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], Jan Wilamowski[2], Shunsuke Teraguchi[2,3], Ayan Sengupta[4], 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]Faculty Data Science, Shiga University, Hikone 522-8522, Japan
[4]Cogent Labs, Tokyo 106-032, Japan

Abstract

Adaptive immune receptor repertoires (AIRRs) have emerged as promising biomarkers for disease diagnosis and clinical prognosis. However, their high diversity and limited sharing between donors pose unique challenges when employing them as features for machine learning based diagnostics. In this study, we investigate the commonly used approach of representing each receptor as a member of a "clonotype cluster". We then construct a feature vector for each donor from clonotype cluster frequencies (CCFs). We find that CCFs are sparse features and that classifiers trained on them do not generalize well to new donors. To overcome this limitation, we introduce a novel approach where we transform cluster frequencies using an adjacency matrix built from pairwise similarities of all receptors. This transformation produces a new feature, termed paratope cluster occupancies (PCOs). Leveraging publicly available AIRR datasets encompassing infectious diseases (COVID-19, HIV), autoimmune diseases (autoimmune hepatitis, type 1 diabetes), and cancer (colorectal cancer, non-small cell lung cancer), we demonstrate that PCOs exhibit lower sparsity compared to CCFs. Furthermore, we establish that classifiers trained on PCOs exhibited improved generalizability and overall classification performance (median ROC AUC 0.893) when compared to CCFs (median ROC AUC 0.714) over the six diseases. Our findings highlight the potential of utilizing PCOs as a feature representation for AIRR analysis in diverse disease contexts.

Main

Medical diagnosis, the process of attributing a disease as the primary cause for a set of symptoms, has been an enduring challenge throughout history. As early as 16th century BCE, Egyptians documented methods resembling modern differential diagnosis algorithms[1]. 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[2]. 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[3], early detection of abnormalities even in the absence of symptoms[4], and personalized diagnoses tailored to an individual's medical history[5]. 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. In this article, we focus on a particularly rich and emerging source of data on disease status that has not traditionally been used in medical diagnosis: adaptive immune receptors (AIRs).

AIRs encompass B cell receptors (BCRs), also known as antibodies in their soluble form, and T cell receptors (TCRs), which, together, constitute a highly complex network of proteins that can distinguish self from non-self at the molecular level (Suppl Materials 1.1.). AIRs are generated combinatorically from gene fragments and selected for their recognition of non-self antigens. The AIR repertoire (AIRR) in a given host thus serves as a comprehensive record of past and ongoing antigen encounters. For this reason, AIRs harbor valuable information about health-disease status that is not encoded in the genome, making AIRR analysis complementary to genome-wide association or gene expression studies. As the primary responders to infections or other perturbations, AIRs possess the unique ability to proliferate in response to the presence of specific disease-associated antigens. Furthermore, considering the abundance of B cells and T cells in the circulating blood, AIRs can be readily accessed by minimally invasive tests. These distinctive properties have garnered significant attention for the potential utilization of AIRs as disease biomarkers[6-10]. Recent evidence also suggests that AIRs play a critical role as early responders to cancer immunotherapy, suggesting their use as companion diagnostics[11].

Moreover, beyond their role as biomarkers, disease-associated AIRs hold promising potential for repurposing as highly specific therapeutics. Antibodies, the soluble form of BCRs, have emerged as a vital and rapidly expanding class of drugs, particularly in the fields of oncology[12] and autoimmunity[13]. However, it is noteworthy that the vast majority of therapeutic antibodies developed thus far have been derived from engineered human-like antibodies in animals or other expression systems, rather than through the exploration of human AIRRs[14]. Therefore, human AIRRs represent an untapped natural resource that could be utilized for both biomarker discovery and therapeutic development[15,16]. However, due to technical limitations, the integration of AIRR data into medical diagnostics has remained an open problem.

The task of diagnosing diseases based on AIRRs can be cast as a classification problem. Here, the input comprises a set of donors and the output is the probability for a given disease. Each donor is assumed to belong to a class. In this study, we focus on binary classes "healthy" and "disease," although additional states such as disease severity or subtype have been explored[9,17,18]. Each donor is characterized by a set of AIRs that represents a snapshot of the donor's AIRR. For simplicity, we assume the presence of a single chain (BCR light or heavy, TCR alpha or beta). However, extending the problem definition to include paired (light-heavy or alpha-beta) AIRR data does not fundamentally alter its nature. A single AIR is traditionally described by the following properties: a V gene name, a J gene name, an amino acid sequence, and the location of the three complementarity-determining regions (CDRs) within this sequence (**Fig. S1**). The objective is to train a binary classifier using features derived from these properties on a subset of donors (training data) with the aim of generalizing the classifier's performance to an unseen set of donors (test data).

A significant challenge in this problem lies in the extensive diversity and limited sharing of AIRs among donors. For instance, a previous study investigated the BCRs of three healthy donors and discovered that only 1-6 percent of BCR clones (defined as identical V gene, J gene and CDR3 amino acid sequence) were shared between any two donors, and mutual sharing across all three donors was as low as 0.3%[19]. 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 AIRRs, 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. Consequently, AIRRs present a distinct challenge that necessitates careful consideration in selecting appropriate features for analysis.

To tackle the donor sharing problem, two strategies have been employed thus far; namely, the "macroscopic" and "microscopic" approaches. In the macroscopic approach, features are derived from the entire AIRR associated with a given donor. This circumvents the need for direct comparison of AIRs between donors, effectively avoiding the problem of sharing. Examples of such features include diversity and evenness[20,21]. On the other hand, the microscopic approach involves representing AIRs in a lower-dimensional coordinate system that brings about convergence of similar but nonidentical sequences. This can be achieved through clustering[22,23] or embedding using natural language processing (NLP) techniques[18,24]. Both macroscopic and microscopic approaches, as well as their combination, have shown promise in disease classification[6]. Moreover, the transformation of AIRs into clusters of antigen or epitope-specific groups constitutes an important and challenging problem in its own right that has primarily been studied in the context of TCRs[18,25-27], but has also been demonstrated in BCRs[28-30]. Because AIRR clustering may shed light on features that are responsible for antigen or epitope specificity, we exclusively concentrate on the microscopic approach in the present study.

Results

We examined three broad disease categories, selecting two representative diseases from each category: infectious disease (COVID-19, HIV), autoimmunity (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). We randomly split donors into training (70%) and test (30%) groups and evaluated the classification performance in two ways: leave-one-out cross validation (LOOCV) using the training set and assessment of the trained classifier on the test set. In the training step, there is one hyperparameter (d_{min}), that represents the minimum required donor diversity within a cluster (see *Cluster Downsizing* in Methods section), which was selected based on the LOOCV evaluation using the average of ROC AUC and PR AUC as a target function. Our primary goal was to compare the performance of classifiers that used paratope cluster occupancies (PCOs) as features with classifiers that used clonotype cluster frequencies (CCF) as features. As

shown in **Fig. 1**, the conversion from CCFs to PCOs involves two steps: the use of paratopes for clustering and the use of adjacency matrices to compute occupancies from these cluster frequencies. We also assessed two alternative features—clonotype cluster occupancies (CCO) and paratope cluster frequencies (PCF)—in which only the occupancy step or only the paratope clustering step was used, respectively. All steps in the methodology other than feature generation were identical. In the following sections we discuss each of the six disease classification assessments separately. Table 1 contains a summary of the performance metrics.

Table 1. Summary of performance metrics

		COVID-19	HIV	AIH	T1D	CRC	NSCLC	Average
ROC AUC	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	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

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. Here, we utilized BCR heavy chain data from a single cell sequencing study of primary blood mononuclear cells (PBMCs) of 44 COVID-19 patients (106,640 BCRs) and 58 healthy donors (174,139 BCRs), all of whom were unvaccinated[31]. After splitting the randomly, 71 donors were used for training and 31 were held out for testing. **Fig. 2A** shows the effect of the d_{min} hyper parameter on the sparsity of the features (see *Cluster Downsizing* in Methods section), as well as the LOOCV target function, reaching a

maximum at 0.7 (pink line). We see very little difference between the four features other than the fact that increasing d_{min} did not reduce the sparsity of CCF features or PCF features without removing all data for one or more donors. Since d_{min} represents the minimum number of donors within a given class (healthy or disease), the CCF and PCF results imply that the clonotype clusters mostly contained BCRs from only a single donor. **Fig. 2B** shows ROC curves for the test data using the PCO features, which achieved an AUC of 0.896 for the holdout (test) donors at the optimal d_{min} value of 0.7. From the qualitative agreement between the LOOCV and test ROC curves, we can say that the classifier was not overfit and generalized well to new donors. In **Fig 2C** the test ROC AUCs for all four features is shown, which indicate that only the model trained on CCF features was unable to classify COVID-19 patients and healthy controls, presumably due to the poor sharing of clonotypes. These observations are consistent with the rather shallow sequencing (2,500 BCRs per donor) that is typical of single cell sequencing studies. In **Fig. 2D** we show precision-recall AUCs, which indicate that the PCO PR AUC was the highest among all trained models (0.896).

HIV

HIV is a retrovirus that infects helper T cells, resulting in a chronic infection that, if untreated, results in acquired immunodeficiency syndrome (AIDS). Broadly neutralizing antibodies against HIV often have distinct paratope features, including long CDRH3 regions. Here, we utilized a large-scale study of BCR heavy chains from PBMCs of 94 HIV+ (5,202,133 BCRs) and 128 healthy uninfected (7,934,405 BCRs) donors[22]. After random splitting, 155 donors were used for training and 67 were held out for testing. Compared with the COVID-19 dataset, the HIV dataset is very large, not only in terms of the number of donors, but also the sequencing depth (roughly twenty times that of the COVID-19 study). This presented a problem for the computation of the adjacency matrix, which has a quadratic time complexity with respect to the number of AIRs. To circumvent this problem, we used cluster downsizing as a pre-processing step. Here, we set d_{min} = 0.05, which reduced the number of AIRs from 13,136,538 to 4,424,328. Larger values of d_{min} would have reduced the number of AIRs further but would have resulted in removal of all AIRs for one or more of the training donors. The reduced HIV dataset was used in all subsequent calculations. **Fig. 3A**, shows that the sparsity of the features is qualitatively lower for CCO and PCO than for CCF or PCF, implying that the occupancy step increases the donor diversity of

clusters, effectively connecting clusters that would otherwise be less diverse, regardless of whether clustering is performed by clonotyping or by paratopes. As with the COVID-19 assessment, the LOOCV target function reached a maximum at a d_{min} value of 0.7. **Fig. 3B** shows that the PCO-based classifier achieved a nearly perfect ROC AUC of 0.985 on test donors, and that the training LOOCV and test ROC curves overlapped well, indicating the classifier was not overfit. In **Figs. 3C-D** we scan see that all four models performed similarly in terms of both ROC AUC and PR AUC.

Autoimmune Hepatitis

Autoimmune hepatitis (AIH) is a chronic liver disease that is thought to be driven by autoreactive T cells[32]. Here, we utilized a study of TCR beta chains from PBMCs of 59 AIH patients (338,541 TCRs) and 59 healthy donors (390,507 TCRs)[17]. After random splitting, 82 donors were used for training and 36 were held out for testing. **Fig. 4A** shows that sparsity of the features at low donor diversity thresholds (d_{min}) separate into two groups (CCF/PCF and CCO/PCO), but that at larger values all but the CCF-based feature sparsity values converge to a single value, reaching a sparsity below 0.1%, and that the LOOCV target function gradually increased with d_{min}. **Fig. 4B** that the PCO classifier achieved a ROC AUC of 0.947 on test donors, and even outperformed the training LOOCV. **Figs. 4C-D** show that the ROC and PR AUCs for the PCO-based model were superior to those of the CCF, PCF or CCO models.

Type 1 Diabetes

Type 1 diabetes (T1D) is an autoimmune disease where T cells target the insulin-producing cells in the pancreas[33]. Unlike the datasets used in the other assessments, which utilized unsorted PBMCs, the T cells used here were enriched by culturing with Human islet antigen. A total of 34 T1D patients (2,161 TCRs) and 11 healthy controls (408 TCRs) from T cells sequenced at the single cell level were used[34]. After splitting randomly, 31 donors were used for training and 14 were used for testing. Similar to what was observed in the COVID-19 case, the low sequencing depth here resulted in highly sparse clonotype clusters and paratope clusters, populated mostly by single donors. However, the occupancy step (CCO, PCO) produced features with greater donor diversity, for which sparsity could be reduced (**Fig. 5A**). The performance on test data even for the PCO-based model only reached a ROC AUC of 0.725, lower than for diseases discussed above

(**Fig. 5B**). However, this was significantly better than the CCF or PCF classifiers, which resulted in ROC AUCs no better than random and PR AUCs that were intermediate between the PCO model and a random predictor (**Figs. 5C-D**). Taken together, the T1D data represents a lower limit on training data, but nevertheless suggests that PCOs provide a more robust feature than CCFs, even at this lower limit.

Colorectal Cancer

Colorectal cancer (CRC) is a cancer of the large intestine that is typically diagnosed by tissue biopsy during a colonoscopy or sigmoidoscopy[35]. T cells are known to infiltrate CRC tumors; however, the extent to which such T cells circulate in the blood is poorly understood. If disease-specific T cells can be detected in the blood, TCR-based diagnosis represents a minimally invasive approach. Here, we utilized TCR beta chain data from a study of 20 CRC patients (137,622 TCRs) and 88 healthy donors (1,100,012 TCRs)[36]. After splitting the studies as described above, there were 75 training donors and 33 held out for testing. **Fig. 6A** shows the sparsity of the resulting features, indicating that PCO and CCO features were less sparse that the other features, and that the differences decreased with increasing values of d_{min}, similar to trends observed previously. The LOOCV target function improved with increasing values of d_{min}, and the maximum value of 0.9 was selected. **Fig. 6B** shows that the ROC curves for LOOCV and test sets agree qualitatively, and that the test ROC AUC achieved a value of 0.82. When we examined the test ROC AUCs for all four classifiers, the PCO features performed better than the other (**Fig. 6C**); however, the PR AUCs indicate that the PCF classifier performed better than the other three (**Fig. 6D**). The discrepancy between ROC and PR AUCs probably derives from the large class imbalance as well as the rather small size of the CRC data.

Non-small Cell Lung Cancer test set

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[37]. Diagnosis traditionally utilizes a wide range of evidence, including symptoms such as persistent cough, imaging data, and tissue biopsy; however, blood-based biomarkers are recently becoming more common[38]. As with CRC, it is known that disease-specific T cells often infiltrate the 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[17,39,40], two for NSCLC training [41,42], one for healthy testing[36] and one for NSCLC testing[43]. This resulted in a total of 204 NSCLC donors (6,734,867 TCRs) and 294 healthy donors (4,120,597 TCRs). After splitting the studies as described above, there were 344 training donors and 152 held out for testing. **Fig. 7A** 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 large values of d_{min}., while the LOOCV target function remained perfect for all values. **Fig. 7B** shows that the PCO-based classifier achieved a ROC AUC of 0.985 for the test donors. This nearly perfect classifier performance would suggest that tumor-specific T cells indeed circulate in the blood. **Figs.7C-D** suggest that only the classifiers trained on occupancy-based features were able to generalize to new donors. This result provides a counterpoint to the HIV assessment, where large training data resulted in nearly perfect classifier performance for all classifiers. Since the sequencing depth was overall similar to that of HIV, we speculate that the occupancies are more robust than cluster frequencies. We explore this question further below.

Classifiers trained on PCO features are robust against batch effects

Use of a large dataset composed of multiple studies provides 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. The remaining studies used for training. This resulted in twelve such splits. These results (**Fig. 8.**) indicated that, in terms of ROC and PR AUC, both PCF and PCO features performed well, but the CCF and CCO features performed inconsistently, in agreement with **Figs. 7C-D**.

It was possible to combine the NSCLC studies in such a way that the trained model failed to generalize to the test data. Out of a total of twelve combinations, there were two for which this problem occurred. Interestingly, these two cases were ones where ones where the size of the training data was much smaller than that of the test data. In spite of these two poor performances,

the performance of the PCO-based classifier achieved a median ROC AUC value of 0.926 and a median PR AUC value of 0.896.

Identifying disease epitope specificity groups from feature importance

In the context disease classification, clusters represent the coordinate system on which we describe the donors. We hypothesized that AIRs that are in particularly important clusters as possessing antigen specificity that is important for the disease in question. To test this hypothesis, we first ranked NSCLC clusters by their XGBoost feature importance. We next identified clusters whose members were strongly biased toward one class using the t-test on the number of members from each class (healthy, disease). In this way, we identified clusters that were biased toward the healthy or the NSCLC class (**Fig. 9**). Next, we focused on clusters that were biased toward the NSCLC class. We searched two TCR databases (McPAS and VDJdb) whose TCRs have been annotated by the targeted peptide-MHC complex, parent antigen and disease. We identified strong hits to NSCLC antigen targeting TCRs. **Fig. 10** illustrates some of these hits.

Discussion

Recent advances in AIR sequencing and machine learning techniques have significantly enhanced our ability to comprehensively observe the immune responses from a routine blood sample. Numerous research studies have demonstrated that AIRs exhibit predictable changes in response to infection[18,44], immunization[45], or immunotherapy[46]. However, in order for these findings to be translated into clinical applications, it is crucial to establish robust methods that can be generalized across different donors, diseases and studies. The standard practice of describing AIRs in terms of clonotypes, using V and J gene names along with CDR3 sequence motifs, has gained wide acceptance in the literature[6,22,23,47,48]. Nonetheless, clonotypes pose several challenges as machine learning features. Firstly, they comprise a combination of categorical (V gene, J gene) and quasi-continuous (CDR3 sequence) data types, resulting in complexities in their analysis. In addition, there is no single agreed upon definition of a clonotype. One approach to constructing features from clonotypes, implemented in the DeepTCR method, utilized autoencoders to transform gene names and CDR3 sequences from categorical variables into a continuous representation[18].

Alternatively, TCR-BERT leveraged self-supervised transfer learning to generate natural TCR CDR3 sequences[49]. Establishing effective and standardized methodologies for representing TCRs will be pivotal for advancing their use in machine learning based diagnostics.

When it comes to BCRs, the underlying rules governing natural sequence generation may be more complex than for TCRs. This complexity arises from the presence of extensive somatic hypermutations driven by antigen binding avidity, which drive amino acid sequences outside the CDR3 region away from their germline gene origins. Consequently, we contend that utilizing paratope pseudo sequences, which encompass all three CDR segments, provides a more natural and comprehensive approach for describing BCRs than do clonotypes. Moreoever, paratope sequences reflect the critical regions involved in antigen recognition and binding and provide a single data type from which a continuous similarity metric can be computed using standard sequence alignment techniques.

What is particularly remarkable is that the utilization of paratope descriptions for AIRs not only proved beneficial for BCR data but also for TCR data. This observation might be attributed to the sparsity cluster-based features which we addressed by transforming cluster frequencies into cluster occupancies using an adjacency matrix. The improved generalizability of occupancies implies the existence of a cohesive network of similarity relationships between AIRs. The cluster occupancy formalism effectively captures these relationships more robustly than cluster frequencies, even when clusters are based on paratopes. These findings may help inform the development of compact and information-rich representations of AIRRs in various contexts. For example, in a recent study of healthy donor responses to Diphtheria Tetanus Pertussis vaccination[29], we observed consistently higher sharing of BCRs across donors when the calculation was based on cluster occupancy compared with clonotype frequency. The occupancy formalism is closely related to network-based AIRR analysis, which has recently been applied to the study of antibody and TCR repertoires[47,48].

Methods

AIR datasets

AIRR data were collected from published studies covering a wide range of diseases, and included both BCRs and TCRs (**Table 2**). 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 in. 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.

Table 2. AIRR Datasets

Disease	Class	Donors	AIRs	Chain	Bioproject	References
Coronavirus disease 2019 (COVID-19)	Disease	44	106,640	BCRH	JGAS000593	[31]
	Healthy	58	174,139			
	Total	102	280,779			
Human immunodeficiency virus (HIV)	Disease	94	5,202,133	BCRH	PRJNA486667	[22]
	Healthy	128	7,934,405			
	Total	222	13,136,538			
Autoimmune hepatitis (AIH)	Disease	59	338,541	TCRB	PRJEB37143	[17]
	Healthy	59	390,507			
	Total	118	729,048			
Type 1 Diabetes (T1D)	Disease	34	2,161	TCRB	PRJNA758093	[34]
	Healthy	11	408			
	Total	45	2,569			
Non-small cell lung cancer (NSCLC)	Disease	204	6,734,867	TCRB	https://doi.org/10.21417/AR2019NC PRJNA544699 PRJNA477518 PRJEB37143 PRJNA767497 https://doi.org/10.21417/ADPT2020V4CD http://doi.org/10.21417/PAS2021STM	[17,36,39-43]
	Healthy	294	4,120,597			
	Total	498	10,855,464			
Colorectal Cancer (CRC)	Disease	20	137,622	TCRB	https://doi.org/10.21417/ADPT2020V4CD	[36]
	Healthy	88	1,100,012			
	Total	108	1,237,634			

AIR representation

Clonotyping is by far the most common way of describing AIRs[6,22,23,47,48]. 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[28] and TCRs[25]. Here, a single pseudo sequence was constructed by concatenating the three CDR segments of the amino acid sequence.

Clustering

We have 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[22] to identify BCRs enriched in HIV-positive donors. Alternatively, another report[23] used a Hamming distance cutoff of 0.15 in the CDR3 sequences to stratify celiac disease patients. In yet another study[48], 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[50].

Paratope-based clustering: Paratope pseudo sequence clustering was carried out using the MMseqs2 (version 13.45111) cluster method[51] with a sequence identity threshold of 80% and a coverage threshold of 90%. We observed that 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 sequence alignment.

AIR adjacency matrices

AIR adjacency matrices were computed from paratope pseudo sequences using the MMseqs2 search method[51], 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 in section 2.3. 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, so 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. S1C**). Furthermore, each clone 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. S2A**.

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 have long tails, resulting in high-dimensional feature spaces. 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, 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. S2B**.

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 population 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 $max(r_k^{y_1}, r_k^{y_2}) < d_{min}$ 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. 1**. The results of the four workflows are discussed in the Results section.

Figure Captions

Figure 1. **Flowchart for classification algorithm.** A, 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. B, 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.

Figure 3. **Assessment of HIV diagnosis based on BCRs.** Formatting follows that of Figure 2.

Figure 4. **Assessment of AIH diagnosis based on TCRs**. Formatting follows that of Figure 2.

Figure 5. **Assessment of T1D diagnosis based on sorted TCRs.** Formatting follows that of Figure 2.

Figure 6. **Assessment of CRC diagnosis based on TCRs.** Formatting follows that of Figure 2.

Figure 7. **Assessment of NSCLC diagnosis based on TCRs.** Formatting follows that of Figure 2.

Figure 8. **Summary of 12 assessments of NSCLC using different studies for training and testing.** A, ROC AUC values on test sets. B, PR AUC values on test sets.

Figure 9. **Assessment of cluster features in NSCLC patient classification**. **A**, 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. **B**, Feature Importance. The Y-axis shows XGBoost feature importance in descending order. **C**, 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 (***$p \leq 0.001$).

Figure 10. **Functional annotation of important NSCLC clusters**. **A**, Potential antigen distribution of NSCLC patient derived sequences annotated by McPAS and VDJdb cancer databases. Each count represents one database hit. After ranking the total hits by annotated antigen in descending order, the top 50 antigens are shown. **B**, Sequences that exactly match with records in McPAS and VDJdb cancer datasets. **C**, Sequences from important features used in XGBoost training also match with records in McPAS and VDJdb cancer databases. Sequences within a cluster with identical CDR3 length are depicted as sequence logos indicating conserved residues.

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, the distribution of BCRs and TCRs in each donor follow a longtail distribution, with the vast number of clones having very low frequencies. C, CDRs 1-3 are arranged to form a continuous molecular surface called a "paratope" (in blue). D, Contacts with antigen (light blue) are primarily restricted to the CDRs compared with the background assessable surface (RSA).

Figure S2. **Cluster frequencies and cluster occupancies.** A, Cluster frequencies are illustrated by a simple example of two donors (A and B) whose AIRRs 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.

References

1 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

2 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

3 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

4 Liu, S. *et al.* in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI).* 1015-1018.

5 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

6 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

7 Gordin, M. *et al.* Breast cancer is marked by specific, Public T-cell receptor CDR3 regions shared by mice and humans. *PLoS Comput Biol* **17**, e1008486 (2021). https://doi.org:10.1371/journal.pcbi.1008486

8 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

9 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

10 Zhang, H. *et al.* Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers. *Clin Cancer Res* **26**, 1359-1371 (2020). https://doi.org:10.1158/1078-0432.CCR-19-3249

11 Luoma, A. M. *et al.* Tissue-resident memory and circulating T cells are early responders to pre-surgical cancer immunotherapy. *Cell* **185**, 2918-2935 e2929 (2022). https://doi.org:10.1016/j.cell.2022.06.018

12 Zhuravlev, V. S., Vashurina, E. A., Orlova, L. D., Kolpikova, L. A. & Mokeeva, R. A. [Serum immunoglobulins in patients with malignant lymphomas]. *Probl Gematol Pereliv Krovi* **27**, 22-24 (1982).
13 Chan, A. C. & Carter, P. J. Therapeutic antibodies for autoimmunity and inflammation. *Nat Rev Immunol* **10**, 301-316 (2010). https://doi.org:10.1038/nri2761
14 Nelson, A. L., Dhimolea, E. & Reichert, J. M. Development trends for human monoclonal antibody therapeutics. *Nat Rev Drug Discov* **9**, 767-774 (2010). https://doi.org:10.1038/nrd3229
15 Boyd, S. D. & Crowe, J. E., Jr. Deep sequencing and human antibody repertoire analysis. *Curr Opin Immunol* **40**, 103-109 (2016). https://doi.org:10.1016/j.coi.2016.03.008
16 Spindler, M. J. *et al.* Massively parallel interrogation and mining of natively paired human TCRalphabeta repertoires. *Nat Biotechnol* **38**, 609-619 (2020). https://doi.org:10.1038/s41587-020-0438-y
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 Sidhom, J. W. & Baras, A. S. Deep learning identifies antigenic determinants of severe SARS-CoV-2 infection within T-cell repertoires. *Sci Rep* **11**, 14275 (2021). https://doi.org:10.1038/s41598-021-93608-8
19 Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398-402 (2019). https://doi.org:10.1038/s41586-019-0934-8
20 Bashford-Rogers, R. J. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* **23**, 1874-1884 (2013). https://doi.org:10.1101/gr.154815.113
21 Venturi, V., Kedzierska, K., Turner, S. J., Doherty, P. C. & Davenport, M. P. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* **321**, 182-195 (2007). https://doi.org:10.1016/j.jim.2007.01.019
22 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
23 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
24 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
25 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
26 Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94-98 (2017). https://doi.org:10.1038/nature22976
27 Yokota, R., Kaminaga, Y. & Kobayashi, T. J. Quantification of Inter-Sample Differences in T-Cell Receptor Repertoires Using Sequence-Based Information. *Front Immunol* **8**, 1500 (2017). https://doi.org:10.3389/fimmu.2017.01500

28 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

29 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

30 Xu, Z. *et al.* Functional clustering of B cell receptors using sequence and structural features. *Molecular Systems Design & Engineering* **4**, 769-778 (2019). https://doi.org:10.1039/C9ME00021F

31 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

32 Mieli-Vergani, G. *et al.* Autoimmune hepatitis. *Nature Reviews Disease Primers* **4**, 18017 (2018). https://doi.org:10.1038/nrdp.2018.17

33 DiMeglio, L. A., Evans-Molina, C. & Oram, R. A. Type 1 diabetes. *The Lancet* **391**, 2449-2462 (2018). https://doi.org:https://doi.org/10.1016/S0140-6736(18)31320-5

34 Linsley, P. S. *et al.* Autoreactive T cell receptors with shared germline-like alpha chains in type 1 diabetes. *JCI Insight* **6** (2021). https://doi.org:10.1172/jci.insight.151349

35 Biller, L. H. & Schrag, D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. *JAMA* **325**, 669-685 (2021). https://doi.org:10.1001/jama.2021.0106

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

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

38 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

39 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

40 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

41 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

42 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

43 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

44 Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* **49**, 659-665 (2017). https://doi.org:10.1038/ng.3822

45 Kotagiri, P. *et al.* B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. *Cell Rep* **38**, 110393 (2022). https://doi.org:10.1016/j.celrep.2022.110393

46 Lundberg, A., Li, B. & Li, R. B cell-related gene signature and cancer immunotherapy response. *Br J Cancer* **126**, 899-906 (2022). https://doi.org:10.1038/s41416-021-01674-6

47 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

48 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

49 Wu, K. *et al.* TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. *bioRxiv*, 2021.2011.2018.469186 (2021). https://doi.org:10.1101/2021.11.18.469186

50 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

51 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

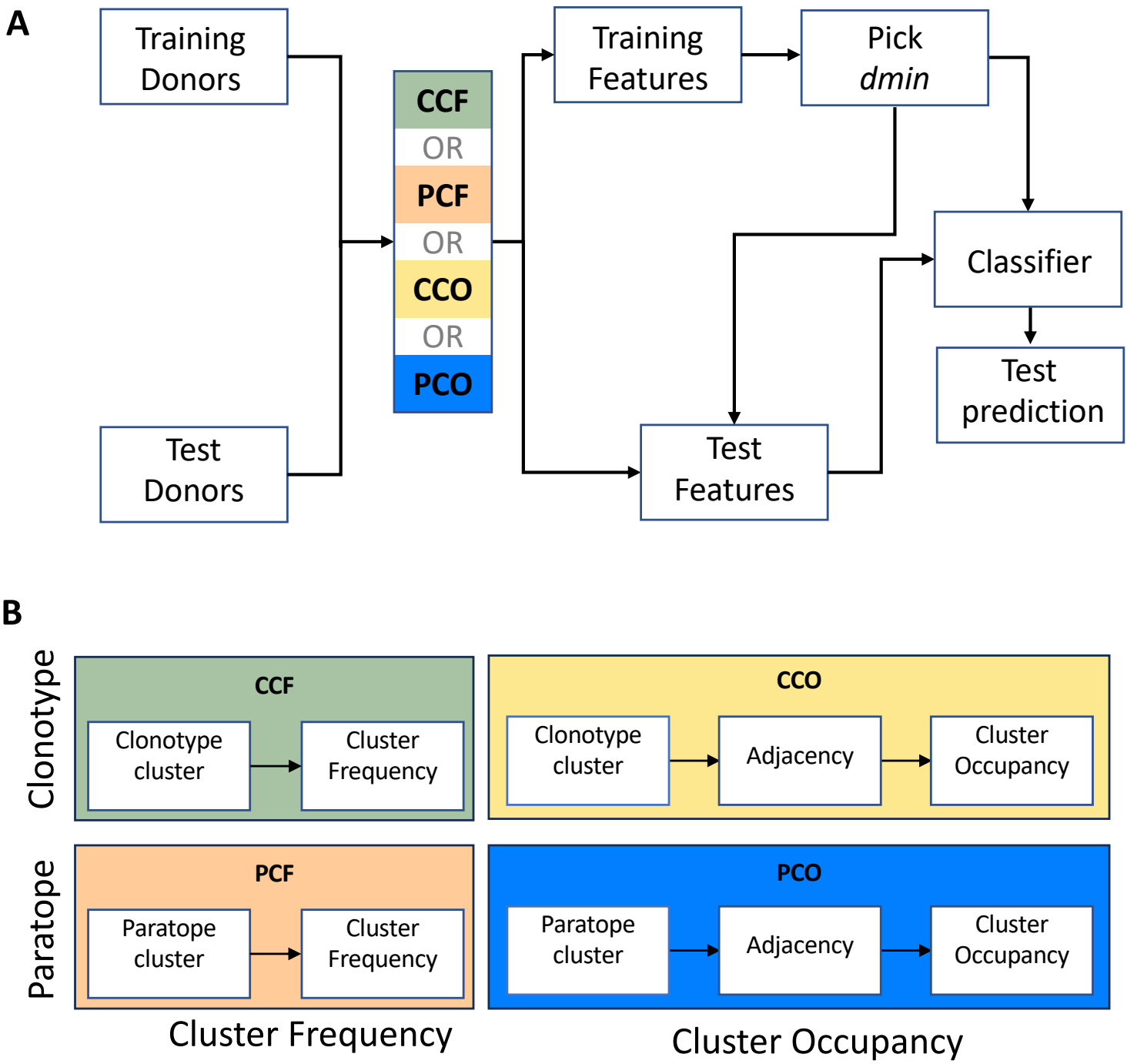

A
Training Donors
Test Donors
CCF
OR
PCF
OR
CCO
OR
PCO
Training Features
Pick dmin
Test Features
Classifier
Test prediction
B
Clonotype
Paratope
CCF
Clonotype cluster
Cluster Frequency
CCO
Clonotype cluster
Adjacency
Cluster Occupancy
PCF
Paratope cluster
Cluster Frequency
PCO
Paratope cluster
Adjacency
Cluster Occupancy
Cluster Frequency
Cluster Occupancy

Figure 1

COVID-19

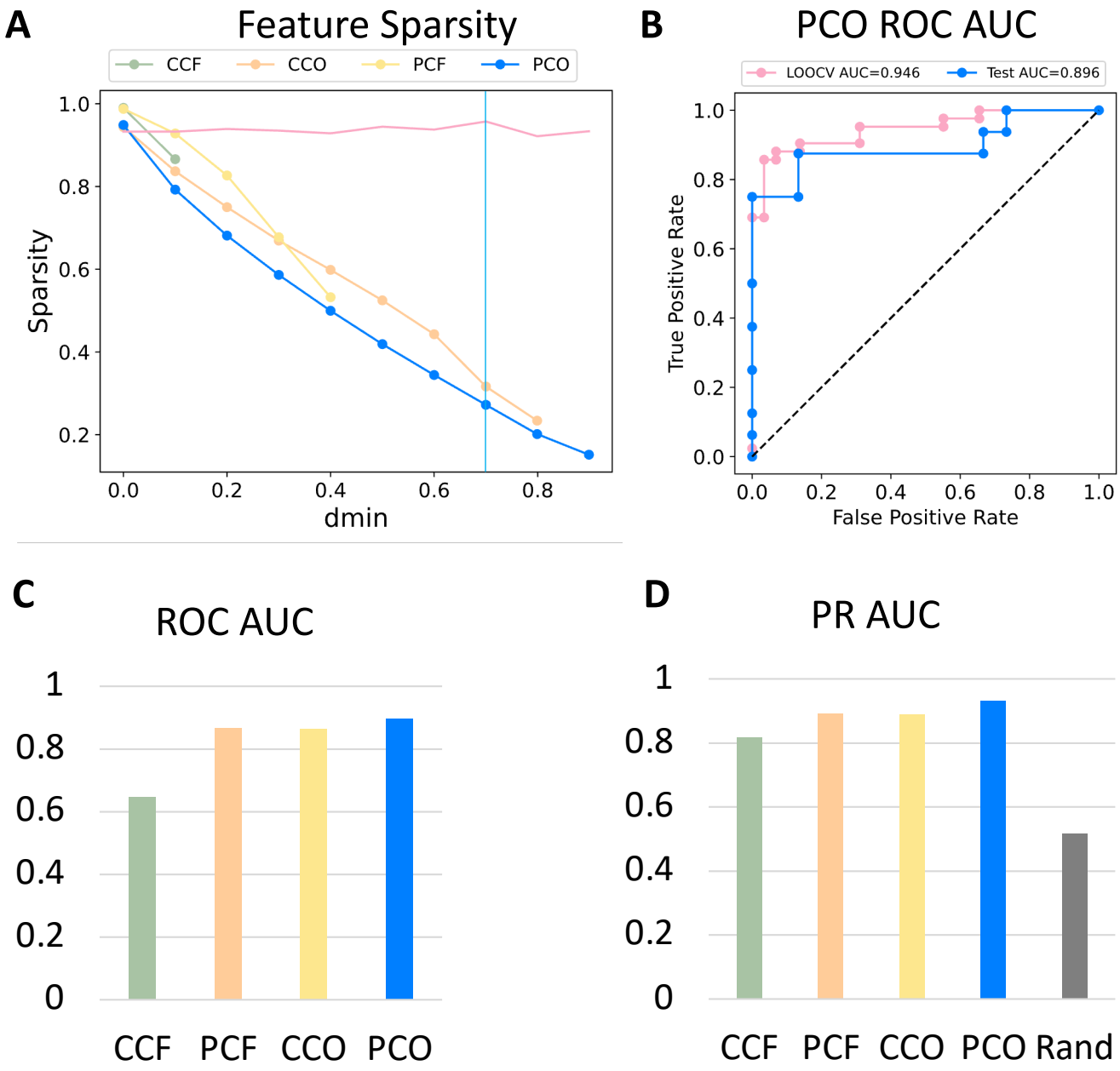


Figure 2

HIV

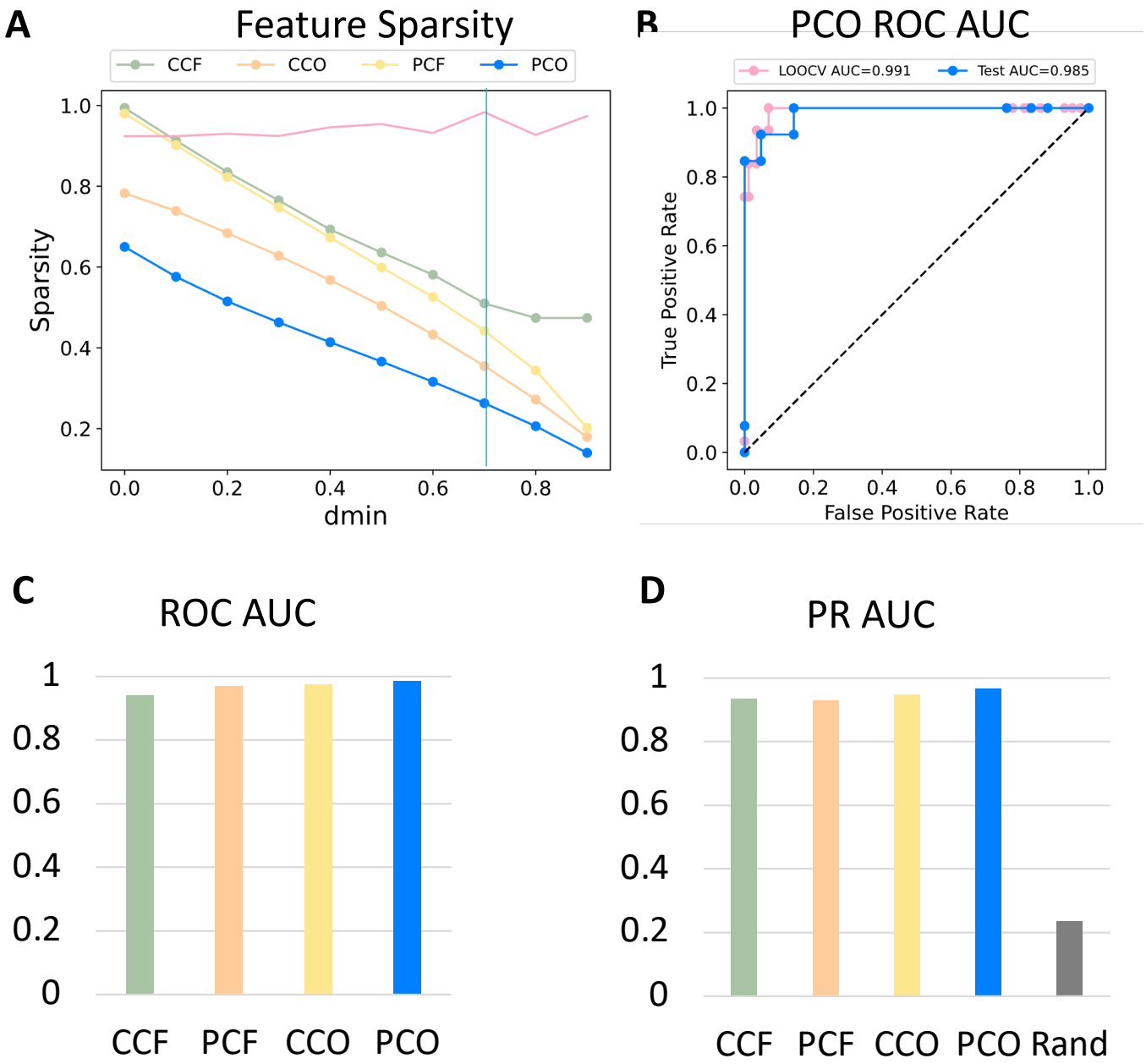

A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC AUC
LOOCV AUC=0.991
Test AUC=0.985
True Positive Rate
False Positive Rate
C
ROC AUC
CCF
PCF
CCO
PCO
D
PR AUC
CCF
PCF
CCO
PCO
Rand

Figure 3

AIH

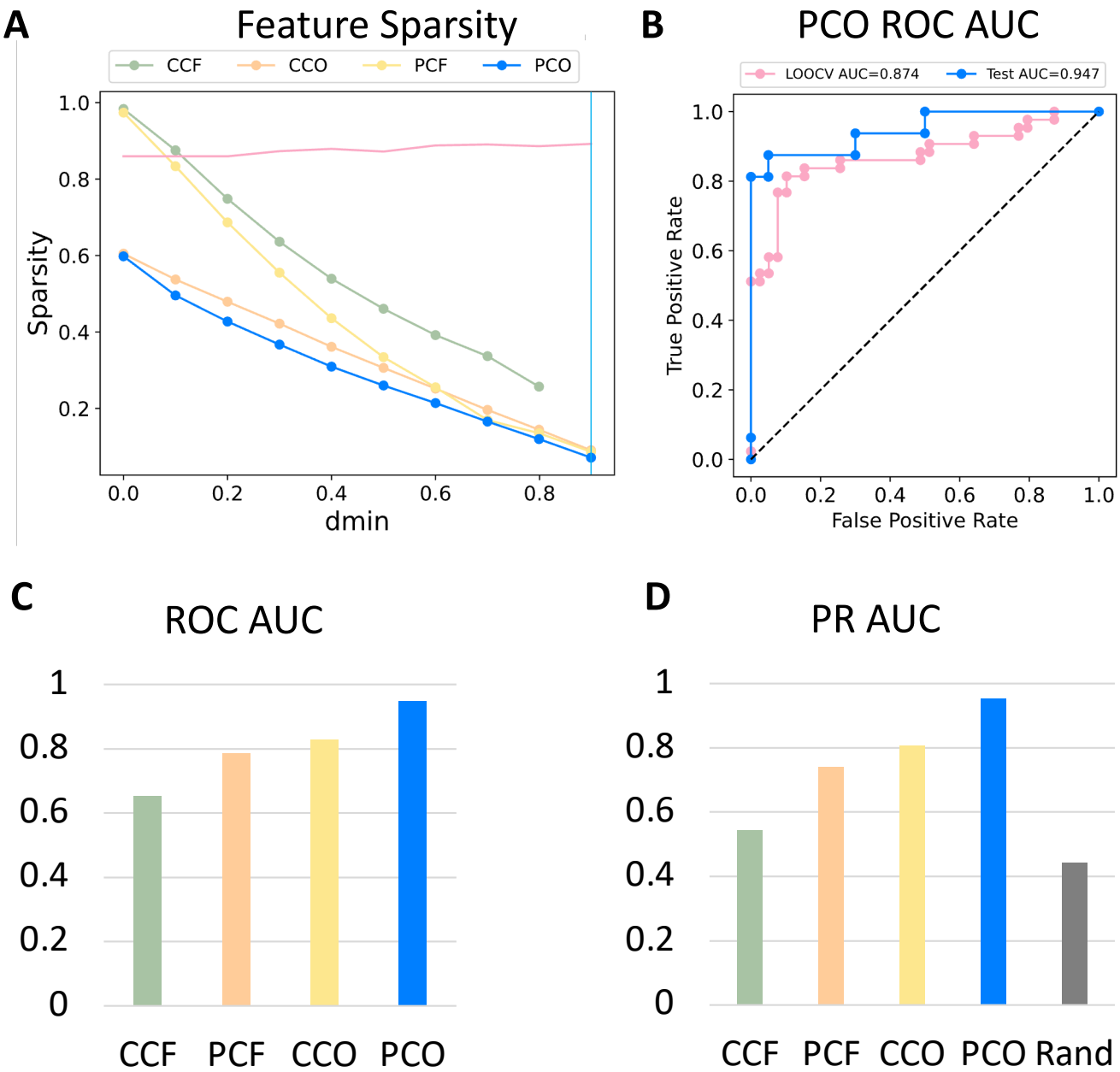

A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC AUC
LOOCV AUC=0.874
Test AUC=0.947
True Positive Rate
False Positive Rate
C
ROC AUC
CCF
PCF
CCO
PCO
D
PR AUC
CCF
PCF
CCO
PCO
Rand

Figure 4

T1D

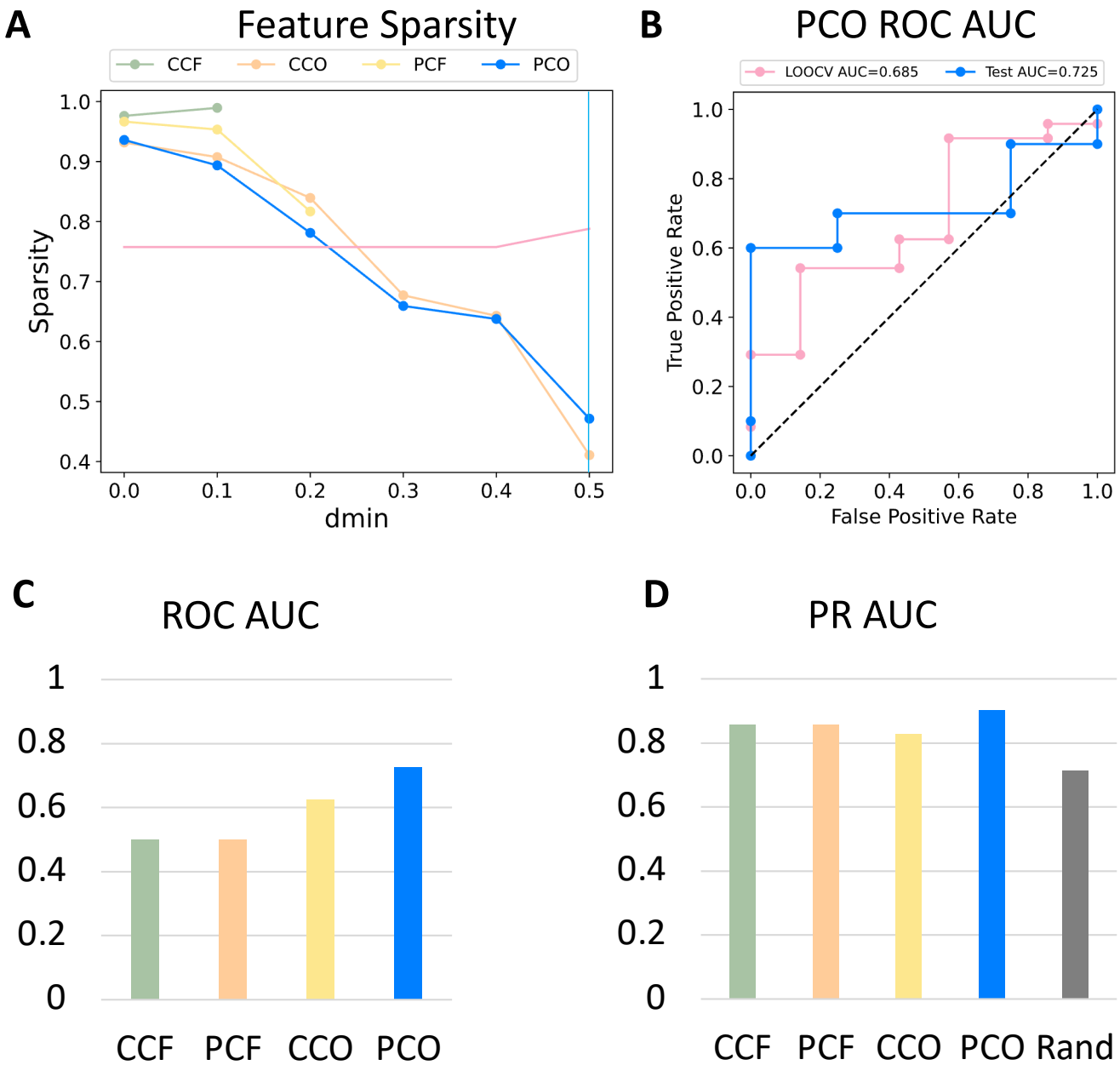

A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC AUC
LOOCV AUC=0.685
Test AUC=0.725
True Positive Rate
False Positive Rate
C
ROC AUC
CCF
PCF
CCO
PCO
D
PR AUC
CCF
PCF
CCO
PCO
Rand

Figure 5

CRC

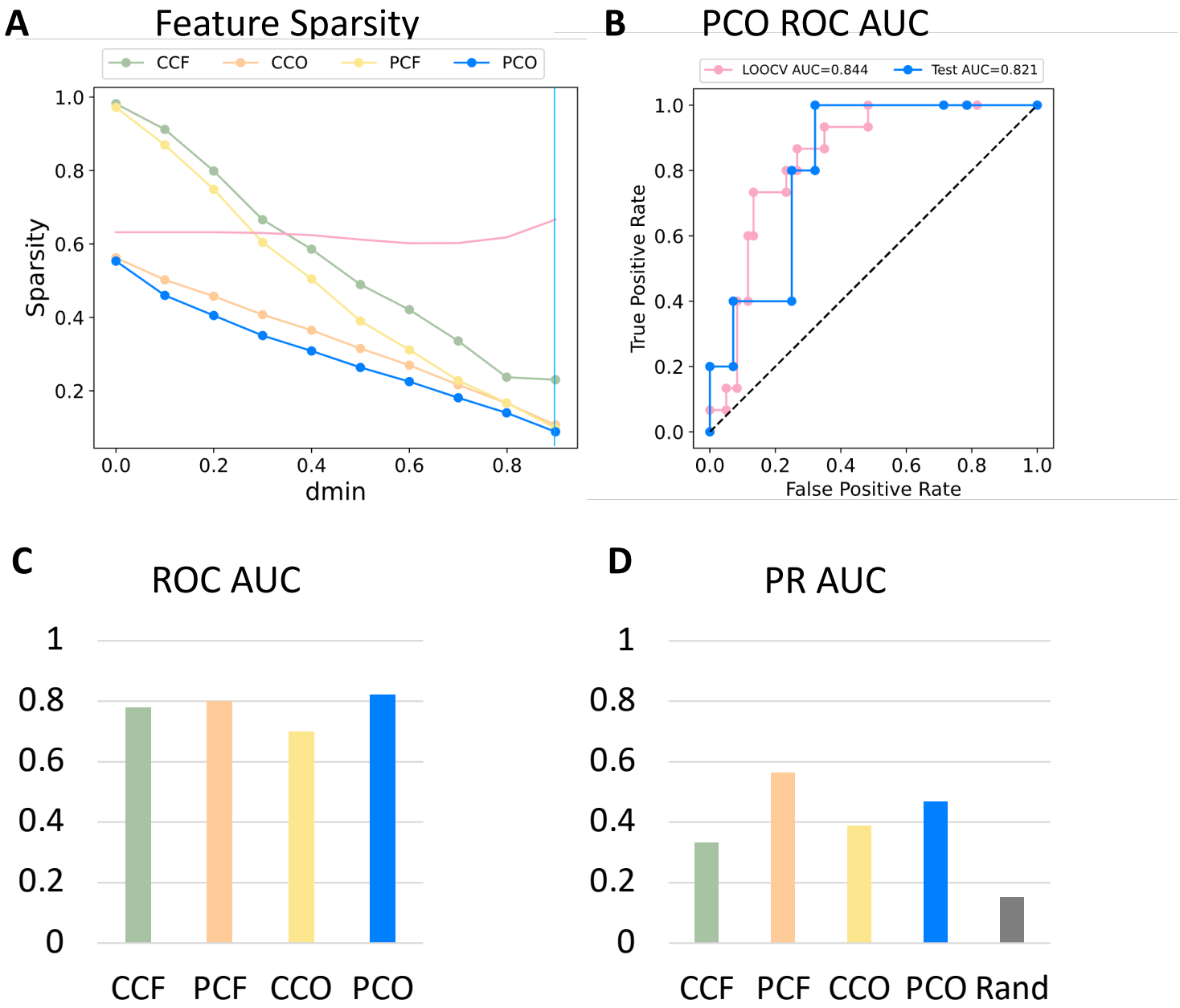

A
Feature Sparsity
CCF
CCO
PCF
PCO
Sparsity
dmin
B
PCO ROC AUC
LOOCV AUC=0.844
Test AUC=0.821
True Positive Rate
False Positive Rate
C
ROC AUC
CCF
PCF
CCO
PCO
D
PR AUC
CCF
PCF
CCO
PCO
Rand

Figure 6

NSCLC

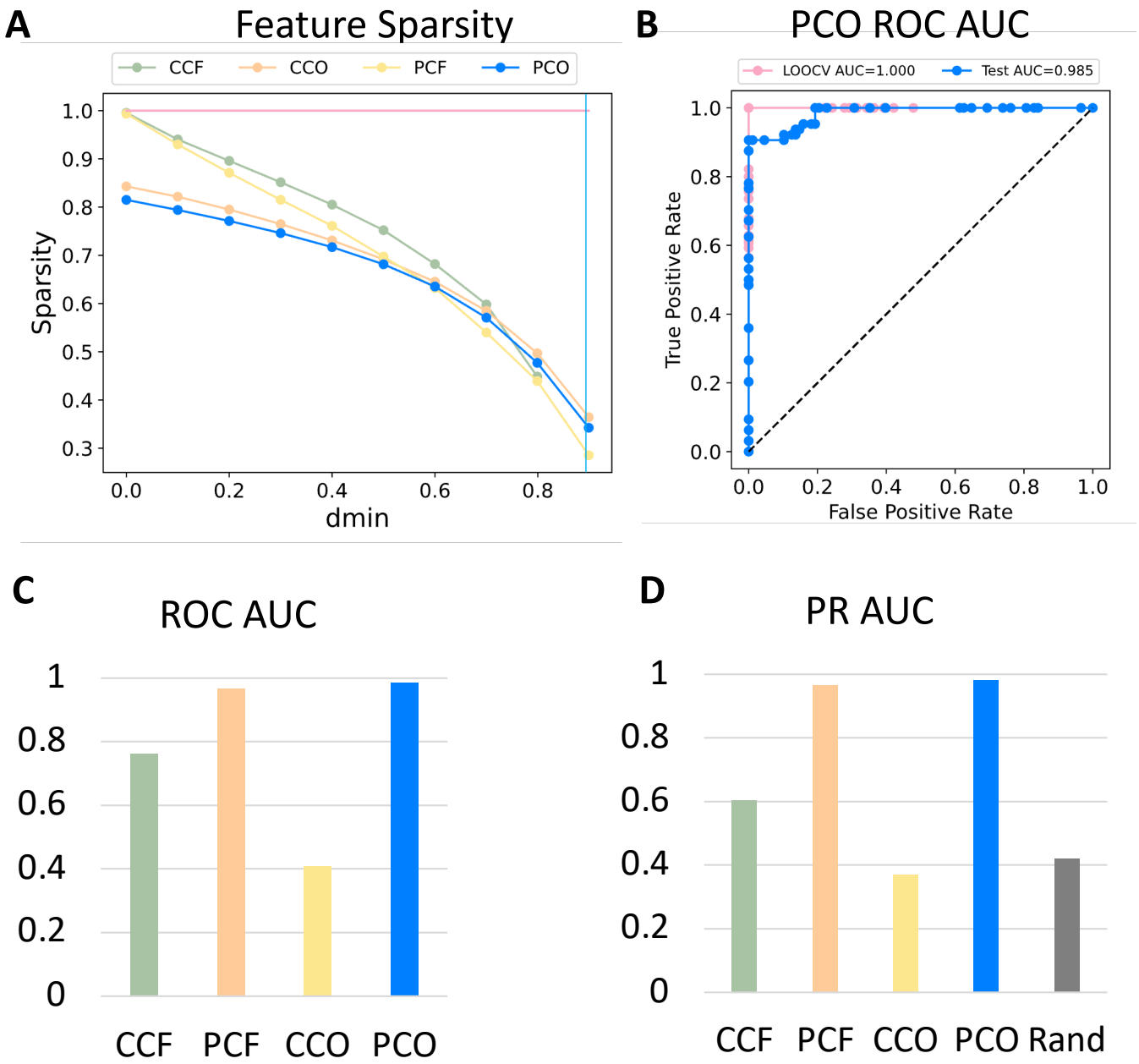


Figure 7

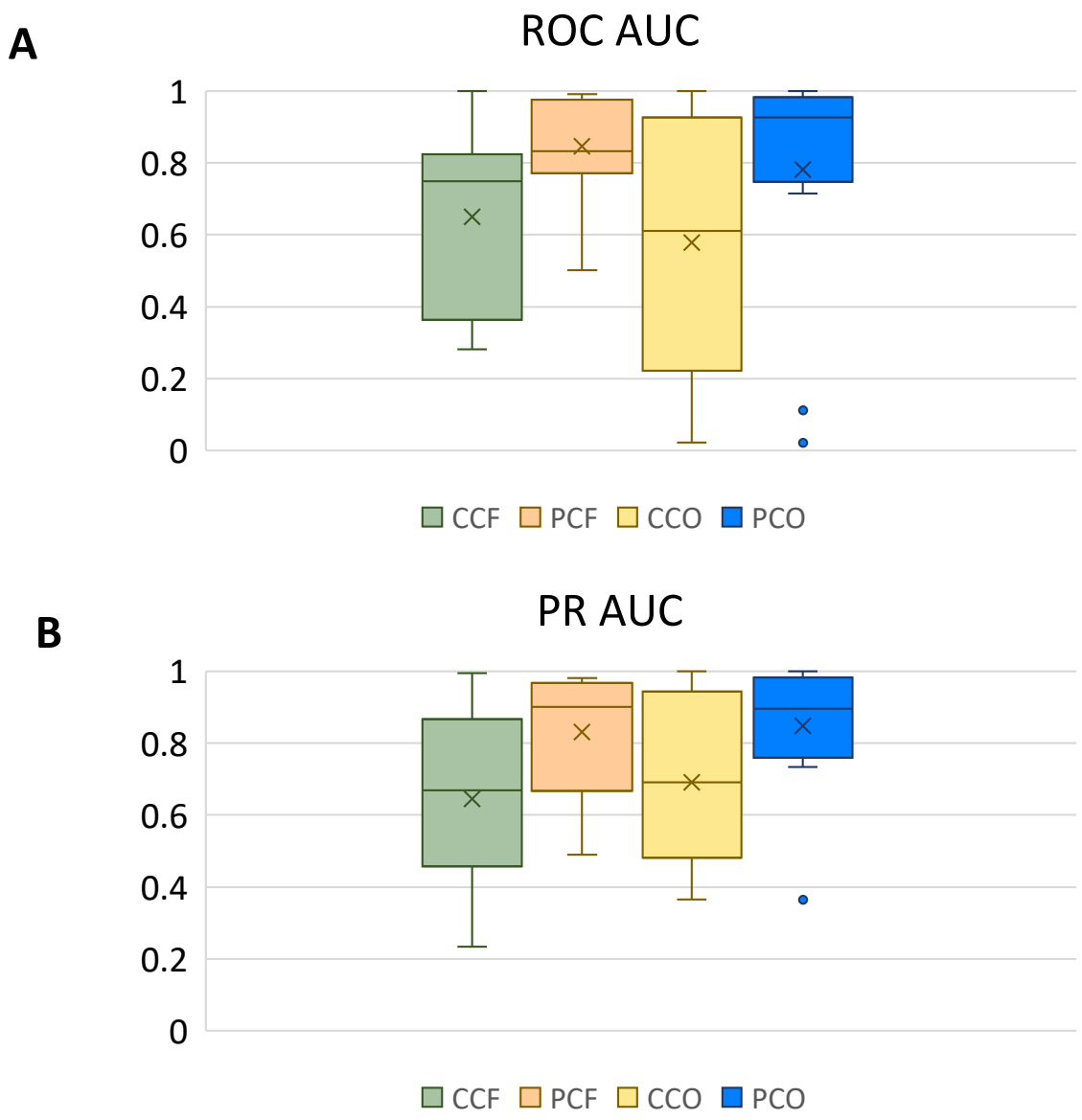

A
ROC AUC
1
0.8
0.6
0.4
0.2
0
CCF
PCF
CCO
PCO
B
PR AUC
1
0.8
0.6
0.4
0.2
0
CCF
PCF
CCO
PCO

Figure 8

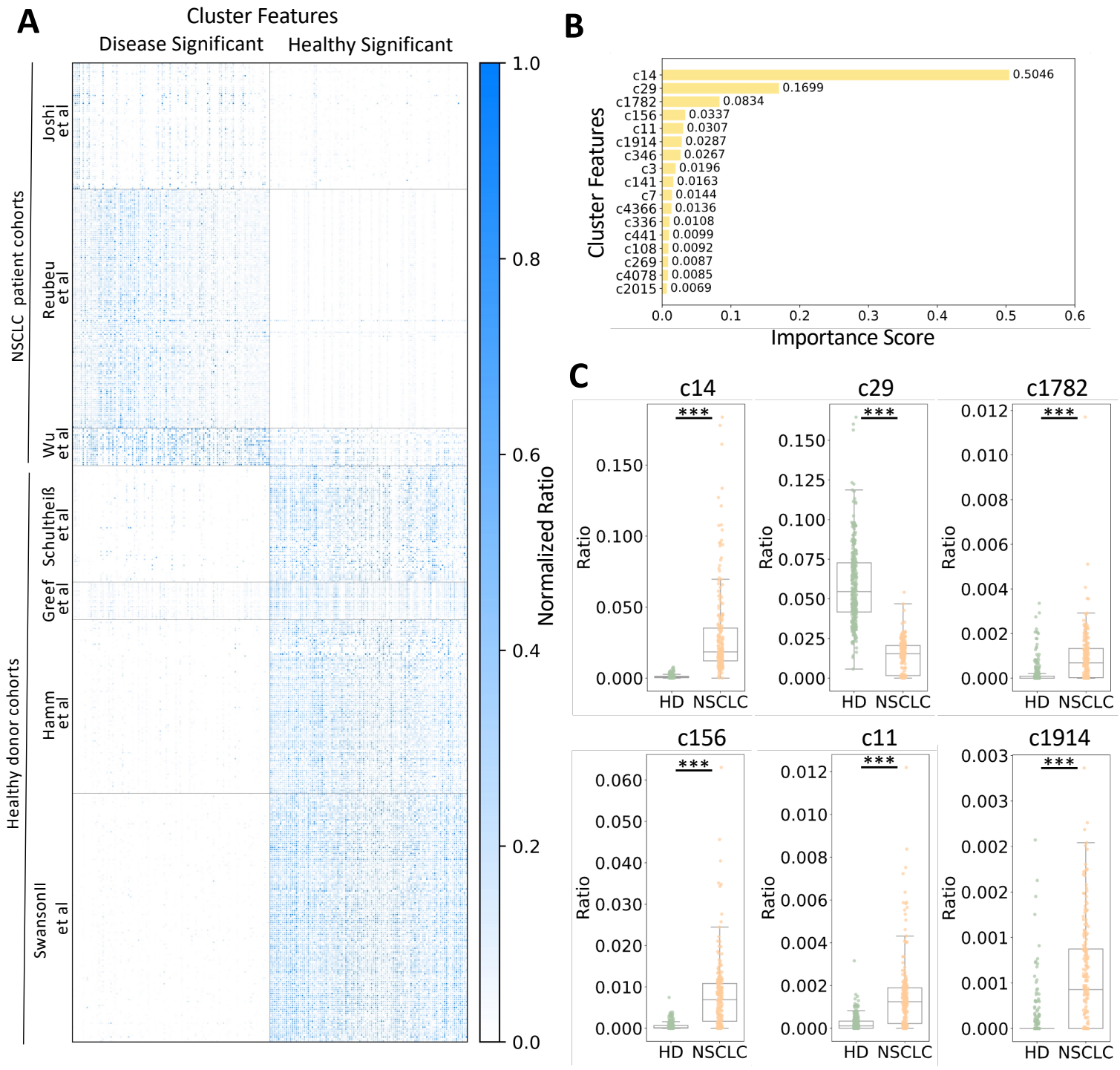
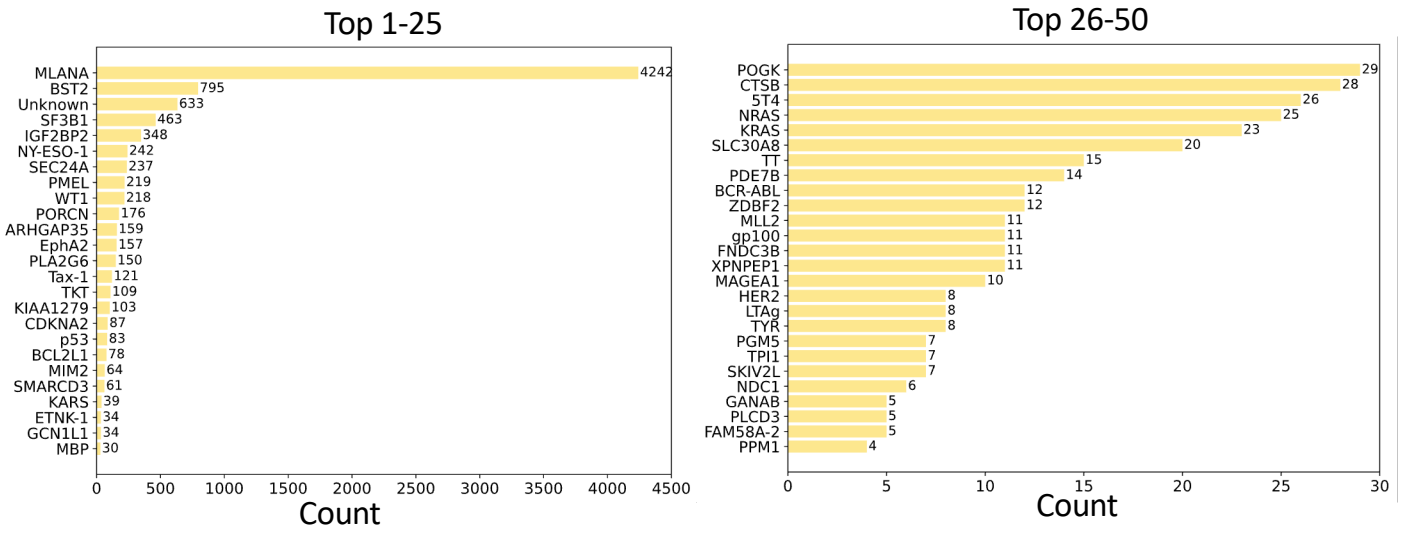

A
Cluster Features
Disease Significant
Healthy Significant
NSCLC patient cohorts
Healthy donor cohorts
Joshi et al
Reubeu et al
Wu et al
Schultheiß et al
Greef et al
Hamm et al
Swansonll et al
Normalized Ratio
1.0
0.8
0.6
0.4
0.2
0.0
B
Cluster Features
Importance Score
c14 0.5046
c29 0.1699
c1782 0.0834
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
C
c14
c29
c1782
c156
c11
c1914
Ratio
HD
NSCLC
***

Figure 9

A



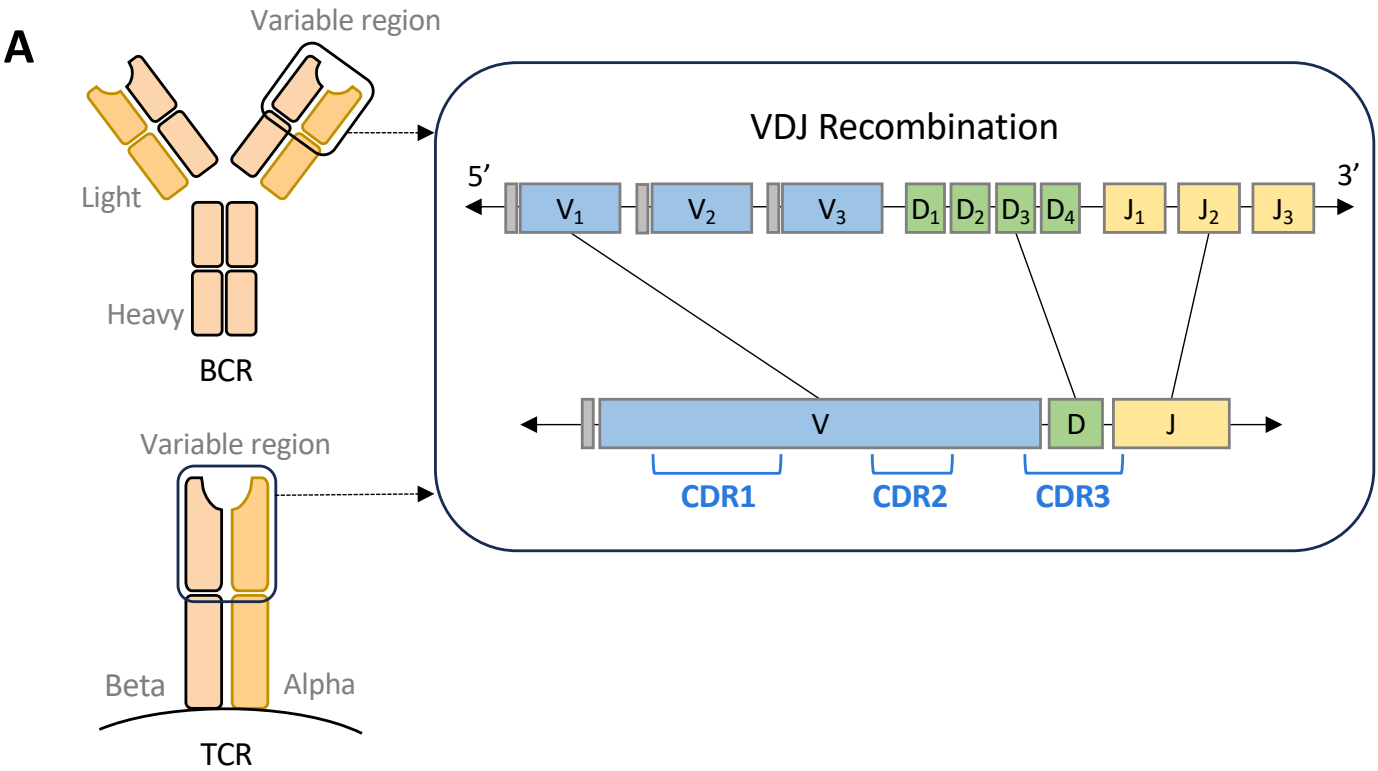
B

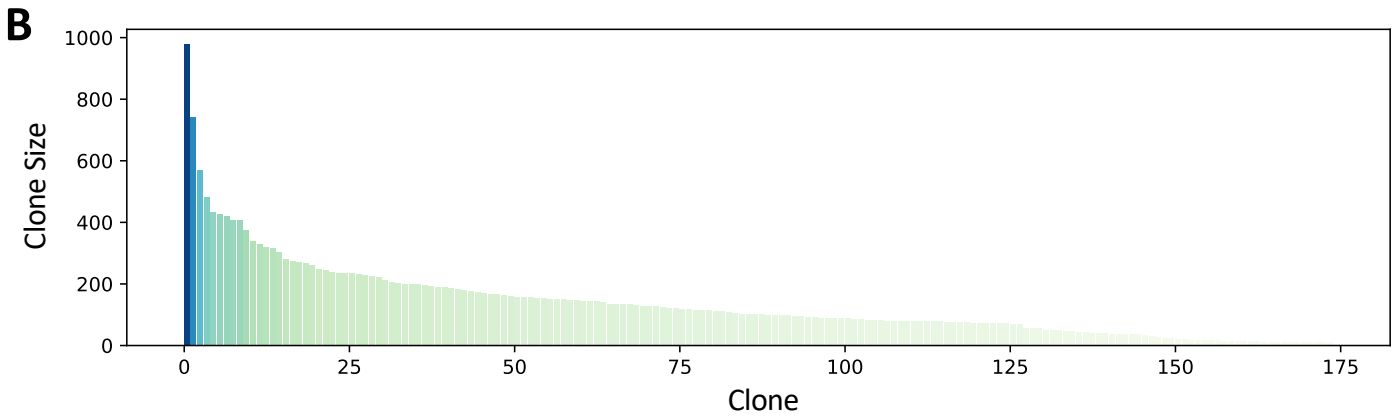
Cluster		V/ J gene	CDR3 (length)	Queries /Donors	Template Antigen	Cluster CDR3 Logo
c134	Queries Template	TRBV6-1/TRBJ1-1 TRBV6-1/TRBJ1-1	ASSETGGTEAF (11) ASSETGGTEAF (11)	6/6	MLANA	
c1801	Queries Template	TRBV9/TRBJ2-3 TRBV9/TRBJ2-3	ASSVGGGVTDTQY (13) ASSVGGGVTDTQY (13)	3/3	ARHGAP35	
c370	Queries Template	TRBV2/TRBJ2-7 TRBV2/TRBJ2-7	ASSVIYEQY (9) ASSVIYEQY (9)	2/2	NY-ESO-1	
c370	Query Template	TRBV2/TRBJ2-7 TRBV2/TRBJ2-7	ASSIIYEQY (9) ASSIIYEQY (9)	1/1		
c6996	Queries Template	TRBV9/TRBJ2-7 TRBV9/TRBJ2-7	ASSPGGGAYEQY (12) ASSPGGGAYEQY (12)	2/2	NY-ESO-1	
c6401	Query Template	TRBV7-8/TRBJ2-5 TRBV7-8/TRBJ2-5	ASSLAGGGETQY (12) ASSLAGGGETQY (12)	1/1	IGF2BP2	

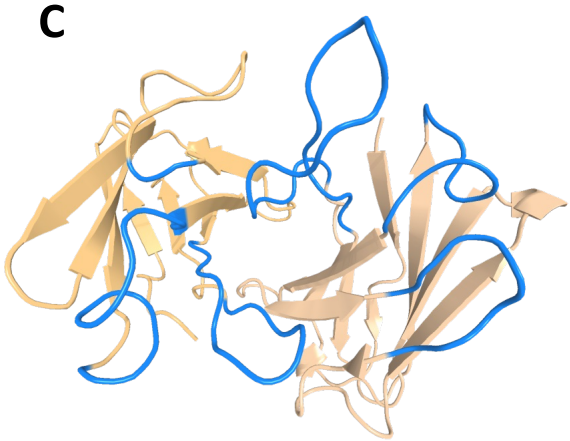
C

Cluster		V/ J gene	CDR3 (length)	Identity	Template Antigen	CDR3 Logo
c14	Query Template	TRBV9/TRBJ1-1 TRBV9/TRBJ1-1	ASSVD-GGGTEAF (12) ASSVDSGAGTEAF (13)	84%	MLANA	
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	

Figure 10


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


C

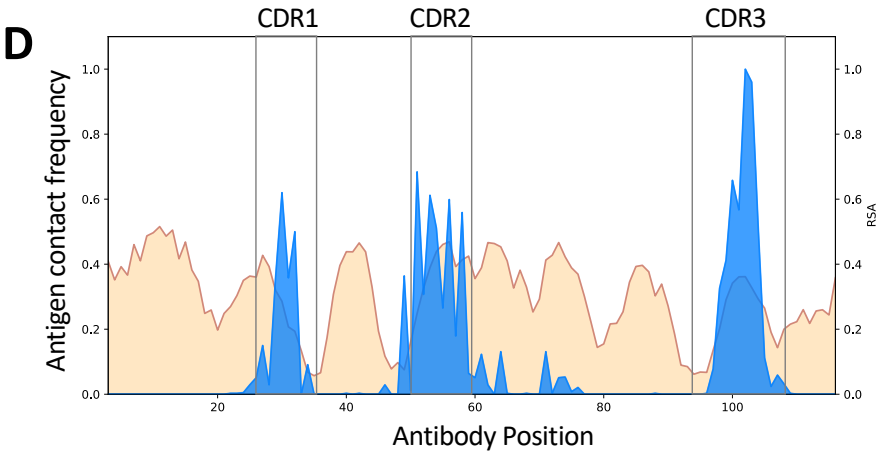

D
CDR1
CDR2
CDR3
Antigen contact frequency
RSA
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
Antibody Position

Figure S1

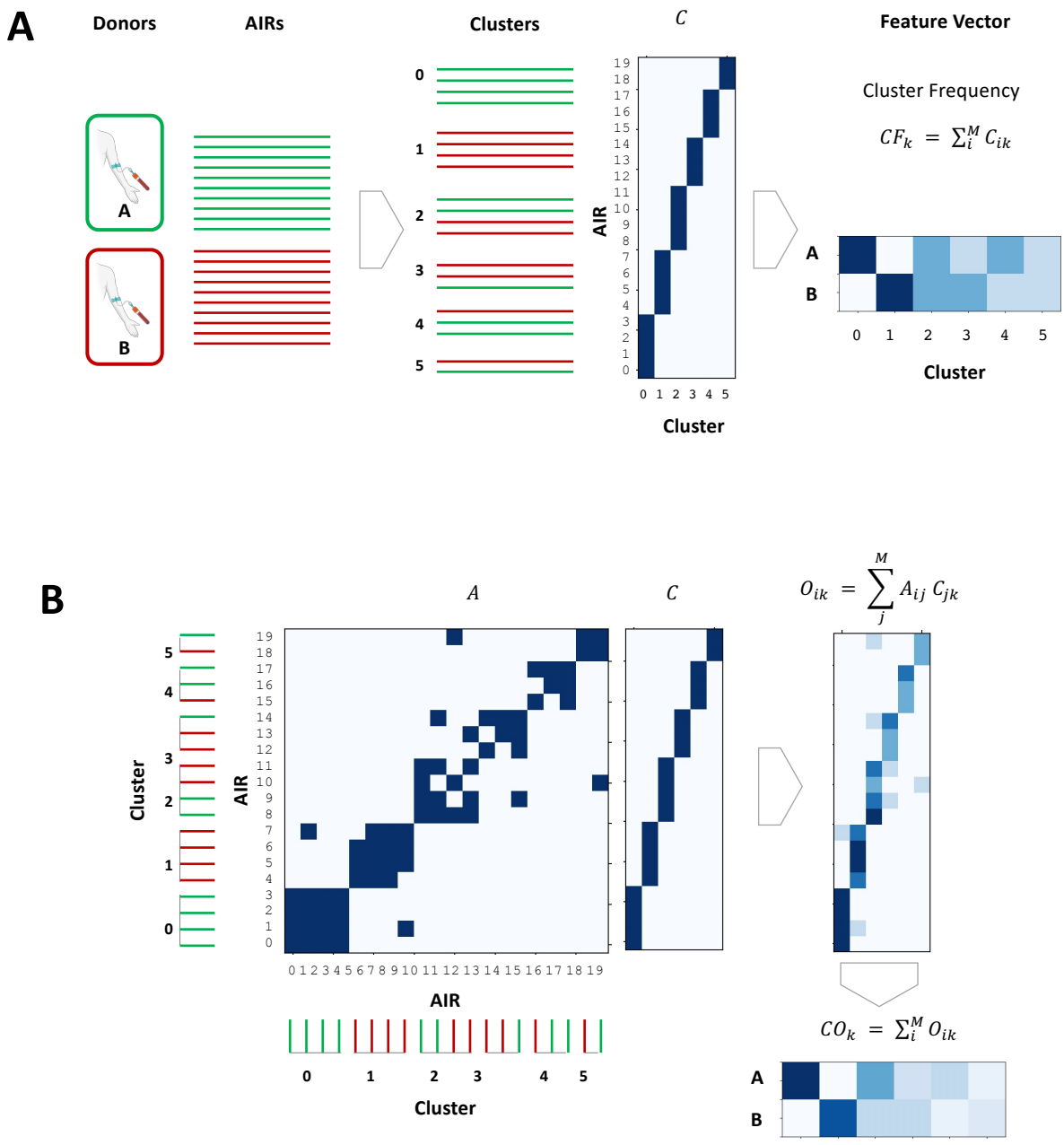

A
Donors
AIRs
Clusters
C
Feature Vector
Cluster Frequency
$CF_k = \sum_i^M C_{ik}$
A
B
0
1
2
3
4
5
AIR
Cluster
B
A
C
$O_{ik} = \sum_j^M A_{ij} C_{jk}$
$CO_k = \sum_i^M O_{ik}$

Figure S2