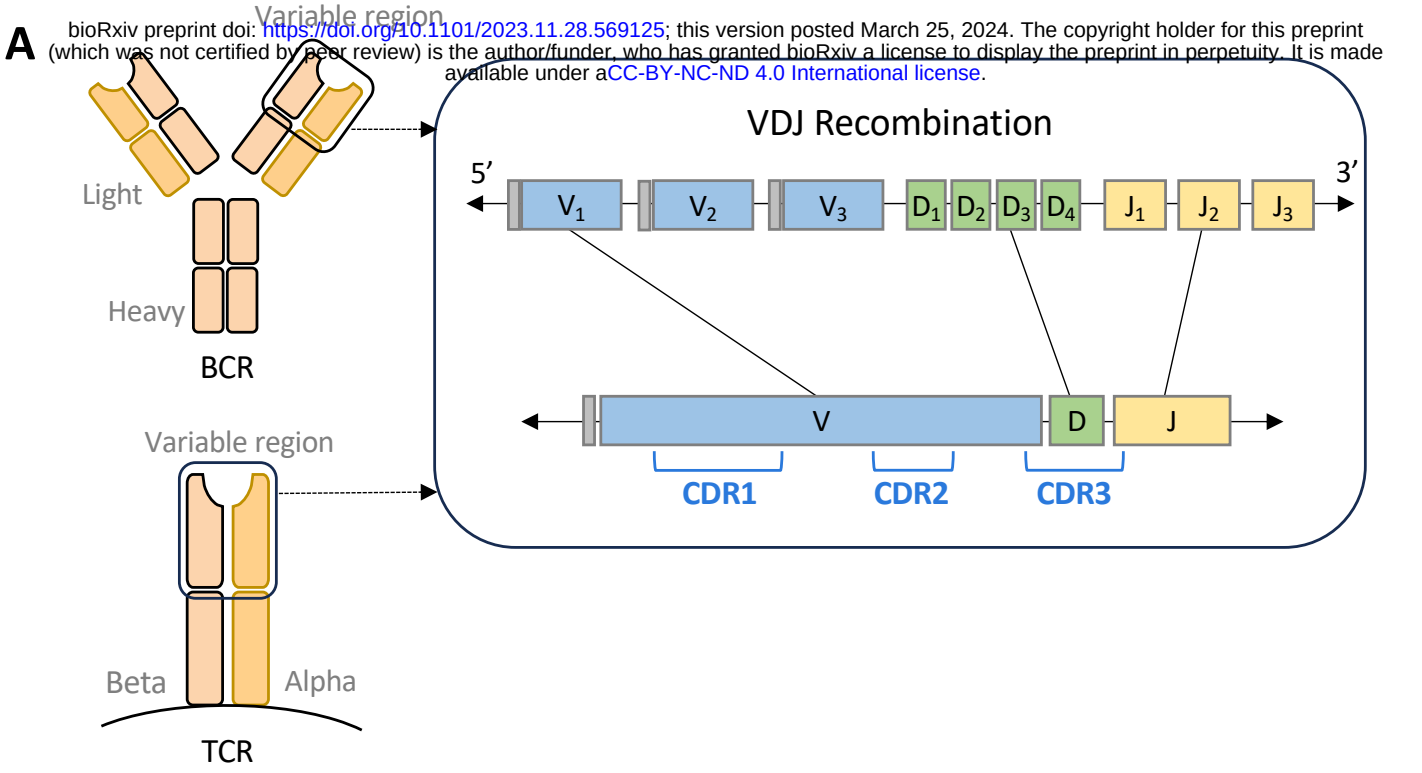
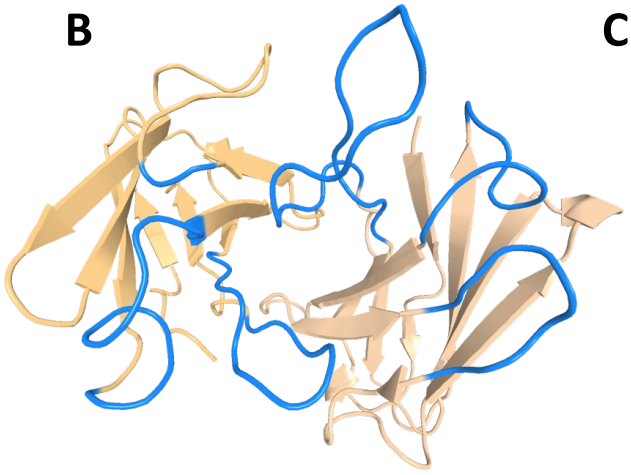
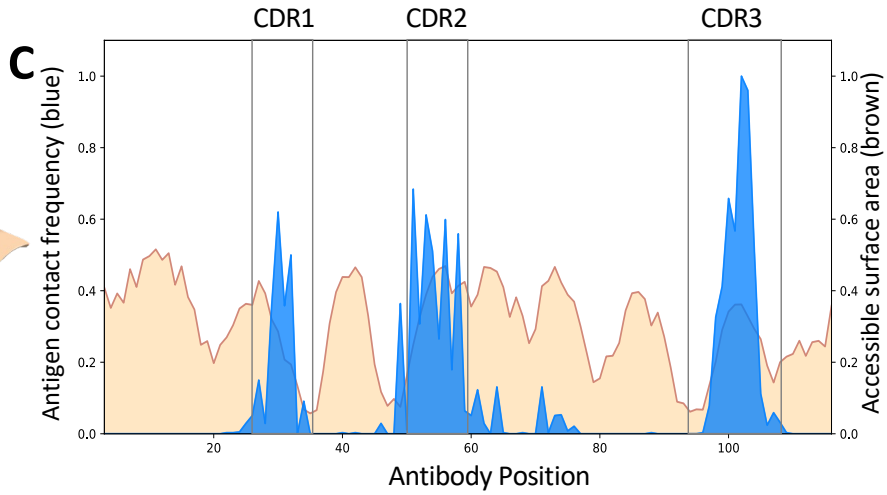


Figure 5







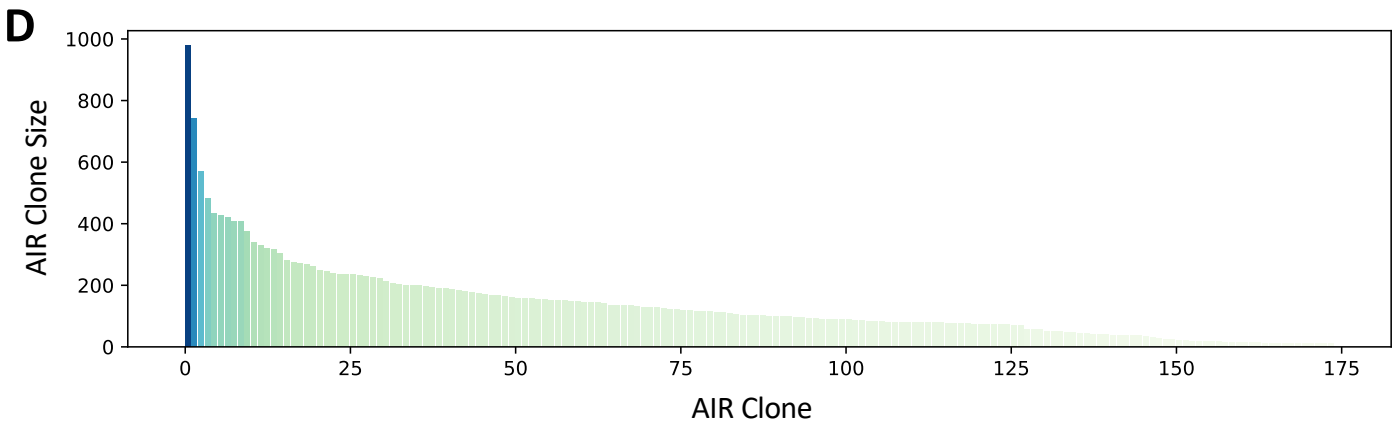


Figure S1

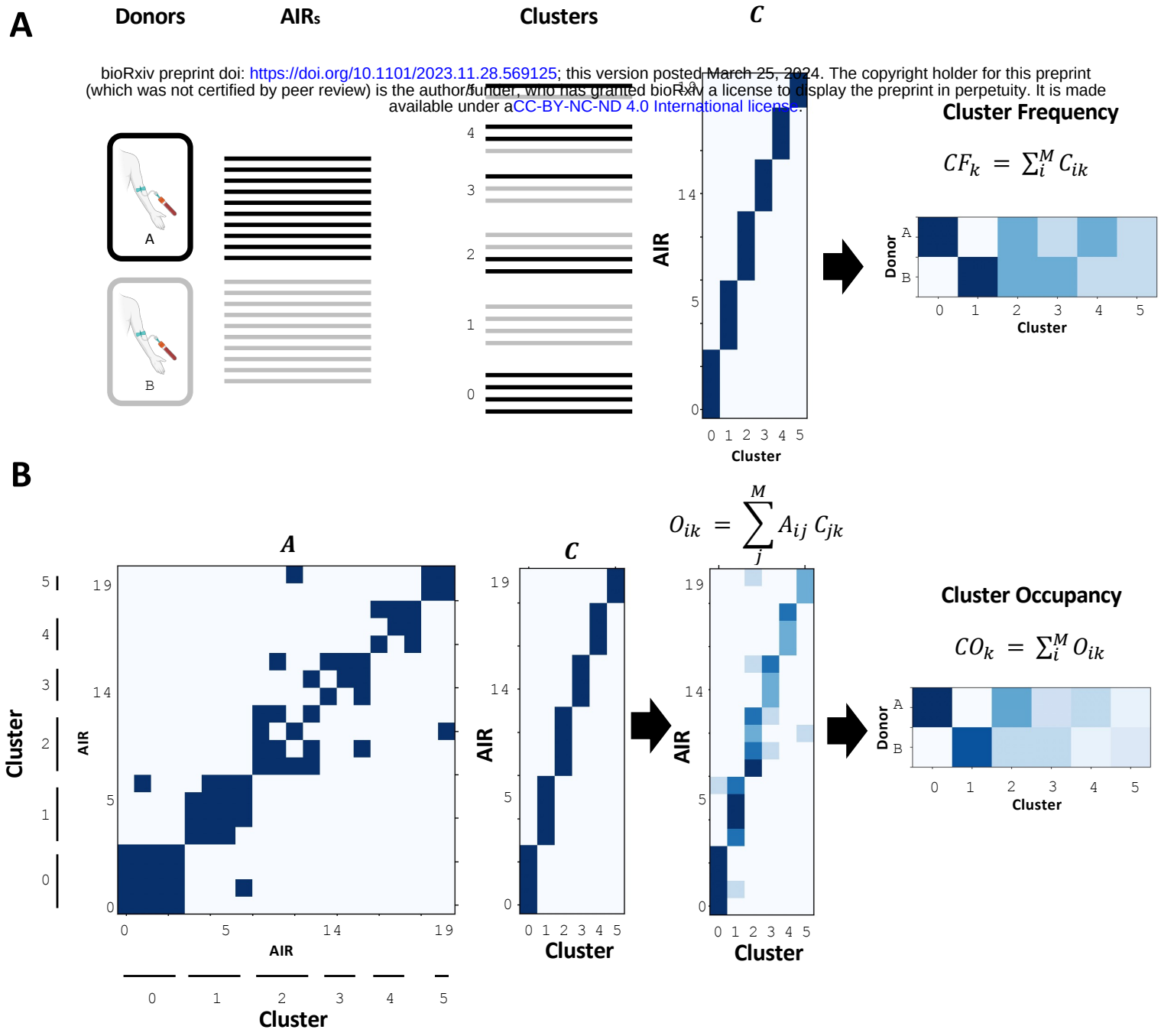


Figure S2

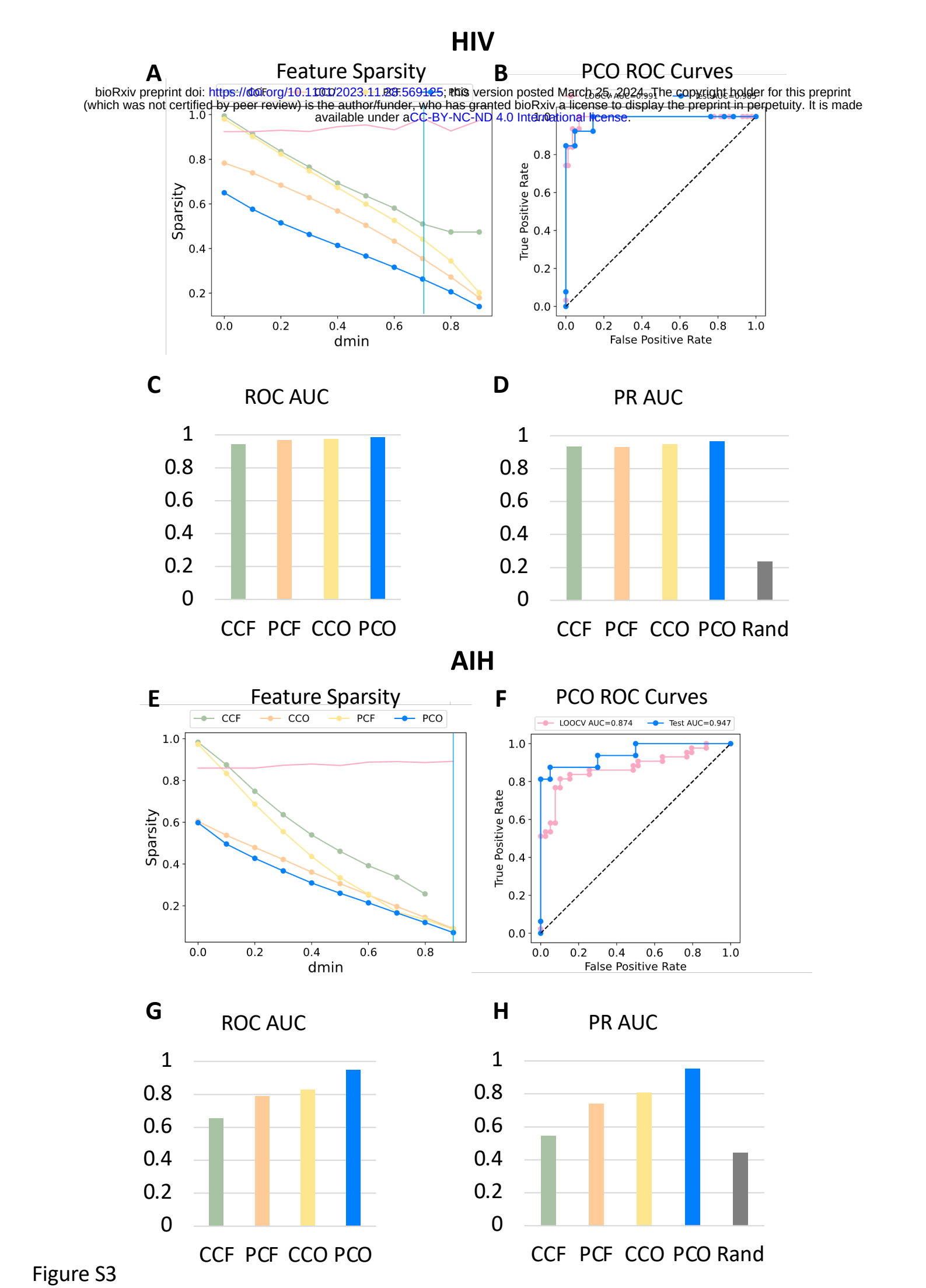


Figure S3

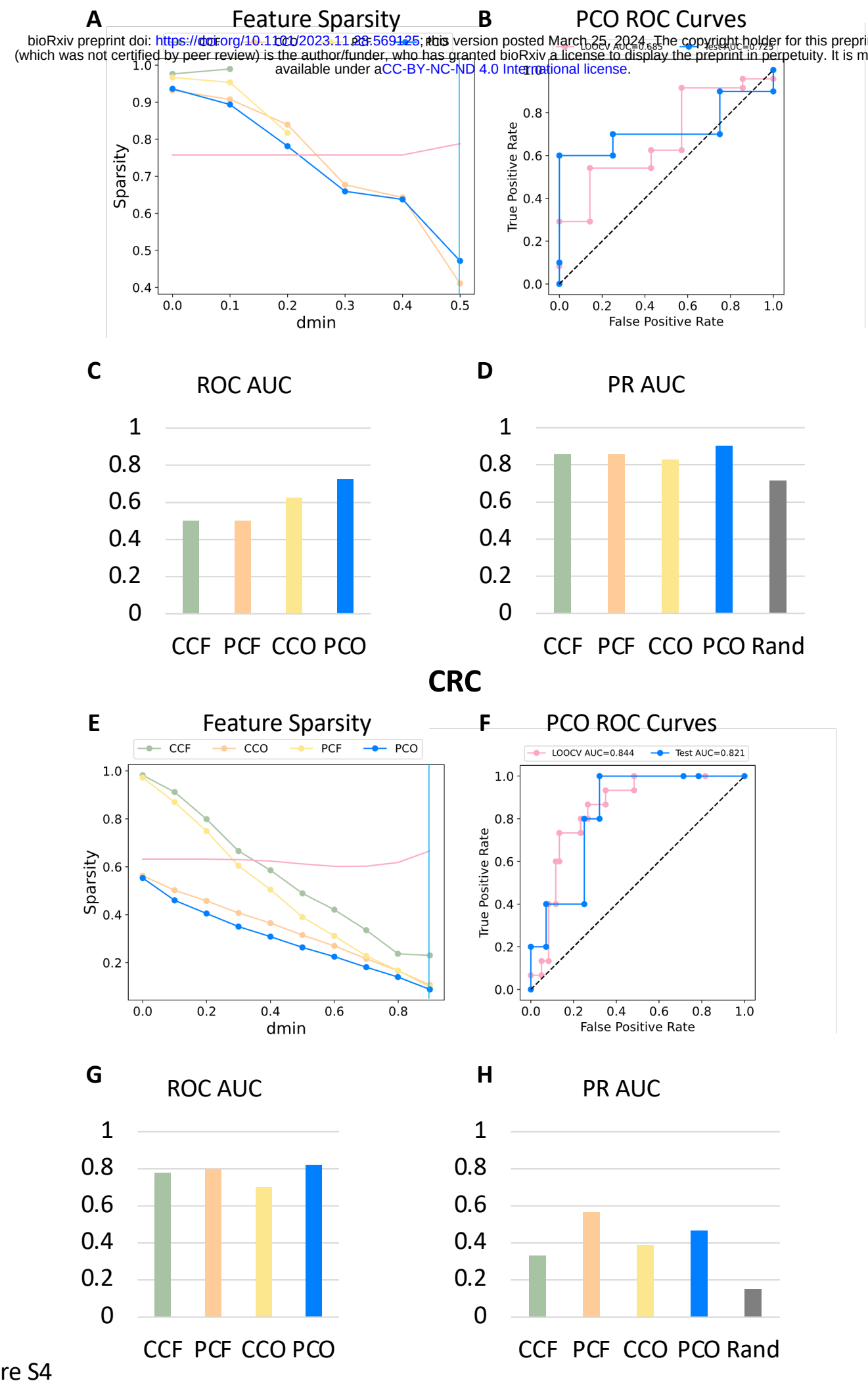


Figure S4

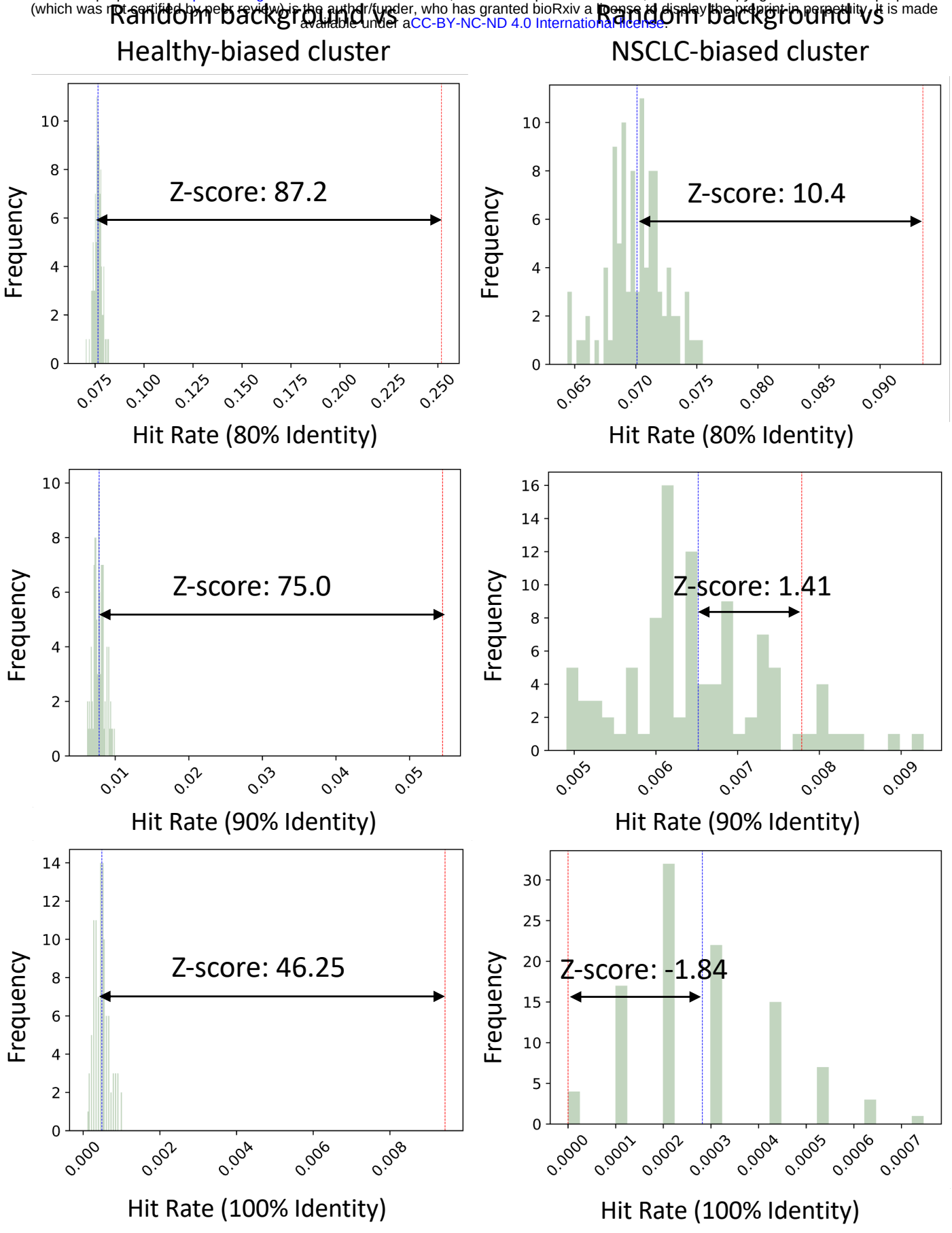
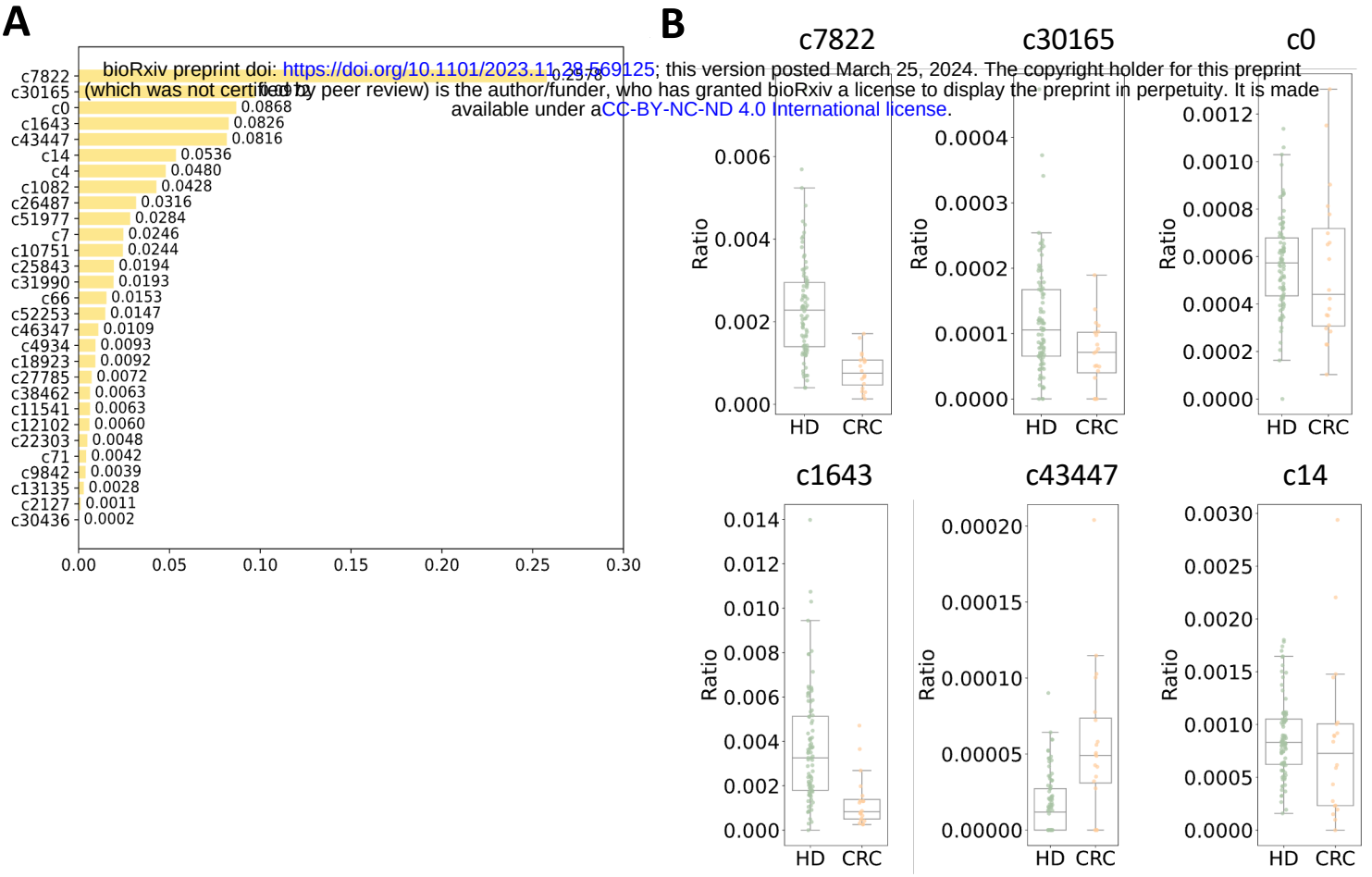


Figure S5



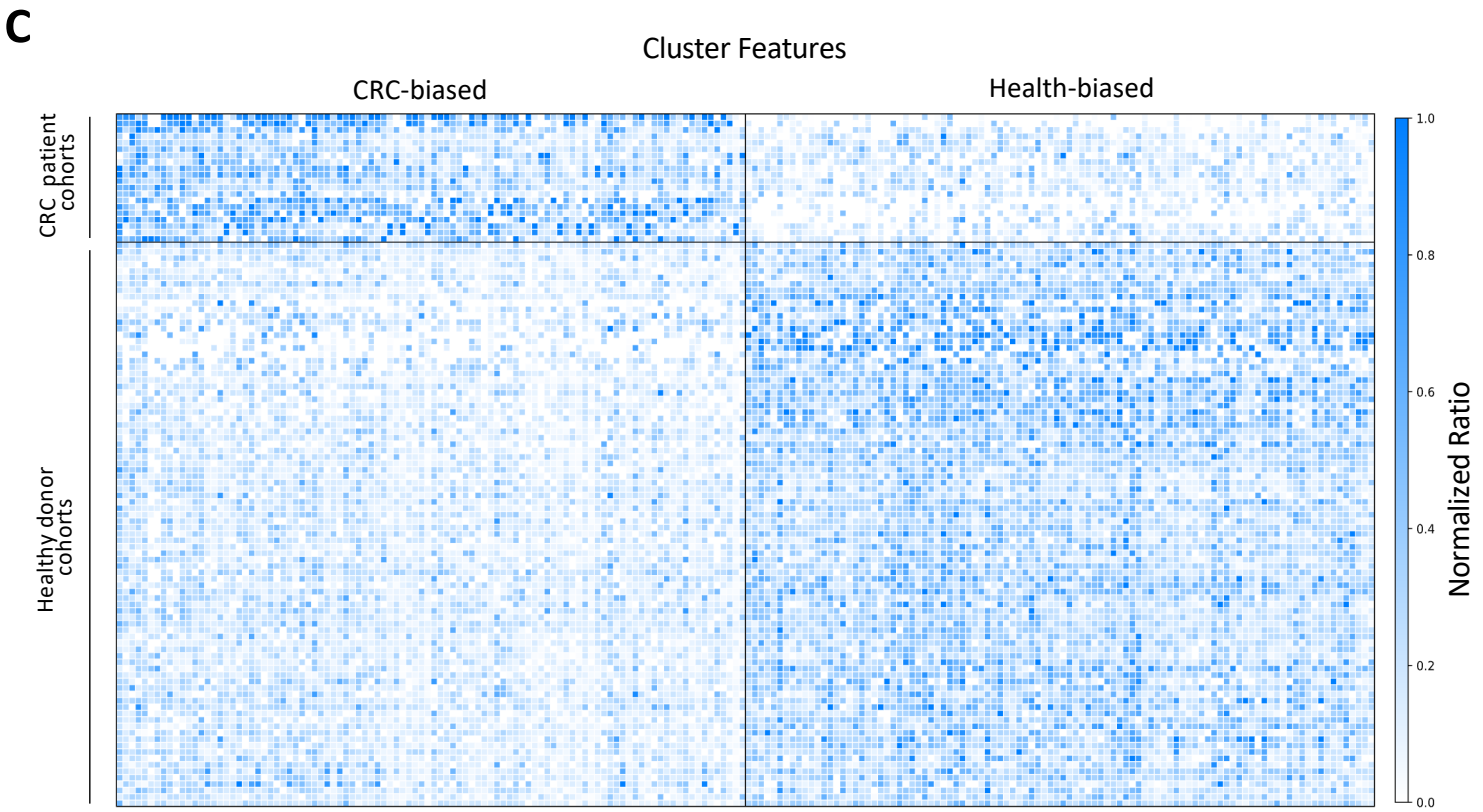
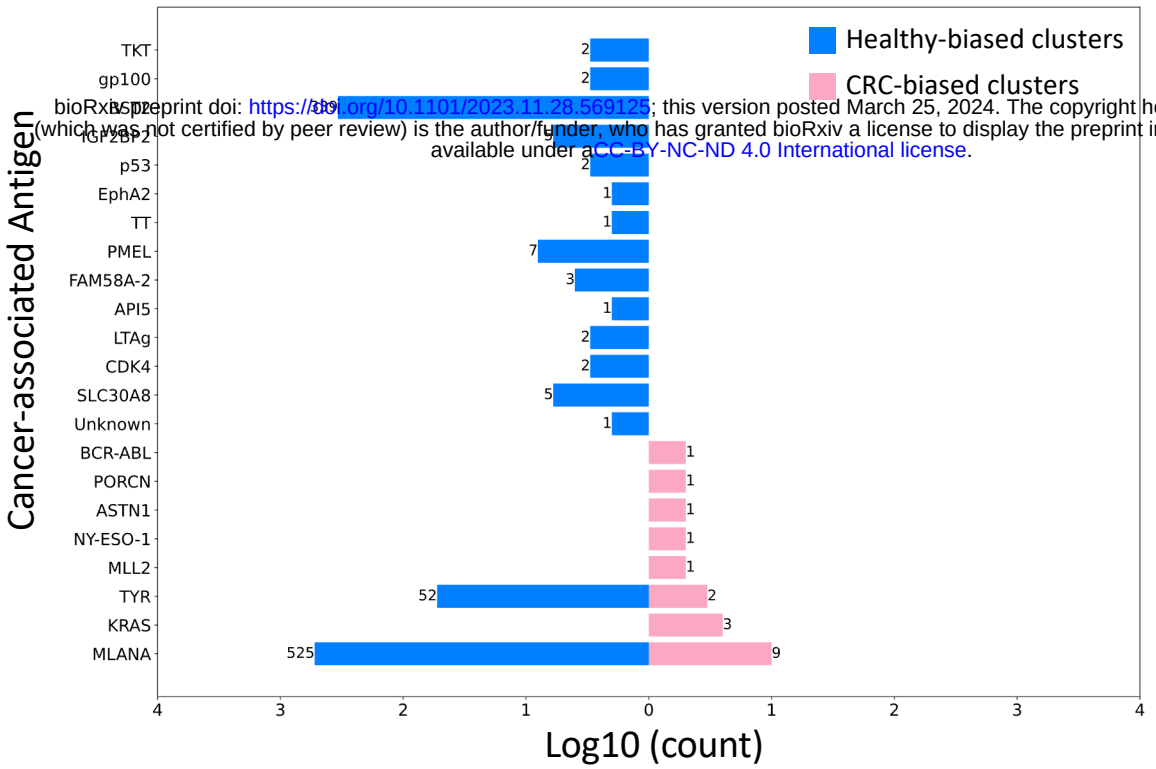


Figure S6

A



B

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

C

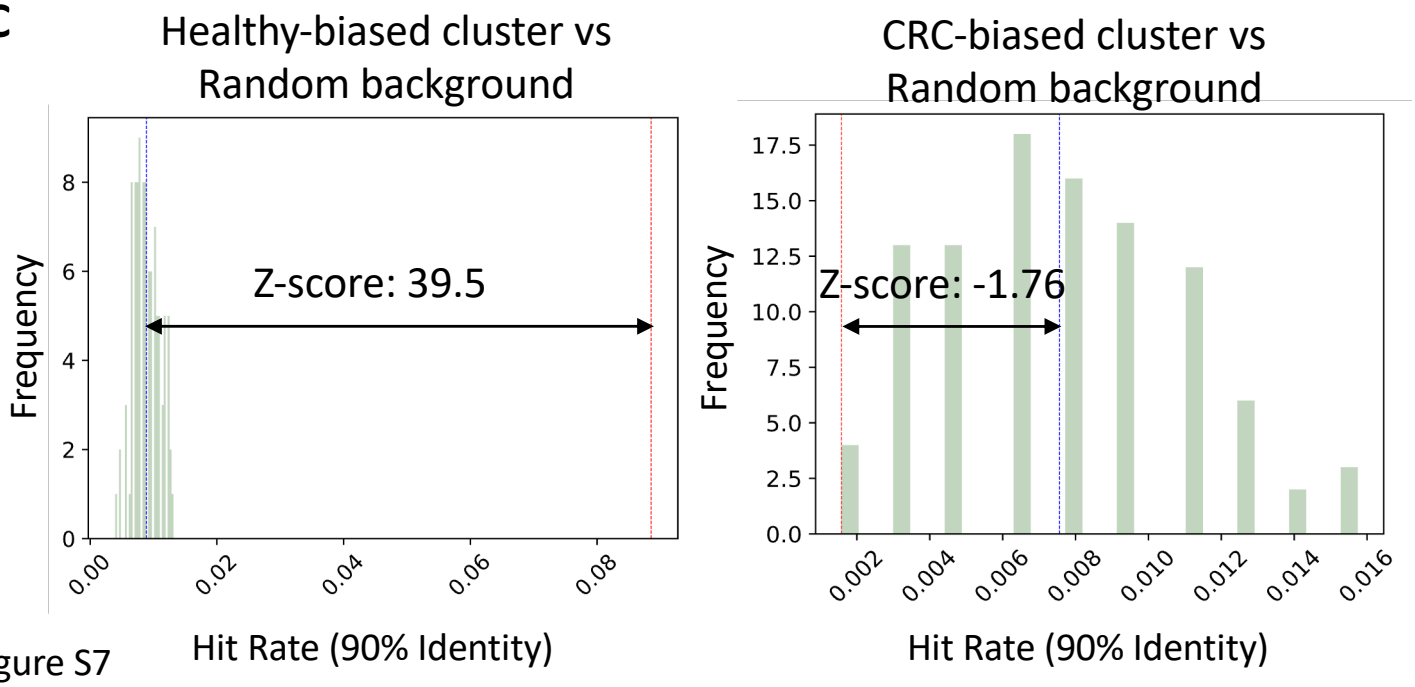


Figure S7

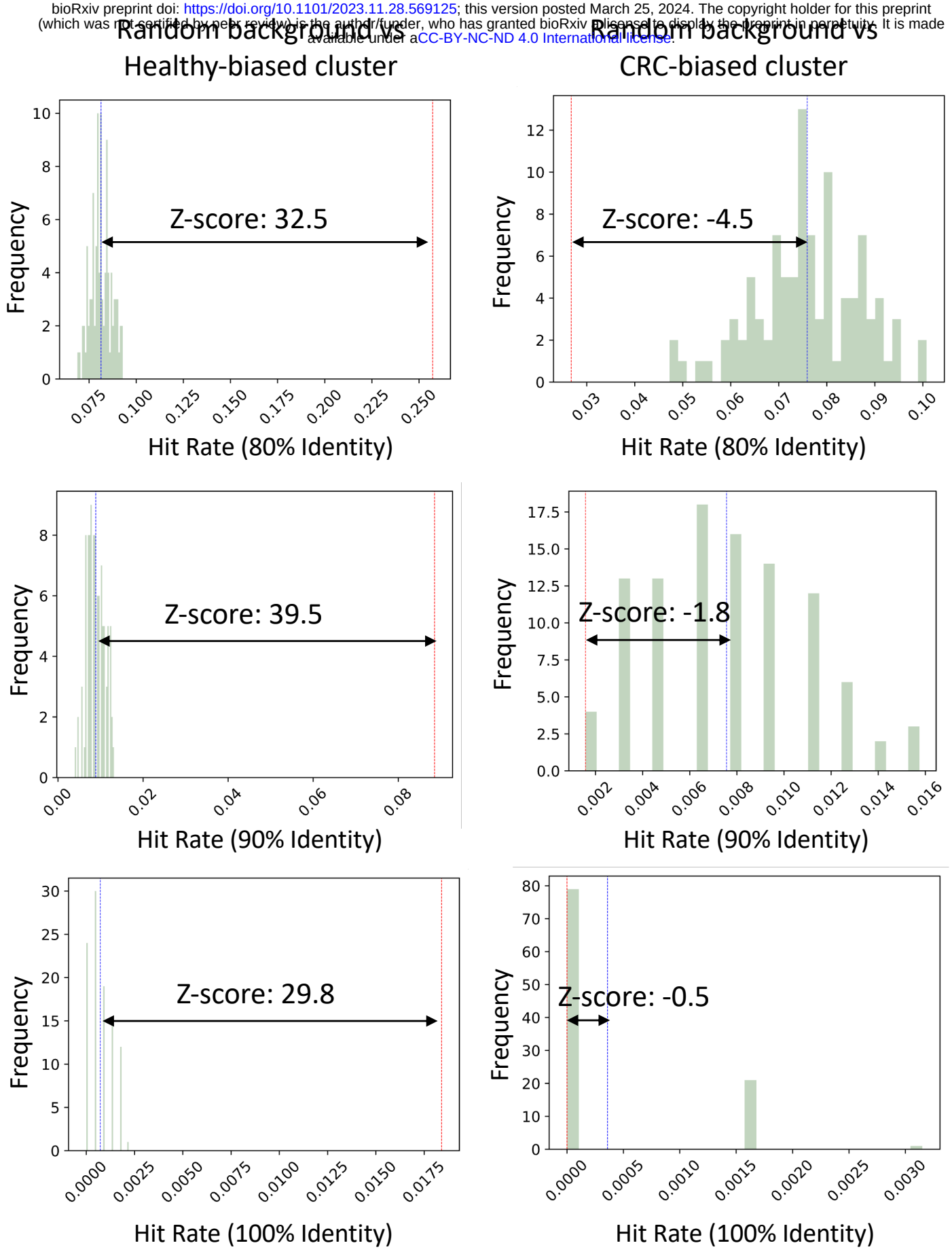
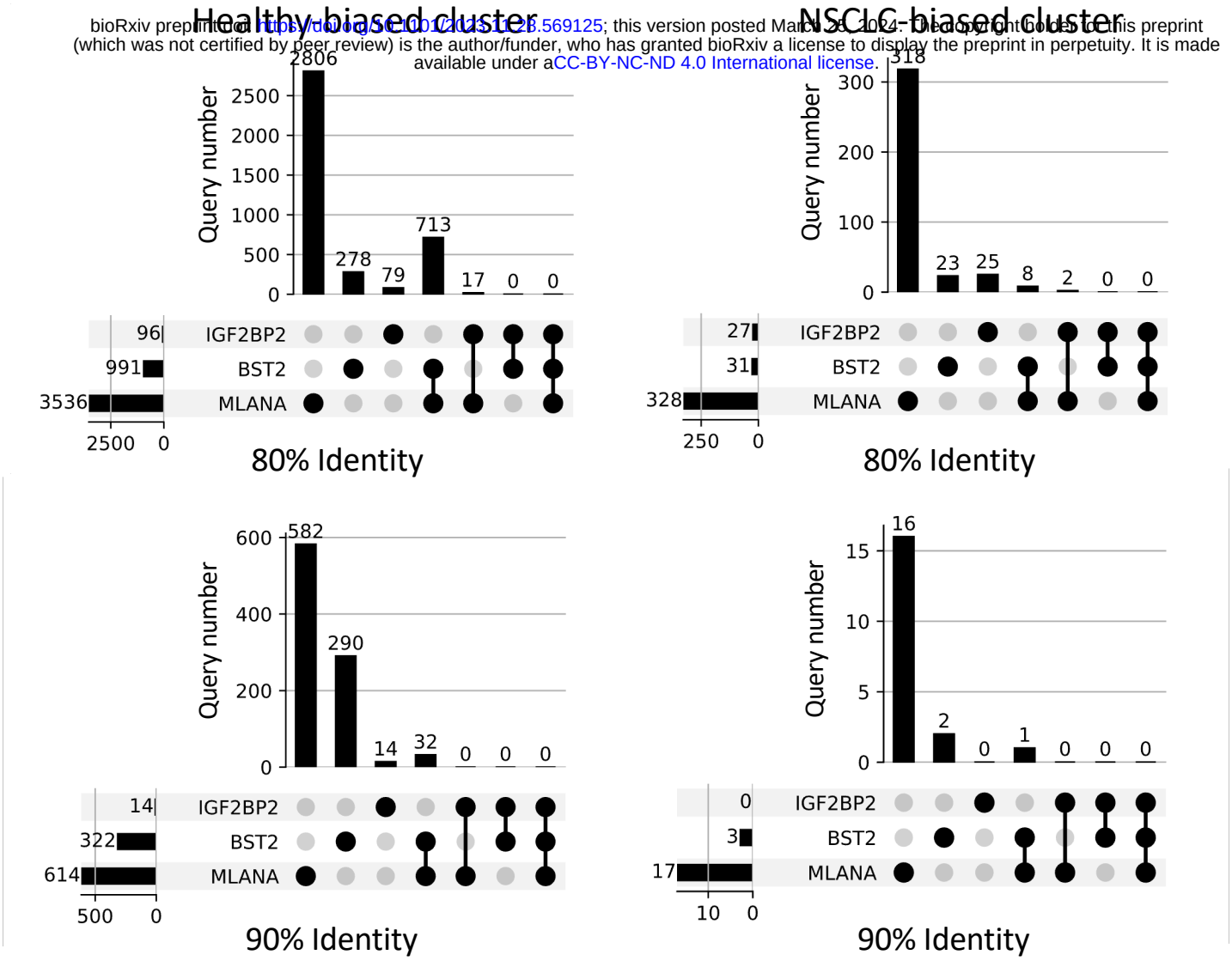


Figure S8

A

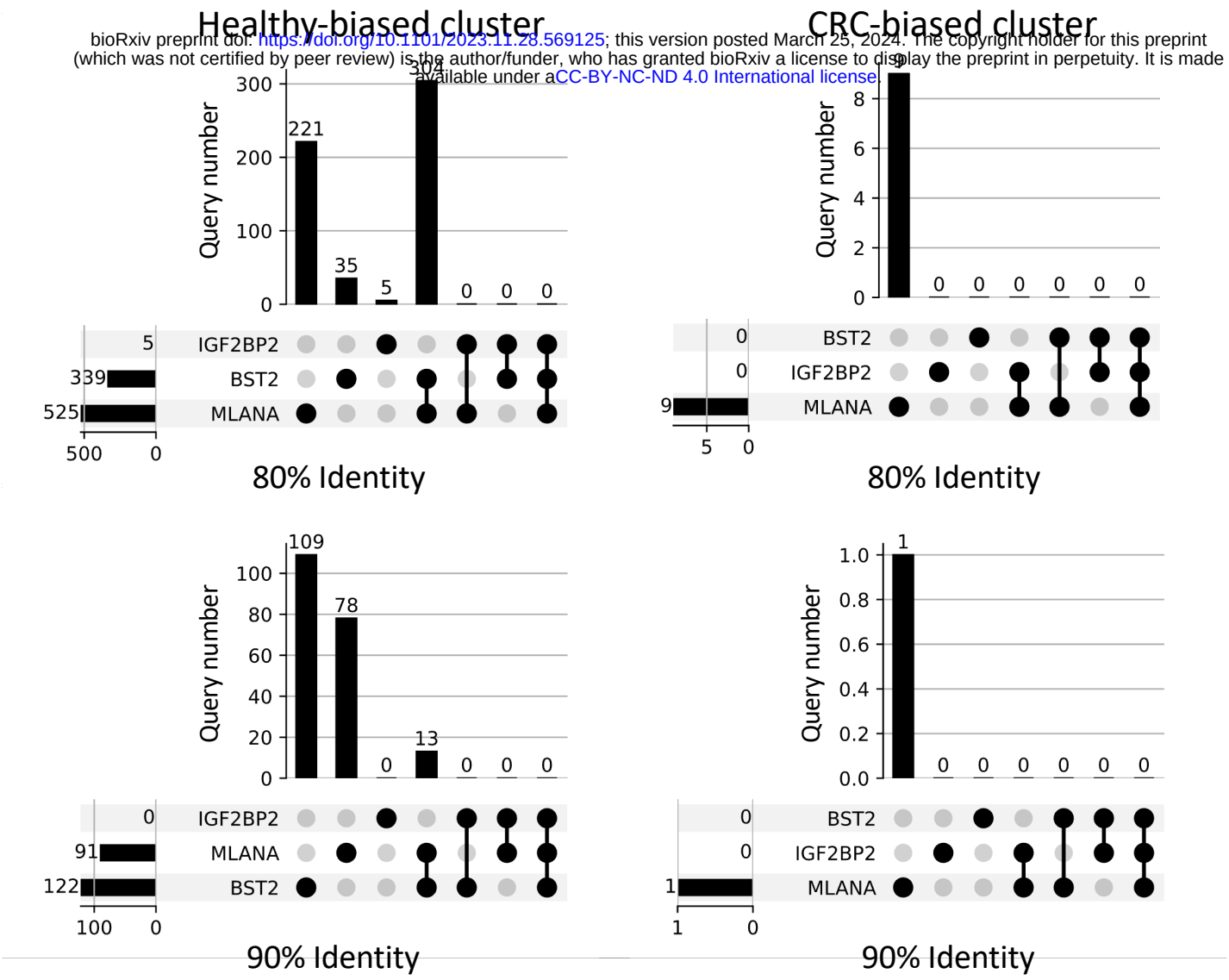


B

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 S9

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 S10