Motifier: An IgOme Profiler Based on Peptide Motifs Using Machine Learning

Haim Ashkenazy[†] **Oren Avram**[†] **Arie Ryvkin**[†] **Anna Roitburd-Berman, Yael Weiss-Ottolenghi, Smadar Hada-Neeman, Jonathan M. Gershoni**[*] **and Tal Pupko**[*]

The Shmunis School of Biomedicine and Cancer Research, *George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel*

Correspondence to Jonathan M. Gershoni and Tal Pupko:Corresponding authors. Fax: +972 3 6422046. *gershoni@tauex.tau.ac.il (J.M. Gershoni), talp@tauex.tau.ac.il (T. Pupko)*

https://doi.org/10.1016/j.jmb.2021.167071
Edited by Michael Sternberg

Abstract

Antibodies provide a comprehensive record of the encounters with threats and insults to the immune system. The ability to examine the repertoire of antibodies in serum and discover those that best represent "discriminating features" characteristic of various clinical situations, is potentially very useful. Recently, phage display technologies combined with Next-Generation Sequencing (NGS) produced a powerful experimental methodology, coined "Deep-Panning", in which the spectrum of serum antibodies is probed. In order to extract meaningful biological insights from the tens of millions of affinity-selected peptides generated by Deep-Panning, advanced bioinformatics algorithms are a must. In this study, we describe *Motifier*, a computational pipeline comprised of a set of algorithms that systematically generates discriminatory peptide motifs based on the affinity-selected peptides identified by Deep-Panning. These motifs are shown to effectively characterize antibody binding activities and through the implementation of machine-learning protocols are shown to accurately classify complex antibody mixtures representing various biological conditions.

Introduction

"Filamentous Fusion Phages" are described in the seminal article published by George Smith in 1985, laying the foundations of what in the fullness of time has become known as "phage display".[1] The essence of this platform is that foreign DNA is cloned into the genome of a bacteriophage that consequently expresses and displays the corresponding peptide on its surface. This affords affinity purification of that phage by an antibody that specifically binds the displayed peptide in a procedure now referred to as "bio-panning". The impact of these innovations has been enormous, and as a result has led to the 2018 Nobel Prize in Chemistry.

Today, phage display is widely used to interrogate protein–protein interactions.[2–6] Most commonly, random peptide libraries displayed on filamentous bacteriophages are screened with antibodies. Next, the foreign DNA inserts in the affinity-selected bacteriophages are sequenced, thus revealing panels of their corresponding peptides.[3,5] In theory, each monoclonal antibody (mAb) is associated with a specific set of peptides it is able to bind.[7,8] It is generally assumed that these peptides collectively reflect the epitope naturally recognized by the mAb being studied.

This assumption, however, is only partially correct, as the situation is markedly more complex. Antibody binding sites are comprised of six "complementarity determining regions" (CDR loops); three of the heavy chain and three of the light chain. The surface of the binding site, i.e., the antibody's paratope, is comprised of some 50 amino acid residues, however, only some of these actually participate in antigen recognition and physically contact the epitope.[9] In fact, much of the antibody's paratope surface remains unoccupied as is readily observed when one examines antibody–antigen co-crystals.[10]

When we interrogate an antibody with short random peptides (e.g., 6–12 amino acids), the entire surface of the paratope is free to interact with the vast collection of different peptides. Some peptides can associate with aspects of the paratope that are directly involved in epitope binding. Other peptides, however, might associate with paratope surfaces that have nothing to do with epitope recognition *per se*. In theory, affinity-selected peptides can be highly epitope-specific while others might be more related to irrelevant aspects of the paratope.

Studying affinity-selected peptides can be informative about an antibody's corresponding epitope.[11,12] Indeed, we and others developed computational algorithms to predict an antibody's epitope, based on a panel of phage-displayed affinity-selected peptides along with the atomic structure of the antigen as input.[13–19] Furthermore, affinity-selected peptides can be informative for characterizing more biologically complex situations, such as changes in the antibody repertoire following the exposure to a specific pathogen.[20,21]

While not all peptides are directly informative for epitope discovery *per se*, the collective panel of affinity-selected peptides can be taken as a detailed surrogate molecular signature for a given antibody and a reflection of its binding capacity. In this study we use the affinity-selected peptides to critically investigate the antibodies themselves. Studying the comprehensive collection of antibody affinity-selected peptides offers a systematic approach to profiling mixtures of antibodies such as those present in an individual's serum.[22,23] To that end, we have introduced a number of modifications to the bio-panning platform in the analysis of serum antibodies. A major adaptation, coined "Deep-Panning", has been to merge phage display with Next-Generation Sequencing (NGS).[19,24,25] This Deep-Panning method has enabled the sequencing of tens of millions of affinity-selected peptides that correspond to the antibodies used to screen phage display peptide libraries. The analysis of the phage displayed peptides has revealed some aberrations in the composition of the libraries used, showing biases in peptide representation and deviations from *bona fide* randomness.[26] These distortions have been addressed and corrected.[27] The application of such Deep-Panning systems has proven useful in addressing a variety of biological questions. For example, Qi *et al.* have devised a combined phage display/NGS high throughput system to map hundreds of linear epitopes,[19] Liu *et al.* combined NGS with phage display to detect peptide ligands that target murine M2 macrophages,[28] Ernst *et al.* applied NGS combined with phage display to study the evolution of protein recognition,[29] and Lövgren *et al.* used a similar approach for raising antibodies against high-density lipoprotein particles.[30] Combining NGS with phage display has also opened the way to sample the entire set of peptides that can be recognized by the serum of an individual at a specific time point (profiling the IgOme,[24]). Characterizing such peptides has many applications, e.g., it can be used to classify individuals as either sick or healthy, to discriminate between variants of a specific disease, and to evaluate a patient's prognosis.[12] For example, such analyses were recently used to study tumor-associated antigens in ovarian cancer,[31] identify antibodies associated with autoimmune Celiac disease,[32] to determine peptides that can be used to diagnose norovirus infections,[33] and to identify HIV specific epitopes in vaccinated rhesus macaques.[23] Importantly, while classic diagnostic tests are based on a single marker, analyzing the entire set of peptides that can be recognized by the serum of an individual at a specific time point can be informative regarding an array of diseases using a single blood test. Analyzing such large datasets comprised of millions of peptides and extracting the most informative and discriminatory markers is computationally intensive.[34,35]. As a result, most previous analyses were focused only on a relatively small subset of peptides (e.g., those that are most amplified) while largely ignoring information captured by the vast majority of the peptides.[11,22]

In the present study we describe "*Motifier*", a computational methodology, designed to analyze, profile, and classify different biological conditions, based on the collections of affinity-selected peptides obtained through Deep-Panning of random peptide libraries. We show how peptide-motif representation, inferred from the numerous affinity-selected peptides, is an efficient and informative approach for analyzing Deep-Panning data. Next, we demonstrate how these motifs can be utilized to accurately classify biological conditions. We start with the analysis of four mAbs and then, we demonstrate the power of *Motifier* by comparing the serum profile of antibodies in HIV-1 positive *vs.* negative individuals.

Results and Discussion

A combined experimental-computational approach

The *Motifier* algorithm relies on affinity-selection of peptides generated by Deep-Panning a phage display random peptide library against monoclonal or polyclonal antibodies that represent various biological conditions, such as diseased *versus* healthy individuals. The algorithm aims to discriminate between different conditions, based on comparative analysis of the affinity-selected peptides. For this, we first infer peptide motifs that characterize each specific biological condition and then utilize these motifs to build models for classifying new samples with respect to their (unknown) biological condition. The combined experimental-computational platform is illustrated in Figure 1.

Motifier is comprised of three main modules: (I) NGS quality assurance; (II) motif inference from affinity-selected peptides; (III) machine-learning model training for accurate classification of unseen samples based on their affinity-selected peptides (see Steps 4–6, respectively, in Figure 1). The application of *Motifier* is described in detail for the analysis of four example mAbs and subsequently for the discrimination of human polyclonal serum representing two biological conditions.

Discrimination of four biological conditions: analysis of four mAbs

As a first simple example of the *Motifier* methodology, four mAbs were analyzed, each taken to represent a different "biological condition". The specific four mAbs were selected as they have been extensively studied and their epitopes have been determined at the atomic level by X-ray diffraction analyses of antibody-antigen co-crystals. The four mAbs bind highly conformational discontinuous epitopes and thus the affinity selected peptides, albeit informative, were not expected to simply correspond to linear segments of the antigens. The four mAbs were

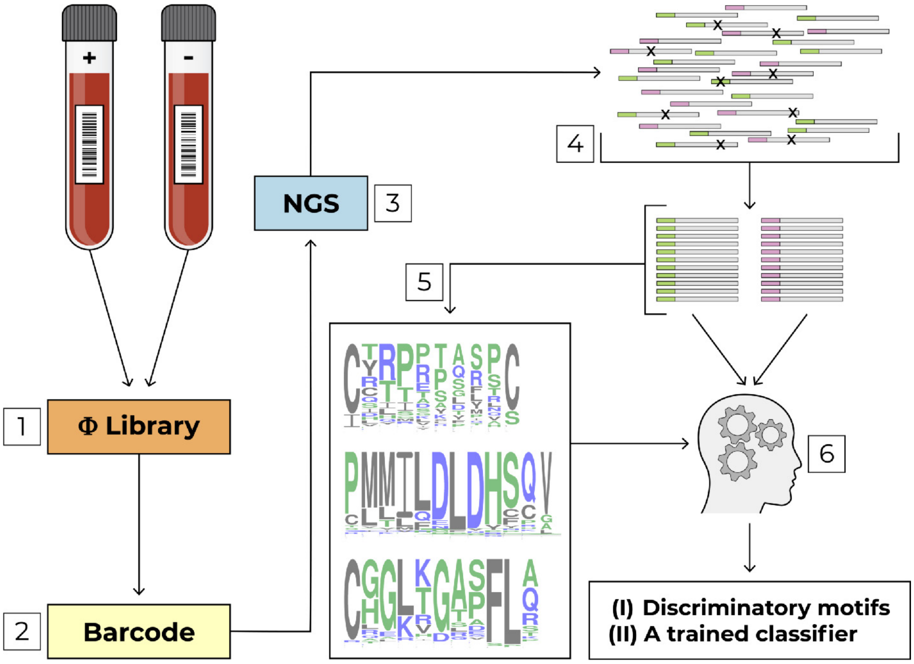


Figure 1. A schematic depiction of the combined experimental-computational platform for IgOme profiling and classification. The experimental part (Steps 1–3) entails the screening of the samples representing two (or more) biological conditions. In this case, sera from infected (+) *vs*. non-infected (−) individuals are used to screen a combinatorial phage display peptide library (Step 1). Sample-index barcodes are introduced by PCR (Step 2, pink and green "barcodes"). Then, the affinity-selected phage-displayed peptides are sequenced by NGS (Step 3). This is followed by computational analysis using the "*Motifier*" pipeline. *Motifier* consists of three main modules (Steps 4–6). First, reads undergo quality filtering, de-multiplexation, and *in-silico* translation (Step 4) yielding a curated set of affinity-selected peptides for each sample. Then, (Step 5) peptide motifs (position-specific scoring matrices) are inferred using a clustering algorithm (for each biological condition), followed by the unification of similar motifs, from repeats or multiple samples representing the same biological condition. The third module implements machine-learning modeling and classification. Each motif dictates a feature for machine learning, in which the value for the feature measures the congruence between a set of peptides in a sample to that motif. Discriminatory motifs are those for which there are different levels of congruency between biological conditions. A random-forest classifier is then trained, to classify unlabeled sera based on their peptides (Step 6). The output of the platform is: (I) a set of discriminatory motifs that can be used for further experimental analysis; and (II) a random-forest model that is able to classify new unseen samples of affinity-selected peptides. For further details see Methods and Results.

b12,[36] 17b,[37] 21c,[38,39] and Herceptin.[40] Herceptin binds the human HER_2/neu receptor, while the other three antibodies target overlapping, yet different, epitopes on the HIV-1 envelope protein gp120[39,41] (Supplementary Figure S1) and thus pose a challenge for *Motifier*, in its ability to discriminate between mAbs that recognize some common surfaces of the same antigen (HIV-1 gp120). Each mAb was used to Deep-Pan the phage random peptide library five independent times, resulting in a total of 20 index-barcoded samples. The total reads for each of the 20 samples are given in the Supplementary Table S1A. In the following experiment, *Motifier* was applied to four of the five repeats of each mAb, which served as a training set for machine learning. The machine-learning algorithm resulted in a set of discriminatory features (motifs), which were then used to classify the fifth sample of each mAb to determine the accuracy of classification of the four "unseen" samples (one for each mAb), based on their affinity-selected peptides.

NGS quality assurance

Motifier first de-multiplexes the NGS reads according to sample indexing-barcodes. For the four mAbs, we de-multiplexed the NGS data into 20 different barcodes (five replication samples for each of the four mAbs). Reads with erroneous barcodes were filtered out. Next, DNA reads that deviated from the expected sequence configuration were removed from the data set (see Methods). Then, the inserts within each DNA read (corresponding to the affinity-selected peptides) were translated to amino acid sequences. The copy number for each unique peptide within a sample was normalized by the sample total number of peptides, to balance differences in sample sizes (see Methods). The average percentage of peptides filtered out among the five samples was 13.48% and the total number of unique peptides for each sample ranged from values 18,284 to 269,430 (the total number of filtered peptides and the number of unique peptides for the 20 samples are given in Supplementary Table S1A).

Motif inference

Antibody recognition of a given peptide is mediated through its ability to associate with a defined set of chemical moieties satisfied by the specific amino acid sequence of the peptide. Surprisingly however, comparing the sets of peptides selected for the four training repeats of each of the mAbs tested, revealed that, on average, less than 3% of the unique peptides were shared by all four mAb-specific samples and more than 91% of the unique peptides were exclusively found in only one of the repeats (Figure 2(A)). This result could be explained due to: (I) the vastness of the peptide library and hence the probability of ever selecting identical peptides with each sampling of the library is extremely low, and (II) the strict definition of peptide uniqueness, even conservative exchanges, e.g., leucine for isoleucine, are taken as categorically different. Consequently, we conclude that a unique peptide sequence was not necessarily an effective mAb-defining discriminating feature. We therefore postulated that mAb recognition of short peptides is less amino-acid sequence specific but rather more dependent on a chemical "motif" that might be represented by a large collection of different member peptides affinity selected and sequenced in the Deep-Panning procedure. Hence for example, an antibody may require a positive charge in close proximity to aliphatic methyl groups situated near an aromatic residue. These functional groups need to be oriented in space so to complement the shape and chemistry of the binding surface of the paratope. Obviously, these physico-chemical constraints can be satisfied by many combinations of amino acids present in the affinity-selected peptides. Valine, leucine, isoleucine and alanine can all provide a required methyl group, and lysine might easily replace an arginine where a positive charge is sought. Thus, one expects that even a highly specific mAb should be able to cross react and bind a diversity of different peptides that all fulfill the generic pattern of functional groups and their spatial orientation by a "motif".

Discovering and constructing a sequence-motif based on a set of millions of different sequences is computationally challenging, and currently none of the developed methods can efficiently handle such large datasets of bio-panned peptides. For example, it is impractical to analyze large amounts of peptides with the widely used MEME tool[34] due to computational limitations. In addition, MEME is restricted to peptides of at least eight amino acid residues, while in our methodology, shorter peptides are often affinity-selected. The MUSI algorithm was suggested as a more efficient algorithm for motif inference.[35] The first step of MUSI involves the alignment of all sequences. Due to running-time constraints, multiple sequence alignment is limited to only a few thousand sequences, thus making the application of MUSI to analyze Deep-Panning data impractical. An alternative approach would be to construct motifs using only the most abundant affinity-selected peptides.[11,22] However, such an approach would be at the expense of considerable information loss, even when analyzing a single mAb. This becomes markedly more problematic when studying polyclonal sera that contain dozens to hundreds of distinct antibodies, each with unique specificities and present at different titers.[42–47]

In view of the above, we devised a multistep procedure in order to effectively cluster the affinity-selected peptides into sample defining

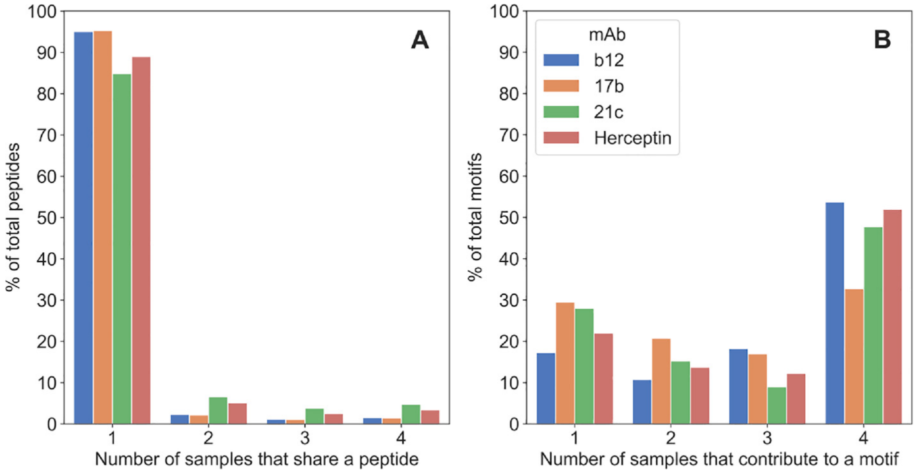


Figure 2. Comparison of unique peptides and the motifs they support among different samples. All the unique peptides for each repeat were listed. For each peptide, we counted how many replicates share it, and recorded the percentage of peptides sharing 1, 2, 3, or 4 samples (Panel A). It is clear that there is very weak overlap between the replicates as the percentage of peptides shared among 2, 3, or 4 different replicates (Y axis) is less than 5% for mAbs b12 and 17b, and no more than 10% for mAbs 21c and Herceptin. The vast majority of (unique) peptides were found in only one out of the four replicates. We also computed the percentage of motifs that are highly similar among different samples. To this end, we clustered similar motifs to a united motif (see Methods). A united motif is considered to be supported by a sample if it includes motifs from that sample. Shown in panel B is the distribution of (united) motifs supported by *i* different samples (*i* = 1, 2, 3, 4). In contrast to Panel A, there is a strong motif-overlap among the different sample replicates (Panel B).

motifs (see Methods). Briefly, peptides from each sample were first clustered using CD-HIT[48] and each cluster was aligned using MAFFT.[49] Then, from each set of aligned peptides we computed a corresponding Position-Specific Scoring Matrix (PSSM).[50,51] Comparing the motifs generated for each of the four “training” repeats for a given mAb, we observed that a large number of motifs were highly similar among samples of the same “biological condition”: on average, more than 46% of the peptides could be associated to motifs in all samples while less than 25% of the motifs were exclusive to one sample of a kind (Figure 2(B)). The high prevalence of motifs showing high similarity among samples representing a given biological condition, stood in marked contrast to the relatively low overlap of individual affinity-selected peptides described above. Therefore, we united similar motifs within a biological condition, resulting in a consolidated list of motifs that characterized each condition (see Methods). Intuitively, such a consolidated list, recapitulates information on most of the peptides that are able to interact with the antibodies of a specific biological condition, in our case a specific mAb. An example demonstrating the unification of similar motifs recognized by mAb 21c is shown in Figure 3. Note that in the fifth position of the motif, both D and E are allowed, but there is a very strong preference for D in that position. Following numerous similar observations, we decided not to a-priori group together amino acids with similar physicochemical characteristics. Nevertheless, similarity in the physicochemical characteristics of amino-acids are implicitly accounted for when constructing the motifs, e.g., in the BLOSSUM matrix used for motif construction (see Methods). Following motif unification of a given mAb, we used its set of motifs as input for machine-learning classification and identification of the most discriminating features (i.e., motifs). To this end, each motif contributed a feature to a Random-Forest classifier, as detailed below.

The machine-learning approach for classification

In the four mAbs experiment, we produced four sets of consolidated motifs. The ultimate objective of the machine-learning classification was to determine the identity of unseen samples (test data) using the discriminating features inferred above for any given condition (mAb). In this experiment four different classifiers were independently trained, one for each mAb. Thus, for the classification of the biological condition, b12 mAb for example, we aimed to determine whether peptides from an unseen sample were congruent with b12 motifs.

To that end, the above motif analysis procedure was applied to positive training samples only (i.e., motifs were not inferred for the negative samples nor for the test data). Each such inferred motif contributed a feature for the machine-learning algorithm. Consider a single motif out of *m* motifs.

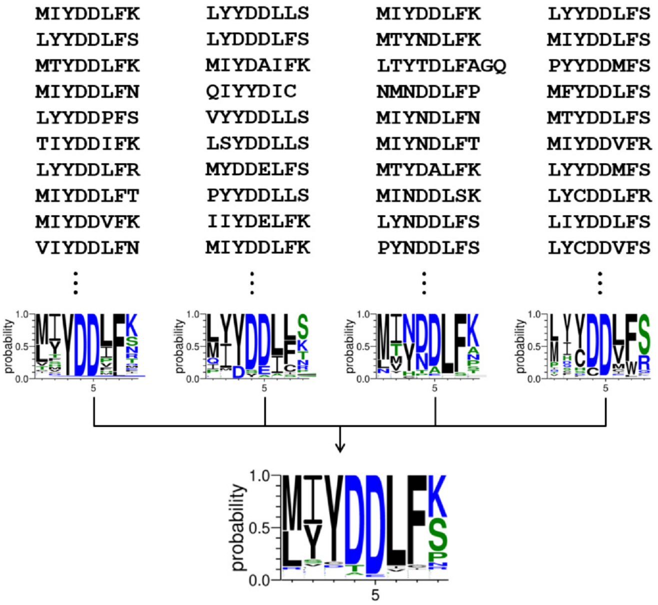


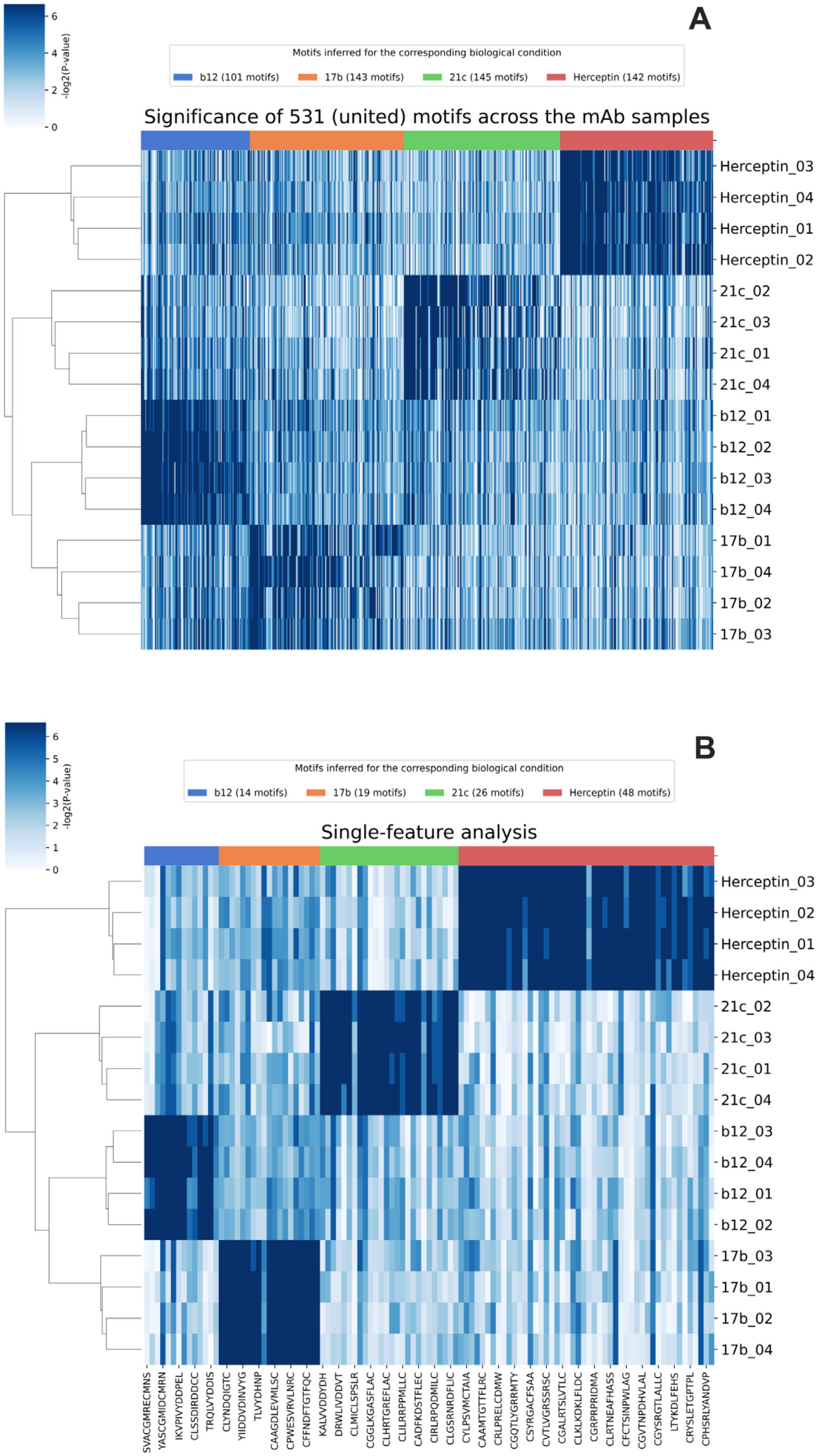
Figure 3. Motif inference. Shown is an example for the motif inference process, from a set of peptide clusters in different mAb 21c replicates. A motif is generated from clustered peptides in each sample. A final united motif is inferred from the sample-derived motifs through the process of motif unification as described in the Methods.

Intuitively, a sample is congruent with this motif if many peptides in the sample “fit” this motif. To test this, we first quantified how well each peptide “fits” a given motif (see Methods). Next, we counted the total number of peptides in the sample that fit this motif and determined whether this value was higher than the value expected by chance (see Methods). This provided a p-value for the motif in question. The lower the p-value, the stronger the fit is between the peptides in a sample and this specific motif. This procedure was repeated for each motif, resulting in an *m*-dimensional vector containing *m* p-values (one for each motif), per sample (both positive and negative). Thus, for a dataset containing *n* samples, an *n* X *m* table was generated and provided to the machine-learning algorithm for training. A toy example of such a table is demonstrated in Supplementary Table S2. For the training datasets analyzed in this experiment, the corresponding tables are given as heat-maps in Figure 4. Next, using Random-Forest machine-learning training and classification (see Methods), motifs were identified as discriminatory (i.e., informative) or not. Given *n* p-value vectors corresponding to *n* training samples and their true labels (positive or negative), we trained a Random-Forest model that would be able to classify a set of new unknown samples based on their cognate p-value vectors (see Supplementary Table S2). For the mAbs analyzed in this experiment, we trained four classifiers: one that could classify whether a sample was b12 positive or not, and similarly for 17b, 21c, and Herceptin.

Experimental testing of the machine-learning output by ELISA

In this experiment, discriminating features for each mAb were identified by machine learning. However, it should be emphasized that these features were inferred *in silico*. In order to confirm that they indeed represent biologically relevant antigenic features recognized by their corresponding mAb, we conducted ELISA tests using selected representative peptides. For this, we chose three representative motifs, one for each of the three gp120-specific mAbs, b12, 17b, and 21c. Three member-peptides for each motif were cloned and expressed as Protein VIII fusions displayed via the fth1 phage. Note, that both high and low copy number peptides were tested (Figure 5).

The ELISA results demonstrate the specific binding of all nine peptides tested, suggesting that the inference of motifs was biologically relevant. As illustrated in Supplementary Figure S1, there is some overlap between the epitopes of the three gp120-specific mAbs. Whereas eight of the peptides proved highly specific for their cognate mAb, the 21c peptide, MIYDDLFK, did cross react with 17b slightly and with b12 considerably more, which concurs with the overlap of the 21c epitope with those of b12 and 17b. No cross reactivity at all was detected with Herceptin (not shown). Taken together, the ELISA results suggest that our methodology was able to correctly detect both discriminatory and biologically meaningful motifs.


A
6
4
2
0
-log2(P-value)
Motifs inferred for the corresponding biological condition
b12 (101 motifs)
17b (143 motifs)
21c (145 motifs)
Herceptin (142 motifs)
Significance of 531 (united) motifs across the mAb samples
Herceptin_03
Herceptin_04
Herceptin_01
Herceptin_02
21c_02
21c_03
21c_01
21c_04
b12_01
b12_02
b12_03
b12_04
17b_01
17b_04
17b_02
17b_03
B
6
5
4
3
2
1
0
-log2(P-value)
Motifs inferred for the corresponding biological condition
b12 (14 motifs)
17b (19 motifs)
21c (26 motifs)
Herceptin (48 motifs)
Single-feature analysis
Herceptin_03
Herceptin_02
Herceptin_01
Herceptin_04
21c_02
21c_03
21c_01
21c_04
b12_03
b12_04
b12_01
b12_02
17b_03
17b_01
17b_02
17b_04
SVACGMRECMNS
YASCGMIDCMRN
IKVPIVYDDPEL
CLSSDIRDDCC
TRQLVYDDIS
CLYNDQIGTC
YIIDDVDINVYG
TLVYDHNP
CAAGDLEVMLSC
CPWESVRVLNRC
CFFNDFTGTFQC
KALVVDDYDH
DRWLIVDDVT
CLMICLSPSLR
CGGLKGASFLAC
CLHRTGREFLAC
CLILRRPPMLLC
CADFKDSTFLEC
CIRLRPQDMILC
CLGSRNRDFLIC
CYLPSVMCTAIA
CAAMTGTTFLRC
CRLPRELCDMW
CGQTLYGRRMTY
CSYRGACFSAA
CVTLVGRSSRSC
CGALRTSLVTLC
CLKLKDKLFLDC
CGRPRPRIDMA
CLRTNEAFHASS
CFCTSINPWLAG
CGVTNPDHVLAL
CGYSRGTLALLC
LTYKDLFEHS
CRYSLETGPTPL
CPHSRLYANDVP

Machine-learning classification using discriminatory motifs

Our next goal was to determine whether the machine-learning classification model could be used to classify an unseen dataset. For this, at the start, one sample from each mAb was set aside and treated as an unknown label (i.e., these samples were not used for motif inference nor for model training). Given a machine-learning model trained for a specific mAb, e.g., Herceptin, there are four possible outcomes (binary classification): (1) the model successfully classifies a sample obtained by panning against Herceptin as “Herceptin positive” (a true positive prediction); (2) the sample is misclassified as “Herceptin negative” (false negative); (3) the model erroneously classifies a sample of peptides, obtained by panning against another antibody (e.g., 17b), as “Herceptin positive” (false positive); (4) the machine learning classifier successfully recognizes a sample originating from a different antibody as “Herceptin negative” (true negative). We trained a classifier for each mAb, and evaluated it on the test set comprised of four unseen samples (see Methods). Using a Random-Forest classifier, we obtained perfect classification (100%) on the test datasets for all mAbs, i.e., there were neither false negative nor false positive predictions. Of note, Random Forest was selected based on its performance for the learning data, as it outperformed several other classification algorithms, including KNN, LDA, SVM with several kernels, Naïve Bayes, and logistic regression, with mean accuracy ranging from 0.562 to 1.0 (Supplementary Table S3)

Our machine-learning algorithm allows combining signals from many peptide motifs in order to obtain optimal classification. However, we can also train a model based on a single motif and determine its classification accuracy (single-feature analysis). We applied this single-feature analysis for each motif in each of the four mAbs. There were 14, 19, 26, and 48 motifs for b12, 17b, 21c, and Herceptin, respectively, which perfectly discriminated their corresponding mAbs from the other three (Figure 4(B)).

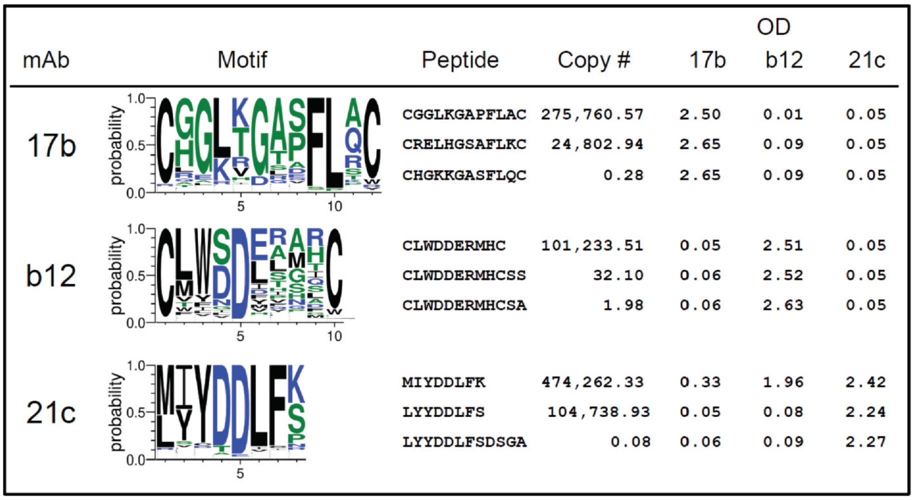
Encouraged by these results, and by the strength of the “signal” in the mAbs data (see Figure 4(A)), we next aimed to test the sensitivity of the analysis. For this, we diluted each sample, *in-silico*, in a ratio of 25% drawn from the original sample and 75% of unrelated peptides. In total we had 16 diluted samples, four per each mAb. Here too, we obtained perfect classification. We repeated this analysis with increasing levels of dilution (10%, 1%, and 0.1% drawn from the original sample, again 16 diluted samples for each dilution level). We obtained perfect classification for the 10% and the 1% dilutions. For the 0.1% dilution we were unable to detect sufficient signal for training (as the training error of each classifier was high), nor for testing.

To test the sensitivity of our approach in a more challenging setup, we repeated the dilution experiments, this time, mixing two mAbs in each case. The rationale was that for the three gp120 specific mAbs, there may be some degree of peptide cross reactivity and thus the question was whether the classifiers would be able to discriminate and identify signals derived from both similar and distinct mAbs. Both the training and test datasets were set by mixing, *in-silico*, 25% of peptides selected by one mAb (e.g., Herceptin) with 25% of peptides derived from a second mAb (e.g., b12) with 50% irrelevant peptides. As before, for the test set we used peptides from the (unseen) fifth sample of each mAb. This test was repeated twelve times to cover all six possible pairs of mAbs and test whether we were able to correctly classify each of the two mAbs represented in each peptide mixture. We repeated this with 10%, 1%, and 0.1% dilutions for each of the test mAbs (i.e., 80%, 98%, and 99.8% irrelevant peptides, respectively). Once again, we obtained perfect classification for all dilution levels except for the 0.1%.

Motifier analysis of polyclonal serum

In the four mAbs experiment described above, each mAb was tested separately. The dilution experiment illustrated that a mixture of two different peptide panels could be analyzed successfully based on discriminating features previously identified by machine learning. More relevant however, is to test *Motifier* in a realistic setting in which the signal-generating antibodies are present among a vast collection of otherwise irrelevant antibodies. Such is the case comparing genuine serum samples taken from individuals representing two distinct biological conditions. For

Figure 4. Four mAbs experiment: motif significance represented as heat-maps, before and after machine learning. Four mAbs were used to affinity-select peptides and motifs were inferred for which p-values were calculated for each sample. Each column corresponds to a motif, represented by its consensus, each row corresponds to a given sample, and the *i,j* entry is a p-value quantifying the congruence of sample *i* with motif *j*. (A) The 531 statistically significant motifs that were used as input to the machine learning; (B) single-feature analysis yielded 107 motifs, each of which classifies the samples with 100% accuracy in 4-fold cross validation. Selected consensus sequences of the motifs are shown.



mAb	Motif	Peptide	Copy #	OD 17b	 b12	 21c
17b		CGGLKGAPFLAC	275,760.57	2.50	0.01	0.05
		CRELHGSAFLKC	24,802.94	2.65	0.09	0.05
		CHGKKGASFLQC	0.28	2.65	0.09	0.05
b12		CLWDDERMHC	101,233.51	0.05	2.51	0.05
		CLWDDERMHCSS	32.10	0.06	2.52	0.05
		CLWDDERMHCSA	1.98	0.06	2.63	0.05
21c		MIYDDLFK	474,262.33	0.33	1.96	2.42
		LYYDDLFS	104,738.93	0.05	0.08	2.24
		LYYDDLFSDSGA	0.08	0.06	0.09	2.27

probability
1.0
0.5
0.0
5
10

Figure 5. mAb binding to phage-displayed selected peptides. mAbs b12, 17b, and 21c bind different epitopes on HIV-1 gp120 that partially overlap (Supplementary Figure S1). In order to confirm that member-peptides of the clustered motifs for each mAb are actually recognized and bind their corresponding antibodies, three peptides from each mAb-motif were cloned and expressed as Protein VIII fusions on filamentous phages using the fth1 vector. The phage-displayed peptides were then used in ELISA tests (see Methods). The motifs and peptide sequences along with their corresponding copy numbers (per million) are shown. The O.D. values for the nine peptides for the three gp120 specific mAbs are given. Note that except for cross reactivity of the MIYDDLFK peptide of the mAb 21c, all other peptides proved highly specific for their corresponding mAbs. A tenth peptide derived from a Herceptin motif (YASTIVVDLDHT) as well as the fth1 vector alone served as negative controls and bound less than 0.1 O.D. The HIV-1 envelope protein, gp120, served as the positive control for the three mAbs being studied and produced signals greater than 2.5 O.D. for each mAb.

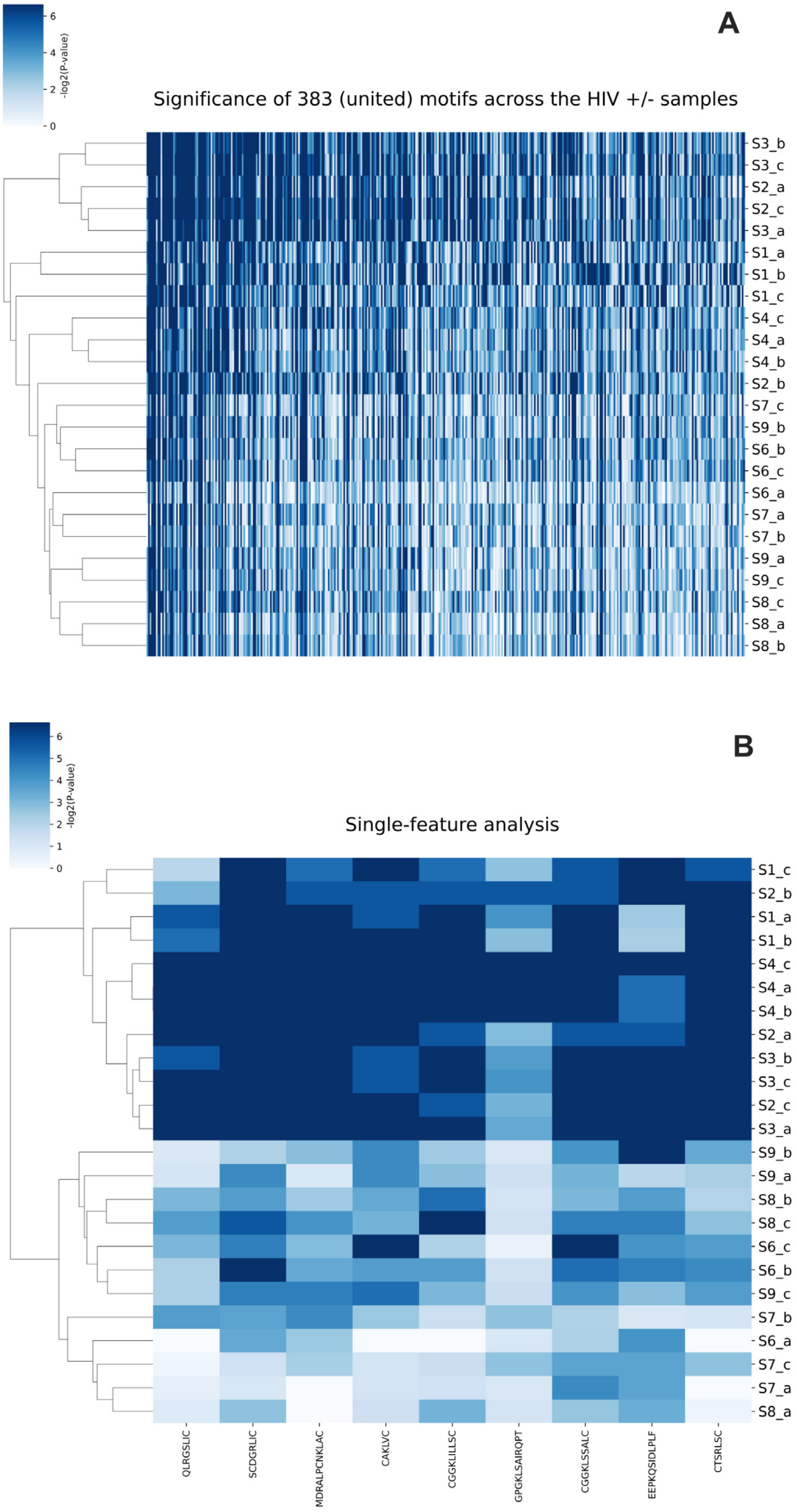
this we Deep-Panned ten different serum samples, five from HIV-1 positive individuals and five from HIV-1 negative people. Each serum sample was deep-panned in triplicate representing one of the two biological conditions, namely HIV-1 positive *vs*. negative. Four triplicates for each condition were used for training while the fifth triplicate was set aside as test data.

After panning and NGS a total of 89,226,224 sample-indexed barcoded peptides were subjected to quality filtration during which a total of 10,527,727 peptides (11.8%) were discarded. The total number of filtered peptides and unique peptides for the 30 samples are given in Supplementary Table S1B. Motifs were inferred for the four HIV-1 positive triplicates. After motif unification the input for the machine learning is a set of p-values for each motif, which is represented as a heat-map in Figure 6(A). There were a total of 383 statistically significant HIV-1 defining motifs (p-value < 0.05, see Methods). Careful examination of the data shows that the two biological conditions can be separated, namely HIV-1 positives (S1-S4) compared with HIV-1 negatives (S6-S9) (Figure 6(A)). Following training of a machine-learning classifier on these data yielded perfect classification on an unseen test set, namely the fifth HIV-1 positive and negative triplicates that were set aside. Figure 6(B) gives the heat-map generated using nine features selected by the single-feature analysis, in which each of these features had accuracy greater than 90%. The crispness and contrast between the heat-maps before and after the application of the machine learning illustrates the improvement in classification ability introduced by the learning process.

We next repeated the noise analyses described above for the mAbs for the polyclonal serum data. For this, we diluted each sample, *in-silico*, in ratios of 25, 10, 1, and 0.1 percent, drawn from the original sample and the remaining peptides sampled from unrelated peptide dataset. Similar to the results above for the mAbs, the signal deteriorates for high dilution ratios: for both the 1% and the 0.1% dilutions, the accuracy obtained was 83.3% (with ROC AUC = 0.89 and ROC AUC = 0.83, respectively). In contrast, a perfect classification was obtained for the 25% and 10% dilutions (ROC AUC = 1 for both). These results suggest that one of the factors contributing to the perfect performance is the high number of relevant peptides.

Machine-learning characteristics and potential limitations

Each machine-learning algorithm could suffer from over-fitting (when the trained model fits the analyzed data but fails to generalize to


A
-log2(P-value)
0
2
4
6
Significance of 383 (united) motifs across the HIV +/- samples
S3_b
S3_c
S2_a
S2_c
S3_a
S1_a
S1_b
S1_c
S4_c
S4_a
S4_b
S2_b
S7_c
S9_b
S6_b
S6_c
S6_a
S7_a
S7_b
S9_a
S9_c
S8_c
S8_a
S8_b
B
-log2(P-value)
0
1
2
3
4
5
6
Single-feature analysis
S1_c
S2_b
S1_a
S1_b
S4_c
S4_a
S4_b
S2_a
S3_b
S3_c
S2_c
S3_a
S9_b
S9_a
S8_b
S8_c
S6_c
S6_b
S9_c
S7_b
S6_a
S7_c
S7_a
S8_a
QLRGSLIC
SCDGRLIC
MDRALPCNKLAC
CAKLVC
CGGKLILLSC
GPGKLSAIRQPT
CGGKLSSALC
EEPKQSIDLPLF
CTSRLSC

independent unseen datasets). This is especially true when the number of observations is relatively small. In our analyses, special care was taken to avoid over-fitting: all training and model-selection procedures were conducted on the training data using cross validation and the test data were only used to evaluate the performance of the selected model. Within the cross-validation scheme, over-fitting was controlled for by a feature-selection procedure (on the training data), in which we iteratively reduced the number of features to the minimum number of features required to accurately classify the data. Finally, we tested several machine-learning classifiers and performed noise analyses. When using these classifiers, or when substantial noise was introduced, performance accuracy declined. This suggests that the high accuracy obtained on the test data is a combination of using the Random-Forest classifier that fits well for the type of data analyzed here as well as the large number of peptides available for analysis. We also note that the noise analysis suggests that the high performance is partially due to the large number of relevant peptides available for analysis. This is a direct result of the incorporation of NGS following the panning procedure as well as our ability to incorporate the signal in these peptides into motifs, rather than discarding large numbers of peptides. Nevertheless, the performance of the algorithm should be better estimated in the future when data with more observations are available.

Conclusions

In this work we developed a computational framework to analyze Deep-Panning experimental data. Our computational pipeline can be divided into three stages: NGS data quality control and translation, biological condition profiling (motifs inference), and predictive model construction to classify new unseen samples. We demonstrated the application of the algorithm to distinguish between peptide data panned against four different mAbs and sera of HIV-1 infected and non-infected individuals. We provided experimental data (ELISA tests) to validate the results of our analyses and have shown the capability to uncover biologically meaningful features of the biological condition being profiled. Our ability to correctly classify the four mAbs (even when the mAbs experimental data were diluted 100-fold) and the HIV-1 positive *vs.* negative samples demonstrates the benefits of using peptide motifs when analyzing next-generation phage-display data and suggests that the platform developed here can extract differentiating signals from polyclonal sera obtained from different biological conditions. In our case, we first developed the pipeline using model “biological conditions” i.e., four distinct mAbs probed in five repeats each. We then applied the system to examine genuine separate biological conditions, namely sera taken from HIV-1 positive and negative individuals, five different people for each biological condition. The same approach can easily be generalized to other pathogens. In addition, it is of interest to test the ability of the methodology to distinguish between more subtle biological conditions, such as response to a treatment, age related immunological differences, and disease progression. Further research is needed to answer such questions as how many samples are sufficient for accurate classification in such cases and what will be the performance of our approach on such challenging tasks.

Materials and Methods

Reagents

The combinatorial random peptide libraries used in this study were produced in house at Tel Aviv University using the fth1 filamentous bacteriophage system.[52,53] fth1 is an 8 + 8 filamentous fd bacteriophage, harboring a second *protein VIII* gene in which recombinant DNA oligonucleotides are cloned into the *Sfi1* sites flanking the cloning cassette.[52] The random peptide library used in this study contained a mixture of peptides 6, 8, 10, and 12 amino acids long, with or without flanking cysteine residues, thus displaying a vast collection of linear and cysteine constrained loops (total complexity > 10^{10} recombinant peptides) as Protein VIII fusions. In order to ensure optimized peptide randomness, the construction of the library was conducted using preferred phosphoramidite ratios for both the N and K positions (N = G:1.0, A:1.5, T:1.5, C:1.6, and K = G:1.0, T:1.5, The Midland Certified Reagent Company, Inc.) and the libraries were produced in *SupE* positive bacteria.[27] Moreover,

◀

Figure 6. HIV-1 positive *vs.* negative sera: motif significance represented as heat-maps, before and after machine learning. HIV-1 positive (S1-S5) and negative (S6-S10) serum samples were used to affinity-select peptides. Motifs were inferred for the HIV-1 positive samples (S5 was not used for motif inference and model training) and p-values were calculated for all samples. Heat-maps were generated in which each column corresponds to a motif, each row corresponds to a given sample, and the *i,j* entry is a p-value quantifying the congruence of sample *i* with motif *j*. (A) The 383 statistically significant motifs that were used as input to the machine learning; (B) single-feature analysis yielded nine motifs, each of which classifies the samples with at least 90% accuracy in 4-fold cross validation. Consensus sequences of the motifs are shown.

repeated NGS analysis of the optimized library showed that no fortuitous “parasitic phages” were present.[26] See Ryvkin et al for complete details on the library construction, characterization and use.[27]

The human monoclonal antibodies used as bait in this study were Herceptin[40] and three mAbs, b12,[36] 21c[38,39] and 17b[37] that have distinct epitopes within HIV-1 gp120. Co-crystal analyses illustrate the relative overlap and proximities of the three epitopes bound by these antibodies (see Supplementary Figure S1).

All the serum samples were collected under informed consent and by IRB approval. The HIV-1 positive sera were kindly provided by Dr. Bart Haynes of Duke University and Dr. Dan Turner of the Sourasky Medical Center, Israel. The HIV-1 negative serum samples were obtained from the Rabin Medical Center, Petach Tikva, Israel or the Israeli National Blood Bank, Tel Hashomer, Israel.

All other chemicals, kits and reagents used throughout this study were of analytical grade and purchased and used as indicated.

Experimental procedures

The* “*Deep Panning*” *of random peptide libraries. The Deep-Panning procedure was essentially as previously described.[24] Shortly, ELISA 8 well strips (Costar® Corning Incorporated, Corning, NY, USA) were coated with protein G (70 μg/ml, Sigma-Aldrich, P4689) diluted in TBS (50 mM Tris-HCl pH 7.5, 150 mM NaCl) over night at 4 °C. The wells were blocked with TBST-BSA (TBS completed with 0.5% of Tween20 and 0.5% BSA) for 1 hour at room temperature. In parallel, 15 μg/ml of purified human antibodies or 1:100 dilution of polyclonal sera were incubated with 10^{11} phages of the optimized random peptide fth1-phage display library, suspended in TBST-BSA, for 1hr at room temperature. After blocking, the wells were washed twice with TBST (TBS completed with 0.5% of Tween20) and incubated for 1 hour with the phage library-mAbs/sera mixture, at room temperature. Subsequently, the plate was washed ten times and the bound phages were eluted (100 mM HCl-Glycine, 1 mg/ml BSA, pH 2.2) and neutralized (1 M Tris-HCl-NaOH, pH 9.1) (capture #1). For captures #2 and #3 additional rounds of amplification and biopanning were carried out. In this study the analyses were performed on elution #3, i.e., after capture #3. Finally, the eluted phages were prepared for Illumina NGS.

Sequencing. Preparations for NGS were conducted as previously described.[27] In short, following bio-panning, the eluted phages were directly used as template (2 μl) for a 60 μl PCR reaction using the Taq polymerase (Larova GmbH, cat. no. PCR-108) and forward (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT<u>NNNNN</u>CAACGTGGC) and reverse (CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCCCCAGAGGC) primers. The thermal profile was: (1) 94 °C, 5 min; (2) 94 °C, 30 sec; (3) 60 °C, 30 sec; (4) 72 °C, 30 sec; (5) go back to step 2 × 34; (6) 72 °C, 5 min.

Adapters A and B for Illumina sequencing and five nucleotide sample index-barcodes to allow multiplexing (underlined NNNNN in the forward primer) were introduced during PCR. The amplified PCR products were validated for size by running in 2% agarose gels. PCR samples were purified by Agencourt AMPure XP - PCR Purification (Beckman Coulter, A63881). The concentration of the PCR cleaned products was measured using a Qubit 2.0 fluorometer (Life Technologies, Q32866), diluted to 2 nM and sent for Illumina sequencing at the Technion Genome Center, Haifa, Israel (Illumina HiSeq V4). The NGS data generated as part of this study were deposited at Dryad and are available at https://datadryad.org/stash/share/FgZa1t2KuWn8dM-BY99ArcL_T9vfgxI4Iy394dojdeU.

ELISA. Phages expressing specific peptides were prepared by cloning DNA inserts corresponding to selected peptides into the *Sfi1* cloning cassette of the fth1 phage display vector.[20] To test the binding of mAbs to these peptides, ELISA was conducted as follows: wells of a standard 96-well ELISA plate (Corning Inc. Life Sciences, Tewksbury, MA) were coated with any one of the four mAbs used in this study (7.5 μg/well) in Phosphate-buffered saline (PBS). After washing with PBS completed with 0.05% Tween 20 (PBST), the wells were blocked with 5% skim milk/PBST and 20% horse serum, 1×10^{10} phages of each peptide in blocking solution were added to the wells of all four mAbs (for 1 hour at room temperature) to test specific and non-specific binding to the mAbs. After washing in PBST buffer, phage binding was detected using a polyclonal rabbit anti-M13 antibody diluted 1:5,000, followed by incubation with anti-rabbit HRP-conjugated antibody (Jackson, West Grove, PA) diluted 1:5,000. Finally, the wells were washed three times and reacted with TMB/E ELISA substrate (Merck Millipore, Billerica, MA). Absorbance was measured at 650 nm using a micro-plate reader (BioTek, Winooski, VT, USA). HIV-1 gp120 was captured as a positive control for antibodies b12, 17b, and 21c. Wild type fth1 was used as a negative control for mAb detection.

Computational pipeline. The computational pipeline was divided into three modules: (I) Quality Control; (II) Motif Inference; (III) Machine Learning (Steps 4–6 in Figure 1). The source code of *Motifier* was written in Python and C++ and is freely available at https://github.com/orenavram/IgomeProfiling.

Quality control. The total number of reads was 133,080,430 and 89,226,224 for the mAb and

HIV-1 experiments, respectively. Reads were demultiplexed into 20 (mAb) and 30 (HIV-1) different samples according to their sample index barcodes (Supplementary Tables S1A and S1B). Four "Quality Control" criteria were applied as follows: (1) the sample index barcodes must be 100% correct; (2) the left and right constant regions (upstream and downstream to the encoded random peptide sequence and including the *Sfi1* cloning sites) contain, in total, no more than k mismatches (k = 1 by default); (3) the random insert is consistent with the specification of the random library (in our case, each codon is sampled according to the NNK rule, i.e., any nucleotide in the first and second positions and G or T in the third position); (4) the total random insert length ranges from 4 to 12 amino acids, namely, 12 to 36 DNA base pairs, in addition to the possibility of flanking cysteine residues. Finally, all sequences that passed quality control criteria were *in-silico* translated, collapsed into a set of unique sequences, and normalized into reads-per-million to balance the different number of reads in each sample.

Motifs inference of a biological condition. The set of peptides that were generated for each sample were used as input for motif inference. For example, in the "mAbs experiment", each mAb was screened against the phage display library five independent times, thus five "samples", each with its list of peptides. The five samples of a given mAb, together are taken to represent a single "biological condition". Hence, the total 20 samples represent four biological conditions, which correspond to the four different mAbs of the experiment.

Motif inference was conducted in two steps, first motifs per individual sample were defined and then, similar motifs of different samples within a given biological condition were united as follows. For each individual sample, peptides were first clustered according to the pairwise sequence identity using CD-HIT with 50% sequence identity cutoff, i.e., all pairwise sequences with 50% sequence identity or higher relative to a cluster's seed, were clustered together.[48,54] For each cluster of sequences, the amino-acid multiple sequence alignment was then inferred using MAFFT (version 7.149).[49,55] To reduce computational times, if a cluster included more than p_1 different peptides (p_1 = 1,000 by default), only the p_1 unique peptides that are most frequent were aligned. To further reduce running times and to remove potential noise, only the c_1 (c_1 = 100 by default) most abundant clusters from each sample were retained. The abundance of each motif was defined as the number of peptides (not necessarily unique) from which it was comprised. Finally, from each aligned cluster, a Position-Specific Scoring Matrix (PSSM,[51]) was computed.

Next, identical or similar motifs from samples of the same biological condition were united. For each sample, we collected the c_1 most abundant motifs and applied the following procedure to determine which motifs should be united.

A pair of PSSMs (motifs) to be analyzed for similarity was aligned using the consensus sequence of each PSSM. The global alignment between the two consensus sequences was computed.[56] This resulted in paired PSSMs of equal lengths. Next, the Pearson Correlation Coefficient (PCC) between a pair of aligned PSSMs was calculated (gapped positions were not included in the computation), and was united when the PCC was higher than 0.6 (see[57,58]). Pairwise unifications started with the most abundant PSSM in the consolidated list of motifs of a given biological condition. Hence, we computed the PCC between the most abundant motif and each of the other motifs and marked all those with a PCC value above the 0.6 threshold. The most abundant motif and the marked motifs (if any) were united and removed from the list. This process was repeated until all motifs were removed from the motif list. In case several motifs were united, we realigned all the p_2 (100 by default) peptides within the united motif (most abundant peptides within each united motif were aligned and the others were discarded) and generated a new united PSSM from these aligned peptides.

Assigning peptides to specific motifs – setting a motif-specific threshold. We simulated 1,800,000 random peptides, 200,000 for each peptide length between 4 and 12 (total 9), half of which were simulated with flanking cysteine residues. Each peptide followed the frequency of codons in an NNK codon table. The amino acid frequencies dictated by the NNK codon table are: {A, G, P, Q, T, V} = 0.0625; {C, D, E, F, H, I, K, M, N, W, Y} = 0.03125 and {L, R, S} = 0.09375 (Q = 0.0625 is a result of SupE suppression of the UAG stop codon[59]).

Next, we computed for each random peptide its likelihood score for a given PSSM.[50,51] This produced a distribution of scores for random peptides for the given PSSM. This process was repeated for all united PSSMs of a given biological condition.

Finally, we needed to set a discriminating threshold-score for each PSSM, to be used for determining whether a peptide is congruent with a PSSM or not (score above or below the threshold, respectively, see below). The threshold-score (X) was set as the score for which α / M of the peptides obtained a higher grade than X and the remaining peptides obtained a lower grade (we used α = 5% and divided by the total number of motifs (M) in the sample, as a Bonferroni correction). For example, in our Herceptin dataset, we had 142 motifs, thus $\alpha / M \approx 0.035\%$. Hence, ~99.965% of the random peptides did not satisfy membership for each motif.

Associate peptides to motifs. In the first experiment, four mAbs were used to screen the library five times each, resulting in 20 samples, representing the four biological conditions. Consequently, four sets of united PSSMs were inferred. Here we assigned the entire collection of unique peptides from each of the 20 samples to each and all possible PSSMs of the experiment using the corresponding *X* threshold-scores defined above. In this manner each PSSM, across all the biological conditions, was assigned with a discrete number of associated peptides for each sample. In other words, a matrix was formed in which the rows were the 20 samples and the columns were all the possible PSSMs of the entire experiment, and the *i,j* entry is the number of peptides (not necessarily unique) from sample *i* associated with motif *j*. Obviously, one expected more peptides of a sample from a biological condition to be associated with motifs defined for that biological condition as opposed to motifs from different biological conditions. Next the statistical significance of the number of peptides associated with a given PSSM was determined.

Calculate motif significance. The above procedure allowed running a set of peptides against a given motif, and detection of all the peptides associated with that motif. Let *np* be the number of peptides (as stated above, not necessarily unique) associated with this motif. We would like to determine if *np* is larger than the value expected by chance. If so, this would allow determining whether a given sample is enriched with peptides associated with a specific motif. To this end, we shuffled the motif's PSSM *N* times (here, $N = 100$) and recorded the number of associated peptides *N* times, once for each shuffled PSSM, using the same set of peptides and the same threshold-score for the motif. This resulted in a distribution of expected *np* values under a null model, in which the association between the peptides and the PSSM was due to chance alone. Next, we determined an empirical p-value from the observed *np* values and the null distribution. The output of this stage was a p-value for each motif for each biological condition. Intuitively, if for a given sample a motif receives a low (significant) p-value, we conclude that the sample contains peptides characterized by this motif (see toy example in Supplementary Table S2).

For the subsequent machine-learning step, the features we collected were the p-values for each motif. For a given biological condition, we were only interested in positively discriminating features. Thus, for the learning step, we kept only motifs that had at least one significant p-value (<=0.05) across all samples of the examined biological condition. If for example, we had a motif that was inferred for 17b but its p-value was higher than 0.05, in all the 17b samples, it indicated that this motif was as specific to 17b as many random shuffles of that motif, and thus this feature was excluded.

Machine-learning classification. In order to predict the label (biological condition) of an unseen sample, we built a binary Random-Forest classifier[60] for each biological condition. Random-Forest algorithm holds a few advantages for the task of identifying the most relevant motifs and correctly classifying unknown samples. Notably, it was suggested to avoid over-fitting, especially when the number of features (in our case motifs) are much larger than the number of samples.[61,62] This is achieved by computing an ensemble of decision trees models, each of them using a bootstrap sample of the data and a random set of features for each tree-split.[63] In turn, this allows us also to estimate the importance of each variable (motif) and select the most discriminatory motifs allowing for accurate classification. These advantages made the Random-Forest algorithm a useful tool for biomarker discovery.[63–66] Specifically, we used the Random-Forest algorithm implemented as part of the Python SciKit-learn package.[67] To avoid overfitting, that is highly likely to occur when the number of features is larger than the number of observations, we sampled 1,000 hyper-parameter configurations (e.g., number of decision trees, trees depth, *etc.*) from a hyper-parameters grid. For each configuration, we trained a model using the feature-selection procedure demonstrated in the study of Svetnik *et al.*[68] with 4-fold stratified cross validation, i.e., the same fraction of positive and negative cases in each fold (in the four mAbs experiment each fold included one mAb sample representing the biological condition and three other samples). Briefly, the feature selection procedure sorts the features (motifs) by their importance for correctly classifying the sample, as determined by the Random-Forest algorithm. Next, a model is trained and evaluated for different subsets composed of decreasing numbers out of the most important features (here, in each iteration half of the features were used). Eventually, we chose the model with the lowest error rate, where ties were broken by the lower number of features.

CRediT authorship contribution statement

Haim Ashkenazy: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Oren Avram:** Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Arie Ryvkin:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - review &

editing, Visualization. **Anna Roitburd-Berman:** Methodology, Validation, Investigation, Resources, Writing - review & editing, Visualization. **Yael Weiss-Ottolenghi:** Methodology, Validation, Investigation, Resources, Writing - review & editing, Visualization. **Smadar Hada-Neeman:** Methodology, Validation, Investigation, Resources, Writing - review & editing, Visualization. **Jonathan M. Gershoni:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Tal Pupko:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

This research was supported by the Israel Ministry of Science, Technology and Space, Grant #47133 to T.P. and J.M.G. and by a United States – Israel Bi-national Agricultural Research and Development Fund (BARD) No. IS-4287-10 to J. M.G. and the National Institute of Health [4 R21 AI096940] to J.M.G. O.A. and H.A. were supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. H.A. was supported in part by Humboldt Research Fellowship for Postdoctoral Researchers from the Alexander von Humboldt Foundation. O.A. was supported in part by a fellowship from the Dalia and Eli Hurvits foundation. A.R-B., Y.W-O., S.H-N. and A.R. were all recipients of Jakov, Miriana, and Jorge Saia Fellowships. J.M.G. is grateful for the generous support of Dr. Peter Kraus. The Authors wish to thank Dr. Bart Haynes and Dr. Dan Turner for the provision of the HIV-1 positive serum samples. The authors thank Or Bialik for technical support.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.167071.

Received 18 November 2020;
Accepted 22 May 2021;
Available online 28 May 2021

Keywords:
deep-panning;
phage display;
random peptide libraries;
next-generation phage display

† These authors contributed equally.

References

1. Smith, G.P., (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315–1317.
2. Sundell, G.N., Ivarsson, Y., (2014). Interaction analysis through proteomic phage display.. *Biomed. Res. Int.*, **2014**, https://doi.org/10.1155/2014/176172 176172.
3. Hamzeh-Mivehroud, M., Alizadeh, A.A., Morris, M.B., Bret Church, W., Dastmalchi, S., (2013). Phage display as a technology delivering on the promise of peptide drug discovery. *Drug Discov. Today*, **18**, 1144–1157. https://doi.org/10.1016/j.drudis.2013.09.001.
4. Potocnakova, L., Bhide, M., Pulzova, L.B., (2016). An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction. *J. Immunol. Res.*, **2016**, 1–11. https://doi.org/10.1155/2016/6760830.
5. Gershoni, J.M., Roitburd-Berman, A., Siman-Tov, D.D., Tarnovitski Freund, N., Weiss, Y., (2007). Epitope Mapping. *BioDrugs*, **21**, 145–156. https://doi.org/10.2165/00063030-200721030-00002.
6. Pande, J., Szewczyk, M.M., Grover, A.K., (2010). Phage display: Concept, innovations, applications and future. *Biotechnol. Adv.*, **28**, 849–858. https://doi.org/10.1016/j.biotechadv.2010.07.004.
7. Scott, J.K., Smith, G.P., (1990). Searching for peptide ligands with an epitope library. *Science*, **249**, 386–390.
8. Aghebati-Maleki, L., Bakhshinejad, B., Baradaran, B., Motallebnezhad, M., Aghebati-Maleki, A., Nickho, H., Yousefi, M., Majidi, J., (2016). Phage display as a promising approach for vaccine development. *J. Biomed. Sci.*, **23**, 66. https://doi.org/10.1186/s12929-016-0285-9.
9. Stave, J.W., Lindpaintner, K., (2013). Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure. *J. Immunol.*, **191**, 1428–1435. https://doi.org/10.4049/jimmunol.1203198.
10. Gohain, N., Tolbert, W.D., Acharya, P., Yu, L., Liu, T., Zhao, P., Orlandi, C., Visciano, M.L., et al., (2015). Cocrystal Structures of Antibody N60–i3 and Antibody JR4 in Complex with gp120 Define More Cluster A Epitopes Involved in Effective Antibody-Dependent Effector Function against HIV-1. *J. Virol.*, **89**, 8840–8854. https://doi.org/10.1128/jvi.01232-15.
11. Ibsen, K.N., Daugherty, P.S., (2017). Prediction of antibody structural epitopes via random peptide library screening and next generation sequencing. *J. Immunol. Methods*, **451**, 28–36. https://doi.org/10.1016/j.jim.2017.08.004.
12. Paull, M.L., Daugherty, P.S., (2018). Mapping serum antibody repertoires using peptide libraries. *Curr. Opin.*

Chem. Eng., **19**, 21–26. https://doi.org/10.1016/j.coche.2017.12.001.

13. Bublil, E.M., Freund, N.T., Mayrose, I., Penn, O., Roitburd-Berman, A., Rubinstein, N.D., Pupko, T., Gershoni, J.M., (2007). Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins Struct. Funct. Bioinforma*, **68**, 294–304. https://doi.org/10.1002/prot.21387.
14. Dekhtyar, M., Morin, A., Sakanyan, V., (2008). Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. *BMC Bioinf.*, **9**, 233. https://doi.org/10.1186/1471-2105-9-233.
15. Halperin, I., Wolfson, H., Nussinov, R., (2003). SiteLight: Binding-site prediction using phage display libraries. *Protein Sci.*, **12**, 1344–1359. https://doi.org/10.1110/ps.0237103.
16. Moreau, V., Granier, C., Villard, S., Laune, D., Molina, F., (2006). Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics*, **22**, 1088–1095. https://doi.org/10.1093/bioinformatics/btl012.
17. Mayrose, I., Penn, O., Erez, E., Rubinstein, N.D., Shlomi, T., Freund, N.T., Bublil, E.M., Ruppin, E., et al., (2007). Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics*, **23**, 3244–3246. https://doi.org/10.1093/bioinformatics/btm493.
18. Mayrose, I., Shlomi, T., Rubinstein, N.D., Gershoni, J.M., Ruppin, E., Sharan, R., Pupko, T., (2007). Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res.*, **35**, 69–78. https://doi.org/10.1093/nar/gkl975.
19. H. Qi, M. Ma, C. Hu, Z. Xu, F. Wu, N. Wang, D. Lai, Y. Li, et al., Antibody binding epitope Mapping (AbMap) of hundred antibodies in a single run, Mol. Cell. Proteomics. (2020) https://doi.org/10.1074/mcp.ra120.002314 (in press).
20. Enshell-Seijffers, D., Smelyanski, L., Vardinon, N., Yust, I., Gershoni, J.M., (2001). Dissection of the humoral immune response toward an immunodominant epitope of HIV: a model for the analysis of antibody diversity in HIV+ individuals. *FASEB J.*, **15**, 2112–2120. https://doi.org/10.1096/fj.00-0898com.
21. Siman-Tov, D.D., Navon-Perry, L., Haigwood, N.L., Gershoni, J.M., (2006). Differentiation of a passive vaccine and the humoral immune response toward infection: Analysis of phage displayed peptides. *Vaccine*, **24**, 607–612. https://doi.org/10.1016/j.vaccine.2005.08.039.
22. Liu, X., Hu, Q., Liu, S., Tallo, L.J., Sadzewicz, L., Schettine, C.A., Nikiforov, M., Klyushnenkova, E.N., et al., (2013). Serum Antibody Repertoire Profiling Using In Silico Antigen Screen. *PLoS ONE*, **8**, https://doi.org/10.1371/journal.pone.0067181 e67181.
23. Bachler, B.C., Humbert, M., Palikuqi, B., Siddappa, N.B., Lakhashe, S.K., Rasmussen, R.A., Ruprecht, R.M., (2013). Novel Biopanning Strategy To Identify Epitopes Associated with Vaccine Protection. *J. Virol.*, **87**, 4403–4416. https://doi.org/10.1128/jvi.02888-12.
24. Ryvkin, A., Ashkenazy, H., Smelyanski, L., Kaplan, G., Penn, O., Weiss-Ottolenghi, Y., Privman, E., Ngam, P.B., et al., (2012). Deep Panning: steps towards probing the IgOme. *PLoS ONE*, **7**, https://doi.org/10.1371/journal.pone.0041469 e41469.
25. Matochko, W.L., Chu, K., Jin, B., Lee, S.W., Whitesides, G. M., Derda, R., (2012). Deep sequencing analysis of phage libraries using Illumina platform. *Methods*, **58**, 47–55. https://doi.org/10.1016/j.ymeth.2012.07.006.
26. Matochko, W.L., Cory Li, S., Tang, S.K.Y., Derda, R., (2014). Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.*, **42**, 1784–1798. https://doi.org/10.1093/nar/gkt1104.
27. Ryvkin, A., Ashkenazy, H., Weiss-Ottolenghi, Y., Piller, C., Pupko, T., Gershoni, J.M., (2018). Phage display peptide libraries: Deviations from randomness and correctives. *Nucleic Acids Res.*, **46**, https://doi.org/10.1093/nar/gky077 e52.
28. Liu, G.W., Livesay, B.R., Kacherovsky, N.A., Cieslewicz, M., Lutz, E., Waalkes, A., Jensen, M.C., Salipante, S.J., et al., (2015). Efficient Identification of Murine M2 Macrophage Peptide Targeting Ligands by Phage Display and Next-Generation Sequencing. *Bioconjug. Chem.*, **26**, 1811–1817. https://doi.org/10.1021/acs.bioconjchem.5b00344.
29. Ernst, A., Gfeller, D., Kan, Z., Seshagiri, S., Kim, P.M., Bader, G.D., Sidhu, S.S., (2010). Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. BioSyst.*, **6**, 1782–1790. https://doi.org/10.1039/c0mb00061b.
30. Lövgren, J., Pursiheimo, J.-P., Pyykkö, M., Salmi, J., Lamminmäki, U., (2016). Next generation sequencing of all variable loops of synthetic single framework scFv—Application in anti-HDL antibody selections. *N. Biotechnol.*, **33**, 790–796. https://doi.org/10.1016/J.NBT.2016.07.009.
31. Frietze, K.M., Roden, R.B.S., Lee, J.-H., Shi, Y., Peabody, D.S., Chackerian, B., (2016). Identification of Anti-CA125 Antibody Responses in Ovarian Cancer Patients by a Novel Deep Sequence-Coupled Biopanning Platform. *Cancer Immunol. Res.*, **4**, 157–164. https://doi.org/10.1158/2326-6066.CIR-15-0165.
32. Pantazes, R.J., Reifert, J., Bozekowski, J., Ibsen, K.N., Murray, J.A., Daugherty, P.S., (2016). Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. *Sci. Rep.*, **6**, 1–11. https://doi.org/10.1038/srep30312.
33. Hurwitz, A.M., Huang, W., Kou, B., Estes, M.K., Atmar, R. L., Palzkill, T., (2017). Identification and characterization of single-chain antibodies that specifically bind GI noroviruses. *PLoS ONE*, **12** https://doi.org/10.1371/journal.pone.0170162.
34. T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, in: Proc. Second Int. Conf. Intell. Syst. Mol. Biol., AAAI Press, Menlo Park, California, 1994, pp. 28–36.
35. T. Kim, M.S. Tyndel, H. Huang, S.S. Sidhu, G.D. Bader, D. Gfeller, P.M. Kim, MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets, Nucl. Acids Res. 40 (2012) e47–e47. https://doi.org/10.1093/nar/gkr1294.
36. Burton, D.R., Pyati, J., Koduri, R., Sharp, S.J., Thornton, G. B., Parren, P.W.H.I., Sawyer, L.S.W., Hendry, R.M., et al., (1994). Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science*, **266**, 1024–1027. https://doi.org/10.1126/science.7973652.
37. Thali, M., Moore, J.P., Furman, C., Charles, M., Ho, D.D., Robinson, J., Sodroski, J., (1993). Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120-CD4 binding. *J. Virol.*, **67**, 3978–3988. https://doi.org/10.1128/jvi.67.7.3978-3988.1993.

38. Xiang, S.H., Doka, N., Choudhary, R.K., Sodroski, J., Robinson, J.E., (2002). Characterization of CD4-induced epitopes on the HIV type 1 gp120 envelope glycoprotein recognized by neutralizing human monoclonal antibodies. *AIDS Res. Hum. Retroviruses*, **18**, 1207–1217. https://doi.org/10.1089/08892220260387959.
39. Diskin, R., Marcovecchio, P.M., Bjorkman, P.J., (2010). Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. *Nat. Struct. Mol. Biol.*, **17**, 608–613. https://doi.org/10.1038/nsmb.1796.
40. H.M. Shepard, Biomarker-Driven Drug Discovery in Cancer - Trastuzumab Development: 2019 Lasker-DeBakey Clinical Medical Research Award, JAMA - J. Am. Med. Assoc. 322 (2019) 1249–1250. https://doi.org/10.1001/jama.2019.13963.
41. Zhou, T., Xu, L., Dey, B., Hessell, A.J., Van Ryk, D., Xiang, S.H., Yang, X., Zhang, M.Y., et al., (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, **445**, 732–737. https://doi.org/10.1038/nature05580.
42. Manz, R.A., Hauser, A.E., Hiepe, F., Radbruch, A., (2005). Maintenance of serum antibody levels. *Annu. Rev. Immunol.*, **23**, 367–386. https://doi.org/10.1146/annurev.immunol.23.021704.115723.
43. Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G.R., Ni, I., Mei, L., et al., (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 20216–20221. https://doi.org/10.1073/pnas.0909775106.
44. Shapiro-Shelef, M., Calame, K., (2005). Regulation of plasma-cell development. *Nat. Rev. Immunol.*, **5**, 230–242. https://doi.org/10.1038/nri1572.
45. Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., et al., (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.*, **1**, 12ra23.
46. Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., et al., (2011). High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. *PLoS ONE*, **6**, https://doi.org/10.1371/journal.pone.0022365 e22365.
47. Hershberg, U., Luning Prak, E.T., (2015). The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **370**, 20140239. https://doi.org/10.1098/rstb.2014.0239.
48. Li, W., Godzik, A., (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
49. Katoh, K., Standley, D.M., (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780. https://doi.org/10.1093/molbev/mst010.
50. Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., et al., (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134. https://doi.org/10.1038/nbt.2486.
51. Stormo, G.D., Schneider, T.D., Gold, L., Ehrenfeucht, A., (1982). Use of the “Perceptron” algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997.
52. Enshell-Seijffers, D., Smelyanski, L., Gershoni, J.M., (2001). The rational design of a “type 88” genetically stable peptide display vector in the filamentous bacteriophage fd. *Nucleic Acids Res.*, **29**, https://doi.org/10.1093/nar/29.10.e50 e50.
53. N.T. Freund, D. Enshell-Seijffers, J.M. Gershoni, Phage display selection, analysis, and prediction of B Cell epitopes, Curr. Protoc. Immunol. 86 (2009) 9.8.1-9.8.30. https://doi.org/10.1002/0471142735.im0908s86.
54. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
55. Katoh, K., Misawa, K., Kuma, K., Miyata, T., (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
56. Needleman, S.B., Wunsch, C.D., (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
57. Pietrokovski, S., (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845. https://doi.org/10.1093/nar/24.19.3836.
58. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W., (2007). Quantifying similarity between motifs. *Genome Biol.*, **8**, R24. https://doi.org/10.1186/gb-2007-8-2-r24.
59. Bossi, L., (1983). Context effects: Translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. *J. Mol. Biol.*, **164**, 73–87. https://doi.org/10.1016/0022-2836(83)90088-8.
60. Breiman, L., (2001). Random forests. *Mach. Learn.*, **45**, 5–32. https://doi.org/10.1023/A:1010933404324.
61. Chen, X., Ishwaran, H., (2012). Random forests for genomic data analysis. *Genomics*, **99**, 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003.
62. T. Hastie, R. Tibshirani, J. Friedman, Random Forests, in: Elem. Stat. Learn., Springer New York, New York, New York, USA, 2009: pp. 587–604. https://doi.org/10.1007/978-0-387-84858-7_15.
63. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., Sacha van Hijum, A.F.T., (2013). Data mining in the life science swith random forest: A walk in the park or lost in the jungle?. *Brief. Bioinform.*, **14**, 315–326. https://doi.org/10.1093/bib/bbs034.
64. Díaz-Uriarte, R., Alvarez de Andrés, S., (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinf.*, **7**, 3. https://doi.org/10.1186/1471-2105-7-3.
65. Toth, R., Schiffmann, H., Hube-Magg, C., Büscheck, F., Höflmayer, D., Weidemann, S., Lebok, P., Fraune, C., et al., (2019). Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin. Epigenetics*, **11**, 148. https://doi.org/10.1186/s13148-019-0736-8.
66. Hada-Neeman, S., Weiss-Ottolenghi, Y., Wagner, N., Avram, O., Ashkenazy, H., Maor, Y., Sklan, E., Shcherbakov, D., et al., (2020). Domain-Scan: combinatorial sero-diagnosis of infectious diseases using machine learning. *Front. Immunol.*, **11**, 3898. https://doi.org/10.3389/FIMMU.2020.619896.

67. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. http://scikit-learn.sourceforge.net.

68. V. Svetnik, A. Liaw, C. Tong, Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship, 2000. https://www.csie.ntu.edu.tw/~b88052/tmp/vietri.pdf.